ALLDATA 2023

The Ninth International Conference on Big Data, Small Data, Linked Data and Open Data

ISBN: 978-1-68558-041-4

April 24th – 28th, 2023

Venice, Italy

ALLDATA 2023 Editors

Guadalupe Ortiz Bellot, Universidad de Cádiz, Spain

ALLDATA 2023

Forward

The Ninth International Conference on Big Data, Small Data, Linked Data and Open Data (ALLDATA 2023), held between April 24th and April 28th, 2023, continued a series of events bridging the concepts and the communities devoted to each of data categories for a better understanding of data semantics and their use, by taking advantage from the development of Semantic Web, Deep Web, Internet, non-SQL and SQL structures, progresses in data processing, and the new tendency for acceptance of open environments.

The volume and the complexity of available information overwhelms human and computing resources. Several approaches, technologies and tools are dealing with different types of data when searching, mining, learning and managing existing and increasingly growing information. From understanding Small data, the academia and industry recently embraced Big data, Linked data, and Open data. Each of these concepts carries specific foundations, algorithms and techniques, and is suitable and successful for different kinds of applications. While approaching each concept from a silo point of view allows a better understanding (and potential optimization), no application or service can be developed without considering all data types mentioned above.

We take here the opportunity to warmly thank all the members of the ALLDATA 2023 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to ALLDATA 2023. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the ALLDATA 2023 organizing committee for their help in handling the logistics of this event.

We hope that ALLDATA 2023 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of Big Data, Small Data, Linked Data and Open Data.

ALLDATA 2023 Chairs

ALLDATA 2023 Steering Committee

Yoshihisa Udagawa, Tokyo University of Information Sciences, Japan
Dong Quan Ngoc Nguyen, University of Notre Dame, USA
Fernando Perales, JOT INTERNET MEDIA, Madrid, Spain

ALLDATA 2023 Publicity Chairs

Laura Garcia, Universitat Politecnica de Valencia, Spain
Javier Rocher Morant, Universitat Politecnica de Valencia, Spain

ALLDATA 2023
Committee

ALLDATA 2023 Steering Committee

Yoshihisa Udagawa, Tokyo University of Information Sciences, Japan
Dong Quan Ngoc Nguyen, University of Notre Dame, USA
Fernando Perales, JOT INTERNET MEDIA, Madrid, Spain

ALLDATA 2023 Publicity Chairs

Laura Garcia, Universitat Politecnica de Valencia, Spain
Javier Rocher Morant, Universitat Politecnica de Valencia, Spain

ALLDATA 2023 Technical Program Committee

Hugo Alatrista-Salas, Universidad del Pacífico, Peru
Farah Alshanik, Clemson University, USA
Houda Bakir, Datavora, Tunisia
Syed Raza Bashir, Toronto Metropolitan University, Canada
Gábor Bella, University of Trento, Italy
Sandjai Bhulai, Vrije Universiteit Amsterdam, Netherlands
Ayan Biswas, Los Alamos National Laboratory, USA
Jean-Yves Blaise, CNRS (French National Centre for Scientific Research) | UMR CNRS/MC 3495 MAP, France
Doulkifli Boukraa, LaRIA Lab | University of Jijel, Algeria
Didem Gurdur Broo, KTH - Royal Institute of Technology, Sweden
Ozgu Can, Ege University, Turkey
Rachid Chelouah, Ecole Internationale des Sciences du Traitement de l'Information (*EISTI*), Cergy, France
Haihua Chen, University of North Texas, USA
Esma Nur Cinicioglu, Istanbul University - School of Business, Turkey
Cinzia Daraio, Sapienza University of Rome, Italy
Subhasis Dasgupta, University of California San Diego, USA
Bidur Devkota, Asian Institute of Technology (AIT), Thailand
Chen Ding, Toronto Metropolitan University, Canada
Christian Dirschl, Wolters Kluwer Deutschland GmbH, Germany
Gokila Dorai, School of Computer & Cyber Sciences - Augusta University, USA
Ricardo Eito Brun, Universidad Carlos III de Madrid, Spain
Mounim A. El Yacoubi, Telecom SudParis / Institut Mines Telecom / Institut Polytechnique de Paris, France
Mahmoud Elbattah, Université de Picardie Jules Verne, France
Denise Beatriz Ferrari, Instituto Tecnológico de Aeronáutica, São José dos Campos - SP, Brazil
Munehiro Fukuda, University of Washington Bothell, USA
Panorea Gaitanou, University of Alcala, Spain
Chiara Gallese Nobile, Eindhoven University of Technology, Netherlands / Carlo Cattaneo University - LIUC, Italy
Fausto Pedro Garcia Marquez, University of Castilla-La Mancha, Spain

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

A Framework for Digital Data Quality Assessment in Digital Biomarker Research

Hui Zhang[1], Regan Giesting[1], Leah Miller[1], Guangchen Ruan[1], Neel Patel[1], Ju Ji[2], Tianran Zhang[1,3], Yi Lin Yang[1,4]

[1]*Digital Health Office, Eli Lilly & Company, Indianapolis, Indiana, USA*

[2]*Advanced Analytics and Data Science, Eli Lilly & Company, Indianapolis, Indiana, USA*

[3]*Department of Computer Science, Brown University, Rhode Island, USA*

[4]*Carmel High School, Carmel, Indiana, USA*

email: {zhang_hui, rgiesting, miller_leah, ruan_guangchen, patel_neel_k, ji_ju yang_yi_lin}@lilly.com, tzhang96@cs.brown.edu

all authors contributed equally

***Abstract*—Digital Health Technology (DHT) utilizes a combination of computing platforms, connectivity, software, and sensors for healthcare-related uses. Today, these technologies collect complex digital data from participants in clinical investigations, including wearable sensor signals and electronic Patient-Reported Outcomes (ePRO)s. These collected data are used to develop digital biomarkers (dBMs), which can act as health outcomes indicators for diagnosing and monitoring disease state and life quality. One essential step towards realizing the full potential of these complex digital data is to define the fundamental principles and methods to demonstrate sufficient data quality and fidelity needed for the research. This paper aims to develop a digital data quality assessment framework across the complete data life cycle in dBM research, including data quality metrics and methods to derive, visualize, and report digital data quality. Aggregating and reporting digital data quality is often challenging and error-prone. We developed a data quality assessment and reporting tool that defines data compliance criteria and views automatically generated quality reports at different levels in a consumable fashion. Combining all these methods helps to establish our digital data quality assessment framework to facilitate dBM research.**

***Keywords*—*digital health technology*; *connected clinical trial*; *sensor data*; *data quality assessment*; *data visualization*; *digital biomarker*.**

I. INTRODUCTION

Digital biomarkers (dBMs) are patient-generated physiological and behavioral measures collected through connected digital devices. The collected data are then used to explain, influence, or predict an individual's health-related outcomes (see [1]). While the development of dMBs invests heavily in advanced analytics, effective results depend on trusted and understood data collected from digital devices. An established data quality assessment framework is thus needed to define the expectation of data, monitor the data for conformance to expectations throughout the trials, and report various measures to assess the data quality (see, e.g., [2]). Establishing a meaningful data quality function will help reduce the risk throughout the dMB research activities and ultimately ensures the success criteria are met.

Today, we use DHT (see, e.g., [3]) to collect some of the most complex digital data from patients for dBM research. There has been an overall need for better data understanding and easier access to quality and trusted digital data to support operational and analytical activities in the research. Establishing a data quality assessment framework and building tools to facilitate the assessment is an emerging industry capability, and some unique challenges for this class of data quality strategy include:

- **Complexity of digital data** — We collect some of the most complex digital data in the dBM context, including sensor signals from wearables, patient-reported outcomes from hand-held devices, and labels and annotations processed and used as ground truth information for algorithm development and model building. Handling these data could be a big data problem. For example, with a sampling frequency of 50Hz, over 4 *million* 3-axial data points are collected from an accelerometer sensor for a single day to understand a patient's daily activities. Similar sensor data streams include, e.g., continuously collected photoplethysmography (PPG) and electrocardiogram (ECG) signals from trial participants.
- **Full-spectrum quality expectations** — Defining quality expectations for digital data and monitoring their conformance to expectations are full-spectrum in the data life cycle. For example, given that data can be collected in a free living environment, scanning the invalid values and noises in wearable sensor signals is often the first profiling step. Identifying the wearable sensor signal's useable (wear-compliant) portions is also a leading data quality function. The ultimate answer to the digital data quality question is the extent our digital data satisfies the specific dBM analysis requirement.
- **Aggregation and reporting** — Generating various measures to assess digital data quality is not trivial. For example, aggregating compliance information from signal level to the number of analyzable digital measures at the visit and study levels can often be tedious and error-prone. Equally challenging is to report data quality in an efficient and effective means across the data life-cycle.

Our task in this paper is to present a data quality assessment

 ISBN: 978-1-68558-041-4

framework and demonstrate a reporting platform to facilitate dBM research. The paper starts with an overview of the typical categories of digital data that our research is concerned with, then focuses on the metrics we use to profile digital data quality and later for aggregation at different levels. We also demonstrate how we put together all functions into a data quality reporting platform to support the work of all data quality assessment functions. Then, we elaborate on how the data quality reporting platform associates data stewards and quality analysts with particular study data, allowing them to run processes via interactive workflows and pull out consumable data quality reports in a central location.

This paper is organized as follows. Section II presents the related work. We present our digital data quality assessment framework and platform in Section III and Section IV, respectively. In Section V, we showcase digital endpoints and, finally, we conclude the paper in Section VI.

II. Related Work

Developing dBM requires conducting studies in a lab or free-living settings to collect raw sensor data, often with appropriate labels and annotations (*e.g.*, reported patient outcomes). Collection and analysis of wearable sensor data, together with other digital data sets, has thus become an emerging capability needed in dBM development. Industry players have begun exploring cost-effective and purpose-built solutions in the past few years. For example, the Medidata sensor cloud [4] is used to manage wearable sensor and digital health technology data for clinical trials. The Koneksa platform [5] provides support to improve compliance monitoring and patient engagement, and other representative efforts to store and deliver raw or processed data from devices in trials, including Evidation [6] and DHDP [7]. Furthermore, good data is more important than big data in dBM development. Given that data are collected in a free-living environment, noise in wearable sensor signals is inherent. To make sensor data useful, we need to monitor the quality and eventually standardize and process them to support dBM discovery, as digital data quality is of fundamental importance to developing algorithms for new dBMs (see, e.g., [8] [9] [10]). In this paper, we are mainly concerned with digital data sets that fall into four general categories:

1) *Raw Sensor Signals*. A device typically collects data from multiple sensor signals at varied pre-configured sampling frequencies to minimize study participants' burden under free living conditions. In most cases, the sensor signals are collected in a nonstop $24 * 7$ fashion throughout the entire study, which generally runs between weeks to months. Therefore, assessing potential issues, such as sensor malfunctioning, or wear non-compliance due to participants' behaviors, is critical to ensure data quality can satisfy the downstream analytics needs. Meanwhile, the quality and coverage of sensor data directly correlate to the dBM derivation, which will be discussed in the later sections of this paper.
2) *Scored Data, or Digital Biomarkers*. In addition to raw sensor signals, device companies usually have their proprietary algorithms to analyze sensor data and derive dBMs from it. For example, heart rate and blood volume pulse can be derived from the raw photoplethysmography (PPG) sensor signal. Derived dBMs are at a much lower resolution than the sensor signal, often at the minute or half-minute level.
3) *Labels/Annotations*. As algorithms and machine learning models used in developing dBMs become more complex, requirements for large annotated data sets grow. Annotating data for machine learning applications is especially challenging in the biomedical domain as it requires the domain expertise of highly trained specialists to perform the annotations. Annotations can come as interval-based events, with precise timestamps to label the onset and offsets of disease events.
4) *Clinical Records*. Apart from raw sensor data and derived dBMs, one yet important piece of data is clinical records that provision key mappings, *e.g.*, device ID to participant ID, participant ID to the treatment cohort, visit dates to treatment phases, *etc*.

Unique challenges arise from these digital data and have made a case for us to develop a data quality assessment framework to define the expectation of these digital data (e.g., completeness, uniqueness, validity, integrity), to monitor the data for conformance to expectations throughout the dBM trials, and, finally, a user interface to display the findings to support operational and analytical activities.

III. Digital Data Quality Assessment Framework

The key functions in our data quality assessment framework should now be clear in Figure 1. The logical series of modeling steps, the problems they induce, and the ultimate resolution of the problems are in the rest of this section as follows.

A. Signal Data Quality Metrics

In the pre-study phase, we establish the Data Transfer Agreement (DTA), to clearly define data quality metrics regarding signal data, including raw sensor signals and dBMs. Below we list the typical quality metrics, and Table I gives an example of the data quality metrics table we find in a DTA document, where $acce_x$, $accel_y$, $accel_z$ and ec are raw sensor signals, st, po (categorical) are derived dBMs (or, scored data) from accelerometry data, and hr and re are the scored ones from ec.

- **Sampling Frequency** — For raw sensor signals, it is the preconfigured average number of samples obtained in one second. For derived dBMs, it is the resolution of resultant features from analyzing raw sensor data.
- **Valid Range** — For numerical variables (*i.e.*, sensor signals and dBMs), a valid range is indicated by minimum and maximum values that can be measured. For enumerated variables, it is a list of predefined categorical values. One example is the rest classification biomarker

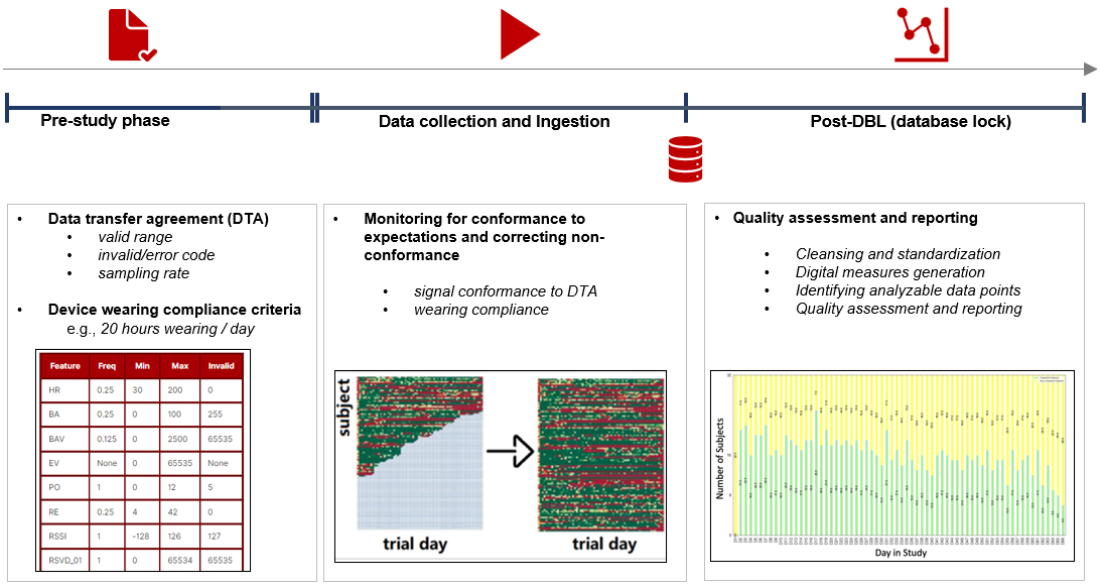


Figure 1: The overall data quality assessment scenario — from establishing DTA in the pre-study phase, to compliance monitoring in the live phase, and finally to the quality assessment and reporting in the post-Database Lock (DBL) phase.

TABLE I: EXAMPLE OF A SIGNAL DATA QUALITY METRICS TABLE FOUND IN A TYPICAL DTA DOCUMENT.

Channel	Description	Units	Min Value	Max Value	Invalid Value	Sampling Frequency (Hz)
$accel_x$	Accelerometer X Vector	gravity/1024	-32768	32767	None	50
$accel_y$	Accelerometer Y Vector	gravity/1024	-32768	32767	None	50
$accel_z$	Accelerometer Z Vector	gravity/1024	-32768	32767	None	50
ec	ECG signal	μV	-10000	10000	32767	125
st	Step count	Steps	0	65535	None	1
hr	Heart rate	beats/min	30	200	0	0.25
re	Respiration rate	beats/min	4	42	0	0.25
po	Posture • Laying Down = 0 • Standing = 2 • Walking = 3 • Running = 4 • Unknown = 5 • Leaning = 11	Enum	0	11	5	1

which has the following classes: “awake”, “sleep”, “toss and turn” and “interrupted”.

- **Invalid Value/Error Code** — In addition to the valid range, devices often provision specific invalid values or error codes to indicate different statuses of malfunctioning, which helps pinpoint the underlying issue.

B. Signal Data Quality Assessment

Connected clinical trials for dBM research often are conducted under a free living condition, *i.e.*, participants wear sensor devices on a best effort basis using instructions communicated during study enrollment. Inevitably, the free living conditions, device wearing compliance, potential device failure, or device malfunction introduce data issues such as missing data or invalid data collected when participants do not wear or incorrectly wear the devices. Figure 2 illustrates how valid signals (*i.e.*, correctly worn signals) can mix with invalid signals

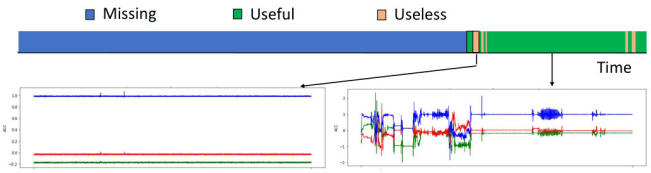


Figure 2: Illustration of sensor signal data issue. Visualized sensor data show different patterns when correctly versus incorrectly worn.

(*i.e.*, incorrectly or not worn signals) in the data collection and how they differ when plotted. Therefore, a *qualitative* means is needed to tell whether a device was operating normally and worn correctly (*i.e.*, **data usefulness**).

To fulfil this goal, the quality assessment is performed in two stages, as discussed in the following.

- **Validity Check**. Data validity check leverages signal data metrics, as discussed in Section III-A. We immediately know how many valid data points we expect to receive for a sensor signal or dBM using its pre-configured sampling frequency. We can filter out invalid values with a valid value range to get valid data coverage, *i.e.*, coverage of valid data points.
  Since raw sensor signal directly correlates with derived dBMs, we can perform a validity check against the two independently and then align their valid data coverage to check the consistency. We may further overlay device incident events to understand the root cause of observed issues better.
- **Non-wear Detection**. After dropping out invalid data through the validity checking process, the subsequent task is to detect moments when the devices were not correctly worn. The non-wear detection can be challenging as data

 ISBN: 978-1-68558-041-4

from such moments can be entirely valid in terms of falling within its valid data range. Instead of reinventing the wheel, we rely on Biobank [11] [12], an accelerometer data processing pipeline whose non-wear detection module is widely adopted as a standard. Below are two key concepts in non-wear detection.

  - *Epoch* — Although data points are collected initially at a high resolution, *e.g.*, 50Hz sampling frequency, the processing is conducted on aggregated values (*e.g.*, 1 or 5 second *short epochs* or 15 minutes *long epochs*) due to the following reasons: (1) collapsing data to epoch summary measures helps to standardize differences in sample frequency across studies; (2) there is little evidence that raw data is an accurate representation of body acceleration, and all scientific evidence so far has been based on epoch averages; (3) collapsing data to epoch summary measures also helps to average out different noise levels making results more comparable across sensor brands.
  - *Non-wear Detection* — Accelerometer non-wear time is estimated based on the standard deviation and the value range of the raw data from *each* accelerometer axis. Classification is done per 30-second epochs based on the characteristics of a larger window centered at these 30-second epochs. Specifically, Biobank identifies stationary periods in 10-second windows where all three axes have a standard deviation of less than $13.0mg$ ($1mg = 0.0098\,m \cdot s^{-2}$). These stationary periods are then used to define whether a window is stationary or not.

C. Signal Data Quality By Granularity

In addition to *qualitative* assessment as discussed in Section III-B, *quantitative* measures that define how much usable data is in a specific period (*i.e.*, **data quality** at different levels) are required before statisticians can begin analysis.

The Data Quality Model. Based on Biobank's non-wear classification on 30-second epoch level, we can further generate data quality that can be used for analysis at different time resolutions. Each phase in our data quality derivation flow is illustrated in Table II to Table V and expanded upon below. Column name "Cvge." is the abbreviation for "Coverage in Minutes".

- **Epoch Level** — This table is generated from Biobank's 30-second epoch classification. It serves as the working basis for subsequent data quality tables. Note that we have one additional column, "Subject," to indicate participant ownership of an epoch.
- **Hourly Level** — From the epoch quality table, we can apply a filter to only keep correctly worn epochs and in turn infer hourly data coverage in terms of compliant minutes. This hourly data quality table is the source for data quality reporting at the finest granularity.
- **Daily and Intraday Window Level** — From the hourly data quality table we can summarize the total coverage for each day and produce daily level data quality tables. In addition, for analysis purposes, we are often interested in specific intraday windows from which digital endpoints are derived — for instance, walking time or step count during the daytime (*i.e.*, daily **p**hysical **a**ctivity) and sleep hours during the nighttime. Thanks to the "Hour" column in the hourly quality table, intraday window coverage can be easily derived by applying filters.
- **Extended Quality with External Mappings** — We can further extend the data quality table with additional mappings when they become available as the study progresses, for instance, mapping between patients and sites/visits, as reported from the clinical operation site. These extra fields allow analysis-specific filtering and aggregation, *e.g.*, to find out which participants have sufficient data and set up individual baselines. We use this table to look for the patients with at least three valid days ($>= 20$ hours of data for a day to be qualified as a valid day) during a pre-treatment visit.

D. Representing Digital Data Quality

Fully understanding the quality of a large dataset, especially one that contains data from wearable device sensors, is not always a trivial undertaking. With numerous considerations to be cognizant of, as discussed in Section III-C, the most logical first step is to present the data with visualizations. Thoroughly understanding the data coverage and quality requires more than one visualization, simply because there is more than one aspect to check. This section presents a family of commonly used visualization examples in our data quality strategy.

- **Identifying Outliers and Missing Data.** Certain metrics must fall between threshold ranges depending on the study and associated data sources. One example is heart rate, which falls within a specified range of 30 to 200 beats/minute for one study. This range is outlined in the DTA for the study and must be applied to all heart rate data points collected. By plotting these signals against the specified thresholds, outliers can be immediately detected by viewing a plot. If outliers exist, further investigation will be completed for that participant's data to see if there are outliers for other metrics. Further, gaps in data can be identified within the same visualization, as demonstrated in Figure 3(a). Detailed data quality reports are generated in conjunction with the visualizations created for displaying outliers and missing data. For example, we convert the signal data from 3(a) to a sequence of colored blocks in Figure 3(b), with green blocks indicating valid sensor signal value in the corresponding period and red indicating missing or invalid signal value identified. In Figure 3(c), we compute the valid data ratio, and therefore can represent the data quality with a numeric value, or with a color from the color palette keyed to the valid data ratio (see e.g., Figure 3(d)).
- **Data Quality Map with Levels of Detail.** The quality of sensor signal data must be examined on various levels, each offering a specific level of detail. While certain levels are more useful for identifying distinct patterns, we will

 ISBN: 978-1-68558-041-4

TABLE II: EPOCH LEVEL QUALITY.

Subj	Timestamp	Non-wear
1002	2021-09-15 19:15:00	false
...	...	...
1005	2021-10-18 09:45:30	true

TABLE III: HOURLY LEVEL QUALITY.

Subj	Date	Hr	Cvge. $(min.)$
1002	2021-09-15	19	45
...	...	...	...
1005	2021-10-18	09	60

TABLE IV: DAILY AND INTRADAY LEVEL QUALITY.

Subj	Date	Cvge. $(min.)$	Window
1002	2021-09-15	1440	pa_daily
...	...	...	...
1005	2021-10-18	720	sleep_night

TABLE V: EXTENDED QUALITY WITH EXTERNAL MAPPINGS.

Site	Subj.	Date	Trial Day Index	Visit	Cvge. $(min.)$	Window
101	1002	2021-09-15	1	0 $(PreTreatment)$	1440	pa_daily
...	...	...	...	...	...	...
103	1005	2021-10-18	32	4	720	sleep_night

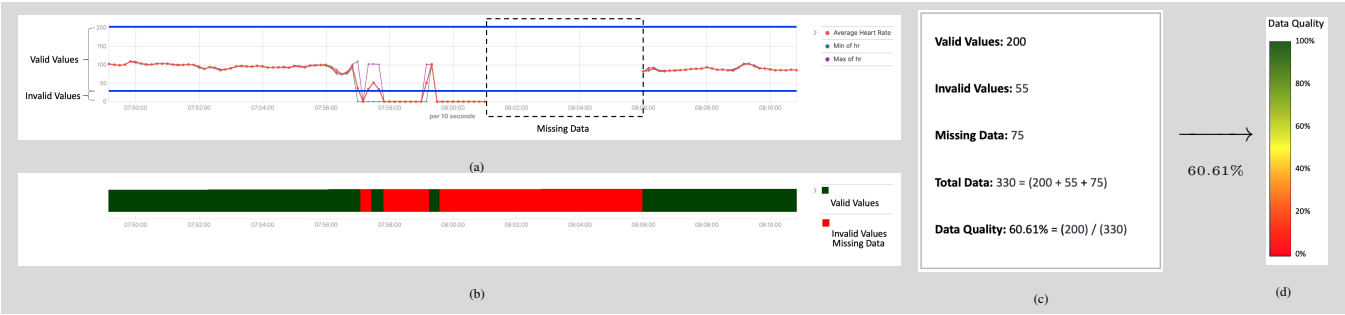


Figure 3: Visualization for sensor data quality. (a) Heart rate data (beats/minute) observed for one participant between 2021-02-15 07:49:00.000 and 2021-02-15 08:11:00.000. Valid range between 30 - 200 beats/minute, as denoted by threshold lines. Invalid data was observed multiple times. Missing data was observed between 2021-02-15 08:01:08.994 and 2021-02-15 08:06:09.000 with nearly 5 minutes of no data. (b) Use colored blocks to represent sensor signal data quality. (c) Deriving numeric representation of the data quality, *i.e.*, valid data ratio. (d) Interpreting data quality with color.

focus on the hourly, daily, and study levels on both a patient and population level:

- *Minute-by-Minute Quality Map for a Day* — Examining signals on a minute level can help to identify the minutes where a device may have intermittent connectivity, or more minor issues can be identified and further inspected, as seen in Figure 4(a).
- *Hour-by-Hour Quality Map for a Trial* — Zooming out, we can look at each hour across all days in the study. The hourly level aggregation mentioned in Section III-C is used to configure the day level plot, shown in Figure 4(b). This figure shows minutes of data coverage for each hour across all study days. This type of visualization allows us to look at compliance trends for a patient that may persist during certain hours of each day. Figure 4(b) shows an interesting device wearing pattern for the participant — taking off the wearable device to charge the battery for a couple of hours in the middle of each day of the trial has resulted in *missing data*, visualized as a sequence of red blocks in the center area of the map.
- *Day-by-Day Population-level Quality Map for a Trial* — Plotting data quality for all hours, days, and participants in a study yields the observation of data quality patterns seen in Figure 4(c). This study-level visualization can help us gain insights into the overall data quality at the population level and the compliance trends at the participant level throughout the trials.
- *Compliant Days Throughout a Trial* — In addition to the number of hours per day, it is also useful to view the number of *compliant* In addition to the number of hours per day, it is also useful to view the number of *compliant* days throughout the study, with a definition of compliance dependent on a study's protocol. One can recognize device-wearing patterns by plotting the number of patients compliant daily in a given study. As seen in Figure 4(d), the number of compliant days in a study decreased due to reduced device wearing as the study progressed.

- **Identifying and Aligning Data Issues.** In many clinical trials, it is a requirement that patients visit a site periodically. Whether it be for receiving dosing of a

Figure 4: Plots showing (a) minute-level quality representation throughout a participant day, (b) hourly-level quality representation for a participant throughout an entire trial, (c) daily-level quality for a population throughout the entire trial, (d) number of compliant days across all days in a study and (e) data coverage and device wearing issues observed throughout a study.

drug, having their vitals checked, or obtaining a device, information is collected by the sites and stored in various reports. One type of report, device reports, are used during data processing and can help understand the device's overall performance, specifically if any device issues exist. Additionally, information derived from these reports can be used to populate visualizations such as Figure 4(e). By combining this visualization with the information received in site reports, patterns specific to potential device issues and wearing patterns can be derived.

From the aforementioned data visualizations, various issues and patterns can be identified. When these are paired with actionable recommendations and delivered to the study team promptly, the study team can notify the corresponding site and participant to ensure the issue is rectified. This process leads to a quick turnaround time for potential improvements to data collection and can resolve the challenges that create low compliance in studies.

E. Generating Compliance Reports

Visualizing data is key to understanding data quality, as discussed in Section III-D. However, it is equally important to have a standardized reporting system for compliance to distribute quality and compliance information. Such systems generate reports that outline compliance on three levels: trial, site, and patient. In addition, automated generation allows systems to be configured at the start of a trial and run at set cadences to produce consistent quality assessment reports efficiently.

For each report, regardless of the level or contents, the thresholds used to configure and derive data metrics and visualizations are based on the expectations outlined in the study protocol. Each report aims to give insights into the population's compliance behavior:

- **Trial Summary**: A single comprehensive trial report can be generated and contains metadata regarding the number of patients, sites, and overall compliance percentages.
- **Study-Level Compliance**: A study-level report, such as Figure 5(a), will typically contain metrics displaying overall enrollment and compliance on a site level. These can allow a clinical trial team to gauge the progress of a specific study easily, *i.e.*, the number of patients who have completed their time in the study and the number of patients still in progress.
- **Site-Level Compliance**: Generating reports based on sites, as seen in Figure 5(b), allows clinical teams to efficiently

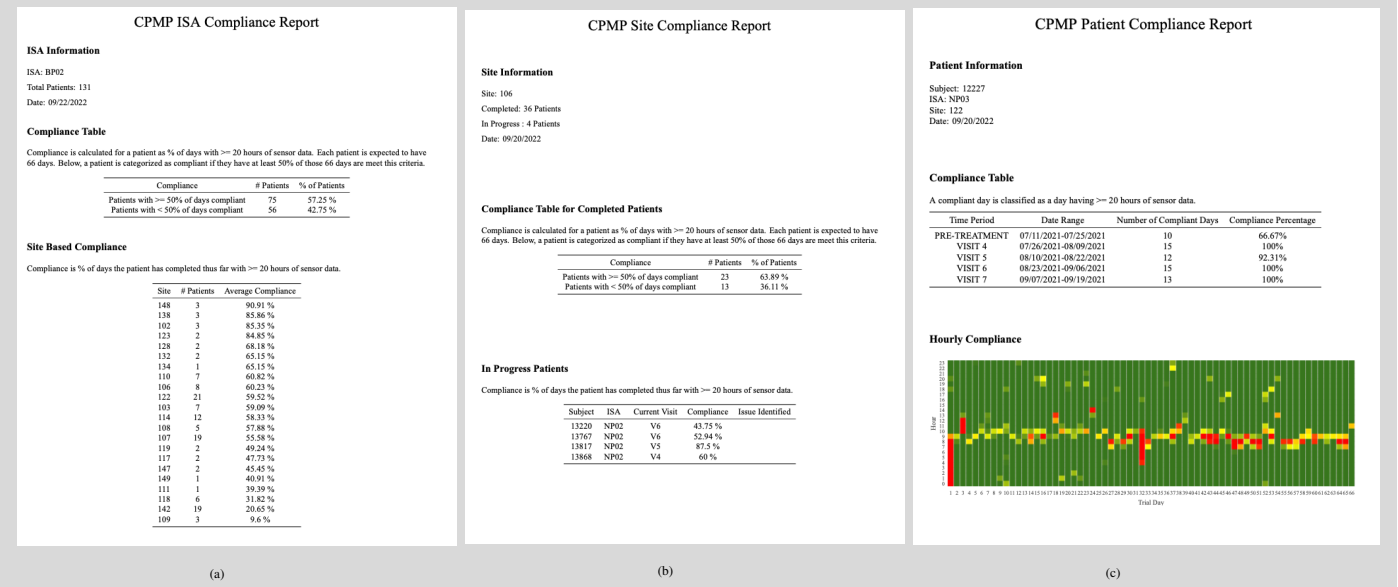


Figure 5: Putting together compliance reports for Intervention-Specific Appendices (ISAs) under Chronic Pain Master Protocol (CPMP). (a) Generated compliance reports on the patient level. (b) Compliance by visit. (c) Customizable compliance report at patient level.

identify which sites may be experiencing issues regarding low compliance across their assigned patients. Typically, site reports contain information for overall performance, with specifics for patients that may fall below a set compliance threshold. The patients with low compliance are labeled with a potential issue- such as low compliance during the nighttime. The potential issues are derived from the hourly compliance for that patient. From here, sites can identify which of their patients contribute most to low compliance and attempt to resolve the issues linked to the low compliance.

- **Patient-Level Compliance**: Reports on a patient level can give insight into their specific patterns of device wearing. In these reports, as seen in Figure 5(c), the number of visits, compliant days within each visit, and compliance percentage per visit are displayed. In addition, an hourly compliance heatmap is visible, allowing for further understanding of when patients wear their devices across the study duration.

F. Data Quality in Novel Digital Endpoint Development

For novel digital endpoint development, raw sensor signals are collected along with annotations or labels, considered the ground truth. Annotations describe events explaining the status of the patient. As such, it is critical to assess the data quality of annotations and sensor signals to identify and address as many defects as possible.

Assessing Annotation Quality. Annotations are typically collected through patient reporting via a survey system or are labeled via software by trained clinicians who observe patient behavior. We first check for defects in the annotations. Defects may include improper data structure, invalid label categories, incomplete annotations, duplicates, and impossibly overlapping annotations. Defects could be caused by bugs in the annotation software or improper training on how to label.

Assessing Annotation Quality with Sensor Signals. Evaluating annotation quality in isolation is insufficient because digital endpoint development requires both annotations and raw sensor signals. So, we must also assess the data quality of annotations and raw sensor signals in conjunction. Therefore, we plot annotated time segments along with raw sensor signals (e.g., Figure 6) to facilitate the data quality assessment.

Discrepancies in the alignment of annotations and raw sensor signals can vary considerably due to time tracking configurations and device properties in each step of the data collection process. Misalignment between annotation and raw sensor signals can be caused by improper device time configuration or the precision of the sensor device's initial time configuration. In addition, if the sensor device's time tracking is not periodically synced, the device's internal Real-time clock (RTC) will slowly drift over time. We measure drift using the sensor signal overlaid with annotation plots. Once the misalignment from the initial configuration time and RTC drift are measured, we align the raw sensor signals to the annotations.

After the annotations and sensor signals have been properly aligned, we observe the plots to identify possible defects in annotation quality. Defects could include improper labels, annotated events that are not apparent in the sensor signals, and time segments that appear to be missing annotations or sensor signals. Specific time segments of concern are selected and validated with the source to determine if further action is needed.

Lastly, depending on study-specific requirements, we may apply other methods to assess data quality. For example, output from movement detection algorithms can be compared to

 ISBN: 978-1-68558-041-4

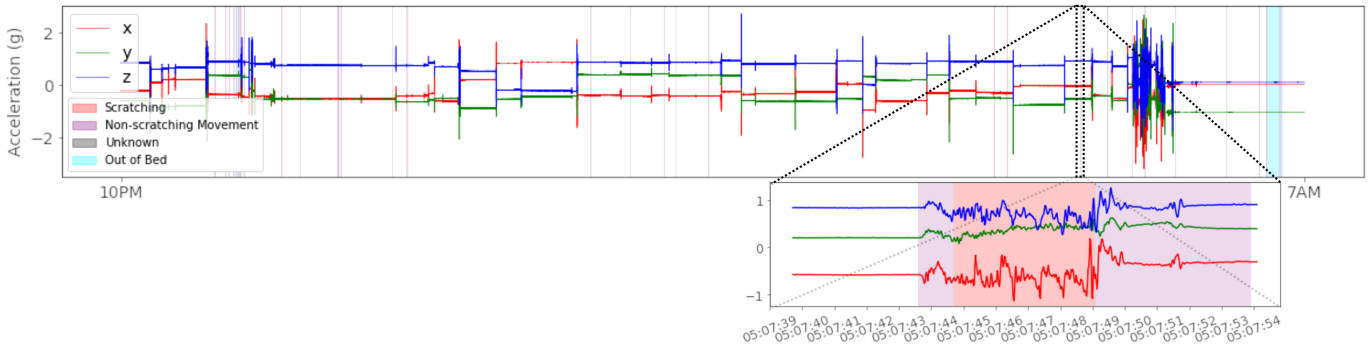


Figure 6: A plot of sensor signals overlaid with annotation labels is used to assess the data quality of annotations in conjunction with sensor signals.

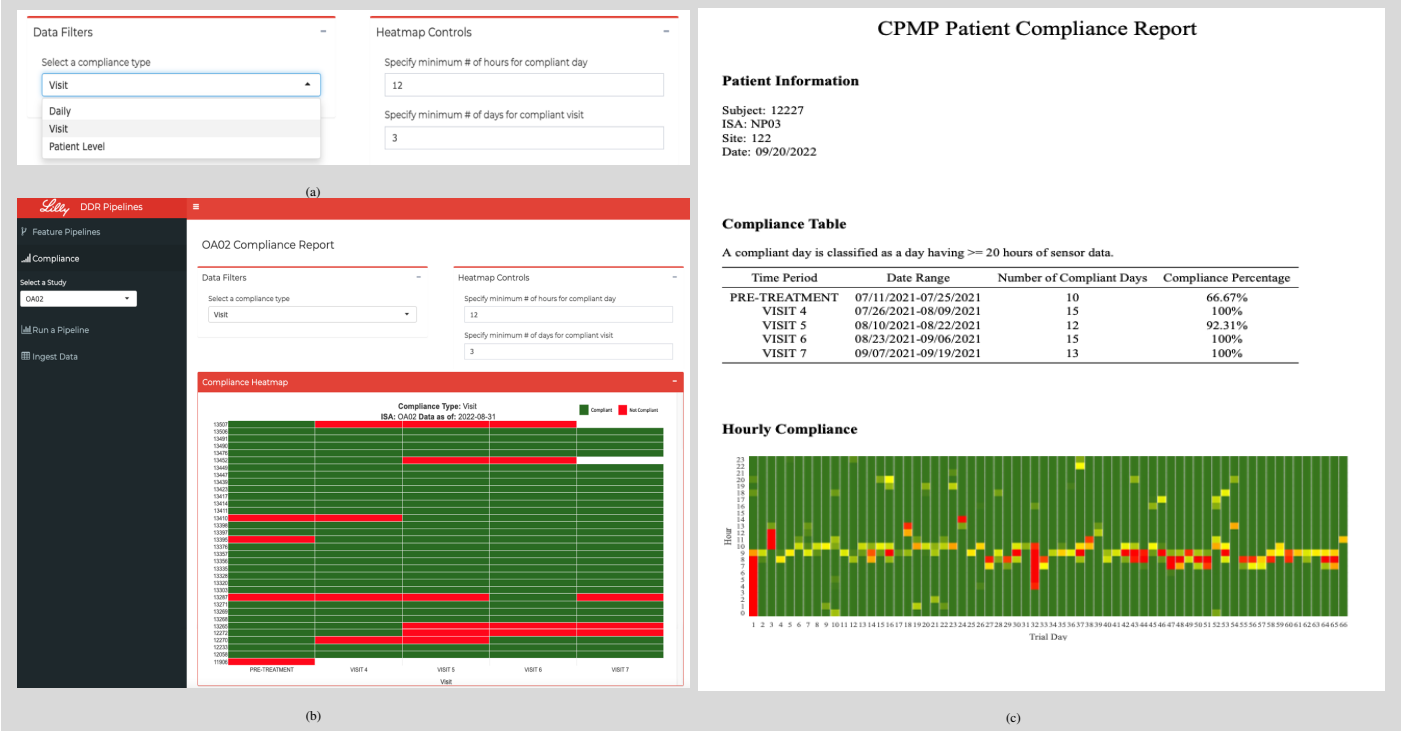


Figure 7: The platform features displaying (a) filters for customizable compliance reports, (b) compliance by visit, and (c) generated compliance reports on the patient level.

annotated time segments that describe the movement to check annotation validity and coverage. Using various methods to assess data quality from different approaches is essential to maintain the data quality needed for novel digital endpoint development.

IV. The Data Quality Assessment Platform

Throughout a clinical trial, accessing data quality metrics is critical to upholding our outlined principles. Therefore, in addition to the compliance reports generated, an interactive data quality assessment platform is used to monitor data quality throughout a trial continuously.

The platform design allows users to customize the plots and view data quality through various lenses, utilizing filters and user controls. For example, users may want to view compliance on a day, visit, or patient level. As seen in Figure 7(a), they can select the level and the metric for which the visualization will show, as discussed below.

Let us take configuring and viewing compliance visualizations as an example. A user wants to view compliance for all patients in a study on the visit level, as seen in Figure 7(b). They define *compliance* as having at least 12 hours of data daily, with 3 days each visit comprising a compliant visit. By selecting the compliance type, which in this case is visit, and inputting the number of hours and days for defining compliance, the user can see the population's compliance for these specific thresholds, as seen in Figure 7(a). Additionally, they can easily compare and contrast different levels and compliance thresholds

 ISBN: 978-1-68558-041-4

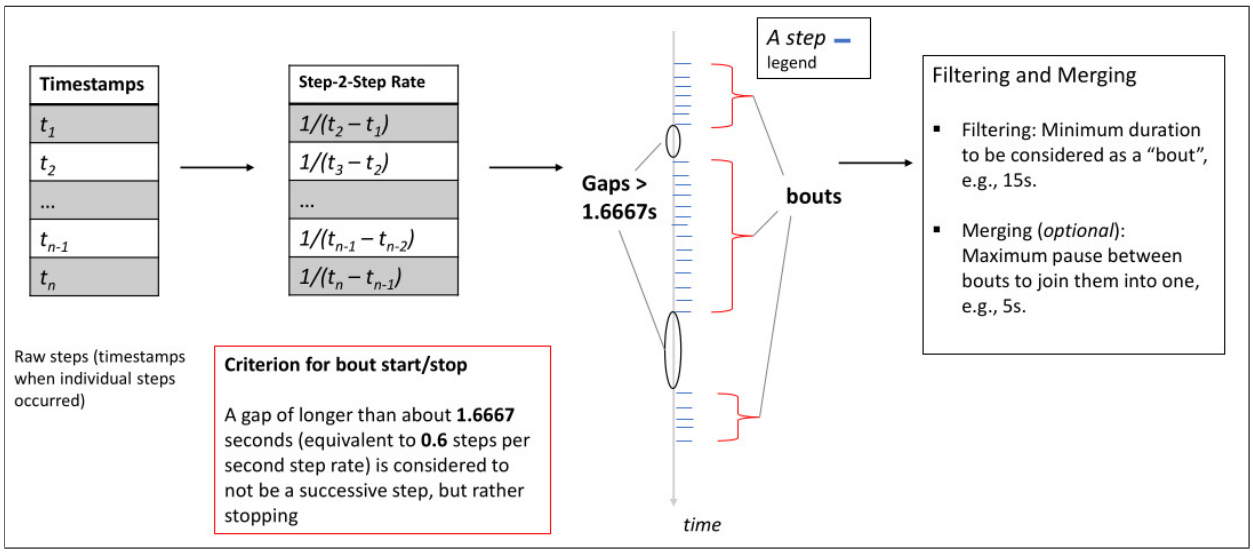


Figure 8: Process of transforming raw steps into bouts.

within the data quality assessment platform.

In addition to the compliance assessment, data quality visualizations, such as Figure 4, are created and customized within the platform. For example, as seen in Figure 7, a user can select a specific time range or time level to view the data. This zoom in and out can be used to identify and trace patterns of device wearing.

The data quality assessment platform allows for customizable, real-time, informative visualizations that enable insights into patient compliance and device-wearing data patterns. The study team can process and act upon these key insights with these visualizations housed in a centralized, consistent, and efficient platform.

V. Digital Endpoints

With out data quality assessment platform, we are able to derive digital endpoints from two categories: **P**hysical **A**ctivity (PA) [13] [14] and sleep [15] [16]. Typical PA features include duration of daily light/moderate/vigorous activities, steps count and gait features. For sleep features they are night sleep duration and **W**akeup **A**fter **S**leep **O**nset (WASO).

Gait features are a unique set of physical activity endpoints that unveil fine-grained walking characteristics, for which we see a significant distinction between health and chronic pain cohorts. Due to their importance, we detail our effort in deriving gait features in this section.

Determining bouts is the most fundamental step since all gait features are based upon bouts. Figure 8 illustrates this process: (1) raw individual steps with their timestamps are obtained from an open source step detection algorithm; (2) derive step rate for every two consecutive steps; (3) since bout by definition is a short period of intense walking activity with less than 1.6 seconds of stop between two steps, we can apply this gap threshold to detect individual gaps; and (4) depending on specific settings of a study (*e.g.*, profile of participating cohorts), we apply a constraint on minimum bout duration (*e.g.* filtering to keep $>= 15s$ bouts) and optionally merge bouts with small gaps in between into a single bout.

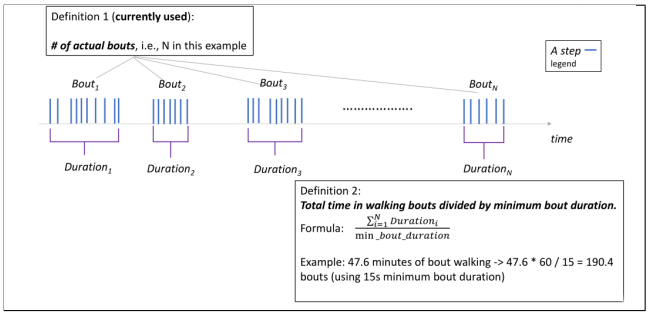


Figure 9: Bout count.

Once bouts are identified, we can derive bout and gait-related features. Below we summarize the derivation process.

- **Bout Count**. We currently use the definition of actual number of identified bouts. Another meaningful definition is the count in terms of minimum duration bout , *i.e.*, $\sum_{i=1}^{N} \frac{Duration_i}{15} = 1$, where 15s is used as minimum bout duration. Figure 9 illustrates the two definitions.
- **Bout Duration**. Bout duration is the average duration across all bouts, *i.e.*, $\frac{\sum_{i=1}^{N} Duration_i}{N}$.
- **Steps per Bout**. Steps per bout is average of the count of steps across all bouts, *i.e.*, $\frac{\sum_{i=1}^{N} StepCount_i}{N}$.
- **Cadence**. A single $bout_i$'s cadence is the number of steps per its duration, *i.e.*, $\frac{StepCount_i}{Duration_i}$, we can then use the averaged cadence across all bouts for the cadence feature, *i.e.*, $\frac{\sum_{i=1}^{N} Cadence_i}{N}$, as shown in Figure 10.
- **Gait Rate**. For a single $bout_i$ with $M+1$ steps, its mean step rate is defined as $\frac{\sum_{i=1}^{M} stepRate_i}{M}$, where $stepRate_i = \frac{1}{t_{i+1} - t_i}$ is the step rate between $step_{i+1}$ and $step_i$, whose occurring timestamps are t_{i+1} and t_i respectively. The gait rate feature is then derived as the average of the

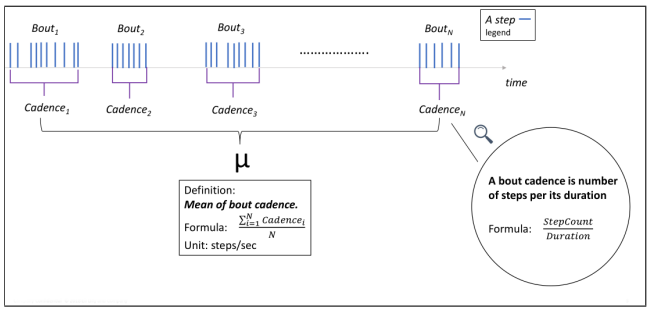


Figure 10: Cadence.

mean step rate across all bouts, *i.e.*, $\frac{\sum_{i=1}^{N} MeanStepRate_i}{N}$. Figure 11 illustrates this process.

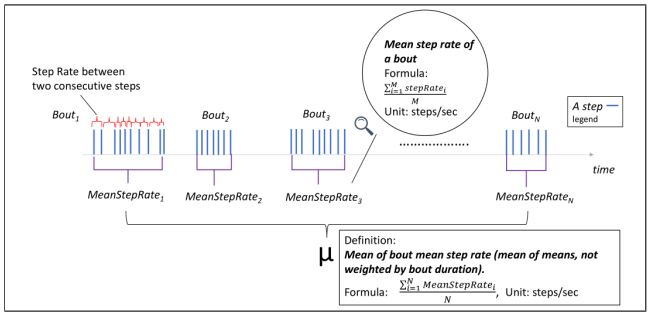


Figure 11: Gait rate.

- **Gait Rate Standard Deviation**. Similar to mean step rate, for a single $bout_i$ with $M+1$ steps, we can calculate standard deviate over the M steps rates, *i.e.*, $\sigma(StepRate_i)$, $i = 1 \cdots M$. The feature is then derived as the mean of standard deviation in step rate from each bout, *i.e.*, $\frac{\sum_{i=1}^{N} StepRateStd_i}{N}$.
- **Step Rate Change**. As shown in Figure 12, a bout's step-to-step rate change is the difference of step rate from the first set of steps (*i.e.*, steps 6 to 8) to steps 23 to 25 on any period of walking with at least 25 steps long. Therefore for $bout_i$ with 25 or more steps, its step to step rate change can be calculated as $\mu(\sum_{i=23}^{25} StepRate_i) - \mu(\sum_{i=6}^{8} StepRate_i)$. In turn, the feature is the mean of step rate change from each eligible bout (≥ 25 steps), *i.e.*, $\frac{\sum_{i=1}^{N} StepRateChange_i}{N}$.

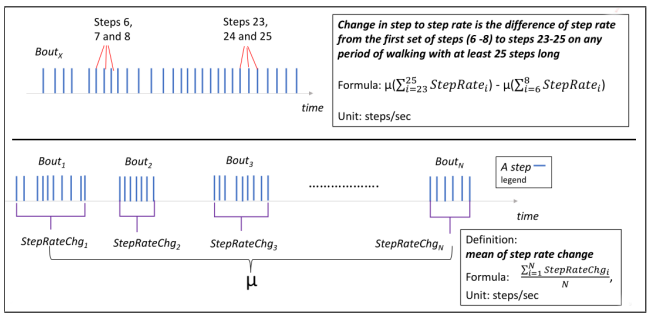


Figure 12: Step rate change.

VI. Conclusion and Future Work

As DHT continues to evolve and collect more complex digital data in clinical trials, the need for a digital data quality assessment platform is increasing. By defining and implementing the fundamentals of data quality into the digital data quality framework and platform, we can generate automated compliance reports, customizable visualizations, and real-time quality metrics. In addition, the methods for facilitating dBM research have been simplified with the centralized digital data quality assessment platform. As dBM research continues, so will the use of the digital data quality assessment platform. Future directions include the use of visual mining and data mining technologies to help identify data quality in a novel way to facilitate data quality assessment.

References

[1] J. M. Wright *et al.*, "Evolution of the digital biomarker ecosystem," *Digital Medicine*, vol. 3, no. 4, pp. 154–163, 2017.
[2] R. Y. Wang, V. C. Storey, and C. P. Firth, "A framework for analysis of data quality research," *IEEE transactions on knowledge and data engineering*, vol. 7, no. 4, pp. 623–640, 1995.
[3] A. Sharma *et al.*, "Using digital health technology to better generate evidence and deliver evidence-based care," *Journal of the American College of Cardiology*, vol. 71, no. 23, pp. 2680–2690, 2018.
[4] R. Lyons, G. R. Low, C. B. Congdon, M. Ceruolo, M. Ballesteros, S. Cambria, and P. DePetrillo, "Towards an extensible ontology for streaming sensor data for clinical trials," in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2021, pp. 1–6.
[5] C. M. Rey, "Wearable data revolution: Digital biomarkers are transforming research, promising a revolution in healthcare," *Clinical OMICs*, vol. 6, no. 2, pp. 10–13, 2019.
[6] I. Clay, "The future of digital health," *Digital Biomarkers*, vol. 4, no. 1, pp. 1–2, 2020.
[7] M. Chen and M. Decary, "Artificial intelligence in healthcare: An essential guide for health leaders," in *Healthcare management forum*, vol. 33, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2020, pp. 10–18.
[8] S. M. Hossain *et al.*, "Mcerebrum: A mobile sensing software platform for development and validation of digital biomarkers and interventions," ser. SenSys '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1–14.
[9] A. Dillenseger *et al.*, "Digital biomarkers in multiple sclerosis," *Brain Sciences*, vol. 11, no. 11, pp. 1519–1544, 2021.
[10] M. M. Rahman *et al.*, "Towards reliable data collection and annotation to extract pulmonary digital biomarkers using mobile sensors," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2019, pp. 179–188.
[11] A. Doherty *et al.*, "Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study," *PloS one*, vol. 12, no. 2, p. e0169649, 2017.
[12] C. Sudlow *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, no. 3, p. e1001779, 2015.
[13] V. T. Van Hees *et al.*, "Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity," *PloS one*, vol. 8, no. 4, p. e61691, 2013.
[14] S. Sabia *et al.*, "Association between questionnaire-and accelerometer-assessed physical activity: the role of sociodemographic factors," *American journal of epidemiology*, vol. 179, no. 6, pp. 781–790, 2014.
[15] V. T. van Hees *et al.*, "Estimating sleep parameters using an accelerometer without sleep diary," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.
[16] A. Doherty *et al.*, "Gwas identifies 14 loci for device-measured physical activity and sleep duration," *Nature communications*, vol. 9, no. 1, pp. 1–8, 2018.

 ISBN: 978-1-68558-041-4

An Industrial Manufacturing Dataset together with Anomaly Detection Results integrated in an Open & Stand Alone Sharing Platform for Sustainable Replication

Gerold Hoelzl, Jonas Zausinger, Matthias Kranz
Chair of Embedded Systems
University of Passau
Passau, Germany
email: first.last@uni-passau.de

Bastian Fleischmann, Sebastian Soller
Almanara Research GmbH
Ruhstorf, Germany
email: first.last@almanara-research.de

***Abstract*—We aim at systems that make sense out of occurring anomalies to autonomously learn to predict and detect possible occurring machine drifts, failures and deviations, and the corresponding errors in the machines and products itself. To assess our prediction and classification methods, we collected data from a fully automated industrial machinery including 3 internal sensors in a large-scale dataset (> 87000 manufactured pieces with 39 different product types, in a timespan of nearly 7 months). We present the scenario and describe the collected data and the sensors. We describe the machine data and the corresponding errors, and present a generic tool that allows visualization, scripting, etc., especially when datasets have to be shared, as it gives an insight into the complexity of the data and the algorithms and make experiments as described in the paper reproducible. We argue to be currently in a replication crisis in data analysis that makes it close to impossible to replicate empirical findings due to the lack of the availability of the underlaying data and the implemented algorithms. We reached a point where we need to question if the results can be believed and how the datasets for evaluation are designed and recorded. To support an inevitable fundamental change towards the full openness of published results in collected data and the used algorithmic processing with minimum effort, we present and make publicly available (i) a large-scale dataset for IoT (Internet of Things) based predictive maintenance in an industrial setting combined with (ii) artificial intelligence algorithms used by our group, elaborated on the dataset embedded in (iii) a general tool to foster easily sharing of both for replicating results.**

***Keywords*—sensor based manufacturing dataset; industry; machine learning; anomaly detection; defect detection; industry 4.0; data sharing; toolset for result replication.**

I. Introduction

Industry 4.0 has become an important topic for researchers in the industrial domain. With the availability of sensors, controllers and communication networks, a vast amount of data can be collected to improve aspects of the industrial production process [1]. Depending on the data, it can be utilized for various application scenarios adapted from techniques used in e.g. in human activity recognition architectures [2][3][4]. Important applications are transparency of the production process, a highly customizable and dynamic production process and smart manufacturing using machine learning [1][5].

One aspect of smart manufacturing is predictive maintenance and machine fault detection [6][7]. Machine learning in combination with sensor data collected beforehand is used to predict the health of the machinery or detect deviations from the normal state. When detecting a deviation from the normal state a technician can be notified to take suitable action. To this end, potential damages can be reduced by suggesting maintenance beforehand or detecting defects when they occur. For this, anomaly detection is successfully applied by researchers in the industrial domain [8][9]. An example for the application of fault detection is the early detection of machine defects by observing the vibration of machine parts using specifically placed vibration sensors [10][11].

In the industrial setting, anomaly detection is often applied in areas where the machine executes similar steps for prolonged periods of time. Deviations from the normal operation are expected to be induced machine issues. However, another important goal of the advancements in industrial production is a highly dynamic production that adjusts itself at any given time. Different products are produced interchangeably as the machinery adapts the operation mode according to the desired final product. Therefore, the operation mode and the notion of normal behavior can also rapidly change. This poses new challenges for fault detection. Changing a product type can be falsely identified as a defect. Likewise, types produced in small volume can be identified as anomalous, as they are insufficiently represented in the training data. Additionally, comparing results for such an industrial process is challenging due to the high variability of the process. Often each research group collects their own data - some of which may not be publicly available - using a custom set of specifically placed sensors to perform their experiments. This makes replicating and comparing results, and as a consequence, improving the methods more challenging.

To this end, we present a dataset gathered from a highly dynamic real-world industrial process and intended to be used for fault detection, using already available internal sensors that are part of the machine by default to increase the technical applicability, in this paper. These sensors collect internal information on the movement and electrical current of machine parts. We provide intrinsic sensor measurements of a CNC (Computerized Numerical Control) machine that is part of a larger production line. There, the produced product type, and configuration of the machine changes on the fly. The dataset spans over a time period of nearly 7 months and contains the production of 87650 workpieces from 39 different types.

 ISBN: 978-1-68558-041-4

These product types share similar basic traits, but can differ in characteristics such as size, design, and the presence of certain traits. In addition to the sensor data, we also collect the occurrence of machine events that are labeled by workers as a ground truth. Together with the dataset we introduce a tool to work with the data. This tool aims to facilitate the usage of the data by providing a simple playground for experimenting.

We show first results from using product type-aware anomaly detection to detect machine faults by performing anomaly detection both globally and in the context of the product type using well-known anomaly detection techniques. These results serve as a baseline for future work to make machine fault detection in a highly dynamic environment more robust.

The remaining paper is structured as follows. Section II describes the collected Dataset used for the Evaluation of the Anomaly Detection Algorithms and our developed Sharing Platform. In Section III we present the Evaluation of the Anomaly Detection Algorithms in our Application Case. Challenges and Future Developments are highlighted in Section IV. The paper is closed with Section V, that summarizes and recaps the achievements and contributions. Section VI links to the online sources where the collected Dataset and the developed Tools are available for download.

II. Dataset

A. Data

We obtain our dataset from an ongoing real-world industrial CNC production line. This production line operates fully automated and produces a variety of different products depending on customer demand and ad hoc supply. Further, the production is performed in a mixed fashion. Products within a configured set of possible product types are produced in a nearly arbitrary order and quantity. Fig. 1 shows the total number of products within the configured set of possible product types for the day, while the products in the configured set are produced in arbitrary order. The product types can differ on properties such as size, design, and weight. Therefore, the processing of the workpiece is adjusted depending on the desired result.

For this dataset we use various internal sensors to observe a single CNC machine that is part of this production line. These sensors are part of the machine's standard equipment. By using the internal sensors, the transferability of results to similar production lines is increased due to reduced requirements for the sensor setup. An overview of the measured machine properties is shown in Table I. We observe the speed and electric current of the milling spindle and the electric current of the servomotor. The electric current of the servomotor also has multiple channels for the current in each direction. We collect the data for this dataset over a period of nearly 7 months, spanning from November 2020 to May 2021. In that time frame, a total of 87650 workpieces from 39 different product types are observed during production.

The workpieces are processed in a sequential order and the production can differ depending on the product type.

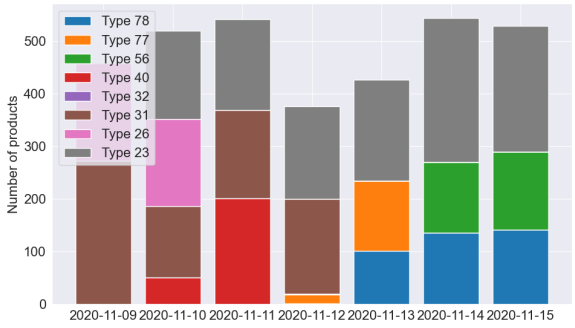


Fig. 1. Number of manufactured products per product type during a seven-day period. Value denotes the total number of products for this day, as configured product types are produced interchangeably in a nearly arbitrary order.

TABLE I: Overview of the observed features in the dataset

Feature	Channels	Samples	Time Unit	Sampling Rate	HDF5 File
Spindle Speed	1	87650	ms	7,8125Hz	spindle.h5
Spindle Current	1	87650	ms	7,8125Hz	spindlemeter.h5
Electric Motor Current	2	87650	ms	7,8125Hz	servometer.h5

Therefore, the measurements are segmented into time series for each individual product. A measurement starts when a new unprocessed workpiece enters the CNC machine. Once the product is finished and exits the machine, the measurement is stopped. In between this period, we collect data of the aforementioned properties with each sensor aiming to measure their respective property every 128 milliseconds. On average we measure around 1100 time steps per sensor channel during the production of a single workpiece. An example for the resulting time series is shown in Fig. 2. There, the rough shape, and value range of the time series for a single product during normal production is illustrated.

The results of the measurements are written into CSV (Comma-Separated Values) files, as shown in Fig. 3. Each property is in a separate file as the sensors collect the data independent of each other. The rows of the CSV files are structured as follows:

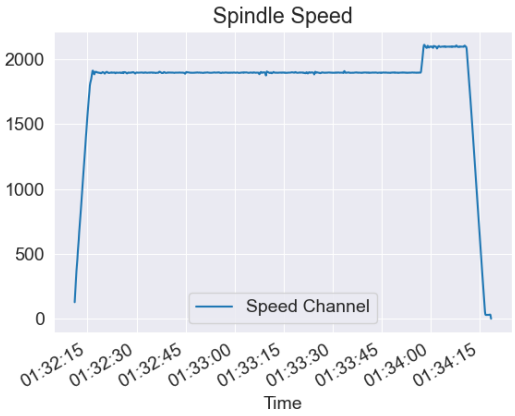
<timestamp>,<channel 1>,<channel 2>,...,<channel n>

The first column is the time at which the measurement point was collected. The following columns are the values of the respective channels at that time.

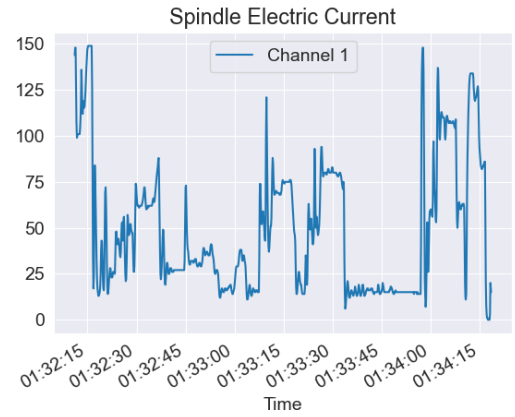
Each product type has an individual program that defines how the production process is performed. Therefore, the measured values can deviate, when comparing the processing of different product types. Distinct types can differ in properties like duration and shape of the time series. So, the data collection system also queries what product is produced by the machine. With this we can attribute each segment to a distinct product type.

To make our data accessible and to work with it we convert each individual CSV file to a *pandas* Data Frame. The structure of a Data Frame corresponds to the structure of the respective CSV file. The columns of the Data Frame

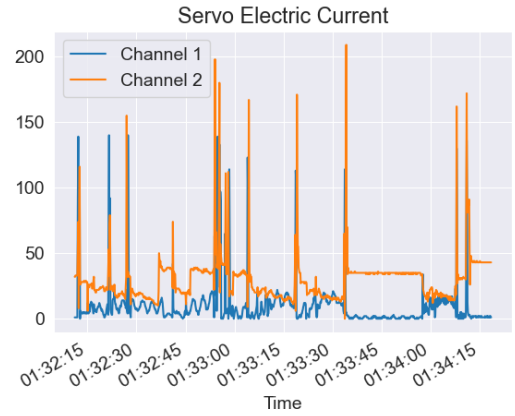
 ISBN: 978-1-68558-041-4



(a) Speed of the milling spindle



(b) Electric current of the milling spindle



(c) Electric current of the servomotor

Fig. 2. Measured values of the three observed properties during the processing of a single workpiece.

are the channels of the property, while the index is the time of measurement. Since we have many CSV files, we collect the Data Frames in HDF5 (Hierarchical Data Format version 5) files. Each property is stored in an individual HDF5 file. The names of the files for the respective properties are shown in Table I. The hierarchical structure of the HDF5 files

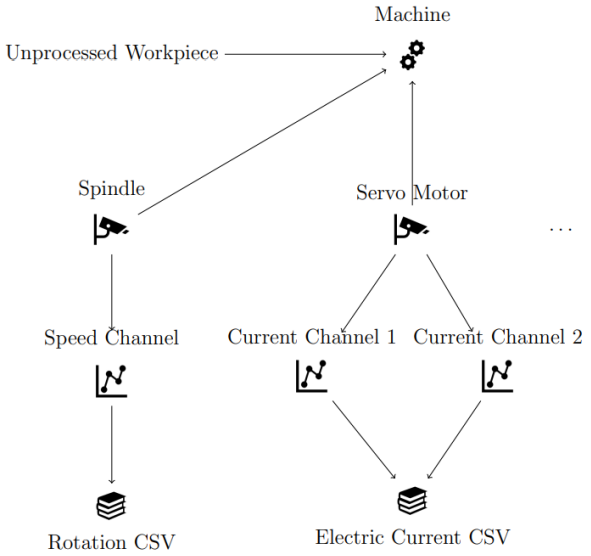


Fig. 3. Sensor collection setup.

themselves looks as follows:

/<property>/t<product type>/<measurement>

The ID of the product type, that corresponds to the measurement, is encoded within the hierarchy as additional information.

B. Ground Truth

The ground truth data consists of machine defects and production issues during the time we collect the measurements. It is in the *Events.xlsx* file and contains labels for events when the machine is down or transitions into a state that requires human intervention or repair. Machine downtimes are only included when the machine is turned on and has enough available material. In our dataset, we have 1033 instances of such events.

The entries in the ground truth have three timestamps. The first timestamp - *date* - stands for the time the event is detected by a worker or a monitoring system. The monitoring system detects events when the production time of a workpiece surpasses a certain threshold. The other two timestamps denote the time period of the event, with the *start* and *end* timestamps. These are inserted by the workers once they perform a checkup or fix the machine. The events are non-overlapping and only one event should occur at a time. During this period either no products exit the production line, or production runs at a limited capacity.

Each entry has a label for the type of event that occurs. The labels are inserted by the workers after they resolve the issue. We have five broader groups of events denoted by numbers: Critical-(1), Major-(2), Minor-(3), Organizational-(4) and Unknown-(5).

Critical events have a severe impact on the production and the machine. They usually require the replacement of machine parts and not responding timely can cause even

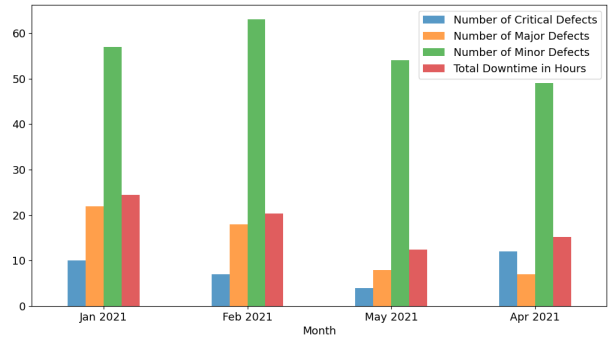


Fig. 4. Number of defects for the three most severe events and total downtime due to these three types of events in the first four months of 2021.

more damage. Critical events are usually when an important machine part breaks. Major events are less severe, but still have a high impact on the operationality of the machine and might require intervention of trained personnel. They are usually problems with the internals of the machine. Minor issues mostly interrupt the production but can be fixed with little effort and technical expertise. Common cases for minor issues are jammed workpieces or loose parts, defects on minor parts of the machine and incorrect operation of the machine. Organizational events are intentional disruptions of the production, such as performing changes on the machine. Unknown events are issues with an unknown source that could not be linked to a certain machine in the production line.

Fig. 4 shows the occurrence of critical, major and minor events in the first four months of 2021, along with the total downtime in hours due to these events. Minor events are the most frequent types of events, while critical and major events occur more rarely. In most cases, critical defects also occur less often than major defects. In total one day to half a day of production time is lost due to these types of defects every month.

The groups have a set of issues that are assigned to them. Each kind of issue is denoted by a categorical number that uniquely identifies them. Following is a list of the groups and the issues that are assigned to them: *Critical*: 1, 2; *Major*: 3, 4; *Minor*: 5, 6, 7, 8, 9, 10; *Organizational*: 11, 12; *Unknown*: 13. The ground truth contains all detected issues of these types that occur during the time of our measurements. It includes issues that originate from normal machine operation as well as defects that originate from external factors, such as human interference.

C. Data Access and Distribution

We developed a browser tool to facilitate access to and experimentation with the data by allowing users to visualize datasets and perform and reproduce data processing steps. The purpose of this tool is to reduce the burden of entry of working with this data by providing the ability to quickly run small tests, view the code of the algorithms along with the visualizations of the data, and thus facilitate the step to own experiments/applications with the data. The goal of this tool is to provide both the data and the code for the experiments in the same environment.

The tool can be populated with custom algorithms in Python code and custom data. It allows experiments to be executed on the data and then displays the visualization of the data and results.

The experiments are comprised of data processing steps combined into a pipeline. For each pipeline step, the users can define how data and intermediate results are to be displayed and visualized by customizing the executed code. For this purpose, the users are provided with a web interface in which they can add and edit scripts for each pipeline step containing the algorithms and the definitions of the visualizations. After the execution of the pipeline, the results and defined visualizations are displayed on the web interface.

The web interface consists of a starting page for an initial overview of the dataset, separate tabs to view and edit each individual data processing steps and the functionality to execute the data processing pipeline. Furthermore, the resulting visualizations and output of the data processing steps are displayed in the tab of the respective data processing step once the pipeline is executed along with the respective code. In addition, users can create additional data processing steps, edit and delete the existing ones and define how final and intermediate results are to be displayed for each step. Therefore, the tool provides a plug and play playground to adjust the data processing for further work. Further, the code for the data processing steps can be extracted to a different environment once a playground is no longer required, as it is contained as python scripts within the too data.

The created visualizations together with data and data processing steps can be passed on to other users so that they can quickly execute the already preset application and thus immediately execute the Algorithms and view the results.

In Fig. 5, the usage flow and the concept behind the application is shown. Researchers from the publishing side (Group 1) can make their data and scripts available in a way the data can be easily accessed, viewed, and executed experiments can be replicated in a local environment only requiring python and relevant libraries for the data processing. This allows other researchers (Group 2) to work with the data, inspect results and perform further work on the data that can yet again be made available.

We bundle this tool together with our raw dataset to make the data more accessible, enable replication and facilitate future work.

D. Data Quality

With the design of the sensor collection and label collection setup we aim to ensure the quality of the provided data on a level that correctly reflects the industrial process, but also inherits challenges of the real-world processes. This might even include unknown errors upcoming algorithmic solutions have to cope with.

 ISBN: 978-1-68558-041-4

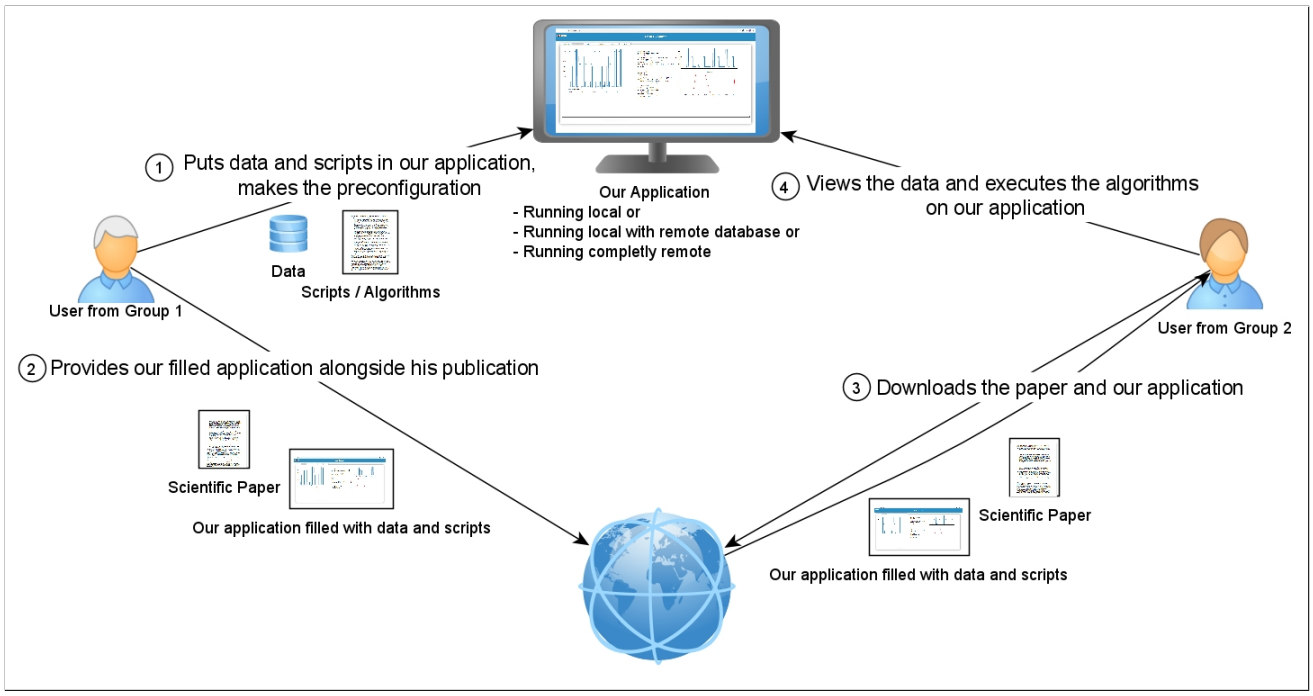


Fig. 5. Schematic of the application workflow and deployment.

To collect the dataset, we made sure that workers are trained and experienced in labelling events that happen during production. The manual labelling of the data by workers during the production is an already long-time established process and the workers are used to it. Therefore, potential mistakes are reduced as workers are already accustomed to the system. We validated the semantic correctness of the sensor and label data by performing multiple experiments [12][13][14][15] on the data to predict upcoming errors in the production line.

The challenging and unique part of the dataset on the other hand is, that we provide a dataset that faithfully represents the available data in an ongoing industrial production under real-world circumstances. This can possibly conflict with the goal of providing a ”clean” dataset where any unexplainable data is filtered out but offers the advantage of still containing each little and possible latent piece of data. We accept minor limitations regarding the quality (i.e. sensor failures) of the data that remain in the dataset on purpose. We see this as a tremendous advantage compared to clean, sometimes even artificial datasets, as our approach leads to more robust algorithms that must deal with imperfect circumstances in a non-perfect world. In addition to sensor failures, sensor noise and further inaccuracies can be present. It can occur that the completion of a product is recognized too early or too late. In the latter case, measurements of other products are attributed to a single product. As the properties are measured independently of each other, the time sensors values are sampled can deviate between different properties. While we aim to collect measurements every 128 millisecond, the actual period between sensor measurements is variable. Reasons are latency, limited processing power and throughput of the industrial network.

The ground truth consists of events that could be detected as they had an impact on the production. Therefore, labels exist for the most important events that occur, but labels are not all encompassing. As the labels are generated by observing the production instead of generating them from the sensor signal, anomalies found in the sensor signals that have no perceivable long-term impact on the production remain unlabeled. Labels for anomalies like a temporary drop-off in the production speed are unavailable. These events would only be labelled indirectly if they result in a noticeable production issue later. As it is a non-isolated running system there are also several external factors that can induce anomalies unrelated to defects. For example, the machine can be stopped or slowed down to perform trials and visual checkups. Such events are also included in the ground truth and might not have early indicators. Therefore, events prior indicators or anomalies can exist in the ground truth. As previously noted, there also exist instances of labels with an unknown source when the workers were unable to identify the defect or were absent during the occurrence of the defect. This also means it is unknown if the labels are relevant to defects of the machine.

III. Application Case - Evaluation

A major objective to achieve with this data is the detection of machine faults and defects during production. Observing the ongoing stream of sensor data, a machine problem should be reliably detected either when it occurs or in advance to take possible countermeasures and reduce damages. On the other hand, false positives - due to the dynamic production - should be minimized as appropriate responses require capacities from trained technicians. Therefore - following our previous work

 ISBN: 978-1-68558-041-4

on similar machines [12][13][16] - we execute a baseline experiment for machine fault detection with this dataset using anomaly detection methods.

For this baseline approach we perform anomaly detection on the level of products. As samples we use all sensor data obtained during the processing of an individual item. Each sample consists of multiple time series that form the feature vector. The number of time steps in these time series can be very high and variable. Therefore, we first reduce the size of the time series to a fixed length. Piecewise Aggregate Approximation [17] is applied to transform each time series into a time series with 100 time steps. We then combine all time series to a single feature vector that is used as representation for anomaly detection. From the set of events, we aim to detect events from the critical and major groups. These issues have the biggest impact on the production and the machine. Therefore, detecting these issues has the highest priority for us. Other less critical events are ignored for this experiment.

As the objective of the fault detection is to detect problems in the ongoing production, we setup the anomaly detection to reflect an on-line application. We use Holdout Cross Validation to tune the algorithms. Therefore, the training, validation and test sets only contain data that is measured in the respective time frames. During the test stage, we also use both the training and validation set to train the anomaly detection models.

Before training the anomaly detection models, we first clean the training set by removing all samples that were measured in a time frame of 4 hours before a critical or major event. Then we use the cleaned training set to train a model for the global production context. This model should detect deviations from previously seen production across all different product types. To also take deviations in the context of the produced product type into account, we subdivide the training set by the product types. Each resulting subset only contains measurements of a single product type and is also used to train models. So, we also build models that evaluate the deviation of measurements compared to their respective peer group with the same product type. A deviation from other measurements of the same product type should be detected by these models. As machine learning techniques for the models we used: Isolation Forest [18], One-Class Support Vector Machine (SVM) [19], Autoencoder [20] and Variational Autoencoder (VAE) [21], k-Nearest Neighbors (KNN) [22], Minimum Covariance Determinant (MCD) [23], and Histogram-based Outlier Score (HBOS) [24].

For a new measurement anomaly detection is performed with both the global model and the model of the respective product type. The state of the machine during the measurement is then deemed anomalous if both models detect it as an anomaly. Otherwise, it is considered normal.

To evaluate the performance of the anomaly detection in such a scenario we calculate the precision, recall and the scores using the scoring method by Lavin et al. [25]. We use a window of 4 hours before the actual event for all metrics, as at the time of the event there is already an impact on the machine and in the best-case events should be detected before they have an impact. Therefore, all detections within that window are considered true positives. For the method by Lavin et al. [25] we also calculate the scores for all three proposed profiles, giving different weights to false positives, false negatives, and true positives. The standard profile of this method is also used as the metric during parameter tuning.

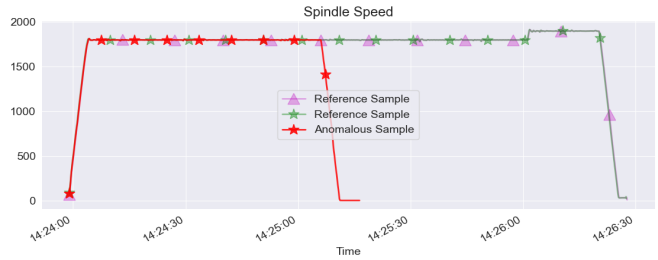
TABLE II: RESULTS OF THE ANOMALY DETECTION EXPERIMENTS.

Method	Recall	Precision	Standard[25]	Low FN[25]	Low FP[25]
One-Class SVM	86.6	41.5	59.89	68.85	53.03
Isolation Forest	86.6	44.1	61.31	69.80	54.90
Autoencoder	86.6	41.2	60.23	69.08	53.71
VAE	86.6	40.6	60.04	68.96	53.34
KNN	30.0	50.0	24.25	26.17	23.51
MCD	53.3	23.8	33.66	40.24	25.52
hbos	83.3	37.8	58.16	66.58	51.54

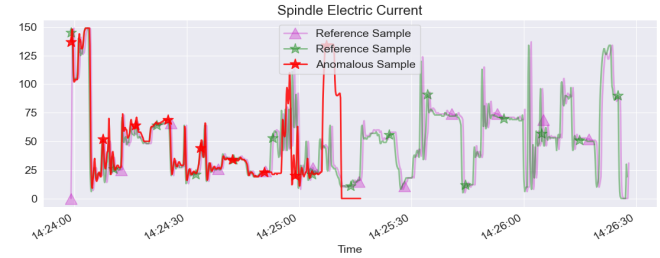
We show the results using the baseline approach to detect machine defects in Table II. In this table, we show the recall, precision, and scores of the aforementioned scoring method with the three default profiles for the corresponding machine-learning method. One-Class SVM, Isolation Forest, Autoencoder and Variational Autoencoder already perform quite well and can detect or find early indicators for 86% of the detects during the test period. Isolation Forest performs slightly better than the others as it has fewer detections without corresponding labels. In Fig. 6, we show the first anomaly detected by the four aforementioned methods during the test period. This detection is compared to two other randomly selected reference samples of the same product types by overlapping and aligning them by their start. Visually the anomalous measurement is distinct from the other measurements. The curve of the electric current of the anomalous sample starts to deviate 45 seconds into the measurements, while the movement drops off after 60 seconds. The production then stops too early shortly after that and no more data is received, while the production in the reference samples continues normally. This anomaly also coincides with a critical defect that occurs at the same time. In terms of precision the scores are lower. There we have a precision of 40% in most cases, except for KNN where the recall is also very low. Generally, the detection of false positives is a challenge for all methods. All methods have a better recall than precision. This also reflects in the other scoring metrics. The scores are generally lower when putting an emphasis on few false positives. Therefore, a relatively high recall is already achievable and thus the considered types of events are detectable. On the other hand, detections that cannot be attributed to the considered events exist and a higher precision would be desirable. One thing to note in that context, is that only critical and major labels created by factory workers are used for evaluation. However, anomalies could also exist outside of these labels, as there are also other types of events and the labels are not created by analyzing the sensor signals themselves but by observing the production of the machine. Therefore, the false positives are only the context of the available labels for critical and major events. As a baseline, we manage to achieve a recall of up

 ISBN: 978-1-68558-041-4

to 86% and a precision of up to 40% to 50% on major and critical events by performing anomaly detection in the global context and the context of the concrete product type.



(a) Speed of the milling spindle for one anomalous sample and two randomly selected reference samples



(b) Electric current of the milling spindle for one anomalous sample and two randomly selected reference samples

Fig. 6. Speed and electric current of the spindle during the first anomaly detected by the One-Class SVM in the test set compared to two randomly selected reference samples at other times. The start of the reference samples is aligned with the start of the anomalous sample to visualize differences.

IV. Challenges with Dataset/Future

The experiment and results in this paper are the foundation for future work and intended to be a basis for assessing new approaches and experiments. Therefore, there is potential to refine the approach or find new approaches that perform better in terms of the raw scores (e.g., by using different representations of the sensor data or machine learning techniques). On the other hand, we only use the most critical events as target labels. While we can already achieve decent results in terms of recall, the performance in terms of precision lower. This suggests there could be other events, in addition to the considered ones, that could be identified reliably. Consequently, exploring the recognizability of different event types is another aspect to look at.

During our experiments and first tests for an online application, collaborating with technical experts by communicating the occurrence of anomalies posed a big challenge. Usually, the occurrence of an anomaly alone is insufficient information for them. On the other hand, the technical experts usually have extensive knowledge about the machine on a technical level. Communicating an explanation in how the sensor data deviates or the expected type of problem that will occur could help to resolve machine problems more efficiently. So, after performing anomaly detection, explaining the detection, or linking it to a concrete type of problem would be another goal. Connected to explaining the detection, is the automatic labelling of the defects. Currently workers must manually label downtimes of the machine when they occur. This means when workers are not present during the downtime or lack training, information can be lost about the cause. As the labelling is also performed manually, there is always the potential that mistakes can occur. This can potentially hinder performing special measures against the systematic occurrence of certain kinds of defects. When certain problems arise regularly or in a high frequency, more throughout inspection and maintenance needs to be performed to eliminate the cause. Therefore, another aspect to improve the uptime of the machinery would be to use historic information from the dataset and create an automatic classification system to also label the downtimes automatically. Lastly, the labels in this dataset only capture the potential effect of anomalies and the labels are not directly linked to samples. Anomalies in the sensor data are not labeled. This poses a challenge when evaluating new algorithms as commonly used metrics are only applicable to a limited extent. In our experiment we dealt with this by using time windows around the events. However, this requires a parameter that influences semantics of the anomaly detection. As labels for anomalies in sensor data are often unavailable in real-world industrial processes, exploring non-parametric methods to evaluate anomaly detection with fuzzy labels can help to improve the applicability in industrial settings.

Currently our approach is based on raw sensor data streams and the hypothesis that an unexpected event, named anomaly, happens in the near future, and with a causal relation to a critical event. Given this systematic, our approach can be generalized to cases, where a given signal characteristic and its future outcome is known, but it's unclear when the signal characteristics itself began the diverge from the expected one, finally resulting in an unexpected or unwanted system behavior.

V. Conclusion

Detecting failures and defects of industrial machines by observing deviations from the normal operations is an important aspect to increase the availability of machinery and the efficiency of industrial production lines. By providing information about possible (upcoming) defects to machine operators and technical personnel actions can be taken to avoid or mitigate possible damages. One challenging scenario, that becomes more important as industrial production advances, is the detection of failures in a highly dynamic production, where the production can rapidly change depending on the requirements for the desired product. In this paper, we present a real-world dataset collected from a single machine that is part of a dynamic production line. In this production, line configured product types can be produced interchangeably depending on demand. The dataset consists of intrinsic sensor data, collected by internal sensors that are part of the default equipment of the machine. These sensors measure the movement and electric current of different machine parts. In addition to the sensor data, we also provide the occurrences of observed machine downtime - manually labelled by workers - as a form of

 ISBN: 978-1-68558-041-4

ground truth. Along with the dataset, we also provide a tool to access and work more easily with the data by providing a playground to test new approaches. We show initial results of an approach that uses a majority vote between anomaly detection in a global context and in the context of the concrete product type to detect machine defects. There, we achieve a recall up to 86% and a precision of around 40% to 50%. This shows that, while being able to detect already a high number of defects, the precision should still be improved for technicians to effectively use the information. These results serve as a baseline for future work and improvements to detect defects more reliably in a dynamic real-world process. Further, we also outline additional challenges - aside from detecting the machine defects - that operators of the machine have when interacting with the machine to gain insight on the machine, increase uptime and take correct measures. This dataset could also help to tackle these challenges and further improve industrial production.

VI. Availability of the Dataset and Tools

The full dataset, the scripts, and the sharing platform for this paper are made publicly available at https://www.hasisaurus.at/DataSet.html. When using our Dataset please cite this work and/or one of [12][13][14][15][16].

References

[1] G. Aceto, V. Persico, and A. Pescapé, "A survey on information and communication technologies for industry 4.0: State-of-the-art, taxonomies, perspectives, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3467–3501, 2019.

[2] G. Hoelzl, "A personalised body motion sensitive training system based on auditive feedback," in *Proceedings of the* 1^{st} *Annual International ICST Conference on Mobile Computing, Applications, and Services (MobiCASE09)*, ser. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, T. Phan, A. Montanari, and P. Zerfos, Eds., vol. 35, ICST. San Diego, California, USA: Springer, October 26-29 2009, ISBN: 978-3-642-12606-2.

[3] F. Huppert, G. Hoelzl, and M. Kranz, "Guidecopter - a precise drone-based haptic guidance interface for blind or visually impaired people," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021.

[4] A. Adel, M. A. Seif, G. Hoelzl, M. Kranz, S. Abdennadher, and I. S. M. Khalil, "Rendering 3D virtual objects in mid-air using controlled magnetic fields," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, Canada*. IEEE/RSJ, 2017, pp. 349 – 356.

[5] D. Preuveneers and E. Ilie-Zudor, "The intelligent industry of the future: A survey on emerging trends, research challenges and opportunities in industry 4.0," *Journal of Ambient Intelligence and Smart Environments*, vol. 9, no. 3, pp. 287–298, 2017.

[6] A. Angelopoulos *et al.*, "Tackling faults in the industry 4.0 era—a survey of machine-learning solutions and key aspects," *Sensors*, vol. 20, no. 1, p. 109, 2020.

[7] J. Lee, H.-A. Kao, and S. Yang, "Service innovation and smart analytics for industry 4.0 and big data environment," *Procedia Cirp*, vol. 16, pp. 3–8, 2014.

[8] B. Luo, H. Wang, H. Liu, B. Li, and F. Peng, "Early fault detection of machine tools based on deep learning and dynamic identification," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 1, pp. 509–518, 2018.

[9] D. Y. Oh and I. D. Yun, "Residual error based anomaly detection using auto-encoder in smd machine sound," *Sensors*, vol. 18, no. 5, p. 1308, 2018.

[10] R. M. Souza, E. G. Nascimento, U. A. Miranda, W. J. Silva, and H. A. Lepikson, "Deep learning for diagnosis and classification of faults in industrial rotating machinery," *Computers & Industrial Engineering*, vol. 153, p. 107060, 2021.

[11] D. Fernández-Francos, D. Martínez-Rego, O. Fontenla-Romero, and A. Alonso-Betanzos, "Automatic bearing fault diagnosis based on one-class ν-svm," *Computers & Industrial Engineering*, vol. 64, no. 1, pp. 357–365, 2013.

[12] S. Soller, G. Hoelzl, and M. Kranz, "Predicting machine errors based on adaptive sensor data drifts in a real world industrial setup," in *Proceedings of the* 18^{th} *Annual IEEE International Conference on Pervasive Computing and Communications (PerCom2020)*. IEEE, March 23-27 2020, pp. 1–9.

[13] S. Soller, B. Fleischmann, M. Kranz, and G. Hölzl, "Evaluation and adaption of maintenance prediction methods in mixed production line setups based on anomaly detection," in *International Workshop on Pervasive Information Flow (PerFlow'21), at 2021 IEEE International Conference on Pervasive Computing and Communications (Percom 2021), Kassel, Germany*, 2021, pp. 520–525.

[14] S. Soller, M. Kranz, and G. Hoelzl, "Adaptive error prediction for production lines with unknown dependencies," in *Proceedings of the 5th International Conference on Real-time Intelligent Systems (RTIS'20), Biarritz, France, Best Paper Award*, ser. WIMS 2020. New York, NY, USA: Association for Computing Machinery, 2020, pp. 227–234.

[15] G. Hoelzl, S. Soller, and M. Kranz, "Detecting seasonal dependencies in production lines for forecast optimization," *Big Data Research*, vol. 30, p. 100335, 2022.

[16] S. Soller, G. Hoelzl, T. Greiler, and M. Kranz, "Analysis of common prediction models for a fuzzy connected source target production based on time dependent significance," in *Proceedings of the 2022 International Conference on Embedded Wireless Systems and Networks*, ser. EWSN '22. USA: Junction Publishing, 2022, pp. 226–231.

[17] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," *Knowledge and information Systems*, vol. 3, no. 3, pp. 263–286, 2001.

[18] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[19] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Advances in neural information processing systems*, 2000, pp. 582–588.

[20] N. Japkowicz, C. Myers, and M. Gluck, "A novelty detection approach to classification," in *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1*, 1995, pp. 518–523.

[21] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[22] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 427–438.

[23] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[24] M. Goldstein and A. Dengel, "Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm," *KI-2012: Poster and Demo Track*, pp. 59–63, 2012.

[25] A. Lavin and S. Ahmad, "Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 38–44.

 ISBN: 978-1-68558-041-4

Investigating the Potential for Open Government Data (OGD) in Qatar

Ali Albinali
Department of Computer Science
Loughborough University
Loughborough, UK
e-mail: A.Albinali@lboro.ac.uk

Russell Lock
Department of Computer Science
Loughborough University
Loughborough, UK
e-mail: R.Lock@lboro.ac.uk

Iain Phillips
Department of Computer Science
Loughborough University
Loughborough, UK
e-mail: I.W.Phillips@lboro.ac.uk

***Abstract*—Data are growing rapidly, and technological breakthroughs are adding new methods and strategies to tackle big data. Thus, research challenges exist to explore the value of big data in supporting improved decision-making and public service delivery across various industries. In recent years, many countries have started to utilize Open Government Data (OGD) in developing open environments and platforms to improve their economies, enabling small and medium enterprises (who may lack the resources to undertake their independent market analysis) to use them. Qatar is no exception; it has begun to promote and make it available to everyone, including small and medium-sized companies. That is due to the potential economic returns and enhanced transparency. This research looks at aspects of OGD and the situation in Qatar. In addition, we collected data from various organizations and stakeholders using two questionnaires. The key goals of the two independent questionnaires were to investigate the amount of knowledge of OGD as a concept, how Small and Medium Enterprises (SMEs) in Qatar perceive present OGD platforms, and how OGD may be improved to fulfill the requirements of SMEs. Further, regardless of gender or educational level, most Qatari youth and adults reacted positively to using OGD.**

Keywords- Big data; Open Government Data (OGD); Small and Medium Enterprises (SMEs); Data Analytics; Data Analytics Framework (DAF).

I. INTRODUCTION

SMEs adopt data analysis and management tools, enabling them to visualize information by synthesizing data from multiple value chain processes [1]. In addition, introducing and utilizing Big Data in any SME requires establishing a model to gather and analyze the data. A good data analysis model will help analyze raw data sets to understand and realize the information they contain and discover patterns in the data to derive valuable insights. Furthermore, adopting a data analysis model aims to overcome data handling difficulty and exploit the volume, variety, velocity, and veracity elements that necessitate a significant data flow in a constantly evolving system [2]. It generates a convenient analysis to solve the challenges caused by a growing amount of data. Further, with the appropriate data analysis model, SMEs can maximize their performance effectively, cutting costs by developing more efficient ways to conduct business and maintaining enormous volumes of data. SMEs using readily available data to provide high-quality services will advance in a competitive climate. Therefore, utilizing Big Data is an opportunity for SMEs to innovate and offer value-added services to their customers.

Furthermore, governments are seeking new ways and methods to support diversification and promote SMEs in the economy. Using data analysis and management platforms allows governments to understand the needs of their citizens better, eliminate systemic flaws, and improve operations, cutting costs and boosting the services provided by any government entity [3]. OGD can help maximize the benefits of using big data. In addition, OGD is a subset of open data and is government-related data open to the public [4]. It refers to publicly accessible information about the government [2]. Multiple datasets, including those related to finance, population, geography, the public, transportation, traffic, education, etc., may be included in government data.

The economy of Qatar is among the most robust in the region and one of the most promising in the world. Qatar maintained balanced growth rates in the face of global concerns, and its GDP increased by more than 12% in the third quarter of 2021 compared with the first. According to the policies put in place by the Qatar National Vision 2030, which aims to lay the groundwork for a competitive and diversified knowledge-based economy, Qatar has succeeded in bolstering its economic standing on the global stage over the past few years [5]. Moreover, the Qatari government promotes using Open Data to encourage openness, public participation, better governance, inclusive economic growth, and innovation in Qatar [5].

SMEs utilize OGD for different reasons. However, due to limited resources, SMEs, particularly small start-ups, need help to use big data analytics effectively [4]. SMEs throughout the Gulf Cooperation Council (GCC) and Qatar are still in the early stages of using OGD [6]. They encounter a number of difficulties when implementing OGD, including the fact that the datasets are not updated frequently and are only available in unprocessed formats, there is no interpretation or clear description for these datasets, the majority of the datasets are only available in Arabic, some GCC countries lack a clear OGD policy and classify OGD under other topics, users are discouraged from adding to the datasets, and the formats and interactive maps and usability are limited.

The current OGD platforms, generally and in GCC countries mainly, only employ data from government agencies; they do not mix data from other sources, such as third-party data and social media, with analytics. The OGD lifecycle classification was one of the most significant categories lacking. In other words, these difficulties should be categorized according to the various stages/phases of the OGD lifecycle. Furthermore, they might be classified based on the actions performed at each step of the OGD lifecycle. As a result, the decision maker or end user may effectively track, monitor, and address the difficulties.

As a result, the Qatari government needs to provide SMEs with an OGD platform to use big data analytics to achieve

 ISBN: 978-1-68558-041-4

more innovation and growth. The required OGD platform must give more than the available platforms and state-of-the-art currently allows. The government, companies, and the general public via social media are the most known data sources. With the help of these various data sources, SMEs will be able to capture the information needed that will aid them in analyzing trends and seeking new business opportunities [7]. Information acquired from social networks and other sources should be merged with public data.

This paper provides research into the use of OGD and the deployment of OGD in Qatari SMEs to investigate the Qatari government's and SMEs' readiness for employing OGD. Two separate surveys were undertaken to measure the amount of understanding of OGD as a concept, as well as how SMEs in Qatar perceive the present OGD platforms and how they could be improved to fit their needs.

To do this, we developed the following research questions, which will be addressed in this paper:

- What is the level of awareness in the public and private sectors about OGD?
- Are there policies and practices to encourage the use and publication of GD in Qatar?
- Is data mainly supplied directly, indirectly, or both on Qatar's Open Data Portal?

In this research, firstly, we selected different types of organizations to participate in the data collection. Organizations may be governmental, semi-governmental, or private. We assume that the selection of varying organization types may affect the data collection method and the procedure for requesting accessibility for both organization's stakeholders and data; secondly, the choice of stakeholders that will participate from these organizations. The organization's type and the stakeholder's background and experience may impact their responses, recommendations, decisions, and selections.

The rest of the paper is structured as follows. Section 2 presents the related background for OGD, its terms, OGD pros and cons. Section 3 introduces the methodology used in our study and the surveys that were undertaken. Section 4 presents the evaluation process and the analysis of the results. Finally, we give our conclusions in Section 5.

II. BACKGORUND AND RELATED WORK

A. OGD Terminology

Many concepts are related to OGD, such as open data, public data, e-government, linked data, and data portal. Firstly, there is a need to differentiate between open data, public data, and OGD.

1) Open data: The open definition states the principles and guidance that open data should conform to regarding data and content [8]. Open indicates how anyone can freely access, use, reuse, and redistribute data for any purpose regarding the requirements that preserve provenance and openness. Therefore, data is published in open data format, machine-readable, platform-independent, and open to the public without restrictions or under an open license [8]. Therefore, open data refers to data that is free of charge to the public without limitations [9]. Open data is considered a key enabler of open government [2].

2) Public data: public data is made freely available to the public but only sometimes open. An example of public data is the archive of legal documents, which are accessible freely. On the other hand, if these public data are organized in a digital format, sorted and indexed, and made available online in a standard format. This public data will be open also. Open data contains heterogeneous data from several sources. So, there is a need for a body to host these data centrally, and the government is a clear choice.

3) OGD: OGD is a subset of open data and is government-related data open to the public [2]. Government data might contain multiple datasets such as finance, population, geographical, public, transportation, traffic, education, etc. [10]. [11] defined another definition for OGD as follows: "Open data is data that anyone can access, use, or share. Simple as that. When big companies or governments release non-personal data, it enables small businesses, citizens, and researchers to develop resources which make crucial improvements to their communities."

Secondly, other terms such as e-government, linked data, and data portal are defined, which also relate to OGD.

1) E-government: there are many definitions of e-government existing in the literature; the one related to the government's use of technology is used to improve its offered services to other entities, including citizens, employees, partners, suppliers, and other government agencies [12]. Therefore, by supporting the connection between citizens and their government, e-government can develop better relationships and deliver information and services more efficiently. While initially, e-government just referred to the presence of government on the Internet as an informative website; the concept has since evolved. With the introduction of the 'open government' concept, open government data initiatives are considered a subset or an extension of e-government [13].

2) Linked data: it is the process of following a set of best practices for publishing and connecting structured data on the web [14]. Linked data refers to data that is published on the web, and it is also connected to other external datasets.

3) Data portal: the open data movement targets making data open for government and public sector information to boost its reuse. A typical implementation is to gather and publish datasets into central data portals or data catalogs to provide a "one-stop-shop" for data consumers [15]. While a data catalog would act as a registry of data sources, providing links, a data portal is more commonly a single entry point hosting the actual data, where end users can, search and access the published data and interact with it suitably [16]. One of the main functions of a data portal is the administration of metadata for the datasets, potentially including metadata harmonization. Different tools are

 ISBN: 978-1-68558-041-4

enabled on government data portals, for example, data format conversion, visualization, query endpoints, etc. Therefore, Open Data Portals (ODP) are essential, and the solution will provide an ODP with a data analytics framework for SMEs.

III. USE SMES MOTIVATIONS FOR THE UTILIZATION OF OGD

A. Generated Economic Value Through Open Data

Open data is already contributing to the economic growth of countries worldwide [23][24]. They also support creating and strengthening new markets, companies, and jobs [19]. Government plays a vital role in creating value from open data, not only in its publication stage. Organizations can create value with open data in various companies and industries in three ways [20]: for traditional companies or new non-technological startups to make decisions, in the same way as their governments use open data to improve decision-making, the same can happen for the vast private sector; to generate new products or services that create value for the clients of the companies; and to be accountable in a market where consumers require more information and reward transparent companies: by releasing data, companies can guarantee that their actions are transparent [13].

B. Promote Greater Openness of Public Data

One of the mechanisms available to the government to encourage the use of open data by the private sector is to strengthen the supply of these data in quantitative and qualitative terms. In South Korea, the government has promoted a series of measures to promote open data, allowing the development of many digital applications from public open data. One of the most active Open Data sites is the Seoul Open Data Plaza (data.seoul.go.kr), managed by the metropolitan government of Seoul [21]. In 2012, Seoul initiated an open data initiative sharing public information to create diverse business opportunities for the private sector and develop IT industries. This portal is an online channel to share and provide citizens with all public data of Seoul, such as real-time bus schedules, subway schedules, Wi-Fi public service places, and facilities for disabled people, among others [22], [23].

C. Promote or Regulate the Opening of Data in Other Sectors

Governments, international organizations, and civil associations have been at the forefront of open data proliferation and openness [18]. As governments have adopted the open data agenda, citizens and consumers demand transparency in other sectors, such as business, academia, and government organizations [24]. As noted in this document, data has become the currency of modern economics. A recent study published by the "Future of Privacy Forum" 30 projects that the global data volume will grow from approximately 0.8 zettabytes (ZB) in 2009 to more than 35 ZB in 2020 [25]. Likewise, the government can play a role in encouraging companies to share their data safely and respectfully regarding the privacy of consumers and citizens [26]. Universities and academia (in their various institutions such as science and technology agencies) also have to take a step forward to publicize and give access to their data in different formats to other societal actors [27].

D. Promote Data Entrepreneurship

The recent success story of ODINE (Open Data Incubator Europe) 32 and the well-known emergence of data ventures in the United Kingdom and the United States have demonstrated the opportunity to generate value, scalability, and profits with open data ventures. In Latin America, there are also cases of successful data ventures that have received foreign investment and have grown in the last 5 years, for instance, OPI, Data4, and Atlantia [25]. Moreover, in the United Kingdom, the government has offered open government data of the highest quality through data.gov.uk. The Open Data Institute (ODI), in its Open Data Means Business research, has analyzed 270 companies in the United Kingdom that use, produce, or invest in open data as part of their business strategy. These companies (also called "open data companies") invoice more than 110 billion dollars a year and employ more than 500 thousand people [26].

IV. CHALLENGES AFFECTING SMES FULL UTILIZATION OF OGD

A set of classifications and categories for the challenges that prevent SMEs in the GCC region from utilizing the OGD effectively are discussed. Saxena [6] applied the models which have been developed to the status of OGD in the GCC countries [28]. Sieber and Johnson have introduced four models of open data that define the relationship between citizens and government (also called the Citizen engagement model of OGD) [29]; for more details about the benefits and costs of these four models (See [29]).

1) Data over the wall - Government publishing of open data: the government publishes open data via an online open data portal that acts as a unidirectional conduit from the data owner/collector (government, community, organization) to the end user (citizen, community organization, or private sector).

2) Code exchange - Government as open data activist: the government supports the reuse of open data to directly extract or create value from its offering, e.g., through app development contests.

3) Civic issue tracker - Data from citizen to government: the government accepts direct feedback from citizens on a limited range of issues in a crowdsourcing paradigm. Data may or may not also come from the government.

4) Participatory open data - Open data as open government: the government-citizen co-production of data where open data becomes a direct conduit between citizen and government, where citizen contributions are dynamic, and the government becomes responsive to demand-side requests for data.

According to [6], most GCC countries fall into the first model outlined above, “Data over the wall”. Therefore, all GCC is still at an early stage or phase for utilizing and implementing OGD [6]. They face a number of challenges in

 ISBN: 978-1-68558-041-4

the OGD implementation, such as the datasets are not regularly updated, the available datasets are in unprocessed format, there is no interpretation or clear description for these datasets, and most of the datasets are available in the Arabic language only, some GCC countries have no clear OGD policy and classify OGD under other topics, discouraged users from contributing to the datasets, and limited formats and interactive maps and user-friendly formats.

Saxena discussed drivers and barriers to reusing OGD in Oman as one of the GCC countries [30]. A qualitative approach has been applied to the national OGD portal of Oman (https://data.gov.om). The national data portal of Oman is a free data-sharing portal where anyone can access data relating to the Sultanate of Oman. The data portal provides datasets from different entities for everyone - citizen, investor, researcher, or developer [31]. The national OGD portal of Oman has published over 56 data sets across 12 sectors. Moreover, there are 17 data providers or entities and three mobile apps which may be used by different users [30], [32]. Saxena concluded that Oman's OGD initiative could be classified as a hybrid of the three models [29]; Data over the wall, Code exchange, and Participatory open data [30].

Having discussed the challenges and barriers of SMEs so far, especially for GCC, various challenges/issues and classifications still need to be mentioned. For example, the OGD platforms use exclusive data from governmental entities; they do not incorporate data from different sources, such as third-party data and social media combined with analytics. Therefore, one of the essential classifications that needed to be included in the OGD lifecycle classification. Moreover, it could be categorized by the different activities in each stage of the OGD lifecycle. Therefore, the end user could track, monitor, and tackle the challenges effectively.

V. METHODOLOGY

Data collection facilitates and improves the decision-making process and the quality of those decisions. Thus, to answer our research questions related to the role of SMEs in Qatar in utilizing and spreading the use of OGD, we used a mixture of quantitative and qualitative data collection methods, such as surveys [33]. The primary objectives of the two independent surveys were to examine the level of awareness of OGD as a concept, how SMEs in Qatar view the current OGD platforms, and how OGD may be enhanced to meet the needs of SMEs.

We targeted two categories of stakeholders for the survey. The first category is the public, i.e., citizens and residents, and the second category is SMEs and Investors. So, we have designed a survey for each category as follows.

The first survey was called the " OGD Awareness Survey", which aims to evaluate the awareness of citizens and residents in Qatar of the Open Government Data. Further, the second survey was called " OGD – SMEs and Investors Survey", which aims to evaluate the awareness and utilization of OGD by SMEs and Investors in Qatar. The surveys were divided into several sections.

VI. RESULTS AND EVALUATION

A. Selection of Organizations and Stakeholders

We have selected various organizations and stakeholders participating in the data collection process. Examples of organizations are the Ministry of Interior (MOI), Ministry of Justice (MOJ), Qatar Development Bank (QDB), Qatar International Court and Dispute Resolution Centre (QICDRC), Ministry of Commerce and Industry and Hukoomi.

B. Survey Data Analysis Procedure and Used Tools

The surveys were introduced in English to manage the different terms included in the survey. Two channels for the targeted stakeholders of the first survey (OGD Awareness Survey) were used. The first channel used was a request for an official email list of consumers or end users of the Ministry of Interior (MoI) and State of Qatar services, either citizens or residents, through an SMS application. After receiving the list of four hundred emails from MoI, these emails were added to an email group and sent an email containing the survey's purpose and URL. The second channel was conducted through a series of visits to MoI service locations, connecting us with participants willing to participate in our study. Then, we met with another one hundred participants to complete the survey using an iPad. Therefore, this survey was distributed to 500 participants. The survey was designed to be online using a tool called Microsoft Forms, which we have authorized access to through Loughborough University within Microsoft Office 365. As a result, 422 responses received a return rate of 84%. The output of Microsoft Forms is a Microsoft Excel file containing participant records.

In the distribution plan of the second survey (OGD – SMEs and Investors Survey), we requested an authorized email list of SMEs from the MoI and State of Qatar services. After receiving the list of 125 emails from MoI, we added them to an email group and sent them an email containing the survey's purpose and URL. The survey was designed to be completed online using Microsoft Forms. Finally, we received 101 responses, a return rate of 81%.

Moreover, we performed an initial analysis in Microsoft Excel files to check which responses should be included in the statistical analysis. For the first survey, we found that 94 records of the total records answered "No" in the consent section, which means they did not participate in our survey. As a result, we found that 328 were valid for analysis. For the second survey, only one record of the total records answered "No" in the consent section, which means they did not participate in our survey. Thus, 100 SMEs or investors were considered valid for analysis. Finally, cleaning and transformation steps were necessary to prepare an Excel file for statistical analysis. Firstly, we changed or renamed the names of columns in the first row or header to meaningful names. Secondly, we carried out two main transformation steps in Microsoft Excel using a Power Query Editor such as the following:

- Remove a set of columns related to Microsoft Forms, such as Start time, completion time, and email.

 ISBN: 978-1-68558-041-4

- Replace the null values in columns with the Not Available (NA) value.

Furthermore, the transformed Microsoft Excel files for the two surveys were imported into IBM Statistical Package for the Social Sciences (SPSS) to provide further analysis of the survey results. SPSS is one of the best-known statistical techniques researchers use to deliver advanced statistical analyses, including the Chi-Square Test, which provides a reliable estimation of research results and can define the relationships between research variables [40][41].

C. OGD Awareness Survey - Data Analysis Findings

We selected all questions from Section A: Demographic Information and questions Q8, Q9, and Q10 from Section B: OGD Awareness. First, general analysis shows the number of responses that express the relationship between two variables using cross-tabulation analysis. Then, we performed an in-depth study that shows the dependency or independency between two questions or variables using the Chi-Square as statistical analysis.

TABLE I. IN-DEPTH ANALYSIS OF RELATIONSHIPS BETWEEN OGD AWARENESS SURVEY QUESTIONS

Question	Expected Related Questions	Chi-Square	P-value	Relationship Status
Q2 - Age Group	Q8.OGD-Reaction	39.681	21.026	Dependent
	Q9. OGD Qatar Gov Usability	29.023		Dependent
	Q10. OGD Qatar Third Parties Usability	22.790		Dependent
Q3 - Gender	Q8.OGD-Reaction	9.878	9.488	Dependent
	Q9. OGD Qatar Gov Usability	5.589		Independent
	Q10. OGD Qatar Third Parties Usability	2.158		Independent
Q4 - Nationality	Q8.OGD-Reaction	4.062	9.488	Independent
	Q9. OGD Qatar Gov Usability	9.595		Dependent
	Q10. OGD Qatar Third Parties Usability	3.979		Independent
Q5 - Highest Qualification	Q8.OGD-Reaction	38.548		Dependent
	Q9. OGD Qatar Gov Usability	22.812	21.026	Dependent
	Q10. OGD Qatar Third Parties Usability	13.270		Independent
Q6 – Computer Knowledge Level	Q8.OGD-Reaction	30.359		Dependent
	Q9. OGD Qatar Gov Usability	18.237	21.026	Independent
	Q10. OGD Qatar Third Parties Usability	19.323		Independent
Q7 - Employment Situation	Q8.OGD-Reaction	45.495		Dependent
	Q9. OGD Qatar Gov Usability	47.567	26.296	Dependent
	Q10. OGD Qatar Third Parties Usability	55.700		Dependent

1) We examined the relationship between Q2- Age Group and the three questions: Q8- OGD Reaction, Q9- OGD of QatarGov Usability and Q10- OD Qatar Third Parties Usability. Figure 1 shows the responses of the Q2- Age Group to question Q9 (OGD of QatarGov Usability). Both responses from age groups range 18-30 and 30-44 indicate that they are mostly Extremely likely to use OGD. These numbers indicate that OGD from the Qatar government is trusted and could be used or utilized by youth and adults. Moreover, we performed the Chi-Square analysis using IBM SPSS between questions Q2, Q8, Q9 and Q10. For example, the Chi-Square value is 29.023 for the relation between Q2 and Q9. Moreover, the degree of freedom "df" is 12, mapped to a p-value of 21.026 with a confidence of 0.95 according to the Chi-Square distribution. Therefore, there is a dependency between the two questions because the Chi-Square value is greater than the p-value. We found that there is an indication that a high percentage of youth and adults show their interest in OGD as a positive topic in responses to Q8, and they will utilize both governmental and third-party's open data. Moreover, these findings are confirmed by performing the Chi-Square analysis between Q2 and the three questions as in Table I.

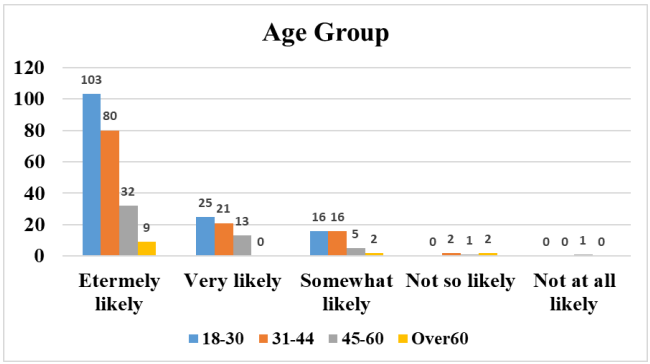


Figure 1. OGD Usability per Age Group.

2) We checked the relationship between Q3- Gender and the same three questions Q8, Q9 and Q10. Figure 2 shows the responses to question Q8. Both responses from gender, either male or female, mostly range between Very positive 117 for males and 73 for females, and somewhat positive 52 for males and 60 for females. These numbers represent a good indication that OGD is something required by both genders. Moreover, the Chi-Square value between the questions Q3-Gender and Q8, Q9 is 9.878, and the degree of freedom "df" is 4, which is mapped to a p-value of 9.488 with confidence 0.95 according to the Chi-Square distribution. Therefore, there is a dependency between the two questions because the Chi-Square value is greater than the p-value. Truly, there is an indication that a high percentage of males and females show their interest in OGD, and they may utilize both governmental and third-parties open data as in responses to OGD of QatarGov Usability and OGD QatarThirdParties Usability, respectively. Moreover, these findings are

 ISBN: 978-1-68558-041-4

confirmed by performing the Chi-Square analysis between Q3- Gender and the three questions, as shown in Table I.

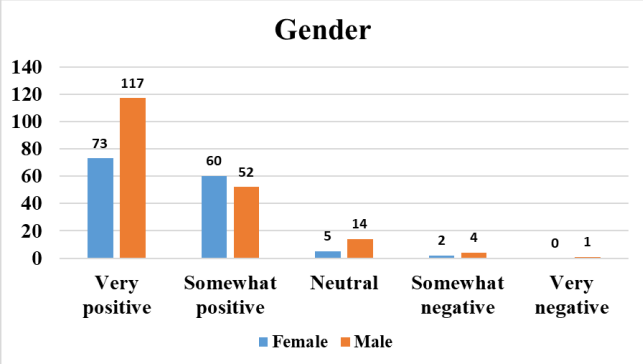


Figure 2. OGD Reaction per Gender.

3) We examined the relationship between Q4- Nationality and the three questions Q8, Q9 and Q10. Figure 3 shows the responses of Q4- Nationality to the variable. Both responses from Qatari or Non-Qatari are mostly Extremely likely to use OGD (140 Qatari and 84 Non-Qatari), and the other values have a slight difference for citizens. These numbers indicate that OGD from the Qatar government is trusted and could be used or utilized by citizens and residents. Furthermore, there is evidence of significant interest in using OGD from citizens and residents. These findings are presented in Table I.

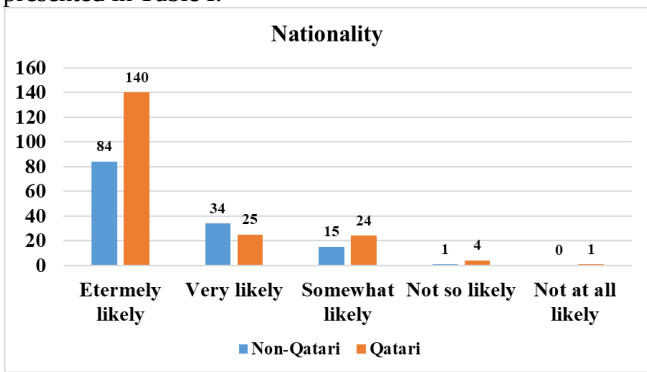


Figure 3. OGD Usability per Nationality.

4) We checked the relationship between Q5- Highest Qualification and the three Q8, Q9 and Q10. Figure 4 shows the Q5- Highest Qualification responses to the Q9- OGD of QatarGov Usability. Both responses for a Bachelor's degree or a Postgraduate degree mostly range between Extremely likely 99 for a Bachelor's degree and 87 for a Postgraduate degree, and Very likely 26 for a Bachelor's degree and 32 for a Postgraduate degree. These numbers indicate that OGD from the Qatar government is trusted and could be used or utilized by highly educated people in Qatar. Moreover, the Chi-Square value is 22.812, and the degree of freedom "df" is 12, which is mapped to a p-value of 21.026 with a confidence of 0.95 according to the Chi-Square distribution. Therefore, there is a dependency between the two variables because the Chi-Square value is greater than the p-value, as in Table I.

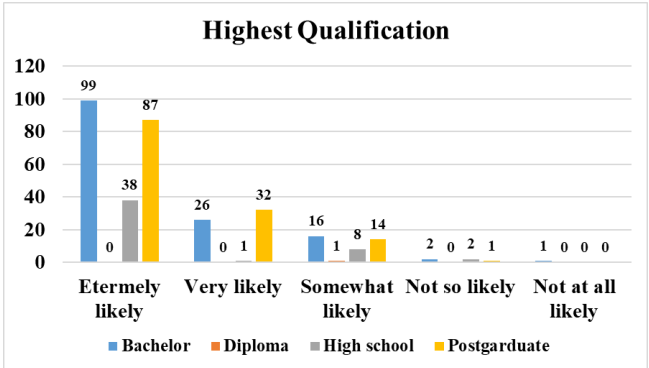


Figure 4. OGD Usability per Highest Qualification.

5) We examined the relationship between Q6- Computer Knowledge Level and questions Q8, Q9 and Q10. Both responses from Computer Knowledge Level Ranges Expert or Intermediate mostly range between Very positive 91 for Expert and 52 for Intermediate, and Somewhat positive 62 for Expert and 39 for Intermediate, as illustrated in Figure 5. These numbers represent a good indication that OGD is something required by people who have expert or intermediate computer knowledge levels. Furthermore, facilitating data access for persons with limited computer abilities is a barrier and should be addressed in the available OGD platforms. The Chi-Square value is 30.359, and the degree of freedom "df" is 12, mapped to a p-value of 21.026 (see Table I) with a confidence of 0.95 according to the Chi-Square distribution. Therefore, there is a dependency between the two variables because the Chi-Square value is greater than the p-value of expert or intermediate computer knowledge levels interested in utilizing advanced features from government and third-parties open data.

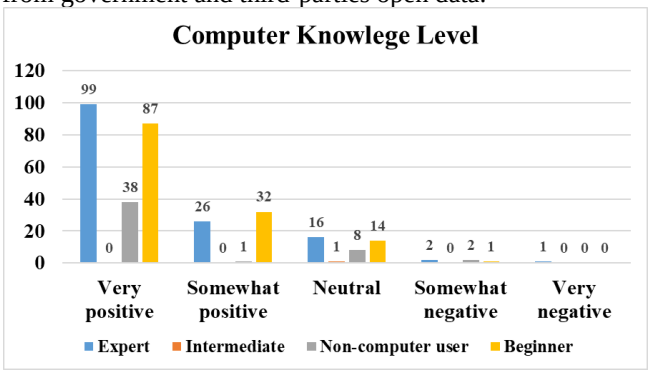


Figure 5. OGD Reaction per Computer Knowledge Level.

6) We checked the relationship between Q7- Employment Situation and the three questions Q8, Q9 and Q10. Figure 6 shows the responses of Q7- Employment Situation to question Q10 - OD QatarThirdParties Usability. Both responses from qualification ranges Expert or Intermediate mostly range between Very likely as 127 for

Working-Full-Time and 41 for in Education. These numbers indicate that open data from Qatar Third-parties are trusted and could be used or utilized by people working Full-Time or in Education. Moreover, the Chi-Square analysis between the question Q7- Employment Situation and the three questions (Q8, Q9 and Q10). For instance, the Chi-Square value for Q10 is 55.700, as shown in Table I, and the degree of freedom "df" is 16, which is mapped to a p-value of 26.296 with confidence 0.95 according to the Chi-Square distribution. Therefore, there is a dependency between the two variables because the Chi-Square value is greater than the p-value.

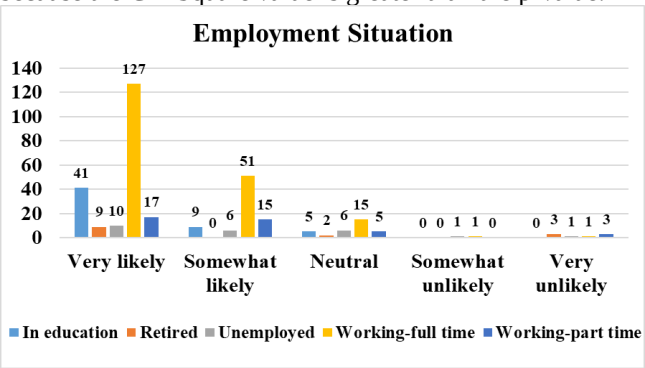


Figure 6. OGD Usability per Employment Situation.

D. *OGD – SMEs and Investors Survey - Data Analysis Findings*

This section discussed how we performed the statistical analysis for the second survey, " OGD – SMEs and Investors Survey," General Analysis, and In-Depth Analysis. We selected Q3, Q6, and Q7 from Section A (Demographic Information), question Q9 from Section B (OGD Awareness), Q12 and Q13 from Section C (OGD Organizational Information), Q18, Q19, Q21, and Q23 from Section D (OGD Technical Information) as in Table V. Through our analysis both general and In-depth, we used keywords of each question to express its meaning for simplicity.

TABLE II. IN-DEPTH ANALYSIS OF RELATIONSHIPS BETWEEN OGD SMES SURVEY QUESTIONS

Question	Expected Related Questions	Chi-Square	P-value	Relationship Status
Q3 - Age Group	Q9. SMEs Utilization Published OGD	2.539	12.592	Independent
	Q12. Verify Qatar Gov OGD Policy	4.397	12.592	Independent
	Q13. Qatar Gov Verify OGD Access	20.747	12.592	Dependent
	Q18. Qatar ODP Usability Purpose	4.392	16.919	Independent
	Q19. Qatar ODP Registration Provided Data Method	6.756	21.026	Independent
	Q21. Qatar ODP Registration Dissuade Status	4.614	12.592	Independent
	Q23. Qatar ODP Verify Data Analytics Usability	3.493	12.592	Independent
Q6 – Highest Qualification	Q9. SMEs Utilization Published OGD	10.743	9.488	Dependent
	Q12. Verify Qatar Gov OGD Policy	3.658	9.488	Independent
	Q13. Qatar Gov Verify OGD Access	10.379	9.488	Dependent
	Q18. Qatar ODP Usability Purpose	13.608	12.592	Dependent
	Q19. Qatar ODP Registration Provided Data Method	38.066	15.507	Dependent
	Q21. Qatar ODP Registration Dissuade_ Status	7.608	9.488	Independent
	Q23. Qatar ODP Verify Data Analytics Usability	6.925	9.488	Independent
Q7 – Computer Knowledge Level	Q9. SMEs Utilization Published OGD	13.132	9.488	Dependent
	Q12. Verify Qatar Gov OGD Policy	16.909	9.488	Dependent
	Q13. Qatar Gov Verify OGD Access	22.256	9.488	Dependent
	Q18. Qatar ODP Usability Purpose	17.619	12.592	Dependent
	Q19. Qatar ODP Registration Provided Data Method	31.504	15.507	Dependent
	Q21. Qatar ODP Registration Dissuade Status	8.729	9.488	Independent
	Q23. Qatar ODP Verify Data Analytics Usability	10.199	9.488	Dependent

1) We examined the relationship between Q3- Age Group and the seven questions: Q9- SMEs Utilization Published OGD, Q12- Verify QatarGov OGD Policy, Q13- QatarGov Verify OGD Access, Q18- Qatar ODP Usability Purpose, Q19- Qatar ODP Registration Provided Data Method, Q21- Qatar ODP Registration Dissuade Status and Q23- Qatar ODP Verify Data Analytics Usability. Figure 7 shows the responses of Q3- Age Group to question Q13- QatarGov Verify OGD Access. Both responses from age groups range 18-30 and 31-44 are mostly free of charge, as 37 for (18-30) and 35 for (31-44). These numbers indicate that responses from these age groups need free-of-charge access to Qatar government OGD. Moreover, the Chi-Square analysis between the question Q3- Age Group and the seven questions mentioned above are presented in Table II. For example, the Chi-Square analysis between Q3 and Q13- QatarGov Verify OGD Access is 20.747. According to the Chi-Square distribution, the degree of freedom "df" is 6, mapped to a p-value of 12.592 with confidence 0.95. Therefore, there is a dependency between the two questions as the Chi-Square value is greater than the p-value. After we performed the analyses between Q3 and these questions, we found no difference between age groups. Q9, Q12, Q13, Q18, Q19, Q21 and Q23. Furthermore, there is an indication that a high percentage of youth and adults need free-of-charge

access to Qatar government OGD as in responses to QatarGov Verify OGD Access. Thus, access to Qatar government OGD should be free of charge according to youth and adults' feedback.

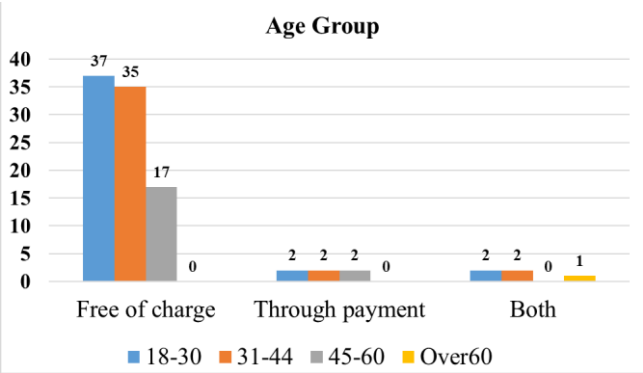


Figure 7. OGD Access per Age Group.

2) We examined the relationship between Q7- Computer Knowledge Level and the seven questions (Q9, Q12, Q13, Q18, Q19, Q21 and Q23). Figure 8 shows the responses of Q7 - Computer Knowledge Level to question Q23- Qatar ODP Verify Data Analytics Usability. Both responses from computer knowledge levels range between Expert and Intermediate or mostly between Yes 34 for Expert and 10 for Intermediate, and 33 No for Expert and 16 for Intermediate. These indicate that responses from these computer knowledge levels had no difference, which may have happened because they did not use an analytics platform. Moreover, the Chi-Square analysis using SPSS between the variables Q7 and the seven questions is illustrated in Table II. For example, the Chi-Square analysis between Q7 and Q23- Qatar ODP Verify Data Analytics Usability is 10.199, and the degree of freedom "df" is four which is mapped to a p-value of 9.488 with confidence 0.95 according to the Chi-Square distribution. Therefore, there is a dependency between the two variables because the Chi-Square value is greater than the p-value. After we performed both general and In-Depth analysis between Q7 and these questions, we found no difference between computer knowledge levels in their responses to the registration process ODP will dissuade the SMEs from utilizing the OGD.

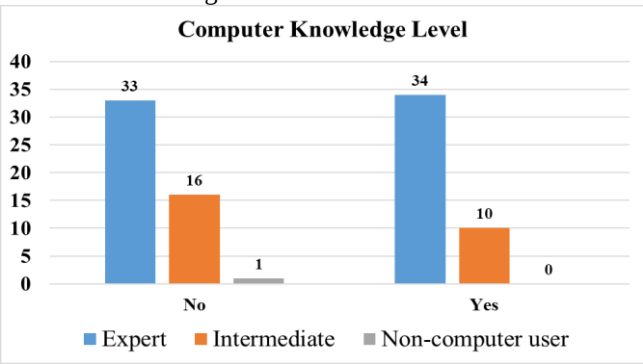


Figure 8. Data Analytics Usability per Computer Knowledge Level

3) We checked the relationship between Highest Qualification and the seven questions. Figure 9 shows the responses of Q6- Highest Qualification to question Q9- SMEs Utilization Published OGD. Both responses from the highest qualification range from a Postgraduate degree and a Bachelor's degree are between 35 Yes for Postgraduate degrees and 16 for Bachelor degrees, and 20 No for a Postgraduate degree and 21 for Bachelor's degree. These numbers indicate that SMEs' highest qualifications (Postgraduate and Bachelor degrees) utilize the published OGD. Moreover, the Chi-Square analysis using SPSS between question Q6 and the seven questions Q9, Q12, Q13, Q18, Q19, Q21, and Q23 are shown in Table II. For example, a Chi-Square analysis between Q6 and Q9 - SMEs Utilization Published OGD is 10.743, and the degree of freedom "df" is four which is mapped to a p-value of 9.488 with confidence 0.95 according to Chi-Square distribution. Therefore, there is a dependency between the two variables because the Chi-Square value is greater than the p-value. After we performed the analysis between Q6 and these questions, we found that there is an indication that a high percentage of highly educated people will utilize Published OGD, need free-of-charge access for Qatar government OGD, believe in the full or semi-full utilization of the open data portal and believe in that Qatar government OGD is provided from both direct (i.e., through ODP) and indirect (i.e., through the website of the ministry or the OGD source).

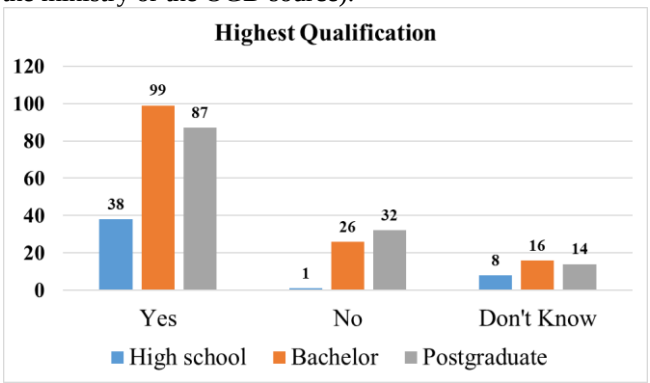


Figure 9. SMEs' Utilisation of Published OGD per Highest Qualification.

VII. Conclusion and Future Work

This paper explored the existing literature regarding the definition of OGD, its history, benefits that could be brought by using it, and any drawbacks. As a result, this paper paved the way for future researchers to develop prominent data analytics theories that would benefit the Qatari economy and company. Furthermore, we have conducted intensive surveys to gather information regarding the readiness to use and enjoy the fruits of OGD. The results showed how this open data could impact society and SMEs. The analysis is concluded with a set of key findings as follows:

 ISBN: 978-1-68558-041-4

1) Most youth and adults in Qatar reacted positively to utilizing OGD, as seen in their responses to the OGD Reaction and how they will use governmental and third-party open data. Therefore, there is a need for a plan to increase awareness of OGD between different age groups.

2) Gender appears not to be a factor in the willingness to utilize OGD in Qatar.

3) It does not matter what nationality is (i.e., Qatari or Non-Qatari) in Qatar towards the OGD, i.e., residents and citizens show interest in OGD as a positive topic with no distinction.

4) Regardless of education, gender, and employment status, people in Qatar have a positive reaction and are keen on OGD. For this reason, the OGD portal must be simple and easy to use by all people, regardless of their education level.

5) According to the replies to Qatar Gov Verify OGD Access, many youths and adults who are SME investors or owners want free access to OGD. As a result, OGD should be made available to SMEs at an accessible cost to help the Qatari private sector, as they may not utilize it if they have to pay.

6) Owners and investors of SMEs of different ages did not differ regarding the following questions: SMEs Use Published OGD, Verify QatarGov OGD Policy, Qatar ODP Usability Purpose, Qatar ODP Registration Provided Data Method, Qatar ODP Registration Dissuade and Qatar ODP Verify Data Analytics Usability.

7) Highly educated investors and owners of SMEs will utilize Published OGD, need to have free access to Qatar government OGD, and believe in full utilization (i.e., accessed, downloaded and used) or semi-full utilization (i.e., accessed and downloaded) of the ODP, and believe in that Qatar government OGD is provided from both direct (i.e., through ODP) and indirect (i.e., through the website of the ministry or the OGD source).

8) All qualifications are the same regarding their belief that the policy of OGD and the registration process of ODP will dissuade SMEs from utilizing the OGD.

9) Many expert and intermediate computer knowledge levels who are SME investors or owners will utilize OGD. They believe that the Qatar government OGD policy should exist and need free-of-charge access for Qatar government OGD. Additionally, they believe in the open data portal's full utilization (i.e., accessed, downloaded, and used) or semi-full utilization (i.e., accessed and downloaded). Moreover, they think that the Qatar government OGD is provided directly (i.e., through ODP) and indirectly (i.e., through the ministry's website or the OGD source).

10) There is no difference between computer knowledge levels in their responses to the registration process of the ODP will dissuade the SMEs from utilizing OGD.

The Qatari government should create a national-level, centralized service where SMEs utilize big data analytics tools and examine open data supplied by the government and others to improve business decision-making and discover new chances for expansion and innovation. We are now working completion completing the development of an OGD platform that encourages the use of OGD and overcomes the issues raised in this article. With big data analytics, the planned OGD platform would assist SMEs in achieving more innovation and growth. Additionally, it establishes a centralized, national service where SMEs may use big data analytics tools and techniques and evaluate open data provided by the government and others to enhance corporate decision-making and find new opportunities for growth and innovation. Additionally, data from social networks and outside sources will be combined with open data to give additional information to SMEs, reflecting the whole economic picture and assisting decision-makers in delivering better conclusions.

REFERENCES

[1] M. Van Rijmenam, T. Erekhinskaya, J. Schweitzer, and M.-A. Williams, "Avoid being the Turkey: How big data analytics changes the game of strategy in times of ambiguity and uncertainty," in *Long Range Planning*, 2019, vol. 52, no. 5, p. 101841, doi: https://doi.org/10.1016/j.lrp.2018.05.007.

[2] J. Kučera, D. Chlapek, and M. Nečaský, "Open Government Data Catalogs: Current Approaches and Quality Perspective," in *Technology-Enabled Innovation for Democracy, Government and Governance*, 2013, pp. 152–166.

[3] R. Fernandez and S. Ali, "SME contributions for diversification and stability in emerging economies - an empirical study of the SME segment in the Qatar economy," *J. Contemp. Issues Bus. Gov.*, vol. 21, no. 1, pp. 23–45, Jul. 2015, [Online]. Available: https://search.informit.org/doi/10.3316/informit.103635030363705.

[4] S. Akter, S. F. Wamba, A. Gunasekaran, R. Dubey, and S. J. Childe, "How to improve firm performance using big data analytics capability and business strategy alignment?," *Int. J. Prod. Econ.*, vol. 182, pp. 113–131, 2016, doi: https://doi.org/10.1016/j.ijpe.2016.08.018.

[5] T. Mohammed, "Economic Policy," *Government Communication Office*, 2022. https://www.gco.gov.qa/en/focus/economic-policy/.

[6] S. Saxena, "Significance of open government data in the GCC countries," *Digit. Policy, Regul. Gov.*, vol. 19, no. 3, pp. 251–263, Jan. 2017, doi: 10.1108/DPRG-02-2017-0005.

[7] C. A. Ardagna, P. Ceravolo, and E. Damiani, "Big data analytics as-a-service: Issues and challenges," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 3638–3644, doi: 10.1109/BigData.2016.7841029.

[8] "Open Definition," 2019. https://opendefinition.org/ (accessed May 23, 2022).

[9] K. J. Reiche and E. Höfig, "Implementation of Metadata Quality Metrics and Application on Public Government Data," in *IEEE 37th Annual Computer Software and*

Applications Conference Workshops, 2013, pp. 236–241, doi: 10.1109/COMPSACW.2013.32.

[10] White house, "Open Government Partnership," *Open Government Intiative*, 2011. https://obamawhitehouse.archives.gov/open/partnership#:~:text=President Obama launched the Open,hundreds of civil-society organizations.

[11] ODI, "ODI," *Open Data Institute*, 2020. https://www.theodi.org/ (accessed Mar. 01, 2020).

[12] K. Layne and J. Lee, "Developing fully functional E-government: A four stage model," *Gov. Inf. Q.*, vol. 18, no. 2, pp. 122–136, 2001, doi: https://doi.org/10.1016/S0740-624X(01)00066-1.

[13] T. Jetzek, M. Avital, and N. Bjørn-Andersen, "Generating Sustainable Value from Open Data in a Sharing Society BT - Creating Value for All Through IT," in *International Working Conference on Transfer and Diffusion of IT*, 2014, pp. 62–82.

[14] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - The story so far," *Int. J. Semant. Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009, doi: 10.4018/jswis.2009081901.

[15] M. Osorio-Sanabria, J. Brito-Carvajal, H. Astudillo, F. Amaya-Fernàndez, and M. Gonzàlez-Zabala, "Evaluating Open Government Data Programs: A Systematic Mapping Study," in *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*, 2020, pp. 157–164, doi: 10.1109/ICEDEG48599.2020.9096755.

[16] C. Alexopoulos, L. Spiliotopoulou, and Y. Charalabidis, "Open Data Movement in Greece: A Case Study on Open Government Data Sources," in *Proceedings of the 17th Panhellenic Conference on Informatics*, 2013, pp. 279–286, doi: 10.1145/2491845.2491876.

[17] A. Stott, "Open data for economic growth," 2014. [Online]. Available: https://www.worldbank.org/content/dam/Worldbank/document/Open-Data-for-Economic-Growth.pdf.

[18] M. S. Altayar, "Motivations for open data adoption: An institutional theory perspective," *Gov. Inf. Q.*, vol. 35, no. 4, pp. 633–643, 2018, doi: 10.1016/j.giq.2018.09.006.

[19] G. Michener and O. Ritter, "COMPARING RESISTANCE TO OPEN DATA PERFORMANCE MEASUREMENT: PUBLIC EDUCATION IN BRAZIL AND THE UK," *Public Adm.*, vol. 95, no. 1, pp. 4–21, 2017, doi: https://doi.org/10.1111/padm.12293.

[20] G. V. Pereira, M. A. Macadar, E. M. Luciano, and M. G. Testa, "Delivering public value through open government data initiatives in a Smart City context," *Inf. Syst. Front.*, vol. 19, no. 2, pp. 213–229, 2017, doi: 10.1007/s10796-016-9673-7.

[21] I. Safarov, A. Meijer, and S. Grimmelikhuijsen, "Utilization of Open Government Data: A Systematic Literature Review of Types, Conditions, Effects and Users," *Info. Pol.*, vol. 22, no. 1, pp. 1–24, Jan. 2017, doi: 10.3233/IP-160012.

[22] F. Welle Donker and B. van Loenen, "How to assess the success of the open data ecosystem?," *Int. J. Digit. Earth*, vol. 10, no. 3, pp. 284–306, Mar. 2017, doi: 10.1080/17538947.2016.1224938.

[23] A. Zuiderwijk and C. C. Hinnant, "Open data policy-making: A review of the state-of-the-art and an emerging research agenda," *Inf. Polity*, vol. 24, pp. 117–129, 2019, doi: 10.3233/IP-190160.

[24] J. P. Tennant, F. Waldner, D. C. Jacques, P. Masuzzo, L. B. Collister, and C. H. J. Hartgerink, "The academic, economic and societal impacts of Open Access: an evidence-based review.," *F1000Research*, vol. 5, pp. 1–39, 2016, doi: 10.12688/f1000research.8460.

[25] S. S. Dawes, L. Vidiasova, and O. Parkhimovich, "Planning and designing open government data programs: An ecosystem approach," *Gov. Inf. Q.*, vol. 33, no. 1, pp. 15–27, 2016, doi: https://doi.org/10.1016/j.giq.2016.01.003.

[26] J. Bates, "The strategic importance of information policy for the contemporary neoliberal state: The case of Open Government Data in the United Kingdom," *Gov. Inf. Q.*, vol. 31, no. 3, pp. 388–395, 2014, doi: https://doi.org/10.1016/j.giq.2014.02.009.

[27] A. Meijer and S. Potjer, "Citizen-generated open data: An explorative analysis of 25 cases," *Gov. Inf. Q.*, vol. 35, no. 4, pp. 613–621, 2018, doi: https://doi.org/10.1016/j.giq.2018.10.004.

[28] A. M. Alsukhayri, M. A. Aslam, I. H. Khan, R. A. Abbasi, and A. Babour, "Toward Building a Linked Open Data Cloud to Predict and Regulate Social Relations in the Saudi Society," *IEEE Access*, vol. 10, pp. 50548–50561, 2022, doi: 10.1109/ACCESS.2022.3174090.

[29] R. E. Sieber and P. A. Johnson, "Civic open data at a crossroads: Dominant models and current challenges," *Gov. Inf. Q.*, vol. 32, no. 3, pp. 308–315, 2015, doi: https://doi.org/10.1016/j.giq.2015.05.003.

[30] S. Saxena, "Drivers and barriers towards re-using open government data (OGD): a case study of open data initiative in Oman," *foresight*, vol. 20, no. 2, pp. 206–218, Jan. 2018, doi: 10.1108/FS-10-2017-0060.

[31] "National Centre for Statistics and Information - Sultanate of Oman. Sultanate of Oman - DATA PORTAL," 2017. https://data.gov.om/.

[32] "Information Technology Authority of Oman (ITA). Open Data - Omanuna Portal," 2022. https://www.oman.om/wps/portal/index/aboutportal.

[33] J. W. Creswell and J. D. Creswell, *Research Design Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. SAGE Publications, Inc, 2018.

[34] A. Bryman and D. Cramer, *Quantitative data analysis with IBM SPSS 17, 18 and 19: A guide for social scientists.* United Kingdom: Routledge-Cavendish/Taylor & Francis Group, 2011.

[35] M. L. Mchugh, "The Chi-square test of independence Lessons in biostatistics," *Biochem. Medica*, vol. 23, no. 2, pp. 143–9, 2013, [Online]. Available: http://dx.doi.org/10.11613/BM.2013.018.

Seeking Higher Performance in Real-Time Data Processing through Complex Event Processing

Guadalupe Ortiz, Adrián Bazan-Muñoz, Pablo Caballero-Torres, Jesús Rosa-Bilbao, Inmaculada Medina-Bulo, Juan Boubeta-Puig
Department of Computer Science and Engineering
UCASE Software Engineering Group, University of Cadiz, Spain
{guadalupe.ortiz, adrian.bazan, pablo.caballero, jesus.rosa, inmaculada.medina, juan.boubeta}@uca.es

Alfonso Garcia-de-Prado
Computer Architecture and Technology Department
UCASE Software Engineering Group, University of Cadiz, Spain
e-mail: alfonso.garciadeprado@uca.es

***Abstract*—Today, data processing has become a key functionality of multiple diverse applications. Large amounts of data from disparate sources must be processed in streaming in order to have real-time knowledge of the domain in question and thus be able to make the most appropriate decisions at each instant of time. This streaming processing has been successfully achieved by introducing Complex Event Processing (CEP) techniques into the solutions provided. Although these solutions have proven their effectiveness in various software architectures and application domains, there is still a need for further research on how to achieve better performance depending on the needs of the application. This paper attempts to shed some light in this area by comparing various configurations of a CEP engine, aiming for better performance in real-time data processing.**

Keywords-Complex Event Processing; Event-driven Service-oriented Architecture; Internet of Things; Data Processing.

I. INTRODUCTION

Today, data processing has become a key functionality of all applications in general and those related to the Internet of Things (IoT) and smart cities, in particular. Large amounts of data are generated from multiple sources at a high speed, which must be processed promptly to have real-time knowledge of the domain in question and thus be able to make the most appropriate decisions at each instant of time. In this context, multiple applications and architectures emerge that address big, small and open data processing, for decision making in various domains, with special emphasis on IoT and smart cities [1].

According to Rahmani et al. [2], Complex Event Processing (CEP) has become a key part of the IoT; indeed multiple publications endorse CEP as a successful technology for streaming data processing in the IoT [3]–[6], including a wide variety of works, in diverse application domains. This integration of CEP with the IoT not only takes place in the cloud, but also at levels closer to the device, such as the fog or the edge [7]. Although when we need to integrate multiple communication protocols and application technologies the use of an Enterprise Service Bus (ESB) in an event-driven service-oriented application facilitates the implementation and maintenance of the architecture [8][9]; in production environments where integration needs are lower, lighter and more efficient architectures can be achieved without using the ESB [10][11]. An architecture that integrates the CEP engine without the ESB can face with greater guarantee of success scenarios that demand higher performance, especially in the current situation where the amount and velocity of data is growing at a vertiginous rate year after year.

For all the above, we need to analyze which configurations of CEP engines can provide us with better performance in the most common scenarios of big data processing in IoT and/or smart cities; where many of the implementations are or could be limited to the integration of data sources through an inbound messaging broker with a data processing engine and an output also channeled through an outbound messaging broker. For performance analysis it is necessary to adjust to a particular implementation and given the wide use of Esper, this is going to be our CEP engine. On the other hand, given the widespread use of RabbitMQ and the immediate integration of AMQP 0.91; these are going to be the broker and protocol for both inbound and outbound messaging used in this research.

As discussed in Section III, in the past several studies on performance for CEP engines were done, such as [14][15] and [16], but we could not find particularly a comparison of 2 opposite mechanism of Esper engine to subscribe to complex events: subscriber and listener, nor the comparison of configuring CEP engines to execute with different number of threads. In this sense, this paper focused on doing the tests needed to analyze such options to check which can provide us with better performance and therefore to complement other existing research on CEP performance analysis.

The rest of the paper is organized as follows. Section II introduces CEP technology. Then, Section III explains the related work and motivates the need for further CEP testing and evaluation. Afterwards, the evaluation scenario proposed as well as the configurations of the test performed are presented in Section IV. Consequently, Section V explains the

 ISBN: 978-1-68558-041-4

results obtained from the tests performed and, finally, Section VI presents the conclusions.

II. Background on Complex Event Processing

CEP [12] is a technology by which we can capture, analyze and correlate in real time huge amounts of data, coming from different application domains and in different formats, to detect relevant situations as they occur [13]. The incoming data to be processed by the system are called *simple events*, while the detected situations are called *complex events*.

To detect these complex events, it is necessary to have previously defined an event pattern that will be responsible for analyzing and correlating one or several simple events in a given period of time. These patterns must be deployed in a CEP engine, i.e., the software in charge of capturing the simple events, analyzing in real time if some of the patterns deployed on the simple input event stream are fulfilled, and creating the complex events.

In this work, we have adopted the Esper CEP engine and its EPL pattern language, because of its recognized prestige in terms of performance and applicability.

III. Related Work and Motivation

We have found several works which provide CEP performance evaluation. For instance, Rosa et al. [14] present a comparative study of several Esper engines for security event management. Esper CEP engine is among the engines evaluated; in their analysis we can see that Esper engine has a very good performance with a high throughput and the authors consider it to be the most suitable taking into account performance and configuration flexibility. We have also found a comparison of the Esper engine with the Sidhi CEP engine [15], in both cases integrated with an ESB and the Mosquito broker [16]. Ortiz et al. also evaluate the time it takes to transfer events in a microservice-based architecture and to process them in the Esper CEP engine [10]. Besides, Corral et al. evaluate how the integration of Esper with Kafka behaves with up to 32 partitions [17] demonstrating that the system is highly scalable under these simple conditions, but not evaluation on the CEP engine isolated, which is our main objective in this paper. Also in [11] an evaluation and comparison of Esper CEP engine in an event-driven architecture with the use of an ESB compared to the use of Data-Flows is provided, which might be complementary to the research done in this paper.

Thus, we can conclude that, to our knowledge, there is no work comparing some particular configurations of Esper CEP engine, such as the use of subscriber and listener in the engine, nor the use of several threads in its execution configuration. Such gap motivated this work which can help us to better understand Esper CEP performance and compliment other existing related works. Particularly, we expect to deploy the architecture evaluated in this paper in a water management company and we need to check which is the most efficient solution for this purpose beforehand.

IV. Evaluation Scenario

This section explains the software architecture used for the performance tests and the machines involved in it, the key performance indicators selected to be measured from the tests and the configuration prepared for the tests.

A. Architecture

The software architecture, as represented in Figure 1, consists in a synthetic data simulator (nITROGEN [18]), which submits data to a RabbitMQ broker; both deployed in *Machine 1*. The CEP application in *Machine 2* is then subscribed to the queue in the RabbitMQ broker to receive the simple events. After the simple events are processed by the CEP engine, the detected complex events are sent to an output RabbitMQ queue in *Machine 3*. The three are server machines with an Intel Xeon Silver 4110 processor and 32 GB of RAM.

B. Key Performance Indicators

To analyze in detail the processing times in each component of the architecture, we have added a series of timestamps along the life of the processed message, from its generation to the end of its processing, as explained in the following lines and shown in Figure 1.

- Let t1 be the timestamp corresponding to when the synthetic data is generated in the simulator; in this case we have used nITROGEN simulator [18].
- Let t2 be the timestamp corresponding to when the simple event (the generated synthetic data) is going to enter the CEP engine; that is, it has already been sent from the simulator to the broker and from the broker to the CEP engine.
- Let t3 be the timestamp that adds Esper CEP to the message when the complex event is detected.
- Let t4 be the timestamp corresponding to the time when the complex event leaves the CEP engine and is sent to the output queue.

Thus, the difference of t_2-t_1 indicates the time it takes for the simple event to be sent from the simulator to the messaging broker and from this to the CEP engine; that is, the sum of the sending time and the processing time in the broker. From now on we will call T_{subm} as this time difference.

On the other hand, t_3-t_2 is the time difference from the reception of the simple event in the CEP engine until the detection of the complex event in the CEP engine, i.e., the processing time of the event in the CEP engine, hereafter t_{proc}.

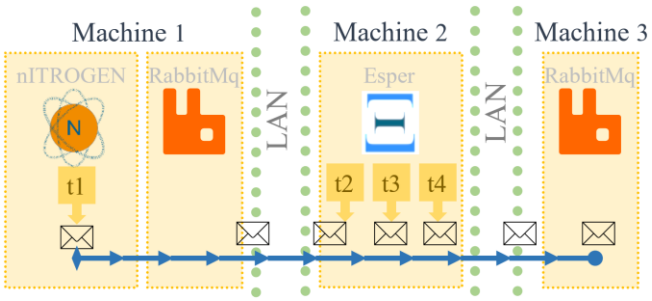


Figure 1. Software Architecture and timestamps taken.

Finally, t_4-t_2 gives us the time difference from when the simple event is going to enter the CEP engine until the complex event leaves the CEP engine to be sent to the output queue; that is, it includes not only the processing time in CEP of the simple event, but also the management of the complex event in the CEP engine. From now on this time will be called T_{man}.

Such three times (T_{subm}, T_{proc}, T_{man}) together with the CPU usage and memory consumption will be the key performance indicators in our evaluation tests.

C. Test Configuration

The objective of these tests is not to evaluate several instructions of the EPL syntax to build patterns, as we did in the past [11], but to evaluate several ways to handle the complex event and configure the CEP engine to process a simple pattern and check specifically if the use of the *listener* and the *subscriber* clauses in a pattern, as well as the configuration of the CEP engine execution using 1 or 10 threads in the processing, influence the performance of the whole architecture.

Let us explain that a *listener* can subscribe to complex events already posted by a pattern match. However, a *subscriber* object receives statement results via method invocation. A subscriber is expected to have performance advantages, but we will have to see if this holds true in the performance tests. We used a very simple event scheme and event pattern for the evaluation with the aim of insolating the behavior of the listener and the subscriber and make it independent of any pattern clause.

The simple events reaching the CEP engine will consist on a JSON element containing i) the timestamp of the instant of creation of the event (*t1*); ii) the timestamp of when it reaches the CEP engine (*t2*), which will be generated empty by default and added when such data is known and iii) a boolean — *shouldTrigger*—that will cause the pattern to be met randomly or not for each incoming event; which is represented as follows in the Esper CEP engine:

```
@public @buseventtype create json schema Dummy as (t1 long, t2 long, shouldTrigger boolean)
```

The pattern will simply add the timestamp of the instant in which the complex event is detected (*t3*) and select the other timestamps that were already in the simple event (*t1* and *t2*).

```
insert into DummyComplexEvent SELECT current_timestamp as t3, t2, t1 FROM Dummy(shouldTrigger)
```

As previously mentioned, the reason for using such a simple event pattern is because we want to focus on the different behavior of the system using the subscriber and listener, as well as executing with one or more threads. It is not our aim to evaluate a wide range of Esper operators as this was done by other works, but to complement such works with this novel tests.

Every test was run for simple events incoming rates of 1 000, 10 000 and up to 50 000 incoming events per second, and each test was run for 10 minutes. As previously said, T_{subm}, T_{proc}, T_{man}, CPU usage and memory usage were measured for every performed test.

V. Results and Discussion

In this section we show and analyze the results of the tests performed both with subscriber and listener and for 1 and 10 threads for the CEP engine execution under the conditions described in Section III.

As we can see in Table I and Table II, the use of one thread, either with listener or with subscriber, seems to have an average consumption of memory and CPU quite similar for any of the tested incoming rates. Also, the submission times from the message queue to the CEP engine are quite similar, as they should be. We can also see that the processing time is also almost the same in both cases, but we note some differences in the management time: even though the subscriber seems to be more efficient than the listener when we have an input rate of 1 000 events per second (0.264 ms the listener versus 0.02 ms of the subscriber) when we reach the input rate of 50 000 events per second, the listener is the one being more efficient (0.007 ms the listener versus 0.33 ms of the subscriber). It is important to point out that the high values reached for T*man* with the input rate of 50 000 events per second are due to the fact that the system collapses and therefore does not process all the messages properly and may give inconsistent values.

Again, as we can see in Table III and Table IV, average consumption of memory and CPU are also quite similar when using ten threads with any of the tested incoming rates and the submission times from the message queue to the CEP engine are quite similar, as well. In this occasion we can see that the processing time is again similar for both the subscriber and listener options. This time, the management time for the listener behaves better (0.14 versus 0.69 ms) at a 10 000 events per second incoming rate, as well as the rate of 50 000 incoming events per second (0.78 ms of the listener versus 0.99 ms of the subscriber).

TABLE I. Test Results with Listener Configuration and 1 Thread.

Incoming Rate (events/s)	Medium Memory Usage (MB)	Medium CPU Usage (%)	T_{subm} (ms)	T_{proc} (ms)	T_{man} (ms)
1 000	466.5	0.37	10.77	0.007	0.264
10 000	518.2	1.35	58.05	0.005	0.58
50 000	520.3	3.38	4 853	0.0034	0.07

TABLE II. Test Results with Subscriber Configuration and 1 Thread.

Incoming Rate (events/s)	Medium Memory Usage (MB)	Medium CPU Usage (%)	T_{subm} (ms)	T_{proc} (ms)	T_{man} (ms)
1 000	463.3	0.34	10.78	0.007	0.02
10 000	535.8	1.35	63.70	0.005	0.86
50 000	524.4	3.53	4 387	0.0038	0.33

 ISBN: 978-1-68558-041-4

TABLE III. TEST RESULTS WITH LISTENER CONFIGURATION AND 10 THREADS.

Incoming Rate (events/s)	Medium Memory Usage (MB)	Medium CPU Usage (%)	T_{subm} (ms)	T_{proc} (ms)	T_{man} (ms)
1 000	485.8	0.67	9.30	0.89	0.86
10 000	517.3	3.62	58.26	0.14	7.66
50 000	7 884.8	20.51	257.18	0.78	85 114.69

TABLE IV. TEST RESULTS WITH SUBSCRIBER CONFIGURATION AND 10 THREADS.

Incoming Rate (events/s)	Medium Memory Usage (MB)	Medium CPU Usage (%)	T_{subm} (ms)	T_{proc} (ms)	T_{man} (ms)
1 000	483.7	0.65	7.90	0.7	0.9
10 000	516.1	2.86	58.36	0.69	4.63
50 000	8 089.6	19.71	264.31	0.99	87 522

Up to this point of the comparison we can say that for simple events there are no big differences between using a listener or a subscriber because although there are some differences at some rates of incoming events per second, they are not significant, not reaching the millisecond.

There are differences between the use of 1 or 10 threads in the execution of the CEP engine, although perhaps not the expected ones. To better observe these differences, we have represented in Figure 2 three graphs with the values taken by T_{subm}, T_{proc} and T_{man}, respectively, for each input rate with the listener and the subscriber and the execution in 1 thread; and these same three graphs but using 10 threads for the execution in Figure 3.

For the time of submission (t_{subm}) we do not appreciate big differences (as expected). However, for the time of processing in the CEP engine (t_{proc}), when using a single thread, the processing time increases as the input rate of simple events increases; however, the processing time decreases when using 10 threads (until it collapses at 50 incoming events per second). On the other hand, if we take as a reference the input rate 10 000 events per second, in which the engine is not collapsed but it is not as fluid as with 1 000 input events, we see that we obtain better times with 1 thread than with 10; possibly due to the greater management involved in the distribution of tasks among the threads and the resolution of the final results. Finally, the processing and management time (t_{man}) increases in both cases as we increase the input rate of simple events, until it saturates at 50000 input events per second; but it remains in any case lower for the execution with 1 thread compared to the one using 10 threads.

VI. THREADS TO VALIDITY

A limited number of tests have been performed in this work. As previously mentioned, a single pattern has been tested, but to better validate the results, perhaps a varied set of operators or domain specific patterns could be tested. On the other hand, generating a greater or lesser number of complex events for each simple input event may yield other results. It should also be noted that the use of 1 thread has been compared with the use of 10 threads, but other intermediate options such as 2, 3, 4, etc. threads have not been tested. Tests with different numbers of threads could lead to other conclusions in addition to those explained in this paper.

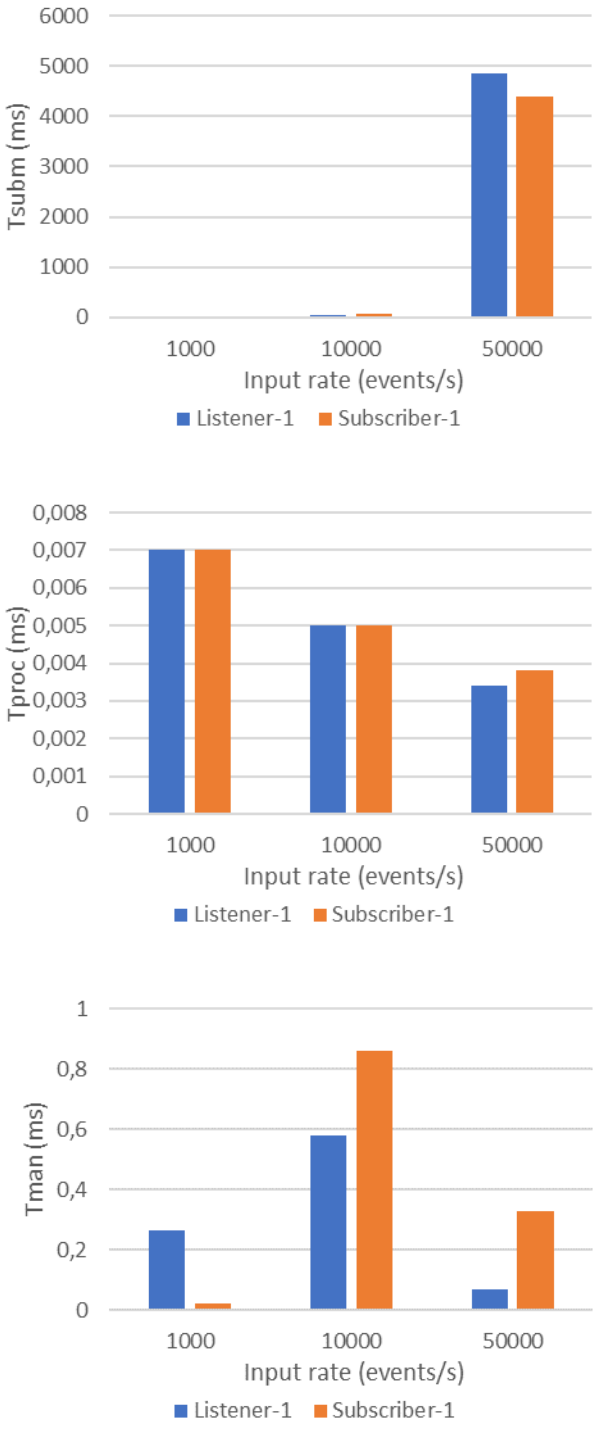


Figure 2. T_{subm}, T_{proc} and T_{man} for each input rate with the listener and the subscriber and the execution in 1 thread.

On the other hand, it is important to bear in mind that in real systems, when measuring processing time, a lack of synchronisation in the clocks of the systems involved may imply a mismatch in the measurement of these times. However, when implementing a real system, our main goal will not be to measure performance time, but rather for the patterns to detect the situations of interest in the domain in question. In this case, clock times have no influence, since the events are processed the instant they arrive, regardless of the timestamp they may contain, and it is the CEP engine that assigns the timestamps necessary for the internal management of the times and time windows.

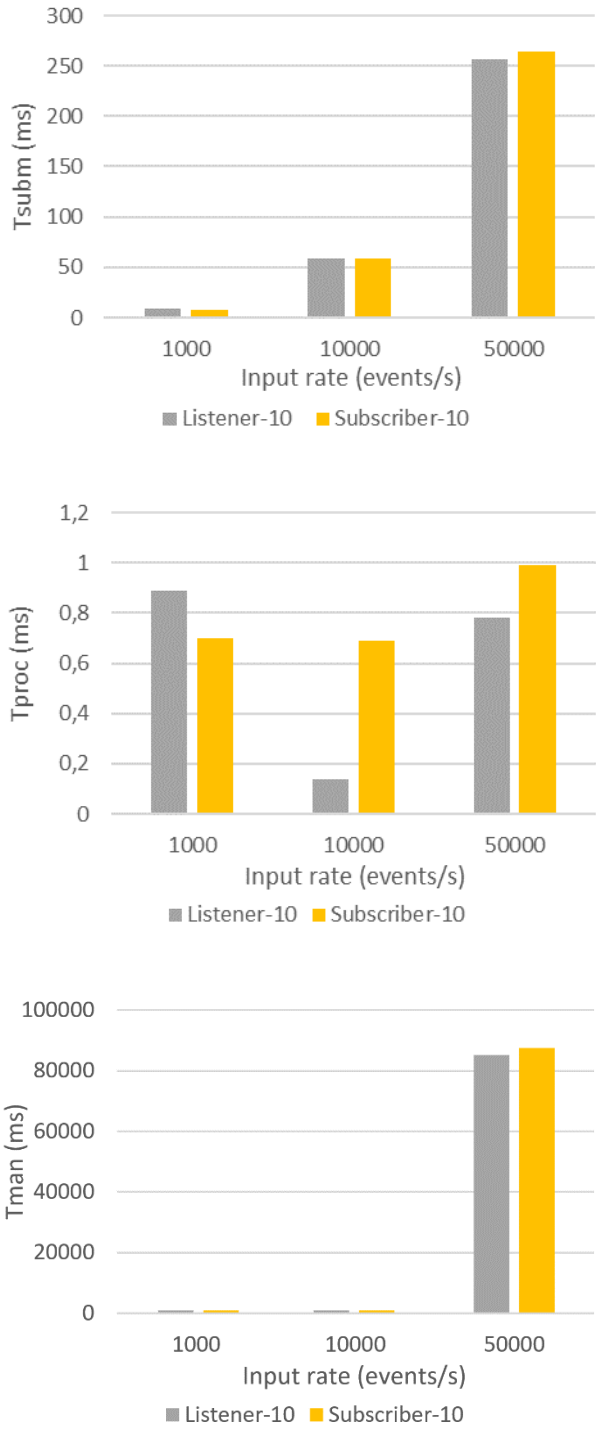


Figure 3. Tsubm, Tproc and Tman for each input rate with the listener and the subscriber and the execution in 1 thread.

VII. Conclusion and Future Work

In light of the results of the tests performed, we can conclude that both the use of subscriber and listener to capture the complex events detected by the CEP engine provide similar behaviour at different rates of incoming events per second. We can also conclude that configuring the CEP engine to use more threads might not be useful when we have large amounts of incoming events as it is more time consuming to distribute and assign the tasks for the different threads than the time required to do it in a single thread.

For future work, we expect to perform further performance tests with the particular patterns developed for the water management company where we will test the architecture evaluated in this paper.

Acknowledgment

This work was partly supported by grant PDC2022-133522-I00 (ASSENTER project) funded by MCIN/AEI /10.13039/501100011033 and by the" European Union Next GenerationEU/ PRTR" and partly by the grant program for R&D&i projects, for universities and public research entities qualified as agents of the Andalusian Knowledge System, within the scope of the Andalusian Plan for Research, Development and Innovation (PAIDI 2020). Project 80% co-financed by the European Union, within the framework of the Andalusia ERDF Operational Program 2014-2020 “Smart growth: an economy based on knowledge and innovation”. Project funded by the Ministry of Economic Transformation, Industry, Knowledge and Universities of the Andalusian Regional Government. DECISION project with reference P20_00865. We are also grateful for the collaboration of the water supply network management company GEN.

References

[1] S. D. Liang, "Smart and Fast Data Processing for Deep Learning in Internet of Things: Less is More", *IEEE Internet Things J.*, vol. 6, no. 4, pp. 5981–5989, Aug. 2019, doi: 10.1109/JIOT.2018.2864579.

[2] A. M. Rahmani, Z. Babaei, and A. Souri, "Event-driven IoT architecture for data analysis of reliable healthcare application using complex event processing", *Clust. Comput.*, vol. 24, no. 2, pp. 1347–1360, Jun. 2021, doi: 10.1007/s10586-020-03189-w.

[3] A. Akbar, A. Khan, F. Carrez, and K. Moessner, "Predictive Analytics for Complex IoT Data Streams", *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1571–1582, Oct. 2017, doi: 10.1109/JIOT.2017.2712672.

[4] R. Mayer, B. Koldehofe, and K. Rothermel, "Predictable Low-Latency Event Detection With Parallel Complex Event Processing", *IEEE Internet Things J.*, vol. 2, no. 4, pp. 274–286, Aug. 2015, doi: 10.1109/JIOT.2015.2397316.

 ISBN: 978-1-68558-041-4

[5] A. Garcia-de-Prado, G. Ortiz, and J. Boubeta-Puig, "CARED-SOA: A Context-Aware Event-Driven Service-Oriented Architecture", *IEEE Access*, vol. 5, pp. 4646–4663, 2017, doi: 10.1109/ACCESS.2017.2679338.

[6] A. Garcia-de-Prado, G. Ortiz, and J. Boubeta-Puig, "COLLECT: COLLaborativE ConText-aware service oriented architecture for intelligent decision-making in the Internet of Things", *Expert Syst. Appl.*, vol. 85, pp. 231–248, Nov. 2017, doi: 10.1016/j.eswa.2017.05.034.

[7] G. Mondragón-Ruiz, A. Tenorio-Trigoso, M. Castillo-Cara, B. Caminero, and C. Carrión, "An experimental study of fog and cloud computing in CEP-based Real-Time IoT applications", *J. Cloud Comput.*, vol. 10, no. 1, p. 32, Dec. 2021, doi: 10.1186/s13677-021-00245-7.

[8] H. Derhamy, J. Eliasson, and J. Delsing, "IoT Interoperability—On-Demand and Low Latency Transparent Multiprotocol Translator", *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1754–1763, Oct. 2017, doi: 10.1109/JIOT.2017.2697718.

[9] A. Massaro *et al.*, "Production Optimization Monitoring System Implementing Artificial Intelligence and Big Data", in *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, Roma, Italy, Jun. 2020, pp. 570–575. doi: 10.1109/MetroInd4.0IoT48571.2020.9138198.

[10] G. Ortiz *et al.*, 'A microservice architecture for real-time IoT data processing: A reusable Web of things approach for smart ports', *Comput. Stand. Interfaces*, vol. 81, p. 103604, Apr. 2022, doi: 10.1016/j.csi.2021.103604.

[11] G. Ortiz, I. Castillo, A. Garcia-de-Prado, and J. Boubeta-Puig, "Evaluating a Flow-based Programming Approach as an Alternative for Developing CEP Applications in the IoT", *IEEE Internet Things J.*, vol. 9, no. 13, pp. 11489–11499, 2021, doi: 10.1109/JIOT.2021.3130498.

[12] D. C. Luckham, *Event processing for business: organizing the real-time enterprise*. Hoboken, N.J, USA: John Wiley & Sons, 2012.

[13] C. Inzinger, W. Hummer, B. Satzger, P. Leitner, and S. Dustdar, "Generic event-based monitoring and adaptation methodology for heterogeneous distributed systems: event-based monitoring and adaptation for distributed systems", *Softw. Pract. Exp.*, vol. 44, no. 7, pp. 805–822, Jul. 2014, doi: 10.1002/spe.2254.

[14] L. Rosa, P. G. Alves, T. J. Cruz, and P. Simoes, "A Comparative Study of Correlation Engines for Security Event Management", presented at the Int. Conf. on Cyber Warfare and Security, Kruger National Park, South Africa., 2015.

[15] WSO2, "Siddhi", 2022. http://siddhi.io/ (accessed Mar. 20, 2023).

[16] J. Roldán, J. Boubeta-Puig, J. Luis Martínez, and G. Ortiz, "Integrating complex event processing and machine learning: An intelligent architecture for detecting IoT security attacks", *Expert Syst. Appl.*, vol. 149, p. 113251, Jul. 2020, doi: 10.1016/j.eswa.2020.113251.

[17] D. Corral-Plaza, G. Ortiz, I. Medina-Bulo, and J. Boubeta-Puig, "MEdit4CEP-SP: A model-driven solution to improve decision-making through user-friendly management and real-time processing of heterogeneous data streams", *Knowl.-Based Syst.*, vol. 213, p. 106682, Feb. 2021, doi: 10.1016/j.knosys.2020.106682.

[18] A. Garcia-de-Prado, "nITROGEN: Internet of Things RandOm GENreator", 2020. https://ucase.uca.es/nITROGEN/ (accessed Mar. 20, 2023).

 ISBN: 978-1-68558-041-4

Small Dataset Acquisition for Machine Learning Analysis of Industrial Processes with Possible Uncertainties

Xukuan Xu
Technische Hochschule Aschaffenburg
Aschaffenburg, Germany
e-mail: xukuan.xu@th-ab.de

Felix Conrad
Technische Universität Dresden
Dresden, Germany
e-mail: felix.conrad@tu-dresden.de

Andreas Gronbach
Fraunhofer-ISC
Würzburg, Germany
e-mail: andreas.gronbach@isc.fraunhofer.de

Michael Möckel
Technische Hochschule Aschaffenburg
Aschaffenburg, Germany
e-mail: michael.moeckel@th-ab.de

***Abstract*—As the algorithms mature, the bottleneck in applying Machine Learning (ML) to process analysis, monitoring and control is often caused by the availability of suitable data and the cost of data acquisition. For many ML projects, datasets have been collected independently of subsequent analysis. In industrial production, data acquisition and coverage of possible process uncertainties pose challenges to the preparation of suitable datasets. This article discusses dataset generation for ML from scratch under the constraint of limited resources with process uncertainties. A new approach towards an adapted Design Of Experiments (DOE) is proposed with the aim of sampling data more efficiently. In this way, we contribute to the challenge of preparing datasets for ML applications.**

Keywords-Small-data; Process uncertainty; Design Of Experiments(DOE); Machine learning.

I. INTRODUCTION

ML makes it possible to efficiently excavate valuable information from data with its powerful data analysis capabilities. With the prosperous advancement of algorithm research, model building is no longer a challenge limiting ML applications [1]. In fact, according to a survey from Crowdflower in 2016 [2], the efforts of data scientists are mainly (60%) consumed by data organizing and data cleaning. After this, 19% of the time is spent collecting datasets. This shows that data preparation is the bottleneck of ML applications in the current stage. However, this difficulty is often overlooked by the informatics community. In most cases, the datasets are unthinkingly pre-existing. With this standpoint, they simply optimate the algorithm at the software side for data analysis. However, the dataset's quality determines the upper limit of data analysis. Therefore, in some cases, it may be unfeasible to look at a solution only from the ML model side.

It is both a challenge and an advantage to look at data preparation from the perspective of a production engineer. Collecting a single element of the dataset requires that a product is physically produced and the relevant data is measured during the manufacturing process. In practice, an extra number of products is required to account for deficient outcomes. This limits the amount of usable data for ML analysis. The overall amount of data is often constrained by cost considerations. However, pre-existing knowledge, experience or even intuition of the process often allows an engineer to focus the data generation on particularly relevant subsets of an overly complex parameter space.

Purpose-built datasets for ML modeling may address two possible directions [3]:

I. Finding the control variables and their optimal values that give rise to an optimal response
II. Exploring the neighborhood around the optimal values to generate knowledge for monitoring, anomaly detection and control

We investigate the latter under the constraint of limited resources (e.g., time, budget) for data acquisition and fixed overall statistical process uncertainty. Based on the data obtained from the Lithium Ion Battery (LIB) production line in the KIproBatt project [4], we describe the practical difficulties in preparing datasets for industrial production in Section 2. In Section 3, existing DOE approaches are described. A set of experimental design schemes suitable for ML modeling is proposed. In addition, we propose a new Small-Data DOE (SD-DOE) suitable for ML modeling with process uncertainty.

II. DESCRIPTION OF SMALL-DATA CONTEXT

A. Small data problem

Small-batch production is often unavoidable in laboratory research, on a pilot production stage prior to upscaling, or in customer-specific (individualized) manufacturing [5]. Often, data acquisition is limited by budget or time constraints to datasets with less than one thousand elements. The particular choice of selected data points affects the outcomes of subsequent analysis. For illustration, we consider the project KIproBatt as an example of a typical small-scale data generation: a total of ca. 500 Li-ion battery cells is to be produced with a semi-automatic production line in a laboratory environment. Research questions include the imp-

 ISBN: 978-1-68558-041-4

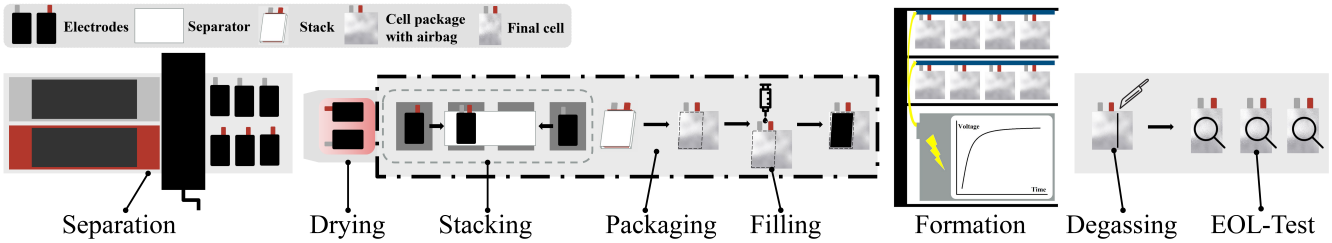


Figure 1. LIB cell assembly process from separation to EOL-test

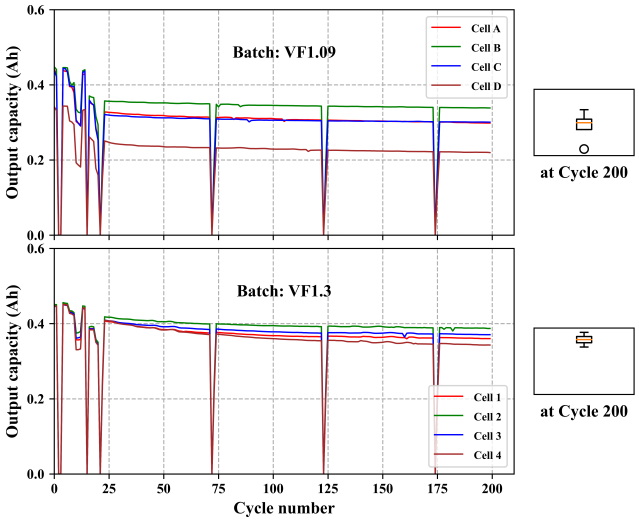


Figure 2. Cell output capacity related to cycle number in a cycling test

act of process deviations on the quality of final cells as well as the exploration of complex correlations among process parameters. Note that one cannot define the "small-data problem" by sole reference to a fixed amount of data. Instead, the characteristics and complexity both of the research objectives and the applied ML methods have to be considered.

B. *Lack of process knowledge & complexity of the production process*

The number of required data depends on the complexity of the process. A large number of features, non-linear relationships and interactions between features increase the complexity of the process and thus the number of data points required. These conditions are often found in industrial production processes [6]. The assembly process of a LIB pouch cell is an example of such a complex process and is depicted in Figure 1: cell assembly starts with electrode separation. Then, the anodes and cathodes are dried and fed into a glove box with a controlled atmosphere. Next, a stacking machine assembles the electrodes with a separator into cell stacks (Z-fold stacking). After the packaging, sealing and electrolyte filling, the cell is activated by the first charge and discharge (formation). The gas generated in this procedure is removed and the cell is finally sealed.

The complexity of this multi-step process leads to manifold variable interdependencies. Hence, an effective analysis should be based on an ML approach. However, it is challenged by limited data, which may lead to undersampling of the parameter space and a lack of convergence of the ML models. We define this as the fundamental characteristic of small-data context.

C. *Process uncertainty*

Complex processes are normally investigated for a limited set of process parameters only. While the remaining parameters are, in theory, assumed to remain constant, their unavoidable fluctuations contribute to statistical uncertainty in all measured data. Other sources for uncertainties lie, for instance, in the measurement uncertainties of the used sensors. This uncertainty is manifested in the data as identical input parameters will lead to a statistical spreading in the target responses.

In the KIproBatt project, using the injected electrolyte volume as the only tunable factor with two levels, we produced four cells at each level while ensuring that the rest of the process parameters were consistent. Each cell was then tested according to the same cycling protocol to evaluate its performance. The cycling protocol also includes non-cycling tests such as pulse, c-rate, and quick charge tests. As reflected in Figure 2, the troughs that occur every 50 cycles indicate the pulse test. The results, using Output Capacity (OC) as an indicator, are shown in Figure 2. It can be seen that the performance of the battery cells within each batch varies. As the box plot illustrates, the process uncertainty is so evident in batch VF1.09 that cell D is judged to be an outlier (box plot).

The reasons for this might be processing errors due to human operations, a lack of process understanding that leaves some potential variables uncontrolled, or measurement errors in the hardware. However, in the end, what emerges is the uncertainty of the OC.

No direct conclusion can be derived when the process uncertainty exceeds the variation imposed on control variables.

Usually, uncertainty reduction could be achieved either by optimizing hardware or by repeated measurement and averaging. However, for fixed measurement capacity, the latter implies a reduced ability for parameter space exploration. Therefore, DOE strategies can be developed further to find new compromises between resource allocation for uncertainty reduction and for parameter space sampling.

III. DOE STRATEGY

A. *Existing DOE strategies*

DOE is an established approach to systematically collect

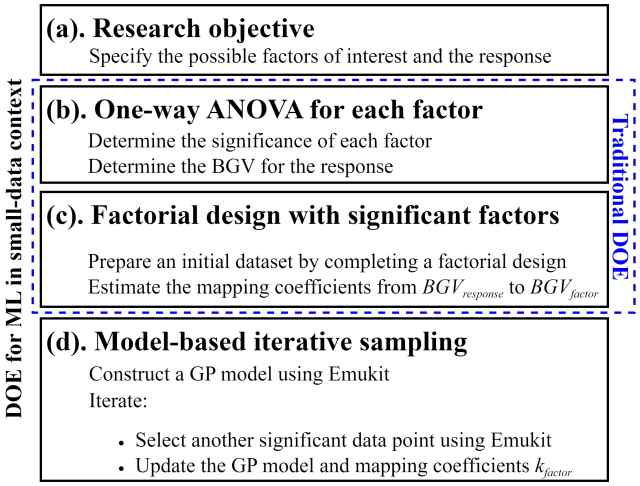


Figure 3. Proposed DOE workflow in small-data context

TABLE I. ANOVA: OC VERSUS EV

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Electrolyte	1	0.011542	0.011542	8.16	0.029
Error	6	0.008482	0.001414		
Total	7	0.020024			

information about a system or process. It aims at delivering the most relevant experimental data for addressing a given research objective. The origin of classical DOE can be traced back to the Analysis Of Variance (ANOVA) proposed by FISHER in the 1920s [7]. Traditional DOE has a set of proven paradigms: screening design for identifying relevant parameters and response surface design for detailed investigation of optimal parameter configurations. With the development of data science and easier access to data, ML tools have been successfully applied to many data analysis problems. ML has unparalleled efficiency advantages in analyzing big data (compared to the volume of data in traditional DOE) with complex interdependencies.

However, little attention has been paid to the interplay of data set generation and ML-based data analysis. A series of studies have conducted the generation of datasets for ML based on traditional DOEs in the past five years [8][9]. In addition, motivated by some ML algorithm developments, iterative data acquisition schemes have been discussed.

Emukit [9] provides such a model-based iterative DOE scheme within a Bayesian optimization framework. The Emukit DOE tool starts from a set of given initial data points and iterates the following three steps to generate sample points in a given input space:

- fit a prediction model to the existing data
- find the next point with the highest marginal predictive variance as predicted by the prediction model
- add this new data point to the existing dataset

Such iteration allows for the most efficient allocation of a limited number of data points based on certain metrics, such as marginal predictive variance of the model. This model-based scheme works well with ML data analysis since a prediction model (e.g., gaussian process model, GP model) is used to predict the target response and calculate the variance during each iteration of data acquisition.

The Emukit approach has shown excellent performance in relevant papers and simulation experiments [10][11] but would profit from further practical validation. In addition, uncertainties are not taken into account for the collected data points. Therefore, we use the Emukit method within the framework of traditional DOE and extend its iterative sampling strategy to account for data uncertainties. The resulting approach is particularly suited for the small-data context with comparatively large uncertainties.

B. Iterative DOE for small-data context

The proposed workflow in a small-data context is shown in Figure 3. We first present the first two steps (a) and (b):

First, factors of interest and their ranges are specified w.r.t. the research objective. In the second step, the range of each factor can be divided into at least two levels. Then, we perform a one-way ANOVA for each factor. ANOVA is performed on adjacent pairs with comparable variance to evaluate each pair's significance. Depending on the upper limit of the data volume, at least two replicate trials at each level are required to determine the significance of the factors (p-value). A level of significance is fixed, e.g., $p_0 = 0.05$ or $p_0 = 0.1$. If $p > p_0$, the considered factor is not significant within the interval defined by the adjacent levels.

We illustrate the subsequent DOE procedure for the example case of process analysis in battery cell production, as performed by the project KIproBatt.

A large number of factors may influence the battery cell performance. Initially, we determine the Electrolyte Volume (EV) as the only varying factor of interest and specify its range between 1.09 gram and 1.3 gram. We have produced 4 battery cells (data points) at each considered level (EV = 1.09 gram, 1.3 gram). For simplicity, we only illustrate the use of analysis of variance for this single factor. We specify the response as the lithium battery cell's OC at cycle 200. We perform ANOVA and obtain the following results (cf Table I).

For a given level of significance of $p_0 = 0.05$, we identify the electrolyte volume as a significant factor for the response ($0.029 < 0.05$) in this interval. Adjusted Mean Squares (Adj MS) are calculated by dividing the adjusted sum of squares by the Degrees Of Freedom (DF). From the Adjusted Mean Square Error (Adj MSE), we obtain the within variance as an estimate for the data uncertainties Δ_{OC}. Root mean square error allows this estimate to have the same units as the response. Thus, we have:

$$Adj\ MSE = {\Delta_{OC}}^2 = 0.001414 \quad (1)$$

With the existing within variance in this example, our requirement for significance can be relaxed until $p = p_0$. Assuming such an extreme $p = 0.05$, we can calculate the minimum required Between-group Variance (BGV) of the factor on the response such that the factor can still be determined as a significant factor. The corresponding F value can be taken from the tabulated F-distribution with group = 2, number of observations = 8 ($F\ value_{2-1,8-2,\alpha=0.05}$).

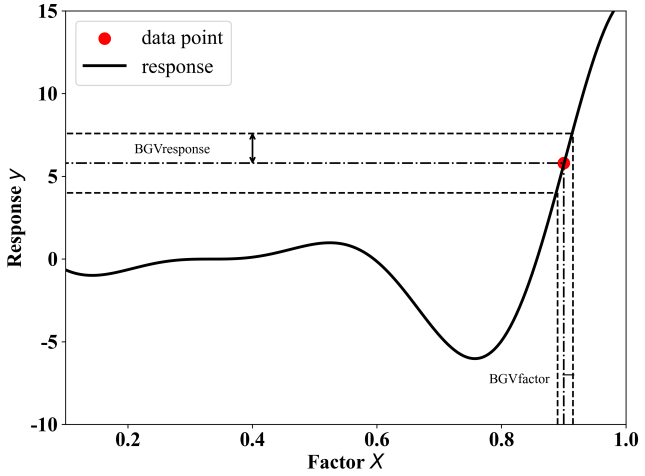


Figure 4. $\boldsymbol{BGV_{factor}}$ defines the regions unsuitable for sampling

$$BGV_{response,min} = \sqrt{F\ value_{1,6,a=0.05} * Adj\ MSE} \quad (2)$$

With the existing Δ_{OC} the critical choice of factor levels in a significance test is the limiting choice for level setting in data generation. Assuming that the levels we set for the EV are too close to each other, then the statistical spreading due to within variance may limit the distinguishability between neighboring levels. If this principle is applied to select the next data point, it can be determined whether this data point represents an additional significant level for the considered factor.

So far, we have only identified an $BGV_{response,min}$. We still need to map this $BGV_{response,min}$ to the corresponding factor $BGV_{factor,min}$, e.g., the electrolyte volume. This will be addressed in step (c) in Figure 3.

The response depends on multiple factors. For the battery cell production, next to the electrolyte volume, experts believe [6] that Drying Time (DT), Wetting Time (WT) after filling, Coating Defects (CD) on electrodes, and Stacking Accuracy (SA) also have considerable impact on the output capacity. A preliminary predictive model for the response $OC_{cycle\ 200}(X_{EV}, X_{DT}, X_{WT}, X_{CD}, X_{SA})$ can be built with the data collected in this factorial design. This model allows calculating the derivative of the response w.r.t each factor

$$k_{factor} = \frac{\partial OC_{cycle\ 200}}{\partial X_{factor}} \quad (3)$$

for local inversion. By using a simple multilinear regression model, e.g.:

$$OC_{cycle\ 200} = \sum_{i=1}^{5} k_i X_i + b \quad (4)$$

the coefficients k_i for each factor X_i in (4) are the mapping coefficients to linear order. Thereby, we can map the $BGV_{response,min}$ (on the responses) to the $BGV_{factor,min}$ (on the factors) and thus determine the minimum required between-group variance $BGV_{factor,min}$ for each factor.

$$BGV_{factor,min} = BGV_{response,min}/k_{factor} \quad (5)$$

As reflected in Figure 4, the $BGV_{factor,min}$ defines an environment around each factor value where no significant data points can be chosen. Each new data point will be used to update the model and the mapping coefficients to determine a more accurate estimate for $BGV_{factor,min}$. Under this framework, we can proceed the iterative sampling in step (d) until all data points for a machine learning dataset have been collected.

IV. CONCLUSION

This article discussed the characteristics of small data problems with process uncertainties. A new approach towards an adapted DOE is proposed with the aim of sampling data more efficiently under such circumstances. This DOE approach is applied to the battery cell production for the project KIprobatt and we are looking forward to presenting our following results.

REFERENCES

[1] F. Conrad, M. Mälzer, M. Schwarzenberger, H. Wiemer, and S. Ihlenfeldt, "Benchmarking AutoML for regression tasks on small tabular data in materials design", Sci Rep, vol. 12, no. 1, Art. no. 1, pp. 19350, Nov. 2022, doi: 10.1038/s41598-022-23327-1.

[2] Figure Eight. *CrowdFlower: Data science report*. [Online]. Available from: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf [retrieved: 02, 2023].

[3] A. Dean, D. Voss and D. Draguljić, Design and Analysis of Expermients, 2nd Edition. New York, NY: Springer, 2017.

[4] X. Xu et al., *KIproBatt: Exploring smart battery cell production based on a generic system architecture and an AI-enhanced process monitoring*. [Online]. Available from: https://doi.org/10.13140/RG.2.2.11573.76006 2021.11.07

[5] J. Fleischer, G. Lanza and K. Peter, "Quantified Interdependencies between Lean Methods and Production Figures in the Small Series Production," Manufacturing Systems and Technologies for the New Frontier, pp. 89–92, 2008, doi: 10.1007/978-1-84800-267-8_17.

[6] M. Westermeier, *Qualitätsorientierte Analyse komplexer Prozessketten am Beispiel der Herstellung von Batteriezellen.* [online]. Available from: https://www.mec.ed.tum.de/fileadmin/w00cbp/iwb/Institut/Dissertationen/322_Westermeier_Markus.pdf [retrieved: 02, 2023].

[7] R.A. Fisher, The Arrangement of Field Experiments in Breakthroughs in Statistics. New York, NY: Springer, 1992.

[8] L. Salmaso et al., "Design of experiments and machine learning to improve robustness of predictive maintenance with application to a real case study", Communications in Statistics - Simulation and Computation, vol. 51, no. 2, pp. 570–582, Feb. 2022, doi: 10.1080/03610918.2019.1656740.

[9] A. Paleyes et al., "Emulation of physical processes with Emukit". arXiv, Oct. 25, 2021. doi: 10.48550/arXiv.2110.13293.

[10] M. Zhang, A. Parnell, D. Brabazon, and A. Benavoli, "Bayesian Optimisation for Sequential Experimental Design with Applications in Additive Manufacturing". arXiv, Nov. 23, 2021. doi: 10.48550/arXiv.2107.12809.

[11] Z. Liu et al., "Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing", Joule, vol. 6, no. 4, pp. 834–849, Apr. 2022, doi: 10.1016/j.joule.2022.03.003.

 ISBN: 978-1-68558-041-4

On Factorizing Million Scale Non-Negative Matrices using Compressed Structures

Sudhindra Gopal Krishna, Aditya Narasimhan, Sridhar Radhakrishnan
School of Computer Science
University of Oklahoma
Norman, USA
email:{sudhi, adinaras, sridhar}@ou.edu

Chandra N Sekharan
Department of Computing Sciences
Texas A&M University, Corpus Christi
Corpus Christi, TX, USA
email: csekharan@tamucc.edu

***Abstract*—Non-negative Matrix Factorization (NMF) is one of the algorithms with a wide range of applications, from dimensionality reduction and computer vision to text mining. The dimensions of these matrices can be of the order of several hundreds of thousands to millions, which is a raw format that would not fit in the main memory. Additionally, while performing matrix factorization on these extremely large matrices, the algorithms involving matrix operations such as transpose, multiplication, and subtraction; demand more storage for intermediate resultant matrices. In this paper, we store the matrices in compressed structures (Compressed Binary Tree CBT and Compressed Sparse Row CSR) that allow factorization without decompression. We also perform factorization CBT without using any intermediate structures by performing a virtual transpose and streaming the intermediate resultant matrices of a sequence of matrix multiplications directly into the compressed structure for every iteration. As an example, for an input matrix A of dimension $65,536 \times 65,536$ with $1.46M$ number of non-zero elements, the peak storage in any iteration of the multiplicative update factorization algorithm is $32.98GB$ when using a 2D array, $200MB$ when using CSR and $14.8MB$ for CBT. The ability to stream (add and delete) into the CBT structure without reallocation is why CBT performs the best. Furthermore, we provide a heuristic to reduce memory usage that also aids in faster convergence.**

***Keywords*—Compression; Matrix Operations; Matrix Factorization**

I. Introduction

Non-negative Matrix Factorization (NMF) can be formally defined as follows: Given a non-negative matrix $A \in \mathbb{R}_+$ of dimension $m \times n$ and an inner dimension $k > 0$, find the factor matrices if any, $W \in \mathbb{R}_+$ of dimension $m \times k$ and $H \in \mathbb{R}_+$ of dimension $k \times n$ such that:

$$A = WH$$

The factor matrices W and H are also non-negative in nature. The rank of the input matrix A gives a lower bound for the inner dimension k. This inner dimension k is referred to as the Non-negative rank of a matrix. This problem of finding the factors that satisfy condition $A = WH$ with $rank(A) = k$ has proved to be an NP-hard problem [1] [2]. The short proof of [2] tries to reduce the graph coloring problem and equates the NP-hardness of the graph chromatic number with the non-negative ranks of the input matrix, which is the smallest inner dimension for NMF.

There are various applications [3] that use NMF from computer vision, text mining/information retrieval, email, and pattern recognition to clustering in machine learning [4], face recognition [5] and data mining [6], [7]. Another application of NMF is that it can be used as a lossy compression algorithm to compress a large matrix. If the inner dimension k is small enough, then the input matrix A can be factored in $W \times H$, resulting in a lower number of elements in total. The number of elements in A to be stored will be $m \times n$, but if factorized, the number of elements to be stored will be $m \times k + k \times n$. The latter is assumed to be smaller when k is small.

The correctness of the factorization is calculated using the Frobenius norm suggested by [8] [9]. Now, the problem can be rewritten as:

$$\min_{W \geq 0, H \geq 0} \|A - WH\|_F$$

Some of the well-known sequential algorithms to solve the non-negative factorization are, *Multiplicative Update Algorithms [8] [10], Gradient Descent Algorithms and Alternating Least Squares Algorithms [11] [12]*. There are several approaches as defined in [13] that can be taken to solve this problem. In this paper, we will evaluate the Multiplicative Update Algorithm defined by Lee & Seung [8].

To solve any of the sequential algorithms mentioned above for large matrices, the algorithms require a system

 ISBN: 978-1-68558-041-4

configuration that can handle a huge number of gigabytes of data at a time. We present two state-of-the-art compressed structures (CBT [14] and CSR [15]) that are used to store these matrices and used for operations and algorithms. The input matrices and the factor matrices are all stored in either of these compressed structures. The matrices used for the analyses are both real-world and synthetically generated. We have also shown that there is a space-time trade-off between the two structures CBT and CSR. CBT taking lesser space and CSR having a shorter query time [15].

Our contribution is as follows:

- We provide a method for factorizing matrices with the least memory footprint per iteration using compressed structures.
- Sections III-A and III-B, explain various value-based matrix−matrix operations that are performed without decompression.
- We provide a matrix-transpose multiplication algorithm (Section III-C) that provides results without transposing, by streaming the result directly into compressed structures.
- In Section III-D, we explain how we sequence 3 or more matrix multiplication operations, without storing any intermediate matrices.
- Proposed a heuristic (Section III-E) that eliminates unnecessary rows/columns that leads to lower memory usage and faster convergence.

II. Related Work

The foundation for the Non-negative matrix factorization was laid by Lee and Seung [8] in 1999, opening the opportunity to hundreds of research journals. Before Lee and Seung, few other notable contributions were made in the area of NMF, but none came close to the fame of Lee and Seung. Paatero and Tapper, 1994 [16], produced the work on positive matrix factorization. Lee and Seung cite the work of Paatero and Tapper in their work. Articles have shown the significance of Paatero's work prior to Lee and Seung but have gone unnoticed.

Since Lee and Seung's NMF was one of the first ones to be popular, it became a baseline for many research. Several researchers have proven that the multiplicative update algorithm proposed by Lee and Seung [8], is slower to converge, which means that it takes many more iterations to complete compared to the gradient descent method and the alternating least squares. Each implementation required a total of 12 matrix operations, of which six require $O(n^3)$ matrix-matrix multiplication, and the rest require $O(n^2)$ matrix-matrix element-wise operations.

To overcome this issue, other researchers, such as Gonzalez and Zhang in 2005 [10], proposed an alteration to the multiplicative update, but it ended up having the same convergence issue. Another researcher named Lin [17] in 2007 proposed a modification that ended with earlier convergence but at the cost of more operations per iteration.

Theoretically performing 12 matrix operations on a matrix is time- and space-consuming; performing the same operation on larger matrices would require a great deal of memory. For example, a $65,536 \times 65,536$ requires about 32 GB of storage in its raw format. To overcome this, in this paper, we use our novel CBT [18], which works well with binary matrices and the bit-packing algorithm proposed in [15] to store integer values. We also propose to store the matrix in CSR [19], a common data structure for storing matrices.

To perform factorization or any operation on large sparse matrices, one must efficiently store the matrices so that the entire data can be loaded onto the main memory in one go. Given a matrix of size n rows and m column, the total number of possible elements in the matrix is the size of the matrix itself, which $m \times n$, therefore, the cost of storing a matrix in raw format would require $(m \times n) \times 64$ number of bits, where 64 is the number of bits required to store a number. But in a sparse matrix, this number tends to be very small, where the number of non-zero elements is extremely less compared to the number of zeros.

Therefore, the sparsity of a matrix is defined as the ratio of the number of non-zero elements to the number of all possible elements that can be in the matrix.

$$Sparsity = \frac{nnz}{m \times n},$$

where nnz is the number of non-zero elements in the matrices, n is the number of rows and m is the number of columns.

This type of behavior in the matrices are found in the real-world, such as social networks, biological network, topological network, and so on. The cost of storing zeros in such cases becomes expensive and redundant to an extent, as they do not contribute to the analysis.

Therefore, to store large sparse matrices, in this paper, we are using existing structures such as CSR [15] [19], and CBT [20].

III. Matrix Factorization

There are several approaches that can be taken to factorize a given matrix. To mention a few of the popular ones, multiplicative update, gradient descent,

and alternating least squares [8] [11]. Here, we take the updated rules provided by Lee and Seung [8].

$$H \leftarrow H \frac{(W^T V)}{(W^T W H)}, \quad W \leftarrow W \frac{(V H^T)}{(W H H^T)}$$

```
begin
    W = rand(m, k)
    H = rand(k, n)
    for i : maxiter do
        H ← H .* (W^T A) ./ (W^T W H + 10^-9)
        W ← W .* (A H^T) ./ (W H H^T + 10^-9)
```

Figure 1. Multiplicative Update algorithm for NMF using the Frobenius norm as a cost function

Algorithm 1, shows the workings of how to factorize the given large matrix using the multiplicative update algorithm. The algorithm involves a series of operations to obtain the desired result of W and H. To clarify the various matrix element-wise operations, the $.*$ operation represents an element-wise multiplication, and a $./$ represents an element-wise division and matrix-based operations such as matrix-matrix multiplication. So, we continue this section by providing the algorithms for the various operations that are the building blocks of 1.

A. Matrix-Matrix Multiplication

One of the first and most important operations to be performed during the factorization process is matrix-matrix multiplication. The work on matrix-matrix multiplication has been published in [21], which explains the working of how two matrices stored in either of the data structures CSR and CBT are multiplied without the need for an intermediate data structure.

B. Element-Wise Matrix Operation

The multiplicative update algorithm consists of several element-wise matrix operations. The operations involved in the algorithm are element-wise multiplication $.*$, element-wise division $./$, and element-wise subtraction $-$ to find the Frobenius norm. Apart from these three, we can also extend the algorithm for element-wise addition $+$.

Algorithm 2, explains the working of the element-wise matrix operation. The operation to be performed, "Op," is specified as input. The algorithm first checks if the dimensions of the two matrices are equal and, if not, throws an error. It then loops through each row of the matrices, and for each row, it checks if the size of the row is zero in either matrix. If it is, it appends a zero to the corresponding row of the resultant matrix C. If both matrices have a row of size zero, it also appends a zero to the corresponding row of C. If only one matrix has a row of size zero, it copies the elements from the non-zero row and appends them to the corresponding row of C. If neither matrix has a row of size zero, the algorithm performs the specified operation on each element of the corresponding rows of A and B and appends the result to the corresponding row of C. Finally, the algorithm returns the resultant matrix C.

```
Input: Matrix A, Matrix B, Operation Op
Output: resultan_matrix C
if A.rowSize != B.rowSize or A.colSize != B.colSize then
    Error: Matrix dimensions should be the same for both the matrices
for i in numberofRows do
    if A[i].rows == 0 and B[i].rows == 0 then
        C[i] = 0
        continue to the next row
    else if A[i] == 0 then
        C[i] = B[i]
        continue to the next row
    else if B[i] == 0 then
        C[i] = A[i]
        continue to the next row
    for aIndex in A[i] do
        for bIndex in B[i] do
            C[i][j] = A[i][j] "Op" B[i][j]
            Where "Op" = "+ or - or .* or ./"
return C
```

Figure 2. Element-wise matrix Addition, Subtraction, Multiplication, and Division

C. Matrix Transpose

Another important operation required to perform matrix factorization is to transpose a given matrix. There are two ways we have handled this situation in this paper, one way is to transpose the given matrix and store it as another matrix that occupies extra space, and another way to do it is to incorporate transpose during the required operation.

 ISBN: 978-1-68558-041-4

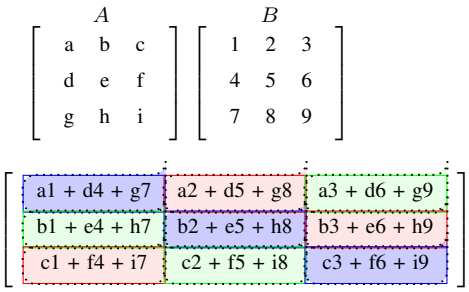


Figure 3. The working of $A^T \times B$, by storing the result in a pattern to eliminate the need to transpose the actual matrix.

The multiplicative update algorithm contains matrix-matrix multiplication where either one of the matrices needs to be transposed. A way to achieve this operation would be to transpose the required matrix and use the algorithm mentioned in [21], but this requires additional memory; here the additional memory is the transposed matrix. To avoid this issue, we perform an in-place transpose multiplication. This can be achieved by accessing the matrices with a different access pattern.

$$A \times B^T = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 5 & 0 & 2 & 3 \\ 3 & 0 & 0 & 5 \\ 0 & 0 & 2 & 4 \\ 0 & 1 & 2 & 0 \end{bmatrix} \end{array} \times \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 2 & 4 & 3 & 1 \\ 2 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 3 & 0 & 2 & 0 \end{bmatrix} \end{array} \tag{1}$$

$$\Rightarrow \begin{array}{l} r_0(A) \rightarrow \\ \\ c_0(B) \rightarrow \end{array} \begin{pmatrix} 5 \\ \times \\ 2 \end{pmatrix} + \begin{pmatrix} 5 \\ \times \\ 4 \end{pmatrix} + \begin{pmatrix} 5 \\ \times \\ 3 \end{pmatrix} + \begin{pmatrix} 5 \\ \times \\ 1 \end{pmatrix}$$

$$\Rightarrow c_0[C] = \{10\ 20\ 15\ 5\} \tag{2}$$

Equation 2 shows an example of $A \times B^T$, where the partial resultant of column $c_0[C]$, is obtained after multiplying the first row $r_0[A]$ of A, and virtually transposed the first column of B, in this case, it is still $r_0[B]$.

Figure 3 shows the multiplication of $A^T \times B$ by virtually transposing A. Here, the colors along the diagonal show the order in which the resultant is obtained. Multiplying $r_0[A]$ with all rows of B, we obtain the main diagonal; continuing the process to the farther rows of A, we move the resultant to the upper triangle and wrap it around to the lower triangle, as shown in **red**, and **green**.

D. Sequence of matrix multiplications

Revisiting algorithm in [21], where algorithms take two matrices as input and multiply them to produce the resultant matrix. However, data structures such as our novel versions of CBT and CSR are amenable to multiplying multiple matrices without storing the intermediate resultant matrix.

Algorithm 5 shows multiple matrix-matrix multiplication. Line 5 takes the output of line 3, the intermediate resultant row, and computes the resultant row on the third matrix. This process can be repeated through any number of input matrices. Therefore, this can be scaled to k as the number of matrices.

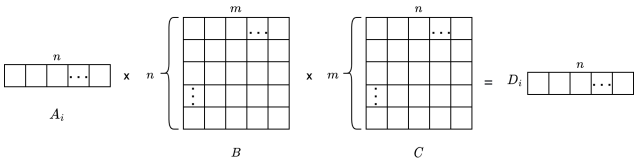


Figure 4. The working of sequential matrix multiplication.

Figure 4 shows the pictorial representation of sequential multiplications of multiple matrices. A row of matrix A, A_i is multiplied by matrix B using the partial sum algorithm to obtain the intermediate resultant row Z_i, then Z_i is multiplied with the next matrix C to obtain the final resultant row D_i.

E. Heuristic for faster convergence

One of the drawbacks of the multiplicative update approach is the convergence time and the iterations it

Input: Matrix A, Matrix B, Matrix C
Output: Resultant_Matrix D

```
1  for row = 0 to numberOfRows do
2      aRow = getRow(A, row)
3      tempArray = copmutePartialSum(aRow,B)
4      /* Call Algorithm in Sec III-A */
5      tempArray = computePartialSum(tempArray,
         C)
6      /* Call Algorithm in Sec III-A */
7      for nnzElements in tempArray do
8          D[row].streamEdge(nnzIndex)
9          D.values.bitPack(nnzValues)
10     /* If we are performing the matrix
         multiplication in CSR, then the number of
         non-zero elements in the resultant data for
         each row should be stored in C */
11 return D
```

Figure 5. Matrix-Matrix Multiplication in Sequence

takes to find an optimal solution. One of the ways to make the algorithm faster would be to reduce the number of non-zero values in the input matrix. If we are given a threshold number of index positions per row that can be made zero, we can come up with a heuristic approach to make specific values zero so that our compression is more efficient. One way to approach this is to remove the noise in the data; that is, we remove the data that do not contribute to the overall solution. This may lead to more loss, but the threshold will dictate the metric of the percentage of loss added to this already lossy factorization approach if we had not taken the heuristic approach. This will be a heuristic approach and will not be optimal. But it will lead to reduced resource utilization. Space is reduced in the already compressed structure and time to query the smaller CBT structure.

IV. Experimental Results

This section evaluates matrix factorization on various matrices. For this experiment, we considered the variety of matrices with variable sparsity.

To factorize the matrices, we must first choose low-rank dense random W and H matrices. Choosing a low-ranking matrix leads to the formation of a smaller resultant matrix, which in turn consumes less space. Finding an optimal rank for factorization is a hard problem, as the algorithm has to go through the process of finding the number of orthogonal rows in the matrix. It is also more likely that the larger the inner dimension of the factors that we compute (W and H), the sparser these matrices will be, in which case CBT outperforms CSR even in terms of the storage of dense matrices. Therefore, in this paper, we perform a brute-force analysis to obtain a minimal rank that would satisfy the criteria to reproduce the almost original matrix when W and H are multiplied.

Table I shows the overall result of the computation performed in this paper. The first set of columns in the table explains the basic details of the input, matrix dimensions in the first column, the number of non-zero elements in the second, matrix size when represented by using the $2-D$ matrix in the third, and the compressed sizes in the fourth and fifth respects. In the next part of the table, we present the inner rank of the factored matrices, followed by the result of $W \times H$ for both CBT and CSR, and the amount of memory required to process factorization at each iteration by CBT, CSR, and $2-D$ representation of the matrix.

In the results, one can notice that the memory required by the $2-D$ matrix is the highest. Still, the majority of the size is just the A matrix. Since the resultants can be streamed into a matrix in O(1) (constant), the extra memory used is very minimal. Still, as the inner rank increases, memory usage will increase accordingly.

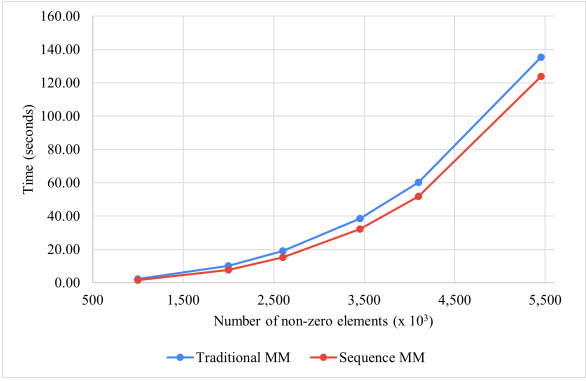


Figure 6. Comparison between the time taken to multiply three matrices in traditional two steps and uses our novel sequence multiplication in a single step for a Million-by-Million matrix.

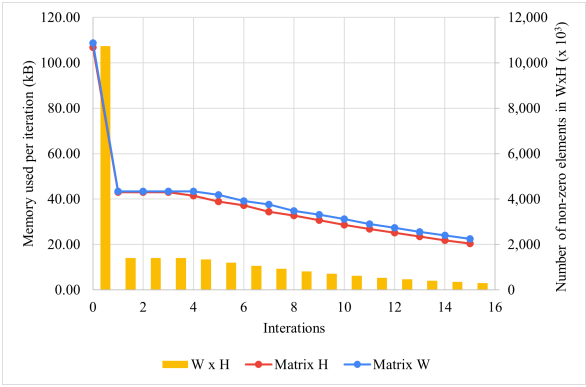


Figure 7. The evolution of W and H during the factorization for Matrix of size (21,504× 21,504, 1.36M nnz elements)

However, when considering the two compressed structures, the proportion of memory consumed by CSR is much greater compared to the memory consumed by CBT [15]. This is due to the inability of CSR's to stream (add/delete), as the arrays need to resize, whereas CBT's ability to perform in-line operations, the advantage of one such operation is shown in the Figure 6, the figure compares the time taken to multiply three matrices in traditional two steps and uses our novel sequence multiplication in a single step for a Million-by-Million matrix of various levels of sparsity ranging from 1M to 5.4M elements. This memory usage will have a significant impact for a very large matrix, as shown in [21].

Figure 7 shows the decrease in the memory required to store W and H as the iteration progresses, with the number of non-zero elements represented in the bars.

 ISBN: 978-1-68558-041-4

TABLE I. THE FACTORIZATION RESULT USING CBT AND CSR AND THE MEMORY REQUIRED TO PROCESS THE FACTORS.

						W × H		Avg Mem/Iter		
Matrix A	**NNZ**	**Matrix Size**	**CBT**	**CSR**	**Inner Rank**	***CBT***	***CSR***	***Matrix***	***CBT***	***CSR***
2688×2688	23,089	55.12 MB	217.36 KB	216.23 KB	448	216.58 KB	216.51 KB	73.5 MB	0.54 KB	0.67 MB
5376×5376	57,752	220.5 MB	547.53 KB	546.68 KB	255	513.87 KB	526.46 KB	241.41 MB	0.29 KB	30 MB
21504×21504	1,385,198	3.44 GB	12.7 MB	12.98 MB	512	12.65 MB	12.95 MB	3.6 GB	13.1 MB	150 MB
43008×43008	998,531	13.78 GB	9.45 MB	9.53 MB	670	9.1 MB	9.98 MB	14.21 GB	9.92 MB	87 MB
65536×65536	1,460,048	32 GB	14.23 MB	14.05 MB	665	13.45 MB	14.12 MB	32.64 GB	14.80 MB	200 MB

All experiments were run on an Intel(R) Xeon(R) W-2295 CPU @ 3.00GHz (16 Cores) with 64 GB of RAM, and the programs were written in GNU C/C++.

V. CONCLUSION AND FUTURE WORK

This paper shows that the given million-scale matrix can be factorized directly on the compressed structure. We also show that the intermediate result obtained in the matrix factorization process can be eliminated using sequential matrix operations. In this paper, we also introduced element-wise matrix multiplication, division, subtraction, addition, and sequential multiple matrix multiplications on top of the existing work of matrix multiplication. We have also shown that traversing through the matrix in the pattern can avoid an explicit transpose operation during the matrix factorization. We also provide the heuristic relationship between inner rank and the sparsity of the factor matrices, and we have also shown in the results that the lower the rank, the smaller the factors W and H. In the future, we would expand the computation to the Alternating Least Squares and Gradient Descent approach to factorize matrices. Our compression algorithms mentioned in this paper natively support binary matrices. Hence, we would also expand our work toward Binary Matrix Factorization.

REFERENCES

[1] S. A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1364–1377, 2010.

[2] Y. Shitov, "The nonnegative rank of a matrix: Hard problems, easy solutions," *SIAM Review*, vol. 59, no. 4, pp. 794–800, 2017.

[3] N. Gillis, "The why and how of nonnegative matrix factorization," *Connections*, vol. 12, no. 2, pp. 257–291, 2014.

[4] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 267–273, 2003.

[5] D. Guillamet and J. Vitria, "Non-negative matrix factorization for face recognition," in *Catalonian Conference on Artificial Intelligence*, pp. 336–344, Springer, 2002.

[6] M. W. Berry and M. Browne, "Email surveillance using non-negative matrix factorization," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 249–264, 2005.

[7] S. Zhang, W. Wang, J. Ford, and F. Makedon, "Learning from incomplete ratings using non-negative matrix factorization," in *Proceedings of the 2006 SIAM international conference on data mining*, pp. 549–553, SIAM, 2006.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, pp. 556–562, 2001.

[9] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor, "Mahnmf: Manhattan non-negative matrix factorization," *arXiv preprint arXiv:1207.3438*, 2012.

[10] E. F. Gonzalez and Y. Zhang, "Accelerating the lee-seung algorithm for nonnegative matrix factorization." http://www.caam.rice.edu/tech_reports/2005/TR05-02.ps, 2005.

[11] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational statistics & data analysis*, vol. 52, no. 1, pp. 155–173, 2007.

[12] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM journal on matrix analysis and applications*, vol. 30, no. 2, pp. 713–730, 2008.

[13] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Transactions on knowledge and data engineering*, vol. 25, no. 6, pp. 1336–1353, 2012.

[14] M. Nelson, S. Radhakrishnan, and C. N. Sekharan, "Billion-scale matrix compression and multiplication with implications in data mining," in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 395–402, IEEE, 2019.

[15] S. Gopal Krishna, M. Nelson, S. Radhakrishnan, A. Chatterjee, and C. Sekharan, "On Compressing Time-Evolving Networks," in *ALLDATA 2021, The Seventh International Conference on Big Data, Small Data, Linked Data and Open Data*, pp. 43–48, 2021.

[16] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[17] C.-J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1589–1596, 2007.

[18] M. Nelson, S. Radhakrishnan, and C. Sekharan, "Queryable Compression on Time-Evolving Social Networks with Streaming," in *Big Data (Big Data), 2018 IEEE International Conference on*, IEEE BigData '18, pp. 146–151, IEEE Computer Society, 2018.

[19] R. A. Snay, "Reducing the profile of sparse symmetric matrices," *Bulletin Géodésique*, vol. 50, no. 4, pp. 341–352, 1976.

[20] M. Nelson, S. Radhakrishnan, A. Chatterjee, and C. Sekharan, "Queryable Compression on Streaming Social Networks," in *Big Data (Big Data), 2017 IEEE International Conference on*, IEEE BigData '17, pp. 988–993, IEEE Computer Society, 2017.

[21] S. G. Krishna, A. Narasimhan, S. Radhakrishnan, and R. Veras, "On large-scale matrix-matrix multiplication on compressed structures," in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 2976–2985, 2021.

 ISBN: 978-1-68558-041-4

Software Monitoring of an IoT Chain Communicating over LoRaWAN

Follow-up on Planned Data Collection

Mohamed El Kharroubi
Algorithmic, Complexity and Logic Laboratory
UPEC University
Creteil, France
email: mohamed.el-kharroubi@u-pec.fr

Fabrice Mourlin
Algorithmic, Complexity and Logic Laboratory
UPEC University
Creteil, France
email: fabrice.mourlin@u-pec.fr

***Abstract* — In this work, we are interested in the problem of authentication badge usage and its control through a low energy message exchange protocol. This problem is interesting because the monitoring of these badges provides incident prevention and reporting on usage. Past work has only signaled the loss of a professional badge by a ringing tone, which is not consistent with current usage. Some authentication phases require the badge to be read over a long period. Our solution is based on sending a small number of messages on a specific protocol for each change of state of the badge. Our main result is the construction of a report on the status of all the badges in use and especially the detection of real loss of badge with notification to their owner. The gain is appreciable compared to the false detections that were previously reported.**

Keywords-loose coupled component; Big Data; LoRaWan protocol; NoSQL Database.

I. INTRODUCTION

Monitoring an Internet of Things (IoT) system necessarily involves a data collection phase. It is often complex because it depends on the deployment of the IT system: storage sites, communication nodes, choice of sensors, signal strength and of course security. This last feature brings the ultimate facet of technicality that transforms the IoT system into a hard-to-observe one.

Today's applications offer uses where personal information is obvious and the understanding of information collection is a key point. Even more so, the planning of data collection is itself monitored. Malicious interventions not only harm application results but also introduce biases that affect the overall application usage. If this duration is initially associated with the duration of deployment of the application, we observe more and more that the use of access to local or remote clouds prolong the impact of these biases and harm the observations on the long term.

These accesses to persistence systems are generally done by using the classic http or https protocol. It seems natural today to favor local storage over remote storage to ensure a good reactivity to incidents. Our past studies have also shown the contribution to data security. On the other hand, imposing data schemas remains too strong a constraint to respect when collecting data on a widely distributed system. The use of log message schema is a step forward to apply aggregation operations on messages of different origins. It is also an action to simplify the understanding of data collection. The typing is very simple and underlines the time stamp of the message and the locality of its production. There are different syntaxes for timestamp format and time zones. The use of data formats customizes the analysis of timestamps.

New information is taken into account such as the signature of the device or the language or the country. In relation to the current execution, useful information is given about the current process, or even the name of the current thread, the name of the lock acquired or returned, the name of the signal sent or processed. In short, the operating system aspects are taken into account, but there still has to be one, which is not necessarily the case in the context of IoT chains.

This information often represents functional states. Thus, a control device connected to the network can read the battery status, mains voltage, storage device status, etc. In addition, it can also collect and transmit statistics about the storage capacity used. Hence, the need for richer data schema to explain what is involved in data collection. The limit of this situation remains the consideration of business data in software monitoring with the evolution that this implies throughout the life of the software. We present our work on the collection of data from IoT chains for the application of Big Data processing.

The rest of this paper is organized as follows. In Section 2, we detail the work related to the project we are addressing, emphasizing their contributions first and the features we wish to push back. Section 3 discusses our use case and the software architecture we are implementing. Section 4 describes our Big Data processing and the provision of information related to the previous collections. Section 5 discusses the current analysis results and the impact on the NoSQL database storage. Finally, Section 6 describes the future work to achieve a better understanding of the computations. The acknowledgement and conclusions close the article.

II. RELATED WORK

In the Internet of Things (IoT) domain, each use case relies on rich and sometimes complex electronic devices for which it is important to design an ad-hoc software approach. The following studies show that monitoring is an integral part of the IT system, like mascarpone in tiramisu.

 ISBN: 978-1-68558-041-4

H. Lv et al. [1] designed a monitoring system for an urban environment that relies on a dedicated communication network to complement on-the-fly data collection. In addition to the monitoring network, there is a remote station for collection and simulation control. In this context, an observer station is identified to ensure live monitoring with the possibility to react in case of unexpected events, but this remains precarious especially in the case of long duration simulation. Events can occur in a cascade mode and it could be difficult to have a suited reaction.

W. T. Hartman et al. [2] work aims to build, test, and implement a low-cost energy monitoring and control system using IoT devices. They have designed a complete system from input to output that includes a mobile app, cloud database, application-programming interface, and hardware development. Equipped with these tools, they propose developments for each energy consuming equipment and thus observe the uses to limit the losses. If the main principles implemented are clearly explained, it remains that the development of monitoring, as an additional layer to a device already in place, is delicate here and it is almost impossible to provide business information.

S. Nocheski et al. [3] were interested in monitoring to maintain the sustainable habitat environment for certain fish species inside fishing ponds through distributed machine-to-machine communication. Their goal is to reduce the chain of responsibility for some basic actions. Their IoT system consists of sensors that measure crucial water quality factors, such as temperature, light intensity, or water level, and an on-board computer that processes the data and sends audio and visual notifications to the fish farm manager. They lead to an experimental development where their action/reaction concept is implemented. The Wivity modem allows the user to communicate with the IoT system through Wi-Fi, cellular, Long Range Wide Area Network (Lora WAN) or satellite communication, all in one product. In this context, a protocol is dedicated to the monitoring and communication of minimal data. Each message is defined by a basic grammar.

G. Yang et al. [4] presented their work on the detection of behavioral signs of pain. Their goal is to evaluate this pain at work in real situation rather than using self-reporting which can only be practiced in delayed mode. They propose a wearable device with a bio sensing face mask to monitor a patient's pain intensity using facial surface electromyogram (sEMG). The wearable device is integrated with the Internet for remote pain monitoring. It transmits data to the server via an http gateway. The cloud-accessible persistence system manages the wireless communication between the server and the web application. In this context, the data security aspects remain a concern but among the assets of the system, the patient is free to move and his mobility is taken into account.

The work of K. Sabanci et al. [5] relies on an electronic device (a camera network) to identify people in a store and count them by calculating their path through an enclosed space. Mobile clients running on a nomadic terminal such as a phone can read in a shared cloud the paths of these people to monitor their theme of interest. The authors want to distinguish between stationary and mobile people. In addition, they apply a background subtraction technique to isolate moving people. This counter will inform users of the areas left empty and then make this information available in a cloud. Users have a mobile application to track changes in the meter. The collection of information from cameras is the source of information, while the sharing is obtained with a shared persistence system. Specific clients come to consult the evolutions of this data.

The use of data schema is a more recent topic, as it introduces difficulties about data representation. J. Wang et al. [6] were interested in the control of the greenhouse environment. The acquisition and control parameters, network protocols are different for various greenhouse feedback. These factors are the keys to the abilities to communicate effectively and transfer meaningful data in IoT infrastructures. In order to realize the adaptation of data communication between the gateway and server in a greenhouse IoT system, an XML-based data encapsulation method was designed to enable data interoperability in a distributed greenhouse IoT system. A multi-agent system is used to fuse heterogeneous information and responses for data synchronization in the greenhouse IoT system. The results show the importance of patterns when communicating data over the network and this independently of the chosen protocol. From BlueTooth, RFID to HTTP, the principle is to format the data flow to better communicate with an external station.

T. M. Bandara et al. [7] make a new contribution with a large area data collection where wireless sensors are accurately distributed over an agricultural area. Thus, the data returned are all localized and this allows representing the parameters with reinforced semantics. Examples of parameters are soil moisture, temperature sensors, water volume sensors, etc. The notion of data schema is introduced in a format specific to the project, but it allows the unification of representations. The use of schema brings the validation of data before its use. Thus, during the analysis, we can discard data that does not respect the associated schema. Moreover, data extraction is of better quality when it is applied to validated data. A schema describes not only the order of the tokens but also the accuracy of the data. On the other hand, data communication process is based on energy-intensive telecommunication technologies.

In the same field of agricultural monitoring, D. Davcev et al. [8] uses a low-energy cost Long Range Wide Area Network (LoRa WAN) protocol for data communication. Of course, this introduces an additional risk because new properties appear such that the message body is encrypted, which means an additional cost for the processing but an increased security against network intrusions. Furthermore, the range of the antennas and their placement play a crucial role in the software architecture of the proposed prototype. The reading of these works underlines the difficulty of adopting the same approach.

 ISBN: 978-1-68558-041-4

TABLE I. CONTRIBUTION AND LIMITATION OF EXISTING SOLUTIONS

Contribution	Limitations
[1] point authentication of user	authentication is missing for a certain period of time
[2] monitoring mixed with business code	need for structured monitoring broken down into monitoring layers
[3] the Wivity modem does not allow the event specific monitoring	there is a need to configure the monitoring by event family
[4] the collected data are stored in a cloud and there is safety problem	there is a need to store data locally at the monitoring system without exposing it to the outside of the system
[5] the persistence system is shared ans tha data access is shared	shared access introduces security issues that must be managed or prohibited.
[6] it brings the use of XML data schema	this concept is to be propagated to all data collection phases, JSON, YAML, etc
[7] the scope od data transmission	the use of specific data transmission is an asset for the separation of concepts

The work presented in this Section represents a significant excerpt that shows the importance of data collection and its difficulties. Among other things, it highlights problems of explanation in the case of non-compliant or erroneous data collection. How monitoring can provide answers to these questions. We are also confronted with this need for explanation in our work. There are many reasons for this. On the one hand, it can come from the collection process itself, but the works presented also show sensor failures, data transfer problems, lack of energy to allow the signal to reach its objective. This list is not exhaustive, and it is important that these incidents are time-stamped and recorded.

III. SOFTWARE ARCHITECTURE

A. *Use Case*

Our case study is based on the use of authentication cards that all the people of the university have in order to enter the different premises of our site. To be more precise, these cards are nominative and are associated with accounts configured in such a way that certain premises are accessible and not others. These cards are also useful when using the classic printers and 3D printers present in different rooms. Thus, a print initiated in an office by a given person does not require the person to choose the destination printer but only the type of printer. After moving to an available printer, authentication with the business card triggers the previously initiated output.

The authentication card can be used in two ways: a simple contact with a sensor or the insertion of the card in a dedicated reader. This is necessary especially on 3D printers or large format printers whose working time can exceed one minute and leads to a new authentication request at regular period. In these cases, it is frequent that a user forgets his card when taking the documents out of the printer. Many users discover that their card is missing the next time they use it. But in this case, where did it stay? For this purpose, we asked for the design of a LoRa WAN cardholder. We chose to use the LoRa Wan protocol because on the one hand, it is a low-cost radio protocol that we master in the laboratory, and on the other hand, it is a protocol that ensures that the body of the messages is encrypted and thus ensures better security of the collected data.

This cardholder has a sensor to detect the presence or absence of card. Moreover, when the card is absent, it emits a LoRa WAN message to indicate the absent card and the identifier of the place of emission. This message is kept in the memory of the cardholder and will be sent again one hour later with an increment. In the end, the transmissions will stop only once the card is back in its holder. The use of such a protocol is explained by the autonomy of the cardholder in energy, but also by the fact that we have a LoRa WAN antenna on the site and an IP/LoRa WAN gateway per floor.

The messages collected by the network monitoring show the loss of a card. They allow the application of a script in order to send a message to the person concerned. In order to continue the validation phase of this device, we have planned several phases of monitoring data collection. From these data, a Big Data batch allows to extract the useful information to insert it in a NoSQL database. A report on the activity of this database is then published for the security department.

B. *System Architecture*

The authentication system is already deployed in the university's buildings, so we are considering the structure of a typical floor, such as those in our own building. We represent one device per device family and the users each have a cardholder containing initially a business card. The entrance and exit of the laboratory being controlled by the use of these badges, each door has two readers that we assume are placed on each side of the door. Each area requiring particular rights is thus considered as a laboratory, i.e., with a minimum entry/exit door. For reasons of simplicity, these zones can be nested one inside the other, but there is no door that allows you to leave two zones simultaneously, you will need two successive doors for that.

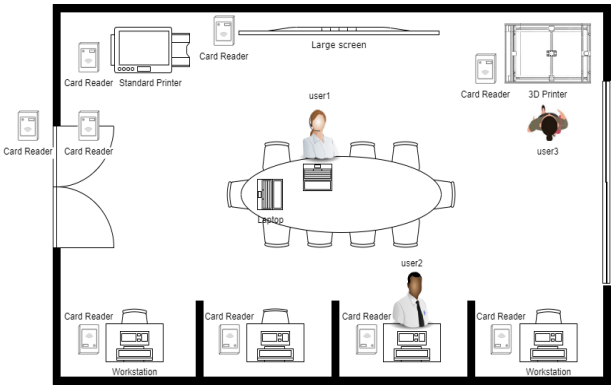


Figure 1. Physical deployment diagram of authentication system

A typical physical deployment diagram is shown in Figure 1. It helps to understand the entities that come into play in our scenario. All devices used in the lab are attached to a card reader including video conferencing equipment. This allows, among other things, to have additional information in case of an incident. The main element of this diagram is the user: three instances are present in Figure 1. Each of them carries a badge

reader containing their professional badge. In order to be connected, each of them inserted their professional badge in the reader embedded in the laptop for user1 or in the external reader at the workstation for user2 and user3. This means that they have taken their badge out of the cardholder and now it is empty. For users1 and user2, authentication is one-time and they can put away their badge as soon as the operation is finished. On the other hand, for user3 who uses the 3D printer, security constraints require his presence during the printing process. In addition, she cannot put away her badge because her presence is periodically evaluated during the printing process. This means that her badge will remain empty during the printing process and that warning messages will be issued.

A diagram of components Figure 2 applies to our first diagram and includes the software part, which produces the data, the collections and especially triggers the extraction of data towards a database. The management of badges is done in parallel with the use of badges, if they are used for the laboratory equipment, our software is monitoring badges is deployed on several materials: the cardholder, the LoRa/IP gateway and a Big Data component (Apache Spark) for data processing. Finally, actions are performed following the events of the NoSQL database.

Two distinct applications are represented in this diagram, Figure 2. The gray components belong to the user authentication application within the laboratory, while the yellow components belong to the badge monitoring application. The latter includes a data collection phase, followed by an information routing using the Kafka framework [9] and a Big Data analysis component based on the Apache Spark framework [10]. These two frameworks are considered standards in the world of Big Data for their ability to partition data in memory and especially to distribute calculations in relation to the data.

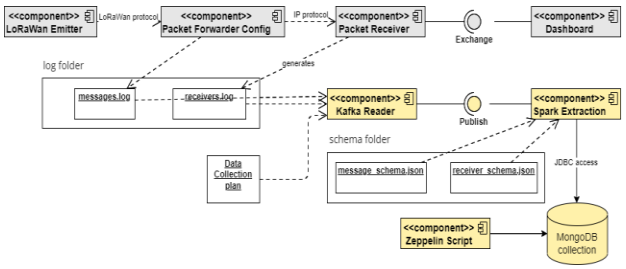


Figure 2. Component Diagram Structure

A log file directory receives all files generated by the packet forwarder or by the packet receiver. These files contain text in JSON format which is a standard text-based format for representing structured data. That is why we also have a JSON schema directory. This allows us to validate the data read by the Kafka component. On the other hand, it is then possible to search for information by using the JSONPath query language [11]. This library is used by the Spark SQL component. The management of the collections is parameterized via a configuration file used by the Kafka component, which manages the data sampling.

IV. BIG DATA WORKFLOW

The design of our Big Data workflow is based on sampling data, extracting information, and then enriching this data before saving it in a persistence system. Thus, it will be easy to build models or reports on data movements.

Figure 2 shows a breakdown into three components where the Kafka component is the entry point of the workflow; it has a large parameterization and initiates the data distribution.

A. *Kafka Reader Component*

Apache Kafka is an open-source framework for streaming messages to other components. It is a must-have tool because its outbound data rate is very high, and its low computational resource consumption is much better than competing tools. The objective of this component is to deliver, analyze and index millions of log messages from professional badge monitoring activity over LoRaWan protocol. The Kafka framework is designed for this type of high-volume, distributed message flow. It ensures that our component named Kafka Reader can process every event without being overwhelmed and without putting pressure on the log message producers.

Log message producers send events to Kafka whose content respects a data schema, which divides them evenly into partitions. We have defined a message schema for each sending source. This makes it easier to perform searches and transformations. Kafka organizes messages into topics, which are themselves divided into partitions. Each partition is an ordered queue of events assigned to our component named Spark Extractor. In our case, the more partitions there are the higher the throughput, at the expense of memory. Of course, this number of partitions has a strong impact on the distribution of computation that is done by Spark Extractor.

Each partition has several replicas that mirror the given partition as closely as possible. In case of a damaged partition, Kafka automatically switches to a replica, which provides a fault tolerance system without us having to implement it. The analysis of processing times highlights the use of these replicas. The synchronization of this distributed component relies on Apache Zookeeper, which allows the construction of an Apache Kafka node cluster [12]. Our configuration file also includes information about the automatic restart in case of connection loss, as well as data about the multiple user components of the Kafka nodes.
The configuration choices are strongly influenced by the work of B.R Hiraman [13].

B. *Spark Extractor Component*

Supplying data from a Kafka topic means that the Kafka Reader component controls the data sampling: from the duration of the reading window to the overlap time between two consecutive windows, including the number of data partitions to be consumed by the Spark component. Depending on the consumed topic, the latter determines the data schema to apply in order to validate the data. Then, the use of JSON Path queries allows us to extract the data we want to keep in a document inserted in MongoDB. An example of a message coming from the packet forwarder is given in Table 2.

 ISBN: 978-1-68558-041-4

TABLE II. SIMPLE ERROR MESSAGE FROM PACKET FORWARDER

Content of the body part
{ "timestamp": "2022-12-23T12:34:56Z", "level": "error", "message": "There was an error forwarding the packet", "request_id": "4402574329", "user_id": "pflrw1" }

The role of the data schema is essential because it is used for our indexing process of the log messages before insertion. Moreover, if the data schema evolves over time, its externalization to the Spark Extractor component ensures an adaptation of the treatments performed on the analyzed log messages.

We chose the MongoDB server because it is very elastic and allows us to combine and store multivariate data without compromising indexing options, data access and validation rules. We use the plugin named mongo-spark-connector in order to establish a connection with the MongoDB server as in [14][15]. We provide the connection URI to MongoDB in the SparkConf object. From the data frames built in memory with Spark Engine, we save each of them via the mongo-spark-connector. Because the value of the message key is rich. We use our data schemas to parse the value associated with the message key [Table II] and thus enrich the JSON document and add new keys/values not present in the original document. This is made possible by the use of dynamic schema in MongoDB.

As an example, table 3 details a log message from the Packet Receiver. Each event from the badge holder is transmitted via a packet forwarder to our Packet Receiver.

The log messages from this component are essential to understand the actions that are triggered.

TABLE III. SIMPLE MESSAGE FROM PACKET RECEIVER

Content of the body part
{ "timestamp": "2022-12-20T10:31:51Z", "level": "success", "message": "badge out of the badgeholder at 2022-12-20T10:25:22Z located at the reader_id r10f1b4", "request_id": "4402171112", "user_id": "pflrw1" }

Many messages are issued to track the monitoring behavior of business badges. The collected information is aggregated to provide richer messages for the observer. Thus, a first message notifies the exit of the badge from the badge holder. Then, a second message is sent when the badge is inserted in the badge reader. Finally, the badge reading event for authentication. Of course, these messages do not have the same component as their origin, but they offer a follow-up of the badges uses overall laboratory. The configurations shown in Figure 2 allow controlling the format and the details of the transmitted messages.

From our component, Spark Streaming engine defines micro-batches that initially guarantee a latency of around 100 ms at execution and a unique processing of Spark events. With Spark 2.3 came the continuous processing model with a reduced latency of 1ms, Spark events can nevertheless be duplicated. The first step is to create a Spark Session. We then retrieve our streaming data from Kafka topics where the messages are already partitioned. The *startingOffsets* option is set to *latest*, forcing us to restart the data stream in Kafka when the Spark application is expecting data. We select only the *value* column, containing our data in string form. The other columns contain metadata that might be useful in a production environment.

The MongoDB logs show the events associated with our insertions. They contribute to a first control during our test phase.

V. ANALYSIS RESULTS

A. Data Visualization

We wrote SQL and scala scripts with the Zeppelin tool to access the MongoDB database. The SQL queries are primarily useful to extract information and then use the Zeppelin graphical library to visualize the results. Examples of queries allow knowing which badges are currently in use, which is to say currently in use by a badge reader. Figure 3 shows badge usage over a 4-week period. On the x-axis is the number of operations while on the y-axis is the time of use of these badges in reading.

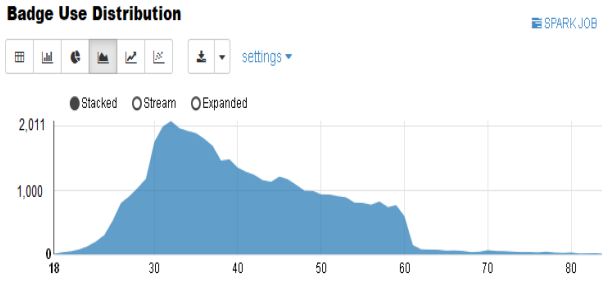


Figure 3. Representation of the time of use of the badges compared to the number of operations made

The NoteBook concept is particularly interesting in the case of simple data visualization. On the one hand, it allows validating the validity of a representation with respect to the expected users, and on the other hand, it ensures that the query language has sufficient expressive power to quickly build this representation.

Figure 4 shows a query to calculate the number of badges that are not in their badge holder.

 ISBN: 978-1-68558-041-4

Once the query is validated, it is easy to include it in a Scala script. The interest is major because the Scala script is compiled, which means that the query is not interpreted for

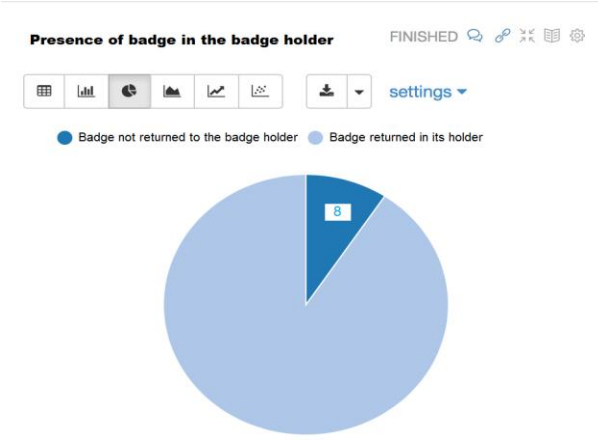


Figure 4. Calculation of the number of badges not returned to the badge holder over a period

each data set but its evaluation is reused for all the date sets, i.e., for each sampling window. We use the Zookeeper tool in conjunction with Spark because it allows more flexibility between data engineers and developers. Several works have shown the interest of such a tool for experience sharing [16]. It is also possible to install additional processors and this is what we did for JSON querying directly on MongoDB. The JsonPath interpreter is not installed by default and there is no such tool in the Zookeeper marketplace. That is why we customized our installation with an ad-hoc development based on the JsonPath library in Scala.

B. Streaming Configuration

Our experiments allowed us to externalize in the configuration of the Spark Extractor component several useful coefficients to manage data sampling. We can cite:

- The duration of a sampling window (ms),
- The overlap time of two consecutive windows (ms),
- The size of the data collected during a window (MB),
- The quota of data not respecting a data format compared to the read data (%).

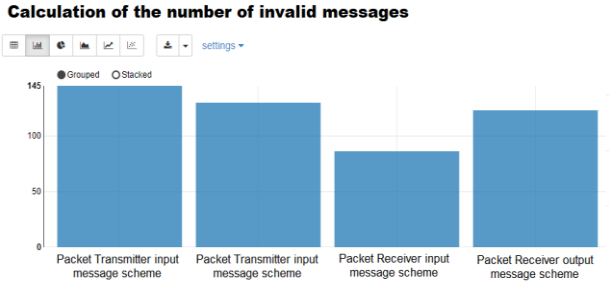


Figure 5. Use of the streaming parameters

An adapted configuration allows a better exploitation of the computing resources, in particular on clusters where the computing time is paying. Even if it was not our case, we share this cluster with other users, and it is advisable not to overload a resource that has become crucial for other simulations. Finally, from our results, we can illustrate these parameters comparisons between several types of configurations (see Figure 5).

VI. CONCLUSION AND FUTURE WORK

We presented our work around the planning of data collection and analysis. We illustrated our work with the analysis of log messages from LoRa WAN transmission tools. We have shown that the use of data schema is an asset to have more accurate collection results. Moreover, we have created highly configurable Big Data analysis components. By precisely defining these parameters, we are able to make the best use of our computational resources and, above all, we have the means to search for the best-defined set of micro-batches.

Among the future work ahead of us, we plan to continue collecting our LoRa WAN message use case with there being other post processing that needs better monitoring. Thus, the use of these business cards are developing and if some sides are humanly delicate to accept, others like the understanding of the uses of resources of the laboratory are important. Indeed, in a policy of maintenance or renewal of materials, not all are equivalent. To prioritize certain actions, it is necessary to understand that all share some tools while others are more specialized. To illustrate this point, we will return to the use of 3D printers, for which we were able to calculate the time of use and thus show a high level of use by a large number of users. This underlines that the need for new functionality, such as a reduction in printing time, is a positive feature for all.

REFERENCES

[1] Z. Lv, B. Hu and H. Lv, "Infrastructure monitoring and operation for smart cities based on IoT system. IEEE Transactions on Industrial Informatics", vol. 16 no. 3, pp. 1957-1962, 2019.

[2] W. T. Hartman, A. Hansen, E. Vasquez, S. El-Tawab, and K. Altaii, "Energy monitoring and control using Internet of Things (IoT) system. In 2018 Systems and Information Engineering Design Symposium (SIEDS)", pp. 13-18, IEEE, 2018.

[3] S. Nocheski and A. Naumoski. "Water monitoring iot system for fish farming ponds. Industry 4.0", 2018.

[4] G. Yang, M. Jiang, W. Ouyang, G. Ji, H. Xie, A. M. Rahmani and H. Tenhunen, "IoT-based remote pain monitoring system: From device to cloud platform. IEEE journal of biomedical and health informatics", vol. 22 no. 6, pp. 1711-1719, 2017.

[5] K. Sabancı, E. Yigit, D. Üstün, A. Toktaş and Y. Çelik, "Thingspeak based monitoring IoT system for counting people in a library. In 2018 International Conference on Artificial Intelligence and Data Processing (IDAP)", pp. 1-6, IEEE, 2018.

[6] J. Wang, M. Chen, J. Zhou and P. Li, "Data communication mechanism for greenhouse environment monitoring and control: An agent-based IoT system. Information Processing in Agriculture", vol. 7 no. 3, pp. 444-455, 2020.

[7] T. M. Bandara, W. Mudiyanselage and M. Raza, "Smart farm and monitoring system for measuring the Environmental condition using wireless sensor network-IOT Technology in farming. In 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA) ", pp. 1-7, IEEE, 2020.

[8] D. Davcev, K. Mitreski, S. Trajkovic, V. Nikolovski and N. Koteli, "IoT agriculture system based on LoRaWAN. In 2018 14th IEEE International Workshop on Factory Communication Systems (WFCS) ", pp. 1-4, IEEE, 2018.

[9] B. Leang, S. Ean, G. A. Ryu and K. H. Yoo, "Improvement of Kafka streaming using partition and multi-threading in big data environment. Sensors", vol.19 no. 1, pp. 134, 2019.

[10] Y. Lou and F. Ye, "Research on data query optimization based on SparkSQL and MongoDB. In 2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES) ", pp. 144-147, IEEE, 2018

[11] P. Bourhis, J. L. Reutter and D. Vrgoč, "JSON: Data model and query languages. Information Systems", pp. 89, 101478, 2020.

[12] T. Dunning and E. Friedman, "Streaming architecture: new designs using Apache Kafka and MapR streams. O'Reilly Media, Inc.", 2016.

[13] B. R. Hiraman, "A study of apache kafka in big data stream processing. In 2018 International Conference on Information, Communication, Engineering and Technology (ICICET) ", pp. 1-3, IEEE, 2018.

[14] M. Armbrust, T. Das, J. Torres, B. Yavuz, S. Zhu, R. Xin and M. Zaharia, "Structured streaming: A declarative api for real-time applications in apache spark. In Proceedings of the 2018 International Conference on Management of Data", pp. 601-613, 2018.

[15] P. Sangat, M. Indrawan-Santiago, and D. Taniar, “Sensor data management in the cloud: Data storage, data ingestion, and data retrieval. Concurrency and Computation: Practice and Experience", vol. 30, no. 1, e4354, 2018.

[16] A. MadhaviLatha and G. V. Kumar, "Streaming data analysis using apache cassandra and zeppelin. IJISET-International Journal of Innovative Science, Engineering and Technology", vol. 3 no. 10, 2016.

 ISBN: 978-1-68558-041-4