Discovering DataOps: A Comprehensive Review of Definitions, Use Cases, and Tools

Kiran Mainali
KTH Royal Institute of Technology
Stockholm, Sweden
e-mail: mainali@kth.se

Lisa Ehrlinger
Software Competence Center Hagenberg GmbH
Hagenberg, Austria,
Johannes Kepler University Linz
Linz, Austria
e-mail: lisa.ehrlinger@scch.at,

Johannes Himmelbauer
Software Competence Center Hagenberg GmbH
Hagenberg, Austria
e-mail: johannes.himmelbauer@scch.at

Mihhail Matskin
KTH Royal Institute of Technology
Stockholm, Sweden
e-mail: misha@kth.se

***Abstract*—Data management approaches have changed drastically in the past few years due to improved data availability and increasing interest in data analysis (e.g., artificial intelligence). The volume, velocity, and variety of data requires novel and automated ways to "operate" this data. In accordance with software development, where DevOps is the de-facto standard to operate code, DataOps is an emerging approach advocated by practitioners to tackle data management challenges for analytics. In this paper, we uncover DataOps from the scientific perspective with a rigorous review of research and tools. As a result, we make the following three-fold contribution: we (1) outline definitions of DataOps and their ambiguities, (2) identify the extent to which DataOps covers different stages of the data lifecycle, and (3) provide a comprehensive overview on tools and their suitability for different stages of DataOps.**

***Keywords*—DataOps; Data lifecycle management; Data analytics.**

I. Introduction

The increasing volume, velocity, and variety of data in recent years opened the possibility for enhanced analytics, e.g., in artificial intelligence applications [1]. Along with these possibilities, companies are putting higher effort into data analytics projects, which are becoming increasingly complex due to data characteristics, sophisticated tools, changing business needs, varied interests among stakeholders, and a lack of a standardized process [2]. While data analytics projects are still struggling with a standardized process, software development has successfully employed DevOps [3], which is an efficient and practical approach for delivering software applications at a higher pace using a combinations of cultural philosophies, practices, and tools [4][5]. There is a recent trend to adopt DevOps practices for data analytics projects [6].While collecting data, DevOps can reduce the effort and time for data retrieval [7]. Furthermore, in the data transformation and analysis, DevOps can maintain and update scripts and manage tools and technologies effectively and collaboratively using a Continuous Integration (CI) server and a central code repository. However, data analytics projects differ from software development in many aspects (e.g., the data and analytics pipeline, stateful data stores, and process controls) and therefore bear more similarities with data integration and business analysis projects [8]. The significant difference is the creation of an analytics pipeline, which copies operational data from business, performs business-rule-based data transformations, and populates the data in a central storage from which analysts can extract business information. This challenge cannot be simply solved by exploiting DevOps practices, but requires a more adjusted approach: DataOps.

In the process of establishing DataOps as a data analytics methodology, people and organizations supporting the concept derived 18 principles of DataOps in the manifesto [9]. The DataOps principle summarizes the best practices, goals, philosophies, mission, and values for DataOps practitioners. The manifesto puts team communication over tools and the process. Experimentation, iteration, and feedback are more important than designing and developing the whole pipeline upfront. Sense of responsibility and cross-functional collaboration is advocated to increase the project efficiency reducing individual soiled responsibilities and heroism.

DataOps is a method to automatically manage the entire data life cycle from data identification, cleaning, integration to analysis and reporting [10]. Its primary goal is business value maximization of data. It borrows proven practices from DevOps in the software development lifecycle. While DevOps is a mature field in software development, DataOps is still in its infancy stage. However, there is very little research to establish DataOps as a methodology.

In 2018, Ereth et al. [11] contributed with a working definition of DataOps. The authors conclude that further research is required to elaborate on this new discipline by investigating the process, related technologies, and tools, as well as the value

 ISBN: 978-1-61208-891-4

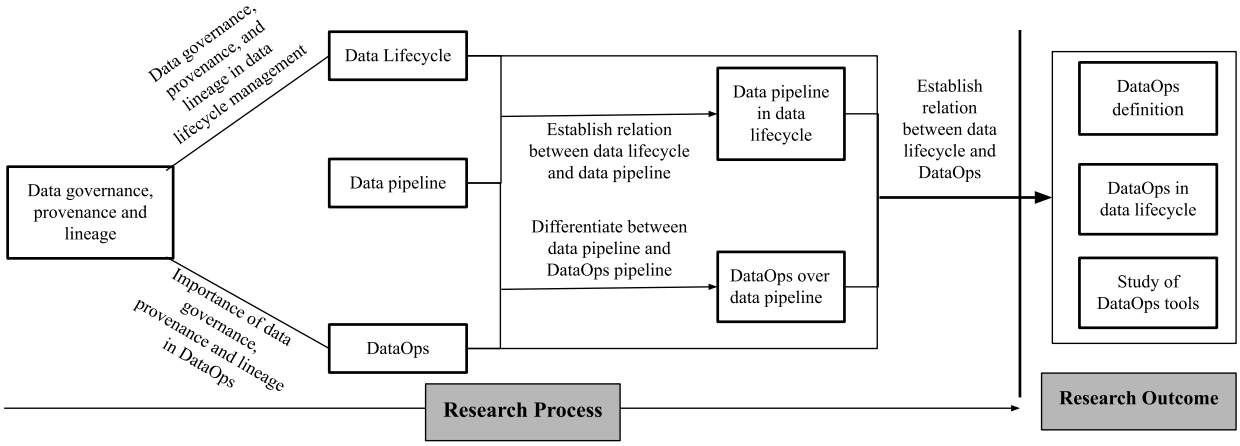


Figure 1. Illustration of the Exploratory Research Method.

proposition DataOps brings to business [11].

In 2020, Raj et al. [12] performed an extensive literature review of DataOps with a focus on the usability. Their paper presents a case study of a large telecommunication company and how their processes evolved by supporting DataOps [12].

Despite the few scientific papers, numerous resources from practitioners are available. Several companies strongly advocate DataOps and deliver a product to support the DataOps principles, e.g., IBM [13], DataKitchen [14], iCEDQ [15], and Eckerson [16]. IBM is a pioneer of the term DataOps and offers numbers of products for different data lifecycle stages (e.g., data collecting, analysis, storage, organization, and publishing) under the umbrella of their cloud service. DataKitchen is one of the leading solution providers for DataOps and is continuously working on establishing DataOps as a methodology with industrial research, e.g., the DataOps manifesto [9] and DataOps implementation guidelines [17]. iCEDQ provides a data monitoring platform and several whitepapers [18] and blogs [19] to help in understanding the implementation of DataOps in practice. Eckerson is a global research and consulting firm and published several reports to define DataOps [20], exploring the use of DataOps [21], and selection criteria for DataOps tools [22], amongst others.

In this paper, we contribute to the discipline of DataOps with a rigorous review of research as well as tools to bridge the gap between theory and practice. The results of the study can be divided into the following three contribution: (1) a summary of DataOps definitions and their ambiguities, (2) an investigation on how DataOps covers the stages of the data lifecycle, and (3) a comprehensive overview on tools and their suitability for different stages of DataOps.

The rest of the paper is organized as follows. Section II presents the research method used, while Section III presents the result. Finally Section IV concludes the paper and discusses possible future work.

II. Research Method

In this work, we investigate DataOps with explorative qualitative research using literature review and online research, illustrated in Figure 1. We started with collecting scientific articles from Google [23], Google scholar [24], ResearchGate [25], IEEE [26], KTH Library [27], KTH-Diva [28], and Semantic Scholar [29], as well as whitepapers and reports from several companies practicing DataOps. In total, we used 71 out of 157 analyzed research articles and 39 out of 112 accessed online resources.

Figure 1 shows the detailed process of research and the outcome of the study. The plain rectangle box represents topics covered, and the line denotes tasks performed to get the result.

III. Results

This section presents the result of our work to (1) define DataOps and point-out ambiguities, (2) investigate DataOps in data lifecycle management and (3) explore state-of-the-art tools for different stages of the data lifecycle.

A. DataOps Definition – What is DataOps?

DataOps is a consequence of three emerging trends: process automation, digital-native companies pressure on traditional industry, and the essence of data visualization and representation of results [30]. There is no commonly agreed definition of DataOps till now. The first time the term DataOps was used in [31] where the importance of executing data analytics task rapidly with ease of collaboration and assured quality outcome in diverse big data and cloud computing environment is discussed. However, the term DataOps gained its popularity only after Andy Palmer's contribution [32], where he describes DataOps as communication, collaboration, integration, and automation enabler practiced with cooperation between data engineers, data scientists and other stakeholders. [33] considers DataOps goal as taking data from the source and delivering to the person, application or system where it produces business value. Some other definitions describe DataOps as "analytic process which spans from data collection to delivery of information after data processing" [34], "develop and deliver data analytics projects in a better way" [43], "is combination of value and innovation pipeline" [35] or "data management approach to improve communication and

integration between previously inefficient teams, system and data" [36].

Our analysis shows that different perspectives inspired DataOps definitions. Some definitions are more goal oriented [37]–[39] while some are activities oriented [38][40] and furthermore some are process and team oriented [6][41][42]. From a goal oriented approach, DataOps is viewed as a process to eliminate errors and inefficiency in data management, reducing the risk of data quality degradation and exposure of sensitive data using interconnected and secure data analytics models. From a process and team-oriented perspective, DataOps is a way of managing activities of data lifecycle with a high level of data governance, collaborating data creators and consumers using digital innovations.

From application perspective, DataOps is a set of practices in the data analytics field that takes proven practices from other industries [43]. It is the combination of proven methodologies that helped to grow other industries: DevOps and Agile methodology from the software industry and lean manufacturing from the automotive/manufacturing industry [10]. DataOps combines the speed and flexibility of Agile and DevOps and quality control of Statistical Process Control (SPC). Agile helps to deliver analytics results in faster ways, DevOps automates the process of analysis and SPC from lean manufacturing tests and monitors the data flow quality in the entire data analytics lifecycle.

DataOps has its own approaches on top of derived processes from other methodologies to tackle the challenges in the field due to the heterogeneity of data analysis projects. Separating the production environment from development gives room for data workers to experiment with the changes and altogether remove the fear of failure. With two different environments, product quality can be assured by continuous testing and cross-environment monitoring. Including customers and other stakeholders in data analytics project sets communication and feedback loop to minimum iteration. With this, changes and improvements in the pipeline can deliver faster results without affecting current pipeline production. Also, role-based task distribution fosters the responsibility of everyone while maintaining the coalition of a team effort.

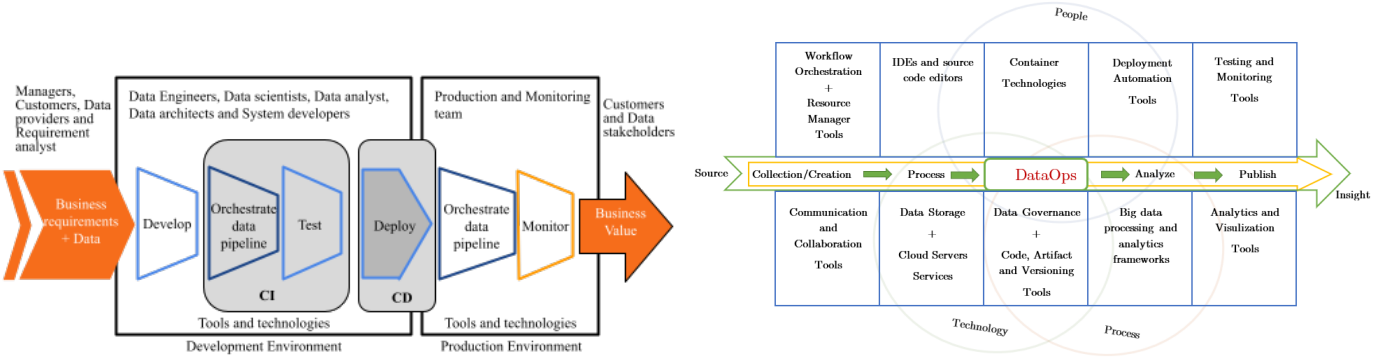
DataOps pipeline (shown in Figure 2a) starts with gathering data and business requirements. Active involvement of managers, data providers, and analysts creates the baseline for pipeline development. Once business requirements and data are finalized, the development of the data pipeline starts. The developed data pipeline is orchestrated by orchestration tools and tested before deploying to the production environment. There could be multiple development environments for each involved data worker. However, the deployment will not be done without assembling all individual work to make the whole pipeline fulfilling all test requirements. Testing and orchestrating of data pipeline will be supported by Continuous Integration (CI) tools and deployment is done through Continuous Development (CD) tools. Deployment task automation reduces the workload of reconfiguration and reworks on the pipeline in another environment. With a combination of CI and CD, data pipeline moves swiftly from the innovation stage to the production stage. In the production phase, pipeline runs in an orchestrated environment as in a development environment. Continuous monitoring follows the pipeline input, performance, and output and cross-validates the monitoring outcomes with test results from the development environment and business requirements. The production team and monitoring teams are responsible for carrying out tasks in a production environment. Teams are composed of people with a different areas of expertise and interests to deliver quality performance. Finally, results will be shared with customers and stakeholders with the expectation of feedback and comments.

Figure 2b illustrates the DataOps ecosystem where various categories of tools aligned in order with people to match the process of converting input to generate insights as output through series of data lifecycle movement in between. Depending on project goal and level of automation, tools and technologies from the stacks are chosen. It is not always necessary to apply all the tools categories listed above. DataOps' primary objective is to deliver quality results in improved time and low cost. If that can be fulfilled by using one or a few tools from the list above, then the project can be delivered with those tools. DataOps is also about continuous improvement, so people working in the project should never give up on experimenting with new technologies and delivering better project results.

1) Ambiguities in DataOps Practices: DataOps is an emerging concept. In recent years information collection and work contributions are progressing in the DataOps through the involvement of DataOps practitioners and enthusiasts. But, there are some prevalent misconceptions in DataOps, which are listed and explained below by observing the industry implementation use cases and scenarios [21][22][35][44] provided by DataOps practitioners:

- **DataOps is just DevOps applied in data analytics.** DataOps is not DevOps for data. It takes best practices from DevOps and Agile methodology and combines with lean manufacturing's SPC and data analytics specific tasks to streamline data lifecycle and provide quality results. Data analytics projects and software development projects have significant differences.
- **DataOps is all about using tools and technology in the data pipeline.** DataOps is not about automating everything using tools and technologies and keeping human involvement away. DataOps advocates a balanced involvement of people along with tools and technology. Communication and collaboration are highly focused on DataOps to turn data into value for all involved parties.
- **DataOps is an expensive methodology.** Acquiring and running different tool always comes with a price. Data analytics projects will cost to an organization, whether they follow DataOps or not. One should compare their investment with the value going to receive in the near future. Furthermore, proper research on tools and technology before implementing on data pipeline can help make informed decision to cut the cost to a minimal.

 ISBN: 978-1-61208-891-4



(a) DataOps pipeline. (b) DataOps framework.

Figure 2. Illustration of DataOps Pipeline and Framework.

- **With DataOps, there is no need for coding.** Without writing code data pipeline task cannot be formed at all. So coding is always a baseline of data analytics projects. With DataOps, even the coding process can be reduced to minimal by reusing and versioning of codes, algorithms and configuration. IDEs and source code editors provide easy writing and debugging of codes.
- **DataOps can only use on data analysis tasks.** DataOps is not just generating reports and delivering fancy charts, templates, bars, and figures. It is about covering the whole data lifecycle from the collection of data to disposal. Moreover, it is not just about covering the data lifecycle; it is also about creating a data-driven organization culture that emphasizes collaboration, communication, transparency, and quality on organizational tasks.
- **DataOps and data pipeline are two different ways of data analytics task propagation.** DataOps is an approach to implement a data pipeline. We apply the DataOps principle and practices while developing and executing data pipelines. Data pipeline with DataOps methodologies is also called DataOps pipeline. DataOps is not an entirely new way of performing data analytics tasks; rather, it redesigns the data pipeline to deliver quality results in a short time with minimal cost and effort.

2) Challenges in DataOps implementation: For an organization to succeed with DataOps, there is a need to consider potential issues. Challenges that need to be considered when implementing DataOps practices are listed below.

- **Changing the organization's culture:** DataOps is all about delivering analytics result faster, and the only way to make it happen is to encourage communication and collaboration across all departments. Data scientists, data engineers, managers, data analysts, system architects, system developers, customers and other data stakeholders all need to come together to break the status quo. DataOps can bring significant change, and for its success, everyone needs to be on board. This includes top executives, IT and business managers, data workers and everyone involved in data analytics project.
- **Innovation with low risk:** DataOps advocates continuous improvement in the product and cycle time, which means lesser time for development, test, and deployment. Teams need to move quickly without compromising quality. Not just quality but also complying with company policies and standards. Automation gives extra space to reduce cycle time by reducing the manual task of testing, monitoring and deployment. With automation on the deployment cycle, there is little time for reviews increasing the risk of missing out details and pieces of information. So initially, it will take time to implement for total confidence in ensuring data and process quality.
- **Cost of DataOps:** The initial cost of introducing new tools and technology, employee training and moving from the old system can be substantial, and it is easy to get discouraged at the beginning when there are no immediate benefits to realize. Nevertheless, in the long run, DataOps will pay off by reducing cycle time and standardizing the analytics product and process quality.
- **Transition from expertise-based team to cross-functional teams:** DataOps succeeds with cross-functional team collaboration and communication. Creating integrated data analytics teams will bring employees together from different departments and with varied expertise to solve a specific problem. Nevertheless, the challenge of structural change is enormous. One should include all related and required members in the team with proper authorities and responsibilities. There should always be a trust-based environment among team members and between analytics teams, management, and customers.
- **Managing multiple environments:** DataOps, with multiple environments, provides freedom of innovation and improvement but also creates the necessity of proper management of those environments. Without an appropriate system management plan, it can quickly go out of hand and create cost and performance exhaustion instead of

 ISBN: 978-1-61208-891-4

benefits.

- **Sharing knowledge:** Tribal knowledge creates a big problem, and DataOps can make it even worst: new tools and technologies, change in processes and execution of data analytics projects in different platforms than before. Without useful documentation or creation of knowledge base, teamwork can be a challenging task to accomplish.
- **Tools and technology diversity:** In DataOps, several tools and technologies are used to accomplish the required tasks. This brings the challenges of maintaining and matching the performances of tools individually and collectively. One tool should not impact and restrict the performance of others. So careful selection of tools is always emphasized.
- **Security and quality:** With multiple environments and team players in project, security and quality is crucial to maintain. Data privacy, system security, data codes and insights quality, data workers and stakeholder's authority should be well described and implemented in DataOps from the beginning. Otherwise, it will be hard to enforce when things go out of hand.

B. DataOps in the Data Lifecycle

DataOps minimizes the analytics cycle time by covering the entire stages of data analysis. The data lifecycle relies on people and tools [10], and DataOps collaborates with people and tools to better manage the data lifecycle. Data analytics pipeline alters data through a series of tasks. Whether it is the ETL (Extraxt, Transform, Load) / ELT (Extract, Load, Transform) pipeline or analysis pipeline, the output will always be different from the input. In data pipelines, one of the challenging tasks is to track data. Data goes through a series of transformations while going from one stage to another. In DataOps, data lifecycle management is unavoidable because of the need to monitor the quality of processes and products. Data governance and data lineage are part of DataOps to assure process and product quality. Quality assurance and the DataOps principle of reproducible and reuse are highly dependent on managing and maintaining data lifecycle change events. Data governance and data lineage is not an easy task to address; it starts with managerial level planning and flourishes with the tools and approach we use to implement our plans.

DataOps applies to the entire data lifecycle [45], from data collection to publishing the result, all data preparation and analysis stages can implement DataOps methodology. It provides the significant advantage of easy management of data lifecycle by applying the intrinsic approach to handle data throughout the analytics cycle. Data pipeline transports data from one stage of the lifecycle to another. DataOps restructures data pipelines and take them out of the black box making them measurable, maintainable through collaboration, communication, integration, and automation. As a result of the restructuring , data lifecycle management becomes more straightforward. DataOps support all stages in the data lifecycle; with the right people and technology in use, data will flow from one stage to another seemingly. With DataOps, a

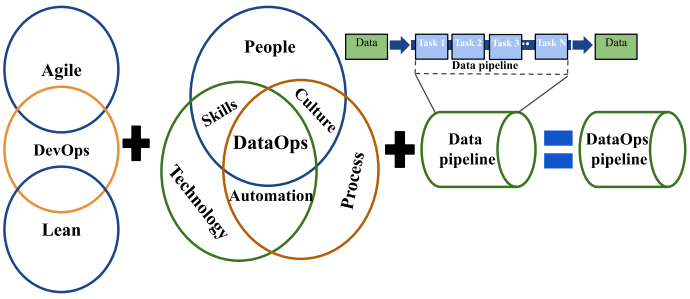


Figure 3. Data Pipeline to DataOps.

published result from the analysis can be trackback to a raw data source, decomposing each transformation task performed over them. DataOps acknowledges the interconnected nature of data engineering, data integration, data quality, and data security [32] and combines all these aspects of data analytics to form an interspace of data movement between data lifecycle stages.

Figure 3 illustrates the DataOps pipeline advancement from data pipeline by adding Agile, DevOps and Lean functionalities with inclusion of people, process, and technology for a designated task. In DataOps, there is no necessity of creating separate pipelines for different stages. Preferably, DataOps utilizes the technical modularity of orchestration, workflow management, and automation tools to provide flexible and customized transformation process when needed.

In the following Section C, the coverage of different stages of data lifecycle in data analytics projects is compared. The comparison is done whether tools used in DataOps can be used in data lifecycle stages: Creation/ Collection, Process, Analyze, Storage, and Publish. These stages of lifecycle are identified by doing extensive literature review of CRUD lifecycle [46], IBM lifecycle [38], USGS lifecycle [39] and DataOne lifecycle [40].

C. Evaluation of DataOps Tools and Technologies

This section provides an overview on the most popular DataOps tools as baseline for further research and to support practitioners in the selection of proper tools for DataOps tasks. Since there are numerous tools with the same features and functionality, it is hard to cover every tool in detail. We picked some of the popular tools and compared them categorically based on the evaluation criteria. Selecting tools and technology in DataOps is a rigorous process and needs detailed research and planning before selecting a particular tool for the designated task [22]. Tools presented in Tables 2-8 are based on their mass user base, relevant features to support the given functionality and popularity of the product in data analytics project execution. Tools presented in the feature-based comparison table are picked after extensive online research by listing and comparing other tools from the same functional categories.

Evaluation Criteria: The criteria for the DataOps tool evaluation and comparison are detailed in Table I and their

 ISBN: 978-1-61208-891-4

selection is based on installation easiness, operation simplicity, integration support to other technologies, and general applicability of the tools. Criteria present a general overview of tools and technologies to non-technical data workers and decision-makers to give a general idea and a starting point to research tools before using them in the project.

TABLE I. EVALUATION CRITERIA

Criteria	Measures
Complexity	Measures how complex is the installation and implementation process of the presented tool. Evaluation is based on the code complexity and dependencies that need to be setup **-HIGH**: Need a high level of coding and configuration to install the product. **-MEDIUM**: Moderate level of coding and configuration required. **-LOW**: Easy to install with no line of code or a few lines of code.
Usability	Measures how simple the tool is to use after the installation, especially for nontechnical data workers. **-HIGH**: Easy to use with little or no technical, coding, or system-related knowledge. **-MEDIUM**: Moderate knowledge of the system, code architecture, or technical detail is required. **-LOW**: High level of technical expertise and/or coding knowledge is required.
Compatibility	Measures the integration capacity of the tool with different operation environments, other tools, databases, data types and/or programming languages. **-HIGH**: Supports a wide range of tools, operation environment, database, data types, and programming languages. **-MEDIUM**: Have some level of support either explicit (in the number of specific tools, languages, databases, data types and/or programming languages declared officially) or have implicit partial support provided through unofficial projects. **-LOW**: Little or no support available.
Application	Provides the information related to the tools' applicability to arrays of projects, data analysis use cases, and industries. **-GENERIC**: Can be used in a variety of projects based on the nature of tools. **-SPECIFIC**: Industry/project-specific usage.
Lifecycle	Lists in which data lifecycle stage the tool can mostly be used.
License	Describes whether the tool is commercial, opensource, freemium, free + commercial and other pricing forms.

We use the following abbreviations in the comparison tables.

Abbreviations for data lifecycle stages: C - Creation/collection; P - Process; A - Analyze; S - Storage; Pu - Publish.

Abbreviations for pricing modules: O - Open-source; Co - Commercial; F - Free; N - Non-profit; Fm - Freemium; U - User based pricing.

1) Categorization of tools and technologies: Tools used in DataOps are categorized in the following functional categories and presented categorically from Table 2 to Table 8 (all references to the tools presented in comparison Tables can be found in the external Dropbox hosted file [47]). The categorization is based on the tools' purpose in the data pipeline. Some tools are uncategorized and kept under "Other tools and technologies" because they do not fall under the first seven categories listed below.

Workflow orchestration tools: Workflow orchestration or pipeline orchestration defines the logical flow of tasks from start to end in the data pipeline. In DataOps, orchestration tools create a logical flow of data analytics task and assemble other tools and technologies, infrastructures, and people to accomplish the job. Several orchestration tools are available with similar design principles targeted to various users and use cases. Choosing them for pipeline workflow management is a thorough job. Orchestration tools include resource provisioning, data movement, data provenance, workflow scheduling, fault tolerance, data storage, and platform integration in the data pipeline [48]. However, all orchestration tools do not have all features inbuilt to support every task in a data pipeline. So, choosing the right orchestration tool is essential to manage tasks in the data pipeline. There has been a practice of developing custom-built workflow orchestration tools for a specific project [49]–[51]. In Table II, comparison of some of the existing popular pipeline orchestration tools is presented by using the comparison criteria presented in Table I.

Testing and monitoring tools: Continuous testing and monitoring are the principal mission of DataOps. With these, performance, quality of input and result, code, and toolchain performance throughout of data pipeline is ensured. Testing and monitoring applies in entire stage of the data lifecycle. In DataOps, testing and monitoring start from the top management by setting the criteria of project quality, and test cases are developed according to the proposed criteria. After the development of test cases and monitoring criteria, suitable existing tools or custom-built test and monitoring framework can be integrated into the data pipeline. Some of the testing and monitoring tools are presented in Table III.

Deployment automation tools: DataOps continuously moves code and configurations from the development environment to the production environment after test cases are satisfied. The deployment automation applied through the process of continuous integration and continuous deployment. Representative tools widely used in deployment automation are presented in Table IV.

Data governance tools: Testing and monitoring are keeping a record of the principles of data governance. Where testing and monitoring are more focused on tracking the whole DataOps pipeline performance measures, data governance is related to data change management and data lineage tracking. Some Tools used in data governance are presented in Table V.

Code, artifact, and data versioning tools: Code, artifacts, and data versioning tools (some presented in Table VI) provide a platform to store different versions of codes, data sets, docker images, and other related documents like logs, user manuals, system manuals, and configurations. With the use of the right tool, accessing and reusing different versions of stored artifacts becomes easier.

Analytics and visualization tools: The importance of visual presentation is always high while demonstrating results. Customers and non-data workers always relish on fined tuned

 ISBN: 978-1-61208-891-4

TABLE II. WORKFLOW ORCHESTRATION TOOLS

Tools	Lifecycles	Complexity	Usability	Compatibility	Applications	License
Airflow	C, P, A	HIGH	MEDIUM	HIGH	GENERIC	O
Apache Oozie	C, P, A	HIGH	MEDIUM	LOW	GENERIC	O
Reflow	P, A	HIGH	LOW	LOW	SPECIFIC	O
Data Kitchen	P, A	LOW	HIGH	HIGH	GENERIC	Co
BMC Control-M	P, A	MEDIUM	MEDIUM	HIGH	GENERIC	Co
Argo Workflows	P, A	HIGH	LOW	LOW	GENERIC	O
Apache NIFI	C, P, A	MEDIUM	MEDIUM	MEDIUM	SPECIFIC	O

TABLE III. TESTING AND MONITORING TOOLS

Tools	Lifecycles	Complexity	Usability	Compatibility	Applications	License
iCEDQ	C, P, A	LOW	HIGH	HIGH	GENERIC	Co
Data Band	P	HIGH	LOW	MEDIUM	GENERIC	O, Co
RightData	S, A, P	MEDIUM	MEDIUM	HIGH	GENERIC	Co
Naveego	C, P, S	HIGH	HIGH	LOW	SPECIFIC	Co
DataKitchen	C, P, S	HIGH	MEDIUM	HIGH	GENERIC	Co
Enterprise Data Foundation	S, A, P	HIGH	LOW	LOW	SPECIFIC	F, N

TABLE IV. DEPLOYMENT AUTOMATION TOOLS

Tools	Lifecycles	Complexity	Usability	Compatibility	Applications	License
Jenkins	C, P, A, S, Pu	MEDIUM	HIGH	HIGH	GENERIC	O
DataKitchen	C, P, A, S, Pu	HIGH	MEDIUM	HIGH	GENERIC	Co
Circle CI	C, P, A, S, Pu	MEDIUM	MEDIUM	MEDIUM	GENERIC	F, Co
GitLab	C, P, A, S, Pu	MEDIUM	MEDIUM	HIGH	GENERIC	O, Co
Travis CI	C, P, A, S, Pu	MEDIUM	HIGH	HIGH	GENERIC	F, Co
Atlassian Bamboo	C, P, A, S, Pu	LOW	HIGH	HIGH	GENERIC	Co

TABLE V. DATA GOVERNANCE TOOLS

Tools	Lifecycles	Complexity	Usability	Compatibility	Applications	License
Apache Atlas	C, P, A, S, Pu	HIGH	MEDIUM	MEDIUM	GENERIC	O
Talend	C, P, A, S, Pu	MEDIUM	MEDIUM	MEDIUM	SPECIFIC	O, Co
Collibra	C, P, A, S, Pu	LOW	LOW	LOW	SPECIFIC	Co
IBM	C, P, A, S, Pu	MEDIUM	HIGH	MEDIUM	GENERIC	Co
OvalEdge	C, P, A, S, Pu	LOW	HIGH	HIGH	GENERIC	Co

TABLE VI. CODE, ARTIFACT AND DATA VERSIONING TOOLS

Tools	Lifecycles	Complexity	Usability	Compatibility	Applications	License
GitLab	C, P, A, S, Pu	MEDIUM	HIGH	MEDIUM	GENERIC	F, Co
Github	C, P, A, S, Pu	MEDIUM	HIGH	MEDIUM	GENERIC	F, Co
DVC	C, P, A, S, Pu	MEDIUM	HIGH	MEDIUM	GENERIC	O
DockerHub	C, P, A, S, Pu	MEDIUM	HIGH	MEDIUM	GENERIC	F, Co

TABLE VII. ANALYTICS AND VISUALIZATION TOOLS

Tools	Lifecycles	Complexity	Usability	Compatibility	Applications	License
Tableau	A	LOW	MEDIUM	HIGH	GENERIC	Co
Power BI	A	LOW	MEDIUM	MEDIUM	GENERIC	Co
QlikView	A	LOW	MEDIUM	LOW	GENERIC	Co

TABLE VIII. COLLABORATION AND COMMUNICATION TOOLS

Tools	Lifecycles	Complexity	Usability	Compatibility	Applications	License
Slack	C, P, A, S, Pu	LOW	HIGH	HIGH	GENERIC	Fm, U
Jira	C, P, A, S, Pu	LOW	HIGH	HIGH	GENERIC	Fm, U
Trello	C, P, A, S, Pu	LOW	HIGH	HIGH	GENERIC	Fm, U

 ISBN: 978-1-61208-891-4

quality results. Data visualization and analytics tools play a big part in understandably presenting results with assured quality. With the support of analytics and visualization tools presented in Table VII, data workers can better communicate results.

Collaboration and communication tools: To better coordinate among team members, communication and collaboration tools (some presented in Table VIII) are necessary. Tools can be simple, from email applications to advance communication tools that have fancy features to automate and record most of the routine tasks.

Other tools and technologies: Other tools and technology include containers technology, resource managers, data storage services, IDEs and source code editors, cloud servers and Big data processing and analytics frameworks. All tools and technologies are integrated among or with above presented tools where as Big data and analysis framework can also be used independently. In [47], popular tools and services under the other tool and technologies section are listed by categorically dividing them to present general use of such tools and technologies in DataOps.

IV. Conclusion

Since the rise of the term DataOps, significant contribution to its definition and practical applications can be observed. DataOps enthusiasts collaborate to create a common principle for uniformly applying the methodology in the heterogeneous data operation environments. Despite all these efforts, still certain ambiguities remain in the applicability of DataOps due to the diverse nature of the data analysis process. Data analysis itself is a broad field, where numerous tools, approaches, and technologies can lead to the same result. However, DataOps advocates collaboration, quality control, and fast delivery of analysis tasks by extending proven DevOps methodology from SDLC as well as Agile and Lean Manufacturing's SPC. With these three reference methodologies and the advantage of existing tools and technologies, DataOps is continuously evolving to an efficient and reliable methodology for data management.

When implementing DataOps principles, it is key to select the right tool for a given use case. DataOps tools and technologies can be used in several stages of the data lifecycle based on their functionalities. Some tools are particularly useful for certain stages, e.g., analysis and visualization tools. Others, like communication and collaboration tools, are independent of the data lifecycle. Deployment automation tools, data governance tools, code, artifact, and data versioning tools, containers, resource manager, IDE and source codes editors, and cloud servers are used across all stages of the data lifecycle. However, workflow orchestration and testing and monitoring tool support over data lifecycle stage depends on their features. Some tools can provide support to the entire pipeline process, whereas others are specifically tailored to certain tasks. Furthermore, big data processing and analysis frameworks provide a complete solution even though the focus is more on data processing and analysis.

In summary, there are numerous tools available on the market with similar features and functionalities. This paper compares their features with respect to the data analysis lifecycle and therefore supports a practitioner in selecting the proper tool for a given use case. Using suitable tools allows to cover all stages of the data lifecycle with the DataOps methodolgy. Eventually, every stage of the data lifecycle (i.e., from data collection, processing and analyzing, to publishing) can be covered by one or a combination of tools and technologies. It is up the the DataOps engineer and to the respective use case which combination of tools are most suitable for which tasks.

This paper focuses on the exploration of existing concepts in DataOps and aims at sheding light to the large variety of tools and technologies. Thus, it acts as starting point for further research on the successful implementation of the DataOps methodology. For future work, we plan to experiment with DataOps by implementing it in different data analysis projects and to validate (1) on the one hand the efficacy of the methodology itself, and (2) on the other hand the performance of different tools for different use cases. The second step can be achieved by implementing tools for the same functionality and to test their performance on a specific industry use case. We also claim that a compatibility rating (based on combined performance when used together in data analytics tasks) of one tool from one functional group to other functional groups would help DataOps practitioners make informed decisions.

Acknowledgement

The research in this paper has been funded by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by FFG and is further supported by the EC H2020 project *"DataCloud: Enabling the Big Data Pipeline Lifecycle on the Computing Continuum"* (Grant nr. 101016835).

References

[1] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.

[2] H. Baars and J. Ereth, "From data warehouses to analytical atoms - The internet of things as a centrifugal force in business intelligence and analytics," in *24th ECIS 2016*. Association for Information Systems, 2016.

[3] L. Fischer *et al.*, "AI System Engineering—Key Challenges and Lessons Learned," *Machine Learning and Knowledge Extraction*, vol. 3, pp. 56–83, 2021.

[4] M. Artac, T. Borovssak, E. Di Nitto, M. Guerriero, and D. A. Tamburri, "DevOps: Introducing infrastructure-as-code," in *Proceedings - 2017 IEEE/ACM 39th ICSE-C 2017*. IEEE, jun 2017, pp. 497–498.

[5] L. E. Lwakatare *et al.*, "DevOps in practice: A multiple case study of five companies," *Information and Software Technology*, vol. 114, pp. 217–230, oct 2019.

[6] Z. Zhang, "DevOps for Data Science System," Master's thesis, KTH, 2020.

[7] K. Kontostathis. (2017) Collecting Data, The DevOps Way. [Retrieved: 2021-07-30]. [Online]. Available: https://insights.sei.cmu.edu/devops/2017/11/collecting-data-the-devops-way.html

[8] S. Ward-Riggs, "The Difference Between DevOps and DataOps – Altis Consulting," [Retrieved: 2021-08-22]. [Online]. Available: https://altis.com.au/the-difference-between-devops-and-dataops/

[9] The DataOps Manifesto, "The DataOps Manifesto ," [Retrieved: 2021-08-19]. [Online]. Available: https://www.dataopsmanifesto.org/

 ISBN: 978-1-61208-891-4

[10] C. Bergh, G. Benghiat, and S. Eran, *The DataOps Cookbook*, 2nd ed., 2019.

[11] J. Ereth, "DataOps – Towards a Definition," in *LWDA*, 2018, pp. 104–112.

[12] A. Raj, D. I. Mattos, J. Bosch, H. H. Olsson, and A. Dakkak, "From Ad-Hoc data analytics to DataOps," in *Proceedings - 2020 IEEE/ICSSP*, 2020, pp. 165–174.

[13] "Dataops — ibm," [Retrieved: 2021-08-19]. [Online]. Available: https://www.ibm.com/analytics/dataops

[14] "Datakitchen — the complete enterprise dataops platform," [Retrieved: 2021-08-19]. [Online]. Available: https://datakitchen.io/

[15] "Dataops platform — etl testing and monitoring — icedq," [Retrieved: 2021-08-19]. [Online]. Available: https://icedq.com/

[16] "Eckerson group - data analytics consulting research," [Retrieved: 2021-08-19]. [Online]. Available: https://www.eckerson.com/

[17] DataKitchen, "DataOps in Seven Steps," 2017, [Retrieved: 2021-08-23]. [Online]. Available: https://medium.com/data-ops/dataops-in-7-steps-f72ff2b37812

[18] H. Crocket, "Fundamental Review of the Trading Book: Data Management Implications," iCEDQ, Tech. Rep., 2018.

[19] S. Gawande, "DataOps Implementation Guide," iCEDQ, Tech. Rep., 2019.

[20] J. Ereth and W. Eckerson, "DataOps: Industrializing Data and Analytics Strategies for Streamlining the Delivery of Insights," Eckerson Group, Tech. Rep., 2018.

[21] W. Eckerson, "Best Practices in DataOps: How to Create Robust, Automated Data Pipelines," Eckerson Group, Tech. Rep. June, 2019.

[22] W. W. Eckerson, "The Ultimate Guide to DataOps: Product Evaluation and Selection Criteria," Eckerson Group, Tech. Rep., 2019.

[23] "Google," [Retrieved: 2021-04-03]. [Online]. Available: https://www.google.com/

[24] "Google Scholar," [Retrieved: 2021-04-03]. [Online]. Available: https://scholar.google.com/

[25] "ResearchGate," [Retrieved: 2021-04-03]. [Online]. Available: https://www.researchgate.net/

[26] "IEEE," [Retrieved: 2021-04-03]. [Online]. Available: https://ieeexplore.ieee.org/Xplore/home.jsp

[27] "KTH Library," [Retrieved: 2021-04-03]. [Online]. Available: https://www.kth.se/en/biblioteket

[28] "KTH-Diva," [Retrieved: 2021-04-03]. [Online]. Available: https://kth.diva-portal.org

[29] "Semantic Scholar," [Retrieved: 2021-04-03]. [Online]. Available: https://www.semanticscholar.org/

[30] M. Stonebraker, N. Bates-Haus, L. Cleary, and L. Simmons, *Getting Data Operations Right*, 1st ed., R. Roumeliotis and J. Bleie, Eds. O'Reilly Media, Inc., 2018.

[31] L. Lenny, "3 reasons why DataOps is essential for big data success — IBM Big Data and Analytics Hub," jun 2014, [Retrieved: 2021-08-12]. [Online]. Available: https://www.ibmbigdatahub.com/blog/3-reasons-why-dataops-essential-big-data-success

[32] A. Palmer, "From DevOps to DataOps - DataOps Tools Transformation — Tamr," may 2015, [Retrieved: 2021-08-24]. [Online]. Available: https://www.tamr.com/blog/from-devops-to-dataops-by-andy-palmer/

[33] E. Jarah, "What is DataOps? — Platform for the Machine Learning Age — Nexla," [Retrieved: 2021-08-01]. [Online]. Available: https://www.nexla.com/define-dataops/

[34] "DataOps and the DataOps Manifesto — by ODSC - Open Data Science — Medium," 2019, [Retrieved: 2021-07-30]. [Online]. Available: https://medium.com/@ODSC/dataops-and-the-dataops-manifesto-fc6169c02398

[35] DataKitchen, "DataOps is NOT just DevOps for data," 2018, [Retrieved: 2021-08-25]. [Online]. Available: https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7

[36] G. Anadiotis, "DataOps: Changing the world one organization at a time — ZDNet," 2017, [Retrieved: 2021-08-18]. [Online]. Available: https://www.zdnet.com/article/dataops-changing-the-world-one-organization-at-a-time/

[37] A. Wahaballa, O. Wahballa, M. Abdellatief, H. Xiong, and Z. Qin, "Toward unified DevOps model," in *Proceedings of the IEEE ICSESS*. IEEE Computer Society, nov 2015, pp. 211–214.

[38] IBM, "Wrangling big data: Fundamentals of data lifecycle management," *IBM Managing data lifecycle*, 2013.

[39] J. L. Faundeen *et al.*, "The United States Geological Survey Science Data Lifecycle Model: U.S. Geological Survey Open-File Report 2013–1265," Tech. Rep., 2013.

[40] S. Allard, "DataONE: Facilitating eScience through Collaboration," *Journal of eScience Librarianship*, vol. 1, no. 1, pp. 4–17, 2012.

[41] J. Densmore, *Data Pipelines Pocket Reference*, 1st ed., 2020.

[42] B. Plale and I. Kouper, "The Centrality of Data: Data Lifecycle and Data Pipelines," in *Data Analytics for Intelligent Transportation Systems*. Elsevier Inc., apr 2017, pp. 91–111.

[43] S. Gibson, "Exploring DataOps in the Brave New World of Agile and Cloud Delivery," Tech. Rep., 2020.

[44] A. Palmer, M. Stonebraker, N. Bates-Haus, L. Cleary, and M. Marinelli, *Getting DataOps Right*, 1st ed. O'Reilly Media, Inc., 2019.

[45] Margaret Rouse, "What is DataOps (data operations)? - Definition from WhatIs.com," 2019, [Retrieved: 2021-08-25]. [Online]. Available: https://searchdatamanagement.techtarget.com/definition/DataOps

[46] X. Yu and Q. Wen, "A view about cloud data security from data life cycle," in *2010 International Conference on Computational Intelligence and Software Engineering, CiSE 2010*, 2010.

[47] M. Kiran, E. Lisa, H. Johannes, and M. Matskin, "Comparison of dataops tools and technologies," [Retrieved: 2021-08-21]. [Online]. Available: https://www.dropbox.com/s/9nqo3r72ce7nix5/DataOps_tools_ComparisonKiran.pdf

[48] M. Barika, S. Garg, A. Y. Zomaya, L. Wang, A. V. Moorsel, R. Ranjan, S. Garg, L. Wang, and A. Van Moorsel, "Orchestrating big data analysis workflows in the cloud: Research challenges, survey, and future directions," *ACM Computing Surveys*, vol. 52, no. 5, 2019.

[49] Y. D. Dessalk, "Big Data Workflows: DSL-based Specification and Software Containers for Scalable Executions," Master's thesis, KTH, 2020.

[50] H. Chen, J. Wen, W. Pedrycz, and G. Wu, "Big Data Processing Workflows Oriented Real-Time Scheduling Algorithm using Task-Duplication in Geo-Distributed Clouds," *IEEE Transactions on Big Data*, vol. 6, no. 1, pp. 131–144, oct 2018.

[51] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

 ISBN: 978-1-61208-891-4