Enterprise Data Solution Leveraging Data Warehousing for Big Data Veracity at Saudi Aramco

Muhammad Shehryar Khakwani
Upstream Database Services Division
Saudi Aramco
Dhahran, Saudi Arabia
e-mail: muhammad.khakwani@aramco.com

***Abstract*— This paper deals with the challenge faced by large organizations of establishing Big Data veracity, for maximizing their return on investment. Large enterprises, whether commercial or government, are collecting data at an unprecedented rate. Nowadays, data entry is not limited to traditional back office transactions; increasingly, data is gathered from sensors installed in physical assets and transmitted in real-time over large networks. Data technology professionals approach this new wealth of data at various levels defining innovative techniques and applying them in areas of data storage, data warehousing, data mining, semantic logic algorithms, complex event processing etc. For business stakeholders, the benefit is not in the increasing speeds of data acquisition, nor in the variety of data gathered. Return on investment is realized when the data is transformed into information which decision-makers can trust for making operational and analytical decisions, giving the organization a competitive advantage. Data Warehousing plays a critical role in assessing Big Data veracity, identifying problem areas, and building the trust needed for decision-making.**

Keywords- Big Data; Veracity, Data Warehouse; Oil and Gas.

I. ENTERPRISE DATABASES

Enterprises today retain more data than they did a decade ago. Databases supporting large enterprises no longer obtain data only from traditional back office data entry systems. Businesses realized the competitive advantage that can be provided by useful and timely information, and fuelled the drive for building efficient data gathering systems. These systems succeeded in streamlining and automating data gathering, in many cases eliminating manual data entry, providing businesses quicker data access, thereby giving rise to Big Data. Gartner Inc. defines Big Data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [1]."

Enterprises moved towards a host of best-of-breed technologies, adopting ones best suited to address a particular need, and in their wake left large volumes of data, in varying formats, arriving at various frequencies. The resulting complexity creates new data management issues for businesses to deal with today. More companies than ever want practical application of Big Data concepts to data sets which are too large to handle with traditional systems [2]. Businesses serious about successfully extracting value from Big Data initiatives will no longer have the luxury of treating data stores with outdated data management policies. Successful enterprises will incorporate policies and practices for dealing with Big Data as part of their comprehensive data governance strategies.

This paper is based on real occurrences in the industry. It describes the situation at Saudi Aramco before Big Data, the introduction of real-time data gathering which resulted in the arrival of Big Data, the problem posed by data veracity, and the critical role Data Warehousing played in dealing with it. It concludes with the use of data warehousing which provides a feasible and implementable solution towards establishing veracity for Big Data.

II. BIG DATA VERACITY CURRENT SITUATION

The industry today is struggling with Big Data for its volume and velocity, and veracity has emerged more recently. It is only recently that the big data is characterized not only along the three "V"s, volume, variety and velocity, but also along a fourth "V", veracity [3]. IBM recognizes this as establishing trust presenting a huge challenge as the variety and numbers of source grows [4].

III. SAUDI ARAMCO UPSTREAM DATA

Saudi Aramco has produced oil since the 1930s, and is currently the world's prevalent oil producer. It remains committed to serving the world's energy needs, and maintaining its standing among the leading energy companies. In order to achieve its objectives, conduct safe operations, and optimize production strategies, Saudi Aramco committed to investing in Information Technology several decades ago.

Production and Reservoir engineers have been gathering and analyzing scientific data for decades at Saudi Aramco. Data acquisition can be categorized in two broad categories, traditional data gathering, and more recently, real-time streaming data gathering.

 ISBN: 978-1-61208-334-6

A. Traditional Data Gathering

Oil companies have collected data to manage their upstream assets, and Saudi Aramco has been on the forefront of such efforts. Engineers require geo-scientific data such as fluid production rates, well and formation pressure readings, and rock properties to make decisions for harvesting hydrocarbons properly. This data is typically acquired by production engineers conducting operations on wells in oilfields, using various gauges to record measurements. The data gathered during operations is entered in a master database, where it is verified for accuracy by ranking engineers. An engineer compares the readings taken to previous measurements, checks nearby wells, and takes into account subsurface factors before approving the values for use.

These processes for gathering and validating data have been in place for years, and have matured over time. Tried and tested, dependable systems upload and retrieve data. Consequently, the trust, or for lack of a better term, *feel* for these values, has also been intrinsically built over the years.

Geoscientists and decision-makers who rely on geo-scientific data to make critical decisions feel far more comfortable when the data they are considering has been reviewed, validated, and signed-off by a qualified engineer. This is especially true when dealing with scientific data related to subsurface measurements; earth can be quite finicky to deal with at the best of times!

This sentiment of having trust in numbers before taking appropriate action is not surprising; I myself trust medical lab test results for my annual checkup after my doctor has looked them over, considered similar tests from years past, my condition at the time, and given his approval for the lab work.

The key point here is that trusting data quality, accuracy, reliability, i.e., its *veracity*, is built into processes for traditional data gathering.

B. Real-Time Streaming Data Gathering

Real-time monitoring of well performance is a complex, multi-departmental and very expensive undertaking at Saudi Aramco [5]. Real-time data gathering in the digital oilfield is based on the premise that getting information faster will alert to potential problems earlier, enable timely intervention, thereby saving costs, and inevitably resulting in safer operations.

For oil and gas producers, real-time streaming data is gathered during drilling and producing operations. Sensors may be installed on surface, for example on pipelines to measure the rate of fluid passing through it, or installed deep inside oil and gas wells to check subsurface temperature and pressure at reservoir conditions. These sensors provide a constant data feed for the ever-changing big picture of a modern day digital oilfield.

This real-time, or *near* real-time, monitoring of assets generated a lot of interest within the user community. There was a real push from the business to unlock the potential from getting information faster.

IV. DIGITAL OILFIELDS GENERATE BIG DATA

Real-time data is transmitted from oilfields as a continual stream, and consumed by software systems for alerting, reporting, analysis, and ideally providing a visual up-to-the-minute representation. As more and more oilfields were equipped with sensors, data started to flow in steadier, stronger streams; Big Data had arrived. Tens of thousands of readings started streaming in from sensors installed across thousands of assets.

A. Determining Data Properties

To establish whether indeed it is Big Data, let us examine the properties of incoming data. The data is machine generated and arrives as key-value pairs at a very high frequency, so it has high *velocity*. The numerous types of sensors transmit different readings such as reservoir pressure, flow rates, temperature etc., so it has a diverse *variety* of data types. At Saudi Aramco, this activity certainly generates an increasingly large *volume* of data.

B. Determining Critical Success Factors

Projects which introduce new technology, with an aim to transform traditional business practices, normally encounter unanticipated problems. Success requires recognizing problems early, and reacting effectively. This holds true when executing Big Data projects in larger, mature enterprises. Unforeseen challenges will need to be addressed during project execution.

1) Technical Factors

For Information Technology (IT) professionals, focus quickly shifts to IT related concerns like calculating storage capacity for incoming streams, efficiently indexing key-value pairs, and minimizing retrieval time for display on large monitors. While these and other activities like procuring the right hardware, accurately estimating and accounting for scaling volumes of data, configuring the systems to optimally run with low latency, are all important factors, they do not, by themselves, declare a successful Big Data implementation.

2) Business Factors

It is vital that success factors be defined from a business perspective. Our measure of success is whether the engineers and geoscientists are able to utilize the information, transform their processes, conduct better analyses, and add value to the bottom line.

C. Problems Encountered

The first set of problems to overcome tended to be technical in nature. The networks needed to be extended to reach all the assets; oilfields tend not be in the most hospitable areas. The bandwidth capacity needed to be upgraded to handle large volumes of incoming data, and clusters of servers procured to receive and store the data.

Careful capacity planning, savvy cost estimates, and forthright communication with management to secure funding are essential parts of providing a satisfactory technical infrastructure. These are reasonably predictable issues, and experienced project managers are able to address these given adequate resources.

 ISBN: 978-1-61208-334-6

Where Big Data projects differ are when dealing with uncertainty and dealing with earth sciences compounds the problem. Uncertainty is an undeniable reality when dealing with Big Data. Managing uncertainty and establishing trust are key to extracting value from Big Data [3].

The problem which proved to be critical turned out to be an intangible one. The single most important factor that lies between success and failure, and the hardest to achieve, turns out to be data veracity: *establishing trust*.

V. PROBLEMS DETERMINING VERACITY IN DATA

It soon became evident that real-time data could not be treated as "the same thing a lot faster". Though Big Data was similar in its look and feel to conventional data, it needed to be treated differently.

Traditional data collection allows engineers to study and validate the data, determine its usefulness and only keep the most reliable data. Those using the data trust its quality and base their decisions on it. It is not possible to give the same level of care to high velocity data. It is unreasonable to continue to try and monitor that amount of data using manual processes.

Following are a few select examples from real life which show the different types of problems encountered when dealing with data veracity in real-time:

A. Validity Domains and Ranges

One of the first data validation actions when dealing with data is to establish validity ranges for incoming values. Valid value domains, and ranges, were determined after studying existing patterns of data gathered traditionally. Subject matter experts were consulted and acceptable values defined. For example, fluid pressure readings should lie within 1,500 to 4,000 psi for flowing wells.

In reality, it was never as simple as an absolute range. What the users really wanted were ranges which were put in context after considering various factors. For instance, consider the geographical locations of wells, take into account the preceding 24 hours, examine some other relevant factors and then apply a range of values for determining validity. These were all the checks that the engineer carried out when taking readings manually.

For Big Data this can be problematic since the *volume* of data arrival can prevent extensive checking – if everything is not checked very quickly, a lot more data is waiting in the queue. At a technical level one has to be very careful in writing extremely efficient code, and make smart choices about what can be verified in a very short period of time.

B. Good Spikes Bad Spikes

In general values that are out of a given range are excluded when reporting. The sensors are installed in harsh conditions, above and below the surface; they may need to be re-calibrated etc. However, what may look erroneous at first, may turn out to be correct. While conducting a data review for streaming data with senior field engineers, a set of wells was chosen and data analyzed. An analyst noticed some spikes in production rates and pointed them out. The field engineers asked for the time of day when the spikes occurred, and then mentioned that the wells during that time were being adjusted and tested for choke settings which would have caused these spikes in rates. A choke is a heavy steel nipple inserted in production tubing used to restrict flow and control pressure [6]. Had these readings been taken through traditional methods, the engineer taking the readings would have been aware of the operations being conducted and taken that into account when validating the data.

Clearly there is a dependency between production rates and choke settings on a well. If the well has sensors for choke settings, then there is a dependency between streaming data readings. If the well is sending production rates only, then engineers must provide the data for choke settings using traditional gathering means. However, by the time data is entered by the engineer using traditional methods, it may be too late for verifying the streaming high *velocity* data.

C. Conflicting Values

In situations, similar to the ones mentioned above, where there is a dependency, applying verification rules can be problematic. There are several instances of inter-dependent streaming values. Consider two sensors, one indicating whether the well is flowing, and another measuring the fluid rate. The majority of the time these will be consistent. There are cases when one indicates the well is flowing, and the other shows a production rate of zero. Which one is now correct? Which sensor value do you trust?

For a decision-maker who wants to determine the volume of fluid produced and take appropriate action, this poses a confidence problem. Increasing or decreasing volumes of fluid require different actions. Conflicting results immediately cause a loss of confidence in the entire data set. Yes, there are techniques where running mathematical algorithms, or checking a third or fourth set of variables determines the veracity among the conflicting variables, but decision-makers hesitate before authorizing expensive action if results are murky. This is understandable especially where expensive decisions are at stake. Consider this: would you make an investment and buy a stock if you were told the streaming price you see is 90% correct, only a 10% chance that it is not the actual stock price?

D. Suspiciously Correct

Another case which crops up when dealing with Big Data shows values that seem perfectly valid, fall within defined validity ranges, but raise suspicion. These may come from a sensor which is transmitting the same value for a period of time, or the value is within a very narrow range over an extended period. Typically this unsettles the data consumer, because he is not sure whether the sensor is malfunctioning or 'stuck' and transmitting the same pattern repeatedly. It is hard to distinguish if the fluid rate really did remain constant, or whether the sensor transmitting data malfunctioned during the period of time. If the values are repetitive for a short period, that may be acceptable, but if it is absolutely constant for a longer period of time then something is likely wrong. When determining veracity, it is difficult to determine exactly when incorrect readings began being transmitted?

 ISBN: 978-1-61208-334-6

How far back in time does the problem go with readings from a sensor, and whether to fix the data or discard the readings?

E. Multiple Truths

As strange as this statement may sound, real life deals with multiple truths. When dealing with the scale of enterprise level data, the Big Data picture is usually composed of smaller pieces of information obtained from different data stores. Resolving these differences does not necessarily identify only one trusted source.

For instance, analyzing data elements from a well, and cross checking against spreadsheets and charts recorded by engineers, does not show one source always being correct over the other. There are cases where the sensors showed conditions which the engineers missed, and vice versa where the engineers had more accurate information and better observations.

In one case, while visiting engineers in the field for this project, sensors were transmitting very low fluid rates. The field engineer saw this and instructed the values be ignored, since the well had been shut-in the previous day, and therefore the fluid flow rate should be recorded as zero. In other cases, sensors showed more accurate choke setting information than data entered manually.

Clearly establishing criteria for acceptance is the key to data veracity. In real life, especially when dealing with large databases which input data from a *variety* of sources, defining a comprehensive set of rules which can always be applied programmatically is not straightforward.

VI. DATA WAREHOUSING– CRITICAL FIRST STEP

Large enterprises deal with various facets of data management simultaneously. Problems do not lie neatly within a single Data Governance domain such as Data Ownership, Data Quality, Data Mining, Data Integration, or High-volume Streaming. An enterprise solution must draw on various disciplines within data management to provide an effective solution. For a large enterprise like Saudi Aramco, spread out over a vast geographical area, with physical and data assets handled by multiple organizations, the challenge was to gain an effective overall understanding.

Data warehousing provides an all-inclusive view of data. The introduction of a data warehouse into the Big Data architecture fulfilled a critical need to bring disparate data sets together in order to establish veracity.

Data warehousing provides an overall understanding of data associations by placing facts together and along a consistent time dimension. A data mart which pulls data from various sources and lets the users check for accuracy goes a long way towards dealing with data veracity. Fig. 1 shows a conceptual view of a data warehouse bringing together real-time and traditionally collected mastered databases to provide a unified view.

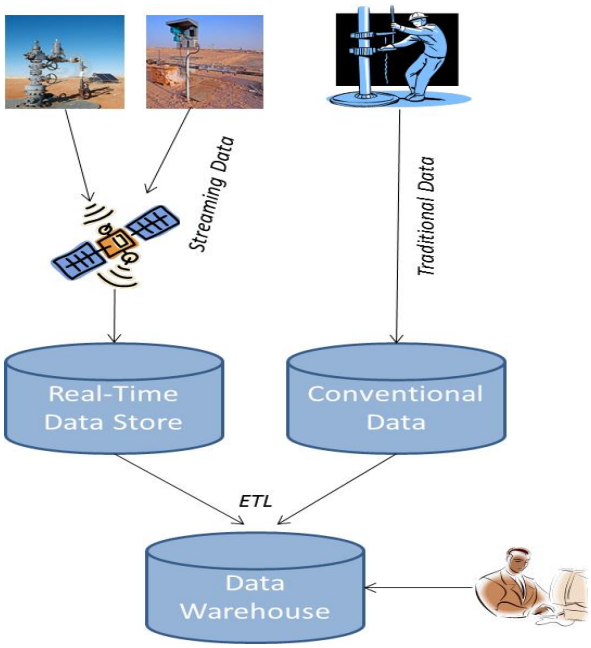

Figure 1. Consolidate data sources for understanding

A. Data Transformation: Smoothing Out Time

At first look, data warehousing seems exactly the wrong approach to the problem. Data warehouses are centered on data cubes with well-defined dimensions, where time plays a crucial role in setting up a data warehouse. When dealing with big projects generating big data arriving from multiple data sources, the time grain is completely askew because the *velocity* of the two data sources (conventional and real-time) is so different. In large industrial applications, data origins spanning multiple organizations cannot be taken in isolation. The two data sets need to be considered to supply context, from a user's perspective.

The time problem is tackled by agreeing on a common baseline for time on the fact table. In our case, it turned out to be a day since users wanted to construct a daily picture for each well. These meant taking readings coming in at various intervals; some arriving every few minutes, others ever few seconds, and aggregating them into a daily summary in the data warehouse.

A key point here is to include only those factors that establish *veracity* and are utilized for analysis. It would be exhaustive to do this for every type of reading. For instance only flow rates, pressures and choke settings were considered, and many other readings like pump vibrations and temperature sensor readings were ignored. Many data elements do not add value to the process of establishing veracity. It is very important to consult subject matter experts from users for determining what is important to gain their confidence – *remember the end goal is to have them use this for better decision-making*.

We defined our fact table to contain information for every calendar day, complementing each streaming value aggregate with the latest values from the master data store, and essentially constructed a "big picture of big data".

 ISBN: 978-1-61208-334-6

B. Advantages of Data Warehouse

There are several advantages to introducing a data warehouse in the architecture and leveraging it to gain a better understanding of Big Data. These include:

- Consolidating Information: The advantage of a Relational Online Analytical (ROLAP) design is that it puts the relevant data *on one row* in the fact table. The Extract Transform Load (ETL) which populates the data warehouse processes data for each well, every day, consolidating data collected from sensors displayed side by side with the latest values available from traditional data sources.
- Consistent Timeline: When ETL constructs a daily picture, it solves a major problem of what to do with data collected over time using different methods. Enterprises that have collected data over years, and choose to modernize sanctioning mega-projects using the latest technology cannot simply discard the values they obtained earlier. All valid data collected over time must be made available to users. The data warehouse can construct data from when it first becomes available in the traditional data store, and continue to add information from when the wells are instrumented to provide streaming data.
- Consolidating Sources: Once the data is available in a data warehouse, it becomes much easier to spot data anomalies. Data from different sources, collected by multiple professionals, and under the ownership of various organizations is now available in one place to look over. This goes a long way towards establishing data veracity.
- Perform Analysis: Having the data in a data warehouse lends itself naturally to running analytical queries. Aggregated values from streaming data which seem suspect are easily picked up, and when necessary one can drill down to the actual readings and analyze a manageable set of information. For instance, it is easy to run SQL queries and check for pressure drops over a threshold, say 5 or 10% in daily averages and look closely at the days where data patterns do not align. The advantage gained is that while one can scan real-time data stores, base lining and transforming data into a warehouse makes it far more manageable. This makes it easier to scan for missing data, and search for patterns where data does not fit with the rest of the readings.

C. Visual Representation

If a picture is worth a thousand words, then visualization must be worth at least a thousand data points. There are several tools available which allow quick visualization using charts and graphs to analyze data in data warehouses. Having the data in one unified data source makes it possible for applications to perform better data analysis and generate comprehensive reports readily. Reporting tools do not have to run exclusively either on streaming, or master data stores, placing the onus on the user to merge the results. Instead reports are built on a correlated data set, stored in fact tables in a data warehousing structure, and presented as a unified view to the end user. Fig. 2 shows a visual representation of streaming data (shown in green, and traditionally gathered data shown in red) placing them next to each other. This makes it easier to identify the outlier scattered values, and those that should be considered.

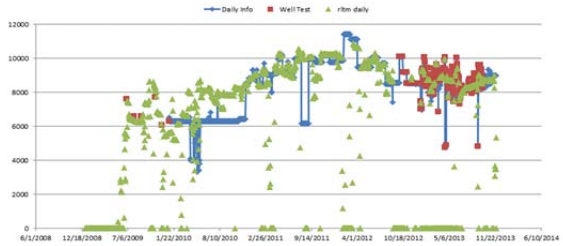

Figure 2. Plotting Traditional and Streaming Data[7]

VII. FIXING PROBLEMS SHOWN BY WAREHOUSING

Data warehousing will make it easier to identify, but not resolve the data issues. Business does not get the intended return on investment until the underlying problems are fixed. When dealing with Big Data, getting the data actually fixed turned out to have its own peculiarities. Some of these characteristics of Big Data fixes include:

A. Technical Challenges

Applying a data fix on voluminous large data sets poses technical problems. Unlike traditional data where incorrect values are updated easily to reflect the correct value, typically, one does not change or update streaming data values generated from sensors, even if they are wrong. They are left alone; instead validation or logic is applied to determine which values to ignore. Sometimes even the validation logic is modified. It does not happen often, but occasionally, engineers do change their minds. If at first they allowed pressure values that fell within a range of 3,000 to 3,500 psi, they may ask us to relax validation rules to include values up to 4,000. In this case, we needed to back up and re-process an awful lot of data. Such re-processing, or re-summarization, crunches through a lot of numbers, and sometimes takes days to complete. In such cases, process data fixes going backwards in time, fixing the latest values first, and keep walking back in history until all the required re-processing is complete. Users tend to work mostly with the latest data, so it is prudent to fix that first.

B. Replacing Sensors

Not all data problems are software related, some involve replacing hardware. Given enough time, nearly all machines malfunction. Once a particular sensor is giving faulty data, it needs to be replaced. For energy companies, dealing with earth and subsurface structures, equipment is often installed in harsh physical conditions; thousands of feet below the

surface, or remote locations with extreme climate conditions. Therefore, swapping faulty equipment might not be a simple operation to carry out. Frequently, the equipment is not easily accessible. Surface sensors are relatively easy to replace, but those installed sub-surface in wells may require a rig to be sent on site, shut the well, and perform a work over. For this reason, sensor packs lately have redundant sensors installed; in case of a faulty sensor, just ignore the data it sends, until the work can be carried out, preferably with other work being done on the well, to minimize the cost of intervention.

C. Changing Practices

Validating high volume, machine-generated data, in order to build the level of trust necessary for achieving business goals requires changing some practices, and developing new procedures. Before users make decisions based on this data for optimizing daily operations, or use the data for metrics in key performance indicators, or perform analysis, or use it for simulation modeling, or for creating production profiles, they must have confidence in the numbers they are dealing with.

Engineers who are validating data may need to be a little more meticulous. At first, it may require spending more time on data validation than they are used to, or have to do it more frequently than they were used to doing it in the past. This is difficult to sell, because people expect modernization and smart sensors to *reduce* their workload, not add more tasks to an already busy day.

Modernization is a transition. It is a commitment, and will result in a better way of doing things, but not without growing pains. For Big Data projects, anticipate changing how work is carried out, and prepare for it by defining a process for updating business practices. This makes it easier to implement improvements on how organizations function and results in new streamlined practices, moving them towards the overall objective of greater efficiency.

VIII. Conclusion and Future Work

Leveraging the power of data warehousing for helping establish data veracity for Big Data is a feasible and implementable solution. Most large organizations already have some data warehousing implemented; the important step to take is to make it part of the information architecture.

Digitizing an oilfield is not simply a matter of installing sensors to receive data. Equally important is strategy which recognizes the business goals and defines a framework on achieving them. In order to maximize a return on investment, and realize the benefits, information technology must play a role of shielding the end user from technical jargon and deliver information in a way that is easy to use, and integrates seamlessly with traditionally collected data. Such a solution would provide a holistic strategy and encompass applications which integrate wholly into an upstream engineer's analytical processes.

There is plenty of work planned in the future after combining high frequency machine generated data with traditional structured data. There still remains the added value of combining the structured dataset now stored in a data warehouse with unstructured or semi-structured data. Upstream geoscientists in the oil and gas sector complement the structured data gathered in databases, with documents, observer logs, surface maps, subsurface maps, and a wide variety of schematics showing geological formations, subsurface contours, fluid migrations etc. A future Big Data solution envisions including all of these unstructured data items with the structured data gathered.

Large enterprises sanction Big Data projects and in doing so commit significant investment and resources to the effort. Some benefits are realized early for operational alerting of any outage or anomaly that requires immediate attention. The gain is fully realized only when this data is filtered and transformed into information which helps analyze trends to manage core assets such as hydrocarbon reservoirs, and wells over a long term, enhancing value in operational, tactical and strategic planning. For large enterprises, the road to harnessing the full potential of Big Data is long, but worth the journey.

.

Acknowledgment

The author would like to thank Yasir Rafie for his valuable comments.

References

[1] Gartner Inc. http://www.gartner.com/it-glossary/big-data/ , retrieved on March 10, 2014.

[2] D. Bartik, "Big Data: Dead by Definition, Alive in Practice" Information Week, http://www.informationweek.com , retrieved on Feb. 13, 2014.

[3] T. Lukoianova, and V. Rubin, "Veracity Roadmap: Is Big Data Objective, Truthful and Credible?" Advances In Classification Research Online, 2014, doi:10.7152/acro.v24i1.14671.

[4] IBM Inc. http://www.ibm.com/developerworks/bigdata/karentest/veracity.html , retrieved on March 10, 2014.

[5] W. Wolfe, "Real Time Well Data Gathering and Analysis" unpublished.

[6] R. D. Langenkamp, The Illustrated Petroleum Reference Dictionary, PennWell Publishing, Tusa, 1980, p. 28.

[7] U. Nahdi, "Integrating Live and Conventional Data" unpublished.

 ISBN: 978-1-61208-334-6