SPE 167844

Geographically-Distributed Databases: A Big Data Technology for Production Analysis in the Oil & Gas Industry

Aymeric Preveral, SPE, and Antoine Trihoreau, SPE, IDMOG, and Nicolas Petit, SPE, MINES ParisTech

This paper was prepared for presentation at the SPE Intelligent Energy Conference and Exhibition held in Utrecht, The Netherlands, 1–3 April 2014.

Abstract

The paper discusses some reported shortcomings of state-of-the-start IT technologies currently employed in the data management of Oil & Gas production operations. Most current IT infrastructures connect historian databases, production databases and application servers. This creates complex issues of data consistency between these systems. In the discussion, a particular focus is put on the geographically-distributed nature of the network which suffers from low-bandwidth limitations and un-reliabilities, e.g. due to satellite communication links.

Taking the production engineers' viewpoint, an example of production allocation using Data Validation and Reconciliation (DVR) serves to stress the malicious impacts of the described architecture. Production allocation represents one of the various monitoring and analysis tasks that are performed, on a daily basis, at the centralized level of data management systems. A quantitative study shows that the problem of mis-synchronization of databases is of great practical importance.

We propose solutions to improve the robustness to communication outages. To improve data consistency across sites in a decentralized manner, the paper exposes the key concepts of distributed storage, message-based communication, and clustering. More generally, the paper proposes to shine a light on the potential relevance of several recent advances in the scientific field of "big-data" to the world of Oil & Gas upstream industry. These off-the-shelf technologies must be specifically tailored to geographically-distributed networks. The specificities are detailed, the necessary development work is outlined, and the potential qualitative benefits are estimated. A possible implementation is sketched.

Introduction

On a daily basis, production engineers monitor and analyze production data. For example, among their daily tasks, production allocation is essential to determine each well production level, detect possible production network leak and perform hydrocarbon accounting. Various experiences detailed in the literature (Knabe et al. 2008; Cruz Villanueva et al. 2012) report that production engineers spend a large part (sometimes up to 60%) of their monitoring time doing data gathering.

In an effort to increase productivity, a recent trend in major companies has been to invest in Information Technology (IT) to reduce tedious and time-consuming data management burdens, with the long-term goal of enabling real-time production analysis and optimization. State-of-the-art IT considers centralized approaches connecting the heart of the system (a centralized database/application server), historian databases, and specialized databases. At the centralized level, various monitoring and analysis tasks can be performed (production monitoring, real-time well modeling, among others). So far, the integration and maintenance complexity of such IT systems have kept them away from general deployment in affiliates where they would fully leverage the value of assets. One culprit is the geographically-distributed network between the production sites and the offices of an Oil & Gas company, some parts of which are low-bandwidth and sometimes unreliable (e.g. satellite communication links between production sites and offices).

On the other hand, big-data technologies, that have emerged in numerous other fields of applications (retail and logistics, among others), are ideal candidates for bringing significant improvements. In fact, these are specifically designed to deal with massive amounts of data in a distributed system, provide robustness to communication breakdowns, and guarantee eventual consistency across sites in a decentralized manner. In the presented context, they would yield availability of a consistent dataset everywhere in the connected network of production sites and offices. This article exposes the principles of one IT system built around these off-the-shelf technologies and specifically tailored to the geographically-distributed networks which are ubiquitous in the Oil & Gas upstream industry.

The article is organized as follows. First, we describe current centralized IT architecture employed in the Oil & Gas production operations. The performance and limitations of these systems are discussed, while the role of network failures and bandwidth limitations is stressed. Next, we outline a possible implementation of an innovative architecture and database technology which could handle the described issues. A particular focus is put on the necessity of introducing the concepts of distributed data store, message-based communication, and clustering. Then, we expose typical detrimental effects of the current ITs limitations for real-time production analysis. An example of real-time production allocation based on Data Validation and Reconciliation (DVR) using mis-synchronized data serves as illustration.

State-of-the-art IT architecture employed in Oil & Gas production

A typical IT system dedicated to Oil & Gas production operations integrates several coexisting concurrent systems. This kind of implementation is reproduced in Figure 1.

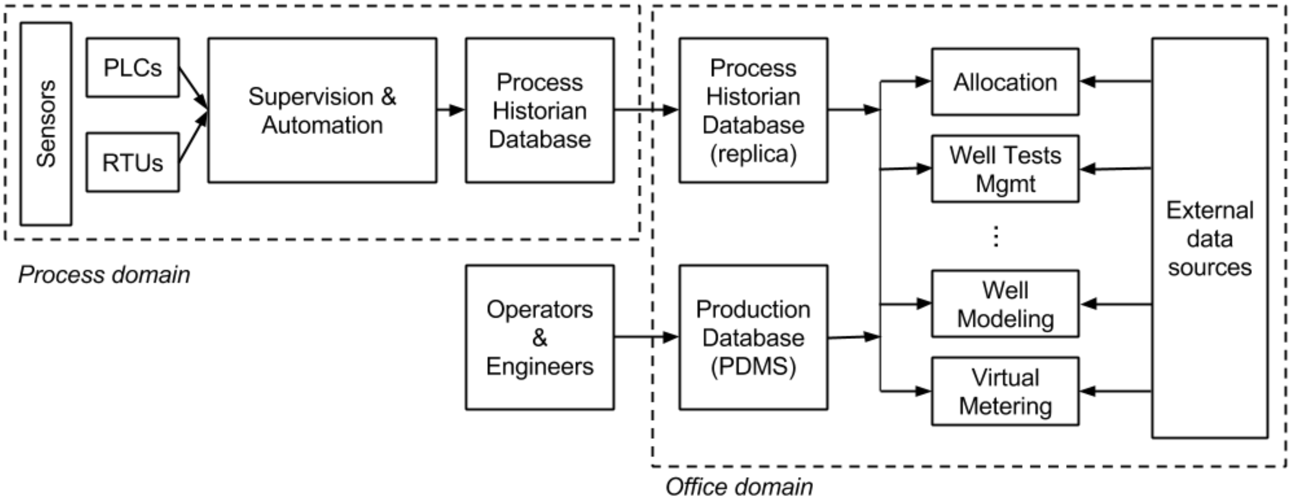

Fig. 1 - An example of a typical production IT System

Sensors are connected to the SCADA system through PLCs and RTUs. Historian databases are used to store sensor data. The historian databases are part of the process-domain infrastructure where OPC standards are often used to exchange data. Security is a key feature of the process-domain systems. This process-domain architecture is designed for real-time on-site operations surveillance.

Production database management systems (PDMS) are used to store production data. PDMS is related to the office-domain data infrastructure. These systems, mostly based on SQL technologies, are dedicated to production monitoring, analysis and planning in relation with petroleum, reservoir and financial software. The database contains various types of documents and data: wells and production network descriptions, field events (shut-ins, well test reports) and daily production reports (allocated and fiscal daily production). A replica of the historian database or a tag-based storage solution for sensor data is often used in between, to link the process domain and the office domain.

At the very large scale of the corporate level, the large amount of data necessary to perform real-time monitoring and analysis exeeds data volumes that existing data management system architectures can collect, store, replicate, and deliver. Some reasons for this are as follows:

i) SCADA systems control processes. Therefore, security must be an absolute concern to avoid disasters. Consequently, extending SCADA systems to build company-wide distributed systems induces serious IT security issues and the related costs.

ii) Some production sites have a limited network bandwidth and unreliable network (e.g. VSAT). Therefore, accessing a central database for daily applications yields to high latency responses and limits data usage. A distributed storage approach can solve this problem but most PDMS are not ready for large scale data replication across sites. PDMS are mostly based on SQL databases which limit data availability because of locks issues in case network partition.

iii) Historian databases and related technologies in the process domain are mostly not aware of an Oil & Gas specific data model such as PPDM (Professional Petroleum Data Management) or exchange format. Therefore, production engineering analysis requires a PDMS and most production engineers have to work both with the historian database and the PDMS to perform analysis. In addition, historian databases are designed to store time series. Consequently, data generated by distributed sensor, which are multidimensional, do not fully fit in a database designed for time series.

Some operating fields, such as the ones reported in Perrons (2010) and Khan et al. (2012), have production databases or historian database replicated to a production support site (subsidiary HQ, corporate HQ) based on the aforementioned technologies. However, integration of this kind of architecture causes high overhead, which in turn prevents Oil & Gas companies from near real-time, high throughput, back and forth data communication between field operation teams and remotely located production experts. Initial design and workload characteristics of these legacy technologies are not fully optimized for digital oilfields systems or Integrated Operations. Building a companywide field/production data management system without technologies designed for purpose would probably induce high integration costs and a heavy maintenance burden.

A proposition: geographically-distributed architecture and related database

To generalize the use of complex production applications such as real-time allocation or any other digital oilfields applications, the production IT architecture has to be adapted. Field data and production data management for the Digital Oilfield is an industry challenge requiring dedicated solutions. Exchange of data among different platforms operating different types of information systems (field process domain and office domain) can be handled by a specific architecture referred to as middleware, a layer enabling interoperability and distributing services. The purpose of this middleware is to fill the field-to-office gap in the operational IT infrastructure.

A short list of desirable properties of this middleware is as follows:

- Provide direct access to all sensors data (e.g. Modbus-based integration) or through automation systems (e.g. OPC-based integration with the Process Historian Database),
- Guarantee local and remote availability of field data even in cases of network outages (e.g. VSAT down),
- Limit bandwidth usage for services communication and data exchange,
- Enable development of user friendly front-end for visualization, validation or appending production data tasks in a collaborative manner,
- Respond to request from applications running on top, with predictable (as much as possible) latency for data ingestion, retrieval, and processing operations,
- Capabilities to deal with main types of data generated during operations: time series, multidimensional data (e.g. DTS, DAS) and contextual data. Times series and multidimensional data represent most of the dataflow from sensors. Contextual data (text messages, tables...) correspond to SCADA alarms and oilfield operation reports such as well tests reports, shut-in and incidents reports, workovers and maintenance,
- Deliver data to other applications and front-end applications with an easy-to-use and uniform Application Programming Interface (API). This API can be based on data exchange standard (e.g. ProdML) an in-house API (e.g. a built-for-purpose REST API).

A more exhaustive list of desirable properties of such a system can be found in Cramer et al. (2012).

In the solution we propose, cross-applications and cross-site interactions are structured using a Message Oriented Middleware (MOM), as described in Mahmoud (2004), which facilitates integration of highly heterogeneous and decoupled systems. An oil field dedicated middleware consists in acquiring data from the process domain, managing interaction between software components and includes a distributed storage system for large production data volumes. We use a *big-data* distributed data store as storage layer for field and production data.

The MOM paradigm is based on asynchronous message passing. The system orchestrates data access and exchanges between applications sharing the same data sources. The MOM allows message passing across applications and message publish/subscribe on distributed systems. It also features:

- Asynchronous communication mechanisms to support widely distributed architectures
- Transformation of data format to fit the receiving application (e.g. from OPC-based data to ProdML-based data)

- Loose coupling among applications to enable multi-vendor systems

The MOM provides asynchronous and highly scalable many-to-many communication model from senders to receivers. In our application, the publishers can be the field data sources or the algorithms and applications outputs. The subscribers are others algorithms/applications using the data published as inputs. The publish/subscribe scheme provides systems decoupled in terms of space, time and synchronicity. These features are well suited for systems with large geographic distribution and unreliable communications. MOM is also well suited to integrate disparate software component paving the way for a better interoperability.

Minimization of the bandwidth footprint for communication and data synchronization is achieved by an efficient encoding of data. The overall bandwidth usage depends on the throughput of the messages broker and the size of each message. Data transmission format used in the middleware is conceived to minimize data message sizes. Compression and encryption systems are used to reduce the weight of data transmission. XML format, widely used in office environment, are not well suited for communication with field operations since it induces a large overhead on transmitted data. Other message format dedicated to MOM should be considered (DDS, AMQP, MQTT...).

In addition, we propose to fully leverage this distributed MOM layer with a distributed database. In the recent years, distributed databases such as Google's BigTable (Chang et al. 2008), and Amazon's Dynamo (DeCandia et al. 2007) solved *big-data* large scale storage challenges in the Web industry. Contrary to traditional Relational Database Management Systems (RDBMS), these databases use redundant storage on clusters of commodity hardware. This architecture is essential to scale to very large datasets of petabytes of data across several storage locations. In these distributed systems, any storage node is considered as master node which avoids "single point of failure" issues. In addition, writes operations can be done on any storage node. An efficient conflict resolution strategy is required which can be considered as a serious drawback.

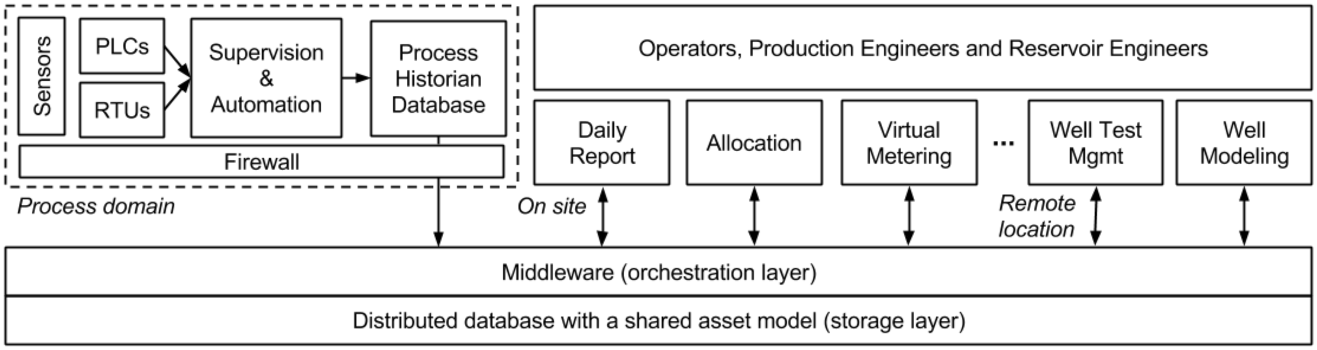
Sensors
PLCs
RTUs
Supervision & Automation
Process Historian Database
Firewall
Process domain
Operators, Production Engineers and Reservoir Engineers
Daily Report
Allocation
Virtual Metering
...
Well Test Mgmt
Well Modeling
On site
Remote location
Middleware (orchestration layer)
Distributed database with a shared asset model (storage layer)

Fig. 2 - Proposed architecture: a Message-Oriented Middleware and a Distributed database

The benefits of the proposed system are as follows. It allows one to use distributed applications and storage across heterogeneous operating systems, programming language, computer architectures, networking protocols, while reducing the complexity on the interconnection functionalities. Several key issues of distributed architectures are addressed: scalability, high availability, components decoupling and interoperability. They are all of importance in the context of Digital Oilfield applications.

Massive datasets such as data generated by high-frequency sensors and distributed sensors (DTS, DAS...) can be stored along with operations events data at a minimal hardware costs at large company scale. These large datasets can be queried with specific techniques such as MapReduce engines with a reasonable latency to extract information. Therefore this system is capable of handling data of dozens of fields, each with hundreds of wells that can be operated & monitored cooperatively.

The use of clusters of commodity hardware is also very interesting for remote production sites where logistics is a concern. Data storage is redundant on several servers so, if one server goes down, the system can stay up and running. While the server is fixed or a new server is shipped on site, the system would probably be less efficient but it would remain available. Storage redundancy on multiple servers also enables a secure storage without any specific backup and data loss is unlikely.

High-availability, promoted in the Web industry to always serve clients, has an interesting echo in the Oil & Gas industry where geographically-distributed locations can easily suffer from network failures and bandwidth limitations between sites. For example, serving a request through a satellite-based communication system can result in a poor user experience and limit data usage. In addition, frequent network failures can be observed during field operations and prevent access to data stored in a remote location. Therefore, distributed databases that gracefully handle storage across sites and resolve conflicts after a partition network are an interesting solution to ease simultaneous data access both on site and on remote locations. With this approach, field and production data are located where they are used in near real-time and can serve applications with very low latency even in case of communication network disruption.

Detrimental effects of heterogeneous systems for real-time production analysis

One can easily illustrate the negative effects of poor data management in everyday's oil production tasks. For this purpose, the production allocation task serves as example below. Another example can be found in Magnis-Petit (2013).

Production allocation is traditionally performed on a daily basis. The most commonly used method is based on well tests, uptime factors and a back allocation algorithm computing an estimate of each well oil and/or gas production. This very robust method is suboptimal for several reasons. It uses averaged and sometimes outdated data to production rough estimates. For example, well tests results are considered valid during days after the test. It assumes some steadiness of the production flow. Yet, this heuristic solution is still widely used in the industry because it requires very limited data.

Others methods can be used when enough instrumentation is available on the well such as model-based allocation and Data Reconciliation and Validation (DVR). Virtual flow meters can be computed from data measured at the well level using ESP sensors, wellhead pressure, and choke position among others. These model-based methods provide real-time estimates of the well rates and can be tuned or calibrated to obtain precise results. An overview of several of these well models is given in Wu et al. (2012). DVR requires the availability of redundant real-time data coming from flow rate, pressure, and temperature sensors across the production network. These techniques originally used in downstream plants are tested in upstream production operations as explained in Couput et al. (2008). The advantages of these methods are:

- Significantly reduced allocations factors
- Early leak or production change detection
- More accurate individual well production estimates

A common aspect of most of these model based allocation methods is that they need to retrieve, in near real-time, data from the several sources of the IT architecture. As detailed earlier, these are PDMS (for well tests results, shut-ins and any other production events) and historian databases (for sensor data). Use of distributed sensors data would require integration of the large volume of data generated by the distributed pressure/temperature sensors.

The principal difference between Data Reconciliation and other estimation techniques is that Data Reconciliation explicitly uses process models to obtain estimates. The reconciled estimates are expected to be more accurate than measurements and, more importantly, are also consistent with the known relationships between process variables.

Applied to an oil production network, data reconciliation can be used to perform production allocation. Formally, we consider the oil rates in the production network:

$$x = \binom{x_m}{x_u}$$

where x_m are the measured flow rates and x_u are the unmeasured flow rates.

Assuming no major leak is present in the production network, the material balance in the network gives:

$$Ax = 0$$

It can also be writing as:

$$A_m x_m + A_u x_u = 0$$

where A is a network representation matrix, A_m corresponds to the measured variables, and A_u corresponds to the unmeasured variables. We also consider y the vector of flow rates measurements:

$$y = x_m + e$$

where e is the vector of measurement errors (noises).

Data Reconciliation is based on the assumption that measurements are not biased. Commonly, it is assumed that e follows a Gaussian distribution with a zero mean and a diagonal covariance matrix. The accuracy of each measurement is characterized by its standard deviation σ_i. The allocation problem can be formulated as a constrained minimization problem:

$$\min_x \sum_i \left(\frac{y_i - {x^e}_i}{\sigma_i}\right)^2$$
$$s.t.\ Ax = 0$$

We consider $\varphi = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_m \end{pmatrix}$ the diagonal matrix of measurement standard deviations.

As mentioned in Narasimhan (1999), QR decomposition is an interesting approach to solve this minimization problem. Noting QR the QR decomposition of A_u, we have:

$$A_u = QR = (Q_{u1} \quad Q_{u2}).\begin{pmatrix} R_{u1} & R_{u2} \\ 0 & 0 \end{pmatrix}$$

The analytic solution of the optimization problem above is:

$$x_m = y - \varphi S^T (S\varphi S^T)^{-1} Sy$$

where

$$S = {Q_{u2}}^T A_x$$

We now treat an example. The oil field under consideration is pictured in figure 3. Consider 5 production wells and 2 water injectors. There are 2 stations with a separator at each station.

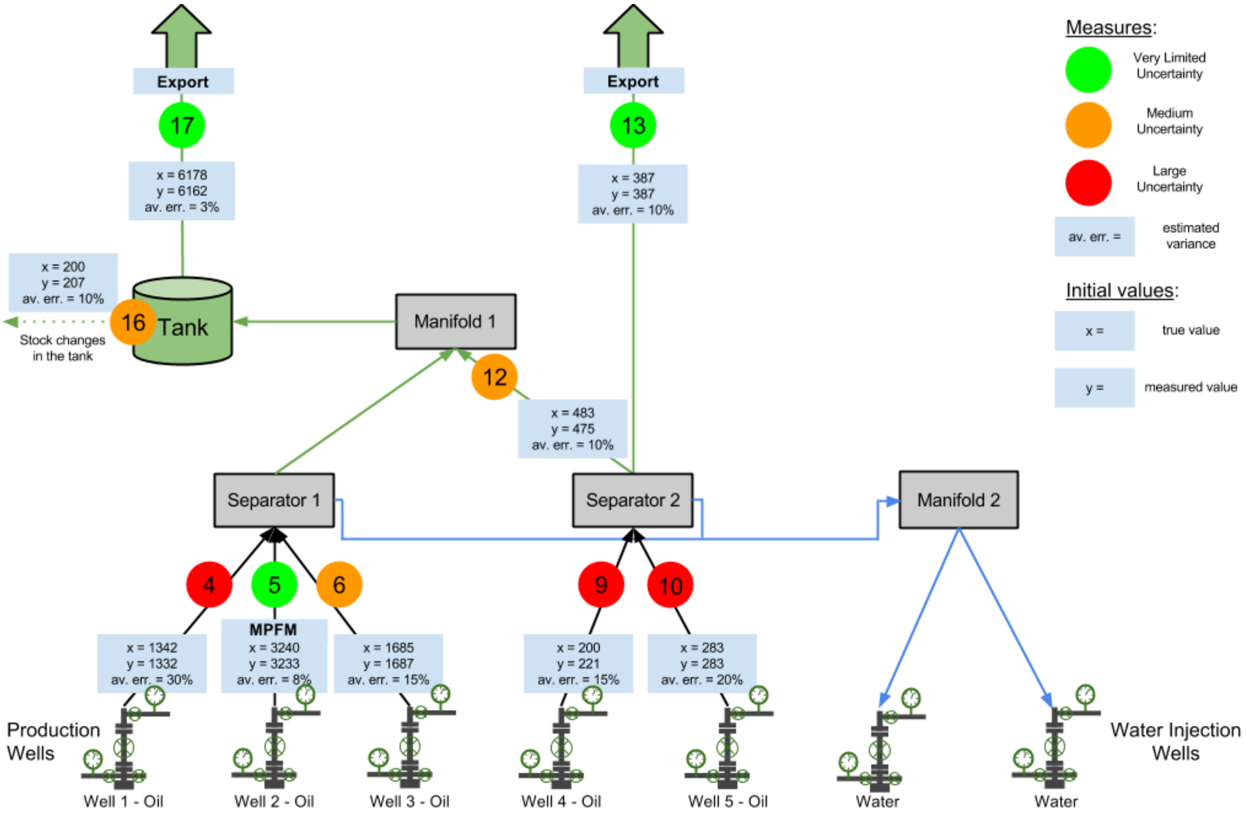

Fig. 3 - A production network

Well 5 is equipped with a Multiphase Flow Meter (MPFM). Two export stations are used on this field and oil exports are measured with very limited uncertainty. Several sensors are deployed along the production network. For conciseness, the matrix A is given in appendix. We implemented the DVR equation with well synchronized and mis-synchronized data, respectively.

Typical synchronization errors due to heterogeneous databases are representative of experimentally observed conditions. Some flow rates are down sampled and averaged of to daily values and delays exist on some flow rate measurements. A priori information on some flow rates measurements range from good to very poor. In Fig. 4, we report flow rates estimates for well 2 and well 4. The averaged estimates are consistent, but the mis-synchronized estimate fail to warn of a flow rate reduction (about time 180 hour), which could be caused by equipment failure. A similar behavior can be observed at time 600. Mis-synchronisation reveals false warnings and does not pay justice to the smooth production operated on the field: the estimate is noisy, while the true value is not. It is very difficult to discard the erroneous data produced by a mis-synchronized estimator. Results show that some data mis-synchronization can be critical and highly detrimental to the produced estimates, yielding spurious oscillations resulting in undesired warnings.

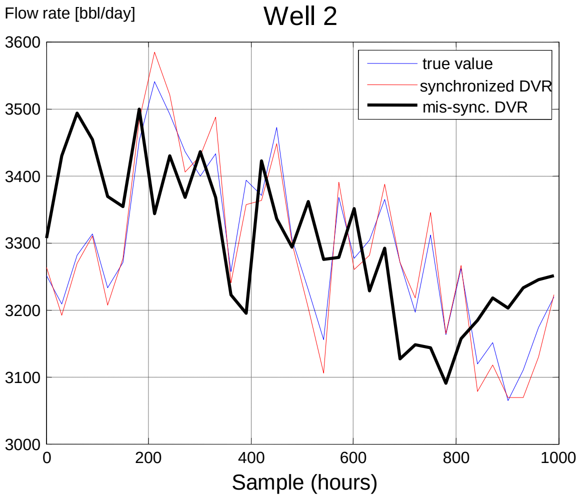

Fig. 4a - Flow rate estimate for well 2.

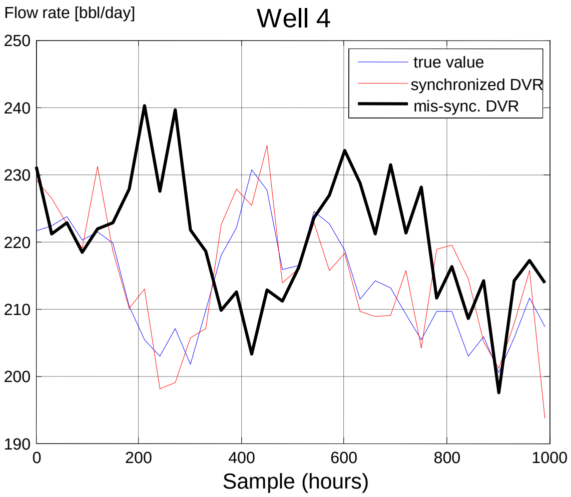

Fig. 4b - Flow rate estimate for well 4.

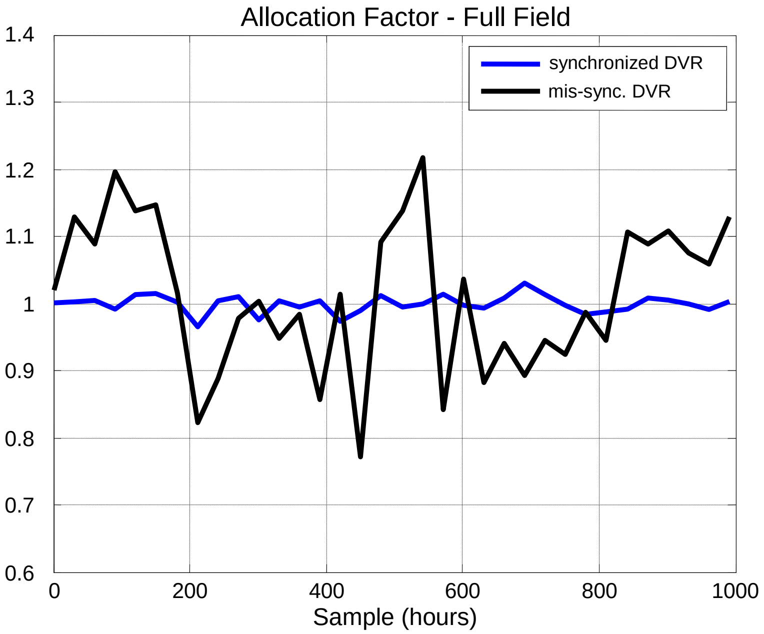

Fig. 5 - Full-field Allocation factor (estimated vs export measures).

The preceding results were obtained following a specific implementation of data management. This example illustrates how a heterogeneous IT system not designed for the Digital Oilfield seriously limits benefits of real-time allocation implementation efforts. Digital Oilfield applications require easy-to-access and consistent data. A dedicated IT architecture designed to manage services and applications across sites is required to serve consistent data across sites. To run our allocation process, we access all required data from the middleware API. The results of the allocation are then push back to the middleware and stored in the distributed database. They are ready to use by other applications such as reporting tools or reservoir simulators. These results are also available for visualization, for example, through a web based monitoring interface.

Conclusion

In this paper we have underlined the benefits of implementing the recent “big-data” technologies and related architectures: Message-Oriented Middleware and clustered storage on commodity hardware. These are: high availability over failure-prone networks, scalability at a company scale and storage security. The induced cleansing of uniformed data, and in particular, the possibility of working with well-synchronized and consistent data, has appeared as instrumental for the daily tasks of petroleum engineers. These foundation technologies open the way to easy development of user-friendly applications. A web based access to all production data ease the readability and analysis. Latest web technologies (HTML5/CSS3) for user interface can be employed to leverage the value of production data with interactive data visualization methods.

In addition, the near real-time data consistency between sites enable a true collaboration across the organization. All daily tasks can be performed therein: morning daily reporting, production monitoring, read and write of field events, production trend analysis, preparation of weekly and monthly reporting and production accounting. The described middleware could provide built-in monitoring and reporting functionalities, however, complex petroleum, reservoir or financial analysis should be delegated to tools specifically designed for these applications. As a consequence, interoperability is a major challenge for this middleware. To reduce the costs of ownership of their software and optimize the value of production data, each Oil & Gas company should be able to orchestrate together the applications they require for operational and development tasks. Consequently, this middleware should comply with industry standards to enable straightforward communication with existing and future software through an easy to use API (Application Programming Language).

References

Knabe S., Shaw, D.C., et al. 2008. Intelligent Continual Right-Time Analysis of Field Data as a Service. Paper SPE 111342 presented at the SPE Intelligent Energy Conference and Exhibition, Amsterdam, The Netherlands, 25-27 Febrary.

Cruz Villanueva, C., Tapia, C., et al. 2012. Workflow Automation (WFA) for Integrated Production Operations in the Macuspana Field. Paper SPE 152234 presented at the SPE Latin America and Caribbean Petroleum Engineering Conference, Mexico City, Mexico, 16-18 April.

R. K. Perrons. 2010. Perdido: The First Smart Field® in the Western Hemisphere. Paper SPE 127858 presented at the SPE Intelligent Energy Conference and Exhibition, Utrecht, The Netherlands, 23-25 March.

Khan, F., Goujard, V., et al. 2012. Well-Performance Monitoring (WPM): Creating Added Value From Raw Data and Application to the Girassol Deepwater-Field Case. *SPE Economics & Management* 4 (3): 182-189. SPE 128557-PA.

Cramer, R., Krebbers, J., et al. 2012. Establishing a Digital Oil Field Data Architecture Suitable for Current and Foreseeable Business Requirements. Paper SPE 149959 presented at the SPE Intelligent Energy Conference and Exhibition, Utrecht, The Netherlands, 27-29 March.

Mahmoud, Qusay H., ed. 2004. *Middleware for communications*. J. Wiley & Sons.

Chang, F., Dean, J., et al. 2008. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26 (2): 4.

DeCandia, G., Hastorun, D., et al. 2007. Dynamo: amazon's highly available key-value store. *SOSP* (7): 205-220.

Magnis, L. and Petit, N., 2013. Impact of imprecise dating of measurements for bulk material flow network monitoring. in Proc. of 12th IFAC/IE E Conference on Programmable Devices and Embedded Systems PDeS.

Wu, X., Humphrey, K., and Liao, T. 2012. Enhancing Production Allocation in Intelligent Wells via Application of Models and Real-Time Surveillance Data. Paper SPE 155031 presented at the SPE International Production and Operations Conference and Exhibition, Doha, Qatar, 14-16 May.

Couput, J.-P., Louis, A., and Danquigny, J. 2008. Transforming E&P Data into Knowledge: Applications of an Integration Strategy. Paper SPE 112517 presented at the SPE Intelligent Energy Conference and Exhibition, Amsterdam, The Netherlands, 25-27 Febrary.

Narasimhan, S., and Jordache, C. 1999. *Data reconciliation and gross error detection: An intelligent use of process data*. Access Online via Elsevier.

Appendix

The matrix A is reported below where the rows represent the edges of the production network (wells, separators, manifolds, export) and the columns represent the vertices of the production network (only for oil).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**18**	**19**
Well 1	1	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Well 2	0	1	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0
Well 3	0	0	1	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0
Sep. 1	0	0	0	1	1	1	0	0	0	0	-1	0	0	0	0	0	0	0
Well 4	0	0	0	0	0	0	1	0	-1	0	0	0	0	0	0	0	0	0
Well 5	0	0	0	0	0	0	0	1	0	-1	0	0	0	0	0	0	0	0
Sep. 2	0	0	0	0	0	0	0	0	1	1	0	-1	-1	0	0	0	0	0
Mani. 1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	-1	0	0	0
Export 1	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0
Tank 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	0
Export 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1