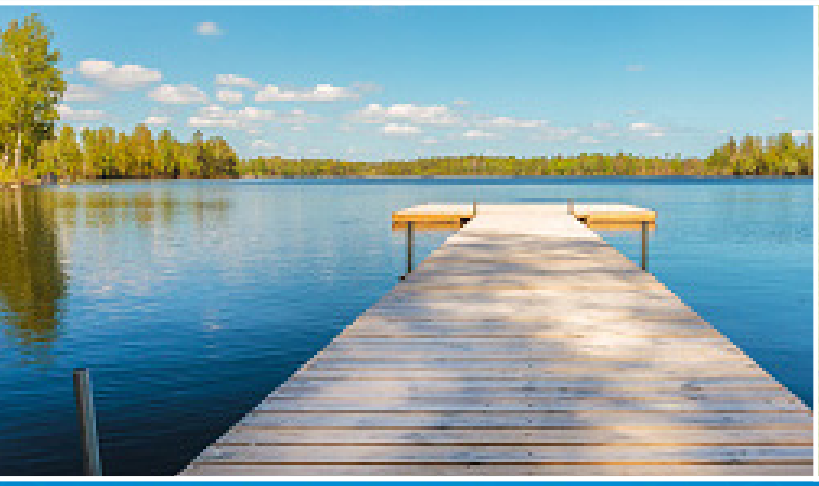
KSG SOLUTIONS
BRINGING SPEED TO YOUR VALUE GENERATION

Data Lake or Data Swamp?



Keeping the Data Lake from Becoming a Data Swamp.

INTRODUCTION

Increasingly, businesses of all kinds are beginning to see their data as an important asset that can help make their operations more effective and profitable. As our ability to gather time-series data grows, more technologies are becoming available to help us make sense of it. How do we choose the right technology and approach for our business problems?

ABOUT THE AUTHOR

John de Koning, success advisor in industrial data processing, created his roots in the oil and gas industry. As a technology and innovation manager for Shell, John was focused on generating $500+ million value annually by introducing innovative ways of processing manufacturing and production data. He became an industry leader by introducing architectures to contextualize, integrate and aggregate manufacturing and production data at a corporate level. The experience and understanding gained has been used as the foundation for this white paper.

The paper is focused on helping industry leaders understand the characteristics of the various data processing techniques, and how they link together to form an optimum solution architecture for processing time-series data in combination with enterprise data lake initiatives.

MANAGEMENT SUMMARY

Data lakes are a simple way for businesses to collect and store raw data from a variety of inputs, without having to know in advance exactly how the data will be used. But in order for data to drive business outcomes, it must be organized and accessible. Without structure, the data lake becomes a swamp. A variety of advanced real-time software systems are available that integrate with enterprise data lake software and can help collect and structure data, so it can be used effectively.

Various solutions are available for processing time-series data. Some of them even pretend to be the Holy Grail for data management and advocate streaming sensor and machine data straight to a data lake and the cloud – and suggest organizing it later. But what about the nature of industrial data streams and the legacy automation equipment that is already out there? Especially in the area of industrial data environment, automation systems have a life cycle up to 20 years and replacement is a serious investment. Sending the raw data from these sources to the data lake is not even an option as interfaces for these legacy sources do not exist.

Access to the data should be simple and affordable, but still enable enterprise wide reporting and analyses.

The solution architecture for time-series data should follow a few strict rules:

1. Connectivity

Ensure the corporate solution is able to connect to the variety of (legacy) data sources and potential new sources.

2. Time-Series Capacity

The system should be able to deal with time-series data (high fidelity, time indexing, and time synchronization).

3. Context

Systems should have easy-to-understand asset/equipment-based relations between the individual data streams, to enable business users to easily compare, view and analyze data on an equipment level without being an IT specialist or data scientist.

4. Accessibility

Process users should be able to analyze and visualize the data to help optimize the use of production facilities.

5. Security

Keep your production facility safe and secure! Don’t allow unintended back-door access to your automation system.

Often positioned as the one size fits all, the currently available data lake technologies do not have the ability yet to handle the above key rules in an effective and efficient way. To assure data from a large variety of (legacy) source systems is landing in the cloud with the correct timestamp, synchronized in time and having the right context, it is important to add an **infrastructure layer** specially designed for this purpose. This combination of time-series and data lake technologies (cloud or on premise) will bring the flexibility and criticality in the various levels of the organization: (a) on the production level, ensuring that data will be secure but accessible; and (b) on the corporate level, allowing data to be contextualized, integrated, and aggregated for better business decision making.

Various solutions are available in the combination of real-time infrastructure and data lake technologies. Based on the above rules, the technology combination of OSIsoft PI System™ toolkit with supporting Data Context Automation tools, like those delivered by Element Analytics, is a leading strategy for a solution architecture that supports both the time-series data needs within operations and enterprise data lake initiatives. Dedicated integration tools are available to easily integrate with Enterprise Data Warehouses and data lake technologies from Microsoft, SAP, or Hadoop, both in the cloud or on premise.

Large companies, like global energy enterprises, have proven this technology combination can easily drive $500+ million benefit per year by introducing enterprise tools and processes for Proactive Monitoring, Exception Based Surveillance, Rotating Equipment Monitoring, Condition Based Maintenance, Margin visualizing, etc. All of these will result in better uptime and higher efficiency of the facilities.

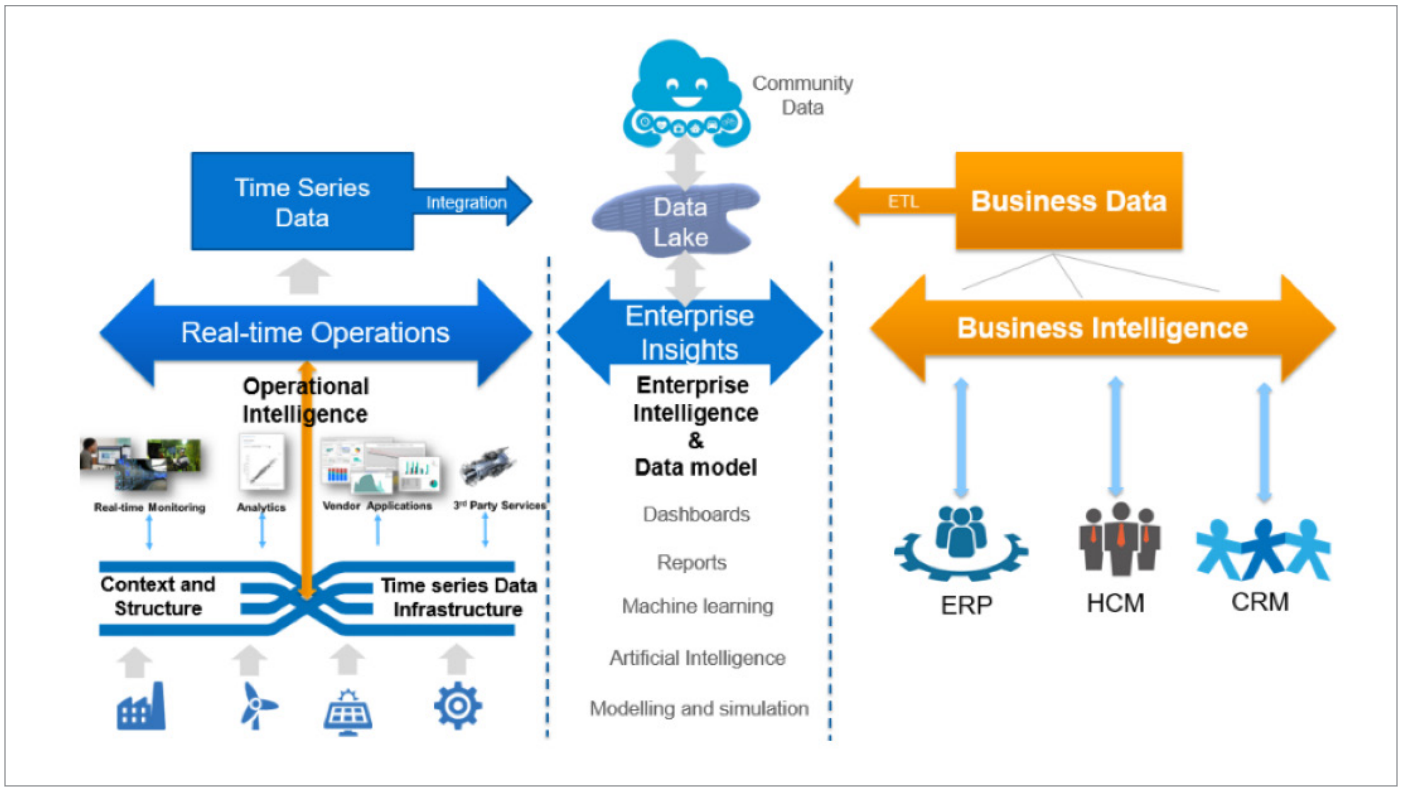


Figure 1 : Hybrid Environment

TRANSITION TO DATA LAKES

Traditional data warehouse technologies use predefined data models to describe the database. The advantage is that you know upfront what the data structure looks like. The downside is that data warehouses are inflexible. A traditional data warehouse cannot keep up with rapid changes in the data model due to the proliferation of new data sources and new questions people want to ask of the data. This overwhelming rate of change is preventing the traditional way of working by first building a data model and a database schema. In addition, the traditional way of (data) change management will not work anymore, as version control will be hard with a fast changing data model.

In a data lake environment, raw data is pushed to the store in their original state. This can be structured, unstructured, blobs, etc. Instead of predefining how the data elements are related to each other (data model), as with a data warehouse, you create the relationships once you need to retrieve the data from the data lake. This is also the major downside of a data lake. With databases and warehouses, it was possible for business professionals (non-IT) to query the data as the complex data modeling when it was already done upfront by IT specialist. In the case of a data lake, you need to be a data scientist to be able to analyze the various chunks of data and link them together to make sense. Table 1 summarizes key characteristics of data warehouses versus data lakes.

Data Warehouse vs. Data Lake		
structured, processed	**Data**	structured / semi-structured / unstructured, raw
schema-on-write	**Processing**	schema-on-read
expensive for large data volumes	**Storage**	designed for low-cost storage
less agile, fixed configuration	**Agility**	highly agile, configure and reconfigure as needed
mature	**Security**	maturing
business professionals	**Users**	data scientists et. al.

Table 1: Data Warehouse vs. Data Lake

THE “PERFECT WORLD” IN AN INDUSTRIAL ENVIRONMENT

The “perfect world” is very simple. You want to have access to all the data that is available (internal and external), query the data in any combination, run integrated analytics to find the missing pieces and visualize the information you are looking for with the tool of your preference. However, the reality is often different. When combined with a real-time, time-series environment, the core concerns are related to the diversity of (legacy) data sources, network latency and reliability, data latency, time synchronization of data streams, and the context or relationship between data streams.

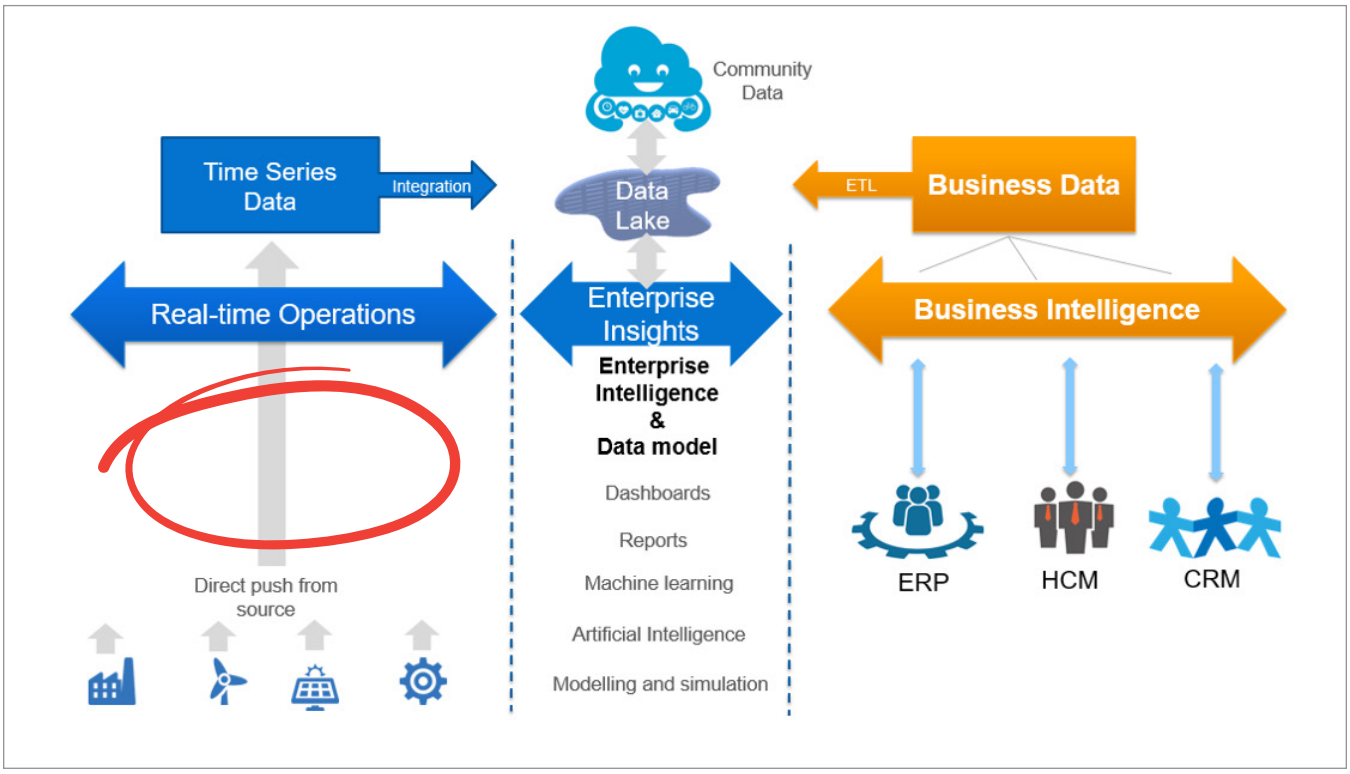


Figure 2: “Perfect World” of Data Processing. What’s Missing?

A GREAT ALTERNATIVE TO THE “PERFECT WORLD” IN AN INDUSTRIAL ENVIRONMENT

A hybrid model delivered by an ecosystem of suppliers will help to bridge the gap between the ‘Perfect World’ and the technology constraints.

Depending on the company size and the equipment used for production, the variety in time-series data sources can be significant. There will be a legacy of control and automation systems, especially in older companies with various production locations that may include systems from various brands, various types per brand and various versions per type. Sending the raw data from these sources to the data lake is not even an option as interfaces for these legacy sources do not exist. Also, the facility location can introduce significant data reliability concerns. Remote facilities connected via low bandwidth connections, like satellites, need additional functionality to avoid data loss. Another important aspect is security. To assure the integrity and safe operations of your facility, the interface technology must be very secure. The following table shows the benefits of adding advanced real-time, time-series systems to the hybrid model to address to key concerns from the data lake technology.

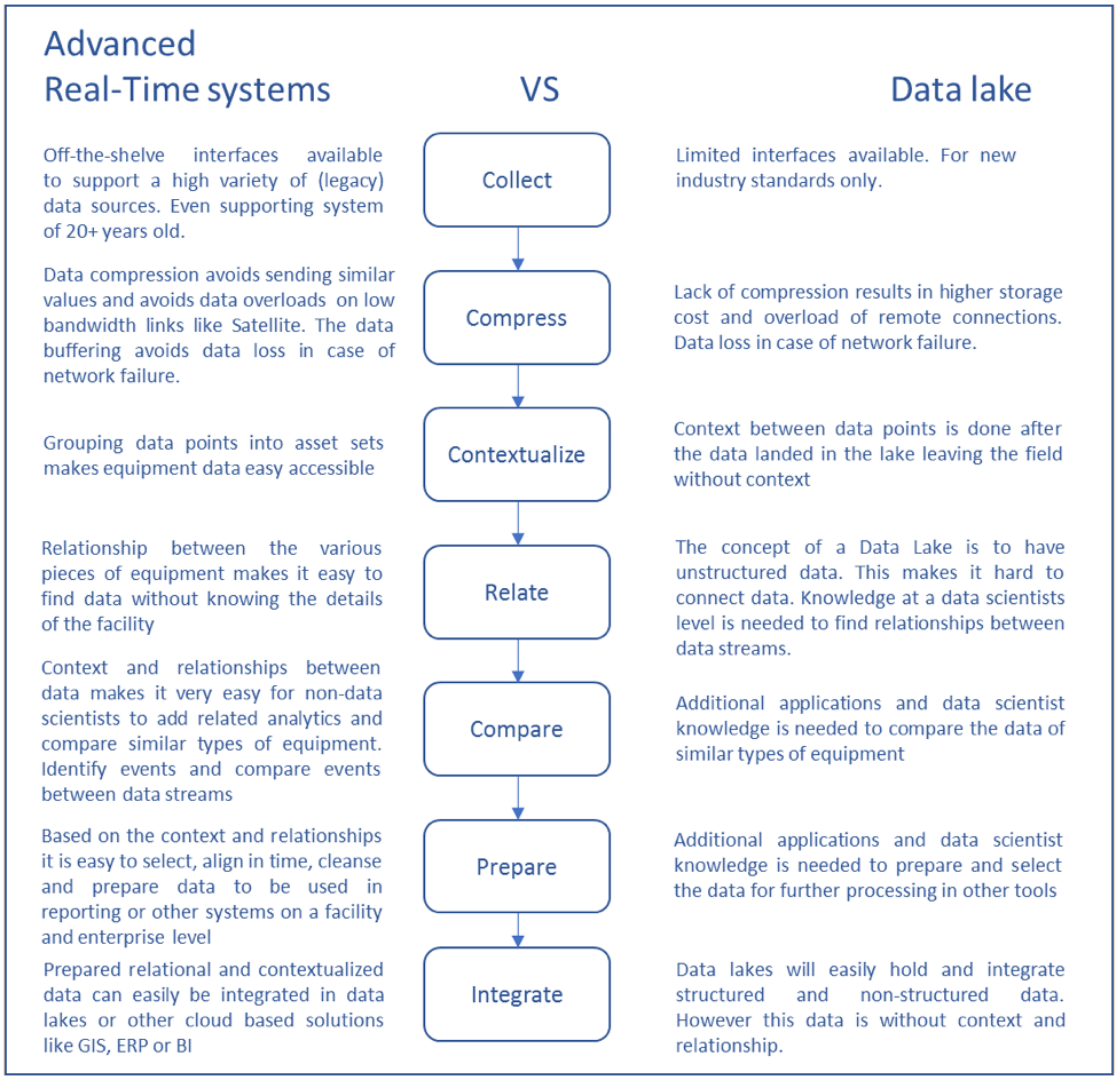


Figure 3: Overview of system characteristics

THE OPTIMUM OF DATA PROCESSING IN THE REAL-TIME, TIME-SERIES WORLD

The combination of data lake technology and time-series infrastructure will help to address the core concerns of the “Perfect World”. In this situation, time series data infrastructure will collect all the data from the field.

The time-series data infrastructure will also assure the availability of data in the field for local viewing, processing and reporting (Edge computing) or feeding data to (near) real-time optimization or advanced control. This Edge computing will assure the data and system availability needed to run and monitor the equipment in the production process itself by avoiding network availability and data latency issues.

BENEFIT OF INTEGRATING AND STANDARDIZING ACCESS TO DATA

The integrated combination of time-series systems and data lakes delivers a ‘One-Stop-Shop’ model for data access across the business and operations enterprise. This enables enterprise wide reporting, enterprise big data analytics and the delivery of enterprise applications across a broad spectrum of use cases. These enterprise applications and reports can be reused throughout the enterprise from one single platform. As the definition of a piece of equipment is the same throughout the company, it is very easy to reuse use cases throughout the company. Best practices from one location can be re-deployed with very low effort at other locations to rapidly generate value. In the case of an IT architecture with a consistent way of accessing data and with a consistent way of building a data model, it will be very easy to build one consistent set of analytics per equipment type and deploy this to all facilities throughout the enterprise. It avoids reinventing the wheel at the various facilities; development and deployment of applications will become very Agile; and most important, the time to value is very short. Large companies like global energy enterprises can easily drive $500+ million benefit per year by introducing enterprise tools for Proactive Monitoring, Exception Based Surveillance, Rotating Equipment Monitoring, Condition Based Maintenance, Margin visualizing, etc. This all will result in better uptime and higher efficiency of the facilities.

In the Energy world, the use of heat exchangers is quite common. Fouling of heat exchangers is a serious concern as it slows down production or forces unplanned outages. This concern is addressed at all facilities by technologists, all of whom try to invent a way to predict the fouling of ‘their’ heat exchangers. However, at the end, such efforts often result in a huge amount of rework by reinventing the same wheel.

TIME-SERIES DATA INFRASTRUCTURE WITH DATA LAKE INTEGRATION

The choice of time-series or real-time infrastructure technology will depend on the enterprise characteristics and requirements. The market of real-time infrastructure systems varies in a few groups:

- Automation vendor-based like Honeywell PHD or Yokogawa Exaquatum
- Open source-based like InfluxDB, Graphite, and Prometheus
- Large Equipment vendor-based like Siemens XHQ
- Vendor independent systems like the OSIsoft PI System

Automation vendor-based time-series data infrastructure

Automation vendors like Honeywell and Yokogawa deliver their own dedicated real-time infrastructure. These tools integrate very well in their automation toolkit. The downside is that these tools have limited analytical capabilities compared to other toolkits, and don't integrate well in a big data environment.

Open source time-series data infrastructure

Systems like InfluxData have their origin in collecting real-time information from online systems for performance monitoring and alerting. Soon after the introduction of InfluxData in 2013, the interfaces for collecting real-time data rapidly extended in the world of social media. Use cases continued to extend in the IoT world. InfluxData is an integration of various open source initiatives: Telegraf for interfacing, InfluxDB for time-series storage, Chronograf for visualization, and Kapacitor for detecting and alerting.

Equipment based time-series data infrastructure

Equipment vendors like Siemens need systems to optimize the service they deliver. They need time-series systems for remote monitoring of their large rotating equipment, like wind turbines. The growth of this turbine market pushed the development of these platforms forward.

Independent vendor based time-series data infrastructure

Independent vendors started to address the various gaps in data collection, analyses and visualization. Two vendors stand out in this area: AspenTech with the InfoPlus 21 system, and OSIsoft with the PI System time-series data infrastructure. Where InfoPlus 21 is more focused on smaller scale, MES-like functionality and local plants, the OSIsoft PI System is designed to be an all-purpose real-time infrastructure from a single set of assets like wind turbines, to a whole plant, an enterprise, or even a community of enterprises, vendors and regulators who need to capture, share and analyze data. The broad variety of interfaces to various types of data sources (450+) is one of the major advantages of the OSIsoft PI System toolkit. There is no restriction in getting data into the system. This means no additional development or unexpected IT cost to connect data sources. Meanwhile, a full context engine with streaming analytics enables the huge volume, variety and diversity of captured data, and turns it into valuable information in real-time that anyone can consume, from a plant engineer to a data scientist working within a data lake.

OVERVIEW OF CAPABILITIES

		Automation Vendor	Open Source	OSIsoft PI	Remarks
Collect	OPC UA	✓	✓	✓	Limited OPC UA support by many automation vendors
	OPC Legacy	✓	✗	✓	
	Industry wide	✗	✗	✓ (1)	(1) OSI PI has 450+ different types of interfaces
	IT equipment	✗	✓	✓	
	Varying Sampling rates	✓	✗	✓	
	Secure connections	✓	✗	✓	
Compress & Buffer	Raw data compression	✓	✗	✓	
	Buffer at the source (manage network outages)	✓ (2)	✗	✓	(2) Only in case history is available in automation system
	Handle sensor failures	✓	✗	✓	
	Redundancy	✓	✓ (3)	✓	(3) Available at a cloud level not at the data source level
Contextualize	Build data hierarchies	✗	✗	✓	
	Equipment definitions	✗	✗	✓	
	(semi) automated hierarchy creation	✗	✗	✓ (4)	(4) With Element Analytics tool integration
	(semi) automated tag mapping in hierarchy	✗	✗	✓ (5)	(5) With Element Analytics tool integration
	Event labeling	✗	✗	✓	
Relate	Ordering by time	✓	✓ (6)	✓	(6) Time synchronization of data is a manual process.
	Ingest other data	✗	✓	✓	
	Templates per equipment type	✓	✗	✓	
	Template based calcs	✗	✗	✓	
	Template based Events	✗	✗	✓	

Table 2: Comparison of Infrastructure Capabilities

		Automation Vendor	Open Source	OSIsoft PI	Remarks
Compare	Analyze equipment of the same type	✗	✗	✓	
	End user visualization	✓	✓ (7)	✓	(7) Limited end-user capabilities
	Dashboards	✗	✓ (8)	✓	(8) API's available for software developers
	Alerting				
	Real time streaming calculations	✓	✗	✓	
	Real time event detection	✗	✗	✓	
	Advanced Statistical Analytics	✗	✗	✗	
	Advanced analytics	✓	✗	✓ (9)	(9) With Element Analytics integration
Prepare	Context based data selection	✗	✗	✓	
	Data cleansing	✗	✗	✓	
	Scheduled data forwarding	✗	✗	✓	
	Streaming data forwarding	✗	✗	✓	
Integrate with Cloud	Data lake integration	✗	✓	✓	
	Geospatial	✗	✗	✓	
	Cloud hosting	✗	✓	✓	
	SaaS	✗	✓	✗	
Edge Computing	Data collection	✓	✗	✓	
	Data context	✗	✗	✓	
	Data analytics	✗	✗	✓	
	Data visualization	✗	✗	✓	

Table 2: Comparison of Infrastructure Capabilities - continued

DATA CONTEXT IS THE KEY TO SUCCESS

In order for data to drive business outcomes, it must be organized and accessible. Without structure, the data lake becomes a swamp. Individual data points have value for engineers very close to the production facility. Engineers usually know in detail how the facility is built and how to find each data point. However, as soon as reporting, monitoring or analyses happen outside of the local environment, it becomes important to add structure, governance, and context to the huge amount of available data points. Knowing the data individually by name is not an option anymore.

Figure 4: Streaming Operational Data to Multiple Applications

Example: Consider the contextual data that surrounds a single lube oil pump in a large facility. Each pump will have a parameter for pump name, power consumption, outlet pressure, outlet flow, outlet temp, and filter differential pressure. Furthermore, anyone in the organization should know where the pump resides, where in a process, and what may flow through the pump. Given the diversity of pumps and their various applications and processes, simply comparing all “pumps” is meaningless for analytics without this context.

A template approach makes this complex data context more accessible to all users. With templates, users don't have to search for multiple tag names for a data stream or need to know the name of the tag. All they need to know is the pump name. For the other parameters, you don't need to know the data stream names anymore. You make this connection between a specific pump and the actual data streams for this pump at the time you add (instantiate) the pump to your system.

Once all your assets are modeled on asset templates, access to the data is very easy. This makes it simple for non-IT staff to use the data, but also building applications and reports will become very fast and easy to deploy. However, one of the gaps for all available systems is the high amount of manual labor needed to make the connections between the data streams and the asset definitions. The problem is not with building the templates themselves, but with connecting instances of the templates to measuring points in the field. With larger systems of 100k+ data streams, this can become quite labor intensive and costly.

In the case of OSIsoft's PI System, a toolkit is available to automate and significantly reduce the effort needed to build templates and map data streams into a structure. This toolkit is delivered by Element Analytics and reduces the time to value by 80%.

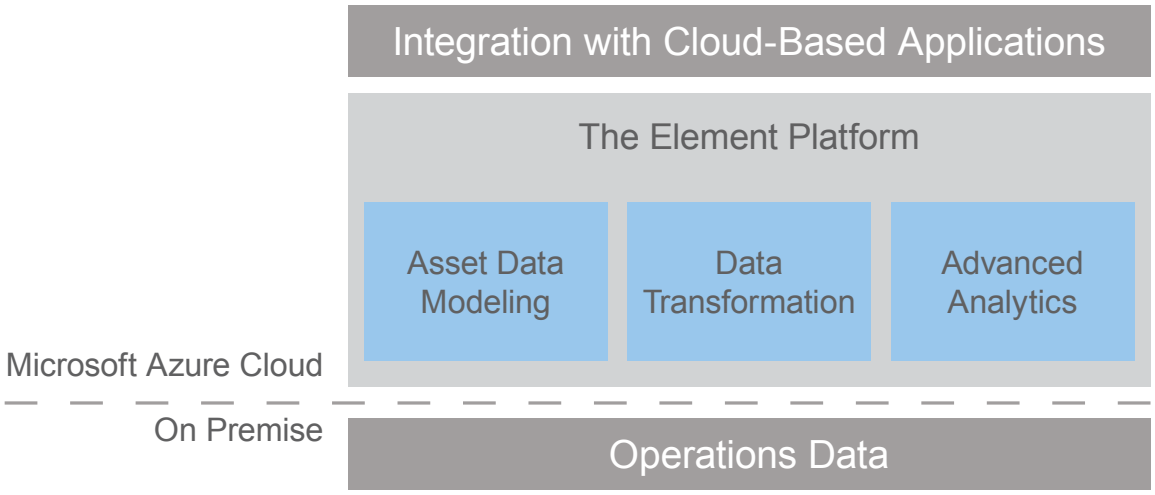


Figure 5: Leveraging Element Analytics for accelerating implementation of data structure

Data scientists need to consider not only the context of data, but also the data preparation. This is where most of the effort can be spent. Data scientists need to prepare the data by selecting the data set, cleansing the data, aligning data in time and formatting it in the correct layout. This is the greatest challenge to data scientists looking to utilize time-series data for advanced analytics. Agile self-service data preparation tools like the OSIsoft's Business Integrators, in combination with tools like Element Analytics, help to open up big data analytics for business users who are not IT specialists or data scientists. Companies like Cemex have shown that a traditional time-series data preparation that would take six months before it's ready for analytics can be reduced to four minutes of preparation time with the right tools. This agile and user-friendly way of working with the OSIsoft toolkit will reduce the time to value for business ideas significantly. In addition, less involvement is needed from IT specialists and data scientists, and it reduces the Total Cost of Ownership (TCO) for the same business value significantly.

CONCLUSION

The perfect world for industrial data processing does not exist yet. Pushing all production and operational data in a raw format to a central big data store will result in a data swamp instead of a data lake. Only specialized data scientists will be able to make sense out of the data. In an industrial environment, pre-processing of all real-time data is essential. Bringing context to the data is a must to assure that business users can leverage the data to optimize operations.

This means that in an industrial environment the combination of a data lake with a real-time infrastructure will bring all the benefits of Big Data processing like:

- Connectivity to the very diverse production and automation world
- Data Scientists will be able to find the big value items by combining all the data
- Enterprise application development and reporting is enabled by having a 'One-Stop-Shop' for data with a standardized data model for all assets
- Operational staff will have direct access to real-time data in a structured and agile way to optimize day-to-day operations

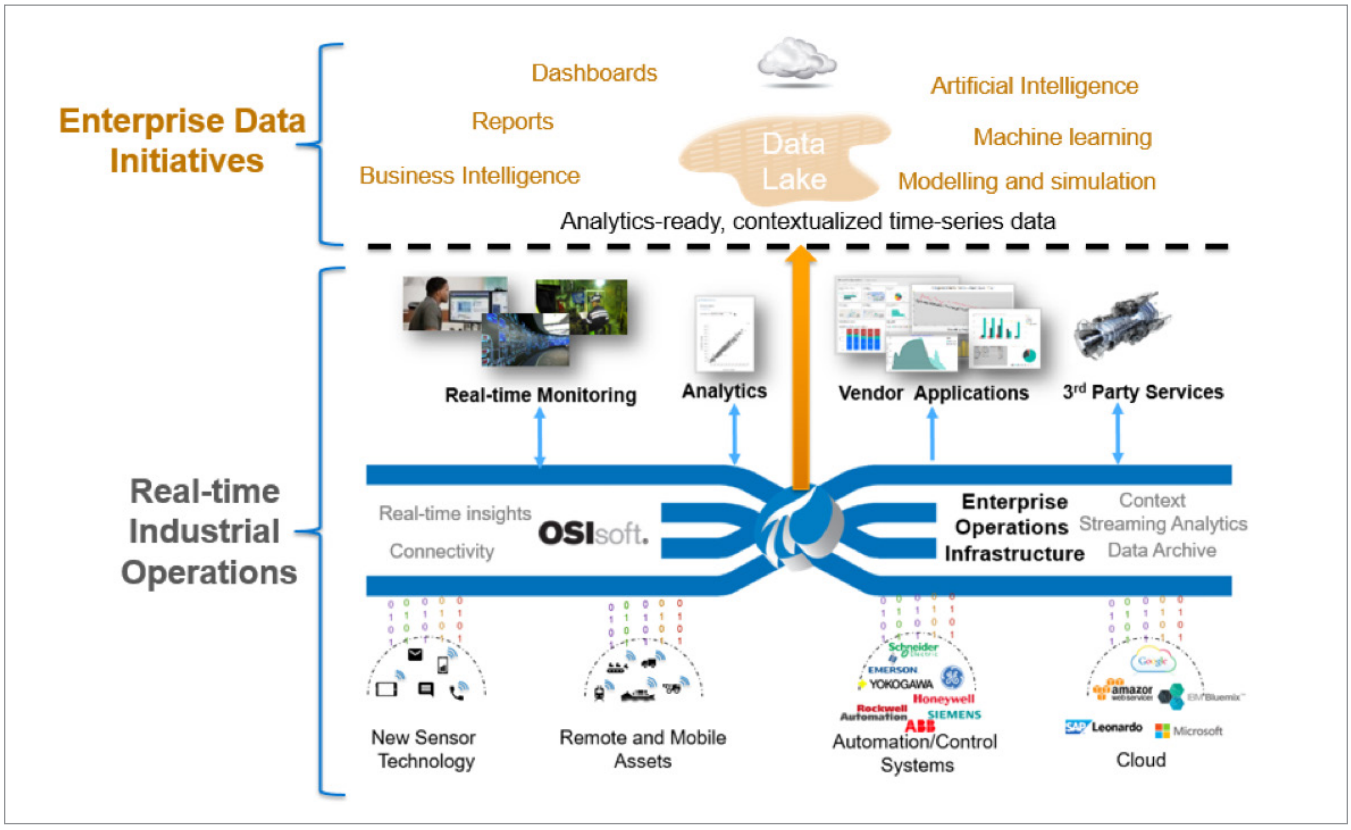


Figure 6: An Enterprise Operations Infrastructure provides the foundation to ensure analytics-ready data for data initiatives.

The seamless integration of OSIsoft's PI System on the production and automation level (interfaces and connectors) and the seamless integration on the Business Intelligence level with cloud and data lake integration makes the OSIsoft's PI System infrastructure a very popular product to bridge the gap between production and data lakes.

In addition, there is no need for software development and complex IT infrastructure to run the PI System which is built on a self-service model. This reduces the need for large (costly) IT teams to make an OSIsoft's PI System implementation a success. The majority of business innovation can be done by key business users (subject matter experts) themselves. Simple integration, no additional development, and simplicity in use will drive down the TCO for this type of infrastructure significantly. The combination of OSIsoft's PI System, with capabilities supported by vendors like Element Analytics, for data modeling and analyses, and the integration of all enterprise data in a data lake platform, will provide an environment for easy implementation and fast realization of value from big data processing.

ABOUT KSG SOLUTIONS

KSG-Solutions is a service and consultancy company with focus on industrial information systems. KSG-Solutions is founded with the objective to help industrial companies to generate more value out of their installed assets by implementing smart solutions based on off-the-shelve IT products. These solutions will help to drive higher asset availability, increased integrity, lower energy consumption, and higher overall productivity. 40+ years of experience with Oil & Gas majors and 30+ years' experience in Real-Time data processing and MES systems, forms the basis for the services provided by KSG-Solutions.

For information, please visit our website at **www.ksg-solutions.nl**

WPLSEN-102617