Architecting an Industrial Sensor Data Platform for Big Data Analytics

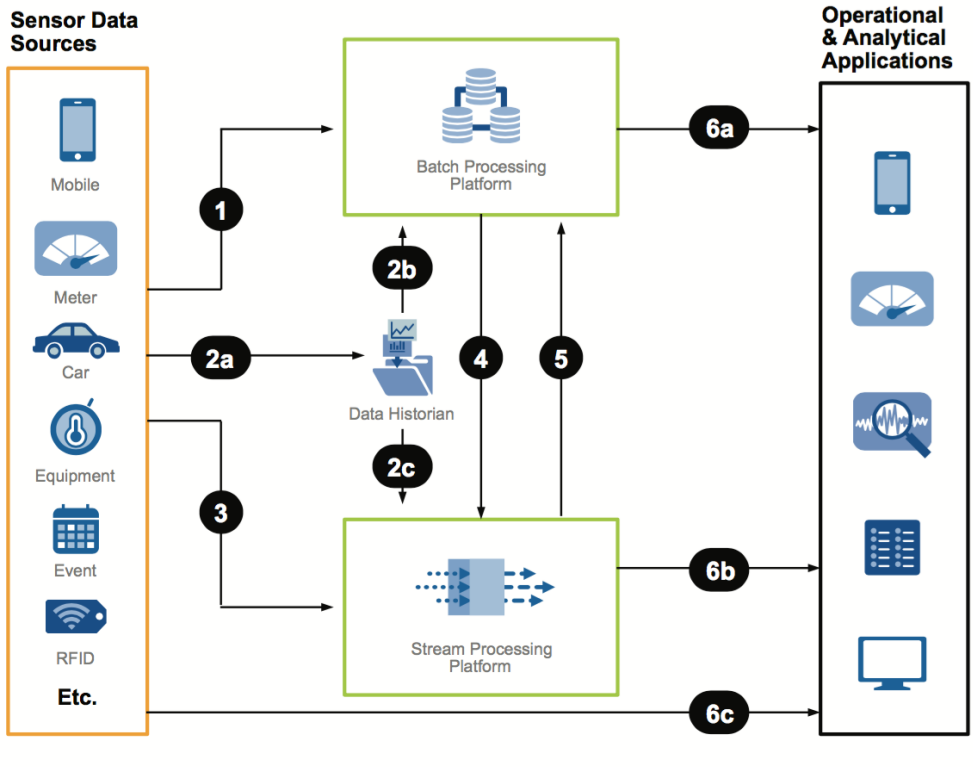
OSIsoft®

Welcome

For decades, organizations have been evolving best practices for IT (Information Technology) and OT (Operation Technology). With the evolution of the IIoT (Industrial Internet of Things) promoting machines with more sensors and corresponding data, there is a greater propensity to apply Big Data and analytics strategies to OT (Operations Technology) centric processes and discrete sensor-based operations technology and data. The amalgamation of Big Data and OT strategies, approaches and data is now revealing novel operational and business insights. Insights that are delivering transformational process, asset health, energy, safety, regulatory and quality improvements.

At the foundation of any Big Data architecture that leverages sensor-based data is a data historian that delivers one version of the truth for sensor data. We know this foundation enables sensor-based data to be consistently, securely and reliably streamed or batch processed into Big Data.

Figure 1 - Blueprint Diagram



Source: Gartner Research Note G00263936 (June 2014)

- Presents one version of the truth to analytics engines
- Captures high fidelity time series data
- Harmonizes time series data for consistency
- Performs common calculations close to source
- •Delivers contex of the source and relationship of the sensor providing the data stream
- Delivers context of events within a time series
- Supports export of data to Big Data engines
- Supports data cleansing augmentation and shaping of data upon transfer to Big Data engines
- Recaptures and associates Big Data analytics findings with source data
- Infrastructure for today's and tomorrow's time series data sources

When investing in a sensor-based data architecture, we know it's critical to understand the questions the Big Data strategy is trying to address. This will provide the governance for the type, quality and volume of data that the architecture will need to handle and which parts of the architecture needs to handle the time series data. For instance, if long-term historical records need to be interpreted to answer the question, then the historian provides an ideal environment to capture and later share the data for analytics. With a historian in place, the data processing strategy for the Big Data analytics will be dependent on the questions being asked. If the question requires large volume of feeds in real-time and to present real-time analytics for immediate decisions then an infrastructure on top of the historian is required to aggregate and present data to streaming process platform for Big Data analytics. Alternatively, if the persistence of the data is not important, and an analysis of a large number of feeds from multiple different sources is needed over a relatively short amount of time, a batch processing platform may be applied.

This paper draws on over 35 years of company history managing sensor based data across Fortune 500 enterprises. The paper discusses a strategic approach for sensor data archival and utilization in Big Data analytics. Extending the need for a historian, this paper discusses the use of a sensor-based Data Infrastructure that represents "consistent and current", one version of the truth and critical capabilities required for successful Big Data analytics.

The Evolution Of Historians For Sensor Data

Historians have been used for over 25 years to capture time series data from sensors. For decades engineers have wrestled with trying to store time series data using conventional relational databases and non-relational databases. Today, the leading historians utilize non-relational database approaches to storing large amounts of time series data over a long period of time. However with new technologies and scalability improvements in relational databases one of the first dilemmas is identifying the time series data archive technology for the historian for Big Data architecture.

Ultimately the choice of relational or non-relational database technology depends on:

- Data Fidelity
- Data volume and diversity
- Data longevity and persistence
- Data analysis needs

For industrial operations it is very important to recognize that the volume and frequency of data is hugely different to typical business feeds. In an industrial process data at the sub-second level can have a significant meaning, the number of sensors for an operation can run in the millions and there's a need for long term archival, indexing and reporting.

As a result for localized, low-fidelity applications a relational database approach may be sufficient. However for long term, high fidelity enterprise- wide industrial time series data management, a non-relational database approach with optimized indexing for time series data is recommended.

Application	Assets	System Age	Stored Events	Data Size	Events Stored/sec
Syncro Phasors	4.8K data streams, 120Hz	3 years	55 Trillion	430TB	>500k eps
Data Center	100k cells 2M breakers	10 years	105 Trillion	840TB	>20k eps
Automated Metering	20M Meters, 5 min/ reads	7 years	177 Trillion	1,400 TB	>1.3 MM eps
Fleet Monitoring	1K, 1 M points	10 years	6,307 Trillion	50,500TB	<1MM eps

In addition to non-relational database approaches, the other focus for the historian over the last 25 years is to drive centralization and context of time series data. Today, historians can access over 450 sources of sensor-based data across multiple applications within the enterprise, and these sources are continuing to grow. Without a strategic approach to data management, the value of the data would be diminished. As sensors are installed that aren't connected to control systems, the IIoT will explode the number of potential streams and sources. To overcome the challenges of multiple silos of data, consistency between data streams and the lack of context to help people and applications interpret and share data, today's historian has transformed into a sensor Data Infrastructure. A Data Infrastructure that layers on top of the physical industrial infrastructure to connect people, assets and data.

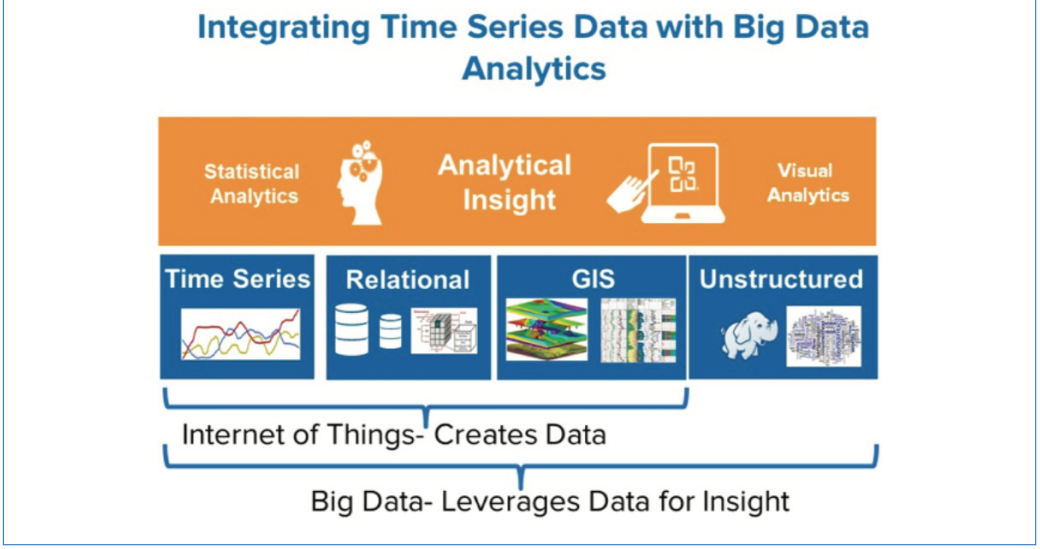
The historian has evolved beyond a system of record for sensor-based data to an enterprise Data Infrastructure that enables not just visibility but insight across traditionally silo'd systems. As part of this evolution, the historian has evolved data best practices to support advanced analytics that allow data streams to be compared between assets, between systems, between processes and between operations. This has largely been driven through incorporating metadata, context and data processing and enrichment capabilities.

Introducing A Data Infrastructure Approach For Sensor Data

One of the greatest challenges faced with Big Data analytics of sensor data is acquiring the sensor data in a scalable, reliable and consistent manner. Applying unified rules, logic filters to assure data quality and accurate representation of what really happened, in context, consistently, across systems is essential for Big Data analytics. In many companies there are multiple repositories of sensor-based data (EAM, MES, CBM for example) and continuous streams of data coming from process, control and automation system interfaces (SCADA, DCS, PLC etc.). This leads to several issues:

- Availability of historical and current data for analysis
- Data silo's from disparate systems that need integrating
- Inconsistency of data preventing data aggregation and comparison
- Missing data context preventing times series data being reliably compared
- Inconsistency of metadata content preventing reliable data interpretation

Figure 2 - Integrating Time Series Data with Big Data Analytics



Furthermore, many Big Data strategies rely on creating data lakes prior to analysis. An inherent challenge with the data lake concept is that lakes grow stale unless fed with fresh input. It's therefore imperative that if data lakes are utilized for analysis, they are either considered disposable or a mechanism is put in place to keep the data lake fresh.

Finally, as new ideas, conclusions and insights are discovered, it's important that there is a strategy to apply these insights back to the data feeds feeding future analytics initiatives. These insights can be in the form of set rules, calculations, templates, or patterns that will continue to feed future analytics.

To overcome these challenges, OSIsoft recommends extending the approach of a traditional historian. Using a common Data Infrastructure for supporting sensor-based time series data that provides Big Data analytics engines with:

1. One current version of the truth for sensor based data
2. Data context to support analytics
3. Two way integration of sensor-based analytics

1. Sensor-based Data Infrastructure: One current version of the truth for all sensor data analytics

Today's industrial environment can contain hundreds of control, process and automation systems all capable of streaming time series data from sensors. At the foundation of every Big Data architecture is a data archive for capturing time series data that can be leveraged for Big Data analytics.

This data archive serves as a layer to present one version of the truth for time series data streaming from sensors to Big Data engines. Traditionally this system of record has been referred to as a historian. However to integrate into a successful Big Data analytics strategy a traditional historian system of record approach for sensor-based data has deficiencies that need to be addressed to support the best practices mentioned above.

Critical to successful Big Data analytics is completeness of data, having centrally available and consistent data and having the most current data that represents one version of the truth for time series data. To ensure one version of the truth is presented to the Big Data analytics engines, several capabilities are often overlooked:

- Capture of sensor data from a diversity of sources
- High fidelity data capture
- Primary calculations and rules near source
- Time series event frames
- Predicted data enrichment
- Hybrid ecosystem of today and future sources of data

2. Sensor-based Data Infrastructure: Data context to support Big Data analytics

With sensor data streaming from thousands of devices and hundreds of processes, streams of data can quickly get unmanageable and subject to incorrect analysis and interpretation without governance. Essential for sensor-based data management to support Big Data analytics is a context model that not only defines the context of a sensor within a process and operation but also provides context management that allows data to be reliably rolled-up and compared to other data streams across operations. To deliver the governance and enable integration the Data Infrastructure requires three essential capabilities:

- Sensor and Asset Data Context
- Sensor and Asset Context Management
- Sensor and Asset Data Calculations and Translation

3. Sensor-based Data Infrastructure: Two way integration of sensor-based analytics

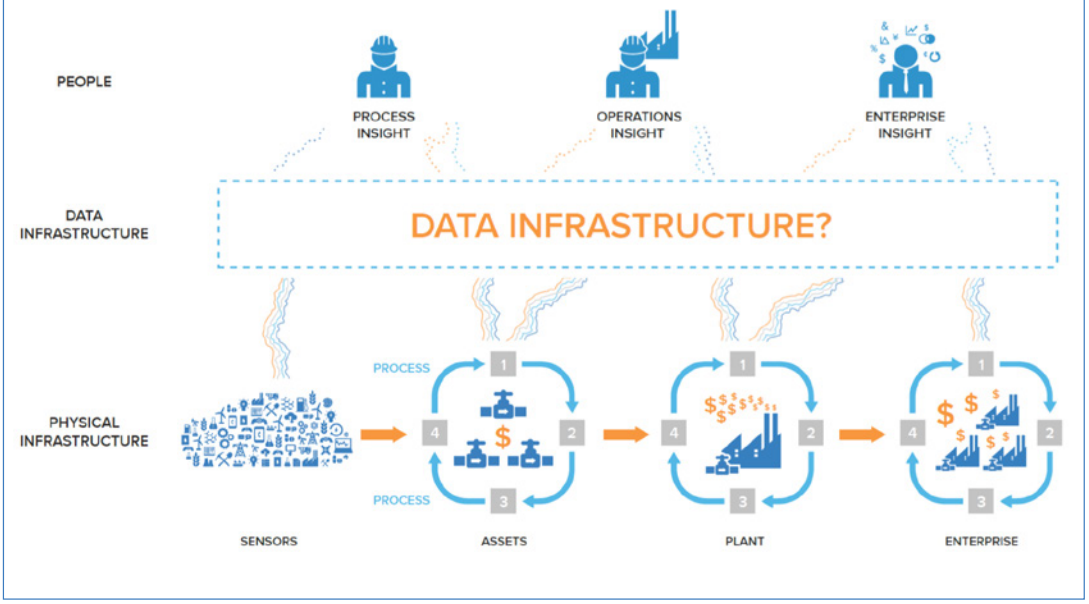
While a historian is essential for capturing and making available time series data to Big Data engines, the value of data is diminished if the data cannot be easily extracted in an utilizable way. Importantly valuable analytical findings from Big Data engines should be returned and captured with the original sensor data so others can benefit from the added insights gained. These insights could

be in the form of rules, calculations, templates, or patterns that could be used in-situ with incoming data into the sensor archive, and out-going data to analytics engines.

To support a Data Infrastructure approach, it is recommended that the infrastructure is able to deliver more than the traditional historian and support out of the box integration with the Big Data analytics environment. To ensure data can be utilized in a reliable and scalable way the integration must provide means upon export to:

- Cleanse
  - Remove surplus or irrelevant data points as part of the data streams being shared for analytics
- Augment
  - Add external data, metadata or context to improve interpretation of the data during analytics
- Shape
  - Present the data in the right format and relational aspect to support the business questions being asked
- Transmit
  - Move the data into batch or streaming analytical processes within the Big Data engine

Figure 3 - Integration with Analytics



Upon completion of analysis and to ensure continuity of the analytical investment, it is important that key data analytics and findings are re-associated with the time series data. This approach enriches the source of time series data for future analytics and at the same time prevents future duplication of effort.

In Summary

Commerce has invested billions of dollars in people, physical assets and processes. Today the promise of more affordable and better connected equipment and machines with sensor data opens new opportunities to deliver data intelligence for Big Data analytics to maximize the value of these investments. Often missing in Big Data strategy is a complementary sensor Data Infrastructure layered on top of the sensors, machines and systems to provide governance and connectivity to deliver one consistent and current version of the truth to fuel Big Data analytics.

At the foundation of any sensor-based architecture for Big Data strategies is a historian that captures and manages time series data derived from sensors. As historians have been capturing sensor-based data in the industrial world for over 25 years they offer an ideal opportunity to apply Big Data analytics to the huge volumes of legacy data to enable novel operational insights into industrial processes.

As the diversity and volume of sensors and time series data continues to expand, the role and capabilities of the historian within the industrial world have been advancing. Today, the historian has evolved beyond a system of record to a common sensor-based Data Infrastructure for industrial operations. An infrastructure that provides governance to data. Governance that includes consistency, context and centralization of data that ultimately improves the utilization of existing and future sensor data within Big Data engines.

With the IIoT hype becoming a reality the sources of sensors based data and locations is set to explode. There will be more data in more formats and in more locations that ever before. In essence regressing data management decades unless the right architecture is adopted. An architecture that provides a single, sensor-based Data Infrastructure for capturing and presenting time series data to the enterprise and Big Data analytics engines.

About OSIsoft, LLC

OSIsoft is a global leader in enabling operational intelligence for over 35 years, delivering an industry standard sensor-based software infrastructure, the PI System, deployed to:

- 125 countries
- 17,000 sites
- Nearly 1 Billion sensor data streams

OSIsoft empowers companies across a range of industries in activities such as exploration, extraction, production, generation, process and discrete manufacturing, distribution and services to leverage streaming data to optimize and enrich their businesses. OSIsoft customers have embraced the PI System to deliver process, quality, energy, regulatory compliance, safety, and security and asset health improvements across their operations.

OSIsoft provides over 450 connections for process, automation and control systems. These connectors facilitate the streaming, processing and storage of high fidelity time series data into a central archive ready for Big Data analytics. Meanwhile an Asset Framework provides a context and computation layer for enriching sensor data for further utilization. New advanced data processing capabilities support the cleaning, augmentation, shaping and transforming of existing, new and predictive data to be Big Data ready.

With over 200 partnerships OSIsoft is actively working with technology leaders and hardware vendors to embrace IIoT and extend its offering to leverage cloud-based technologies and deliver interfaces at the edge to stream sensor data to centralized data archives and processing tools ready for Big Data batch and streaming analytics.

Founded in 1980, OSIsoft is a privately-held company, headquartered in San Leandro, California, U.S.A. with offices around the world. For more information visit www.osisoft.com.

Figure 1 - Blueprint Diagram