Getting Results Using the PI System and Big Data

Gregg Le Blanc
Director of Research and Innovation
OSIsoft

TABLE OF CONTENTS

1 OVERVIEW

In a series of explorations pertaining to the use of the PI System™ and various types of "NoSQL" type solutions, great emphasis was placed upon the implementation and scope of the target problem space. Specific attention was placed on producing better quality data so that validated, trustworthy data was produced from a single, reliable infrastructure. This helped maximize the reusability of the effort involved in delivering trustworthy, instantly usable, operational data in context.

Because of the similarities of certain elements of the PI System to certain NoSQL technology types, the qualities that differentiate the results of an implementation will substantially rely upon dividing the tasks of maintaining a rich, durable "system of record" and an agile analysis platform. The long term total cost of ownership involved in solving the initial problem, maintaining the solution, and solving new use cases as they arise in an "infrastructure-oriented" (rather than point-solution) manner will guide architecture choices as these implementations cross operational and business boundaries.

While the PI System can be considered a "NoSQL" database in some cases, its user-orientation, purpose-built data acquisition interfaces, extensibility, and analytical tools for time-series data set it apart from generic columnar data storage tools.

This strength lends tremendous value to any layered application that builds upon the PI System at an infrastructure-level. Gartner deemed the vast array of "Operational Technology" data found in manufacturing facilities, plants, and other data-rich, "non-IT controlled" entities as "dark data" due to its opacity to the business. The PI System helps collect and put that data in context for users everywhere.

The overlap of the business and operations is where the PI System and Big Data will provide the most value to the entire business. However, IT and OT data are "shared space" and business or operational needs do not exist within a vacuum. Therefore, the one-way transmission of raw data to the business (or vice versa) serves little purpose because it lacks the rich context and expertise embedded in its source.

1.1 Creating a shared data space

Today's businesses are increasingly interested in doing a more flexible type of analytics on larger data sets using "NoSQL" technologies and Map/Reduce type analytics. The stated goal is quite similar to the PI System' goal for operations: "gather all the data into a common format for later discovery" - essentially formulating questions and developing insights later.

A key difference between many NoSQL systems and the PI System is that the PI System usually serves as both a "system of record" as well as an analysis engine. Some companies treat NoSQL installations as additional systems for analysis purposes that do not replace (or introduce redundant) systems of record.

In either case, users in the IT or OT domain use tools in either suite to derive insight into their pressing issues or untapped opportunities. The shared space between operational and business domains opens up an area where questions may involve several time spans, data shapes, and systems of record.

Big Data systems can serve these types of purposes when underlying systems of record provide decision-ready, trustworthy, high quality data at query-execution time. These necessarily data-hungry applications look to aggregate almost any source available so that useful correlations can be produced and consumed by users elsewhere within the organization. By minimizing the transformation and refining burden required within the Big Data layer, a data scientist's focus can be maintained on asking better questions using better data curated by domain experts.

These powerful analytical and technical tools can provide new insight to a multitude of audiences. Maintaining the agility of the Big Data platform means dividing the data shaping and scrubbing labor intelligently between the strengths of each architecture participant so the resulting applications lead to better, quicker results.

Additionally, performing these tasks closer to the originating data domain means preserving the contextual information, expertise, and using *better data* in everyone's applications (not just Big Data) - that mean *trustworthy data* rather than *raw data* - which will lead to better insights everywhere.

Sharing analysis results back to the originating systems of record (where possible), or manufacturing *better quality data* at the system of record will benefit the operations and / or business as a whole in the face of the forthcoming business and technology changes we face today.

This paper will apply a broad brush to the architecture of a couple common "Big Data" use patterns and attempt to discuss where the hand-off between the PI System and common Big Data type systems make the most sense. These are simply heuristics to start discussions that pertain to leveraging the strengths of in-place technologies and skills.

2 THE PI SYSTEM AS PURPOSE-BUILT NOSQL

Before the term "NoSQL" was actually a term, the PI System was storing data in its columnar format (one of the so-called "4 Pillars of NoSQL", but not quite a Key, Value store). The suite of supporting applications, interfaces, and user features around the PI System have been purpose-built to assist in the unification of a broad, proprietary information space that typically was either opaque or tremendously difficult to correlate between disparate vendors (or both).

A key, additional difference in the basic design of the PI System is its orientation around time-series and "streaming events" processing. While some "NoSQL" technologies will play in this arena, they are not necessarily built with this purpose in mind. The PI System also serves as an aligned, harmonized, durable system of record for events that cannot be persisted for long periods of time in other underlying systems.

Issues such as "out of order" data, updates, repeatable events, fault tolerance, and calculations were designed into the architecture of the PI System as continuous inputs. The system was built with the expectation that continuously updating values would stream into the database.

The raw data from PI System sensors, manual inputs, and other historians is normally enhanced, tested, and streamlined into "better data" (i.e. via "data quality" tests) through the application of rules created by process engineers, experts, and other data scientists. These are applied within the PI System through various calculation tools and can be exposed to consumers so that they get the required data and the fidelity needed for the questions they develop.

Examples include:

- Flow rates are converted to production rate "per day" or "per shift"
- Production downtime is accounted for at the production line itself
- Sensor outages are handled through logic and updated after repair
- Repairs etc. are accounted for by staff nearest to the activity in the facility
- Second by second data may be rolled up to averages at an appropriate fidelity for target systems/consumers
- Engineering unit conversions are made between systems

When designing any solution that involves PI System data, the target consumer's data environment must be taken into account. A necessary conversion is made when transitioning between a system where "event flow" becomes aggregated into "transactions" - and ultimately, fidelity or alignment between raw data will be changed.

A streaming, time-series system, such as the PI System, operates natively in the "time domain" and thus, calculates results differently. Consider the daily totals previously mentioned. The averages, and "standard deviations" that are calculated in the PI System are "time weighted" (because they are actually samples) vs. "count weighted" (i.e. evenly spaced over a representative distribution) because of its inherent calculation capabilities.

Additionally, considering how to best apply the strengths of a PI System to a Big Data problem will require the proper alignment of metadata. Constructing a "metadata model" in the PI System, using the correct units of measure, and the appropriate per-entity roll-ups can make a tremendous difference in streamlining upstream analytics and reducing unneeded (and confusing) data duplication.

3 COMPLEMENTARY USE AND SCENARIOS

Dividing the analytical labor between the time-series and transactional realms until the point where the "alignment" and "harmonization" make the most sense, and the "domain expertise" is closest to the data source will provide the best grounds for an agile application.

Truly, Big Data isn't a "data problem" at all (i.e. correlation is not causation). It is really more like a conversation between operational entities or a "Big Question" problem - and asking a valid question is still the hardest part of any problem. When a good question is developed, the tasks that are created from the question should span (and leverage) underlying systems. Those systems should become enriched to support the question if it is a recurring pattern of inquiry.

The PI System should simply be one of these systems. If an important calculation needs to be created in the time-series domain (or extra metadata needs to be created in the Asset Framework (AF) Database to enable querying), then the benefit of having that result "on tap" for other users increases the value of the infrastructure for later.

Treating the PI System as an authoritative "system of record" for time-series data and analytics derived from its domain keeps "Big Data" agile because the answers can always be reliably be reconstituted from the source.

What's more - when new sensors come online, performing more textured and detailed year-over-year analysis become easier if the PI System remains an authoritative source in the face of changing Big Data technology. This allows the Big Data customer to adapt as their needs change.

A few common Big Data patterns and practices follow that will illustrate the role of the PI System.

3.1 Metadata

When businesses turn to Big Data solutions for insight into patterns, they are typically crossing and developing correlations between many silos of data within the organization. When this includes "Operational Technology" data (OT Data - as Gartner says), the PI System may be one of several sources. The key component of finding and traversing the disparate systems is the development of a singular view of the problem space and resulting information space.

That "singular view" must encompass a few requirements, such as:

- A shared reference or key to align the data
- A common nomenclature or model for reference
- Consistent engineering units
- A common time scale

- An efficient query execution plan to enable data gathering and analysis completion in a timely manner

The PI System can manage a portion, or all of an operation's organizational structure from a topological perspective by assigning logical groupings to flat sets of measurement data streams. These structures can be augmented with calculations, engineering units, aliases, etc. in order to create *better quality, trustworthy data* for use in other applications.

Metadata structures and the analyses that references them can also be built by referencing a master model that is built and maintained in another system. By adapting frameworks or industry standard organization principles, some OSIsoft customers have elected to implement a queryable structure using a "Resource Description Framework" model and "Web Ontology Language" for rationalizing aliases to common assets with duplicate names (these nomenclatures are known as "RDF and OWL"). These queries are often processed in a SQL-like syntax called "SPARQL" and allow a user to traverse many systems via a common model such as the "Common Information Model" or CIM, ISA-88, or ISA-S95.

Systems that house RDF and OWL models can be based on various technologies, but are usually considered a "library" or "metadata system of reference" for a business. They will authoritatively update dependent systems and templatize interactions at a metadata and infrastructure level so that every system will benefit from reusable work. These systems will have a storage mechanism, a programmatic surface for extensibility, and a query interpreter to help develop and update relationships between systems.

Through the implementation of a common query syntax in SPARQL, asset names in the PI System can be created and updated properly. This also means that business systems, such as Big Data implementations (e.g. Hadoop) can find the information required for analytics without knowing site-specific information. By applying RDF and OWL principles, site specific nomenclature can be kept alongside a more consistent naming convention understood by the business without impeding the work done at a site level (and without duplication of data).

3.2 Data Quality and Alignment

While the data from various systems can be combined easily within a "Big Data" tool, the preparation of the data and its transformation into a usable state is something that requires expertise in both the domain of the operational process and the tool set.

A common challenge with time-series data is dealing with real-world behavior of "outages" (whether due to nature, technical faults, or just delays). Additionally, detection of "outlier" values, stuck instruments, or faulty measurements will require both an acute action plan if the data is required immediately, as well as a recourse for eventual recalculation once replacement values arrive.

The fundamental practice of preparing better quality data for use in applications distills down to whether the logic is best kept with the system of record (nearest the aggregation point), or within the analytical query execution.

The architecture that promotes the most re-use of this work should promote the best data quality and best future-proofing of any analytical tool.

Usually, the rules around data quality are automated, but as new instrumentation and "edge-cases" arise at the site, these rules will be updated. As such, keeping the set of rules consistent with the time period in the system of record over which they are valid is paramount when rolling up data for use in higher level systems. The domain expertise required to keep these moving parts aligned should be carefully considered (as well as the roles and responsibilities involved) when developing the correct data quality rules for use by an application.

In summary, alignment and transformation of the data from a time-series realm into a more "transactional" realm should happen at the last possible point. This ensures the least amount of overhead is incurred and the most "texture" and domain expertise is preserved within the systems of record.

The goal of this practice is to lower the potential data duplication, redundancy, and maintain a "single version of the truth."

3.3 Pattern Development

A common "Big Data" practice can be described as "pattern recognition". This can manifest in several ways. Once the metadata layer of the Operational Technology data is no longer opaque, and the data quality being produced from the operations side of the business is trustworthy, the patterns that can emerge may fall into two common zones:

- Customer behavior
    - Typical use pattern identification (i.e. profiling)
    - Outlier identification
    - Product or service call resolution
    - Customer improvement opportunities (e.g. underserved areas, outages, etc.)
    - And more...
- System or business unit behavior
    - Hourly / Shift-wise / Daily / Weekly / Monthly / Yearly trends
    - What-if analysis
    - Performance evaluation
    - Meta-analysis of analysis effectiveness
    - And more...

Patterns emerge by segmenting an information space by population or customer demographics and then building a graph that helps separate it be common characteristics. Geography is a fairly simple place to start, but refining that graph begins to uncover the types of data required to develop an answer.

For instance, examine this question:

- How does the average customer within <Zip Code, Neighborhood>, who paid their bill <late once last year>, use energy in the winter <Last Year> compared to the low demand months?

A company might ask this to identify targets for a new rate program for high demand months (due to house insulation, poorly efficient home appliances or insulation, etc.).

The data would come from several systems (these are for illustration purposes only, actual implementation would vary):

- CRM for the customer data
    - To identify specific target customers
- GIS / CRM for the layout of billing and location data
    - To get key indices for meters
- The PI System for meter data
    - PI AF Metadata for meters needs to be tagged and Geo-Fenceable by Zip Code / meter area ID etc. to facilitate query
        - Totalized data per month for customers (pre-calculated, high quality data only)
    - Transform based on correct units, time scale of request interval
    - Totalized data per month per Zip Code area (pre-calculated, high quality data for each zone)
        - Transform based on correct units, time scale of request interval
- Hadoop / Big Data query
    - Link together a table of each data source
    - Perform Map/Reduce
        - Select customers where late payment was within "winter" months
        - GroupBy on Neighborhood within Zip Code
    - Visualize data to see which of these customers appear like they may benefit from a targeted discount program.

Clearly this is an over-simplification of the problem. However, the totalization of the data in the PI System can manage and maintain readings in a time-compatible format up until the time of query. If the metadata remains aligned, then the query should be faster and easier to maintain and perform than doing multiple transformations within the Big Data application using raw data from the PI System.

3.4 Increased Texture

Once patterns have emerged in Big Data, that doesn't mean that everything is known or static. New data constantly arrives and sources expand their footprint. Sensors are added, upgrade projects happen, so data is constantly on the move.

The PI System is there to help add more "texture" to any query. This is one reason "data replication" is to be avoided - one copy becomes obsolete almost immediately.

Like any "living system," PI System is constantly being updated, and when the metadata and sensor inputs get a set of fresh data, any Big Data queries that can adapt to these changes will benefit.

The time intervals over which Big Data analyses produce answers (e.g. CRM system time intervals) and the operations interval (i.e. sensors, devices, etc.) may not always be in harmony. But the presence of new sensors can augment learning algorithms, analytics, or demographically segmented graphs and influence models (e.g. Mahout, Numenta, etc.).

Here is how an "increased texture" use case could manifest:

- As part of a "continuous improvement program," the company is deploying new instrumentation and collecting new data for key customers in order to create better services on their behalf.

This could be for a number of different types of companies. A machinery vendor might attempt to deploy specific sensors around equipment to track and improve performance of equipment in the field (while also building a generic, anonymized asset performance reference database). A utility company might perform a "building audit" or "residence audit" to help customers identify inefficient appliances or equipment. A vehicle company might deploy a set of onboard diagnostics to a fleet of vehicles to help profile common issues.

The Big Data question might manifest in two ways:

- Identify <target customers that own asset with improved sensors> in order to offer them a new service that helps improve that asset's performance. Provide baseline performance data and projected savings based on a new data model.
- Rebuild the profile of <the asset with improved sensors> and compare it to a previous comparable data set to determine the program's efficacy. Monitor customer uptake of new service offering to determine if it <avoided costs/added revenue>.

The capture of new data into the PI System is a fairly traditional use pattern. The augmentation of the metadata structures, data quality, and supporting analytics need to be rolled out in a way that supports the Big Data queries.

The key difference is the augmentation of the metadata and sensor data over time periods where Big Data queries might already be built. If the Big Data queries can "inherit" new information (or with minimal effort) from the underlying infrastructure automatically, then the improvement project gains should be realized with less effort (and this should be measurable).

Building the appropriate logic into the systems of record may allow for a lower total cost of ownership because these systems will return appropriate data quality and metadata structures that support the Big Data analyses without customizing the higher level applications.

4 SUMMARY

Through the application of best practices and appropriate division labor at for maintaining a better quality system of record nearer each data domain, the union of the operational and business data worlds can result in deeper understanding of the data they share. The augmentation of Big Data with OSIsoft's PI System infrastructure can lead to better insight that evolves in an agile manner.

To achieve this, the analytics and data quality must be performed at the appropriate level so that work done on the PI System infrastructure remains reusable by a wide variety of consumers. The focus of Big Data should remain formulating and answering good questions. These answers start with the production of *trustworthy, better quality data* at the system of record level so the Big Data level remains flexible.

Being able to separate the scrubbing, alignment, and transformation of the time-series data from the operational domain before it reaches the business domain keeps Big Data users agile and focused on developing business insights. The blending of these data types should be a conversation between operations and the business that creates better quality data in the PI System. It enriches the enterprise as a whole while reducing the opacity of "dark data" - allowing it to benefit users in the entire organization.

5 ABOUT OSISOFT, LLC

OSIsoft (www.osisoft.com) delivers the PI System, the industry standard in enterprise infrastructure, for management of real-time data and events. With installations in more than 110 countries spanning the globe, the PI System is used in manufacturing, energy, utilities, life sciences, data centers, facilities, and the process industries. This global installed base relies upon the PI System to safeguard data and deliver enterprise-wide visibility into operational, manufacturing, and business data. The PI System enables users to manage assets, mitigate risks, comply with regulations, improve processes, drive innovation, make business decisions in real-time, and to identify competitive business and market opportunities. Founded in 1980, OSIsoft is headquartered in San Leandro, California, with operations worldwide and is privately held.

For more information please visit **www.osisoft.com**

A white paper produced by: