The Rise of Industrial Big Data

Leveraging large time-series data sets to drive innovation, competitiveness and growth—capitalizing on the big data opportunity

Introduction

Industrial businesses have entered the age of "big data," whereby the volume, variety and complexity of data they manage is exploding at record rates. According to McKinsey & Company, "manufacturing stores more data than any other sector — close to 2 exabytes of new data stored in 2010."

Big data is the proliferation of data from various systems, devices and applications whose size makes it challenging to capture, manage, and process within a tolerable period of time using traditional software solutions. Big data sizes can range from a few dozen terabytes to many petabytes of data in a single data set.

Massive amounts of operational data are coming online with the ever-increasing set of advanced devices and equipment, a movement often referred to as the Industrial Internet[1]. Forward-thinking businesses are leveraging this data for operational excellence and predictive analysis to create a competitive advantage and accelerated growth.

This white paper discusses what big data means for the industrial sector and the significant implications it will have in the near future; how historians provide commercial-off-the-shelf (COTS) solutions to address big data challenges as companies accumulate larger and larger data sets; and how critical insights enabled by big data can significantly improve operational performance.

Leveraging big data is imperative as information is at the heart of competition and growth for industrial businesses. Data-driven strategies based on real-time and historical process information will help companies optimize performance.

[1] www.forbes.com/sites/ciocentral/2011/11/23/the-industrial-internet-like-facebook-for-things/; http://bits.blogs.nytimes.com/2011/11/21/81057/

The industrial data challenge

Today, the creation and use of big data expand beyond large web companies like Yahoo, Google, and Facebook. Businesses everywhere, including industrial enterprises, face mounting pressure to stay competitive with data-driven strategies—requiring increasingly more data, which results in the accumulation of larger and larger data sets. In addition, evolving and evermore stringent regulatory requirements necessitate the collection of more information as proof for audit and compliance purposes.

Manufacturing companies record tremendous amounts of process data, and this growing volume is becoming ubiquitous. For example, a CPG company that produces a personal care product generates 5,000 data samples every 33 milliseconds, resulting in:

- 152,000 samples per second
- 9 million samples per minute
- 545 million samples per hour
- 4 billion samples per shift
- 13 billion samples per day
- 4 trillion samples per year

Clearly the volume of data from which to extract value is beyond the capability of a traditional data management system. What is more, the challenge of managing big data for industry goes beyond the sh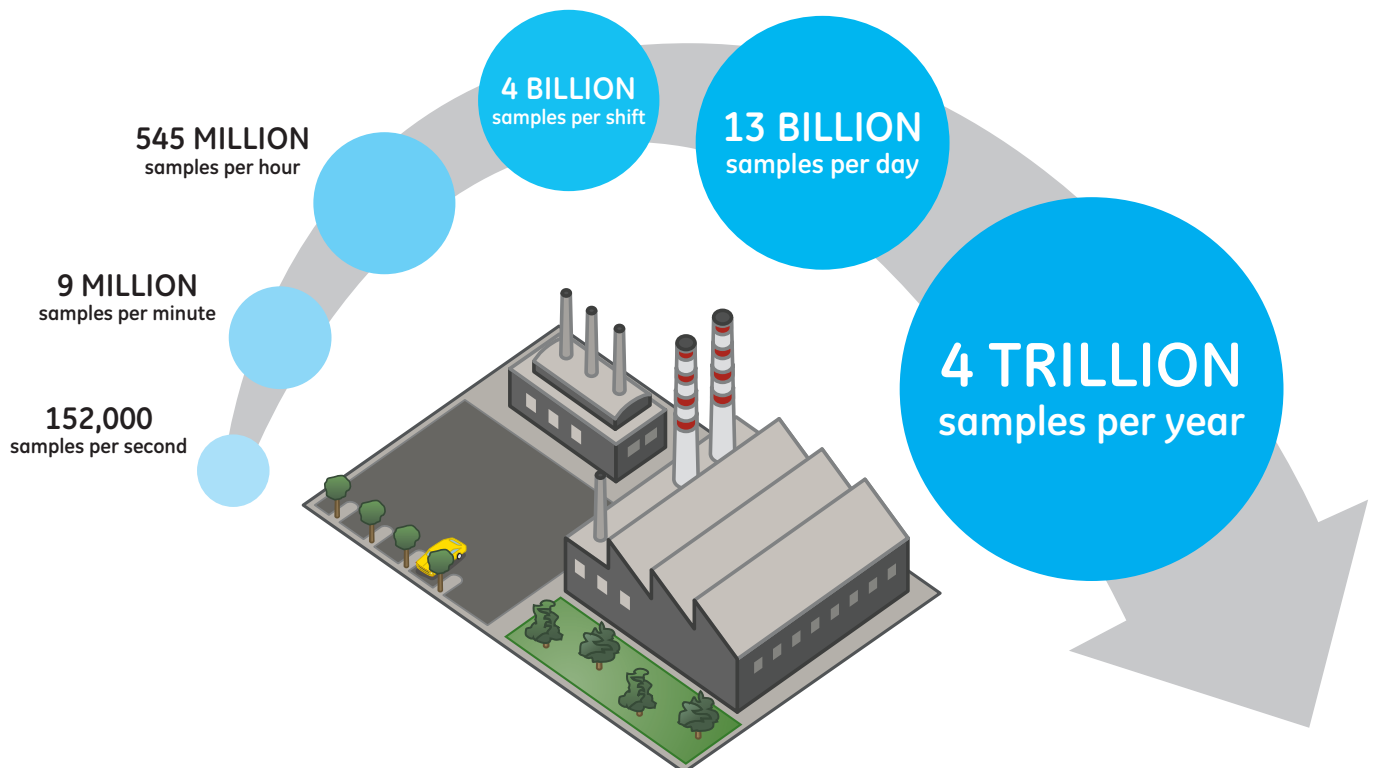eer volume of information; there is the diversity and complexity of data, which comes in various formats and from disparate sources. There are typically "islands" of process information that must be aggregated, stored, and analyzed to derive context and meaningful value.

To leverage big data, industrial businesses need the ability to support different types of information, the infrastructure to store massive data sets, and the flexibility to leverage the information once it is collected and stored—enabling historical analysis of critical trends to enable real-time predictive analysis. As businesses increasingly realize that much more of their value proposition is information-based, technologies that can address big data are quickly gaining traction.

Seeking an industrial data solution

Luckily for industrial companies, Google, Yahoo and Facebook are pushing the envelope on big data needs. Their desire to analyze clickstreams, web logs, and social interactions has forced them to create new tools for storing and analyzing large data sets. The groundwork these companies have created can also be leveraged in the industrial sector to manage the explosion of data that will only continue to grow.

For example, Hadoop is a tool that enables data storage scale through the use of commodity hardware, distributing data across many low-cost computers. Once distributed, new challenges

545 MILLION samples per hour

9 MILLION samples per minute

152,000 samples per second

4 BILLION samples per shift

13 BILLION samples per day

4 TRILLION samples per year

These figures reflect the data generated from just one of many machines that produce a particular personal care product, underscoring the sheer volume of data created by industrial companies.

The Rise of Industrial Big Data

arise in locating and processing the data, which are addressed by MapReduce, a framework where data is processed in parallel across many nodes in a cluster. It allows processing to be mapped to the data across many locations, and then reduces the outputs for similar data elements into a single result.

Hadoop is an Open Source technology that is rapidly evolving. New tools are needed to simplify the low-level access of data through standard query mechanisms and standard orchestration techniques.

Nearly every analytical insight a business would be interested in involves an element of time, calling on solutions that are specifically designed to leverage large time-series data sets for critical insight to capitalize on the value of their data.

While Hadoop may have big promise for handling large data sets, the complexity involved and the specialized skillset needed to create a Hadoop environment is often beyond the ability of industrial businesses. Yet these businesses still need to scale across the enterprise to handle large sets of time-series data generated in manufacturing and other industrial operations.

For example, a manufacturing manager may want to understand the significance of temperature variation on quality as the rate of flow of materials varies through a production line; or a power plant supervisor may want to analyze five years of past data to examine anomalies and variations to understand whether they were followed by subsequent outages to enable predictive analysis.

This level of operational insight requires the ability to quickly run a query against large data sets for specific time periods—a unique and powerful capability that calls for an industrial data solution.

The power of advanced historians

While historian software may not yet be top of mind when it comes to industrial big data solutions, what many companies may not realize is that these advanced, out-of-the-box solutions are specifically designed to efficiently collect, store and manage large volumes of time-series process data, which is precisely the industrial big data challenge.

As data sets grow larger and more complex, advanced historians offer an effective, simple, and easy way for companies to efficiently leverage vast amounts of real-time and historical process data, a critical need for optimized decision support. They help companies easily connect and collect their data from various systems and devices, making it accessible to uncover intelligence that would otherwise be locked away in the data.

The time-series-friendly data structures of advanced historians enable them to vastly outperform traditional relational or key-value data structures, efficiently running queries across large data sets and associated periods of time. Historians provide a much faster read/write performance and "down to the microsecond" resolution for true real-time data, capturing the value of information at the process level to continuously drive improvements.

Furthermore, advanced historians connect to process data sources and acquire the data directly—aggregating data enterprise-wide and compressing it for efficient storage, which greatly reduces the volume of data needed to accurately reproduce the time-series signals.

For the CPG company referenced earlier, a historian can reduce the amount of disk space required per sample by 85% versus using traditional database approaches. Historians inherently store time-series data much more efficiently than traditional methods through intelligent logging that "leans" out non-value added data points, which can take up significant disk space, yet still represents the "whole truth."

For example, the temperature profile of a storage tank over a month that remains constant is not interesting; what is interesting and value-added data is the time when it spikes up five degrees.
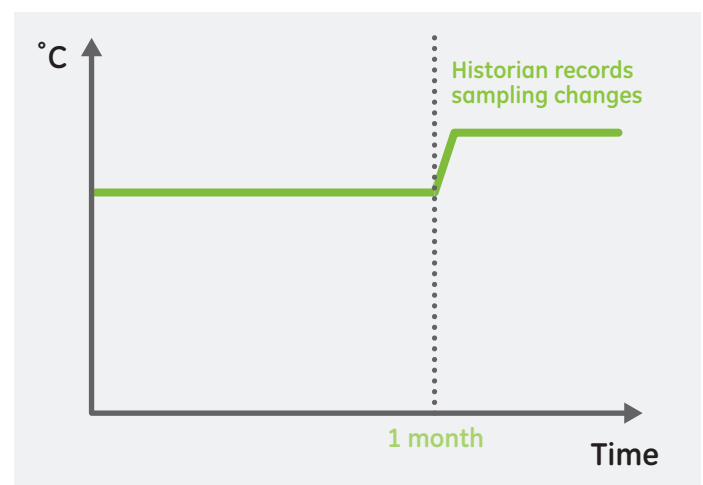
Historian's compression will compare the last recorded sample to the current and not log a sample if it has not significantly changed, greatly reducing storage resources.

Advanced historians can eliminate the non-value-added data—avoiding unnecessary consumption of resources and still replicate the truth of the temperature profile over that period of time.

Driving innovation, competitiveness and growth

With rich historian capabilities in place, industrial companies can fully leverage advanced analytics—efficiently running queries against years of historical data to identify trends and patterns for real-time decision support. They can gain insights to drive decisions that affect key areas such as product quality or lost production time, in addition to enabling continuous improvements across the enterprise such as:

- Proving quality to a trading partner or customer
- Maximizing yields - ensuring that processes in a factory can be tuned to the specific qualities of materials at each stage
- Recovering capacity – equipping plants with deep insight into all of the factors that cause process/equipment instability, so they can "engineer out" causes of capacity loss
- Facilitating supply chain execution with objective insights into things that affect completion times and material usage, which are key to managing purchase and logistics costs
- Simplifying and controlling the response to external parties' demands for data (regulators, NGOs, consumer advocates)

Case in point: Leveraging big data saves millions at GE Energy

Industrial enterprises stand to gain from big data only if they have the capabilities in place to make their time-series process data easily accessible so it can be analyzed to uncover critical business trends. With such insight, companies can improve their operational responsiveness and agility, using information as a competitive differentiator to set themselves apart from industry peers.

As an example, GE Energy's Monitoring & Diagnostics (M&D) Center in Atlanta, Georgia collects data from thousands of gas turbines in more than 50 different countries around the world—collecting data for its customers on the order of 10 gigabytes per day. Having to organize and interpret a constant flow of data on vibration and temperature signals delivered by sensors across its fleet, big data is part of the Center's everyday operations.

High data compression and real-time data access

To collect and manage its continuous data stream, the Center relies on GE's secure Proficy Historian software. Its powerful data compression capabilities have enabled extremely efficient collection, storage and centralization of massive volumes of data. For instance, it has reduced the amount of storage required from 60 terabytes per year to 10 terabytes per year—resulting in significant cost savings given the cost per terabyte to manage stored data.

Prior to implementing Proficy Historian, the Center could store only three months of data online on its legacy applications, which were based on multiple relational databases, limiting the ability to optimize its data. Fulfilling data requests would take days or weeks because it would have to pull data from the archives, manually load offline data, and then run queries against it—a time-consuming and often challenging task.

Now with Historian, it can store ten years of data online, enabling it to efficiently run queries against much larger sets without having to manually move data in and out of systems for near real-time analysis. It can quickly get answers to critical issues that impact operational performance such as degradation of equipment since installation. As a result of faster identification of issues, it can make timelier decisions and drive quicker corrective actions.

Faster analytics and predictive diagnostics

The Center today runs approximately a hundred different algorithms continuously across its data, whereby analytics can run much faster against historical data to bring meaning and context to its real-time operations systems, providing a key competitive advantage. It can also predict failure and downtime of assets weeks in advance by comparing its historical data against current performance—looking at trends and patterns for signs of deterioration—to detect, diagnose and predict issues before they occur.

For instance, the Center avoided multiple failures caused by servo and actuator problems on its valves, saving millions for its customers in downtime by leveraging historian and advanced analytics, which delivered data contextualization and actionable intelligence. Overall, through the use of big data across its fleet, it estimates a cost savings and cost avoidance of $75 million per year, a 2X increase in its ability to deliver value to customers since Historian was installed. It's the power of big data at work!

With Proficy Historian, GE's M&D Center manages a constant flow of data across its fleet—leveraging big data for better, faster decisions to optimize operational and financial performance.

Conclusion

Business and IT leaders need to ask themselves whether their industrial enterprise is maximizing the full potential value of their process data and using that insight to drive real-time improvements. As data volumes continue to expand, information-driven strategies will only become more pervasive as a source of competitiveness—making the use of big data in the industrial space ever more imperative.

A closer look at advanced historians demonstrates how such technologies can help enterprises leverage their time-series process data by providing the ability to efficiently run real-time analytics within massive sets of historical data. These solutions have the potential to revolutionize the way enterprises do business by providing critical insights for timelier operational decisions while also enabling continuous improvements across the enterprise.

Going forward, as information increasingly empowers enterprises to understand their businesses better and to foresee what is possible, those that capitalize on the value of big data will gain insights to improve performance beyond their competitors. They will be positioned to better innovate, compete, and drive value—all of which will significantly accelerate business growth and continuously drive optimized performance for long-term success.