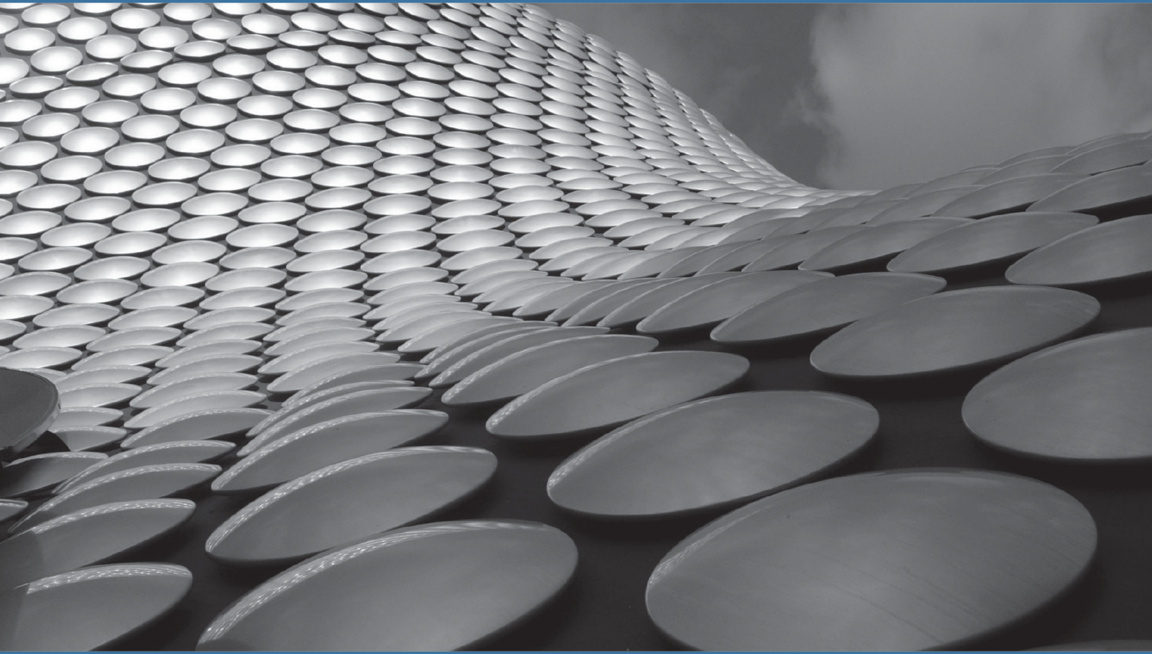
O'REILLY®

Scaling Data Science for the Industrial IoT

Advanced Analytics in Real Time



Andy Oram

Welcome to the Intelligence of Things

LEARN. PREDICT. OPTIMIZE THE POWER OF IoT.

Each “thing” in the Internet of Things can generate up to millions of data points every day. Tackling the volume, velocity, and variety of data from IoT is key to delivering even more powerful solutions. Thingworx ushers in the new age of Intelligence.

Thingworx Analytics unlocks business intelligence you can use to make an impact, and enables enterprises to find the true value in their IoT data – to learn from past data, understand and predict the future, and make decisions that will enhance outcomes. It provides an easy way to use advanced analytics methods without requiring expert training in data science, complex mathematics, or machine learning. With ThingWorx Analytics you can:

- Watch: Monitor edge devices and provide real-time pattern and anomaly detection on real-time data streams.
- Predict: Provide automated predictive modeling and operationalization for a variety of different outcomes and pattern and anomaly detection on real-time data streams.
- Adapt: Deliver prescriptive and simulative intelligence that identifies factors that contribute to an outcome and explains how to change a predicted outcome.
- Optimize: Automatically operationalize and maintain predictive and simulative intelligence to deliver to end-users.

Learn more at www.thingworx.com/analyze/

Scaling Data Science for the Industrial Internet of Things

Advanced Analytics in Real Time

Andy Oram

Beijing · Boston · Farnham · Sebastopol · Tokyo O'REILLY®

Scaling Data Science for the Industrial Internet of Things
by Andy Oram

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Brian Jepson
Production Editor: Kristen Brown
Proofreader: Kristen Brown

Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

December 2016: First Edition

Revision History for the First Edition
2016-12-16: First Release

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-98017-0

[LSI]

Table of Contents

Scaling Data Science for the Industrial Internet of Things

Few aspects of computing are as much in demand as data science. It underlies cybersecurity and spam prevention, determines how we are treated as consumers by everyone from news sites to financial institutions, and is now part of everyday reality through the Internet of Things (IoT). The IoT places higher demands on data science because of the new heights to which it takes the familiar "V's" of big data (volume, velocity, and variety). A single device may stream multiple messages per second, and this data must either be processed locally by sophisticated processors at the site of the device or be transmitted over a network to a hub, where the data joins similar data that originates at dozens, hundreds, or many thousands of other devices. Conventional techniques for extracting and testing algorithms must get smarter to keep pace with the phenomena they're tracking.

A report by ABI Research on ThingWorx Analytics predicts that "by 2020, businesses will spend nearly 26% of the entire IoT solution cost on technologies and services that store, integrate, visualize and analyze IoT data, nearly twice of what is spent today" (p. 2). Currently, a lot of potentially useful data is lost. Newer devices can capture this "dark data" and expose it to analytics.

This report discusses some of the techniques used at ThingWorx and two of its partners—Glassbeam and National Instruments—to automate and speed up analytics on IoT projects. These activities are designed for high-volume IoT environments that often have real-time requirements, and may cut the time to decision-making by orders of magnitude.

Tasks in IoT Monitoring and Prediction

To understand the demands of IoT analytics, consider some examples:

Farming

A farm may cover a dozen fields, each with several hundred rows of various crops. In each row, sensors are scattered every few feet to report back several measures, including moisture, temperature, and chemical composition of the soil. This data, generated once per hour, must be evaluated by the farmer's staff to find what combination works best for each crop in each location, and to control the conditions in the field. Random events in the field can produce incorrect readings that must be recognized and discarded. Data may be combined with observations made by farmers or from the air by drones, airplanes, or satellites.

Factory automation

Each building in a factory campus contains several assembly lines, each employing dozens of machines manipulated by both people and robots. A machine may have 20 sensors reporting its health several times a second in terms of temperature, stress, vibration, and other measurements. The maintenance staff want to determine what combination of measurements over time can indicate upcoming failures and need for maintenance. The machines come from different vendors and are set up differently on each assembly line.

Vehicle maintenance

A motorcycle manufacturer includes several sensors on each vehicle sold. With permission from customers, it collects data on a daily basis from these sensors. The conditions under which the motorcycles are operated vary widely, from frigid Alaska winters to sweltering Costa Rican summers. The manufacturer crunches the data to determine when maintenance will be needed and to suggest improvements to designers so that the next generation of vehicles will perform better.

Health care

A hospital contains thousands of medical devices to deliver drugs, monitor patients, and carry out other health care tasks. These devices are constantly moved from floor to floor and attached to different patients with different medical needs. Changes in patient conditions or in the functioning of the devices must be evaluated quickly and generate alerts when they indicate danger (but should avoid generating unnecessary alarms that distract nursing staff). Data from the devices is compared with data in patient records to determine what is appropriate for that patient.

In each of these cases, sites benefit by combining data from many sources, which requires network bandwidth, storage, and processing power. The meaning of the data varies widely with the location and use of the plants, vehicles, or devices being monitored. A host of different measurements are being collected, some of which will be found to be relevant to the goals of the site and some of which have no effect.

The Magnitude of Sensor Output

ThingWorx estimates that devices and their output will triple between 2016 and 2020, reaching 50 billion devices that collectively create 40 zetabytes of data. A Gartner report (published by Datawatch, and available for download by filling out a form), says:

- A single turbine compressor blade can generate 500GB of data per day.
- A typical wind farm may generate 150,000 data points per second.
- A smart meter project can generate 500 million readings of data per day.
- Weather analysis can involve petabytes (quintillions of bytes) of data.

What You Can Find in the Data

The concerns of analysts and end users tend to fall into two categories, but ultimately are guided by the goal to keep a system or process working properly. First, they want to catch *anomalies*: inputs that lie outside normal bounds. Second, in order to avoid the crises implied by anomalies, they look for *trends*: movements of specific variables (also known as *features* or *dimensions*) or combinations of variables over time that can be used to predict important outcomes.

Trends are also important for all types of planning: what new products to bring to market, how to react to changes in the environment, how to redesign equipment so as to eliminate points of failure, what new staff to hire, and so on.

Feature engineering is another element of analytics: new features can be added by combining features from the field, while other features can be removed. Features are also weighted for importance.

One of the first judgments that an IoT developer has to make is where to process data. A central server in the cloud has the luxury of maintaining enormous databases of historical data, plus a potentially unlimited amount of computing power. But sometimes you want a local computer on-site to do the processing, at least as a fallback solution to the cloud, for three reasons. First, if something urgent is happening (such as a rapidly overheating motor), it may be important to take action within seconds, so the data should be processed locally. Second, transmitting all the data to a central server may overload the network and cause data to be dropped. Third, a network can go down, so if people or equipment are at risk, you must do the processing right on the scene.

Therefore, a kind of triage takes place on sensor data. Part of it will be considered unnecessary. It can be filtered out or aggregated: for instance, the local device may communicate only anomalies that suggest failure, or just the average flow rate instead of all the minor variations in flow. Another part of the data will be processed locally. Perhaps it will also be sent into the cloud, along with other data that the analyst wants to process for predictive analytics.

Local processing can be fairly sophisticated. A set of rules developed through historical analysis can be downloaded to a local computer to determine the decisions it makes. However, this is static analysis. A central server collecting data from multiple devices is required for dynamic analysis, which encompasses the most promising techniques in modern data science.

Naturally, the goal of all this investment and effort is to take action: fix the broken pump, redesign a weak joint in a lever, and so on. Some of this can be automated, such as when a sensor indicates a problem that requires a piece of machinery to shut down. A shutdown can also trigger the start of an alternative piece of equipment. Some operations are engineered to be self-adjusting, and predictive analytics can foster that independence.

Characteristics of Predictive Analytics

In rising to the challenge of analyzing IoT's real-time streaming data, the companies mentioned in this report have had to take into account the challenges inherent in modern analytics.

A Data Explosion

As mentioned before, sensors can quickly generate gigabits of data. These may be reported and stored as thousands of isolated features that intersect and potentially affect each other. Furthermore, the famous V's of big data apply to the Internet of Things: not only is the *volume* large, but the *velocity* is high, and there's a great deal of *variety*. Some of the data is structured, whereas some may be in the form of log files containing text that explains what has been tracked. There will be data you want to act on right away and data you want to store for post mortem analysis or predictions.

You Don't Know in Advance What Factors are Relevant

In traditional business intelligence (BI), a user and programmer would meet to decide what the user wants to know. Questions would be quite specific, along the lines of, "Show me how many new customers we have in each state" or "Show me the increases and declines in the sales of each product." But in modern analytics, you may be looking for unexpected clusters of behavior, or previously unknown correlations between two of the many variables you're tracking—that's why this kind of analytics is popularly known as *data mining*. You may be surprised which input can help you predict that failing pump.

Change is the Only Constant

The promise of modern analytics is to guide you in making fast turns. Businesses that adapt quickly will survive. This means rapidly recognizing when a new piece of equipment has an unanticipated mode of failure, or when a robust piece of equipment suddenly shows problems because it has been deployed to a new environment (different temperature, humidity, etc.).

Furthermore, even though predictive models take a long time to develop, you can't put them out in the field and rest on your laurels.

New data can refine the models, and sometimes require you to throw out the model and start over.

Tools for IoT Analytics

The following sections show the solutions provided by some companies at various levels of data analytics. These levels include:

- Checking thresholds (e.g., is the temperature too high?) and issuing alerts or taking action right on the scene
- Structuring and filtering data for input into analytics
- Choosing the analytics to run on large, possibly streaming data sets
- Building predictive models that can drive actions such as maintenance

Local Analytics at National Instruments

National Instruments (NI), a test and measurement company with a 40-year history, enables analytics on its devices with a development platform for sensor measurement, feature extraction, and communication. It recognizes that some calculations should be done on location instead of in the cloud. This is important to decrease the risk of missing transient phenomena and to reduce the requirement of pumping large data sets over what can get to be quite expensive IT and telecom infrastructure.

Measurement hardware from NI is programmed using LabVIEW, the NI software development environment. According to Ian Fountain, Director of Marketing, and Brett Burger, Principal Marketing Manager, LabVIEW allows scientists and engineers without computer programming experience to configure the feature extraction and analytics. The process typically starts with sensor measurements based on the type of asset: for example, a temperature or vibration sensor. Nowadays, each type of sensor adheres to a well-documented standard. Occasionally, two standards may be available. But it's easy for an engineer to determine what type of device is being connected and tell LabVIEW. If an asset requires more than one measurement (e.g., temperature as well as vibration), each measurement is connected to the measurement hardware on its own channel to be separately configured.

LabVIEW is a graphical development environment and provides a wide range of analytical options through function blocks that the user can drag and drop into the program. In this way, the user can program the device to say, “Alert me if vibration exceeds a particular threshold.” Or in response to a trend, it can say, “Alert me if the past 30,000 vibration readings reveal a condition associated with decreasing efficiency or upcoming failure.”

NI can also transmit sensor data into the cloud for use with an analytical tool such as ThingWorx Analytics. Because sensors are often high bandwidth, producing more data than the network can handle, NI can also do feature extraction in real time. For instance, if a sensor moves through cycles of values, NI can transfer the frequency instead of sending over all the raw data. Together with ThingWorx, NI is exploring anomaly detection as a future option. This would apply historical data or analytics to the feature.

Extracting Value from Machine Log Data With Glassbeam

Glassbeam brings a critical component of data from the field—log files—into a form where it can be combined with other data for advanced analytics. According to the Gartner report cited earlier, log files are among the most frequently analyzed data (exceeded only by transaction data), and are analyzed about twice as often as sensor data or machine data.

Glassbeam leverages unique technology in the data translation and transformation of any log file format to drive a differentiated “analytics-as-a-service” offering. It automates the cumbersome multi-step process required to convert raw machine log data into a format useful for analytics. Chris Kuntz, VP of Marketing at Glassbeam, told me that business analysts and data scientists can spend 70-80 percent of their time working over those logs, and that Glassbeam takes only one-twentieth to one-thirtieth of the time.

Glassbeam’s offering includes a visual data modeling tool that performs parsing and extract, transform, load (ETL) operations on complex machine data, and a highly scalable big data engine that allows companies to organize and take action on transformed machine log data. Binary and text streams can also be handled. As its vertical industry focus, Glassbeam’s major markets include storage networking, wireless infrastructure, medical devices, and clean energy.

Log files are extremely rich in content and carry lot of deep diagnostics information about machine health and usage. However, sifting through varied log formats and parsing to uncover the hidden nuggets in this data is a headache every administrator dreads. Does this field start at the same character position in every file? Does it occupy a fixed number of positions or end with a marker? Are there optional fields that pop up in certain types of records?

Figuring all this out is the kind of task that's ripe for automation. Instead of coding cumbersome logic in traditional approaches like regular expressions, Glassbeam's Semiotic Parsing Language (SPL) can define and run analytics that compare fields in different records and perform other analytics to figure out the structure of a file.

Note that the analytics can also be distributed: some parts can run right at the edge near the device, feeding results back to a central server, and other parts can run on the cloud server with access to a database of historical information.

Glassbeam can also perform filtering—which can greatly reduce the amount of data that has to be passed over the network—and some simple analytics through technologies such as correlations, finite state machines, and Apache Spark's MLlib. For instance, the resulting analytics may be able to tell the administrator whether a particular machine was used only 50% of the time, which suggests that it's underutilized and wasting the client's money.

ThingWorx and Glassbeam exchange data in several ways, described in this white paper. ThingWorx can send messages one way to an ActiveMQ broker run by Glassbeam. ThingWorx can also upload data securely through FTP/SSH. Glassbeam also works directly with ThingWorx Analytics by plugging directly into the ThingWorx Thing Model, making it easier and faster to build advanced analytics, predictions, and recommendations within ThingWorx mashups and ThingWorx-powered solutions.

In addition to providing a powerful data parsing and transformation engine, Glassbeam has a suite of tools that allow users to search, explore, apply rules and alerts, and generate visualizations on data from machine logs, as well as take action on analytics results from

ThingWorx. These powerful user and administrative tools complement and feed into the analytics offered by ThingWorx Analytics.

Analytics in ThingWorx

As resources grow, combining larger data sets and more computer processing power, you can get more insights from analytics. Unlike the three companies profiled earlier in this report, ThingWorx deals with the operations on devices directly. It's a cloud-based solution that accepts data from sensors—it currently recognizes 1,400 types—or from intermediate processors such as the companies seen earlier. Given all this data, ThingWorx Analytics (a technology that PTC brought in by purchasing the company Coldlight) then runs a vast toolset of algorithms and techniques to create more than just predictive models: causal analysis, relationship discovery, deep pattern recognition, and simulation.

According to Joseph Pizonka, VP of Product Strategy for ThingWorx Analytics, all the user needs to do is tell the system what the objective is—for instance, that a zero flow rate indicates a failure—and what data is available to the system. The analytics do the rest.

For instance, analytics can generate a *slgnal*, the characteristics of a device that show that it is high-performing or low-performing. Various features you want to track—the model and age of the machine, for instance—form a *profile*. Pizonka describes profile generation as "finding a needle in a haystack."

Internally, ThingWorx Analytics performs the standard machine learning process of injecting training data into model building and validating the results with test data, and later in production with data from the field. Model building is extremely broad, using a plethora of modern techniques like neural networks to search for the most accurate prediction model.

ThingWorx Analytics also runs continuously and learns from new input. It compares its prediction models to incoming data and adjusts the models as indicated. Pizonka says that the automated system can generate in minutes to hours what a human team would take months to accomplish. It's good for everyday problems, letting data scientists focus on more meaty tasks.

Prerequisites for Analysis

Some common themes appear in the stories told by the various companies in this article.

- You need to plan your input. If you measure the wrong things, you will not find the trends you want.
- You may need to train your model, even if it evolves automatically upon accepting data.
- You need a complete view of your machine data, and you need that data in the right format. If it's not in the right format, you'll need to parse it into a structure suitable for analysis.
- The value of analytics increases dramatically as data sizes increase—but they may need to increase logarithmically.

The variety of analytical techniques available to modern data crunchers is unprecedented and overwhelming. The examples in this report show that it is possible to marshal them and put them to use in your IoT environment to find important patterns, make predictions, and guide humans and machines toward better outcomes.

About the Author

Andy Oram is an editor at O'Reilly Media. An employee of the company since 1992, Andy currently specializes in programming topics. His work for O'Reilly includes the first books ever published commercially in the United States on Linux, and the 2001 title *Peer-to-Peer*.