Harnessing Machine Data with Data-Driven Machine Learning

(White Paper)

Executive Summary

Petabytes of machine data are generated by billions of machines every single day. Traditional data processing techniques and infrastructure can no longer effectively handle such gigantic and complex data sets. The limelight is now on Big Data technologies to take the lead in this space. One major innovation here is Data-Driven Machine Learning, which provides the intelligence to translate data into insights and value. Popular use cases include text analytics, pattern recognition and predictive analytics. Also equally important, several key technologies have been developed for the infrastructure of Big Data systems, which helps to augment the computing capabilities required for such heavyweight processing. The result is a real-time distributed computation framework with high fault tolerance, also known as the lambda architecture.

Background

Machine data (or machine-generated data) is information produced by the activity of machines, which commonly include servers, embedded systems and other networked devices. Process logs and sensor data are some examples of machine data. Machine data affects a wide range of domains, including energy, transportation, IT infrastructure, industrial applications and data centers.

In the last decade, the rise of the *Internet-of-Things* and other telematic technologies have drastically increased the size, complexity and number of computer networks across the world. These networks generate huge amounts of machine data, which is expected to increase 15 times between 2012 and 2020. (EMC Corporation, n.d.) The explosive growth of machine data is one major factor contributing to the Big Data phenomenon we often hear of today.

Big Data is generally understood as the use of huge data sets for information mining, which was previously unseen of because traditional data processing technologies were unable to handle such complexity. Big Data seems to be the gateway to transformational business value, with huge success stories from the likes of Amazon and Netflix. And it is not restricted to digital businesses - in 2013, a survey done across a dozen global industries revealed that companies that invested heavily in Big Data projects have received huge returns and built competitive advantages, thereby putting other competitors at risk of obsolescence. (Ramaswamy, 2013)

Big Data Brings Big Challenges

However, harnessing Big Data is easier said than done. In another survey conducted, although 75% of the respondents agree that better Big Data management would lead to smarter business decisions, a significant 37.5% also regard Big Data as their biggest organizational challenge. (Bertolucci, 2012) Most organizations, especially those without skilled talent, have difficulties building the intelligence and infrastructure to manage Big Data, which are explained further below.

Challenge 1: Building Big Data Intelligence

The first reaction many organizations have towards Big Data is to collect as much data as they can. This *collect-first-and-think-later* mindset often leads to a growing pile of buried data that not only provides poor Return On Investment (ROI), but also makes it increasingly difficult to manage. Collecting data is but a small part of the equation, and people soon realize that "data does not speak, it only responds". (Solanki, 2014) The real challenge (and opportunity) lies in using the right techniques to draw insights to the right questions. This is the Big Data intelligence that could make or break any Big Data project.

Challenge 2: Building a Big Data Infrastructure

Big Data is commonly characterized by four V's - Volume, Velocity, Variety and Veracity. Each quality has its share of technical considerations as discussed in the table below. Considerable amount of technical expertise and resources are thus required to design and build the infrastructure for an effective Big Data management system.

Fortunately, much research and work have been done in this space to resolve these challenges. The next few sections discuss some prominent infrastructure technologies that can help organizations better manage their massive collection of machine data.

Analyzing the Four V's of Big Data

Quality	Description	Technical Considerations (for a Big Data Management System)
Volume	Quantity of data	• Scaling to accommodate large and variable amounts of data • Fast and parallel data processing • Analyzing data in high-dimensional spaces
Velocity	Flow of streaming data	• Real-time and batch-level capabilities to manage different flow of data • Handling shocks in data streams
Variety	Forms and sources of data	• Handling data of different forms, structured or unstructured • Interoperability with various data sources
Veracity	Quality of data	• Minimizing data loss • Cleaning data to minimize noise and biases • Validating accuracy of data and analytics

Drawing Insights from Machine Data with Data-Driven Machine Learning (DDML)

Targeting Challenge 1 of building Big Data intelligence, some platforms have employed the concept of Data-Driven Machine Learning (DDML). DDML is a method of data analysis that automatically builds its analytical models using data. For Big Data applications, DDML has proven to be far more effective than traditional data analytics techniques.

Traditional techniques rely mainly on a trial-and-error approach, which does not scale well to handle huge and heterogeneous data sets. The potential correlations within the data are too complex for any analyst to derive and test with reasonable time and effort. (Galetto, 26)

Conversely, machine learning is ideal for this scenario. Its ability to conduct comprehensive analyses with minimal human intervention enables it to uncover value buried in the masses of data. In fact, it thrives with more data since the analytical models automatically learn from data and adapt to produce better insights. (Skytree, n.d.)

This allows DDML techniques to *think* and process beyond human scale. These algorithms need to also be scalable with parallel processing abilities to support the heavy lifting.

The next section describes several popular DDML techniques that have been used to analyze machine data for different tasks. Some example case studies from Neuro10 are also discussed below to aid in understanding.

(A) Text analytics

Text analytics aims to derive latent information from both structured and unstructured data found in log messages generated by machines. It often integrates different machine learning and pattern recognition techniques such as log linear model, hidden Markov model and n-gram. These techniques parse and index the data, before further processing is done for performance analysis, data visualization, predictive analytics and so on. For the analysis of unstructured text data, it usually requires more complex parsing to derive linguistic and knowledge patterns that help to understand the message conveyed.

Case Study: Log Simplification (Text Analytics)

Machines running for just an hour could generate up to billions of log records. Instead of having network administrators read through all of them for analysis and reporting purposes, Neuro10's log simplification feature automatically summarizes these records into reports with mere few paragraphs that are human-readable. This feature uses a statistical modeling technique called Topic Modeling to cluster sentences according to abstract "topics" discovered in document collections, without the need for human intervention.

(B) Network analytics

Network analytics aims to analyze network operations for performance indication and anomaly detection. Such real-time capabilities enable predictive analytics in IT infrastructure management services, which accelerates the identification and resolution of issues before they cause any serious impact on the network's operations. For instance, the machine learning technique, Eigenvector centrality analysis, can help analyze the criticality of network nodes. Other DDML tasks in this category include network signature generator, community pattern discovery, time-series based differential anomaly analysis, email flow management and network visualization.

(C) Pattern discovery and recognition

Pattern discovery and recognition aims to mine patterns of diverse forms from massive machine data, including negative patterns, multi-level patterns, multi-dimensional patterns, sequential patterns and sub-graph patterns. These patterns help to extract meaningful information from semi-structured data that can be used for identity management, user authentication and IT infrastructure analysis, such as identifying meaningful fields to be reported on the fly. DDML techniques for such tasks could use well-established scalable pattern discovery methods such as éclat, deep learning, Bayesian network analysis, Markov random field and association rule-mining.

Case Study: Failure Analysis (Network Analytics)

In managing large and complex networks, it is often a huge challenge to identify and resolve failures. Neuro10's failure analysis feature is able to isolate fault regions within the network and provide valuable information pointing to the potential root cause of the issue, which improves mean-time-to-repair (MTTR). This is made possible with the Network Community Discovery technique, which dynamically groups network nodes based on the behavior of machine-machine interactions.

Case Study: Behavior Profiling (Pattern Recognition)

One of the common cyber security threats faced by organizations is intrusion attack, i.e. when intruders gain trusted access into a secured environment. Such attacks can be very costly to combat and recover from as they involve both technical and human factors. Neuro10's behavior profiling feature identifies potential intrusion threats by analyzing user behavior patterns. Such a high-level approach can effectively complement existing authentication mechanisms to improve security.

(D) Time series analysis

Time series analysis aims to filter and analyze events that occur over time, so as to explore correlation and causation between events. For instance, temporal-spatial mining can be used to monitor and trigger alerts based on time and location signals. The models developed can also forecast failure trends and facilitate pre-emptive intervention. DDML techniques that could help to build the model's backbone processing engine include hidden Markov model and various other computational intelligence techniques such as Artificial Neural Networks (ANN), genetic programming and Support Vector Machine (SVM).

Case Study: Anomaly Detection (Time Series Analysis)

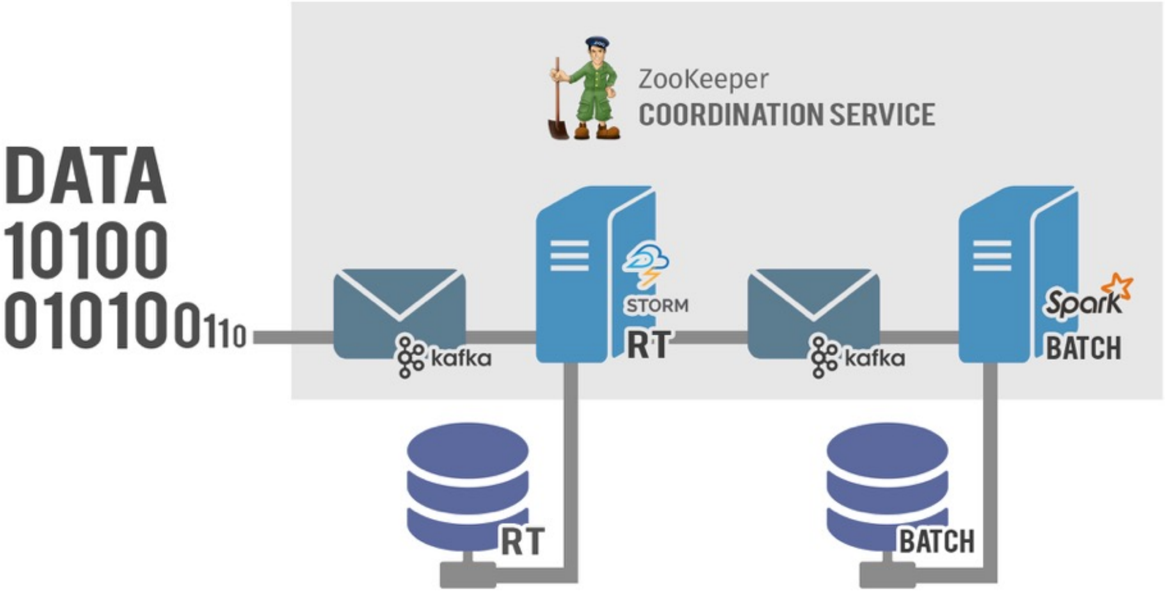
An important task in system management is detecting anomalies, i.e. deviations in system behavior. Traditionally, anomalies are identified based on thresholds that are manually identified and entered. But with the Big Data phenomenon, data is getting huge, complex and very dynamic. It is no longer within human ability to effectively monitor the data at such a pace of change. Neuro10's anomaly detection feature can continuously ingest time-series data (e.g. t*emperature* at each *unit time),* and self-adapt its statistical model with new thresholds as the data changes over time.

Designing a Distributed Infrastructure for DDML Workflows

Besides building Big Data intelligence, organizations need to also plan for an infrastructure (i.e. technology pipeline) to support these heavyweight DDML workflows. A Big Data infrastructure like this needs to account for the technical considerations in managing the four V's of Big Data as discussed in Challenge 2 above. The solution is a real-time distributed computation framework with high fault tolerance, also known as the lambda architecture. This section discusses one such open source implementation that has been gaining popularity in recent years.

Architecture Overview

The distributed computation framework illustrated below uses open source technologies that have been widely adopted to support enterprise Big Data and machine learning workflows. Individually, each technology has good performance and high usability. When integrated, they also complement each other very well to enable advanced analytics on fast-moving and voluminous data.

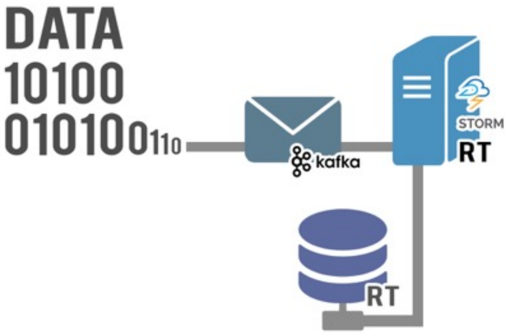


Key Technologies for an Open Source DDML Infrastructure

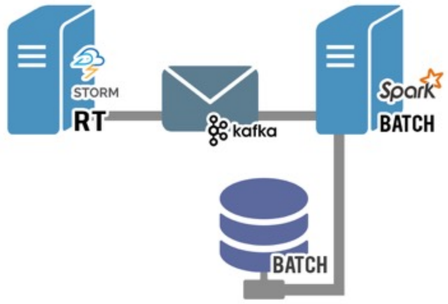
Name	Description	Features
Apache Kafka	Distributed real-time publisher-subscriber messaging system for general purpose	High throughput, reliable delivery, horizontal scalability, message durability
Apache Storm	Distributed real-time computation system for processing large volumes of high-velocity data	Fast computation, parallel and scalable computing, fault tolerant, reliable processing
Apache Spark	In-memory data processing engine on the Apache Hadoop platform	Supports batch, real-time and advanced Big Data analytics with fast and efficient iterations
Apache ZooKeeper	Distributed configuration and synchronization service	High availability, fast and scalable cluster coordination

System Interactions

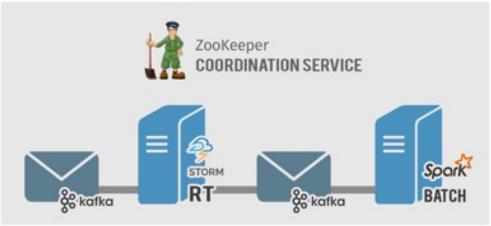
The full lambda architecture can be broken down into three main subsystems - real-time stream processing, batch processing and coordination service. The components and behavior of each subsystem is elaborated below, with illustrations.



(1) *Real-time stream processing* - When data streams enter the system, a messaging service like Kafka forwards the events-of-interest as tuple messages to Storm clusters for stream processing tasks at real-time. The results are then persisted in databases optimized for such real-time operations, such as HBase, Cassandra and Solr.



(2) *Batch processing* - Relevant data from the real-time processing tasks is also sent to Kafka for forwarding to Spark and/or Hadoop clusters for batch-level analytics. The results are then persisted in databases optimized for such batch operations, such as the Hadoop Distributed File System (HDFS).



(3) *Coordination service* - While the system is running, ZooKeeper maintains and coordinates the state of data, processes and machines across the various clusters. This helps to ensure the distributed framework behaves correctly and is fail tolerant.

Conclusion

The magnitude of data collected is growing and will continue to grow even more as we embrace technology adoption (and hence data generation) across the world. Many organizations have faced challenges managing this growth, primarily due to the lack of agility, technical expertise and resources. This has driven decision makers to turn to solution providers for quick and effective remedy, so as to ensure the organization's survival and competitiveness. Fortunately, the results have been positive. There exist notable cases of companies flourishing under a data-driven strategy, such as the likes of Amazon and Netflix. Such a speed of innovation would not have been possible without the rise of Machine Learning in the Big Data space. In this new revolution, Machine Learning will provide unprecedented opportunities across all data-generating fields, including those churning out petabytes of machine data.

References

TechTarget. (n.d.). Machine Data. Retrieved February 2016, from IoT Agenda: http://internetofthingsagenda.techtarget.com/definition/machine-data

EMC Corporation. (n.d.). New Digital Universe Study Reveals Big Data Gap: Less Than 1% of World's Data is Analyzed; Less Than 20% is Protected. Retrieved February 2016, from EMC Corporation: http://www.informationleap.com/about/news/press/2012/20121211-01.htm

Ramaswamy, S. (2013, June 25). What the Companies Winning at Big Data Do Differently. Retrieved February 2016, from Harvard Business Review: https://hbr.org/2013/06/what-the-companies-winning-at

Bertolucci, J. (2012, August 16). Big Data Development Challenges: Talent, Cost, Time . Retrieved February 2016, from InformationWeek: http://www.informationweek.com/big-data/big-data-analytics/big-data-development-challenges-talent-cost-time/d/d-id/1105829

Solanki, A. (2014, November 2). "Data doesn't speak, it only responds". Retrieved April 2016, from LinkedIn: https://www.linkedin.com/pulse/20141102011806-108356215--data-doesn-t-speak-it-only-responds

Galetto, M. (26, February 2016). Machine Learning and Big Data Analytics: The Perfect Marriage. Retrieved March 2016, from NGDATA: http://www.ngdata.com/machine-learning-and-big-data-analytics-the-perfect-marriage/

Skytree. (n.d.). Why do Machine Learning on Big Data? Retrieved February 2016, from Skytree: http://www.skytree.net/machine-learning/why-do-machine-learning-big-data/

Published on: April 2016