Available online at www.elixirpublishers.com (Elixir International Journal)

Information Technology

Elixir Inform. Tech. 82 (2015) 32260-32262

ISSN: 2229-712X

Aggrandizement of Big Data Analytics beyond the Batch to Real time

Abhishek Khare
Shri Vaishnav Institute of Management, Indore India.

ARTICLE INFO

Article history:
Received: 9 December 2014;
Received in revised form:
25 April 2015;
Accepted: 1 May 2015;

Keywords
Hadoop,
MapReduce,
Batch processing,
Big Data,
Real-time processing.

ABSTRACT
Data worldwide is growing 40 percent per year, a rate of growth that is daunting for any organization. Big Data Analytics is becoming an increasingly important part of every aspect, batch processing system like Hadoop had evolved for excellent offline data processing platform for Big Data. There are many use cases across various domains which require real-time / near real-time response on Big Data for faster decision making. This paper describes beyond batch processing system towards Real-Time and Streaming.

Introduction

Constant data emission by sensors, machines, vehicles, mobile phones, social media networks, and other real-time sources are compelling organizations to rethink their data and analytics strategy beyond batch-processing. The big data revolution is well under way. True real-time analytics are critical for many analytic applications. For example, true real time is required for the continuous monitoring of a process, network, or facility. Consider that a few minutes on a rail or truck schedule can affect customer satisfaction, and they can add up, amounting to major delays. Additional time to access content on a mobile device can lead to serious customer satisfaction issues and unwanted churn. Organizations can act on the information by immediately sending an alert to the appropriate manager, updating a management dashboard, offering an incentive to a churning customer, adjusting machinery, or preventing fraud. It is impractical to build planning application with reasonable response time because they have challenges like retrieving large size of data from the database will take excessive I/O and network overhead. Distribution and exception detection operations require large data update in a database resulting in severe performance problems. It is common to spend a lot of time and effort tuning relational databases for big-data analytics-type applications, which often need to scan through a large portion of data from several tables. Relational databases are not reliable in terms of consistent performance. Performance could be further degraded due to compound SQL statements or large multi-way joins. The recent development of NoSQL is to address such performance issues.

The MapReduce framework has made complex large-scale data processing easy and efficient. Despite this, MapReduce is designed for batch processing of large volumes of data, and it is not suitable for recent demands like real-time and online processing. MapReduce is inherently designed for high throughput batch processing of big data that take several hours delivered to batch system to be processed. Some MapReduce implementations especially real-time ones like Spark and Grid Gain support this technique. However, this technique is not adequate for demands of a true stream system. Furthermore, the MapReduce model is not suitable for stream processing. [2]

Currently, there are two notable stream processing frameworks that are inherently designed for big data streams are Storm from Twitter, and S4 from Yahoo [6]. Both frameworks run on the Java Virtual Machine and both process keyed streams.

Real time processing

Real time analytics is a process of delivering information about events as they occur. Standard relational databases and SQL simply can't store or process big data, and are reaching fundamental capacity and scaling limits.[5] Not only are the data formats outside the scope of relational databases, but much of the processing requires iterative logic, complex branching, and special analytic algorithms. SQL is a declarative language with a powerful but fixed syntax. Big data generally needs procedural languages and the ability to program arbitrary new logic [7]. Real time processing requires capture and process data in seconds or milliseconds from multiple sources, including streaming data, event streams, and message queues that provide complete views of business entities and situations based on both real-time and latent data, presented in terms that are business friendly and actionable. True real-time analytics are critical for many analytic applications it requires continuous monitoring of a process, network, or facility.

Challenges and Obstacles with Big Data

Before Top big data challenges, we found that data infrastructure and the ability of their data centre to provide the scalability, low latency, and performance needed to handle big data projects, data governance and policy challenges for defining the data that will be stored, analyzed, and accessed, along with determining its relevance. The results are clear with growing data volumes, it is essential that real-time information that is of use to the business can be extracted from its IT systems, otherwise the business risks being swamped by a data deluge. Meanwhile, competitors that use data to deliver better insights to

Tele:
E-mail addresses: abhishekkhare@rocketmail.com

decision-makers stand a better chance of thriving through the difficult economy and beyond. [3]

As data volume increases, the ability to collect and present data in a way that the business can understand, so it can make decisions faster than the competition, will be the key to keeping the business competitive

Analytics Delivery: Moving from Batch to Real-Time

IT industries are using both batch and real-time delivery of analytics equally, with a 0-50 split. However, that's changing—companies are expecting to do almost two-thirds of their analytics in real time by 2016. [4] As organizations aim to become more efficient, they need to work on getting more light on the meaning of data to increase market share, but it is no good if they can't take information from reviews and on Twitter and translate that into revenue. The best way of succeeding is to be able to look at all the information available in real time in one place. Our highly engaged IT industries are already conducting significantly more of their data analytics in real time and plan to do even more.

Advanced technologies to enable real-time big data analytics

- **Operational Intelligence**: The sources are an eclectic mix of new ones like streaming data, machine data, social media data, NoSQL data platforms such as Hadoop and traditional enterprise sources like enterprise applications, relational databases. Unlike solutions exclusively for structured data, Operational Intelligence also deals with datasets that are schema free, metadata free and evolving structurally.
- **Streaming data:** As more organizations move deeper into monitoring operations in real time, there's a growing need to quickly capture and process events expressed as messages or events in a stream that generates and delivers data almost continuously. At the same time, the number of data streams is increasing because many new forms of big data are communicated via streams especially sensor data and machine data. In many ways, analytic correlation across multiple data streams is the epitome of operational intelligence.
- **Event processing**: Technologies for event processing have been around for years, but most are designed to monitor only one stream of events at a time. Even if users monitor multiple streams, they end up with multiple silos views into real-time business operations. Operational intelligence creates a more unified view that correlates events from multiple streams and other information sources, arming businesses with better insights.
- **Actionable analytic outcomes**: Operational Intelligence's combination of leading-edge analytic technologies helps users and their organizations act immediately on the results of real-time analyses. [5]
- **Predictive analytics:** Predictive analytics enables organizations to move to a future-oriented view of what's ahead and offers organizations some of the most exciting opportunities for driving value from big data. Real-time data provides the prospect for fast, accurate, and flexible predictive analytics that quickly adapt to changing business conditions. The faster you analyze your data, the more timely the results, and the greater its predictive value.[18]

Prolegomenon of Cloud Computing

The way the economy is going, the world is becoming more finance focused, whether that is gaining competitive advantage or increasing revenue for many organizations, finding a practical and cost-effective use for big data is challenge but to use cloud services which offer flexibility. Big data is a very scalable model. Now organizations start off the big data journey without having large capital investment because there are interfaces for organizations to build big data clusters on the cloud.

No longer scattered in multiple federated databases throughout the enterprise, Big Data consolidates information in a single massive database stored in distributed clusters and can be easily deployed in the cloud to save costs and ease management. Companies may also move Big Data to the cloud for disaster recovery, replication, load balancing, storage, and other purposes. A popular example of the benefits of cloud supporting big data can be noted at both Google and Amazon.com. Both companies depend on the capability to manage massive amounts of data to move their businesses forward. These providers needed to come up with infrastructures and technologies that could support applications at a massive scale. [14]

Big Data Security Analytics

Big Data tools and techniques are now bringing attention to analytics for fraud detection in healthcare, insurance, and bank. Experimental research in cyber security is rarely reproducible because today's data sets are not widely available to the research community and are often insufficient for answering many open questions. The goal of Big Data analytics for security is to obtain actionable intelligence in real time. [15]Securing the massive amounts of data that are inundating organizations can be addressed in several ways. A starting point is to basically get rid of data that are no longer needed. If you do not need certain information, it should be destroyed, because it represents a risk to the organization. That risk grows every day for as long as the information is kept. Of course, there are situations in which information cannot legally be destroyed; in that case, the information should be securely archived by an offline method. [16]

The real challenge may be determining whether the data are needed—a difficult task in the world of Big Data, where value can be found in unexpected places. For example, getting rid of activity logs may be a smart move from a security standpoint. Big Data analytics can be leveraged to improve information security and situational awareness. For example, Big Data analytics can be employed to analyze financial transactions, log files, and network traffic to identify anomalies and suspicious activities, and to correlate multiple sources of information into a coherent view. [17]

Streaming Analytics Platform for Real-time Big Data

Enterprises can use StreamAnalytix to take advantage of the steady stream of innovation flowing from the worldwide Open Source movement. Based on Apache Storm, StreamAnalytix brings together a proven technology stack to provide massive scalability, dynamic routing of data, and a best-in-class visual pipeline designer for rapid application development and deployment. The platform is designed to empower enterprises across industry verticals to address a wide range of applications and use cases. Included in this are analytics of sensor based data, machine generated (M2M) data, logs, click streams, advertising data, and processing of financial data and transactions in real-time.[13]

Acknowledgment

With how much data we've got and how fast it's generated, we can hit bandwidth constraints. Naturally the move is towards more of a cloud model, but we need to understand the trade-off between processing in the cloud and moving data around, versus

Analysis of available real time tools

TABLE 1. ANALYSIS OF AVAILABLETOOLS FORREAL TIME DATA ANALYTICS

Esper	Storm	Flume	Queue messaging system	Stream Analytix
-Processing engine for data streams -SQL-Like support–run queries on data stream -Sliding windows(time or length) -Pattern matching -Executes large number of queries simultaneously [9]	- Data Carrier for Esper -Facilitates data transfer -Continuous computation -Distributed, fault tolerant -Scalable, no data loss -Provides parallelism -Acking & replay capability [10]	- Stream oriented data flow - Log streaming from various sources - Collect, aggregate and move data to centralized data store - Distributed reliable - Failover and recovery mechanism [11]	- Bridging the gap between flume and storm -Robust messaging - Flexible routing - Highly available -Makes flume and storm integration loosely coupled [12]	- Reliable - Manageable - Vendor Support - Open source community innovation - No Vendor Lock-in - Low Cost - Future proof - Open, flexible, extensible, and easy-to-use - top of any standard Hadoop stack and is integrated with Apache Storm. [13]

the cost of storage of data and in-house analytics. Financial industry-Fraud detection, trading, E-commerce-Recommendations, Telecom industry- Machine to Machine communication, Supply chain management, Business Activity Monitoring requires real time analysis to helps organizations to stay ahead of competition and has to improve the return on investment for ourselves and their customer. The anxiety over which tools to use will decrease over coming years and a mix of both traditional relational database and SQL tools with big data technologies and NoSQL tools will be used, as they converge and take on each other's qualities. Enterprises are moving to add real-time streaming analytics engines or platforms to their Big Data architecture stack. To do that, they have two sub-optimal options - expensive, proprietary, commercial products or they can "Do-it-yourself" using raw Open Source.

References

[1] Hadoop. http://hadoop.apache.org/.

[2] Jimmy Lin and Chris Dyer. Data-Intensive Text Processing with Mapreduce.Morgan and Claypool Publishers, 2013.

[3] Yang, H., and Callan, J. 2009. "OntoCop: Constructing Ontologies for Public Comments," IEEE Intelligent Systems (24:5), pp.70-75.

[4] P. Institute, "Third Annual Benchmark Study on Patient Privacy and Data Security," Ponemon Institute LLC,2012.

[5] R. Cattell, "Scalable sql and nosql data stores," ACM SIGMOD Record, vol. 39, no. 4, pp. 12–27, 2011.

[6] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari, "S4: Distributed Stream Computing Platform," in Proceedings of IEEE International Conference on Data Mining Workshops (ICDMW), 2010, pp. 170–177.

[7] http://vip.informatica.com/?elqPURLPage=8808

[8] http://esper.codehaus.org

[9] https://github.com/nathanmarz/storm

[10] https://github.com/tomdz/storm-esper

[11] http://archive.cloudera.com/cdh/3/flume/UserGuide/#_architecture

[12] http://www.rabbitmq.com/

[13] http://www.streamanalytix.com

[14] D. Kossmann, T. Kraska, and S. Loesing, "An evaluationof alternative architectures for transaction processing in the cloud," in Proceedings of the 2010 international conference on Management of data. ACM, 2010, pp. 579–590.

[15] Camp, J. (2009). Data for Cybersecurity Research: Process and "whish list". Retrieved July 15, 2013, from http://www.gtisc.gatech.edu/files_nsf10/data-wishlist.pdf.

[16] Bryant, R., R. Katz & E. Lazowska. (2008). Big-Data Computing: Creating revolutionary breakthroughs in commerce, science and society. Washington, DC: Computing Community Consortium.

[17] https://www.cloudsecurityalliance.org/

[18] http://www.intel.com