CLOUD COMPUTING SECURITY AND BIG DATA ANALYTICS

BASAKY, FREDERICK DUNIYA, PhD.
Information Technology Department
Salem University Lokoja,
Nigeria

Abstract

Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. Big Data is the term for a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. Today Big Data draws a lot of attention in the IT world. The rise of the Internet and the digital economy are responsible for the rapid growth in the demand for data storage and analytics. So far, IT department are facing enormous challenges in protecting and analyzing these increased volumes of information. The type of information being created is no more traditional structured data rather it is unstructured data or Big Data that include documents, images, audio, video, and social media contents. Big Data Analytics is a way of extracting value from these huge volumes of information, and it drives new market opportunities and maximizes customer retention. This paper focuses on discussing the various technologies that work together as a Big Data Analytics system that can help predict future volumes, gain insights, take proactive actions, and give way to better strategic decision-making. This paper further adopts the usage and impact of big data analytics to the business value of an enterprise in order to improve its competitive advantage using a data sets such as Hadoop and Map Reduce.

Keywords: Cloud Computing, Big Data, Analytics, Hadoop, Map Reduce

Introduction

Data is everywhere. The amount of digital data that exists is growing at a rapid rate, doubling every two years, and changing the way we live. According to IBM, 2.5 billion gigabytes (GB) of data was generated every day in 2012.

Forbes in their articles, states that data is growing faster than ever before and by the year 2020, about 1.7 megabytes of new information will be created every seconds for every human being on the planet.

For organization of all sizes, data management has shifted from an important competency to a critical factor that determine market winners. Over a 1000 companies and government agencies are benefiting from the innovations of the web pioneers. This organization are defining new initiatives.

Big data analytics is where advanced analytic techniques operate on big data sets. Hence, big data analytics is really about two things—big data and analytics—plus how the two have teamed up to create one of the most profound trends in business intelligence (BI) today.

According to TDWI 2009 survey, 38% of organizations surveyed reported practicing advanced analytics, whereas 85% said they would be practicing it within three years. Why the rush to advanced analytics? First, change is rampant in business, as seen in the multiple “economies” we have gone through in recent years. Analytics helps us discover what has changed and how

we should react. Second, as we crawl out of the recession and into the recovery, there are more and more business opportunities that should be seized. To that end, advanced analytics is the best way to discover new customer segments, identify the best suppliers, associate products of affinity, understand sales seasonality, and so on. For these reasons, user organizations have been implementing analytics in recent years.

The rush to analytics means that many organizations are embracing advanced analytics for the first time, and hence are confused about how to go about it. Even if you have related experience in data warehousing, reporting, and online analytic processing (OLAP), you will find that the business and technical requirements are different for advanced forms of analytics. To help user organizations select the right form of analytics and prepare big data for analysis, this report will discuss new options for advanced analytics and analytic databases for big data so that users can make intelligent decisions as they embrace analytics.

Note that user organizations are implementing specific forms of analytics, particularly what is sometimes called advanced analytics. This is a collection of related techniques and tool types, usually including predictive analytics, data mining, statistical analysis, and complex SQL. We might also extend the list to cover data visualization, artificial intelligence, natural language processing, and database capabilities that support analytics (such as Map Reduce, in-database analytics, in-memory databases, columnar data stores).

Instead of “advanced analytics,” a better term would be “discovery analytics,” because that’s what users are trying to accomplish. (Some call it “exploratory analytics.”) In other words, with big data analytics, the user is typically a business analyst who is trying to discover new business facts that no one in the enterprise knew before. To do that, the analyst needs large volumes of data with plenty of detail. This is often data that the enterprise has not yet tapped for analytics.

For example, during the recent economic recession, companies were hit by new forms of customer agitation. To discover the root cause of the newest form of agitation, a business analyst would grab several terabytes of detailed data drawn from operational applications to get a view of recent customer behaviors.

The analyst might mix that data with historic data from a data warehouse. Dozens of queries later, the analyst would discover a new churn behavior in a subset of the customer base. With any luck, that discovery would lead to a metric, report, analytic model, or some other product of BI, through which the company could track and predict the new form of churn. Discovery analytics against big data can be enabled by different types of analytic tools, including those based on SQL queries, data mining, statistical analysis, fact clustering, data visualization, natural language processing, text analytics, artificial intelligence, and so on. It’s quite an arsenal of tool types, and savvy users get to know their analytic requirements before deciding which tool type is appropriate to their needs. All these techniques have been around for years, with many of them appearing in the 1990s. The difference today is that far more user organizations are actually using them. That’s because most of these techniques adapt well to very large, multi-terabyte data sets with minimal data preparation. That brings us to big data.

Big Data is an important concept, which is applied to data, which does not conform to the normal structure of the traditional database. Big Data consists of different types of key technologies like Hadoop, HDFS, NoSQL, Map Reduce, MongoDB, Cassandra, PIG, HIVE, and HBASE that work together to achieve the end goal like extracting value from data that would be

previously considered dead. According to a report published by Transparency Market Research, the total value of big data was estimated at $6.3 billion as of 2012, by 2018, it's expected to reach the staggering level of $48.3 billion that's almost a 700 percent increase. It has been observed that organizations utilize less than 5 percent of their available data. This is because the rest is simply too expensive to deal with. Big Data is derived from multiple sources. It involves not just traditional relational data, but all unstructured data sources that are growing at a significant rate. For instance, machine-derived data multiplies quickly and contains rich, diverse content that needs to be discovered. Another example, human-derived data from social media is more textual but the valuable insights are often overloaded with many possible meanings.

Big Data Analytics reflect the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods. From businesses and research institutions to governments, organizations now routinely generate data of unprecedented scope and complexity. Gleaning meaningful information and competitive advantages from massive amounts of data has become increasingly important to organizations globally. Trying to efficiently extract the meaningful insights from such data sources quickly and easily is challenging. Thus, analytics has become inextricably vital to realize the full value of Big Data to improve their business performance and increase their market share. The tools available to handle the volume, velocity, and variety of big data have improved greatly in recent years. In general, these technologies are not prohibitively expensive, and most of the software are open source. Hadoop, the most commonly used framework, combines commodity hardware with open source software. It takes incoming streams of data and distributes them onto cheap disks; it also provides tools for analyzing the data. However, these technologies do require a skill set that is new to most IT departments, which will need to work hard to integrate all the relevant internal and external sources of data. Although attention to technology isn't sufficient, it is always a necessary component of a big data strategy. This paper discusses some of the most commonly used big data technologies mostly open source that work together as a big data analytics system for leveraging large quantities of unstructured data to make more informed decisions.

Review of Literature

Big Data is a data analysis methodology enabled by recent advances in technologies that support high-velocity data capture, storage and analysis. Data sources extend beyond the traditional corporate database to include emails, mobile device outputs, and sensor-generated data where data is no longer restricted to structured database records but rather unstructured data having no standard formatting. Since Big Data and Analytics is a relatively new and evolving phrase, there is no uniform definition; various stakeholders have provided diverse and sometimes contradictory definitions. One of the first widely quoted definitions of Big Data resulted from the Gartner report of 2001. Gartner proposed that, Big Data is defined by three V's volume, velocity, and variety. Gartner expanded its definition in 2012 to include veracity, representing requirements about trust and uncertainty pertaining to data and the outcome of data analysis. In a 2012 report, IDC defined the 4th V as value—highlighting that Big Data applications need to bring incremental value to businesses. Big Data Analytics is all about processing unstructured information from call logs, mobile-banking transactions, online user

generated content such as blog posts and tweets, online searches, and images which can be transformed into valuable business information using computational techniques to unveil trends and patterns between datasets.

Another dimension of the Big Data definition involves technology. Big Data is not only large and complex, but it requires innovative technology to analyze and process. In 2013, the National Institute of Standard and Technology (NIST) Big Data workgroup proposed the following definition of Big Data that emphasizes application of new technology; Big Data exceed the capacity or capability of current or conventional methods and systems, and enable novel approaches to frontier questions previously inaccessible or impractical using current or conventional methods. Business challenges rarely show up in the appearance of a perfect data problem, and even when data are abundant, practitioners have difficulties to incorporate it into their complex decision-making that adds business value. In 2012, McKinsey & Company conducted a survey of 1,469 executives across various regions, industries and company sizes, in which 49 percent of respondents said that their companies are focusing big data efforts on customer insights, segmentation and targeting to improve overall performance. An even higher number of respondents 60 percent said their companies should focus efforts on using data and analytics to generate these insights. Yet, just one-fifth said that their organizations have fully deployed data and analytics to generate insights in one business unit or function, and only 13 percent use data to generate insights across the company. As these survey results show, the question is no longer whether big data can help business, but how can business derive maximum results from big data.

Characteristics of Big Data

i. Volume – The name 'Big Data' itself is related to a size which is enormous. Size of data plays very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon volume of data. Hence, 'Volume' is one characteristic which needs to be considered while dealing with 'Big Data'.

ii. Variety – Another aspect of 'Big Data' is its variety. Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Now days, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. is also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analysing data.

iii. Velocity – The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

iv. Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Benefits of Big Data Processing

Ability to process 'Big Data' brings in multiple benefits, such as-

- Businesses can utilize outside intelligence while taking decisions Access to social data from search engines and sites like Facebook, twitter are enabling organizations to fine tune their business strategies.
- Improved customer service
  Traditional customer feedback systems are getting replaced by new systems designed with 'Big Data' technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.
- Early identification of risk to the product/services, if any.
- Better operational efficiency 'Big Data' technologies can be used for creating staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of 'Big Data' technologies and data warehouse helps organization to offload infrequently accessed data.

Big Data Technologies

The Big Data landscape is dominated by two classes of technology: systems that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored; and systems that provide analytical capabilities for retrospective, complex analysis that may touch most or all of the data. These classes of technology are complementary and frequently deployed together.

Operational and analytical workloads for Big Data present opposing requirements and systems have evolved to address their particular demands separately and in very different ways. Each has driven the creation of new technology architectures. Operational systems, such as the NoSQL databases, focus on servicing highly concurrent requests while exhibiting low latency for responses operating on highly selective access criteria. Analytical systems, on the other hand, tend to focus on high throughput; queries can be very complex and touch most if not all of the data in the system at any time. Both systems tend to operate over many servers operating in a cluster, managing tens or hundreds of terabytes of data across billions of records.

Operational Big Data

For operational Big Data workloads, NoSQL Big Data systems such as document databases have emerged to address a broad set of applications, and other architectures, such as key-value stores, column family stores, and graph databases are optimized for more specific applications. NoSQL technologies, which were developed to address the shortcomings of relational databases in the modern computing environment, are faster and scale much more quickly and inexpensively than relational databases.

Critically, NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational Big Data workloads much easier to manage, and cheaper and faster to implement.

In addition to user interactions with data, most operational systems need to provide some degree of real-time intelligence about the active data in the system. For example, in a multi-user game or financial application, aggregates for user activities or instrument performance are displayed to users to inform their next actions. Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

Analytical Big Data

Analytical Big Data workloads, on the other hand, tend to be addressed by MPP database systems and Map Reduce. These technologies are also a reaction to the limitations of traditional relational databases and their lack of ability to scale beyond the resources of a single server. Furthermore, Map Reduce provides a new method of analysing data that is complementary to the capabilities provided by SQL.

As applications gain traction and their users generate increasing volumes of data, there are a number of retrospective analytical workloads that provide real value to the business. Where these workloads involve algorithms that are more sophisticated than simple aggregation, Map Reduce has emerged as the first choice for Big Data analytics. Some NoSQL systems provide native Map Reduce functionality that allows for analytics to be performed on operational data in place. Alternately, data can be copied from NoSQL systems into analytical systems such as Hadoop for Map Reduce.

Combining Operational and Analytical Technologies; Using Hadoop

New technologies like NoSQL, MPP databases, and Hadoop have emerged to address Big Data challenges and to enable new types of products and services to be delivered by the business.

One of the most common ways companies are leveraging the capabilities of both systems is by integrating a NoSQL database such as MongoDB with Hadoop. The connection is easily made by existing APIs and allows analysts and data scientists to perform complex, retroactive queries for Big Data analysis and insights while maintaining the efficiency and ease-of-use of a NoSQL database.

NoSQL, MPP databases and Hadoop are complementary: NoSQL systems should be used to capture Big Data and provide operational intelligence to users, and MPP databases and Hadoop should be used to provide analytical insight for analysts and data scientists. Together, NoSQL, MPP databases and Hadoop enable businesses to capitalize on Big Data.

Competitive Advantages

Thomas H. Davenport was perhaps the first to observe in his Harvard Business Review article published in January 2006 (“Competing on Analytics”) how companies who orientated themselves around fact based management approach and compete on their analytical abilities considerably out-performed their peers in the marketplace. The reality is that it takes continuous improvement to become an analytics-driven organization. In a presentation given at the Strata New York conference in September 2011, McKinsey & Company showed the eye opening; 10-year category growth rate differences between businesses that smartly use their big data and those that do not.

Amazon uses Big Data to monitor, track and secure 1.5 billion items in its inventory that are laying around 200 fulfilment centers around the world, and then relies on predictive analytics for its ‘anticipatory shipping’ to predict when a customer will purchase a product, and pre-ship it to a depot close to the final destination. Wal-Mart handles more than a million customer transactions each hour, imports information into databases to contain more than 2.5 petabytes and asked their suppliers to tag shipments with radio frequency identification (RFID) systems that can generate 100 to 1000 times the data of conventional bar code systems. UPS deployment of telematics in their freight segment helped in their global redesign of logistical

networks. Amazon is a big data giant and the largest online retail store. The company pioneered e-commerce in many different ways, but one of its biggest successes was the personalized recommendation system, which was built from the big data it gathers from its millions of customers' transactions.

The U.S. federal government collects more than 370,000 raw and geospatial datasets from 172 agencies and sub agencies. It leverages that data to provide a portal to 230 citizen-developed apps, with the aim of increasing public access to information not deemed private or classified. Professional social network LinkedIn uses data from its more than 100 million users to build new social products based on users' own definitions of their skill sets. Silver Spring Networks deploys smart, two-way power grids for its utility customers that utilize digital technology to deliver more reliable energy to consumers from multiple sources and allow homeowners to send information back to utilities to help manage energy use and maximize efficiency. Jeffrey Brenner and the Camden Coalition mapped a city's crime trends to identify problems with its healthcare system, revealing services that were both medically ineffective and expensive.

Conclusion

Today's technology landscape is changing fast. Organizations of all shapes and sizes are being pressured to be data driven and to do more with less. Even though big data technologies are still in a nascent stage, relatively speaking, the impact of the 3V's of big data, which now is 5v's cannot be ignored. The time is now for organizations to begin planning for and building out their Hadoop-based data lake. Organizations with the right infrastructures, talent and vision in place are well equipped to take their big data strategies to the next level and transform their businesses. They can use big data to unveil new patterns and trends, gain additional insights and begin to find answers to pressing business issues. The deeper organizations dig into big data and the more equipped they are to act upon what's learned, the more likely they are to reveal answers that can add value to the top line of the business. This is where the returns on big data investments multiply and the transformation begins. Harnessing big data insight delivers more than cost cutting or productivity improvement but it definitely reveals new business opportunities. Data-driven decisions always tend to be better decisions.

References

Apache Software Foundation. (2010).

Apache ZooKeeper. Retrieved April 5, 2015 from https://zookeeper.apache.org Chae, B., Sheu, C., Yang, C. & Olson, D. (2014). The impact of advanced analytics and data accuracy on operational performance: A contingent resource based theory (RBT) perspective, Decision Support Systems, 59,119-126.

Bo Li, (2013). Survey of Recent Research Progress and Issues in Big Data.

Chambers, C., Raniwala, A., Adams, S.,Henry, R., Bradshaw, R., and Weizenbaum, N. (2010). Flume Java: Easy, Efficient Data-Parallel Pipelines. Google, Inc. Retrieved April 1, 2015 from http://pages.cs.wisc.edu/~akella/CS838/F12/838-CloudPapers/FlumeJava.pdf

Cisco Systems. Cisco UCS Common Platform Architecture Version 2 (CPA v2) for Big Data with Comprehensive Data Protection using Intel Distribution for Apache Hadoop. Retrieved March 15, 2015, from

http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/Cisco_UCS_CPA_for_Big_Data_with_Intel.html.

DATASTAX Corporation. (2013, October). Big Data: Beyond the Hype - Why Big data Matters to you [White paper]. Retrieved March 15, 2015 from https://www.datastax.com/wp-content/uploads/2011/10/WP-DataStax-BigData.pdf.

Davenport, T & Patil, D. (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review,90, 70-76.

Dhawan, S. & Rathee, S. (2013). Big Data Analytics using Hadoop Components like Pig and Hive. American International Journal of Research in Science, Technology, Engineering & Mathematics, 88, 13-131. Retrieved from http://iasir.net/AIJRSTEMpapers/AIJRSTEM13-131.pdf. https://www.mongodb.com/big-data-explained.