International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 5 Issue 1 ǁ January. 2017 ǁ PP.16-21

Moving Toward Big Data: Challenges, Trends and Perspectives

Ming-Hsing Chiu[1], Joshua Mcadams[2]
[1]*Department of Computer Science, Dillard University, New Orleans, LA, USA*
[2]*Google-in-Residence Scholar, Dillard University, New Orleans, LA, USA*

Abstract: Big data refers to the organizational data asset that exceeds the volume, velocity, and variety of data typically stored using traditional structured database technologies. This type of data has become the important resource from which organizations can get valuable insightand make business decision by applying predictive analysis. This paper provides a comprehensive view of current status of big data development,starting from the definition and the description of Hadoop and MapReduce – the framework that standardizes the use of cluster of commodity machines to analyze big data. For the organizations that are ready to embrace big data technology, significant adjustments on infrastructure andthe roles played byIT professionals and BI practitioners must be anticipated which is discussed in the challenges of big data section. The landscape of big data development change rapidly which is directly related to the trend of big data. Clearly, a major part of the trend is the result ofthe attempt to deal with the challenges discussed earlier. Lastly the paper includes the most recent job prospective related to big data. The description of several job titles that comprise the workforce in the area of big data are also included.

Keywords: Big Data, MapReduce and Hadoop, Data Scientist,

I. INTRODUCTION

Big data technology has revolutionize the way organizations collect, manage and utilize the explosive amount of data that has been generated in recent years. According to the report by International Data Corporation (IDC), from the beginning of time through 2011, the world contained nearly 1 zettabyte (10^{21} bytes) of data. Then the floodgate opened. 2011 saw the generation of twice (1.8 zettabytes) the amount of all-through-history number in just one year.It was estimated by IDC that by 2020, just 4 years from now, we'll be generating 35 zettabytes per year. This amount of data is clearly far exceeding that traditional structured database technologies can handle. Owing to the development of new data processing technology such as Hadoop and MapReduce, organizations are now able to utilize predictive analytics on the vast amount of data to obtain insightful information. Many organization have been using the data asset to gain the valuable insight that are not considered possible before.

Famous examples include: 1.IBM were using electronic health record data to predict heart disease [4], and using data from the World Health Organization to predictthe location of future malaria outbreaks [5]. 2. Target uses a wealth of customer data to predict future purchasing habits. Specifically, pregnancy kicks off a chain of purchases that are fairly distinctive. Target's data collection is spookily prescient, sending one teen customer nappy vouchers before her own father knew she was pregnant [3]. 3. WeatherSignal App works by repurposing the sensors in Android devices to map atmospheric readings [6]. Handsets such as the Samsun S4, contain a barometer, hygrometer (humidity), ambient thermometer and lightmeter. Obviously, the prospect of millions of personal weather stations feeding into one machine that will average out readings is exciting, and one that has the potential to improve forecasting. 4. One of the most exciting examples is using big data to predict the ongoing 2016 presidential election. OpenText's U.S. Election Tracker reads, analyzes and visualizes key U.S. Election big data, every day, from hundreds of news sources, helping the U.S. voters make more informed decisions. The analysis includes natural language processing, semantic processing, sentiment analysis and opinion algorithms [7].

In the following section, 5Vs of big data isdescribed. The frameworkof Hadoop and MapReduce that enables the big data analytics is introduced in section 3. Followed by the discussion of challenges of big data. Section 5 lists the trend of big data development. Section 6 covers the job perspectives related to big data. Finally, section 7 contains some concluding remarks.

II. 5Vs OF BIG DATA

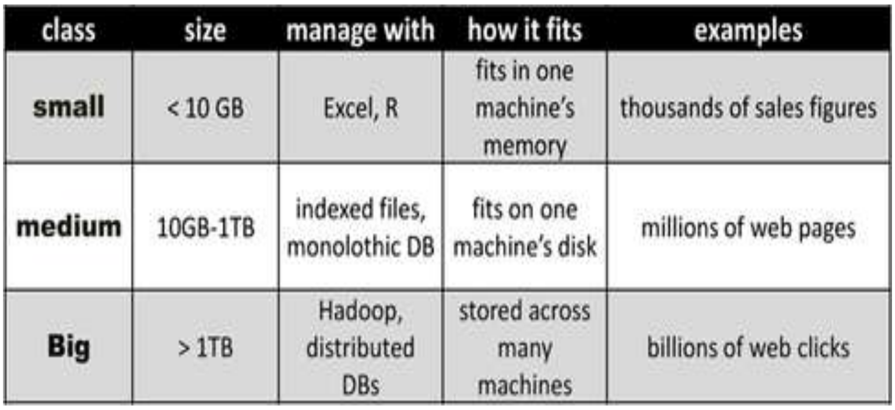
Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. Volume describes the amount of data generated by various sources. Big data is usually associated with this characteristic. Velocity describes the frequency at which data is generated, captured and shared. Variety describes the diversity of data, which include structured and unstructured data. Structured data refers to the type of data found in spreadsheet or in the relational database, which comprise only about 5% of all active data. All other data are considered unstructured data,including but not limiting to audio, video, log files, social media streams, and sensor data from the “internet of things”.To illustrate how fast data can be generated, Twitter produces over 90 million tweets per day. And Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data, the equivalent of 167 times the information contained in all the books in the US Library of Congress.

Veracity is the 4th V coined by IBM, which symbolizes the untrustworthy nature in some sources of data.Due to the integration of disparate data systems, the correctness and accuracy of information is often questionable. Value, the 5th V, was introduced by Oracle as a crucial characteristic of Big Data. The ability to understand and manage the vast amount of data, and then integrate them into the larger Business Intelligence ecosystem can provide previously unknown insights associated with the organizations.

III. MAPREDUCE AND HADOOP

The underliningengine that allows the quick processing of big data is the framework of Hadoop and MapReduce. The concept of MapReduce was first proposed by Dean and Ghemawat from Google in 2004 [1],which became the primary programming framework for processing big data. It first splits the input data into a number of chunks that is manageable by a single processor, then goes through the mapping, shuffling, and reducing stages to get the final result.Apache Hadoop was implemented in 2006 by a group of software developers headed by Cutting [2]. Hadoop is an architecture that enables MapReduce to be run on clusters of commodity machines in parallel that are dynamically scalable. In other word, the number of machines can be added or reduced on demand, depending on the load of data. The Distributed file system among the machines in the clusters is called Hadoop Distributed File System (HDFS).

Table 1 depicts the size of data that would require the use of Hadoop. In 2009, Yahoo demonstrated the power of Hadoop that ran 17 cluster of 24,000 machine, sorting one Terabyte of data in 62 second. In 2011, Yahoo announced its new humongous Hadoop cluster, reaching 42K Hadoop nodes and hundreds of petabytes of storage.



class	size	manage with	how it fits	examples
small	< 10 GB	Excel, R	fits in one machine’s memory	thousands of sales figures
medium	10GB-1TB	indexed files, monolothic DB	fits on one machine’s disk	millions of web pages
Big	> 1TB	Hadoop, distributed DBs	stored across many machines	billions of web clicks

Table 1: Size of Data and Corresponding Content

IV. THE CHALLENGES OF BIG DATA

Conventionally, the relationship between IT (information technology) practitioners and BI (business intelligence) professionals in an organization is relatively well defined: IT processes organizational data and generatesstructured data such as relational database or spreadsheet, which then used by BI to make decisions.As big data technology matures to a point in which more organizations are prepared to pilot and adopt it as a core component of the information management and integrated analytics, IT and BI alike will bump up against a number of challenges that must be addressed before any big data program can be successfully implemented. Below is the discussion of the five most common faced challenges[8, 9].

4.1 Talent Gap Challenge

There is a growing community of application developers who are relying on leveraging Hadoop MapReduce framework. The promotion of these technologies creates a group of experts that are experienced in tool implementation and its use as a programming model, rather than the data management aspects. This suggests that many big data tool experts remain somewhat immature on the practical aspects of data modeling, data architecture, and data integration.

There is no doubt that as more data practitioners become engaged, the talent gap will gradually narrow down, but when developers are not skilled at addressing these fundamental data architecture and data management issues, the ability to achieve and maintain a competitive edge through technology adoption will be severely impaired. Section 6 includes a description of all job titles that comprise the talents needed for the complete big data operation.

4.2 Infrastructure/Platform Challenge

Big Data is a much talked about technology across businesses today. The majority of organizations are convinced of its usefulness, however the implementations primarily focus on applications rather than infrastructure. Most organizations have their own legacy computer systems. Some may have storage system based on NetApp, EMC or other major players. Moving toward Hadoop environment may prove to be daunting and challenging.

The organizations opt to take advantage of Big Data technology must answer a number of questions. How do you apply big data to business applications? What kind of storage system and hardware architecture do you attempt to build? Are you going to build it all by yourself? Unpredictability of business world must also be taken into consideration. The underlying platform chosen must be robust and scalable in a cost effective manner. Cloud service can be a viable solution for organizations choose to avoid installing their own infrastructures.

4.3 Data Sources Integration Challenge

The volume, velocity, and variety of big data has direct impact on data integration and management. Organizations must first deal with the variety of data which may be generated by IT, Internet of things, or users. It takes a lot of skills to put data in the right platform so that you can use it properly. For example, if the data comes from social media content, you need to know who the user is in a general sense, such as a customer using a particular set of products, and figure out what it's you are trying to “visualize” out of the data.

Another aspect is due to the velocity of the data which make data synchronization necessary across the data sources under certain circumstances. There are two types of synchronizations. From a data currency perspective, it implies that the data coming from one source is not out of date with data coming from another source. From a semantics perspective, synchronization implies the similarity of data concepts, definitions, metadata, and the like. One of the promise of big data is that it's possible to reduce the costly upfront data preparation and data engineering which typically constitute 80% of the time and cost of data management.

4.4 The Governance Challenge

Data governance is the process of ensuring the compliance of organizational data policy, which can be a difficult task considering the variety and reliability of the unstructured data. The objective of data governance is to institute the right levels of control to achieve one of three outcomes: identify data issues that might have negative business impact; prioritize those issues in relation to their corresponding business value drivers; and have data stewards take the proper actions when alerted to the existence of those issues [10].

It is naive to assume that when it comes to big data governance the approach to data quality is the same as traditional approaches. When we examine the key characteristics of big data analytics, the analogy with the conventional approaches to data quality and data governance starts to break down.Big data applications look at many input streams originating within and outside the organization, some taken from a variety of social networking streams, syndicated data streams, news feeds, preconfigured search filters, public or open-sourced datasets, sensor networks, or other unstructured data streams. Such diverse datasets resist singular approaches to governance.

4.5 Big Data Syndication Challenge

Most of the practical cases for big data involve data availability, ranging from augmenting existing data storage to providing access to end-users employing business intelligence tools for the purpose of value discovery. These BI tools not only must be able to connect to one or more big data platforms, they must provide transparency to the data consumers to reduce or eliminate the need for custom coding. At the same time, as the number of data consumers grows, it can be anticipated a need to support a rapidly expanding collection of many simultaneous user accesses. That demand may spike at different times or the day or in reaction to different aspects of business process cycles. Ensuring right-time data availability and reusability to the community of data consumers becomes a critical success factor.

Big data syndication provides a supply chain to the different types of downstream applications in a way that is seamless and transparent to the consuming applications while elastically supports demands. Finally, it's important to realize that funding is a challenge that always plays a dominant role at every stage of big data integration because these days most budget is very tight. As a result, every penny must be spent wisely. Clearly, in establishing infrastructure, what we are looking for is predictability in cost. One of the reason developers of big data move to Hadoop is because it's complying with industrial standard hardware, and scalable. In addition, the long term cost of managing the storage facilities, analytic platforms, and all the related infrastructure must be taken into consideration.

V. THE TREND OF BIG DATA

The year 2015 was an important one in the world of big data. What used to be hype became the norm as more businesses realized that data, in all forms and sizes, is critical to making the best possible decisions. In 2016 and beyond, we'll see continued growth of systems that support unstructured data as well as massive volumes of data [11]. These systems will evolve and mature to operate well inside of enterprise IT systems and standards, enabling both business users and data scientists to fully realize the value of big data. A number of trends are discussion below. Note that most of the trends included in the discussion are stimulated by the need to tackle the challenges covered in previous section.

5.1 The NoSQL Take 0ver

We noted the increasing adoption of NoSQL technologies, which are commonly associated with unstructured data, as the conventional relational database management systems become inadequate. Going forward, the shift to NoSQL databases becoming a leading piece of the enterprise IT landscape as the benefits of schema-less database concepts become more pronounced. According to Gartner's Magic Quadrant [12], the developers of Operational Database Management Systems, which in the past were dominated by Oracle, IBM, Microsoft and SAP, have shifted to more than a dozen NoSQL developers, including MongoDB, SQLFire, Casandra, HBase, and Amazon Web Services (with DynamoDB).

5.2 Apache Spark Lights Up Big Data

Apache Spark has moved from being a component of the Hadoop ecosystem to the big data platform of choice for a number of enterprises. Spark provides dramatically increased data processing speed compared to Hadoop and is now the largest big data open source project, according to Spark originator and Databricks co-founder, MateiZaharia [13]. We see more and more compelling enterprise use cases around Spark, such as at Goldman Sachs where Spark has become the major platform of big data analytics.

5.3 Hadoop continues to evolve

In a recent survey of 2,200 Hadoop customers, only 3% of respondents anticipate they will be doing less with Hadoop in the next 12 months. 76% of those who already use Hadoop plan on doing more within the next 3 months and finally, almost half of the companies that haven't deployed Hadoop say they will within the next 12 months.

As further evidence to the growing trend of Hadoop becoming a core part of the enterprise IT landscape, we'll see investment grows in the components surrounding enterprise systems such as security. Apache Sentry project provides a system for enforcing fine-grained, role based authorization to data and metadata stored on a Hadoop cluster. These are the types of capabilities that customers expect from their enterprise-grade RDBMS platforms and are now coming to the forefront of the emerging big data technologies, thus eliminating one more barrier to enterprise adoption. We also see a growing demand from end users for the same fast data exploration capabilities they've come to expect from traditional data warehouses.

5.4 The Number Of Options For Preparing Data Grows.

As the demand to reduce the cost of preparing raw data for analytics mounts, a number of developments have been underway. Self-service data preparation tools are exploding in popularity. This is in part due to the shift toward generated data discovery tools such as Tableau that reduce time to analyze data. Business users also want to be able to reduce the time and complexity of preparing data for analysis, something that is especially important in the world of big data when dealing with a variety of data types and formats. We've seen a host of innovation in this space from companies focused on end user data preparation for Big Data such as Alteryx, Trifacta, Paxata and Lavastorm.

5.5 Data Warehouse Growth Is Heating Up In The Cloud

Data warehouse have not been phasing out as were expected. But it's no secret that growth in this segment of the market has been slow. But we now see a major shift in the application of this technology to the cloud where Amazon led the way with an on-demand cloud data warehouse in Redshift. Redshift was AWS's fastest growing service but it now has competition from Google with BigQuery. Offerings from long time data warehouse power players such as Microsoft (with Azure SQL Data Warehouse) and Teradata along with new start-ups such as Snowflake, winner of Strata + Hadoop World 2015 Startup Showcase, also are gaining adoption in this arena. Surveys indicate 90% of companies who have adopted Hadoop will also keep their data warehouses. With these new cloud offerings, those customers can dynamically scale up or down the amount of storage and compute resources in the data warehouse relative to the larger amounts of information stored in their Hadoop data lake.

5.6 IoT, Cloud and Big Data Come Together.

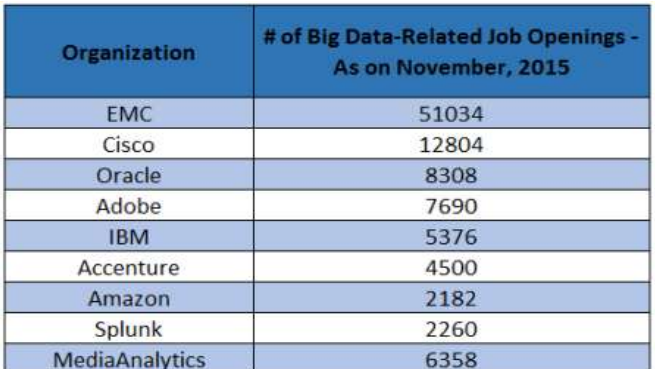
The technology is still in its early days, but the data from devices in the Internet of Things will become one of the “killer apps” for the cloud and a driver of petabyte scale data explosion. For this reason, we see leading cloud and data companies such as Google, Amazon Web Services and Microsoft bringing Internet of Things services to life where the data can move seamlessly to their cloud based analytics engines.

VI. JOB PERSPECTIVESRELATED TO BIG DATA

According to a Big Data Survey by Knowledgent, there are clear indicators that Big Data is moving out of the experimental stage. Most respondents of this survey firmly believe that many organizations would be utilizing Big Data in the production environment and over 60% of them responded that Big Data initiatives are very or extremely important to their organizations**.** In 2015, 1.9 million jobs related to Big Data were created in the United States to work on the IT-side of big data projects,but each of these jobs is supported by 3 new jobs outside of IT.In general, several professionals with different skills are required to complete any project. For example, as mentioned earlier,around 80 % of effort for big data project are spent on cleaning, integrating, and transforming the data before it can be utilized.

EMC, IBM, Cisco, Oracle are just a few of the top companies who are looking for Big Data skills.

Table 2 is a distribution of job demands of the top Big Data employers today, according to Forbes [14]



Organization	# of Big Data-Related Job Openings - As on November, 2015
EMC	51034
Cisco	12804
Oracle	8308
Adobe	7690
IBM	5376
Accenture	4500
Amazon	2182
Splunk	2260
MediaAnalytics	6358

Table 2: Big Data Jobs Demand from Major Employers

Jobs that require big data expertise pay well, too. The average salary, according to a report by Forbes last year, is $124,000. Below is the list of 7 most common job titles related to big data [15].

Data Scientist

Sometimes referred as “the sexiest job in the 21 century”, data scientists are responsible for analyzing data and extracting useful information out of it.Data scientists thrive on solving real world problems with real data. They are very good at using different techniques for analyzing data from different sources to help business make intelligent decisions. They need to have both skills of a software engineer and an applied scientist.

Data Analyst

The data analyst collects, processes and performs statistical data analyses. A data analyst needs to understand R, Python, HTML, JavaScript, C/C++, SQL and NoSQL databases.

Data Architect

Data architects are responsible for the design, structure, and maintenance of data, ensuring the accuracy and accessibility of data.

Data Engineer

Big data engineers are all-purpose workforce of a big data analytics operation. They often come from programming backgrounds, and are experts in big data frameworks, such as Hadoop. They’re called on to ensure that data pipelines are scalable, repeatable, and secure, and can serve multiple constituents in the enterprise.

Statistician

Collect, analyze and interpret qualitative and quantitative data with statistical theories and methods. Statisticians have to be mathematicians, who can churn out reports basis some data provided to them. The second part is about being able to use a big data technology to automate data processing that provides insight on a real-time basis.

Database Administrator

Big data and databases go hand-in-hand. A database administrator ensures that the database is available, is performing properly, and is secure.

Business Analyst

Business analysts are primarily responsible for generating insights that convert big data into business value. This can include translating insights and patterns embedded in data assets into language understood by the business administrator.

VII. CONCLUSION

Big data technology has revolutionize the way organizations collect, manage and utilize the explosive amount of data that has been generated in recent years, including but not limiting to audio, video, log files, social media streams, and sensor data from the “internet of things”. This paper provides a comprehensive view of current status of big data development. This paper has described and discussed the 5Vs of big data, the framework of Hadoop and MapReduce that enables the big data analytics, the challenges of big data, the trend of big data development, and the job perspectives related to big data. This paper will prove most beneficial to the planner of the organizations ready to adopt big data technology as well as for the newly grads that are looking for jobs in the area of big data.

REFERENCES

[1]. MapReduce: Simplified Data Processing on Large Clusters, Dean &Ghemawat. OSDI 2004, San Francisco, CA, December, 2004.

[2]. Cutting, Doug (28 Jan 2006). "new mailing lists request: hadoop". issues.apache.org.

[3]. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=3&hp&pagewanted=all&

[4]. http://cloudtimes.org/2013/10/14/ibm-wants-to-predict-heart-disease-through-big-data-analytics/

[5]. http://venturebeat.com/2013/09/29/ibm-uses-big-data-to-predict-outbreaks-of-dengue-fever-and-malaria/#3QW6mxLOYOPrZTpC.99

[6]. http://blogs.scientificamerican.com/guest-blog/weathersignal-big-data-meets-forecasting/

[7]. http://www.forbes.com/sites/adigaskell/2016/03/11/using-big-data-to-predict-the election/#c07b7b776642

[8]. https://www.sas.com/resources/asset/five-big-data-challenges-article.pdf

[9]. https://www.progress.com/docs/default-source/default-document-library/Progress/Documents/Papers/Addressing-Five-Emerging-Challenges-of-Big-Data.pdf

[10]. http://data-informed.com/data-governance-for-big-dta-analytics-considerations-for-data-policies-and-processes/

[11]. http://www.tableau.com/about/blog/2015/12/top-8-trends-big-data-2016-47846

[12]. https://www.gartner.com/doc/reprints?id=1-2XXUR6C&ct=160204&st=sb

[13]. http://spark.apache.org/

[14]. http://www.forbes.com/sites/louiscolumbus/2015/11/16/where-big-data-jobs-will-be-in-2016/#438e0339f7f1

[15]. http://talkincloud.com/cloud-computing/7-big-data-jobs-you-need-know#slide-7-field_images-54391