GRASPING THE FUNDAMENTALS OF BIG DATA IN DISTRIBUTED COMPUTING

[1]MOHANA RAMAKRISHNAN, [2]ELAMMARAN JAYAMANI

[1]Faculty of Engineering, Computing and Science, [2]Swinburne University of Technology, Sarawak Campus
E-mail: mramakrishnan@swnburne.edu.my

Abstract- This paper exploringa number of big data contents, which selected in research for example data tools, data types and storage techniqueused in the real world applications. An intellectual revision made from the basics of big data.Big Data is the new era, giving new challenging environment in day-today life to understand or get rid of big data as fast. The choice of big data is often a fairly difficult one, and it is not uniquely determined by the facilities available.It must always be taken in the light of the operations that are performed on the large amount of data. It is an abstraction of reality that certain properties and characteristics of the real objects are ignored because they are peripheral and irrelevant to particular problems. In this context, the significance of various big data projects becomes apparent. The importance of using big data that offer a convenient set of basic abstractions common to most problems of data processing lies mainly in the area of reliability of the resulting programs. This paper describes big data tools,data types and techniques used in real time authentic applications.

Keywords- Big Data, Tools and Techniques, Data Types, Characteristics, Distributed Computing**.**

I. INTRODUCTION

There are three distinguishable characteristics of data size: condition, location, and population. Condition refers to the state of readiness of the data for analysis. Location refers to where the data reside. Big data reside in relational databases consosting of a set of data tables. Population refers to the group of individuals having qualities or characteristics in common and related to the study under consideration. Big data is not a new type of data, rather it defining its characteristics and affords more innovative techniques for best challenging environment. Data always needed for accuracy and efficiency, best-giving man machine system to promote objectives with understanding logic. Lisa Arthur in a Forbes defined big data as “a collection of data from traditional and digital sources inside and outside of our organization that represents a source for on going discovery and analysis”. Big data and its research topics are more in reality of distributed computing systems. The distributed dataare growing at a huge speed, so it leads to difficulty to manage massive data. Big data are using the techniques called “information retrieval” which defines a quadruple collection of documents, set of queries, framework and a ranking model, function associated with the document query. There are two sides of big data. On one side is classical statistics that considers big as simply not small. Theoretically, big is the sample size after asymptotic properties of the method “kick in” for valid results. On the other side is contemporary statistics that considers big in terms of lifting observations and learning from the variables. Although it depends on who is analyzing the data, a sample size greater than 50,000 individuals can be considered big. Thus, calculating the average income from a database of 2 million individuals requires heavy-duty lifting(number cruching). In terms of learning or uncovering the structure among the variables, big can be considered 50 variables or more. Regardless of which the data analyst is working, Exploratory Data Analysis(EDA) scales up for both rows and colums of the data table.

II. TYPICAL BIG DATA TYPES

In general, Big Data is defined by three dimensions:
1. Volume
2. Variety
3. Velocity

A. Volume

Data volume can be measured by the sheer quantity of transactions, events, or amount of history that creates the data volume, but the volume is often further exacerbated by the attributes, dimensions, or predictive variables. Typically, analystics have used smaller data sets called samples to create predictive models. Its similar to seeing the iceberg that sits above the waterline but not seeing the huge iceberg that lies beneath the surface. By removing the data volume constraint and using larger data sets, enterprises can discover subtle patterns that can lead to targeted actionable micro-decisions, or they can factor in more observations or variables into predictions that increase the accuracy of the predictive models.

B. Variety

Data variety is the assortment of data. Traditional data, especially operational data, is “structured” as it is put into a database based on the type of data(ie., char, numeric, floating point etc.,). Over the past couple of decades, data has increasingly become “unstructured” as the sources of data have proliferated beyond operational applications. Often times, text, audio, video, image, geospatial, and Internet data are considered “unstructured data”. Since many of the sources of this data are often a combination of

different types of data that has some pattern or structure that is not as strictly defined as structured data.

C. Velocity

Data velocity is about the speed at which data is created, accumulated, ingested,and processed. The increasing pace of the world has put demands on business to process information in real-time or with near real-time responses. This may mean that data is processed on the fly or while "streaming" by to make quick, real-time decisions or it may be that monthly batch process are run interday to produce more timely decisions.

In a recent study, two more dimensions are added as variability, data flows can be highly inconsistent with periodic peaks and complexity, when dealing with huge volumes of data, it comes from multiple sources. It is quite an undertaking to link, match,cleanse and transform across systems. However, it is necessary to connect ad correlate relationships, hierarchies and multiple data linkages or our data can quickly spiral out of control.

III. MULTI-TOOLS TO ANALYZE BIG DATA

Multiple tools are used to analyze big data, which if often an iterative process in real-time environments. These processes are divided into five phases given by Smith in his research. Each phase defines in terms of "real-time"and "big data", they are "context-dependent". Smith's five phases of the process as follows, a) Data distillation b) Model development c) validation and deployment d) Real-time scoring and finally e) Model refresh.

A. Data Distillation

In this phase, the contents of unstructured data files like audio, video, text documents are extracted features to structure data with usage of disparate data sources, and outcomes for modelling, and exporting sets of detailed to a local data structure, the detailed process is given in figure1.

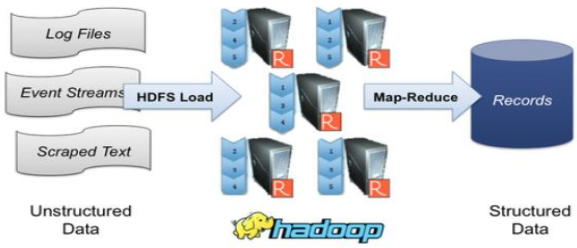


Figure 1:Big-Data Analysis from David Smith

B. Model Development

In this phase, the main goal is to create a predictive model which typically constructs, refine and compare dozens of models. It would be more powerful, strong, comprehensive and implementable. The processes which include feature selection,sampling and aggregation; variable transformation; model estimation; model refinement and model benchmarking.

C. Validation and deployment

In this phase, make sure that all the processes are working well with real-world entities. The term validation refers that re-extracting fresh data, running against the model, getting outcomes with good results would be deployed into the production department.

D. Real-time scoring

In this phase, the actions are performed on the basis of making decisions. Actual communication or integration made through the brokered systems. The deployed scoring rules aredivorced from the data in a data structure mart.

E. Model Refresh

Data always need to change to the world for different purposes, should refresh often the original data and build re-used data or modified data for analysis. For example, RevoDeployR is a web-based automated revolution service.

IV. BIG DATA TECHNIQUES

The big data foundation is composed of two major systems, one is data storing techniques and other is data processing techniques.

Big Data storage is often synonymously interchanged with the Hadoop Distributed File System, but traditional data warehouses can also house Big Data. HDFS is a distributed data storage that has become the de facto- standard because, we can store any type of data without limitations on the type or amount of data. One of the reasons HDFS has become so popular is that we don't have to any "set up" to store the data. We have to understand the data that will be housed in the database and set up by schemas. Another reason HDFS is used to complement traditional data warehouses are due to limitations on the type of data the database supports the size of data that the database can store. Additionaly, there is a wide – range of variance in how database software exploits the underlying network architecture. Datawarehouse appliances have varying degrees of performnaces advantages due to its tight coupling.

Today the de facto processing software for HDFS is MapReduce, which defined as a fault-tolerant parallel programming framework that was desined to harness distributed processing capabilities. MapReduce automatically divides the processing workload into smaller workloads that are distributed.

When working with HDFS, data manipulation and calculations are programmed using the high level programming for example Java.

There are few different typical workflows that become processing bottlenecks as data gets increasingly larger. The first one is the spped of ingesting or locating the data. The second one is the speed of analytics computational processing and the third is speed at which we can perform on analytics on demand and respond to business.

A. Big Data Computing Platforms

While there were many different techniques available today, parallel computing platforms are the only platforms suits to handle the speed and volume of data being produced today.There are three prominent parallel computing options available today:

1. Clusters or grids
2. Massively Parallel Processing(MPP)
3. High Performance Computing(HPC)

Clusters and grids are types of computing where servers are loosely coupled and networked together for distributing workloads. Massively parallel processing platforms are essentially a grid inside of a box, it combines storage, memory and compute to create platforms. Examples are EMC Greenplum and ParAccel. High Performance Computing environments are designed for high speed floating point processing and much of the calculations are done in memory, which renders the highest computational performance possible. Examples are Cray computers and IBM Blue Gene.

B. Big Data Limitations

Typical Big Data, computing platform limitations encountered are,

1. Disk bound
2. I/O bound
3. Memory bound
4. CPU bound

Disk bound means not enough memory storage to do the processing in our system. For example, having 2PB of data and there is only 500TB of storage capacity. I/O bound not having enough bandwidth to move around the data to meet the needs of the business. Memory is limited on various computing platforms because memory is still expensive. For example, lets us had 500TBs of data and we are running an analytic across 100TBs. As the analytic computing platform, it is trying to shove 100TBs into memory do to the necessary crunching. The analytic process may get to 90TBs and then become bounded by the available memory on the computing platform, which means the analytic process can't perform its task. Similarly, some analytics are more computationally intensive, which means that they use of the capacity of the CPU. For example, Monte Carlo simulation, when executing more calculation with different staring conditions or parameters. In this huge “what if ”process, there is more computational intensive calculations being performed concurrently that can simply overwhelm and exhaust the CPU capacity.

V. MODERN TECHNIQUE IN BIG DATA

NoSQL and Hadoop Distributed File System is scaled open source clustered file system that extended a global namespace, distributed front end,i.e., hundreds of petabytes with no problems. It allows software-based data security to the cluster and removes single –point of failure issue.The advantages of this software system is more scalabale and available to manage unstructured data. Also added advantage on high performance value hardware, automatic data reproduction across three nodes carries off the need for backup, then write once and read many times, dynamic and flexible schema design structure, simple programming model with simple queries to get access to data such s Java Application Programming Interface(API).

CONCLUSION

In this paper, we highlighted significance of big data types, data tools and techniques deeply and added with present trend big data in the business world. Big data techniques is radically changing the way data is produced, processes, analysed and consumed. On the other hand technology helps evolve new and more effective data sources. On the other, as more and more data gets captured, technology steps in to help process this data quickly, efficiently and visualize it to drive informed decisions. Now, more than any other time in the short history of analytics, technology plays an increasingly pivotal role in the entire process of how we gather and use data.

REFERENCES

[1] D. D. Fasel. (2012). Potentials of Big Data for Governmental Services.

[2] D. Gonzalez, "Chapter 5 - Big Data," in Managing Online Risk, D. Gonzalez, Ed., ed Boston: Butterworth-Heinemann, 2015, pp. 101-125.

[3] A. Katal, M. Wazid, and R. H. Goudar, "Big data: Issues, challenges, tools and Good practices," in Contemporary Computing (IC3), 2013 Sixth International Conference on, 2013, pp. 404-409.

[4] K. D. C. A.B. Manwar; Hemant S.Mahalle, Dr.Vinay Chavan. (2012). A VECTOR SPACE MODEL FOR INFORMATION RETRIEVAL: A MATLAB APPROACH.

[5] B. Saha and D. Srivastava, "Data quality: The other face of Big Data," in Data Engineering (ICDE), 2014 IEEE 30th International Conference on, 2014, pp. 1294-1297.

[6] F. Tekiner and J. A. Keane, "Big Data Framework," in Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, 2013, pp. 1494-1499.

[7] D. Loshin, "Chapter 7 - Big Data Tools and Techniques," in Big Data Analytics, D. Loshin, Ed., ed Boston: Morgan Kaufmann, 2013, pp. 61-72.

[8] Z. Du, "Inconsistencies in big data," in Cognitive Informatics & Cognitive Computing (ICCI*CC), 2013 12th IEEE International Conference on, 2013, pp. 61-67.

[9] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of "big data" on cloud computing: Review and open research issues," Information Systems, vol. 47, pp. 98-115, 1// 2015.

★★★