Big Data Analytics: Challenges, Tools

Ms. Ashwini Mandale, Prof.Shriniwas Gadage

*Abstract***—**

The big data have various challenges like heterogeneity, scale, timeliness, complexity, privacy problem. This paper addresses these challenges. As Data is being collected at huge amount of scale, in a broad range of application areas. Previously decisions were based on guesswork, or on painstakingly constructed models of reality but now they made based on the data itself. Nowadays Big Data analysis drives every aspect of our modern the world, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences. To take full advantage of visual analytics, organizations will need to address various challenges related to visualization. This paper addressing is done about various big data analytics opportunities, challenges and tools.

***Index Terms*— Analytics, Apache, Big data, heterogeneity, SQL,**

I. Introduction

The assurance of data-driven decision-making is being recognized broadly, and there is growing enthusiasm for the notion of Big Data. The term “Big Data” is about the use of skills to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. The platform, tools and software used for big data analytics purpose are collectively called “Big Data technologies”.

The various problems related to big data like heterogeneity, scale, timeliness, complexity and privacy prevent progress at all phases of the pipeline that can create value from data. There are various other foundational challenges like data organization, modeling, analysis and retrieval.

Finding relevant data distributed across multiple databases, database tables and/or files were very time consuming. Organizations often lacked sufficient documentation or search capabilities to enable efficient identification of desired data. System experts asked queries to database administrators or others [9].

II. LITERATURE SURVEY

Solutions for handling Big Data:

a. The availability of Cloud based solutions has dramatically lowered the cost of storage, amplified by the use of commodity hardware. Virtual file systems, either open source or vendor specific, helped transition from a managed infrastructure to a service based approach;

b. When dealing with large volumes of data, it is necessary to distribute data and workload over many servers. New designs for databases and efficient ways to support massively parallel processing have led to a new generation of products like the so called noSQL databases and the Hadoop map-reduce platform.

A. Opportunities with big data:-

We are awash in a flood of data today.

Scientific research

It has been revolutionized by Big Data [1]. The Sloan Digital Sky Survey has today become a central resource for astronomers the world over. Now a day’s astronomer’s job to one where the pictures are all in a database already and the astronomer’s task is to find interesting objects and phenomena in the database.

Education

In the world of education there is access to a huge database and collection of every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. Nowadays there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about student’s performance.

Other fields

Urban planning (through fusion of high-fidelity geographical data), intelligent transportation (through analysis and visualization of live and detailed road network data)[2], environmental modeling (through sensor networks ubiquitously collecting data) [4], energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative , computational social sciences 2 a new methodology fast growing in popularity because of the dramatically lowered cost of obtaining data) , financial systemic risk analysis (through integrated analysis of a web of contracts to find

Manuscript received April 8th, 2015.

Ms. Ashwini Mandale, ME Computer science and engineering Student, Savitribai Phule Pune University, G. H. Raisoni College of Engg & Management Wagholi,Pune, 8407975547,

Prof.Shriniwas Gadage, Manager Alkonsys, Adj.faculty Computer Engg, Savitribai Phule Pune University ,G.H.Raisoni College of Engg and Management Wagholi, Pune.

dependencies between financial entities) , homeland security (through analysis of social networks and financial transactions of possible terrorists), computer security (through analysis of logged information and other events, known as Security Information and Event Management (SIEM)), Two main sources of health big data are genomics-driven big data (genotyping, gene expression, sequencing data) and payer–provider big data (electronic health records, insurance records, pharmacy prescription, patient feedback and responses)[7] and so on.

B.Phases in the Processing Pipeline:-

The analysis of Big Data involves multiple distinct phases [4] as shown in the figure below, each of which introduces challenges.

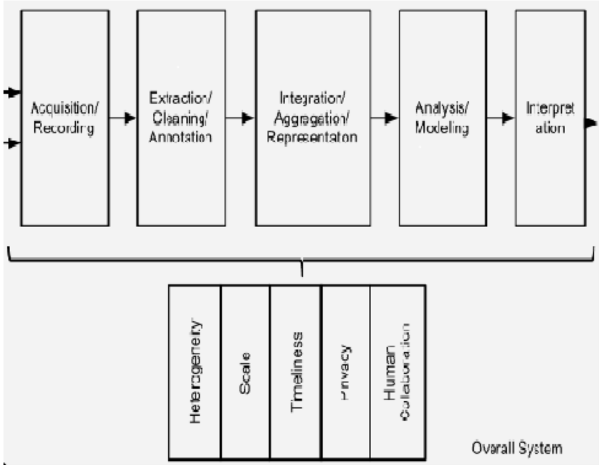

Fig.1 big data analysis phases.

a. Data Acquisition and Recording

Big Data is recorded from some data generating source. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information. The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured.

b. Information Extraction and Cleaning

An information extraction process pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis.

c. Data Integration, Aggregation, and Representation

The professionals (domain scientists) must create effective database designs [3], either through devising tools to assist them in the design process or through forgoing the design process completely and developing techniques so that databases can be used effectively in the absence of intelligent database design.

d. Query Processing, Data Modeling, and Analysis

Methods for querying and mining Big Data (noisy, dynamic, heterogeneous, inter-related and untrustworthy) are fundamentally different from traditional statistical analysis. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments. Data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions. Big Data is also enabling the next generation of interactive data analysis with real-time answers. A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses. A tight coupling between declarative query languages and the functions of such packages will benefit both expressiveness and performance of the analysis.

e. Interpretation

A decision-maker, provided with the result of analysis, has to interpret these results. Interpretation involves examining all the assumptions made and retracing the analysis.

Big Data and Advanced Analytics

The term big data has come to be used to describe multi-terabytes of data sets Social media channels, websites; automatic censors at the workplace and robotics are producing a plethora of structured, unstructured and semi-structured data. Advanced analytics based on big data is the art of putting all these fragmented and often disconnected pieces together and generate actionable insights for the enterprise.

KloudData understands the new ways in which enterprises interact with their customers and business partners and appreciates the need to adjust business analytics to include the variety of formats in which data is being generated and captured across multiple channels today. KloudData's big data offering is designed to accelerate the development and adoption of big data analytics solutions that leverage from extended platforms like Sybase IQ and Hadoop.

KloudData aims to provide customers with an efficient and cost-effective means for storing, mining, extracting, analyzing and presenting big data in highly consumable formats to empower business users to take better informed decisions. Our objective is to present highly available and scalable analytics solutions to transform decision making at the fastest moving enterprises of today.

C. Challenges in Big Data Analysis:-

a. Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. But the machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis.e.g. A patient who has multiple medical procedures at a hospital. One could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. Computer systems work most efficiently if they can store multiple items that are all identical in size and structure.

Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis.

b.Scale

Nowadays modification happening is that data volume is scaling faster than compute resources, and CPU speeds are static. Parallel data processing techniques that were applied in the past for processing data across nodes don't directly apply for intra-node parallelism, since the architecture looks very different as there are many more hardware resources such as processor, caches and processor memory channels that are shared across cores in a single node. This needs to rethink how to we design, build and operate data processing components.

The second dramatic modification happening is that the move towards cloud computing[5], which now aggregates multiple disparate workloads with varying performance goals (e.g. interactive services demand that the data processing engine return back an answer within a fixed response time cap) into very large clusters. This level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger clusters (that are required to deal with the rapid growth in data volumes).

A third modification happening is that the transformative change of the traditional I/O subsystem. HDDs (store persistent data and had slower random I/O performance) are increasingly being replaced by solid state drives, Phase Change Memory are. These newer storage technologies do not have the same large spread in performance between the sequential and random I/O performance, which requires a rethinking of how to design storage subsystems for data processing systems.Fig.2 shows various challenges in big data analytics

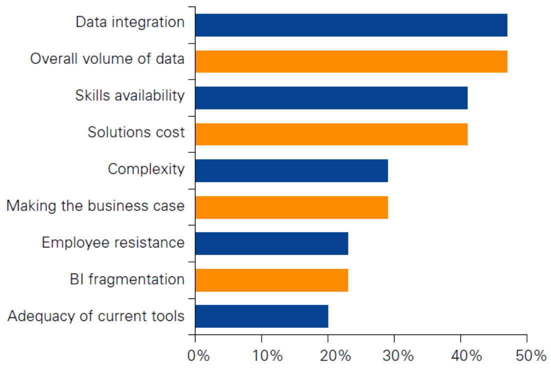

Fig.2 Biggest Challenges for Success in Big Data and Analytics. Source: TM Forum, 2012.

c.Timeliness

The larger the data slower the speed to analyze. Some situations are there where result of the analysis is required immediately.e.g. If a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed For Scanning the entire large data set in data analysis for finding suitable elements is obviously impractical. Index structure is there but it supports only some classes criteria. With new analyses desired using Big Data, there are new types of criteria specified, and a need to devise new index structures to support such criteria. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

d.Privacy

There is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

There are many additional challenging research problems.e.g. one do not know yet how to share private data while limiting disclosure and ensuring sufficient data utility in the shared data, security for information sharing in Big Data use cases.

e.Human Collaboration

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding.e.g.CAPTCHA.

A Big Data analysis system must support input from multiple human experts (separated in space and time) and shared exploration of results. A popular new method of harnessing human ingenuity to solve problems is through crowd-sourcing (Wikipedia, the online encyclopedia). The extra challenge here is the inherent uncertainty of the data collection devices [6]. The fact that collected data are probably spatially and temporally correlated can be exploited to better assess their correctness.

f. Skills availability

Big Data is being harnessed with new tools and is being looked at in different ways. There a shortage of people with the skills to bring together the data, analyze it and publish the results or conclusions [7].

g. Analytics Architecture

It is not clear yet how an optimal architecture of an analytics system should be to deal with historic data and with real-time data at the same time. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, and extensible, allows ad hoc queries, minimal maintenance, and debuggable.

h. Statistical significance.

It is important to achieve significant statistical results, and not be fooled by randomness; it is easy to go wrong with huge data sets and thousands of questions to answer at once.

i. Distributed mining

Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.

j. Time evolving data.

Data may be evolving over time, so it is important that the Big Data mining techniques should be able to adapt and in some cases to detect change first.

k. Compression

Dealing with Big Data, the quantity of space needed to store it is very relevant. There are two main approaches: compression where doesn’t loose anything or sampling where choose what is the data that is more representative. Using compression, may take more time and less space, so one can consider it as a transformation from time to space. Using sampling loosing information, but the gains in space may be in orders of magnitude.

l. Visualization.

A main task of Big Data analysis is how to visualize the results. As the data is so big, it is very difficult to end user-friendly visualizations. New techniques and frameworks to tell and show stories will be needed.

m.Hidden Big Data

Large quantities of useful data are getting lost since new data is largely untagged and unstructured data. The 2012 IDC study on Big Data explains that in 2012, 23% (643 Exabyte) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

D. Tools: Open Source Revolution

The Big Data phenomenon is intrinsically related to the open source software revolution. Large companies as Facebook, Yahoo!, Twitter, LinkedIn benefitt and contribute working on open source projects. Big Data infrastructure deals with Hadoop, and other related software as:

a. Apache Hadoop

It is the software for data-intensive distributed applications, based in the MapReduce programming model and a distributed file system called Hadoop Distributed Filesystem (HDFS). Hadoop allows writing applications that rapidly process large amounts of data in parallel on large clusters of compute nodes. A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job.

b. Apache Hadoop related projects [10]

Apache Pig, Apache Hive, Apache HBase, Apache ZooKeeper, Apache Cassandra, Cascading, Scribe and many others.

c. Apache S4 [11]

It is the platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time.

d. Storm

It is the software for streaming data-intensive distributed applications, similar to S4, and developed by Nathan Marz at Twitter.

III. Conclusion

In the begin of this paper we discuss various opportunities with big data how one can use it. In this paper many technical challenges described such as scale, heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. Big data needs support and encourage fundamental research towards addressing these technical challenges to achieve the promised benefits of Big Data.

Acknowledgment

I would like to thanks to Prof.Shriniwas Gadage, Manager, Alkonsys, Adj.faculty Computer Engg-G.H.Raisoni College of Engg and Management, Pune for his help, Encouragement and intellectual influence, which made this paper possible. His invaluable guidance in the successful completion of this paper work.

References

[1] Ji, Changqing, et al. "Big data processing in cloud computing environments." Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on. IEEE, 2012.

[2] Keim, Daniel, Huamin Qu, and Kwan-Liu Ma. "Big-Data Visualization." Computer Graphics and Applications, IEEE 33.4 (2013): 20-21.

[3] Madden, Sam. From databases to big data. Internet Computing, IEEE 16.3 (2012): 4-6.

[4] Ferguson, Mike. "Architecting a Big Data Platform for Analytics." A Whitepaper Prepared for IBM (2012).

[5] Shekhar, Shashi, et al. Spatial big-data challenges intersecting mobility and cloud computing. Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access. ACM, 2012.

[6] Chittaranjan, Gokul, Jan Blom, and Daniel Gatica-Perez. "Mining large-scale smartphone data for personality studies." Personal and Ubiquitous Computing 17.3 (2013): 433-450.

[7] “Big Data Analytics in Biomedical Research,” *Biomedical Computation Review* (August 2, 2012).

[8] “The Top Challenges in Big Data and Analytics “by lavastorm analytics.

[9] Enterprise Data Analysis and Visualization: An Interview Study, Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer.

[10] P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis. IBM Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data.McGraw-Hill Companies, Incorporated, 2011.
[11] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari.S4: Distributed Stream Computing Platform. In ICDM Workshops, pages 170{177, 2010.
[]2] Big Data Analytics _ Big Data Applications _ Big Data Management.html http://www.nytimes.com/
[13]Apache Hadoop, http://hadoop.apache.org.

Ms. Ashwini Mandale, Ms. Ashwini Mandale, ME Computer science and engineering Student, Savitribai Phule Pune University,G.H.Raisoni College of Engg & Management Wagholi,Pune.Her area of interest is data mining, cloud security, big data analytics.

Prof.Shriniwas Gadage, Manager, Alkonsys, Adj.faculty Computer Engg, Savitribai Phule Pune University ,G.H.Raisoni College of Engg and Management, Pune. He has been completed post-graduation He guided various post graduate projects. He published various papers in international journals. His area of interest is data mining, cloud computing, image processing, cloud computing.