International Journal of Information Technology and Management

Vol. IX, Issue No. XIV, November-2015, ISSN 2249-4510

AN ANALYSIS ON VARIOUS TRENDS OF BIG DATA COMPUTING AND CLOUDS: ISSUES AND CHALLENGES

AN INTERNATIONALLY INDEXED PEER REVIEWED & REFEREED JOURNAL

www.ignited.in

International Journal of Information Technology and Management
Vol. IX, Issue No. XIV, November-2015, ISSN 2249-4510

An Analysis on Various Trends of Big Data Computing and Clouds: Issues and Challenges

Charles R.[1] Dr. P. Selva Kumar[2]

[1]Research Scholar, Madurai Kamaraj University, Tamilnadu

Abstract – This paper discusses approaches and environments for carrying out analytics on Clouds for Big Data applications. It revolves around four important areas of analytics and Big Data, namely (i) data management and supporting architectures; (ii) model development and scoring; (iii) visualization and user interaction; and (iv) business models.

The amount of data that is traveling across the internet today, not only that is large, but is complex as well. Companies, institutions, healthcare system etc., all of them use piles of data which are further used for creating reports in order to ensure continuity regarding the services that they have to offer. The process behind the results that these entities requests represents a challenge for software developers and companies that provide IT infrastructure. The challenge is how to manipulate an impressive volume of data that has to be securely delivered through the internet and reach its destination intact. This paper treats the challenges that Big Data creates.

- - - - - - - - - - - - - - X - - - - - - - - - - - - - -

INTRODUCTION

Data is the collection of values and variables related in some sense and differing in some other sense. In recent years the sizes of databases have increased rapidly. This has led to a growing interest in the development of tools capable in the automatic extraction of knowledge from data. Data are collected and analyzed to create information suitable for making decisions. Hence data provide a rich resource for knowledge discovery and decision support. A database is an organized collection of data so that it can easily be accessed, managed, and updated. Data mining is the process discovering interesting knowledge such as associations, patterns, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses or other information repositories. A widely accepted formal definition of data mining is given subsequently. According to this definition, data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data . Data mining uncovers interesting patterns and relationships hidden in a large volume of raw data. Big Data is a new term used to identify the datasets that are of large size and have grater complexity . So we cannot store, manage and analyze them with our current methodologies or data mining software tools. Big data is a heterogeneous collection of both structured and unstructured data. Businesses are mainly concerned with managing unstructured data. Big Data mining is the capability of extracting useful information from these large datasets or streams of data which were not possible before due to its volume, variety, and velocity.

The extracted knowledge is very useful and the mined knowledge is the representation of different types of patterns and each pattern corresponds to knowledge. Data Mining is analyzing the data from different perspectives and summarizing it into useful information that can be used for business solutions and predicting the future trends. Mining the information helps organizations to make knowledge driven decisions. Data mining (DM), also called Knowledge Discovery in Databases (KDD) or Knowledge Discovery and Data Mining, is the process of searching large volumes of data automatically for patterns such as association rules . It applies many computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining extract only required patterns from the database in a short time span. Based on the type of patterns to be mined, data mining tasks can be classified into summarization, classification, clustering, association and trends analysis.

Enormous amount of data are generated every minute. A recent study estimated that every minute, Google receives over 4 million queries, e-mail users send over 200 million messages, YouTube users upload 72 hours of video, Facebook users share over 2 million pieces of content, and Twitter users generate 277,000 tweets . With the amount of data growing exponentially, improved analysis is required to extract information that best matches user interests. Big data refers to rapidly growing datasets with sizes beyond the capability of traditional data base tools to store, manage and analyse them. Big data is a heterogeneous collection of both structured and unstructured data. Increase of storage capacities, Increase of processing power and availability of data are the main reason for the appearance and growth of big data. Big data refers to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients in decision making. The data may be enterprise specific or general and private or public. Big data are characterized by 3 V's: Volume, Velocity, and Variety.

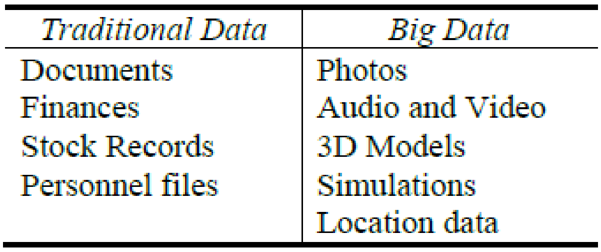
Big Data mining refers to the activity of going through big data sets to look for relevant information. Big data samples are available in astronomy, atmospheric science, social networking sites, life sciences, medical science, government data, natural disaster and resource management, web logs, mobile phones, sensor networks, scientific research, telecommunications. Two main goals of high dimensional data analysis are to develop effective methods that can accurately predict the future observations and at the same time to gain insight into the relationship between the features and response for scientific purposes. Big data have applications in many fields such as Business, Technology, Health, Smart cities etc. These applications will allow people to have better services, better customer experiences, and also to prevent and detect illness much easier than before.

The rapid development of Internet and mobile technologies has an important role in the growth of data creation and storage. Since the amount of data is growing exponentially, improved analysis of large data sets is required to extract information that best matches user interests. New technologies are required to store unstructured large data sets and processing methods such as Hadoop and Map Reduce have greater importance in big data analysis. To process large volumes of data from different sources quickly, Hadoop is used. Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It allows running applications on systems with thousands of nodes with thousands of terabytes of data. Its distributed file system supports fast data transfer rates among nodes and allows the system to continue operating uninterrupted at times of node failure. It runs Map Reduce for distributed data processing and is works with structured and unstructured data.

Economic entities and not only, had developed over the years new and more complex methods that allows them to see market evolution, their position on the market, the efficiency of offering their services and/or products etc. For being able to accomplish that, a huge volume of data is needed in order to be mined so that can generate valuable insights.

Every year the data transmitted over the internet is growing exponentially. By the end of 2016, Cisco estimates that the annual global data traffic will reach 6.6 zettabytes. The challenge will be not only to "speed up" the internet connections, but also to develop software systems that will be able to handle large data requests in optimal time.

To have a better understanding of what Big Data means, the table below represents a comparison between traditional data and Big Data (Table 1. Understanding Big Data).



Traditional Data	*Big Data*
Documents	Photos
Finances	Audio and Video
Stock Records	3D Models
Personnel files	Simulations
	Location data

Table 1. Understanding Big Data.

This example provides information about the volume and the variety of Big Data. It is difficult to work with complex information on standard database systems or on personal computers. Usually it takes parallel software systems and infrastructure that can handle the process of sorting the amount of information that, for example, meteorologists need to analyze.

The request for more complex information is getting higher every year. Streaming information in real-time is becoming a challenge that must be overcome by those companies that provides such services, in order to maintain their position on the market.

By collecting data in a digital form, companies take their development to a new level. Analyzing digital data can speed the process of planning and also can reveal patterns that can be further used in order to improve strategies. Receiving information in real-time about customer needs is useful for seeing market trends and forecasting.

The expression "Big Data" also resides in the way that information is handled. For processing large quantities of data that is extremely complex and various there needs to be a set of tools that are able to navigate through it and sort it. The methods of sorting data differ from one type of data to another. Regarding Big Data, where the type of data is not singular, sorting is a multi-level process.

Big Data can be used for predictive analytics, an element that many companies rely on when it comes to see where they are heading. For example, a telecommunication company can use data stored from length of call, average text messages sent, average bill amount to see which customers are likely to discard their services.

CLOUD AND BIG DATA

Cloud delivery models offer exceptional flexibility, enabling IT to evaluate the best approach to each business user's request. For example, organizations that already support an internal private cloud environment can add big data analytics to their in-house offerings, use a cloud services provider, or build a hybrid cloud that protects certain sensitive data in a private cloud, but takes advantage of valuable external data sources and applications provided in public clouds.

Using cloud infrastructure to analyze big data makes sense because:

1. Investments in big data analysis can be significant and drive a need for efficient, cost-effective infrastructure. The resources to support distributed computing models in-house typically reside in large and midsize data centers. Private clouds can offer a more efficient, cost-effective model to implement analysis of big data in-house, while augmenting internal resources with public cloud services. This hybrid cloud option enables companies to use on-demand storage space and computing power via public cloud services for certain analytics initiatives (for example, short-term projects), and provide added capacity and scale as needed.

2. Big data may mix internal and external sources. While enterprises often keep their most sensitive data in-house, huge volumes of big data (owned by the organization or generated by third-party and public providers) may be located externally—some of it already in a cloud environment. Moving relevant data sources behind your firewall can be a significant commitment of resources. Analyzing the data where it resides—either in internal or public cloud data centers or in edge systems and client devices—often makes more sense.

3. Data services are needed to extract value from big data. Depending on requirements and the usage scenario, the best use of your IT budget may be to focus on analytics as a service (AaaS)—supported by your internal private cloud, a public cloud, or a hybrid model.

Cloud computing models can help accelerate the potential for scalable analytics solutions. Clouds offer flexibility and efficiencies for accessing data, delivering insights, and driving value. However, cloud-based big data analytics is not a one-size-fits-all solution.

Organizations using cloud infrastructure to provide AaaS have multiple options. By weighing factors of workload, cost, security, and data interoperability, IT can choose to utilize their private cloud to mitigate risk and maintain control; use public cloud infrastructure, platform, or analytics services to further enhance scalability; or implement a hybrid model that combines private and public cloud resources and services.

The bottom line: No matter which cloud delivery model makes the most sense, businesses with varying needs and budgets can unlock the potential of big data in cloud environments.

BIG DATA TRENDS

What makes cloud computing such a cost-effective delivery model for big data analytics? How are big data and cloud technologies converging to make big data analytics in clouds a reasonable option? For big data analytics:

Data is becoming more valuable. Today the conversation is shifting from "What data should we store?" to "What can we do with the data?" Enterprises are looking to unlock data's hidden potential and deliver competitive advantage. Gartner predicts that enterprise data will grow by 800 percent from 2011 to 2015, with 80 percent unstructured (for example, e-mails, documents, video, images, and social media content) and 20 percent structured (for example, credit card transactions and contact information).

With the potential for so much data to reveal insights that can boost competitiveness, companies must find new approaches to processing, managing, and analyzing their data—whether it's structured data typically found in traditional relational database management systems (RDBMSs) or more varied, unstructured formats. Plus, combining diverse data sources and types has the potential to uncover some of the most interesting unexplored patterns and relationships.

Data analytics is moving from batch to real time. Intel's 2012 survey of 200 IT managers in large enterprises found that while the amount of batch versus real-time processing is split evenly today, the trend is toward increasing real time to two-thirds of total data management by 2015.2 At the same time, the technology for processing real-time or near-real-time information is moving past hype to early stages of maturity.

Real time supports predictive analytics. Predictive analytics enables organizations to move to a future-oriented view of what's ahead and of fers 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.

The scope of big data analytics continues to expand. Early interest in big data analytics focused primarily on business and social data sources, such as e-mail, videos, tweets, Facebook* posts, reviews, and Web behavior. The scope of interest in big data analytics is growing to include data from intelligent systems, such as in-vehicle infotainment, kiosks, smart meters, and many others, and device sensors at the edge of networks—some of the largest-volume, fastest-streaming, and most complex big data. Ubiquitous connectivity and the growth of sensors and intelligent systems have opened up a whole new storehouse of valuable information. Interest in applying big data analytics to data from sensors and intelligent systems continues to increase as businesses seek to gain faster, richer insight more cost-effectively than in the past, enhance machine-based decision making, and personalize customer experiences.

ISSUES AND CHALLENGES

Big data analysis is the process of applying advanced analytics and visualization techniques to large data sets to uncover hidden patterns and unknown correlations for effective decision making. The analysis of Big Data involves multiple distinct phases which include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modeling and analysis and Interpretation. Each of these phases introduces challenges. Heterogeneity, scale, timeliness, complexity and privacy are certain challenges of big data mining.

Heterogeneity and Incompleteness-

The difficulties of big data analysis derive from its large scale as well as the presence of mixed data based on different patterns or rules (heterogeneous mixture data) in the collected and stored data. In the case of complicated heterogeneous mixture data, the data has several patterns and rules and the properties of the patterns vary greatly. Data can be both structured and unstructured.

80% of the data generated by organizations are unstructured. They are highly dynamic and does not have particular format. It may exists in the form of email attachments, images, pdf documents, medical records, X rays, voice mails, graphics, video, audio etc. and they cannot be stored in row/ column format as structured data. Transforming this data to structured format for later analysis is a major challenge in big data mining. So new technologies have to be adopted for dealing with such data.

Incomplete data creates uncertainties during data analysis and it must be managed during data analysis. Doing this correctly is also a challenge. Incomplete data refers to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values. While most modern data mining algorithms have inbuilt solutions to handle missing values (such as ignoring data fields with missing values), data imputation is an established research field which seeks to impute missing values in order to produce improved models (compared to the ones built from the original data). Many imputation methods exist for this purpose, and the major approaches are to fill most frequently observed values or to build learning models to predict possible values for each data field, based on the observed values of a given instance.

Scale and complexity-

Managing large and rapidly increasing volumes of data is a challenging issue. Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, organization, retrieval and modeling are also challenges due to scalability and complexity of data that needs to be analysed.

Timeliness-

As the size of the data sets to be processed increases, it will take more time to analyse. In some situations results of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed by preventing the transaction from taking place at all. Obviously a full analysis of a user’s purchase history is not likely to be feasible in real time. So we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. In such cases Index structures are created in advance to permit finding qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criteria.

TECHNIQUES FOR BIG DATA MINING

Big data has great potential to produce useful information for companies which can benefit the way they manage their problems. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules. These massive data sets are too large and complex for humans to effectively extract useful information without the aid of computational tools. Emerging technologies such as the Hadoop framework and MapReduce offer new and exciting ways to process and transform big data, defined as complex, unstructured, or large amounts of data, into meaningful knowledge.

Hadoop - Hadoop is a scalable, open source, fault tolerant Virtual Grid operating system architecture for data storage and processing. It runs on commodity hardware, it uses HDFS which is fault-tolerant high bandwidth clustered storage architecture. It runs MapReduce for distributed data processing and is works with structured and unstructured data . For handling the velocity and heterogeneity of data, tools like Hive, Pig and Mahout are used which are parts of Hadoop and HDFS framework. Hadoop and HDFS (Hadoop Distributed File System) by Apache is widely used for storing and managing big data.

Hadoop consists of distributed file system, data storage and analytics platforms and a layer that handles parallel computation, rate of flow (workflow) and configuration administration.

HDFS runs across the nodes in a Hadoop cluster and together connects the file systems on many input and output data nodes to make them into one big file system. The present Hadoop ecosystem, as shown in Figure 1, consists of the Hadoop kernel, MapReduce, the Hadoop distributed file system (HDFS) and a number of related components such as Apache Hive, HBase, Oozie, Pig and Zookeeper and these components are explained as below :

- HDFS: A highly faults tolerant distributed file system that is responsible for storing data on the clusters.

- MapReduce: A powerful parallel programming technique for distributed processing of vast amount of dataon clusters.

- HBase: A column oriented distributed NoSQL database for random read/write access.

- Pig: A high level data programming language for analyzing data of Hadoop computation.

- Hive: A data warehousing application that provides a SQL like access and relational model.

- Sqoop: A project for transferring/importing data between relational databases and Hadoop.

- Oozie: An orchestration and workflow management for dependent Hadoop jobs.

Figure 2 gives an overview of the Big Data analysis tools which are used for efficient and precise data analysis and management jobs. The Big Data Analysis and management setup can be understood through the layered structured defined in the figure. The data storage part is dominated by the HDFS distributed file system architecture and other architectures available are Amazon Web Service, Hbase and CloudStore etc. The data processing tasks for all the tools is Map Reduce and it is the Data processing tool which effectively used in the Big Data Analysis.

For handling the velocity and heterogeneity of data, tools like Hive, Pig and Mahout are used which are parts of Hadoop and HDFS framework. It is interesting to note that for all the tools used, Hadoop over HDFS is the underlying architecture. Oozie and EMR with Flume and Zookeeper are used for handling the volume and veracity of data, which are standard Big Data management tools .

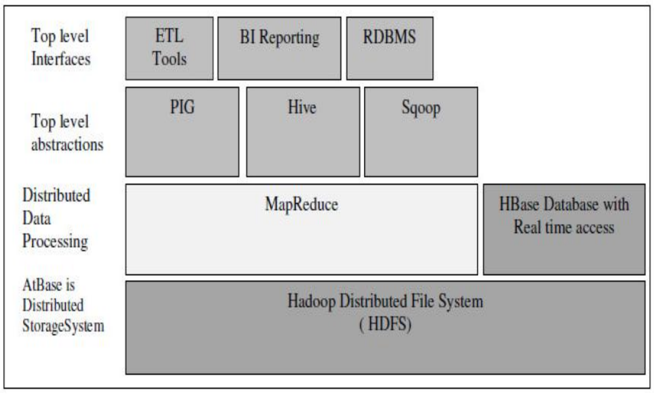

Top level Interfaces
ETL Tools
BI Reporting
RDBMS
Top level abstractions
PIG
Hive
Sqoop
Distributed Data Processing
MapReduce
HBase Database with Real time access
AtBase is Distributed StorageSystem
Hadoop Distributed File System (HDFS)

Figure 1 : Hadoop Architecture Tools.

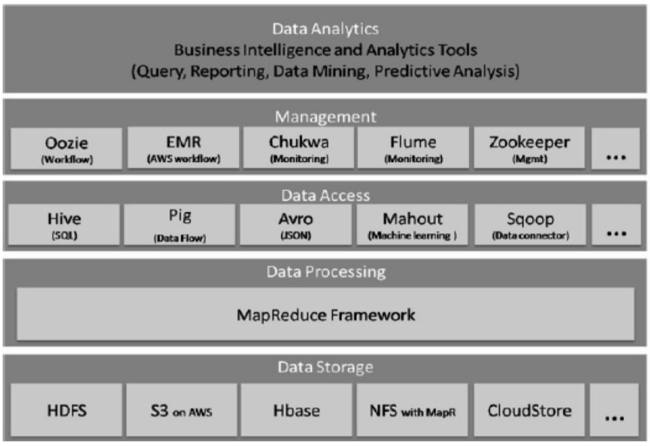


Figure 2: Big data analysis tools.

MapReduce -

MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. Hadoop MapReduce is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes.

The MapReduce consists of two functions, map() and reduce(). Mapper performs the tasks of filtering and sorting and reducer performs the tasks of summarizing the result. There may be multiple reducers to parallelize the aggregations . Users can implement their own processing logic by specifying a customized map() and reduce() function. The map() function takes an input key/value pair and produces a list of intermediate key/value pairs. The MapReduce runtime system groups together all intermediate pairs based on the intermediate keys and passes them to reduce() function for producing the final results. Map Reduce is widely used for the Analysis of big data.

METHODOLOGY

Organizations are increasingly generating large volumes of data as result of instrumented business processes, monitoring of user activity, web site tracking, sensors, finance, accounting, among other reasons. With the advent of social network Web sites, users create records of their lives by daily posting details of activities they perform, events they attend, places they visit, pictures they take, and things they enjoy and want. This data deluge is often referred to as Big Data; a term that conveys the challenges it poses on existing infrastructure in respect to storage, management, interoperability, governance, and analysis of the data.

In today's competitive market, being able to explore data to understand customer behaviour, segment customer base, offer customized services, and gain insights from data provided by multiple sources is key to competitive advantage. Although decision makers would like to base their decisions and actions on insights gained from this data, making sense of data, extracting non obvious patterns, and using these patterns to predict future behaviour are not new topics. Knowledge Discovery in Data (KDD) aims to extract non obvious information using careful and detailed analysis and interpretation. Data mining, more specifically, aims to discover previously unknown interrelations among apparently unrelated attributes of datasets by applying methods from several areas including machine learning, database systems, and statistics. Analytics comprises techniques of KDD, data mining, text mining, statistical and quantitative analysis, explanatory and predictive models, and advanced and interactive visualization to drive decisions and actions.

Figure 3 depicts the common phases of a traditional analytics workow for Big Data. Data from various sources, including databases, streams, marts, and data warehouses, are used to build models. The large volume and different types of the data can demand pre-processing tasks for integrating the data, cleaning it, and filtering it. The prepared data is used to train a model and to estimate its parameters. Once the model is estimated, it should be validated before its consumption. Normally this phase requires the use of the original input data and specific methods to validate the created model.

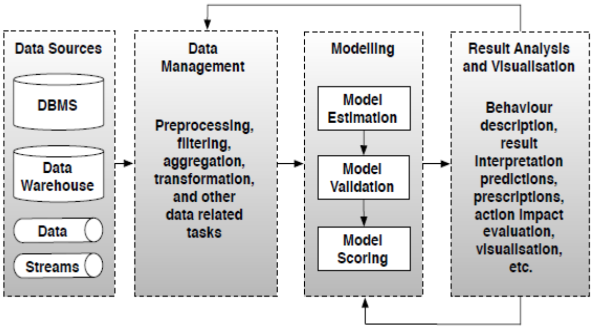

Data Sources
DBMS
Data Warehouse
Data
Streams
Data Management
Preprocessing, filtering, aggregation, transformation, and other data related tasks
Modelling
Model Estimation
Model Validation
Model Scoring
Result Analysis and Visualisation
Behaviour description, result interpretation predictions, prescriptions, action impact evaluation, visualisation, etc.

Figure 3: Overview of the analytics workow for Big Data.

Finally, the model is consumed and applied to data as it arrives. This phase, called model scoring, is used to generate predictions, prescriptions, and recommendations. The results are interpreted and evaluated, used to generate new models or calibrate existing ones, or are integrated to pre-processed data.

Analytics solutions can be classified as descriptive, predictive, or prescriptive as illustrated in Figure 4. Descriptive analytics uses historical data to identify patterns and create management reports; it is concerned with modeling past behaviour. Predictive analytics attempts to predict the future by analysing current and historical data. Prescriptive solutions assist analysts in decisions by determining actions and assessing their impact regarding business objectives, requirements, and constraints.

Despite the hype about it, using analytics is still a labour intensive endeavor. This is because current solutions for analytics are often based on proprietary appliances or software systems built for general purposes. Thus, significant effort is needed to tailor such solutions to the specific needs of the organization, which includes integrating different data sources and deploying the software on the company's hardware (or, in the case of appliances, integrating the appliance hardware with the rest of the company's systems).

Such solutions are usually developed and hosted on the customer's premises, are generally complex, and their operations can take hours to execute. Cloud computing provides an interesting model for analytics, where solutions can be hosted on the Cloud and consumed by customers in a pay-as-you-go fashion. For this delivery model to become reality, however, several technical issues must be addressed, such as data management, tuning of models, privacy, data quality, and data currency.

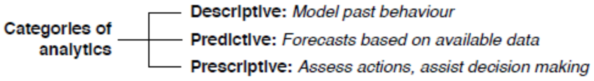

Categories of analytics
Descriptive: Model past behaviour
Predictive: Forecasts based on available data
Prescriptive: Assess actions, assist decision making

Figure 4: Categories of analytics.

This work highlights technical issues and surveys existing work on solutions to provide analytics capabilities for Big Data on the Cloud. Considering the traditional analytics workow presented in Figure 3, we focus on key issues in the phases of an analytics solution. With Big Data it is evident that many of the challenges of Cloud analytics concern data management, integration, and processing.

CONCLUSION

Building a viable solution for large and complex data is a challenge that companies in this field are continuously learning and implementing new ways to handle it. One of the biggest problems regarding Big Data is the infrastructure's high costs. Hardware equipment is very expensive for most of the companies, even if Cloud solutions are available. Each Big data system requires massive processing power and stable and complex network configurations that are made by specialists. Besides hardware infrastructure, software solutions tend to have high costs if the beneficiary doesn't opt for open source software. Even if they chose open source, to configure there is still needed specialists with the required skills to do that. The downside of open source is that maintenance and support is not provided as is the case of paid software. So, all that is necessary to maintain a Big Data solution working correctly needs, in most cases, an outside maintenance team.

We predict cloud computing will grow and with the age of BigData, the survey report proposed some of the key challenges existing in the field of cloud computing. With the existing tool and techniques it is not sufficient to adhere all the challenges relating to big volume of data. It is not feasible to provide better data quality with the existing technology and again privacy is big problem with cloud data.

The amounts of data is growing exponentially worldwide due to the explosion of social networking sites, search and retrieval engines, media sharing sites, stock trading sites, news sources and so on. Big Data is becoming the new area for scientific data research and for business applications. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules. Big data analysis helps companies to take better decisions, to predict and identify changes and to identify new opportunities. In this paper we discussed about the issues and challenges related to big data mining and also Big Data analysis tools like Map Reduce over Hadoop and HDFS which helps organizations to better understand their customers and the marketplace and to take better decisions and also helps researchers and scientists to extract useful knowledge out of Big data. In addition to that we introduce some big data mining tools and how to extract a significant knowledge from the Big Data. That will help the research scholars to choose the best mining tool for their work.

REFERENCES

- Albert Bifet, (2013), "Mining Big data in Real time", Informatica 37, pp15-20
- B. Franks, Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, 1st Edition, Wiley and SAS Business Series, Wiley, 2012.
- D. Loshin, Chapter 5 – data governance for big data analytics: considerations for data policies and processes, in: D. Loshin (Ed.), Big Data Analytics, Morgan Kaufmann, Boston, 2013, pp. 39–48.
- M. D. Assuncao, R. N. Calheiros, S. Bianchi, M.A. Netto, R. Buyya, "Big Data Computing and Clouds: Challenges,Solutions, and Future Directions," arXiv preprint arXiv:1312.4722, .2013.
- M. Schroeck, R. Shockley, J.Smart, D. Romero-Morales, P. Tufano, "Analytics: The real-world use of bigdata," in, IBM Global Business Services, 2012.
- O. Tene, J. Polonetsky, "Privacy in the age of big data: a time for big decisions," Stanford Law Review Online 64, 63, 2012.
- Peer Research: Big Data Analytics: Intel's IT Manager Survey on How Organizations Are Using Big Data. Intel IT Center (August 2012).
- Priya P. Sharma, Chandrakant P. Navdeti, (2014), " Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution", IJCSIT, 5(2), pp2126-2131
- R. Akerkar, Big Data Computing, CRC Press, 2013.
- Richa Gupta, (2014), "Journey from data mining to Web Mining to Big Data", IJCTT, 10(1),pp18-20

- Richa Gupta, Sunny Gupta, Anuradha Singhal, (2014), “Big Data:Overview”, IJCTT, 9 (5)