ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

A Study on Big-Data Approach to Data Analytics

Ishwinder Kaur Sandhu[#1], Richa Chabbra[2]

[1]M.Tech Student, Department of Computer Science and Technology, NCU University, Gurgaon, Haryana, India

[2]Assistant Professor, Department of Computer Science and Technology, NCU University, Gurgaon, Haryana, India

ABSTRACT: The internet has made new sources which produces huge amount of data that can be available to business enterprise. If business organizations want to stand in competition with other firms, it should adopt the new technology and techniques that are emerging due to big data.
Big data consists of structured, semi structured and unstructured data. Structured data is formatted data and it is used in managing databases. Semi structured and unstructured data are unformatted data and consist of all kinds of unformatted data including multimedia and social media content.
The paper consists of big data analytics – the process of analyzing, mining big data, and 1010 data.
1010 data is held by private company that provides a cloud based software and its associated services for business analytics.

KEYWORDS: Structured Data, Unstructured Data, Semi structured Data, 1010data.

I. INTRODUCTION

Big Data allude to data management over a large-scale and those analytical technologies that exceed the capability of traditional data processing. Big Data is distinguished from traditional data storage technologies in three ways: volume, velocity and variety.Analytics of big data is the process of analyzing and extracting data. It tends to produce operational knowledge at a very high scale. Need to analyze the collected data is one of the main drivers of Big data analysis tools.
Big data includes following advancements in Big data technology:(a) Decrease in the cost of storage and CPU power. (b) Flexibility in terms of data nodes for storage and computation. (c)Evolution of frameworks such as Hadoop, which allows parallel processing.
These advances tend to create dissimilarities among traditional and big data analytics.Few years ago one of the major technical issues was data storage and its scalability. However, a new technology which is rather efficient and scalable has been subsumed which tends to solve the data management and storage issue. An approximate value of 2.5 Exabyte which is equivalent to 2.5 billion gigabytes (GB) of data was produced every day according to IBM's statistics in 2012. Among the value 75% of data was in unstructured form which came through multiple sources such as text, voice and videos.Though this occurrence has speeded up and would further tend to increase the connectivity with devices. Remarkable internet competitors such as Google, Facebook, Amazon, and Twitter were first facing the increased data problems, which is now solved by designing an ad-hoc network to manage the situation. This was the initiating step of big data trend with relatively cheap solutions.Meanwhile, a leap forward has helped increase the assumption of panacea for managing big data. Cloud based results availability has further decreased the storage cost with the help of virtual hardware. It is mandatory to disperse the data and its tasks over multiple servers so as to deal with large amount of it.

Key characteristics of Big Data:-*(a)Volume:*A lot needs to be done.(b)*Velocity:*The need for speed.*(c)Variety:*Plethora of options.*(d)Value*: How much is it worth.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Advantages:-(a) Helps redevelop your product. (b) Performs risk analysis. (c) Keeps your data safe.(d) Customizes your website in real time. (e) Reduces the maintenance costs. (f) Offers customized healthcare.

Disadvantages: - (a)Deal with nearly all the elements of big data. (b)Getting right information requires special computing power.(c) Big data skills are not applicable for large supply[8].

Challenges to big data :- (a) Special computing power is required, such as, special version of Hadoop for parallel processing. (b) Use of real time insight may require different way of working within any organization, that is, its aim should be to construct an information centric organization [8].

II. REFERENCE ARCHITECTURE OF BIG DATA ANALYTICS

Bottom of the Architecture consists of Infrastructure layer which provides hardware and platform to help run the components of big data during its analysis. Due to its shared nature, it is used to support diverse synchronous practices.Infrastructure under this layer supports traditional and specialized database management system and has been optimized for analytics[1][5].

Second layer, the information layer includes components which manages the particulars incorporating data stores in addition to units to capture, move, merge, process and virtualize data. Bottom of this layer comprises of data stores for specific purposes including operational data stores, data management systems, etc. Furthermore, these stores represent the sources of data that are taken into Logical Data Warehouse (LDW) [7][10]. A collection of data is represented by LDW that has been indulged for historical and analytical purposes.Components above the Logical Data Warehouse perform functions including processing and event handling. Lastly components on the top of this layer virtualize the data for user consumption.

Third layer, the service layer comprises of components that provides general services. These services involve presentation and information services which are parts of Services oriented Architecture. These services are defined as used and are shared among different solutions. Following are the services provided by the layer (a)Monitoring Business Activity, (c)Handling multiple events and (d) Services over information.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

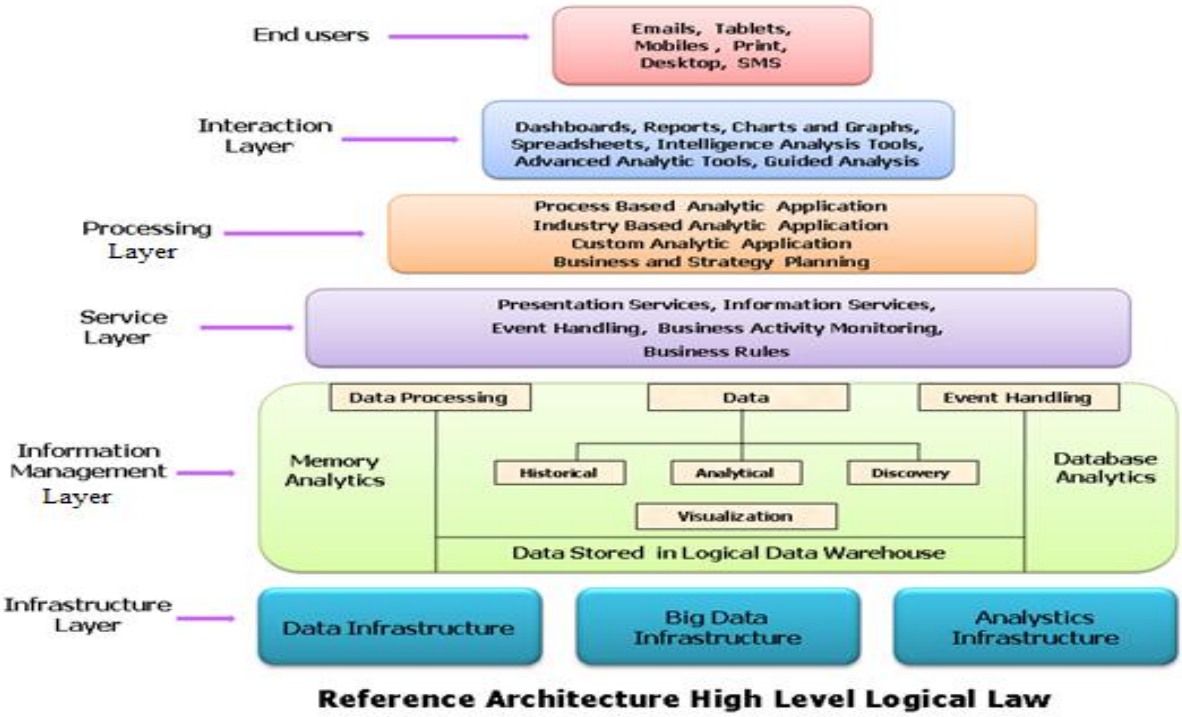

Fig.1. Reference Architecture

Process layer represents segments which performs high level activities for processing. Several applications supporting analytical, performance management processes and intelligence gathering are called by this layer for the purpose of big data analytics.
Interaction layer comprises of segments which are used for enhancing interaction with the end users. Its components include charts, graphs, spreadsheets, reports and dashboards. Addition to these properties, this layer provides tools to analyze the performance and discovery activities.

III. INTERACTIVE ANALYTICS ON BIG DATA

Approaches to analyze big data:
(a) Analysis using a high performance Analytical Database. (b) Analysis of data subsets using interactive Analytical Software. (c) An Interactive Analytical Database as provided by 1010data[4].

We illustrate each of these possibilitiesin order to compare them. There are two aspects for these possibilities. Firstly,capabilities experienced by the user. Secondly, work involved to make the capability available.

The analysis diagram shown below illustrates how different analytical databases are used. This analytical database is accessible using Structured Query Language to lodge analytical functions in addition to the data selection and manipulations. Some analytical software might be used by users that has been tailored to work along database or they can simply write the queries. Both the cases needs query to be constructed to deliver a specific answer.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

This database is built to respond in a good speed and not specifically to respond at bilateral speeds. Also after the quick response, it is not specified that the analysis tool will be communal. By resubmitting the query, user may proceed. This happens when database repeats much of the work done at that time.The user would likely to create a batch of queries that they'd like to answer and submit at once, as the interface is not interactive. At the moment theyreceive the answers, they might den concentrate some of those queries to proceed further. Though the way of processing is different, in order to support this kind of capability, the IT center would have to prepare their analytical databases and associated data fields. This process is known to be extensive which includes data tuning and modeling activities.

The second option gives user a real-time interactive interface. In here, a subset of suitable data is extracted from database or the warehouse holding large amount of it. Furthermore, the IT Department would organize data for extraction and its refreshing [6][7].

Limitation for users is that there is a limited amount of data that can be processed and analyzed in a given session. In case of analyzing larger volumes of data, they are either forced to work in a batch manner or in manner to analyze the data in a gradual way. Neglecting the large volume of data, 1010data allows the user to be interactive and demands zero effort from IT sector. It is dramatically faster for the users firstly because the elementary database has been built for speed and secondly because the interposed results are worked upon by the database. Users can easily merge manipulation activities with the analytical as the user is not limited to single query language like SQL. The result sets can be saved at instance and different methods can be applied to experiment with it. Team work can also be done in order to work collaboratively passing intermediated result sets to the other. The approachis highly flexible.

The third approach is based on the resilience of 1010data:
1010data basically runs within its own data center. Its data centers are built on highly reliable cloud infrastructure. Servers in their data centers run with RAID disks. These disks provide reliability ensuring rare server failures. In case any failure does occur, 1010data is built to briskly recover, regardless of the type of failure. Data duplication tends to high level of resilience, further reflecting data in a similar way that RAID data configuration is made across multiple disks[4]. Therefore, at the event of data failure, the same data is present for use on all other servers.

Process inside the analysis states that the current user's session state is saved at that instance, which furthers requires only the current query to be recomputed, if it was interrupted. At the time of such failure the user receives an error message which pops the user to resubmit the query my just clicking of a mouse. The present master process that was responsible for implementing slave processing on the server which failed abruptly simply fires up new slave processes on other server nodes that holds necessary data and tends to continue as before. Failure of a master process tends to lost the session and the user is asked to redo the whole process. However a save session property is provided by the software, allowing user to save the session. Therefore recovery of data will initiate from the last saved pointif the user has made use of it.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

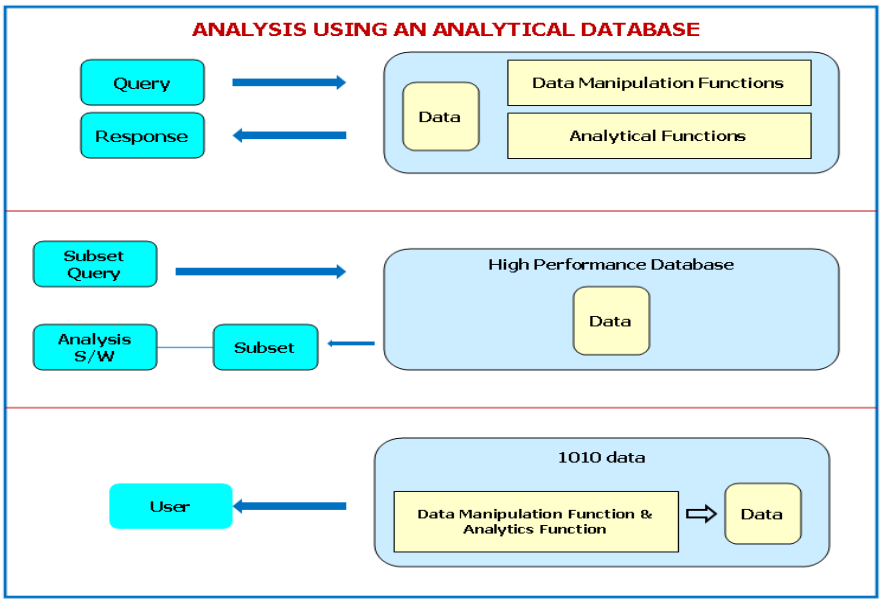

Fig.2. Analytics on Big data

Need of updating the database:
A simultaneous updating process runs on a regular basis reinforcing more amount of data into the database. The updating process runs unconventionally to all other processes and hence the master process is also unaware of the updated data. This feature of data updating effects in such a way that data is kept unchanged for the user's until user's session life.Even thoughsession of current user lastsmore than a day which is though a sparse event, the same concept of data updating is applied.
Therefore, refreshing of data is done at the start of every new session from the user's perspective at the time of new master process creation. But, from the database's point of view, it is refreshed all the time. At some places user generates interest to know when a high level refresh of data takes place. This might come off when 1010data tend to be updated with full day's data at an instance. To receive this information, System message board generates 1010data further displaying reports on such updates [6].On other hand, users could create a session to inspect the previous records, until and unless the session is closed.

IV. CONCLUSION

1010data implements an interactive feature to quickly access and analyze large amount of data, may be up to trillions of rows. It consists of huge analytical capabilities sets of data which has vast scope of application, which are afar the usual scope of any analytical product.

REFERENCES

1. Big Data & Analytics Reference Architecture,An Oracle White Paper, Oracle Enterprise Transformation SolutionsSeries
2. Big Data -A New World of Opportunities,NESSI White Paper, December 2012
3. Big data Analytics for security intelligence, Big data working group,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

4.Robin Bloor, Big data analytics
5.Oracle Reference Architecture Information Management
6. Oracle Reference Architecture Business, Analytics Foundation
7.Oracle Reference Architecture BusinessAnalytics Infrastructure
8. The advantages and disadvantages of real-timeBig data analytic ,Datafloq
9. Analytics: The real world use of Big data, IBM Institute for Business Value
10. Logical Warehousing for Big Data, Gartner, Issue 1

BIOGRAPHY

Ishwinder Kaur Sandhu is a M.Tech student of Computer Science Department, NCU University, Gurgaon, Haryana, India. Her research interests are Big-Data, Cloud Computing and Database.

Richa Chabbra is a Assistant Professor of Computer Science Department, NCU University, Gurgaon, Haryana, India. Her research interests are Big-Data, Cloud Computing and Database.