Big Data Technologies

Thabit Zatari

Abstract- Big Data! The word is quiet not as simple as it sounds like. Big Data is a combination of many activities that involve elicitation of data, organizing the data on the basis of prioritization and usage and finally performing analysis of large sets cluster and collection of data; these all activities combine together to be known as BIG DATA (Fiedler, 2014). This is done so that further enhancements can be performed in discovering patterns by which it is easy to work with things in a flow (Mongodb.com, 2015).

Introduction

There are analytics out there that help the companies to understand the concept and information that can be extracted from the data. Analysts can easily identify which data is important or useful and which is not. These people are called Big Data Analysts, who extract the knowledge that is gained by analyzing the data. This is a challenge for most of the companies and data warehouses. Because it is a sheer number of data and then it is of different forms and formats. It is a developing term that depicts any voluminous measure of organized, semi-organized and unstructured information that can possibly be worked on for data (Webopedia.com, 2015).

Big data is basically of two forms:

- Structured
- Unstructured

The challenge is to collect analyze combine and process the kinds of data.

Now, coming towards the concept of Big Data technologies, while the term is new and it has a scope of latest developments (Fiedler, 2014). There are several technologies used to handle big data. These tools are evolving day by day and most of the tools are not even explored at most because of the time span of development. Big data technologies use several tools to incorporate with today's need of data warehousing. Big data innovation must include the feature of well-built searching, advancement, organization and analysis approaches for all types of sensors and data that is used for social, pictorial and geospatial information etc.

Big data is always measured in the form of: (Fiedler, 2014)

- Volume
- Velocity
- Variety

1. **Volume**
   A normal computer system may have a storage capacity of 500 GB to 1 TB and so on. Social media like Facebook adds almost 500TB of data into its storage every day that means 15000 TB per month. That is a large amount of data to be handled that one can imagine of.
2. **Velocity**
   A usual ad gets an attention of millions of users based on its behavior and rate of processing at millions of actions per second. Similarly, a high-frequency algorithm depicts changes within nano-seconds like in stock markets. Playing games over the internet takes users at the same time; the reason of bringing up these examples is to look at the responsiveness of the data that is being used so the users don't have to wait.
3. **Variety**
   Big data isn't simply information or figures. Big data is additionally intelligent information, 3D information, sound and feature, and unstructured content; including log and online networking.

   Conventional database frameworks were intended to address data that was in the small quantity of organized information, fewer upgrades, and non-changing information structure. Customary database frameworks are additionally intended to work on an alone server, making the possibility of the limit to exceed. As applications have developed to serve extensive volumes with the client's data, the normal utilization of the social database has turned into a risk for some organizations instead of being a support to the firms.

   For such problems we have databases like, SQL server, Oracle 10g, MongoDB, Column-oriented databases, Schema-fewer databases, or NoSQL databases, MapReduce, Hadoop, hive, PIG, WibiData, PLATFORM etc (Fiedler, 2014).

   With the usage of these databases, companies can regain their profits and generate revenue. We will further look

at the Big Data Technologies. (Fiedler, 2014)

Big Data Technologies

Column-oriented databases

Normal tools that save data in the form of rows, these kinds of databases are best for online transactions. Because they provide high-speed processing's. But they lag a bit on the query processing; query means the client's request. The explanation for this is that the data is in a huge amount, illustration: stock exchange information. As the data size surpasses, its processing time is decreased; they are conversely corresponding to one another. Column wise arrangement of DBMS store data in the form of columns that makes the large amount of data accessible and data compression. Quick query times are generated. They only allow group revisions, so the updating time is much less than traditional models (Rodrigues, 2012).

Schema-less databases, or NoSQL databases

There are some databases types which will fit into this schema-less DB's. They focus on the retrieval and storage of huge volume data either organized or unorganized. Performance is gained by putting on some restrictions on regular and conventional databases, for example, Volatile DBMS that only has read-write reliability state and it is in exchange of scalability and parallel processing, this means that we can trade off some features that we can survive without and lessen the load (Rodrigues, 2012).

MapReduce

MapReduce will allow large queries that will run scalability instead thousands of servers. MapReduce basically has two responsibilities: (Rodrigues, 2012)

- The Map task, which means mapping of data input, is done in to the data sets which are converted to a different set of data with values or a single row relational database.
- The task to Reduce includes: mapping of outputs combined from "MAP" part of MapReduce. This will create a reduced set of single row relational databases.

Hadoop

Hadoop is so far the most reliable execution of MapReduce, being an open source stage for taking care of Big Data. It is sufficiently adaptable to have the capacity to work with various information sources, either collecting numerous clusters of information keeping in mind the end goal to do considerable big scale set up, or performing analysis or examination of BIG DATA from a database to run processor-centralized machine learning job that will intelligently perform all the processing. It has a few distinctive applications, but creating use cases is done for bigger volumes of data that is continuously in use (Data, 2015).

For example, Area based information from climate or activity sensors, electronic or online networking information (Data, 2015).

HIVE

A similar bridge as of SQL database that allows the applications to run queries against the Hadoop data sets. Facebook is the founder of hive and is an open source DB structure. It has achieved a higher level of abstraction of Hadoop framework with which queries can be processed easily against any data set that is in the cluster (Rodrigues, 2012).

PIG

Also a Bridge that connects Hadoop to business users, it is quite identical to Hive. The language used in PIG lets the queries surpass the data sets and run easily, it does not use SQL language. And is also an open source relational DB. (Data, 2015)

WIBIDATA

WibiData is a blend of web analysis with Hadoop, being based on top of H Base that is also a layer of DB on the top of Hadoop. It permits sites to better analyze and work with their client information, empowering continuous communication to client side processing's (Rodrigues, 2012).

For example**, serving customized data, providing suggestions and choices to the users**

Storage Technologies

The data size and volume increases so does the needed storage techniques that are effective well-organized and reliable. The main development is related to data compression and data storage on the cloud (Rodrigues, 2012).

Big Data in the cloud

Big data has a lot of technologies nowadays; every other company is using technology that suits them. Most of which are open source that allows them to easily tangle with the areas that are rough and never allowed changes in the traditional DBMS's. Not all technologies are like cloud storage; let us first see what is the cloud? Cloud is server based, this means that one gets to use an online server and put data there so it is accessible online. The reason behind this is that user's servers do not have enough space to maintain their own data warehouses do BIG DATA is now taking a whole new direction. There are several vendors that would allow for cloud storage and most of them are already offering online server side hosted applications like Hadoop group or cluster of data, it will also provide the processing on the cloud side so user's computer does not have to wait for the

long processing times, all users have to do is send the query online and they will get the results. The Hadoop data sets are managed on demand according to the user's usage. (Rodrigues, 2012)

Conclusion

The platforms of BIG DATA TECHNOLOGIES discussed have their cloud versions as well. BIG DATA and cloud computing can work parallel and combine to give a unique and ultimately progressive result that will take 50 % of the dependency away from the processing time. Distributed computing lets organizations of all sizes to get more revenue from their data than any other time in recent memory, by empowering full quick analysis at a few past expenses (Harris, 2012). This drives organizations to gain and store significantly more data, making it the actual BIG DATA organization. Because its handling power is way more than someone can imagine (Rodrigues, 2012).

Companies using Big Data in the cloud are:

- Google
- Infochimps
- Medio
- Metamarkets
- Microsoft
- Xignite

Many other companies are using Big Data in the cloud but are still progressively working on it and will reveal new technologies related to it for improving efficiency (Harris, 2012).

References

J,Clufia,D,Bunzil&S,Scruggs.(2014). Consumer Marketers: Digging Too Deep With Data Mining?Digitalwave, Available at:http://digitalmediaix.com/consumer-marketers-digging-too-deep-with-data-mining/#.VXMgI8-qqko[accessedjune6,2015]

R,Dholakia.(2013). Scholarly Research in Marketing: Trends and Challenges in the Era of Big Data, International Encyclopedia of Digital Communication & Society, Available at: http://web.uri.edu/business/files/Encycl-Communication-DataMining-n-Marketing-.pdf

C,Meadows.(2013). Amazon data mining to find customer tastes , Teleread, Available at:http://www.teleread.com/ebooks/amazon-mines-data-to-find-customer-tastes/[AccessedJune6,2015]

M,Usman& R,Peers.(2010). Integration of Data Mining and Data Warehousing: a practical methodology, International Journal of Advancements in Computing Technology, vol.2(3), pp.31 – 46.Available at: http://aut.researchgateway.ac.nz/handle/10292/1684[AccessedJune6,2015]