Big Data

AnasAlarfaj

Introduction

The modern world is digital and huge amount information is being generated on millisecond basis. The generated information has to be stored in digital format for further processing and analysis. Big data is the data bank with data getting accumulated as a continuous process. This accumulated data can be structured or unstructured. This data can be relevant and useful to the company only upon proper analysis. The size of the big data is in ranges of petabytes and exabytes (Steve Lohr, 2012). There are many forms of big data. Some forms of big data are data generated through social network sites, smartphone data, e-mails, climate predictions, data from intelligent devices and many others. Data generated through the mentioned means are normally unstructured and are not easy to analyze. These databases have fields that are not properly ordered, have no well defined rules, varying data rate and many others. Analysis of this data can give fruitful information and can uplift a company in multiple ways. Analysis of big data is said to be "Big Return on Investment" (Steve Lohr, 2012). Through big data analysis companies have achieved 20% decrease in patient mortality, 99% increase in placement of power generation resources and 92% decrease in processing time of the production industries (IBM, (2014).

What is Big Data?

Big data can be defined as the continuous collection of data in the database that needs to be analyzed using dedicated tools to generate powerful results. Big data is generally complex in nature i.e. unstructured with unorganized and text heavy information, multi-structured with different data formats and huge image contents, not following any database model and many others. The analysis tool must be capable enough to manage the haphazard data and provide productive information. Many IT companies like IBM, SAP and Oracle are working hard to analyze big data and provide valuable information to the customers.

Understanding Big Data

Big data is a term that represents the explosive growth of digital data. In order to manage this essential data, IT engineers are finding issues with respect to data storage and tools for data analysis. According to IBM, the people around the world are generating 2.5 quintillion bytes of data on a daily basis and this rate is on the rising side (IBM, (2014). Some of the generated data is not only enormous but also complex to analyze. Companies are finding it difficult to keep their pace with this rapidly increasing data. IT giants have come up with a dedicated platform called Big data platform for supporting industries to have powerful information for achieving an upper edge over their competitors.

3 'V's' of Big Data

Although big data is generally big, there are three V's for identifying about their types and behavior (Steve Lohr, 2012). The first V is Volume of the available big data. This parameter is very important and it gives the idea about requirements for database and analysis tools. The second V is the Variety. It gives information on the complexity of the data whether it is structured or unstructured, the number of fields and many others about the internal data layout. The third V is the velocity (Pinal Dave, October 2). It indicates the rate at which the data is being available for analysis. It gives the information on how quickly the data can be analyzed using standard tools.

Big Data Platform

This platform is mainly concerned with analysis and extraction of maximum information from the digital data that is stored in different databases. This platform works on predefined objectives and has different coding for different customer requirements. The big data platform works on pipeline processing with various phases of analysis to extract the useful information. To begin explaining about big data platform, the first step is to understand a user's need. Based on the user needs, objectives are set. Relevant information pertaining to the predefined objectives are recorded, anything irrelevant is filtered out. This brings about clarity in the data and helps in easy analysis. During the data recording process, one cannot expect to have clear and error free data at all times. In order to ensure error free data, the recorded data is extracted and cleaned in a periodic manner (Eva Tse, 2014). Cleaning of the data involves usage of queries to check if the data is valid and correct. Using of data without proper cleaning may end up in erroneous outputs. Cleaning of data is the second phase in the pipeline processing. Information in the recorded data may be scattered and stored at different database locations. Depending on the objectives, the relevant information is aggregated in a common area at nearby locations. The databases are so created for performing easier aggregation of the relevant information. The aggregated data is then integrated together and represented in a proper manner to achieve the objective (Eva Tse, 2014). This process of data aggregation, integration and representation is the third phase of the pipeline process. The integrated data is error free and data mining is carried out by implementing various methods such as query processing, data modeling and various other methods of data analysis. Data mining of the error free data is the fourth phase of the big data pipeline process and it generates some useful values, figures and graphs. Interpretation of this generated information is carried out by

the user for the purpose of decision making and framing policies.

Applications of Big Data

Results of the big data analysis find its application in many areas (IBM, (2014), (Andy Hayler, 2013). Some of the applications of big data information are:

- Automotive Sector: It helps in determining the insight on customer needs and other information such as optimization of production methods, sales prediction, vehicle design requirements and many others.

- Banking Sector: It brings out methods to have better customer service, investigation and detection of frauds, increasing efficiency of banking software and others.

- Government Sector: It is highly beneficial in the government and helps in prediction of natural calamities and threats, controlling crime rates, benefits of social programs, population growth rate and others.

- Healthcare: In health care big data analysis helps in multiple ways. It helps to study the behavior of cell growth, patient monitoring and diagnosis, determining patient's mortality rate for a particular disease and others.

- Communication Sector: It helps effective ways of customer support, analysis of networks, area of investments, acceptance of offers by the customers and many others.

Benefits of Big Data

- Big data is helpful in achieving optimization. It gives an insight into the focused area and greatly helps in increasing system efficiencies and system productivity (Ericsson white paper, 2013).

- It helps industries to understand the dynamic behavior of a particular product. For example, in a telecommunication company, annual information about the customer's choice will be helpful in predicting customer behavior in any part of the year.

- It helps in giving new ideas on improvement to the industries. The outcomes of big data analysis can be used by the experts for development in many areas such as space technologies, development of medicines and many others.

- Industries can have real time information about the trends of various parameters and customer behavior that are affecting the business (Ericsson white paper, 2013). .

- Big data can assist to develop multi sided business models. Companies can use the outcome of the big data analysis to develop appropriate models for the business.

Big data and Cloud Computing

Cloud computing is one of the recent time developments in IT sector where the computing services are sold by certain service providers. In short it can be said that cloud computing has tremendous storage capacity, networking capabilities and high bandwidth. It is held by certain service providers such as amazon. A user can pay to use the required memory space to store his/her information (MangodB, 2014).

Industries at present are relying on cloud computing for storing the big data. Among various benefits of using cloud computing some are explained as follows. Cloud computing is inexpensive and user needs to pay only for the memory space that is used (MangodB, 2014). Cloud computing is extremely flexible and user can adjust the usage space dynamically. It is reliable and the scope for data loss is very less.

Big Data tools

Among various available tools for big data processing some are explained as follows (Peter Wayner, 2012).

- Jaspersoft BI Suite: This tool is mainly for user defined report generation from the big data. It performs the analysis and generates reports in the form of pdf's. It uses storage platforms such as MangoDB, Cassandra, Redis and Riak.

- Pentaho Business Analytics: This tool can perform operations like sorting, shifting and other necessary data movement activities. It has a graphical

programming interface and can generate reports with graphs and pictures.

- Karmasphere studio and analyst: This tool is built on Eclipse software and can generate used defined interface for generating reports and deriving other essential parameters. It can run Handoop jobs.

- Talend open studio: This tool is also developed in Eclipse based IDE. It has a canvas and provides wide range of capabilities to the users. Each of the capabilities is in the form of small icons (Cynthia Harvey, 2012). The user can select the desired icon and drag them over the canvas for further processing works.

- Skytree server: This tool is rugged in construction and is designed using a number of classic proven algorithms. It has a user interface and the algorithm runs in the backend. This tool claims that it is 10,000 times faster than all the other available tools.

- Splunk: Unlike other data analysis tools, splunk has a collection of sub routines to suit for a particular application. It generates an index of the available data and uses them for performing analysis. It also updates the index every time a new record is added.

Challenges of Big Data

Big data analysis being a pipeline process with five different phases has got many challenges. Each of the processing phases has one or more associated challenges. Some of the big data challenges are as explained below (Steve Lohr, 2012)

- Incomplete information: Big data is a kind of unpredictable data which at any point of time is incomplete (Global Pulse, 2012). This data is highly dynamic in some cases and always adds up new information to the existing data. Analysis of such an heterogeneous data is a difficult task.

- Data scaling and processing: The big data is voluminous and requires powerful processors and other managing equipments. Development of the required processing speed is a big challenge. Many processing methods such as parallel processing methods, intra node parallelism and others are being adopted to overcome this issue.

- Time consumption: Processing of these huge data takes its own time. The available processor must not only process the available data but also process the incoming data at a faster rate to maintain pace with the input data (Global Pulse, 2012). The time required for processing is a big challenge.

- Data security threats: The big data analysis poses threats to the customers. For example, a telecommunication company would like to make certain predictions and it gives its customer information over a certain period to the analysis team. Under such a situation there is no guarantee on the safety of the customer data.

- Requirement of human intervention: Big data analysis may some time require experts from different fields to interpret the results and performing operations. It is required that experts from different fields must contribute to define the work that needs to be done. The timely human intervention is a big challenge.

Big Data and Netflix

Netflix is on of the many companies that analyze the big data and uses its outcome for its decision making. This company started its business in late 90's by performing home delivery of DVD's of customer's choice. In recent times, (2007) it has started live streaming of video on customers demand on internet (Daniel Nippes, 2014). The company charges its customer for showing videos and gets their ratings for providing better services. Some of the facts about Netflix are as follows (Daniel Nippes, 2014). This company has more than 25 million users world wide, they stream videos and record customer actions such as pause, fast forward, rewind and many others. They get around 4 millions ratings on a daily basis and 3 million people searche for this site. In earlier times, Netflix stored the customer information and other details on the data warehouse and were performing data mining to extract information. With tremendous increase in data Netflix has switched over to cloud for data storage.

Netflix uses Cassandra, an open source NoSQL database for processing of the data (Eva Tse, 2014). IT experts at Netflix have used Cassandra to create Aegisthus for the purpose of huge information pipelining. It uses cloud to store the data and perform processing at the required stage (Daniel Nippes, 2014).

In this way Netflix has come up with a cost effective solution for analysis of big data.

Conclusion

It is understood that big data apart from being voluminous is also complex in nature. The big data can be available in any form, it can be structured, unstructured, multi structured, text heavy, presence of erroneous data, presence of irrelevant information and many others. Big data as a raw material is worthless and is of no use whereas, proper analysis of the big data can give very useful information for decision making and bringing out optimizations in the industry.

Big data are analyzed through pipeline processing and it follows five important phases to provide useful results. Big data in recent times has found its application in many areas. Processing speed and processing capabilities are important challenges for big data. Many things have been done to use the big data information effectively, but many more are still to be done.

References

Peter Wayner. (2012, April 18). 7 top tools for taming big data. http://www.infoworld.com/article/2616959/big-data/7-top-tools-for-taming-big-data.html

Andy Hayler. (2013). 'Big data' applications bring new database choices, challenges. http://www.computerweekly.com/feature/Big-data-applications-bring-new-database-choices-challenges

Eva Tse. (2014, October 7). Using Presto in our big data platform on AWS. http://techblog.netflix.com/2014/10/using-presto-in-our-big-data-platform.html

MangodB (2014). Big Data Explained. http://www.mongodb.com/big-data-explained

IBM. (2014). Bringing big data to the enterprise. http://www-01.ibm.com/software/in/data/bigdata/industry.html

Global Pulse. (2012, May). Big data for development: Challenges & Opportunities. http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf

Pinal Dave. (2013, October 2). Big data – What is big data – 3 Vs of Big data. http://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/

Ericsson whte paper, (2013, August). Big data analytics. www.ericsson.com/res/docs/whitepapers/wp-big-data-v7.pdf

Daniel Nippes. (2014, May 2015). Netflix's big data architecture. http://dataconomy.com/netflix-big-data-architecture/

Cynthia Harvey. (2012, June 4). 50Top open source tools for Big Data. http://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-1.html

Steve Lohr. (2012). Challenges and Opportunities with big data. https://www.purdue.edu/discoverypark/cyber/assets/pdfs/BigDataWhitePaper.pdf.