International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

A Review paper on BIG Data

Sushmita, Simranjeet Kaur, Simranjot Kaur

Department CSE, College AIT Gharuan, Punjab , India

ABSTRACT

Big data is a data or data sets so large or complex that traditional data processing applications are inadequate and distributed databases are needed. Firms like Google, eBay, LinkedIn, and Face book were built around big data from the beginning. It is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, real time data etc. Challenges include sensor design, capture, data curation, sharing, storage, analysis, visualization, information privacy etc. Big data refers to datasets high in variety and velocity, so that very difficult to handle using traditional tools and techniques. The process of research into massive data to reveal secret correlations named as big data analytics. Big Data is a data whose complexity requires new techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. We need a different platform named Hadoop as the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes.

Keywords: Big Data, Parallel Programming, Data Buck Map Reduce.

I. INTRODUCTION

Every digital process and social media exchange produces Big data. The Systems, sensors and mobile devices transmit. The arrival of big data is from multiple sources at a frightening velocity, volume and variety. We need optimal processing power, analytics capabilities and skills to extract meaningful value from big data. More confident decision making can do with accurate big data. Good decisions lead to greater operational efficiency, cost reduction and reduced risk. Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, and combat crime and so on Scientists, business executives, practitioners of media, and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks[2][3][4] . Big data "size" is a constantly moving target, ranging from a few dozen terabytes to many peta byte of data. (1 petabyte is 1000 terabytes)

Here are some real-world examples of Big Data in action:

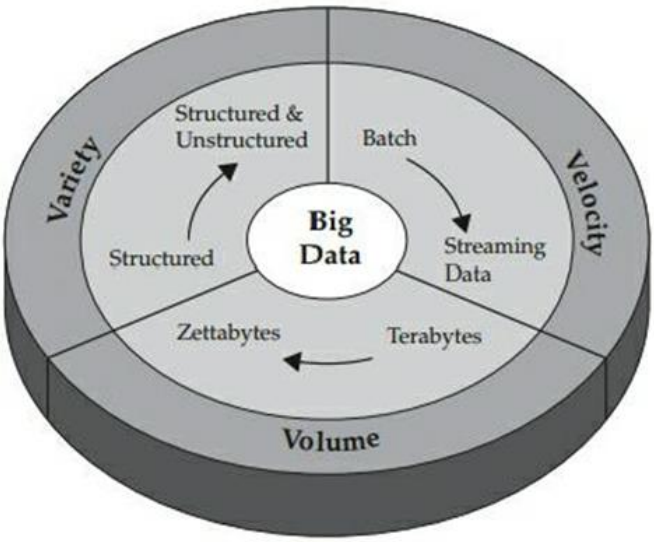
- ✓ Consumer product companies and retail organizations are monitoring social media like Face book and Twitter to get an unprecedented view into customer behavior, preferences, and product perception.
- ✓ Manufacturers are able to monitor minute vibration data from their equipment, which changes slightly

as it wears down, to predict the optimal time to replace or maintain. Replacing it too soon wastes money and replacing it too late triggers an expensive work stoppage.

- ✓ Manufacturers are also monitoring social networks, but with a different goal than marketers: They are using it to detect aftermarket support issues before a warranty failure becomes publicly detrimental.
- ✓ The government is making data public at the national level, state level, and city level for users to develop new applications that can generate public better.
- ✓ Financial Services organizations are taking data mined from customer interactions to slice and dice their users into finely tuned segments and enables these financial institutions to create increasingly relevant and sophisticated offers.
- ✓ Advertising and marketing agencies are tracking social media to Insurance companies are using Big Data analysis to see which home insurance applications can be immediately processed, and which ones need a validating in-person visit .
- ✓ Retail organizations are engaging brand advocates, changing the perception of brand antagonists, and even enabling enthusiastic customers to sell their products. All these things are doing by embracing social media.
- ✓ Hospitals predict those patients that are likely to seek readmission within a few months of discharge by analyzing medical data and patient records . The hospital can then preventing another costly hospital stay.
- ✓ To offer more appealing recommendations and more successful coupon programs the , Web-based businesses are developing information products that combine data gathered from customers
- ✓ Sports teams are using data for tracking ticket sales and are using big data for tracking team strategies also.

a. Three Vs of big data: volume, velocity & Variety

Volume. The size available data has been growing At an increasing rate. This applies to companies and To individuals. A text file is a few killo bytes,a sound file is a few mega bytes while a full length movie is a few giga bytes. More sources of data with a larger size of data combine to increase the volume of data that has to be analyzed. This is a major issue volumes and how to use analytics to create value from relevant data. Volume referred as amount of data.



Velocity. Thortan may says ”Initiatives such as the use of RFID tags are smart metering are driving an ever greater need to deal with the torrent of data in near-real time”. This couple with the need and drive to be more agile and deliver insight quicker is putting tremendous pressure on organization to build the necessary infrastructure and skill base to react quickly enough. Variety. Today data comes in different types of formats. Structured and numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio and financial transactions. Managing, merging and governing different varieties of data are something many organizations still struggle with. Different types and sources of data are there. Data variety exploded from structured and legacy data stored in enterprise storages to unstructured, semi structured, audio, video etc.

We consider two additional dimensions when thinking about big data: Variability. With the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. It is trending in social media. Everyday seasonal and event-triggered peak data loads cannot be able to manage. Even more unstructured data involved. The inconsistency the data can show at times—-which can hamper the process of handling and managing the data properly. The inconsistency the data can show at times can hamper the process of handling and managing the data properly.

Complexity. Difficulties dealing with data increase with the expanding universe of data sources and are compounded by the need to link, match and transform data across business entities and systems organization need to understand relationships, such as complex hierarchies and data linkages , among all data.

Veracity, The quality of captured data, which vary so high. The Accurate analysis of data depends on the veracity of source data.

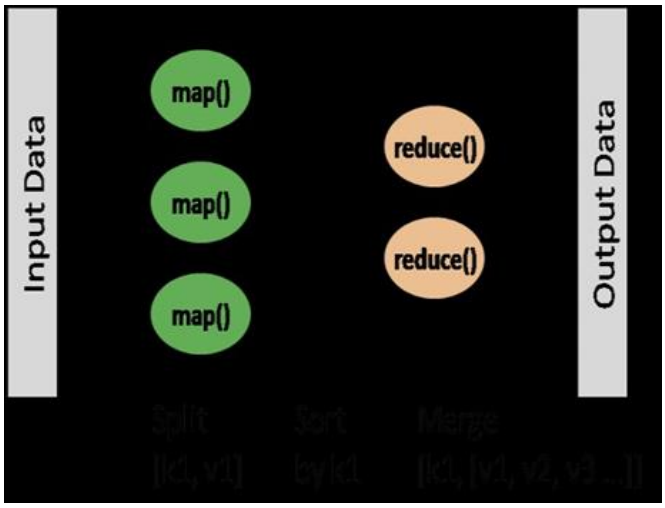
II. PARALLEL PROGRAMMING & MAPREDUCE

Data analysis software parallelizes fairly naturally. Many programmers are interested to building programs on the parallel model. The parallel research had the most success in the field of parallel databases. Rather than requiring the programmer to unknot an algorithm into separate threads to be run on separate cores, parallel databases let them break up the input data tables into pieces, and pump each piece through the same single-machine program on each processor. This "parallel dataflow" model makes parallel programming as easy as programming a single machine. nd it works on "shared-nothing" clusters of computers in a data center: The machines involved can communicate via simple streams of data messages, without a need for an expensive shared RAM or disk infrastructure. [6]

Famous big data analysis tool is Hadoop. Apache Hadoop is an open-source software framework . it is written in Java for distributed storage and distributed processing of big data on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines or racks of machines) are commonplace and thus should be automatically handled in software by the framework.[7]

The heart of Hadoop isMapReduce. It is this programming paradigm that allows for massive scalability across thousands of servers in a Hadoop cluster. It is useful for batch processing on petabytes or zeta bytes of data stored in Apache Hadoop. If we are familiar with clustered scale-out data processing solutions. Then the MapReduce concept is simple to understand. MapReduce programming model has twisted a new page in the parallelism story. The MapReduce framework is a parallel dataflow system that works by dividing data across machines. Each of which runs the same single-node logic. MapReduce asks programmers to write traditional code, in languages like C, Java, Python and Perl. In addition to its familiar syntax, MapReduce allows programs to be written to and read from traditional files in a file system, rather than requiring database schema definitions.

MapReduce refers to two separate and distinct tasks. The first is the job of map, which takes a set of data and converts it into another set of data. Individual elements are broken down into value pairs. The reduce job takes the output from a map as input and combines those data values into a smaller set of values. The reduce job is always performed after the map job. So the sequence of the name MapReduce.



MapReduce

III. BEST BIG DATA ANALYTICS USE CASES

3.1. Sentiment Analysis

Sentiment analysis offers powerful business intelligence to enhance the customer experience, revitalize a brand, and gain competitive advantage. The key to successful sentiment analysis lies in the ability to dig for multi- structured data pulled from different sources into a single database.

3.2. 360-Degree View of Customer A 360-degree customer view offers a deeper understanding of customer behavior and motivations. Obtaining a 360-degree customer review requires analysis of data from different sources like social media, data collecting sensors, mobile devices etc. From there, more effective

micro-segmentation and real-time marketing are getting as result.

3.3Ad Hoc Data Analysis

Ad-hoc analysis only looks at the data requested or needed, providing another layer of analysis for data sets that are becoming larger and more varied. Big data ad-hoc analytics can help in the effort to gain greater insight into customers by analyzing the relevant data from unstructured sources, both external and internal.

3.4 Real-Time Analytics

Systems that offer real-time analytics quickly decipher and analyze data sets, providing results even as data is being generated and collected. This high-velocity method of analytics can lead to immediate reaction and changes allows for better sentiment analysis, split testing, and improved targeted marketing.

3.5 Multi-Channel Marketing Multi-channel marketing creates a seamless across different types of media like company websites, social media, and physical stores. During all stages of the buying process multi-channel marketing requires an integrated big data approach.

3.6 Customer Micro-Segmentation

Customer micro-segmentation provides more tailored and targeted messaging for smaller groups. This personalized approach requires analysis of big data collected through sources like customers' online interactions, social media etc.

3.7 Ad Fraud detection

Ad fraud detection requires data analysis of fraud strategies by recognizing patterns and behaviors. Data that shows irregularity of group behavior make it so ad fraud is find out and blocked before it is spread.

3.8 Click stream analysis

Click stream analysis helps to grow the user experience by optimizing company websites, and offering better insight into customer segments. Click stream analysis helps to personalize the buying experience, getting an improved return on customer visits with big data.

3.9 Data Warehouse Modernization

Integrate big data and data warehouse capabilities to boost operational efficiency. Optimize your data warehouse to enable fresh types of analysis. Use big data technologies to set up a staging area or landing zone for your new data before formative what data should be moved to the data warehouse. divest infrequently accessed or aged data from warehouse and application databases using in sequence integration software and tools.

3.10 Big Data and Predictive Modeling

The most common uses of big data by companies are for tracking business processes and outcomes, and for building a wide array of predictive models. Amazon and Netflix recommendations rely on predictive models of what book or movie an individual might want to purchase. Google's search results and news feed rely on algorithms that predict the significance of particular web pages or articles. pple's auto- complete function tries to forecast the rest of one's text or email based on past convention patterns. Online advertising and marketing rely greatly on automated predictive models that aim individuals who might be particularly likely to answer to offers.

The application of predictive algorithms extends well ahead of the online world. In health care, it is now common for insurers to adjust payments and quality measures based on "risk scores," which are resulting from predictive models of human being health expenses and outcomes. n individual's risk score is naturally a weighted sum of health indicators that recognize whether an individual has different persistent conditions, with the weights chosen based on a statistical analysis. Credit card companies use predictive models of default and repayment to guide their underwriting, pricing, and marketing actions.

IV. KEY BIGDATA CHALLENGES

Data frequently loses its trustworthiness due to

- ✓ Undetected error in incoming data
- ✓ Multiple data source that get out of sync over time
- ✓ Structural change to data in upstream processes not expected downstream and,

- ✓ Presence of multiple IT platforms (Hadoop, DW, Cloud).unexpected errors creep in when data resides in a system or it moves between a Data warehouse to a Hadoop environment, or No SQL database or the cloud . fault process ad hoc data polices ,poor discipline in capturing and storing data and lack of control over some data sources(example: External data providers) all contribute to data changing unexpectedly.

V. DATA BUCK

An autonomous, self learning ,big Data quality and integrity validation and data reconciliation tool. Designed to simplify elaborated and complex validation and reconcilations. Machine learning capabilities built into the tool autonomously set 100,000s of validation checks w/o manual intervention. Data buck learns about your data quality behaviour and models it using advanced Machine learning techniques. Its develops extensive Data Quality Fingerprints at multiple hierarchical incoming data for reasonableness.

Big data skills are in short supply There's already a shortage of data scientists in the market. This includes a shortage of people who know how to labor well with large volumes of data and big data sets. Companies need the right merge of people to help make sense of the data streams that are coming into their organizations. This includes skills for applying prophetic analytics to big data, a skill set that even most data scientists be short of.

VI. CONCLUSION

The availability of Big Data, low-cost commodity hardware, and analytic software has shaped a unique moment in the history of data analysis. The union of these trends means that we have the capabilities required to analyze amazing data sets quickly and cost-effectively for the first time in history. All these capabilities are neither theoretical nor trivial. They represent a real leap forward and a clear chance to realize enormous gains in terms of efficiency, productivity, income, and profitability. Requirements for dealing out that may seem unbelievable today will soon be routine when big data systems are available. We learn how to exploit them. Not very many years ago, systems the scale of Face book and Google would have seemed like science fiction. At that time 100 transactions per second for airline and banking systems was a stretch. Several new requirements will also combine data from many sources, not all of which will be company-owned. For instance, some will make use of "open data" from government. Lots of opening for innovators!

VII. REFERENCES

[1]. "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.

[2]. "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.

[3]. "Community cleverness required". Nature 455 (7209): 1.4 September 2008. doi:10.1038/455001a.

[4]. "Sandia sees data management challenges spiral". HPC Projects. 4 August 2009.

[5]. META Group. "3D Data Management: Controlling Data Volume, Velocity, and Variety." February 2001.

[6]. Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". Gigaom Blog.

[7]. "Welcome to pacheP Hadoop®!". hadoop.apache.org. Retrieved 2015-09-20.