Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey on Data Mining with Big Data

V. Vadivu
Department of Information Technology
Bharathiyar University, Tamil Nadu, India

Abstract— Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques.Big Data Mining is the capability of extracting useful information from these large datasets or streams of data,that due to its volume,variability and velocity,it was not possible before to do it.This study paper includes the information what is big data,importance of big data,issues of big data,techniques and data mining with big data.

Keywords— Big data,Importance of Big data,Issues of Big Data,Techniques,Data Mining with Big Data

I. INTRODUCTION

In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity. Despite these problems, big data has the potential to help companies improve operations and make faster, more intelligent decisions.Basically,the Big Data is stored at different places and also the data volumes may get increased as the data keeps on increasing continuously.So,to collect all the data stored at different places is that much expensive.The typical data mining methods are used for mining the small scale data in our personal computer systems and its maintain the privacy.

II. IMPORTANCE OF BIG DATA

A. Big Data

Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.The storage capacity of Big Data is stared from terabyte to exabyte.

B. Importance

When big data is effectively and efficiently captured, processed, and analyzed, companies are able to gain a more complete understanding of their business, customers, products, competitors, etc. which can lead to efficiency improvements, increased sales, lower costs, better customer service, and/or improved products and services.
The effective use of big data exist in the following areas

- Using information technology (IT) logs to improve IT troubleshooting and security breach detection, speed, effectiveness, and future occurrence prevention.
- Use of voluminous historical call center information more quickly, in order to improve customer interaction and satisfaction.
- Use of social media content in order to better and more quickly understand customer sentiment about you/your customers, and improve products, services, and customer interaction.
- Fraud detection and prevention in any industry that processes financial transactions online, such as shopping, banking, investing, insurance and health care claims.
- Use of financial market transaction information to more quickly assess risk and take corrective action.

III. ISSUES OF BIG DATA

A. Problems of Big Data

IBM's Big Data estimates conclude that "each day we create 2.5 quintillion bytes of data." The exponential growth of data means that 90 percent of the data that exists in the world today has been created in the last two years. "This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, e-commerce transaction records, and cell phone GPS coordinates, to name a few."
To put the data explosion in context, consider this. Every minute of every day we create:

- More than 204 million email messages
- Over 2 million Google search queries
- 48 hours of new YouTube videos
- 684,000 bits of content shared on Facebook
- More than 100,000 tweets
- 3,600 new photos shared on Instagram
- Nearly 350 new WordPress blog posts

B. Barriers

1) Unstructured data. There are two types of data in storage, structured and unstructured data. Structured data has a high degree of organization, and is typically stored in a relational database that can be easily searched.Unstructured data is, obviously, not structured in any meaningful way, including such things as photographs, videos, MP3 files, etc. Unstructured data is difficult to search and analyze.

2) I/O barriers. If you're dealing with something like mapping genomes, gathering information from the Mars Rover or running sophisticated weather simulations, the transaction volumes of these data sets challenge traditional storage systems, which don't have enough processing power to keep up with the huge number of I/O requests.

3) Management. There are a million and one storage management tools out there. The most basic one – and one still in wide use even in business, believe it or not, is a simple Excel spreadsheet – but vendors from EMC to Hitachi Data Systems to NetApp offer solid storage management solutions. The trouble is, though, that data-sharing standards are still lacking and escaping vendor-lock is a never-ending challenge.

4) The WAN. As cloud computing becomes mainstream, the simplest way to break down data silos is to leverage the cloud to help with everything from search to backups to raw processing. However, as more storage moves into the cloud, the more the WAN will impede on Big Data progress. The WAN, unfortunately, isn't keeping up with Moore's Law, nor with the storage-specific analog Kryder's Law. Any Big Data storage solution must include some combination of redundant MPLS links, WAN optimization and CDN services.

5) Security. As you break down data barriers, certain people may get access to data (say HR records) that they should never, ever see. Thus, authentication, access and security in general are a major Achilles heel of Big Data storage.

IV. TECHNIQUES

The traditional databases, such as SQL, weren't designed with Big Data in mind, eventually a Big Data alternative emerged are list out below

A. Emerging Technologies in Big Data

1) Schema-less databases, or NoSQL databases

There are several database types that fit into this category, such as key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some (or all) of the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

2) MapReduce

This is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers. Any MapReduce implementation consists of two tasks:

- The "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples;
- The "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples (hence the name).

3) Hadoop

Hadoop is by far the most popular implementation of MapReduce, being an entirely open source platform for handling Big Data. It is flexible enough to be able to work with multiple data sources, either aggregating multiple sources of data in order to do large scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs. It has several different applications, but one of the top use cases is for large volumes of constantly changing data, such as location-based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data.

4) Hive

Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users.

5) PIG

PIG is another bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.

6) WibiData

WibiData is a combination of web analytics with Hadoop, being built on top of HBase, which is itself a database layer on top of Hadoop. It allows web sites to better explore and work with their user data, enabling real-time responses to user behavior, such as serving personalized content, recommendations and decisions.

7) PLATFORA

Perhaps the greatest limitation of Hadoop is that it is a very low-level implementation of MapReduce, requiring extensive developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, eliminating the interactivity that users enjoyed with conventional databases. PLATFORA is a platform that turns user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

8) Storage Technologies

As the data volumes grow, so does the need for efficient and effective storage techniques. The main evolutions in this space are related to data compression and storage virtualization.

9) SkyTree

SkyTree is a high-performance machine learning and data analytics platform focused specifically on handling Big Data. Machine learning, in turn, is an essential part of Big Data, since the massive data volumes make manual exploration, or even conventional automated exploration methods unfeasible or too expensive.

10) Big Data in the cloud

As we can see, from Dr. Kaur's roundup above, most, if not all, of these technologies are closely associated with the cloud. Most cloud vendors are already offering hosted Hadoop clusters that can be scaled on demand according to their user's needs. Also, many of the products and platforms mentioned are either entirely cloud-based or have cloud versions themselves.Big Data and cloud computing go hand-in-hand. Cloud computing enables companies of all sizes to get more value from their data than ever before, by enabling blazing-fast analytics at a fraction of previous costs. This, in turn drives companies to acquire and store even more data, creating more need for processing power and driving a virtuous circle.

V. DATA MINING WITH BIG DATA

To maintain the privacy is one of the main aims of data mining algorithms.Presently,to mine information from Big Data,by using any one of the adopted technology.In such algorithms,large data sets are divided into number of subsets and then,mining algorithm are applied to those subsets.Finally,summation algorithms are applied to the results of mining algorithms,to meet the goal of Big Data mining.

A. Data Mining

Data mining involves six common classes of tasks:

1) Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

2) Association rule learning (Dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

3) Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

4) Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

5) Regression – attempts to find a function which models the data with the least error.

6) Summarization – providing a more compact representation of the data set, including visualization and report generation.

VI. CONCLUSIONS

Big Data is going to continue growing during the next years,and each data scientist will have to manage much more amount of data every year.This This data is going to more diverse,larger,faster and is becoming the new scientific data research and for business applications.Big datamining is new era which is help to discover knowledge.

REFERENCES

[1] D.Howe et al,"Big Data:The Future of Biocuration," Sept.2008.
[2] A.Labrinidis and H.Jagadish,"Challenges and Opportunities with Big Data,"Proc.VLDB Endowment, ,2012.
[3] Y.Lindell and B.Pinkas,"Privacy Preserving Data Mining",J.Cryptology,2001

