CS 6350 Big Data Management and Analysis

Assignment 1

Name: Dharav Bhatt

UTD ID: 2021599509

I. What is Big Data?

Ans. There is variety of data being generated around us. Such as:

- Genomic Data
  DNA data
- Data from various Social Media platforms
  Audio, Video, GIF, tags, embedded links, etc.
- Space Data
- IOT Data
- Mobile data
- Streaming Data
- Network Data and many more.

All these data in collection is known as Big Data.

Big data can also be defined as a data that is growing and presenting challenges in all the three dimensions, i.e. increasing Volume, Velocity, and Variety.

Section 1.1 of the paper

Q1. What does the term Big Data (BD) refer to? How is BD different from traditional datasets?

Ans. Big data is mainly described as an enormously large dataset. Traditional datasets consist of only structured data which is comparatively smaller in size. Whereas big data typically includes masses of unstructured data that need more real-time analysis.

Q2. What challenges have emerged because of the rise of BD?

Ans. The main challenge raised due to the rise of the BD is collecting and integrating massive data from the widely distributed data source. For example, Google processes data of hundreds of Petabytes (PB), Facebook generates log data of over 10 PB per month, on average, and 72 hours of videos are uploaded to YouTube every minute. The increasingly growing data cause a problem of how to store and manage such huge heterogeneous datasets with moderate requirements on hardware and software infrastructure.

Section 1.2 of the paper

Q1. This section presents several definitions and features of BD. Write down in pointwise fashion the features of BD. Pay special attention to the 3V definition proposed by Laney and understand what each term means.

Ans. The features of BD are as follow:

1. Volume: Big Data is increasing rapidly with the generation and collection of masses of Data.

2. Velocity: Analysis of the data collected must be rapidly and timely conducted in order to maximumly utilize the commercial value of BD.

3. Variety: BD includes various types of semi-structured and unstructured data such as audio, Video, webpage, and text, as well as traditional structured data.

Characteristics of BD:

Q1. What is meant by volume of BD? How has it changed over the time?

Ans. Volume of BD simply refers to the amount of data that is to be stored and analyzed. In the year 2000, 800,000 petabytes (PB) of data were stored in the world. This number is expected to reach 35 zettabytes (ZB) by 2020. Twitter alone generates more than 7 terabytes (TB) of data every day, Facebook 10 TB, and some enterprises generate terabytes of data every hour of every day of the year.

Q2. How has increased volume created a "blind zone" for organizations?

Ans. As the amount of data available for the enterprise increases, the percent of data it can process and analyze decrease, thereby creating the blind zone.

Q3. What is meant by variety of BD? What are the various types of data that large organizations acquire today?

Ans. BD includes various types of unstructured and semi-structured data. The data is more complex as compare to traditional data. It includes web pages, web log files (including click-stream data), search indexes, social media forums, e-mail, documents, space data, and sensor data from active and passive systems, traditional data and so on.

Q4. How is velocity of data applied to data in motion? What are the advantages of streams computing?

Ans. Velocity of data can be applied to data in motion as the speed at which the data is flowing. In order to find an insight in this data, organizations must be able to analyze this data in near real time. With streams computing, you can execute a process similar to a continuous query that identifies people who are currently "in the New Jersey flood zones," but you get continuously updated results, because location information from GPS data is refreshed in real time.

II. What is the value of Big Data

Department of Homeland Security (DHS) is using big data in many areas such as:

- Defense/Military Intelligence.
- Government Intelligence Agencies.
- Law Enforcement agencies and 1st Responders.
- Financial Services Industry & fraud Detection.
- Cyber security agencies.
- Border & customs Control.
- Mass Transportation (air, sea & land).
- Intelligence fusion centers.
- Critical Infrastructure Security.
- Other vertical homeland security and public safety.

The federal government is using big data in order to combat terrorism. In 2012, the government launched two pilot programs, Neptune and Cerberus. Neptune is a data lake of unclassified information coded with data tags in order to control access and protect personal privacy. Cerberus is a data lake of classified information that has more stringent security. Taken together, the Department of Homeland Security is able to run analytics in order to identify threat patterns and predict potential sources of domestic terrorism.

The primary mission of DHS is, among other things, to prevent terrorist attacks within the United States, reduce the vulnerability of the United States to terrorism, minimize the damage and assist in the recovery from terrorist attacks that do occur within the United States, support the missions of its legacy components, monitor connections between illegal drug trafficking and terrorism, coordinate efforts to sever such connections, and otherwise contribute to efforts to interdict illegal drug trafficking.

During the Cerberus Pilot the system ingests certain DHS component unclassified data about individuals and maintains the data in a DHS owned and controlled cloud computing environment on a Top Secret/Sensitive Compartmented Information (TS/SCI) network, where it is available for classified searches and evaluation using various analytical tools

III. Challenges of Big Data

- **Data Presentation:** Big data includes traditional structured data as well as semi-structured and unstructured data such as audio, video, GIF, genomic data, network data, space data, etc. Because of these variety of data, it has become difficult for the organizations to represent the useful data for the analysis. The primary purpose of data presentation is to make data more meaningful for the computer analysis and user interpretation. Lack of meaningful data will reduce the value of the original data and may even greatly affect the data analysis.

- **Redundancy reduction and data compression:** Data generated by sensor network possess high level of data redundancy. Redundancy reduction and data compression compresses this redundant data in such that the potential value of the data are not affected.

- **Data Life Cycle Management:** Big Data is generating at such a rapid velocity that the current storage system could not support such a massive amount of data. Therefore, a data importance principle related to the analytical value should be developed to decide which data shall be stored and which data shall be discarded.

- **Analytical Mechanism:** Non-relational databases are proved advantageous in the processing of unstructured data and started to become mainstream in big data analysis. But still it has some limitations. We shall find a compromising solution between RDBMSs and non-relational databases such that system has ability to utilize a mixed database architecture that integrates the advantages of both types of database.

- **Data confidentiality:** Due to huge volume of the big data, the service providers can't maintain and analyze them effectively and have to rely on tools to analyze such data. Which increase the potential safety risks. Such data contains details of the lowest granularity and some sensitive information such as credit card numbers. Therefore, analysis of big data may be delivered to a third party for processing only when proper preventive measures are taken to protect such sensitive data, to ensure its safety.

- **Energy management:** Because of increase of data volume and analytical demands energy consumption of mainframe computing systems has increased to a great extent. As a result of which, it becomes necessary to develop a system-level power consumption control and management mechanism while the expandability and accessibility are ensured.

- **Expendability and scalability:** Due to variety characteristic of the big data, the analytical algorithm must be able to process increasingly expanding and more complex datasets.

- **Cooperation:** As big data has data of various types, sometimes its management and analysis requires experts in different fields cooperate to cooperate to complete the analytical objectives. One of the solution can be we can develop a comprehensive big data network architecture to help scientists and engineers in various fields access different kinds of data and fully utilize their expertise to complete the analytical objectives.

IV. Storage for Big Data

Q1. What factors should you take into account when using distributed storage for Big Data?

Ans. Following should be taken into account when using distributed storage for big data:

- **Consistency:** Data stored in a distributed storage system is divided into multiple pieces to be stored at different servers to ensure availability in case of server failure. However, server failures and parallel storage may cause inconsistency among different copies of the same data. A distributed storage system must ensure that the data stored at multiple servers of the system must be identical.

- **Availability:** A distributed storage system must be such that even if the server failures it must not affect the entire system and satisfy customer's requests in terms of reading and writing.

- **Partition Tolerance:** This property ensures that even if any network partition occurs, the distributed storage system still works well. The distributed system should have a certain level of tolerance that it persists link/node failures or temporary congestion.

Fill in the blanks

1. Hadoop is top level **Apache** project written in **Java** programming language.

2. Hadoop was inspired by **Google's work on Google File System and MapReduce programming paradigm.**

3. Hadoop is different from transactional system in the following ways:

   Hadoop is designed to scan through large data sets to produce its results through a highly scalable, distributed batch processing system. Hadoop is not about speed-of-thought response times, real time warehousing, or blazing transactional speeds; *it is* about discovery and making the once near-impossible possible from a scalability and analysis perspective.

4. Two parts of Hadoop are:
   File system (Hadoop Distributed File System)
   Programming paradigm (MapReduce)

5. Why is redundancy built into Hadoop environment?

   Not only is the data redundantly stored in multiple places across the cluster, but the programming model is such that failures are expected and are resolved automatically by running portions of the program on various servers in the cluster.

Components of Hadoop

1. The three pieces of Hadoop project are:

   Hadoop Distributed File System (HDFS)
   Hadoop MapReduce model,
   Hadoop Common.

Hadoop Distributed File System

1. How is it possible to scale Hadoop cluster to hundreds of nodes?

Ans. Data in a Hadoop cluster is broken down into smaller pieces (called *blocks*) and distributed throughout the cluster. In this way, the map and reduce functions can be executed on smaller subsets of your larger data sets, and this provides the scalability that is needed for Big Data processing.

2. Each server in a Hadoop cluster uses **inexpensive** disk drives.

3. What is data locality? What does it achieve?

Ans. The goal of Hadoop is to use commonly available servers in a very large cluster, where each server has a set of inexpensive internal disk drives. For higher performance, MapReduce tries to assign workloads to these servers where the data to be processed is stored. This process is known as data locality.

4. What are the benefits of breaking a file into blocks and storing these blocks with redundancy?

5. The default size of a block in HDFS is **64** MB.

6. What are the advantages of large block sizes in HDFS?

Ans. Hadoop was designed to scan through very large data sets, so it makes sense to use a very large block size so that each server can work on a larger chunk of data at the same time. For larger files, a higher block size will greatly reduce the amount of metadata required by the NameNode.

7. What is a NameNode in HDFS? What are its functions?

Ans. NameNode is a special server that manages all of the Hadoop's data placement logic. This NameNode server keeps track of all the data files in HDFS, such as where the blocks are stored, and more. All of the NameNode's information is stored in memory, which allows it to provide quick response times to storage manipulation or read requests.

8. All of NameNode's information is stored in **Memory**.