CS 6350.002 S22: Big Data Management and Analytics

BIG DATA INTRODUCTION

Submitted by Venkata Kowsik Temididapathi (VXT200001)

I. What is Big Data

Section 1.1 of the paper:

1. What does the term Big Data (BD) refer to? How is BD different from traditional datasets?
Big data is generally used to describe enormous dataset. Big data typically includes masses of unstructured data that need more real-time analysis.

2. What challenges have emerged because of the rise of BD?
The challenges that have emerged because of the rise of Big Data are:

- We must collect and integrate massive data from widely distributed data sources.
- Since the data is increasingly growing, we must find ways to store and manage such huge datasets with the present hardware and software infrastructure.
- Since the data is unstructured, we have to mine the data to find better understanding to improve the decision making.

Section 1.2 of the paper:

1. 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.

- Big data shall mean the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time.
- Datasets which could not be acquired, stored, and managed by classic database software.
- Datasets volumes that conform to the standard of big data are changing and may grow over time or with technological advances.
- Datasets volumes that conform to the standard of big data in different applications differ from each other.
- Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis.
- Big data shall mean the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies.

- The 3V's of Big Data:
    - Volume means, with the generation and collection of masses of data, data scale becomes increasingly big.
    - Velocity means the timeliness of big data, specifically, data collection and analysis, etc. must be rapidly and timely conducted, to maximumly utilize the commercial value of big data.
    - Variety indicates the various types of data, which include semi-structured and unstructured data such as audio, video, webpage, and text, as well as traditional structured data.

Characteristics of BD (Chapter 1 of book):

Read and understand the 3V character of BD. Answer the following questions:

1. What is meant by volume of BD. How has it changed over time?
The amount of data that is created. For the year 2020, 40 ZB of data was created which is 300 times increase from the data in 2005.

2. How has increased volume created a "blind zone" for organizations?
The "Blind zone" states that the percent of data that is useful is decreasing. But to find if a data is useful, you must see it. In the blind zone, it is unknown if the data is useful or not.

3. What is meant by variety of BD? What are the various types of data that large organizations acquire today?
Variety represents all types of data (structured, semi structured and unstructured).
The various types of data that are acquired are:
Twitter data, weather pattern data, audio data, and video data.

4. How is velocity of data applied to data in motion. What are the advantages of streams computing?
The speed at which the data is flowing. We have to analyze and store this data before another set of data is received.
The advantages of stream computing are to get results as the data is in flow. The results keep updating based on the new data.

II. What is the value of Big Data

Section 1.3 of the paper and chapter 2 of the book

1. Read section 1.3 of the paper and chapter 2 of the book. They list several industries (e.g. US medical industry, retail industry, government operations, public health, etc) that can benefit enormously by using Big Data techniques. Choose any one such industry and do research about Big Data applications in that industry. Write a brief 2-3 paragraph report.

Big data is useful in many industries, one of which is the retail industry. In retail industry, we can use big data to optimize pricing, supply chain movement, and improve customer loyalty. It can be used to predict the patterns/trends of the purchases. It will also help in the attraction of new customers. It is also useful to create recommendations and to improve the customer service.

In addition to big data, some companies use social media data and web browsing data to predict the next big thing. Another example is the use of weather to forecast the demand of a product. Brands like Walgreens and Pantene worked with the Weather Channel to account for weather patterns in order to customize product recommendations for consumers. For example, with an increase in humidity, the use of anti-frizz products will increase. So, to capitalize on this, Walgreens served up ads and in-store promotions to drive sales. There was an increase in 10% of sales on the Pantene products over 2 months. So, retail forecasting and retail projections are used to properly allocate their resources the most effectively throughout different parts of the year.

We know that big e-commerce companies use recommendations to make more sales. Amazon uses customer data to recommend items for you based on your past searches and purchases. This is known to most of us, but this generates 29 percent of sales through their recommendations engine which analyzes more than 150 million accounts. This has led to big profits for the company.

III. Challenges of Big Data

Section 1.5 of the paper

1. Read section 1.5 of the paper and summarize in your own words the challenges of developing and managing Big Data applications.

The key challenges of developing and managing Big data applications are:

- Data Representation: It makes the data more meaningful and interpretable for the data analyst. Improper data representation makes the data useless and meaningful.
- Redundancy reduction and data compression: Generally due to the availability of data, there is a lot of redundancy on the data. Since more data use space and resources, it is better to compress or remove some of the redundancy data.
- Data life cycle management: Generally, the value of a data from a system is the freshness of the data. So, we need to know which data is useful and which is not.
- Data confidentiality: Many companies are able to receive and store the data, but they are sent to an external company for inspection. This causes a lot of security issues.
- Energy management: The extensive use of data uses a lot of energy to maintain this data.

IV. Storage for Big Data

Section 4.2 of the paper

1. What factors should you take into account when using distributed storage for Big Data?
Consistency, Availability and Partition Tolerance are the factors when using distributed storage of big data.
Consistency refers to assuring that multiple copies of the same data are identical.
Availability refers to accepting customer's requests of reading and writing.
Partition Tolerance is that it is desirable that the distributed storage still works well when the network is partitioned.

Chapter 4 of the book

Fill in the blanks / Short answer questions:

1. Hadoop is top level ___Apache___ project written in __Java___ programming language.

2. Hadoop was inspired by __Google's work on the Google File System (Distributed) and Map Reduce programming paradigm__ .

3. Hadoop is different from transactional systems in the following ways:
Unlike transactional systems, Hadoop is designed to scan through large data sets to produce its results through a highly scalable, distributed batch processing system.

4. Two parts of Hadoop are:
A file system (the Hadoop Distributed File System) and a programming paradigm (MapReduce)

5. Why is redundancy built into Hadoop environment?
In the Hadoop environment, data is redundantly stored in multiple places throughout the cluster. This redundancy provides a fault tolerance and the capability for the Hadoop cluster to correct itself. This also allows Hadoop to scale out some of its workload across large clusters.

Components of Hadoop:

1. The three pieces of Hadoop project are:
Hadoop Distributed File System (HDFS), the Hadoop MapReduce model, and Hadoop Common.

Hadoop Distributed File System:

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

The data in the cluster is broken into smaller pieces and are distributed in the cluster. Whereas the map and reduce functions are executed on smaller subsets thus providing scalability, which is necessary for big data processing.

2. Each server in a Hadoop cluster uses __inexpensive___ (inexpensive / expensive) disk drives.

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

Data locality in Hadoop is the process of moving the computation close to where the actual data resides instead of moving large data to computation. MapReduce tries to assign workload to these servers where the data is stored. This minimizes overall network congestion. This also increases the overall throughput of the system.

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

The benefits are higher availability, better scalability, and data locality.

5. The default size of a block in HDFS is ____64______ MB.

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

The ability to work on large chunk of data in a single server is one of the advantages. Since data is not sent to other nodes, it improves both the performance and the overhead to real work ratio.

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

Hadoop’s data placement logic is managed by a special server called NameNode. It keeps track of the data files and where the blocks are stored. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories.

8. All of NameNode's information is stored in ___memory_____ (disk / memory).