Lam Phan

CS4301.5U1

Dr. Nagar

05/30/2018

Homework 1

//I have highlighted my answer for faster readability.

1. **What is big data?**

Section 1.1

1. What does the term Big Data (BD) refer to?

The term big data prefer to information that cannot be processed or analyzed using traditional processes or tolls.

How is BD different from traditional datasets?

Big Data can be raw form, semi-structured, unstructured that cannot be stored in traditional systems because traditional systems can struggle to store and perform the required analytics to gain understanding from these logs since much of the information being generated doesn’t lend itself to traditional database technologies.

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

a) Organizations do not know how to get value out of huge information they have access to because the information are in its raw form or in a semi-structured or unstructured format. The result is that the organizations do not know if the information is worth keeping or not.

b) Half of the business leaders today realize they do not have access to the insights they need to do their jobs according to the IBM survey.

Section 1.2

1. What is meant by **volume** of BD. How has it changed over time?
2. Volume is one of the characteristics of Big Data. It is exploding and changing over time.
3. The volume has changed over time by companies, organizations and the people since we all generating data and store everything such as: environment data, financial data, surveillance data, etc. some enterprises even generate terabytes of data every hour of everyday of the year.

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

The volume of data available to organizations today is on the rise, while the percent of data they can analyze is on the decline therefore the issue created a blind zone for organizations.

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

a) Variety of Big Data could mean complex since it represents all types of data from traditional structured data to raw, semi-structured, and unstructured data.

b) The various types of data that large organizations acquire today are: raw, semi-structured, and unstructured data.

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

a) Velocity of data applied to data in motion by the speed at which the data is flowing.

b) The advantages of streams computing:

* We can execute a process similar to continuous query that identifies people who are currently in a specific place, but we get continuously updated results because location information from GPS data is refreshed in real time.

1. **What is the value of Big Data?**
2. 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.

In 2011, IBM’s successfully created Watson a technology that allow a computer to process and understand natural language. Watson also implemented a deep-rooted learning behavior that understand correct and incorrect decisions, and it could even apply risk analysis to future decisions and domain of knowledge. Watson use Hadoop to store and pre-process its corpus of knowledge.

Watson incorporates massively parallel analytical meaning behind words, distinguish between relevant and irrelevant content, and demonstrate confidence to deliver precise final answer. It has the ability of the discovery of understanding and insight not just processing data.

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

We must be able to analyze the data in near real time if we hope to find insights in the data.

We need to consider where the data should be stored, and how data is store.

1. **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?

Redundancy: because it offers multiple benefits. It has higher availability. It allows Hadoop cluster to break work up into smaller chunks and run those jobs on all the servers in the cluster for better scalability. It gives the benefit of data locality which is critical when working with large data sets.

Chapter 4 of the book One of the most popular distributed storage mechanisms for Big Data is Hadoop. Chapter 4 of the book presents a very good introduction to it. 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 File System and the MapReduce Programming paradigm\_\_\_\_\_\_\_\_\_\_\_\_ .

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

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