BIG DATA INTRODUCTION

Reading sources:
1. Book - Understanding Big Data by IBM
http://www.utdallas.edu/~axn112530/cs6350/Understanding_BigData.pdf

2. Paper
Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. *Mobile Networks and Applications*, *19*(2), 171-209.
http://mmlab.snu.ac.kr/~mchen/min_paper/BigDataSurvey2014.pdf
or from the ACM Digital Library http://dl.acm.org/citation.cfm?id=2843712

I. What is Big Data

Answer the following questions by reading the sources mentioned.

Section 1.1 of the paper:

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

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

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.

Here is a start:

- Datasets which could not be captured, managed, and processed by computers within a reasonable time frame [Hadoop]

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?

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

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

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

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 <u>any one such industry</u> and do research about Big Data applications in that industry. Write a brief 2-3 paragraph report.

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.

IV. Storage for Big Data

We will spend a significant amount of time discussing the storage mechanism of Big Data, so it's good to be familiar with the storage mechanism for Big Data.

Section 4.2 of the paper

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

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 ______________ project written in ______________ programming language.

2. Hadoop was inspired by _______________ .

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

4. Two parts of Hadoop are:

5. Why is redundancy built into Hadoop environment?

Components of Hadoop:

1. The three pieces of Hadoop project are:

Hadoop Distributed File System:

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

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

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

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 __________ MB.

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

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

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