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?

Ans.) Large and complicated data volumes that are difficult to handle or process with conventional data processing tools are referred to as "big data." The three Vs: volume, variety, and velocity, define it. The terms volume, variety, and velocity describe the sheer amount of data, the many sorts of data, and the speed at which the data must be processed. Big data has grown in significance as businesses look to mine the enormous volume of data being created daily for insightful information. It can be used in a variety of domains, such as science, technology, and business.

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

Ans.) Big Data presents a number of challenges, including handling the volume, diversity, and velocity of data, as covered in the provided text. It also highlights how important it is to connect Big Data technologies with current enterprise infrastructure and extract value from the data.

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]

- Big data is the next frontier in productivity, competitiveness, and innovation.
- Big data is similar to data collections that conventional databases are unable to collect, store, or manage.
- A 3V model defines big data.(Amount, Speed, and Diverseness)
- New Origination for technologies that are intended to efficiently extract values from vast amounts of various data types through high-velocity capture, discovery, and analysis.
- Data whose volume, processing speed, and representation make it impossible to examine using conventional techniques.

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?

Ans.) Volume refers to the amount of data and this volume is exploding in the organization because Volumes of data has changed from Terabytes to Petabytes, and it is going at rapid increase.

2. How has **increased volume** created a "blind zone" for organizations?
Ans.) Organizations are now in a "blind zone" as a result of the growing volume of data; while the enterprise has access to more data than ever before, the proportion of data that it can process, comprehend, and analyze is decreasing. Organizations may be missing out on important insights as a result of this condition, which has created a blind zone where they are unaware of what they do not know. Terabytes to petabytes and eventually zettabytes are the terms used to describe data sizes that are no longer compatible with conventional systems. Many organizations today are overwhelmed by the sheer amount of data being kept, and those who are ill-equipped to handle this data find themselves swamped.

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

Ans.) Means Bigdata includes too many different types of data in the datasets. There are 3 types of data: Raw, semi-structured and unstructured.

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

Ans.) For data, which is in motion, velocity refers to how quickly arrived data is processed and analyzed and put back to the motion. I think the usage of stream computing allows organizations to process and analyze the data in a real-time manner.

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.

Ans) Let's take an example from the social media industry, specifically focusing on Instagram. Instagram is a popular photo and video sharing platform that utilizes advanced algorithms and big data to deliver personalized content to its users. While there are other social media platforms like Facebook and Twitter, Instagram stands out for its visually engaging interface and seamless user experience.

One of Instagram's standout features is its Explore page, where users can discover new content tailored to their interests based on their previous interactions and behaviors on the platform. Similar to Spotify's personalized recommendations, Instagram analyzes users' interactions with posts, stories, and accounts to suggest content that aligns with their preferences.

Moreover, Instagram's algorithm continuously learns from users' engagement patterns to refine its recommendations over time, ensuring that users are exposed to content that resonates with them. This personalized approach has contributed to Instagram's growing user base and increased user engagement.

Another noteworthy feature on Instagram is the annual "Instagram Year in Review" or "Instagram Highlights," which provides users with a summary of their activity on the platform over the past year. This feature utilizes data analysis and big data techniques to curate personalized highlights, such as the user's most liked posts, top hashtags, and most interacted-with accounts.

By leveraging big data, Instagram enhances user satisfaction by delivering relevant content and personalized experiences tailored to individual preferences. This emphasis on user-centric features and data-driven insights contributes to Instagram's continued success and popularity among social media users.

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.

Ans)The challenges of data representation, redundancy reduction with data compression, data life cycle management, analytical mechanisms, data confidentiality, energy management, expendability, scalability, and cooperation are present in today's BD app development and management.

Conventional Relational database management systems (RDBMS) are overly dependent on costly technology and are unable to manage large volumes of diverse data. Eventually, cloud computing will allow us to get around this choice. Furthermore, NoSQL and distributed file systems are beneficial.

Proper data representation is essential to avoid undervaluing the original data and perhaps leading to incorrect analysis. Additionally, decide how to store important data for analysis. Sometimes using relational and non-relational databases may give good results in big data analysis. Data should be kept protected too. Making use of some energy reduction helps to make the cost lower. Creation of common architecture should be there to do research and analyze the huge varieties of data.

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?

Ans.) Consistency, Availability and Partition Tolerance

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 a top level _____ project written in _____ programming language.

Ans.) Apache, Java

2. Hadoop was inspired by _____ .
Ans.) Google's GFS and MapReduce programming Paradigm

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

Ans.) Hadoop is designed to produce results from highly scalable system like Distributed batch processing systems and Hadoop is built around a function-to-data model.

4. Two parts of Hadoop are:
HDFS and MapReduce

5. Why is redundancy built into the Hadoop environment?

Along with data stored in multiple places across the cluster and programming model is also containing failures in such cases and those failures will be resolved so that's why Hadoop is redundancy built.

Components of Hadoop:

1. The three pieces of Hadoop project are: HDFS, Hadoop MapReduce model and Hadoop Common

Hadoop Distributed File System:

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

Ans.) In a Hadoop cluster, Data is broken down into smaller pieces(Blocks), Followed by distributing these blocks into the cluster.

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

Ans.) Inexpensive

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

Ans.) MapReduces tries to assign workloads to the available servers where data which is about to process is being stored. Maybe Storage Area network/ Network attached storage, not sure :)

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

Ans.) In-built Fault tolerance and Fault compensation capabilities, Higher availability. In addition, redundancy allows to break work into smaller chunks.

5. The default size of a block in HDFS is _____ MB.
Ans.) 64 MB

6. What are the advantages of large block sizes in HDFS?
Ans.) To reduce the amount of metadata required by the NameNode.

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

Ans.) Data placement logic is managed by a special server titled NameNode in HDFS. It keeps track of data files like Blocks are stored in which locations in HDFS. NameNode's info is stored in memory, so it allows quick response time to storage manipulation/ Read requests.

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

Ans.) Memory