**Thien Dang Nguyen**

**DXN210021**

**CS6350**

Reading lab1

**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?**

The term Big Data applies to information that can’t be processed or analyzed using traditional processes or tools

Differences from Traditional Datasets:

* Traditional datasets are typically structured and stored in relational databases (RDBMS), which efficiently handle structured data.
* Big Data includes semi-structured and unstructured data such as audio, video, webpages, and text.
* Traditional databases manage data using fixed schemas and indexed queries, whereas BD requires new architectures for fast processing.

1. **What challenges have emerged because of the rise of BD?**

* Data Acquisition: The ability to collect massive data from diverse sources (IoT, sensors, social media)​
* Storage & Management: Traditional databases struggle with scalability due to hardware and software limitations​
* Real-Time Processing: The increasing velocity of data generation demands near-instant analytics (e.g., fraud detection in banking)​
* Privacy & Security Risks: The large-scale collection of personal and sensitive data introduces risks that traditional security measures cannot fully mitigate​

**Section 1.2 of the paper:**

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

**Key Features of BD:**

* **Datasets that cannot be processed reasonably using traditional systems**​.
* **Highly scalable storage and processing models are required** (e.g., distributed computing, cloud-based storage).

**The 3Vs Model (Doug Laney, 2001)**​:

* **Volume:**
  + Refers to the massive scale of data being generated.
  + Data volume grows **exponentially** over time (e.g., daily TBs to PBs of data from IoT, social media, and enterprise applications).
* **Velocity:**
  + Describes the **speed** at which data is generated and processed.
  + Example: Stock market transactions and IoT sensors produce high-speed data streams.
* **Variety:**
  + Represents the **different types of data**—structured (databases), semi-structured (JSON, XML), and unstructured (videos, images, social media posts).
  + Traditional databases handle structured data, but BD must process text, images, audio, and real-time sensor data.

**Extended 4Vs Model (IDC, 2011)**​:

* **Value:** Extracting meaningful insights from raw data for decision-making.
  + Example: Social media analytics providing consumer sentiment insights.

**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 volume of Big Data refers to the sheer scale of data being generated, stored, and processed by organizations.
* The amount of data were stored is increased day by day:
  + In the year 2000, the total amount of data stored worldwide was 800,000 petabytes (PB).
  + By 2020, this number was projected to reach 35 zettabytes (ZB)
  + Social media and enterprise applications significantly contribute to this growth, with Twitter generating 7 terabytes (TB) of data per day and Facebook generating 10 TB per day​.

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

* As data volume increases, organizations struggle to process, analyze, and extract insights from all available data, creating what is known as the "blind zone"​
* This means the percentage of data that organizations can effectively analyze is decreasing, leaving valuable insights hidden within the data.
* The blind zone creates both risks and missed opportunities:
  + Organizations may not detect important trends in customer behavior.
  + Cybersecurity threats or financial fraud may go unnoticed due to a lack of data visibility.
  + Valuable insights remain untapped, reducing competitive advantages.

1. **What is meant by the variety of BD? What types of data do organizations acquire today?**

* Variety refers to the different types and formats of data that organizations collect.
* Variety refers Types of data acquired today:
  + Structured Data: Relational databases, financial transactions, ERP systems.
  + Semi-Structured Data: JSON, XML, log files, emails.
  + Unstructured Data:
    - Text (social media posts, customer reviews, blogs).
    - Multimedia (images, videos, audio recordings).
    - Sensor Data (IoT, weather tracking, surveillance cameras).
    - Web Data (clickstreams, browsing history, search engine logs). to the different types and formats of data that organizations collect.
* 80% of today’s enterprise data is unstructured.

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

* Velocity refers to the speed at which data is generated, processed, and analyzed​
* Examples of data in motion:
  + Financial transactions: Fraud detection systems analyze transactions as they happen.
  + IoT sensors: Smart cities track real-time traffic data to optimize traffic lights​.
  + Stock markets: Trading algorithms analyze live market data to make instant decisions.
* Advantages of stream computing (IBM InfoSphere Streams)​:
  + Real-time insights: Organizations react instantly to trends, security threats, and opportunities.
  + Scalability: Handles millions of events per second with massively parallel processing.
  + Continuous data flow: Unlike batch processing, streaming provides a constant, up-to-date analysis of changing data.
  + Reduced storage costs: Irrelevant data is discarded before being saved, reducing unnecessary storage expenses.

**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.**

* U.S. Healthcare Industry
  + Potential Value: Big Data could generate over $300 billion in value, reducing U.S. healthcare expenditure by 8%​
  + Applications:
    - Predictive analytics for early disease detection and personalized treatment plans.
    - Real-time patient monitoring through wearable sensors and IoT devices.
    - Fraud detection in insurance claims​
  + Example:
    - Google Flu Trends used Big Data analytics on search queries to track influenza outbreaks faster than CDC reports
* Retail Industry
  + Potential Value: Retailers using Big Data can increase profits by over 60%​
  + Applications:
    - Personalized recommendations based on customer browsing and purchase history.
    - Dynamic pricing strategies using real-time market demand.
    - Supply chain optimization to reduce inventory waste.
  + Example:
    - Amazon processes millions of customer interactions daily to improve its product recommendations and logistics efficiency
* Government and Public Sector
  + Potential Value: Efficient use of Big Data could save over €100 billion in developed European economies, excluding fraud and tax savings​
  + Applications:
    - Smart city initiatives to improve water management, traffic flow, and energy use​
    - Crime prediction and prevention through data-driven law enforcement.
    - Tax fraud detection using AI-powered analytics.
  + Example:
    - Miami-Dade County & IBM: Big Data analytics helped detect water leaks, reducing water bills by $1 million
* Financial Services & Fraud Detection
  + Applications:
    - Real-time fraud detection in banking transactions.
    - Customer credit risk assessment based on transaction history.
    - Algorithmic trading using market data to predict stock trends.
  + Example:
    - IBM InfoSphere Streams helps banks detect fraudulent transactions within milliseconds​
* Transportation & Logistics
  + Applications:
    - Route optimization to reduce fuel consumption and delivery time.
    - Predictive analytics for vehicle maintenance scheduling.
    - Traffic congestion prediction for smart city planning.
  + Example:
    - UPS uses Big Data analytics to optimize delivery routes, saving millions of miles and reducing fuel costs

Report about US. Health Industry:

The U.S. healthcare industry has embraced Big Data analytics to transform patient care, operational efficiency, and medical research. Healthcare providers and organizations utilize vast amounts of data from electronic health records (EHRs), medical imaging, genetic sequencing, wearable devices, and insurance claims to derive actionable insights. These data sources, when analyzed through advanced algorithms and machine learning models, help in early disease detection, personalized treatment planning, and population health management.

One of the most impactful applications of Big Data in healthcare is predictive analytics for patient outcomes. Hospitals and healthcare systems analyze historical patient data to identify individuals at high risk for various conditions, enabling preventive interventions before health issues become severe. For example, some healthcare organizations use predictive models to forecast patient readmission risks, allowing care teams to adjust discharge planning and follow-up care accordingly. Additionally, Big Data analytics supports precision medicine initiatives by analyzing genetic information alongside clinical data to determine the most effective treatments for individual patients based on their specific genetic makeup and medical history.

**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.**

* Data representation: Datasets have varying levels of heterogeneity in type, structure, semantics, organization, granularity, and accessibility. Proper data representation is crucial to make data meaningful for computer analysis and user interpretation. Poor representation can reduce the value of the original data and hinder effective analysis.
* Redundancy reduction and data compression: There is typically high redundancy in datasets. Effective redundancy reduction and compression techniques are needed to reduce system costs while preserving the potential value of the data. For example, sensor network data often contains significant redundancy that can be filtered and compressed.
* Data lifecycle management: The rapid growth of data generation far outpaces advances in storage systems. This creates challenges in deciding what data to store versus discard based on its analytical value. A data importance framework needs to be developed to manage the lifecycle of data based on its value over time.
* Analytical mechanism: Big data systems need to process massive heterogeneous data within limited timeframes. Traditional relational databases lack the scalability needed. While non-relational databases show advantages for unstructured data, they still have limitations. A hybrid approach combining aspects of both may be needed.
* Data confidentiality: Many organizations lack the internal capability to effectively analyze huge datasets and must rely on third parties, creating potential security risks. Sensitive data needs to be properly protected when shared for analysis.
* Energy management: Processing, storing and transmitting big data consumes significant energy. System-level power management mechanisms are needed while maintaining scalability and accessibility.
* Expendability and scalability: The analytical systems must be able to handle continuously expanding datasets with increasing complexity.
* Cooperation: Big data analysis requires interdisciplinary collaboration between experts in different fields. A comprehensive architecture needs to be established to help scientists and engineers access and analyze data effectively.

**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?**

* Consistency
  + Multiple servers need to cooperatively store data, which increases the probability of server failures
  + Data is typically divided into multiple pieces and stored across different servers for availability
  + Server failures and parallel storage can cause inconsistency between different copies of the same data
  + Need to ensure multiple copies of the same data remain identical
* Availability
  + The distributed storage system operates across multiple sets of servers
  + When more servers are used, server failures become inevitable
  + The system needs to remain functional and satisfy customer requests for reading and writing data even when some servers fail
  + The entire system should not be seriously affected by individual server failures
* Partition Tolerance
  + Servers in a distributed storage system are connected by a network
  + Networks can experience link/node failures or temporary congestion
  + The distributed system needs to maintain a certain level of tolerance to problems caused by network failures
  + The system should continue working effectively even when the network is partitioned

**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’s work on the Google File System (GFS) and the MapReduce programming model​.

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

* Hadoop is designed for large-scale batch processing, while transactional systems handle real-time transactions.
* Hadoop processes massive amounts of unstructured or semi-structured data, while transactional systems typically work with structured data in relational databases.
* Hadoop follows a function-to-data model, meaning it moves computation to where data is stored rather than bringing data to the computation

**4. Two parts of Hadoop are:**

* Hadoop Distributed File System (HDFS) – A scalable, fault-tolerant storage system.
* MapReduce – A programming model for distributed data processing

**5. Why is redundancy built into Hadoop environment?\**

* Redundancy is built into Hadoop to provide fault tolerance and high availability.
* Data is replicated across multiple nodes, ensuring that failure of one or more nodes does not result in data loss​.
* This redundancy also improves data locality, enabling Hadoop to distribute workloads efficiently across the cluster.

**Components of Hadoop:**

**1. The three pieces of Hadoop project are:**

* Hadoop Distributed File System (HDFS) – A distributed storage system.
* Hadoop MapReduce – A programming model for distributed data processing.
* Hadoop Common – The set of shared libraries and utilities required by other Hadoop modules​

**Hadoop Distributed File System:**

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

* Hadoop divides data into smaller blocks and distributes them across multiple nodes.
* The MapReduce model processes these blocks in parallel, ensuring scalability.
* Each node performs computation where data is stored, minimizing network overhead​

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

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

* Data locality means that Hadoop processes data on the same node where it is stored.
* This reduces network traffic, increases processing speed, and improves efficiency​

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

**redundancy?**

* Fault Tolerance: If one node fails, data is still available from replicated blocks.
* Parallel Processing: Hadoop can process different parts of a file simultaneously.
* Scalability: Large files can be handled without a single machine running out of storage​

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

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

* Faster Processing: Reduces the number of disk seeks, making data reads more efficient.
* Lower Metadata Overhead: The NameNode manages fewer blocks, reducing memory usage.
* Better Performance: Large block sizes suit workloads with sequential reads​.

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

* The NameNode is the master node in HDFS that manages metadata about file locations.
* Functions:
  + Tracks where each data block is stored.
  + Directs clients to the correct DataNodes for data access.
  + Ensures data replication for fault tolerance​

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