Kushal Choudhary (kxc240000)

CS 6350: Reading Lab Assignment 1

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**

1. 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 refers to large, complex, and fast-growing datasets that cannot be handled by traditional data tools. It differs from traditional datasets in three major ways:

- Volume: Enormous size (TBs to PBs)
- Velocity: Real-time or fast streaming data
- Variety: Structured, semi-structured, and unstructured formats

Traditional datasets are structured and can typically fit into relational databases. They are relatively static and limited in volume, velocity, and variety.

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

Key challenges include:

- Storage & Scalability: Managing huge, growing datasets
- Processing Speed: Real-time analysis is hard with traditional systems
- Integration: Combining varied data types (text, video, logs)
- Privacy & Security: Protecting sensitive data
- Data Quality & Skills Gap: Cleaning noisy data and hiring experts

(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.

1. Big Data refers to datasets that traditional database systems cannot capture, manage, and process within a reasonable time frame.
2. Doug Laney's 3V Model:
    - Volume: Refers to the sheer amount of data generated from various sources (e.g., social media, sensors, logs).
    - Velocity: Refers to the speed at which new data is generated and moves around (e.g., real-time processing of tweets or stock trades).
    - Variety: Refers to the different formats of data (structured, unstructured, semi-structured like text, images, audio, video, etc.).
3. Additional features include:
    - Complexity: Due to the heterogeneous and interrelated nature of data.
    - Scalability Issues: Existing systems struggle with the speed and scale at which data grows.

- Not just size, but also difficulty: The challenge is not only in storage but in how to analyze and extract value from the data.

4. The term Big Data is relative, what qualifies as "big" changes over time as technology evolves.

Characteristics of Big Data
(Chapter 1 of book)

1. **What is meant by volume of BD. How has it changed over time?**
 Volume refers to the huge amount of data being generated every second. Over the years, this has grown from gigabytes to terabytes, petabytes, and even more. The growth comes from sources like mobile apps, social media, online transactions, sensors, and connected devices. As more people and systems go digital, the volume keeps growing rapidly.

2. **How has increased volume created a "blind zone" for organizations?**
As the volume of data increases, organizations often find themselves overwhelmed by the sheer amount of information. This leads to a "blind zone," where valuable insights remain hidden simply because companies lack the tools, infrastructure, or skills to process and analyze all the data. They might only analyze a small portion and ignore the rest, which can cause missed opportunities or even bad decisions due to incomplete views.

3. **What is meant by variety of BD? What are the various types of data that large organizations acquire today?**
Variety refers to the different types and formats of data that organizations deal with. Unlike traditional structured data (like tables in relational databases), today's data comes in many forms, emails, social media posts, images, audio files, videos, GPS signals, clickstreams, and sensor data. These can be structured, unstructured, or semi-structured. Big Data systems are built to handle this variety, making it possible to gather insights from sources that were previously hard to process.

4. **How is velocity of data applied to data in motion. What are the advantages of streams computing?**
Velocity describes the speed at which data is generated, transmitted, and processed. With the rise of IoT devices, real-time social media, and live transaction systems, data now flows continuously rather than being collected in batches. This is where "data in motion" comes in, it means handling data while it's being created, not afterward. Stream computing is useful here because it allows for real-time analysis and decision-making. For example, fraud detection in banking, live traffic updates, or instant customer feedback processing all rely on fast data handling through stream computing.

2. 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.
The healthcare industry has seen a transformative impact through the adoption of Big Data. Traditional health systems often relied on paper records and manual reporting, making it difficult to quickly analyze patient outcomes or detect patterns. However, with the rise of electronic health records (EHRs), wearable devices, and real-time patient monitoring, enormous volumes of structured and unstructured data are now generated every day. Big Data techniques allow this data to be mined and analyzed for patterns that can improve diagnostics, optimize treatment plans, and prevent diseases.

For example, predictive analytics models can identify patients who are at risk of chronic conditions like diabetes or heart failure even before they show major symptoms. By analyzing historical data combined with genetic, lifestyle, and environmental factors, healthcare providers can offer personalized treatment and early interventions. Hospitals also use Big

Data to optimize operations, such as reducing wait times in emergency rooms by forecasting patient inflow and optimizing staffing schedules. Public health agencies can track and respond to disease outbreaks more effectively by monitoring real-time data streams from clinics and social media.

Big Data doesn't just improve clinical outcomes, it also brings down costs. As highlighted in both Chapter 2 of the book and Section 1.3 of the survey paper, the U.S. healthcare system, which has historically suffered from inefficiencies and fragmented data, can potentially save hundreds of billions of dollars annually through better data integration, fraud detection, and evidence-based decision-making. The ability to process massive datasets quickly and accurately helps doctors, researchers, and policymakers make informed decisions that ultimately lead to a healthier population and a more efficient healthcare system.

3. 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.

- Developing and managing Big Data applications is tough because of the nature and complexity of the data involved. One of the biggest challenges is data integration. Data comes from different sources like social media, sensors, mobile devices, and web logs. These are often in different formats, making it hard to combine them into one system that can be analyzed properly.
- Another issue is data quality and preprocessing. A lot of the data is noisy, messy, incomplete, or inconsistent. Before it can be analyzed, it needs to be cleaned and processed. This step is time-consuming but necessary to get meaningful insights.
- Storage and scalability are also major concerns. Traditional databases can't handle the massive size of Big Data, so distributed systems like HDFS are used. But managing these systems requires a strong understanding of distributed computing and infrastructure.
- Security and privacy are big concerns too. Since Big Data often includes sensitive information like medical or financial records, it's important to protect this data from unauthorized access and follow legal guidelines for data privacy.
- Lastly, real-time processing is challenging. While batch processing works in many cases, some applications need real-time insights, such as fraud detection or targeted advertising. This requires faster, more responsive systems that can handle both large volumes and high speed.

4. Storage for Big Data
(Section 4.2 of the paper)

1. What factors should you consider when using distributed storage for Big Data?

- Volume and Scalability
  Distributed storage systems must handle huge volumes of data efficiently. The system should be scalable, allowing seamless expansion by adding more nodes without disrupting existing services.
- Fault Tolerance
  Since distributed systems run on multiple machines, there is a high probability of node failures. The storage architecture should ensure data replication and quick recovery mechanisms so that no data is lost when a node crashes.
- Data Locality and Computation
  Bringing computation to the data rather than transferring massive datasets across the network is a key optimization. Systems like Hadoop's HDFS enable this by ensuring that processing occurs close to where the data is stored, reducing network latency and improving performance.

- Consistency and Availability Trade-offs
  In distributed environments, CAP theorem plays a major role. It's essential to understand and balance between Consistency, Availability, and Partition Tolerance based on the specific use case.
- Storage Formats
  The choice of data format (e.g., columnar, row-based, or object storage) impacts compression efficiency, query speed, and usability. Systems must support flexible formats tailored to different types of workloads, such as structured, semi-structured, or unstructured data.
- Metadata Management
  Efficient metadata handling is essential for tracking data location, structure, and access rights. Centralized or distributed metadata servers should manage this without becoming performance bottlenecks.
- Security and Access Control
  With data being distributed across various nodes and possibly geographic regions, ensuring data privacy, encryption, and access control mechanisms is crucial.
- Cost and Resource Optimization
  Distributed storage must balance performance with cost-effectiveness. This includes optimizing resource usage such as CPU, memory, disk I/O, and network bandwidth, as well as minimizing energy consumption.

(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 File System (GFS) and the MapReduce programming model.**

3. **Hadoop is different from transactional systems in the following ways:**
   - Hadoop is designed for batch processing, not real-time transactions.
   - It handles large-scale unstructured data, unlike transactional systems which typically deal with structured, relational data.
   - Hadoop follows a write-once, read-many model, whereas transactional systems support random reads/writes and updates.
   - Hadoop systems prioritize throughput and scalability over low-latency transactions.
   - It is built to run on commodity hardware with built-in fault tolerance, rather than expensive, high-availability systems.

4. **Two parts of Hadoop are:**
   - HDFS (Hadoop Distributed File System)
   - MapReduce (Programming model for processing)

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

Redundancy is built into Hadoop to ensure fault tolerance and high availability. Since Hadoop runs on clusters of commodity hardware where node failures are common, data is replicated across multiple nodes so that even if one node fails, the system can continue functioning without data loss.

Kushal Choudhary (kxc240000)

(Components of Hadoop)

1. The three pieces of Hadoop project are:

- HDFS (for storage)
- MapReduce (for processing)
- Hadoop Common (shared utilities/libraries)

(Hadoop Distributed File System)

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

Hadoop is designed from the ground up to support scalability. It works by breaking data into blocks and distributing those blocks across many different machines in the cluster. Each machine handles a portion of the storage and processing, which makes it easy to increase capacity just by adding more nodes. The Hadoop Distributed File System (HDFS) manages storage efficiently, while the MapReduce framework ensures that tasks run in parallel across the cluster. This setup allows Hadoop to handle huge volumes of data and grow to hundreds or even thousands of nodes without any major architectural changes.

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

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

Data locality means moving computation to where the data resides, instead of moving data across the network to the computation. This reduces network congestion and improves processing speed, making data processing much faster and efficient.

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

- Breaking files into blocks allows parallel processing across the cluster.
- Redundancy (usually via replication) ensures fault tolerance, if one copy of a block is lost, others are available.
- It allows scalability and reliability even with hardware failures.

5. The default size of a block in HDFS is **128** MB.

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

- Fewer blocks means less metadata overhead for NameNode.
- Better throughput for large files since fewer reads and seeks are required.
- Reduces the burden on NameNode and makes it easier to process large datasets efficiently.

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

NameNode is the master server in HDFS.
It manages:

- The namespace of the file system (like a directory tree).
- Metadata (file names, block locations, permissions).
- It does not store actual data, only metadata about where blocks are located.

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