جامـــــعـة المـــــسـتـقـبـل
AL MUSTAQBAL UNIVERSITY

كلية العلوم

قســـم الانظمة الطبية الذكية

Lecture (1 & 2): Introduction to Big Data and Hadoop

Subject: Big Data Analysis in Healthcare
Level: Fourth
Lecturer: Asst. Lecturer Qusai AL-Durrah
Duration: Two hours

1. Introduction

Healthcare is entering an era defined by an unprecedented surge in digital data. Modern hospitals, research institutes, and public-health agencies generate massive records every second. Examples include:

- **Electronic Health Records (EHRs):** Millions of clinical transactions and patient histories recorded daily.
- **Diagnostic Imaging:** High-resolution MRI, CT, and PET scans that can produce gigabytes of data per patient.
- **Genomics and Proteomics:** Next-Generation Sequencing (NGS) experiments producing terabytes in a single run.
- **Wearable and IoT Devices:** Continuous real-time sensor streams such as heart-rate monitors and glucose trackers.

This explosion of data offers tremendous benefits—enhanced clinical decision-making, personalized medicine, early outbreak detection, and more efficient hospital operations.
However, it also presents formidable challenges. Traditional relational database systems, built for gigabytes of structured data and transactional workloads, cannot scale to the petabytes of mixed-format medical data generated at high velocity.

The discipline of **Big Data Analytics** provides the conceptual and technological foundation to meet these challenges. Within this domain, **Apache Hadoop** stands out as an open-source framework that supports distributed storage and parallel computation. This lecture introduces the essential concepts of Big Data and explores Hadoop as a cornerstone technology for advanced healthcare analytics.

2. Learning Outcomes

By the end of this lecture, students will be able to:

1. **Define Big Data** and explain its five key characteristics: Volume, Velocity, Variety, Veracity, and Value.
2. **Describe** the technological challenges of storing, protecting, and analyzing massive healthcare datasets.
3. **Explain** the architecture and components of the Hadoop framework (HDFS , MapReduce, and YARN).
4. **Compare** Hadoop with relational database management systems (RDBMS), high-performance grid computing, and volunteer computing.
5. **Identify** healthcare applications where Hadoop provides decisive advantages.

3. The Big Data Phenomenon

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size. However, there are certain basic tenets of Big Data that will make it even simpler to answer what is Big Data:

● It refers to a massive amount of data that keeps on growing exponentially with time.

● It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.

● It includes data mining, data storage, data analysis, data sharing, and data visualization.

● The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.

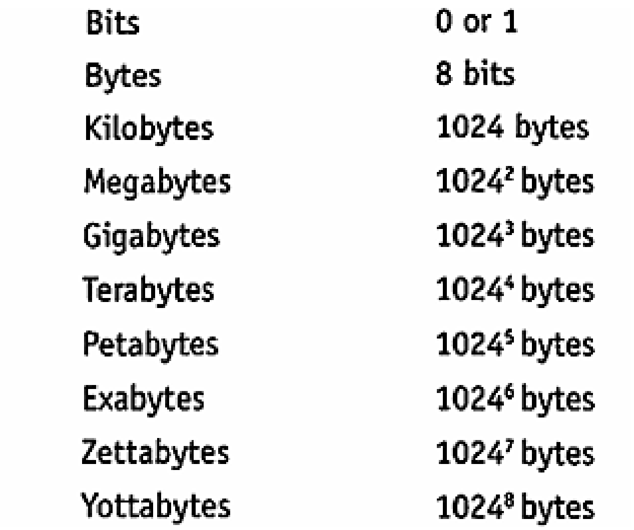
3.1 Global Data Explosion

The world's "digital universe" expanded from **0.18 zettabytes in 2006** to **1.8 zettabytes in 2011**, a tenfold increase in just five years. Healthcare is a major contributor to this growth through imaging archives, continuous patient monitoring, and genomic research.

3.2 Characteristics of Big Data: The Five V's

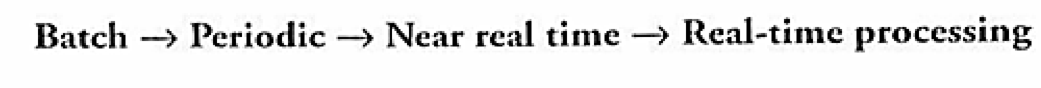
Big Data is typically defined by five dimensions:

1. **Volume**: refers to the size of the data represented in the form of terabytes (TB), petabytes (PB), zettabytes (ZB), etc..



Bits	0 or 1
Bytes	8 bits
Kilobytes	1024 bytes
Megabytes	1024^2 bytes
Gigabytes	1024^3 bytes
Terabytes	1024^4 bytes
Petabytes	1024^5 bytes
Exabytes	1024^6 bytes
Zettabytes	1024^7 bytes
Yottabytes	1024^8 bytes

2. **Velocity**: This alludes to the data generation frequency, the measurement of which may be in milliseconds, seconds, minutes, hours, days, weeks, months, or years. The processing frequency may also vary according to the needs of the user. While certain data may call for real-time processing, others need only be processed when required.



Batch → Periodic → Near real time → Real-time processing

3. **Variety**: Multiple formats: structured EHR tables, semi-structured HL7 messages, and unstructured radiology images or clinical notes.

4. **Veracity**: Data quality and trustworthiness, addressing issues such as noise, missing values, and conflicting records.

5. **Value**: The actionable insights derived to improve patient outcomes and operational efficiency.

3.3 Why Big Data?

The more data we have for analysis, the greater will be the analytical accuracy and also the greater would be the confidence in our decisions based on these analytical findings. This will entail a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and innovating on new products, new services, and optimizing existing services.

More data → More accurate analysis → Greater confidence in decision making → Greater operational efficiencies, cost reduction, time reduction, new product development, and optimized offerings, etc.

3.4 Benefits of big data analytics

There are quite a few advantages to incorporating big data analytics into a business or organization. These include :

Cost reduction: Big data can reduce costs in storing all the business data in one place. Tracking analytics also helps companies find ways to work more efficiently to cut costs wherever possible .

Product development: Developing and marketing new products, services, or brands is much easier when based on data collected from customers' needs and wants. Big data analytics also helps businesses understand product viability and keep up with trends .

Strategic business decisions: The ability to constantly analyze data helps businesses make better and faster decisions, such as cost and supply chain optimization .

Customer experience: Data-driven algorithms help marketing efforts (targeted ads, as an example) and increase customer satisfaction by delivering an enhanced customer experience .

Risk management: Businesses can identify risks by analyzing data patterns and developing solutions for managing those risks .

Entertainment: Providing a personalized recommendation of movies and music according to a customer's individual preferences has been transformative for the entertainment industry (think Spotify and Netflix) .

Education: Big data helps schools and educational technology companies alike develop new curriculums while improving existing plans based on needs and demands .

Health care: Monitoring patients' medical histories helps doctors detect and prevent diseases .

Government: Big data can be used to collect data from CCTV and traffic cameras, satellites, body cameras and sensors, emails, calls, and more, to help manage the public sector .

Marketing: Customer information and preferences can be used to create targeted advertising campaigns with a high return on investment (ROI) .

Banking: Data analytics can help track and monitor illegal money laundering.

4. Technical Challenges of Healthcare Big Data

1. Storage vs. Access-Speed Gap

- Disk capacity grows much faster than read/write speed.

- Example: a 1990 hard drive stored 1.37 GB at 4.4 MB/s (full read ≈ 5 minutes), while a modern 1 TB drive reads at ~100 MB/s (full read > 2.5 hours).

2. Reliability and Fault Tolerance

- Large clusters increase the probability of node failure.
- Replication and automated recovery are essential for clinical data integrity.

3. Data Integration and Heterogeneity

- Combining genomic, sensor, and clinical data demands flexible schema-on-read approaches.

4. Security and Privacy

- Regulations such as HIPAA (or local equivalents) require encryption, strict access controls, and auditing to safeguard patient information.

5. A Brief History of Hadoop

Understanding Hadoop's origins clarifies its design principles and community-driven growth.

5.1 Origins in Nutch

Hadoop began in the early 2000s within the **Apache Nutch** project, an open-source web search engine created by **Doug Cutting** and **Mike Cafarella**. The team needed to scale Nutch to index billions of web pages while maintaining reliability across many inexpensive servers.

5.2 Inspiration from Google Papers

In 2003 and 2004, Google published two influential papers describing the **Google File System (GFS)** and **MapReduce**. These provided a blueprint for large-scale distributed storage and parallel computation. Inspired by these designs, the Nutch developers implemented their own distributed filesystem (NDFS) and a MapReduce-like processing framework.

5.3 Creation of Hadoop

By 2006, these modules had broad utility beyond web search. Cutting and Cafarella separated them into an independent project named **Hadoop**, after a yellow toy elephant belonging to Cutting’s son. The name was chosen for its simplicity and memorability.

5.4 Yahoo! Adoption and Early Milestones

Yahoo! recognized Hadoop’s potential and hired Doug Cutting in 2006, investing significant resources to scale the platform. Milestones included:

- **2006:** Official launch as an Apache subproject.
- **2007–2008:** Yahoo! built large clusters and used Hadoop to generate its production search index.
- **2008:** Yahoo! operated a 10,000-core Hadoop cluster and set a record by sorting one terabyte of data in 209 seconds.

5.5 Mainstream Growth

By 2008, Hadoop had become a top-level Apache project and attracted adopters such as **Facebook**, and **The New York Times**. A notable example was the Times’ use of Amazon EC2 and Hadoop to convert four terabytes of scanned archives into PDFs in less than 24 hours—an achievement impossible without Hadoop’s scalability.

6. Apache Hadoop as a Scalable Solution

6.1 Core Components

- **HDFS (Hadoop Distributed File System):**
    - Splits large files into blocks and replicates them across nodes.

- Provides high throughput and automatic fault tolerance.

- **YARN (Yet Another Resource Negotiator):**
  - Manages cluster resources, schedules jobs, and monitors execution across nodes.
  - Determines how data is processed and allocates resources like RAM for each data block.
- **MapReduce:**
  - Uses *map* and *reduce* functions to process key–value pairs in parallel.
  - Executes computation where the data reside (*data locality*).

6.2 Key Advantages

- **Linear scalability:** Adding nodes nearly proportionally increases capacity.
- **Automatic job scheduling and failure recovery.**
- **Commodity hardware:** Reduces infrastructure cost.

7. Hadoop Ecosystem

The Hadoop ecosystem provides complementary tools critical for healthcare analytics:

- **Hive:** SQL-like querying for large clinical datasets.
- **Pig:** High-level scripting for complex data transformations.
- **HBase:** Column-oriented NoSQL database for real-time patient record access.
- **ZooKeeper:** Cluster coordination and synchronization.
- **Sqoop and Oozie:** Bulk data transfer and workflow orchestration.

These components together create a comprehensive platform for smart medical data science.

8. Comparison with Other Computing Paradigms

Feature	Traditional RDBMS	Hadoop / MapReduce
Typical Scale	Gigabytes	Petabytes
Schema	Fixed	Flexible
Update Pattern	Frequent	Write-once, read-many
Scaling Strategy	Vertical	Horizontal
Best Use Case	Transactions	Large-scale batch analytics

- **Grid Computing:** High CPU power but network bottlenecks for data-intensive tasks.
- **Volunteer Computing:** Useful for CPU-heavy problems but unsuitable for secure, bandwidth-intensive healthcare data.

9. Healthcare Applications of Hadoop

Hadoop enables a wide range of smart-medical innovations:

- **Genomics:** Parallel alignment and variant analysis of massive DNA datasets.
- **Medical Imaging:** Distributed storage and retrieval of MRI/CT scans for AI-driven diagnostics.
- **Real-time Patient Monitoring:** Stream processing to detect anomalies in vital signs.
- **Public Health Surveillance:** Integration of EHR and social-media feeds for epidemic prediction.

10. Conclusion

Big Data in healthcare is defined not merely by size but by speed, diversity, and the imperative to extract clinical value. Hadoop, with its distributed storage (HDFS), parallel processing (MapReduce), and rich ecosystem, provides a robust and cost-effective foundation for advanced analytics, empowering the Smart Medical Systems Department to drive innovation in patient care and biomedical research.

Homework Assignment 1 (Google Classroom):

Answer the following question in one short, well-organized essay:

Question:
Explain how the rise of Big Data created the need for new data-management solutions and how the historical development of Hadoop—from the early Nutch project to Yahoo!'s large-scale adoption—addressed these challenges. In your answer, make sure to:

- Describe the five V's of Big Data and their impact on healthcare data systems.
- Identify at least two major technical challenges of traditional storage or processing methods.
- Summarize the key milestones in Hadoop's early history that made it a turning point in Big Data analytics.

Submission Details:

- **Format:** Word or PDF file
- **Length:** 300 – 400 words

References

- Tom White, *Hadoop: The Definitive Guide*, 3rd Edition, O'Reilly Media, 2012 – Chapter 1 "Meet Hadoop".