Unit # 1
Introduction to Big data

Introduction to Big Data

Data

Data refers to any collection of facts, figures, or statistics that can be used for analysis, decision-making, or other purposes. It can be in various forms such as numbers, text, images, or videos.

Characteristics

Volume: Generally smaller in scale, often manageable by traditional database systems.
Variety: Can include structured data (like databases), semi-structured data (like JSON or XML files), or unstructured data (like text documents, images).
Velocity: The rate at which data is generated and processed can vary but is usually within the capacity of standard systems.
Complexity: Often less complex and can be easily stored, processed, and analyzed using conventional tools.

Big Data

Big Data refers to extremely large and complex datasets that are challenging to process and analyze using traditional data processing tools. It is characterized by the "3 Vs" (Volume, Velocity, and Variety) and often requires advanced tools and technologies to handle.

Introduction to Big Data (contd)

Characteristics

Volume: Massive amounts of data, often in petabytes or exabytes, which exceed the storage and processing capabilities of traditional databases.
Variety: Includes a wide range of data types—structured, semi-structured, and unstructured—such as social media posts, sensor data, transaction logs, and video streams.
Velocity: Data is generated at a high speed, often in real-time or near-real-time, requiring fast processing to extract actionable insights.
Veracity: Refers to the quality and trustworthiness of the data, which can be variable and uncertain in Big Data scenarios.
Value: The potential insights and business value that can be extracted from Big Data through advanced analytics.

Conventional Data refers to general information that can be easily processed and analyzed using conventional tools.
Big Data refers to much larger, more complex datasets that require specialized tools and techniques to manage, process and extract value.

Introduction to Big Data (contd)

Twitter users generate about 500 million tweets per day i.e 347,220 tweets per minute , 5,787 tweets per second.

VOLUME

Average : 300 bytes (Text & Metadata) / 150 gigabytes (GB) of raw data per day / 54 terabytes (TB) of raw tweet data per year.

VARIETY

Beside Text , Multi media data , Meta data , System logs Structured (user profiles, tweets) , Semi-structured (JSON) & Unstructured data (images, videos)

VELOCITY

Twitter processes data in real-time to update feeds, monitor trending topics, detect spam, and more. Handle spikes during major events .

VERACITY

Data includes noise, spam, bots to increase trending tag, misinformation, and incomplete or ambiguous content.

Introduction to Big Data (contd)

Netflix users stream over 140 million hours of content per day i.e 5,833,333 hours of content per hour , 97,222 hours of content per min.

IRCTC can handle over 1.3 million bookings in a single day. i.e 54,167 bookings per hour , 902 bookings per minute per min.

Consider National level bank like SBI. It has ~ 420 million customers. SBI handles ~ 10 million transactions in a day. This boils down to Approximately 1,000,000 transactions per hour. Approximately 16,667 transactions per minute.

For a moderately small bank or organizations , there can be ~ 250 transactions per hour. Approximately 4 transactions per minute.

For instance, an average smartphone user might consume anywhere from 2 to 10 GB per month.

Total Data = 100,000,000 users × 5 GB (Average data) = 500,000,000 GB

Total Data = 500,000 TB (or 500 PB)

Total data per min = 11,570 gigabytes

How Big Data Works ?

Big Data operates through a combination of **Technologies**, **Processes** , **Methodologies**

1. Data Generation and Collection

- Source: Big Data is generated from multiple sources, including social media, sensors, websites, transactions, mobile devices, and more. This data can be structured (databases), semi-structured (XML, JSON), or unstructured (text, images, videos).
- Volume: The data generated is massive in scale, often measured in terabytes, petabytes, or even exabytes.

2. **Data Ingestion**

- Batch Processing: Data is collected over a period of time and then processed in large batches. Tools like Hadoop's HDFS (Hadoop Distributed File System) are often used for this purpose.
- Real-Time Processing: Data is processed as it arrives, allowing for real-time analytics. Technologies like Apache Kafka, Apache Storm, and Apache Flink are commonly used to handle streaming data.

3. Data Storage

- Distributed Storage: Big Data is typically stored in distributed file systems or cloud storage, which allows data to be spread across multiple servers or locations. This ensures scalability and fault tolerance.

How Big Data Works ? (contd)

- Data Lakes: A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. Technologies like Amazon S3, Azure Data Lake, and Google Cloud Storage are often used.

4. Data Processing

- Apache Hadoop: An open-source framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
- Map Reduce: A programming model used for processing large data sets with a distributed algorithm. It divides the task into smaller sub-tasks (Map) and then combines the results (Reduce).

5. Data Visualization

- Reports: Automatically generated reports that summarize the insights from Big Data analysis, making it easier for decision-makers to understand the findings.
- Real-Time Monitoring: Visualization tools that provide real-time insights, enabling businesses to react immediately to changes or trends.

6. Data Analysis

- Machine Learning: Applying algorithms that can learn from data to make predictions or decisions without being explicitly programmed.

How Big Data Works ? (contd)

- Text Analytics: Processing unstructured text data to extract meaningful information, often used in sentiment analysis, natural language processing (NLP), etc.
- Data Mining: Extracting patterns and knowledge from large datasets using statistical and computational techniques.

7. Decision Making

- Actionable Insights: The final goal of Big Data is to provide actionable insights that can drive business decisions. This could range from optimizing operations, personalizing customer experiences, predicting market trends, or improving products and services.
- Automation: In some cases, the insights derived from Big Data can lead to automated decisions, where systems make decisions or take actions without human intervention, such as automated trading systems or personalized marketing campaigns.

8. Feedback

- Continuous Improvement: The process is iterative, with feedback from decisions and actions being fed back into the system to refine algorithms, improve models, and adapt to new data.

Unstructured Data in Big Data

Unstructured data refers to information that doesn't fit neatly into traditional row-column databases like relational databases which is highly organized and easily searchable.

Unstructured data includes a variety of formats, such as text, images, videos, social media posts, emails, logs, and sensor data, making it difficult to analyze with traditional tools.

Unstructured data lacks a pre-defined data model or schema making it difficult to analyze with traditional tools. However , it often contains rich information that can be mined for insights.

Real-World Examples

1. Social Media Posts:

Social media platforms like Twitter, Facebook, and Instagram generate vast amounts of unstructured data in the form of text posts, comments, images, and videos. The text, images, and videos in posts are unstructured, while user ID and post date are structured.

*Thank You … . This **Site** Is Extremely Helpful For All Of Us. Especially In The **Stressful** Exam Days, It Helped A Lot And Everything Is So Organized. Good Job Team.*

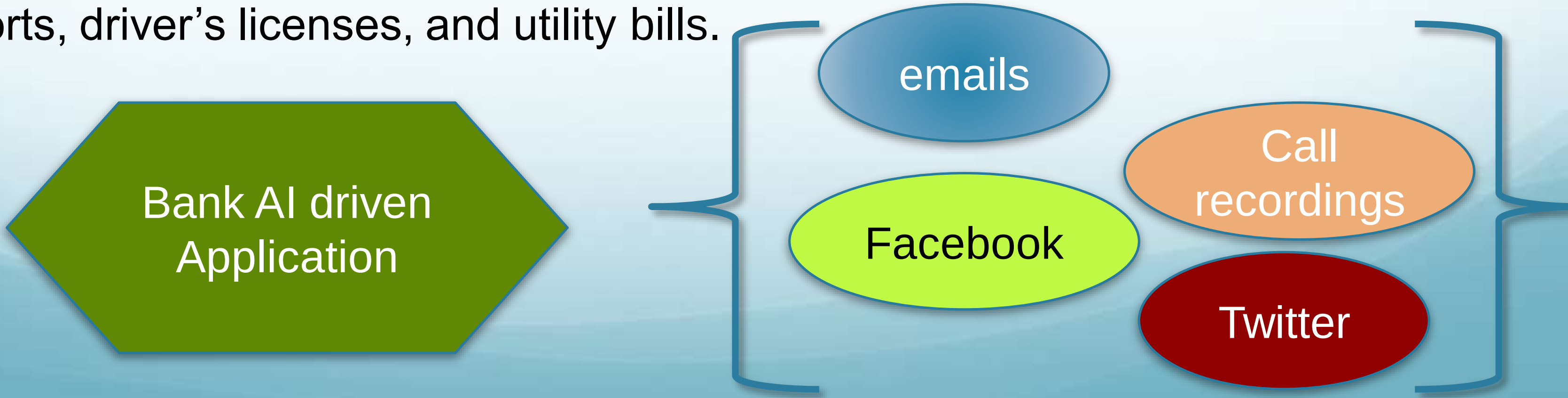
talk about igdtu resource community , about exam condition , Exam Guidance, Overall sentiment …

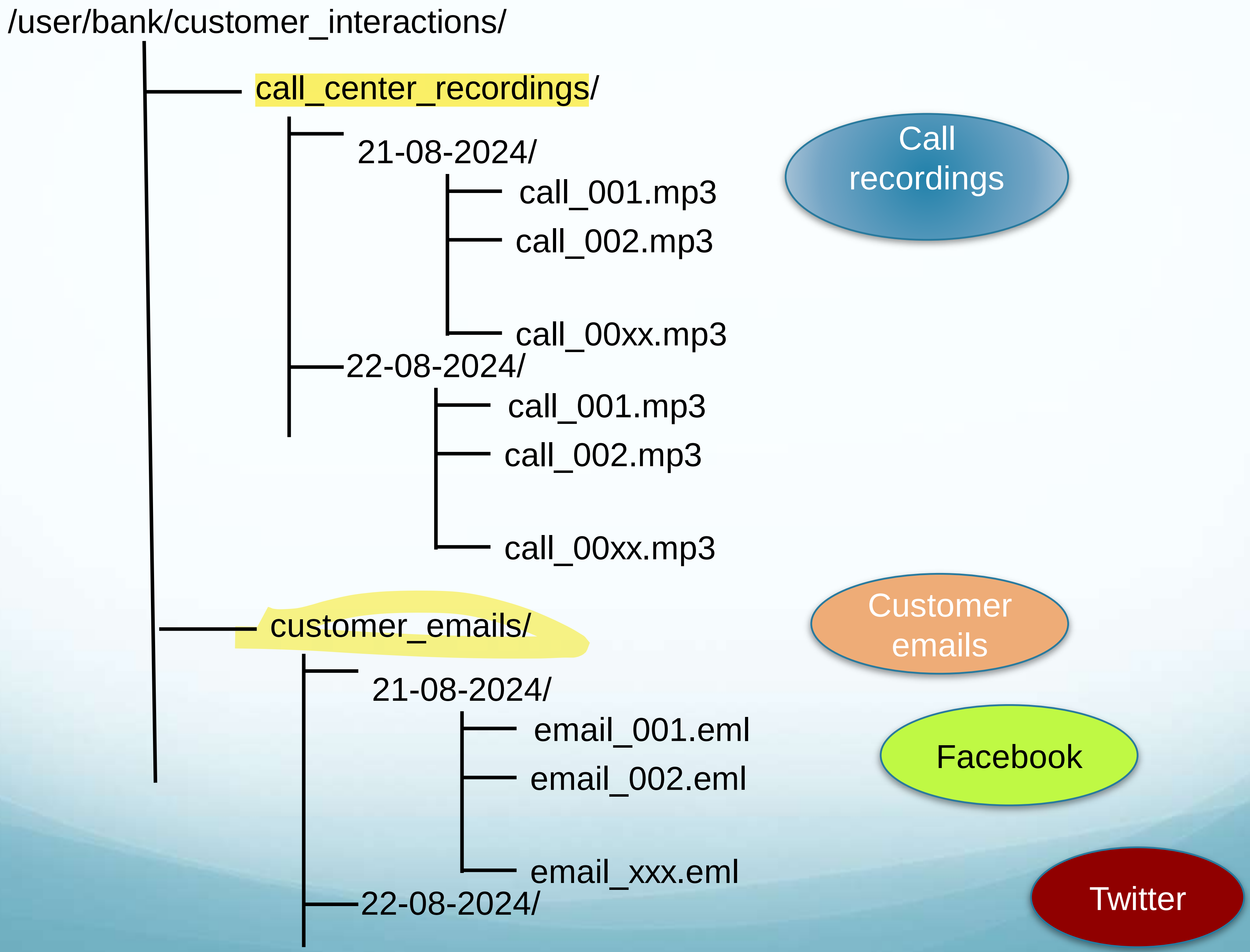
Classify as a positive or Negative Review – Sentiment Analyser

Unstructured Data in Big Data (contd)

2. Financial Services :

- Financial institutions receive millions of customer emails that may include inquiries, complaints, transaction disputes, or service requests. The body of these emails contains unstructured data, while metadata like timestamps and sender/receiver information are structured.
- Customer service interactions over the phone are recorded and transcribed into text. These transcripts are unstructured data that can be analyzed for customer sentiment, frequently asked questions, and service quality.
- The data in Annual , Legal , Contractual & Financial reports. Companies produce detailed financial reports that include narrative descriptions, management discussions, and footnotes.
- Financial institutions track news from various sources to stay informed about market conditions, economic indicators, and company-specific events.
- Images and scanned documents for Eg. When checques are deposited, financial institutions often scan them into the system. Personal verification docs like passports, driver's licenses, and utility bills.





/user/bank/customer_interactions/

call_center_recordings/

21-08-2024/

call_001.mp3

call_002.mp3

call_00xx.mp3

22-08-2024/

call_001.mp3

call_002.mp3

call_00xx.mp3

customer_emails/

21-08-2024/

email_001.eml

email_002.eml

email_xxx.eml

22-08-2024/

Call recordings

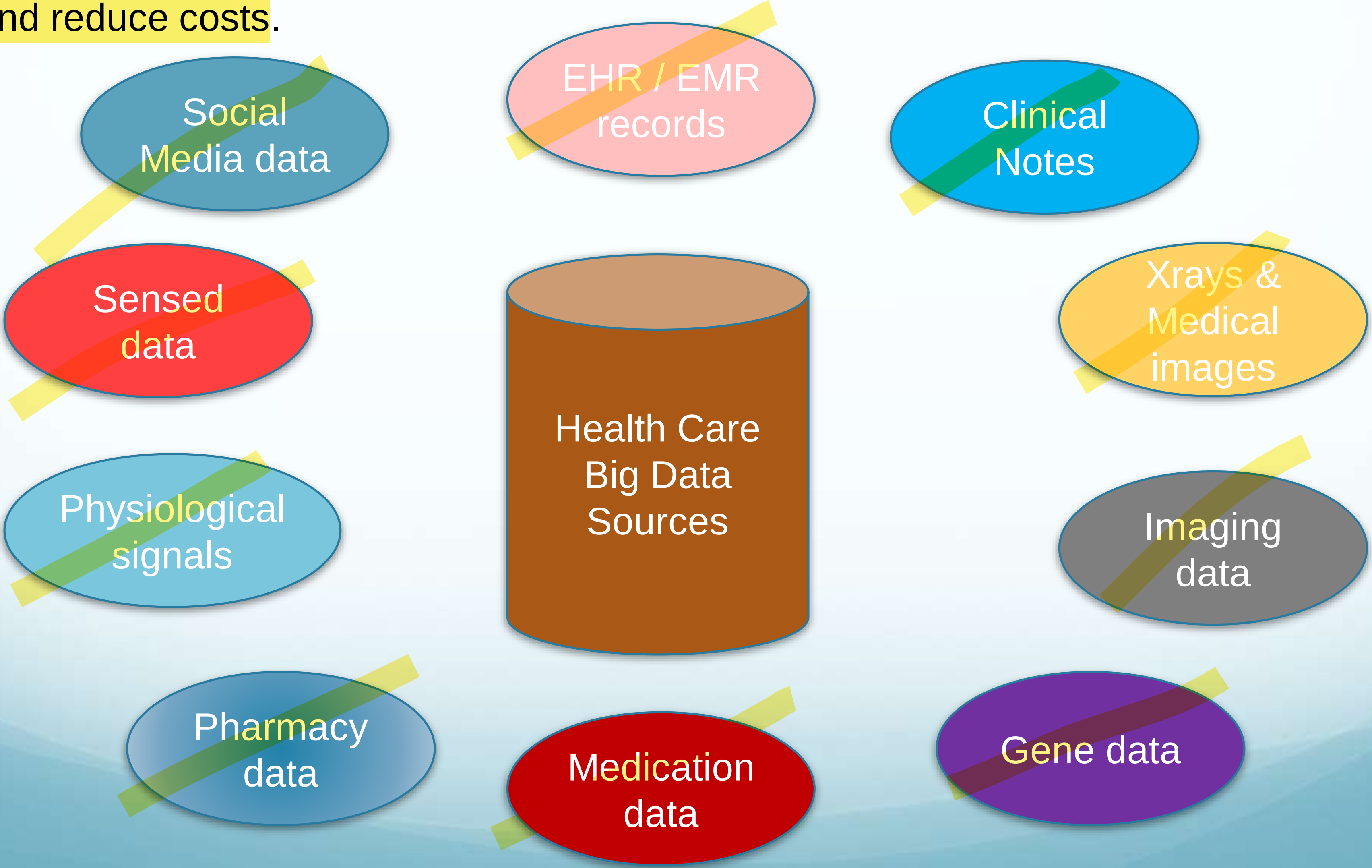
Customer emails

Facebook

Twitter

Big data in healthcare

The healthcare industry generates vast amounts of data, much of which is considered "Big Data" due to its volume, variety, velocity, and veracity. This data originates from various sources and can be used to improve patient outcomes, streamline operations, and reduce costs.



Social Media data

EHR / EMR records

Clinical Notes

Sensed data

Health Care Big Data Sources

Xrays & Medical images

Physiological signals

Imaging data

Pharmacy data

Medication data

Gene data

Big data in healthcare (contd)

1. Electronic Health Records (EHRs)

Electronic Health Records (EHRs) are digital versions of a patient's paper chart, and they play a critical role in modern healthcare systems. In a Big Data context, EHRs generate vast amounts of data, which include structured data (like patient demographics) and unstructured data (like doctor's notes).

Structured Data:

Patient Demographics:
Name, age,
gender,
address,
contact information.

Clinical Data:
Diagnosis codes (ICD-10),
treatment codes (CPT),
vital signs (blood pressure, heart rate),
and medication prescriptions.

Lab Results:
Blood tests,
cholesterol levels,
glucose levels, etc.

Billing Information:
Insurance details,
billing codes,
payment history

Big data in healthcare (contd)

Un Structured Data:

Doctor's Notes:
Free-text descriptions of visits
symptoms,
recommendations.

Imaging Data:
X-rays
MRI , PET SCAN
CT scans stored in digital files

Transcriptions:
Audio recordings of Doctor-Patient
Transcribed in to text.

Patient Reports:
Patient survey data colelcted ,
Habits , pain levels etc

Semi Structured Data:

Sensor data :
Data extracted from wearable devices , Padded Glucometers , ECG recordings in JSON or XML formats, Output of Medical instruments like EEG machines etc

Big data in healthcare (contd)

1. Patient Information Table

PatientID	FirstName	DoB	Gender	Address	Phone				
P001	Ram	-	-	-	-				
P002	Sham	-	-	-	-				

2. Clinical Data Table

PatientID	VisitId	VisitDate	DiagnosticCode	DiagDescription	Treatement Code
P001	V1001	-	E11.9	Diabetese	-
P002	V1002	-	I10	Dislocation	-

3. Doctor Note Table

PatientID	VisitId	VisitDate	Doctor Comments
P001	V1001	-	Patient presents with elevated blood glucose levels
P002	V1002	-	Patient's blood pressure is stable.Continue medication.

Big data in healthcare (contd)

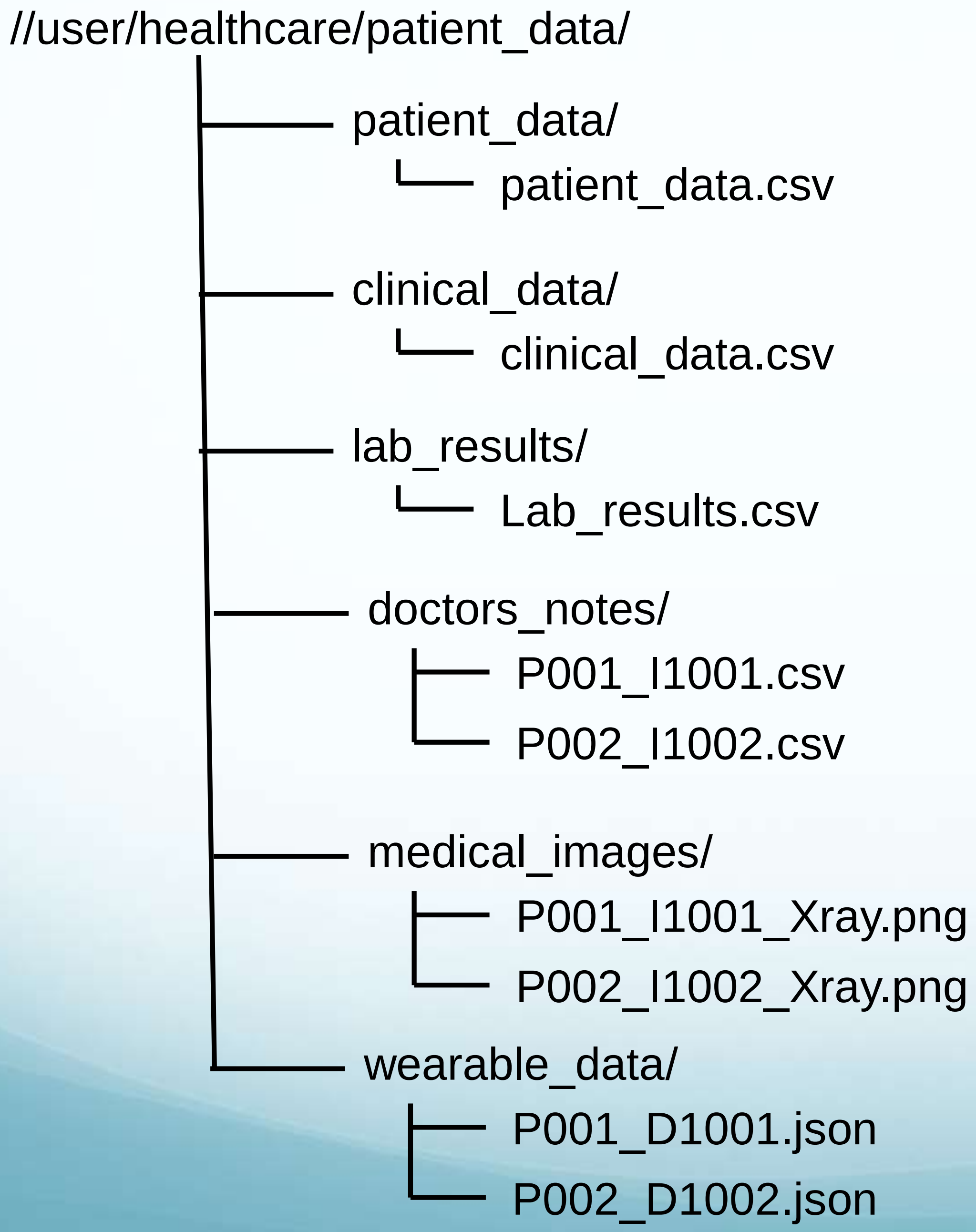
4. Medical Imaging Table

PatientID	VisitId	ImagingDate	ImagingType	ImageFilePath
P001	V1001	-	xray	/hdfs/path/to/images/P001_I1001_Xray.png
P002	V1002	-	mri	/hdfs/path/to/images/P002_I1002_MRI.dcm
P001	V1001	-	ultras	/hdfs/path/to/images/P001_I1001_ULT.dcm

5. Wearable Device Table

PatientID	VisitId	ReadingDate	ReadData
P001	V1001	-	/{"heart_rate": 75, "steps": 1200, "calories": 80}
P002	V1002	-	{"heart_rate": 82, "steps": 1450, "calories": 95}

Big data in healthcare (contd)



```
//user/healthcare/patient_data/
    ├── patient_data/
    │     └── patient_data.csv
    ├── clinical_data/
    │     └── clinical_data.csv
    ├── lab_results/
    │     └── Lab_results.csv
    ├── doctors_notes/
    │     ├── P001_I1001.csv
    │     └── P002_I1002.csv
    ├── medical_images/
    │     ├── P001_I1001_Xray.png
    │     └── P002_I1002_Xray.png
    └── wearable_data/
          ├── P001_D1001.json
          └── P002_D1002.json
```

Big data in healthcare (contd)

2. Imaging in Health Care

Imaging centers in Health care institutions provides ability to store, process, and analyze large volumes of X-rays, MRIs, CT scans, and ultrasounds imaging data . These images, are stored in formats like DICOM or can be even proprietary formats in some cases. The images contains rich information but are also large and complex, requiring advanced techniques and methods for efficient management and analysis.

Structured Data:

Patient Demographics:
Name, age,
gender,
address,
contact information.
Imaging Data:
X-rays , Ultrasound
MRI , CT scans
Study type , Technician Notes

Billing Information:
Insurance details,
billing codes,
payment history

Un Structured Data:

Metadata:
Study Information
Technician Notes
File formats , Image resolution
Storage path

Annotations :
Radiologist Notes
Segmentation notes

Big data in healthcare (contd)

Semi Structured Data:

Diagnostic : Probability scores , Recommendations , AI Model used

1. Patient Information Table

PatientID	FirstName	DoB	Gender	Address	Phone	Bill_code	Insurance
P001	Ram	-	-	-	-		Star
P002	Sham	-	-	-	-		ICICI

2. Imaging Table

PatientID	ImageId	Model	Probability score	ImageType	FilePath
P001	V1001	CNN	0.97	DICOM	/hdfs/path/to/images/P001_S001_Brain.dcm
P002	V1002	RNN	0.86	DICOM	/hdfs/path/to/images/P002_S002_Chest.dcm

3. Radiology / Diagnostic Table

PatientID	ImageId	Condition	RadiologistComments
P001	V1001	Tumor	Lesion detected in the frontal lobe.
P002	V1002	Pneumonia	Suspected pneumonia; further tests needed.

Big data in healthcare (contd)

3. Genomics Big data

Genetics involves managing, analyzing, and interpreting large-scale genomic data to understand genetic variations, diagnose genetic disorders, and develop personalized medicine. It involves sequencing the genomes of thousands of patients and analyzing this data in combination with their clinical records to identify genetic markers associated with these diseases.

Structured Data:

Patient Demographics:
Name, age, gender,
Ethnicity,
Family history.

Genome Study Data:
DNA Sequence from NGS
RNA sequence , Identified mutation
Phenotypic data

Billing Information:
Insurance details,
billing codes,
payment history

Un Structured Data:

Metadata:
Information about genes
Chromosome details
Pathways associated with Gene
Mapping quality

Annotations :
NGS Notes
Prediction models
Risk scores

Big data in healthcare (contd)

Semi Structured Data:

Diagnostic : Risk scores , Recommendations , AI Model used

1. Patient Information Table

PatientID	FirstName	DoB	Gender	Ethnicity	Family history
P001	Ram	-	-	South	Diabetese
P002	Sham	-	-	Central	Cardio in Father lineage

2. Genome Sequence Table

PatientID	SeqId	Gene	Chromosome	ModelUsed	FilePath
P001	S001	BRCA1	17	LipidRiskModelV2	/hdfs/path/to/images/P001_S001_Brain.dcm
P002	S002	LDLR	19	CardiogenomicsModelV3	/hdfs/path/to/images/P002_S002_Chest.dcm

3. Diagnostic Table

PatientID	ImageId	Condition	RadiologistComments
P001	V1001	Heart Disease	DNA repair and tumor suppression.
P002	V1002	LDL disorder	Cholesterol metabolism mild disturbances

Thanks