FDA: Unit 2 – Dealing with Different Types of Data

Data: Raw facts and figures without any context or meaning on their own.
Example: Sales numbers.

Metadata: Data that provides information about other data.
Example: For a photo – resolution, file size, data taken.

ETL (Extract, Transform, Load): ETL is a process used in data integration for analytics and reporting. It involves extracting data from various sources, transforming it into a usable format, and loading it into a target system.

Steps:

1. Extract: Pull data from various sources such as databases, files, and APIs.
2. Transform: Clean, filter, and format the data to meet analysis or business requirements.
3. Load: Store the transformed data into a target system such as a data warehouse or database.

Types of Analytics:

1. Text Analytics
2. Web Analytics
3. Marketing Analytics

Text Analytics: Text analytics is the process of analysing and understanding written or spoken language using computer algorithms. It extracts valuable information, patterns, and insights from unstructured text data like emails, customer feedback, social media posts, and online reviews.

Techniques Used:

- Natural Language Processing (NLP): Understands text meaning.
- Sentiment Analysis: Detects the emotional tone (positive, negative, neutral).
- Keyword Extraction: Finds the most frequent or relevant terms.
- Topic Modeling: Identifies major themes or discussion topics.

7 Steps of Text Analysis:

1. Language Identification: Detect the language of the text.
2. Tokenization: Break the text into words or small parts.
3. Sentence Breaking: Split the text into individual sentences.
4. Part-of-Speech Tagging: Identify each word’s role (noun, verb, adjective, etc.).
5. Chunking: Group words into meaningful phrases (e.g., "big data analytics").
6. Syntax Parsing: Understand sentence grammar and relationships. Example*:* "The cat sat on the mat" → Subject: "cat", Verb: "sat", Object: "mat".
7. Sentence Chaining: Connect multiple sentences to understand overall context and meaning.

Common Use Cases / Techniques:

1. Sentiment Analysis: Understand customer opinions to improve service.
2. Customer Feedback Analysis: Extract insights from reviews or surveys.
3. Social Media Monitoring: Track public sentiment and trends in real time.

Web Analytics: Web analytics is the process of collecting, measuring, analyzing, and reporting website data. It helps understand user behavior, improve user experience, and achieve business goals.

Key Questions Answered:

- How many people visit the website?
- Where do visitors come from?
- Which pages are most/least visited?
- Why do users leave without completing an action?

Components:

1. Data Collection: Tools collect user interactions from the website.
2. Metrics:
   - Page Views: Number of times a page is loaded.
   - Sessions: A user's entire visit to the site.
   - Bounce Rate: % of visitors who leave after viewing one page.
   - Conversion Rate: % of users who complete a goal (purchase, signup, etc.).
3. Dimensions (to segment data):
   - Source: (Google, Direct, Facebook, etc.)
   - Location: (Country, City)
   - Device: (Mobile, Desktop, Tablet)
4. Analysis & Insights:
   - Identify high-performing pages
   - Detect drop-off points in user flow
   - Analyze top-performing traffic sources
5. Reporting & Visualization: Dashboards and charts to present data clearly for decision-makers.

Benefits:

- Improve site performance (e.g., reduce load times).
- Enhance user experience (e.g., better layout/navigation).
- Increase conversions (e.g., reduce cart abandonment).
- Optimize marketing efforts (e.g., identify best channels).

Example:

An e-commerce site finds:

- 70% of users are on mobile
- Mobile bounce rate is 50% higher than desktop
  Insight: The mobile site needs optimization for better sales.

Marketing Analytics: Marketing analytics is the use of data to measure, manage, and improve marketing performance. It helps businesses understand which marketing efforts are working and where to improve.

Key Questions:

- Which marketing channels drive the best results?
- What's the ROI of each campaign?
- How can we reach the right audience more effectively?

Components:

1. Data Collection: From websites, campaigns, ads, social media, CRM, etc.
2. Metrics:
   - Cost per Click (CPC)
   - Return on Ad Spend (ROAS)
   - Click-through Rate (CTR)
   - Customer Lifetime Value (CLTV)
3. Types of Analysis (DDPP):
   - Descriptive – What happened?
   - Diagnostic – Why did it happen?
   - Predictive – What will happen?
   - Prescriptive – What should be done?
4. Tools Used:
   - Google Analytics
   - Adobe Analytics
   - HubSpot, Salesforce Marketing Cloud
   - Google Ads, Facebook Ads Manager

    - Tableau, Power BI (for data visualization)
5. Analysis & Insights:
    - Identify high-performing campaigns
    - Reduce spend on low-performing channels
    - Segment audiences for better targeting
6. Reporting & Decision-Making:
    - Use dashboards to track KPIs
    - Adjust campaigns in real-time for better results

Benefits:

- Better Decision-Making – Data-driven strategies
- Optimized Budgets – Focus on what works
- Higher ROI – Improve outcomes through continuous improvement
- Improved Targeting – Reach the right customers with the right message

Example:

A retail brand compares ads:

- Instagram ads = 3x higher conversions than Google Ads
  Action: Increase Instagram ad spend, reduce Google Ads budget.

Types of Data:

Introduction: Data is the foundation of analytics. Different types of data require different techniques for analysis.

There are different types of data:

1. Numerical data
2. Categorical data
3. Structured data
4. Unstructured data

Quantitative Data or Numerical Data

Data that represents numbers and can be measured or counted.

Types:

1. Discrete:
    - Countable values.
    - Discrete data only involves integers.
    - Eg: The number of participants in an event; The number of students in a school.
2. Continuous:
    - Has floating point values generally within the range.
    - Eg: Temperature.

Advantages:

1. Easy to analyse with statistics.
2. Helps to summarize large datasets quickly.
3. Allows mathematical operations.
   Eg: Calculate Total Sales = Price * Quantity.
4. Used for predictive modelling.

Disadvantages:

1. Misses quantitative speech.
   Eg: "Customer happiness" is hard to quantify.
2. Can be misleading without context.
   Eg: Average salary of 50,000.

Qualitative Data or Categorical Data

This type of data is descriptive and cannot be measured in numbers, it is often divided into two categories.

Types:

1. Nominal: Data used for labelling or categorizing without any order or ranking.
   Eg: Gender, Color
2. Ordinal data: Data that involves order or ranking.
   Eg: Ratings (Good, Better, Best), Educational level.

Advantages:

1. Simple to collect & classify.
2. Useful for grouping & segmentation.
3. Easy to visualize with the help of charts.

Disadvantages:

1. Limited mathematical analysis.
2. Categories can be subjective.
3. Groups may hide the imp details.

Structured Data

Data that is organized in a predefined format, usually in rows and columns (like in databases or spreadsheets). It is easy to search, filter, and analyse using tools like SQL or Excel.

Advantages:

1. Easy to store, search and analyze.
2. Compatible with many tools (Eg: Power BI).
3. High accuracy and reliability.

Disadvantages:

1. Limited flexibility.
2. Requires predefined schema.
3. Not for messy data like text messages or chats or voice notes.

Unstructured data

Data that does not follow a specific format or structure. It is not organized in rows or columns.

Advantages:

1. Handles complex and diverse data types (Eg: Text, Images, Videos, etc.).
2. Covers majority of real-world data (~80%).
3. Helps in advanced analytics.

Disadvantages:

1. Hard to process and analyze.
2. Requires expensive storage and tools.
3. Time consuming to extract meaning.

Questions

1. Define data
2. List the different categories of data.
3. Explain numerical data.
4. Explain categorical data.
5. Explain structured data.
6. Explain unstructured data.
7. Difference between structured & unstructured data (5-7 points).
8. Difference between numerical data & categorical data (5-7 points).

Data Types & Formats

- Data in analytics comes in different forms and must be stored in specific formats for processing.
- Data types define the kind of values stored in a data set (Eg: Numbers, Text, Dates).
- Data formats define the structure and representation of data files for storage, sharing, and analysis (Eg: CSV, JSON, XML).

Data types:

1. Integer:

- Number with decimal points.
- Eg: 10, -25.

2. Float:
   - Decimal values.
   - Eg: 3.14, 9.99
3. String:
   - Sequence of characters.
   - Eg: “Hello”, “Mumbai”.
4. Boolean:
   - Binary logic.
   - Eg: True/False
5. Date or Time:
   - Dates

Data Formats:

1. CSV (Comma Separated Values): Data stored in rows and columns, separated by commas.
   Advantages:
   1. Simple and lightweight format.
   2. Supported by almost all software.
   3. Easy to edit with text editor.
   4. Small file size.

   Disadvantages:
   1. Cannot handle complex hierarchical data.
   2. No formulas.
   3. Errors may occur if data contains commas, line breaks, or special characters.
   4. Not suitable for large data set.

2. Excel: Spreadsheet supporting rows and columns, formulas, charts, and formatting.
   Advantages:
   1. User friendly interface.
   2. Supports charts, formulas, etc.
   3. Easy for non-technical users.

   Disadvantages:
   1. Limited capacity for large data etc.
   2. Not suitable for automation at scale
   3. Proprietary format
   4. Difficult to track changes in collaborative environment

3. JSON (JavaScript Object Notation): Lightweight text-based format using key values pairs.
   Advantages:
   1. Human readable and easy to parse.
   2. Supports nested structure.
   3. Language independent and widely supported.
   4. Compact and efficient for data transmission.

   Disadvantages:
   1. Larger in size compared to Binary formats.

4. XML (Extensive Markup Language):
   Advantages:
   1. Highly structured and self-descriptive.
   2. Supports complex and hierarchical data.
   3. Good for data validation with schemas.

   Disadvantages:
   1. Verbose → large file size.

2. Slower to parse compared to JSON.
3. More complex to read or write for humans.

5. Multimedia format:

Questions:

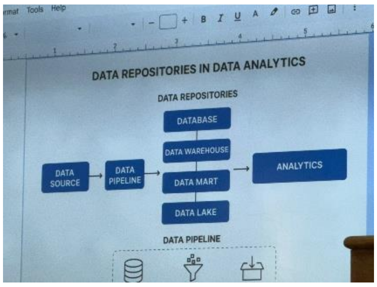
1. Define datatype.
2. Explain following types of datatypes.
3. What is the need of datatypes and data formats.
4. Explain the following types of data formats.
5. Differentiate between different data formats.

What is a Data Repository?

A data repository is a central place where data is collected, stored, and managed. It helps organizations keep data safe, organized, and ready for analysis or any future use.

Types of Data Repositories

1. Database
2. Data Warehouse
3. Data Mart
4. Data Lake
5. Data Pipeline



1. **Database**

A database is an organized collection of structured information (usually in tables), managed using software called a Database Management System (DBMS).

Key Characteristics:

1. Follows ACID properties (ensures data is reliable and consistent)
2. Good for storing structured data (like tables)
3. Supports CRUD operations (Create, Read, Update, Delete)
4. Great for handling everyday transactions

Types:

1. Relational Database (RDBMS): Uses tables (SQL, MySQL, PostgreSQL)
2. NoSQL Database: Stores data in formats like documents, key-values, graphs (MongoDB, Cassandra)

Advantages:

1. Fast queries on structured data
2. Ensures data integrity and security
3. Uses standard query languages (e.g., SQL)

Disadvantages:

1. Not ideal for unstructured data (like photos or videos)
2. Scaling up to handle huge data can be complex or expensive

Examples & Real-life Use Cases:

1. Banking software for accounts & transactions
2. Student database for college enrollments
3. Social media user profiles

Extra Insight:
Modern databases can also handle some types of semi-structured data (like JSON fields), making them more flexible than before.

2. **Data Warehouse**
A data warehouse is a large, central storehouse for integrating and storing historical, structured data from many sources. It's mainly for analysis and reports, not day-to-day operations.
Characteristics:

1. Optimized for analytical queries (OLAP)
2. Uses ETL (Extract, Transform, Load) to clean and move data into the warehouse

Advantages:

1. Combines data from different departments, creating a big-picture view
2. Enables complex, historical trend analysis

Disadvantages:

1. Expensive to build and maintain
2. Poor for unstructured or real-time data

Examples & Use Cases:

1. Healthcare data analysis (patient history, treatments)
2. Company-wide sales and marketing analysis

Extra Insight:
Modern warehouses (cloud-based, like Snowflake, BigQuery) can scale up or down easily and handle both structured and semi-structured data.

3. **Data Mart**
A data mart is a smaller section of a warehouse, focused on just one department (like sales, finance, or HR).
Characteristics:

1. Department/domain-specific
2. Quick access, as its smaller in size
3. Built "top-down" from a warehouse or "bottom-up" from department data

Advantages:

1. Fast, relevant data for specific teams
2. More affordable than full warehouses

Disadvantages:

1. Can lead to data silos (different teams have their own separate data)
2. Doesn’t cover the whole organization

Examples:

1. Marketing data mart for campaigns
2. Finance data mart for tracking expenses

Extra Insight:
Many companies start with a data mart and eventually build up to a full data warehouse.

4. **Data Lake**
A data lake stores massive amounts of raw data in its original format (structured, semi-structured, or unstructured). You organize it (“add a schema”) only when you use it.
Characteristics:

1. Schema on read: Structure is applied only when data is read, not stored.
2. Stores all data:
3. Scalability: designed to handle petabytes of data.
4. Cost-Effective: Uses inexpensive cloud or distributed storage.
5. Supports all advanced analytics: Enables AI, ML, and Predictive modeling.

Advantages:

1. Flexible for many types of data, including big data and machine learning
2. Scalable and cheap compared to warehouses
3. Can store massive data at low cost

Disadvantages:

1. Can become a “data swamp” (disorder, hard to find value) if not managed properly
2. Raw data needs more processing to be useful

Examples:

1. IoT sensor data collection
2. Storing massive logs for cybersecurity or website analytics

Extra Insight:

More companies now combine lakes and warehouses using a “lakehouse” approach—getting the flexibility of both systems.

Component:

1. Raw data zone:
   - Stores unprocessed data from multiple sources.
   - Example: log files, sensor data, raw transactions.
2. Cleansed /processed zone
   - Data is cleaned, filtered or enriched for analytics.
3. Curated zone
   - Ready – to use datasets, optimized for reporting and BI.
4. Sandbox Zone
   - Experimental area for data scientist to test ML models.

Applications:

1. IOT data storage:
   - Sensor and device-generated
2. Machine learning
3. Fraud detection
4. Healthcare analysis

5. **Data Pipelines (Bonus)**

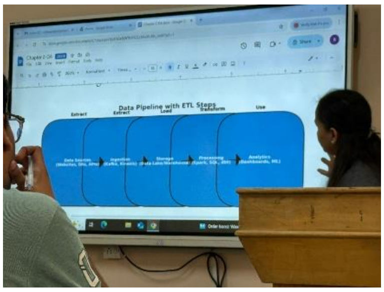
A data pipeline is not a repository, but the process that moves data between systems (e.g., from a database into a warehouse).

- Cleans, transforms, and loads data automatically
- Supports real-time analytics

Data pipelines Are important because data comes from many sources (app, sensor, social media) and must be collected in 1 place for better decision making

Practical Tips:

1. Choose a database for daily transactions and fast lookups
2. Use a data warehouse for long-term trends and big-picture analysis
3. Data marts are best for department-specific needs
4. Data lakes are ideal for big data, machine learning, and storing raw or varied information


Data Pipeline with ETL Steps
Extract
Extract
Load
Transform
Use
Data Sources (Websites, DBs, APIs)
Ingestion (Kafka, Kinesis)
Storage (Data Lake/Warehouse)
Processing (Spark, SQL, dbt)
Analytics (Dashboards, ML)

Key components of data pipeline:

1. Data Sources:
   - Where raw data originates.
   - Eg: Databases, APIs, IOT devices, Social media, Spreadsheets, Log files.

2. Ingestion Layer:

- Collects and brings data into pipelines.

3. Storage Layer:
   - A central repository where raw data is stored.
   - Options: Data warehouse, Data lakes.

4. Processing and Transformation Layer:
   - Cleans, formats, and transforms raw data into an analyzable form.
   - Tasks: Removing duplicates, Handling missing values, Aggregations, Normalization.
5. Analytics Layer:
   - Processed data is made available for analysis, reporting or machine learning.
   - Tools: Power BI, Tableau, Looker.

6. Orchestration and Workflow Management:
   - Ensures data flows smoothly through stages.

7. Monitoring and Login:
   - Tracks pipeline performance, errors and data quality.
   - Tools: Prometheus, Grafana, ELK Stack

Use cases:

1. Business Intelligence
2. Machine Learning
3. Fraud Detection
4. IOT Analytics

Advantages:

1. Automation of data flow
2. Scalability
3. Consistency and Reliability
4. Data Integration
5. Improves productivity
6. Reusability

Disadvantages:

1. High initial setup cost and complexity
2. Maintenance overhead
3. Data quality issues
4. Scalability challenges
5. Security and compliance risks

Q. Write the comparison between all 5 repositories (Homework) (Atleast 6 points) (Jay Maharashtra).