full one-shot videos on :**JK Coding Pathshala YouTube channel**

JK Coding Pathshala

https://youtube.com/@jayeshkande9215?feature=shared

Unit IV	BIG DATA ANALYTICS	(06 Hrs)
Big Data Analytics- Architecture and Life Cycle, Types of analysis, Analytical approaches, Data Analytics with Mathematical manipulations, Data Ingestion from different sources (CSV, JSON, html, Excel, mongoDB, mysql, sqlite), Data cleaning, Handling missing values, data imputation, Data transformation, Data Standardization, handling categorical data with 2 and more categories, statistical and graphical analysis methods, Hive Data Analytics.		

Q3) a) What is data wrangling? Why do you need it? Explain data wrangling methods? **[9]**

b) What is categorical variable? Why do you need categorical variable encoding? With an example, explain one-hot encoding. **[9]**

OR

Q4) a) Draw and explain Architecture of HIVE. **[8]**

b) How missing values and categorical variables are preprocesses before building model? Explain with suitable example. **[4]**

c) Explain z-score normalization. For the following dataset carry out z-score normalization (standardization), X = 23, 29, 52, 31, 45, 19, 18, 27. **[6]**

P.T.O.

Q3) a) Explain different types of Big Data Analysis techniques. **[8]**

b) i) Explain Different Data Transformation techniques. **[3]**

ii) What is dataset? Explain with python syntax of 2 different types of dataset used in Big data. **[6]**

OR

Q4) a) i) Explain Mean, Mode and variance and standard deviation with suitable example. **[8]**

ii) Explain Data Standardization. **[3]**

b) Draw and explain Architecture of HIVE. **[6]**

P.T.O.

Q3) a) Compare HBASE and HIVE with suitable parameters. **[8]**

b) How missing values are filled in Pandas Data Frame with zeros? Assume suitable data. **[3]**

c) Explain Min-max scaling. For the following dataset carry out min-max Scaling, X = 24,28,53,30,40,18,15,21. **[6]**

OR

Q4) a) What is categorical variable? Why do you need categorical variable encoding? With an example, explain one-hot encoding? **[8]**

b) What is data wrangling? Why do you need it? Explain data wrangling methods? **[9]**

P.T.O.

Q3) a) Explain Mean, Mode and variance and standard deviation with suitable example. **[9]**

b) Draw and explain Architecture of HIVE **[8]**

OR

Q4) a) Explain Min-max scaling. For the following dataset carry out min-max Scaling, X=24, 28, 53, 30, 40, 18, 15, 21 **[9]**

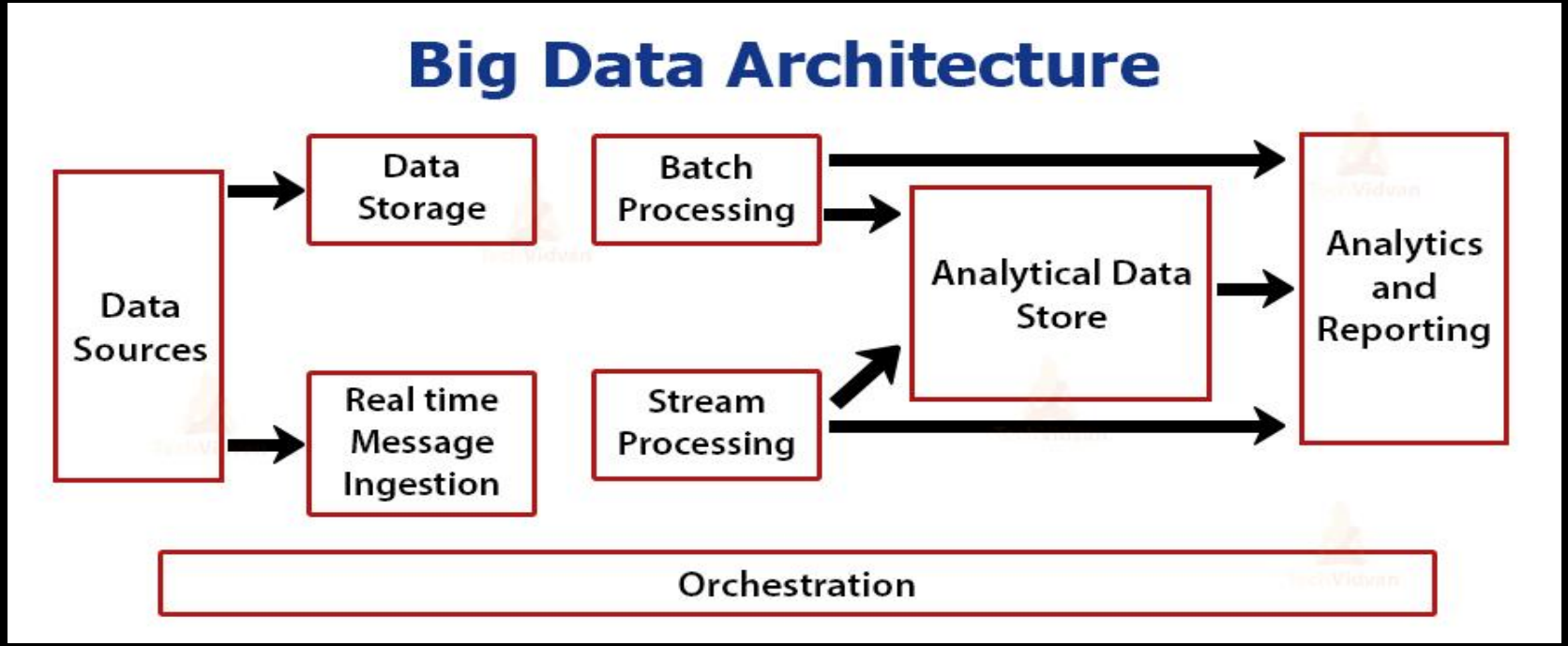
b) What is data Wrangling? Why do you need it? explain data Wrangling methods? **[8]**

Q3)	a)	Explain different steps in Data Analytics Project Life cycle	**[7]**
	b)	Draw and explain Architecture of HIVE.	**[7]**
	c)	Explain different data transformation techniques.	**[3]**

OR

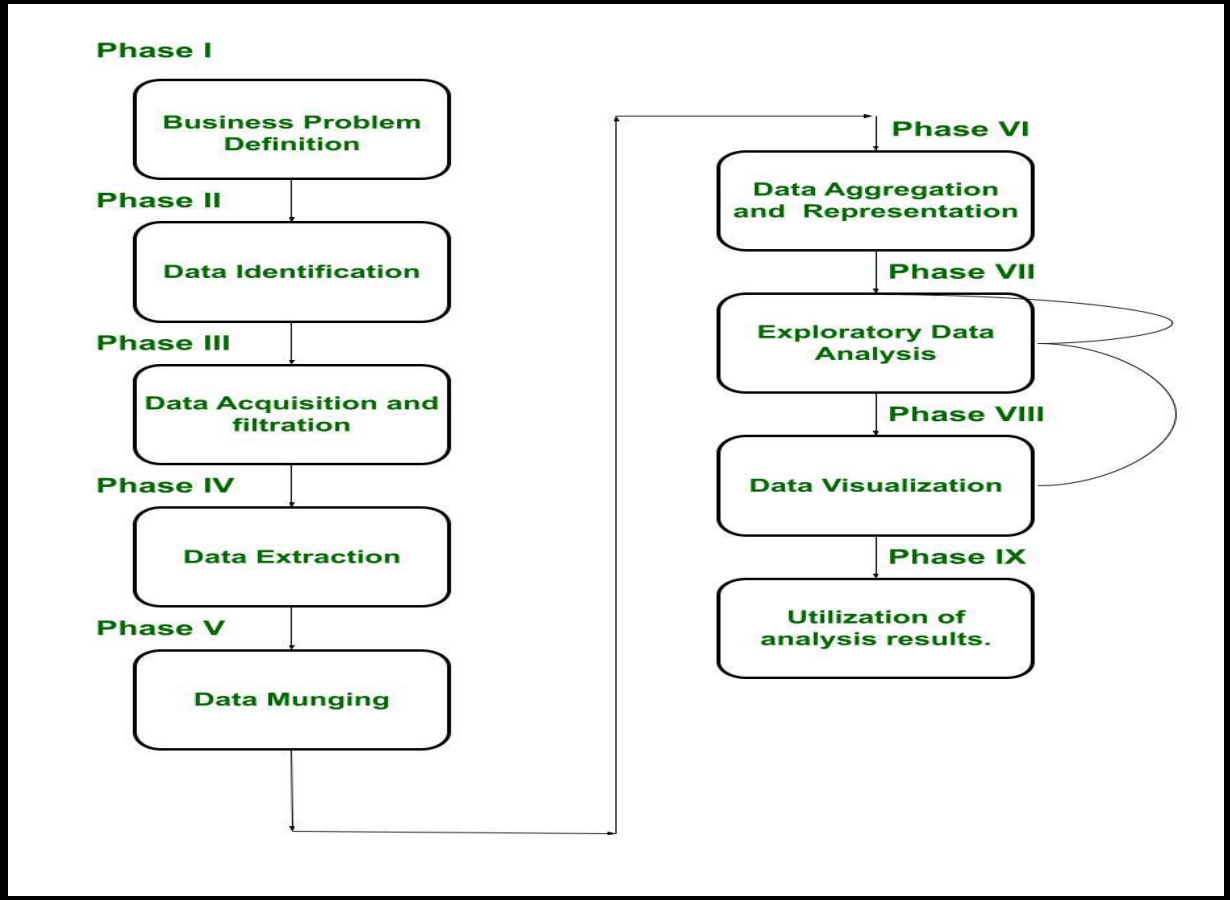
Q4)	a)	Explain different kinds of Big Data Analysis.	**[7]**
	b)	How data can be ingested in python. Write syntax in python for the same.	**[7]**
	c)	Explain role of visualization in big data analytics.	**[3]**

Big Data Analytics Architecture



Component	Explanation (Hinglish)	Example / Tools
Data Sources	Jahan se data aata hai — multiple sources ho sakte hain like databases, sensors, logs, social media, etc.	Databases, IoT devices, Social media feeds
Data Storage	Raw data ko store karne ke liye jagah — large scale storage systems.	HDFS, Amazon S3, Google Cloud Storage
Batch Processing	Data ko batches mein process karna — jab data accumulated ho jaye tab process karte hain.	Hadoop MapReduce, Apache Spark
Real-time Message Ingestion	Real-time data ko continuously collect karna from streaming sources.	Apache Kafka, Flume
Stream Processing	Real-time data ko turant process karna, continuously analyze karna.	Apache Flink, Apache Storm, Spark Streaming
Analytical Data Store	Processed data ko ek jagah store karna jahan se analytics ke liye access ho sake.	Data warehouses like Redshift, BigQuery
Analytics and Reporting	Data analysis karna aur business ko reports, dashboards provide karna.	Tableau, Power BI, Looker
Orchestration	Sare processes ko manage karna, automate karna, taaki smooth workflow chal sake.	Apache Airflow, Oozie

Big Data Analytics Life Cycle



The Big Data Analytics Life cycle is divided into nine phases, named as :

1.Business Case/Problem Definition

2.Data Identification

3.Data Acquisition and filtration

4.Data Extraction

5.Data Munging(Validation and Cleaning)

6.Data Aggregation & Representation(Storage)

7.Exploratory Data Analysis

8.Data Visualization(Preparation for Modeling and Assessment)

9.Utilization of analysis results.

Phase No.	Phase Name	Hinglish Explanation	Example / Tools
I	Business Problem Definition	Sabse pehle business problem ko clearly define karna — **kya solve karna hai?**	"Customer churn kam kaise karein?"
II	Data Identification	Decide karna ki **kaunsa data chahiye** problem solve karne ke liye.	Customer demographics, transaction history
III	Data Acquisition and Filtration	Data collect (acquire) karna aur **unwanted/noisy data ko filter** karna.	APIs, databases, web scraping
IV	Data Extraction	Data ko extract karke ek centralized jagah le jaana — jaise data warehouse ya data lake.	Data warehouse, data lake
V	Data Munging (Cleaning & Preparation)	Raw data ko clean aur prepare karna taaki analysis ke layak bane.	Missing values fill karna, formats standardize
VI	Data Aggregation and Representation	Similar data points ko group karke summarize karna — jaise **average spending, total sales**, etc.	Summarization techniques
VII	Exploratory Data Analysis (EDA)	Data explore karna taaki patterns, trends, aur outliers milein.	Graphs, statistics
VIII	Data Visualization	Insights ko **charts, graphs, dashboards** mein dikhana.	Tableau, Power BI
IX	Utilization of Analysis Results	Insights ko real business decisions mein use karna.	"Offer discounts to high-risk churn customers"

📊 Types of Big Data Analysis

Types of Big Data Analysis

Descriptive Analysis

Kya hua?

Yeh analysis batata hai ki past me kya hua tha tha

Example: “Last week kitne logon ne website visit kiya?

Diagnostic Analysis

Kyu hua?

Yeh samjhata hai ki koi cheez kyu hui.

Example: “Sales suddenly down kyu hua?”

Predictive Analysis

Aage kya ho sakta hai?

Yeh future ka prediction karta hai based on past data

Example: “Agle mahine ka

Prescriptive Analysis

Ab kya karna chahiye?

Yeh best action suggest karta hai for better results

Example: “Sales badhane ke

Type of Analysis	Basic Question	Explanation (Hinglish)	Example
Descriptive Analysis	Kya hua?	Yeh batata hai ki past mein kya activity ya event hua.	“Last week kitne logon ne website visit kiya?”
Diagnostic Analysis	Kyu hua?	Yeh analysis samjhata hai ki koi particular event ya change kyu hua.	“Sales suddenly down kyu hua?”
Predictive Analysis	Aage kya ho sakta hai?	Yeh batata hai ki future mein kya hone ke chances hain, past data ke base par.	“Agle mahine ka demand kya hoga?”
Prescriptive Analysis	Ab kya karna chahiye?	Yeh suggest karta hai best actions jo outcome improve karne mein help karein.	“Sales badhane ke liye konsi strategy follow karein?”

Analytical Approaches

Approach	Description (Hinglish)	Example
Quantitative Analysis	Numbers, statistics aur mathematical models ka use.	Sales data analyze karna using average, trend, regression.
Qualitative Analysis	Non-numerical data jaise ki feedback, reviews ya interviews ka analysis.	Customer feedback se product ki quality samajhna.
Statistical Analysis	Data ke patterns aur relationships ko discover karne ke liye statistical tools ka use.	Correlation between age & buying behavior.
Diagnostic Analysis	Identify karta hai issues ya problems ke root cause.	Kyun users ne product cancel kiya uska analysis.
Predictive Analysis	Future ke outcomes predict karta hai using machine learning ya past trends.	Customer churn prediction.
Prescriptive Analysis	Decision-making mein help karta hai by suggesting next best actions.	Inventory reorder karna based on demand forecast.
Exploratory Data Analysis (EDA)	Data ke andar patterns, outliers ya interesting points explore karta hai.	Pehli baar data visualize karke trends find karna.
Text Analysis (NLP)	Textual data jaise emails, reviews, tweets ka analysis karta hai.	Sentiment analysis of product reviews.
Diagnostic Analytics	Drill-down karta hai to understand “kyu” kuch hua.	Traffic drop ka root cause find karna.
Real-Time Analytics	Live ya instantly aane wale data ka analysis.	Fraud detection during online transactions.

Data Analytics with Mathematical manipulations

Mathematical Manipulations ka Matlab?
Data ko samajhne ke liye kuch mathematical operations karte hain, jaise:

- **Summation (jodna)**
- **Average (ausat nikalna)**
- **Variance aur Standard Deviation (data ka spread samajhna)**
- **Normalization (data ko scale karna)**
- **Correlation (do variables ke beech relation dekhna)**

Kuch Important Mathematical Manipulations in Data Analytics:

1. Mean (Average)

Definition:

Mean ya Average data ka ek aisa central value hota hai, jisme saare data points ko jodkar unki total sankhya se divide kiya jata hai.

Formula:

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Explanation:

- x_i = har ek data point
- n = total data points ki sankhya
- $\sum$ matlab sum karna (sab points ko jodna)

2. Variance (Vichalan)

Definition:

Variance data ke spread ko measure karta hai. Matlab data kitna average se door hai, iska pata chalta hai.

Formula:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Explanation:

- x_i = har ek data point
- μ = mean (average)
- $(x_i - \mu)^2$ = har data point aur mean ke beech ka fark ka square
- Sum kar ke total data points se divide karte hain

3. Standard Deviation (Pramanik Vichalan)

Definition:

Standard Deviation variance ka square root hota hai, jo data ke spread ko original units mein batata hai.

Formula:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

Explanation:

Data points ka average spread measure karta hai. Jitna zyada SD, utna zyada data scatter.

4. Normalization (Sadharanikaran)

Definition:

Normalization data ko ek fixed range (jaise 0 se 1 ke beech) mein laane ka process hai, taaki data easily compare kiya ja sake.

Formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Explanation:

- x = original data value
- $\min(x)$ = data ka minimum value
- $\max(x)$ = data ka maximum value
- x' = normalized value (0 se 1 ke beech)

5. Correlation Coefficient (Sambandh Gunank)

Definition:

Correlation do variables ke beech ke linear relationship ko measure karta hai, jo -1 se +1 ke beech hota hai.

Formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Explanation:

- x_i, y_i = variables ke data points
- $\bar{x}, \bar{y}$ = un variables ke means
- $r = +1$ matlab perfect positive correlation
- $r = -1$ matlab perfect negative correlation
- $r = 0$ matlab no linear relation

Question:

Aapke paas do variables ke data points hain:

- $X = [2, 4, 6, 8, 10]$
- $Y = [5, 9, 12, 15, 20]$

Tasks:

1. X ka **Mean** nikaalo
2. X ka **Variance** aur **Standard Deviation** calculate karo
3. X ke data ko **Normalization** karo (0 se 1 ke range mein)
4. X aur Y ke beech **Correlation coefficient** nikaalo

Data:

$X = [2, 4, 6, 8, 10]$
$Y = [5, 9, 12, 15, 20]$

---

1. Mean of X

Formula:

$$\text{Mean} = \frac{\sum x_i}{n}$$

Calculate sum of X:

$2 + 4 + 6 + 8 + 10 = 30$

Number of points, $n = 5$

Mean:

$$\frac{30}{5} = 6$$

Answer: Mean of X = 6

2. Variance and Standard Deviation of X

Formula for Variance:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$$

Calculate each term $(x_i - \mu)^2$:

x_i	$x_i - \mu$	$(x_i - \mu)^2$
2	2 - 6 = -4	(-4)^2 = 16
4	4 - 6 = -2	(-2)^2 = 4
6	6 - 6 = 0	0
8	8 - 6 = 2	4
10	10 - 6 = 4	16

Sum of squares:

$$16 + 4 + 0 + 4 + 16 = 40$$

Variance:

$$\sigma^2 = \frac{40}{5} = 8$$

Standard Deviation:

$$\sigma = \sqrt{8} \approx 2.83$$

Answer:

Variance = 8

Standard Deviation ≈ 2.83

3. Normalization of X

Formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Find $\min(x) = 2, \max(x) = 10$

Calculate normalized values:

x	$x' = \frac{x-2}{10-2} = \frac{x-2}{8}$
2	$\frac{2-2}{8} = 0$
4	$\frac{4-2}{8} = \frac{2}{8} = 0.25$
6	$\frac{6-2}{8} = \frac{4}{8} = 0.5$
8	$\frac{8-2}{8} = \frac{6}{8} = 0.75$
10	$\frac{10-2}{8} = \frac{8}{8} = 1$

Answer: Normalized X = [0, 0.25, 0.5, 0.75, 1]

4. Correlation Coefficient between X and Y

Formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- Mean of X, $\bar{x} = 6$ (already calculated)
- Mean of Y, $\bar{y} = \frac{5+9+12+15+20}{5} = \frac{61}{5} = 12.2$

Calculate $(x_i - \bar{x})$, $(y_i - \bar{y})$, their products, and squares:

x_i	$x_i - 6$	y_i	$y_i - 12.2$	$(x_i - 6)(y_i - 12.2)$	$(x_i - 6)^2$	$(y_i - 12.2)^2$
2	-4	5	-7.2	(-4) * (-7.2) = 28.8	16	51.84
4	-2	9	-3.2	(-2) * (-3.2) = 6.4	4	10.24
6	0	12	-0.2	0 * (-0.2) = 0	0	0.04
8	2	15	2.8	2 * 2.8 = 5.6	4	7.84
10	4	20	7.8	4 * 7.8 = 31.2	16	60.84

Now sum these up:

- $\sum(x_i - \bar{x})(y_i - \bar{y}) = 28.8 + 6.4 + 0 + 5.6 + 31.2 = 72$
- $\sum(x_i - \bar{x})^2 = 16 + 4 + 0 + 4 + 16 = 40$
- $\sum(y_i - \bar{y})^2 = 51.84 + 10.24 + 0.04 + 7.84 + 60.84 = 130.8$

Finally,

$$r = \frac{72}{\sqrt{40 \times 130.8}} = \frac{72}{\sqrt{5232}} = \frac{72}{72.34} \approx 0.995$$

Answer: Correlation coefficient $r \approx 0.995$ (Strong positive correlation)

Calculation	Result
Mean of X	6
Variance of X	8
Standard Deviation of X	2.83
Normalized X	[0, 0.25, 0.5, 0.75, 1]
Correlation between X and Y	0.995 (Strong +ve)

Data Ingestion from different sources (CSV, JSON, html, Excel, mongoDB, mysql, sqlite)

PYQ (Previous Year Question) Example:
Q. How can data be ingested in Python? Explain with syntax for CSV, JSON, HTML, Excel, MongoDB, MySQL, and SQLite data sources.

📥 DATA INGESTION IN PYTHON

"Data ingestion" ka matlab hota hai **data ko different sources se Python mein load karna** — analysis, visualization ya ML ke liye.

1. **CSV File Se Data Ingestion**

```
import pandas as pd
data = pd.read_csv("filename.csv")
```

➡️ read_csv() function se CSV file ka data DataFrame mein load hota hai.

2. **JSON File Se Data Ingestion**

```
import pandas as pd
data = pd.read_json("filename.json")
```

➡️ JSON (JavaScript Object Notation) file se structured data ko pandas read karta hai.

🔹 3. HTML Table Se Data Ingestion

```
import pandas as pd
data = pd.read_html("https://example.com")[0]
```

➡️ Web page par HTML tables ko read_html se list of DataFrames mein convert karte hai.

🔹 4. Excel File Se Data Ingestion

```
import pandas as pd
data = pd.read_excel("filename.xlsx")
```

➡️ Excel ke .xlsx format ko pandas read_excel function se read karta hai.

🔹 5. **MongoDB Se Data Ingestion**

```
import pymongo
client = pymongo.MongoClient("mongodb://localhost:27017/")
db = client["mydatabase"]
collection = db["mycollection"]

data = list(collection.find())
```

➡️ MongoDB se data fetch karne ke liye pymongo package ka use hota hai.

🔹 6. MySQL Se Data Ingestion

```
import mysql.connector
import pandas as pd

conn = mysql.connector.connect(
    host="localhost",
    user="root",
    password="password",
    database="testdb"
)

query = "SELECT * FROM tablename"
data = pd.read_sql(query, conn)
```

➡️ MySQL se connect karne ke baad pandas read_sql() function se data fetch karta hai.

🔹 7. **SQLite Se Data Ingestion**

```
import sqlite3
import pandas as pd

conn = sqlite3.connect("mydatabase.db")
query = "SELECT * FROM tablename"
data = pd.read_sql(query, conn)
```

➡️ SQLite lightweight DB hota hai, Python ke sqlite3 module se data fetch kiya jata hai.

Sr	Source	Python Library	Syntax (Example)	HiEnglish Note
1	CSV File	pandas	pd.read_csv("file.csv")	CSV file read karne ke liye sabse common function
2	JSON File	pandas	pd.read_json("file.json")	JSON file ko DataFrame mein convert karta hai
3	Excel File	pandas	pd.read_excel("file.xlsx")	Excel (.xls/.xlsx) file ko load karta hai
4	HTML Table	pandas	pd.read_html("https://example.com")[0]	Website se table read karta hai
5	MongoDB	pymongo + pandas	pd.DataFrame(list(collection.find()))	MongoDB ka data list mein laake DataFrame banate hai
6	MySQL	sqlalchemy + pymysql	pd.read_sql("SELECT * FROM table", conn)	SQL query se data read karta hai
7	SQLite	sqlite3 + pandas	pd.read_sql("SELECT * FROM table", conn)	Local DB (SQLite) se data fetch karta hai

Extra Note:
•Har source ke liye database connection banana zaroori hota hai (specially MySQL, SQLite, MongoDB).
•pandas is the core library for structured data ingestion.

Data Cleaning

PYQ: "What is Data Cleaning? Explain common techniques with Python code."

How missing values are filled in Pandas Data Frame with zeros? Assume suitable data. [3]

🧹 DATA CLEANING IN PYTHON

Data Cleaning ka matlab hota hai — **galat, missing, duplicate ya irrelevant data ko theek karna ya hataana** taaki analysis sahi ho sake.

🔑 Common Data Cleaning Steps

🔹 1. Missing Values ko Handle Karna

```
import pandas as pd
df = pd.read_csv("data.csv")

# Missing values check
print(df.isnull().sum())

# Option 1: Fill with mean
df["column_name"].fillna(df["column_name"].mean(), inplace=True)

# Option 2: Drop rows with missing values
df.dropna(inplace=True)
```

Agar kisi column mein missing value ho, toh usse **mean, median, ya mode** se fill kar sakte hai.

- 2. **Duplicate Values ko Remove Karna**

```
# Check duplicates
print(df.duplicated().sum())

# Remove duplicates
df.drop_duplicates(inplace=True)
```

Duplicates data ko analysis se pehle remove karna zaroori hota hai.

- 3. **Wrong Data Types Ko Fix Karna**

```
# Convert string to int
df["age"] = df["age"].astype(int)
```

Example: Age column string mein ho toh usse int mein convert karna chahiye.

- **4. Inconsistent Text ko Clean Karna**

```
# Convert to lowercase
df["name"] = df["name"].str.lower()
```

Example: "Jayesh" aur "jayesh" ko same treat karne ke liye **lowercase** ya **strip** use karte hai.

- 5. **Outliers Ko Detect/Handle Karna**

```
# Using IQR method
Q1 = df["column"].quantile(0.25)
Q3 = df["column"].quantile(0.75)
IQR = Q3 - Q1

# Filter out outliers
df = df[(df["column"] >= Q1 - 1.5*IQR) & (df["column"] <= Q3 + 1.5*IQR)]
```

Outliers ko hataane se model better train hota hai.

Step	Python Function	Hindi Tip
Missing Values	fillna(), dropna()	Mean/median se bharo ya hatao
Duplicates	drop_duplicates()	Ek hi row baar-baar ho toh hatao
Wrong Data Types	astype()	String ko int/float mein convert karo
Inconsistent Format	str.lower(), strip()	Text ko uniform banao
Outliers	IQR Method	Bahut jyada choti/badi values hatao

How missing values are filled in Pandas Data Frame with zeros? Assume suitable data. [3]

In a DataFrame, missing values are typically represented as NaN (Not a Number).
The method fillna(0) is used to replace all missing values with a specified value (in this case, 0).

```
import pandas as pd
import numpy as np

# Creating a DataFrame with missing values
data = {'Name': ['Alice', 'Bob', 'Charlie', np.nan],
        'Age': [25, np.nan, 30, 22],
        'City': ['New York', 'Los Angeles', np.nan, 'Chicago']}

df = pd.DataFrame(data)

# Filling missing values with 0
df_filled = df.fillna(0)

# Display the result
print(df_filled)
```

	Name	Age	City
0	Alice	25.0	New York
1	Bob	0.0	Los Angeles
2	Charlie	30.0	0.0
3	0.0	22.0	Chicago

In this example:

•**Missing values (NaN)** are replaced by zeros in all columns (Name, Age, City).

🔍 What is Data Imputation?

Data Imputation ka matlab hota hai:
"Missing ya NULL values ko kisi suitable value se replace karna."
Yeh ek **data cleaning technique** hai jo machine learning models ke liye data ko complete banata hai.

💡 Why Impute Data?

- Algorithms NaN values ko samajh nahi paate
- Incomplete data → Wrong results
- Isiliye, imputation se hum **data ko usable banate hain**

Method	Description (HiEnglish)	Syntax (Pandas)
1. Mean Imputation	Missing value ko column ke average se replace karna.	df['Age'] = df['Age'].fillna(df['Age'].mean())
2. Median Imputation	Missing value ko column ke median se fill karna.	df['Age'] = df['Age'].fillna(df['Age'].median())
3. Mode Imputation	Categorical columns ke liye, sabse common value se fill karna.	df['Gender'] = df['Gender'].fillna(df['Gender'].mode()[0])
4. Constant Imputation	Missing values ko ek fixed value (e.g. 0, 'Unknown') se bharna.	df.fillna(0) ya df.fillna('Unknown')
5. Forward Fill (ffill)	Upar wali value se missing value fill karna.	df.fillna(method='ffill')
6. Backward Fill (bfill)	Neeche wali value se missing ko fill karna.	df.fillna(method='bfill')

```
import pandas as pd
import numpy as np

data = {
    'Name': ['Amit', 'Sita', 'Raj', 'Priya'],
    'Age': [25, np.nan, 30, np.nan],
    'City': ['Mumbai', np.nan, 'Delhi', 'Pune']
}

df = pd.DataFrame(data)

# Mean Imputation for Age
df['Age'] = df['Age'].fillna(df['Age'].mean())

# Mode Imputation for City
df['City'] = df['City'].fillna(df['City'].mode()[0])
```

Tip:

- Numerical values → Mean / Median
- Categorical values → Mode
- Zyada missing ho → Row ya Column ko drop kar do (df.dropna())

What is Data Transformation?
Data Transformation ka matlab hota hai **raw data ko clean, structured aur model-friendly format me convert karna**. Isse machine learning models ko train karna easy ho jaata hai.

S.No.	Technique	Explanation (HiEnglish)
1.	**Normalization** (Min-Max Scaling)	Data ko **0 to 1 range** me convert karta hai. 📌 Formula: (X - min) / (max - min) 🧠 Use when features have **different ranges**.
2.	**Standardization** (Z-score Scaling)	Data ko aise transform karta hai ki uska **mean = 0 aur standard deviation = 1** ho jaye. 📌 Formula: (X - mean) / std 🧠 Use when data has **outliers** or follows normal distribution.

PYQS

Z-Score Normalization (Standardization)

📘 What is Z-Score Normalization?

Z-score normalization is a method of standardizing data by transforming the values so that they have:

- **Mean (μ) = 0**
- **Standard deviation (σ) = 1**

It is useful when features have different units or scales.

📌 Formula:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- Z is the z-score,
- X is the data point,
- μ is the mean of the dataset,
- σ is the standard deviation of the dataset.

Explain z-score normalization. For the following dataset carry out z-score normalization (standardization), X = 23, 29, 52, 31, 45, 19, 18, 27. **[6]**

Given Dataset:

$$X = \{23, 29, 52, 31, 45, 19, 18, 27\}$$

Step 1: Calculate the Mean (μ)

$$\mu = \frac{23 + 29 + 52 + 31 + 45 + 19 + 18 + 27}{8}$$

$$\mu = \frac{244}{8} = 30.5$$

Step 2: Calculate the Standard Deviation (σ)

$$\sigma = \sqrt{\frac{1}{n}\sum(X_i - \mu)^2}$$

X	X - μ	$(X - \mu)^2$
23	-7.5	56.25
29	-1.5	2.25
52	21.5	462.25
31	0.5	0.25
45	14.5	210.25
19	-11.5	132.25
18	-12.5	156.25
27	-3.5	12.25

$$\sigma = \sqrt{\frac{56.25 + 2.25 + 462.25 + 0.25 + 210.25 + 132.25 + 156.25 + 12.25}{8}}$$

$$\sigma = \sqrt{\frac{1032.0}{8}} = \sqrt{129} \approx 11.36$$

Step 3: Apply Z-Score Formula

$$Z = \frac{X - 30.5}{11.36}$$

X	Z = (X - 30.5)/11.36	Z (approx)
23	(23 - 30.5)/11.36	-0.66
29	(29 - 30.5)/11.36	-0.13
52	(52 - 30.5)/11.36	1.89
31	(31 - 30.5)/11.36	0.04
45	(45 - 30.5)/11.36	1.28
19	(19 - 30.5)/11.36	-1.02
18	(18 - 30.5)/11.36	-1.10
27	(27 - 30.5)/11.36	-0.31

Final Z-Score Normalized Values:

$$\{-0.66, -0.13, 1.89, 0.04, 1.28, -1.02, -1.10, -0.31\}$$

Min-Max Scaling

📘 What is Min-Max Scaling?

Min-Max Scaling (also called **Normalization**) is a technique used to scale data between a **fixed range**, usually **[0, 1]**.

📌 Formula:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Where:

- X = original data point
- X_{min} = minimum value in the dataset
- X_{max} = maximum value in the dataset

Given Dataset:

$$X = \{24, 28, 53, 30, 40, 18, 15, 21\}$$

Step 1: Identify Minimum and Maximum

$$X_{\text{min}} = 15, \quad X_{\text{max}} = 53$$

Step 2: Apply the Min-Max Scaling Formula

$$X_{\text{scaled}} = \frac{X - 15}{53 - 15} = \frac{X - 15}{38}$$

X	(X - 15) / 38	Scaled Value (approx)
24	(24 - 15)/38 = 9/38	0.237
28	13/38	0.342
53	38/38	1.000
30	15/38	0.395
40	25/38	0.658
18	3/38	0.079
15	0/38	0.000
21	6/38	0.158

Final Min-Max Scaled Dataset:

$$\{0.237, 0.342, 1.000, 0.395, 0.658, 0.079, 0.000, 0.158\}$$

📌 **What is a Categorical Variable?**

Categorical variable wo hota hai **jisme limited set of values** hote hain, jaise:

- **Gender** → Male, Female
- **City** → Mumbai, Delhi, Pune
- **Grade** → A, B, C

Yeh variables **numerical nahi hote**, lekin unhe model ke liye **numerical format me convert** karna padta hai.

❓ Why Do We Need Categorical Variable Encoding?

Machine Learning models **text ko directly samajh nahi sakte**, isiliye categorical variables ko **numbers me encode karna** padta hai so that:

- Model easily calculation kar sake
- Better accuracy mile
- Data ML-friendly ban jaye

🧠 **One-Hot Encoding Explained (with Example)**

One-Hot Encoding har category ke liye **ek separate binary column** create karta hai (0 ya 1).

🔸 **Example:**

Original Column (City):

Mumbai
Delhi
Pune

One-Hot Encoded:

Mumbai	Delhi	Pune
1	0	0
0	1	0
0	0	1

Python Syntax (Pandas):

```
import pandas as pd

df = pd.DataFrame({'City': ['Mumbai', 'Delhi', 'Pune']})
encoded = pd.get_dummies(df['City'])
print(encoded)
```

🔁 **Handling 2 or More Categories:**

- **Binary (2 categories)** → Use **Label Encoding** (e.g., Male = 0, Female = 1)
- **More than 2 categories** → Use **One-Hot Encoding** (as shown above)

Statistical and Graphical Analysis Methods

&

📊 Role of Visualization in Big Data Analytics

1. Statistical Analysis Methods

Statistical analysis ka matlab hota hai data ka **summary nikaalna using numbers**.
Isme hum use karte hain:

Method	Use (HiEnglish)
Mean	Average value batata hai
Median	Beech ka value, outliers ko ignore karta hai
Mode	Sabse zyada baar repeat hone wala value
Standard Deviation	Data kitna spread hai, yeh batata hai
Correlation	Do variables ke beech relation check karta hai

🖼️ 2. Graphical Analysis Methods

Yeh methods data ko **visual form me represent karte hain** jisse data ko samajhna easy ho jaata hai.

Graph Type	Use (HiEnglish)
Histogram	Frequency distribution dikhata hai
Boxplot	Outliers aur spread dikhata hai
Bar Chart	Category-wise comparison karta hai
Scatter Plot	Do variables ke beech relation show karta hai
Heatmap	Correlation matrix dikhata hai visually

🎯 Role of Visualization in Big Data Analytics

Visualization big data me **insights nikalne ka sabse powerful tool** hai.

Point	Explanation (HiEnglish)
🔍 Pattern Identification	Data ke andar hidden patterns ko jaldi samajh sakte hain
🚨 Anomaly Detection	Outliers aur errors ko identify karna easy ho jaata hai
📈 Trend Analysis	Time series ya progress ko visually dekh sakte hain
📊 Easy Communication	Non-technical log bhi visualizations ke through data samajh lete hain
⚡ Faster Decision-Making	Visuals help in making quick business decisions

Hive Architecture

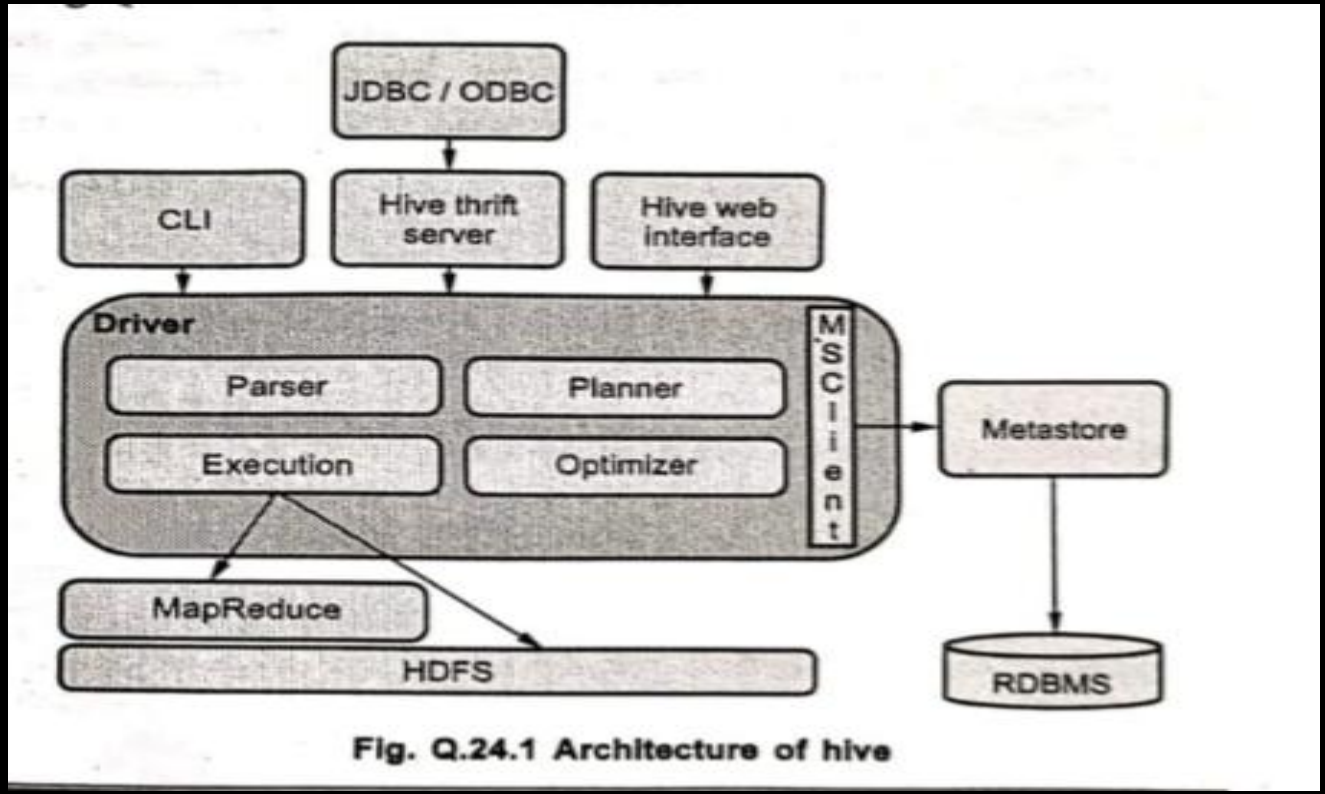


Fig. Q.24.1 Architecture of hive

📋 Hive Architecture Components

Component	Explanation (HiEnglish)
User Interfaces	Hive ke sath interact karne ke 3 tareeke – CLI (Command Line), JDBC/ODBC, Web UI
Driver	Yeh pura query process karta hai (like a controller)
→ Parser	Query syntax check karta hai
→ Planner	Logical plan banata hai (kaise query run hogi)
→ Optimizer	Query ko optimize karta hai best performance ke liye
→ Executor	Optimized query ko execute karta hai
Metastore	Metadata store karta hai (tables, columns, data types) — Usually stored in RDBMS
Hive Clients	JDBC/ODBC, CLI, Hive Thrift Server — yeh client layer provide karte hain
HDFS (Storage)	Hive ka main storage system — Hadoop Distributed File System
MapReduce Engine	Hive ke queries MapReduce me convert hoke parallel execute hote hain

🎯 Example of Hive Use

Use-case: Suppose a company wants to analyze **monthly sales data** stored in HDFS.

•They write:

```
SELECT region, SUM(sales) FROM monthly_sales GROUP BY region;
```

•Hive steps:

- **Parser**: Syntax check karega
- **Planner**: Query plan banayega
- **Optimizer**: Best way decide karega query run karne ka
- **Executor**: Query ko MapReduce jobs me convert karke HDFS pe run karega
- **Result**: User ko summarized sales by region mil jaayega

Parameter	**HBase** (NoSQL)	**Hive** (SQL-like on Hadoop)
Type	NoSQL database (Column-oriented)	Data warehouse tool (SQL-like)
Data Model	Table → Row → Column family → Columns	Tables with rows and columns (like SQL)
Query Language	HBase Shell / Java API	HiveQL (similar to SQL)
Use Case	Real-time read/write of big data	Batch processing, data analysis
Speed	Fast for real-time access	Slow (used for batch queries)
Storage	HDFS (Hadoop Distributed File System)	HDFS
Data Format	Non-relational (unstructured/semi-structured)	Structured data (like Excel tables)
Schema	Schema-less (flexible)	Fixed schema (defined columns)
Processing Type	Real-time	Batch processing
Integration	Works with Apache Hadoop	Built on top of Hadoop + MapReduce
Best For	Applications like Facebook messages	Reports, Analytics, Large scale SQL queries

🔹 **What is Data Wrangling? (Data Wrangling kya hota hai?)**

Data Wrangling ka matlab hota hai:

Raw (ganda, unclean) data ko clean, organize, aur usable format mein convert karna.

👉 Jab aapko real-world data milta hai (jaise CSV, Excel, databases), woh aksar incomplete, messy, ya inconsistent hota hai.

Data wrangling use clean aur structured banata hai so that aap us par analysis ya machine learning kar sako.

◆ Why Do You Need Data Wrangling? (Hume Data Wrangling kyun chahiye?)

Reason	Explanation (HiEnglish)
1 Better Accuracy	Clean data se analysis ya prediction zyada accurate hota hai.
2 Remove Errors	Raw data mein errors, null values hote hain — unhe remove karte hain.
3 Consistency	Sabhi data ek format mein ho — isse process karna easy hota hai.
4 Time Saving	Wrangled data se reports jaldi ban jati hain, analysis fast hota hai.

🔹 Data Wrangling Methods (Steps ya Techniques)

Step	Method Name	HiEnglish Explanation
1️⃣	**Data Collection**	Data ko different sources se gather karna (CSV, APIs, DBs).
2️⃣	**Data Cleaning**	Missing values, duplicate rows, spelling mistakes ko fix karna.
3️⃣	**Data Transformation**	Data ka format change karna (jaise dates ko dd-mm-yyyy mein lana).
4️⃣	**Data Enrichment**	Extra useful info add karna (jaise "Age" se "Age group" banana).
5️⃣	**Data Validation**	Check karna ki data correct hai ya nahi (rules ke according).
6️⃣	**Data Formatting**	Final format mein data ko arrange karna (columns rename, sort karna).

🔹 **What is a Dataset? (Dataset kya hota hai?)**

Dataset ek collection hoti hai **data** ki — rows aur columns ke form mein. Aap ise **Excel sheet**, **CSV file**, **database table**, ya koi bhi **structured data format** samajh sakte ho.

👉 Dataset is used in **Big Data**, **Data Science**, and **Machine Learning** to train models, analyze trends, and more.

🔹 2 Common Types of Datasets in Big Data (with Python Syntax)

1. Structured Dataset

•Ye data rows/columns ke format mein hota hai, jaise Excel sheet ya SQL table.

•Example format: .csv, .xls, SQL databases

```
import pandas as pd

# CSV file se structured dataset read karna
df = pd.read_csv("data.csv")

# First 5 rows print karna
print(df.head())
```

📌 **Use Case:** Sales data, Employee records, Bank transactions, etc.

2. Unstructured Dataset

•Ye data structured format mein nahi hota — jaise text, images, videos, logs.

•Example: Tweets, YouTube comments, images, etc.

- **Python Syntax for Text Dataset (Unstructured):**

```
# Unstructured text data example
text_data = [
    "Big Data is amazing!",
    "Hadoop and Spark are powerful tools.",
    "Data Wrangling is necessary for analysis."
]

# Convert text data to structured format using pandas
import pandas as pd
df = pd.DataFrame(text_data, columns=["Comments"])
print(df)
```

jayesh_kande_

Jayesh Kande

16	275	276
posts	followers	following

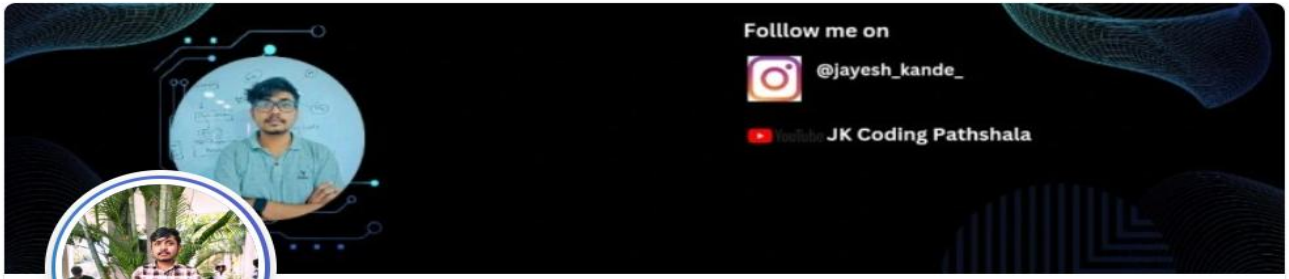
23

रास्ते बदलो, मंजिल नहीं

yt.openinapp.co/0y0qd

LinkedIn

Articles | People | Learning | Jobs | Games



Jayesh Kande

Third-Year IT Engineering Student | Aspiring Web Developer | Java Enthusiast | Data Structures & Algorithms Learner | Proficient in C, C++, Java, and MERN Stack | AI + Web Development Project Enthusiast

Nashik, Maharashtra, India · Contact Info

494 followers · 495 connections

See your mutual connections

Kbt engineering college nashik

Join to view profile | Message

✨ Thank You for Watching! ✨

📲 Follow us on Instagram:**@jayesh_kande_**
🔗 Connect with us on LinkedIn:[**Jayesh Kande**]