**Vid 1**

**Big Data :**

Big data

Big Data: Large, complex data sets requiring advanced processing techniques.

**Big Types :**

Structured Data – Structured data: organized, fixed-format data in databases, easily searchable.

Unstructured Data – Unstructured data: raw, unorganized information without predefined format or schema.

Semi-Structured Data – Semi-structured data: flexible format, partial organization, like JSON, XML.

Quasi-Structured Data – Quasi-Structured Data woh data hota hai jo partially structured hota hai, jaise JSON ya XML, jaise log files, clickstream data, aur sensor data.

**Distributed storage systems**

Distributed storage systems, multiple servers par data store karne ka system hota hai.

**How to store and process Big Data ek sentence mein**

Big Data ko store karne ke liye distributed storage systems (jaise Hadoop HDFS, AWS S3) use kiya jata hai,

aur process karne ke liye parallel computing frameworks (jaise Apache Spark, MapReduce) ka use hota hai.

**Data generated per minute**

Har minute duniya mein lagbhag 1.7 MB data per person generate hota hai, jo social media posts, online transactions, searches aur sensors se aata hai.

**Big Data ka application**

Data Engineering mein Big Data ka application data ko store, process aur analyze karne mein hota hai.

**Healthcare:** Data se treatment personalize aur disease pehle predict kiya jaata.

**Finance:** Data se fraud pakadte aur market risks ko manage karte.

**Retail:** Customer ka behavior samajhkar inventory aur sales ko manage karte.

**Manufacturing:** Machines ka data dekhar repair aur supply chain ko optimize karte.

**Social Media & Marketing:** Data se ads target aur customer ka mood samjha jaata hai.

**Transportation:** Traffic control, route planning aur smart vehicles ko data guide karta.

**Energy Sector:** Data se electricity use predict aur waste ko control kiya jaata.

**Education:** Students ka data padhkar unka study plan better banate hain.

**Government:** Public service aur law improve karne ke liye data ka use.

**Sports:** Players aur fans ka data padhkar strategy aur experience better.

Vid 2

**storage unit**

storage unit data ko store, manage aur access karne ke liye organize kiya jata hai.

**YARN**

YARN (Yet Another Resource Negotiator) ek resource management layer hai

jo Hadoop ecosystem mein alag-alag applications ke liye system resources allocate aur manage karta hai.

**Hadoop ka use**

Hadoop ka use large-scale data ko efficiently store, process aur analyze karta hai.

**Hadoop**

Hadoop ek open-source framework hai, jo big data ko distributed tareeke se store aur process karta hai.

**Hadoop Applications :**

बिलकुल! यहाँ Hadoop applications को simple और clear एक-एक line में समझाया है:

1️⃣ Big Data Storage and Processing — Hadoop बड़े data को store और जल्दी process करने में मदद करता है।  
 2️⃣ Data Warehousing — Hadoop बहुत बड़े data को traditional warehouses की जगह सस्ते में store करता है।  
 3️⃣ Log and Event Data Analysis — Hadoop logs और events का analysis करके user behavior को समझने में मदद करता है।  
 4️⃣ Recommendation Systems — Hadoop का use personalized recommendations देने वाले systems बनाने में होता है।  
 5️⃣ Fraud Detection — Hadoop suspicious transactions पकड़ने और fraud alerts देने में मदद करता है।  
 6️⃣ Search Engines — Hadoop billions webpages को crawl और index करने में search engines को help करता है।  
 7️⃣ Social Media Analytics — Hadoop social media data (posts, comments) का analysis करके trends दिखाता है।  
 8️⃣ Healthcare and Genomics — Hadoop medical data और DNA information को analyze करने में doctors को support करता है।  
 9️⃣ Retail and E-Commerce — Hadoop customer के खरीदने के patterns और market trends समझने में use होता है।  
 🔟 IoT and Sensor Data Processing — Hadoop IoT devices से आए massive data को efficiently process करता है।

Vid 3

**HDFS**

(Hadoop Distributed File System) ek distributed storage system hai jo bade data ko multiple machines par store karke efficiently manage karta hai.

**MapReduce**

MapReduce data ko process aur analyze karne ke liye use hota hai.

MapReduce ek programming model hota hai jo large data ko parallel processing ke through "Map" (filtering & sorting) aur "Reduce" (summarizing) steps mein process karta hai.

**Flume**

Data engineering mein Flume ek tool hai jo log data ko easily collect karke Hadoop jaise system tak pahunchata hai.

**Sqoop**

Data engineering mein Sqoop ek tool hota hai jo Hadoop aur relational databases ke beech data transfer karne ke liye use hota hai.

**Flume and Sqoop**: Used for data collection and injection. Sqoop handles structured data, while Flume is for unstructured or semi-structured data

**HBase**

Data engineering mein HBase ek distributed, scalable, NoSQL database hai jo large datasets ko real-time mein store aur access karne ke liye Hadoop ke upar kaam karta hai.

**Cloudera**

Data engineering mein Cloudera ek platform hai jo big data ko store, process aur analyze karne ke liye Hadoop-based tools ka use karta hai.

**Hive**

Data engineering mein Hive ek data warehouse tool hai jo Hadoop ke upar SQL jaisa interface provide karta hai taaki hum large datasets ko easily query aur manage kar saken.

**Pig**

Data Engineering mein Pig ek high-level platform hai jo Hadoop par large data sets ko process karne ke liye scripting language (Pig Latin) ka use karta hai.

**Mahout**

Data engineering mein Mahout ek open-source library hai jo scalable machine learning algorithms ko big data par apply karne ke liye use hoti hai, khas kar Apache Hadoop ke sath.

**Oozie**

Oozie ek workflow scheduler hai jo Hadoop ecosystem mein data pipelines ko automate karne ke liye use hota hai.

**ZooKeeper**

Data engineering mein ZooKeeper ek open-source tool hota hai jo distributed systems ke coordination, configuration management, aur synchronization ke liye use hota hai.

VID 4

Data Engineering mein Map Phase ek process hota hai jisme data ko key-value pairs mein tod kar transform kiya jaata hai, taaki uspar efficiently processing aur grouping ki jaa sake.

data engineering mein Reduce Phase kya hota hai sentence mein

Data Locality ka matlab hota hai — data ko us jagah rakhna jahan use process karna hai, taki data ko door se lana na pade aur kaam jaldi ho jaye.

Data Engineering mein Map and Reduce core components hote hain jo data ko process karne ke liye — Map data ko filter aur transform karta hai, aur Reduce summarized output generate karta hai.

Bilkul! Main ab simple aur samajhne layak (understanding) way mein explain karta hoon — hinglish mein, jaise teacher samjhaye — step by step:

1️⃣ MapReduce Definition:  
 MapReduce ek programming model hai jo large datasets ko parallel aur distributed way mein process karta hai — yani data ko chhote-chhote parts mein tod kar multiple machines pe fast process karwana.

2️⃣ Key Components:  
 Isme 2 main parts hote hain — Mapper aur Reducer.  
 👉 Mapper: Data ko small pieces mein todta hai aur key-value pair banata hai.  
 👉 Reducer: In pairs ko combine karke final result nikalta hai.

3️⃣ Divide and Conquer Strategy:  
 MapReduce problem ko break karta hai — "divide" karta hai tasks ko chhote-chhote parts mein aur alag-alag systems pe "conquer" karta hai yani solve karta hai, isse processing fast ho jati hai.

4️⃣ Mapper Class Function:  
 Mapper input data ko line by line read karta hai, usko logical keys aur values mein todta hai — jaise Word Count mein "Word, 1" ka pair banana.

5️⃣ Reducer Class Function:  
 Reducer same keys wale data ko collect karta hai aur unko add ya merge karke meaningful output deta hai — jaise total count batana for each word.

6️⃣ Core Daemons:  
 MapReduce ko handle karne ke liye 2 main helper background processes hote hain — Job Tracker aur Task Tracker.

7️⃣ Job Tracker Role:  
 Job Tracker ka kaam hota hai complete job ko plan karna, kaha pe kaunsa task jayega ye decide karna aur ye check karna ki sab kuch theek se execute ho raha hai ya nahi.

8️⃣ Task Tracker Role:  
 Task Tracker, Job Tracker se assigned task ko real mein data node pe execute karta hai aur result aur status wapas bhejta hai.

9️⃣ Data Locality Principle:  
 Data ko transfer karne se time waste hota hai, isliye MapReduce logic yeh follow karta hai ki "code ko data ke paas le jao" taaki minimum network usage ho aur speed fast mile.

🔟 Word Count Example:  
 Agar humare paas ek bada text file hai, to:  
 ➡️ Mapper har word ko dekh kar uske liye (word, 1) pair banayega.  
 ➡️ Reducer in sab pairs ko add karke batayega ki har word kitni baar aaya.

✅ Bas is process se large data ko bhi asaani se fast aur scalable way mein process kiya jata hai!

Agar chaaho to main iska ek simple diagram bhi bana kar bhej sakta hoon, aur aur easy example bhi share kar sakta hoon! Batau? 🎯📊