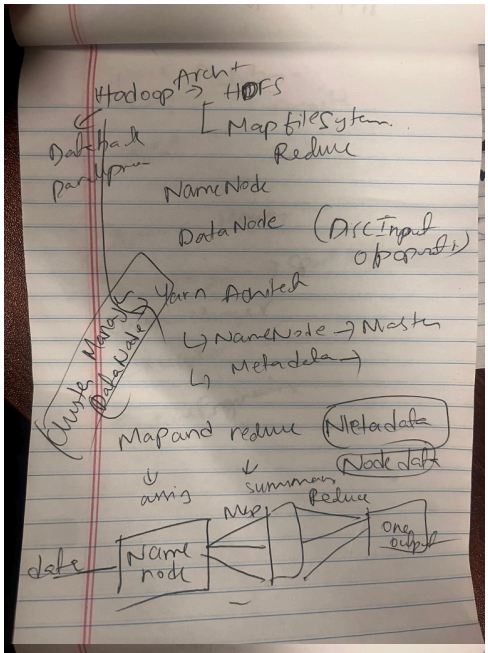
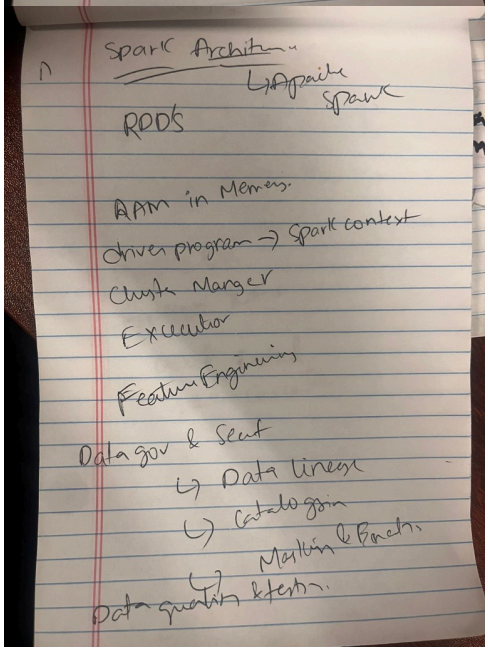
Day-2

Wednesday, January 21, 2026 7:28 PM



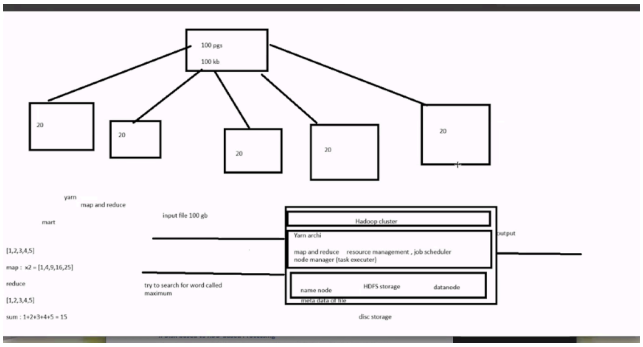


Hadoop Architecture

- **HDFS** → Distributed storage (NameNode, DataNode)
- **YARN** → Resource management
- **MapReduce** → Batch processing engine

Spark Architecture

- **Driver** → Creates DAG, schedules tasks
- **Executors** → Run tasks in parallel
- **Cluster Manager** → YARN / Kubernetes
- In-memory → **Faster than MapReduce**



5. Hadoop vs Spark Comparison

Aspect	Hadoop (Disk-Based)	Spark (RDD-Based)
Data Storage	Stored and processed from disk (HDFS)	Stored temporarily in memory (RDDs)
Intermediate Results	Written to disk after each stage	Kept in memory until needed
Performance	Slower due to disk I/O	Faster due to in-memory caching
Fault Tolerance	Replication-based	RDD lineage-based recomputation
Use Cases	Batch ETL and archival processing	Interactive analytics, ML, and real-time processing

Data Governance & Security

- **Data Lineage**: Track data origin and flow
- **Cataloging**: Tools like AWS Glue Catalog
- **Masking & Encryption**: Secure sensitive data

. **Data Lineage** = *"Where did the data come from?"*
Data Cataloging = *"Index or Google for your data"*

Tools like **AWS Glue Data Catalog** or **Apache Atlas** help:

- List all your data tables/files
- Show what each table/column means
- Who owns the data
- When it was last updated

Masking & Encryption = ***"Hide or lock the secret data"***

- **Masking**: Hide sensitive data like phone numbers or credit card info.
  - Example: Show ****1234 instead of full card number

Hadoop is still used, m
companies are moving

mapping is used thro
systems to target sy