Cluster:- grop of virtual computer

Apache Spark Streaming

Processing
Data Storage
Dstream in Spark

Apache Spark MLib (ML)

- Spark Context
- DAG Scheduler
- Task Scheduler
- Scheduler Backend

→ x ←

(i) What is spark apache?
(ii) Why apache spark? what problem does it solve?

Apache Spark is a unified computing engine and set of libraries for parallel data processing

(iii) What is unified?
(iv) what is computing engine?
(v) what is libraries?
(vi) what is parallel data processing?

Unified

⇒ Spark is designed to support wide range of task over the same computing engine for e.g. Data scientist, Data engineer, Data analyst all can use the same platform for

their Transformation or modelling.

Computing Engine Storage

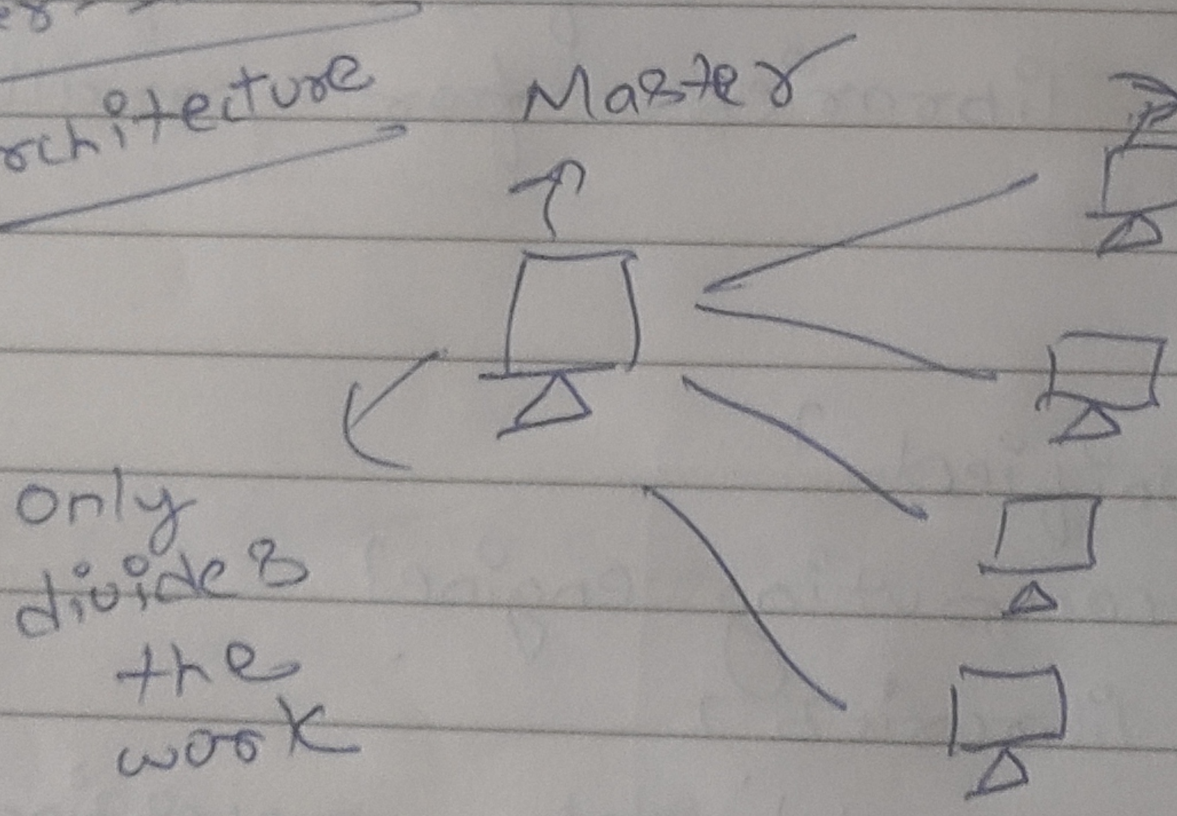
→ Spark is limited to a computing Engine.
It doesn't store the data.

→ Spark can connect with different data source
S3, HDFS, JDBC/ODBC, Azure storage etc.

→ Spark work with almost all data storage
system

Computer Cluster

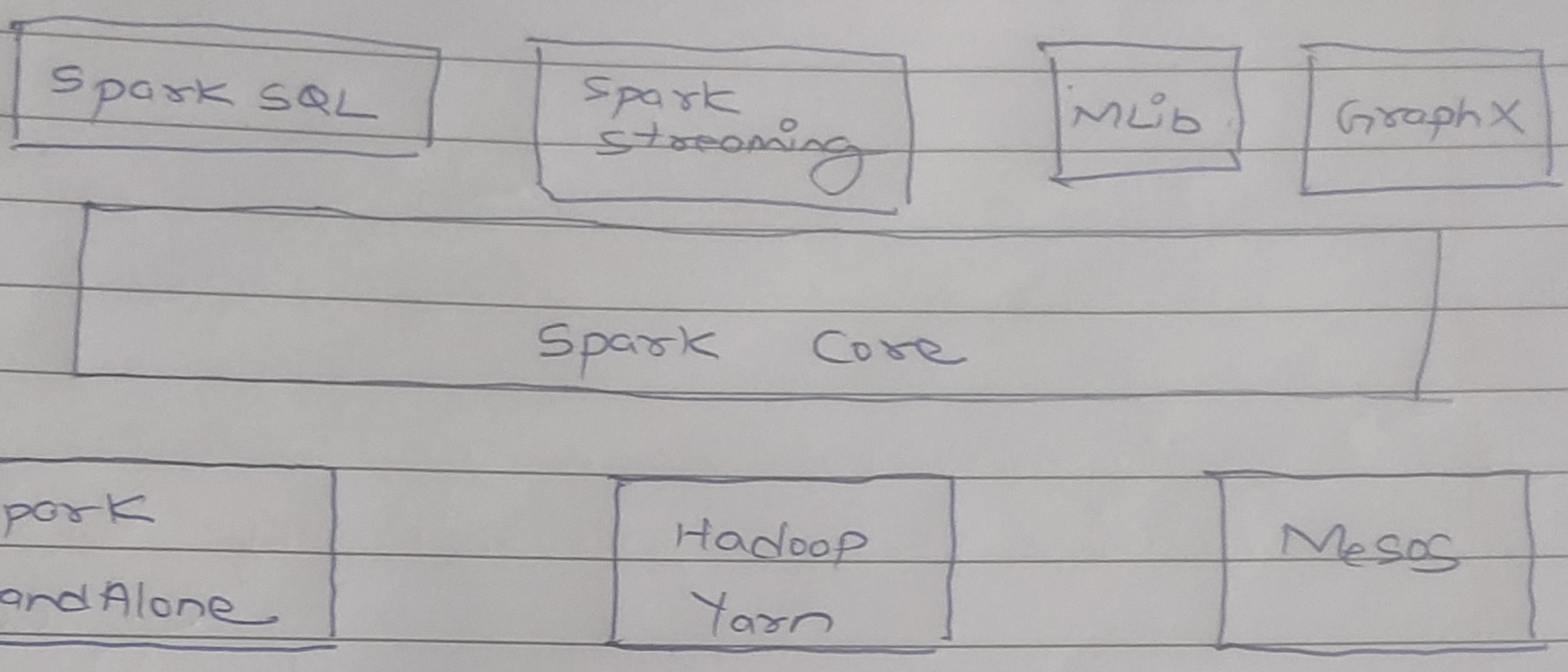
Master-Slave Architecture



Spark features

- Spark written in Scala, runs in JVM
- API : Scala, JAVA, Python, R
- Interactive shell : Scala and Python
- Data source : SQL, NoSQL, S3, HDFS, Local File System, etc.
- Good fit for iterative task like ML

Spark Components



How Spark works