What is Big Data?

5 Vs of Big Data

Stages of Big Data Analytics

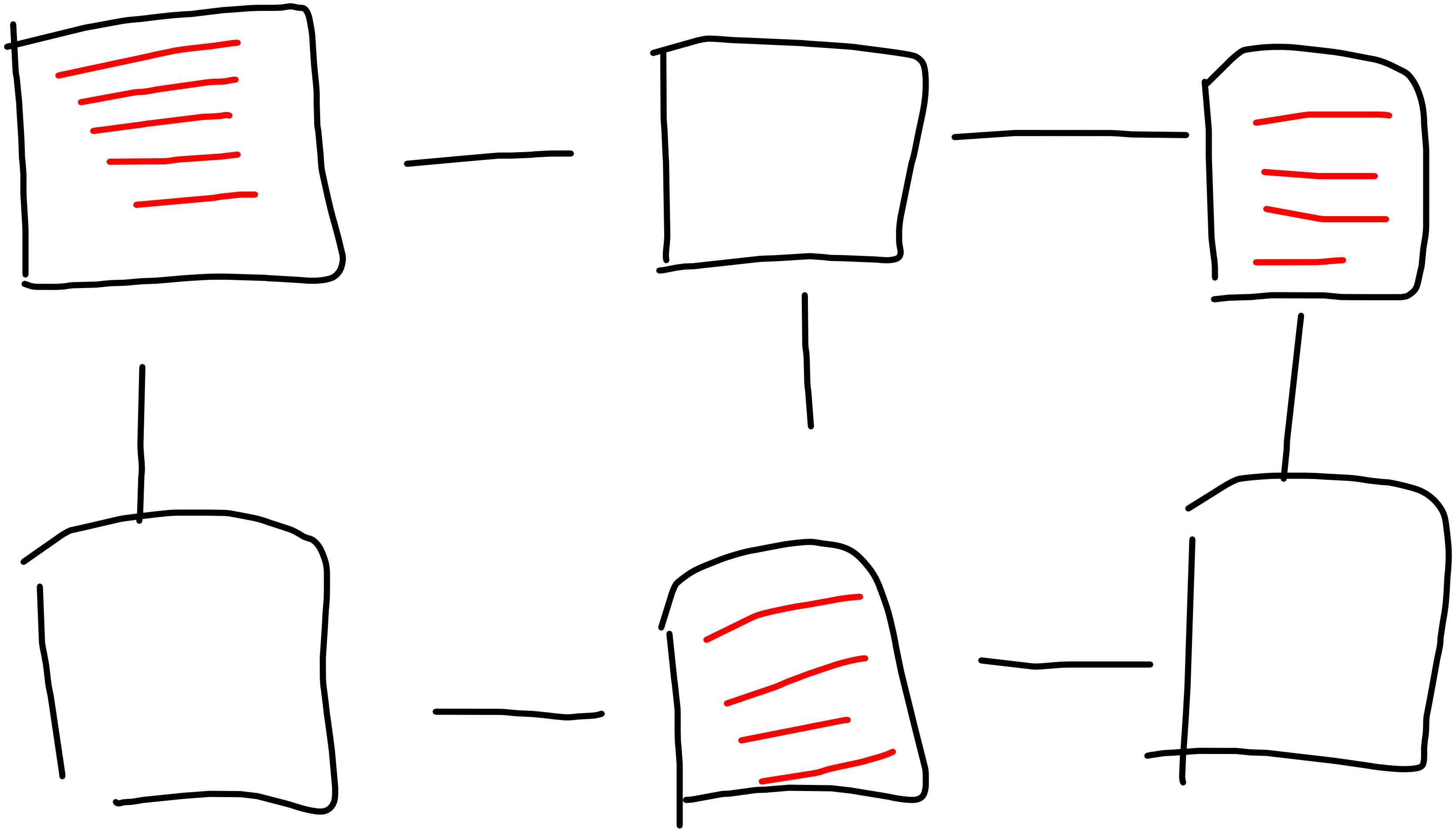
Types of Big Data Analytics

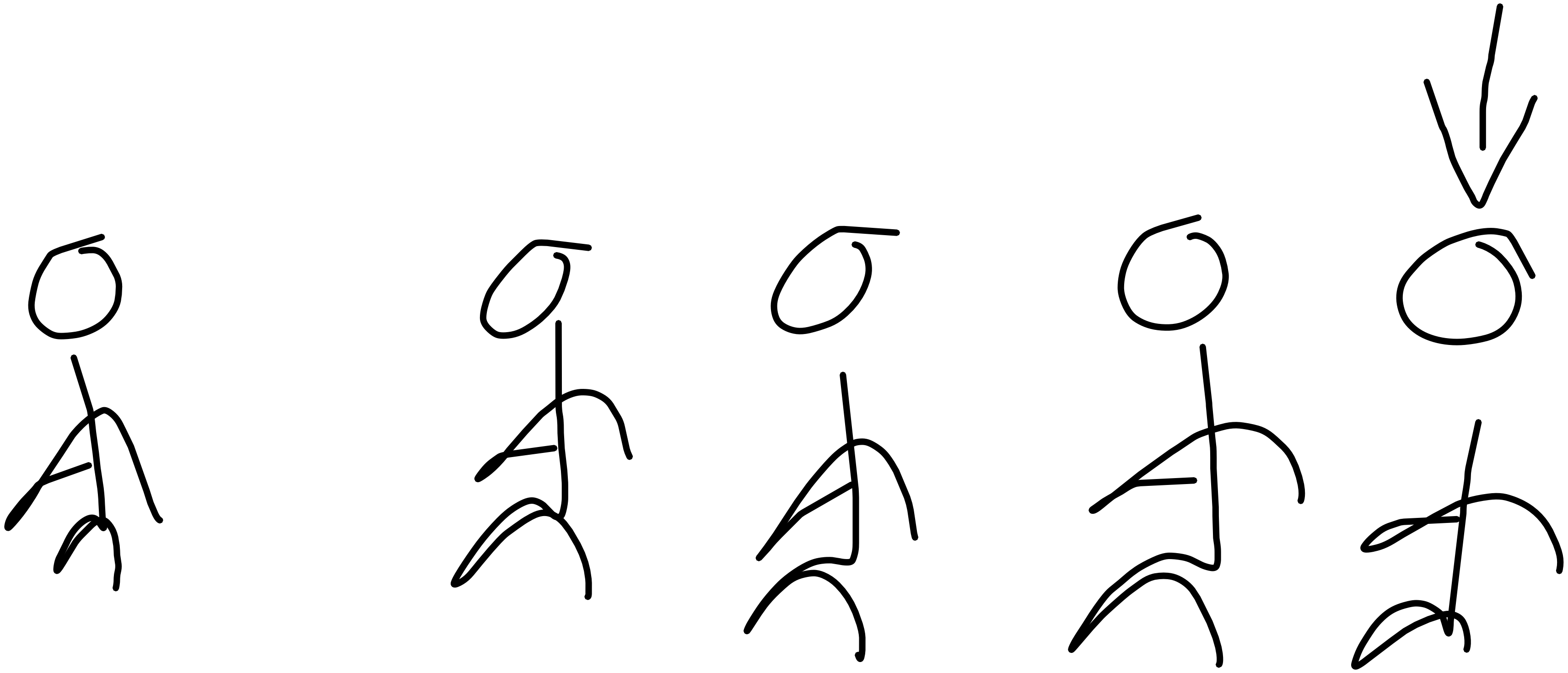
Applications of Big Data Analytics

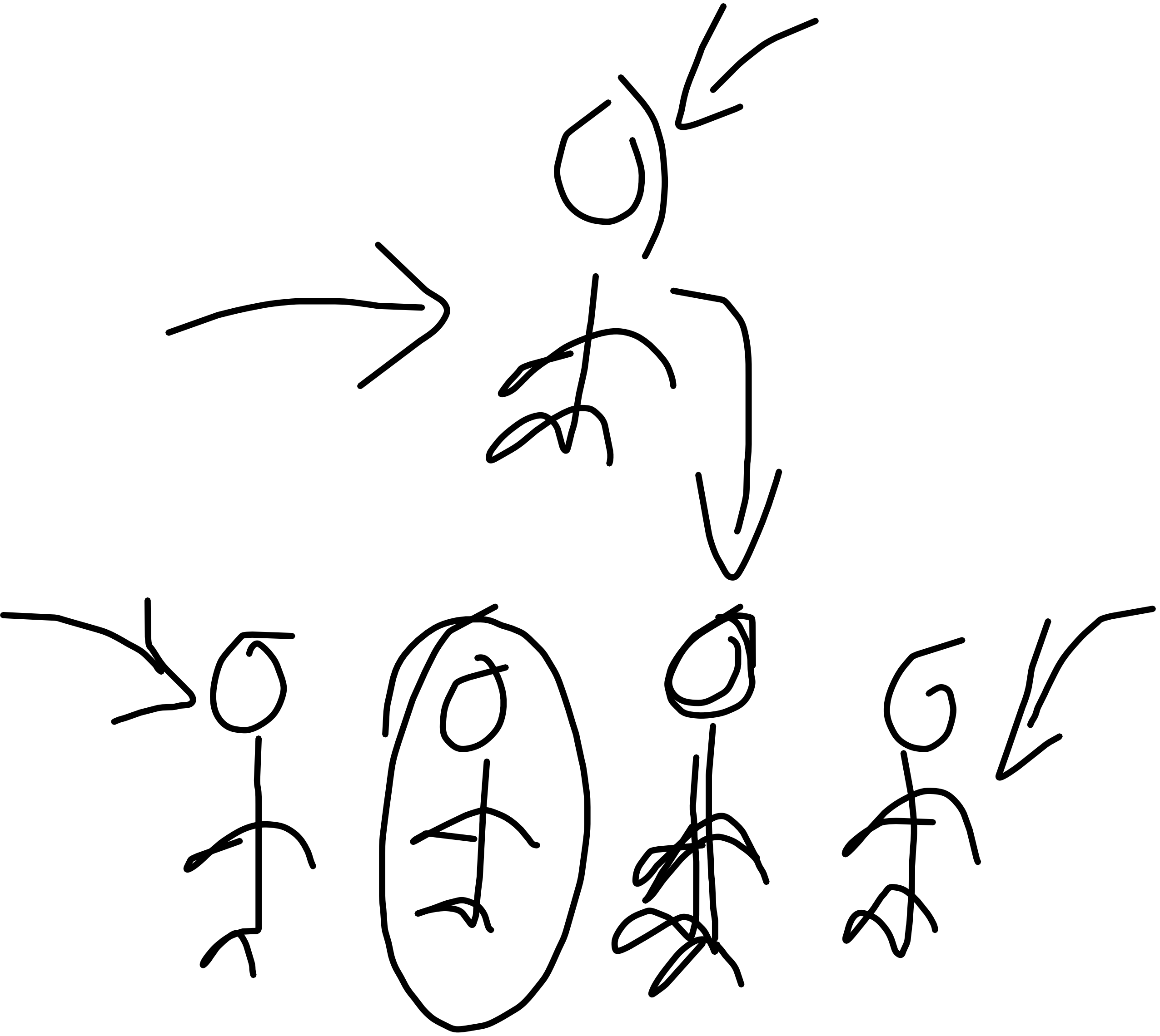
Problems of Big Data Analytics

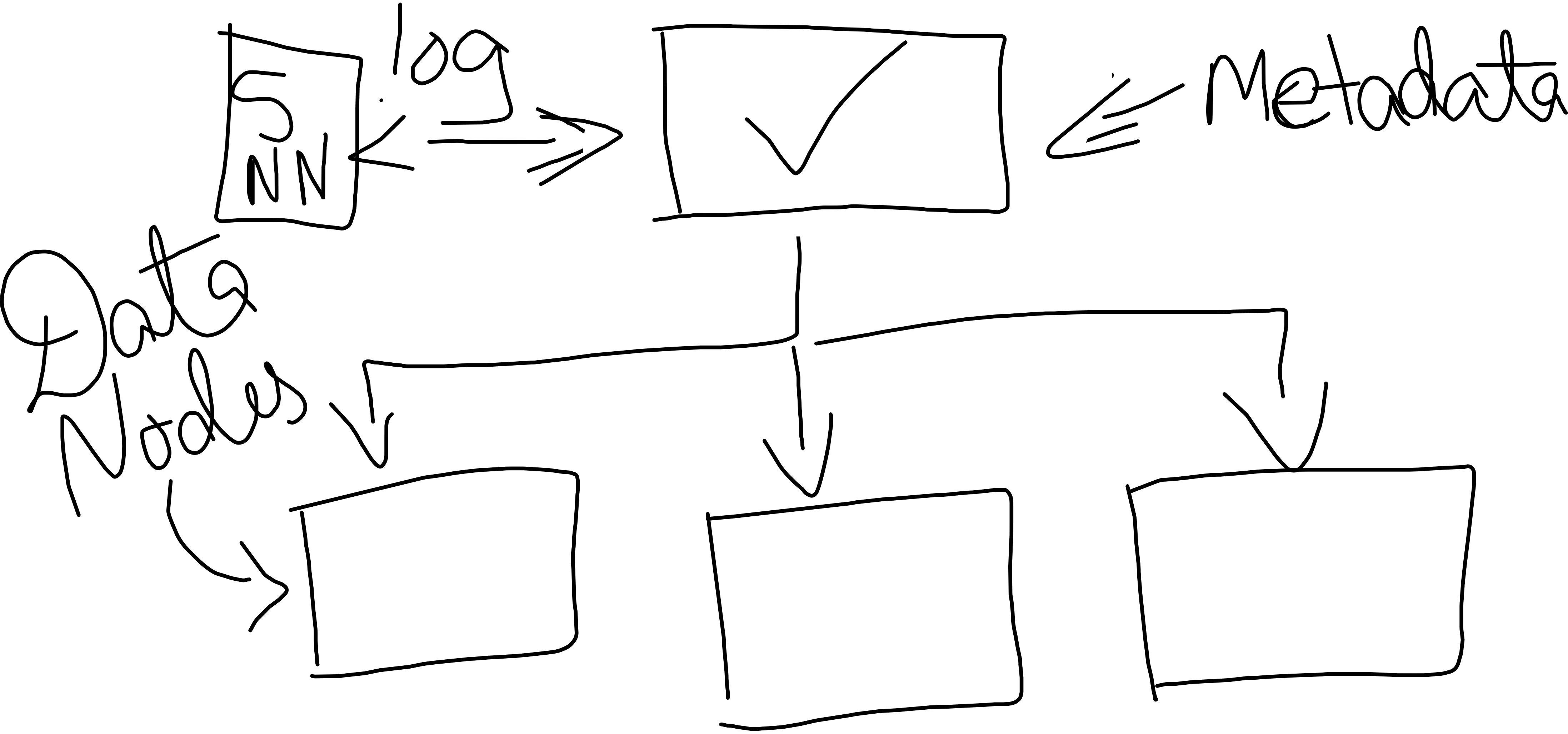
What is Hadoop?

Components of Hadoop

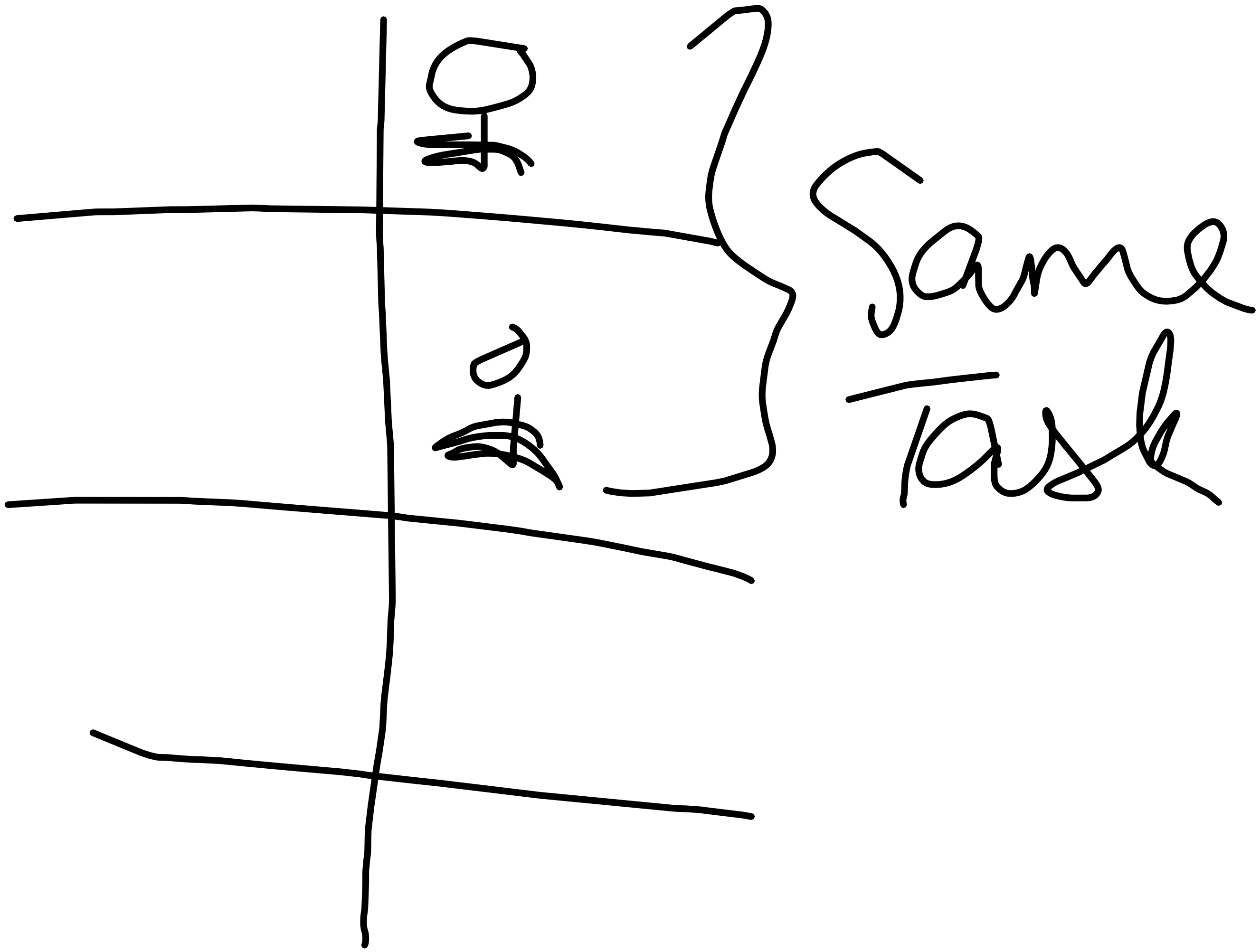




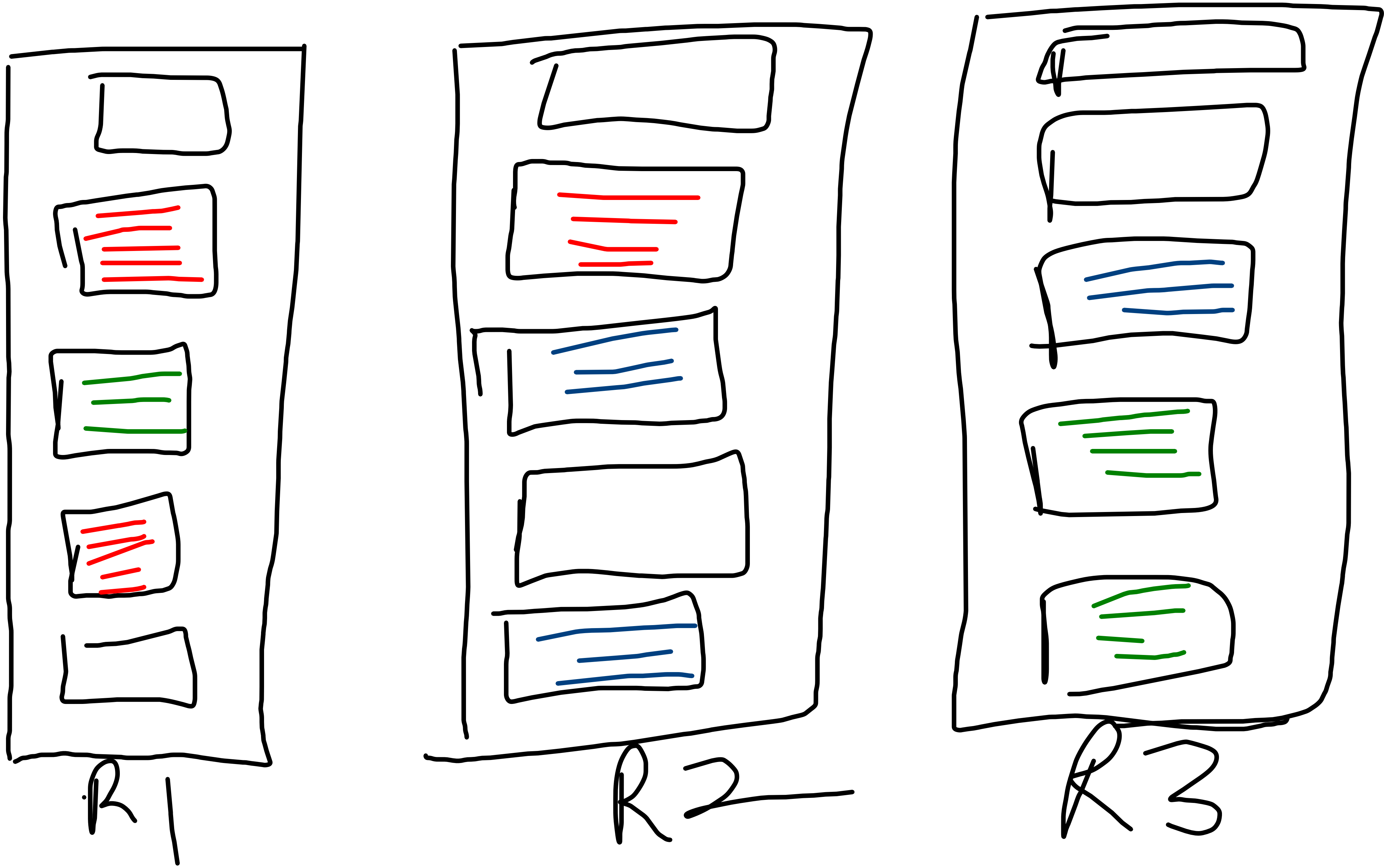



SNN
log
✓
Metadata
Data Nodes

NameNode
Master Nodes
Metadata
High Quality Hardware
DataNode
Slave Nodes
Data
Commodity Hardware
Secondary NameNode (Maintain logs)


Same
Task

Rack Awareness



Advantages of Hadoop

Highly Available

Cost Effective

Fault Tolerance

Horizontally Scalable

Open Source

Variety of Data Sources

High Throughput

Disadvantages of Hadoop

Vulnerability

Security (Kerberos Authentication)

Processing Overhead (Read/Write Operations)

Only Batch Processing

Hive

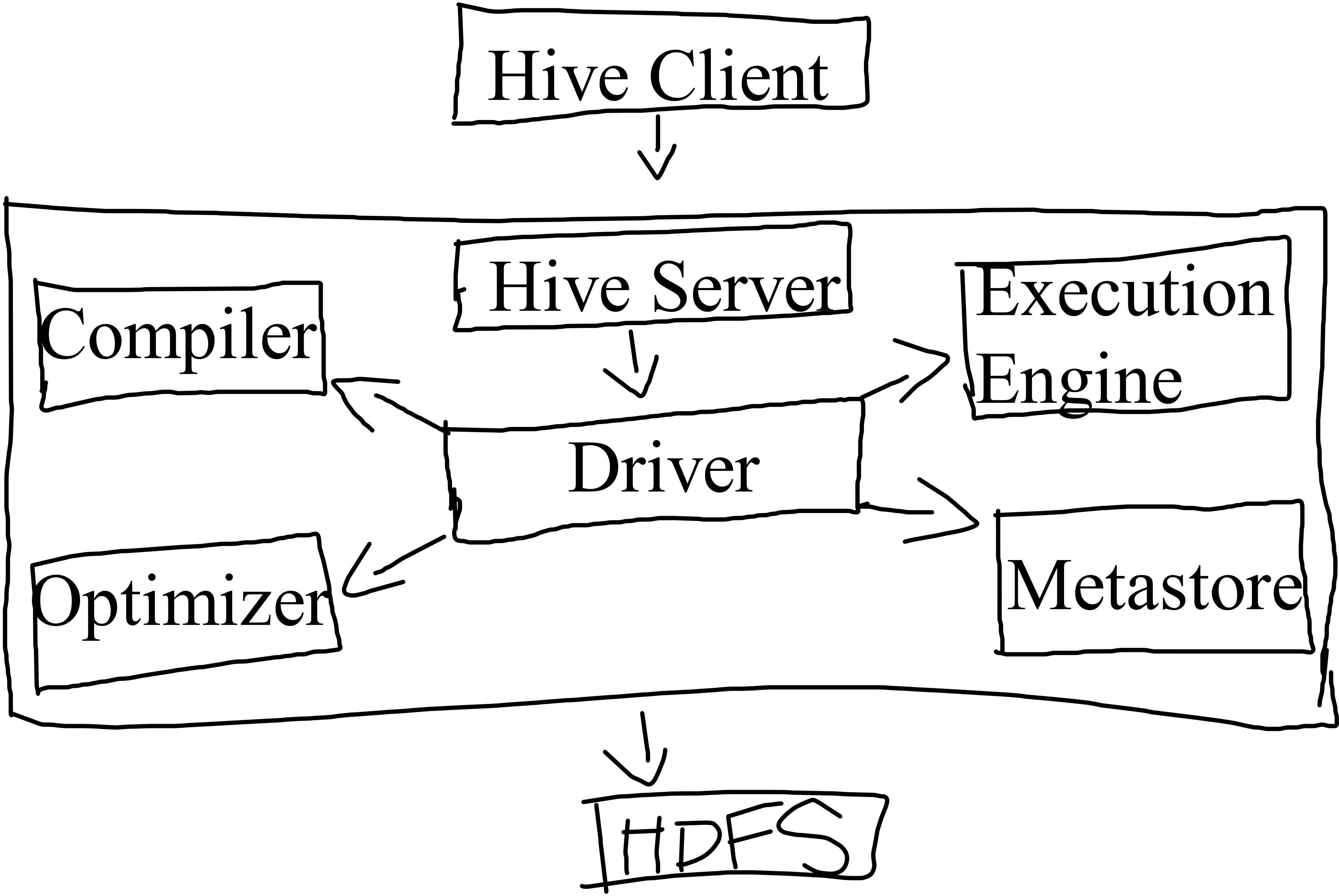
Before big data => SQL
After big data => MapReduce => Java

Employee
CEO

SQL => Hive => MapReduce => HDFS
Hive Query Language (HQL)

Hive gives an SQL like interface to query data stored in HDFS.

Developed by Facebook.
Contributions from Netflix, FINRA.



Hive Client

Hive Server

Compiler

Execution Engine

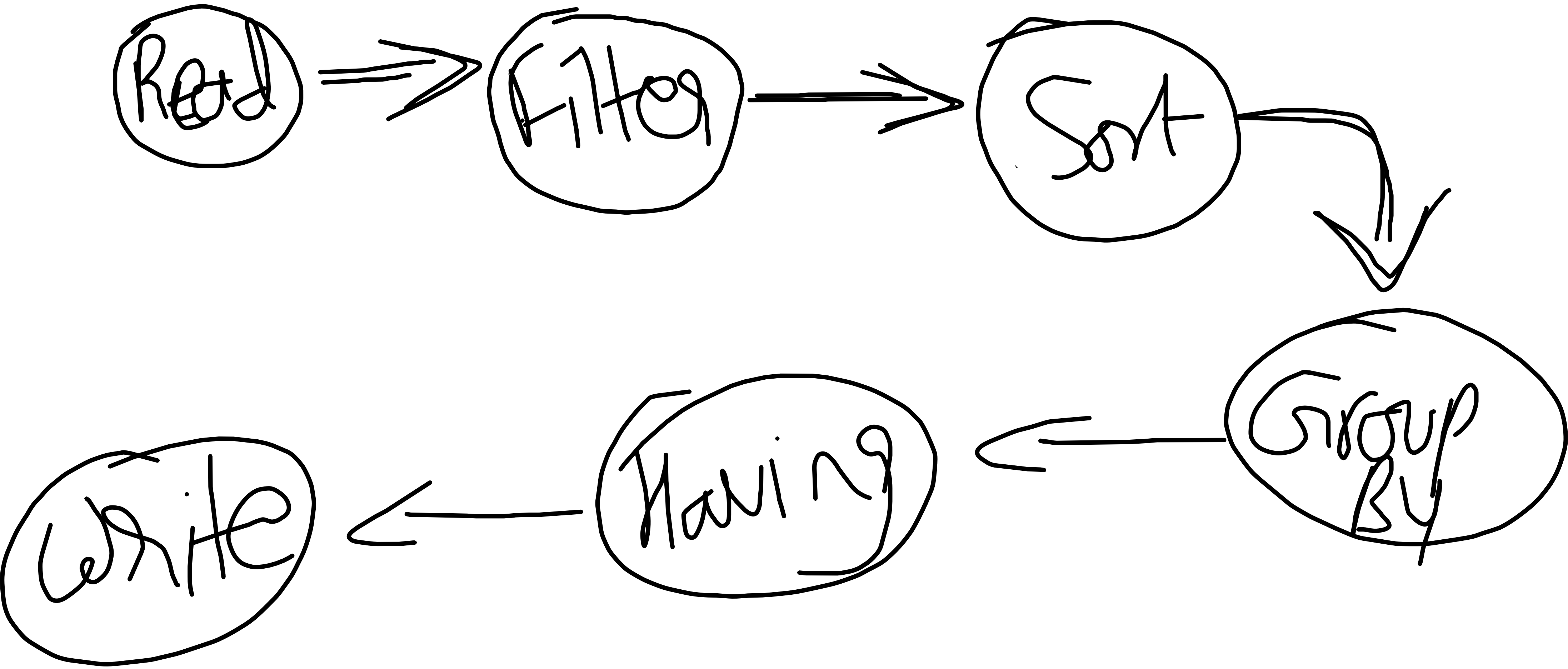
Driver

Optimizer

Metastore

HDFS

Directed Acyclic Graph (DAG)



Data Model of Hive

Tables

Partitions

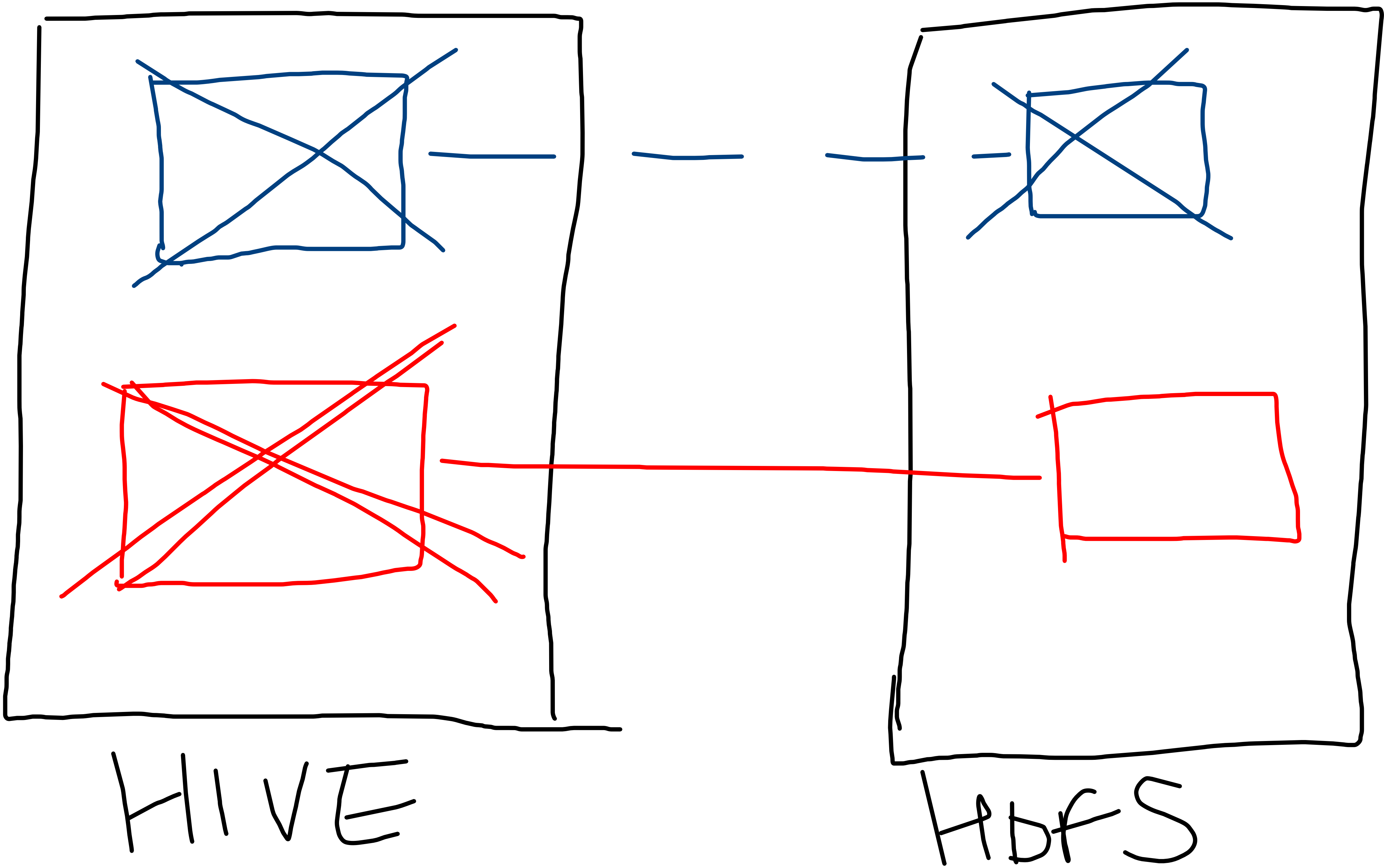
Buckets

Managed Tables (Default)

Do not control the creation and deletion of data

External Tables

Control the creation and deletion of data


HIVE
HDFS

Partitions

Country	Sales
India	$10
Japan	$7
India	$8
Japan	$5
China	$17

```
Select *
from table
where
Country='India';
```

COUNTRY	Sales
India	$10
India	$7
India	$5
Japan	$17
Japan	$7
Japan	$9

select *
from table
where
County='India'

Apache Hive organizes tables into partitions for grouping same type of data together based on a column or partition key.

Faster and efficient queries.

Buckets

In Hive, Tables or partitions are subdivided into buckets based on the hash functions of a column in the table to give extra structure to the data that may be used for more efficient queries.

Apache Pig

Pig Latin => Pig => MapReduce => HDFS

Developed at Yahoo Research

```
result = Filter data by country=="India";
```