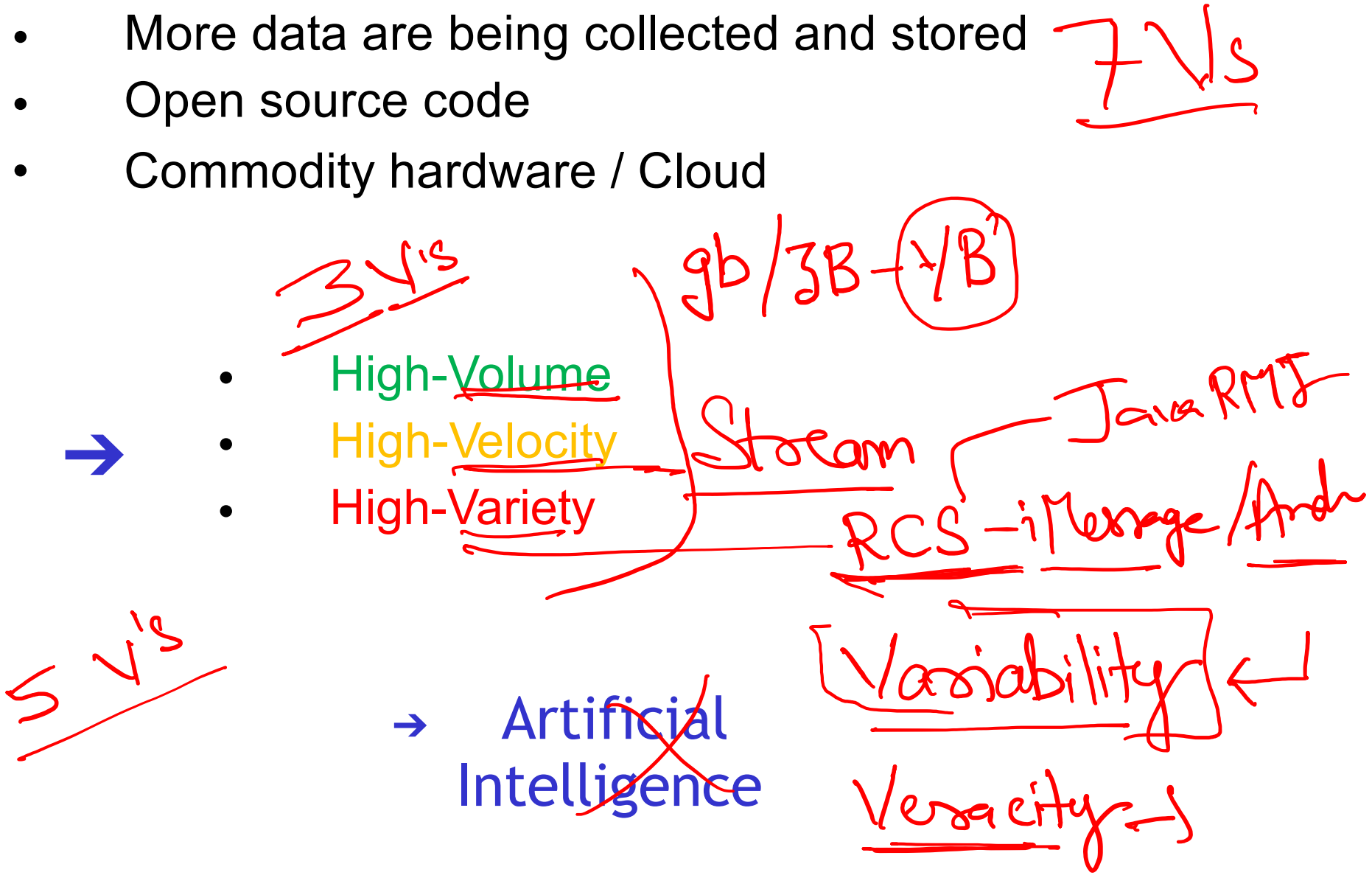
Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

→
- High-Volume
- High-Velocity
- High-Variety

→ Artificial Intelligence



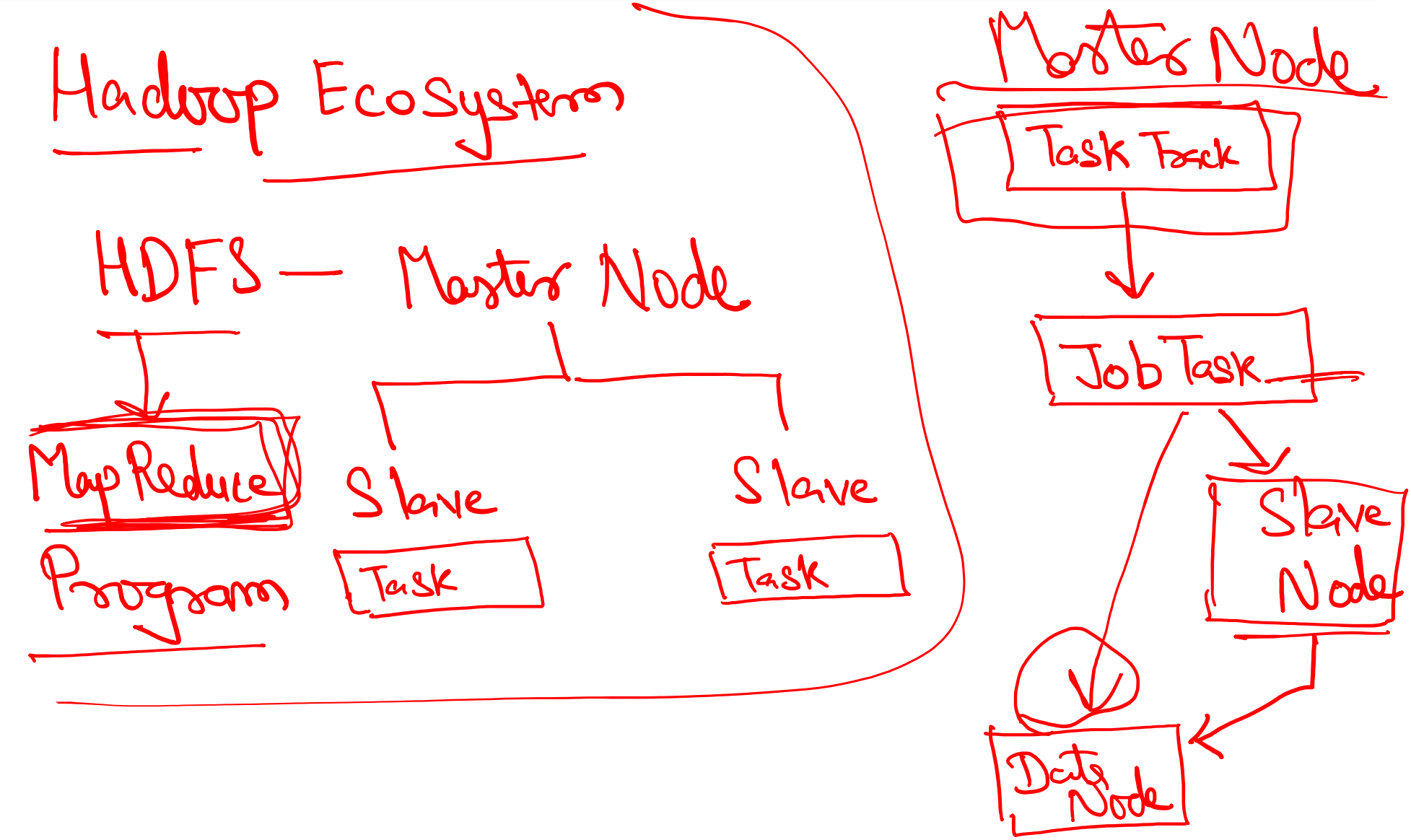
Volume
Velocity
Variety
} 3V's

So important

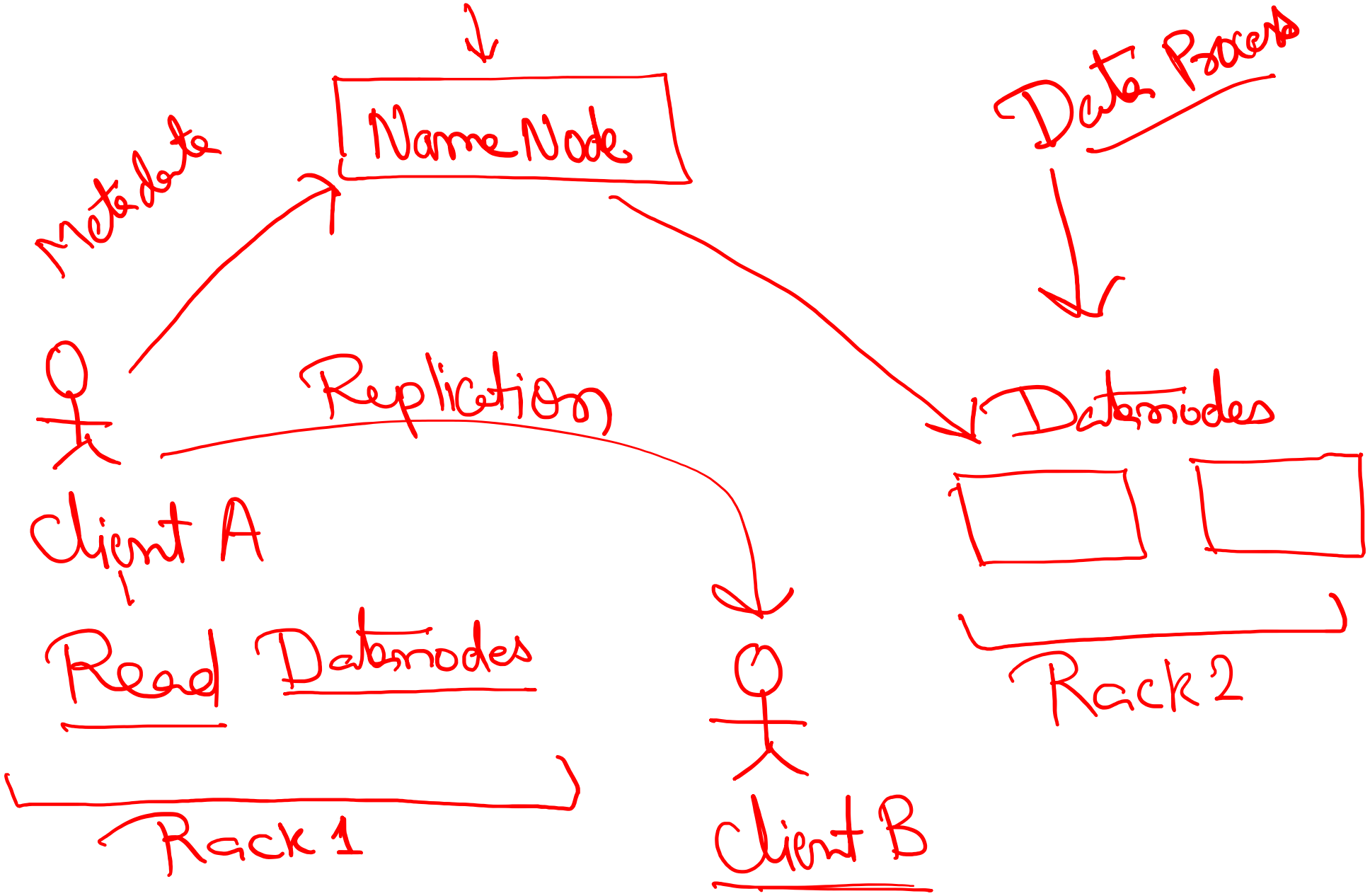
Variability
Veracity
} 5V's

Visualization
} 7V's

Value is the end game of everything.


Hadoop EcoSystem
HDFS – Master Node
Map Reduce
Program
Slave
Task
Slave
Task
Master Node
Task Track
Job Task
Slave
Node
Data
Node

HDFS Architecture



Course Main Thrust 1: Apache Hadoop and Big Data

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

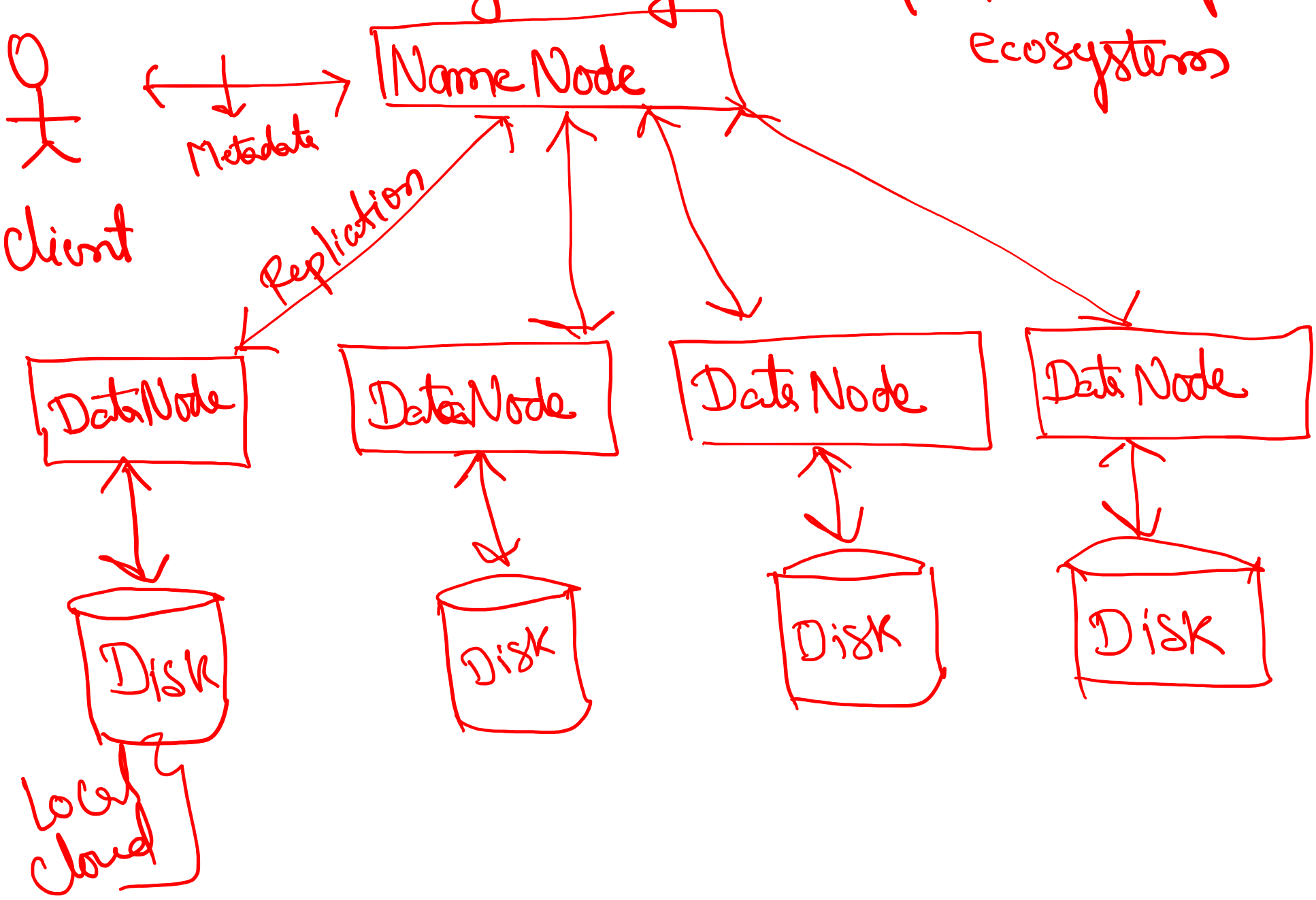
The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

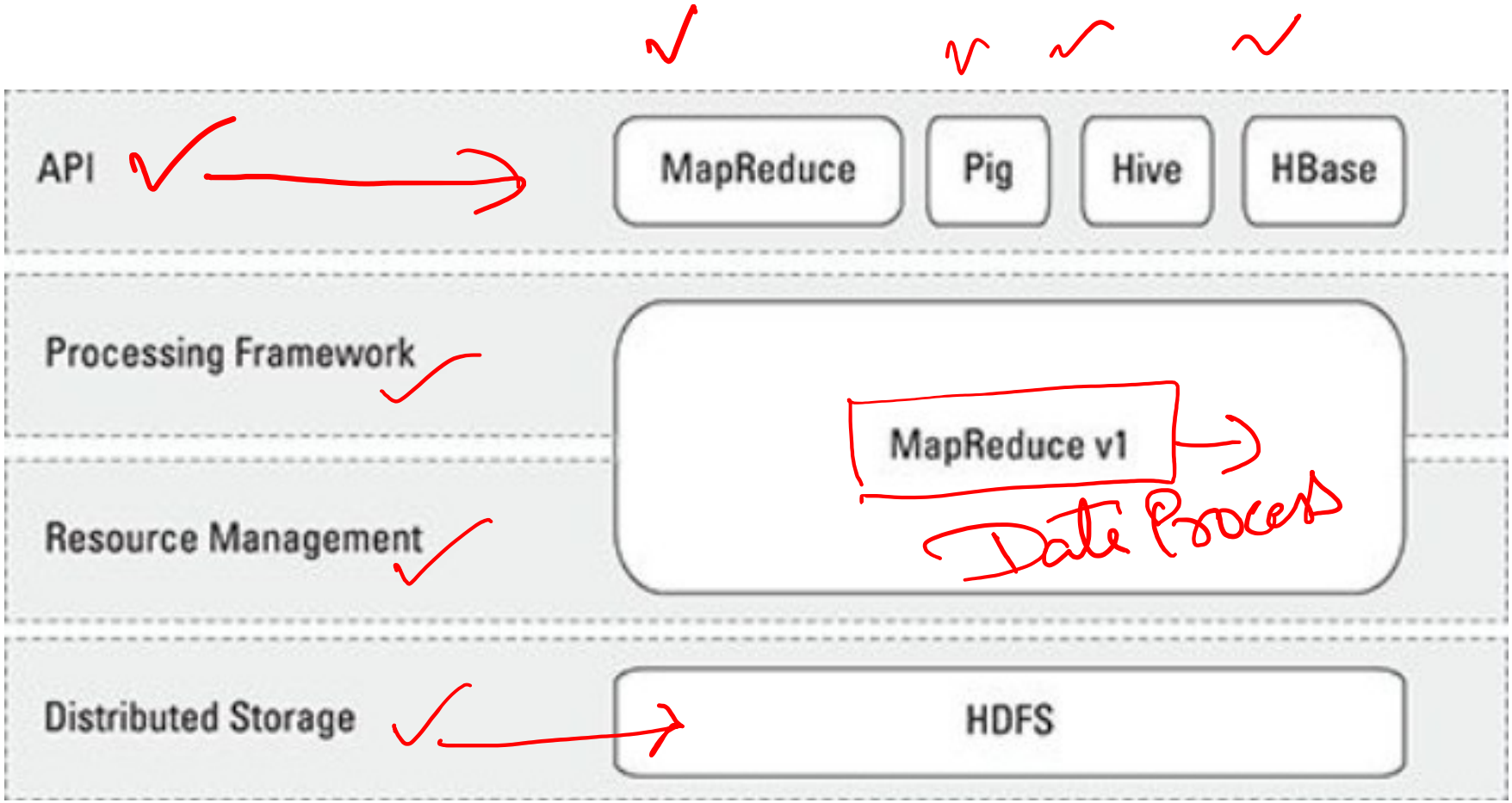
- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

http://hadoop.apache.org

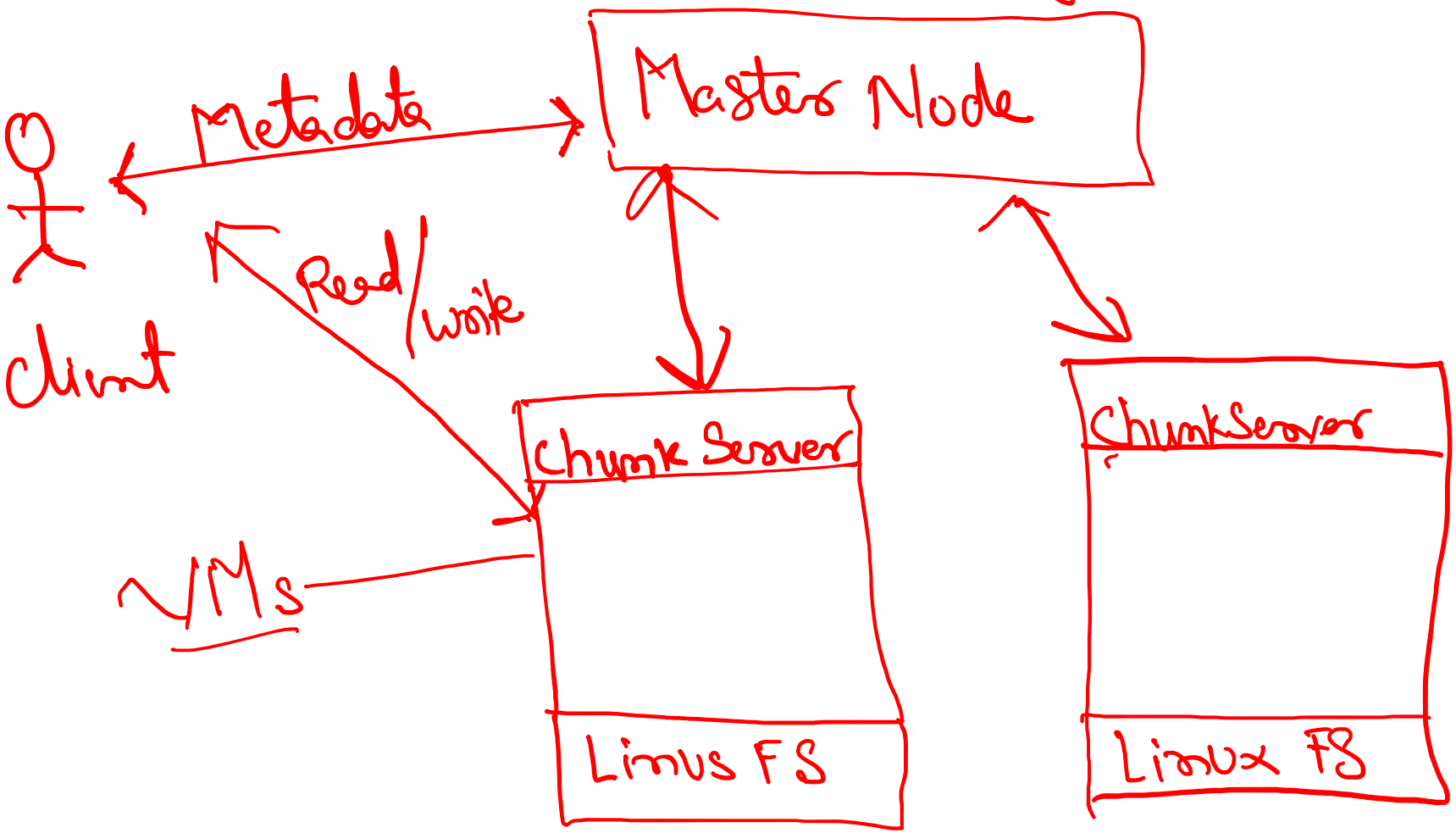
HDFS – Storage layer → Apache Hadoop ecosystem



Four distinctive layers of Hadoop



Google file System → GCP → google cloud Platform



Apache Hadoop Architecture Eco System Diagram.