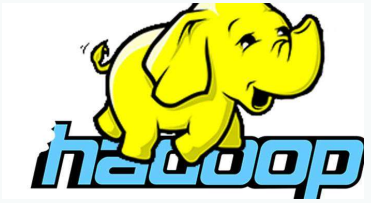
HDFS

What is Hadoop?



An open-source framework

Created to make it easier to work with big data.

It provides a method to

- Access data
- Process data
- manage resources across the computing and network resources

Hadoop Core Modules

Hadoop Distributed File System (HDFS)

- Provides access to application data.
- Can work with other file systems - FTP, Amazon S3 etc

Hadoop YARN

- Provides the framework to schedule jobs and manage resources

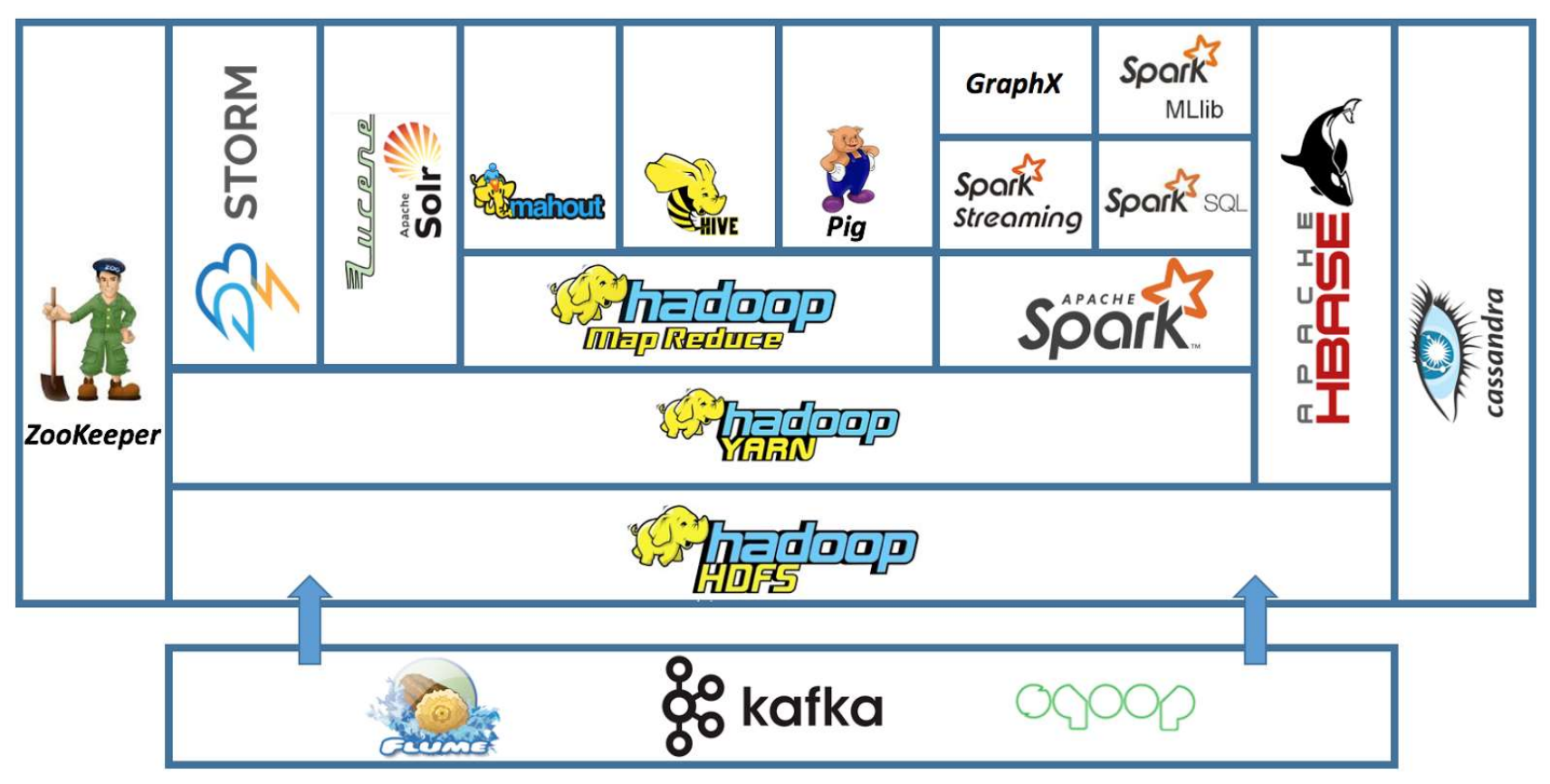
Hadoop MapReduce

- Parallel processing system for large data sets.

Hadoop Common

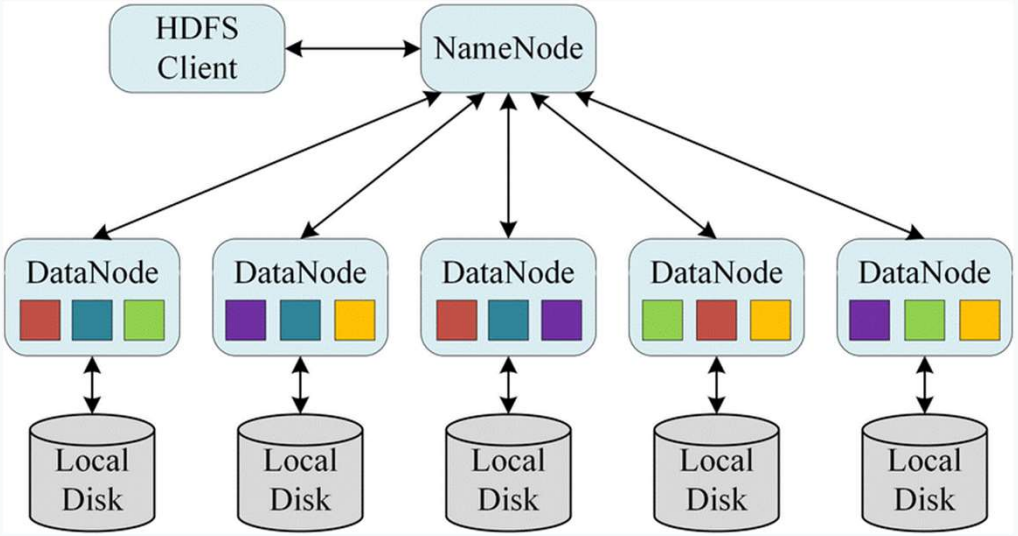
- A set of utilities that supports the three other core modules.

Hadoop Ecosystem

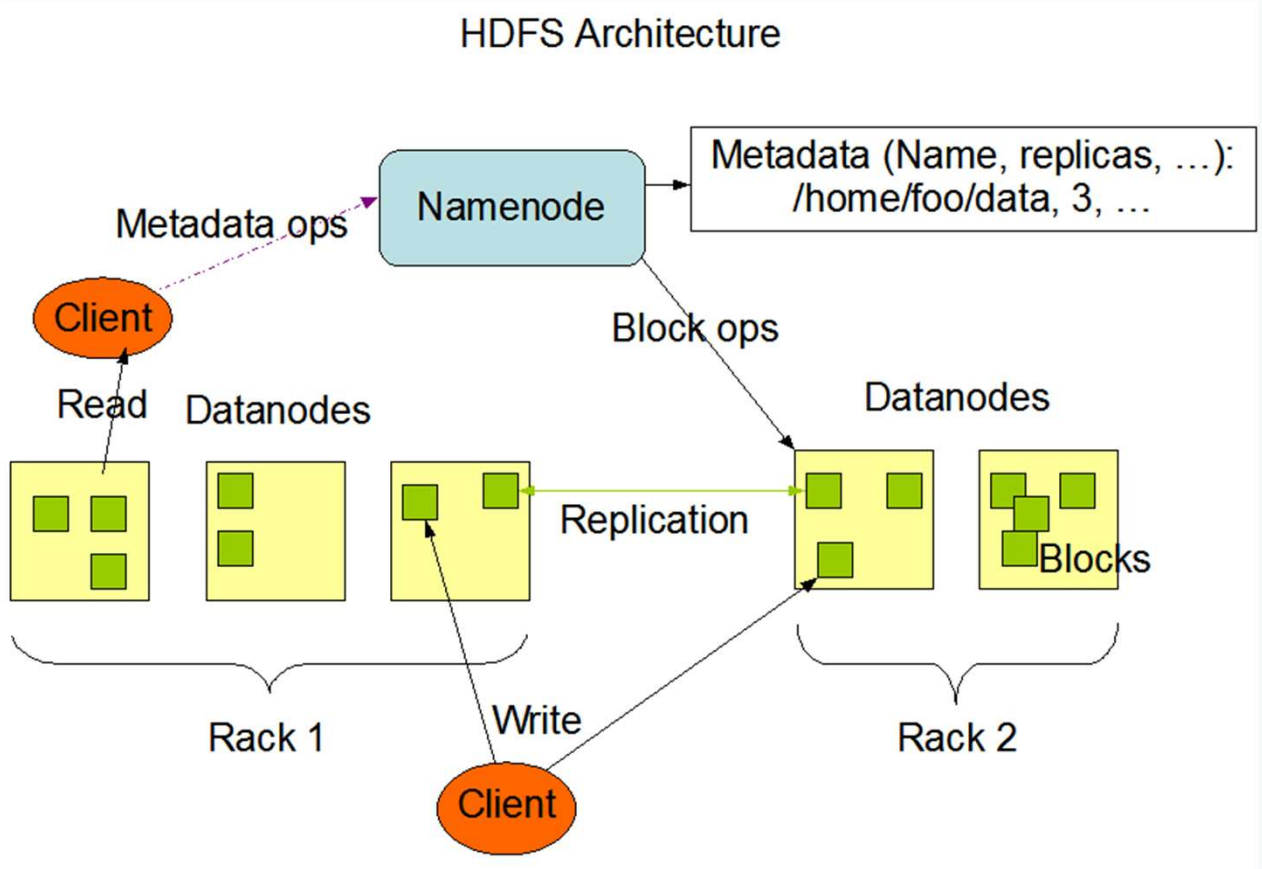


HDFS

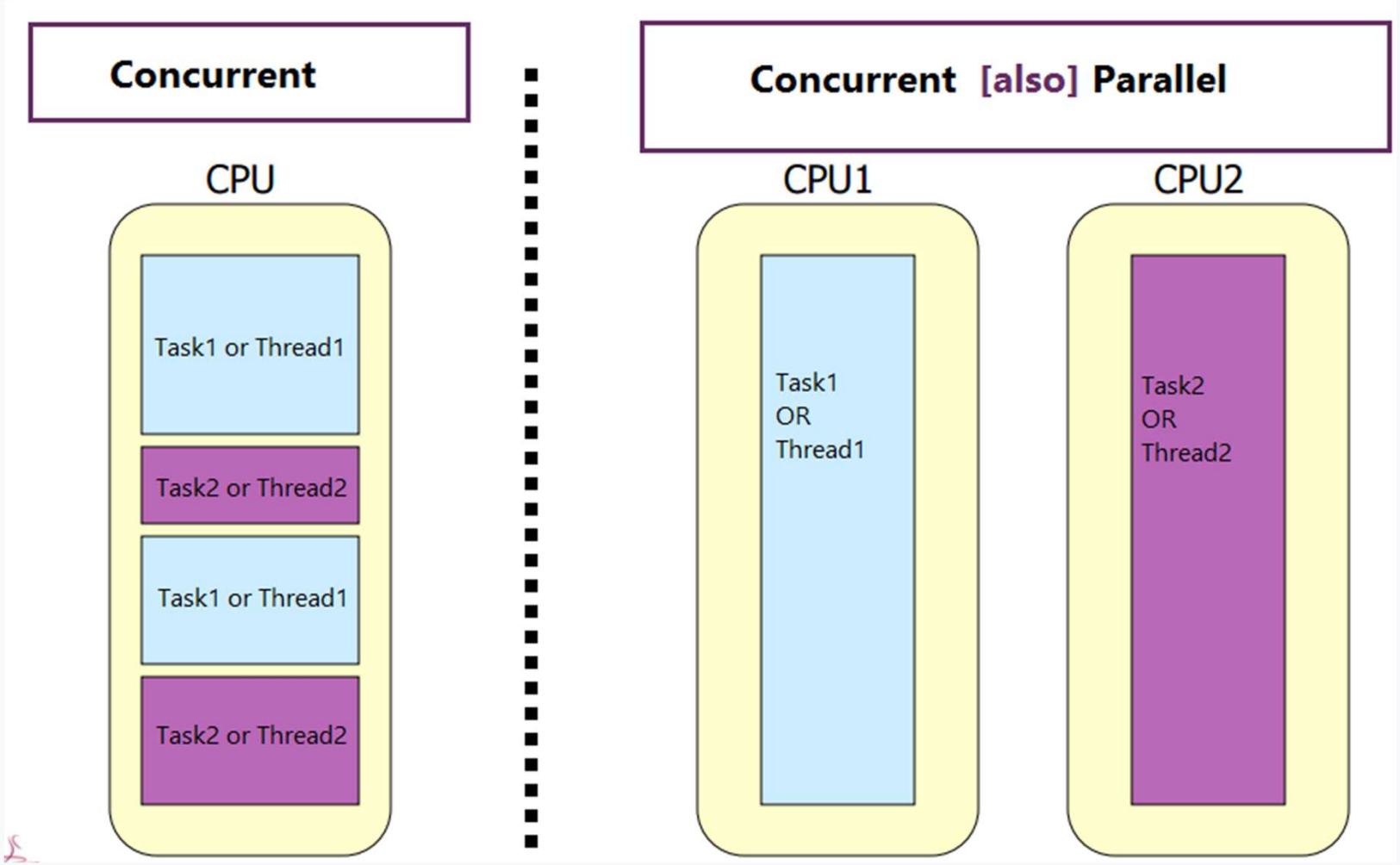
- Manages how data files are divided and stored across the cluster.
- Data is divided into blocks
- Each server in the cluster contains data from different blocks
- There is also some built-in redundancy.



HDFS Architecture

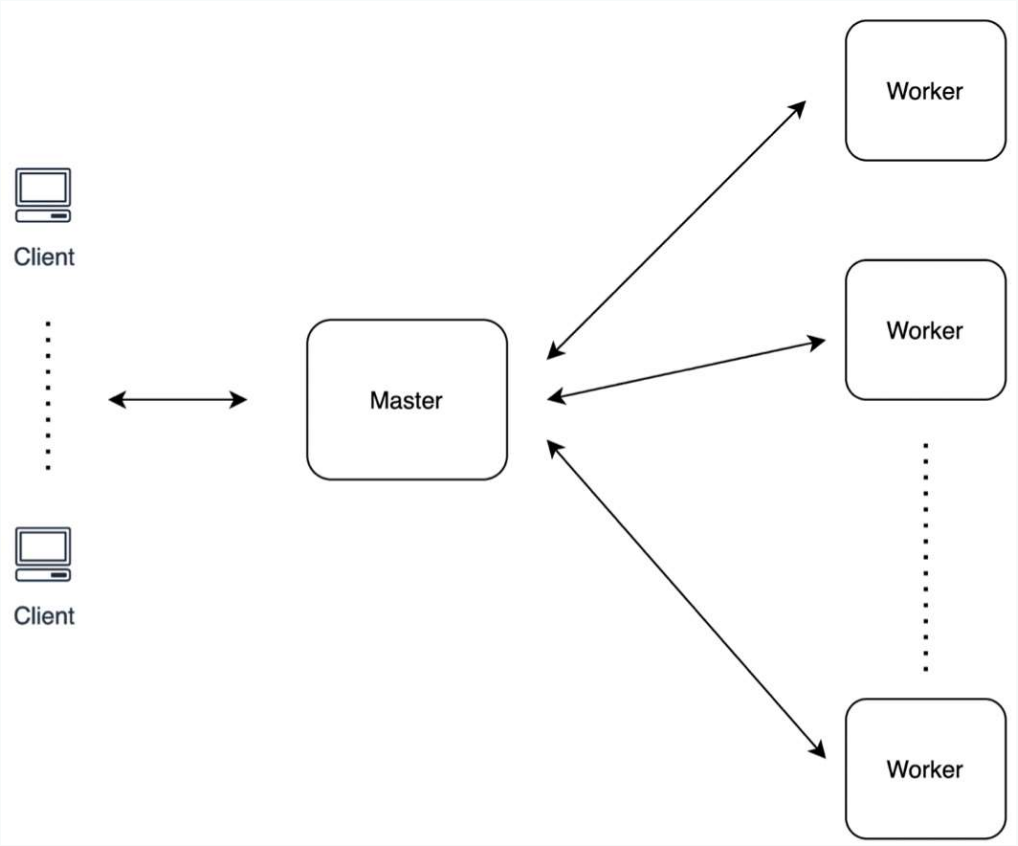


Concurrency and Parallalism

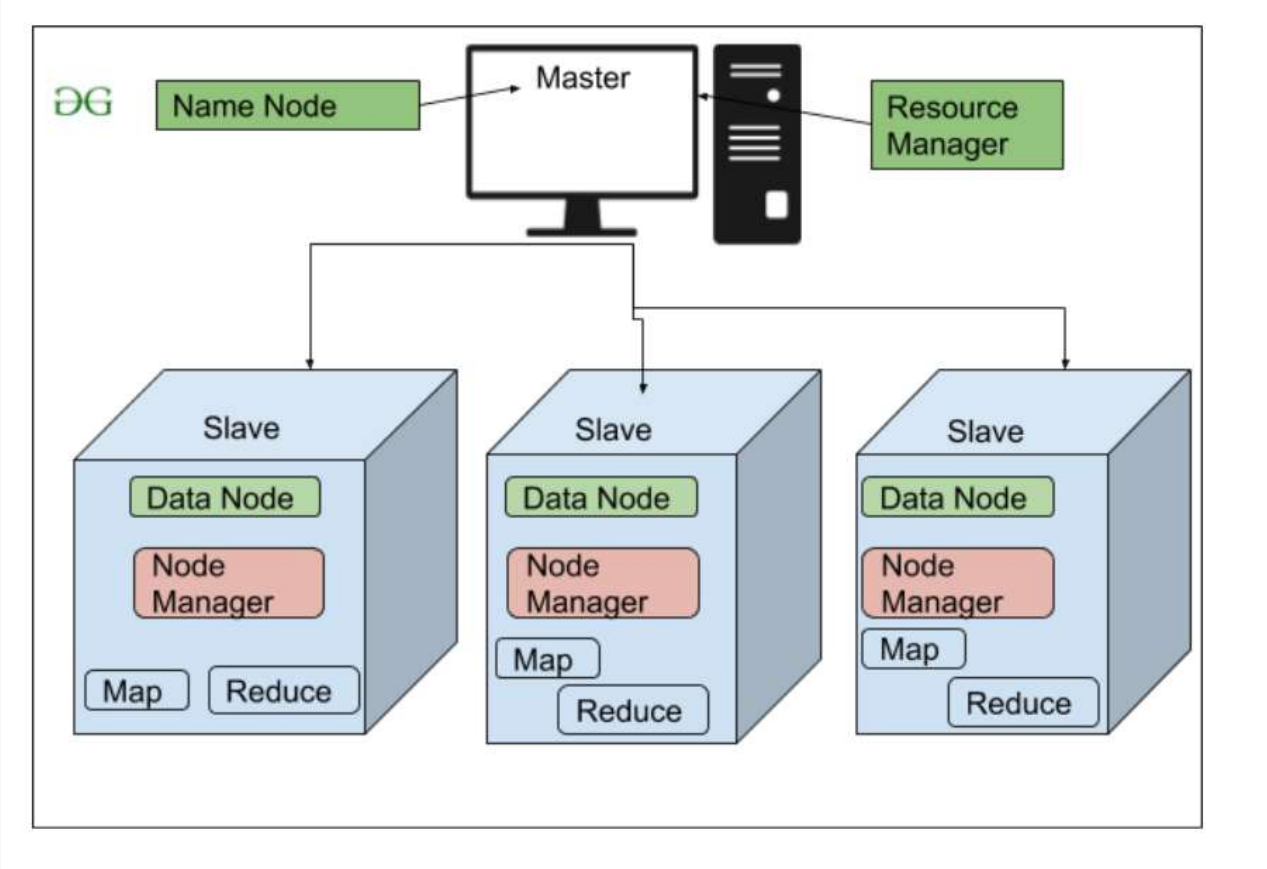


Distributed Processing

- Is the technique of linking together multiple computer servers over a network into a cluster



HDFS Architecture



Thank You