# HDFS

Hadoop Distributed File System

# HDFS (Hadoop Distributed File System)

- HDFS or Hadoop Distributed File System, which is completely written in Java programming language, is based on the Google File System (GFS). Google had only presented a white paper on this, without providing any particular implementation. It is interesting that around 90 percent of the GFS architecture has been implemented in HDFS.

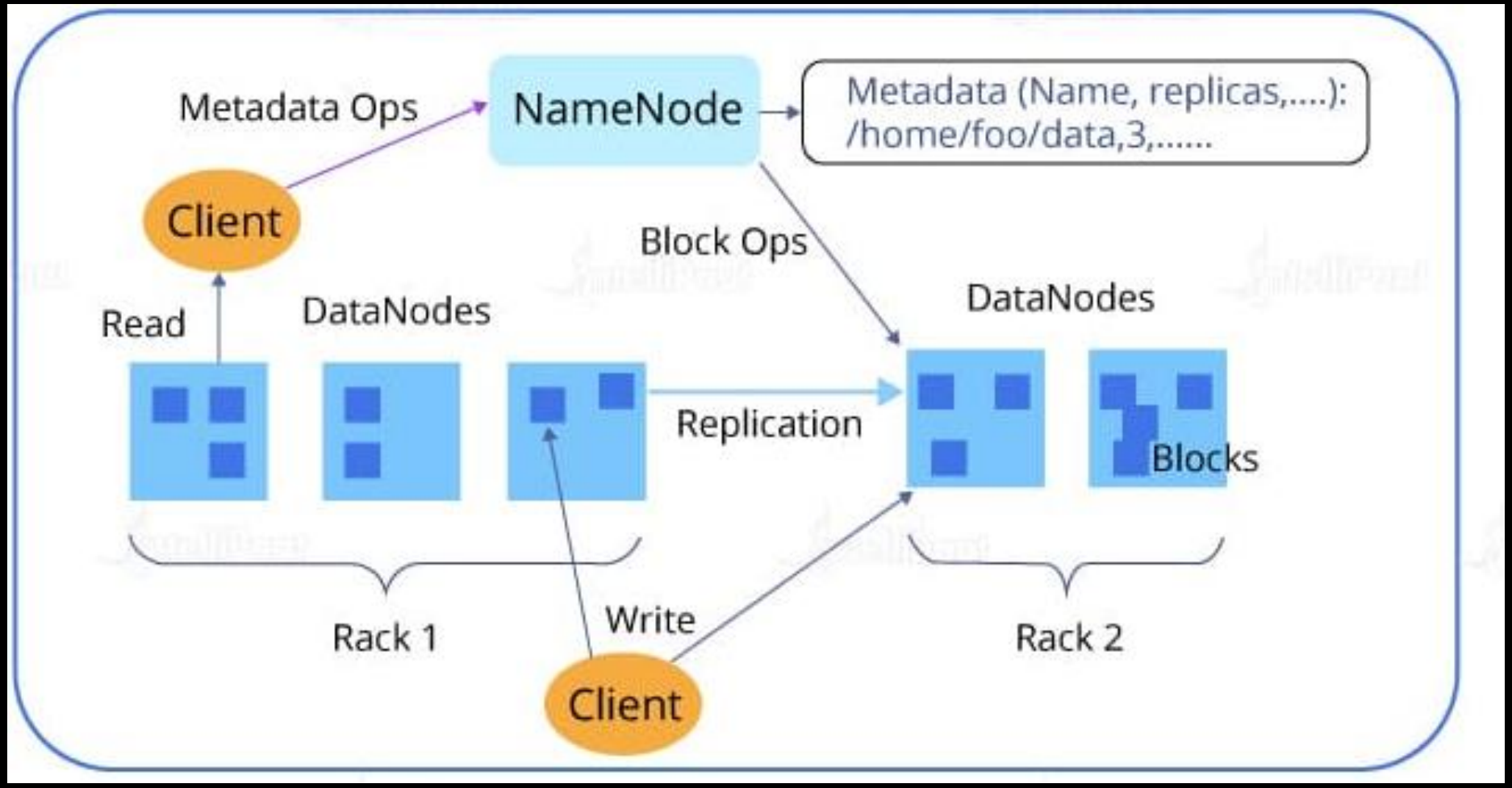
# Why does HDFS work very well with Big Data?

- HDFS is a perfect tool for working with big data. The following list of facts proves it.
- HDFS uses the MapReduce method for accessing data, which is very fast.
- HDFS follows the data coherency model, in which the data is synchronized across the server. It is very simple to implement and is highly robust and scalable.
- HDFS is compatible with any kind of commodity hardware and operating system processors
- As data is saved in multiple locations, it is safe enough.
- It is conveniently accessible to use a web browser which makes it highly utilitarian.

# HDFS Block

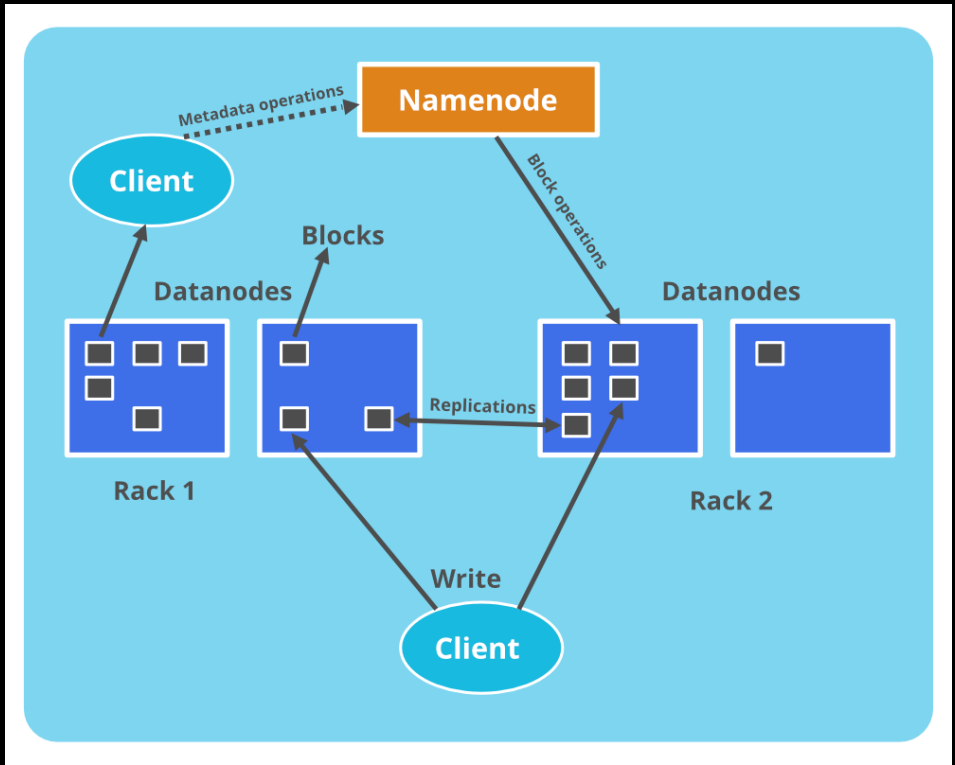
When you upload a file into HDFS, it will automatically be split into 128 MB fixed-size blocks (In the older versions of Hadoop, the file used to be divided into 64 MB fixed-size blocks). So basically, it takes care of placing the blocks in three different Data Nodes by replicating each block three times.

# HDFS Architecture



# Data Node

- Nodes wherein the blocks are physically stored are known as Data Nodes. Since these nodes hold the actual data of the cluster, they are termed as Data Nodes.



# Name Node

- A Name-Node keeps track of all the files or datasets in HDFS. It knows the list of blocks that are made up of files in HDFS, not only the list of blocks but also the location of them.