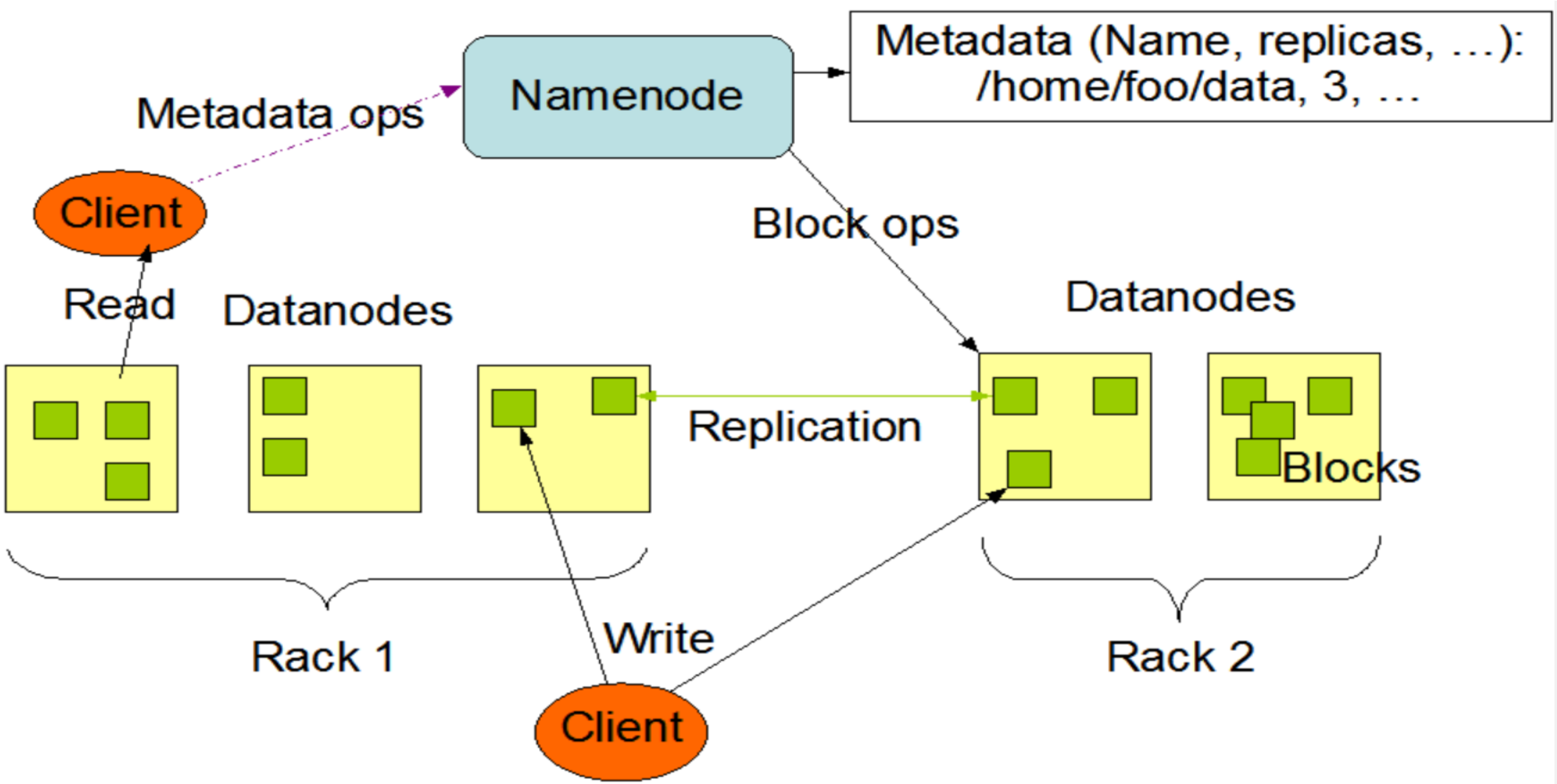
Hadoop Distributed File System (HDFS)

- Distributed file system designed to run on commodity hardware
- Highly fault-tolerant
- HDFS applications need a write-once-read-many access model for files
- A MapReduce application or a web crawler application fits perfectly with this model
- Portability across heterogeneous hardware and software platforms

Hadoop Distributed File System (HDFS)

- Hive, Impala and Spark
- SQL-on-Hadoop
- Commonly used by NoSQL databases
- Applications that run on HDFS have large data sets

HDFS Architecture



Block Replication

Namenode (Filename, numReplicas, block-ids, …)
/users/sameerp/data/part-0, r:2, {1,3}, …
/users/sameerp/data/part-1, r:3, {2,4,5}, …

Datanodes