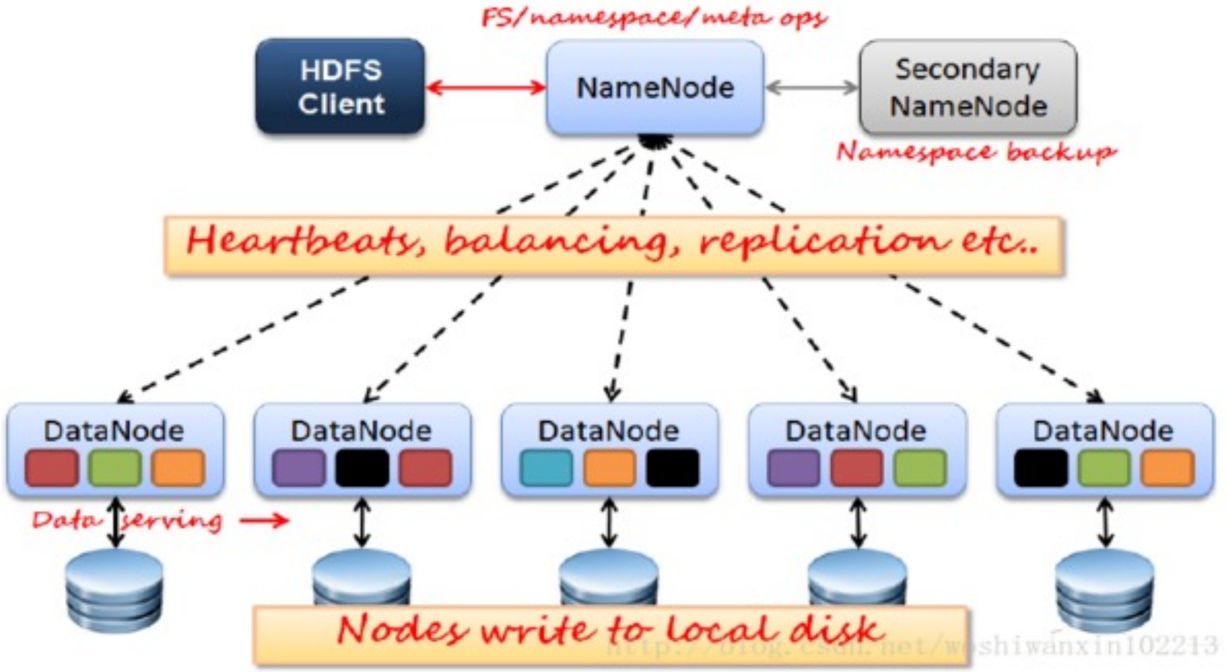
HDFS

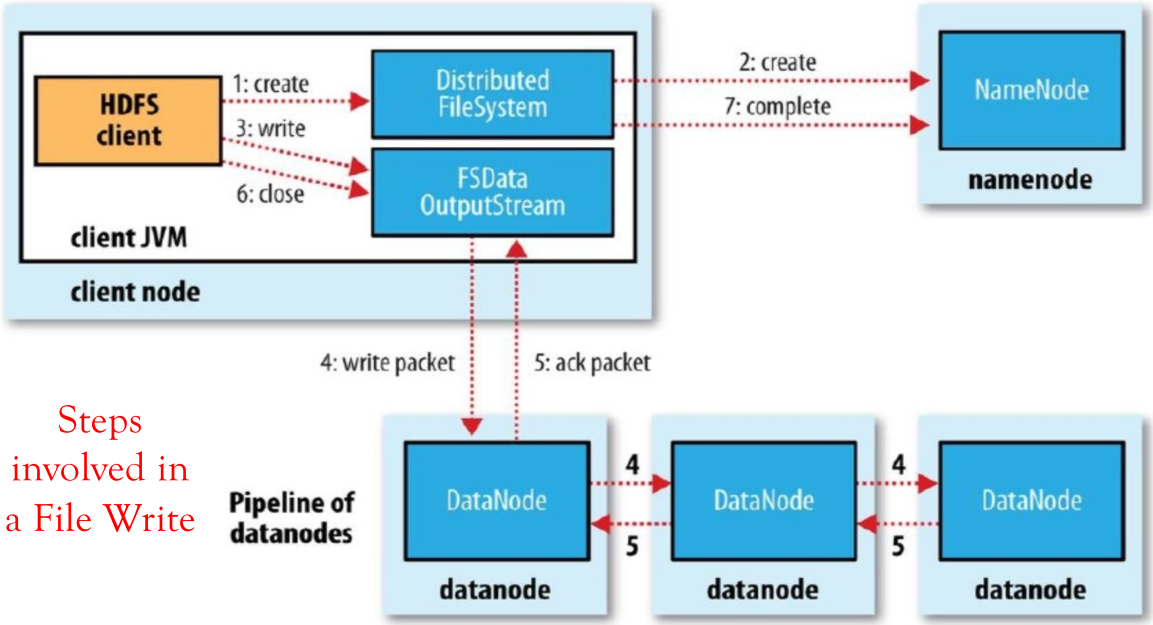
- All nodes in the cluster share the same namespace
- Write-once-read-many - Clients can only append to existing files
- Clients can find location of blocks and directly accesses data from the DataNodes
- Entire metadata is in the main memory of NameNode and is never swapped to disk
- Metadata comprises of filenames, blocks for each file, replication factor, DataNodes for each block, and File attributes
- DataNodes store the blocks in their local OS file system, can directly serve the data to the clients, and can send data to other DataNodes directly
- Rack awareness for fault tolerance



This Photo by Unknown Author is licensed under CC BY-SA

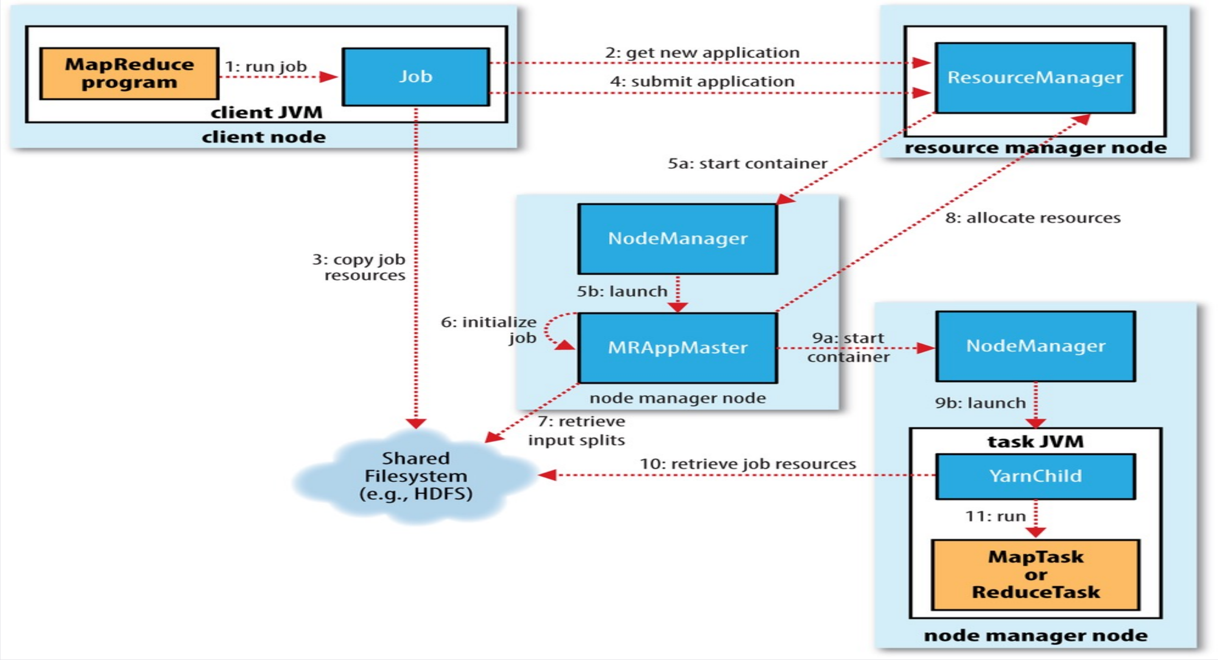
SOME HDFS COMMANDS

```
hdfs dfs -ls /path/to/directory
hdfs dfs -mkdir /path/to/new_directory
hdfs dfs -copyFromLocal /local/path/to/file /hdfs/path/to/destination
hdfs dfs -copyToLocal /hdfs/path/to/file /local/path/to/destination
hdfs dfs -cp /hdfs/src/path /hdfs/destination/path
hdfs dfs -chown <user>:<group> /hdfs/path/to/file_or_directory
```



Steps involved in a File Write

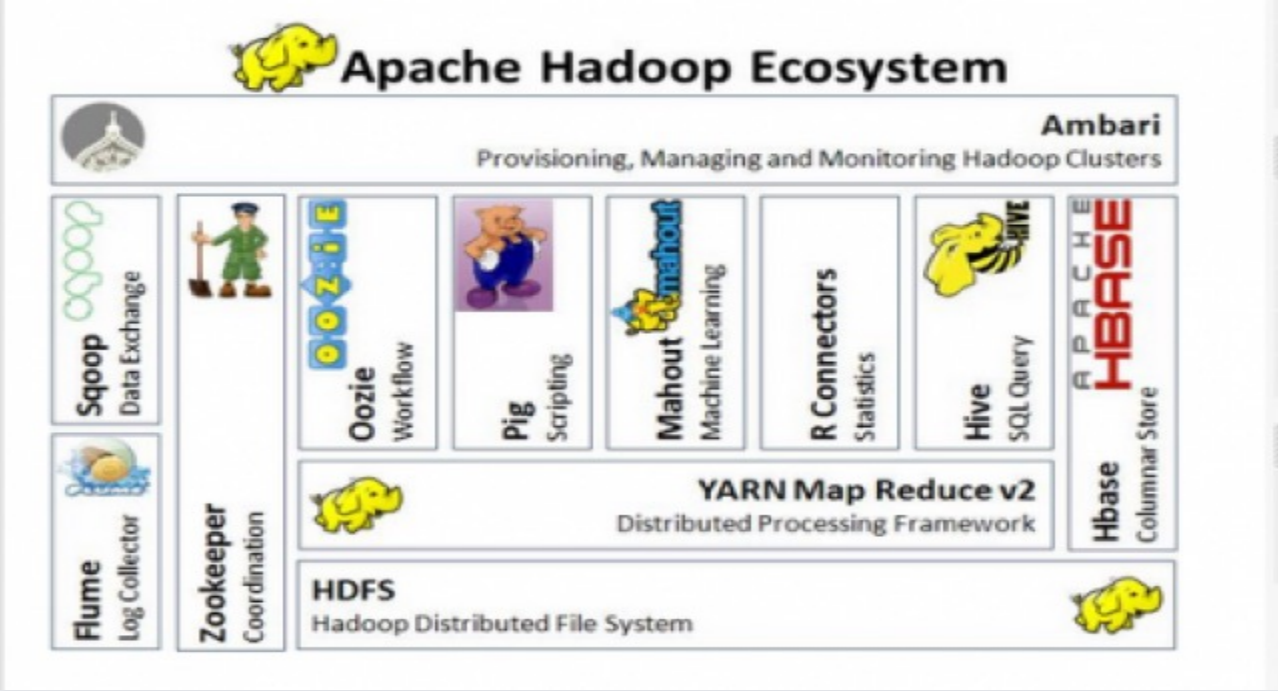
This Photo by Unknown Author is licensed under CC BY-SA



Running Hadoop

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar \
  -mapper <MapperScript>   -reducer <ReducerScript> \
  -input <InputDirectory>      -output <OutputDirectory>
```

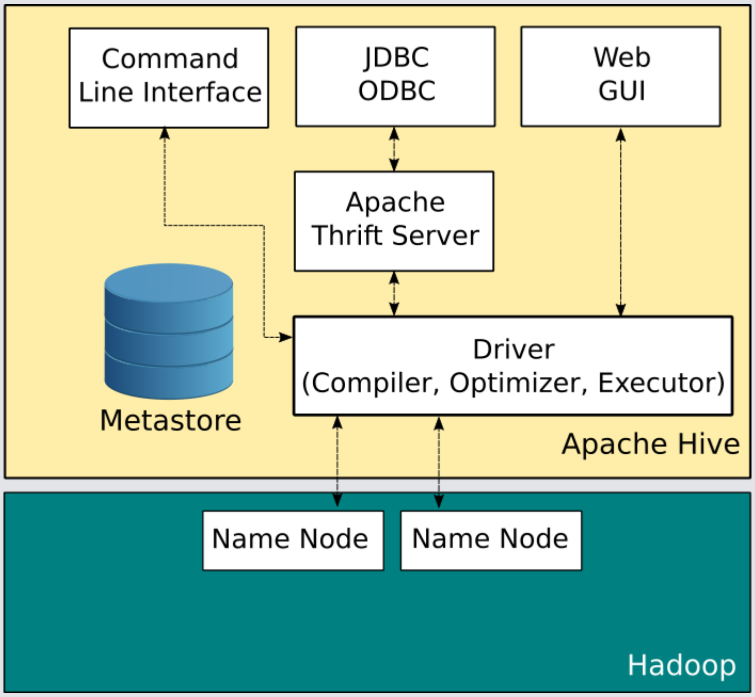
- Hadoop’s default mapper is the Identity Mapper



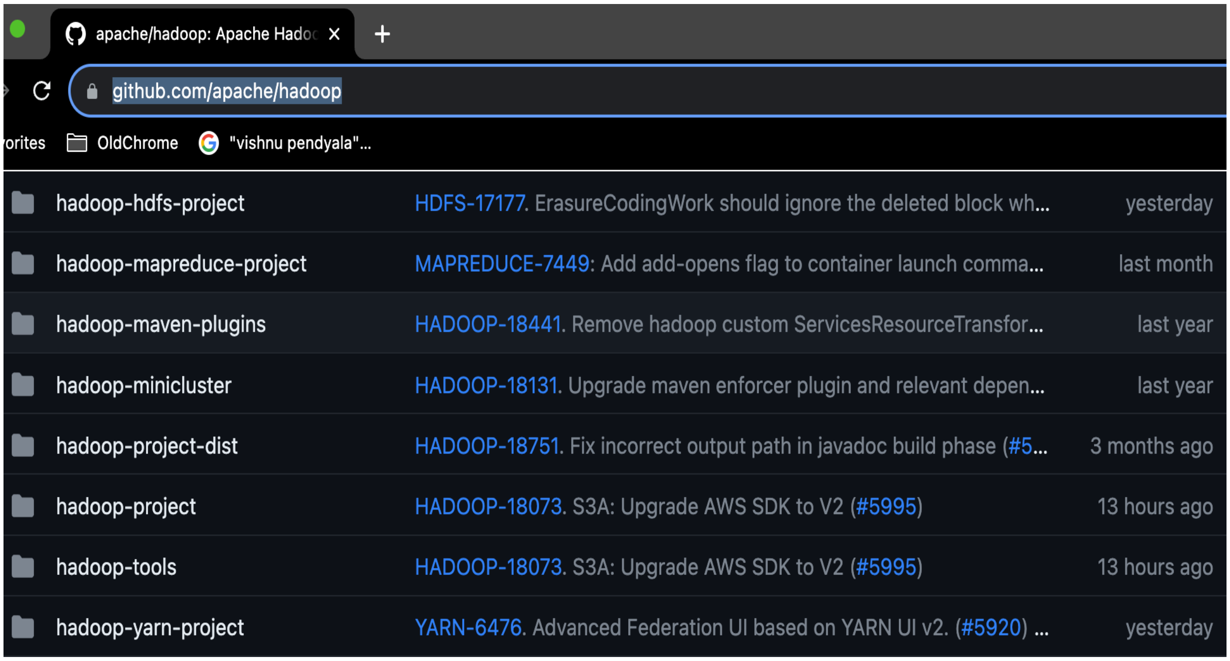
This Photo by Unknown Author is licensed under CC BY-SA

This Photo by Unknown Author is licensed under CC BY-SA

- SQL is increasingly the lingua franca of data
- Schema on read
- HiveQL similar to MySQL's SQL
- No transaction support
- Converts a query into one or more stages



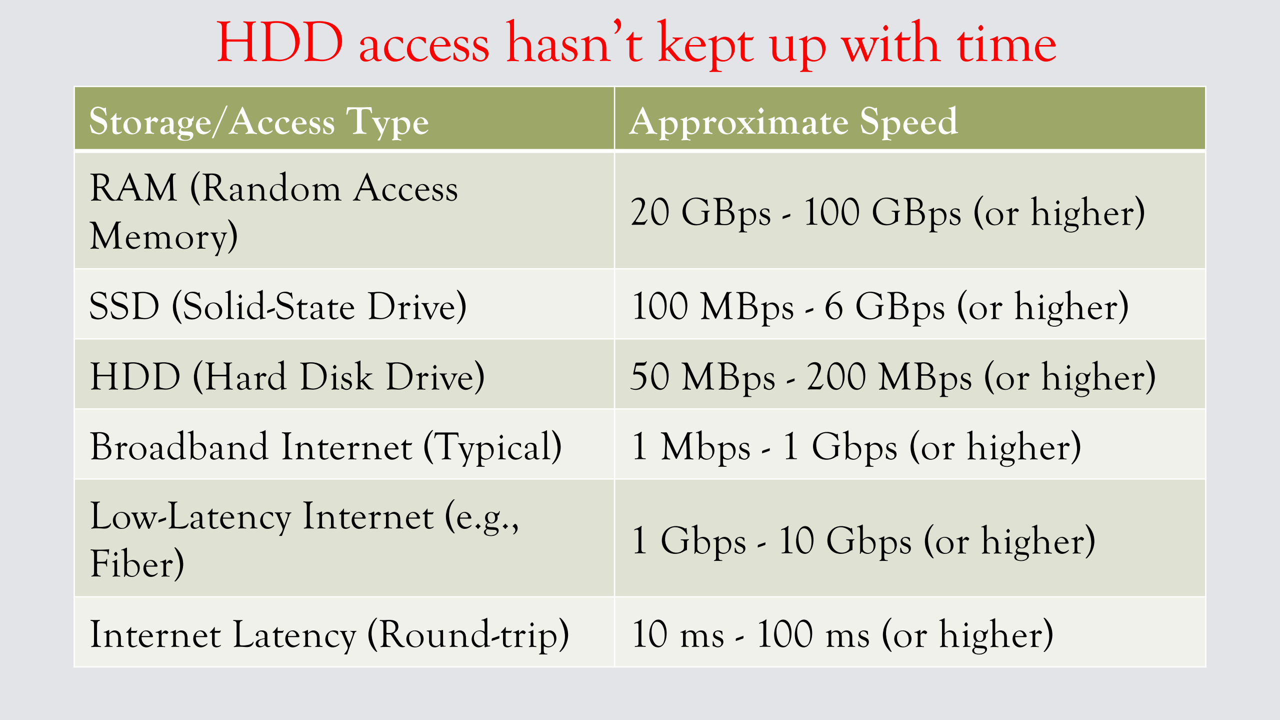
This Photo by Unknown Author is licensed under CC BY-SA



hadoop-hdfs-project	HDFS-17177. ErasureCodingWork should ignore the deleted block wh...	yesterday
hadoop-mapreduce-project	MAPREDUCE-7449: Add add-opens flag to container launch comma...	last month
hadoop-maven-plugins	HADOOP-18441. Remove hadoop custom ServicesResourceTransfor...	last year
hadoop-minicluster	HADOOP-18131. Upgrade maven enforcer plugin and relevant depen...	last year
hadoop-project-dist	HADOOP-18751. Fix incorrect output path in javadoc build phase (#5...	3 months ago
hadoop-project	HADOOP-18073. S3A: Upgrade AWS SDK to V2 (#5995)	13 hours ago
hadoop-tools	HADOOP-18073. S3A: Upgrade AWS SDK to V2 (#5995)	13 hours ago
hadoop-yarn-project	YARN-6476. Advanced Federation UI based on YARN UI v2. (#5920) ...	yesterday

What are the issues with Hadoop?

- Users must think in “map” and ”reduce” terms and provide functions that align with the paradigm
- Persisting intermediate results helps with fault tolerance, but HDD access between iterations slows it down substantially
- Waiting for the map and reduce phases can also slow it down



HDD access hasn't kept up with time

Storage/Access Type	Approximate Speed
RAM (Random Access Memory)	20 GBps - 100 GBps (or higher)
SSD (Solid-State Drive)	100 MBps - 6 GBps (or higher)
HDD (Hard Disk Drive)	50 MBps - 200 MBps (or higher)
Broadband Internet (Typical)	1 Mbps - 1 Gbps (or higher)
Low-Latency Internet (e.g., Fiber)	1 Gbps - 10 Gbps (or higher)
Internet Latency (Round-trip)	10 ms - 100 ms (or higher)

SUMMARY?

Sample Footer Text

9/14/24

19