(Big) Data Engineering In Depth

From Beginner to Professional

Moustafa Alaa

Senior Data Engineer at Onfido, London, UK

The Definitive Guide to Big Data Engineering Tasks

Previous video recap!

Hadoop Core Components

Hadoop Core Components

Hadoop Core Components

- HDFS.

Hadoop Core Components

- HDFS.
- Map-Reduce.

Hadoop Core Components

- – HDFS.
- – Map-Reduce.
- – YARN.

HDFS

- HDFS is responsible for storing the data on the Hadoop cluster.

HDFS

- HDFS is responsible for storing the data on the Hadoop cluster.

- Data is split into blocks with configurable block size, for example, 64MB, 128MB, and 512MB.

HDFS

- HDFS is responsible for storing the data on the Hadoop cluster.

- Data is split into blocks with configurable block size, for example, 64MB, 128MB, and 512MB.

- Each data block is replicated and distributed across the cluster data node. This replication is configurable, and by default, three replica (folds).

HDFS

- HDFS is responsible for storing the data on the Hadoop cluster.

- Data is split into blocks with configurable block size, for example, 64MB, 128MB, and 512MB.

- Each data block is replicated and distributed across the cluster data node. This replication is configurable, and by default, three replica (folds).

- Each block is stored in three different nodes. It is recommended to have two nodes in the same rack and the third one in a different rack.

HDFS

- HDFS is responsible for storing the data on the Hadoop cluster.

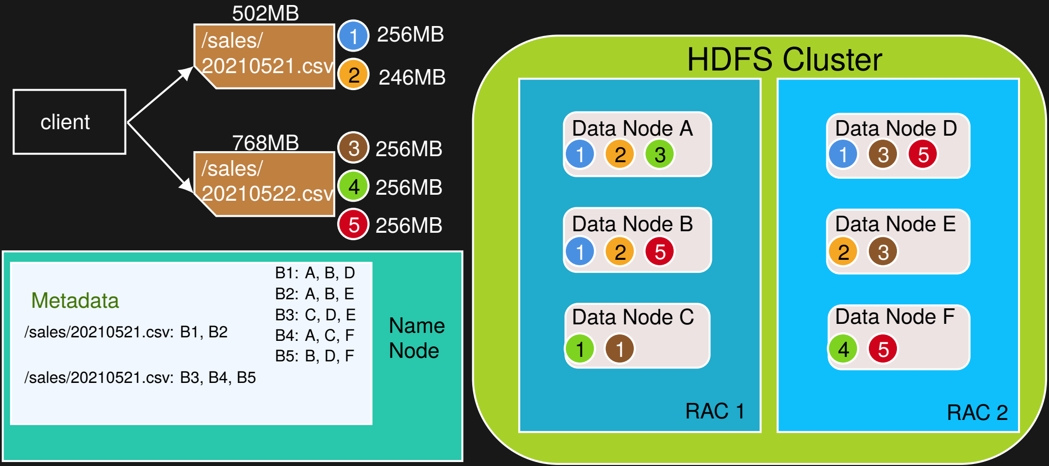
- Data is split into blocks with configurable block size, for example, 64MB, 128MB, and 512MB.

- Each data block is replicated and distributed across the cluster data node. This replication is configurable, and by default, three replica (folds).

- Each block is stored in three different nodes. It is recommended to have two nodes in the same rack and the third one in a different rack.

- A *NameNode* keeps track of the location of the blocks and which blocks make up these files. These details known as *metadata*.

HDFS



Metadata

/sales/20210521.csv: B1, B2

/sales/20210521.csv: B3, B4, B5

B1: A, B, D
B2: A, B, E
B3: C, D, E
B4: A, C, F
B5: B, D, F

Figure: HDFS

HDFS

- Hadoop cluster contains NameNodes and DataNodes.

[1] Small file problems https://blog.cloudera.com/the-small-files-problem/

HDFS

- Hadoop cluster contains NameNodes and DataNodes.

- NameNodes daemon must be running at all times. A daemon is simply a program running on a node.

[1] Small file problems https://blog.cloudera.com/the-small-files-problem/

HDFS

- Hadoop cluster contains NameNodes and DataNodes.
- NameNodes daemon must be running at all times. A daemon is simply a program running on a node.
- Hadoop cluster contains at least two NameNodes Active/Standby nodes.

[1] Small file problems https://blog.cloudera.com/the-small-files-problem/

HDFS

- Hadoop cluster contains NameNodes and DataNodes.
- NameNodes daemon must be running at all times. A daemon is simply a program running on a node.
- Hadoop cluster contains at least two NameNodes Active/Standby nodes.
- HDFS files are *write one*, so we can't do any random writes.

[1] Small file problems https://blog.cloudera.com/the-small-files-problem/

HDFS

- Hadoop cluster contains NameNodes and DataNodes.
- NameNodes daemon must be running at all times. A daemon is simply a program running on a node.
- Hadoop cluster contains at least two NameNodes Active/Standby nodes.
- HDFS files are *write one*, so we can't do any random writes.
- HDFS is optimized for large files. If we have many small files we could face a problem *Hadoop small files problem*

[1] Small file problems https://blog.cloudera.com/the-small-files-problem/

Access HDFS

- To acess HDFS we use Hadoop APIs.

[1] HDFS Commands https://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-common/FileSystemShell.html

Access HDFS

- To acess HDFS we use Hadoop APIs.
- These APIs provide various functionality over HDFS .

[1] HDFS Commands https://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-common/FileSystemShell.html

Access HDFS

- To acess HDFS we use Hadoop APIs.
- These APIs provide various functionality over HDFS .
- We can use the command line ”FsShell” or call the API through MapReduce, Spark, or Other Restful interfaces.

[1] HDFS Commands https://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-common/FileSystemShell.html

Thank you for watching!

See you in the next video ☺