Zewail City of Science and Technology

Communications and Information Engineering (CIE) Program

Big Data Analytics (CIE 427) - Fall 2020

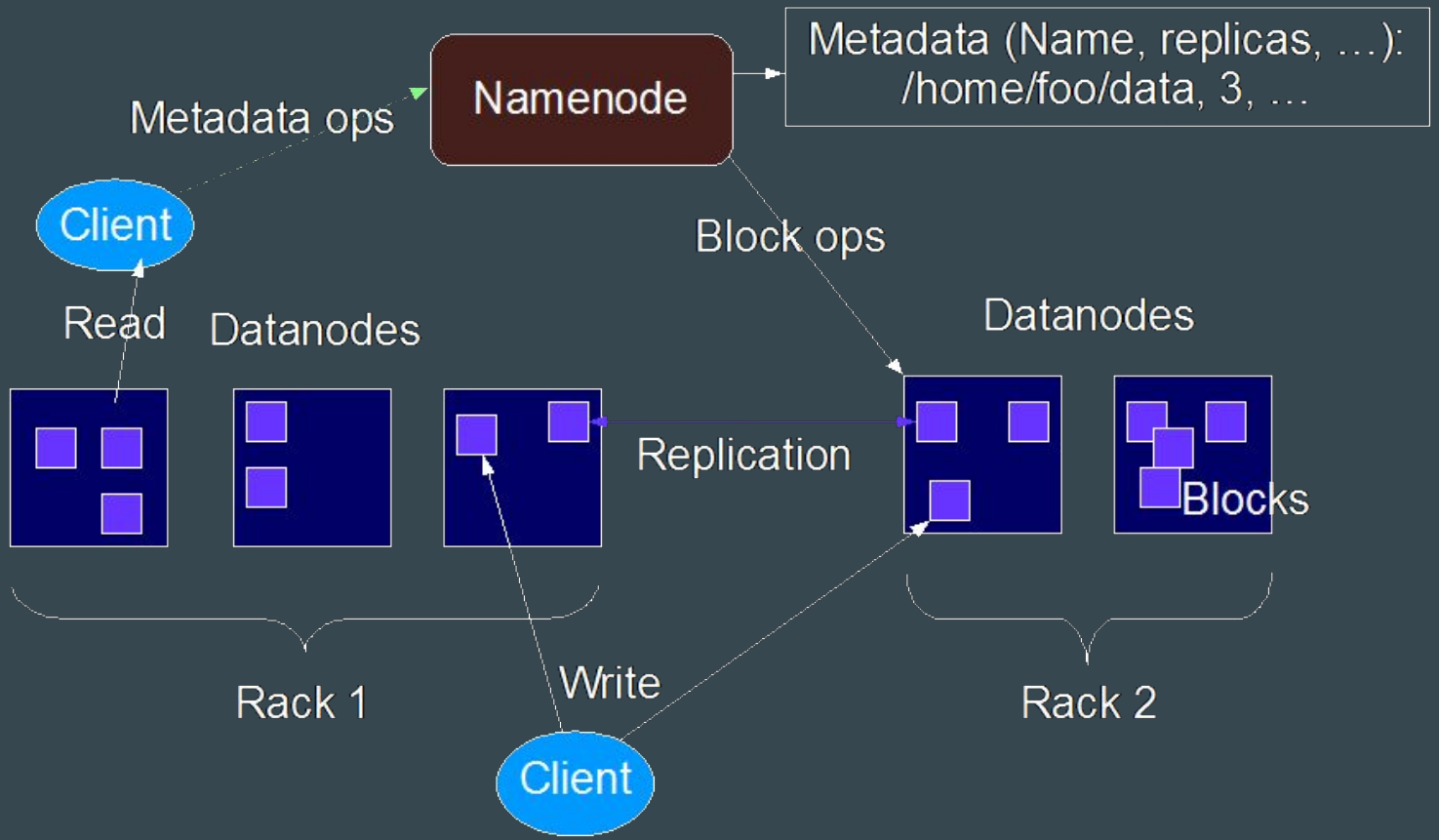
Hadoop Distributed File System

...

Original by: Muhammad Hamdy AlAref

Quick Talk About The Cluster

HDFS Architecture



Hadoop Pseudo-Distributed Operation

Prerequisites: Passphraseless `ssh`

Try `ssh` to the `localhost` without a passphrase!

```
$ ssh localhost
```

If that did not work, run these commands!

```
$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
$ chmod 0600 ~/.ssh/authorized_keys
```

Prerequisites: Replication

In the file `etc/hadoop/hdfs-site.xml`, set the replication factor,

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
</configuration>
```

Prerequisites: Hadoop Filesystem

In the file `etc/hadoop/core-site.xml`, set HDFS the default filesystem for Hadoop,

```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

Prerequisites: Format Filesystem

Before HDFS runs for the first time, it must be formatted!

```
$ bin/hdfs namenode -format
```

Run the Daemons

```
$ sbin/hadoop-daemon.sh --script hdfs start
namenode
$ sbin/hadoop-daemon.sh --script hdfs start
datanode
```

This should start local NameNode and DataNode daemons!
You should be able to connect to the NameNode's web interface at
http://localhost:50070/!

Setup Home

After setting HDFS as the default Hadoop filesystem, all of the relative `--input` and `--output` paths passed to MapReduce will be referenced to `hdfs:///user/$USER` instead of the local filesystem. So, let's setup these up!

```
$ bin/hdfs dfs -mkdir /user
$ bin/hdfs dfs -mkdir /user/$USER
```

Filesystem API

HDFS offers a UNIX-like filesystem API that is very similar to the Linux commands we already discussed! To view the available commands, just call the script with no commands!

```
$ bin/hdfs
$ bin/hdfs dfs
```

The web interface can be used to do most file operations as well!

Pseudo-Distributed Word Count Demo

Requirement: Pseudo-Distributed Pseudo-PageRank

Repeat the Pseudo-PageRank requirement form last time, but this time with HDFS!

- Upload the input to HDFS!
- Run the same MapReduce job but with `--input` and `--output` directories in HDFS!

Acknowledgement

The graph and the material introduced in HDFS architecture were adapted from the official Apache Hadoop documentation.

A lot of the material introduced in running Hadoop Pseudo-Distributed Operation were adapted from the official Apache Hadoop documentation.

Thank you!