Big Data Technologies

Agenda

- Hadoop Multi-node cluster
- Replication

Q & A

- Does NameNode needs higher configuration?
    - Hadoop can be installed on commodity hardware.
    - If any data node fails, its replica can be accessed from other data nodes.
    - However if name node fails, we cannot start secondary namenode immediately. This makes NameNode as SPOF.
    - Here NameNode is expected to be machine with good configuration.
        - NameNode should not fail (ideally).
        - All metadata is loaded in NameNode memory and served from there to workers/clients. It needs higher memory configuration.
- Hadoop ports
    - NameNode -- 9000 -- IPC port (worker/client communication)
    - ResourceManager -- 8032 -- Submit MR job
    - NameNode -- 9870 -- HDFS Web UI
- Can we modify files uploaded in HDFS?
    - NO. HDFS is "Write once, Read multiple times" file system.
    - In Hadoop 2.x, appending to existing file is allowed (-appendToFile).

Access Hadoop cluster from Client machine

- step 1: Install java, hadoop (.bashrc) on client.
- step 2: Use generic option -fs to deal with master node (e.g. 172.18.4.102) of Hadoop cluster.

```
hadoop fs -fs hdfs://172.18.4.102:9000 -ls /user/hduser
hadoop fs -fs hdfs://172.18.4.102:9000 -put file.txt /user/hduser
```

Generic options

- terminal> hadoop fs -help
  - -fs --> specify namenode to communicate with (hdfs).
  - -jt --> specify resource manager to communicate with (map reduce).
  - -D --> specify configuration e.g. dfs.replication, dfs.blocksize, ...
- Syntax: hadoop fs 'generic options' -command args

getmerge command

- Merge output of the files given by a pattern and download them on client machine.
- terminal> hadoop fs -usage getmerge
- terminal> hadoop fs -getmerge /user/osboxes/file*.txt myfile.txt
- terminal> cat myfile.txt

Hadoop Java API documentation

- $HADOOP_HOME/share/doc/hadoop/api/index.html