Big Data Technologies

Agenda

- Hadoop
  - HDFS APIs
  - HDFS Read/Write Internals
  - HDFS NameNode & SecondaryNameNode working
  - HDFS Standby NameNode
  - HDFS NameNode Federation

Maven troubleshooting

- method 1:
  - Project -> Maven -> Update project -- Force update (check mark) -- Ok.
- method 2:
  - Close eclipse.
  - In your home directory (e.g. /home/sunbeam), delete ".m2" directory.
  - Ensure that you are connected to stable internet.
  - Reopen eclipse and build the project.
- Maven Tutorial: https://youtu.be/lMXBrlVFYA0
- Maven Tutorial: https://jenkov.com/tutorials/maven/maven-tutorial.html

To run Hadoop application jar from command line

- step 1: Create Runnable Jar from eclipse (Project -> Export -> Java - Runnable Jar).
- step 2: Run with "hadoop jar"
  - terminal> hadoop jar app.jar
  - Internally this command adds all hadoop classes/jars into the java CLASSPATH and then execute the your jar's main class.

Hadoop Docs (all)

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

Java File System APIs

- java.io.File class represents File or Directory.
  - Get attributes/information about file/directory.
  - Get directory contents (file listing).
- WAP to get a path from user. If it is path of file, then display its metadata. If it is path of directory, then display its contents.

HDFS APIs

- DistributedFileSystem -- represents HDFS.
- FsDataInputStream -- represents a file (on hdfs) for reading.
- FsDataOutputStream -- represents a file (on hdfs) for writing.
- Configuration -- represent hadoop config/settings.
- Path -- represent a path on hdfs.
- FileStatus -- represent file on hdfs -- get metadata/directory listing.

HDFS -- File append

- Hadoop is "Write-Once Read-Multiple Times" File System.
- New files can be uploaded into HDFS, but existing files data cannot be edited.
- Hadoop 2.x added feature of appending the files.

```
vim hi.txt
# add some contents into the file
hadoop fs -appendToFile hi.txt /user/nilesh/welcome.txt
# needs heavy processing (MR) -- not recommended -- will raise error if not configured
hadoop fs -head /user/hduser/hello.txt
```

HDFS commands

- User commands
  - hadoop fs -command ...
- Admin commands
  - hdfs namenode -command ...
  - hdfs dfsadmin -command ...
- File metadata can be modified (upto some extent) using various HDFS commands.
  - hadoop fs -setrep ...
    - To change replication factor.
  - hadoop fs -touch ...
    - To change timestamp.
  - hadoop fs -chmod ...
    - To change the permissions.
  - hadoop fs -chown ...
    - To change the owner.
  - hadoop fs -mv ...
    - To change file location (move to other dir in hdfs) or to rename.
- Most of HDFS commands are similar to corresponding Linux commands. Refer help.

Hadoop cluster and client

- Typical Hadoop cluster includes one master node, one backup master node and multiple worker nodes.
- This cluster can be on-premise (computers in local network) or on cloud (AWS, GCP, Azure, etc).
- The client machine is any other computer that has access to hadoop cluster.

Safe mode

- While starting HDFS, all namenode data is loaded from its disk into RAM. It also verifies integrity of the data.
- This process takes significant amount of time (depending on size of the data). At this time hadoop is said to be in safe mode.
- Any operations done on HDFS during this period will fail. Once all metadata is loaded and verified, safemode is automatically OFF.