databrickstest_hadoop

Lab 3 - Hadoop on Databricks

Student's name: Khanh Nguyen

Student G#: 00536860

Class section #: AIT 614 – 002

References:

Dr. Liao's tutorial: **Liao_Hadoop_Databricks**

Installing Hadoop

NOTE: No need to install Java for Hadoop on Databricks since it has already been installed when the new cluster is created.

```
# Download the hadoop from Apache website if the file does not exist.
!wget -N https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
```

```
--2022-03-09 19:15:40--  https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.95.219, 2a01:4f8:10a:201a::2,
...
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 500749234 (478M) [application/x-gzip]
Saving to: ‘hadoop-3.3.0.tar.gz’


hadoop-3.3.0.tar.gz   0%[                    ]       0  --.-KB/s
```

```
hadoop-3.3.0.tar.gz   0%[                    ]  56.00K   252KB/s
hadoop-3.3.0.tar.gz   0%[                    ] 168.00K   378KB/s
hadoop-3.3.0.tar.gz   0%[                    ] 328.00K   493KB/s
hadoop-3.3.0.tar.gz   0%[                    ]   1.01M  1.13MB/s
hadoop-3.3.0.tar.gz   0%[                    ]   3.41M  3.13MB/s
hadoop-3.3.0.tar.gz   1%[                    ]   6.38M  4.86MB/s
hadoop-3.3.0.tar.gz   1%[                    ]   9.37M  6.11MB/s
hadoop-3.3.0.tar.gz   2%[                    ]  12.34M  7.02MB/s
hadoop-3.3.0.tar.gz   3%[                    ]  15.32M  7.75MB/s
hadoop-3.3.0.tar.gz   3%[                    ]  18.30M  8.32MB/s
```

```
# Unzip the hadoop tar file
!tar -xf hadoop-3.3.0.tar.gz
```

```
# Copy the hadoop to the /usr/local/ if updated.
!cp -ru hadoop-3.3.0/ /usr/local/
```

```
# Find the default Java path
!readlink -f /usr/bin/java | sed "s:bin/java::"
```

```
/usr/lib/jvm/zulu8-ca-amd64/jre/
```

Running Hadoop & Operating HDFS Use hadoop fs or hdfs dfs command when interacting with HDFS. Note:

1. hdfs dfs works for all the operations related to HDFS only.
2. hadoop fs is used for handling different file systems, such as Local file systems, (S)FTP, S3, and others.

```
# Run Hadoop
!/usr/local/hadoop-3.3.0/bin/hadoop
```

```
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
 or    hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
  where CLASSNAME is a user-provided Java class

  OPTIONS is none or any of:

tput: unknown terminal "unknown"
--config dir                     Hadoop config directory
--debug                          turn on shell script debug mode
--help                           usage information
buildpaths                       attempt to add class files from build tree
hostnames list[,of,host,names]   hosts to use in slave mode
hosts filename                   list of hosts to use in slave mode
loglevel level                   set the log4j level for this command
workers                          turn on worker mode

  SUBCOMMAND is one of:


    Admin Commands:
```

```
# Hadoop Help commands
!/usr/local/hadoop-3.3.0/bin/hadoop fs -help put get cat rm
```

```
-put [-f] [-p] [-l] [-d] <localsrc> ... <dst> :
  Copy files from the local file system into fs. Copying fails if the file already
  exists, unless the -f flag is given.
  Flags:

  -p  Preserves access and modification times, ownership and the mode.
  -f  Overwrites the destination if it already exists.
  -l  Allow DataNode to lazily persist the file to disk. Forces
         replication factor of 1. This flag will result in reduced
```

```
        durability. Use with care.

  -d  Skip creation of temporary file(<dst>._COPYING_).
-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst> :
  Copy files that match the file pattern <src> to the local name.  <src> is kept.
  When copying multiple files, the destination must be a directory. Passing -f
  overwrites the destination if it already exists and -p preserves access and
  modification times, ownership and the mode.
-cat [-ignoreCrc] <src> ... :
  Fetch all files that match the file pattern <src> and display their content on
  stdout.
```

```
# HDFS DFS commang - help
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -help put get

-put [-f] [-p] [-l] [-d] <localsrc> ... <dst> :
  Copy files from the local file system into fs. Copying fails if the file already
  exists, unless the -f flag is given.
  Flags:

  -p  Preserves access and modification times, ownership and the mode.
  -f  Overwrites the destination if it already exists.
  -l  Allow DataNode to lazily persist the file to disk. Forces
        replication factor of 1. This flag will result in reduced
        durability. Use with care.

  -d  Skip creation of temporary file(<dst>._COPYING_).
-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst> :
  Copy files that match the file pattern <src> to the local name.  <src> is kept.
  When copying multiple files, the destination must be a directory. Passing -f
  overwrites the destination if it already exists and -p preserves access and
  modification times, ownership and the mode.
```

```
# Display the Hadoop version
!/usr/local/hadoop-3.3.0/bin/hadoop version

Hadoop 3.3.0
Source code repository https://gitbox.apache.org/repos/asf/hadoop.git -r aa96f1871bfd858f9bac59cf2a81ec470da6
49af
Compiled by brahma on 2020-07-06T18:44Z
Compiled with protoc 3.7.1
From source with checksum 5dc29b802d6ccd77b262ef9d04d19c4
This command was run using /usr/local/hadoop-3.3.0/share/hadoop/common/hadoop-common-3.3.0.jar


# Check the HDFS storage
!/usr/local/hadoop-3.3.0/bin/hadoop fs -df

Filesystem          Size         Used     Available  Use%
file:///    157459890176  10113486848  147346403328    6%


# List the files using hadoop fs
!/usr/local/hadoop-3.3.0/bin/hadoop fs -ls

Found 7 items
drwxr-xr-x   - root root       4096 1970-01-01 00:00 conf
drwxr-xr-x   - root root       4096 2022-03-09 18:39 eventlogs
drwxr-xr-x   - root root       4096 2022-03-09 19:30 ganglia
drwxr-xr-x   - 1001 1001       4096 2020-07-06 19:50 hadoop-3.3.0
-rw-r--r--   1 root root  500749234 2020-07-15 17:30 hadoop-3.3.0.tar.gz
drwxr-xr-x   - root root       4096 2022-03-09 19:04 logs
-r-xr-xr-x   1 root root    1307386 1970-01-01 00:00 preload_class.lst


# List the files in the HDFS using hdfs dfs
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -ls
```

```
Found 7 items
drwxr-xr-x   - root root       4096 1970-01-01 00:00 conf
drwxr-xr-x   - root root       4096 2022-03-09 18:39 eventlogs
drwxr-xr-x   - root root       4096 2022-03-09 19:30 ganglia
drwxr-xr-x   - 1001 1001       4096 2020-07-06 19:50 hadoop-3.3.0
-rw-r--r--   1 root root  500749234 2020-07-15 17:30 hadoop-3.3.0.tar.gz
drwxr-xr-x   - root root       4096 2022-03-09 19:04 logs
-r-xr-xr-x   1 root root    1307386 1970-01-01 00:00 preload_class.lst
```

```
# Create a new file and add the content into the file
!echo "This is a test file created by Khanh Nguyen." > myfile.txt
!date >> myfile.txt
! whoami >> myfile.txt
```

```
# List the files and found this new file
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -ls
```

```
Found 8 items
drwxr-xr-x   - root root       4096 1970-01-01 00:00 conf
drwxr-xr-x   - root root       4096 2022-03-09 18:39 eventlogs
drwxr-xr-x   - root root       4096 2022-03-09 19:30 ganglia
drwxr-xr-x   - 1001 1001       4096 2020-07-06 19:50 hadoop-3.3.0
-rw-r--r--   1 root root  500749234 2020-07-15 17:30 hadoop-3.3.0.tar.gz
drwxr-xr-x   - root root       4096 2022-03-09 19:04 logs
-rw-r--r--   1 root root         79 2022-03-09 19:31 myfile.txt
-r-xr-xr-x   1 root root    1307386 1970-01-01 00:00 preload_class.lst
```

```
# Print the file
!cat myfile.txt
```

```
This is a test file created by Khanh Nguyen.
Wed Mar  9 19:31:17 UTC 2022
root
```

```
# Create a new sub folder
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -mkdir myTest
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -ls

Found 9 items
drwxr-xr-x   - root root       4096 1970-01-01 00:00 conf
drwxr-xr-x   - root root       4096 2022-03-09 18:39 eventlogs
drwxr-xr-x   - root root       4096 2022-03-09 19:30 ganglia
drwxr-xr-x   - 1001 1001       4096 2020-07-06 19:50 hadoop-3.3.0
-rw-r--r--   1 root root  500749234 2020-07-15 17:30 hadoop-3.3.0.tar.gz
drwxr-xr-x   - root root       4096 2022-03-09 19:04 logs
drwxr-xr-x   - root root       4096 2022-03-09 19:32 myTest
-rw-r--r--   1 root root         79 2022-03-09 19:31 myfile.txt
-r-xr-xr-x   1 root root    1307386 1970-01-01 00:00 preload_class.lst


# Put the file onto the new HDFS subfolder:
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -put myfile.txt  myTest
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -ls myTest

Found 1 items
-rw-r--r--   1 root root         79 2022-03-09 19:32 myTest/myfile.txt


# Copy the created file onto the new HDFS subfolder and list the files
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -cp myfile.txt  myTest/myTestfile.txt
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -ls myTest

Found 2 items
-rw-r--r--   1 root root         79 2022-03-09 19:32 myTest/myTestfile.txt
-rw-r--r--   1 root root         79 2022-03-09 19:32 myTest/myfile.txt


# List the files in the local folder
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -ls myTest
```

```
Found 2 items
-rw-r--r--   1 root root         79 2022-03-09 19:32 myTest/myTestfile.txt
-rw-r--r--   1 root root         79 2022-03-09 19:32 myTest/myfile.txt


# Remove the file in the HDFS
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -rm myfile.txt
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -ls

2022-03-09 19:33:43,548 INFO Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use df
s.bytes-per-checksum
Deleted myfile.txt
Found 8 items
drwxr-xr-x   - root root       4096 1970-01-01 00:00 conf
drwxr-xr-x   - root root       4096 2022-03-09 18:39 eventlogs
drwxr-xr-x   - root root       4096 2022-03-09 19:30 ganglia
drwxr-xr-x   - 1001 1001       4096 2020-07-06 19:50 hadoop-3.3.0
-rw-r--r--   1 root root  500749234 2020-07-15 17:30 hadoop-3.3.0.tar.gz
drwxr-xr-x   - root root       4096 2022-03-09 19:04 logs
drwxr-xr-x   - root root       4096 2022-03-09 19:32 myTest
-r-xr-xr-x   1 root root    1307386 1970-01-01 00:00 preload_class.lst


# Remove the folder
!/usr/local/hadoop-3.3.0/bin/hdfs dfs -rm -r myTest

2022-03-09 19:33:54,069 INFO Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use df
s.bytes-per-checksum
Deleted myTest

Found 7 items
drwxr-xr-x   - root root       4096 1970-01-01 00:00 conf
drwxr-xr-x   - root root       4096 2022-03-09 18:39 eventlogs
drwxr-xr-x   - root root       4096 2022-03-09 19:30 ganglia
drwxr-xr-x   - 1001 1001       4096 2020-07-06 19:50 hadoop-3.3.0
-rw-r--r--   1 root root  500749234 2020-07-15 17:30 hadoop-3.3.0.tar.gz
```

```
drwxr-xr-x   - root root       4096 2022-03-09 19:04 logs
-r-xr-xr-x   1 root root    1307386 1970-01-01 00:00 preload_class.lst
```