Extras/Working with BDFS Alluxio

Alluxio (BDFS)

https://docs.oracle.com/en/cloud/paas/big-data-compute-cloud/csspc/big-data-file-system-bdfs.html (https://docs.oracle.com/en/cloud/paas/big-data-compute-cloud/csspc/big-data-file-system-bdfs.html)
and
https://www.alluxio.org/docs/master/en/index.html (https://www.alluxio.org/docs/master/en/index.html)
and
http://www.alluxio.org/docs/master/en/Configuration-Settings.html (http://www.alluxio.org/docs/master/en/Configuration-Settings.html)

Note: if running BDC on OCI (as compared to OCI-Classic), BDFS will not work until version 18.2.2

Display the alluxio command line help...

```
%sh
alluxio fs
```

```
Usage: java AlluxioShell
       [cat <path>]                                          Prints the file's contents to the console.
       [checksum <Alluxio path>]                             Calculates the md5 checksum of a file in the Alluxio filesystem.
       [chgrp [-R] <group> <path>]                           Changes the group of a file or directory specified by args. Specify -R to change the group recursively.
       [chmod [-R] <mode> <path>]                            Changes the permission of a file or directory specified by args. Specify -R to change the permission recursively.
       [chown [-R] <owner> <path>]                           Changes the owner of a file or directory specified by args. Specify -R to change the owner recursively.
       [copyFromLocal <src> <remoteDst>]                     Copies a file or a directory from local filesystem to Alluxio filesystem.
       [copyToLocal <src> <localDst>]                        Copies a file or a directory from the Alluxio filesystem to the local filesystem.
       [count <path>]                                        Displays the number of files and directories matching the specified prefix.
       [cp [-R] <src> <dst>]                                 Copies a file or a directory in the Alluxio filesystem. The -R flag is needed to copy directories.
       [createLineage <inputFile1,...> <outputFile1,...> [<cmd_arg1> <cmd_arg2> ...]]   Creates a lineage.
       [deleteLineage <lineageId> <cascade(true|false)>]     Deletes a lineage. If cascade is specified as true, dependent lineages will also be deleted.
       [du <path>]                                           Displays the size of the specified file or directory.
       [fileInfo <path>]                                     Displays all block info for the specified file.
       [free <path>]                                         Frees the space occupied by a file or a directory in Alluxio.
       [getCapacityBytes]                                    Gets the capacity of the Alluxio file system.
       [getUsedBytes]                                        Gets number of bytes used in the Alluxio file system.
       [leader]                                              Prints the current leader master host name

ExitValue: 255
```

An example of listing the alluxio (BDFS) file system

```
%sh
alluxio fs ls -R -f /citibike/
```

```
1.00B     05-10-2018 19:34:19:082  Directory      /citibike/raw
130.33MB  05-10-2018 19:34:19:082  In Memory      /citibike/raw/201612-citibike-tripdata.csv
1.00B     05-11-2018 14:12:48:033  Directory      /citibike/modified
130.33MB  05-11-2018 14:12:53:683  In Memory      /citibike/modified/201612-citibike-tripdata.nh.csv
```

Explicitly load the data we want to work with into BDFS

```
%sh
alluxio fs load /citibike/modified/201612-citibike-tripdata.nh.csv
alluxio fs ls -R /citibike
```

```
1.00B     02-03-2018 15:54:03:180  Directory      /citibike/raw
130.33MB  02-03-2018 15:54:03:181  In Memory      /citibike/raw/201612-citibike-tripdata.csv
1.00B     02-03-2018 18:51:31:613  Directory      /citibike/modified
130.33MB  02-03-2018 18:51:31:630  In Memory      /citibike/modified/201612-citibike-tripdata.nh.csv
```

An example of listing the alluxio files system using hadoop fs

```
%sh
hadoop fs -ls swift://journeyC.default/citibike/modified
# use the below LOGGER setting to avoid lots of INFO logging from alluxio by default
export HADOOP_ROOT_LOGGER=WARN
hadoop fs -ls bdfs://localhost:19998/citibike/modified
```

```
18/05/11 14:21:45 WARN httpclient.HttpMethodBase: Going to buffer response body of large or unknown size. Using getResponseBodyAsStream instead is recommended.
Found 1 items
-rw-rw-rw-   1  136661199 2018-02-21 19:13 swift://journeyC.default/citibike/modified/201612-citibike-tripdata.nh.csv
Found 1 items
-rw-rw-rw-   3  136661199 2018-05-11 14:12 bdfs://localhost:19998/citibike/modified/201612-citibike-tripdata.nh.csv
```

An example of using alluxio (BDFS) versus standard object store (swift)

```
%spark

// If you get this error message:
// java.lang.IllegalStateException: Cannot call methods on a stopped SparkContext.
// Then go to the Settings tab, then click on Notebook.  Then restart the Notebook.  This will restart your SparkContext
```

```
swift_df: org.apache.spark.sql.DataFrame = [528: string, 2016-12-01 00:00:04: string ... 13 more fields]
bdfs_df: org.apache.spark.sql.DataFrame = [528: int, 2016-12-01 00:00:04: timestamp ... 13 more fields]
# of rows: 812191
Swift Count Elapsed time: 6s
..
# of rows: 812191
BDFS Count Elapsed time: 1s
..
```

Create an external hive table against BDFS (alluxio)

```
%sh

/u01/bdcsce/opt/alluxio/bin/alluxio fs chmod 777 /citibike/modified/

hive  <<EOF
DROP TABLE bike_trips_objectstore_bdfs;

CREATE external TABLE bike_trips_objectstore_bdfs (
TripDuration int,
StartTime timestamp,
StopTime timestamp,
StartStationID string,
StartStationName string,
StartStationLatitude string,
StartStationLongitude string,
EndStationID string,
EndStationName string,
EndStationLatitude string,
EndStationLongitude string,
BikeID int,
UserType string,
BirthYear int,
Gender int
)
ROW FORMAT delimited
FIELDS TERMINATED BY ','
location 'bdfs://localhost:19998/citibike/modified/';


exit;

EOF
```

```
Changed permission of /citibike/modified to 777
WARNING: Use "yarn jar" to launch YARN applications.
Logging initialized using configuration in file:/etc/hive/2.4.2.0-258/0/hive-log4j.properties
hive> DROP TABLE bike_trips_objectstore_bdfs;
OK
Time taken: 0.876 seconds
hive>
    > CREATE external TABLE bike_trips_objectstore_bdfs (
    > TripDuration int,
    > StartTime timestamp,
    > StopTime timestamp,
    > StartStationID string,
    > StartStationName string,
    > StartStationLatitude string,
    > StartStationLongitude string,
    > EndStationID string,
    > EndStationName string,
    > EndStationLatitude string,
    > EndStationLongitude string,
    > BikeID int,
    > UserType string,
    > BirthYear int,
    > Gender int
```

```
    > )
    > ROW FORMAT delimited
    > FIELDS TERMINATED BY ','
    > location 'bdfs://localhost:19998/citibike/modified/';
OK
Time taken: 1.881 seconds
hive>
    >
    > exit;
```

Compare the performance of Spark SQL tables on object store (swift) versus bdfs versus hdfs

```
%spark

val swift_df=spark.sql("select * from bike_trips_objectstore")
val bdfs_df=spark.sql("select * from bike_trips_objectstore_bdfs")
val hdfs_df=spark.sql("select * from bike_trips")

{
var t0 = System.nanoTime()
println("# of rows: %s".format(
  swift_df.count()
))
var t1 = System.nanoTime()
println("Swift Count Elapsed time: " + (t1 - t0)/1000000000 + "s")
println("..")

t0 = System.nanoTime()
println("# of rows: %s".format(
  bdfs_df.count()
))
t1 = System.nanoTime()
println("BDFS Count Elapsed time: " + (t1 - t0)/1000000000 + "s")
println("..")

t0 = System.nanoTime()
println("# of rows: %s".format(
  hdfs_df.count()
))
t1 = System.nanoTime()
println("HDFS Count Elapsed time: " + (t1 - t0)/1000000000 + "s")
println("..")

}
```

```
swift_df: org.apache.spark.sql.DataFrame = [tripduration: int, starttime: timestamp ... 13 more fields]
bdfs_df: org.apache.spark.sql.DataFrame = [tripduration: int, starttime: timestamp ... 13 more fields]
hdfs_df: org.apache.spark.sql.DataFrame = [tripduration: int, starttime: timestamp ... 13 more fields]
# of rows: 812192
Swift Count Elapsed time: 6s
..
# of rows: 812192
BDFS Count Elapsed time: 0s
..
# of rows: 812192
HDFS Count Elapsed time: 3s
..
```

Test Alluxio via the Spark Thrift Server

```
%jdbc(sts)
select usertype, count(*) from bike_trips_objectstore_bdfs group by usertype
```

usertype	count(1)
	5388
Subscriber	774278
Customer	32526

View the Alluxio Web UI (port 19999)

The suggested way is to ssh into BDC and tunnel port 19999. Then point your local browser to http://127.0.0.1:19999/configuration (http://127.0.0.1:19999/configuration)

Example of using the Alluxio interpreter in zeppelin

```
%alluxio
help
```

```
Commands list:
        [help] - List all available commands.
        [cat <path>] - Prints the file's contents to the console.
        [chgrp [-R] <group> <path>] - Changes the group of a file or directory specified by args. Specify -R to change the group recursively.
        [chmod -R <mode> <path>] - Changes the permission of a file or directory specified by args. Specify -R to change the permission recursively.
        [chown -R <owner> <path>] - Changes the owner of a file or directory specified by args. Specify -R to change the owner recursively.
        [copyFromLocal <src> <remoteDst>] - Copies a file or a directory from local filesystem to Alluxio filesystem.
        [copyToLocal <src> <localDst>] - Copies a file or a directory from the Alluxio filesystem to the local filesystem.
        [count <path>] - Displays the number of files and directories matching the specified prefix.
        [createLineage <inputFile1,...> <outputFile1,...> [<cmd_arg1> <cmd_arg2> ...]] - Creates a lineage.
        [deleteLineage <lineageId> <cascade(true|false)>] - Deletes a lineage. If cascade is specified as true, dependent lineages will also be deleted.
        [du <path>] - Displays the size of the specified file or directory.
        [fileInfo <path>] - Displays all block info for the specified file.
        [free <file path|folder path>] - Removes the file or directory(recursively) from Alluxio memory space.
        [getCapacityBytes] - Gets the capacity of the Alluxio file system.
        [getUsedBytes] - Gets number of bytes used in the Alluxio file system.
        [listLineages] - Lists all lineages.
```

```
%alluxio
ls -R /citibike
```

```
1.00B     05-10-2018 19:34:19:082  Directory      /citibike/raw
130.33MB  05-10-2018 19:34:19:082  In Memory      /citibike/raw/201612-citibike-tripdata.csv
1.00B     05-11-2018 14:12:48:033  Directory      /citibike/modified
130.33MB  05-11-2018 14:12:53:683  In Memory      /citibike/modified/201612-citibike-tripdata.nh.csv
```

Change Log

May 11, 2018 - Updated for BDC 18.2.2

```
%md
```