Homework 2. Big Data processing

Done by D. Skok m16-502

Business logic: Program which calculate linux syslog priority (is given by 7 – debug, 6 – info, 5 – notice, 4 - warning, warn, 3 - err, error, 2 - crit, 1 - alert, 0 - emerg, panic) count by hours.

Ingest technology: bash.

Storage technology: HDFS.

Computation technology: Spark SQL (DataFrame, DataSet)

Report includes:

1. ZIP-ed src folder with your implementation (attachment to e-mail).

2. Screenshot of successfully executed tests.

```
Results :

Tests run: 2, Failures: 0, Errors: 0, Skipped: 0

[INFO]
```

3. Screenshots of successfully executed job and result (logs).

```
17/12/15 17:51:58 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks have all completed, from pool
17/12/15 17:51:58 INFO CodeGenerator: Code generated in 98.464264 ms
+---+-----+
|_c1|count|
+---+-----+
|  0|  117|
|  1|  133|
|  2|  119|
|  3|  133|
|  4|  121|
|  5|  127|
|  6|  131|
|  7|  119|
+---+-----+

17/12/15 17:51:58 INFO SparkContext: Invoking stop() from shutdown hook
17/12/15 17:51:58 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
17/12/15 17:51:58 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
17/12/15 17:51:58 INFO MemoryStore: MemoryStore cleared
17/12/15 17:51:58 INFO BlockManager: BlockManager stopped
17/12/15 17:51:59 INFO BlockManagerMaster: BlockManagerMaster stopped
17/12/15 17:51:59 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
17/12/15 17:51:59 INFO SparkContext: Successfully stopped SparkContext
17/12/15 17:51:59 INFO ShutdownHookManager: Shutdown hook called
17/12/15 17:51:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-347229d1-5c6f-468c-abff-8e077bf7644b
[cloudera@quickstart target]$
```

4. Quick build and deploy manual (commands, OS requirements etc).

REQUIREMENTS

Cloudera Quickstart VM 5.12.0, Apache Spark 2.1.0, 4+ GiB RAM

SET UP

`git clone https://DariaSkok@bitbucket.org/DariaSkok/bd_hw2.git`

`cd bd_hw2`

mvn clean install

INPUT DATA

input data is automatically generated(bash script)

RUN

cd target

java -jar spark-1.0-SNAPSHOT-jar-with-dependencies.jar OUTPUT_DIRECTORY

5. System components communication diagram (UML or COMET).