Project 1. Analyzing Book- Crossing Data

Steps:

1: Start the Hadoop cluster by using command: ***start-all.sh***

2: Extract the content of *BX-CSV-Dump.zip* to ***~/input*** directory: ***unzip BX-CSV-Dump.zip -d ~/input***

3: Copy the content of **input/** to **hdfs://input/** folder using command:
hadoop dfs -copyFromLocal input/ /input

```
jazhar192@hadoop:~$ ls -l input/
total 117936
-rw-rw-r-- 1 jazhar192 jazhar192 30682276 Dec 23 14:31 BX-Book-Ratings.csv
-rw-rw-r-- 1 jazhar192 jazhar192 77787439 Dec 23 14:31 BX-Books.csv
-rw-rw-r-- 1 jazhar192 jazhar192 12284157 Dec 23 14:31 BX-Users.csv
jazhar192@hadoop:~$ hadoop dfs -copyFromLocal input/ /input
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

15/12/23 14:32:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
jazhar192@hadoop:~$ hadoop dfs -ls /input
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

15/12/23 14:32:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
-rw-r--r--   1 jazhar192 supergroup   30682276 2015-12-23 14:32 /input/BX-Book-Ratings.csv
-rw-r--r--   1 jazhar192 supergroup   77787439 2015-12-23 14:32 /input/BX-Books.csv
-rw-r--r--   1 jazhar192 supergroup   12284157 2015-12-23 14:32 /input/BX-Users.csv
jazhar192@hadoop:~$
```

Note:** To verify execute command: ***hadoop dfs -ls /input

4.1: Execute the bookfrequency.jar on **hdfs://input/BX-Books.csv** data set using command:
hadoop jar bookfrequency.jar BookFrequency /input/BX-Books.csv /usecase1

```
15/12/23 14:32:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
-rw-r--r--   1 jazhar192 supergroup   30682276 2015-12-23 14:32 /input/BX-Book-Ratings.csv
-rw-r--r--   1 jazhar192 supergroup   77787439 2015-12-23 14:32 /input/BX-Books.csv
-rw-r--r--   1 jazhar192 supergroup   12284157 2015-12-23 14:32 /input/BX-Users.csv
jazhar192@hadoop:~$ hadoop jar bookfrequency.jar BookFrequency /input/BX-Books.csv /usecase1
15/12/23 14:38:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
15/12/23 14:38:20 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
15/12/23 14:38:20 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/12/23 14:38:20 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with Too
lRunner to remedy this.
15/12/23 14:38:20 INFO input.FileInputFormat: Total input paths to process : 1
15/12/23 14:38:20 INFO mapreduce.JobSubmitter: number of splits:1
15/12/23 14:38:20 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1978803048_0001
15/12/23 14:38:21 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/12/23 14:38:21 INFO mapreduce.Job: Running job: job_local1978803048_0001
15/12/23 14:38:21 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/12/23 14:38:21 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
15/12/23 14:38:21 INFO mapred.LocalJobRunner: Waiting for map tasks
15/12/23 14:38:21 INFO mapred.LocalJobRunner: Starting task: attempt_local1978803048_0001_m_000000_0
15/12/23 14:38:21 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/12/23 14:38:21 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/input/BX-Books.csv:0+77787439
15/12/23 14:38:21 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/12/23 14:38:21 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/12/23 14:38:21 INFO mapred.MapTask: soft limit at 83886080
15/12/23 14:38:21 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/12/23 14:38:21 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/12/23 14:38:21 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
15/12/23 14:38:22 INFO mapreduce.Job: Job job_local1978803048_0001 running in uber mode : false
15/12/23 14:38:22 INFO mapreduce.Job:  map 0% reduce 0%
15/12/23 14:38:23 INFO mapred.LocalJobRunner:
15/12/23 14:38:23 INFO mapred.MapTask: Starting flush of map output
15/12/23 14:38:23 INFO mapred.MapTask: Spilling map output
15/12/23 14:38:23 INFO mapred.MapTask: bufstart = 0; bufend = 2374140; bufvoid = 104857600
15/12/23 14:38:23 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 25153108(100612432); length = 1061289/6553600
15/12/23 14:38:24 INFO mapred.MapTask: Finished spill 0
15/12/23 14:38:24 INFO mapred.Task: Task:attempt_local1978803048_0001_m_000000_0 is done. And is in the process of committing
15/12/23 14:38:24 INFO mapred.LocalJobRunner: map
15/12/23 14:38:24 INFO mapred.Task: Task 'attempt_local1978803048_0001_m_000000_0' done.
15/12/23 14:38:24 INFO mapred.LocalJobRunner: Finishing task: attempt_local1978803048_0001_m_000000_0
15/12/23 14:38:24 INFO mapred.LocalJobRunner: map task executor complete.
15/12/23 14:38:24 INFO mapred.LocalJobRunner: Waiting for reduce tasks
```

4.2: To see output execute command: ***hadoop dfs -cat /usecase1/p****

```
15/12/23 14:38:25 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=5816772
                FILE: Number of bytes written=9252732
jazhar192@hadoop:~$ hadoop dfs -cat /usecase1/p*
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

15/12/23 14:42:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
0       4589
1376    1
1378    1
1806    1
1897    1
1900    3
1901    7
1902    2
1904    1
1906    1
1908    1
1909    2
1910    1
1911    19
1914    1
1917    1
1919    1
1920    33
1921    2
1922    2
1923    11
1924    2
1925    2
1926    2
1927    1
1928    2
1929    7
1930    13
1931    3
1932    5
1933    4
1934    1
1935    3
```

---------- End of Case 1 ----------

5.1: Execute the maxbook.jar on **hdfs://input/BX-Books.csv** data set using command:
hadoop jar maxbook.jar MaxBook /input/BX-Books.csv /usecase2

```
jazhar192@hadoop:~$ hadoop jar maxbook.jar MaxBook /input/BX-Books.csv /usecase2
15/12/23 14:51:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
15/12/23 14:51:14 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
15/12/23 14:51:14 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/12/23 14:51:14 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with Too
lRunner to remedy this.
15/12/23 14:51:14 INFO input.FileInputFormat: Total input paths to process : 1
15/12/23 14:51:14 INFO mapreduce.JobSubmitter: number of splits:1
15/12/23 14:51:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local712170952_0001
15/12/23 14:51:15 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/12/23 14:51:15 INFO mapreduce.Job: Running job: job_local712170952_0001
15/12/23 14:51:15 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/12/23 14:51:15 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
15/12/23 14:51:15 INFO mapred.LocalJobRunner: Waiting for map tasks
15/12/23 14:51:15 INFO mapred.LocalJobRunner: Starting task: attempt_local712170952_0001_m_000000_0
15/12/23 14:51:15 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/12/23 14:51:15 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/input/BX-Books.csv:0+77787439
15/12/23 14:51:15 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/12/23 14:51:15 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/12/23 14:51:15 INFO mapred.MapTask: soft limit at 83886080
15/12/23 14:51:15 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/12/23 14:51:15 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/12/23 14:51:15 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
15/12/23 14:51:16 INFO mapreduce.Job: Job job_local712170952_0001 running in uber mode : false
15/12/23 14:51:16 INFO mapreduce.Job:  map 0% reduce 0%
15/12/23 14:51:17 INFO mapred.LocalJobRunner:
15/12/23 14:51:17 INFO mapred.MapTask: Starting flush of map output
15/12/23 14:51:17 INFO mapred.MapTask: Spilling map output
15/12/23 14:51:17 INFO mapred.MapTask: bufstart = 0; bufend = 1843494; bufvoid = 104857600
15/12/23 14:51:17 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 25153108(100612432); length = 1061289/6553600
15/12/23 14:51:17 INFO mapred.MapTask: Finished spill 0
15/12/23 14:51:17 INFO mapred.Task: Task:attempt_local712170952_0001_m_000000_0 is done. And is in the process of committing
15/12/23 14:51:17 INFO mapred.LocalJobRunner: map
15/12/23 14:51:17 INFO mapred.Task: Task 'attempt_local712170952_0001_m_000000_0' done.
15/12/23 14:51:17 INFO mapred.LocalJobRunner: Finishing task: attempt_local712170952_0001_m_000000_0
15/12/23 14:51:17 INFO mapred.LocalJobRunner: map task executor complete.
15/12/23 14:51:17 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/12/23 14:51:17 INFO mapred.LocalJobRunner: Starting task: attempt_local712170952_0001_r_000000_0
15/12/23 14:51:17 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
```

5.2: To see output execute command: ***hadoop dfs -cat /usecase2/p****

```
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=77787439
        File Output Format Counters
                Bytes Written=41
azhar192@hadoop:~$ hadoop fs -cat /usecase2/p*
5/12/23 14:54:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ax book published in year 2002 :        17298
azhar192@hadoop:~$
```

---------- End of Case 2 ----------

6.1 Execute the numbook.jar on **hdfs://input/BX-Books.csv** and **hdfs://input/BX-Book-Ratings.csv** data set using command:

hadoop jar numbook.jar NumBook /usecase3

```
jazhar192@hadoop:~$ hadoop jar numbook.jar NumBook /usecase3
15/12/23 15:45:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
15/12/23 15:45:28 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
15/12/23 15:45:28 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/12/23 15:45:28 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with Too
lRunner to remedy this.
15/12/23 15:45:28 INFO input.FileInputFormat: Total input paths to process : 1
15/12/23 15:45:28 INFO input.FileInputFormat: Total input paths to process : 1
15/12/23 15:45:28 INFO mapreduce.JobSubmitter: number of splits:2
15/12/23 15:45:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local973183642_0001
15/12/23 15:45:29 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/12/23 15:45:29 INFO mapreduce.Job: Running job: job_local973183642_0001
15/12/23 15:45:29 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/12/23 15:45:29 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
15/12/23 15:45:29 INFO mapred.LocalJobRunner: Waiting for map tasks
15/12/23 15:45:29 INFO mapred.LocalJobRunner: Starting task: attempt_local973183642_0001_m_000000_0
15/12/23 15:45:29 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/12/23 15:45:29 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/input/BX-Books.csv:0+77787439
15/12/23 15:45:29 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/12/23 15:45:29 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/12/23 15:45:29 INFO mapred.MapTask: soft limit at 83886080
15/12/23 15:45:29 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/12/23 15:45:29 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/12/23 15:45:29 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
15/12/23 15:45:30 INFO mapreduce.Job: Job job_local973183642_0001 running in uber mode : false
15/12/23 15:45:30 INFO mapreduce.Job:  map 0% reduce 0%
15/12/23 15:45:31 INFO mapred.LocalJobRunner:
15/12/23 15:45:31 INFO mapred.MapTask: Starting flush of map output
15/12/23 15:45:31 INFO mapred.MapTask: Spilling map output
15/12/23 15:45:31 INFO mapred.MapTask: bufstart = 0; bufend = 276771; bufvoid = 104857600
15/12/23 15:45:31 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26145208(104580832); length = 69189/6553600
15/12/23 15:45:31 INFO mapred.MapTask: Finished spill 0
15/12/23 15:45:31 INFO mapred.Task: Task:attempt_local973183642_0001_m_000000_0 is done. And is in the process of committing
15/12/23 15:45:31 INFO mapred.LocalJobRunner: map
15/12/23 15:45:31 INFO mapred.Task: Task 'attempt_local973183642_0001_m_000000_0' done.
15/12/23 15:45:31 INFO mapred.LocalJobRunner: Finishing task: attempt_local973183642_0001_m_000000_0
15/12/23 15:45:31 INFO mapred.LocalJobRunner: Starting task: attempt_local973183642_0001_m_000001_0
15/12/23 15:45:31 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
```

5.2: To see output execute command: ***hadoop fs -cat /usecase3/p****

```
        File Input Format Counters
                Bytes Read=110970
        File Output Format Counters
                Bytes Written=72
jazhar192@hadoop:~$ hadoop fs -cat /usecase3/p*
15/12/23 15:46:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
0       16094
1       38
10      1914
2       85
3       155
4       237
5       1053
6       918
7       1956
8       2835
9       1979
jazhar192@hadoop:~$
```

---------- End of Case 3 ----------

***Configuration:**

I. Hadoop v2.6.2

```
jazhar192@hadoop:~$ hadoop version
Hadoop 2.6.2
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r 0cfd050febe4a30b1ee1551dcc527589509fb681
Compiled by jenkins on 2015-10-22T00:42Z
Compiled with protoc 2.5.0
From source with checksum f9ebb94bf5bf9bec892825ede28baca
This command was run using /home/jazhar192/hadoop/share/hadoop/common/hadoop-common-2.6.2.jar
```

II. Added following to ***~/.bashrc*** file

export CLASSPATH=$CLASSPATH:$HADOOP_INSTALL/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.6.2.jar:$HADOOP_INSTALL/share/hadoop/common/hadoop-common-2.6.2.jar

export HADOOP_USER_CLASSPATH_FIRST=true

```
#HADOOP VARIABLES START
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
export HADOOP_INSTALL=/home/jazhar192/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
#HADOOP VARIABLES END

# Set HIVE_HOME
export HIVE_HOME=/home/jazhar192/hive
export PATH=$PATH:$HIVE_HOME/bin
export PATH
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
#export CLASSPATH=$CLASSPATH:/home/realjamshed/hive/lib/*:.
export CLASSPATH=$CLASSPATH:$HADOOP_INSTALL/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.6.2.jar:$HADOOP_INSTALL/share/hadoop/common/hadoop-common-2.6.2.jar
export HADOOP_USER_CLASSPATH_FIRST=true
# HIVE VARIABLES END

#PIG VARIABLE START
export PIG_HOME=/home/jazhar192/pig
export PATH=$PATH:$PIG_HOME/bin

# HBase VARIABLE
export HBASE_HOME=/home/jazhar192/hbase
export PATH=$PATH:$HBASE_HOME/bin

export PIG_CLASSPATH=$PIG_HOME/lib/h2/*:$PIG_HOME/pig-withouthadoop.jar:$HBASE_HOME/lib/*:$HADOOP_HOME/lib/*:$CLASSPATH
                                                                                                                    146,1         Bot
```