BDS Assignment 2 group 2023

Name BITS ID Contribution
CHANDUPATLA ANIRUDH REDDY 2022DA04387 100%
MEDAPATI JAYANAGA SURESH REDDY 2022DA04445 100%
YASH RATHORE 2022DA04617 100%

System details:

```
UUID: 032e02b4-0499-0590-2c06-580700080009
CPU max MHz:          3200.0000
CPU min MHz:          1550.0000
```

```
YARN top - 17:49:28, up 0d, 0:28, 0 active users, queue(s): root
NodeManager(s): 1 total, 1 active, 0 unhealthy, 0 decommissioned, 0 lost, 0 rebooted
Queue(s) Applications: 0 running, 0 submitted, 0 pending, 0 completed, 0 killed, 0 failed
Queue(s) Mem(GB): 8 available, 0 allocated, 0 pending, 0 reserved
Queue(s) VCores: 8 available, 0 allocated, 0 pending, 0 reserved
Queue(s) Containers: 0 allocated, 0 pending, 0 reserved
```

1 Configuration Files:

hdfs-site.xml:

```
<property>
 <name>dfs.replication</name>
 <value>1</value>
 </property>
```

core-site.xml:

```
 <property>
 <name>fs.defaultFS</name>
 <value>hdfs://localhost:9000</value>  </property>
 <property>
<name>hadoop.proxyuser.dataflair.groups</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.dataflair.hosts</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.server.hosts</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.server.groups</name> <value>*</value>
 </property>
```

mapred-site.xml:

```
 <property>
 <name>mapreduce.framework.name</name>  <value>yarn</value>
 </property>
 <property>
 <name>mapreduce.application.classpath</name>
<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:
$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>
 </property>
```

yarn-site.xml:

```
 <property>
 <name>yarn.nodemanager.aux-services</name>
 <value>mapreduce_shuffle</value>
 </property>
 <property>
 <name>yarn.nodemanager.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CON
F_DIR,CLASSPATH_PREP
END_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
 </property>
```

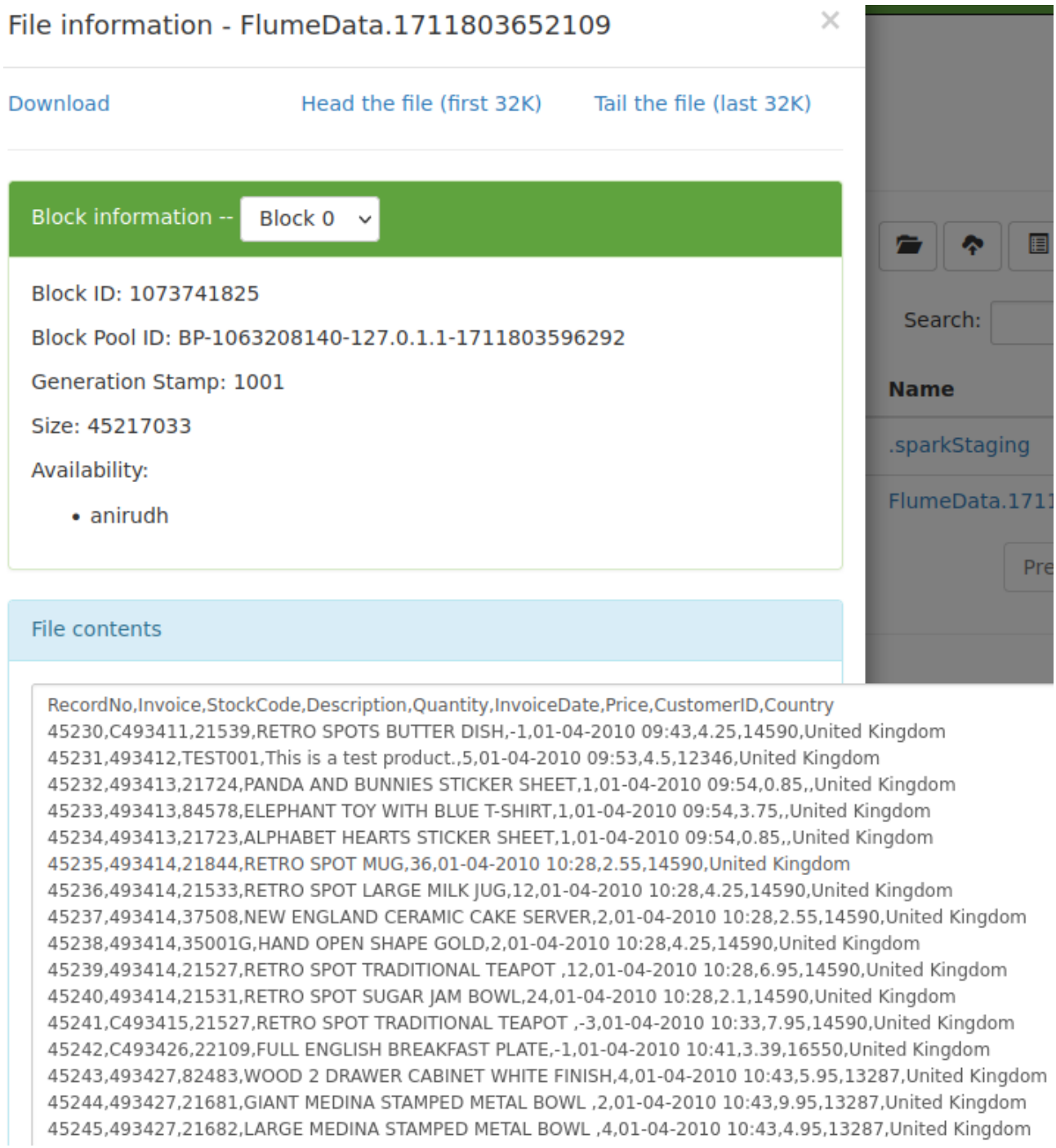
log4j.properties:

```
# Log levels of third-party libraries
log4j.logger.org.apache.commons.beanutils=WARN

# Set root logger level to INFO and its only appender to STDOUT.
log4j.rootLogger=INFO, STDOUT

# Hadoop streaming appender configuration
log4j.logger.org.apache.hadoop.mapreduce.v2.app.MRAppMaster=INFO, STDOUT

# STDOUT appender configuration
log4j.appender.STDOUT=org.apache.log4j.ConsoleAppender
log4j.appender.STDOUT.layout=org.apache.log4j.PatternLayout
log4j.appender.STDOUT.layout.ConversionPattern=%d{yyyy-MM-dd HH:mm:ss} %-5p %c{1}:
%L - %m%n
```

2 flume agent configuration:

```
# Define Source, Channel & Sink
flatfile_agent.sources = splDirSrc
flatfile_agent.channels = memoryChannel
flatfile_agent.sinks = hdfsSink

# Configuring the souce
flatfile_agent.sources.splDirSrc.type = spooldir
flatfile_agent.sources.splDirSrc.spoolDir = /home/anirudh/BDS_assignment/
flatfile_agent.sources.splDirSrc.inputCharset = ISO-8859-9

# Configuring the channel
flatfile_agent.channels.memoryChannel.type = memory
flatfile_agent.channels.memoryChannel.capacity = 1000
flatfile_agent.channels.memoryChannel.transactionCapacity = 1000
flatfile_agent.sources.splDirSrc.fileHeader = false

# Configuring the Sink
flatfile_agent.sinks.hdfsSink.type = hdfs
flatfile_agent.sinks.hdfsSink.hdfs.path = /user/anirudh
flatfile_agent.sinks.hdfsSink.hdfs.round = true
flatfile_agent.sinks.hdfsSink.hdfs.roundValue = 1
```

```
flatfile_agent.sinks.hdfsSink.hdfs.roundUnit = hour
flatfile_agent.sinks.hdfsSink.hdfs.useLocalTimeStamp = true
#flatfile_agent.sinks.hdfsSink.hdfs.path = /user/training/rana
flatfile_agent.sinks.hdfsSink.hdfs.fileType = DataStream

# Write format can be text or writable
flatfile_agent.sinks.hdfsSink.hdfs.writeFormat = Text

flatfile_agent.sinks.hdfsSink.hdfs.batchSize = 100

# Rollover file based on maximum size
flatfile_agent.sinks.hdfsSink.hdfs.rollSize = 0

# Never rollover based on the number of events
flatfile_agent.sinks.hdfsSink.hdfs.rollCount = 0

flatfile_agent.sources.splDirSrc.channels = memoryChannel
flatfile_agent.sinks.hdfsSink.channel = memoryChannel
```



3 code, query commands:

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus
application_1711803610688_0003	anirudh	query 1 & 2 Spark SQL	SPARK		default	0	Sat Mar 30 18:34:25 +0550 2024	Sat Mar 30 18:34:25 +0550 2024	Sat Mar 30 18:35:02 +0550 2024	FINISHED	SUCCEEDED
application_1711803610688_0002	anirudh	query 2 Hadoop Map Reduce	MAPREDUCE		default	0	Sat Mar 30 18:33:25 +0550 2024	Sat Mar 30 18:33:25 +0550 2024	Sat Mar 30 18:33:40 +0550 2024	FINISHED	SUCCEEDED
application_1711803610688_0001	anirudh	query 1 Hadoop Map Reduce	MAPREDUCE		default	0	Sat Mar 30 18:32:35 +0550 2024	Sat Mar 30 18:32:35 +0550 2024	Sat Mar 30 18:32:51 +0550 2024	FINISHED	SUCCEEDED

Showing 1 to 3 of 3 entries

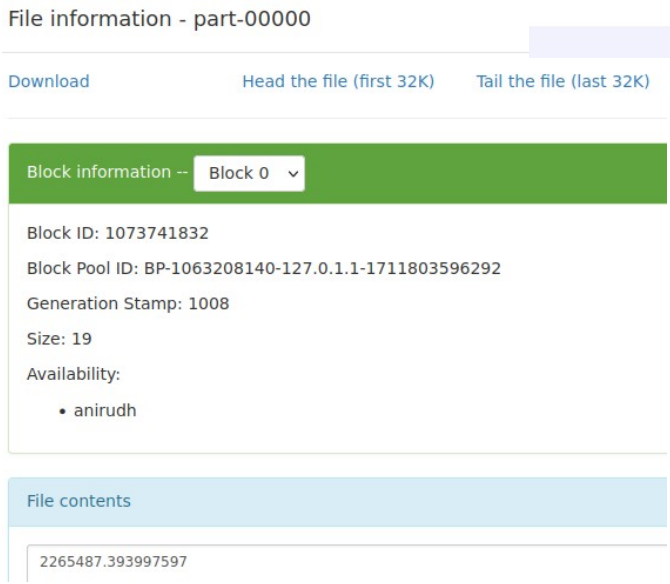
Hadoop mapreduce:
query 1 command:

```
hadoop jar /home/anirudh/hadoop-3.3.2/share/hadoop/tools/lib/hadoop-streaming-3.3.2.jar \
-D mapreduce.job.name="query 1 Hadoop Map Reduce" \
-input /user/anirudh/FlumeData.* \
-output ~/reducer_output \
-mapper ~/hadoop-mapper-query1.py \
-reducer ~/hadoop-reducer-query1.py
```

query 1 output:
2265487.393997597

time taken: 15 seconds

User:	anirudh
Name:	query 1 Hadoop Map Reduce
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Mar 30 18:32:35 +0530 2024
Launched:	Sat Mar 30 18:32:35 +0530 2024
Finished:	Sat Mar 30 18:32:51 +0530 2024
Elapsed:	16sec



query 2 command:

```
hadoop jar /home/anirudh/hadoop-3.3.2/share/hadoop/tools/lib/hadoop-streaming-3.3.2.jar \
-D mapreduce.job.name="query 2 Hadoop Map Reduce" \
-input /user/anirudh/FlumeData.* \
-output ~/reducer_output2 \
-mapper ~/hadoop-mapper-query2.py \
-reducer ~/hadoop-reducer-query2.py
```

time taken: 14 seconds
query 2 output:
Stockcode Quantity
10002 6537
10002R 1
10080 98
10109 -4
10120 -8658
10123C 245
10123G 1228
10124A 122
10124C -5
10124G 21

User:	anirudh
Name:	query 2 Hadoop Map Reduce
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Mar 30 18:33:25 +0530 2024
Launched:	Sat Mar 30 18:33:25 +0530 2024
Finished:	Sat Mar 30 18:33:40 +0530 2024
Elapsed:	14sec

10125	727
10133	851
10134	552
10135	1909
10138	-208
11001	2457
15002	-1000
15030	101
15034	3527
15036	19398
15039	2987
15044A	4484
15044B	1340
15044C	650
15044D	927
15056BL	4614
15056N	4975
15056P	761
15056bl	37
15056n	53
15056p	17
15058A	264
15058B	206
15058C	226
15060A	-8
15060B	899
15060b	2
16008	251
16010	14
16011	1893
16012	1618
16014	12502
16015	568
16016	1025
16020C	31
16033	11322
16043	744
16044	5112
16045	412
16046	5207

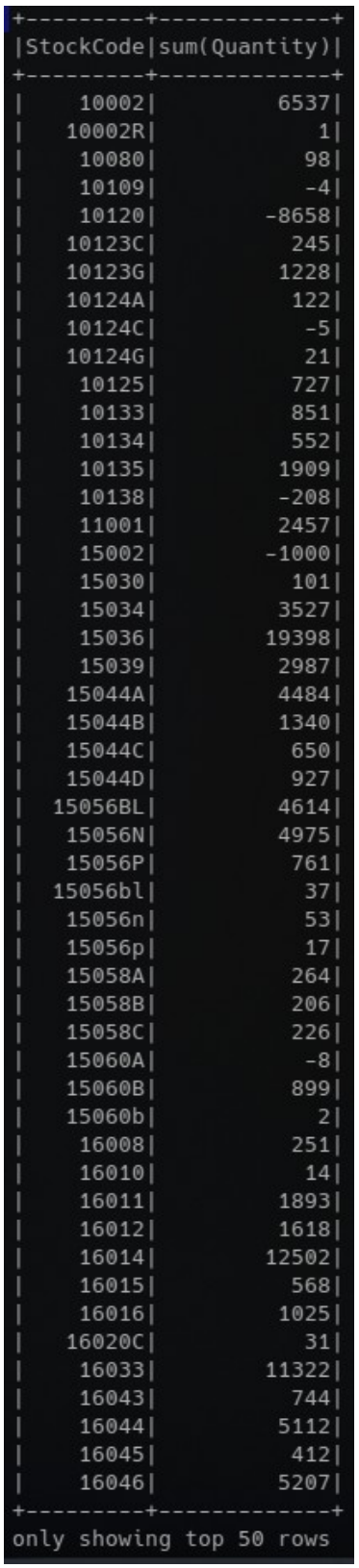
Spark command:

```
spark-submit --master yarn --deploy-mode client --num-executors 4 --executor-memory 4G --executor-cores 4 --driver-memory 4G --driver-cores 2 spark-pyspark-sql.py
```

spark query 1 output:

```
total revenue: 2265487.3939997507
query 1 time: 0.09971427917480469 seconds
```

spark query 2 output and `time: 0.03753304481506348 seconds`



StockCode	sum(Quantity)
10002	6537
10002R	1
10080	98
10109	-4
10120	-8658
10123C	245
10123G	1228
10124A	122
10124C	-5
10124G	21
10125	727
10133	851
10134	552
10135	1909
10138	-208
11001	2457
15002	-1000
15030	101
15034	3527
15036	19398
15039	2987
15044A	4484
15044B	1340
15044C	650
15044D	927
15056BL	4614
15056N	4975
15056P	761
15056bl	37
15056n	53
15056p	17
15058A	264
15058B	206
15058C	226
15060A	-8
15060B	899
15060b	2
16008	251
16010	14
16011	1893
16012	1618
16014	12502
16015	568
16016	1025
16020C	31
16033	11322
16043	744
16044	5112
16045	412
16046	5207

only showing top 50 rows