

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 source

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
```

File information - FlumeData.1711803652109

[Download](#)

[Head the file \(first 32K\)](#)

[Tail the file \(last 32K\)](#)

Block information --

Block 0 ▾

Block ID: 1073741825

Block Pool ID: BP-1063208140-127.0.1.1-1711803596292

Generation Stamp: 1001

Size: 45217033

Availability:

- anirudh

File contents

```
RecordNo,Invoice,StockCode,Description,Quantity,InvoiceDate,Price,CustomerID,Country
45230,C493411,21539,RETRO SPOTS BUTTER DISH,-1,01-04-2010 09:43,4.25,14590,United Kingdom
45231,493412,TEST001,This is a test product.,5,01-04-2010 09:53,4.5,12346,United Kingdom
45232,493413,21724,PANDA AND BUNNIES STICKER SHEET,1,01-04-2010 09:54,0.85,,United Kingdom
45233,493413,84578,ELEPHANT TOY WITH BLUE T-SHIRT,1,01-04-2010 09:54,3.75,,United Kingdom
45234,493413,21723,ALPHABET HEARTS STICKER SHEET,1,01-04-2010 09:54,0.85,,United Kingdom
45235,493414,21844,RETRO SPOT MUG,36,01-04-2010 10:28,2.55,14590,United Kingdom
45236,493414,21533,RETRO SPOT LARGE MILK JUG,12,01-04-2010 10:28,4.25,14590,United Kingdom
45237,493414,37508,NEW ENGLAND CERAMIC CAKE SERVER,2,01-04-2010 10:28,2.55,14590,United Kingdom
45238,493414,35001G,HAND OPEN SHAPE GOLD,2,01-04-2010 10:28,4.25,14590,United Kingdom
45239,493414,21527,RETRO SPOT TRADITIONAL TEAPOT ,12,01-04-2010 10:28,6.95,14590,United Kingdom
45240,493414,21531,RETRO SPOT SUGAR JAM BOWL,24,01-04-2010 10:28,2.1,14590,United Kingdom
45241,C493415,21527,RETRO SPOT TRADITIONAL TEAPOT ,-3,01-04-2010 10:33,7.95,14590,United Kingdom
45242,C493426,22109,FULL ENGLISH BREAKFAST PLATE,-1,01-04-2010 10:41,3.39,16550,United Kingdom
45243,493427,82483,WOOD 2 DRAWER CABINET WHITE FINISH,4,01-04-2010 10:43,5.95,13287,United Kingdom
45244,493427,21681,GIANT MEDINA STAMPED METAL BOWL ,2,01-04-2010 10:43,9.95,13287,United Kingdom
45245,493427,21682,LARGE MEDINA STAMPED METAL BOWL ,4,01-04-2010 10:43,4.95,13287,United Kingdom
```

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

File information - part-00000

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

[Download](#)

[Head the file \(first 32K\)](#)

[Tail the file \(last 32K\)](#)

Block information -- Block 0 ▾

Block ID: 1073741832

Block Pool ID: BP-1063208140-127.0.1.1-1711803596292

Generation Stamp: 1008

Size: 19

Availability:

• anirudh

File contents

2265487.393997597

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	

File information - part-00000

[Download](#)[Head the file \(first 32K\)](#)[Tail the](#)

Block information -- Block 0 ▾

Block ID: 1073741842

Block Pool ID: BP-1063208140-127.0.1.1-1711803596292

Generation Stamp: 1018

Size: 46155

Availability:

- anirudh

File contents

```
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
```

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