**Flume to Hive POC**

**Step 1:** Go to below path from edge node of your cluster and edit flume.conf file

Path- /etc/flume-ng/conf

Edit flume.conf

a1.sources = test\_source

a1.channels = c1

a1.sinks = hdfs-Cluster1-sink

# Describe the source

a1.sources.test\_source.type = exec

a1.sources.test\_source.command = cat /home/cloudera/Desktop/Project1.2/TitanicData.txt

a1.sources.test\_source.batchSize = 2

a1.sources.test\_source.channels = memoryChannel

a1.sources.test\_source.interceptors = itime

a1.sources.test\_source.interceptors.itime.type=timestamp

#Describe the sink

a1.sinks.hdfs-Cluster1-sink.type = hdfs

a1.sinks.hdfs-Cluster1-sink.channel = memoryChannel

a1.sinks.hdfs-Cluster1-sink.hdfs.path = hdfs://172.27.74.95:8020/hive\_data

a1.sinks.hdfs-Cluster1-sink.hdfs.fileType = DataStream

a1.sinks.hdfs-Cluster1-sink.hdfs.useLocalTimeStamp = true

a1.sinks.hdfs-Cluster1-sink.hdfs.rollCount = 0

a1.sinks.hdfs-Cluster1-sink.hdfs.rollSize = 0

a1.sinks.hdfs-Cluster1-sink.hdfs.rollInterval = 60

#use a channel which buffers events in memory

a1.channels.c1.type = memory

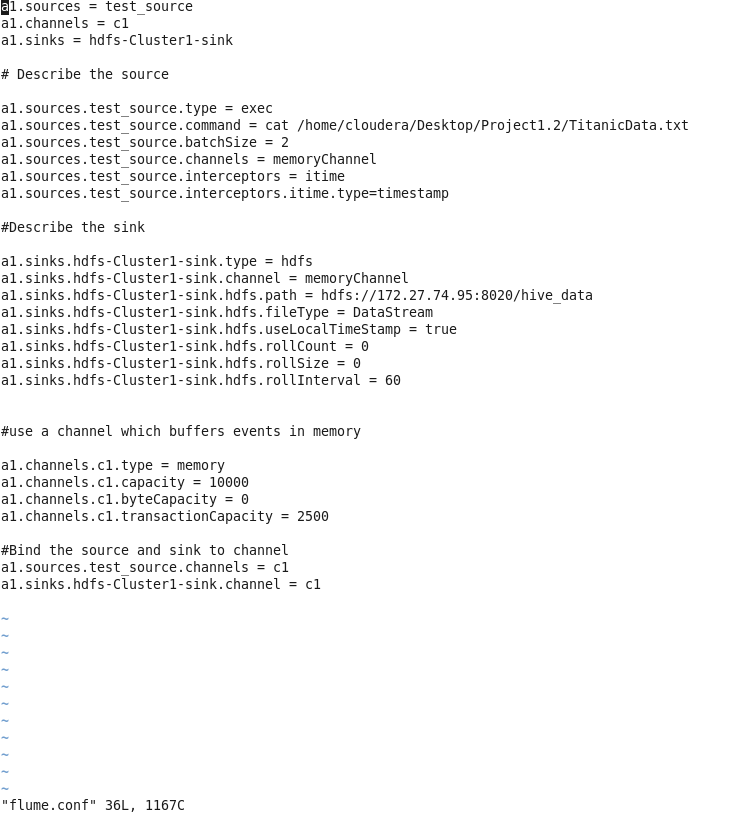
a1.channels.c1.capacity = 10000

a1.channels.c1.byteCapacity = 0

a1.channels.c1.transactionCapacity = 2500

#Bind the source and sink to channel

a1.sources.test\_source.channels = c1

a1.sinks.hdfs-Cluster1-sink.channel = c1

**Step 2:** Execute command from path mentioned below

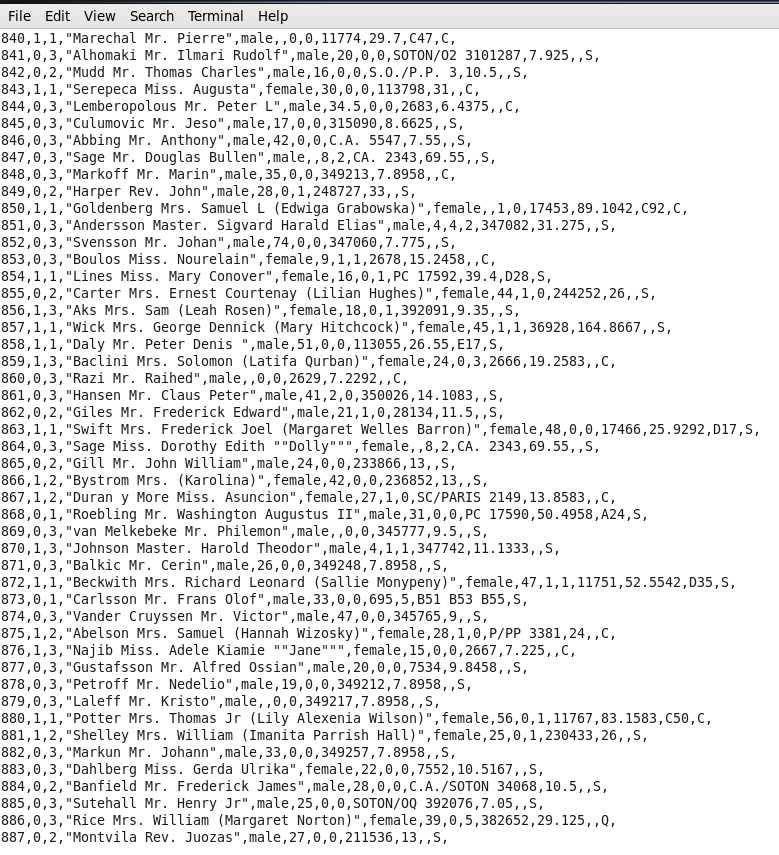
Path: /etc/flume-ng/conf

Command: flume-ng agent --conf conf --conf-file flume.conf --name a1 -Dflume.root.logger=INFO.console



**Step 3:** Check flume data on hdfs

Command: hadoop fs -cat hdfs://172.27.74.95:8020/hive\_data/FlumeData.1495792002572.tmp



**Step 4:** Create managed hive table

create table Titanic891(

PassengerId int,

Survived int,

Pclass int,

Name string,

Sex string,

Age int,

SibSp int,

Parch int,

Ticket bigint,

Fare double,

Cabin string,

Embarked string

)

row format delimited

fields terminated by ','

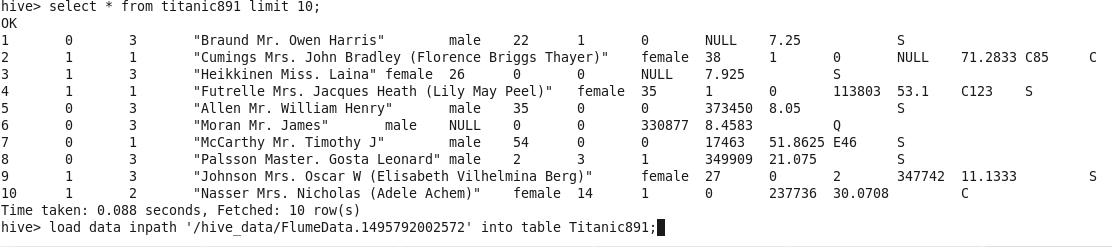
lines terminated by '\n'

stored as textfile;

**Step 5:** Load flume data from hdfs to hive table

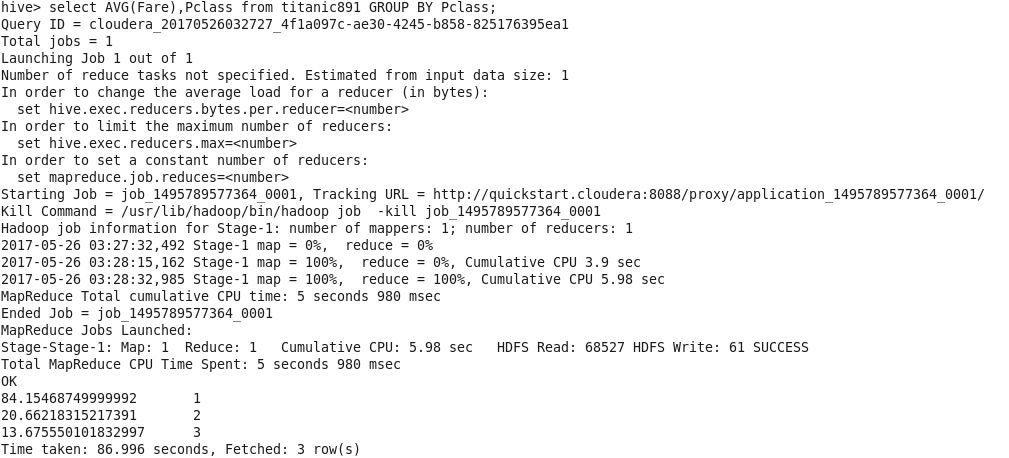
load data inpath '/hive\_data/FlumeData.1495792002572' into table Titanic891;

**Step 6:** Check data in hive table



**Problem Statement 1:** In this problem statement, we will find the average fare of each class.

**Hive Query:** hive> select AVG(Fare),Pclass from titanic891 GROUP BY Pclass;



**Output:**

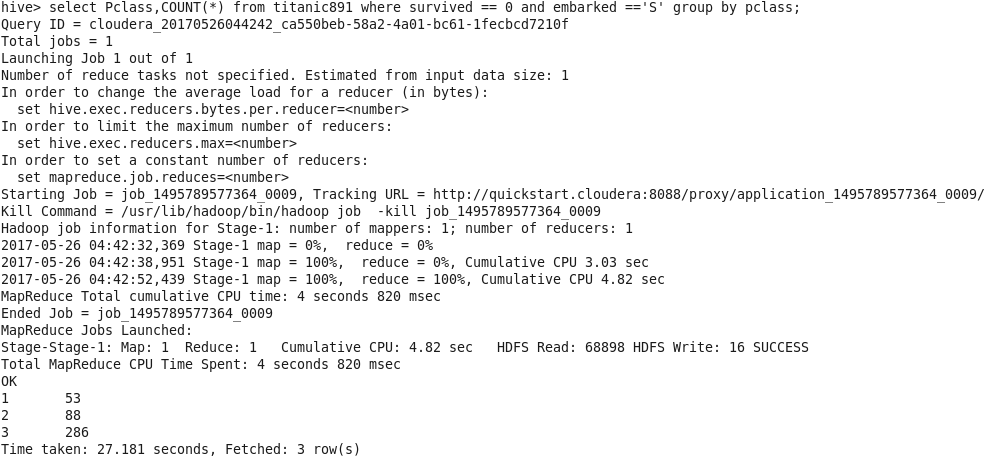
84.15468749999992 1

20.66218315217391 2

13.675550101832997 3

**Problem Statement 2:** In this problem statement, we will find the number of people alive in each class and embarked at Southampton.

**Hive Query:** hive> select Pclass,COUNT(\*) from titanic891 where survived == 0 and embarked =='S' group by pclass;



**Output:**

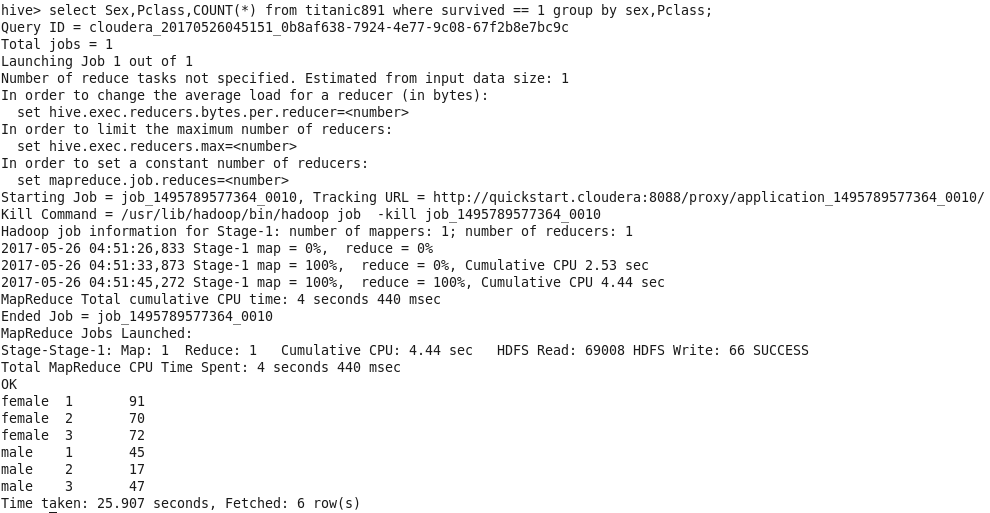
1 53

2 88

3 286

**Problem Statement 3:** In this problem statement, we will find out number of males and females who died in each class.

**Hive Query:** hive> select Sex,Pclass,COUNT(\*) from titanic891 where survived == 1 group by sex,Pclass;



**Output:**

female 1 91

female 2 70

female 3 72

male 1 45

male 2 17

male 3 47

**Concept for hive query writing:**

SELECT [ALL | DISTINCT] select\_expr, select\_expr, ...

FROM table\_reference

[WHERE where\_condition]

[GROUP BY col\_list]

[HAVING having\_condition]

[ORDER BY col\_list]]

[LIMIT number];