**TITANIC DATA ANALYSIS:**

Coping dataset into HDFS using flume:

**flume\_titanic.conf**

agent1.sources=source1

agent1.channels=channel1

agent1.sinks=sink1

agent1.channels.channel1.type=file

agent1.sinks.sink1.channel=channel1

agent1.sources.source1.channels=channel1

agent1.sources.source1.type=spooldir

agent1.sources.source1.spoolDir=/home/acadgild/pig/flume

agent1.sources.source1.fileHeader=false

agent1.sources.source1.fileSuffix= .COMPLETED

agent1.sinks.sink1.type=hdfs

agent1.sinks.sink1.hdfs.path=hdfs://localhost.localdomain:9000/flume\_titanic

agent1.sinks.sink1.hdfs.batchSize=1000000000

agent1.sinks.sink1.hdfs.rollSize=234564324567

agent1.sinks.sink1.hdfs.rollInterval=10000

agent1.sinks.sink1.hdfs.rollCount=150000000

agent1.sinks.sink1.hdfs.fileType= DataStream

start flume agent:

**flume-ng agent –name agent1 –conf-file flume\_titanic.conf**

**Please find screenshots.**

***Loading data:***

titanic\_data= LOAD '/flume\_titanic/FlumeData.1489635005060' USING PigStorage(',') AS (id:int,survived:int,pclass:int,name:chararray,sex:chararray,age:int,sibsp:chararray,parch:chararray,ticket:chararray,fare:int,cain:chararray,embarked:chararray);

**1)**

grp= GROUP titanic\_data BY pclass;

avg= FOREACH grp GENERATE group,AVG(titanic\_data.fare);

DUMP avg;

**Output**:

(1,83.74537037037037)

(2,20.456521739130434)

(3,13.173116089613035)

**2)**

filtered= FILTER titanic\_data BY survived==0 AND embarked=='S';

grp= GROUP filtered BY pclass;

counted= FOREACH grp GENERATE group,COUNT(filtered);

DUMP counted;

**Output**:

(1,53)

(2,88)

(3,286)

**3)**

grouped= GROUP filtered BY (sex,pclass);

count= FOREACH grouped GENERATE group.sex,group.pclass,COUNT(filtered);

DUMP(count);

**Output**:

(male,1,45)

(male,2,17)

(male,3,47)

(female,1,91)

(female,2,70)

(female,3,72)