TITANIC PROJECT 1.2

For this concepts used : Oracle VM, MobaXterm, hadoop, flume, pig

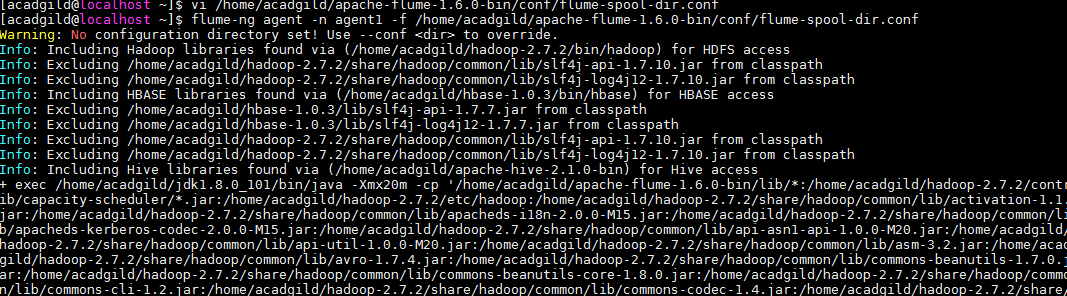
The first step is to copy the file onto HDFS path. Two approaches were tried using flume. Attaching the flume.conf file in both methods in the repository. Flume conf files are :

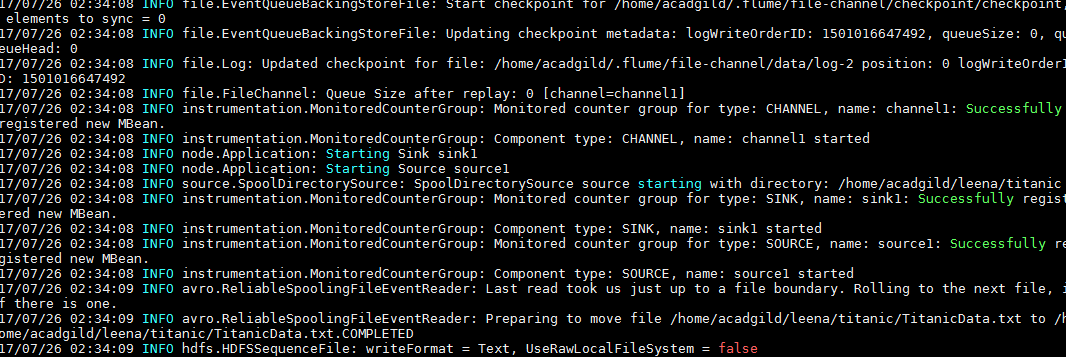
1. ~/apache-flume-1.6.0-bin/conf/flume-spool-dir.conf
2. ~/apache-flume-1.6.0-bin/conf/myflume.conf

Attaching screenshots here for the flume run and the outputs.

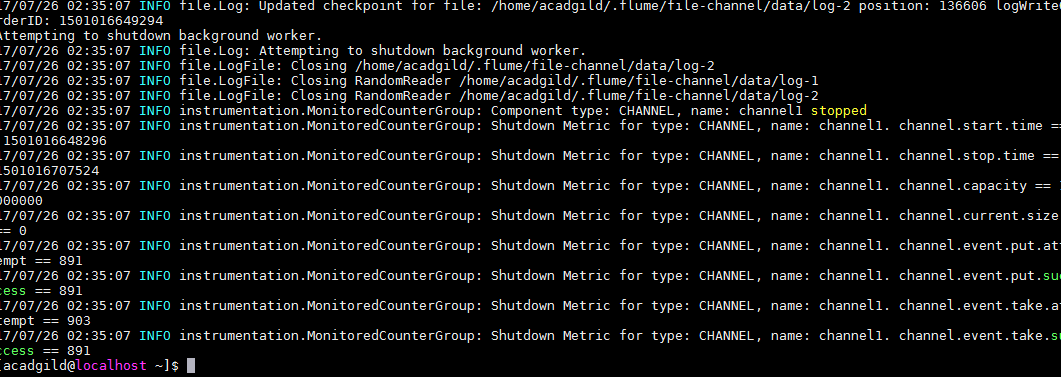
Flume is run using

flume-ng agent -n agent1 -f /home/acadgild/apache-flume-1.6.0-bin/conf/flume-spool-dir.conf

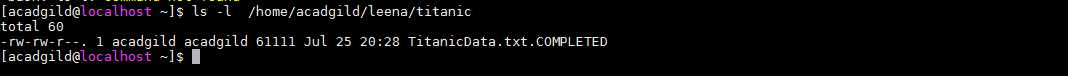




Using ctrl+C this is stopped



The file is marked ‘COMPLETED’.



Have the file “TitanicData.txt” in HDFS at /home/acadgild/ass-hdfs

Start Pig using pig and enter the grunt shell

pig

Next, register

REGISTER '/home/acadgild/pig-0.16.0/lib/piggybank.jar';

A = load 'ass-hdfs/TitanicData.txt' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

B = FOREACH A GENERATE (int)$0 AS PassengerID, (int)$1 AS Survived, (int)$2 AS PClass, (chararray)$4 AS Name, (chararray)$5 AS Sex, (int)$6 AS Age, (int)$7 AS SibSp, (int)$8 AS Parch, (chararray)$9 AS Ticket, (float)$10 AS Fare, (chararray)$11 AS Cabin, (chararray)$12 AS Embarked; B = FOREACH A GENERATE (int)$0 AS PassengerID, (int)$1 AS Survived, (int)$2 AS PClass, (chararray)$3 AS Name, (chararray)$4 AS Sex, (int)$5 AS Age, (int)$6 AS SibSp, (int)$7 AS Parch, (chararray)$8 AS Ticket, (float)$9 AS Fare, (chararray)$10 AS Cabin, (chararray)$11 AS Embarked;

1. Average Fare for each class

C = GROUP B BY PClass;

D = FOREACH C GENERATE B.PClass, AVG(B.Fare);

Output :



1. No of people alive in each class and embarked at Southampton

E = FILTER B BY Embarked == 'S' AND Survived == 0;

F = GROUP E all;

G = FOREACH F GENERATE COUNT(E.Name);



1. No of males and females who died in each class

First get the no of died passengers. Then get them into two relations of males and females. Then group by PClass. Use Count to get the count of males and females on the two relations.

grunt> DIED = FILTER B BY Survived == 1;

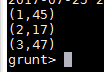
grunt> DIED\_MALES = FILTER DIED BY Sex == 'male';

grunt> DIED\_FEMALES = FILTER DIED BY Sex == 'female';

grunt> GRP\_MALES = GROUP DIED\_MALES BY PClass;

grunt> GRP\_FEMALES = GROUP DIED\_FEMALES BY PClass;

grunt> C\_MALES = FOREACH GRP\_MALES GENERATE group, COUNT(DIED\_MALES.Name);



grunt> C\_FEMALES = FOREACH GRP\_FEMALES GENERATE group, COUNT(DIED\_FEMALES.Name);

