*Project 1.2 - Titanic Data Analysis*

*By: Krushna Chandra Debashram*

*02-Aug-2017*

**Introduction:**

The file ‘TitanicData.txt’ has dataset contains information about passengers who boarded Titanic ship. It contains data points like:

Passenger’s age

Their native place

Details of who survived

Fare details of various travel classes

Number of casualties from various classes etc.

**Dataset Description:**

**DATA SET DESCRIPTION**

Column 1: PassengerId

Column 2: Survived (survived=0 & died=1)

Column 3: Pclass

Column 4: Name

Column 5: Sex

Column 6: Age

Column 7: SibSp

Column 8: Parch

Column 9: Ticket

Column 10: Fare

Column 11: Cabin

Column 12: Embarked

**Objective:**

Analysis the passenger data.

1. Find the average fare of each class.
2. Find the number of people alive in each class and embarked at Southampton.
3. Find out number of males and females who died in each class.

**Prerequisites:**

Hadoop cluster installed in the system.

Flume

**Program used:**

Hive

**Start Hadoop daemon:**

**[**acadgild@localhost ~]$ start-all.sh

[acadgild@localhost ~]$ mr-jobhistory-daemon.sh start historyserver

[acadgild@localhost ~]$ jps

3122 ResourceManager

3301 Jps

3669 JobHistoryServer

2966 SecondaryNameNode

3256 NodeManager

2733 NameNode

2830 DataNode

**Start MySql :**

[acadgild@localhost ~]$ sudo service mysqld start

Starting mysqld: [ OK ]

**Create a directory in Hadoop:**

[acadgild@localhost ~]$ hadoop dfs -mkdir /user/krushnadebashram\_project1\_2

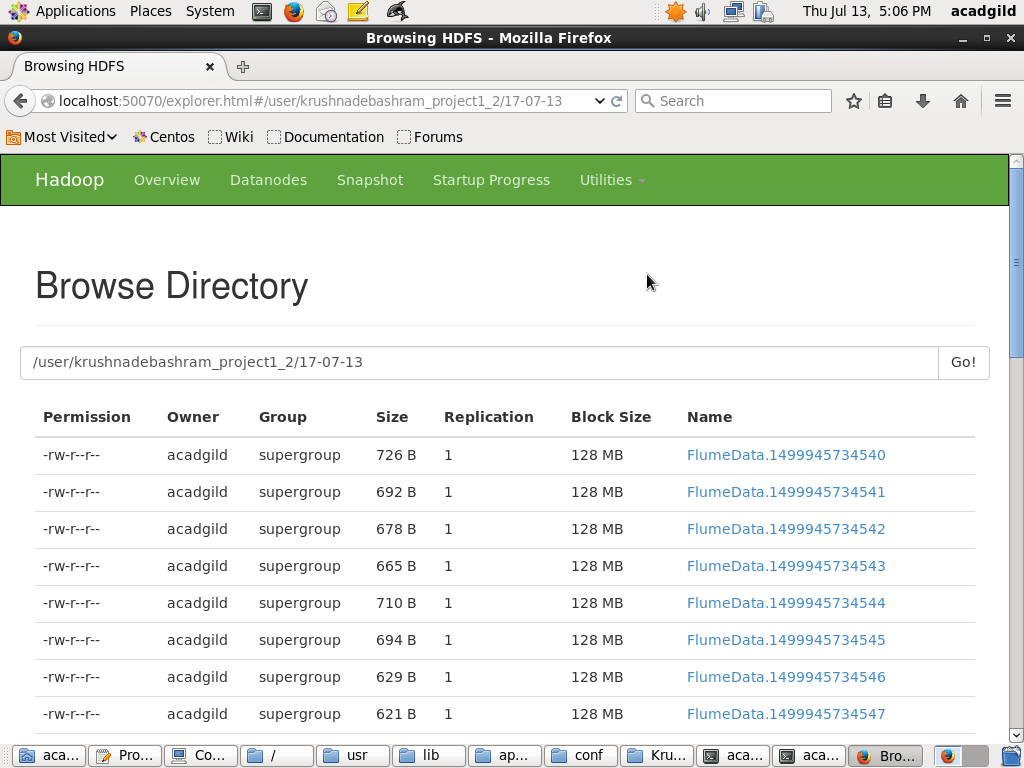
**Create Flume Conf file:**

Create a conf file flume\_Titanic.CONF

|  |
| --- |
| agent.channels = channel1  agent.channels.channel1.type=memory  agent.channels.channel1.capacity = 10000  agent.channels.memorychannel1.transactionCapacity=10000000  # Source definition  agent.sources = source1  agent.sources.source1.channels = channel1  agent.sources.source1.type = exec  agent.sources.source1.command = head -n -1 /home/acadgild/KrushnaDebashram/TitanicData.txt  #sink definition  agent.sinks=sink1  agent.sinks.sink1.channel=channel1  agent.sinks.sink1.type=hdfs  agent.sinks.sink1.hdfs.useLocalTimeStamp = true  agent.sinks.sink1.hdfs.path = /user/krushnadebashram\_project1\_2/%y-%m-%d  agent.sinks.sink1.hdfs.writeFormat=Text  agent.sinks.sink1.hdfs.fileType = DataStream |

**Upload the file in Hadoop using flume:**

[acadgild@localhost ~]$flume-ng agent -n agent -c conf -f /usr/lib/apache-flume-1.4.0-bin/conf/flume\_Titanic.Conf

****

**Go to hive shell:**

Go to hive

[acadgild@localhost ~]$ hive

hive>

**Create a hive table and populate table:**

CREATE TABLE TitanicData\_Hive(PassengerId INT, Survived INT , Pclass INT, Name STRING, Sex STRING, Age INT, SibSp INT,Parch INT,Ticket STRING, Fare DECIMAL(10,5), Cabin STRING, Embarked STRING)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/hive/data/TitanicData';

LOAD data inpath '/user/krushnadebashram\_project1\_2/17-07-13/\*' into table TitanicData\_Hive;

**Problem 1:**

In this problem statement, we will find the average fare of each class.

**Solution 1:**

Select Pclass,Avg(Fare) From TitanicData\_Hive group by PClass;

1 84.1546875

2 20.662183152

3 13.687643061

**Problem 2:**

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

**Solution 2:**

Select Pclass,Count(PClass) From TitanicData\_Hive where Survived=0 and Embarked='S' group by PClass;

1 53

2 88

3 286

**Problem 3:**

In this problem statement, we will find out number of males and females who died in each

class.

**Solution 3:**

Select Pclass,Sex,Count(Sex) From TitanicData\_Hive where Survived=1 group by PClass,Sex;

1 female 91

1 male 45

2 female 70

2 male 17

3 female 72

3 male 47