Project 2

ACADGILD

Project 2 - Titanic Data Analysis

Table of Contents

1. Introduction

The data set 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.

2. Objective

3. Prerequisites

You should have Hadoop cluster installed in your system.

4. Associated Data Files

https://drive.google.com/file/d/0ByJLBTmJojjzNmV0dk1EMmwwQ1U/view?usp=sharing

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

5. Problem Statement

You can use any of the technologies like Map Reduce, Pig or Hive of your choice.

Note:You need to copy the data set into HDFS using flume and send the screen shot of that with the project solution

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

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

3.In this problem statement we will find out number of male and female people died in each class.

6. Approximate Time to Complete Task

300 Minutes