

Problem Statement:

Carry out below two tasks (Task #1 Task #2 and Task #3) as mentioned below

Task # 1:

Using the data set provided at [/user/jayantm/Batch43/CUTe/PartB/Task1/RDD_dataset.csv](#) to your local Linux directory, use **flume** to copy the data from Linux machine to HDFS.

Task # 2:

Construct the data from task #1 as a Spark Rdd and answer the following questions.

1. Read the data as an RDD.
2. If there are any null fields in the dataset, skip the records.
3. Count the total number of records in the datasets.
4. Find the revenue generated by state and product.
5. Calculate the total revenue generated by total weekdays.

Task # 3:

The data for below task has been taken from kaggle and is available in HDFS at the following location.
 ('/user/jayantm/B43CUTe/PartB/Task3/SQL_dataset.csv')

First row in the dataset is the header. Treat fields with the value 'NA' as null values.

Dataset Description:

Column_Name	Description	Type
ROW_ID	Sequence number	Numeric
CASE_STATUS	Status associated with the last significant event or decision. Valid values include "Certified," "Certified-Withdrawn," "Denied," and "Withdrawn".	String
EMPLOYER_NAME	Name of employer submitting labor condition application.	String
SOC_NAME	Occupational name associated with the SOC_CODE. SOC_CODE is the occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System.	String
JOB_TITLE	Title of the job	String
FULL_TIME_POSITION	Y = Full Time Position; N = Part Time Position	String
PREVAILING_WAGE	Prevailing Wage for the job being requested for temporary labor condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer's minimum requirements for the position.	Numeric
YEAR	Year in which the H-1B visa petition was Filed	Numeric
WORKSITE	City and State information of the foreign worker's intended area of employment	String
lon	longitude of the Worksite	Numeric
lat	latitude of the Worksite	Numeric

Perform the below steps:

1. Create a dataframe for above csv data.
(First line in the dataset is header, comma is a part of data in few fields, fields are escaped with double-quote(")).
2. Verify summary of the dataframe (how many rows and columns).
3. Derive the summary statistics.
4. Find the count of distinct values in each column.
5. List EMPLOYER_NAME and YEAR in the descending order of the Approved applications count (Approved applications are obtained using CASE_STATUS = 'CERTIFIED').
6. List the approved applications count in the descending order for the JOB_TITLE = "DATA SCIENTIST" and for each employer and year.
7. Find the null values count in each column.
8. Remove all the rows with null values (in any column/position).
9. Verify the null values count in each column.
10. List the count of applications in each status (CASE_STATUS) in the descending order of the year.
11. Find the mean PREVAILING_WAGE for each year for the approved applications.
12. Find the mean PREVAILING_WAGE for each year for the approved applications for each employer.
13. Find the approved applications count in each year for the full-time positions in the descending order of the year.

NOTE:

1. Submit your answers as .zip file following the convention PartB_Batch43_<ID>.zip in Piazza **CSE7322c** Module for Batch43.

Note: Zip File should consist of 3 files with properly commented.

- a. Flume properties file used for **Task#1**
- b. .py file for **Task#2**
- c. .py file used for **Task#3**