Assignment – Spark

SparkSQL

**Task 0:**

1. Create a directory /tmp/sparkSQL\_DDAnumber
2. Copy the dataset file to this directory

**Task 1:**

Describe the dataset in your words. Also, write down your observations about the available values in each field.

**(50 marks)**

#Write your detailed answer here

Answer: The dataset is UpGrad's ad events dataset, which consists of all the data gather from marketing channels like Social networking websites, ads on search engines, link’s on IIITB’s and UpGrad’s websites. These data are further use by the marketing team to plan which marketing medium or which courses are having high demand that they can focus on.

**Task 2:**

Create table using this file. And display the schema for the same.

**(50 marks)**

#Write your command and specifically explain the commands and parameters used.

Answer: I have create a table in sparksql as below:-

%sql

Create Table Upgrad\_Marketing

using com.databricks.spark.csv

options (path "s3://ughdfsdemo/tmp/sparkSQL\_DDA1610218/ugmixp\_data", header "true",inferSchema "true")

#I have use the below command to see the schema of the table

%sql

describe Upgrad\_Marketing

#so, timestamp and distinct\_id are integer, UserRequestedDataAnalyticsInfo, blank and userstartedDataAnalyticsapplication is Boolean and rest all is string.

**Task 3:**

Perform an exploratory analysis on this data. Find out specifically -

* Number of cities the prospective students are visiting
* Number of cities from each state and name of the state
* Unique mediums, sources
* Number and names of referring domains
* Number and names of courses

**(50 marks)**

#Write your command and comment for your command.

Answer: The 1st query is the number of prospective students from each cities :-

%sql

select distinct (city),count(city) as No\_Of\_Prospective\_Students from Upgrad\_Marketing group by city

**or**

%sql

select count(distinct(lcase(city))) as No\_Of\_Cities from Upgrad\_Marketing

is – 1136 cities

#We see that the highest number of prospective students are from Mumbai followed by Bengaluru.

2nd. We need to find out the number of cities from each states :-

%sql

select distinct (lcase(region)) as States,count(distinct (lcase(city))) as No\_Of\_Cities\_In\_States from Upgrad\_Marketing group by region

#We see that the highest number of States from where the prospective students were visiting is Maharashtra followed by Gujarat.

3rd.We need to find out the number of unique medium and sources used in the campaign-

%sql

select count(distinct (lcase(utm\_medium))) as No\_Of\_Unique\_Communication\_Used, count(distinct (lcase(utm\_source))) as No\_of\_Sources from Upgrad\_Marketing

#We see that about 34 communication medium were used from about 61 different sources.

4th.We need to find out the number of referring domain with their names.

%sql

select count(distinct referring\_domain) as Unique\_No\_Of\_referring\_domain from Upgrad\_Marketing

#we found that there are 232 referring domain.

%sql

select distinct (referring\_domain) as Names\_of\_\_referring\_domain from Upgrad\_Marketing

#from the above command we get all the names of 232 domains.

5th.Now we need to find out number of courses with their names.

%sql

select count(distinct(lcase(course))) as Number\_of\_Courses\_Offered\_at\_UpGrad from Upgrad\_Marketing

#So there are 18 distinct courses offered at UpGrad.

%sql

select distinct lcase(course) as Names\_of\_Courses\_Offered\_at\_UpGrad from Upgrad\_Marketing

#the above query will list all the course names.

**Task 4:**

There is a field in the dataset called “Timestamp”, convert it into date:time format. Explain the the function/method you use. **(Hint: You need to do a research on this to find out suitable method for mysql.)**

**(50 marks)**

#Write your command and comment for your command.

Answer: %sql

select FROM\_UNIXTIME(Timestamp) as TimeStamp from DDA1610218.Marketing\_Upgrad

#I have use FROM\_UNIXTIME function since the value is an non-decimal with 10 digits and does not have any special character.

**Task 5:**

Find out the date on which maximum visits happened.

**(25 marks)**

#Write your command and comment for your command.

Answer: First we have create a table of only date column using below command.

%sql

Create Table Upgrad\_Date

as select substr(FROM\_UNIXTIME(Timestamp),1,10) as Date\_Time from Upgrad\_Marketing

#Next we have use group by on date\_time, to find out the date on which maximum number of visits has occurred using count and order by command, since max is not working on nested command.

%sql

select date\_time as Date,count(date\_time) as Max\_No\_Of\_Visits from Upgrad\_Date group by date\_time order by Max\_No\_Of\_Visits desc limit 1

#we see that 8th Sep 2016 a total of 5,837 visits has occurred.

**Task 6:**

Find out the number of times a page is visited and order it in descending order.

**(25 marks)**

#Write your command and comment for your command.

Answer: So we have use pagetitle column to count the page visits as it will be easy to know from title what the students prefer rather than using URL.

%sql

select PageTitle As Page\_Title,count(PageTitle) as No\_of\_Times\_Page\_Is\_Visited from Upgrad\_Marketing group by PageTitle order by No\_of\_Times\_Page\_Is\_Visited desc

**Task 7:**

Which ad mediums are most effective to reach to courses? Support your answer with suitable commands and your logic behind their use. **(Hint: Multiple commands (nested or individual) may be needed to get the answer).**

**(100 marks)**

#Write your command and comment for your command.

Answer: So communication medium can be answered by initial\_referring\_domain column which contain the name of initial communication, so first we have to standardized the column in lower case and then create a table using lower case column and use the group by command to count the visit through each medium.

%sql

create table UpGrad\_Medium

as select lcase(initial\_referring\_domain) as Communication\_Medium\_Used from Upgrad\_Marketing

%sql

select Communication\_Medium\_Used,count(Communication\_Medium\_Used) as No\_Of\_Visits\_Through\_Each\_CM from UpGrad\_Medium group by Communication\_Medium\_Used order by No\_Of\_Visits\_Through\_Each\_CM desc

#So, we see that search engine Google is the leading communication medium followed by direct marketing.

SparkR

**Task 1:**

Create a dataframe from this file. Describe it using all possible methods.

**(50 marks)**

#Write your command and comment for your command.

Answer: #We will create a DF using read.df command.

%r

S3URL = "s3://ughdfsdemo/tmp/sparkSQL\_DDA1610218/ugmixp\_data"

UpGrad\_Marketing <- read.df(sqlContext,S3URL,"com.databricks.spark.csv",header="true",inferschema="true")

#Now we will describe it using below command which will show us the schema of the DF

%r

printSchema(UpGrad\_Marketing)

UpGrad\_Marketing

%r

SparkR::head(UpGrad\_Marketing)

%r

UpGrad\_Marketing\_Data <- collect(UpGrad\_Marketing)

%r

UpGrad\_Marketing\_Data[1:10,]

%r

str(Clean\_Marketing\_Data)

**Task 2:**

Find out the missing values from this dataset. Also, suggest, which missing value treatment should be used.

**(50 marks)**

#Write your command and comment for your command.

Answer: We see that the datasets consists of both NA values and blank fields, If we use simple is.na function we will get only 281115 missing values.

%r

sum(is.na(UpGrad\_Marketing\_Data))

#but while looking at the datasets we see that there is blank fields which are not recognized as NA values and since sappy and apply function are not working in spark I have created below function to convert blank to NA’s.

%r

Blank\_NA <-

function(x, name = "No\_Blank") {

for (i in 1:ncol(x)) {

x[, i] <- ifelse(is.na(x[, i]) == FALSE & x[, i] == "", NA, x[, i])

}

assign(paste(name), x, environment(Blank\_NA))

}

%r

Blank\_NA(UpGrad\_Marketing\_Data,name = "Clean\_Marketing\_Data")

#now by using the sum(is.na()) function I am getting about 920937 NA values. The missing values treatment which I will be using is Mode since most of the NA’s are present in string values.

**Task 3:**

Plot a bar chart showing number of visits from each os type.

**(50 marks)**

#Write your command and comment for your command.

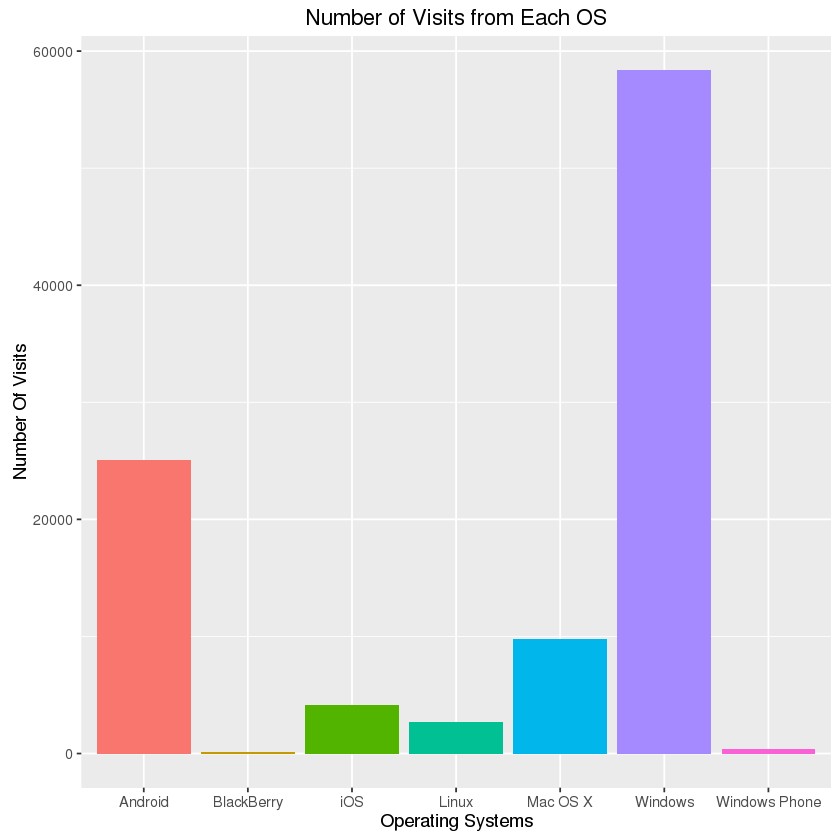
Answer: %r

library(ggplot2)

%r

ggplot(Clean\_Marketing\_Data,aes(x=os,fill=os)) + geom\_bar() + ggtitle("Number of Visits from Each OS") + xlab("Operating Systems") + ylab("Number Of Visits") + guides(fill=FALSE)

#I have use os in x axis and use geon\_bar to count the number of each os visits, ggtitle to name the x and y axis and removes the fill legend using guides(fill=FALSE). Which shows that windows is the leading in visits.



**Task 4:**

Plot a chart showing number of visits for each course. Select a chart type of your choice and defend your selection.

**(50 marks)**

#Write your command and comment for your command.

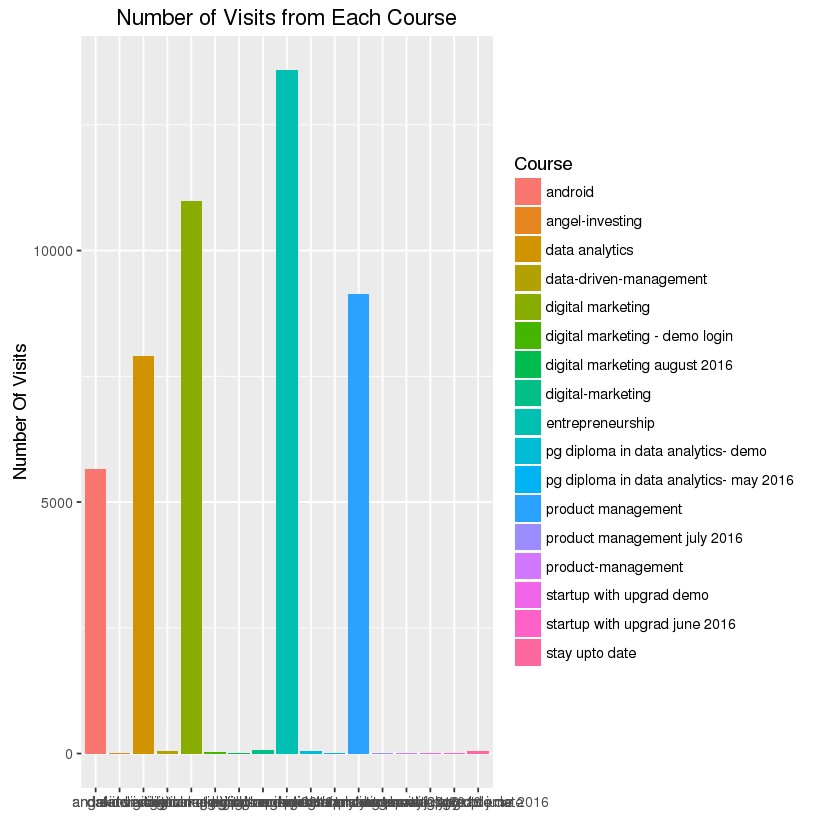
Answer: So, I have selected bar chart to view the courses with their number of visits,

%r

Clean\_Marketing\_Data$Course <- trimws(tolower(Clean\_Marketing\_Data$Course))

%r

ggplot(Clean\_Marketing\_Data,aes(x=Course,fill=Course)) + geom\_bar() + ggtitle("Number of Visits from Each Course") + xlab("Course") + ylab("Number Of Visits") + guides(fill=FALSE)



#we also see that there are 53493 NA values in course which is quite high, so we have not used any impute method, since as per mode we are getting entrepreneurship as the only option to replace the NA’s which will then become an outliers.

**Task 5:**

Which UTMMedium (initial as well as latest) is most effective and for which course?Support your answer with suitable comments.

**(150 marks)**

#Write your command and comment for your command.

Answer: So, First we have to find out the initial utm medium which is most effective with course:-

%r

Initial\_Medium <- aggregate.data.frame(Clean\_Marketing\_Data[,23],by=list(Clean\_Marketing\_Data$utm\_medium,Clean\_Marketing\_Data$Course),FUN = "length")

colnames(Initial\_Medium) <- c("Initial Medium","Course","No\_Of\_Visits")

Initial\_Medium[which(Initial\_Medium$No\_Of\_Visits == max(Initial\_Medium$No\_Of\_Visits)),]

#So we see that initial medium is more effective in “paid\_social” for “product management” Course which has 4,023 visits.

#Now we will check for latest medium on course.

%r

Latest\_Medium <- aggregate.data.frame(Clean\_Marketing\_Data[,15],by=list(Clean\_Marketing\_Data$LatestUTMMedium,Clean\_Marketing\_Data$Course),FUN = "length")

colnames(Latest\_Medium) <- c("latest Medium","Course","No\_Of\_Visits")

Latest\_Medium[which(Latest\_Medium$No\_Of\_Visits == max(Latest\_Medium$No\_Of\_Visits)),]

#We see that in latest medium “thankyoupage” is more effective for course “entrepreneurship” which has 4,071 views.

#So the above observation completely makes sense since initially to make the course known to the public it is must to do some paid social advertisement.