**CASE STUDY -1**

**Project Description:** The overview of the project is to definitely to increase the company value and provide insights which will give or effect the growth of the Microsoft. Here as a data analyst I will go throw the questions posed by different departments and will get useful insights by deriving the given data. To work on this data, I will use chose MYSQL. For case study1 have done the below analysis.

**Tech-Stack Used:** Here I am using MYSQL workbench because it is free and user friendly. I have already had a work experience on this. Not only that but it is recommended also by trainity.



**Creation of tables:**

create database project3;

use project3;

create table job\_data ( job\_id int ,

actor\_id int ,

event varchar(50),

language varchar(100),

time\_spent int,

org char(1),

ds varchar(10)

);

show tables;

#drop table job\_data;

select \* from job\_Data;

show variables like 'sercure\_file\_prev';

load data infile 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/job\_data.csv' into table job\_data fields terminated by ','

enclosed by '"' lines terminated by '\n' ignore 1 rows;

select \* from job\_data;

select count(\*) from job\_data;

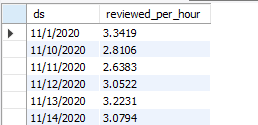
#######################################################

Here I have inserted nearly 5000 records to the existing data.

**Approach:**

**##A.Jobs Reviewed Over Time:** Calculate the number of jobs reviewed per hour for each day in November 2020.

select ds,(sum(time\_spent)/3600) reviewed\_per\_hour from job\_data group by ds order by ds asc;



Here I have used ds as a varchar so in order by did not worked properly. My plan is to use another column by using alter and change it into data format so I have inserted ds as varchar.

To get the number of jobs reviewed per hour in a day in November month I have converted hour into seconds because the data given has seconds of time taken. So I have converted the hour into seconds and used group by ds to get the total time used in a single day.

---------------------------------------------------------------------------------------------------------------------

**##B.Throughput Analysis:**

SELECT

date,

throughput\_per\_second,

seven\_day\_rolling\_avg

FROM

(

SELECT

date,

throughput\_per\_second,

AVG(throughput\_per\_second) OVER (

ORDER BY date

ROWS BETWEEN 6 PRECEDING AND CURRENT ROW

) AS seven\_day\_rolling\_avg

FROM

(SELECT

ds AS date,

COUNT(\*) \* 1.0 / SUM(time\_spent) AS throughput\_per\_second

FROM

job\_data

GROUP BY

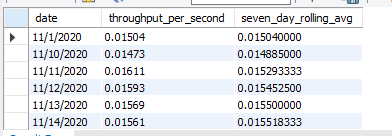
ds )daily\_throughput

)rolling\_avg

ORDER BY

date;

Here through put is nothing but the events for number of seconds. Here I haved tried to calculate it. First of all I have calculated the daily throughput like per second how many events. It is given..



---------------------------------------------------------------------------------------------------------------------

**#C.Language Share Analysis: Language Percentage in last 30 days**

**#select \*,date\_format(str\_to\_date(ds,'%m/%d/%Y'),'%Y-%m-%d %H:%i:%s') as date\_time from job\_data order by date\_time asc;**

**#alter table job\_data add column date\_time datetime;**

**#alter table job\_data drop column date\_time;**

**#update job\_data set date\_time = date\_format(str\_to\_date(ds,'%m/%d/%Y'),'%Y-%m-%d %H:%i:%s') where 1=1;**

**select b.language,**

**(count(b.language)/ (select count(\*) from job\_data a where a.date\_time between DATE\_FORMAT(b.date\_time, '%Y-%m-01') and LAST\_DAY(b.date\_time) ))\* 100 as language\_perc**

**#DATE\_FORMAT(b.date\_time, '%Y-%m-01') AS first\_day\_of\_month,**

**#LAST\_DAY(b.date\_time) Last\_day\_of\_month**

**from job\_data b**

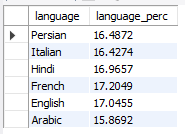
**where b.date\_time between DATE\_FORMAT(b.date\_time, '%Y-%m-01') and LAST\_DAY(b.date\_time)**

**and DATE\_FORMAT(b.date\_time, '%b')= 'Nov'**

**group by b.language order by language desc;**

Here I have added a column to get the ds in date format properly and modified data in ds column into date format.

Here I have calculated the percentage of the laguages by grouping the language and to get only the November month I have used that dateformat %b and to get the last date of the month I have used **LAST\_DAY(b.date\_time). It will calculate properly for everymonth if needed also.**



---------------------------------------------------------------------------------------------------------------------

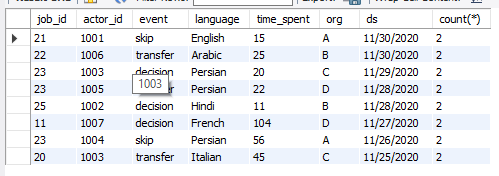
**##D.Duplicate rows detection:** Identify duplicate rows in the data.

#select \* from job\_data;

select job\_id,actor\_id,event,language,time\_spent,org,ds,count(\*) from job\_data

group by job\_id,actor\_id,event,language,time\_spent,org,ds

having count(\*) > 1;



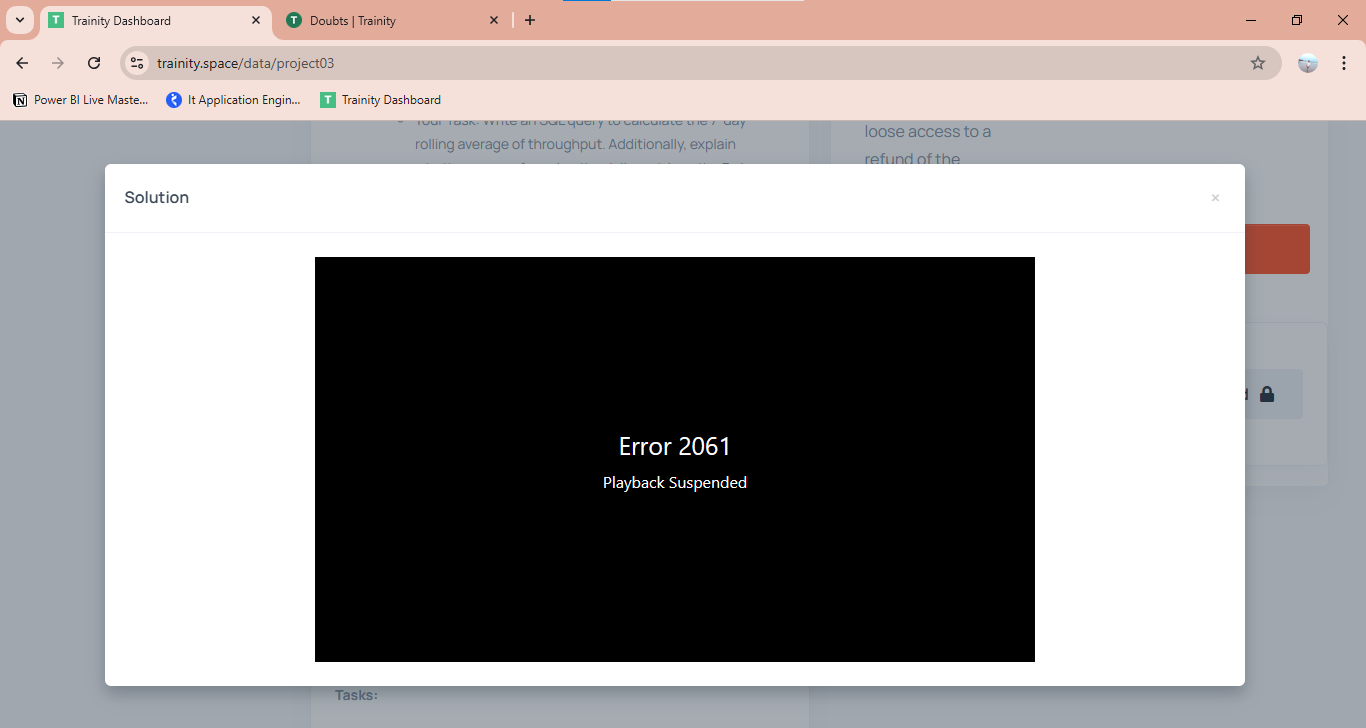
here I have used this query to know if any record has greater than 1 record then we can consider it as the duplicate record. To consider all the rows I have used group by for all the columns.

**Insights:** While working this project I have gained the knowledge of questioning the data and working every pinpoint corner of data. Cause till now I didn’t even know that based on the what is throughput. While working on these issues I again recalled all my SQL knowledge and basics again which we do not work often. I feel good to participate in this, because it improved my confidence than any other thing. In excel sheet also I have learnt so many things.

**Results:** I have found it a bit more tricky than the Instagram project. For this I have seen so many videos. I raised question also in community but I did not found any solution given to my questions. Even I have tried to watch the solution video and that did not worked. But finally I have seen one solution to add the rows and added 5000 lines of data to it. Now I have covered basics and some of advanced sql also with this.

Please let me know if any mistakes are there and any query optimization can be done to improve accuracy and efficiency. \

Please provide the answers to the questions raised and I did not watched the video and it raised error also. Please find it below and one chance of mine has gone please restore it back. Because I did not used it.



Thank you trainity.