

Day32 - March 8th 2024

1. Started my day as usual
2. Packed food and headed to library
3. Started solving leetcode problems

#### 4. Also started learning pyspark practically from youtuber manish kumar

Summary

Outline

Headings you add to the document will appear here.

⑭ Permissive → default  
→ set null value to all corrupted fields

9. We have 3 modes and Permissive mode is default mode..it sets null value to all corrupted fields

10. Now we'll do hands-on

```
flight_df_header_schema = spark.read.format("csv")\
    .option("header", "true")\
    .option("inferSchema", "true")\
    .option("mode", "FAILFAST")\
    .load("/FileStore/tables/flight_data.csv")

flight_df_header.show(5)
```

11.

12. Notes on inferSchema : <https://q.co/gemini/share/43be33546bfa>

13.

community.cloud.databricks.com/?o=4633718999277322#notebook/2840863557306816/command/2840863557306819

read\_csv\_file\_in\_spark Python

File Edit View Run Help Last edit was 3 hours ago New cell UI: OFF

Run all Terminated Share Publish

United States	Saint Maarten	390
Malta	United States	1
Bolivia	United States	46
Anguilla	United States	21
Turks and Caicos ...	United States	136
United States	Afghanistan	2
Saint Vincent and...	United States	1
Italy	United States	390
United States	Russia	156

only showing top 20 rows

Command took 1.54 seconds -- by kaushikvarma958@gmail.com at 3/8/2024, 4:21:33 PM on My Cluster

```
1 flight_df = spark.read.format("csv")\
2     .option("header", "false")\
3     .option("inferSchema", "true")\
4     .option("mode", "FAILFAST")\
5     .load("/FileStore/tables/Flight_data_2010.csv")
6 flight_df.show()
```

3) Spark Jobs

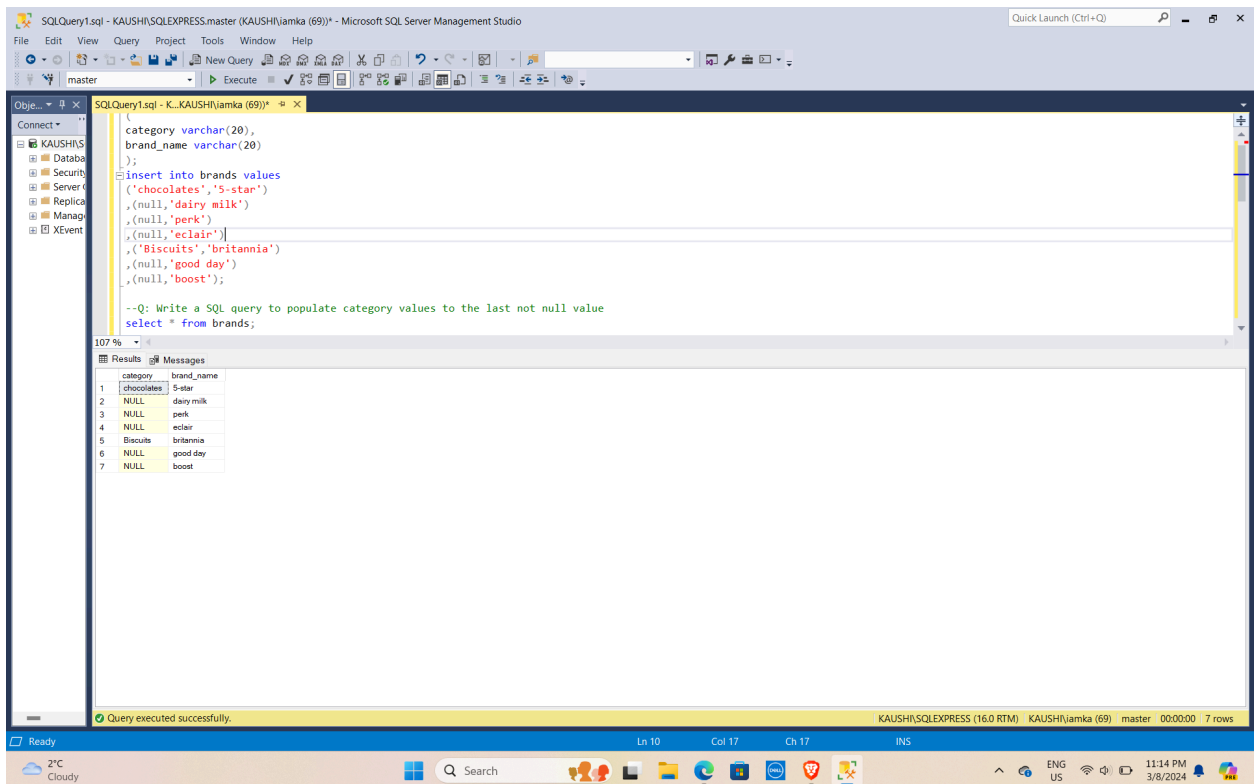
flight\_df: pyspark.sql.dataframe.DataFrame = [c0: string, \_c1: string ... 1 more field]

United States	Ireland	264
United States	India	69
Egypt	United States	24
Equatorial Guinea	United States	1
United States	Singapore	25
United States	Grenada	54
Costa Rica	United States	477
Senegal	United States	29
United States	Marshall Islands	44
Guyana	United States	17
United States	Sint Maarten	53
Malta	United States	1

5. U can refer my day1\_notes here :

-<https://docs.google.com/document/d/1bnS-i3sx1IRpYaxe8LG-Fr59hvpOtNe4LsFqKPA0oKY/edit?usp=sharing>

## 6. Ended my day by solving one complex SQL question



SQLQuery1.sql - K:\KAUSHI\iamka (69) - Microsoft SQL Server Management Studio

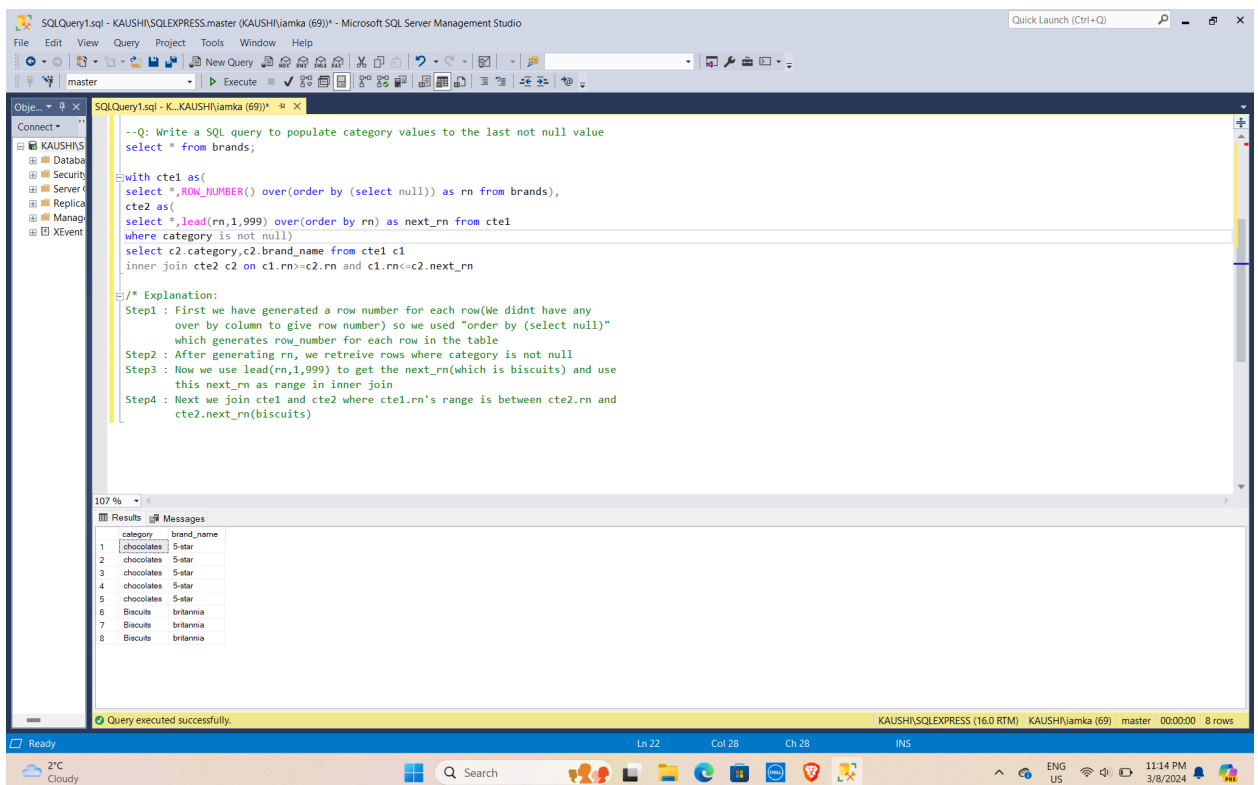
```
category varchar(20),
brand_name varchar(20)
);
insert into brands values
('chocolates','5-star')
,(null,'dairy milk')
,(null,'perk')
,(null,'eclair')
,('Biscuits','britannia')
,(null,'good day')
,(null,'boost');

--Q: Write a SQL query to populate category values to the last not null value
select * from brands;
```

Results

category	brand_name
chocolates	5-star
NULL	dairy milk
NULL	perk
NULL	eclair
Biscuits	britannia
NULL	good day
NULL	boost

Query executed successfully.



SQLQuery1.sql - K:\KAUSHI\iamka (69) - Microsoft SQL Server Management Studio

```
--Q: Write a SQL query to populate category values to the last not null value
select * from brands;

with cte1 as(
select *,ROW_NUMBER() over(order by (select null)) as rn from brands),
cte2 as(
select *,lead(rn,1,999) over(order by rn) as next_rn from cte1
where category is not null)
select c2.category,c2.brand_name from cte1 c1
inner join cte2 c2 on c1.rn=c2.rn and c1.rn<=c2.next_rn

/* Explanation:
Step1 : First we have generated a row number for each row(We didnt have any
over by column to give row number) so we used "order by (select null)"
which generates row_number for each row in the table
Step2 : After generating rn, we retrieve rows where category is not null
Step3 : Now we use lead(rn,1,999) to get the next_rn(which is biscuits) and use
this next_rn as range in inner join
Step4 : Next we join cte1 and cte2 where cte1.rn's range is between cte2.rn and
cte2.next_rn(biscuits)
```

Results

category	brand_name
chocolates	5-star
chocolates	5-star
chocolates	5-star
chocolates	5-star
chocolates	5-star
Biscuits	britannia
Biscuits	britannia
Biscuits	britannia

Query executed successfully.