

Day33 - March 9th 2024

1. Started my day and completed all my activities
2. Solved some leetcode easy questions on arrays,Hashmaps,strings

3. Started learning spark practically from manish kumar

(1) schema in spark | Lec-4 - x

Manish Kumar

YouTube

manish kumar

Possible interview questions

- how to create schema in PySpark?
- what are other ways to creating it?
- what is StructField and StructType in schema?
- what if I have header in my data.

Schema

Struct Type  
Struct Field

DDL

spark practical (DataFrame API)

MANISH KUMAR · 5 / 25

- Read data in spark | Lec-3 | Read modes in spark
- schema in spark | Lec-4
- Handling corrupted records in spark | PySpark | Databricks
- how to read json file in pyspark

docs.google.com/document... x

Spark\_practical\_day2

File Edit View Insert Format Tools Extensions ...

100% Normal text Arial

DEST_COUNTRY_NAME	ORIGIN_COUNTRY_NAME	count
United States	Saint Martin	2

Here we have a null value in the count column and we have defined count column as non nullable, so failfast mode is treating this as malformed

We have defined myschema which includes the header and when reading a file..if we choose .option("header",false) then it gives result like this..because dataframe api is reading header from myschema

```
flight_df.spark.read.format("csv")\n    .option("header","false")\n    .option("skipRows",1)\n    .option("inferSchema","false")\n    .schema(my_schema)\n    .load("/filestore/tables/flight_data.csv")\n\nflight_df.show(5)
```

To avoid that we use ".option('skipRows',1)..It skips the first row in the file

So in the end ..we have answered this four questions in the schema

how to | Gemini

Can the v | CSVs

Shirdi Sai | +

how to | Gemini

Can the v | CSVs

Shirdi Sai | +

manish kumar

spark fundamental (Theory)

Introduction

What is spark

Flatten Nested Json in spark

Lec-23

23:55 / 24:01

spark practical (DataFrame API)

MANISH KUMAR · 7 / 25

1 Introduction

2 how to create databricks community edition account

3 databricks community edition overview

4 Read data in spark | Lec-3 | Read modes in spark

docs.google.com/document...

Spark\_practical\_day2

File Edit View Insert Format Tools Extensions ...

100%

Normal text

Arial

7

7. Here I have uploaded multiple types of json files

```
1 File uploaded to /FileStore/tables/corrupted.json
2 File uploaded to /FileStore/tables/employee-1.json
3 File uploaded to /FileStore/tables/Multi_line_correct.json
4 File uploaded to /FileStore/tables/Multi_line_incorrect.json
5 File uploaded to /FileStore/tables/single_file_json_extra_fields.
  json
```

8. First we will use line\_delimited\_json

```
{ "name": "Manish", "age": 20, "salary": 20000 },
{ "name": "Nikita", "age": 25, "salary": 21000 },
{ "name": "Pritam", "age": 16, "salary": 22000 },
{ "name": "Prantosh", "age": 35, "salary": 25000 },
{ "name": "Vikash", "age": 67, "salary": 40000 }
```

9. Refer :

<https://www.youtube.com/watch?v=M0Kx205dxmM&list=PLTnSGelpGnGjaMSYVlduVWsjkWoRbhr&index=7>

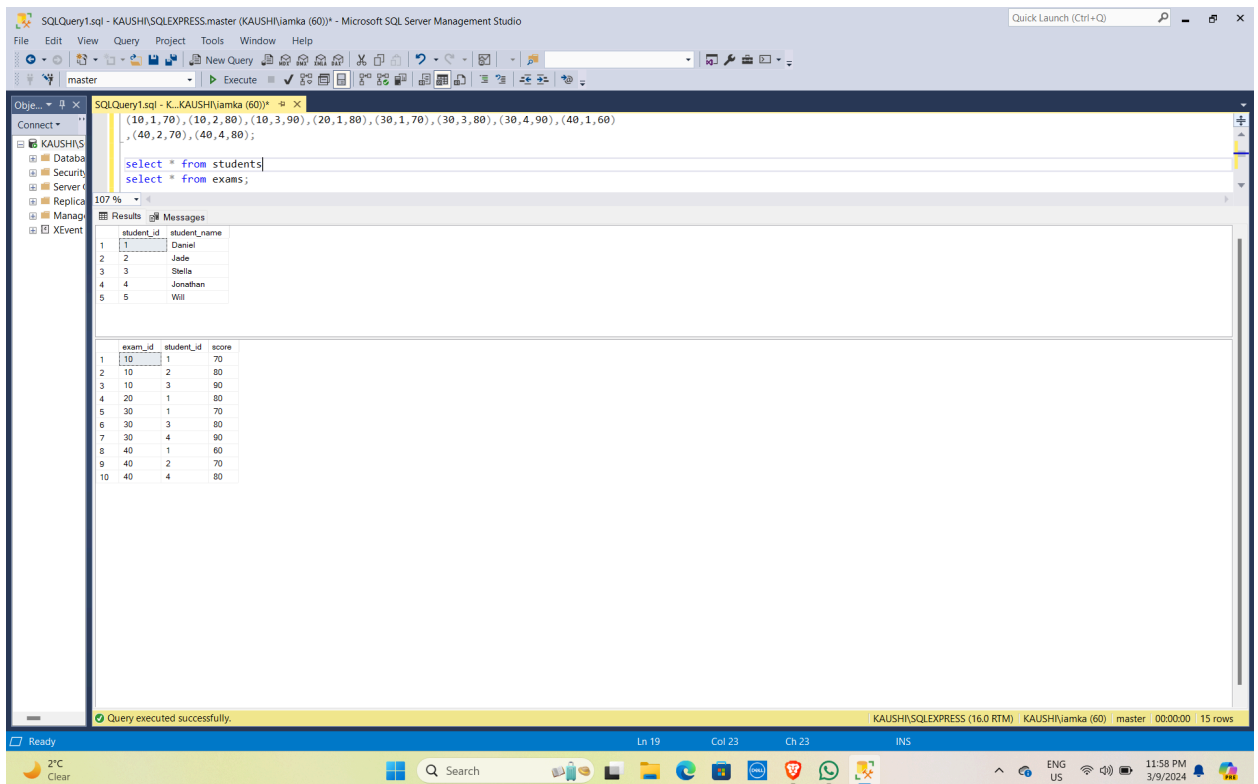
10. Line delimited vs multi line : <https://q.c/gemini/share/0019d1dfc26f>

11.

4. You can find more :

[https://docs.google.com/document/d/1tAOrNPJQXc4sRC9SJeFs6Q8Z-qdQMSR\\_I-laETixAul/edit?usp=sharing](https://docs.google.com/document/d/1tAOrNPJQXc4sRC9SJeFs6Q8Z-qdQMSR_I-laETixAul/edit?usp=sharing)

## 5. Ended my day by solving a complex SQL question



SQLQuery1.sql - K:\KAUSHI\iamka (60)\* - Microsoft SQL Server Management Studio

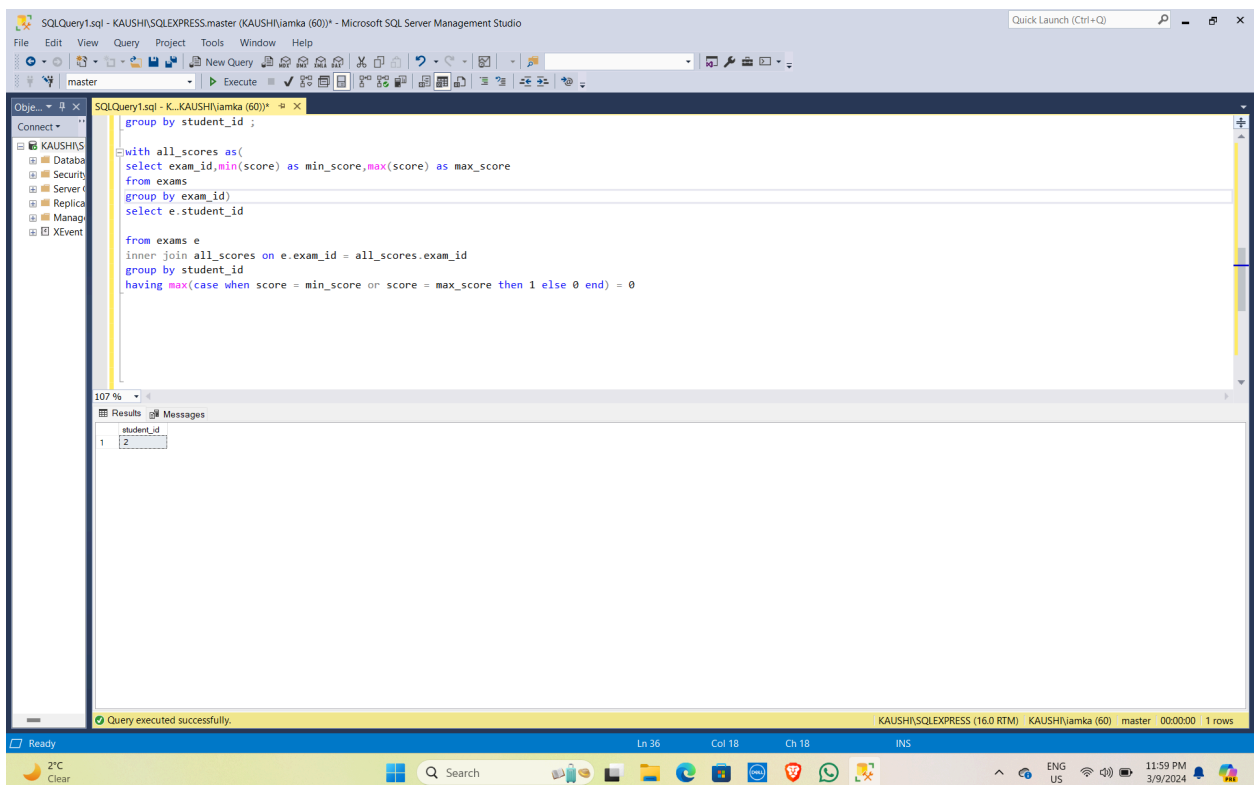
```
(10,1,70),(10,2,80),(10,3,90),(20,1,80),(30,1,70),(30,3,80),(30,4,90),(40,1,60),
(40,2,70),(40,4,80);

select * from students
select * from exams;
```

student_id	student_name
1	Daniel
2	Jade
3	Stella
4	Jonathan
5	Will

exam_id	student_id	score
10	1	70
10	2	80
10	3	90
20	1	80
30	1	70
30	3	80
30	4	90
40	1	60
40	2	70
40	4	80

Query executed successfully.



SQLQuery1.sql - K:\KAUSHI\iamka (60)\* - Microsoft SQL Server Management Studio

```
group by student_id ;

with all_scores as(
select exam_id,min(score) as min_score,max(score) as max_score
from exams
group by exam_id)
select e.student_id

from exams e
inner join all_scores on e.exam_id = all_scores.exam_id
group by student_id
having max(case when score = min_score or score = max_score then 1 else 0 end) = 0
```

student_id
2

Query executed successfully.

## 6.