

Day37 - March 13th 2024

1. Started my day as usual
2. Cooked food for friends and packed it in box and headed to library
3. Started a leetcode problem and solved it in 1.5hr

4. Learning spark theory from manish kumar a youtuber

The image displays a YouTube video player on the left and a Google Docs document on the right, both showing content related to Spark theory.

YouTube Video: The video is titled "spark fundamental (Theory)" by MANISH KUMAR. The video player shows a man speaking, with handwritten notes "JSON, YAML" and "Semistructure" overlaid. Below the video, a playlist is visible with the following items:

- 1. spark series introduction (3:21) MANISH KUMAR
- 2. spark overview | lec-1 (14:09) MANISH KUMAR
- 3. Why Apache spark | Lec-2 (13:43) MANISH KUMAR
- 4. Hadoop vs Spark (13:43) MANISH KUMAR

Google Docs Document: The document is titled "Spark Theory Day1". It contains the following text:

11. From this we've come to know what actually spark does

Why Apache spark?

1. So initially we have multiple databases like

Database → Oracle, Teradata, exoddata, MySQL

and they used to store only relational data.

2. But these days..we are getting data like text,CSV,image and in many other formats
- 3.
- 4.

5.

6.

16. we have executors containers..which are given by YARN as we requested

executor container
↳ w2, w3, w4, w7, w8

Application master container

Python main() → JVM main() → Application Driver

e1 JVM 5 core CPU & 5GB RAM
e2 JVM
e3 JVM
e4 JVM
e5 JVM

17.

18. Now what if we write a python UDF? Our executors will not understand that

19. Lets take a small example and understand

Application master

JVM main() → e1 JVM Python linker
e2 JVM Python worker

5. Please find my doc here :

<https://docs.google.com/document/d/1sr0UGbwFZSRJ5JDUN7BQubTIEg9PQ2XPcFcSYx3oe4U/edit?usp=sharing>

6. Ended my day by solving a complex problem from Ankit's YT

The screenshot shows the Microsoft SQL Server Management Studio interface. The query editor contains the following SQL code:

```
create table tbl_orders (
    order_id integer,
    order_date date
);
insert into tbl_orders
values (1,'2022-10-21'),(2,'2022-10-22'),
(3,'2022-10-25'),(4,'2022-10-25');

select * from tbl_orders

select * into tbl_orders_copy from tbl_orders --Creating a snapshot of the table

insert into tbl_orders
values (5,'2023-10-21'),(6,'2023-10-22'); --inserting two new records

delete from tbl_orders where order_id =1 --deleting a record from the original table

select * from tbl_orders_copy
select * from tbl_orders
```

The Results pane shows the output of the queries:

order_id	order_date
1	2022-10-21
2	2022-10-22
3	2022-10-25
4	2022-10-25

order_id	order_date
2	2022-10-22
3	2022-10-25
4	2022-10-25
5	2023-10-21
6	2023-10-22

The status bar at the bottom indicates "Query executed successfully."

The screenshot shows the Microsoft SQL Server Management Studio interface. The query editor contains the following SQL code:

```
select coalesce(t1.order_id,t2.order_id) as order_id,
case when t2.order_id is Null then 'I'
when t1.order_id is Null then 'D'
end as flag
from tbl_orders t1
full outer join tbl_orders_copy t2 on t1.order_id = t2.order_id
where t1.order_id is Null or t2.order_id is Null

/* Explanation:
Step1 : Initially Given a table with sample data of orders with 4 rows
Step2 : Created a snapshot of the table to store a copy of table version
Step3 : Inserted two new rows in the original table and deleted one row from
the original table
Step4 : Task is to find the delta's of these two tables
Step5 : Used full outer join to get the details from the both the tables
Step6 : Now if a t2.order_id null then it indicates that we have inserted rows
Step7 : And if t1.order_id null then it indicates that we've deleted a row
Step8 : Just see the code and get intuition
```

The Results pane shows the output of the query:

order_id	flag
5	I
6	I
1	D

The status bar at the bottom indicates "Query executed successfully."

7.