

Day34 - March 10th 2024

1. Started my day same as yesterday
2. Cooked food and headed to library
3. Started solving leetcode problems will share my leetcode id on github for proof
4. Learning spark by practically implementing the code

The screenshot shows a Google Docs interface with the document titled "Spark_Practical_day3". The document content includes a heading "How to write data in spark" followed by a section titled "Potential interview question :-". Below this, there is a list of three questions:

- ① what are the modes available in dataframe writer?
- ② what is partitionBy and bucketBy?
- ③ How to write data into multiple partition?

Below the list, the text "1." is visible. Further down, the text "2. Dataframe write general structure" is visible, followed by a partially obscured heading "Dataframe write general structure". The document is viewed in a browser window with multiple tabs open, including "Shirdi Sai Bai", "GDB online", "Feed | LinkedIn", "Spark_Pr", "Read_Parc", "Netflix", "Walmart Lab", "SQL Server D", "Get only the", "date from di", and "concatenati". The Windows taskbar at the bottom shows the date as 3/10/2024 and the time as 11:07 PM.

youtube.com/watch?v...

manish kumar

write_file_in_spark

Command took 2.82 seconds -- by manishkumar@gmail.com at 5/4/2023, 18:10:28 AM on My Cluster

```
1 df.write.format("csv")\n2   .option("header","true")\n3   .option("header","overwrite")\n4   .option("path","FileStore/tables/csv_write")\n5   .save()
```

1

Next: Partitioning and bucketing in Spark | Lec-9 | Practical video
spark practical (DataFrame API) - 9 / 25

How to write dataframe to disk in spark | Lec-8

MANISH KUMAR
14.2K subscribers

Subscribe

129 4 Share

6,196 views May 6, 2023 spark practical (DataFrame API)
In this video I have talked about how can you write your transformed dataframe onto disk in spark. Please do ask your doubts in comment section.
Directly connect with me on: https://topmate.io/manish_kumar25

databricks

Read_Parquet_File&write file Python

File Edit View Run Help Last e... New cell UE OFF

111 Ragu 12.0 35000.0 INDIA f
112 Sweta 43.0 200000.0 INDIA f
113 Raushan 48.0 650000.0 USA =
114 Mukesh 36.0 95000.0 RUSSIA =
115 Prakash 52.0 750000.0 INDIA =

Command took 19.18 seconds -- by kaushikvarma95@gmail.com at 3/18/2024, 11:42:58 PM on My Cluster

Cmd 6

```
1 #write\n2 df.write.format("csv")\n3   .option("header","true")\n4   .option("mode","overwrite")\n5   .option("path","FileStore/tables/csv_write")\n6   .save()
```

1

[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts

youtube.com/watch?v...

manish kumar

write_file_in_spark

Command took 0.31 seconds -- by manishkumar@gmail.com at 5/4/2023, 18:22:33 AM on My Cluster

```
1 dbutils.fs.ls("/FileStore/tables/csv_write_repartition/")\n\nOut[10]: [FileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/_SUCCESS', name='_SUCCESS', size=0, modificationTime=1633779310000),\nFileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/_committed_248503806878423836', name='_committed_248503806878423836', size=285, modificationTime=1633779310000),\nFileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/_started_248503806878423836', name='_started_248503806878423836', size=0, modificationTime=1633779310000),\nFileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/part-00000-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-5-1-c000.csv', name='part-00000-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-5-1-c000.csv', size=184, modificationTime=1710132556000),\nFileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/part-00001-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-6-1-c000.csv', name='part-00001-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-6-1-c000.csv', size=184, modificationTime=1710132556000),\nFileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/part-00002-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-7-1-c000.csv', name='part-00002-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-7-1-c000.csv', size=190, modificationTime=1710132556000)]
```

1

Next: Partitioning and bucketing in Spark | Lec-9 | Practical video
spark practical (DataFrame API) - 9 / 25

How to write dataframe to disk in spark | Lec-8

MANISH KUMAR
14.2K subscribers

Subscribe

129 4 Share

6,196 views May 6, 2023 spark practical (DataFrame API)
In this video I have talked about how can you write your transformed dataframe onto disk in spark. Please do ask your doubts in comment section.
Directly connect with me on: https://topmate.io/manish_kumar25

databricks

Read_Parquet_File&write file Python

File Edit View Run Help Last e... New cell UE OFF

Cluster

Cmd 9

```
1 #to check the file\n2 dbutils.fs.ls("/FileStore/tables/csv_write_repartition/")
```

Out[6]: [FileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/_SUCCESS', name='_SUCCESS', size=0, modificationTime=1710132556000),\nFileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/_committed_8548821234069270882', name='_committed_8548821234069270882', size=285, modificationTime=1710132556000),\nFileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/_started_8548821234069270882', name='_started_8548821234069270882', size=0, modificationTime=1710132556000),\nFileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/part-00000-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-5-1-c000.csv', name='part-00000-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-5-1-c000.csv', size=184, modificationTime=1710132556000),\nFileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/part-00001-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-6-1-c000.csv', name='part-00001-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-6-1-c000.csv', size=184, modificationTime=1710132556000),\nFileInfo(path='dbfs:/FileStore/tables/csv_write_repartition/part-00002-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-7-1-c000.csv', name='part-00002-tid-8548821234069270882-b9915248-989c-4148-bdab-743a384ac5ab-7-1-c000.csv', size=190, modificationTime=1710132556000)]

1

[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts

5. Ended my by solving a complex SQL question from Ankit's YT

SQLQuery1.sql - KAUSHI\SQLEXPRESS.master (KAUSHI\iamka (54)) - Microsoft SQL Server Management Studio

File Edit View Query Project Tools Window Help

Connect to server: master

Object Explorer: KAUSHI\SQLEXPRESS (16.0 RTM) - KAUSHI\iamka (54) - master

```

with cte as (
select Callerid,Recipientid,concat(month(datecalled),'-',day(datecalled)) as dat,row_number() over(partition by callerid,concat(month(datecalled),'-',day(datecalled))
order by Callerid,datecalled) as rn1 from phonelog )
--select * from cte
,cte2 as (
select callerid,dat,min(rn1) as mini,max(rn1) as maxi
from cte
group by callerid,dat)
select c1.callerid,c1.recipientid from cte c1
inner join cte2 c2 on c1.Callerid = c2.Callerid and c1.dat = c2.dat
and (c1.rn1 = c2.mini or c1.rn1 = c2.maxi)
group by c1.Callerid,c1.Recipientid
having count(1) > 1;

```

/* Explanation:
Step1 : First I have retrieved date and month and gave a row_number
partitioned by caller_id and dat and framed it in one cte
Step2 : Now I have retrieved min of rn and max of rn for each caller_id
on a given date(min of rn gives us the first call and max of rn
gives us the last call made by callerid) and framed in cte2
Step3 : Now I have joined cte1 and cte2 on inner join cte2 c2 on
c1.Callerid = c2.Callerid and c1.dat = c2.dat
and (c1.rn1 = c2.mini or c1.rn1 = c2.maxi)
Step4 : Next group by on caller_id and recipient_id, now if a caller_id
made first and last call to same recipient..then their count would be
two and we retrieve them.

Results

	callerid	recipientid
1	2	4
2	2	5

Query executed successfully.

KAUSHI\SQLEXPRESS (16.0 RTM) KAUSHI\iamka (54) master 00:00:00 2 rows

Ready 10°C Clear

Ln 42 Col 21 Ch 21

11:28 PM 3/10/2024

6.