Abhishek_MandaotryAssignment_3

June 25, 2024

```
[1]: // MySQL database connection properties
     val jdbcUrl = "jdbc:mysql://34.100.169.199:3306/project-2"
     val connectionProperties = new java.util.Properties()
     connectionProperties.put("user", "root")
     connectionProperties.put("password", "abhi@iit123")
     connectionProperties.put("driver", "com.mysql.cj.jdbc.Driver")
    VBox()
    Starting Spark application
    <IPython.core.display.HTML object>
    FloatProgress(value=0.0, bar_style='info', description='Progress:',u
     ⇔layout=Layout(height='25px', width='50%'),...
    SparkSession available as 'spark'.
    FloatProgress(value=0.0, bar_style='info', description='Progress:',u
     ⇒layout=Layout(height='25px', width='50%'),...
    jdbcUrl: String = jdbc:mysql://34.100.169.199:3306/project-2
    connectionProperties: java.util.Properties = {}
    res2: Object = null
    res3: Object = null
    res4: Object = null
[2]: def writeToMySQL(df: org.apache.spark.sql.DataFrame, tableName: String): Unit = {
       df.write.format("jdbc").option("url", jdbcUrl).option("dbtable", tableName).
      ⇔option("user", "root").option("password", "abhi@iit123").mode("overwrite").
     ⇒save()
            .mode(SaveMode.Overwrite)
     //
            .jdbc(jdbcUrl, tableName, connectionProperties)
     //
     // totalBenzeneDF.write.format("jdbc").options(
     // Map(
     // "url" -> mysqlUrl,
     // "user" -> mysqlUser,
     // "password" -> mysqlPassword,
```

```
// "dbtable" -> "total_benzene_df"
    // )).mode("append").save()
    VBox()
    FloatProgress(value=0.0, bar_style='info', description='Progress:',_
     ⇔layout=Layout(height='25px', width='50%'),...
    writeToMySQL: (df: org.apache.spark.sql.DataFrame, tableName: String)Unit
[3]: spark
    VBox()
    FloatProgress(value=0.0, bar_style='info', description='Progress:',u
     ⇔layout=Layout(height='25px', width='50%'),...
    res14: org.apache.spark.sql.SparkSession =
    org.apache.spark.sql.SparkSession@7997eeeb
[4]: // S3 paths
    val inputS3Path = "s3://task-3-emr/air-quality-data/"
    val outputParquetS3Path = "s3://task-3-emr/air-quality-data-parquet/"
    VBox()
    FloatProgress(value=0.0, bar_style='info', description='Progress:', u
     inputS3Path: String = s3://task-3-emr/air-quality-data/
    outputParquetS3Path: String = s3://task-3-emr/air-quality-data-parquet/
[5]: val airQualityDF = spark.read.option("header", "true").option("inferSchema", __

¬"true").csv(inputS3Path)
    VBox()
    FloatProgress(value=0.0, bar_style='info', description='Progress:',u
     ⇔layout=Layout(height='25px', width='50%'),...
    airQualityDF: org.apache.spark.sql.DataFrame = [From Date: string, To Date:
    string ... 23 more fields]
[6]: airQualityDF.count()
    VBox()
    FloatProgress(value=0.0, bar_style='info', description='Progress:',u
     ⇔layout=Layout(height='25px', width='50%'),...
    res16: Long = 14296116
```

```
[7]: airQualityDF.write.format("parquet").mode("overwrite").save(outputParquetS3Path)
   VBox()
   FloatProgress(value=0.0, bar_style='info', description='Progress:',u
    →layout=Layout(height='25px', width='50%'),...
[8]: val parquetDF = spark.read.parquet(outputParquetS3Path)
   VBox()
   FloatProgress(value=0.0, bar_style='info', description='Progress:',_
    →layout=Layout(height='25px', width='50%'),...
   parquetDF: org.apache.spark.sql.DataFrame = [From Date: string, To Date: string
   ... 23 more fields]
[9]: // Job 1: Filter by date range
   val filterDF = parquetDF.filter($"From Date" >= "2023-01-01" && $"To Date" <= __
    →"2023-12-31")
   filterDF.show()
   writeToMySQL(filterDF, "filtered_by_date")
   VBox()
   FloatProgress(value=0.0, bar_style='info', description='Progress:',u
    →layout=Layout(height='25px', width='50%'),...
   filterDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [From Date:
   string, To Date: string ... 23 more fields]
   +-----
   ____+__
   ---+---+
                           To Date | PM2.5 (ug/m3) | PM10 (ug/m3) | NO
           From Date
   (ug/m3)|NO2 (ug/m3)|NOx (ppb)|NH3 (ug/m3)|SO2 (ug/m3)|CO (mg/m3)|Ozone
   (ug/m3)|Benzene (ug/m3)|Toluene (ug/m3)|Eth-Benzene (ug/m3)|MP-Xylene (ug/m3)|O
   Xylene (ug/m3)|Temp (degree C)|RH (%)|WS (m/s)|WD (deg)|SR (W/mt2)|BP (mmHg)|VWS
   (m/s)|Xylene (ug/m3)|AT (degree C)|
         ____+__
   ______
   ---+----+
   |2023-02-10 09:00:00|2023-02-10 10:00:00|
                                        40.0|
                                                 139.5
                                                          2.45
   7.95
                                     0.31
          6.25
                  19.45
                           14.55
                                              61.85
                                                           0.1
   0.21
                 NULL
                              NULL
                                           NULL
                                                       56.0|
                                                  NULL
   1.85
         239.5
               219.0
                        NULL
                               25.65
                                        0.01
   NULL|
```

12023-	-02-10 10:	00:00 2023-02	2-10 11:	00:00	0.00	18.75	24 0	92.25	1.95
5.881	4.671	21.62	1.	5.1	0.29	, at	64.9	21	0.01
		NULL							
		236.5	NULLI	26.18	I	0.01		NULLI	
NULL						40.051		00 51	4 501
		00:00 2023-02							
5.351	4.281	23.6	14	.18	0.27	1	66.1	71	0.01
0.21		NULL 264.5	1	NULLI		N	ULLI	1	53.25
		264.51	NULLI	26.9	I	0.01		NULLI	
NULL									
		00:00 2023-02							
5.0	3.97	24.45	13.	05	0.26		70.33	1	0.01
0.2		NULL 175.75		NULL		N	ULL		48.5
		175.75	NULL	27.55	I	0.01		NULL	
NULL									
12023-	02-10 13:	00:00 2023-02	2-10 14:	00:00		18.5		93.5	1.62
5.42	4.23	24.6	11	.62	0.27	1	71.4	7	0.0
0.2		NULL		NULL		N	ULL		45.0
1.95	248.0	266.25	NULL	28.0	l	0.0		NULL	
NULL									
12023-	-02-10 14:	00:00 2023-02	2-10 15:	00:00		12.0		83.0	1.55
4.47	3.62	25.93	8	.72	0.27	1	73.	6	0.0
0.2		NULL		NULL		N	ULL		40.75
1.77	237.75	235.25	NULL	28.45	1	0.01		NULL	
NULL									
2023-	02-10 15:	00:00 2023-02	2-10 16:	00:00		11.75		84.0	1.45
4.75	3.67	27.02		7.2	0.31	1	76.9	2	0.0
0.2		NULL		NULL		N	ULL		39.0
1.73	241.5	185.5	NULL	28.67	I	0.01		NULL	
NULL									
2023-	-02-10 16:	00:00 2023-02	2-10 17:	00:00		10.25		86.75	1.58
4.7	3.77	24.95	7	.7	0.32		79.83	1	0.0
0.2		24.95 NULL		NULL		N	ULL		39.5
		126.0							
NULL									
2023-	-02-10 17:	00:00 2023-02	2-10 18:	00:00		13.5		84.0	1.6
5.75	4.33	23.18	12	.12	0.33	1	80.4	2	0.0
		NULL							
		73.0							
NULL									
12023-	-02-10 18:	00:00 2023-02	2-10 19:	00:00		20.0	1	01.75	1.48
		21.72		1.38					
50.01	0.551	0.2 201.5 53.7	75 l	NULL I	27.5	31	0.01		NULL I
NULL	•		•	•		-	- •		•
	-02-10 19:	00:00 2023-02	2-10 20:	00:001		22.01		132.0	1.3
		19.45							- •
0.03		0.3		0.03	_		0.1		NULL
- •		•		•			•		•

NULL		200.75	53.0	NULL	26.95	0.0	NULL
12023-0	02-10 20	:00:00 20)23-02-10	21:00:00	24.5	5 136.0	2.15
49.15	27.8	88	20.4	6.93	0.83	21.0	
0.12		0.75		0.38		0.45	0.1
55.0	0.35	216.0	53.0	NULL	26.58	0.0	NULL
NULL							
12023-0	02-10 21	:00:00 20)23-02-10	22:00:00	41.75	5 226.25 49.97	1.42
22.15	12.9	3 2	20.15	6.7	0.48	49.97	
0.12		1.82		1.05		1.05 0.0	0.3
54.75	0.4	140.0	53.25	NULL	26.4	0.0	NULL
NULL							
12023-0	02-10 22	2:00:00 20	23-02-10	23:00:00	23.79	5 95.75	2.05
						37.43	
0.01		0.45		0.05		0.1	NULL
63.51	0.1	97.75 l	53.75	NULL I	25.53	0.1 0.0	NULL I
NULL				•			•
		8:00:00120)23-02-11	00:00:001	19.2	5 94.75	2.271
	23 5	51 1	8 17 l	4 75 l	0 641	15 221	
0.08	20.0	0 571	.0.11	0.11	0.01	0.12	0.031
	0 151	130 251	54 01	NIII I I	25 051	0.0	NIII I
NULL		100.201	01.01	NODE	20.001	0.01	NODL
) <u>03</u> _00_11	01.00.001	27	5 129.25	5 051
10 651	30 5	21 1	10 30 I	3 501	0 961	5.83	3.031
						0.42	
75 01	A 1 I	06 751	E4 01	ATTTT T I	04 721	0 01	ATTIT T I
	0.1	96.75	54.0	NULL	24.73	0.0	NULL
NULL							
NULL 2023-0	02-11 01	:00:00 20)23-02-11	02:00:00	44.0	0 150.25	4.85
NULL 2023-0 39.47	02-11 01 24.9	:00:00 20)23-02-11 20.1	02:00:00	44.(0.58	0 150.25 19.35	4.85
NULL 2023-0 39.47	02-11 01 24.9	:00:00 20)23-02-11 20.1	02:00:00	44.(0.58	0 150.25 19.35	4.85
NULL 2023-0 39.47 0.1 74.75	02-11 01 24.9 0.27	:00:00 20)23-02-11 20.1	02:00:00	44.(0.58	0 150.25	4.85
NULL 2023-0 39.47 0.1 74.75 NULL	02-11 01 24.9 0.27	:00:00 20 2 1.23 104.5	023-02-11 20.1 54.0	02:00:00 3.1 0.2 NULL	44.0 0.58 (24.73	0 150.25 19.35 0.57 0.0	4.85 0.2 NULL
NULL 2023-0 39.47 0.1 74.75 NULL 2023-0	02-11 01 24.9 0.27 02-11 02	:00:00 20 2 1.23 104.5	023-02-11 20.1 54.0 023-02-11	02:00:00 3.1 0.2 NULL 03:00:00	44.0 0.58 24.73 27.29	0 150.25 19.35 0.57 0.0 5 98.25	4.85 0.2 NULL
NULL 2023-(39.47 0.1 74.75 NULL 2023-(15.78 1	02-11 01 24.9 0.27 02-11 02 9.5	::00:00 20 2 1.23 104.5 ::00:00 20	23-02-11 20.1 54.0 023-02-11 20.43	02:00:00 3.1 0.2 NULL 03:00:00 2.92	44.0 0.58 (24.73 27.29	0 150.25 19.35 0.57 0.0 5 98.25 32.75	4.85 0.2 NULL 1.45
NULL 2023-(39.47 0.1 74.75 NULL 2023-(15.78 1	02-11 01 24.9 0.27 02-11 02 9.5	::00:00 20 2 1.23 104.5 ::00:00 20	23-02-11 20.1 54.0 023-02-11 20.43	02:00:00 3.1 0.2 NULL 03:00:00 2.92	44.0 0.58 (24.73 27.29	0 150.25 19.35 0.57 0.0 5 98.25 32.75	4.85 0.2 NULL 1.45
NULL 2023-0 39.47 0.1 74.75 NULL 2023-0 15.78 0.03 78.5	02-11 01 24.9 0.27 02-11 02 9.5	::00:00 20 2 1.23 104.5 ::00:00 20	23-02-11 20.1 54.0 023-02-11 20.43	02:00:00 3.1 0.2 NULL 03:00:00 2.92	44.0 0.58 (24.73 27.29	0 150.25 19.35 0.57 0.0 5 98.25 32.75	4.85 0.2 NULL 1.45
NULL 2023-0 39.47 0.1 74.75 NULL 2023-0 15.78 0.03 78.5 NULL	02-11 01 24.9 0.27 02-11 02 9.5	1.23 1.23 104.5 2:00:00 20 7 2.38 76.0	23-02-11 20.1 54.0 023-02-11 20.43 53.5	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL	44.0 0.58 24.73 27.28 0.33 24.38	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0	4.85 0.2 NULL 1.45 0.03 NULL
NULL 2023-(39.47 0.1 74.75 NULL 2023-(15.78 0.03 78.5 NULL 2023-(120	02-11 01 24.9 0.27 02-11 02 9.5 0.1	1.23 1.23 104.5 2:00:00 20 37 0.38 76.0 3:00:00 20	23-02-11 20.1 54.0 023-02-11 20.43 53.5	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL 04:00:00	44.0 0.58 24.73 27.29 0.33 24.38 18.29	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0	4.85 0.2 NULL 1.45 0.03 NULL
NULL 2023-0 39.47 0.1 74.75 NULL 2023-0 15.78 0.03 78.5 NULL 2023-0 36.67	02-11 01 24.9 0.27 02-11 02 9.5 0.1 02-11 03 23.	::00:00 20 :2 1.23 104.5 ::00:00 20 :7 0.38 76.0 ::00:00 20 3	023-02-11 20.1 54.0 023-02-11 20.43 53.5 023-02-11 21.8	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL 04:00:00 2.7	44.0 0.58 24.73 27.28 0.33 24.38 18.28 0.38	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0 5 69.0 11.58	4.85 0.2 NULL 1.45 0.03 NULL 4.67
NULL 2023-(39.47 0.1 74.75 NULL 2023-(36.67 0.05 0.05	02-11 01 24.9 0.27 02-11 02 9.5 0.1 02-11 03 23.	1.23 1.23 104.5 2:00:00 20 37 2:038 76.0 3:00:00 20 3 0.45	23-02-11 20.1 54.0 023-02-11 20.43 53.5 023-02-11 21.8	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL 04:00:00 2.7 0.03	44.0 0.58 24.73 27.29 0.33 24.38 18.29 0.38	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0 5 69.0 11.58 0.12	4.85 0.2 NULL 1.45 0.03 NULL 4.67 NULL
NULL 2023-(39.47 0.1 74.75 NULL 2023-(36.67 0.05 0.05	02-11 01 24.9 0.27 02-11 02 9.5 0.1 02-11 03 23.	1.23 1.23 104.5 2:00:00 20 37 2:038 76.0 3:00:00 20 3 0.45	23-02-11 20.1 54.0 023-02-11 20.43 53.5 023-02-11 21.8	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL 04:00:00 2.7 0.03	44.0 0.58 24.73 27.29 0.33 24.38 18.29 0.38	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0 5 69.0 11.58	4.85 0.2 NULL 1.45 0.03 NULL 4.67 NULL
NULL 2023-(39.47 0.1 74.75 NULL 2023-(15.78 0.03 78.5 NULL 2023-(36.67 0.05 82.25 NULL	02-11 01 24.9 0.27 02-11 02 9.5 0.1 02-11 03 23. 0.15	::00:00 20 2 1.23 104.5 ::00:00 20 7	23-02-11 20.1 54.0 023-02-11 20.43 53.5 023-02-11 21.8 53.0	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL 04:00:00 2.7 0.03 NULL	44.0 0.58 24.73 27.28 0.33 24.38 18.28 0.38 24.1	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0 5 69.0 11.58 0.12 0.0	4.85 0.2 NULL 1.45 0.03 NULL 4.67 NULL NULL
NULL 2023-(39.47 0.1 74.75 NULL 2023-(36.67 0.05 82.25 82.25 NULL 2023-(36.67 0.05 82.25 82.	02-11 01 24.9 0.27 02-11 02 9.5 0.1 02-11 03 23. 0.15	::00:00 20 2 1.23 104.5 ::00:00 20 7	23-02-11 20.1 54.0 23-02-11 20.43 53.5 23-02-11 21.8 53.0 23-02-11	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL 04:00:00 2.7 0.03 NULL 05:00:00	44.0 0.58 24.73 27.28 0.33 24.38 18.28 0.38 24.1 24.28	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0 5 69.0 11.58 0.12 0.0	4.85 0.2 NULL 1.45 0.03 NULL 4.67 NULL NULL 7.57
NULL 2023-(39.47 0.1 74.75 NULL 2023-(36.67 0.05 82.25 NULL 2023-(39.8	02-11 01 24.9 0.27 02-11 02 9.5 0.1 02-11 03 23. 0.15 02-11 04 27.32	::00:00 20 2 1.23 104.5 ::00:00 20 7	23-02-11 20.1 54.0 23-02-11 20.43 53.5 23-02-11 21.8 53.0 23-02-11 3.12	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL 04:00:00 2.7 0.03 NULL 05:00:00 2.22	44.0 0.58 24.73 27.29 0.33 24.38 18.29 0.38 24.1 24.29 0.41	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0 5 69.0 11.58 0.12 0.0 5 90.75 7.27	4.85 0.2 NULL 1.45 0.03 NULL 4.67 NULL NULL 7.57 0.05
NULL 2023-(39.47 0.1 74.75 NULL 2023-(15.78 0.03 78.5 NULL 2023-(36.67 0.05 82.25 NULL 2023-(39.8 0.5 0.5	02-11 01 24.9 0.27 02-11 02 9.5 0.1 02-11 03 23. 0.15 02-11 04 27.32	::00:00 20 2 1.23 104.5 ::00:00 20 7	23-02-11 20.1 54.0 223-02-11 20.43 53.5 223-02-11 21.8 53.0 223-02-11 3.12	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL 04:00:00 2.7 0.03 NULL 05:00:00 2.22 0.2	44.0 0.58 24.73 27.28 0.33 24.38 18.28 0.38 24.1 24.28 0.41	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0 6 69.0 11.58 0.12 0.0 5 90.75 7.27 0.03	4.85 0.2 NULL 1.45 0.03 NULL 4.67 NULL 7.57 0.05 84.0
NULL 2023-(39.47 0.1 74.75 NULL 2023-(15.78 0.03 78.5 NULL 2023-(36.67 0.05 82.25 NULL 2023-(39.8 0.5 0.5	02-11 01 24.9 0.27 02-11 02 9.5 0.1 02-11 03 23. 0.15 02-11 04 27.32	::00:00 20 2 1.23 104.5 ::00:00 20 7	23-02-11 20.1 54.0 223-02-11 20.43 53.5 223-02-11 21.8 53.0 223-02-11 3.12	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL 04:00:00 2.7 0.03 NULL 05:00:00 2.22 0.2	44.0 0.58 24.73 27.28 0.33 24.38 18.28 0.38 24.1 24.28 0.41	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0 5 69.0 11.58 0.12 0.0 5 90.75 7.27	4.85 0.2 NULL 1.45 0.03 NULL 4.67 NULL 7.57 0.05 84.0
NULL 2023-(39.47 0.1 74.75 NULL 2023-(39.8 0.5 NULL) 2023-(39.8 0.5 NULL) 2023-(39.8 0.5 NULL) 2023-(39.8 0.5 NULL)	02-11 01 24.9 0.27 02-11 02 9.5 0.1 02-11 03 23. 0.15 02-11 04 27.32	::00:00 20 2 1.23 104.5 ::00:00 20 7	23-02-11 20.1 54.0 23-02-11 20.43 53.5 23-02-11 21.8 53.0 23-02-11 3.12 NUI	02:00:00 3.1 0.2 NULL 03:00:00 2.92 0.03 NULL 04:00:00 2.7 0.03 NULL 05:00:00 2.22 0.2 LL 24.05	44.0 0.58 24.73 27.29 0.33 24.38 18.29 0.38 24.1 24.29 0.41 0.0	0 150.25 19.35 0.57 0.0 5 98.25 32.75 0.15 0.0 6 69.0 11.58 0.12 0.0 5 90.75 7.27 0.03	4.85 0.2 NULL 1.45 0.03 NULL 4.67 NULL 7.57 0.05 84.0

```
only showing top 20 rows
[10]: // Job 2: Aggregate by average PM10
     val avgPM10DF = parquetDF.groupBy("From Date", "To Date").agg(avg("PM10 (ug/
     avgPM10DF.show()
     writeToMySQL(avgPM10DF, "average_pm10")
    VBox()
    FloatProgress(value=0.0, bar_style='info', description='Progress:',_
     ⇔layout=Layout(height='25px', width='50%'),...
    avgPM10DF: org.apache.spark.sql.DataFrame = [From Date: string, To Date: string
    ... 1 more field]
    +----+
             From Date
                                To Date
    +----+
    |2023-02-12 00:00:00|2023-02-12 01:00:00|152.91847887323948|
    |2023-02-09 14:00:00|2023-02-09 15:00:00|114.16144092219021|
    |2023-01-07 07:00:00|2023-01-07 08:00:00|154.91086956521738|
    |2022-12-05 19:00:00|2022-12-05 20:00:00|223.41452887537994|
    |2022-12-23 20:00:00|2022-12-23 21:00:00| 245.6475076923077|
    |2022-12-25 02:00:00|2022-12-25 03:00:00|
                                             158.1368125
    |2022-11-17 03:00:00|2022-11-17 04:00:00|155.64638095238095|
    |2022-08-07 10:00:00|2022-08-07 11:00:00| 44.20376383763837|
    |2010-10-18 16:00:00|2010-10-18 17:00:00|
                                                   1.861
    |2010-10-28 22:00:00|2010-10-28 23:00:00|
                                                   NULLI
    |2010-12-11 20:00:00|2010-12-11 21:00:00|
                                                  12.955l
    |2010-12-28 10:00:00|2010-12-28 11:00:00|
                                                  59.165 l
    |2011-01-25 14:00:00|2011-01-25 15:00:00|
                                                   1.96
    |2011-02-08 15:00:00|2011-02-08 16:00:00|
                                                  2.085
    |2011-05-04 14:00:00|2011-05-04 15:00:00|
                                                   1.76
    |2011-05-21 14:00:00|2011-05-21 15:00:00|
                                                  35.2351
    |2011-05-22 23:00:00|2011-05-23 00:00:00|
                                                   NULL
    |2011-06-26 20:00:00|2011-06-26 21:00:00|
                                                   2.881
    |2011-10-07 08:00:00|2011-10-07 09:00:00|
                                                  14.81
    |2011-12-06 22:00:00|2011-12-06 23:00:00|
                                                   2.361
    only showing top 20 rows
```

[]:

```
[11]: // Job 3: Count of records by date
     val countByDateDF = parquetDF.groupBy("From Date").agg(count("*").
      →as("record_count"))
     countByDateDF.show()
     // writeToMySQL(countByDateDF, "count_by_date")
    VBox()
    FloatProgress(value=0.0, bar_style='info', description='Progress:',u
      countByDateDF: org.apache.spark.sql.DataFrame = [From Date: string,
    record_count: bigint]
     +----+
              From Date | record_count |
     +----+
     |2023-02-19 17:00:00|
                                4261
     |2022-12-18 00:00:00|
                                402 l
     |2022-12-19 22:00:00|
                                403|
     12022-11-06 08:00:001
                                388 l
     |2022-08-09 19:00:00|
                                366 l
     |2022-08-30 21:00:00|
                                371 l
     |2022-09-10 17:00:00|
                                372
     |2022-10-06 00:00:00|
                                377
     |2010-10-29 15:00:00|
                                 261
     |2010-10-29 19:00:00|
                                 261
     |2011-02-24 00:00:00|
                                 29|
     |2011-03-21 17:00:00|
                                 291
     |2011-03-30 22:00:00|
                                 29|
     |2011-04-11 23:00:00|
                                 29|
     |2011-05-12 11:00:00|
                                 31|
     |2011-06-18 21:00:00|
                                 32|
     |2011-11-24 10:00:00|
                                 32 l
     |2012-01-01 02:00:00|
                                 321
     |2012-02-08 04:00:00|
                                 321
     |2012-03-04 05:00:00|
                                 321
     +----+
    only showing top 20 rows
[12]: countByDateDF.show()
    VBox()
    FloatProgress(value=0.0, bar style='info', description='Progress:',u
      →layout=Layout(height='25px', width='50%'),...
     +----+
              From Date|record_count|
```

```
|2023-02-19 17:00:00|
                            426
    |2022-12-18 00:00:00|
                            4021
    |2022-12-19 22:00:00|
                            403|
    |2022-11-06 08:00:00|
                            388 l
    |2022-08-09 19:00:00|
                            366|
    [2022-08-30 21:00:00]
                            371
    |2022-09-10 17:00:00|
                            3721
    |2022-10-06 00:00:00|
                            377
    |2010-10-29 15:00:00|
                             26 l
    |2010-10-29 19:00:00|
                             26|
    |2011-02-24 00:00:00|
                             29|
    |2011-03-21 17:00:00|
                             29|
    [2011-03-30 22:00:00]
                             291
    |2011-04-11 23:00:00|
                             291
    |2011-05-12 11:00:00|
                             31 l
    |2011-06-18 21:00:00|
                             321
    |2011-11-24 10:00:00|
                             321
    |2012-01-01 02:00:00|
                             321
    |2012-02-08 04:00:00|
                             321
    |2012-03-04 05:00:00|
                             321
    +----+
    only showing top 20 rows
[13]: writeToMySQL(countByDateDF, "count_by_date")
    VBox()
    FloatProgress(value=0.0, bar_style='info', description='Progress:',u
     ⇔layout=Layout(height='25px', width='50%'),...
[14]: // Job 4: Filter by high NO2 levels
    val highNO2DF = parquetDF.filter($"NO2 (ug/m3)" > 100)
    highNO2DF.show()
    writeToMySQL(highNO2DF, "high_no2")
    VBox()
    FloatProgress(value=0.0, bar style='info', description='Progress:',u
     →layout=Layout(height='25px', width='50%'),...
    highNO2DF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [From Date:
    string, To Date: string ... 23 more fields]
    +----+
    ______
    _____
    ______
    ---+----+
```

From Date To Date | PM2.5 (ug/m3) | PM10 (ug/m3) | NO (ug/m3)|NO2 (ug/m3)|NOx (ppb)|NH3 (ug/m3)|SO2 (ug/m3)|CO (mg/m3)|Ozone (ug/m3)|Benzene (ug/m3)|Toluene (ug/m3)|Eth-Benzene (ug/m3)|MP-Xylene (ug/m3)|O Xylene (ug/m3)|Temp (degree C)|RH (%)|WS (m/s)|WD (deg)|SR (W/mt2)|BP (mmHg)|VWS (m/s)|Xylene (ug/m3)|AT (degree C)| +-----______ ---+ |2023-02-13 19:00:00|2023-02-13 20:00:00| 7.95| 60.5 262.0 68.1 115.9 21.65 11.65 2.03 17.57 0.75| 0.1 1.75 0.45 0.321 31.0 | 0.12 | 114.25 | 0.01 NULL 9.01 NULL 24.35 NULL |2023-02-13 20:00:00|2023-02-13 21:00:00| 137.5l 382.751 41.45 129.02 102.33 27.681 10.12 3.391 8.53| 0.17| 2.2 0.71 1.12 0.51 36.5 | 0.12 | 79.75 9.0 NULL 0.0 NULL | 23.85 NULL |2023-02-13 21:00:00|2023-02-13 22:00:00| 178.01 474.51 89.72 120.12 136.88| 32.12 4.8| 8.0 6.67 0.351 2.921 0.931 1.7 0.751 42.25 | 0.12 | 90.0 9.0 NULLI 0.0 NULL 23.58 NULL |2023-02-13 22:00:00|2023-02-13 23:00:00| 200.0| 543.25 102.8 145.2 7.82 32.12 4.9716.47 0.45| 3.8 1.3| 2.47| 1.15 44.25 | 0.15 | 36.0| NULL 9.01 23.3 0.01 NULL NULL |2023-02-13 23:00:00|2023-02-14 00:00:00| 186.75 l 498.01 80.381 104.9 121.13 32.01 6.421 3.991 5.47| 3.05| 0.531 4.35 1.58 1.4 45.0 0.2 81.5 NULL 23.1 0.0 NULL 9.0 NULL | |2023-02-14 19:00:00|2023-02-14 20:00:00| 116.0 433.75 85.25 139.22 31.75 131.47 15.38 4.39 10.1 0.1 2.651 0.531 0.971 0.41 33.5 | 0.15 | 120.5 0.0 NULLI 7.51 NULLI 25.7 NULL |2023-02-14 20:00:00|2023-02-14 21:00:00| 275.75 634.0| 98.85 119.88 42.081 14.78 5.06 7.03| 144.12 0.281 3.55 0.91 1.75 0.77 33.01 52.0 0.01 NULL 0.2 8.01 NULL 25.47 NULL |2023-02-15 08:00:00|2023-02-15 09:00:00| 145.75 399.5 60.371 106.7| 108.3 46.27 10.95 2.56 12.02 0.35| 4.4 1.3| 1.98| 0.85|

47.5 0.2 117.25 99.75 NULL	NULL	24.95	0.01	NULL
2023-02-15 09:00:00 2023-02-15				21.98
102.9 72.62 33.38	21.65	1.44	43.02	0.001
0.4 4.25	1.28	2.02	21	0.88
24.25 0.33 127.75 236.75	NULLI	27.081	0.01	NULLI
NULL				
2023-02-15 10:00:00 2023-02-15				
141.9 95.35 31.7				
0.33 3.37	0.97	1.5	58	0.73
18.5 0.47 166.0 297.5	NULL	28.25	0.0	NULL
NULL				
2023-02-16 10:00:00 2023-02-16	11:00:00	103.0	258.0	20.9
111.92 76.55 36.33				
0.3 2.83	0.88	1.62	2	0.75
0.3 2.83 27.75 0.4 183.25 229.5	NULL	27.55	0.0	NULL
NULL				
2023-02-16 19:00:00 2023-02-16	20:00:00	201.25	750.0	105.95
108.4 143.8 30.17 0.1 2.78	0.551	1.027)	0.571
39.5 0.12 73.5 7.25	NIII I	26 621	0.01	NIII I
NULL	NOLL	20.021	0.01	NOLL
2023-02-16 20:00:00 2023-02-16	21.00.001	420 751	966 AI	240 551
				240.551
106.67 252.33 42.65				1 001
0.28 3.67				
44.5 0.17 195.0 7.25	NULLI	26.251	0.01	NULLI
NULL				
2023-02-17				
106.05 68.0 35.05				
0.22 2.52 24.5 0.4 124.75 300.25	0.78	1.5	55	0.7
	NULL	28.78	0.0	NULL
NULL				
2023-02-17 11:00:00 2023-02-17	12:00:00	89.0	235.5	16.18
116.53 75.13 33.9	46.33	0.85	83.9	
0.2 2.52	0.85	1.4	<u> </u>	0.7
0.2 2.52 20.5 0.6 159.75 347.75	NULL	29.65	0.0	NULL
NULL				
2023-02-17	19:00:00	55.0	287.75	43.05
192.38 137.35 26.65	15.17	2.04	38.1	
0.1 3.02	0.671	1.27	· 7	0.72
26.25 0.15 43.25 NULL	NULL.	28.671	0.01	NULL
NULL				
2023-02-17 19:00:00 2023-02-17	20.00.001	102 5 l	341 01	60 021
166.88 137.57 38.1				00.021
				U 881
0.12 3.27 39.75 0.12 44.0 NULL	V.00 MIII I	26 QOI	0.01	О.ОО МПТ Т
NULL	MOPF (20.30	0.01	MOTF
2023-02-17 20:00:00 2023-02-17	21:00:00	203.25	592.5	139.53

```
151.7
           194.15
                      43.75
                                 14.81
                                          7.21
                                                     7.081
    0.31
                3.92
                                 1.15
                                               2.35
                                                             1.15
    44.25 | 0.18 | 24.75 |
                         NULL
                                  NULLI
                                          26.45
                                                   0.0
                                                               NULL
    NULLI
    |2023-02-17 21:00:00|2023-02-17 22:00:00|
                                           268.751
                                                      734.75
                                                               155.25
    120.77
            190.45
                        49.71
                                 11.53
                                                      6.851
                                          7.471
    0.53
                 4.921
                                  1.7
                                                3.42
                                                              1.62
    46.75 | 0.12
                 183.01
                         NULL
                                  NULLI
                                          25.971
                                                   0.01
                                                               NULLI
    |2023-02-17 22:00:00|2023-02-17 23:00:00|
                                           220.751
                                                      657.75
                                                               132.22
    107.68
            164.78
                       52.52
                                 9.47
                                           6.24
                                                      5.85
    0.75
                                 2.08
                                                4.15
                                                              2.08
                  5.71
    42.75 | 0.25 | 185.75 |
                                                               NULL
                         NULL
                                  NULL
                                          26.12
                                                   0.01
    NULLI
    +----+
       ____+__
    ______
    ---+----
    only showing top 20 rows
[15]: // Job 5: Aggregate by sum of SO2
    val sumSO2DF = parquetDF.groupBy("From Date", "To Date").agg(sum("SO2 (ug/m3)").
     ⇔as("total_so2"))
    sumSO2DF.show()
    writeToMySQL(sumSO2DF, "sum_so2")
    VBox()
    FloatProgress(value=0.0, bar_style='info', description='Progress:',_
     →layout=Layout(height='25px', width='50%'),...
    sumSO2DF: org.apache.spark.sql.DataFrame = [From Date: string, To Date: string
    ... 1 more field]
                               To Date
             From Date
                                             total so2
    +----+
    |2023-02-12 00:00:00|2023-02-12 01:00:00|
    |2023-02-09 14:00:00|2023-02-09 15:00:00| 5872.639999999999
    |2023-01-07 07:00:00|2023-01-07 08:00:00| 4715.539999999999
    |2022-12-05 19:00:00|2022-12-05 20:00:00|
                                               4592.65|
    |2022-12-23 20:00:00|2022-12-23 21:00:00|
                                               4735.41
    |2022-12-25 02:00:00|2022-12-25 03:00:00|
                                               3968.58
    |2022-11-17 03:00:00|2022-11-17 04:00:00| 4462.87000000001|
    |2022-08-07 10:00:00|2022-08-07 11:00:00|
    |2010-10-18 16:00:00|2010-10-18 17:00:00| 81.7400000000001|
    |2010-10-28 22:00:00|2010-10-28 23:00:00|
                                                 66.4I
    |2010-12-11 20:00:00|2010-12-11 21:00:00|117.32000000000002|
```

```
[16]: // Job 6: Select specific columns

val selectColumnsDF = parquetDF.select("From Date", "To Date", "PM10 (ug/m3)",

\( \times \text{"NO (ug/m3)"} \)

selectColumnsDF.show()

writeToMySQL(selectColumnsDF, "selected_columns")
```

VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:',u \(\) \(

selectColumnsDF: org.apache.spark.sql.DataFrame = [From Date: string, To Date:
string ... 2 more fields]

]	 From Date 	To Date	PM10 (ug/m3) NO	(ug/m3)
2023-02-10	09:00:00 2023-02-10	10:00:00	139.5	2.45
2023-02-10	10:00:00 2023-02-10	11:00:00	92.25	1.95
2023-02-10	11:00:00 2023-02-10	12:00:00	96.5	1.78
2023-02-10	12:00:00 2023-02-10	13:00:00	115.0	1.6
2023-02-10	13:00:00 2023-02-10	14:00:00	93.5	1.62
2023-02-10	14:00:00 2023-02-10	15:00:00	83.0	1.55
2023-02-10	15:00:00 2023-02-10	16:00:00	84.0	1.45
2023-02-10	16:00:00 2023-02-10	17:00:00	86.75	1.58
2023-02-10	17:00:00 2023-02-10	18:00:00	84.0	1.6
2023-02-10	18:00:00 2023-02-10	19:00:00	101.75	1.48
2023-02-10	19:00:00 2023-02-10	20:00:00	132.0	1.3
2023-02-10	20:00:00 2023-02-10	21:00:00	136.0	2.15
2023-02-10	21:00:00 2023-02-10	22:00:00	226.25	1.42
2023-02-10	22:00:00 2023-02-10	23:00:00	95.75	2.05
2023-02-10	23:00:00 2023-02-1	00:00:00	94.75	2.27
2023-02-11	00:00:00 2023-02-1	01:00:00	129.25	5.05
2023-02-11	01:00:00 2023-02-1	02:00:00	150.25	4.85
2023-02-11	02:00:00 2023-02-1	03:00:00	98.25	1.45
2023-02-11	03:00:00 2023-02-1	04:00:00	69.0	4.67
2023-02-11	04:00:00 2023-02-1	05:00:00	90.75	7.57

[17]: // Job 7: Filter by CO2 range val filterCO2DF = parquetDF.filter($\mbox{"CO (mg/m3)"} >= 300 \&\& \mbox{"CO (mg/m3)"} <= 400$) filterCO2DF.show() writeToMySQL(filterCO2DF, "filtered_co2") VBox() FloatProgress(value=0.0, bar_style='info', description='Progress:',u ⇔layout=Layout(height='25px', width='50%'),... filterCO2DF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [From Date: string, To Date: string ... 23 more fields] +-----______ ---+----+ From Datel To Date | PM2.5 (ug/m3) | PM10 (ug/m3) | NO (ug/m3)|NO2 (ug/m3)|NOx (ppb)|NH3 (ug/m3)|SO2 (ug/m3)|CO (mg/m3)|Ozone (ug/m3)|Benzene (ug/m3)|Toluene (ug/m3)|Eth-Benzene (ug/m3)|MP-Xylene (ug/m3)|O Xylene (ug/m3)|Temp (degree C)|RH (%)|WS (m/s)|WD (deg)|SR (W/mt2)|BP (mmHg)|VWS (m/s)|Xylene (ug/m3)|AT (degree C)| +-----_____+___ _____ ______ ---+----|2018-08-28 12:00:00|2018-08-28 13:00:00| 20.5 36.8 27.63 44.6 8.15 0.21 17.78 336.01 NULL NULL 58.01 2.731 174.75 484.25 706.0 0.741 28.81 0.01 NULL NULL NULL NULL NULL |2020-08-23 02:00:00|2020-08-23 03:00:00| 2.01 1.88 4.351 3.851 0.931 0.241 10.81 349.3 NULL NULL 82.51 2.071 279.01 2.0 707.01 0.81 NULL NULLI 24.05 86.81 NULL NULL NULL |2020-08-25 02:00:00|2020-08-25 03:00:00| 1.5 2.831 11.38 8.35| 1.33| 0.16 11.65 338.1 NULL NULL 79.25 226.251 2.01 2.0 708.25 0.81 24.621 5.79| NULL NULL NULL NULL| NULL |2020-09-23 05:00:00|2020-09-23 06:00:00| 7.25| 2.01 1.05 4.05| 0.351 0.24 21.7 NULL 373.21 NULL

-----+

only showing top 20 rows

80.75

288.25

2.0

701.5

2.4

	273.42	NULL 1	ULL NULL	NULL	l
				9.66	
16.42 NUL	L 0.0	NULL	315.21	NULL	
				NULL	
	87.05 736	.3 736.3	3 736.3	736.3	736.3
736.3					
2018-05-06 08	:00:00 2018-0	5-06 09:00:00) NULL	9.09	21.65
16.49 NUL	L 0.0	NULL	311.65	NULL	
				NULL	
NULL 20.05	358.3 735.	93 735.93	3 735.93	735.93	735.93
735.93					
2013-02-28 03	:00:00 2013-0	2-28 04:00:00) NULL	. 128.0	128.0
128.0 128.	0.0	1.52	305.32	0.0	
NULL	NULL	NULI	. 1	7.83	0.0
NULL NULL	0.13 0	.O NULI	. NULL	.7.83 NULL	NULL
NULL					
2015-06-22 14	:00:00 2015-0	6-22 15:00:00) NULL	. 27.97	14.53
42.5 0.83	1.37	71.55	357.55	NULL	29.41
80.56	1.51	92	2.13	392.74	745.3
-0.03 190.27	l NULLI	NULLI	NULLI NULL	. NULI	i.
NULL					
	:00:00 2020-1	2-25 19:00:00	255.25	87.83	136.55
144.02 20.	87 4.1	1 65.37	'l 315.131	7.3	•
•	- · ·				
NULL I	94.01	0.3	3 1	41.5	7.01
				41.5	
750.5 -0.11				41.5 NULL	
750.5 -0.11 NULL	18.05 18	.27 NUI	L NULL	.41.5 NULL	NULL
750.5 -0.11 NULL 2018-11-22 20	18.05 18 :00:00 2018-1	.27 NUI 1-22 21:00:00	L NULL 63.33	41.5 NULL 3 175.4	NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL	18.05 18 :00:00 2018-1 37.09	.27 NUI 1-22 21:00:00 26.5	L NULL 0 63.33 326.18	41.5 NULL 3 175.4 40.49	NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL	.27 NUI 1-22 21:00:00 26.5 NU	L NULL 0 63.33 326.18 ULL	.41.5 NULL 8 175.4 40.49 NULL	NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL	.27 NUI 1-22 21:00:00 26.5 NU	L NULL 0 63.33 326.18 ULL	41.5 NULL 3 175.4 40.49	NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0	.27 NUI 1-22 21:00:00 26.5 NU 42.21	L NULL 0 63.33 326.18 ULL 23 NULL	41.5 NULL 3 175.4 40.49 NULL NULL	NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 12018-11-22 23	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1	.27 NUI 1-22 21:00:00 26.5 NU 42.21 :	L NULL 0 63.33 326.18 0 LL 23 NULL 0 53.19	.41.5 NULL 3 175.4 40.49 NULL NULL	NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NUL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68	.27 NUI 1-22 21:00:00	L NULL 0 63.33 326.18 ULL 23 NULL 0 53.19 317.52	.41.5 NULL 3 175.4 40.49 NULL NULL 0 145.55 21.8	NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL	.27 NUI 1-22 21:00:00 26.5 NU 42.21 3 1-23 00:00:00 21.97	L NULL 0 63.33 326.18 ULL 1.23 NULL 0 53.19 317.52	.41.5 NULL 8 175.4 40.49 NULL NULL 0 145.55 21.8	NULL NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL	.27 NUI 1-22 21:00:00	L NULL 0 63.33 326.18 ULL 1.23 NULL 0 53.19 317.52	.41.5 NULL 3 175.4 40.49 NULL NULL 0 145.55 21.8	NULL NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL NULL NULL NULL NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80	.27 NUI 1-22 21:00:00	L NULL 0 63.33 326.18 ULL 23 NULL 0 53.19 317.52 5 1.27	.41.5 NULL 8 175.4 40.49 NULL NULL 0 145.55 21.8 NULL	NULL NULL NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL NULL NULL NULL NULL NULL NULL NULL 2018-11-27 07	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80 :00:00 2018-1	.27 NUI 1-22 21:00:00 26.5 NU 42.21 3 1-23 00:00:00 21.97 NULI .0 33.15	L NULL 0 63.33 326.18 ULL 1.23 NULL 0 53.19 317.52 1.27	.41.5 NULL 8 175.4 40.49 NULL NULL 0 145.55 21.8 NULL NULL	NULL NULL NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80 :00:00 2018-1 L 20.98	.27 NUI 1-22 21:00:00	L NULL 0 63.33 326.18 ULL 1.23 NULL 0 53.19 317.52 1.27 0 66.3 311.88	.41.5 NULL 8 175.4 40.49 NULL NULL 9 145.55 21.8 NULL NULL NULL 8 144.08 25.67	NULL NULL NULL NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80 :00:00 2018-1 L 20.98 NULL	.27 NUI 1-22 21:00:00	L NULL 0 63.33 326.18 ULL 1.23 NULL 0 53.19 317.52 1 5 1.27 0 66.3 311.88	.41.5	NULL NULL NULL NULL NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL 2018-11-27 07 11.87 NULL NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80 :00:00 2018-1 L 20.98 NULL	.27 NUI 1-22 21:00:00	L NULL 0 63.33 326.18 ULL 1.23 NULL 0 53.19 317.52 1 5 1.27 0 66.3 311.88	.41.5	NULL NULL NULL NULL NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL 12018-11-27 07 11.87 NULL NULL NULL NULL NULL NULL NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80 :00:00 2018-1 L 20.98 NULL NULL NULL 110	.27 NUI 1-22 21:00:00 26.5 NU 42.21 3 1-23 00:00:00 21.97 NULI .0 33.18 1-27 08:00:00 10.32 NULI .0 24.28	NULL NULL	.41.5	NULL NULL NULL NULL NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80 :00:00 2018-1 L 20.98 NULL NULL 110 :00:00 2018-1	.27 NUI 1-22 21:00:00 26.5 NU 42.21 3 1-23 00:00:00 21.97 NULI .0 33.15 1-27 08:00:00 10.32 NULI .0 24.25	L NULL 0 63.33 326.18 ULL 1.23 NULL 0 53.19 317.52 1 6 1.27 0 66.3 311.88 1 6 1.45	.41.5	NULL NULL NULL NULL NULL NULL NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL NULL NULL NULL NULL NULL NULL NULL 2018-11-27 07 11.87 NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80 :00:00 2018-1 L 20.98 NULL NULL 110 :00:00 2018-1:	.27 NUI 1-22 21:00:00 26.5 NU 42.21 3 1-23 00:00:00 21.97 NULI .0 33.15 1-27 08:00:00 10.32 NULI .0 24.25 2-20 22:00:00 16.15	NULL NULL	.41.5	NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL NULL NULL NULL NULL NULL NULL 2018-11-27 07 11.87 NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80 :00:00 2018-1 L 20.98 NULL NULL 110 :00:00 2018-1	.27 NUI 1-22 21:00:00 26.5 NU 42.21 3 1-23 00:00:00 21.97 NULI .0 33.19 1-27 08:00:00 10.32 NULI .0 24.29 2-20 22:00:00 16.15 NULI	L NULL 326.18 326.18 326.18 311.23 NULL 317.52 317.52 311.88 311.88 311.88 311.88 311.88	.41.5	NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80 :00:00 2018-1 L 20.98 NULL NULL 110 :00:00 2018-1	.27 NUI 1-22 21:00:00 26.5 NU 42.21 3 1-23 00:00:00 21.97 NULI .0 33.19 1-27 08:00:00 10.32 NULI .0 24.29 2-20 22:00:00 16.15 NULI	L NULL 326.18 326.18 326.18 311.23 NULL 317.52 317.52 311.88 311.88 311.88 311.88 311.88	.41.5	NULL NULL
750.5 -0.11 NULL 2018-11-22 20 5.89 NULL NULL NULL NULL NULL NULL 2018-11-22 23 35.52 NULL NULL NULL NULL NULL NULL NULL NULL 2018-11-27 07 11.87 NULL NULL NULL	18.05 18 :00:00 2018-1 37.09 NULL 65.0 :00:00 2018-1 L 36.68 NULL NULL 80 :00:00 2018-1 L 20.98 NULL NULL 110 :00:00 2018-1: L 31.83 NULL NULL 145	.27 NUI 1-22 21:00:00 26.5 NU 42.21 3 1-23 00:00:00 21.97 NULI .0 33.18 1-27 08:00:00 10.32 NULI .0 24.28 2-20 22:00:00 16.15 NULI .0 40.83	NULL NULL	.41.5	NULL

```
NULL
                  NULL
                               NULL
                                                        NULL
                                            NULL
                               5.09|
                                        NULLI
   NULL
          NULL
                NULL
                        39.14
                                                   NULLI
   NULLI
   |2018-12-23 01:00:00|2018-12-23 02:00:00| 47.34|
                                               127.87
                                                          NULL
           NULL
                    47.721
                             15.4l
                                    385.63
                                               10.11
   NULLI
               NULL
                              NULL
                                           NULL
                                                        NULL
   NULL| NULL|
               NULL
                      NULLI
                             36.461
                                      5.21
                                             NULLI
                                                        NULL
   |2018-12-23 04:00:00|2018-12-23 05:00:00|
                                                 82.241
                                                          NULLI
                                       48.48|
   51.08|
           NULL
                    57.45
                                                7.04
                             11.15
                                    327.01
   NULL
               NULL
                                           NULL
                              NULL
                                                        NULL
   NULL| NULL|
               NULL|
                                             NULL|
                                                        NULL
                      NULL
                             31.29
                                      2.36
   NULL
   |2018-12-23 21:00:00|2018-12-23 22:00:00|
                                        37.65|
                                               155.48
                                                          NULL
          NULLI
                    54.75l
                             16.14
                                    327.01
                                                8.221
   NULL
               NULL
                              NULL
                                           NULL
                                                        NULL
                                             NULL
   NULL| NULL|
               NULL
                      NULL
                             44.64
                                      1.7
                                                        NULL
   NULL|
   |2018-12-23 22:00:00|2018-12-23 23:00:00|
                                        52.821
                                                 157.54
                                                          NULL
                                     392.2
           NULLI
                    55.77
                             16.47|
                                               10.16
   NULL
                                           NULL
               NULL
                              NULL
                                                        NULL
   NULL| NULL|
               NULL| NULL|
                             44.991
                                    4.621
                                           NULLI
                                                        NULL
   NULLI
   |2018-12-23 23:00:00|2018-12-24 00:00:00|
                                        58.32 | 138.62 |
                                                          NULL
                  49.65|
   9.8|
          NULL
                           17.66
                                 372.22
                                              22.12
                                                          NULL
   NULL
                  NULL
                               NULL
                                            NULL
                                                        NULL
                               2.06| NULL|
   NULL
          NULL
                 NULL
                        37.01
                                                   NULL
   NULLI
   |2018-12-26 21:00:00|2018-12-26 22:00:00| 52.2| 159.29|
                                                          NULL
   7.91
           NULLI
                   44.17
                            16.57
                                   366.291
                                              20.341
                                                           NULLI
   NULL
                  NULL
                                NULL
                                            NULL
                                                        NULL
                130.0
                                1.52
                                        NULL
                                                   NULL
   NULL
          NULL
                         49.6
   NULLI
   +-----
   _____
    ______
    ---+---+
   only showing top 20 rows
[18]: // Job 8: Aggregate by max NO
    val maxNODF = parquetDF.groupBy("From Date", "To Date").agg(max("NO (ug/m3)").
    ⇔as("max_no"))
    maxNODF.show()
    writeToMySQL(maxNODF, "max no")
```

NULLI

NULL

75.01

```
VBox()
    FloatProgress(value=0.0, bar style='info', description='Progress:',u
      →layout=Layout(height='25px', width='50%'),...
    maxNODF: org.apache.spark.sql.DataFrame = [From Date: string, To Date: string
     ... 1 more field]
     +----+
              From Date
                                  To Date|max_no|
     +----+
     |2010-01-01 12:00:00|2010-01-01 13:00:00| |
     |2010-01-01 14:00:00|2010-01-01 15:00:00| 6.83|
     |2010-01-05 06:00:00|2010-01-05 07:00:00| 53.31|
     |2010-01-05 17:00:00|2010-01-05 18:00:00| 9.87|
     |2010-01-07 20:00:00|2010-01-07 21:00:00| 7.31|
     |2010-01-08 03:00:00|2010-01-08 04:00:00| 7.25|
     |2010-01-10 08:00:00|2010-01-10 09:00:00| 8.94|
     |2010-01-10 19:00:00|2010-01-10 20:00:00| 7.98|
     |2010-01-10 22:00:00|2010-01-10 23:00:00| 82.13|
     |2010-01-11 06:00:00|2010-01-11 07:00:00| 9.64|
     |2010-01-11 14:00:00|2010-01-11 15:00:00| 7.46|
     |2010-01-11 16:00:00|2010-01-11 17:00:00| 9.37|
     |2010-01-19 19:00:00|2010-01-19 20:00:00| 7.91|
     |2010-01-23 02:00:00|2010-01-23 03:00:00| 88.2|
     |2010-01-23 16:00:00|2010-01-23 17:00:00|
                                             9.5
     |2010-01-24 02:00:00|2010-01-24 03:00:00| 87.68|
     |2010-01-24 22:00:00|2010-01-24 23:00:00| 74.25|
     |2010-01-25 01:00:00|2010-01-25 02:00:00| 5.55|
     |2010-01-25 02:00:00|2010-01-25 03:00:00| 58.25|
     |2010-01-28 14:00:00|2010-01-28 15:00:00| 7.83|
     +----+
    only showing top 20 rows
[19]: // Job 9: Aggregate by min PM10
     val minPM10DF = parquetDF.groupBy("From Date", "To Date").agg(min("PM10 (ug/
      minPM10DF.show()
     writeToMySQL(minPM10DF, "min_pm10")
    VBox()
    FloatProgress(value=0.0, bar style='info', description='Progress:', __
      ⇔layout=Layout(height='25px', width='50%'),...
    minPM10DF: org.apache.spark.sql.DataFrame = [From Date: string, To Date: string
     ... 1 more field]
              From Date
                                   To Date|min_pm10|
```

```
|2010-01-01 12:00:00|2010-01-01 13:00:00|
                                           NULL
    |2010-01-01 14:00:00|2010-01-01 15:00:00|
                                           NULL
    |2010-01-05 06:00:00|2010-01-05 07:00:00|
                                           NULL
    |2010-01-05 17:00:00|2010-01-05 18:00:00|
                                           NULL
    |2010-01-07 20:00:00|2010-01-07 21:00:00|
                                           NULL
    |2010-01-08 03:00:00|2010-01-08 04:00:00|
                                           NULL
    |2010-01-10 08:00:00|2010-01-10 09:00:00|
                                           NULL
    |2010-01-10 19:00:00|2010-01-10 20:00:00|
                                           NULL
    |2010-01-10 22:00:00|2010-01-10 23:00:00|
                                           NULL
    |2010-01-11 06:00:00|2010-01-11 07:00:00|
                                           NULL|
    |2010-01-11 14:00:00|2010-01-11 15:00:00|
                                           NULL
    |2010-01-11 16:00:00|2010-01-11 17:00:00|
                                           NULL|
    |2010-01-19 19:00:00|2010-01-19 20:00:00|
                                           NULL
    |2010-01-23 02:00:00|2010-01-23 03:00:00|
                                           NULL
    |2010-01-23 16:00:00|2010-01-23 17:00:00|
                                           NULL
    |2010-01-24 02:00:00|2010-01-24 03:00:00|
                                           NULL
    |2010-01-24 22:00:00|2010-01-24 23:00:00|
                                           NULL
    |2010-01-25 01:00:00|2010-01-25 02:00:00|
                                           NULL
    |2010-01-25 02:00:00|2010-01-25 03:00:00|
                                           NULL
    |2010-01-28 14:00:00|2010-01-28 15:00:00|
                                           NULL
    only showing top 20 rows
[20]: // Job 10: Filter by low SO2 levels
     val lowSO2DF = parquetDF.filter($"SO2 (ug/m3)" < 50)</pre>
     lowSO2DF.show()
     writeToMySQL(lowSO2DF, "low_so2")
    VBox()
    FloatProgress(value=0.0, bar style='info', description='Progress:',u
     ⇔layout=Layout(height='25px', width='50%'),...
    lowSO2DF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [From Date:
    string, To Date: string ... 23 more fields]
    +-----
    ______
    _____
    ______
    ---+---+
                                 To Date | PM2.5 (ug/m3) | PM10 (ug/m3) | NO
              From Date
    (ug/m3)|NO2 (ug/m3)|NOx (ppb)|NH3 (ug/m3)|SO2 (ug/m3)|CO (mg/m3)|Ozone
    (ug/m3)|Benzene (ug/m3)|Toluene (ug/m3)|Eth-Benzene (ug/m3)|MP-Xylene (ug/m3)|O
    Xylene (ug/m3)|Temp (degree C)|RH (%)|WS (m/s)|WD (deg)|SR (W/mt2)|BP (mmHg)|VWS
    (m/s)|Xylene (ug/m3)|AT (degree C)|
```

		+						
		·						
	·+		-+	+	+	+	+	
•		00:00 2023-02	2-10 10:0	00:00	40.0)	139.5	2.451
		19.45						
		NULL						
1.85	239.5	219.0	NULL	25.65	0.01		NULL	
NULL								
12023-	02-10 10:	00:00 2023-02	2-10 11:0	00:00	18.75	5	92.25	1.95
5.88	4.67	21.62	15	5.1	0.29	64.	92	0.0
0.2		NULL		NULL		NULL		56.25
2.12	241.0	236.5	NULL	26.18	0.0		NULL	
NULL								
		00:00 2023-02						
		23.6						
		NULL						
		264.5	NULL	26.91	0.0		NULL	
NULL								
		00:00 2023-02						
		24.45						
		NULL						
NULL	230.75	175.75	NOLLI	27.55	0.01		MOLL	
	.02-10 13.	00:00 2023-02	2-10 1/1.0	00.001	10 F	:1	03 51	1 621
		24.6						
		NULL						
1.951	248.01	266.25	NUT.I.I	28.01	0.01	NOLL	NUT.T. I	10.01
NULL	210.01	200.201		20.01	0.01			
	02-10 14:	00:00 2023-02	2-10 15:0	00:00	12.0)	83.01	1.55
		25.93						
0.21		NIII I I		MIII I		MIII I I		40 751
1.77	237.75	235.25	NULL	28.45	0.01		NULL	
NULL								
2023-	02-10 15:	00:00 2023-02	2-10 16:0	00:00	11.75	5	84.0	1.45
4.75	3.67	27.02	7	7.2	0.31	76.	92	0.01
		NULL						
1.73	241.5	185.5	NULL	28.67	0.0		NULL	
NULL								
		00:00 2023-02						
		24.95						
0.2		NULL	·	NULL		NULL	··	39.5
		126.0	NULL	28.67	0.0		NULL	
NULL		00 0010000 5	2 40 40 4	00.001	40 -	. 1	04.01	
		00:00 2023-02						
		23.18						
	210 251	NULL	NITIT T I	U.UJ	0.01	NOTF	NITIT T	42./5
0.95	210.25	73.0	NOFF	28.33	0.01		NULL	

NULL							
12023-0	02-10 1	8:00:00	2023-02-10	19:00:00	20.	0 101.75	1.48
						67.52	
0.0		0.2		0.1		0.1	NULL
50.0	0.55	201.5	53.75	NULL	27.53	0.1 0.0	NULL
NULL							
12023-0	2-10 1	9:00:00	2023-02-10	20:00:00	22.	0 132.0	1.3
18.52	10	.91	19.45	7.671	0.471	53.25	
0.03		0.3		0.03		0.1	NULL
52.25	0.38	200.75	53.0	NULL	26.95	0.0	NULL
NULL							
2023-0	2-10 2	0:00:00	2023-02-10	21:00:00	24.	5 136.0	2.15
49.15	27.	88	20.4	6.93	0.83	21.0	
0.12		0.75		0.38		0.45	0.1
						0.01	
NULL							
		1:00:00	2023-02-10	22:00:00	41.7	75 226.25	1.42
						49.97	
						1.05	
54.751	0.41	140.0	53.25	NULLI	26.41	0.01	NULL I
NULL	·			·	·	·	•
	2-10 2	2:00:001	2023-02-10	23:00:00	23.7	['] 5 95.75	2.051
						37.43	
0.01		0.45		0.05		0.1	NULL
63.5	0.1	97.75	53.75	NULL	25.53	0.1	NULL
NULL							
12023-0	2-10 2	3:00:00	2023-02-11	00:00:00	19.2	25 94.75	2.27
40.771	23.	55 l	18.17	4.751	0.641	15.22	
0.08		0.57		0.1		0.12	0.03
70.0	0.15	130.25	54.0	NULL	25.05	0.0	NULL
NULL							
12023-0	02-11 0	0:00:00	2023-02-11	01:00:00	37.	5 129.25	5.05
49.65	30.	53	19.32	3.52	0.86	5 831	
						0.42	
						0.0	
NULL							
12023-0	02-11 0	1:00:00	2023-02-11	02:00:00	44.	0 150.25	4.85
						19.35	
0.1		1.23		0.2		0.57	0.2
74.75	0.27	104.5	54.0	NULL	24.73		NULL
NULL							
		2:00:00	2023-02-11	03:00:00	27.2	25 98.25	1.45
15.78	9.	57	20.43	2.92	0.33	32.75	
							0.03
						0.01	
NULL							
		3:00:00	2023-02-11	04:00:00	18.2	25 69.0	4.67
36.67	23	.3	21.8	2.7	0.38	11.58	

0.03| 0.12| NULL| 24.1| 0.0| 0.05| 0.45| NULL 82.25 | 0.15 | 120.0 | 53.0 NULL NULL |2023-02-11 04:00:00|2023-02-11 05:00:00| 24.25| 90.75| 7.57 27.32| 23.12 | 2.22 | 0.41 | 39.81 7.27 0.05 0.5 0.1 0.2 0.03| 84.0| 24.05| 0.15 | 111.25 | 53.25 NULL 0.0 NULL NULL +----+ ______ _____ ---+----+ only showing top 20 rows

[]: