

EMRDemo

June 25, 2024

```
[18]: // MySQL database connection properties
val jdbcUrl = "jdbc:mysql://34.100.169.199:3306/project2"
val connectionProperties = new java.util.Properties()
connectionProperties.put("user", "root")
connectionProperties.put("password", "abhi@iit123")
connectionProperties.put("driver", "com.mysql.cj.jdbc.Driver")
```

VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:',
↳ layout=Layout(height='25px', width='50%'),...

jdbcUrl: String = jdbc:mysql://your-mysql-host:3306/your-database
connectionProperties: java.util.Properties = {}
res9: Object = null
res10: Object = null
res11: Object = null

```
[ ]: spark
```

VBox()

Starting Spark application

```
[29]: // 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:',
↳ layout=Layout(height='25px', width='50%'),...

inputS3Path: String = s3://task-3-emr/air-quality-data/
outputParquetS3Path: String = s3://task-3-emr/air-quality-data-parquet/

```
[30]: val airQualityDF = spark.read.option("header", "true").option("inferSchema",  
↳ "true").csv(inputS3Path)
```

VBox()

```
FloatProgress(value=0.0, bar_style='info', description='Progress:',
↳layout=Layout(height='25px', width='50%'),...

airQualityDF: org.apache.spark.sql.DataFrame = [From Date: string, To Date:
string ... 23 more fields]
```

```
[31]: airQualityDF.count()
```

```
VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:',
↳layout=Layout(height='25px', width='50%'),...

res24: Long = 14296116
```

```
[32]: airQualityDF.write.format("parquet").mode("overwrite").save(outputParquetS3Path)
```

```
VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:',
↳layout=Layout(height='25px', width='50%'),...
```

```
[33]: 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]
```

```
[36]: // Job 1: Filter by date range
val filterDF = parquetDF.filter($"From Date" >= "2023-01-01" && $"To Date" <=
↳"2023-12-31")
filterDF.show()
```

```
VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:',
↳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]
+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
|          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)|

```

+-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+-----+
-----+-----+
|2023-02-10 09:00:00|2023-02-10 10:00:00|      40.0|      139.5|      2.45|
7.95|      6.25|      19.45|      14.55|      0.3|      61.85|      0.1|
0.2|      NULL|      NULL|      NULL|      NULL|      56.0|
1.85|     239.5|     219.0|      NULL|     25.65|      0.0|      NULL|
NULL|
|2023-02-10 10:00:00|2023-02-10 11:00:00|      18.75|      92.25|      1.95|
5.88|      4.67|      21.62|      15.1|      0.29|      64.92|      0.0|
0.2|      NULL|      NULL|      NULL|      NULL|      56.25|
2.12|     241.0|     236.5|      NULL|     26.18|      0.0|      NULL|
NULL|
|2023-02-10 11:00:00|2023-02-10 12:00:00|      13.25|      96.5|      1.78|
5.35|      4.28|      23.6|      14.18|      0.27|      66.17|      0.0|
0.2|      NULL|      NULL|      NULL|      NULL|      53.25|
1.95|     251.0|     264.5|      NULL|     26.9|      0.0|      NULL|
NULL|
|2023-02-10 12:00:00|2023-02-10 13:00:00|      15.25|      115.0|      1.6|
5.0|      3.97|      24.45|      13.05|      0.26|      70.33|      0.0|
0.2|      NULL|      NULL|      NULL|      NULL|      48.5|
1.75|     238.75|     175.75|      NULL|     27.55|      0.0|      NULL|
NULL|
|2023-02-10 13:00:00|2023-02-10 14:00:00|      18.5|      93.5|      1.62|
5.42|      4.23|      24.6|      11.62|      0.27|      71.47|      0.0|
0.2|      NULL|      NULL|      NULL|      NULL|      45.0|
1.95|     248.0|     266.25|      NULL|     28.0|      0.0|      NULL|
NULL|
|2023-02-10 14:00:00|2023-02-10 15:00:00|      12.0|      83.0|      1.55|
4.47|      3.62|      25.93|      8.72|      0.27|      73.6|      0.0|
0.2|      NULL|      NULL|      NULL|      NULL|      40.75|
1.77|     237.75|     235.25|      NULL|     28.45|      0.0|      NULL|
NULL|
|2023-02-10 15:00:00|2023-02-10 16:00:00|      11.75|      84.0|      1.45|
4.75|      3.67|      27.02|      7.2|      0.31|      76.92|      0.0|
0.2|      NULL|      NULL|      NULL|      NULL|      39.0|
1.73|     241.5|     185.5|      NULL|     28.67|      0.0|      NULL|
NULL|
|2023-02-10 16:00:00|2023-02-10 17:00:00|      10.25|      86.75|      1.58|
4.7|      3.77|      24.95|      7.7|      0.32|      79.83|      0.0|
0.2|      NULL|      NULL|      NULL|      NULL|      39.5|
1.7|     240.5|     126.0|      NULL|     28.67|      0.0|      NULL|
NULL|
|2023-02-10 17:00:00|2023-02-10 18:00:00|      13.5|      84.0|      1.6|

```

5.75	4.33	23.18	12.12	0.33	80.42	0.0
0.2		NULL	0.03		NULL	42.75
0.95	210.25	73.0	NULL	28.33	0.0	NULL
NULL						
2023-02-10 18:00:00	2023-02-10 19:00:00			20.0	101.75	1.48
11.43	7.3	21.72	11.38	0.4	67.52	
0.0	0.2		0.1		0.1	NULL
50.0	0.55	201.5	53.75	NULL	27.53	0.0
NULL						NULL
2023-02-10 19:00:00	2023-02-10 20:00:00			22.0	132.0	1.3
18.52	10.9	19.45	7.67	0.47	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
2023-02-10 20:00:00	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
55.0	0.35	216.0	53.0	NULL	26.58	0.0
NULL						NULL
2023-02-10 21:00:00	2023-02-10 22:00:00			41.75	226.25	1.42
22.15	12.93	20.15	6.7	0.48	49.97	
0.12	1.82		1.05		1.05	0.3
54.75	0.4	140.0	53.25	NULL	26.4	0.0
NULL						NULL
2023-02-10 22:00:00	2023-02-10 23:00:00			23.75	95.75	2.05
25.42	15.2	18.77	5.7	0.51	37.43	
0.0	0.45		0.05		0.1	NULL
63.5	0.1	97.75	53.75	NULL	25.53	0.0
NULL						NULL
2023-02-10 23:00:00	2023-02-11 00:00:00			19.25	94.75	2.27
40.77	23.55	18.17	4.75	0.64	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
2023-02-11 00:00:00	2023-02-11 01:00:00			37.5	129.25	5.05
49.65	30.53	19.32	3.52	0.86	5.83	
0.12	1.32		0.15		0.42	0.17
75.0	0.1	96.75	54.0	NULL	24.73	0.0
NULL						NULL
2023-02-11 01:00:00	2023-02-11 02:00:00			44.0	150.25	4.85
39.47	24.92	20.1	3.1	0.58	19.35	
0.1	1.23		0.2		0.57	0.2
74.75	0.27	104.5	54.0	NULL	24.73	0.0
NULL						NULL
2023-02-11 02:00:00	2023-02-11 03:00:00			27.25	98.25	1.45
15.78	9.57	20.43	2.92	0.33	32.75	
0.03	0.38		0.03		0.15	0.03
78.5	0.1	76.0	53.5	NULL	24.38	0.0
						NULL

```

NULL|
|2023-02-11 03:00:00|2023-02-11 04:00:00|      18.25|      69.0|      4.67|
36.67|      23.3|      21.8|      2.7|      0.38|      11.58|
0.05|      0.45|      0.03|      0.12|      NULL|
82.25|  0.15|  120.0|   53.0|      NULL|   24.1|   0.0|      NULL|
NULL|
|2023-02-11 04:00:00|2023-02-11 05:00:00|      24.25|      90.75|      7.57|
39.8|   27.32|   23.12|   2.22|   0.41|      7.27|      0.05|
0.5|      0.1|      0.2|      0.03|      84.0|
0.15| 111.25|  53.25|      NULL|  24.05|   0.0|      NULL|
NULL|
+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+
---+-----+
only showing top 20 rows

```

```

[41]: // Job 2: Aggregate by average PM10
val avgPM10DF = parquetDF.groupBy("From Date", "To Date").agg(avg("PM10 (ug/
    ↪m3)").as("avg_pm10"))
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|      avg_pm10|
+-----+-----+-----+-----+
|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|      1.86|
|2010-10-28 22:00:00|2010-10-28 23:00:00|      NULL|
|2010-12-11 20:00:00|2010-12-11 21:00:00|    12.955|
|2010-12-28 10:00:00|2010-12-28 11:00:00|    59.165|
|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	35.235
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	2.88
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	2.36

+-----+

only showing top 20 rows

```
[43]: // 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:',
 layout=Layout(height='25px', width='50%'),...

countByDateDF: org.apache.spark.sql.DataFrame = [From Date: string,
 record_count: bigint]

	From Date record_count
--	------------------------

+-----+

2023-02-19 17:00:00	426
2022-12-18 00:00:00	402
2022-12-19 22:00:00	403
2022-11-06 08:00:00	388
2022-08-09 19:00:00	366
2022-08-30 21:00:00	371
2022-09-10 17:00:00	372
2022-10-06 00:00:00	377
2010-10-29 15:00:00	26
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	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
2012-01-01 02:00:00	32
2012-02-08 04:00:00	32
2012-03-04 05:00:00	32

+-----+

only showing top 20 rows

104.9	121.13	32.0	6.42	3.99	5.47	
0.53	4.35		1.58	3.05	1.4	
45.0	0.2	81.5	9.0	NULL	23.1	0.0
NULL						NULL
2023-02-14 19:00:00	2023-02-14 20:00:00			116.0	433.75	85.25
131.47	139.22	31.75	15.38	4.39	10.1	
0.1	2.65		0.53	0.97	0.4	
33.5	0.15	120.5	7.5	NULL	25.7	0.0
NULL						NULL
2023-02-14 20:00:00	2023-02-14 21:00:00			275.75	634.0	98.85
119.88	144.12	42.08	14.78	5.06	7.03	
0.28	3.55		0.9	1.75	0.77	
33.0	0.2	52.0	8.0	NULL	25.47	0.0
NULL						NULL
2023-02-15 08:00:00	2023-02-15 09:00:00			145.75	399.5	60.37
108.3	106.7	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	24.95	0.0
NULL						NULL
2023-02-15 09:00:00	2023-02-15 10:00:00			190.25	508.25	21.98
102.9	72.62	33.38	21.65	1.44	43.62	
0.4	4.25		1.28	2.02	0.88	
24.25	0.33	127.75	236.75	NULL	27.08	0.0
NULL						NULL
2023-02-15 10:00:00	2023-02-15 11:00:00			101.0	255.25	24.45
141.9	95.35	31.7	51.1	1.1	63.0	
0.33	3.37		0.97	1.58	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	45.1	1.14	62.48	
0.3	2.83		0.88	1.62	0.75	
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	11.92	4.02	7.05	
0.1	2.78		0.55	1.2	0.57	
39.5	0.12	73.5	7.25	NULL	26.62	0.0
NULL						NULL
2023-02-16 20:00:00	2023-02-16 21:00:00			429.75	866.0	240.55
106.67	252.33	42.65	15.02	7.96	8.22	
0.28	3.67		1.15	2.27	1.08	
44.5	0.17	195.0	7.25	NULL	26.25	0.0
NULL						NULL
2023-02-17 10:00:00	2023-02-17 11:00:00			77.25	208.25	14.28
106.05	68.0	35.05	32.3	0.99	65.28	
0.22	2.52		0.78	1.55	0.7	
24.5	0.4	124.75	300.25	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|      0.7|
20.5|      0.6| 159.75| 347.75|      NULL| 29.65|      0.0|      NULL|
NULL|
|2023-02-17 18:00:00|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|      0.67|      1.27|      0.72|
26.25| 0.15| 43.25|      NULL|      NULL| 28.67|      0.0|      NULL|
NULL|
|2023-02-17 19:00:00|2023-02-17 20:00:00|     102.5|      341.0|      60.02|
166.88| 137.57|      38.1|      15.0|      4.34|      7.47|
0.12|      3.27|      0.88|      1.6|      0.88|
39.75| 0.12| 44.0|      NULL|      NULL| 26.98|      0.0|      NULL|
NULL|
|2023-02-17 20:00:00|2023-02-17 21:00:00|     203.25|      592.5|      139.53|
151.7| 194.15|      43.75|      14.8|      7.21|      7.08|
0.3|      3.92|      1.15|      2.35|      1.15|
44.25| 0.18| 24.75|      NULL|      NULL| 26.45|      0.0|      NULL|
NULL|
|2023-02-17 21:00:00|2023-02-17 22:00:00|     268.75|      734.75|      155.25|
120.77| 190.45|      49.7|      11.53|      7.47|      6.85|
0.53|      4.92|      1.7|      3.42|      1.62|
46.75| 0.12| 183.0|      NULL|      NULL| 25.97|      0.0|      NULL|
NULL|
|2023-02-17 22:00:00|2023-02-17 23:00:00|     220.75|      657.75|      132.22|
107.68| 164.78|      52.52|      9.47|      6.24|      5.85|
0.75|      5.7|      2.08|      4.15|      2.08|
42.75| 0.25| 185.75|      NULL|      NULL| 26.12|      0.0|      NULL|
NULL|
+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+
---+-----+
only showing top 20 rows

```

```

[45]: // 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%'),...
```

```
sumS02DF: org.apache.spark.sql.DataFrame = [From Date: string, To Date: string
... 1 more field]
```

From Date	To Date	total_so2
2023-02-12 00:00:00	2023-02-12 01:00:00	4854.43
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	4735.4
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	4462.870000000001
2022-08-07 10:00:00	2022-08-07 11:00:00	3346.05
2010-10-18 16:00:00	2010-10-18 17:00:00	81.74000000000001
2010-10-28 22:00:00	2010-10-28 23:00:00	66.4
2010-12-11 20:00:00	2010-12-11 21:00:00	117.32000000000002
2010-12-28 10:00:00	2010-12-28 11:00:00	109.82999999999998
2011-01-25 14:00:00	2011-01-25 15:00:00	203.26999999999998
2011-02-08 15:00:00	2011-02-08 16:00:00	198.1
2011-05-04 14:00:00	2011-05-04 15:00:00	190.06999999999994
2011-05-21 14:00:00	2011-05-21 15:00:00	208.64000000000007
2011-05-22 23:00:00	2011-05-23 00:00:00	154.26
2011-06-26 20:00:00	2011-06-26 21:00:00	55.22999999999999
2011-10-07 08:00:00	2011-10-07 09:00:00	47.169999999999995
2011-12-06 22:00:00	2011-12-06 23:00:00	138.87999999999997

only showing top 20 rows

```
[46]: // Job 6: Select specific columns
val selectColumnsDF = parquetDF.select("From Date", "To Date", "PM10 (ug/m3)",
↳ "NO (ug/m3)")
selectColumnsDF.show()
```

VBox()

```
FloatProgress(value=0.0, bar_style='info', description='Progress:',
↳ layout=Layout(height='25px', width='50%'),...
```

```
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

only showing top 20 rows

VBox()

```
filterC02DF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [From
Date: string, To Date: string ... 23 more fields]
```

11

2018-08-28 12:00:00 2018-08-28 13:00:00	20.5	36.8	27.63
44.6 8.15 0.21 17.78 336.0	NULL	NULL	NULL
58.0 2.73 174.75 484.25		706.0	
0.74 28.8 0.0 NULL NULL NULL		NULL	
NULL			
2020-08-23 02:00:00 2020-08-23 03:00:00	2.0	1.88	4.35
3.85 0.93 0.24 10.8 349.3	NULL	NULL	NULL
82.5 2.07 279.0 2.0		707.0	
0.8 24.05 86.8 NULL NULL NULL		NULL	
NULL			
2020-08-25 02:00:00 2020-08-25 03:00:00	1.5	2.83	11.38
8.35 1.33 0.16 11.65 338.1	NULL	NULL	NULL
79.25 2.0 226.25 2.0		708.25	
0.8 24.62 5.79 NULL NULL NULL		NULL	
NULL			
2020-09-23 05:00:00 2020-09-23 06:00:00	2.0	1.05	7.25
4.05 0.35 0.24 21.7 373.2	NULL	NULL	NULL
80.75 2.4 288.25 2.0		701.5	
0.75 23.95 273.42 NULL NULL NULL		NULL	
NULL			
2018-05-06 07:00:00 2018-05-06 08:00:00	NULL	9.66	20.65
16.42 NULL 0.0 NULL 315.21	NULL		
NULL NULL NULL NULL NULL		30.82	
NULL 183.43 87.05 736.3 736.3 736.3 736.3		736.3	
736.3			
2018-05-06 08:00:00 2018-05-06 09:00:00	NULL	9.09	21.65
16.49 NULL 0.0 NULL 311.65	NULL		
NULL NULL NULL NULL NULL		30.55	
NULL 20.05 358.3 735.93 735.93 735.93 735.93		735.93	
735.93			
2013-02-28 03:00:00 2013-02-28 04:00:00	NULL	128.0	128.0
128.0 128.0 0.0 1.52 305.32	0.0		
NULL NULL NULL NULL 17.83		0.0	
NULL NULL 0.13 0.0 NULL NULL NULL		NULL	
NULL			
2015-06-22 14:00:00 2015-06-22 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 92.13 392.74		745.3	
-0.03 190.27 NULL NULL NULL NULL		NULL	
NULL			
2020-12-25 18:00:00 2020-12-25 19:00:00	255.25	87.83	136.55
144.02 20.87 4.11 65.37 315.13	7.3		
NULL 94.0 0.3 141.5		7.0	
750.5 -0.11 18.05 18.27 NULL NULL NULL		NULL	
NULL			
2018-11-22 20:00:00 2018-11-22 21:00:00	63.33	175.4	NULL
5.89 NULL 37.09 26.5 326.18	40.49	NULL	
NULL NULL NULL NULL NULL		NULL	

NULL	NULL	65.0	42.21	1.23	NULL	NULL
NULL						
2018-11-22 23:00:00	2018-11-23 00:00:00			53.19	145.55	NULL
35.52	NULL	36.68	21.97	317.52	21.8	
NULL	NULL	NULL	NULL	NULL		NULL
NULL	NULL	NULL	80.0	33.15	1.27	NULL
NULL						
2018-11-27 07:00:00	2018-11-27 08:00:00			66.3	144.08	NULL
11.87	NULL	20.98	10.32	311.88	25.67	
NULL	NULL	NULL	NULL	NULL		NULL
NULL	NULL	NULL	110.0	24.25	1.45	NULL
NULL						
2018-12-20 21:00:00	2018-12-20 22:00:00			44.52	136.66	NULL
11.19	NULL	31.83	16.15	305.54	12.51	
NULL	NULL	NULL	NULL	NULL		NULL
NULL	NULL	NULL	145.0	40.83	1.38	NULL
NULL						
2018-12-23 00:00:00	2018-12-23 01:00:00			48.43	137.88	NULL
75.0	NULL	45.32	11.92	305.42	14.78	NULL
NULL	NULL	NULL	NULL	NULL		NULL
NULL	NULL	NULL	39.14	5.09	NULL	NULL
NULL						
2018-12-23 01:00:00	2018-12-23 02:00:00			47.34	127.87	NULL
52.62	NULL	47.72	15.4	385.63	10.11	
NULL	NULL	NULL	NULL	NULL		NULL
NULL	NULL	NULL	NULL	36.46	5.2	NULL
NULL						
2018-12-23 04:00:00	2018-12-23 05:00:00			48.48	82.24	NULL
51.08	NULL	57.45	11.15	327.01	7.04	
NULL	NULL	NULL	NULL	NULL		NULL
NULL	NULL	NULL	NULL	31.29	2.36	NULL
NULL						
2018-12-23 21:00:00	2018-12-23 22:00:00			37.65	155.48	NULL
24.46	NULL	54.75	16.14	327.01	8.22	
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			52.82	157.54	NULL
48.95	NULL	55.77	16.47	392.2	10.16	
NULL	NULL	NULL	NULL	NULL		NULL
NULL	NULL	NULL	NULL	44.99	4.62	NULL
NULL						
2018-12-23 23:00:00	2018-12-24 00:00:00			58.32	138.62	NULL
9.8	NULL	49.65	17.66	372.22	22.12	NULL
NULL	NULL	NULL	NULL	NULL		NULL
NULL	NULL	NULL	37.01	2.06	NULL	NULL
NULL						
2018-12-26 21:00:00	2018-12-26 22:00:00			52.2	159.29	NULL

7.91	NULL	44.17	16.57	366.29	20.34	NULL
NULL		NULL		NULL	NULL	NULL
NULL	NULL	130.0	49.6	1.52	NULL	NULL
NULL						

```

+-----+-----+-----+-----+-----+-----+
|-----+-----+-----+-----+-----+-----+
|-----+-----+-----+-----+-----+-----+
|-----+-----+-----+-----+-----+-----+
|-----+-----+-----+-----+-----+-----+
|-----+-----+-----+-----+-----+-----+

```

only showing top 20 rows

```

[50]: // Job 8: Aggregate by max NO
val maxNODF = parquetDF.groupBy("From Date", "To Date").agg(max("NO (ug/m3)").
    ↪as("max_no"))
maxNODF.show()

```

VBox()

```

FloatProgress(value=0.0, bar_style='info', description='Progress:',
    ↪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	5.8
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

```
[52]: // Job 9: Aggregate by min PM10
val minPM10DF = parquetDF.groupBy("From Date", "To Date").agg(min("PM10 (ug/
    ↪m3)").as("min_pm10"))
minPM10DF.show()
```

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

```
[53]: // Job 10: Filter by low SO2 levels
val lowSO2DF = parquetDF.filter($"SO2 (ug/m3)" < 50)
lowSO2DF.show()
//
// writeToMySQL(lowSO2DF, "low_so2")
```

VBox()

```
FloatProgress(value=0.0, bar_style='info', description='Progress:',
  ↳ layout=Layout(height='25px', width='50%'),...
```

```
lowS02DF: 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)|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
+-----+-----+
|2023-02-10 09:00:00|2023-02-10 10:00:00|          40.0|          139.5|          2.45|
7.95|          6.25|          19.45|          14.55|          0.3|          61.85|          0.1|
0.2|          NULL|          NULL|          NULL|          56.0|
1.85|          239.5|          219.0|          NULL|          25.65|          0.0|          NULL|
NULL|
|2023-02-10 10:00:00|2023-02-10 11:00:00|          18.75|          92.25|          1.95|
5.88|          4.67|          21.62|          15.1|          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|
|2023-02-10 11:00:00|2023-02-10 12:00:00|          13.25|          96.5|          1.78|
5.35|          4.28|          23.6|          14.18|          0.27|          66.17|          0.0|
0.2|          NULL|          NULL|          NULL|          53.25|
1.95|          251.0|          264.5|          NULL|          26.9|          0.0|          NULL|
NULL|
|2023-02-10 12:00:00|2023-02-10 13:00:00|          15.25|          115.0|          1.6|
5.0|          3.97|          24.45|          13.05|          0.26|          70.33|          0.0|
0.2|          NULL|          NULL|          NULL|          48.5|
1.75|          238.75|          175.75|          NULL|          27.55|          0.0|          NULL|
NULL|
|2023-02-10 13:00:00|2023-02-10 14:00:00|          18.5|          93.5|          1.62|
5.42|          4.23|          24.6|          11.62|          0.27|          71.47|          0.0|
0.2|          NULL|          NULL|          NULL|          45.0|
1.95|          248.0|          266.25|          NULL|          28.0|          0.0|          NULL|
NULL|
|2023-02-10 14:00:00|2023-02-10 15:00:00|          12.0|          83.0|          1.55|
4.47|          3.62|          25.93|          8.72|          0.27|          73.6|          0.0|
0.2|          NULL|          NULL|          NULL|          40.75|
```


1.77	237.75	235.25	NULL	28.45	0.0	NULL
NULL						
2023-02-10 15:00:00	2023-02-10 16:00:00			11.75	84.0	1.45
4.75	3.67	27.02	7.2	0.31	76.92	0.0
0.2	NULL	NULL	NULL	NULL	39.0	
1.73	241.5	185.5	NULL	28.67	0.0	NULL
NULL						
2023-02-10 16:00:00	2023-02-10 17:00:00			10.25	86.75	1.58
4.7	3.77	24.95	7.7	0.32	79.83	0.0
0.2	NULL	NULL	NULL	NULL	39.5	
1.7	240.5	126.0	NULL	28.67	0.0	NULL
NULL						
2023-02-10 17:00:00	2023-02-10 18:00:00			13.5	84.0	1.6
5.75	4.33	23.18	12.12	0.33	80.42	0.0
0.2	NULL	0.03	NULL	42.75		
0.95	210.25	73.0	NULL	28.33	0.0	NULL
NULL						
2023-02-10 18:00:00	2023-02-10 19:00:00			20.0	101.75	1.48
11.43	7.3	21.72	11.38	0.4	67.52	
0.0	0.2	0.1	0.1	NULL		
50.0	0.55	201.5	53.75	NULL	27.53	0.0
NULL						
2023-02-10 19:00:00	2023-02-10 20:00:00			22.0	132.0	1.3
18.52	10.9	19.45	7.67	0.47	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						
2023-02-10 20:00:00	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	0.1	
55.0	0.35	216.0	53.0	NULL	26.58	0.0
NULL						
2023-02-10 21:00:00	2023-02-10 22:00:00			41.75	226.25	1.42
22.15	12.93	20.15	6.7	0.48	49.97	
0.12	1.82	1.05	1.05	0.3	NULL	
54.75	0.4	140.0	53.25	NULL	26.4	0.0
NULL						
2023-02-10 22:00:00	2023-02-10 23:00:00			23.75	95.75	2.05
25.42	15.2	18.77	5.7	0.51	37.43	
0.0	0.45	0.05	0.1	NULL		
63.5	0.1	97.75	53.75	NULL	25.53	0.0
NULL						
2023-02-10 23:00:00	2023-02-11 00:00:00			19.25	94.75	2.27
40.77	23.55	18.17	4.75	0.64	15.22	
0.08	0.57	0.1	0.12	0.03	NULL	
70.0	0.15	130.25	54.0	NULL	25.05	0.0
NULL						
2023-02-11 00:00:00	2023-02-11 01:00:00			37.5	129.25	5.05

49.65	30.53	19.32	3.52	0.86	5.83	
0.12	1.32		0.15	0.42		0.17
75.0	0.1	96.75	54.0	NULL	24.73	0.0
NULL						
2023-02-11 01:00:00	2023-02-11 02:00:00		44.0	150.25	4.85	
39.47	24.92	20.1	3.1	0.58	19.35	
0.1	1.23		0.2	0.57		0.2
74.75	0.27	104.5	54.0	NULL	24.73	0.0
NULL						
2023-02-11 02:00:00	2023-02-11 03:00:00		27.25	98.25	1.45	
15.78	9.57	20.43	2.92	0.33	32.75	
0.03	0.38		0.03	0.15		0.03
78.5	0.1	76.0	53.5	NULL	24.38	0.0
NULL						
2023-02-11 03:00:00	2023-02-11 04:00:00		18.25	69.0	4.67	
36.67	23.3	21.8	2.7	0.38	11.58	
0.05	0.45		0.03	0.12		NULL
82.25	0.15	120.0	53.0	NULL	24.1	0.0
NULL						
2023-02-11 04:00:00	2023-02-11 05:00:00		24.25	90.75	7.57	
39.8	27.32	23.12	2.22	0.41	7.27	0.05
0.5	0.1		0.2	0.03		84.0
0.15	111.25	53.25	NULL	24.05	0.0	NULL
NULL						

```

+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+
-----+-----+

```

only showing top 20 rows

[]: