Code Logic - Retail Data Analysis

In this document, you will describe the code and the overall steps taken to solve the project.

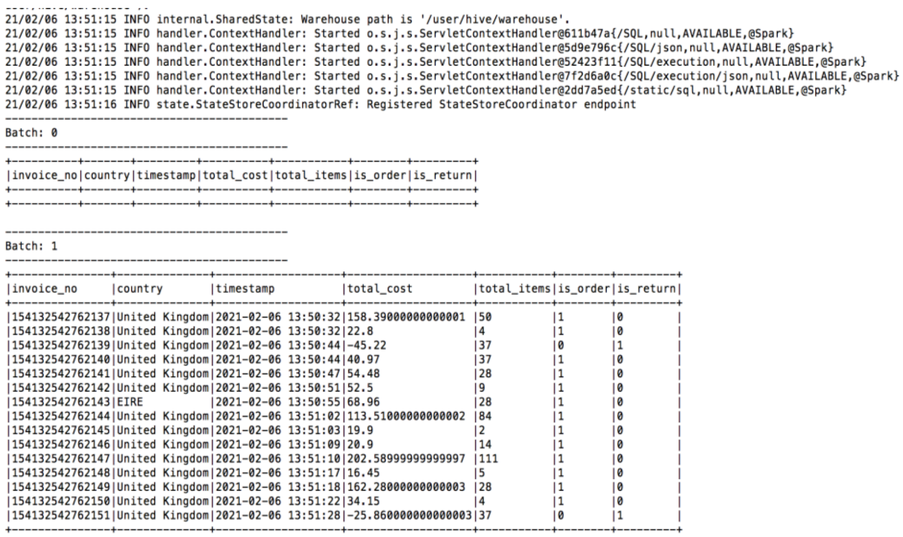
Step 1: Created the spark session and written code to read Data from Kafka

Step 2: Created the schema according to the datatypes, imported the data from the json formatted kafka topic value as the intial data in step 1 is of key value pair format.

Step 3 : Imported the necessary environment variables and created the functions to get the necessary columns of total cost, items

Step 4: Created the necessary columns calling the udf's created in step3.

Step 5: Created the final datastream to pull the required columns and written to the console. Below is the screenshot of the data streams written to console



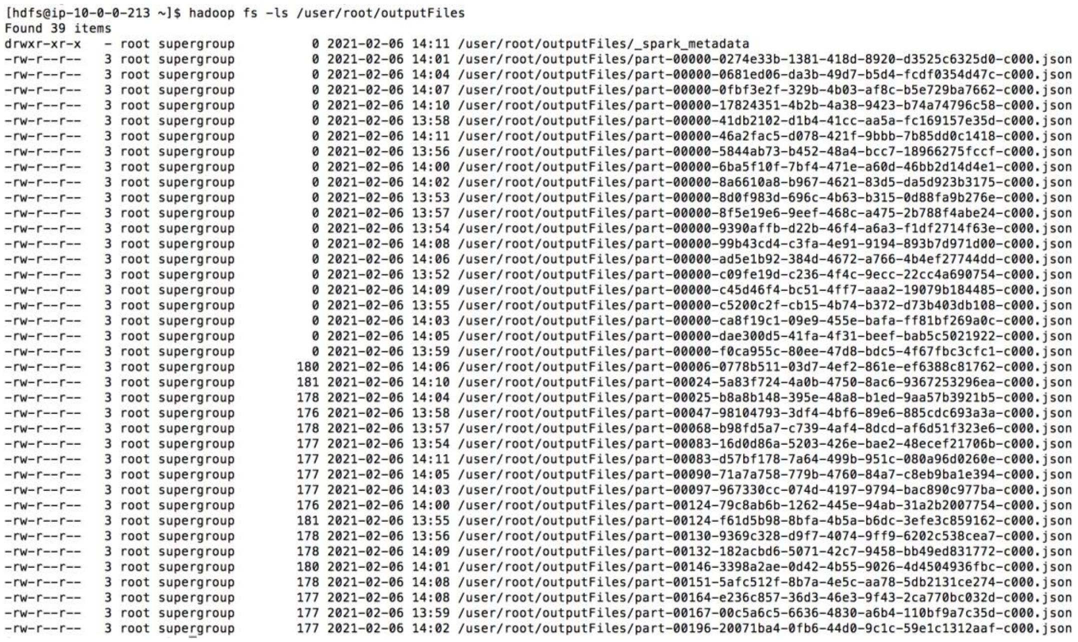
```
21/02/06 13:51:15 INFO internal.SharedState: Warehouse path is '/user/hive/warehouse'.
21/02/06 13:51:15 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@611b47a{/SQL,null,AVAILABLE,@Spark}
21/02/06 13:51:15 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@5d9e796c{/SQL/json,null,AVAILABLE,@Spark}
21/02/06 13:51:15 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@52423f11{/SQL/execution,null,AVAILABLE,@Spark}
21/02/06 13:51:15 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7f2d6a0c{/SQL/execution/json,null,AVAILABLE,@Spark}
21/02/06 13:51:15 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@2dd7a5ed{/static/sql,null,AVAILABLE,@Spark}
21/02/06 13:51:16 INFO state.StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
-------------------------------------------
Batch: 0
-------------------------------------------
+----------+-------+---------+----------+-----------+--------+---------+
|invoice_no|country|timestamp|total_cost|total_items|is_order|is_return|
+----------+-------+---------+----------+-----------+--------+---------+
+----------+-------+---------+----------+-----------+--------+---------+

-------------------------------------------
Batch: 1
-------------------------------------------
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|invoice_no     |country       |timestamp          |total_cost        |total_items|is_order|is_return|
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|154132542762137|United Kingdom|2021-02-06 13:50:32|158.39000000000001|50         |1       |0        |
|154132542762138|United Kingdom|2021-02-06 13:50:32|22.8              |4          |1       |0        |
|154132542762139|United Kingdom|2021-02-06 13:50:44|-45.22            |37         |0       |1        |
|154132542762140|United Kingdom|2021-02-06 13:50:44|40.97             |37         |1       |0        |
|154132542762141|United Kingdom|2021-02-06 13:50:47|54.48             |28         |1       |0        |
|154132542762142|United Kingdom|2021-02-06 13:50:51|52.5              |9          |1       |0        |
|154132542762143|EIRE          |2021-02-06 13:50:55|68.96             |28         |1       |0        |
|154132542762144|United Kingdom|2021-02-06 13:51:02|113.51000000000002|84         |1       |0        |
|154132542762145|United Kingdom|2021-02-06 13:51:03|19.9              |2          |1       |0        |
|154132542762146|United Kingdom|2021-02-06 13:51:09|20.9              |14         |1       |0        |
|154132542762147|United Kingdom|2021-02-06 13:51:10|202.58999999999997|111        |1       |0        |
|154132542762148|United Kingdom|2021-02-06 13:51:17|16.45             |5          |1       |0        |
|154132542762149|United Kingdom|2021-02-06 13:51:18|162.28000000000003|28         |1       |0        |
|154132542762150|United Kingdom|2021-02-06 13:51:22|34.15             |4          |1       |0        |
|154132542762151|United Kingdom|2021-02-06 13:51:28|-25.86000000000003|37         |0       |1        |
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
```

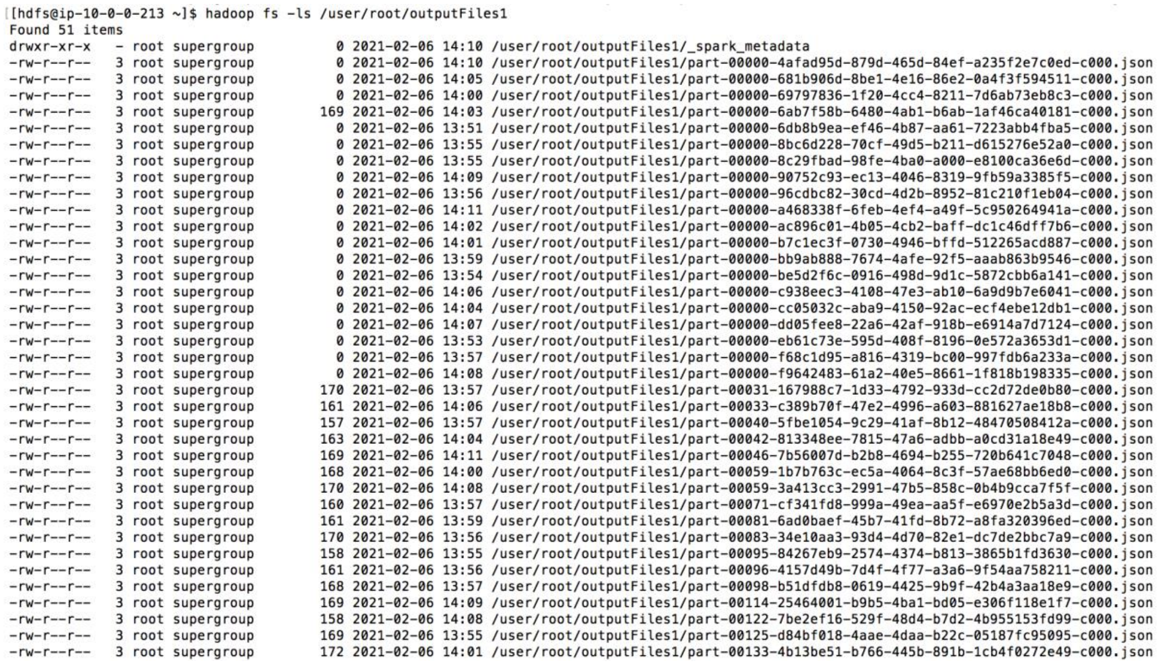
Step 6: Calculated time based KPI's with tumbling window of one minute on orders acroos the globe. Used the necessary aggregations and also the watermark

Step 7: Calculated time and country based KPI's with tumbling window of one minute on orders on a country basis.

Step8: All the above KPI's created in step 6 and step 7 are written to json files on hdfs folders created (outputFiles and outputFiles1). Below is the screenshots of the files created.



```
[hdfs@ip-10-0-0-213 ~]$ hadoop fs -ls /user/root/outputFiles
Found 39 items
drwxr-xr-x   - root supergroup          0 2021-02-06 14:11 /user/root/outputFiles/_spark_metadata
-rw-r--r--   3 root supergroup          0 2021-02-06 14:01 /user/root/outputFiles/part-00000-0274e33b-1381-418d-8920-d3525c6325d0-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:04 /user/root/outputFiles/part-00000-0681ed06-da3b-49d7-b5d4-fcdf0354d47c-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:07 /user/root/outputFiles/part-00000-0fbf3e2f-329b-4b03-af8c-b5e729ba7662-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:10 /user/root/outputFiles/part-00000-17824351-4b2b-4a38-9423-b74a74796c58-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:58 /user/root/outputFiles/part-00000-41db2102-d1b4-41cc-aa5a-fc169157e35d-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:11 /user/root/outputFiles/part-00000-46a2fac5-d078-421f-9bbb-7b85dd0c1418-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:56 /user/root/outputFiles/part-00000-5844ab73-b452-48a4-bcc7-18966275fccf-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:00 /user/root/outputFiles/part-00000-6ba5f10f-7bf4-471e-a60d-46bb2d14d4e1-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:02 /user/root/outputFiles/part-00000-8a6610a8-b967-4621-83d5-da5d923b3175-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:53 /user/root/outputFiles/part-00000-8d0f983d-696c-4b63-b315-0d88fa9b276e-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:57 /user/root/outputFiles/part-00000-8f5e19e6-9eef-468c-a475-2b788f4abe24-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:54 /user/root/outputFiles/part-00000-9390affb-d22b-46f4-a6a3-f1df2714f63e-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:08 /user/root/outputFiles/part-00000-99b43cd4-c3fa-4e91-9194-893b7d971d00-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:06 /user/root/outputFiles/part-00000-ad5e1b92-384d-4672-a766-4b4ef27744dd-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:52 /user/root/outputFiles/part-00000-c09fe19d-c236-4f4c-9ecc-22cc4a690754-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:09 /user/root/outputFiles/part-00000-c45d46f4-bc51-4ff7-aaa2-19079b184485-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:55 /user/root/outputFiles/part-00000-c5200c2f-cb15-4b74-b372-d73b403db108-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:03 /user/root/outputFiles/part-00000-ca8f19c1-09e9-455e-bafa-ff81bf269a0c-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:05 /user/root/outputFiles/part-00000-dae300d5-41fa-4f31-beef-bab5c5021922-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:59 /user/root/outputFiles/part-00000-f0ca955c-80ee-47d8-bdc5-4f67fbc3cfc1-c000.json
-rw-r--r--   3 root supergroup        180 2021-02-06 14:06 /user/root/outputFiles/part-00006-0778b511-03d7-4ef2-861e-ef6388c81762-c000.json
-rw-r--r--   3 root supergroup        181 2021-02-06 14:10 /user/root/outputFiles/part-00024-5a83f724-4a0b-4750-8ac6-9367253296ea-c000.json
-rw-r--r--   3 root supergroup        178 2021-02-06 14:04 /user/root/outputFiles/part-00025-b8a8b148-395e-48a8-b1ed-9aa57b3921b5-c000.json
-rw-r--r--   3 root supergroup        176 2021-02-06 13:58 /user/root/outputFiles/part-00047-98104793-3df4-4bf6-89e6-885cdc693a3a-c000.json
-rw-r--r--   3 root supergroup        178 2021-02-06 13:57 /user/root/outputFiles/part-00068-b98fd5a7-c739-4af4-8dcd-af6d51f323e6-c000.json
-rw-r--r--   3 root supergroup        177 2021-02-06 13:54 /user/root/outputFiles/part-00083-16d0d86a-5203-426e-bae2-48ecef21706b-c000.json
-rw-r--r--   3 root supergroup        177 2021-02-06 14:11 /user/root/outputFiles/part-00083-d57bf178-7a64-499b-951c-080a96d0260e-c000.json
-rw-r--r--   3 root supergroup        177 2021-02-06 14:05 /user/root/outputFiles/part-00090-71a7a758-779b-4760-84a7-c8eb9ba1e394-c000.json
-rw-r--r--   3 root supergroup        177 2021-02-06 14:03 /user/root/outputFiles/part-00097-967330cc-074d-4197-9794-bac890c977ba-c000.json
-rw-r--r--   3 root supergroup        176 2021-02-06 14:00 /user/root/outputFiles/part-00124-79c8ab6b-1262-445e-94ab-31a2b2007754-c000.json
-rw-r--r--   3 root supergroup        181 2021-02-06 13:55 /user/root/outputFiles/part-00124-f61d5b98-8bfa-4b5a-b6dc-3efe3c859162-c000.json
-rw-r--r--   3 root supergroup        178 2021-02-06 13:56 /user/root/outputFiles/part-00130-9369c328-d9f7-4074-9ff9-6202c538cea7-c000.json
-rw-r--r--   3 root supergroup        178 2021-02-06 14:09 /user/root/outputFiles/part-00132-182acbd6-5071-42c7-9458-bb49ed831772-c000.json
-rw-r--r--   3 root supergroup        180 2021-02-06 14:01 /user/root/outputFiles/part-00146-3398a2ae-0d42-4b55-9026-4d4504936fbc-c000.json
-rw-r--r--   3 root supergroup        178 2021-02-06 14:08 /user/root/outputFiles/part-00151-5afc512f-8b7a-4e5c-aa78-5db2131ce274-c000.json
-rw-r--r--   3 root supergroup        177 2021-02-06 14:08 /user/root/outputFiles/part-00164-e236c857-36d3-46e3-9f43-2ca770bc032d-c000.json
-rw-r--r--   3 root supergroup        177 2021-02-06 13:59 /user/root/outputFiles/part-00167-00c5a6c5-6636-4830-a6b4-110bf9a7c35d-c000.json
-rw-r--r--   3 root supergroup        177 2021-02-06 14:02 /user/root/outputFiles/part-00196-20071ba4-0fb6-44d0-9c1c-59e1c1312aaf-c000.json
```



```
[hdfs@ip-10-0-0-213 ~]$ hadoop fs -ls /user/root/outputFiles1
Found 51 items
drwxr-xr-x   - root supergroup          0 2021-02-06 14:10 /user/root/outputFiles1/_spark_metadata
-rw-r--r--   3 root supergroup          0 2021-02-06 14:10 /user/root/outputFiles1/part-00000-4afad95d-879d-465d-84ef-a235f2e7c0ed-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:05 /user/root/outputFiles1/part-00000-681b906d-8be1-4e16-86e2-0a4f3f594511-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:00 /user/root/outputFiles1/part-00000-69797836-1f20-4cc4-8211-7d6ab73eb8c3-c000.json
-rw-r--r--   3 root supergroup        169 2021-02-06 14:03 /user/root/outputFiles1/part-00000-6ab7f58b-6480-4ab1-b6ab-1af46ca40181-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:51 /user/root/outputFiles1/part-00000-6db8b9ea-ef46-4b87-aa61-7223abb4fba5-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:55 /user/root/outputFiles1/part-00000-8bc6d228-70cf-49d5-b211-d615276e52a0-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:55 /user/root/outputFiles1/part-00000-8c29fbad-98fe-4ba0-a000-e8100ca36e6d-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:09 /user/root/outputFiles1/part-00000-90752c93-ec13-4046-8319-9fb59a3385f5-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:56 /user/root/outputFiles1/part-00000-96cdbc82-30cd-4d2b-8952-81c210f1eb04-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:11 /user/root/outputFiles1/part-00000-a468338f-6feb-4ef4-a49f-5c950264941a-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:02 /user/root/outputFiles1/part-00000-ac896c01-4b05-4cb2-baff-dc1c46dff7b6-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:01 /user/root/outputFiles1/part-00000-b7c1ec3f-0730-4946-bffd-512265acd887-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:59 /user/root/outputFiles1/part-00000-bb9ab888-7674-4afe-92f5-aaab863b9546-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:54 /user/root/outputFiles1/part-00000-be5d2f6c-0916-498d-9d1c-5872cbb6a141-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:06 /user/root/outputFiles1/part-00000-c938eec3-4108-47e3-ab10-6a9d9b7e6041-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:04 /user/root/outputFiles1/part-00000-cc05032c-aba9-4150-92ac-ecf4ebe12db1-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:07 /user/root/outputFiles1/part-00000-dd05fee8-22a6-42af-918b-e6914a7d7124-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:53 /user/root/outputFiles1/part-00000-eb61c73e-595d-408f-8196-0e572a3653d1-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 13:57 /user/root/outputFiles1/part-00000-f68c1d95-a816-4319-bc00-997fdb6a233a-c000.json
-rw-r--r--   3 root supergroup          0 2021-02-06 14:08 /user/root/outputFiles1/part-00000-f9642483-61a2-40e5-8661-1f818b198335-c000.json
-rw-r--r--   3 root supergroup        170 2021-02-06 13:57 /user/root/outputFiles1/part-00031-167988c7-1d33-4792-933d-cc2d72de0b80-c000.json
-rw-r--r--   3 root supergroup        161 2021-02-06 14:06 /user/root/outputFiles1/part-00033-c389b70f-47e2-4996-a603-881627ae18b8-c000.json
-rw-r--r--   3 root supergroup        157 2021-02-06 13:57 /user/root/outputFiles1/part-00040-5fbe1054-9c29-41af-8b12-48470508412a-c000.json
-rw-r--r--   3 root supergroup        163 2021-02-06 14:04 /user/root/outputFiles1/part-00042-813348ee-7815-47a6-adbb-a0cd31a18e49-c000.json
-rw-r--r--   3 root supergroup        169 2021-02-06 14:11 /user/root/outputFiles1/part-00046-7b56007d-b2b8-4694-b255-720b641c7048-c000.json
-rw-r--r--   3 root supergroup        168 2021-02-06 14:00 /user/root/outputFiles1/part-00059-1b7b763c-ec5a-4064-8c3f-57ae68bb6ed0-c000.json
-rw-r--r--   3 root supergroup        170 2021-02-06 14:08 /user/root/outputFiles1/part-00059-3a413cc3-2991-47b5-858c-0b4b9cca7f5f-c000.json
-rw-r--r--   3 root supergroup        160 2021-02-06 13:57 /user/root/outputFiles1/part-00071-cf341fd8-999a-49ea-aa5f-e6970e2b5a3d-c000.json
-rw-r--r--   3 root supergroup        161 2021-02-06 13:59 /user/root/outputFiles1/part-00081-6ad0baef-45b7-41fd-8b72-a8fa320396ed-c000.json
-rw-r--r--   3 root supergroup        170 2021-02-06 13:56 /user/root/outputFiles1/part-00083-34e10aa3-93d4-4d70-82e1-dc7de2bbc7a9-c000.json
-rw-r--r--   3 root supergroup        158 2021-02-06 13:55 /user/root/outputFiles1/part-00095-84267eb9-2574-4374-b813-3865b1fd3630-c000.json
-rw-r--r--   3 root supergroup        161 2021-02-06 13:56 /user/root/outputFiles1/part-00096-4157d49b-7d4f-4f77-a3a6-9f54aa758211-c000.json
-rw-r--r--   3 root supergroup        168 2021-02-06 13:57 /user/root/outputFiles1/part-00098-b51dfdb8-0619-4425-9b9f-42b4a3aa18e9-c000.json
-rw-r--r--   3 root supergroup        169 2021-02-06 14:09 /user/root/outputFiles1/part-00114-25464001-b9b5-4ba1-bd05-e306f118e1f7-c000.json
-rw-r--r--   3 root supergroup        158 2021-02-06 14:08 /user/root/outputFiles1/part-00122-7be2ef16-529f-48d4-b7d2-4b955153fd99-c000.json
-rw-r--r--   3 root supergroup        169 2021-02-06 13:55 /user/root/outputFiles1/part-00125-d84bf018-4aae-4daa-b22c-05187fc95095-c000.json
-rw-r--r--   3 root supergroup        172 2021-02-06 14:01 /user/root/outputFiles1/part-00133-4b13be51-b766-445b-891b-1cb4f0272e49-c000.json
```