# PROJECT 5 : Write Query using Synapse Serverless SQL pool to read CSV, JSON and Parquet files.

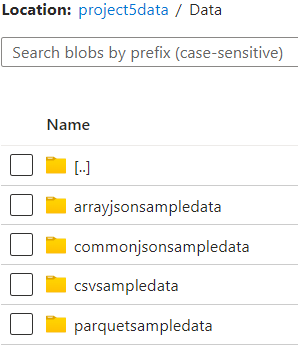
**Project overview:**

This project aims to optimize data ingestion and query performance on Azure Synapse Serverless SQL Pool by effectively handling various file formats (CSV, JSON, Parquet) from Azure Data Lake Storage Gen2.

**Task:** Query CSV, JSON, Parquet format files from Azure Data Lake Storage Gen2 using Synapse Serverless SQL Pool.

**Resources Used:**

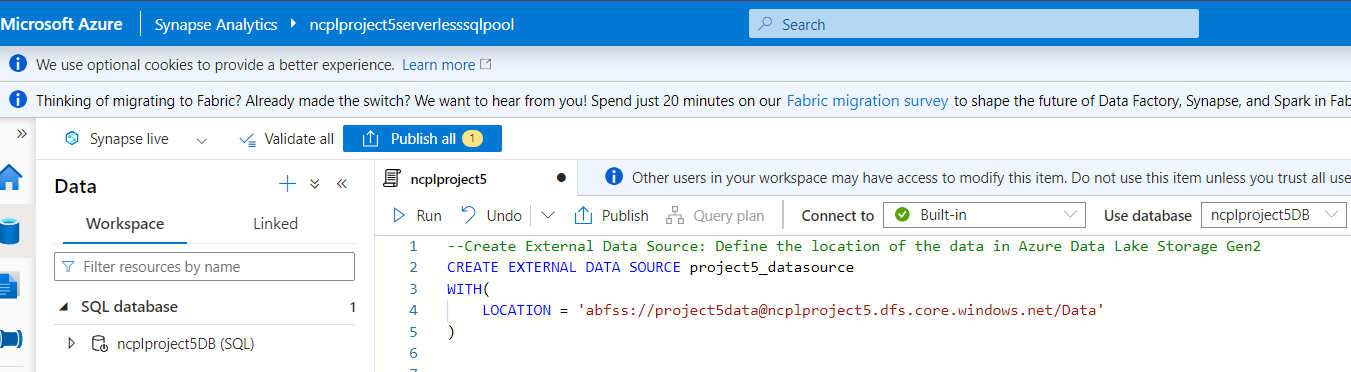
* Azure synapse analytics workspace: ncplproject5serverlesssqlpool
* ADLS: ncplproject5
* Container: project5data
* Directory: Data
* Folders:



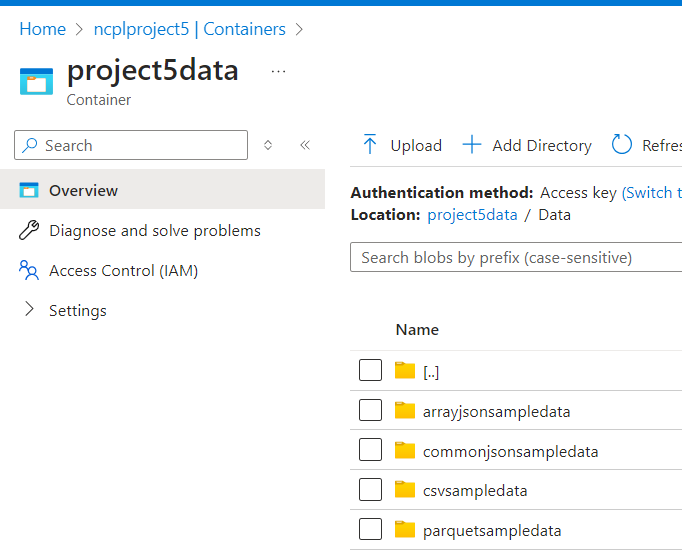
* Database: ncplproject5DB
* External data source: project5\_datasource
* External file formats: CSV, JSON, PARQUET

**Steps:**

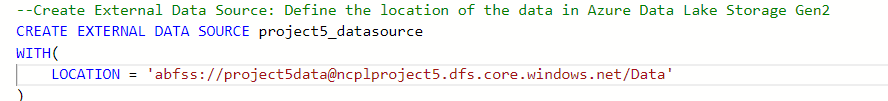
1. Created a workspace named ‘ncplproject5serverlesssqlpool’, an ADLS named ‘ncplproject5’ and a database named ‘ncplproject5DB’.



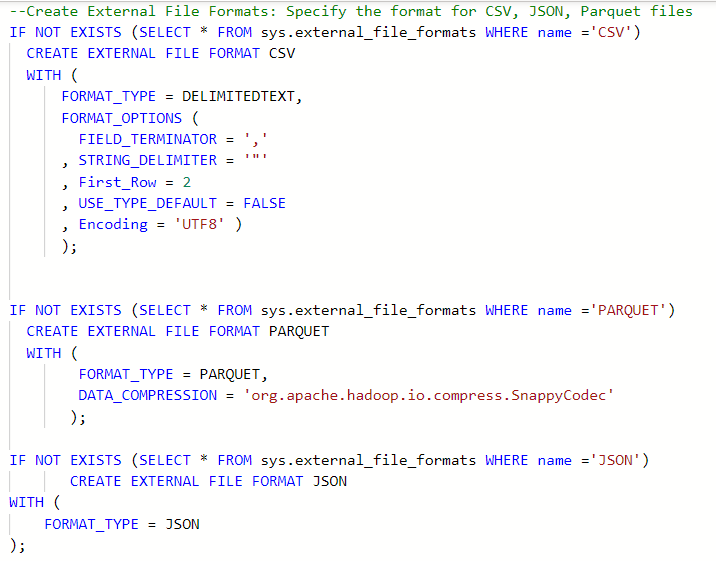
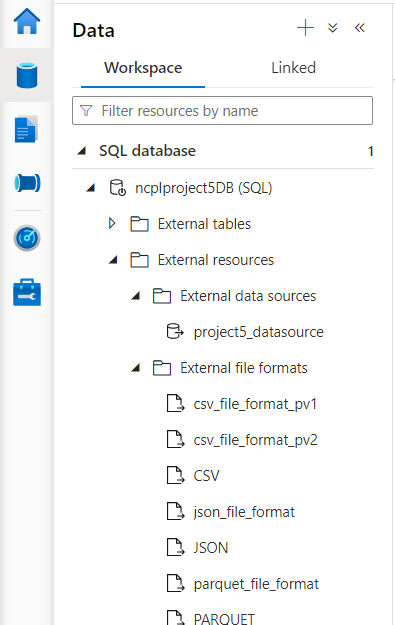
1. Created a container in ADLS and uploaded csv, json and parquet files by creating separate folders.



1. Created External Data Source: project5\_datasource

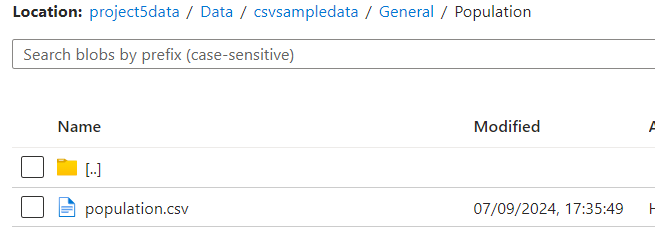


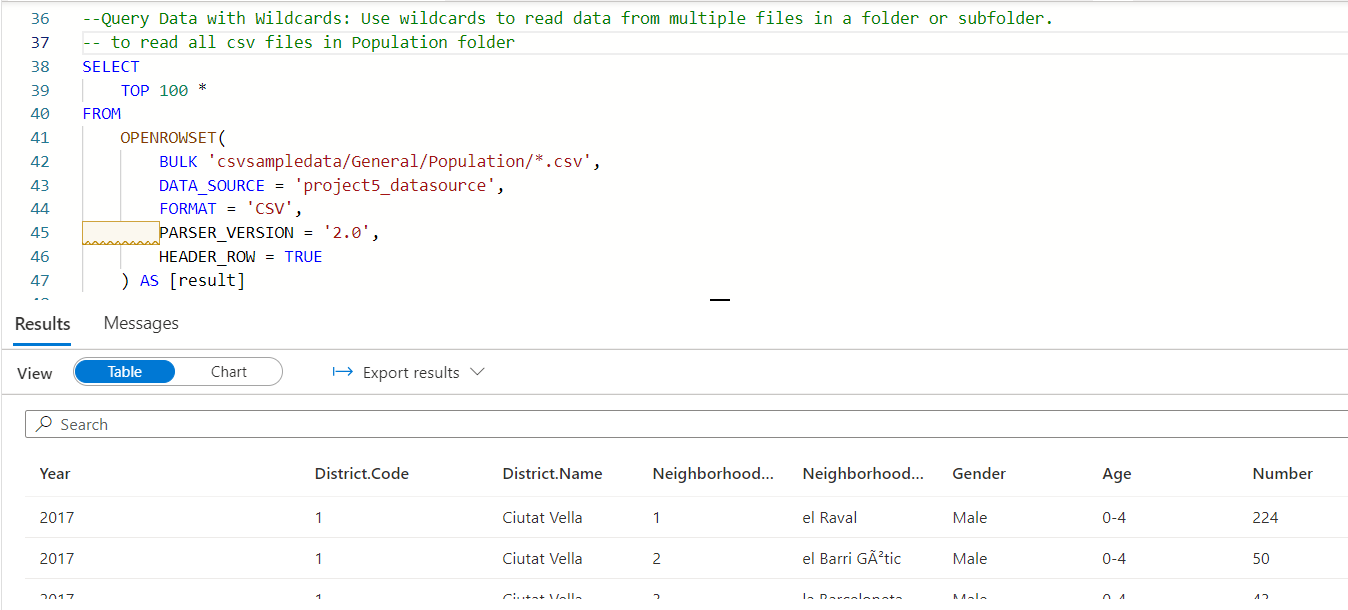
1. Created External File Formats:

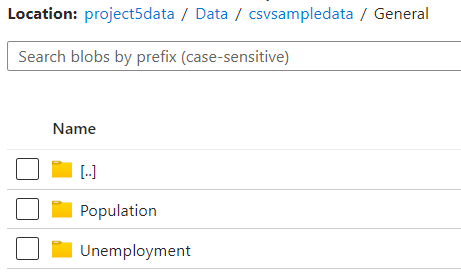
1. Query Data with Wildcards(\* - asterisk): Use wildcards to read data from multiple files in a folder or subfolder.

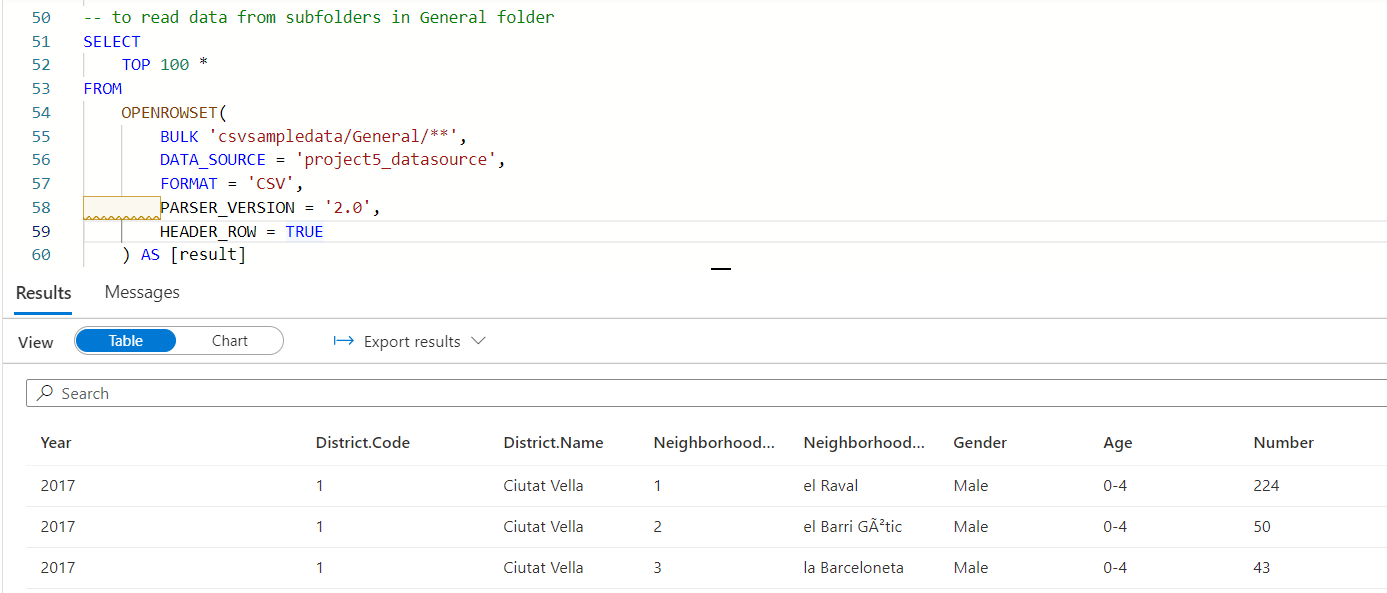
* To read all csv files in Population folder:



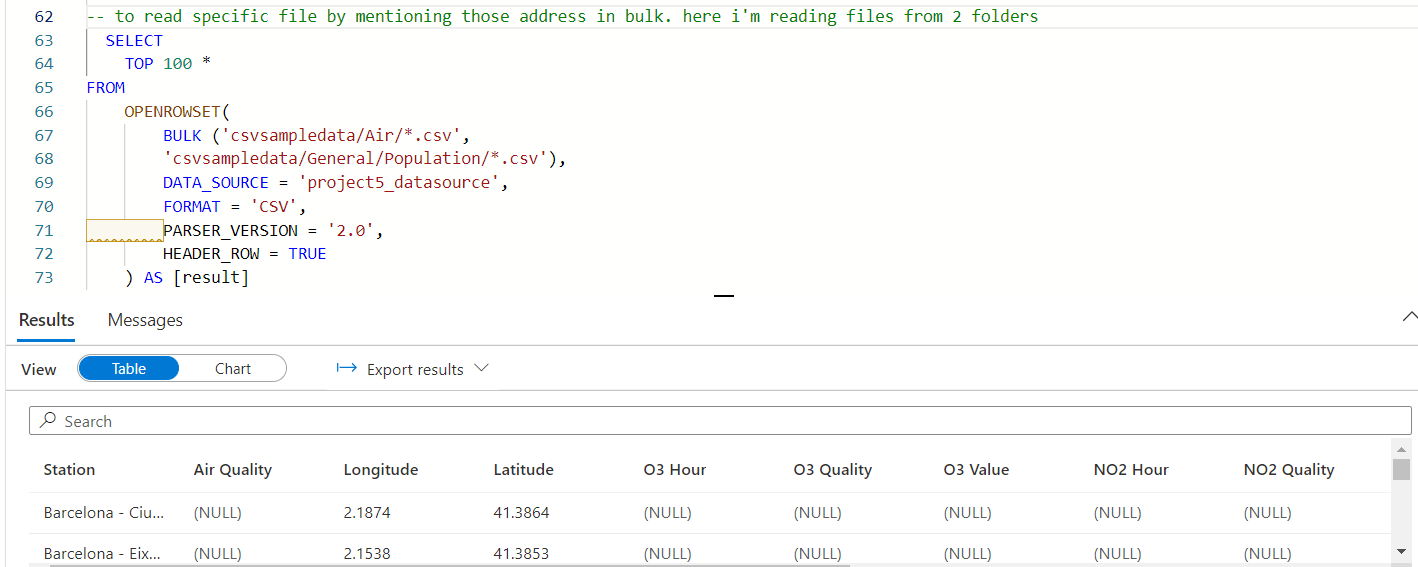


* to read data from subfolders in General folder:



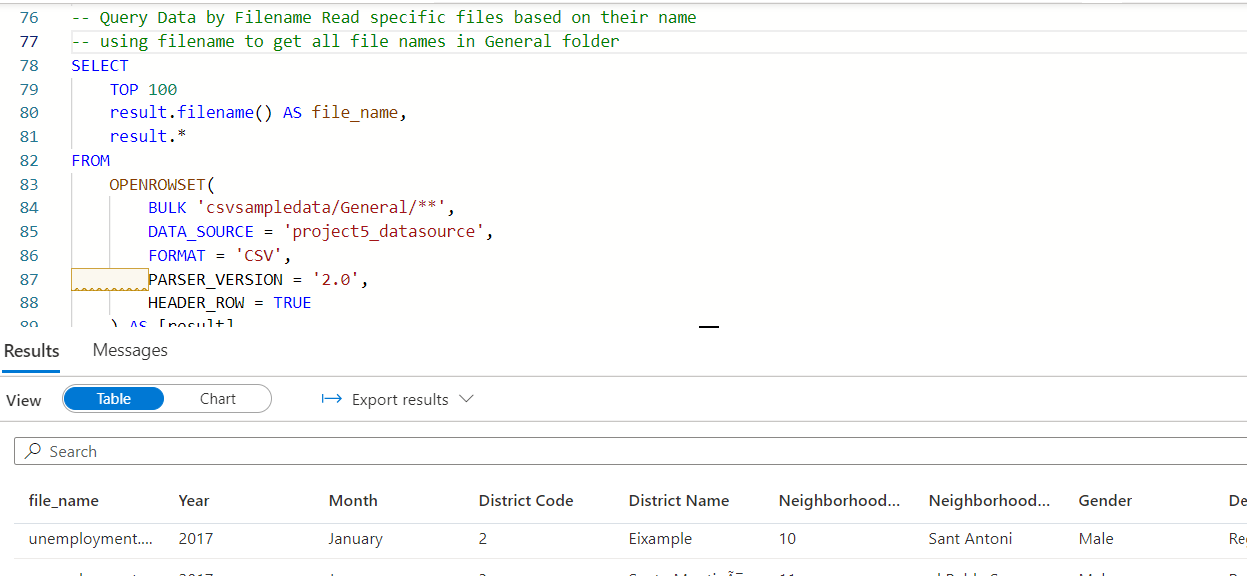


* to read specific files by mentioning those address in bulk. Here I’m reading files from 2 folders:

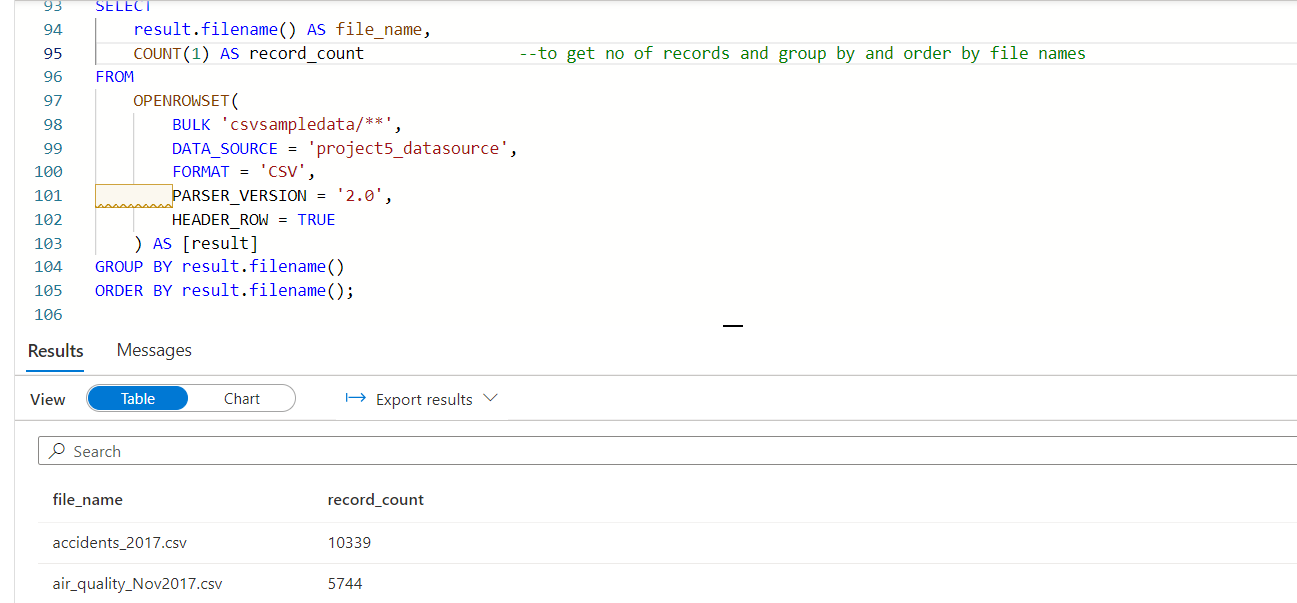


1. Query Data by Filename: Read specific files based on their name.

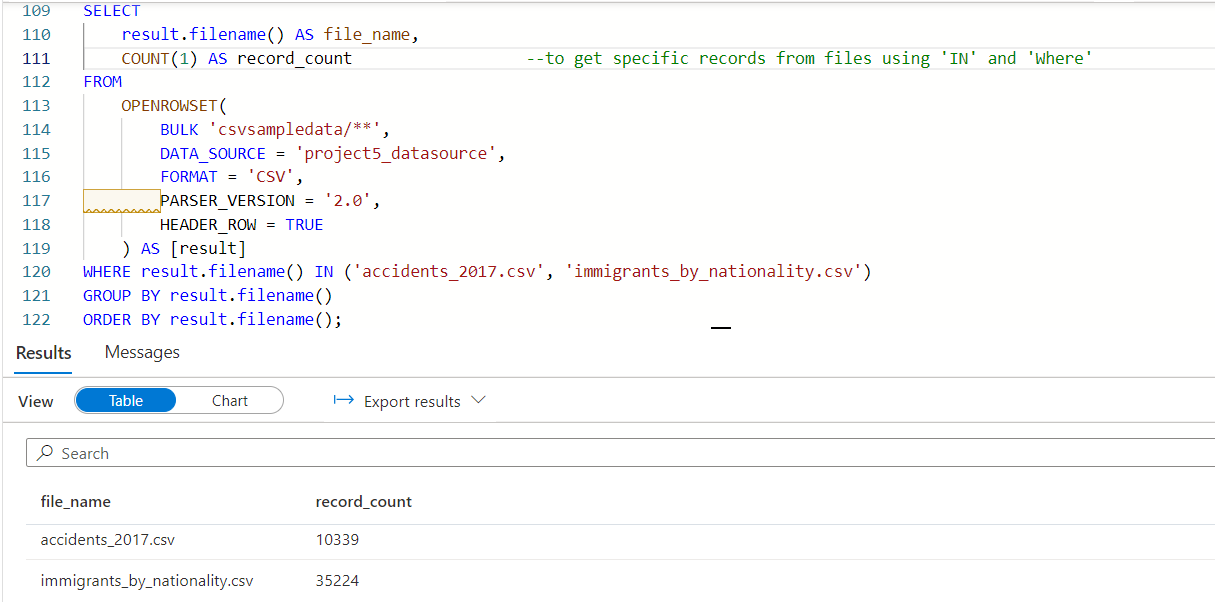
* using filename to get all file names in General folder



* to get no of records and group by and order by using file names

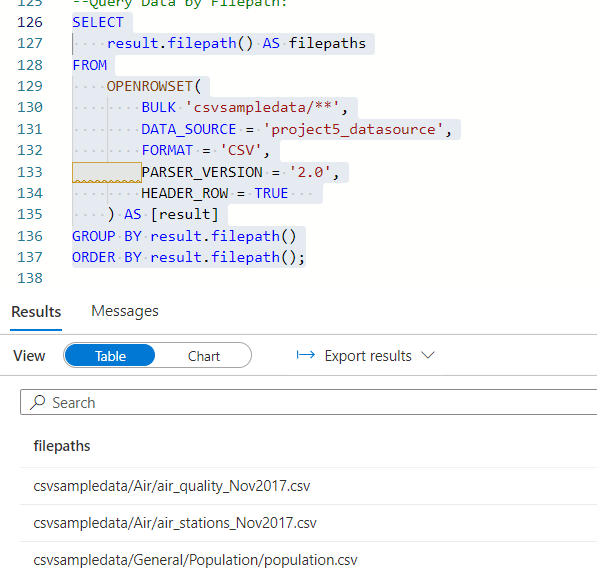


* to get count of specific files using 'IN' and 'Where'

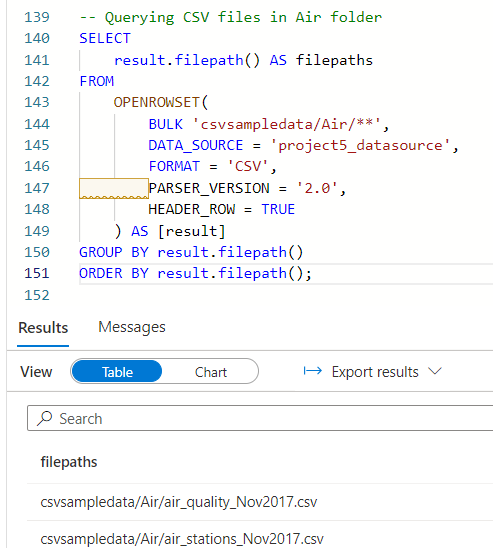


1. Query Data by Filepath:

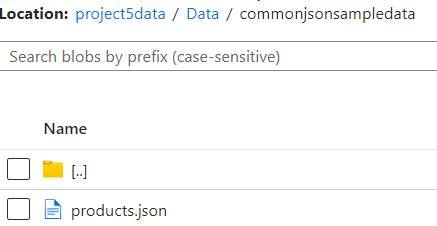
* to get file paths of all csv files.

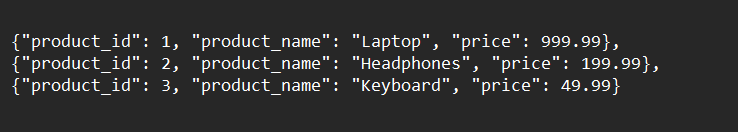


* To get file paths of csv files in specific folder(Air)

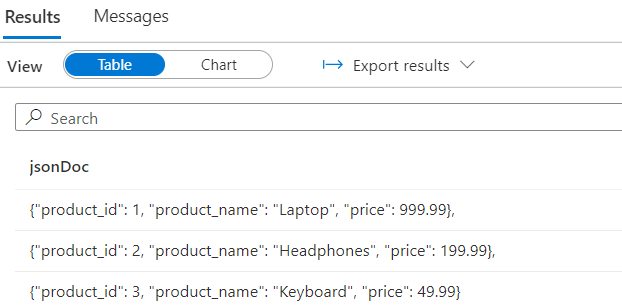
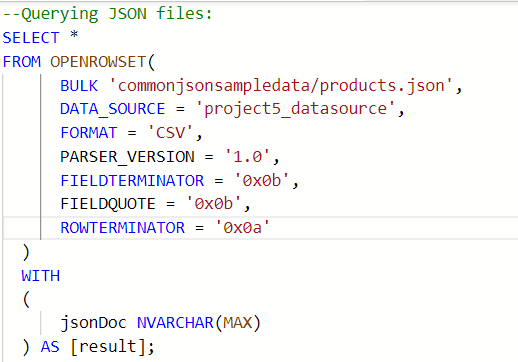


1. Query JSON files: uploaded sample file in ‘commonjsonsampledata’ folder.

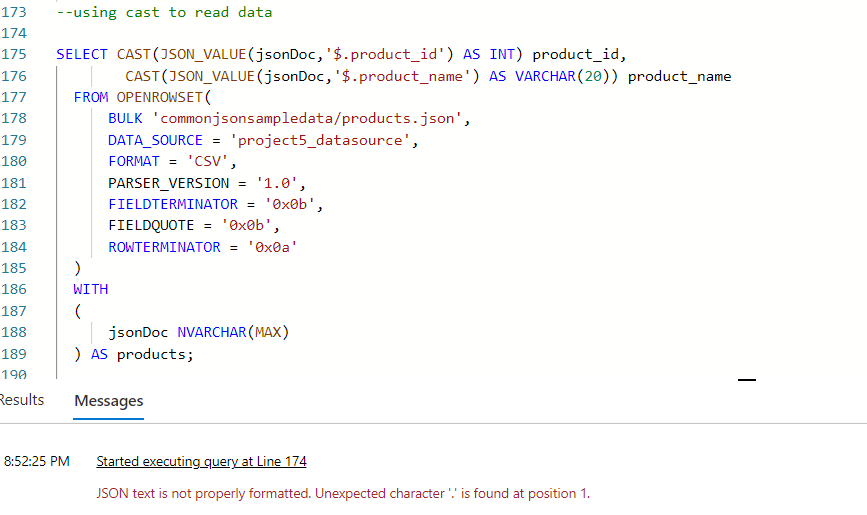




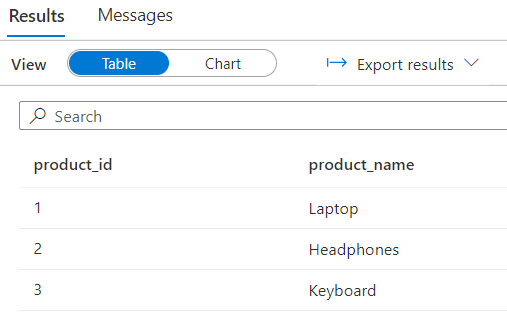
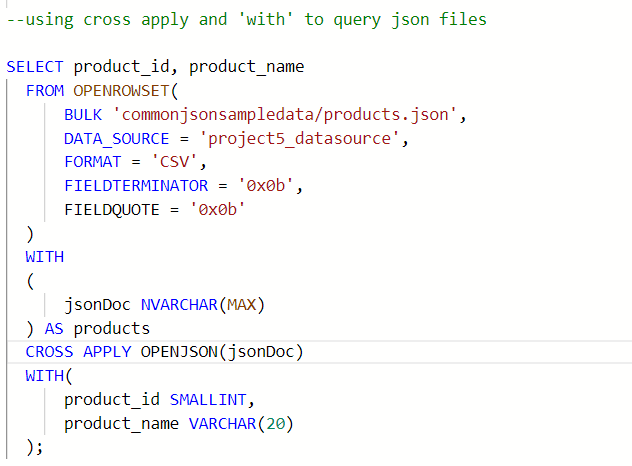
* General method of Querying JSON files: querying same as csv files but with some changes.



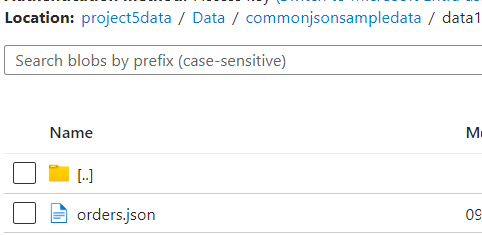
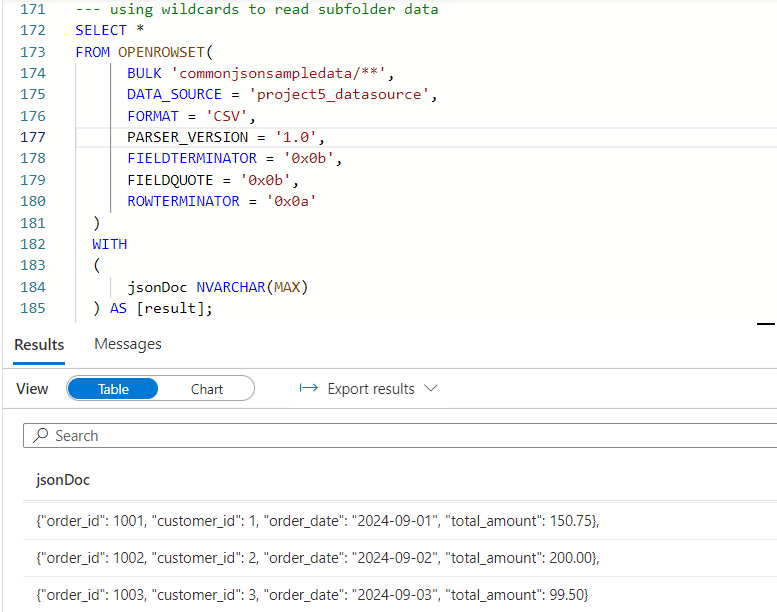
* using CAST to read data as we need: Got some error here, working on that.



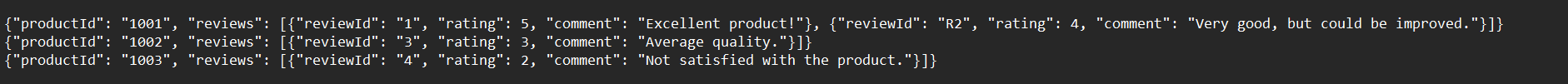
* JSON\_VALUE extracts a scalar value from a JSON string.
* If the no of keys are more in json file, it’s not easy to mention all those keys using CAST which means this method of using cast is not suitable. So we go for Cross apply method.
* using cross apply and 'with' to query json files:

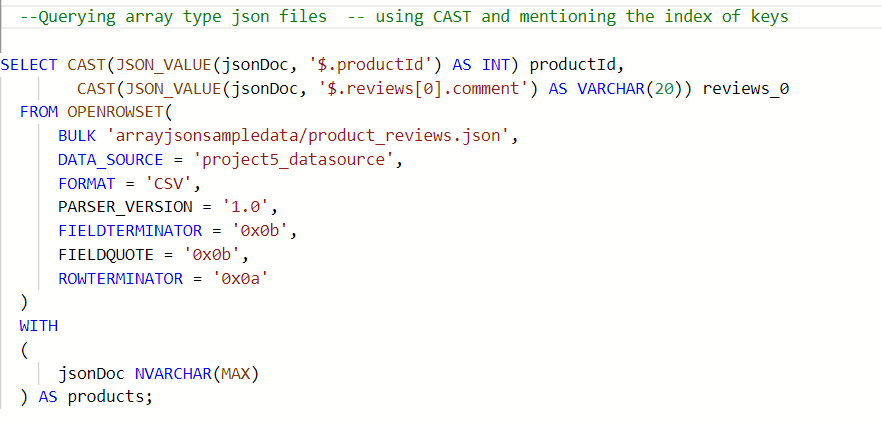


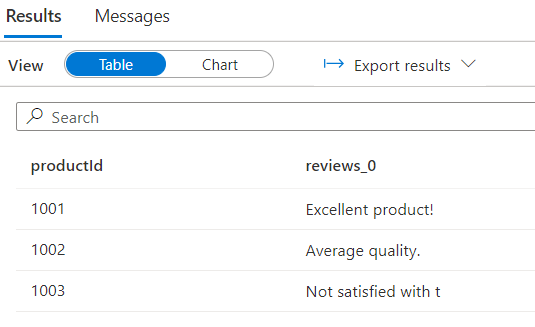
* Query json files: Using wildcards to read the subfolder data :



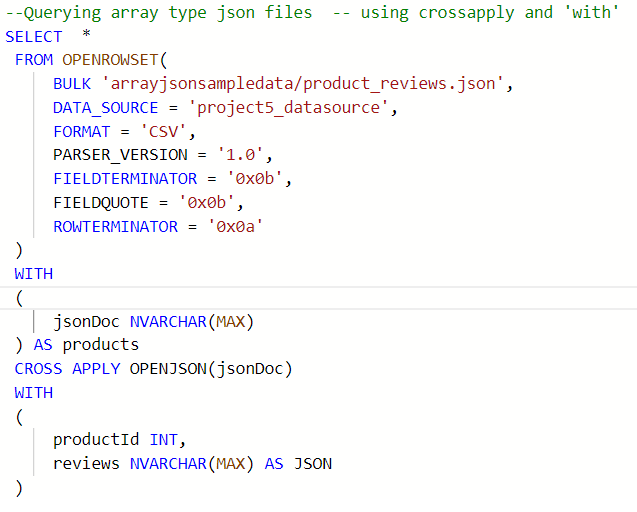
* Querying array type json files -- using CAST and mentioning the index of keys:





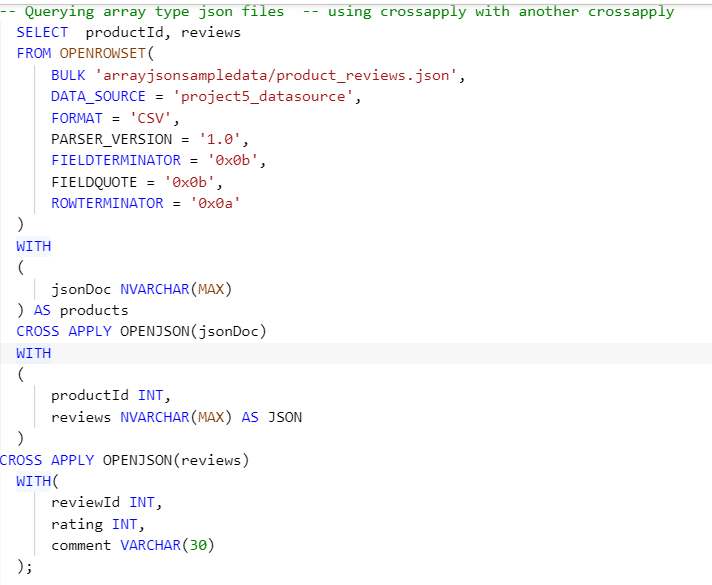


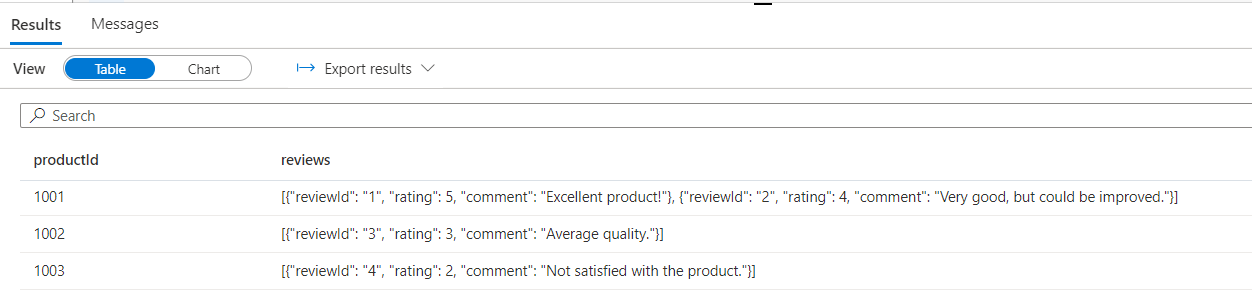
* Querying array type json files -- using crossapply and 'with':





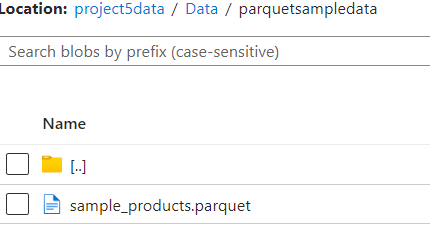
* Querying array type json files -- using crossapply with another crossapply:

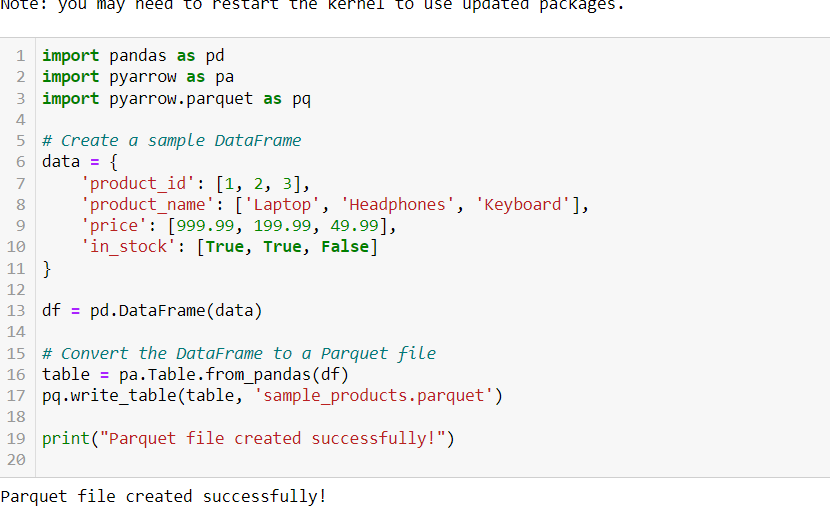




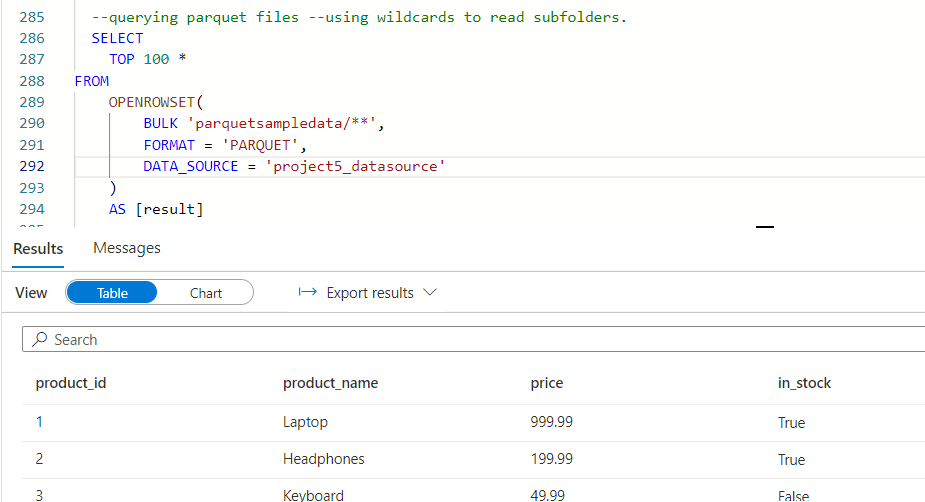
1. Query Parquet files

* Created parquet file using jupyter notebook and uploaded in ‘parquetsampledata’ folder

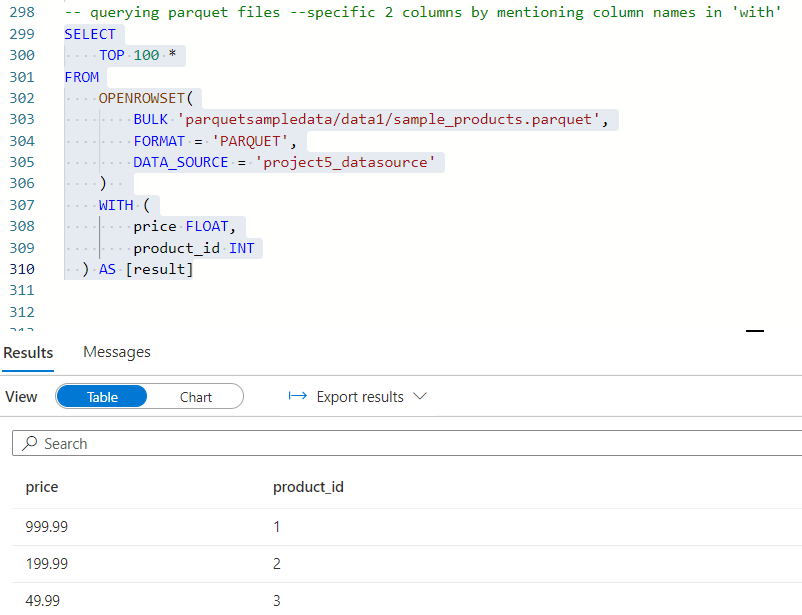




* Querying parquet files: using wildcards to read subfolder data



* querying parquet files –reading specific 2 columns by mentioning column names in 'with’:



ERRORS FACED

|  |  |  |  |
| --- | --- | --- | --- |
| ERROR | DESCRIPTION | RESOLUTION | WHAT CAN BE DONE TO AVOID THE ERROR |
| Incorrect schema of input data | Read wrong data types like there was a column which I tried to read as an int but it showed some error as there was a character in ID column | Corrected the data and it worked. | Validate data before processing it. |
| Row terminator error | Row terminator error while reading json file | Removed it and that worked | Ensure your JSON file is well-formed.  Validate it before using. |
| JSON file reading error | While using CAST to read data from json files got error due to format issue. |  |  |

CONCLUSION:

Read CSV, Json and Parquet Format files using Synapse Serverless SQL pool. Made use of Use wildcard character to read folder and subfolder data, Used Filename and Filepath to read data along with the file name of each partition.