



ABSTRACT

Azure Synapse Analytics (ASA) is a limitless analytics service that combines enterprise SQL data warehousing and big data analytics services.

The main objective of this exploration is to assess the connection between HPCC Systems and leverage their capabilities. This will explore ASA's data warehousing, visualization, and efficiency attributes compared to HPCC Systems features.

Data Collection

Data was collected into the ASA file system through a Secure File Transfer Protocol (SFTP) connection to the HPCC Systems landing zone in supported file formats - XML, Parquet, CSV - for querying and analysis. JSON files were ingested through the SFTP connection and queried through the REST API

OBJECTIVE/METHODS

The value of this project is to evaluate any better approaches for querying, data transformation, and data delivery.

Methods (1 and 2, 3 and 4 are related):

- 1) SFTP connecting ASA to HPCC Systems Landing Zone
- 2) REST API connecting ASA to HPCC Systems Roxie Cluster for JSON
- 3) Integrated data pipelining
- 4) Dataflow for data transformation (Replicating HPCC Systems Data Tutorial)

ACKNOWLEDGMENT/ RESOURCES

I would like to extend my gratitude to my mentors - Xiaoming Wang and Michael Gardner - for their continued support throughout the project. I would also like to thank Azure Support Engineers for helping with troubleshooting issues and debugging components when research and exploration was difficult. Their collective help allowed for this initial project on Synapse Analytics to be as insightful as it has been so far. I hope that the HPCC Systems platform team will further leverage this investigation in the future.

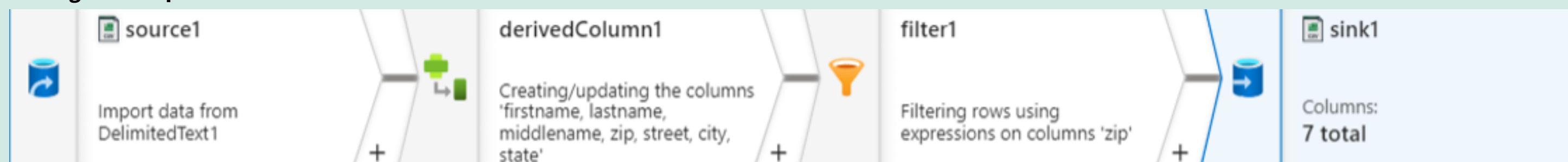
References:

<https://www.youtube.com/watch?v=LLrjaVdBuM0&t=7957s>

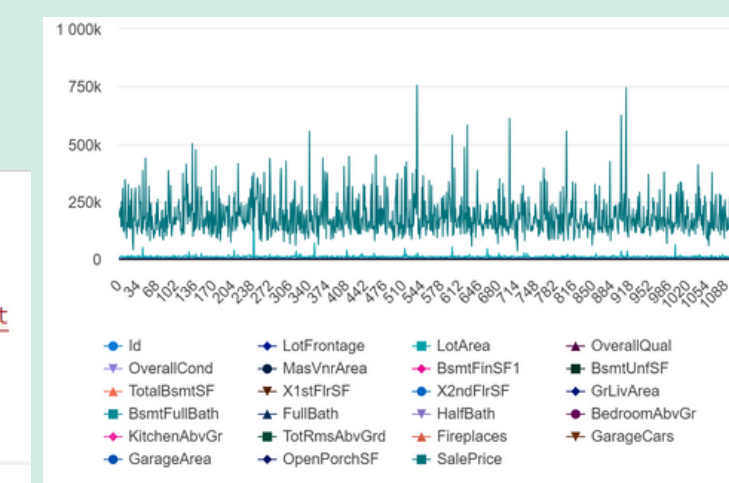
https://www.youtube.com/playlist?list=PLMWaZteqtEaIZxPCw_0AO1GsQESq3hZc6

RESULTS

Integrated Pipeline: Dataflow

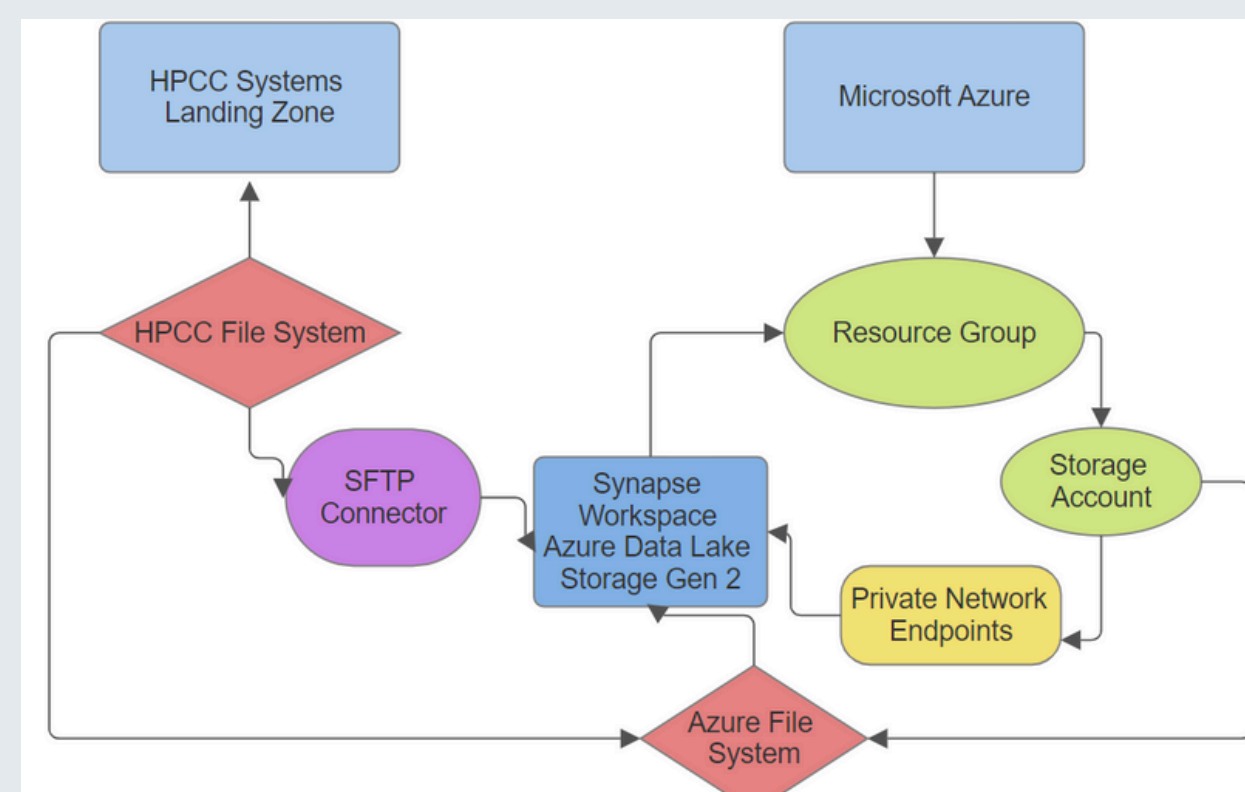


- 1) Loading in a sample file sales_data.csv from SFTP connection to landing zone, querying, and visualizing its data



```
SELECT *
FROM OPENROWSET(
    BULK 'https://dlsshounakj.dfs.core.windows.net/
    /fsshounakj/sales_data.csv',
    FORMAT = 'CSV',
    HEADER_ROW = TRUE,
    PARSER_VERSION='2.0'
) AS [result]
```

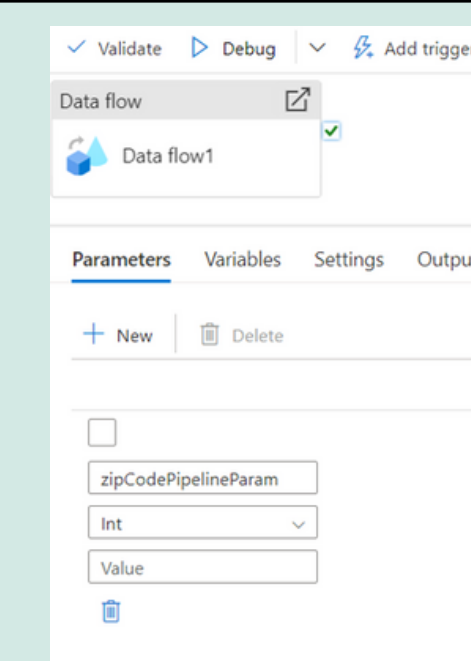
IMPLEMENTATION/ DISCUSSION



This diagram shows deployments and tasks for the SFTP connection. ASA has also proven to be quite cost-effective as costs incurred are primarily part of the deployment process and executing queries, data workloads, and tasks are not as expensive.

- 2) Integrated pipeline triggers data flow to replicate the HPCC Systems Data Tutorial

C1	C2	C3	C4	C5	C6	C7
LOUDONE	ROVEGNO	N	23832	239 E 79TH ST ...	CHESTERFIELD	VA
HOLT	COWIT	D	23832	23 ROBERTS DR	CHESTERFIELD	VA
NIYATI	MACKESY	V	23832	101 HIGHLAND...	CHESTERFIELD	VA
TERECA	ESHUN	(NULL)	23832	3842 N PROGR...	CHESTERFIELD	VA
KAYEANN	SAUBER	(NULL)	23832	1565 MACOPIN...	CHESTERFIELD	VA
VANDY	GUSE	W	23832	4516 BALLY GA...	CHESTERFIELD	VA
ALAZNE	BODO	M	23832	2 BROOKWOO...	CHESTERFIELD	VA
HYDER	MILLIARD	(NULL)	23832	888 8TH AVE A...	CHESTERFIELD	VA
AHARONA	MEWES	(NULL)	23832	540 PACIFIC ST ...	CHESTERFIELD	VA
PERSIAN	NEWBURG	F	23832	12 BESSIE LN	CHESTERFIELD	VA
BUNNINDER	GRIMSICH	T	23832	42 RED BUD RD	CHESTERFIELD	VA



Output for input parameter '23832' as a relational table

CONCLUSION

Azure Synapse Analytics has been a reliable tool for data transformation and processing. Overall, ASA's features that connect it to HPCC Systems allow for comparative analysis.

- 1) ASA has good visualization capabilities. Its dynamic visualization exceeds some of the static capabilities of HPCC Systems data visualization tools. This can be further enhanced in connection with Power BI's dashboarding features as well.
- 2) ASA's development tools are easy to set up for data transformation and querying in the context of the HPCC Systems data tutorial replication, however, ASA is not as time efficient as the Roxie query.
- 3) Built-in integration capabilities for ASA when compared to HPCC Systems. ASA's linked services have a code-free, user-friendly option to connect to external data sources.