**AZURE ETL PROJECT QUESTIONS**

**1.Why should one use Azure Key Vault when working in the Azure environment? What are the pros and cons? What are the alternatives?**

Weuse AKV to protect data in the cloud, and securely store and control access of secrets

Pros: one location to store the application secrets , easier access, Provides integration with other ِAzure services like ADF,GUI was quite easy

For me I did not get the cons yet but may be it pricey

**alternatives :**AWS Key Management Service (KMS)

**2. How do you achieve loop functionality within a Azure Data Factory pipeline? Why would you need to use this functionality in a data pipeline?**

You can use foreach loops to execute the loop functionality within a Azure Data Factory pipeline.set of activities or pipelines can achieve multiple times, with different values each time.

It is useful to use the same pipeline to perform transformations for multiple files instead of a different pipeline for each file if the transformation you want to apply is the same.

**3. What are expressions in Azure Data Factory? How are they helpful when designing a data pipeline? Please explain with an example.**

Expressions are JSON based formulas used to create a sequence involving variables and parameters.

Helpful because you can use set up azure data factory pipeline and pass external values into pipelines, datasets, linked services, and data flows.

Example – to read in a file instead of writing the value as

‘filename.csv’

you can use an expression like *concat* –

Concat(@filename(), ‘.’, @filetype())

the pipeline takes inputPath and outputPath parameters

“@pipeline().parameters.inputPath"

**4. What are the pros and cons of parametrizing a dataset’s activity in Azure Data Factory?**

dynamic content – allows you to create generic dataset into an expression or sequence instead of hardcoding.

Lower maintenance – with parametrizing you don’t need to keep updating the input format if new files are uploaded as long as they follow the same basic naming convention.

Cons – If expressions are complex the parametrized values can become difficult to read and interpret by someone else

**5. What are the different supported file formats and compression codecs in Azure Data Factory? When will you use a Parquet file over an ORC file? Why would you choose an AVRO file format over a Parquet file format?**

Supported file formats: Avro format.

* Binary format.
* Delimited text format.
* Excel format.
* JSON format.
* ORC format.
* Parquet format.
* XML format.

ORC Vs Parquet:

Parquet is more optimized of storing nested data whereas ORC is more capable of Predicate Pushdown – filtering data while processing. Additionally ORC is more compression efficient

Parquet vs Avro:

Parquet is columnar based whereas Avro is row based. This means Parquet is better for reading/ analytical queries whereas Avro is more optimized for writing. Parquet is ideal for querying a subset of columns in a multi-column table – since it’s optimized for nested data while AVRO is ideal for ETL operations where we need to query all the columns.