**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?**

Pros: Consolidate passwords for different servers, dbs etc in one location for easier access

Enables you to easily add/link keys from vault to other

**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?**

Loop functionality can be achieved using the ForEach activity. After adding it to the pipeline, multiple activities that you want to perform on the data during each iteration can be added within the ForEach canvas. For example – adding an execute pipeline activity, copy etc.

The ForEach module works like a for loop In Python to iteratively go through and execute functions over each unit (can be a file, folder, sub-directory etc) within a larger directory. This way you can create 1 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 customize the parametrized value exactly to match your input data without hardcoding. Example – to read in a file instead of writing the value as

‘filename.csv’

you can use an expression like *concat* –

Concat(@filename(), ‘.’, @filetype())

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

Flexible – allows you to generalize the format 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, Binary, Delimited text, JSON, csv, excel, orc, parquet,

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.