Conner McNicholas
Azure Mini Project: End-to-end ETL

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

   Azure Key Vault is a cloud service for securely storing and accessing secrets such as API keys, passwords, certificates, or cryptographic keys. Key Vault greatly reduces the chances that sensitive information in leaked to unauthorized channels

   Alternatives: AWS Key Management service, LastPass, CyberArk, KeePass, and of course creating a novel secrets management solution tailored to your application.

   Pros: Automatically renew SSL certificates, Versioned keys and secrets, Integration with other Azure services

   Cons: Subscription-wide throughput limits, Account-level access policies

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

   The “ForEach” Activity defines a repeating control flow in an Azure Data Factory or Synapse pipeline. This activity is used to iterate over a collection and executes specified activities in a loop.

   For the same reason one would need loops as part of program code, data pipelines also require loops (e.g. pulling source files from a collection of websites).

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

   In mapping data flow, many transformation properties are entered as expressions. These expressions are composed of column values, parameters, functions, operators, and literals that evaluate to a Spark data type at run time.

   They are helpful in designing pipelines by enabling you to customize the behavior of your pipelines in almost every setting and property. For example, a pipeline has to execute the same command for multiple dates, and an expression can set the value of those dates by loading them from a specific file.

4. **What are the pros and cons of parametrizing a dataset in Azure Data Factory pipeline's activity?**
   Parameterizing a dataset minimizes the amount of hard coding and increases the number of reusable objects and processes in a solution.allowing for a highly flexible ETL solution, dramatically reduced costs from solution maintenance, and saving tremendous amount of time. Azure Data Factory currently does not provide for date parameters and so dates must be passed as strings (using format "yyyy-MM-dd").

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/codecs include: binary, excel, delimited text, json, parquet, xml, orc, avro

   Parquet is columnar, so excels at data warehousing solutions where massive columnar data needs to be analyzed, and is better optimized for use within Spark applications than ORC.

   Avro is better equipped for cases where support is required for an evolving schema or file structure modification, combining the advantages of both json and binary to describe data and keep storage size to a minimum.