**Homework**

**Differences between Parquet and CSV?**

**Storage Efficiency**

* CSV files store data in rows, which often leads to redundancy and larger file sizes. For example, if a dataset that consumes 9.5 GB in CSV format may only require 81 MB in Parquet due to its columnar storage and compression
* Parquet's columnar structure allows similar data types in columns to compress more effectively, reducing storage costs significantly

**. Ease of Use**

* CSV is simple, universally supported, and human-readable, making it suitable for small datasets or quick prototyping
* Parquet is machine-oriented and requires specialized libraries or tools (e.g., Apache Spark, Pandas) to read and write files. It is better suited for backend systems or data lakes

**Differences between Azure Data bricks and Azure Synapse?**

**Ans:**

**Azure Data Factory:** Azure Data Factory is a cloud-based data integration and orchestration service that allows organizations to create, schedule, and manage data pipelines for data movement and transformation. It also provides ETL Services to the organizations. ADF has GUI for creating and managing pipelines.

**Azure Databricks:** Its a Data analytics tool which is mainly used for big data analytics and Machine Learning. In this we have advanced data transformations using Apache Spark. Azure data bricks mainly work on Notebooks for data analysis and development. It’s a High scalability with Built-in Spark.

It mainly comprises of 4 Engines:

1. Ingestion
2. Transformation
3. Analysis
4. Power BI

**Azure Synapse:** Azure Databricks is a data analytics tool mainly used for big data analytics and machine learning. We have advanced data transformations using Apache Spark. Azure data bricks mainly work on Notebooks for data analysis and development. It has high scalability with Built-in Spark.