***Case Study :***

***Assignment 1 Definition***

***The XYZ bank has opted to transition away from utilizing Datawarehouse and is embarking on a journey towards employing centralized cloud-based Azure storage. As part of this shift, mechanisms have been established within On-Prem systems and other sources to transmit incremental daily snapshots of data to Azure Datalake Storage.***

* ***Traditionally, the Marketing System (MS) accessed the necessary datasets on a daily basis from the Datawarehouse for the purpose of generating dashboards. However, in light of the bank's strategic pivot, the MS team is now tasked with designing a system to daily ingest the datasets from the centralized storage into their designated storage account. They will also need to execute requisite operations on the datasets to ensure they are readily accessible for dashboarding purposes.***
* ***In our capacity as Senior Azure Engineers, it is imperative that we collaborate to devise a scalable solution for the MS team's requirements.***

Mapa com linhas coloridas

Descrição gerada automaticamente com confiança média***Architecture design:***

In this scenario, we can use Azure Data Factory to orchestrate a pipeline end-to-end (ETL) finalizing with data in curated zone ready for consumption.

Starting with extract data from sources of any kind, like Db’s, files, and other sources to move data into storage account, in the raw zone. For better performance, we use Vm’s with Integration Runtime client (Self-Hosted) for making the connection between on prem and azure cloud.

We can do this incrementally (delta loads), taking new records to the destination. Using date fields, for example, with an Azure SQL to store parameter tables showing when the last successful load was made, and from this date extract new records.

After this, we have a raw data coming in a parquet format in our storage account.

Next step with databricks is cleaning the data organizing them by subjects and entity, partitioning, merging tables and moving them into the enriched zone. In this step the data is ready for analytics consumption.

The next phase will be modeling the data, as the marketing system needs. In this consumption zone, the data will be delivered as a product, and consumption in this architecture will be through virtualization of views from the synapse serverless db.

This pipeline is scalable to any additional data source.

After this we can see this dataset on a DataViz tool such as Power Bi.