### 1)What projects you have done?

Project 1: Build Data pipeline using Azure for WF financial risk application for batch data.

### 2) If you can share what was the company that you did this work for? (You can also say something like Fortune 500 financial services or Healthcare etc if you are not comfortable sharing the name)

- I had previously worked with Wells Fargo and Chase. Total experience : 8+

### 3) What was the size of the team and your role?

In my previous position, I worked as a senior data engineer in a team of 7.

As part of the ongoing cloud migration initiative, we were required to build a data pipeline using Azure.

Specifically, my responsibilities included:

* Early adoption of Microsoft Azure and participation in implementation design of azure data pipeline for the project
* Providing knowledge sharing sessions on Azure and AWS.
* Collaborating with other teams such as data modellers, data analysts, and software engineers to understand their data needs and provide them with the proposed data infrastructure to support their work.
* Designing and implementing a scalable data pipeline to move and transform data from multiple storage containers into the organization's data storage and processing systems using Azure Databricks and ADLS.
* Managing and optimizing the performance of the aforementioned data storage (bronze, silver and gold data layers) to ensure they can handle the organization's current and future data needs.

### 4) What was the business problem?

**Problem statement:** In lieu of the organisation’s initiative to move away from licensed ETL tools, few teams were required to work on early adoption of the cloud. Objective was to build a scalable data pipeline to collect, process, and store data from multiple sources using either AWS or Azure.

### 5) What technology framework did you use - Spark/Flink/others?

Azure databricks, Azure data lake storage (ADLS)

### 6) how were the pipelines orchestrated?

Using Apache Airflow

### 7) how did you do unit and performance testing of these pipelines?

* The transformations were tested by running them on a small sample of data and by verifying the output against expected results.
* Business users were also involved in the testing and the results are compared against the loads performed with our regular ETL loads using Abinitio.
* Used logging and monitoring to track pipeline’s progress and identify issues

### 8) Is this in production?

- We worked with production size data to favour a more realistic testing approach. However, the data pipeline has not been used (yet) for data loading in production.

### 9) what language did you use - Java/python/scala/others?

Spark (and python to some extent)

### 10) what was the Spark/flink/other distributed runtime environment ?

Worked with Apache spark and RDDs

### 11) what was the size of some of the key jobs - number of datasets processsed, size of data ?

We processed risk related data of about 4GB.

12) how big was the data lake / data warehouse / lakehouse?  
- what kind of processing did you work on (ingestion, data processing, validation, data quality, transformation/aggregation, ML algorithms, graph algorithms etc)

* Data Extraction: Data was extracted from the various storage containers and loaded into DataBricks. We processed risk related data of about 4GB spread across 3 files.
* Data Transformation: Data was transformed using Spark and SQL to prepare it for analysis. This includes cleaning, filtering, joining, and aggregating the data.
* Data Loading: The transformed data was loaded into ADLS in Parquet format for efficient storage and query processing.
* Unit testing: The data was validated using automated scripts and tests to ensure accuracy and completeness.
* Data Visualization: The data was visualized using tools such as Power BI to enable business users to gain insights and make informed decisions.

### 13) how was the processed data consumed?

* Data was not consumed for actual reporting since it was a POC project. But the final data has been fetched with PowerBI reports and was validated against the production data.

### 14) any learnings/experiences.

- The recent projects I worked on pushed me into this wonderful world of data engineering where I was not only concerned with the ETL part (which was pretty much how I viewed my role a few years ago) but was also involved at every step of the data processing, starting from the design to data modelling to finally making the data business-consumable. I realised that being able to architect data pipelines is far more interesting that limiting myself to just one language or one tool or technology.  I love the aspect of integrating things together and taking information from so many different sources and helping business build a story out of them.

I have also learnt to work with multiple data formats. The work I did on this project also reinforced the importance of running test cases with varied datasets and the importance of communicating with business users throughout the development and giving demos to the users of the pipeline at various stages.

### 15) If possible a quick visual which includes the above would be awesome. You can use any diagram tool - online would be great like [draw.io](http://draw.io/) or lucidchart etc.

A quick overview of my work can be found at the below link

<https://github.com/m-anusha-DE/aboutme/blob/main/AzureDataPipleine.pdf>