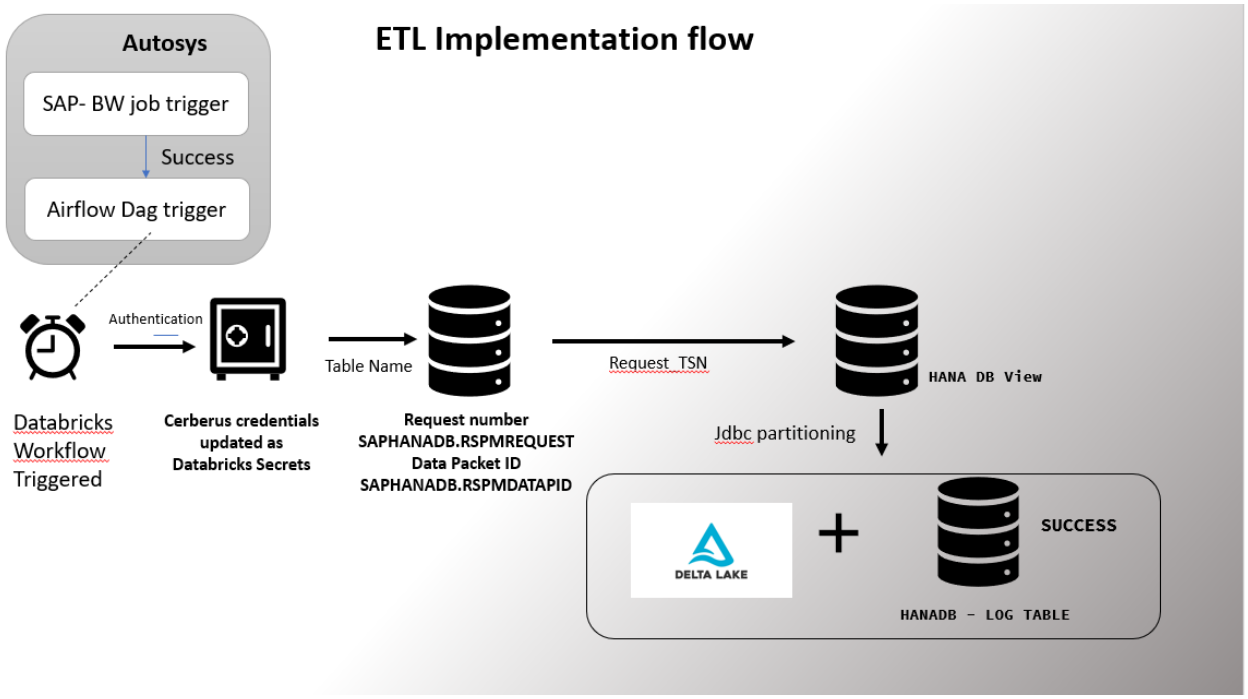

Databricks - PySpark: Databricks provides an optimized platform for running Apache Spark workloads; the ETL logic was implemented for finance and non-finance consumers to transfer huge volumes of data from HANA DB to S3 Locations.

Here's a summary of my experience in Data Engineering:

1. Connection to Cerberus and management to credentials to HANA DB as Databricks secrets with appropriate user management.
2. Designed and implemented Pyspark ETL logic: Connection to HanaDB and data retrieval through SQL join queries from different tables and writing to S3 bucket in delta format using parameters passed from Airflow dags and workflows.
3. Implemented data models that facilitate efficient data storage, data validation, retrieval, and analysis.
4. Optimized Databricks cluster and JDBC partitioning to write over 20 million records in under 30 mins.
5. Implemented alert mechanisms to detect failure (Email, Slack, ServiceNow Tickets) and duration threshold to handle long duration jobs.
6. Created dashboard on Unity Catalog (bronze layer) on the hourly data that gets egressed for tables created in QA and production environment. This dashboard helps the downstream team to validate the received data.
7. Created more than 50 workflows and managed the optimization on cluster management by choosing the right run version, worker type, and spark configuration.
8. Explored and implemented delta live pipelines to simplify the development, management, and reliability of data processing workflows, to efficiently process and analyze data in real-time.

9. Shared the delta table programmatically with the right access to the consumers from different regions and different projects.
10. Trained and mentored the entire team about the ETL development using Databricks as I was a pioneer in implementing this data egress concept in Nike for SAP HANA domain.
11. Developed automation scripts, GitHub code management, recon solution, switching environments within GitHub and Databrick workflows, scheduling jobs, and to ensure data pipelines run smoothly and efficiently.
12. Collaborated with various teams, including data scientists, analysts, and business stakeholders, to understand their data requirements, updated the log retention over 300 days, enabled change data feed and share the required recon stats to ensure that data was not missed before or after the egress.
13. Supported round the clock during production go-live and ensured transparency in communicating delays or issues.

Autosys:

1. Created jobs by working with cross-functional team and convinced the management to implement Databricks plugin to trigger Databricks workflow directly from Autosys.
2. Orchestration implemented by Creating jobs in Autosys to trigger the dags in Airflow which in turn trigger the Databricks workflow.

Airflow:

1. Created dags to trigger the Databricks workflow. The scheduling was implemented in Autosys which is dependent on the BW job completion. The run time was sent as a parameter to Databricks workflow from Airflow.
2. Worked with the airflow team to implement effective cluster configurations for better performance of the dags and addressed the issues with root cause analysis.