Kalyan Sai
Data Engineer
(813) 388 8041
Kalyanchowdary216@gmail.com

Professional Summary

- Experienced data engineer with **5+** years of hands-on experience building and managing data pipelines and infrastructure.
- Expertise in creating and maintaining data pipelines using **GCP**, **Azure**, and **AWS** cloud technologies.
- Experience in designing and implementing data processing systems using GCP services such as **BigQuery**, Dataflow, **Dataproc**, and Pub/Sub.
- Develop data set processes for **data modelling**, and Data mining. Recommend ways to improve data reliability, efficiency, and **quality**.
- Developed **PySpark** Applications using Python and Implemented the Apache **PySpark** data processing project to handle data from various **RDBMS** and Straming sources.
- Skilled in programming languages like **Java, Scala, Python and SQL.**
- Knowledge of **ETL** methods for data extraction, transformation and loading in corporate-wide ETL Solutions and Data Warehouse tools for reporting and data analysis.
- Experience working with **AWS** Data Platform - AWS Cloud Formation, Development Endpoints, AWS Glue, EMR, Athena, Lambda and Jupiter/Sagemaker Notebooks, **Redshift**, **S3**, and **EC2** instances.
- Experience in Microsoft Azure/Cloud Services like SQL Data Warehouse, Airflow, Azure SQL Server, **Azure Databricks**, Azure Data Lake, Azure Blob Storage, and Azure Data Factory.
- Hands on working experience with Databricks and snowflake database.
- Implemented data transformation logic and data cleansing processes using Java based frameworks.
- Experienced in Branching, Tagging, and maintaining the version across the Environments using SCM tools like **Git**, **GitLab**, **GitHub**, and SVN on Linux and Windows platforms.
- Evaluated model output using the uncertainty matrix (Precision, Recall).
- Created IAM Roles and defining Policies and applying to AWS services.
- Worked on complex **SQL Queries**, and **PL/SQL** procedures and converted them to ETL tasks.
- Performed Exporting and importing of data into simple storage service (S3).
- Good knowledge on Importing volumes, launching EC2, RDS, creating security groups, auto-scaling, load balancers (ELBs) in the defined virtual private connection.
- Knowledgeable in building data pipelines and ETL workflows using tools such as Apache Beam, Apache Airflow, and Cloud Composer.
- Knowledgeable in designing and optimizing data models for performance and scalability using Synapse Analytics and SQL Database.
- Skilled in designing data architectures and optimizing data models for performance and scalability using tools such as **Apache** HBase and Apache Cassandra.
- Hands on experience in Test-driven development, Software Development Life Cycle (SDLC) methodologies like **Agile** and **Scrum**.
- Research on membership inference attacks to protect the sensitive information used in ML model training.
- Good analytical, communication skills and ability to work with a team as well as independently with minimal supervision.
- Evaluated model output using the uncertainty matrix (Precision, Recall).

TECHNICAL SKILLS

- **Cloud Platforms:** AWS, Azure, GCP
- **Programming Languages:** Java, Scala, Spark, Python, SQL, C, C++,
- **Databases:** Oracle, SQL Server, MySQL, Cassandra, PostgreSQL, DynamoDB, CosmosDB, BigTable, Spanner

Kalyan Sai
Data Engineer
(813) 388 8041
Kalyanchowdary216@gmail.com

- **Distributed Messaging Systems:** Apache Kafka, AWS Kinesis, GCP Pubsub/Pubsub Lite, Azure Event Hub
- **Distributed Processing Systems:** Apache Spark, Apache Flink, GCP Dataflow
- **Data Visualization Tools:** Tableau, Power BI
- **Data warehouse Systems:** AWS Redshift, Azure Synapse, GCP BigQuery
- **Machine Learning:** AWS Sagemaker, TensorFlow, PyTorch

Client: Deutsche Bank, Cary, NC **Aug 2021 - Present**
Sr. Data Engineer
Responsibilities:

- Collaborated with cross-functional teams to gather requirements, design data models, and implement solutions on Azure **Databricks** Lakehouse, meeting business objectives and adhering to best practices in data engineering and cloud architecture.
- Developed and implemented data pipelines using Azure Data Factory to orchestrate and automate the movement of data between Azure Databricks Lakehouse and Azure Snowflake, ensuring efficient data integration and processing across the ecosystem.
- Designed and optimized data warehouse architecture on **Azure Snowflake**, leveraging features such as clustering, materialized views, and automatic scaling to improve performance and reduce costs for analytical workloads.
- Configured Azure Snowflake to ingest streaming data from IoT devices, implementing real-time data processing and analysis pipelines for immediate insights and decision-making within the Azure Databricks Lakehouse.
- Collaborated with stakeholders to design and implement data governance policies and procedures within Azure Databricks Lakehouse, ensuring data quality, integrity, and compliance across the platform, including integration with Azure Data Factory for metadata management and lineage tracking.
- Implemented data migration strategies from on-premises data warehouses to Azure Snowflake using Azure Data Factory, ensuring seamless transition and minimal downtime for critical business operations while leveraging Snowflake's scalability and performance capabilities.
- Optimized data storage and retrieval processes using **Azure Data Lake Storage** within the Databricks Lakehouse, improving data accessibility, scalability, and cost-effectiveness for large-scale data sets.
- Developed data ingestion pipelines using Azure Databricks and Azure Data Factory to seamlessly integrate data from disparate sources into the Databricks Lakehouse, ensuring data consistency and reliability.
- Conducted performance tuning and optimization of **SQL** queries and Spark jobs within Azure Databricks Lakehouse, leveraging indexing, partitioning, and other techniques to enhance query execution speed and resource efficiency.
- Collaborated with data scientists and analysts to provision and manage dedicated SQL pools and serverless SQL pools within Azure Synapse Analytics, enabling ad-hoc querying and interactive data exploration within the Databricks Lakehouse.
- Provided technical leadership and guidance in the adoption of best practices for data engineering on Azure Databricks Lakehouse, including architectural design reviews, code reviews, and performance optimization sessions, driving continuous improvement and innovation in data engineering processes.
- Utilized Azure Databricks for scalable data processing and advanced analytics within the Lakehouse architecture, enabling real-time insights and predictive modeling for business decision-making.
- Employed **FastAPI** framework to develop efficient and scalable **RESTful APIs** for data retrieval and manipulation within the Azure Databricks Lakehouse environment, ensuring high performance and responsiveness.

- Maintained all developments in **GitHub** repositories using Push, Pull, Commit, and Merge operations, facilitating collaboration and version control within the Azure Databricks Lakehouse project.

Environment: Python, Data Analytics, Cloud Storage, Data Studio, Azure Databricks, Azure Data Lake Storage, Azure Data Factory, Apache Airflow, Cloud Run, Agile, JIRA.

Client: M&T Bank ,Buffalo, NY **Feb 2020 – July 2021**
Sr. Data Engineer

Responsibilities:

- Led the design and implementation of data migration strategies, leveraging Azure native services including Azure Data Lake Storage (**ADLS**), Azure Data Factory, and Azure Synapse Analytics, ensuring seamless transition and optimal performance.
- Led the design and implementation of data migration strategies, leveraging Azure native services including Azure Data Lake Storage (ADLS), Azure Data Factory, and Azure Synapse Analytics for seamless transition and optimal performance in Databricks Lakehouse environment.
- Developed and maintained complex **ETL** pipelines using Azure Data Factory to extract, transform, and load data from various sources into Azure Synapse Analytics, enabling real-time analytics and reporting capabilities within Databricks Lakehouse.
- Utilized Azure Fabric to orchestrate and manage microservices-based applications, ensuring scalability, reliability, and efficient resource utilization within the Azure ecosystem for Databricks Lakehouse.
- Spearheaded the design and implementation of scalable data processing solutions using Azure Synapse Analytics, optimizing query performance and enhancing data processing efficiency through distributed computing in Databricks Lakehouse architecture.
- Demonstrated expertise in **SQL**, crafting efficient schema designs and dimensional data models for diverse analytical requirements in Azure Synapse Analytics, facilitating easy data retrieval and analysis in Databricks Lakehouse.
- Executed Extract, Transform, Load (ETL) operations from various source systems to Azure Data Storage services, employing Azure Data Factory and **T-SQL** for subsequent data processing in Azure Databricks within Databricks Lakehouse setup.
- Customized Scala and Pyspark User Defined Functions (UDFs) to meet specific business requirements in Databricks Lakehouse environment.
- Crafted JSON Scripts to deploy pipelines in Azure Data Factory (ADF), facilitating data processing via SQL Activity for Databricks Lakehouse.
- Conducted data processing and structuring using Spark on Databricks, encompassing data aggregation, joining, and querying via Data Frames with PySpark for Databricks Lakehouse architecture.
- Configured Snowflake Spark Streaming to receive real-time data from Kafka to store the stream data to **HDFS** within Databricks Lakehouse.
- Implemented advanced data compression and partitioning strategies in Azure Synapse Analytics to optimize storage utilization and improve query performance, resulting in cost savings and faster query execution times in Databricks Lakehouse.
- Conducted performance tuning and optimization of Azure Synapse Analytics workloads in Databricks Lakehouse, leveraging distributed query processing and resource management techniques to achieve optimal query performance and resource utilization.
- Designed and implemented disaster recovery strategies for Azure Synapse Analytics in **Databricks** Lakehouse, including backup and restore procedures, ensuring high availability and data integrity in case of system failures or disasters.

- Employed a variety of Python libraries including NumPy, SciPy, pandas, scikit-learn, and Spark libraries **PySpark** and MLlib for diverse model development and analytics purposes within Databricks Lakehouse.
- Enforced agile practices and **CI/CD** methodologies for seamless development to deployment cycles in Databricks Lakehouse environment.

Environment: Python (Scikit-learn/SciPy), Big Data, ETL, SQL Server, Spark, PySpark, Kafka, Azure SQL, AzureML Studio, Jupiter Notebook, Databricks, Power BI, Agile, Git, CI/CD, Linux.

Client: Abbott Park, Illinois **Nov 2018 – Jan 2020**
Data Engineer
Responsibilities:

- Created and implemented ETL procedures to transfer data from **AWS S3** to Snowflake using Snow pipe.
- Utilized **AWS Glue** to establish a batch processing pipeline for patient information and carried out data purification, data manipulation, and mapping transformations utilizing PySpark scripting.
- Established a real-time **Data Analytics**, data pipeline using AWS Kinesis for patient information, extracting it from S3 and storing it to **Dynamo DB**.
- I have utilized Terraform for cloud resource management, which requires access credentials to communicate with the cloud provider's API.
- Constructed a cost-efficient and internally managed serverless email system utilizing AWS Lambda and Simple Email Services (SES).
- Developed automated **CI/CD** data pipelines using AWS, Python, Code Pipeline, and deployed code modifications into **AWS EC2instances** using AWS Code Deploy.
- Implemented a trigger for Lambda functions to build a pipeline for the transfer and transformation of small datasets, resulting in cost savings and time efficiency.
- Created Apache Airflow DAGs (directed acyclic graphs) to schedule the pipeline for transferring data from AWS S3 to **Snowflake**
- Created **dashboards** and reports for business intelligence purposes using **AWS Quick Sight** on data obtained from the DynamoDB.
- Implemented **AWS** Code Pipeline and Created Cloud formation JSON templates in Terraform for infrastructure as code.
- Utilized **GitHub** as a source code management tool for performing tasks such as branching, merging, and tagging, among,
- Automate provisioning and repetitive tasks using Terraform and **Python**, Docker container, Service Orchestration.
- Collaborated in an **agile** workspace and utilized **JIRA** to keep track of and create user stories.

Environment: AWS EC2, S3, Glue, Kinesis, Code Pipeline, Code Deploy, Splunk Cloud, QuickSight, Athena, ETL, GitHub, Apache Airflow, DAGs, Snowflake, Agile, JIRA.