Fayaz K

(469)-502-3454 | fayazkf98@gmail.com | Linkedin

PROFESSIONAL SUMMARY

Experienced Data Engineer with 6+ years of success designing, developing, and optimizing data platforms across AWS, Azure, and GCP. Skilled in building secure, scalable pipelines using tools like Apache Spark, Airflow, Glue, and Databricks while applying strong data modeling and governance principles. Proficient in Python, SQL, and Java, with a proven track record of enabling real-time insights, automating ETL workflows, and delivering data products aligned to business goals. Experienced partnering with cross-functional teams to improve operations, ensure compliance, and make data more accessible across healthcare, telecom and aviation sectors.

TECHNICAL SKILLS

Cloud Platforms: AWS (S3, EMR, Redshift, DynamoDB, SQS, Lambda, EC2, IAM, Key Management Service), Google Cloud Platform (Big Query, Dataflow, Pub/Sub, Google Cloud Storage), Microsoft Azure (Data Lake, Cosmos DB, Azure Databricks, Azure Data Factory, Azure Functions, Logic apps), IAAS **| Big Data Processing:** Apache Hadoop, Apache Spark, Apache Flink, Apache Hive, Apache HBase **|**
Data Streaming: Apache Kafka, Apache Storm, Apache Flink, Apache Beam **| ETL & Data Integration:** Apache NiFi, Talend, Informatica, AWS Glue, Azure Data Factory **| Databases:** MySQL, PostgreSQL, Oracle, SQL Server, Cassandra, MongoDB, HBase, NoSQL, Teradata, IBM DB2, PostgreSQL **| Data Warehousing:** Amazon Redshift, Google BigQuery, Snowflake, Databricks **| Containerization and Orchestration:** Docker, Kubernetes, Apache Airflow **| DevOps & CI/CD:** Jenkins, GitLab CI, DevOps, GitHub, AWS Cloud Formation, Terraform **| Data Visualization:** Tableau, Power BI, Looker **| Version Control & Monitoring tools:** Git, SVN **,** Splunk, Grafana **| Programming Languages:** Python, SQL, Java, Linux, Unix, R, Scala, GO

WORK EXPERIENCE

Senior Data Engineer | American Airlines | Dallas, TX May 2024 – Current

- Led the migration of enterprise data from an on-premise Oracle system to **AWS Redshift**, using Agile methodologies to iteratively improve the process and incorporate stakeholder feedback at each sprint.
- Developed **PySpark**-based **AWS Glue** jobs to extract data from **HDFS** and load it into **Amazon S3**, integrating CI/CD pipelines for automated testing and deployment across environments.
- Implemented data quality checks using PyDeequ and custom Python scripts within AWS Glue jobs to validate data accuracy and completeness, ensuring only trusted data flowed into Redshift and downstream reporting systems
- Loaded and transformed data into **Amazon Redshift**, while setting up **AWS CloudWatch** to monitor **AWS** RDS instances and ensure system health. Designed robust data ingestion frameworks supporting both batch and real-time streaming using **AWS Batch, Kinesis**, and **AWS** Data Pipeline.
- Built scalable, event-driven architectures using **AWS Lambda** in conjunction with API Gateway, **DynamoDB** and **S3** for low-latency data processing and service orchestration.
- Processed **JSON** data using **SparkSQL**, created Schema RDDs, and loaded the structured data into **Hive** tables for analytics and reporting use cases. Automated job orchestration using **Apache Airflow**, scheduling and monitoring Spark and Hive jobs based on data availability and SLAs.
- Created end-to-end **AWS** Data Pipelines to handle **ETL** processes across multiple data sources and visualized results using Matplotlib for business insights.
- Developed Snowflake data models and complex **SQL** queries to analyze investment data, enabling performance tracking and trend analysis. Utilized **Spark** Streaming and **Kafka** to build a near real-time learner data model, writing transformed results to **DynamoDB** for instant access and reporting.
- Employed Git and GitHub for version control, deployed workloads on **Amazon EC2**, and used Python libraries such as NumPy and SciPy for advanced numerical computations.
- Built interactive dashboards and business reports in **Tableau**, applying various visualizations (bar charts, funnel plots, heatmaps, bubble charts, etc.) to empower stakeholders with actionable insights
- Collaborated with cross-functional teams including finance analysts, and product managers to define data contracts, validate KPIs, and implement Redshift materialized views and performance-tuned schemas that reduced dashboard query latency by over 40%.

Data Engineer | T-Mobile | Bangalore, India Dec 2020 – Dec 2022

- Migrated legacy telecom batch CDR (Call Detail Record) workflows from Informatica and on-prem Hadoop systems into a scalable, event-driven architecture on Azure using Azure Data Factory, Azure Event Hubs, Kafka and Azure Stream Analytics, enabling faster data delivery and real-time operational insights.
- Rebuilt Informatica ETL logic into distributed PySpark pipelines in Azure Databricks, processing billions of network logs for advanced telecom use cases like subscriber churn modeling, call drop prediction, and fraud analytics, resulting in a 35% improvement in model execution times.
- Designed a Delta Lakehouse framework on Azure Data Lake Storage Gen2 with Bronze, Silver, and Gold layers, implementing governance and data access control through Unity Catalog, enhancing compliance and secure cross-domain collaboration.
- Re-engineered transformation logic in dbt, translating legacy Informatica mappings into modular SQL-based models within Azure Synapse Analytics, and implemented Slowly Changing Dimensions (SCD) Type 1 & Type 2 for customer and network dimension tracking.
- Built real-time Power BI dashboards visualizing key metrics like network usage, call latency, dropped calls, and churn trends, empowering business stakeholders and customer support teams with live decision-making capabilities.
- Automated the CI/CD pipeline using Azure DevOps, Terraform and Git, enabling seamless deployment of ADF pipelines, Databricks notebooks, and infrastructure as code across development, staging, and production environments.
- Collaborated with data science teams to deploy ML models via MLflow and Azure Machine Learning, enabling real-time scoring of user call behavior and churn risk, improving campaign targeting and retention strategy.
- Established comprehensive monitoring and alerting using Azure Monitor, Log Analytics, and Application Insights, ensuring robust operational visibility and faster issue remediation across streaming and batch data pipelines.

Data Engineer| Optum | Hyderabad| India Jun 2018 – Dec 2020

- Designed and built scalable data pipelines using **Azure Data Factory** and Stream Analytics to automate the collection, transformation, and storage of clinical trial audit data. Ensured secure and compliant data flow into **Azure Blob Storage** and **Synapse Analytics** to support regulatory standards and long-term traceability.
- Utilized SSIS to handle **ETL** tasks and loaded processed data into **Azure Synapse** for **analytics**, while SSRS was used to create detailed reports tracking key clinical trial metrics, improving transparency and compliance monitoring.
- Created automated workflows in Azure Data Factory to process large volumes of raw data efficiently and reliably, enabling seamless integration across various systems.
- Leveraged **Azure** Functions and **Blob Storage** to build a fully serverless architecture for lightweight processing and long-term storage, optimizing cost and operational performance.
- Used **PySpark** within **Azure Databricks** to process and validate terabytes of data across distributed systems. This ensured the accuracy and consistency of clinical data transformations on a scale.
- Integrated predictive analytics tools such as **TensorFlow**, **PyTorch**, and Scikit-learn to experiment with clustering, decision trees, and prototype models leading to improvements in resource scheduling and logistics efficiency.
- Set up **CI/CD** pipelines in **Azure DevOps** and **Jenkins** to automate building, testing, and deploying **PySpark**-based solutions, speeding up delivery cycles and reducing manual intervention.
- Developed interactive dashboards in **Tableau**, allowing stakeholders to track KPIs, spot emerging trends, and make faster, more informed decisions based on real-time visuals.
- Used **Docker** and **Azure Kubernetes Service** (AKS) to containerize and deploy scalable data solutions, ensuring smooth environment transitions and easy management of microservices.
- Proactively monitored pipeline health using **Azure Monitor** and **Grafana**. Set up alerting and built real-time dashboards to visualize ETL performance, error rates, and latency improving incident response and operational reliability.

EDUCATION

The University of North Texas Denton, TX
Masters – Data Science (GPA 3.6/4) Jan 2023 - Dec 2024

CERTIFICATIONS

- Microsoft Certified: Azure Data Engineer Associate
- Microsoft Certified: Power BI Data Analyst Associate
- Snowpro Core Certified: Snowflake