RIYAZ DUDEKULA

Data Engineer

+19132955802 || ✉ riyaz.d1604@gmail.com || Kansas, USA || LinkedIn/DudekulaRiyaz4

CAREER OBJECTIVE

- Data Engineer with **4+** years of expertise in architecting, designing, and implementing end-to-end **data pipelines, ETL processes, and analytics solutions** across **AWS, Azure, GCP, and on-premises Hadoop ecosystems**.
- Hands-on experience with **AWS (S3, Redshift, EMR, Lambda, DynamoDB, CloudWatch)**, **Azure (ADLS, ADF, Databricks, Synapse Analytics)**, and **GCP (BigQuery, GCS, Dataflow, DataProc, Cloud Functions, Stackdriver)** for scalable batch and real-time data processing.
- Strong proficiency in **Informatica ETL, Apache Airflow, Oozie, and workflow schedulers**, building reusable mappings, transformations, and orchestrating data pipelines with DAGs and control flows.
- Developed and optimized **Spark applications using Spark-SQL in Databricks** for large-scale data extraction, transformation, and aggregation across diverse file formats.
- Skilled in **Kafka-based data streaming**, integrating producers and consumers with AWS services for real-time analytics.
- Expertise in **Data Warehousing, Data Modeling (Star Schema, Snowflake), and SQL performance tuning** using DDL/DML, DMV, and DTA for high-performance analytics.
- Proficient in **CI/CD pipelines (Jenkins, Groovy, Docker, Kubernetes)** and scripting with **Python & PowerShell** for automation and workflow optimization.
- Experienced in **data visualization** using Tableau, Matplotlib, and ggplot2 to deliver actionable insights, and in **migrating on-premise workloads to Azure** with Site Recovery and Azure Backups.

RELEVANT SKILLS

- **Cloud Platforms: AWS (**S3, Redshift, EMR, Lambda, DynamoDB, CloudWatch), **Azure (**ADLS, Data Factory, Databricks, Synapse Analytics, Azure Site Recovery), **GCP (**BigQuery, Cloud Storage, Dataflow, DataProc, Cloud Functions, Stackdriver)
- **Big Data & Data Processing:** Hadoop (HDFS, MapReduce, Hive, Pig), Apache Spark (Spark-SQL, PySpark, Databricks), Apache Kafka (Producers, Consumers, Streaming), Oozie, Apache Airflow (Workflow orchestration, DAGs)
- **ETL & Data Integration:** Informatica PowerCenter (mappings, transformations, workflows), Azure Data Factory, AWS Glue, Data Migration (On-Prem to Cloud)
- **Databases & Data Warehousing:** SQL Server, Oracle, MySQL, PostgreSQL, Amazon Redshift, Azure Synapse, Google BigQuery, Data Modeling (Star Schema, Snowflake), Performance Tuning (DDL, DML, DMV, DTA)
- **Programming & Scripting:** Python (ETL, automation, data pipelines), SQL (Advanced queries, stored procedures, optimization), PowerShell (automation), Shell Scripting
- **DevOps & CI/CD:** Jenkins (Pipeline syntax, Groovy), Docker, Kubernetes (Containerization & Orchestration), Git, GitHub, Bitbucket
- **Visualization & Reporting:** Tableau, Power BI, Matplotlib, ggplot2
- **Other Tools & Platforms:** Cisco CloudCenter (multi-cloud deployments), REST APIs (integration with microservices)

WORK EXPERIENCE

Azure Data Engineer

C2FO, Kansas, USA May 2025 - Present

C2FO is a Leawood, Kansas-based financial technology company it operates a working capital finance platform. I am developing and optimizing ETL (Extract, Transform, Load) processes to ingest, process, and store data from various sources using Azure services such as Azure Data Factory and Azure Databricks.

Key Responsibilities:

- Designed and implemented **end-to-end ETL pipelines** using **Azure Data Factory, Databricks, and Synapse Analytics** for seamless ingestion, transformation, and integration.

- Developed and optimized **Spark applications** in **Databricks** using Python, Scala, and Spark-SQL to process and analyze large-scale datasets.
- Implemented **performance tuning techniques** in Azure Data Factory, Synapse Analytics, and SQL queries to improve pipeline efficiency and reduce latency.
- Built and managed **real-time and batch pipelines** leveraging **Azure Data Lake, Stream Analytics, and Data Lake Analytics** for enterprise BI and analytics.
- Automated **job scheduling and orchestration** on **Azure Virtual Machines** using **Control-M** and Airflow, ensuring reliability and optimized resource usage.
- Architected and deployed **Infrastructure as Code (IaC)** using **Terraform and Docker**, integrated with **Jenkins CI/CD** pipelines for automated provisioning and deployments.
- Executed **data migration projects** from SQL Server to **Snowflake** using Python and SnowSQL, staging **API and Kafka streaming data** (JSON) into Snowflake.
- Optimized batch and streaming **data synchronization** by leveraging **GCP Dataflow and Pub/Sub**, enabling near real-time migration from legacy systems.
- Standardized **Elasticsearch usage patterns** and migrated **SAS-based metrics** into Snowflake, improving reporting performance and reducing query-related issues.
- Collaborated across **SDLC phases** (requirements, design, development, deployment) to deliver scalable, secure, and cost-optimized data engineering solutions.

AWS Data Engineer

AMC Theatres, Leawood, Kansas, USA May 2024 – Apr 2025

AMC Theatres is the world's largest movie exhibition company. Utilized AWS storage and database services, including Amazon S3, Redshift, and RDS, to store structured and unstructured data, ensuring efficient data retrieval and storage.

Key Responsibilities:

- Built and optimized **ETL pipelines** using **Hive, Pig, Spark, and AWS S3**, streamlining ingestion, transformation, and pre-aggregation of large datasets.
- Developed **AWS Lambda functions** for real-time data validation, transformation, and loading into **S3**, improving automation and reducing manual effort.
- Achieved **70% faster EMR cluster launches** and **60% faster Hadoop job execution** by optimizing configuration and leveraging **Boto3** for seamless S3 integration.
- Automated **CI/CD pipelines** with **Jenkins, Docker, and Kubernetes**, and provisioned **AWS EC2 and Terraform templates** for scalable cloud deployments.
- Processed and analysed **real-time streaming data** with **Kafka, PySpark, and Storm**, enabling business-critical insights.
- Performed **SQL and Snowflake performance tuning** (indexing, profiling, clustering) to improve query response times and reduce infrastructure costs.

GCP Data Engineer

Accenture, Hyderabad, India Jul 2022 - Jul 2023

State Street Corporation is a global financial services and bank holding company. I implemented data security measures and ensuring compliance with regulatory requirements using GCP security tools.

Key Responsibilities:

- Designed and created **Hive, HBase, and HBase-integrated Hive tables** using ORC format and Snappy compression for optimized storage and query performance.
- Developed **end-to-end ETL and streaming pipelines** using **Dataproc, Dataflow, PySpark, Spark-Scala**, and **Python DAGs in Airflow** for batch and real-time analytics.
- Built **reusable BigQuery views and data marts** to power **Data Studio dashboards** with consistent metrics and definitions.
- Implemented **dynamic cluster provisioning and autoscaling in Dataproc**, reducing idle time and optimizing GCP costs.

- Monitored and troubleshot **GCP services** using Cloud SDK and Cloud Shell, ensuring high system reliability and faster incident resolution.
- Integrated **Cassandra databases** via **DataStax Spark Connector** and migrated data between **GCP and Azure** using **Azure Data Factory** for multi-cloud analytics.

Data Engineer

Merck Group, Hyderabad, India Jan 2021 - Aug 2022

Merck is a global healthcare company that develops and sells innovative health solutions for people and animals. I designed the data models and integrating data from different systems to create unified data platforms that support analytics and reporting needs.

Key Responsibilities:

- Developed **PySpark applications** for ETL pipelines, transforming and loading large-scale data into **HDFS, Snowflake, and cloud storage**.
- Built **Databricks workflows and ADF pipelines**, orchestrating data extraction from SQL Server and processing for cloud and SFTP targets.
- Implemented **Apache Airflow DAGs** for workflow automation, scheduling, and monitoring of batch and streaming pipelines.
- Migrated on-premise SQL Server data to **Azure Synapse Analytics and Azure SQL DB**, performing **query optimization and performance tuning**.
- Developed and maintained **CI/CD pipelines** using Jenkins, Git, Maven, Docker, and Kubernetes for automated deployments.
- Monitored and maintained **Spark clusters, Elasticsearch, and Azure logs** using Log Analytics, Ambari, Metricbeat, and Kibana to ensure system reliability and performance.

EDUCATION

University of Central Missouri Missouri, USA

- Masters / Computer Science (Aug 2023 - May 2025)

CERTIFICATIONS

Udemy – Azure Fundamentals.

Coursera – Building Scalable Applications with AWS