

**Name: Rohan Reddy B**

**Ph.no: +1 8138935627**

**Gmail.ID: brohanreddy24@gmail.com**

**Linkedin: <https://www.linkedin.com/in/rohanreddy-b-aa0b24190/>**

## **PROFESSIONAL SUMMARY**

- Around 4+ years of work experience in IT consisting of Data Analytics Engineering & as a Programmer Analyst.
- Experienced with cloud platforms like Amazon Web Services, Azure, Databricks (both on Azure as well as AWS integration of Databricks).
- Strong background in developing T-SQL queries for data transformation, cleansing, aggregation, good understanding of relational and dimensional data modelling.
- Proficient with complex workflow orchestration tools namely Oozie, Airflow, Data pipelines and Azure Data Factory, CloudFormation & Terraforms.
- Implemented Data warehouse solution consisting of ETLs, On - premise to Cloud Migration and good expertise building and deploying batch and streaming data pipelines on cloud environment.
- Worked on Airflow 1.8(Python2) and Airflow 1.9(Python3) for orchestration and familiar with building custom Airflow operators and orchestration of workflows with dependencies involving multi-clouds.
- Spark for ETL follower, Databricks Enthusiast, Cloud Adoption & Data Engineering enthusiast in Open-source community.
- Developed and Deployed various Lambda functions in AWS with in-built AWS Lambda Libraries and also deployed Lambda Functions in Scala with custom Libraries.
- Experience in design, testing, implementation, maintenance, and control of the organization's physical, relational, and object-oriented databases across multiple platforms and computing environments.
- Proven track record of optimizing Spark application performance for optimal batch interval time, parallelism level, and memory optimization.
- Experienced with SSIS performance tuning on Control flow, Data flow, Error handling, and Event handler, re-running of failed SSIS packages.
- Adept at utilizing BI tools such as Power BI and QlikView for enhancing reporting capabilities and developing BI applications in accordance with client requirements.
- Demonstrated capability to liaise with key stakeholders for delivering compelling business value to senior leadership and clients.
- Well versed in Data warehouse concepts normalization/de-normalization techniques for optimum performance in relational and dimensional database environments and building Referential Integrity Constraints.
- Strong experience in Dimensional Modelling using Star and Snowflake schemas to build Fact and Dimensional tables. Implemented SCD's (Slowly Changing Dimensions) and CDC (CDC Change Data Capture).
- Experience with all stages of the SDLC and Agile Development model right from the requirement gathering to Deployment and production support.
- Hands on experience working on NoSQL databases including HBase and its integration with Hadoop cluster.
- Experience with Py Spark and Azure Data Factory in creating, developing, and deploying high-performance ETL pipelines. Used Kafka to load real-time data from multiple data sources into HDFS.
- Experience in building Real-time Data Pipelines with Kafka Connect and Spark Streaming. Skilled in query optimization, data workflows using Alteryx and data visualization presentations using Power BI, Quick Sight.
- Excellent understanding of best practices of Enterprise Data Warehouse and involved in Full life cycle development of Data Warehousing.
- Developed and executed data pipelines using Dataflow, a fully managed service for real-time and batch data processing, enabling efficient data transformations and streaming analytics.
- Proficiency with Hive and hands on experience in writing HiveQL queries, extensively used Merge, Partitioning, and bucketing to improve performance.
- Good understanding of data modelling (Dimensional & Relational) concepts like Star-Schema Modelling, Snowflake Schema modelling, Fact and Dimension tables.
- Hands on experience in SQL and NOSQL database such as Snowflake, HBase, Cassandra and MongoDB.
- Hands on experience in setting up workflow using Apache Airflow and Oozie workflow engine for managing

and scheduling Hadoop jobs.

- Experience in Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses, as well as data processing like collecting, aggregating and moving data from various sources using Apache Flume, Kafka, PowerBI and Microsoft SSIS.
- Strong experience in working with UNIX/LINUX environments, writing shell scripts.
- Strong skills in analytical, presentation, communication, problem solving with the ability to work independently as well as in a team and had the ability to follow the best practices and principles defined for the team.

#### **TECHNICAL SKILLS :**

- **RDBMS/Databases:** MSSQL Server, Oracle 9i/10g, MS Access, MYSQL, Azure Data Lake
- **ETL Tools:** SQL Server Integration Services (SSIS), Azure Data Bricks, Azure Data Factory, AWS Glue, Terraform, AWS CloudFormation, dbt (data build tool)
- **Real-Time Processing:** Apache Flink, Spark Streaming, Kafka, Kinesis
- **Cloud Technologies:** AWS, Azure, Databricks, AWS CloudFormation, Azure Synapse Analytics
- **Containerization/Orchestration:** Docker, Kubernetes, Azure Kubernetes Services (AKS)
- **CI/CD Tools:** Jenkins, GitLab, CircleCI, GitHub Actions
- **Monitoring & Alerting:** Prometheus, Grafana, AWS CloudWatch

#### **PROFESSIONAL EXPERIENCE:**

**Client: Truist Bank, Doral, FL**

**August 2023 – Present**

**Role: Data Engineer.**

- As a Data Engineer, Maintained data pipelines, data integration, ETL processes and data management delivery within a global data lake and data warehouse environment.
- Installed Kafka on Hadoop cluster and configured producer and consumer to establish connection from source to HDFS with popular hash tags.
- Build Complex distributed systems involving huge amount data handling, collecting metrics building data pipeline, and Analytics. Experience in data transformations using Map-Reduce, HIVE for different file formats. Implemented a proof of concept deploying this product in AWS S3 bucket and Snowflake.
- Created Data models and data dictionaries for projects. Verified data quality between systems, the accuracy of analytical reports and existing data structures to identify logical gaps, data quality issues and business process improvements by utilizing expertise with business rules, system behaviour and the data model
- Worked on complex SQL Queries, PL/SQL procedures and convert them to ETL tasks.
- Worked with Reporting developers to oversee the implementation of reports/dashboard designs in Tableau.
- Experience in data cleansing and data mining. Wrote, compiled, and executed programs as necessary using Apache Spark in Scala to perform ETL jobs with ingested data.
- Created configuration files for building ETL pipelines. Migrated ETL code from Talend to Informatica. Involved in development, testing and post production for the entire migration project.
- Worked closely with Business Analyst and report developers in writing the source to target specifications for Datawarehouse tables based on the business requirement needs.
- Maintained data pipelines, data integration, ETL processes, and data management within a global data lake and data warehouse environment using **AWS Glue, S3, and Redshift**. Orchestrated ETL pipelines using Azure Synapse Analytics, managing big data workflows and advanced queries across distributed environments.
- Implemented Prometheus and Grafana dashboards for proactive monitoring and alerting on pipeline performance in Azure environments.
- Integrated Azure Kubernetes Services (AKS) to automate the deployment and scaling of data processing workloads, leveraging Docker containers for microservice isolation.
- Employed CircleCI for automating continuous integration and deployment of ETL workflows in Azure Data Factory pipelines.
- Installed Kafka on the Hadoop cluster and configured producer and consumer to establish connections from the source to **HDFS**, managing high-throughput **real-time data streaming**.

- Optimized **NoSQL** solutions by integrating **DynamoDB** and **HBase** with AWS data architecture for unstructured data storage and querying.
- Built complex distributed systems handling large volumes of data, focusing on **data quality governance** using **AWS Glue Data Catalog** for metadata management.
- Managed and monitored big data pipelines using **Apache Airflow** for scheduling and orchestration.
- Improved data storage with **schema evolution** techniques using **Parquet** and **Avro** to handle changes in data structure over time.
- Designed and developed **data models** for **Star Schema** and **Snowflake Schema** in a highly scalable AWS Redshift environment for better performance.
- Utilized **IAM roles** and **data encryption** protocols (both in transit and at rest) to enhance data security in the cloud.
- Leveraged **AWS Lambda** for serverless data processing and integrated with **Kinesis Analytics** for real-time data insights.
- Enhanced CI/CD pipelines using **GitLab** and **Jenkins** for continuous deployment and automated testing of ETL processes.
- Integrated **Great Expectations** as a data validation tool to ensure data accuracy, consistency, and quality before loading into the AWS data lake.
- Collaborated with business teams to develop **advanced Tableau dashboards** and reports that pulled data from the data lake for insights.
- Designed and Developed Spark workflows using Scala for data pull from AWS S3 bucket and Snowflake applying transformations on it. Worked on various automation tools like GIT, Terraform, Ansible.
- Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.
- Developed reusable framework to be leveraged for future migrations that automates ETL from RDBMS systems to the Data Lake utilizing Spark Data Sources and Hive data objects.
- Implemented AWS Step Functions to automate and orchestrate the Amazon Sage Maker related tasks such as publishing data to S3, training ML model and deploying it for prediction.
- Used Spark SQL for Scala & amp, Python interface that automatically converts RDD case classes to schema RDD. Deploy the code to EMR via CI/CD using Jenkins.
- Instantiated, created, and maintained CI/CD pipelines and apply automation to environments and applications.
- Prepared scripts to automate the ingestion process using Python and Scala as needed through various sources such as API, AWS S3, Teradata and snowflake.

**Client: Accenture, Hyderabad, India**

**May 2019 – July 2022**

**Role: Data Engineer**

**Responsibilities:**

- Worked with the analysis teams and management teams and supported them based on their requirements.
- Involved in the entire project life cycle, from design discussions to production deployment.
- Good Knowledge of monitoring, managing, and reviewing Hadoop clusters using Cloudera Manage. Investigate Data Quality issues and generate presentable narratives based on biases possible due to incompleteness of data. Was responsible for ETL and data validation using SQL Server Integration Services.
- Good Understanding of Data ingestion, Airflow Operators for Data Orchestration and other related python libraries. Used Kibana visualizations to highlight compliance metrics in the translated report
- Worked on Python APIs calls and landed data to S3 from external sources. Analyse Machine Data and create excel visualization plots as a story narrative for Business Users and Product Owners.
- Used Tableau and Excel for Visualization charts etc. and regularly communicating finding with Product Owners. Expert in Developing SSIS packages to ETL data into heterogeneous data warehouses.
- Worked with Data Engineers and Data Scientists to help and understand the gaps between the Product Integrity Datasets and Digital Manufacturing by leveraging the Advanced Analytics.
- Used python subprocess module to perform UNIX shell commands and extracted data from Agent.
- Optimize and tune SQL queries and data processes to ensure high performance and reliability of the data infrastructure. Ingested data from on-premise systems (MySQL, Cassandra) and external sources using Azure Data Factory, Azure Data Lake, and Azure Synapse Analytics.
- Implemented data governance strategies by integrating Azure Purview to automate the discovery and

management of data assets.

- Developed and optimized data models using Star Schema in Azure SQL Data Warehouse for better reporting and querying performance.
- Managed and monitored Azure Kubernetes Services (AKS) for efficient resource scheduling and scaling of big data clusters.
- Designed and implemented ETL pipelines using Python and PySpark to ingest structured and unstructured data from MongoDB and Blob Storage into Azure Data Lake.
- Orchestrated ETL workflows using Apache Airflow for data transformations and loading into Azure Synapse for analysis.
- Used dbt (data build tool) for transforming data in the warehouse to ensure data integrity and consistency across datasets.
- Implemented data validation and data quality checks using Great Expectations, automating data tests within the pipeline.
- Enhanced security by implementing Azure Key Vault for secure storage of credentials and integration of Azure IAM roles for access control.
- Collaborated with data scientists and product owners to automate machine learning model training and deployment using Azure ML Studio.
- Developed advanced visualizations in Power BI for performance reporting, pulling data from Azure Analysis Services to ensure low-latency queries.
- Worked on Tableau Visualization Charts and daily status Dashboards. Worked in an Agile Environment and indulged in Design Review and End-to-End UATs and assisted QA in automating test cases.
- Unit testing, UAT testing and End-to-End automation design reviews. Demonstrated good communication skills and story narratives while Sprint Demos to leadership and Stake holders.

## **EDUCATION**

- Master of Science in Business Analytics and Information Systems from University of South Florida