Akhil Reddy
Data Engineer

mail2akhilreddy.a@gmail.com **6503837427**

Over 5+ years of experience in Big Data technologies like Spark, Hadoop ecosystem and Data Warehousing. Expertise in designing and implementing data pipelines across diverse sectors, both on-premises and in the cloud (Azure, AWS). Proven ability to use Azure Data Factory (ADF), Databricks, SQL, and Synapse server. Strong skills in Data Analysis, Data Cleansing, Data Scraping, and Data Visualization over large and multi-dimensional datasets

PROFESSIONAL SUMMARY

- Worked with large scale data processing over the data collected from a variety of **structured** and **unstructured** data.
- A good understanding of **dimensional data modeling** like **Star Schema** and **Snowflake Schema** for data warehouse projects, ensuring efficient data storage and retrieval.
- Possessing in-depth understanding and experience in designing, managing, optimizing **OLTP** and **OLAP** databases.
- Successfully optimized data warehouse performance and storage utilization through data **modeling** techniques.
- A good understanding of Spark architecture with **Databricks**. Can set up Microsoft Azure and AWS with Databricks, Workspace, and manage Clusters.
- Extensive experience in working with **Azure BLOB** and **Data Lake storage,** proficiently loading data into **Azure SQL Server** and **Synapse** analytics Data Warehouse.
- Developed and implemented **Azure Key Vaults** and **Secret Scopes** to secure sensitive data, such as passwords, API keys, and certificates.
- Leveraged Azure Data Factory (ADF) to orchestrate and manage **SSIS** packages, ensuring seamless integration and execution within Azure data workflows.
- Experience working with **AWS services** like EC2, EMR, S3, RDS, KMS, Kinesis, Lambda, API gateways, IAM etc.
- Expertise in designing and implementing Data Pipelines using **AWS Glue.**
- Full-fledged knowledge in developing rich interactive reports and Dashboards using visualizations tools like **Power BI**, **Tableau, Python libraries** (Altair, Matplotlib, Ggplot, Seaborn).
- Experience in developing Spark Applications using **Spark RDD**, **Spark SQL** and Data frame APIs.
- Proficient in writing complex SQL queries, Stored Procedures, Database design, Normalization, creating Indexes, User Defined Functions (UDFs) and Triggers.
- Knowledge in Scala programming language, leveraging its object-oriented and functional programming features for building scalable and high-performance data processing applications.
- Expertise in **Normalization** and **Denormalization** techniques in relational and dimensional database environments.
- Developed and tested **Type 2** and **Type 3** dimensions for a data warehouse project to track historical changes.
- Extensively used Apache Spark, spark SQL to perform processing tasks on massive datasets.
- Good knowledge of **Snowflake schema design and architecture**. Leveraged **Snowpipe** to establish a connection between AWS/Azure cloud storage and Snowflake data warehouse, enabling automated data ingestion.
- Strong hands-on experience on **Deep learning** models using Neural Networks.
- Possessing extensive experience in Python **EDA** (Exploratory Data Analysis), data cleaning, data extraction and scraping, text mining, regular expression operations, and working with large and multi-dimensional datasets.

- Developed Spark applications using **PySpark**, **Spark SQL** in **Databricks** for **data extraction**, **transformation, and aggregation** from various file formats (Parquet, delta, Json, CSV, Avro), uncovering valuable customer usage patterns.
- Used **Materialized views** to improve data performance by caching frequently used queries.
- Experienced in using **Airflow** to automate data pipelines and workflows.
- Expertise in DevOps user stories and story points, including gathering requirements, defining user stories, estimating story points, tracking progress, implementing, testing, deploying, and monitoring.
- Involved in release management process to streamline the **Azure DevOps** code repository management and CI/CD process for smooth deployment of code and spinning up the infrastructure.
- Experienced in database **performance tuning**, including index optimization, query rewrite, and schema design.

CATEGORY	TECHNOLOGIES AND TOOLS
Big Data Technologies	Hadoop, Map Reduce, HDFS, Apache Spark, PySpark, Scala, Hive.
Databases/Data Warehouse	MS SQL Server, My SQL, PostgreSQL, Data Modelling, ER Studio, Star Schema, Snowflake schema, Facts and Dimensions tables, Snowflake, OLTP and OLAP.
Cloud Environment	• **Azure** – ADLS/Blob, Azure Data Factory (ADF) (ADF), Azure Databricks (ADB), Synapse Analytics, Azure Synapse, Key-vault, Cosmos DB. • **AWS** – S3, RDS, Amazon Glue, EC2, Kinesis, Kafka, EMR, Step Functions, Lambda Functions, IAM.
ETL	ADF, Glue, SSIS package
Reporting Tools	PowerBI, Tableau, Python Data Visualization tools (Altair, Matplotlib, Ggplot, Seaborn).
Programming languages	SQL, NoSQL, R, Python (Pandas, NumPy, Scipy, Scikit-learn, TensorFlow, Keras), SAP ABAP.
Machine Learning	• Hypothesis Testing, Resampling methods, Time Series, Computer Vision. • **Supervised learning** - Linear and Nonlinear Regression, Model selection and Regularization, Classification, Logistic regression, Discriminant analysis, Decision Trees, Ensemble methods - Bagging, Boosting, Random Forest; SVM, Deep Learning (CNN, RNN, GANs) • **Unsupervised learning** - Dimensionality Reduction, Kalman Filtering and Clustering.
Certifications	Python Certificate by Google. SAP Certified Development Specialist-ABAP for SAP HANA 2.0
ERP	SAP (ISU, PM)

PROFESSIONAL EXPERIENCE

Client :UBS, Chicago, IL **Aug 2022 – Present**

Role :Sr. Azure Data Engineer

Key Responsibilities:

- Structured and **normalized** the data, defined Schemas, created database tables, and **Materialized views**.
- Experienced in applying **dimensional modeling methodologies** to design scalable and flexible data structures for complex data systems.
- Extensively utilized various **Azure Cloud Services**, including Azure SQL server, Data Factory, Azure Analysis Services, Azure Monitoring, Key Vault, and Azure Data Lake.
- **Azure Data Factory (ADF)** is used to build and manage **ETL** data pipelines and leveraged its features for data orchestration, data loading and transformation.
- **Azure Databricks** is used to develop complex data transformations, such as merging, cleaning, transforming, creating features, and aggregating and joining data.
- Created Spark applications with PySpark and Spark SQL on Databricks to extract, transform, and aggregate data from diverse file formats such as Parquet, Delta, JSON, CSV.
- **Optimized** SQL, NoSQL, Scala, Python, PySpark, spark SQL queries and code to improve performance and scalability, using techniques such as data partitioning, indexing, and caching.
- Conducted daily checks of **ETL log files** to monitor the status of the pipelines, and took necessary actions, such as troubleshooting issues, rerunning pipelines, or making code fixes.
- Also monitored ETL log files on Databricks daily to track pipeline statuses and took necessary actions for troubleshooting or making code fixes.
- Built ETL pipelines using Azure Data Factory (ADF) and load the data into **Snowflake** for specific business units.
- Utilized Snowflake SnowSQL and UDFs to perform complex data transformations within the Snowflake environment, reducing data movement and improving efficiency.
- Leveraged Snowflake's cloud-based architecture for high elasticity, scalability, and recovery for data warehouses.
- Built **stored procedures** to connect and extract data from various sources like On-prem & On-cloud Database, EPR applications etc., ingest data into an Azure Data Lake and finally loaded into NoSQL and Azure SQL Database.
- Monitored **database health checks** on a daily basis, and performed necessary maintenance tasks, such as checking storage space, deleting unused data, dropping and rebuilding indexes.
- Developed and implemented **Slowly Changing Dimensions (Type 2 & 3)** and generated **Snapshots** of the dimension to track historical changes in the data to study how dimension attributes have changed over time.
- Provided extensive **Application Support** and addressed all Incidents and Service Requests created through the Service-Now ticketing tool within the given service level agreement (**SLA**), ensuring that no SLAs were missed.
- Deployed ADF pipelines and entities to production using a **Dev-Master-Release** branching strategy and implemented hotfixes using a separate hotfix branch to minimize disruption to production users.
- Leveraged GitHub to manage and version control the source code for data pipelines and data warehousing solutions, and to automate the build, test, and deployment process using GitHub Actions.
- Used **Azure DevOps** to deploy CI/CD pipelines, to move data from Dev to UAT and production environments.
- Conducted comprehensive **Exploratory Data Analysis** using SQL and Python to identify patterns, trends, and anomalies to gain insights about the data. Built ML models to conduct predictions using R Language.
- Developed interactive reports and Dashboards using visualizations tools like **Power BI**, **Tableau, Python libraries.**
- Implemented **Azure Synapse Analytics** for data warehousing for efficiently analyzing large volumes of data.
- Applied best practices for optimizing data loading and transformation within Azure Synapse.
- Leverage **Airflow**'s DAGs (Directed Acyclic Graphs) to create and manage complex data pipelines.

Environments: Azure Data Factory (ADF), Azure Databricks, Azure Monitoring, Key Vault, Azure SQL Server, Azure Data Lake Gen 2, Cosmos DB, Azure Synapse, SSIS, R Language, Python, PySpark, Spark SQL, Scala, SQL Server, PostgreSQL, NoSQL, Hadoop, Power BI, Azure DevOps, file formats (Parquet, delta, Json, CSV), GitHub, Medallion architecture, Snowflake, OLTP, OLAP.

Client : Emids, Franklin, TN Mar 2020- Jul 2022

Role : Azure Data Engineer

Key Responsibilities:

- Responsible for developing the implementation strategy for the Modern Datawarehouse solution – that involves migrating on-prem data to Azure.
- Involved in creating Integration/ETL Architecture and design artifacts using ETL tools like SSIS and Azure Data Factory (ADF) for moving legacy application data to newer environments and then to Azure systems.
- Utilized Azure Data Factory (ADF) for the orchestration and administration of SSIS packages, ensuring smooth integration and execution within Azure's data workflows.
- Transitioned SSIS packages to Azure Data Factory (ADF) for seamless integration and execution within Azure's data workflows.
- Created Stored procedures within client framework that has led to rapid increase in the number of data sources being landed in the Azure Data Lake using modern technologies like Azure Data Factory (ADF), Databricks.
- Designed SIT framework tightly integrated with ETL accelerators which would help comparing data quality results between on-prem and Azure systems.
- Extensive use of Scala, Pyspark, SQL, Python in Azure Databricks to perform data transformations, and loading.
- Designed the batch processing schedules, intraday schedules for processing daily loads and real time data within the Integrated Data Warehouse.
- Build capabilities for streamlining the Support process by adding notification and audit framework by utilizing log analytics and diagnostic setting. Also utilized the notification capabilities of Azure Data Factory (ADF).
- Involved in release management process to streamline the ADO code repository management and CI/CD process for smooth deployment of code and spinning up infrastructure.
- Defined acceptance criteria for each user stories in Azure DevOps Sprint dashboard and perform testing, troubleshooting and maintenance activities for different applications.
- Used Azure Streaming analytics to ingest IIOT data and automating the data load process to Data Warehouse.
- Explored opportunities to automate processes, improve application performance, enhance data quality, and ensure system reliability, leveraging SSIS and ADF functionalities for workflow automation and data governance.
- Administered and managed security through IAM portal groups to add and remove users and authenticate access to relevant environment, resource groups etc.
- Tuned and optimized SQL and NoSQL queries in Azure Synapse, enhancing overall system responsiveness. Integrated Azure Synapse with Power BI for real-time reporting and visualizations, providing stakeholders with up-to-date insights.
- Employed SQL querying and Python scripting skills to conduct a comprehensive EDA, identifying patterns, trends, and anomalies to unlock data-driven insights.
- Developed implementation strategy involving Airflow to orchestrate on-prem data migration to Azure using Data Factory and Databricks. Designed ETL architecture using Airflow DAGs to manage data movement.

Environments: Azure Data Factory (ADF), Azure Databricks, Azure SQL server, Azure Data, Azure Synapse Analytics, Azure Stream Analytics, Cosmos DB, Python, PySpark, Spark SQL, SSIS, SQL Server, PostgreSQL, NoSQL, Hadoop, SAS, Scala, Power BI, Tableau, Azure DevOps, Ansible, Terraform, Airflow, OLTP and OLAP.

Client : Publicis Sapient, **Boston, MA** **Jan 2019 – Feb2020**
Role : Data Engineer

Key Responsibilities:

- Worked on development of data ingestion pipelines using ETL tool, Talend & bash scripting with big data technologies including but not limited to Hadoop, Hive, Spark, Scala, Kafka.
- Experience in developing scalable & secure data pipelines for large datasets.
- Expertise in extracting data from data sources, data quality check, transformations, and metadata enrichment.
- Supported data quality management by implementing proper data quality checks in data pipelines.
- Demonstrated expertise in designing, implementing, and maintaining data pipelines using AWS services such as Amazon S3, RDS , EC2, Amazon Redshift, Amazon Glue, Kinesis, AWS Lambda, Amazon EMR etc.
- Utilized AWS Elastic MapReduce (EMR) for processing large data and implementing batch processing workflows.
- Designed end-to-end data pipelines, integrating various data sources into a centralized data lake using AWS.
- Implemented ETL processes using AWS Glue to cleanse, transform, and enrich raw data for downstream analytics.
- Leveraged Amazon EMR for distributed data processing, performing tasks such as data transformation, aggregation, and analysis on large datasets using frameworks like Apache Spark and Hadoop.
- Designed and maintained data warehousing solutions using Amazon Redshift, optimizing query performance, and ensuring data integrity for reporting and analytics purposes.
- Developed serverless data processing workflows using AWS Lambda, orchestrating data movement and transformation steps while minimizing operational overhead.
- Optimized SQL, NoSQL, Python, PySpark, spark SQL queries using techniques such as data partitioning, indexing.
- Analyzed data using Python, identifying patterns, trends, and anomalies to unlock data-driven insights.
- Exploring DAGs, their dependencies and logs using Airflow pipelines for automation with a creative approach.
- Designed and implemented a fully operational production grade largescale data solution on Snowflake.

Environment: Data Migration, S3, RDS, EC2 , Agile, Kinesis, NoSQL, Kafka, Lambda, Glue, PySpark, EMR, Data bricks, Hive, Athena, Step Functions, Lambda, Airflow, DynamoDB, Scala, Hadoop, CI/CD, SQL, Tableau, Python, PostgreSQL, Agile, Snowflake, OLTP and OLAP systems, Stream and Batch Processing.

Education : Trine university masters in information science