**AMRUTHA UPPU**

**AWS Certified DataEngineer Associate & Data Bricks Certified Data Engineer Associate & SnowPro® Core Certification**

Senior Data Engineer with expertise in building, and maintaining scalable, high- performance data pipelines. Skilled in modern big data technologies such as Apache Spark, Kafka, Hadoop, and Databricks. Experience working on cloud platforms AWS and Azure. Proficient in creating robust data pipelines, ETL / ELT processes to transform raw data into actionable insights. Proven ability to work with complex datasets, ensuring data quality and integrity. Experienced in implementing data warehousing solutions, enabling real-time analytics and large-scale batch processing to support business intelligence and data-driven decision making.

## Professional Summary:

- Data Engineer with 6+ years of experience in big data processing, ETL development, and designing cloud-based architectures on AWS and Azure to deliver scalable, high-performance data solutions.
- Proficient in AWS-based data engineering solutions, leveraging **AWS Glue**, **AWS EMR, Redshift, S3, Lambda, Athena and DynamoDB** to build robust, scalable data pipelines to analyze large datasets efficiently .
- Expertise in ETL development, designing and optimizing workflows using AWS Glue**, Step Functions, and Airflow**, ensuring seamless data transformation, cleansing, and ingestion.
- Proficient in Data Quality, Requirements Analysis, Data Modeling, Master Data Management (MDM), Data Warehousing, Data Analysis.
- Proficient in Utilizing **Apache Airflow** in developing and managing **DAG**s to orchestrate complex data workflows and automate dependency-driven pipeline execution.
- Proficient in real-time streaming solutions with **Apache Kafka, Spark Streaming, Azure Event Hubs and AWS Kinesis** enabling real-time analytics and event-driven processing.
- **Proficient in Azure-based data engineering solutions**, leveraging **Azure Data Factory, Databricks, Synapse Analytics, ADLS, Cosmos DB, Azure Functions and Azure DevOps** to build robust, scalable data pipelines to process large datasets efficiently.
- Experience in working with both **relational** (MySQL, Oracle, PostgreSQL) and **NoSQL** (MongoDB, DynamoDB,Cassandra) databases.
- Experience with **AWS IAM**, security best practices, and access control policies, ensuring compliance
- Experience with Azure Active Directory (AAD), Role-Based Access Control (RBAC), and security best practices, implementing access control policies and ensuring compliance with enterprise and regulatory standards
- Experienced in building robust, scalable data pipelines across diverse platforms, leveraging **Python**, **PySpark, SQL, Snowflake SQL & T-SQL** for efficient data transformation and processing .
- Proficient in data modeling, implementing **star and snowflake schemas**, optimizing query performance in **Snowflake**, **Redshift**, and **Synapse Analytics** for analytical workloads .
- Automated infrastructure provisioning using **Terraform**,enabling deployments of AWS / Azure Resources.
- Optimized AWS S3 and **ADLS(Azure Data lake)** storage by implementing efficient lifecycle management policies, reducing storage costs while maintaining data availability.
- Familiar with **Alteryx** for building visual ETL workflows, data cleansing, and automation of reporting processes.
- Built **Data Marts** to provide faster, domain-specific insights for business teams, and applied **Data Mesh** principles to promote decentralized ownership and improve data accessibility across the organization.
- Experienced in building and managing **CI/CD** pipelines using GitHub, Azure DevOps, BitBucket to enable seamless, version-controlled deployments across development, staging, and production environments
- Solid understanding of data security and compliance best practices, including LDAP authentication, **SSL encryption**, **RBAC**, and **data masking,** ensuring adherence to **GDPR, HIPA**A, and industry regulations.
- Experienced in using python libraries like **Pandas, NumPy, SQLalchemy, PySpark**,**Boto3** and Matplotlib.
- Hands-on experience with **Databricks**, including developing production-grade notebooks for ETL, writing PySpark / Spark SQL transformation logic and managing access through **Unity Catalog**
- Consistently recognized for outstanding contributions, with a proven ability to solve complex infrastructure

**PROFESSIONAL EXPERIENCE:**

**Project Name: Customer 360 (C360)** July 2023 to Present
**Client:Duke Energy, Role: Sr. Data Engineer**

**ROLES & RESPONSIBILITIES:**

- Contributed to Customer 360**,** a unified data product supporting all business portfolios at Duke Energy, leveraging Snowflake as the core data warehousing solution
- Established a Single Source of Truth for Heat Type Detection, enabling classification of consumer energy usage patterns — including heating, gas, and electric vehicle (EV) consumption.
- Implemented data mesh architecture, ensuring distributed data ownership while maintaining governance and compliance
- Built and maintained AWS EMR clusters for large-scale distributed data processing, optimizing resource allocation for cost-effective computation
- Designed and developed modular ETL pipelines using PySpark, SparkSQL, and Python, curating datasets to support targeted energy-saving programs and load management strategies.
- Implemented Data Pipelines sourcing from Kafka and writing the Data into AWS S3 with CDC and SCD Type 2 applied.
- Leveraged Snowflake Time Travel and Zero Copy Cloning for efficient testing, debugging, and recovery without impacting production workloads
- Utilized Snowpipe for continuous ingestion from AWS S3, reducing latency in data availability for downstream BI dashboards
- Developed complex SQL queries, stored procedures, and Python scripts to analyze large datasets, enforce business rules, and generate dashboards for business insight.
- Designed and orchestrated data workflows, integrated with AWS Step Functions, CloudWatch, and Lambda for event-driven automation and monitoring
- Integrated dbt (Data Build Tool) for modular SQL-based data transformations, testing, and documentation, enabling reusable models and version-controlled transformations within Snowflake.
- Implemented a scalable and automated data pipeline from Salesforce to AWS, leveraging AWS Glue, AWS Lambda, and Amazon S3 for efficient data extraction, transformation, and storage
- Feed data into AI/ML models that support load forecasting and customer segmentation
- Exposed key data products and ML insights through RESTful APIs, implemented using Flask, SQLAlchemy ORM, AWS Lambda, and API Gateway.
- Orchestrated and automated workflows using Apache Airflow, scheduling and monitoring pipelines that integrated S3 ingestion, EMR preprocessing, and Snowflake transformations.
- Ensured GDPR and compliance measures were met by implementing data masking, encryption, and RBAC policies within S3
- Worked closely with BI developers and business stakeholders to develop interactive Power BI dashboards using Snowflake as a data source for real-time insights.
- Implemented Data Marts, Semantic Layers, Reporting Layers in at Enterprise Data level in TeraBytes Volume
- Implemented secure, automated CI/CD pipelines using Terraform and Concourse, ensuring versioned repeatable deployments across environments.

**Environment:** AWS EMR, ,Glue, S3, Athena, Lambda, Step Functions, Cloud Watch, AWS KMS ,API Gateway, Lake Formation, Iceberg,DBT, Airflow,Kafka, VPC, Snowflake ,DBT,Terraform, Flask,Power BI,Python, SQL, PySpark,Git

**Project Name: Pandora OSS & DIME** April 2021 – January 2023
**Role: Data Engineer, Employer: CTS**

**ROLES & RESPONSIBILITIES:**

- Worked closely with business users to understand their requirements and delivered data solutions that met their needs effectively.
- Migrated terabytes of data from on-premises systems to the cloud using both full and incremental load strategies.
- Designed and implemented end-to-end data pipelines using Azure Databricks, Azure Data Factory (ADF), Azure Data lake Storage (ADLS) for extracting, transforming, and loading (ETL) large volumes of structured and unstructured data
- Developed and executed unit and functional tests in Databricks, improving data pipeline reliability and reducing production defects by 30%
- Developed ETL pipelines using PySpark, Spark SQL, and Python for transforming high-volume of data
- Worked with Azure SQL Database to develop SQL scripts for extraction, transformation, and validation, ensuring accurate and consistent reporting.
- Developed a data analytics solution using Azure Data Lake and Synapse Analytics to improve data quality, and provide reliable data for downstream teams.
- Optimized large-scale Spark workloads on Databricks by applying broadcast joins, caching, partition tuning, and Z-ORDER clustering improving query performance and reducing compute costs
- Automated data quality profiling and validation in PySpark to detect null values, schema drift, skewed distributions, and date anomalies, ensuring reliable and consistent datasets
- Managed cross-environment testing across Dev, QA, UAT, and Production stages.
- Deployed data pipelines using CI/CD practices in Azure DevOps (VSTS), ensuring reliable and automated delivery
- Prepared detailed documentation covering specifications, requirements, and testing to support smooth development and delivery
- Guided team members through interactive knowledge-sharing sessions in collaboration with the manager

**Environment:** AzureDataLake, Azure Data factory,Data Bricks, Key Vault, PySpark, Python, SQL, T-SQL, Synapse,Azure DevOps, SQL Server Management Studio

**Employer: CTS** Jan 2019 – March 2021
**Role: Data Engineer Location: India**
**Projects: Unilever Data Ingestion, UDL Gen2 Migration**

**ROLES & RESPONSIBILITIES:**

- Moved 400+ TB of data from ADLS Gen 1 to ADLS gen2 Shell Scripts and ADF
- Designing and developing real-time ETL/ELT pipelines using ADF (Azure Data Factory) based on the Source requirements.
- Testing the Data Pipelines using both Standard test templates and custom test cases.

- Performed data quality checks in Azure Databricks (Spark), including null checks, duplicate detection, schema validation, and referential integrity checks, to ensure data accuracy and integrity across pipelines.
- Developed real-time data pipelines using Azure Functions and Event Hubs to ingest streaming data into Azure Data Lake Storage (ADLS) and Azure Data Explorer (ADX), leveraging Stream Analytics for transformations
- Deployed and automated data pipelines using CI/CD practices with Azure DevOps (VSTS), ensuring reliable, repeatable, and faster releases
- Databricks notebooks to process structured Data and Semi Structured Data (CSV, JSON, Parquet etc.) as part of the transformation process.
- Worked on the DevOps lifecycle by validating deployments, monitoring pipelines, rerunning jobs when needed, tuning performance, and fixing defects to keep data pipelines running smoothly.
- Prepared design and technical documentation based on low-level design, following client best practices and standards
- Provided production support by responding to critical events and alerts, and performed in-depth Root Cause Analysis (RCA) and Root Cause Failure Analysis (RCFA) to resolve issues and prevent recurrence.

**Environment:** AzureDataLakeStore, Azure Data factory, Azure DataBricks, Azure Key Vault, Event hubs, kafka, Azure Data Explorer(ADx), Azure Stream Analytics, Azure DevOps ,PySpark, SQL and Python.

**TECHNICAL SKILLS:**

- **Programming Languages**: Python, SQL, PySpark
- **Big Data Technologies:** Apache Spark, Hive, HDFS, MapReduce, Sqoop, YARN
- **BI Tools:** Tableau, Power BI
- **Data Warehousing:** Azure Synapse, Snowflake,Redshift
- **Streaming :** Apache Kafka, Event Hub, Stream Analytics, Kinesis
- **Database Management:** MySQL, HBase, Cosmos DB, Cassandra, Snowflake.
- **Tools and Software:** GitHub, SharePoint, JIRA, Confluence
- **Cloud Platforms:** AWS (S3, EC2, RDS, Redshift, Lambda, EMR, Kinesis, DMS, DynamoDB, API Gateway, cloud watch, Step Functions, IAM, Glue, Athena,Cloud Formation,Lake formation),Microsoft Azure (Data Lake, Data Factory, DataBricks, Azure SQL, Logic apps, Azure Functions, Event hubs, VM, Azure Storage, Azure DevOps, Snowflake)
- **Containerization and Orchestration:** Docker, Kubernetes, Airflow
- **Version Control Systems:** Git, SVN
- **Data Skills:** Data Visualization, Data Modeling, Data Normalization, Data Warehousing, Data Mining, Data Analysis, Data Quality, Data Integration, Data Transformation, Data Cleansing
- **Machine Learning Frameworks:** Supervised, Unsupervised, NLP, Deep Learning, LLM

## Education:

Bachelor of Technology in Electronics and Communication July 2015 – May 2019
CGPA: 8.9
**Vignan University, Vadlamudi ,** India