**Abhijith Singh Thakur**

Austin, TX • (512) 942-7025 • abhijthsnght@gmail.com

# Summary

Results-driven **Data Engineer** with over **6 years of experience** designing, building, and optimizing data pipelines on **AWS cloud infrastructure**. Skilled in **ETL development, data lakes, and real-time streaming**, with hands-on expertise in **AWS Glue, Amazon Redshift, Apache Spark, and Kinesis**. Adept at writing optimized **SQL queries** and leveraging **Python and Terraform** for automation and infrastructure as code. Strong understanding of **data governance, security, and compliance**. Passionate about improving **data quality, performance, and efficiency** to drive business insights.

**Technical Skills:**

* **Programming Languages:** Python, R, Spark, Java, C, C++.
* **Data Engineering:** Apache Kafka, Kinesis, Hadoop, PySpark, Apache Airflow, Informatica PowerCenter, Hive, HDFS, API Integration, Snowflake, Azure Data Lake, Azure Data Factory, AWS Redshift, Data Warehousing, Alteryx,
* **Data Processing Technologies:** ETL/ELT, data replication, message queuing technologies (Apache Kafka, Kinesis), Spark, Flink, API-based data acquisition, Erwin.
* **Scripting Languages:** Python, Linux Shell Scripting.
* **Deployment & Configuration Management:** Terraform, Ansible, Chef, Kubernetes, Docker, Amazon ECS.
* **Data Analysis & Visualization:** Tableau, Power BI, Excel, SciPy, Matplotlib.
* **Machine Learning:** TensorFlow, PyTorch, Scikit-Learn, CNN, ML Model Integration, Feature Engineering.
* **Cloud Platforms:** AWS (S3, EMR, Lambda, Athena, Redshift), Azure (Data Factory, Synapse, Data Lake Gen 2, Logic Apps, Fabric) Google Cloud.
* **Project Management:** Agile, Scrum, JIRA, Confluence.
* **Database Management:** PostgreSQL, MySQL, Oracle Database, Azure SQL Database, Microsoft SQL Server.
* **Version Control:** Git, GitHub.
* **CI/CD:** Jenkins, Azure DevOps.

# EXPERIENCE

**Client – ERS State of Texas Austin, Texas**

## Data Engineer Jan 2024 - Present

***Responsibilities:***

* Assisted in developing and maintaining ETL pipelines for ERP systems used by Texas state agencies, streamlining financial reporting, payroll processing, and procurement data flows.
* Supported data lake management on Amazon S3, implementing basic data organization strategies like folder structuring and lifecycle policies, helping reduce storage costs by 15%.
* Developed batch data ingestion workflows using AWS Glue and AWS Lambda, reducing manual data processing efforts by 40%.
* Wrote and optimized SQL queries for reporting in Amazon Redshift, improving query response time for ERP-related reports by 25%.
* Assisted in setting up real-time data pipelines using Amazon Kinesis, ensuring efficient streaming of financial transactions with reduced latency.
* Monitored and troubleshot daily data workflows using Amazon CloudWatch, identifying and resolving pipeline failures to minimize downtime.
* Implemented basic data validation checks using Python and Pandas, improving data quality and accuracy.
* Gained experience in data security and governance, supporting the setup of AWS IAM roles and S3 access policies to comply with government data regulations.
* Worked closely with cross-functional teams in an Agile environment, assisting in requirement gathering and data workflow optimizations
* Helped integrate Terraform for infrastructure automation, reducing deployment time for AWS resources by 50%.

**Tech Stack:** AWS Glue, AWS Lambda, Amazon Redshift, Amazon S3, AWS Glue Data Catalog, Amazon Kinesis, Python, Pandas,

SQL, AWS IAM, AWS CloudWatch.

**Client - Mars McLean, Viginia**

## Data Engineer Sep 2021 – Dec 2023

***Responsibilities:***

* Assisted in migrating structured and unstructured data from Azure Data Lake and Azure Blob Storage to Amazon S3, improving data retrieval speed by 25% for supply chain and sales analytics.
* Developed and maintained ETL workflows using AWS Glue and AWS Lambda, automating data ingestion processes and reducing manual intervention by 40%.
* Supported the transition of Azure Synapse Analytics workloads to Amazon Redshift, assisting in query optimization and schema design to enhance reporting efficiency.
* Helped implement real-time data streaming solutions using Amazon Kinesis, replacing Azure Event Hubs to support sales transaction tracking and inventory management with 30% lower latency.
* Assisted in migrating batch processing jobs from Azure Databricks to AWS EMR (Apache Spark and PySpark), improving data transformation speed for marketing analytics.
* Helped in the migration of NoSQL databases from Azure CosmosDB to Amazon DynamoDB, ensuring smooth data transformation and integration with existing applications.
* Created basic data validation scripts using Python, Pandas, and SQL, automating post-migration quality checks to ensure data integrity.
* Monitored AWS services and data pipelines using Amazon CloudWatch, setting up basic alerts and tracking performance metrics for supply chain and retail analytics dashboards.
* Assisted in implementing security best practices by supporting AWS IAM role configurations and S3 bucket policies, ensuring compliance with data privacy regulations.
* Worked closely with cross-functional teams, including data analysts and business stakeholders, to align migration goals with business needs.

**Tech Stack:** AWS Glue, AWS Lambda, Amazon Redshift, Amazon S3, AWS Glue Data Catalog, Amazon Kinesis, Python, Pandas, SQL, AWS IAM, AWS CloudWatch.

**Client - Molina HealthCare Irving, Texas**

## Data Engineer Apr 2020 – Aug 2021

***Responsibilities:***

* Assisted in developing and maintaining ETL pipelines using Azure Data Factory, ensuring secure and efficient data movement for patient records, claims processing, and regulatory reporting.
* Helped manage large-scale healthcare data storage on Azure Data Lake, improving data accessibility for medical analytics, insurance processing, and compliance tracking.
* Supported the integration of Electronic Health Records (EHR) and insurance claims data into Snowflake Data Warehouse, optimizing storage and retrieval processes while reducing infrastructure costs by 20%.
* Assisted in building data ingestion workflows for clinical trial data, patient monitoring systems, and hospital inventory management, leveraging Azure Functions and Azure Logic Apps for automation.
* Developed basic data validation scripts in Python and Pandas to ensure HIPAA-compliant data accuracy, reducing errors in medical claims processing.
* Helped enhance data partitioning, columnar compression, and parallel processing techniques in Azure Synapse Analytics, improving query performance for healthcare analytics by 30%.
* Supported GraphQL API development for secure and flexible data retrieval of patient records, diagnostic results, and treatment histories, reducing unnecessary data transfers by 25%.
* Assisted in setting up data governance and security policies in Azure, including role-based access controls (RBAC) and encryption standards, ensuring compliance with HIPAA and GDPR regulations.
* Monitored and troubleshot data pipelines using Azure Monitor and Log Analytics, identifying and resolving pipeline failures to minimize downtime for healthcare analytics dashboards.
* Collaborated with healthcare data analysts, compliance teams, and business stakeholders to ensure accurate and timely data availability for medical research and patient care improvements.

**Tech Stack:** Azure Data Factory, Azure Data Lake, Snowflake, Azure Synapse Analytics, Azure Functions, Azure Logic Apps, Apache Spark, PySpark, Apache Airflow, GraphQL, Python, Pandas, SQL, Azure Monitor, Azure DevOps.

**Client - Zerodha** **Bangalore, India**

## Data Engineer Feb 2019 – Mar 2020

***Responsibilities:***

* Assisted in developing SQL-based ETL workflows to process financial transactions, trade logs, and customer portfolio data, reducing data processing time by 30%.
* Helped design real-time data pipelines using AWS Glue and Lambda, automating data ingestion for market order flows, trade settlements, and risk analytics.
* Developed Python scripts for automated data validation to ensure trade execution accuracy, reducing discrepancies in equity and derivatives trade processing.
* Optimized PostgreSQL and MySQL queries to improve market data retrieval speed, reducing response times for customer trade history and P&L reports by 20%.
* Assisted in integrating stock market APIs to fetch real-time pricing, historical stock data, and order book analytics, improving trading strategy models.
* Monitored and troubleshot trading data pipelines using AWS CloudWatch, reducing job failures and improving data availability for financial risk assessment dashboards.
* Implemented basic data security best practices by assisting in setting up AWS IAM roles and S3 bucket encryption, ensuring compliance with SEBI regulations.
* Supported the migration of historical trading data from on-prem SQL databases to AWS S3, enabling scalable storage and retrieval for trade analytics and back testing.
* Assisted in automating customer risk profiling using Python and SQL, improving fraud detection and compliance monitoring in financial transactions.
* Collaborated with quantitative analysts and trading teams to optimize data pipelines for faster execution of market insights and algorithmic trading strategies.

**Tech Stack:** SQL, PostgreSQL, MySQL, Python, AWS Glue, AWS S3, AWS Lambda, AWS CloudWatch, Stock Market APIs, Pandas, SEBI Compliance, Financial Data Processing

# EDUCATION

**LAMAR UNIVERSITY** Beaumont, Texas

*Master of Science in Computer Science*

# CERTIFICATIONS

* **AWS CERTIFIED DATA ENGINEER – ASSOCIATE**
* **Database programming with SQL - Oracle Academy**
* **DP-203: AZURE DATA ENGINEER**