Aditya Naik

979-633-8551 | adityavnaik97@gmail.com | adityanaik.info | LinkedIn | GitHub

SUMMARY

AWS certified data engineer with 4+ years of experience in building and optimizing ETL data pipelines, data warehouses and data analytics systems using AWS, GCP, Spark, Flink, Kafka and PowerBI to drive business insights and decision-making.

EDUCATION

Texas A&M University – ***College Station, TX, USA*** *September 2022 – May 2024*
Master of Science in Management Information Systems

Pune University – ***Pune, India*** *August 2015 – May 2019*
Bachelor of Technology in Computer Engineering

SKILLS

Certifications: *AWS Certified Developer, Professional Scrum Master*
Languages: SQL, Python (Pandas, NumPy), Java, Scala, Bash, JavaScript, Groovy
Database: NoSQL (MongoDB, DynamoDB, Cassandra), Relational Databases (Postgres, MySQL, MSSQL)
Data: MapReduce, HDFS, Hadoop, HBase, Flume, Airflow, Kafka, SSIS, Snowflake, Dbt
BI Tools: Tableau, Power BI, Grafana, Looker, ClickHouse
Cloud: AWS (Redshift, EMR, Glue), GCP (BigQuery, DataFlow, Pub/Sub)
Other: Git, JIRA, Jenkins, Terraform, Docker, Kubernetes, Agile, Scrum

EXPERIENCE

TAMU Institute of Technology Infused Learning *September 2022 – May 2024*
Data Engineer *– Spark, Hive, HDFS, Databricks, Python, Postgres, Tableau, Power BI, Data Warehouse, Excel*

- Implemented Spark jobs for distributed data processing using Python – Pyspark and SQL – HiveQL in HDFS within Hadoop framework on Databricks, enhancing data cleaning and data integration efficiency by 5x.
- Designed ETL workflows in Spark via DAGs, and integrated 120 GB of customer demographics, transactions and sales revenue data from 50 different Excel sources in PostgreSQL.
- Established data warehouse and data marts in Star and Snowflake schema via dimensional data modeling for MPP databases.
- Crafted Tableau dashboards to visualize effects of marketing campaigns on sales, empowering decision-making and reducing analysis time by 10x.

HSBC Holdings PLC *October 2020 – July 2022*
Data Engineer II *– SQL, ETL, ELT, Kafka, Flink, AWS, S3, Redshift, Airflow, Data Pipeline, Data Lake, DBT, EMR*

- Optimized complex SQL queries and performance tuning through indexes, partitions, stored procedures, triggers and aggregations, resulting in faster generation of financial reports for stakeholders by 59%.
- Constructed scalable ETL /ELT data pipelines using Airflow and Apache Flink in EMR for data processing to migrate legacy order management system (Fidessa) to AWS, saving $40,000/year in license and operational costs.
- Implemented Kafka consumers for data ingestion and real-time streaming of financial transaction events of petabyte scale from topics through RESTful APIs, reducing asynchronous stream and batch processing latency by 65%.
- Collaborated with cross-functional teams of analysts, creating big data architecture through data models and NoSQL schemas for data warehouse – Redshift and data lakes to support 15 analytical reporting systems.
- Utilized SQL for ad-hoc data analysis and reporting, providing actionable insights to stakeholders and supporting financial processes for Equities, Derivatives and Fixed Income products, ensuring data consistency and reliability across systems.

HSBC Technology *July 2019 – October 2020*
Software Engineer *– Java, CI/CD, Jenkins, Docker, Kubernetes, Unix, Bash, Git*

- Built Jenkins CI/CD pipelines and used Kubernetes for orchestrating Docker containers, elevating release frequency by 5x.
- Automated Linux virtual machines provision in GCP via shell scripts, reducing manual intervention to 0% in production servers.
- Coordinated Scrum ceremonies in Agile teams using JIRA and provided code documentation via version control tools.

PROJECTS

Online QnA forum using AWS | *GitHub*

- Created data storage in DynamoDB, S3 and caching in Redis for 90 GB of unstructured data of students and questions.
- Architected ETL data pipelines using AWS Glue, to extract questions, aggregate student marks and load into Redshift.
- Designed microservices as AWS Lambda functions in Python and deployed via Terraform reducing deployment time by 45%.

Gen AI model for disease classification (Hackathon Winner) | *Link*

- Generated synthetic healthcare datasets for preliminary disease using NLP and TensorFlow.
- Performed predictive analytics using Gen AI on historical database of physician comments with 98% accuracy