RAHUL REDDY

• San Jose, CA • 6693040995 • rahulr10reddy@gmail.com • LinkedIn • GitHub

Data Engineer with 5+ years of experience in managing terabytes of data, building batch & streaming data pipelines serving Machine Learning & Analytics use cases.

WORK EXPERIENCE

Gen AI Data Engineer | LEXIM AI Oct 2023 – May 2024

- Orchestrated & built data pipelines for regulatory data, supporting **data modeling & LLM** use cases. Achieved improved **data quality** & model predictions while minimizing cloud costs by **32%.**
- Developed APIs to efficiently retrieve metadata & **2TB** of documents from regulatory websites.
- Architected a self-managed, multi-node **vector database** infrastructure, reducing database costs by **90%** while implementing a **real-time** vector embeddings pipeline for RAG functionalities.
- Collaborated with cross functional team to implement **stored procs** and **fact tables** for web application.

Tech stack: Python, SQL, Airflow, AWS (S3, EC2, AWS Lambda, MySQL, IAM, CloudWatch, CloudFormation), Elasticsearch, Docker, Kafka, JSON, ETL, Flask, CI/CD, IaC, Lang chain, LLM, Linux, Pandas, Regex.

Data Engineer II | Verizon Jun 2021 – Jan 2022

- Led the Network AD project, developed a scalable **cloud pipeline** to process telemetry data of 10 Million+ rows for near real-time **anomaly detection**, minimized manpower required for monitoring by 80%.
- Optimized the **distributed processing** resources from 600 GB to 40 GB, 200 cores to 10 cores, and the processing time from 22 min to 8mins.

Tech stack: Python, Spark, Pyspark, Apache Kafka, EMR, Redshift, Apache Iceberg, PowerBI, Airflow

Data Engineer | Genpact Oct 2018 - Jun 2021

- Designed & Implemented **ETL** pipelines for 2 projects involving procurement spend and Inventory data, processed and transformed 6-7 terabytes of **streaming data** & loading into a Datawarehouse.
- Worked with finance team to implement **CDC pipeline** for getting real-time new users info while stream processing multiple tables on fly, which lowered the waiting time by 94%.
- Architectured and executed Infrastructure as Code (IaC) & **automated** provisioning of server configuration, resulting in a 60% reduction in deployment time.
- Facilitated fault-tolerant **logging solution** for 300+ microservices and critical service applications by replacing Splunk with opensource Elasticsearch on 112 Nodes.

Tech stack: Hive, HDFS, Spark, Confluent Kafka, kstreams, OLAP, ELK, Powershell, Ansible, Terraform.

Data Analytics Engineer | Mahindra Rise Aug 2017 – Aug 2018

- Gathered data from multiple sources and visualized **KPI's & KAI's** of overall plant production at different shop floors and inspection stages, also reduced **inventory reports** from 2hrs to 15mins.
- Fabricated a **DataMart** and composed 25+ ETL Pipelines for Inventory data of all the manufacturing plants.

Tech stack: SQL, SAP, SAP HANA, Tableau, MySQL, MS Excel, Azure data factory

SKILLS

Technologies: AWS, Azure, Spark, Airflow, EMR, Terraform, Docker, Kinesis, Databricks, Jenkins
Storage: RedShift, BigQuery, MySQL, HDFS, Hive, Neo4j, Postgres, snowflake
Programming Languages: Python, Bash, HTML, SQL
ETL and Visualization: DBT, Azure data factory, AWS Glue, PowerBI, Grafana, Tableau, kibana
Tools, Libraries & environments: Git, Jira, Pandas, TensorFlow, scikit-learn, Linux

EDUCATION

San Jose State University, San Jose, California.	MS, Applied Data Science, Data Analytics	**May 2024**
Hindustan University, Chennai, India	Bachelors in Engineering	**April 2017**

PROJECTS

- Fabricated a Comprehensive **Analytical Ecosystem** on AWS for E-commerce (***Flink, DynamoDB, Kinesis, SNS, S3, Quick Sight, CloudWatch, EC2***)(GitHub)
- Big data pipeline for extracting data from diff sources ***(AWS [CDK, Lambda, Glue], API, parquet)*** (GitHub)
- Knowledge Graph extraction from unstructured documents using LLMs and built Graph RAG on its knowledge base ***(Neo4j, Mistral, Langchain, RDF, ETL)*** (GitHub)