Praharsha Prateek More

Boston, MA | 774-473-9096 | Praharsha.m0209@gmail.com | linkedin.com/in/praharsha-p-784186190

SUMMARY

Data Engineer in Ad-tech, Automotive and Banking with 5+ years delivering real-time and batch data products that power ad delivery and personalization. I ship Kafka/Kinesis streams and Airflow pipelines into Snowflake, Redshift, BigQuery, and I've automated Data Subject Access Request and secure cross-cloud migrations with Terraform, RBAC, and lineage.

TECHNICAL SKILLS

- **Languages:** Python, SQL, Java, Scala
- **Data & Streaming:** Spark (PySpark), Kafka, Kinesis, Airflow, Iceberg, Delta Lake, Hive/Impala, dbt
- **Cloud:** AWS (S3, Glue, EMR, Redshift, DynamoDB) Azure (Databricks, Data Factory, Synapse, ADLS) GCP (BigQuery, Dataproc, Dataflow, Composer)
- **Warehousing:** Snowflake, Redshift, BigQuery
- **DevOps/IaC:** Terraform, Docker, Kubernetes, CI/CD (Jenkins, GitLab)
- **BI:** Power BI, Looker, Tableau
- **Governance/Quality:** Data contracts, RBAC/masking, lineage, anomaly checks, resource monitors, MDM (Ataccama)

WORK EXPERIENCE

Data Engineer — **Boston, MA**
Fox Corporation (Contract) — *Sep 2024-Present*

- Architected and maintained ad-tech data pipelines for Adrise, ensuring seamless ad delivery and monetization at scale
- Designed and automated end-to-end DSAR (Data Subject Access Request) workflows with Apache Airflow, integrating APIs and securely interfacing with Redshift, Snowflake, and other cloud databases for data access, deletion, and correction.
- Engineered secure data migration processes between Redshift, S3, and Snowflake, ensuring integrity and consistency across multi-cloud environments.
- Orchestrated robust Airflow pipelines with S3 and Snowflake for quarterly ad-hoc analytics, aggregating data from Redshift and DynamoDB and building reusable tables/stored procedures for data analysts.
- Implemented scheduled jobs to optimize DynamoDB performance by clearing month-old envelope tables, significantly improving pipeline efficiency.
- Built scalable streaming solutions with Apache Kafka and AWS Kinesis for hybrid ingestion, enhancing real-time data availability and observability.
- Deployed Apache Iceberg tables on Amazon S3 with schema evolution and partition pruning, enabling time travel queries and optimizing batch performance via PySpark on EMR.
- Delivering actionable Power BI dashboards, powered by Redshift and S3 data, reducing manual reporting effort and improving business insight delivery by 30%.
- Provided technical guidance to junior engineers on Airflow DAG design and Snowflake query optimization, improving team delivery speed and consistency.

Data Engineer Intern — **Boston, MA**
Nissan Motor Corporation — *Sep 2023-Aug 2024*

- Built Snowflake core objects like tables, views, stages, file formats and integrated AWS S3 with external stages and Snowpipe for auto-ingest.
- Designed Snowflake dimensional schemas and CDC consumer tables, enabling analysts to query 5TB+ datasets with predictable performance.
- Designed and developed AWS Glue ETL to load data from S3 (Parquet/Text) into Redshift; performed architecture assessments across EMR, Redshift and S3.
- Developed EMR, Hive and Impala pipelines: created external Hive tables on S3, wrote reusable ingestion and repair scripts, and built Hive transformations as the baseline layer.
- Wrote complex SQL and built stored procedures plus SSIS packages to support batch integrations and heavy transformations.
- Prototyped dbt models for modular SQL transformations and automated documentation in Snowflake, improving transparency and reusability
- Implemented Python AWS Lambda functions with concurrency and multithreading to speed up processing via asynchronous execution.
- Built CloudWatch dashboards and alerts to monitor ingestion pipelines, proactively resolving failures before SLA breaches
- Established reusable SQL templates, code review practices, and automated data quality checks, ensuring consistent transformations across teams.
- Set up CI/CD with Jenkins and infrastructure as code with CloudFormation; configured EC2 Auto Scaling for elastic capacity.

Project Engineer — **Hyderabad, India**
Wipro Limited — *Jun 2018-Jul 2021*

- Designed and automated a batch ETL pipeline using Azure Functions, to parse and transform 30 GB of daily transaction data from core banking systems and incorporated Change Data Capture (CDC) for incremental updates.
- Optimized data ingestion from 11 in-house databases with volume of 3 TB+ using Python scripts with Pandas, removing duplicate transaction data and achieving a 31% reduction in processing time, enabling faster forecasting of processing loads.
- Developed an automated workflow using PySpark on a distributed Spark cluster to extract risk and fraud indicators from unstructured transaction data, transforming raw financial data into structured risk profiles.
- Authored and optimized SQL queries by implementing indexing and efficient joins in Azure Synapse Analytics to streamline data retrieval, reducing query execution time by 23%.
- Deployed and managed Airflow DAGs using Azure Kubernetes Service and Azure Monitor to orchestrate and monitor data pipelines, ensuring high reliability, and automated troubleshooting workflows to detect and resolve errors.
- Built a centralized PowerBI dashboard that streamlined data aggregation and analysis, resulting in a 37% reduction in reporting time for compliance and risk management teams.

EDUCATION

University of Massachusetts Dartmouth — **North Dartmouth, MA**
Masters in Data Science — *Sep 2021-Aug 2023*

Keshav Memorial Institute of Technology — **Hyderabad, India**
Bachelors in Electronics and Communications Engineering — *Jul 2015-May 2019*

PERSONAL PROJECTS

YouTube Trending Data Pipeline

- Built an end-to-end data pipeline using AWS CLI, S3, Glue, Lambda, and Athena, processing 200K+ daily trending video records into a structured data catalog.
- Optimized Athena queries and Glue jobs to reduce compute cost by 25%, while maintaining sub-second query performance for trending video insights.
- Automated ETL workflows with Glue and SQL-based transformations, enabling fast ad-hoc queries on metadata like views, likes, and categories.
- Delivered a centralized dataset powering insight into audience engagement trends across geographies, reducing manual reporting effort by 50%.

Ad Campaign Analytics Pipeline & Dashboard

- Engineered a PySpark ETL pipeline on AWS EMR to process and aggregate 5M+ daily ad events from S3 into Snowflake, achieving sub-2s query latency for downstream analytics.
- Developed an interactive React and D3.js dashboard to visualize campaign KPIs (CTR, impressions, conversions, ROAS) with drill-down filters, cutting reporting time by 60%.
- Enabled marketing teams to track and optimize ad performance in real time, driving faster decision-making and improving campaign ROI by 18%.

Twitter Data Pipeline

- Extracted tweets in real time via the Twitter API and transformed text data with Python for sentiment and keyword analysis.
- Deployed workflows on Apache Airflow (EC2) to orchestrate ingestion, cleaning, and storage pipelines into Amazon S3 with retry logic and alerts.
- Enabled continuous monitoring of 50K+ tweets/day, providing structured datasets for trend analysis and improving analyst productivity by 35%.

CERTIFICATIONS

- Google Cloud Professional Data Engineer
- AWS Certified Data Engineer Associate