Ranaa Mahveen

862-296-4031 | rmahveen50@gmail.com | https://www.linkedin.com/in/ranaa-mahveen/

Summary

Senior Data Engineer with 8+ years of experience designing, architecting, and operationalizing large-scale data pipelines, distributed compute workflows, and analytics platforms across R&D, healthcare, finance, and enterprise environments. Expert in **Python, SQL, PySpark, Spark, Databricks, Snowflake, and dbt**, with strong cloud experience in **AWS** and working knowledge of **GCP**. Proven ability to build automated, production-grade ETL/ELT systems, optimize high-volume workloads, standardize data models, and enable data platforms that support scientific research, predictive analytics, and enterprise decision-making. Experienced working in **drug discovery and research data domains**, collaborating with scientists and engineers to design trustworthy, governed, AI/ML-ready data assets. Known for delivering robust, observable, CI/CD-integrated data services and mentoring engineering teams.

Technical Skills

- **Languages:** Python, SQL, PySpark, Java, Scala, Pandas, NumPy
- **Cloud Platforms:** AWS (S3, EC2, Lambda, Batch, CloudFormation, Athena), GCP
- **Data Engineering:** Databricks, Spark, dbt, Airflow, Control-M, Delta Lake, Snowflake, AWS Glue
- **Big Data & Streaming:**Kafka, Hadoop ecosystem, Hive
- **Data Warehousing:** Snowflake, Redshift, Oracle, PostgreSQL, MySQL
- **Containers & DevOps:** Docker, Jenkins, GitLab CI/CD, ECS
- **Data Quality:** Great Expectations, Python/SQL test frameworks, monitoring dashboards
- **Visualization:** Tableau, Spotfire, Power BI
- **Other:** JIRA, Confluence, VSCode, GitHub Copilot, Agile/Scrum

Professional Experience

Senior Data Engineer | Amgen Inc. *April 2024 – Present*

Work in **R&D and drug discovery** supporting scientific data workflows and high-volume research datasets. Lead development of Databricks/Spark pipelines, cross-domain transformations, and cloud-enabled data ingestion patterns.

- Designed and maintained **large-scale automated ETL/ELT pipelines** using Databricks (Spark) and Oracle, reducing end-to-end research data latency and increasing reliability across scientific datasets.
- Optimized Oracle SQL packages and materialized views, improving analytical query performance by **40%** for downstream R&D workflows.
- Built standardized, reusable transformation modules and parameterized pipelines, reducing logic duplication and increasing maintainability across research domains.
- Developed Python-based REST APIs enabling automated, governed movement of scientific data between research systems.
- Designed cross-functional transformation layers integrating Oracle and Databricks sources, enabling unified analytics for scientists and informatics teams.
- Participated in architectural discussions and internal POCs for GCP ingestion and scalable research-data patterns.
- Implemented CI/CD using GitLab, Docker, and Jenkins, ensuring consistent deployments and compliance with internal quality practices.
- Built automated **PySpark and SQL data validation** tests, improving accuracy, lineage, and trust of research datasets.
- Tuned Spark clusters for compute efficiency and cost optimization; improved job reliability through observability and structured logging.
- Authored documentation for ETL workflows, data models, and operational procedures; mentored junior engineers on Spark, SQL, and AWS best practices.
- Collaborated with scientists and architects to align data models with evolving research workflows, improving interoperability across drug discovery programs.

Data Engineer | Sinfotech LLC. (Client: USAA) May 2022 – April 2024

Built high-volume financial data transformations using Snowflake and dbt; led modernization of ELT workflows for core analytics.

- Developed modular dbt data models (staging, core) and optimized Snowflake SQL transformations for high-volume datasets.
- Improved Snowflake runtime efficiency through incremental models, warehouse tuning, and optimized table design.
- Orchestrated multi-environment workflows using **Control-M**, ensuring predictable and fault-tolerant execution.
- Built dbt tests, macros, and automated validation frameworks to ensure accuracy and trust in curated analytical layers.
- Automated repetitive engineering tasks using Python, reducing manual development workload by **60%**.
- Conducted comparative evaluations of Snowflake vs. GCP BigQuery pipelines to support cloud modernization strategy.
- Provided rotational on-call support, resolving high-priority pipeline failures under strict SLAs.
- Documented data flows, business rules, and ELT architecture for governance and easier onboarding.
- Trained and onboarded new data engineers, standardizing Snowflake/dbt patterns across teams.

Software Engineer | Move Inc. Mar 2020 – May 2022

Built cloud-native ingestion pipelines and Snowflake/dbt models for multi-domain enterprise analytics.

- Designed and developed ingestion pipelines using **AWS S3, Lambda, EC2, Batch**, and Snowflake.
- Built dbt models, seeds, sources, and automated schema/data/freshness tests, strengthening trust in curated datasets.
- Implemented CI/CD using Jenkins, Docker, and GitHub, enabling automated, reliable environment promotions.
- Designed Snowflake RBAC structures (roles, schemas, warehouses, secure views) ensuring compliant and controlled data access.
- Migrated Airflow workloads to **Amazon MWAA**, improving observability, job stability, and operational overhead.
- Managed schema evolution and ingestion for Parquet, Avro, JSON, and text formats.
- Built containerized data workflows using ECS and AWS Batch to improve scalability and throughput.
- Integrated CloudWatch and Splunk dashboards for monitoring, performance tuning, and pipeline reliability.
- Partnered with BI and engineering teams to translate reporting needs into Snowflake/dbt data models.

Graduate Assistant, West Virginia University Aug 2017 – Dec 2019

- Developed a chronic kidney disease prediction model using Python, scikit-learn, and XGBoost; improved minority-class recall by **20%** using SMOTE.
- Built automated preprocessing pipelines (Pandas, NumPy) on NHANES datasets.
- Designed SQL transformations and structured feature datasets for modeling.
- Collaborated with biomedical researchers to ensure clinical applicability of transformations.

Application Developer | IBM India Private Limited Jan 2014 – Dec 2016

- Developed backend Java (Spring/Hibernate) components and SQL logic for large-scale telecom applications.
- Implemented data validation and transformation logic across XML/JSON workflows.
- Built REST/SOAP services and improved backend performance through SQL optimizations.
- Worked with senior developers to troubleshoot data inconsistencies and improve data access patterns.
- Delivered UI enhancements in AngularJS/JavaScript to improve usability.

Education

Master of Science, Computer Science – West Virginia University

Bachelor of Technology, Computer Science – Jawaharlal Nehru Technological University