VIRAJ RATHOD

virajrathod99@gmail.com | 623-396-9127 | linkedin.com/in/virajrathod | Salt Lake City (Ready to Relocate)

EXPERIENCE

DigiCert – Senior Data Engineer *April 2020 - Present*

- Led a streaming project using PySpark on Databricks to ingest and parse certificates from certificate transparency logs, enabling the BI team to analyze customer metrics.
- Designed and developed automated ETL data pipelines, improving efficiency by 45%.
- Implemented real-time, scalable data pipelines using Kafka and CDC pipelines, enhancing data migration by 33%.
- Created a Python framework for dynamic certificate generation, showcasing strong scripting skills.
- Developed complex business queries for strategic reporting and advanced analytics.
- Designed QA Audit scripts to identify anomalies, outliers, and workflow issues.
- Built real-time dashboards on Grafana for machine tracking and alerting.
- Developed a cost analysis dashboard using AWS metrics in Grafana and MySQL, providing insights into cost optimization strategies.

University of Utah – Data Science Intern *September 2019 – April 2020*

- Engineered data lineage models using Scala and Spark's GraphX, centralizing data availability and visualization.
- Implemented multi-user security for Dataiku using Sentry and Active Directory, enhancing project isolation.
- Developed predictive models with Dataiku, Tableau, and Cloudera, achieving 65% accuracy.

Reliance Jio - Data Engineer *January 2018 - July 2019*

- Constructed data ingestion pipelines using Apache Kafka, HDFS, and Apache Ignite, ensuring data integrity.
- Led the development of company-wide CI/CD framework, reducing deployment time by 90%.
- Managed servers using Ansible, ELK Stack, and maintained Azure infrastructure.
- Demonstrated expertise in big data technologies including Kafka, Hadoop, and Spark.

EDUCATION

University of Utah, David Eccles School of Business August 2019 – July 2020
MS in Information Systems – Data Science | GPA: 3.9/4.0

TECHNICAL SKILLS

Programming Languages: Python, SQL, PySpark, Scala, Bash
SQL Databases: Oracle 12+, MySQL, MariaDB
Tools: Debezium, Databricks, Apache Pulsar
ETL Tools: DOMO, Tableau, Dataiku
Cloud Services: AWS, IBM Cloud, Azure
Big Data Technologies: Kafka, Zookeeper, HDFS, Spark
In-Memory Databases: Apache Ignite
Time-series Databases: InfluxDB, PostgreSQL, Prometheus
DevOps: ELK Stack, TICK Stack, CI/CD (GitLab Runner, Jenkins, Ansible), Grafana, SonarQube

PROFESSIONAL PROJECTS

Streaming Certificate Data

- Managed streaming certificate data from Certificate Transparency Logs using PySpark and SQL on Databricks.
- Handled data ingestion, deduplication, and storage in raw delta lake tables for ETL job consumption.

Metrics and Reporting Pipeline

- Architected and developed a data pipeline on Databricks using PySpark to extract segmented data from Adobe Analytics API.
- Provided the product and BI teams with real-time analytics for data-driven decision making, enhancing marketing strategies and customer engagement.

Complex ETL Pipeline with SCD2

- Developed a Slowly Changing Dimension (SCD2) pipeline using Python, YAML, and SQL.
- Implemented a star schema to manage relationships between dimensions and facts, using SHA256 for change detection.

Real-Time Scalable Data Pipelines

- Implemented real-time data pipelines using Kafka and CDC techniques with MySQL-based Data Warehouse.
- Enhanced data migration processes by 33%, supporting aggregated reporting and analysis.

Cassandra to Databricks Migration

- Migrated data from Cassandra DB to Databricks Delta tables using Apache Pulsar and Kafka.
- Ensured data consistency and real-time processing with CDC enabled in Cassandra.

AWS Projects

- Built real-time dashboards on Grafana integrated with AWS S3 and AWS RDS for machine tracking and alerting.
- Developed a cost analysis dashboard using AWS CloudWatch metrics and MySQL time series database on AWS RDS.