Sarah Chen

Data Engineer | san.francisco@email.com | (415) 555-0123 | linkedin.com/in/sarachen | github.com/sarachen

PROFESSIONAL SUMMARY

Results-driven Data Engineer with 8+ years of experience designing and implementing scalable data pipelines, warehousing solutions, and ETL processes. Expertise in building robust data infrastructure across cloud platforms (AWS, Azure, GCP) supporting analytics and ML applications. Proven track record of optimizing data systems that process petabytes of data daily.

TECHNICAL SKILLS

Languages: Python, SQL, Scala, Java, Bash
Data Processing: Apache Spark, Kafka, Airflow, Databricks, dbt, Flink
Cloud Platforms: AWS (Redshift, EMR, Glue, S3, Lambda), Azure (Synapse, Data Factory), GCP (BigQuery, Dataflow)
Databases: PostgreSQL, MySQL, MongoDB, Cassandra, Redis, Snowflake
Tools & Frameworks: Docker, Kubernetes, Terraform, Git, Jenkins, Tableau, Looker

PROFESSIONAL EXPERIENCE

Senior Data Engineer | TechCorp Inc. | San Francisco, CA | June 2020 – Present

- Architected and deployed real-time streaming data pipelines using Kafka and Spark Streaming, processing 5TB+ daily across 200+ microservices
- Reduced data warehouse query latency by 65% through optimization of Snowflake schemas and implementation of incremental materialized views
- Led migration of legacy ETL processes to modern ELT framework using dbt and Airflow, improving data freshness from 24hrs to 2hrs
- Built automated data quality monitoring framework using Great Expectations, reducing data incidents by 80%
- Mentored team of 4 junior engineers and established best practices for data pipeline development and documentation

Data Engineer | DataFlow Solutions | Seattle, WA | March 2018 – May 2020

- Designed and implemented AWS-based data lake architecture handling 2PB of structured and unstructured data using S3, Glue, and Athena
- Developed 50+ production ETL pipelines using Python and Apache Airflow, ensuring 99.9% SLA compliance
- Optimized dimensional data models in Redshift, reducing storage costs by 40% while improving query performance
- Collaborated with data scientists to build feature stores supporting 15+ ML models in production
- Implemented CI/CD practices for data pipelines using Jenkins and version-controlled SQL/Python code in Git

Data Engineer | Analytics Pro | Austin, TX | January 2017 – February 2018

- Built scalable ETL processes using Python and Spark to integrate data from 20+ external APIs and databases
- Developed data validation frameworks ensuring 99.5% data accuracy across financial reporting systems
- Created automated monitoring and alerting systems for pipeline failures, reducing MTTR by 50%
- Partnered with BI team to design star schema models supporting executive dashboards in Tableau

Junior Data Engineer | StartupXYZ | Boston, MA | July 2015 – December 2016

- Developed Python scripts to automate data extraction from REST APIs and load into PostgreSQL databases
- Assisted in building batch processing pipelines using Apache Spark for customer behavior analytics
- Created SQL queries and stored procedures to support reporting requirements for product and marketing teams
- Maintained data documentation and lineage using internal wiki and metadata management tools

EDUCATION

Master of Science in Computer Science | Stanford University | 2015
Bachelor of Science in Computer Engineering | University of California, Berkeley | 2013

CERTIFICATIONS

AWS Certified Data Analytics – Specialty | Google Cloud Professional Data Engineer | Databricks Certified Data Engineer