

Contact

www.linkedin.com/in/sairaghuadepu (LinkedIn)
sairaghu.hashnode.dev/ (Portfolio)
github.com/sairaghu538 (Personal)

Top Skills

Glue
Shell Scripting
Data Pipelines

Certifications

Data Analytics Essentials
Data Science Orientation
Google Data Analytics Capstone: Complete a Case Study
Accenture North America - Data Analytics and Visualization Job Simulation
AWS Certified Solutions Architect – Associate

Sairaghu Adepu

AWS Certified Data Engineer | Scala/Spark Developer | Big Data Engineer | Apache Spark | Cloudera Hadoop | Data Pipelines | United States

Summary

- ▶ I have experience in building scalable, efficient, and impactful data systems across diverse industries
- ▶ Proven track record in enhancing data pipeline performance and optimizing ETL processes with expertise in Python, SQL, and cloud technologies

CAREER HIGHLIGHTS

- ★ Led the optimization of a retail pipeline, improving efficiency by reducing data shuffling and enhancing Snowflake queries
- ▶ Enhanced data processing workflows for an e-commerce giant, reducing processing time by 30%
- ▶ Successfully migrated multiple data systems to AWS, improving system performance by 20%
- ▶ Spearheaded the end-to-end implementation of personalized recommendation systems using AWS and Snowflake, delivering impactful insights for over 100k daily users

AREAS OF EXPERTISE

- ▶ Cloud Data Engineering & Architecture
- ▶ Data Integration & ETL Optimization
- ▶ Data Security & Compliance
- ▶ Real-time Streaming & Data Pipelines
- ▶ Snowflake & AWS Cloud Technologies
- ▶ Data Performance Tuning & Query Optimization

TECHNICAL SKILLS

- ▶ Languages: Python, SQL, Scala
- ▶ Frameworks: Apache Spark, Apache NiFi, PySpark, AWS Glue, Hadoop CDP/CDH
- ▶ Data Warehousing: Snowflake, Hive

- Cloud Platforms: AWS, Azure
- Data Integration Tools: AWS Glue, Azure Data Factory
- Version Control & CI/CD: Git, Jenkins

CERTIFICATIONS

- AWS Certified Solutions Architect – Associate
- AWS Certified Cloud Practitioner

Experience

TD

Data Engineer

January 2025 - Present (8 months)

Philadelphia, Pennsylvania, United States

- Designed and implemented ETL pipelines using AWS Glue to process and transform banking data from S3 sources, ensuring seamless integration into target S3 buckets for downstream analytics and reporting.
- Leveraged Troposphere to programmatically generate CloudFormation templates (CFT), enabling automated and consistent infrastructure provisioning.
- Utilized AWS Athena for data validation on banking datasets stored in S3, ensuring data accuracy, consistency, and compliance with regulatory standards.
- Deployed Amazon SNS for computerized notifications, improving customer satisfaction by 20% and increasing loan application completion rates by 15%.
- Revised data validation workflows by implementing AWS Athena's query optimization strategies to enhance efficiency and manage costs effectively.
- Configured and monitored Amazon EMR clusters for large-scale banking data processing, utilizing Spark and Hadoop to streamline big data workloads.
- Automated IAM policy creation and deployment using Python scripts, reducing manual intervention and improving security compliance.
- Assisted in the migration of banking data from Db2 to AWS Redshift and S3, ensuring seamless data extraction, transformation, and validation for a successful transition.
- Completed AWS Event Bridge to develop event-driven architectures, automating anomaly detection and job failure alerts through AWS Lambda and SNS for real-time issue resolution.

- Built CI/CD pipelines using Jenkins & Git, reducing deployment time by 40% and improving release cycle efficiency.
- Migrated legacy data warehouses to Snowflake, reducing storage costs by 25% and enhancing query performance for fraud detection and regulatory compliance analytics.
- Advanced optimized SQL queries to generate regulatory reports for Basel III and CCAR compliance, reducing processing time from hours to minutes.
- Automated Spark application deployment using Jenkins and Git, implementing CI/CD to streamline job execution and version control.

Big Lots

Cloud Engineer | Data Engineer

January 2024 - December 2024 (1 year)

Columbus, Ohio, United States

- Developed data pipelines to integrate and transform large-scale sales and inventory data using Python, SQL, and AWS services, enhancing processing efficiency.
- Directed real-time analytics dashboards with Tableau and Power BI, reducing report generation time by 25% and delivering actionable insights into customer behavior and inventory levels.
- Streamlined data extraction from multiple sources using ETL processes and Apache Airflow, ensuring accuracy and consistency while reducing data discrepancies.
- Managed AWS cloud-based data infrastructure, ensuring uptime, scalability, and efficient handling of growing data volumes.
- Programmed data quality checks and validation using Python and SQL, increasing accuracy by 20%, reducing errors, and improving decision-making.
- Led predictive model development for inventory forecasting using Python machine learning, optimizing stock accuracy and supply chain decisions.
- Orchestrated batch and real-time data workflows with Apache NiFi, AWS Glue, and PySpark, reducing processing times and ensuring system stability.
- Appointed scalable big data solutions on AWS EMR using Hadoop, Spark, and Hive, improving data retrieval efficiency by 35%.
- Implemented real-time data streaming pipelines with Apache Kafka, processing millions of retail transactions daily and integrating POS systems with Snowflake for faster insights.
- Progressed Apache Airflow DAGs to automate ETL workflows, improving processing efficiency by 30% and reducing manual intervention in inventory tracking.

- Polished complex SQL queries for customer purchase data, cutting report generation time by 50% and enhancing sales analytics.
- Tuned Spark SQL queries using broadcast joins, predicate pushdown, and dynamic partition pruning, improving data retrieval efficiency.
- Established AWS CloudWatch monitoring solutions to track Spark job performance and optimize cluster resource utilization.

Anthem, Inc.

Big Data Engineer / Hadoop Developer

September 2018 - August 2022 (4 years)

India

- Planned, deployed, and optimized Apache Spark applications on Cloudera Hadoop (CDH/CDP), leveraging HDFS, Hive, Impala, and HBase for large-scale distributed data processing.
- Operated Spark-SQL, Pig, Python and Scala for data transformations, reducing processing time by 20% and streamlining reporting tasks.
- Automated claims processing workflows using Apache Airflow, cutting manual reconciliation efforts by 50% and improving insurance claims approval efficiency.
- Elevated optimized SQL queries for patient record retrieval, reducing latency by 30% and ensuring HIPAA data processing.
- Deployed Cloudera CDP-based data lake solutions for high-volume patient claims and EHR data, improving scalability by 40% while ensuring HIPAA compliance and secure data governance.
- Worked with Cloudera Hadoop ecosystem (CDH), including HDFS, Hive, Impala, and Kafka, for efficient data ingestion, storage, and querying.
- Progressed Spark SQL queries and PySpark transformations to enhance Hive and Impala performance.
- Built and managed data lakes using HDFS, ensuring efficient storage and retrieval of structured and unstructured data.
- Evolved ETL workflows using Apache NiFi and Oozie, ensuring seamless data integration.
- Developed Python and Scala-based ETL jobs by reducing shuffle operations and minimizing wide transformations, improving Spark job efficiency.
- Ensured data integrity, accuracy, and security using access control (IAM) and encryption techniques.
- Utilized Sqoop to transfer data between HDFS, Hive, and relational databases (PostgreSQL, Oracle).
- Automated CI/CD pipelines for data workflows using Git, and Jenkins, ensuring version control and seamless deployment.

- Established high-performance data transformation pipelines using Python, and Scala, leveraging functional programming paradigms.
- Integrated Apache Airflow with Spark workflows to orchestrate ETL pipelines, improving automation and reducing manual intervention.

Exelon

Data Analyst

June 2016 - August 2018 (2 years 3 months)

hyderabad India

- Maintained software applications to automate and streamline business operations, enhancing system effectiveness using Java, Python, and SQL.
- Furnished CI/CD pipelines using Jenkins and Git, reducing deployment time by 40% and improving release cycle speed and consistency.
- Enhanced data management workflows by designing and executing complex SQL queries, improving data processing efficiency.
- Incorporated data management software with internal applications, enabling real-time data synchronization and improving reporting accuracy by 15%.
- Operated Excel for advanced data analysis and reporting, providing actionable insights that informed key business decisions and boosted operational performance.
- Collaborated with teams to design and implement software solutions, ensuring seamless integration with existing systems and improving overall system performance by 20%.
- Provided data validation scripts using Pandas & NumPy, ensuring 99.9% data accuracy across multiple systems.

Education

Wilmington University

Master's degree, Computer and Information Sciences and Support Services · (August 2022 - December 2023)