Krushit Sakaria

+1 (929) 701-6733 | sakariakrushit29@gmail.com |

SUMMARY

- Accomplished Cloud Data Engineer with 7+ years of experience designing, implementing, and optimizing scalable cloud-based data solutions across AWS, GCP, and Azure, ensuring efficient data processing, storage, and analytics.
- Designed and automated ETL pipelines using AWS Glue, Apache Airflow, and Talend, enabling seamless data ingestion, transformation, and orchestration across structured and unstructured data sources.
- Expert in real-time data processing leveraging Apache Flink, AWS Kinesis, and Apache Kafka, facilitating high-velocity data streaming, low-latency insights, and event-driven architectures.
- Proficient in SQL for query optimization, indexing, and partitioning across multiple databases, including AWS Redshift, BigQuery, MySQL, Amazon RDS, and Snowflake, ensuring optimized performance.
- Implemented scalable data lake architectures on AWS S3 and hybrid cloud solutions using Redshift, GCP BigQuery, and Athena, improving data access speeds and analytics capabilities.
- Advanced Python skills for data manipulation and analysis, leveraging Pandas, NumPy, Matplotlib, and SciPy to process large datasets and extract actionable insights.
- Developed and optimized interactive dashboards using Tableau, Power BI, and Looker, enabling real-time business intelligence, KPI tracking, and performance monitorin.
- Integrated AI/ML models into data pipelines using TensorFlow, PyTorch, and Azure ML Studio, driving predictive analytics, automation, and data-driven decision-making.
- Experience in Generative AI (GenAI) integration, leveraging OpenAI API and Hugging Face Transformers to enhance NLP tasks, automate insights, and build AI-powered data solutions.
- Hands-on expertise in distributed data processing with Apache Spark and Hadoop, optimizing ETL execution times and improving data availability and scalability.
- Implemented DevOps best practices using Docker, Kubernetes, Terraform, and CI/CD pipelines (GitLab, Jenkins, Azure DevOps), ensuring automated deployments and scalable infrastructure.
- Strong knowledge of cloud security and compliance, including AWS IAM policies, role-based access control (RBAC), data encryption, GDPR, HIPAA, and SOC 2 compliance standards.
- Architected hybrid cloud solutions by integrating AWS and GCP services, achieving cost-effective, high-performance data storage and real-time querying solutions.
- Built and optimized streaming data pipelines with Apache Kafka, ensuring fault-tolerant, scalable, and highly available real-time analytics solutions.
- Active participant in Agile and DevOps environments, collaborating with cross-functional teams to align data engineering efforts with business goals, CI/CD strategies, and cloud modernization initiatives.
- Collaborated in Agile environments, aligning data solutions with business objectives while actively participating in cross-functional teams.

SKILLS

Scripting Languages	Python, SQL, Bash, PowerShell, Java
Databases	SQL, MySQL, SAP HANA, Amazon RDS, AWS DynamoDB, MongoDB
AWS Services	S3, RDS, Redshift, Glue, Lambda, EMR, Kinesis, EC2, DynamoDB, CloudFormation, IAM
Visualization Tools	Tableau, Power BI, Looker, Microsoft Excel
ETL Tools	Alteryx, Apache NiFi, Apache Airflow, AWS Glue, Azure Data Factory (ADF), Apache Spark, Talend
Packages	Pandas, NumPy, Matplotlib
AI/ML Tools	Azure ML Studio, TensorFlow, PyTorch, Keras
GenAI	OpenAI API, ChatGPT, Hugging Face Transformers
Big Data	Apache Hadoop, Apache Kafka, Apache Flink
Cloud Ecosystem & DevOps	AWS, Azure, GCP, Docker Containers, Kubernetes, Amazon EKS, CI/CD Pipeline, GitLab
Data Warehousing	Amazon Redshift, Google BigQuery, Snowflake
Version Control	Git, Git Hub, Git Lab
Monitoring & Logging	CloudWatch, Datadog, ELK Stack, Prometheus, Grafana
Other	SDLC, Agile Methodologies, Data Cleaning, Automation, Problem Solving, Critical Thinking, Root Cause Analysis, A/B Testing

PROFESSIONAL EXPERIENCE

Key Bank – New York City, NY **March 2022 – Present**
Data Engineer

- Architected and deployed comprehensive analytics platforms utilizing AWS DynamoDB, AWS RDS, and AWS Redshift, ensuring scalable data storage, real-time reporting, and optimized decision-making.
- Designed and implemented robust ETL pipelines using AWS Glue, Apache Spark, and Python to automate data ingestion, transformation, and loading, improving data accessibility and performance.
- Optimized SQL queries and indexing strategies in Redshift, Aurora, and DynamoDB, improving query execution times by up to 40% and enhancing overall database performance.
- Developed real-time data streaming solutions using Apache Kafka, AWS Kinesis, and AWS Lambda, enabling seamless data processing and automated workflows.
- Integrated machine learning models from Hugging Face and AWS SageMaker into data pipelines, enabling AI-driven insights, predictive analytics, and automated decision-making.
- Orchestrated ETL workflows with Apache Airflow, automating scheduling, dependency handling, and monitoring, reducing data pipeline failures by 70%.
- Implemented AWS S3-based data lakes integrated with Snowflake and Redshift, improving analytics performance and enabling efficient big data processing.
- Leveraged Terraform and Infrastructure as Code (IaC) to automate the provisioning of AWS resources, including Lambda functions, S3 buckets, and Redshift clusters.
- Migrated on-premise data systems to AWS, leveraging Redshift, Glue, and Airflow to modernize legacy infrastructure and improve data processing efficiency.
- Developed and optimized machine learning pipelines using AWS SageMaker and Apache Spark, automating model training, deployment, and monitoring.
- Integrated Apache Spark and Hadoop to enhance big data processing efficiency, enabling high-speed, in-memory transformations for large-scale datasets
- Developed APIs and Flask-based applications to ingest and process external data sources, integrating seamlessly with AWS S3 and other cloud services.
- Ensured data security and compliance by implementing IAM policies, encryption techniques, and role-based access control (RBAC) in alignment with GDPR and HIPAA regulations.

Candid Org – New York City, NY **Sep 2020 – Feb 2022**
Cloud Data Engineer

- Architected and developed robust data pipelines using Azure Data Factory and AWS Glue to automate ETL processes, integrating diverse data sources into centralized data lakes and warehouses.
- Designed and implemented a data lake architecture on AWS S3, enabling efficient storage, processing, and retrieval of structured and unstructured data for analytics and machine learning.
- Developed real-time analytics dashboards using AWS Kinesis, Power BI, and Tableau, aggregating data from multiple sources to provide stakeholders with actionable insights.
- Integrated machine learning models into data pipelines using TensorFlow, PyTorch, and Azure ML, enhancing predictive analytics capabilities and data-driven decision-making.
- Orchestrated and automated ETL workflows using Apache Airflow, reducing processing time and improving the reliability of data pipelines with automated retries.
- Optimized large-scale data processing using Apache Spark on Azure HDInsight and AWS EMR, improving transformation efficiency and reducing processing times.
- Developed and optimized SQL-based data validation frameworks using Python, Pandas, and SQL, ensuring the integrity and accuracy of ingested data across hybrid cloud environments.
- Implemented fine-grained security controls using AWS IAM roles and policies, ensuring secure and compliant data access across multiple cloud environments.
- Migrated legacy batch workflows to streaming solutions using Talend, Apache NiFi, and AWS Kinesis, improving data processing speed and enabling near real-time analytics.
- Built scalable data transformation pipelines with Pandas and NumPy, preparing structured and unstructured data for machine learning models and analytics workflows.
- Automated monitoring and alerting systems with AWS CloudWatch and Azure Monitor to ensure high availability, reliability, and continuous operation of data workflows.

NextEra Energy FPL- Miami, FL **Jun 2019 – Aug 2020**
Data Engineer

- Architected and deployed cloud-based data analytics platforms using AWS S3 and GCP BigQuery, enabling scalable storage and real-time analysis of structured and unstructured data.
- Designed and automated ETL pipelines with Apache Airflow and AWS Glue, streamlining data ingestion from MySQL, AWS RDS, and external APIs into centralized data lakes..
- Developed high-performance querying solutions using GCP BigQuery, optimizing hybrid cloud data processing and reducing data retrieval latency.
- Implemented serverless data transformation jobs using AWS Lambda, improving workflow efficiency while reducing operational costs and infrastructure maintenance.
- Utilized AWS DynamoDB for real-time data storage and retrieval, optimizing query performance for high-volume transactional data processing.
- Integrated GCP Dataflow for real-time stream processing, enabling continuous data ingestion workflows and reducing latency for event-driven analytics.
- Optimized ETL workflows and SQL queries, improving data integration efficiency across AWS Glue and Apache Airflow environments.
- Automated A/B testing data collection using Python and SQL, streamlining analysis and enhancing test reporting accuracy for data-driven decision-making.
- Orchestrated infrastructure provisioning with AWS CloudFormation, ensuring scalable and repeatable deployments of data processing pipelines.
- Designed and implemented real-time data streaming solutions using Amazon Kinesis and Apache Flink, enabling low-latency event processing and analytics.
- Developed optimized SQL queries and indexing strategies in GCP BigQuery and AWS Redshift, improving query execution times and accelerating report generation.
- Created interactive data visualization dashboards using Matplotlib and Tableau, delivering actionable insights into key business metrics.
- Enhanced monitoring and logging capabilities using AWS CloudWatch and GCP Stackdriver, proactively identifying performance bottlenecks and improving pipeline reliability.
- Collaborated with cross-functional teams to define key performance indicators (KPIs), aligning data strategies with business goals to optimize data-driven decision-making.

Ultracab – Gujrat, IND **Aug 2016 – Dec 2017**
Data Engineer

- Designed and implemented scalable data integration pipelines using Talend and Apache NiFi, automating ETL workflows for diverse data sources, including SAP HANA, MongoDB, and Amazon RDS.
- Deployed and maintained Apache Kafka clusters to enable high-throughput real-time data streaming and seamless message processing across distributed systems.
- Utilized Apache Flink for distributed stream processing, supporting complex event processing (CEP) and real-time data transformations for critical business applications.
- Automated ETL workflows with Apache Airflow, streamlining dependencies across AWS services (S3, RDS, DynamoDB) to enhance data availability and reliability.
- Developed and fine-tuned Snowflake as a data warehouse, improving storage efficiency, query performance, and real-time analytics for structured and semi-structured datasets.
- Integrated Looker with Snowflake to design interactive dashboards, allowing business users to generate ad-hoc reports without requiring SQL expertise.
- Conducted data quality validation using Python and SQL, ensuring accurate data ingestion and improving the reliability of business analytics and reporting.
- Implemented Talend's data masking features for GDPR and HIPAA compliance, anonymizing sensitive customer data while maintaining analytics integrity.
- Developed and optimized advanced SQL queries and stored procedures, streamlining data access, improving query performance, and enabling efficient reporting.
- Utilized Apache Flink's complex event processing capabilities to detect patterns and anomalies in streaming data, enabling proactive decision-making.
- Automated and optimized ETL processes with Talend, improving data pipeline efficiency and significantly reducing processing times for large-scale data operations.
- Collaborated with cross-functional teams to define key business metrics, aligning data engineering initiatives with strategic objectives to enhance decision-making.

EDUCATION

Masters in Electrical & Computer Engineering — **Jan 2018 - May 2019**

Texas A & M University, Texas, US

Bachelor in Computer Engineering — **May 2012 - May 2016**

Gujarat Technological University, Gujarat, India