Solution Architect   Data Engineer

Rajesh Babu Pasupuleti
Data Engineer | Cloud Data Engineer
Email: pasupuletirajeshbabu1@gmail.com
Phone: +1 937-886-7774
Linkedin: linkedin.com/prajeshbabu
Portfolio : portfolio.com/prajeshbabu.io

Professional Summary:

- **Data Engineer** with **4+ years** of experience architecting high-performance, cloud-native data pipelines that transform raw data into actionable insights, accelerating business decision-making and innovation.
- Engineered data ingestion, transformation, and storage workflows using **AWS services**, including **S3, Glue, Lambda, Redshift, Athena, DynamoDB, RDS, EMR, Step Functions, QuickSight, Kinesis, AWS DMS, CloudWatch, Lake Formation, IAM**, ensuring high-performance data management.
- Developed and automated scalable data processing solutions with **GCP services**, including **BigQuery, Dataflow, Cloud SQL, Cloud Run, Cloud Composer, Vertex AI, and Pub/Sub**, enabling efficient cloud-native ETL pipelines.
- Optimized big data processing using **Apache Spark (PySpark/Scala) and Databricks**, reducing query execution time and improving ETL performance.
- Orchestrated seamless workflow automation using **Apache Airflow and AWS Step Functions**, minimizing manual intervention and improving job scheduling efficiency.
- Designed and optimized data modeling, schema design, and query performance tuning for **Amazon Redshift, Snowflake, BigQuery, and PostgreSQL**, implementing **star/snowflake schema, indexing, and partitioning** to accelerate analytics.
- Built and managed data lakes with **AWS S3, Lake Formation, and Delta Lake**, ensuring optimized storage, governance, and **reducing data retrieval time by 35%**.
- Developed real-time data streaming pipelines using **Apache Kafka, Amazon Kinesis, and Flink**, enabling low-latency, event-driven architectures and **reducing event processing latency by 50%**.

- Engineered and optimized data migration and replication using **AWS DMS and Snowflake**, accelerating data movement across heterogeneous environments by **50%**, enabling incremental data ingestion, and **reducing latency for near real-time updates by 40%**.
- Automated infrastructure provisioning using **Terraform**, **AWS CloudFormation**, and **Helm Charts**, reducing deployment time by 60% while streamlining CI/CD pipelines with GitHub Actions, **Jenkins, AWS CodePipeline,** and **GitLab CI/CD,** improving deployment efficiency by 40%.
- Deployed and managed containerized workloads with **Docker, Kubernetes (EKS, GKE),** and **Amazon ECS**, enhancing system scalability and availability by 99.9% uptime while optimizing resource utilization by 30%.
- Optimized serverless data engineering workflows with **AWS Lambda, AWS Glue, AWS Batch, and Step Functions**, **reducing operational overhead and cutting compute costs by 30%**.
- Enforced data governance, role-based access control (RBAC), and security best practices with **AWS IAM, KMS, AWS Macie, Data Masking (PII compliance), and encryption strategies**, ensuring **data privacy and regulatory compliance**.
- **I**ntegrated and deployed machine learning and AI-driven analytics using **Amazon SageMaker, AWS Bedrock (LLM/RAG), Google Vertex AI, and SageMaker Feature Store**, delivering scalable AI-powered solutions.
- Implemented cost optimization strategies by **partitioning, indexing, caching, materialized views, and query performance tuning** across **Redshift, Athena, Snowflake, and BigQuery**, reducing compute costs by 25% while improving query speed.
- Designed and automated event-driven architectures using **AWS EventBridge, SNS, SQS, and Lambda**, improving workflow automation, enhancing data processing speed, and **reducing manual dependencies**.

Skill Matrix:

Skill Category	**Technologies & Tools**
Cloud Providers	AWS, Google Cloud Platform
Programming	SQL, Python, PySpark, Scala, Bash Scripting
Big Data Processing	Apache Spark, Databricks, HDFS, MapReduce, YARN, Flink, AWS EMR, Snowflake, BigQuery

Data Warehousing	Amazon Redshift, Snowflake, BigQuery
Databases	Amazon RDS, DynamoDB, MySQL,PostgreSQL, Cloud SQL, MongoDB, Cassandra, Firebase Realtime Database
Data Modeling	Star & Snowflake Schema, Fact & Dimension Tables, Indexing, Partitioning,
ETL & Data Pipelines	AWS Glue, GCP Dataflow, DBT, SQLAlchemy
Orchestration & Workflow Tools	Apache Airflow DAGs, AWS Step Functions
API Development & Integration	FastAPI, Flask, REST APIs, GraphQL, API Gateway (AWS, GCP), Postman, Swagger/OpenAPI
Machine Learning & AI Services	Amazon SageMaker, AWS Bedrock (LLM/RAG), Google Vertex AI.
Infrastructure as Code (IaC)	AWS CloudFormation, Terraform, GitHub Actions, Jenkins, AWS CodePipeline, GitLab CI/CD, Docker, Kubernetes, Helm Charts.
Compute Services	AWS EC2, AWS Fargate, AWS Batch, AWS Lambda, Google Cloud Functions, Google Cloud Run.
Security & Governance	AWS IAM, KMS, Role-Based Access Control (RBAC)
Data Visualization	AWS QuickSight, GCP Looker , Tableau, Power BI.

Experience Summary:

Client: Wells Fargo **May 2023 - Present**
Role: Cloud Data Engineer

Project Summary:

As a **Cloud Data Engineer**, I built and optimized **AWS-based data solutions** for **scalable data ingestion, transformation**, and analytics. Designed an **enterprise data lake architecture**, automated **CI/CD deployments**, and **developed ETL pipelines** to process **terabytes of financial data efficiently**. Focused on reducing processing **latency, improving query performance**, and cost optimization while ensuring secure and **scalable cloud data infrastructure**.

Responsibilities:

- Developed high-performance **ETL pipelines** using **AWS Glue & Lambda**, reducing **data transformation time by 50%** while enabling **automated schema evolution**.
- Optimized Glue job execution with **PySpark tuning & DPU scaling**, reducing **costs and improving performance by 40%**.
- Automated workflow orchestration using **AWS Step Functions & Apache Airflow**, eliminating manual intervention **in data processing workflows**.
- Implemented real-time **streaming pipelines** using **Amazon Kinesis & Flink**, reducing **data ingestion latency by 60%** for real-time financial transaction processing.
- Configured event-driven architectures with **Amazon EventBridge & Lambda**, automating **trigger-based workflows** and **reducing operational overhead**.
- Built and optimized a data warehouse in **Amazon Redshift**, implementing **distribution keys, sort keys, and query tuning**, reducing **query execution time by 60%**.
- Designed data marts using **Athena, AWS Glue, and QuickSight**, accelerating business **reporting by 70%**.
- Implemented columnar storage formats **(Parquet/ORC)** in **Amazon S3**, reducing **query latency by 60%**.
- Enforced data security & compliance using **AWS IAM, Lake Formation, and Amazon Macie**, ensuring **GDPR, SOC 2, and PCI DSS compliance**.
- Implemented **fine-grained access control** for **sensitive financial data**, reducing security risks and audit compliance gaps.
- Automated infrastructure deployment using **Terraform & AWS CloudFormation**, reducing **manual setup time by 70%**.
- **Built CI/CD pipelines** for data pipeline releases using **AWS CodePipeline & CodeBuild**, improving **deployment speed by 30%**.
- Reduced AWS compute costs by **25%** by **optimizing Glue job parallelism, leveraging Redshift concurrency scaling, and automating S3 lifecycle policies**.
- Configured logging & real-time monitoring with **AWS CloudWatch, AWS CloudTrail, and Datadog**, ensuring **99.9% pipeline reliability**.
- Developed automated **failure alerts & retries** using **SNS & Step Functions**, reducing **pipeline failures by 50%**.

Client: eClerx **April 2020 - July 2022**
Role: Data Engineer

Project Summary:

Designed and developed **scalable ETL pipelines** using **AWS Glue, Lambda, and Step Functions**, improving **data ingestion and transformation efficiency by 40%**. Built **real-time streaming data pipelines** using **Kinesis, Lambda, and DynamoDB Streams**, enabling **low-latency ingestion with a 50% reduction in processing latency**. Automated **schema evolution and metadata management** with **AWS Glue Data Catalog & Apache Hudi**, reducing **manual schema changes by 30%**. Enforced **data security and governance** using **AWS IAM, AWS Lake Formation, and KMS**, ensuring **100% compliance with security policies**.

Responsibilities:

- Built scalable ETL pipelines using **AWS Glue, Step Functions, and Lambda**, reducing **data processing time by 40%** and **improving automation**.
- Optimized batch data processing workflows with **Glue and PySpark**, increasing **throughput by 35%** while reducing compute costs by **25%**.
- Automated schema evolution in **AWS Glue & Snowflake**, reducing **manual schema interventions by 80%** and ensuring **seamless data updates**.
- **Developed real-time data ingestion pipelines** using **Amazon Kinesis, Lambda, and DynamoDB Streams**, reducing **data processing latency by 50%**.
- Implemented event-driven architectures with **Amazon EventBridge & Step Functions**, automating workflow execution and cutting **manual overhead by 60%**.
- **Optimized streaming data ingestion** with **Apache Flink on AWS Kinesis**, improving **real-time analytics performance**.
- Designed and optimized **a data warehouse** in **Amazon Redshift**, improving **query performance by 60%** through indexing, partitioning, and workload management.
- **Built analytical data marts** using **Redshift, Athena, and Glue**, enabling **faster business reporting and reducing query execution time by 35%**.
- Implemented Parquet & ORC storage formats in **AWS S3**, reducing **data retrieval latency by 40%** and optimizing query efficiency.
- Enhanced **security & data governance** using **AWS IAM, Lake Formation, and KMS**, ensuring **100% compliance with GDPR & SOC 2**.

- **Implemented data masking & encryption policies** in **AWS Glue & Redshift**, securing **sensitive customer data**.
- Configured automated access control policies, reducing **security compliance risks by 40%**.
- **Automated infrastructure deployment** using **Terraform & AWS CloudFormation**, reducing **manual setup time by 70%**.
- Developed **CI/CD pipelines** for **data pipelines** using **AWS CodePipeline & CodeBuild**, improving **deployment efficiency by 50%**.
- Integrated real-time monitoring & logging with **AWS CloudWatch & SNS**, reducing **incident response time by 40%**.
- Reduced AWS costs by 25% by optimizing Glue job execution**, automating S3 lifecycle policies, and leveraging Redshift concurrency scaling**.
- **Optimized resource utilization** by implementing **Auto Scaling for AWS Lambda & Glue**, reducing **compute waste by 20%**.
- Tuned **Redshift & Athena query performance**, reducing **data processing costs** and **improving system efficiency**.