Anwar Shaikh

anwarshaikh078 | anwarshaikh078 | mysite | anwarswork078@gmail.com | +1.404.948.8990

Summary

Solutions-focused Cloud Architect with 6+ years designing and implementing enterprise-scale data platforms across GCP and AWS. Proven track record leading technical consultations, delivering 5+ successful cloud migrations, and architecting solutions on GCP. Expert in translating complex business requirements into scalable, cost-optimized cloud architectures. Google Professional Data Engineer, ML Engineer, GEN AI Leader and Cloud Architect certified. Strong communicator with experience presenting to C-level executives and technical teams.

Work Experience

Dollar General June 2025 - Present
Senior Data AI Engineer Atlanta, USA

- Architected enterprise-scale cloud migration strategy for 400+ legacy Lawson integrations to Oracle ERP on GCP, defining microservices architecture, API gateway strategy (Tyk), and service mesh implementation (Dapr).
- Led technical discovery workshops with 2 business units, translating functional requirements into technical specifications for microservices, reducing integration complexity by **60%**.
- Built **multi-agent** discount search system using Google ADK with 3 parallel agents and **RAG**, deployed to Cloud Run, processing 1000+ QPS with 100-350ms response time. 30+ integrated PDF documents with automatic indexing and semantic search.

Onix Jan 2023 - May 2025
Senior Data Engineer Atlanta, USA

- Authored **Historical Data Migration Framework** using the KubernetesPodOperator on GKE cluster. Migrated around 4 PB of data using tpt export and native BQ load.
- Lead a team of 5 data engineers to successfully migrate 200 ETL jobs to Google Cloud from Teradata. This included analysis, design, development, testing, and deployment to production.
- Designed end-to-end ML solution predicting telecom customer churn using **XGBoost** with **85%+** accuracy, implementing **MLflow** for experiment management, FastAPI microservice deployed on GCP Cloud Run, and interactive Gradio interface, enabling technical demonstrations and architecture discussions with enterprise stakeholders on cloud-native ML deployment patterns
- Introduced Cloud Storage as a persistent volume mount for data processing tasks on **Google Kubernetes Engine (GKE)**, enabling efficient access to a **10** terabyte data lake and optimizing resource utilization.
- Enhanced Spark job performance by **20%** through efficient **memory management** and **optimization techniques**, resulting in reduced processing time by **50%** for large-scale data analytics tasks.
- Increased data ingestion throughput by **30%** by implementing efficient data partitioning and rollup strategies within Druid.
- Led a team of 3 members to develop a Jenkins pipeline for **CI/CD** processes, deploying code from GitLab to Data Platform.

Datametica Jul 2019 - Aug 2022
Data Engineer III Pune, India

- Guided the team to deploy Spark workloads on **Dataproc** serverless, with error reporting in persistent history server.
- Implemented **data partitioning** and **indexing strategies** in big query, yielding substantial performance and meeting SLA targets.
- Fine-tuned performance and optimized SQL queries, resulting in a remarkable **50%** reduction in query execution time.
- Crafted and nurtured data warehouses on **Google Cloud Platform**, managing **10 tb** of data. BI reports on Apache druid.
- Engineered **streaming data pipeline** utilizing Google PubSub and Cloud Functions. Processed **0.3** million messages and loaded to the streaming tables. Persisted the data in Google Cloud Storage for 7 days.
- Developed and deployed data pipelines on **Dataflow**, processing over TBs of data daily and reducing processing time by **60%**.
- Designed and conserved a **data lake** architecture on Google Cloud Platform (GCP), allowing for seamless data access and analysis. This architecture streamlined **data governance** and accelerated insights.
- Leveraged Reservation API to increase BigQuery slots to address the 50-70 min delay in **SLA** time, further reducing by **45 min.**
- Programmed an Audit Framework utilizing **CloudSQL** to facilitate a CDC process, enabling tracking of updates across 300 tables.
- Leveraged **Bigtable** and **Cloud Spanner** for storing frequently accessed metadata and transactional data associated with data assets, enabling efficient retrieval for data governance tasks and ensuring data consistency.
- Launched and maintained real-time data pipelines using **Cloud Pub/Sub and Cloud Spanner**, achieving sub-second latency for ingesting and retrieving 100k transactions daily, facilitating real-time fraud detection.
- Authored an airflow DAG's generation framework in Python, helped create **700+** dags, and saved manual work for **10** people.

Education

2023 Master's in Information Systems(Big Data Analytics) at **Georgia State University**
2019 Bachelor's in Computer Science at **MGM's Jawarhal Nehru College of Engineering**

Skills

Programming Skills:	Python, SQL, Java, Unix, Shell, R, Scala, Pyspark, Apache Beam, PL/SQL.
GCP Services:	GCS, Cloud Functions, PubSub, Composer, Dataflow, Dataproc, Bigquery, CloudSQL, Spanner, Vertex AI Platform, AutoML, Google Kubernetes Engine, IAM, Generative AI, Looker.
AI Skills:	LLMs, Prompt Engineering, RAG, Vector Databases, Embeddings, Fine-tuning, Model Optimization, LangChain, Vertex AI Generative AI, Google Cloud AI APIs, NLP, Model Evaluation, MLOps, Model Monitoring, Feature Engineering, Data Preprocessing.
Data Engineering:	Ingestion, Transformation, Extraction, Data Modeling, Data Architecture, Data warehouse, Change Data Capture, Slowly Changing Dimensions, Normalization, Constraints, ETL, Optimization.
Big Data Technologies:	Hadoop, HDFS, Apache Spark, Hive, Kafka, MapReduce, Sqoop, Druid, Flink.
Machine Learning:	Supervised and Unsupervised Learning, Predictive Modelling, Clustering Techniques, Neural Networks, Regression, Classification Models, Statistical Analysis.
Other tools:	Jenkins, Terraform, Docker, Tidal, Agile,Jira, Scrum, Pycharm, R Studio, Kubernetes, VPC network, IAM, Compute Instance, Persistent Disk, AWS, Github, BitBucket, S3, Redshift, EC2.
Certifications:	• Google Cloud Professional Data Engineer • Google Cloud Professional ML Engineer • Google Cloud Professional Cloud Architect • Google Cloud Generative AI Leader • Build and Deploy a Generative AI Solution using a RAG framework