Data Engineer SeungHo Choi

seungho546@naver.com | GitHub | LinkedIn | Portfolio | Seoul, Korea

Summary

5-year experienced Data Engineer who dreams of and realizes smooth data flow and diverse utilization.
Continuously researches, tests, and implements solutions for building **stable data pipelines**, configuring **cost-efficient data platforms**, and providing **environments focused on data analysis**.
Aims to create business value through data and contribute to organizational alignment toward common goals through effective communication.

Experience

Neowiz | Data Engineer

July 2020 - Present

Responsible for building real-time (CDC) data pipelines and operating data warehouses, achieving the following key results:

- **Integrated 15+ diverse data sources and built 1B+ daily CDC ETL**: Unified fragmented data to establish analytical foundation.
- **Designed Redshift multi-cluster architecture**: Resolved performance bottlenecks and laid the foundation for data mesh structure.
- **Built multi-cloud (AWS ↔ GCP) real-time data pipeline**: Processed 40M+ daily records supporting real-time analytics and FDS.
- **Designed data lake architecture using Trino and Iceberg**: Enhanced data accessibility and distributed DW load.
- **Achieved 90% operational resource reduction through automation and monitoring**: Implemented stable platform operations using IaC, Grafana, Prefect.
- **Reduced fixed costs by 20% ($3,000+) through infrastructure optimization**: Maximized cost efficiency via Graviton migration, serverless architecture, and automated idle resource management.
- **Reduced data extraction requests by 40% with LLM-based Text-to-SQL system**: Realized data democratization and increased team focus on core tasks.

Technical Skills

- **Specialties**: Real-time(CDC) Data Pipeline, Multi-cloud Architecture, Cost Optimization, Data Governance
- **Cloud Platforms**: AWS, GCP
- **Data Engineering**: Prefect, Apache Kafka, Trino, Apache Iceberg
- **Data Warehouse**: Redshift, BigQuery, Snowflake
- **Databases**: MySQL, PostgreSQL, DynamoDB, ElasticSearch, Redis

- **Programming**: Python, SQL, Java
- **Infrastructure**: Terraform, Docker, ECS, Grafana, Serverless Framework
- **AI/ML**: LangChain, Langfuse, OpenAI GPT, SageMaker

Key Projects

1. AWS Multi-Cluster Architecture Implementation (Presented at GAMES ON AWS 2024)

- **Resolved Redshift single-cluster performance limitations with Serverless and Data Sharing-based multi-cluster architecture**. Achieved **50% query performance improvement** through workload distribution and Zero-ETL implementation, established data mesh foundation, and ensured stability by minimizing downtime to under 30 minutes during 180TB data encryption.

2. Multi-Cloud Real-time Data Pipeline (AWS ↔ GCP)

- **Built multi-cloud pipeline transferring 40M+ daily records from AWS (Aurora) to GCP (BigQuery)** using AWS DMS, SQS, and Lambda. Implemented near real-time processing with 1-2 minute average latency to support FDS and reduce ETL management resources.

3. Trino on ECS-based DataLake Platform

- **Built federated query platform for unified querying across diverse data sources** by directly deploying Trino on ECS environment. Enhanced analyst efficiency in exploring raw data, reduced DW load, and **established technical foundation for Lakehouse architecture**.

4. Streaming Data Collection Platform

- **Implemented standard pipeline for stable collection and integration of semi-structured/real-time data** using Amazon MSK as central hub with MSK Connect, DynamoDB Streams, and idempotent UPSERT logic. Established event-driven deep analysis environment and resolved data silos.

5. LLM-based Text-to-SQL System

- **Developed system converting natural language questions to SQL** using RAG, Few-shot prompting, and Langfuse (LLMOps). **Reduced repetitive data extraction requests by 40%** for data team and promoted data democratization culture.

6. Infrastructure Operations and Technical Innovation Support

- **Built IaC (Terraform, Serverless) and monitoring (Grafana) systems** to ensure operational stability, and **reduced fixed costs by 20%+ through Graviton migration and automated idle resource management**. Led **Snowflake PoC** and developed shared libraries contributing to technical innovation and productivity improvement.