데이터 엔지니어 최승호

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

소개

데이터가 잘 흐르고 다양하게 활용될 수 있도록 꿈꾸고 실현하는 **5년차 데이터 엔지니어 최승호**입니다.
어떻게 하면 **안정적인 데이터 파이프라인**을 구축할지, 어떻게 하면 **비용 효율적인 데이터 플랫폼**을 구성할지, 어떻게 하면 **데이터 분석에 집중할 수 있는 환경**을 제공할지 지속적으로 고민하고 테스트하며 도입합니다.
데이터를 통해 비즈니스 가치를 창출하고, 조직이 공통된 목표를 향해 나아갈 수 있도록 커뮤니케이션하며 기여하는 것을 목표로 합니다.

경력

Neowiz | Data Engineer

2020.07 - 현재

실시간(CDC) 데이터 파이프라인 구축 및 데이터 웨어하우스 운영을 담당하며 다음과 같은 핵심 성과를 달성했습니다.

- **15개 이상 다양한 데이터 소스 통합 및 일 10억 건+ CDC ETL 구축**: 파편화된 데이터를 통합하여 분석 기반을 마련했습니다.
- **Redshift 멀티클러스터 아키텍처 설계**: 성능 병목을 해결하고 데이터 매시 구조의 초석을 마련했습니다.
- **멀티클라우드(AWS ↔ GCP) 실시간 데이터 파이프라인 구축**: 일 4,000만 건 이상의 데이터를 처리하며 실시간 분석 및 FDS를 지원했습니다.
- **Trino와 Iceberg를 활용한 데이터 레이크 아키텍처 설계**: 데이터 접근성을 높이고 DW 부하를 분산시켰습니다.
- **자동화 및 모니터링 시스템으로 운영 리소스 90% 절감**: IaC, Grafana, Prefect 등을 활용하여 안정적인 플랫폼 운영을 달성했습니다.
- **인프라 비용 최적화로 고정비용 20% ($3,000+) 절감**: Graviton 전환, 서버리스 아키텍처 도입, 유휴 리소스 자동 관리를 통해 비용 효율을 극대화했습니다.
- **LLM 기반 Text-to-SQL 시스템으로 데이터 추출 요청 40% 감소**: 데이터 민주화를 실현하고 팀의 핵심 업무 집중도를 높였습니다.

보유 기술

- **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

주요 프로젝트

1. AWS 멀티 클러스터 아키텍처 도입 (GAMES ON AWS 2024 발표)

- Redshift 단일 클러스터의 성능 한계를 **Serverless와 Data Sharing 기반의 멀티 클러스터 아키텍처로 해결**했습니다. 워크로드 분산 및 Zero-ETL 도입으로 **쿼리 성능을 50% 향상**하고 데이터 매시의 기반을 마련했으며, 180TB 데이터 암호화 시 다운타임을 30분 내로 최소화하며 안정성을 확보했습니다.

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

- AWS DMS, SQS, Lambda를 활용해 **AWS(Aurora)에서 GCP(BigQuery)로 일 4,000만 건의 데이터를 이전하는 멀티클라우드 파이프라인을 구축**했습니다. 평균 1-2분 지연 시간의 준실시간 처리를 구현하여 FDS를 지원하고 ETL 관리 리소스를 절감했습니다.

3. Trino on ECS 기반 DataLake 플랫폼

- ECS 환경에 Trino를 직접 배포하여 **다양한 데이터 소스를 통합 쿼리할 수 있는 Federated Query 플랫폼을 구축**했습니다. 이를 통해 분석가의 원본 데이터 탐색 효율을 높이고 DW 부하를 줄여, **Lakehouse 아키텍처의 기술적 토대**를 마련했습니다.

4. 스트리밍 데이터 수집 플랫폼 구축

- Amazon MSK를 중앙 허브로 구축하고 MSK Connect, DynamoDB Streams, 멱등성 UPSERT 로직을 활용하여 **반정형/실시간 데이터를 안정적으로 수집 및 연동하는 표준 파이프라인을 구현**했습니다. 이를 통해 이벤트 기반의 심층 분석 환경을 마련하고 데이터 사일로를 해소했습니다.

5. LLM 기반 Text-to-SQL 시스템

- RAG, Few-shot 프롬프팅, Langfuse(LLMOps)를 적용하여 **자연어 질문을 SQL로 변환하는 시스템을 개발**했습니다. 데이터 팀의 **반복적인 데이터 추출 요청을 40% 감소**시키고, 데이터 민주화 문화를 확산시켰습니다.

6. 인프라 운영 및 기술 혁신 지원

- **IaC(Terraform, Serverless) 및 모니터링(Grafana) 시스템**을 구축하여 운영 안정성을 확보하고, **Graviton 전환 및 유휴 리소스 자동 관리** 등으로 고정 비용을 20% 이상 절감했습니다. 또한 **Snowflake PoC를 주도**하고 공용 라이브러리를 개발하여 기술 혁신과 생산성 향상에 기여했습니다.