Seunggab Kim · Data Engineering Team Lead
seunggabi@gmail.com · LinkedIn · GitHub · Blog · 한국어 · PDF

Data engineer with 10 years in software and 8+ years in data engineering. I build and run data platforms that go from CDC and streaming to a lakehouse and BI. When the root cause is in open source, I fix it upstream; my changes are merged into Apache Spark, Superset and Airflow.

Experience

Orange Square (WOWPASS) · Data Engineering Team Lead · Oct 2023 – Present - Designed and run a CDC pipeline that streams changes from 197 production MySQL tables into an Iceberg lakehouse on S3 Tables on a 3-minute commit cycle (6 partitions and replication factor 3 per topic, Parquet with zstd, 128 MB target files), using Debezium and MSK Connect, moving analytics queries off the production database - Built Kafka and Flink streaming and an app event log pipeline on Avro and Schema Registry (about 1.7 million events a day; producers with acks=all, idempotence and lz4) - Run 1,300+ Airflow DAGs (18 standard DAG factories, about 6,000 SQL jobs) and EMR Spark batch; manage 6 MSK and EMR alarm types as Terraform modules - Introduced a DataHub catalog, built an internal KPI dashboard, and help operate Superset on Athena (100+ dashboards, about 50 users) - Built server APIs for external payment and transit card integrations; automated the development workflow with 25 GitHub Actions workflows (review assignment, release promotion, dependency audit)

TMAP Mobility · Data Engineer · Oct 2022 – Oct 2023 - Proposed and adopted a monorepo with a shared core module that brought dozens of batch jobs together and cut new job setup from 3 days to a few hours; moved Airflow DAGs to repository-based versioning - Wrote Kotlin batch jobs: Hive to DynamoDB, JSON to Hive, query results to S3, S3 version recovery, SQL alerts to Slack

Bucketplace (Ohouse) · Data Engineer · Apr 2022 – Oct 2022 - Split ad report aggregation into Spark jobs and Spring Batch; built the build and deploy setup for Kotlin and Spring Boot services - Moved OLAP data sync from API calls to asynchronous Kafka, with zero sync failures after the move; set up a monorepo and Spring Config Server; contributed to the global log schema design

NAVER · Dec 2016 – Apr 2022 - Data Engineering Tech Lead (Apr 2021 – Apr 2022): led the global log system and dashboard projects; NAVER TV personalization API - Data Engineer, incl. backend (Oct 2018 – Apr 2021): extended the music services' (VIBE, LINE MUSIC, NOW) log system, hundreds of millions of events a day, to KR, JP and NOW; Spark Streaming to HDFS; dozens of Spark and Hive metric batches; year-end recap data for millions of users - In the same period: HBase personalization API and Django dashboards; launched karaoke and pay-per-listen charts; personalized video recommendations for LINE MUSIC - Front End Developer (Dec 2016 – Sep 2018): Smart Editor components for Blog and Post (React, MobX)

Open Source

- **Apache Superset** (3 merged): `epoch_us` dataset time format, MCP API key scope limits, CI fix; proposed SIP-230 (download reason logging)
- **Apache Spark**: added `SparkSession.builder.config(Map)`, shipped in 3.4.0
- **Apache Airflow**: `TableauOperator` template fields; Helm chart release name length fix
- **PyAthena**: complex type (MAP, ARRAY, STRUCT) parsing fix · **Apache Iceberg**: Kafka Connect improvement in review

Skills

Area	Tools
Ingestion & streaming	Kafka, Kafka Connect, Debezium, Amazon MSK, Flink, Spark Streaming, Avro
Storage & query	Iceberg, S3 Tables, Hive, Spark SQL, Athena, Presto/Trino, MySQL, HBase, Redis
Batch & orchestration	Spark, Airflow, EMR, Spring Batch
Serving & governance	Superset, DataHub, Spring Boot, Django
Languages & infra	Kotlin, Java, Python, SQL · AWS, Terraform, Kubernetes (Helm), GitHub Actions

Education & Certifications

B.S. in Computer Engineering, Myongji University (GPA 4.18/4.5) · Engineer Information Processing (Korea), Oracle OCP 11g Administration, Linux Master Level 2