전대홍

Data Engineer · Open Source Contributor

daehong770@gmail.com | GitHub | Blog | LinkedIn

Introduce

데이터 플랫폼의 설계부터 개발, 성능 최적화까지 경험해 온 데이터 엔지니어 전대홍입니다.

금융권에서 Hadoop · Spark · Iceberg 기반 배치 데이터 플랫폼을 설계 · 개발하며 **일 평균 25억 건, 수백TB 규모**의 데이터를 처리했습니다. 핀테크에서는 Redshift 기반 DW를 Databricks 레이크하우스로 전환해 **약 1,600개 테이블 · 수십TB**의 데이터를 **적재 누락 0건**으로 이관하고, 준실시간 파이프라인을 구축했습니다.

문제가 생기면 겉으로 보이는 현상에서 멈추지 않고, 실행 계획과 소스 코드까지 따라가며 원인을 찾습니다. 측정으로 근본 원인을 확인하고, 검증된 방법으로 구조를 바꾸는 방식으로 일합니다.

Skills

Language	SQL, Python, Java, Scala
Data Platform	Databricks, Delta Lake, Spark, Iceberg, Hadoop, Hive, Impala
Orchestration	Airflow, Oozie
Cloud	AWS (S3, SQS, DMS, Redshift, Aurora, Glue)
Infra · Deploy	Kubernetes (EKS), Terraform, ArgoCD, Jenkins
Monitoring	Prometheus, Grafana, Loki

Work Experience

PFCT

AI 기반 대출 리스크 관리 핀테크 기업, 크플 앱 운영

데이터플랫폼팀
매니저
2026.01 - 현재

Redshift → Databricks 메달리온 레이크하우스 전환

- **수십TB · 1,600개 테이블** 이관 및 DMS CDC 기반 Bronze 적재 구축, E2E Latency **p99 30초 이내**
- DLT를 단일 Fan-out Job으로 재설계하고 Merge 병목 제거, **비용 39% 절감 · Merge 시간 73% 단축**
- DQ 검증 자동화로 누락 31만 행 · 순서 오류 15만 행 탐지 · 복구, **최종 누락 0건**
- 메달리온 계층별 S3 버킷 · Unity Catalog 구조와 권한 · 라이프사이클 정책 설계

데이터 분석 전용 시맨틱 레이어 구축

- 의미가 겹치는 테이블과 중복 컬럼을 정리해 기준 테이블이 하나인 분석용 Gold 테이블 설계 · 구축
- 기존 사내 온톨로지 대비 질의 **정답률 17% → 98%**, **일관성 30% → 98%**, **응답 시간 91% · LLM 토큰 87% 절감**

AI 기반 인프라 운영 체계 구축

- AWS 리소스 약 1,200개를 LLM 기반으로 Terraform IaC 전환, **plan 변경 0건**으로 무중단 이관
- **팀 공용 AI Workspace** 기반 IaC로 누가 작업해도 같은 결과 · 변경 이력이 남는 운영 체계 구축
- 망분리 환경의 Databricks까지 Workspace로 제어, 장애 분석은 옵저버빌리티부터 보도록 표준화

데이터 플랫폼 운영 · 업무 자동화

- EKS 기반 Airflow 운영, 신규 데이터 파이프라인 구축, DMS · Zero-ETL 태스크 관리
- Grafana · Prometheus · Loki 기반 데이터 플랫폼 모니터링 · 알림 운영
- 신규 데이터 적재, 모기지론 데이터 싱크 등 사내 데이터 업무 자동화

디딤365

금융기관 대상 빅데이터 플랫폼 구축 기업

데이터솔루션팀
선임(대리)
2024.03 - 2025.12

Spark · Iceberg 기반 배치 데이터 플랫폼 설계 · 개발

- 기존 Hadoop ETL 대비 **처리 시간 40% 단축**, S3 · Glue 등 HDFS에 종속되지 않는 구조 확보
- H/A · Failover로 **장애 자동 복구율 95%**, Snapshot · Orphan File 자동 정리로 **스토리지 30% 절감**
- DB Lock 기반 자원 제어로 **일 평균 800개 이상 배치 작업 · 25억 건** 데이터 안정 처리

개인신용정보 파기 시스템 성능 최적화

- 약 1,100개 파기 작업을 Spark 병렬 제출로 전환해 **CPU 코어 사용률 20~30% → 80%, 처리 시간 70% 이상 단축**
- 반복되던 4억 행 백업 테이블 Join과 수작업 집계를 선집계로 대체해 **전체 워크플로우 시간 80% 단축**
- Hive 통계 누락으로 잘못 선택된 Join을 Broadcast로 교정해 **처리량 20% 향상**
- 셔플 파티션 수만큼 쌓이던 Small File을 coalesce로 병합해 NameNode · 조회 부하 감소
- 배치 스케줄러의 힙 · TCP 누수를 Heap Dump로 찾아 제거해 **월 장애 시간 90% 감소**

디세이코리아

AI 솔루션 전문기업, 이커머스 플랫폼 구축 사업

개발팀
사원(연구원)
2022.04 - 2023.08

E-commerce B2C 플랫폼 설계 · 개발

- ngram Full Text Index · 쿼리 튜닝으로 상품 조회 **300ms → 90ms 미만** 단축
- DB Lock · 트랜잭션 재설계로 동시 발급 시 **쿠폰 중복 · 초과 발급 방지**

데이터 처리 성능 최적화

- **수십 GB 규모 대용량 파일을 Chunk 단위로 분할 · 병렬 처리**하고, Wafer LOT ID 기반 데이터 분류 구현

Open Source

오픈소스 한국 커뮤니티 / 멘토(운영진) 2025.05 - 현재

- 멘티 **500명 이상** 대상 오픈소스 기여 멘토링 진행
- Apache Iceberg, Paimon, Gravitino 등 **5개 프로젝트에 PR 20건 이상 merge**

Apache Iceberg / Contributor

- Spark 벡터화 읽기에서 행마다 하던 Position Delete Index 조회를 범위 단위 순회로 바꿔 검사 CPU 사용량 최대 **9.3배 감소** (#18027)
- Position Delete 로딩 실패 시 Loading 상태가 풀리지 않던 무한 대기 문제 수정, 실패 감지 · 재시도 로직 구현 (#3435)
- Delete Vector의 연속 삭제 구간으로 Row Selection을 만들고 불필요한 Row Group 탐색을 생략해 읽기 시간 최대 **48% 단축** (#3454)

Apache Paimon / Contributor

- 정렬 구간이 하나뿐일 때 불필요한 Sort Merge를 생략하고 중복 키 검증 경로를 추가해 읽기 시간 **102ms → 18ms** 단축 (#1164)
- 집계 과정에서 승자 레코드를 복사하지 않고 참조로 처리해 메모리 할당 **19.6%** · 읽기 시간 **13% 감소** (#1338)

Education

한국방송통신대학교 / 컴퓨터과학과 2024.03 - 2026.08

강남대학교 / 국어국문학과 2015.03 - 2019.02