데이터 엔지니어 포트폴리오

이랜드그룹 이랜드몰 데이터파이프라인 및 데이터 웨어하우스 구축 및 설계

YGPLUS 신규 음악사업 데이터플랫폼 구축 및 설계

작성자 : 유원상

Phone : 010-3367-8109

Email : hhttr13@naver.com

CONTENTS

자기소개

학력 (Education)

1) 2009.03 ~ 2012.02 : 상원고등학교 졸업
2) 2012.03 ~ 2016.02 : 경기대학교 컴퓨터과학과 졸업
3) 2025.03 ~ 현재 : 서강대학교 대학원 AI/SW학부 소프트웨어공학 재학

경력 (Career)

1) ㈜아이시스테크놀러지 - 솔루션사업부/팀원

- KDB산업은행 거액 익스포저 산출 프로그램 개발
- NH캐피탈 차세대 시스템 개발
- 하나은행 바젤3 FRTB 산출 프로그램 개발

2) 이랜드이노플 – 커머스플랫폼/데이터파트/파트장

- 이랜드몰 데이터웨어하우스 모델링
- 데이터 파이프라인 개발 및 유지보수
- 이랜드몰 정산데이터 정합성 검증 지원
- 실시간 데이터마트 설계

3) 와이지플러스 – IT기획운영팀/데이터엔지니어

- 와이지플러스 데이터플랫폼 개발
- AWS 클라우드 인프라 환경 설계 및 구축
- 온프레미스 – 클라우드 데이터 파이프라인 설계 및 구축
- 데이터웨어하우스 모델링
- 정산데이터 정합성 검증 지원

사용기술 (Skills)

1) Back-end

- Python : Apache Airflow Dag 및 모듈 개발
- Java : 금융권 재직당시 백엔드 비즈니스로직 개발

2) Database

- AWS Redshift : 데이터웨어하우스 설계 및 구축, 쿼리최적화
- Snowflake : 데이터웨어하우스 설계 및 구축, 쿼리최적화
- Postgresql : 화면조회용 DB 및 프로시저 로직 개발
- Oracle : 배치 및 화면조회 쿼리 개발

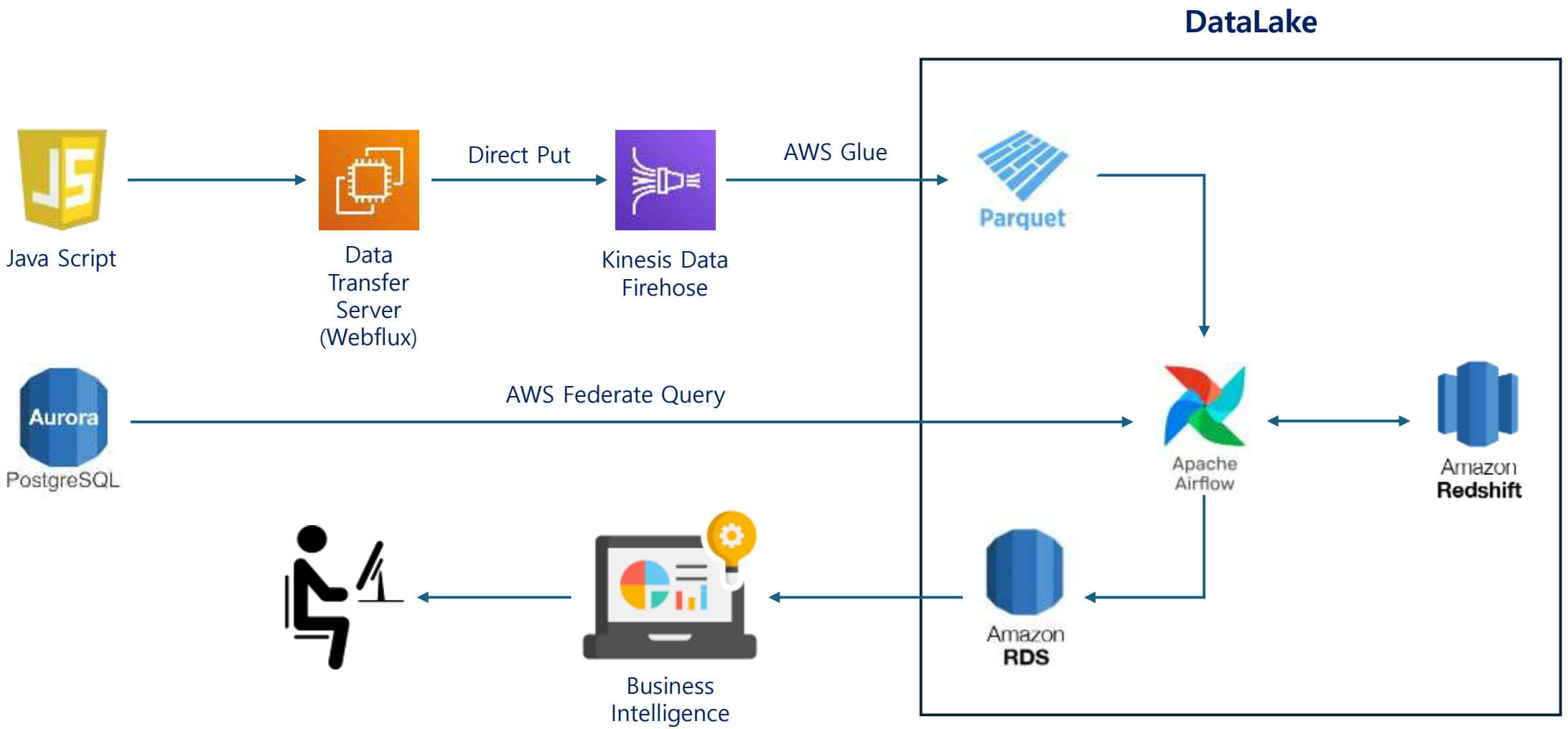
3) ETC

- Airflow : ETL 및 데이터웨어하우스 가공
- Kinesis, Kafka : 데이터의 실시간 수집 및 CDC
- Gitlab & Github : 팀원간의 협업툴로 사용, CI/CD 파이프라인 구축

자격증 (License)

1) 한국사능력검정시험 1급 (2018.08)
2) 정보처리기사 (2015.09)
3) SQL 개발자 (2023.07)

이랜드몰 데이터 파이프라인 구축

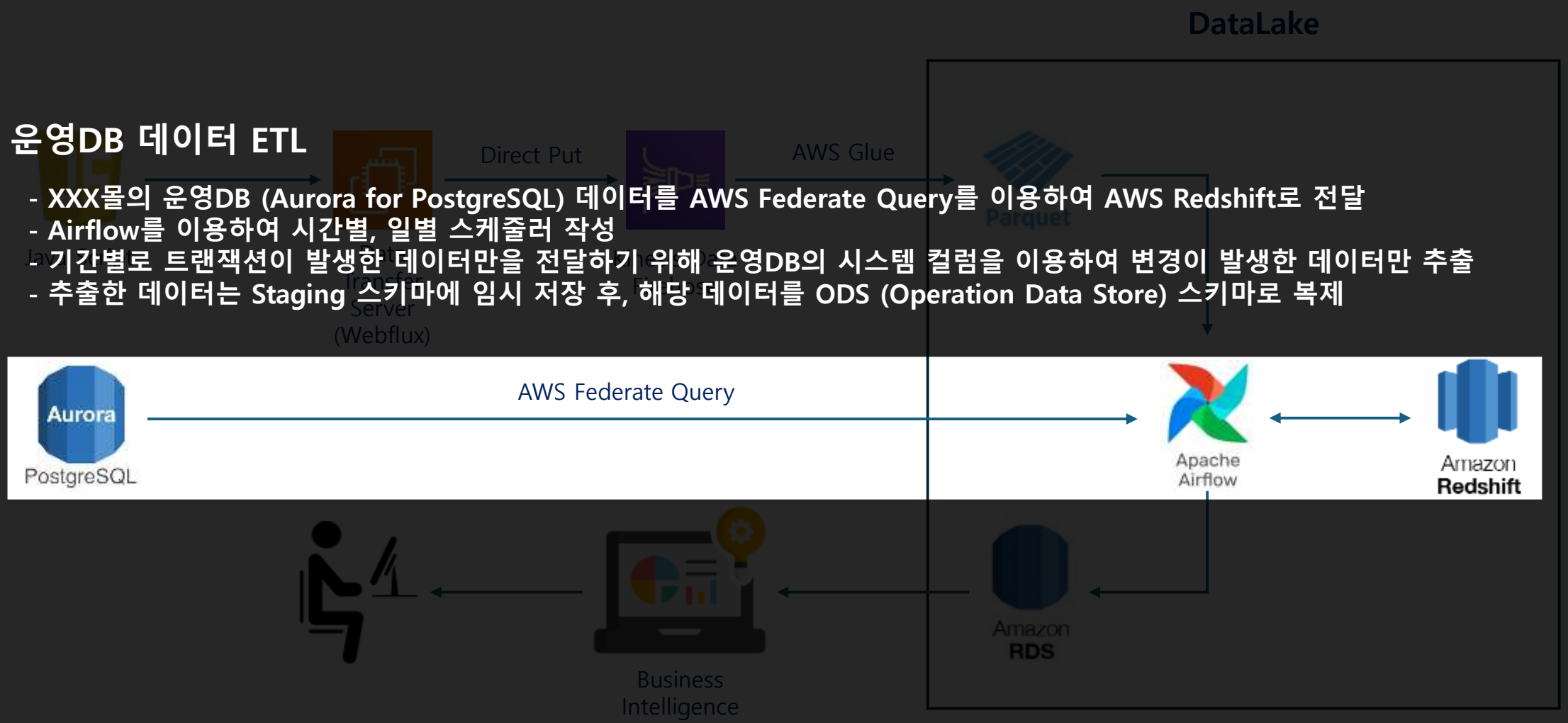


이랜드몰 데이터 파이프라인 구축

고객 방문데이터 수집 프로세스

- XXX몰의 스크립트에 수집서버로 데이터를 전달하는 스크립트 추가
- 페이지가 Request 되는 경우, 특정영역을 클릭하는 경우 해당 스크립트를 호출
- 수집한 데이터의 경우 Spring Webflux를 활용하여 스레드를 비동기식으로 처리
- 처리한 데이터는 Log4j를 통해 수집서버의 특정 디렉토리의 로그파일에 실시간으로 적재
- Firehose로 해당 로그파일에 연결하여 1분단위로 S3의 동적 파티셔닝 된 디렉토리에 적재 (YYYY/MM/DD/HH24로 구성)
- Parquet 형식으로 적재된 데이터를 Airflow를 이용하여 AWS Redshift에 적재

이랜드몰 데이터 파이프라인 구축



이랜드몰 데이터 파이프라인 구축

데이터웨어하우스내에서 가공 및 데이터 제공

- 수집한 방문 및 주문데이터를 이용하여 차원, 팩트, 마트 테이블들을 생성
- 집계가 완료된 데이터마트를 S3를 통해 RDS로 전달
- BI 화면에서 RDS에 적재된 테이블을 조회
- 타 부서 혹은 타 팀에서 데이터 요청 시 S3, GCS 등으로 데이터 전달
- BI 외에도 별도로 Redash 서버를 운영하여 자유롭게 데이터를 조회할 수 있도록 제공

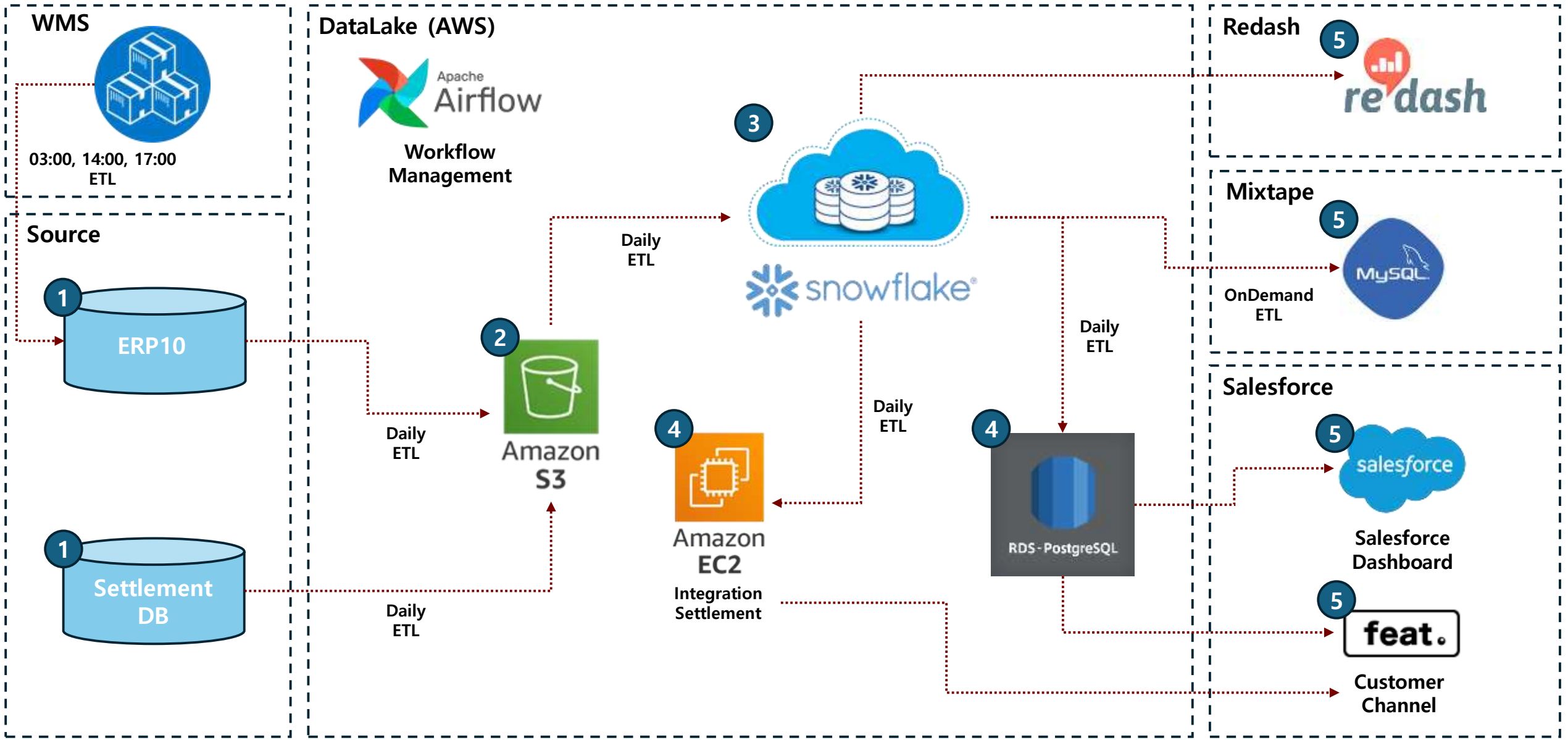
이랜드몰 데이터 웨어하우스 구축

- **팩트테이블**
  - **주문팩트**
    - 주문마스터 >> 주문상품 >> 주문상세 Hierarchy 구조로 반정규화된 팩트 생성
    - 해당 Hierarchy 구조에 맞춰 주문혜택, 포인트발생, 배송비 등의 데이터를 집계하여 팩트에 적재
    - 배송비의 경우 해당 Hierarchy 구조에 맞춰 적재 시 중복집계가 발생하는 경우를 대비하여 금액대비 상품별 가중치를 두어 배송비를 쪼개서 적재
    - 최종적인 팩트테이블을 생성하기 이전의 가공을 위해 사용하는 전처리성 팩트들 존재 (상품, 상세, 포인트, 배송비, 혜택, 결제 등의 전처리팩트 존재)
  - **방문팩트**
    - 페이지 Request 발생을 적재하는 Tracking팩트, 특정 영역 클릭의 발생을 적재하는 Reacting팩트 생성
    - Tracking팩트를 이용하여 PageView, UniqueVisitor, LoginVisitor 등의 데이터 집계 가능
    - Reacting팩트를 이용하여 클릭 수, 클릭율 등의 데이터 집계 가능
- **차원테이블**
  - **SCD-1 차원**
    - 이력 별로 관리하지 않아도 문제가 없거나, 가장 최신화된 값으로 데이터를 관리하고 싶은 경우 SCD-1로 차원테이블 생성
    - 해당 차원을 사용하여 만든 데이터마트의 재집계 혹은 Backfill 발생이 거의 없는 경우 사용 (시간별, 일별로 이력성으로 집계값이 적재되는 경우)
  - **SCD-2 차원**
    - 집계하고자하는 기간내에 원천 데이터의 변경이 여러 번 발생하는 경우
      - ex) 시간단위로 관리하는 SCD-2 차원의 경우 일별 집계하는 데이터의 속성값이 시간단위의 변경이 발생하는 경우 해당 이력을 관리할 수 있음
      - 주간, 월간 등으로 데이터의 재집계가 발생하는 테이블의 경우 재집계 시 최신화된 값을 바라보고 있으면 속성값의 이력관리가 되지 않음
    - StarSchema 구조를 위해 원천테이블이 SnowFlake 구조인 경우 반정규화 하여 관리
    - Depth가 존재하는 차원 혹은 반정규화가 이뤄진 차원의 경우 Depth별 혹은 각 차원의 변경 발생 시 상위 Depth까지 변경하도록 로직 구성

이랜드몰 데이터 웨어하우스 구축

- **데이터마트**
  - **Redshift >> BI**
    - 여러 기준 별 팩트값의 집계 데이터부터 방문팩트를 이용한 퍼널분석, 영역별 히트맵, 인기검색어 등의 데이터 제공
    - 주문팩트를 이용하여 MD별 상품매출 실적, 회원별 매출실적 등의 다양한 기준 별 집계 데이터 제공
    - 주문이 발생할 때, 해당 페이지의 속성을 장바구니에 적재되게 로직을 구성하여 주문데이터와 방문데이터가 서로 매핑될 수 있도록 설계
    - 이를 이용하여 방문데이터를 매출과 연결하여 데이터를 제공 ex) Conversion Rate 등..
  - **Redshift >> Redash**
    - BI 화면과 별도로 고객이 원하는 데이터셋이 존재하고, 이를 자주 사용 하는 경우 마트 테이블로 생성하여 Redash로 제공
    - 테이블이 아닌 쿼리로 데이터셋을 제공하길 원하는 경우, 쿼리로 데이터셋 제공
    - 간단한 시각화 데이터가 필요한 경우 사용
  - **Redshift >> S3**
    - 타 팀 혹은 타 부서에서 요청하는 데이터가 존재하는 경우 데이터마트를 생성하여 S3에 데이터 적재
    - GCS 등의 다른 플랫폼으로의 데이터셋의 전송이 필요한 경우 Airflow를 통해 S3에 데이터 적재 후 Storage Transfer 등을 이용하여 데이터 전달

와이지플러스 데이터 파이프라인 설계 및 구축



와이지플러스 데이터플랫폼 구축

- **AWS 클라우드 인프라 설계 및 구축**
  - **EC2**
    - 여러대의 EC2를 사용하여 Celery Executor 구조의 Apache Airflow 구성 (Redis Sentinel을 사용한 HA 구성)
    - 500 ~ 1000개의 레이블에 전달 할 정산서 파일을 생성하기 위한 서버로 사용 (최대 500만건 가량의 엑셀파일)
    - Redash 대시보드 용도의 EC2 서버 사용
    - EventBridge Scheduler를 통해 몇몇 서버들의 경우 새벽에만 동작하도록 구성
  - **AWS Lambda**
    - Github Actions를 통한 CI/CD Pipeline에 소스코드 배포 시 AWS Lambda 활용
    - Tracking팩트를 이용하여 PageView, UniqueVisitor, LoginVisitor 등의 데이터 집계 가능
  - **RDS for PostgreSQL**
    - 대시보드를 통해 외부 고객들에게 제공 할 데이터 서빙용 DB 구성
    - 초기에는 NoSQL(AWS DynamoDB)로 구성했으나, 화면에서의 집계 조건 등이 계속 확장되어 RDS로 변경
  - **S3 Storage**
    - ETL 데이터 및 배포용 소스코드, 비동기 로직 성공여부 체크 등을 보관하는 용도
    - 각종 데이터들을 보관하는 용도 (S3 Intelligence Tiering 기능 활용)

와이지플러스 데이터웨어하우스 구축

- **팩트테이블**
  - **음원팩트**
    - DSP사로부터 전달받는 음원 매출 데이터를 기반으로 팩트 테이블 생성
    - 음원, 앨범, 스트리밍서비스, 레이블 등을 기준으로 스트리밍수, 저작인접권료, 실연권료, 유통수수료, 정산금 등을 조회
  - **음반/MD(상품) 매출 팩트**
    - ERP에 적재되는 음반 및 MD(상품)의 판매 Raw데이터를 기반으로 팩트 테이블 생성
    - 음반과 상품, 사입과 위/수탁 품목을 구분하여 관리
    - 마스터 & 상세 Hierarchy 구조로 팩트 테이블 반정규화
  - **음반/MD(상품) 입/출고 팩트**
    - WMS로부터 전달받는 음반 및 MD(상품)의 입/출고 Raw데이터를 기반으로 팩트 테이블 생성
    - 입/출고 수불 데이터를 기반으로 별도의 품목 별 재고 테이블을 생성하여 관리
- **차원테이블**
  - **SCD-1 차원**
    - 이력 별로 관리하지 않아도 문제가 없거나, 가장 최신화된 값으로 데이터를 관리하고 싶은 경우 SCD-1로 차원테이블 생성
    - 음원, 앨범, 음반 등의 경우 기준이 되는 속성의 변경이 거의 발생하지 않고, 발생하더라도 해당 속성이 데이터마트의 차원으로 사용되는 경우가 거의 없어 SCD-1으로 관리
  - **SCD-2 차원**
    - 집계하고자하는 기간내에 원천 데이터의 변경이 여러 번 발생하는 경우
    - 계약의 수수료율이나, 각 레이블들의 지분율 등이 변경되는 경우가 간혹 있어 해당 정보들은 SCD-2로 관리
    - 특정 기간별 사입 품목들의 원가 대비 손익 등을 비교 분석하기 위해 판매단가 및 원가 등이 포함된 품목 정보는 SCD-2로 관리

와이지플러스 데이터웨어하우스 구축

- **데이터마트**
  - **Featuring(고객채널 대시보드)**
    - 외부 고객(레이블)들에게 정산금 및 매출 추이를 보여주기 위한 Featuring이라는 이름의 대시보드 서비스 오픈
    - 월별 정산금, 음원/앨범/음반/국가/스트리밍서비스 등 매출 Top 100, 계약 별 매출 및 인접권료 등의 다양한 데이터 제공
    - 최대 500만건, 시트 15개로 이뤄진 통합정산서(xlsx) 다운로드 기능 제공 (해당 파일의 경우 일배치로 새벽에 미리 생성)
    - 데이터 서빙용 DB로 RDS for PostgreSQL을 사용
  - **Salesforce(내부 대시보드)**
    - 내부 현업 및 경영진에게 보여주기 위한 용도로 Salesforce에 내부포탈 대시보드 서비스 오픈
    - 외부 고객들에게 제공하는 것 보다 더 심화 된 데이터 및 모든 레이블들에 대한 정보 제공
    - 주로 시각화 위주의 대시보드 형태로 구성
    - Snowflake에서 생성된 데이터셋을 API 및 배치를 통해 Salesforce로 전달하여, 해당 데이터를 조회
  - **Redash**
    - 기존의 대시보드 화면 외에 즉각적으로 현업이나 경영진이 원하는 시각화 자료가 필요한 경우 Redash로 제공
    - Snowflake와 Redash를 직접 연동하여, 다양한 차원 계층구조로 원하는 조건으로 실시간 집계를 하여 데이터 제공
    - 추후 Tableau나 PowerBI, Quicksite를 도입하기 전 가볍게 사용하기 위한 용도

Thank you