Chap12 데이터 시스템의 미래

앞으로의 데이터 시스템의 미래에 대해서 살펴 본다.

데이터 통합

- 데이터를 사용하는 모든 환경에서 단일 시스템으로는 여러가지의 요구사항을 충족하기는 어렵다.
- 그렇기 때문에 반드시 여러 소프트웨어를 함께 사용해야 한다.
- 그렇다면 이종 데이터 시스템을 통합하는 가장 적절한 방법은?
  - 파생 데이터
    - 결정적 재시도
    - 멱등성
    - 비동기 갱신으로 동시간 갱신 보장을 지원하지 못함
  - 분산 트랜잭션
    - 원자적 커밋으로 변경 효과가 정확히 한번 일어나도록 제어
    - 선형성
- 파생 데이터와 분산 트랜잭션 사이의 중간 지점 찾기

일괄 처리와 스트림 처리

- 입력된 데이터를 소비해 형태를 바꾸고 필터링하고 집계 후 모델을 학습하고 평가하고 마지막엔 출력
- 위의 목표를 달성하기 위한 입력 처리 도구
- 일괄 처리
  - 느리고 정확하다.
  - 일괄 처리시 연산에 참여한 과정중 하나가 실패하면 abort가 일어나고 실패가 확산된다.
  - 버그가 발생할 확률이 적다.
- 스트림 처리

    - 빠르지만 근사치 계산으로 데이터 정확성의 신뢰가 떨어진다.
    - 스트림 처리는 일부의 결함이 국소적으로 남아있긴 하지만 큰 문제로 확산되지 않는다.
    - 내결함성을 확보하기 어렵다.
        - 두번 처리되는것도 데이터 손상의 형태이다.
- 데이터 재처리
    - 새로운 뷰 - 일괄 처리
    - 부분 변경된 뷰 - 스트림 처리
- 람다 아키텍처
    - 이벤트 소싱과 유사하게 불변하는 이벤트를 생성해 증가하기만 하는 데이터셋에 추가하는 방식

데이터베이스 언번들링

- 데이터베이스, 하둡, 운영체제 모두 추상화 수준에서 보면 같은 기능을 수행한다.
- 데이터를 저장하고 처리하며 질의한다.
- 유닉스
    - 논리적, 저수준 하드웨어 추상화 제공
- 관계형 데이터 베이스
    - 디스크 상의 자료구조, 동시성, 장애 복구등 고수준 추상화 제공
- 연합 데이터베이스, 언번들링 데이터베이스
- 언번들링
    - 이종 저장소 시스템의 결합
        - 분산 트랜잭션
        - 이벤트 로그

데이터플로 주변 애플리케이션 설계

- 데이터베이스 인사이드 아웃

    - 애플리케이션 코드로 특화된 저장소와 처리 시스템을 조립하는 접근법
- 엑셀의 스프레드 시트
    - 데이터플로 프로그래밍 능력, 데이터 시스템 수준에서 필요한 기능이자 지향점
- 변환 함수: 데이터셋이 다른 데이터셋으로부터 파생될 때 거치는 함수, 모두 파생 데이터 셋이다.
    - 보조색인
    - 전문 검색 색인
    - 머신러닝 시스템
    - 캐시
- 스트림 연산자를 이용한 최근 애플리케이션 개발
    - 상태 비저장 애플리케이션, 상태 저장 데이터베이스
    - 애플리케이션 코드와 상태의 분리
- 쓰기 경로와 읽기 경로의 경계선을 옮기기 (애플리케이션 코드와 상태 관리 간의 관계 재조정)
    - 데이터를 수집하는 지점에서 데이터를 소비하는 지점까지의 모든 여정
    - 쓰기 경로 : 조급한 평가(eager), 미리 계산하기
    - 읽기 경로 : 느슨한 평가(lazy), 나중에 계산하기
- 읽기도 이벤트
    - 스트림 처리자도 그 자체로 단순한 데이터베이스이다.
- 오프라인 대응 가능 상태저장 애플리케이션
    - 로컬 데이터 베이스를 이용해 많은 일을 함
    - 네트워크 연결시 백그라운드에서 원격 서버와 동기화

정확성

- 스트림을 이용시 데이터의 정확성
    - 정확히 한번(분산트랜잭션)
    - 멱등성 보장

- 연산 식별자
  - 이벤트 id를 도입하여 중복 트랜잭션 방지
- 느슨하게 해석되는 제약 조건
  - 완화된 유일성 개념을 사용해 제약 조건 회피
  - 초과 예약 후 사과하기 (비행기 좌석)
  - 재고보다 많은 상품 판매하고 사과하기
  - 초과 인출 허용하고 초과 인출 수수료 부과하기
- 코디네이션 회피
  - 원자적 커밋, 선형성, 파티션에 걸친 동기 코디네이션 없이 무결성 보장
  - 비동기적으로 지역간 복제
  - 선형성 없음
- 비지니스 방식에 따라 트레이드 오프를 생각해서 결정
- 검증 하는 문화
  - 항상 원하는 만큼 이상적으로 동작하지 않기 때문에 오류는 무조건 발생한다.
  - 구더기 무서워서 장 못담그랴...
  - 정확성 보장이 절대적이라 가정하며 희박한 데이터 손상 가능성은 대비하지 않음

옳은일 하기

- 갑자기..?
- 데이터를 저장하는데 있어 윤리적 책임을 져야한다.
  - 이루다 사건
- 빅데이터를 통해 AI를 학습 시킨다.
  - 학습된 AI는 사람을 판단하는 도구로 쓰인다.
  - 사람에 대한 평가를 고정 관념화 한다.
- 사생활 추적
  - 나의 위치정보, 나의 좋아요 정보, 나의 연령정보가 감시 대상이 될 수 있다.