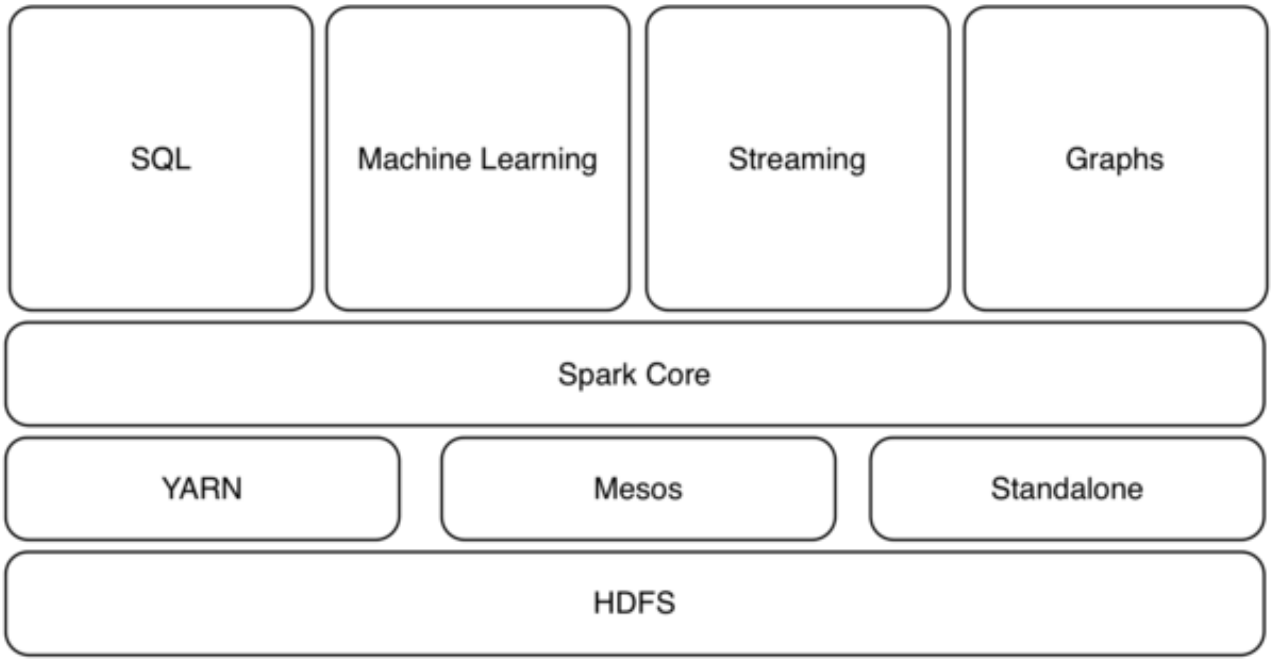
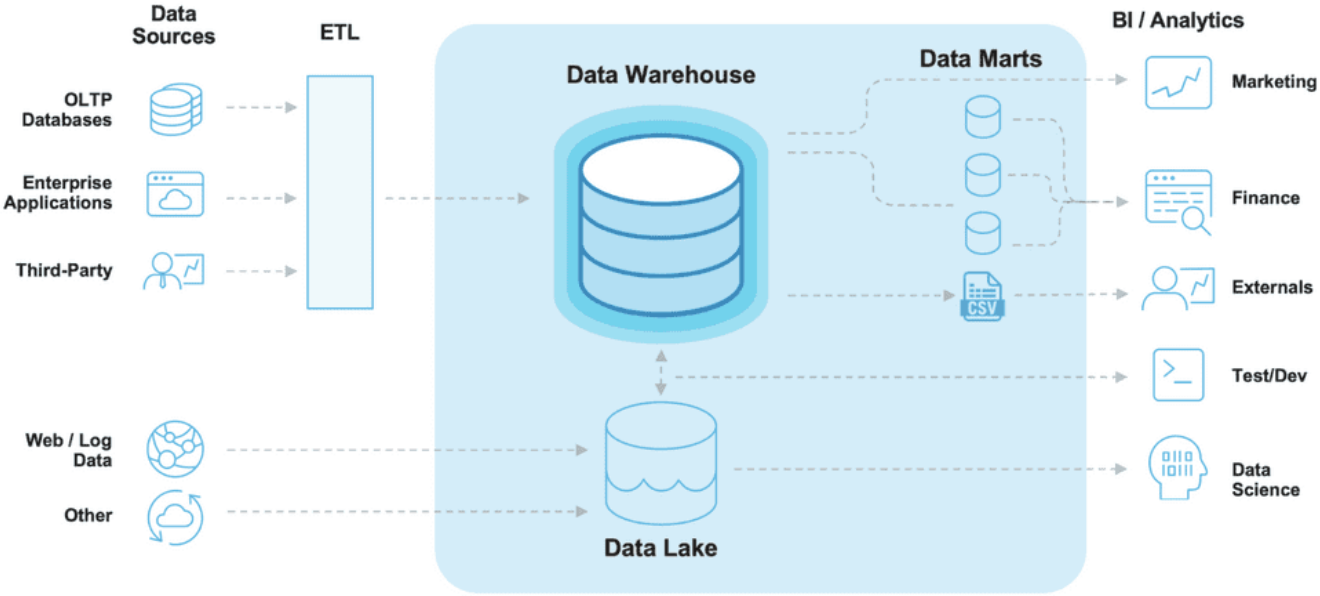
- 분산 시스템의 발전에 따라, hadoop을 이용하게 되었고, 그래서 "빅데이터"라는 키워드가 생기기 시작했다.
- 

- 

용어	설명
nosql	* 고속으로 빈번한 읽기, 쓰기 및 분산 처리가 강점 (SQL 과 비교) * 하둡(모여진 데이터를 나중에 집계가 목적) 인 Hadoop과 달리, NoSQL은 애플리케이션에서 온라인으로 접속하는 데이터베이스이다. * 전통 RDB의 제약을제거하는 것을 목표로하는 데이터베이스의 총칭이다.
데이터 웨어하우스	* 데이터 저장창고인데, 분석을 할 수 있는 물리적인 서버 * 사용자의 의사 결정에 도움을 주기 위하여, 기간시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스를 말한다. 줄여서 DWhttps://bomwo.cc/posts/Datawarehouse/ * 하지만 운영 시스템에 쿼리(SELECT,JOIN)를 기간으로 작성하게 된다면 시스템에 부하가 발생하게 되

	어서 앱이 동작하지 않게 되거나 데이터베이스를 분석하는 동안 사용하지 못하게 됩니다. 따라서 이런 부하를 발생시키지않고 운영 DB를 추출하여 큰 규모의 데이터 조회 및 분석을 위해 만들어 진 것이 데이터 웨어하우스입니다.
하둡	* 빅데이터 기술의 시작점입니다. * 하둡이 HDFS(Hadoop Distributed File System)를 통해 스토리지를 경유하기 때문에 상호작용이 많아져 느리다는 단점이 있다. * 다수의 컴퓨터에서 (마치 하나의 컴퓨터인 양) 대량의 데이터를 처리하기 위한 시스템이다. * 구글에서 개발된 분산 처리 프레임워크인 MapReduce를 참고해 제작됨: 초반 자바 -> 2009년 Hive 소프트웨어 출시: SQL 같은 쿼리 언어를 하둡에서 실행할 수 있다. => 프로그래밍 없이 데이터를 집계할 수 있게 되어 사용자 확대됨. * 모여진 데이터를 나중에 집계가 목적 * 스케일업, 아웃에 유리
스파크	* 스파크는 인메모리 처리를 기본으로 하기 때문에 좀더 빠르고 지연 속도가 낮은 분석 * 처리는 기본적으로 메모리에서 하기 때문에 기계학습이나 차트 계산처럼 반복 계산이 많은 작업에서는 하둡보다 빠르게 수행할 수 있는 것이다.
데이터 레이크	* 빅데이터 이전 축적 장소 * 구체적으로 임의의 데이터를 저장할 수 있는 분산 스토리지가 데이터 레이크로 이용된다. * 모든 데이터를 그대로 저장하고, 나중에 필요한 것만 꺼내 사용한다. > 웹 서버 등에 생성된 데이터는 처음에 RDB와 NoSQL 등의 텍스트 데이터로 저장된다. 이 후, 모든 데이터가 Hadoop으로 모이고, 거기서 대규모 데이터 처리가 실행된다.
데이터 마트	* 데이터 웨어하우스 환경에서 정의된 접근 계층으로 데이터웨어하우스에서 데이터를 꺼내 사용자에게 제공하는 역할을 한다 * 금융, 마케팅 또는 영업과 같은 특정 팀 또는 사업 단위의 요구를 충족시키는 데이터 웨어하우스이다 * 규모가 더 작고, 집중적이며 사용자 커뮤니티에 가장 잘 맞는 데이터 요약을 포함할 수 있다. * 데이터 마트는 데이터 웨어하우스의 일부일 수 있습니다.
애드혹 (AdHoc) 분석	자동화 등을 생각하지 않고 수작업으로 데이터를 집계할 수 있으면 충분하다. 이것을 일회성 데이터 분석이라는 의미로 애드혹(ad hoc analysis) 분석이라고 한다.
BI(Business Intelligence) 도구	* Business Intelligence의 약자로, 데이터를 통합/분석하여 기업 활동에 연관된 의사결정을 돕는 프로세스
oltp olap 차이	https://too612.tistory.com/511
맵리듀스	구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크
Hive	sql을 이용해 맵리듀스를 실행할 수 있게 해주는 기술입니다.
기간계 시스템과 정보계 시스템	**기간계 시스템**: 신중해야함. 비즈니스에 근간이 되는 중요한 시스템. 정지되면 업무가 멈춰서 완벽한 테스트 필요 **정보계 시스템**: 사내 커뮤니케이션, 의사결정을 위함. 보다 덜 엄격 두 개가 혼합 안되게 하자. 사내에만 필요한 정보를 기간계 시스템에 통합하면 업데이트가 어려워 분리한

	다.
KPI	**핵심성과*지표*.** KPI(key performance indicator) 모니터링 * **DAU**: Daily Active User : 서비스를 이용한 1일 유저 수 * **계속률** : 서비스를 계속해서 이용하고 있는 유저의 비율 * **ARPPU** : Average Revenue Per Paid User: 유료 고객 1인당 평균 매출 * **CTR**: Click Through Rate : 광고의 표시 횟수에 대한 클릭 비율 * **CPC:** Cost Per Click : 1회 클릭에 대해 지불한 광고 * **CPA**: Cost Per Acquisition: 1건의 고객 취득을 위해 지불한 광고비

- `OLTP` or `OLAP`
  -

구분	OLTP	OLAP
목적	비즈니스 활동 지원	비즈니스 활동에 대한 평가,분석
주 트랜잭션 형태	SELECT, INSERT, UPDATE, DELETE	SELECT
속도	수초 이내	수초 이상 수분 이내
데이터 표현 시간	실시간	과거
관리단위	테이블	분석된 정보
최적화 방법	트랜잭션 효율화, 무결성의 극대화	조회 속도, 정보의 가치, 편의성
데이터의 특성	트랜잭션 중심	정보 중심
예시	회원정보 수정	1년간의 주요 인기 트랜드
	상품주문	한달간의 항목별 수입, 지출
	댓글 남기기 및 수정	10년간 A회사의 직급별 임금 상승률

- 분산디비를 SQL 로 집계할 때는 두 가지 방법이 있다.
  Hive라는 쿼리 엔진 사용 (현재는 대화형 쿼리 엔진)
  데이터웨어하우스 제품 이용 : ETL(extract-transform-load) 프로세스

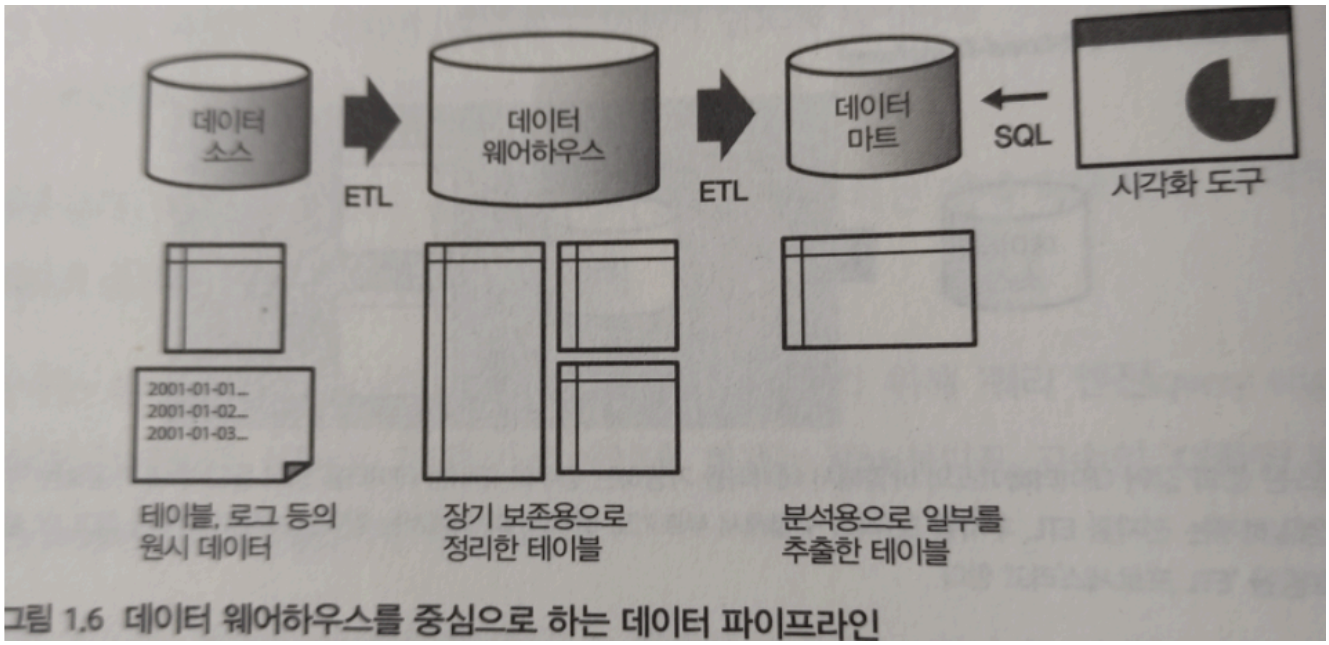
- 

그림 1.6 데이터 웨어하우스를 중심으로 하는 데이터 파이프라인

참고 문헌

- https://bomwo.cc/posts/Datawarehouse/
- https://github.com/Team-Neighborhood/I-want-to-study-Data-Science/wiki/데이터-엔지니어
- 빅데이터 - 하둡, 하이브로 시작하기: https://wikidocs.net/book/2203