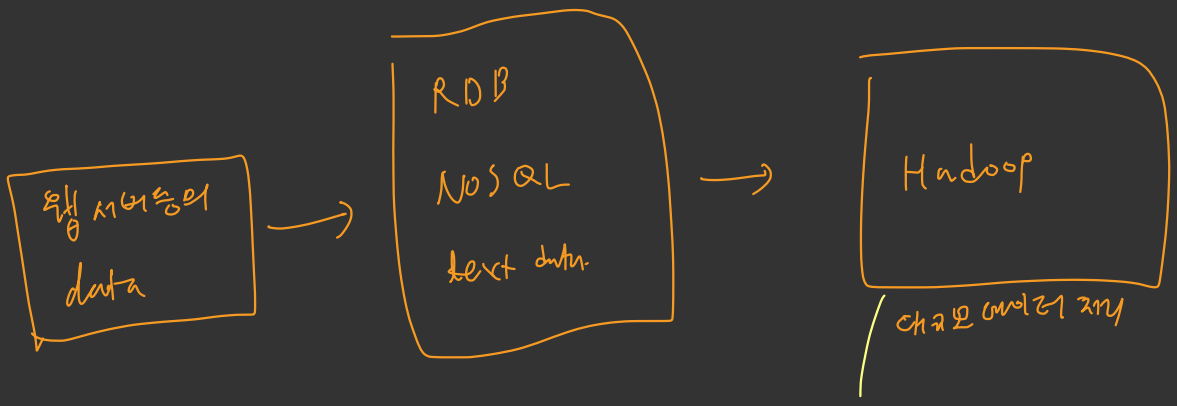
해커톤: Raw 데이터 가공



다수의 컴퓨터에서 대량의 데이터 처리 시스템.

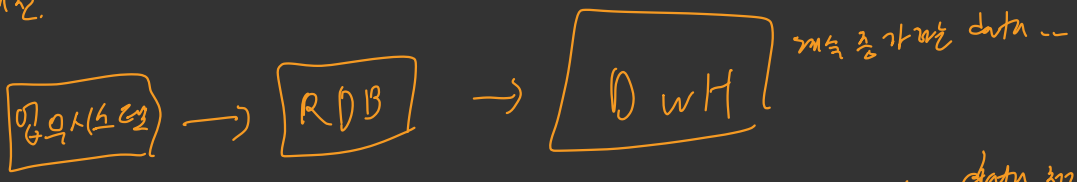
(ex) 전세계 웹페이지로부터 검색엔 모든시 방대한 data 정리할까 필요.

Hive - Hadoop에서 SQL같은 고차언어로 실행하기위한 software

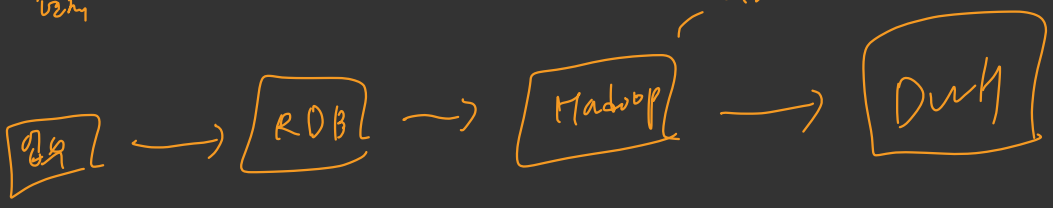
Hadoop - 오픈 데이터로 대중에 공개

DW 단점 - 소프트웨어 하드웨어가 결합..
하드웨어도 교체해야.

① 예전.



② 현재



Google Big Query
Amazon Redshift ⇒ 클라우드상에서 DWH구축가능.
DWH

∴ 데이터 마스터 - DWH 에 저장된 data로 대화형으로 분석.
(BI 도구.)

빅데이터의 등장은, 기존 DWH 만으로 다루기 힘든 시스템을 효율적
확장성이 뛰어난 데이터 처리로 만듦

★ 데이터 전송 방식.

에전 → 배크업 : 어딘가에 존재하는 데이터를 정해진 주기, DB등에서 가져올

실시간 스트리밍 방식 : 지속적으로 생성되는 데이터를 끊임없이 보내는 방법.

↳ 모바일 app등에서 보내는 이벤트 data 및, embedded 장비

↳ 시계열 (time series) 에 최적화되고 지금 특정일에 일어나는지 알 수 있다

나중을 고려 → 확장성 이 높은 것을 사용해야 한다

블록스토리지 → S3, NoSQL도 가능

객체스토리지

분산 Data 처리 (컴퓨팅, ETL 프로세스)

↑ MapReduce, Data 처리량↑, 많은 컴퓨터자원

나중에 분석할 목적 데이터가공해서 타부 DB저장

★ 방식

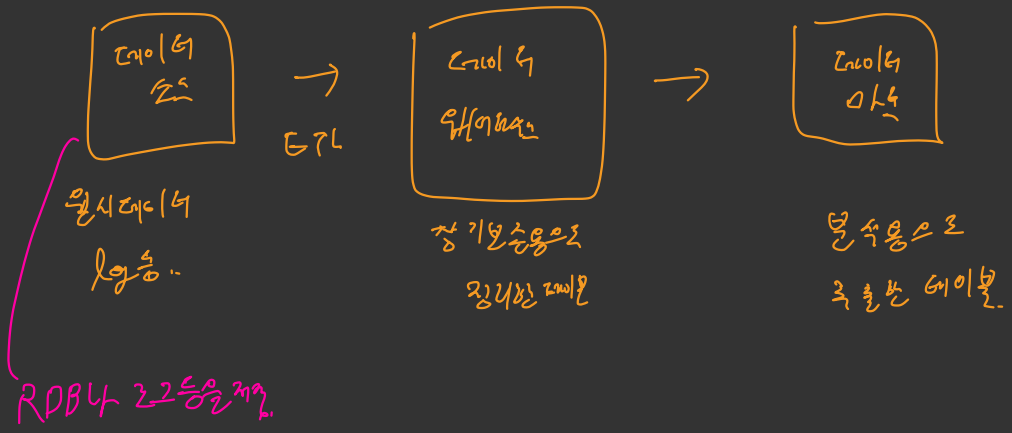
① SQL 사용

② ETL (대부분 데이터 웨어하우스 제품사용)

↑ 이 타부 DW에 많게 파일형식으로

데이터 웨어하우스는 "대량의 데이터를 장기 보관하는 것에 최적화되어있다.

└ 소량의 데이터를 자주쓰고 읽는데는 적합하지 않다



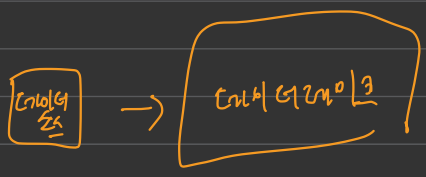
※데이터레이크

모든데이터가 데이터 웨어하우스를 거쳐서 만들어지지 않음.

ex) Binary data.

모든데이터를 그대로 저장하고

나중에 필요한 것들만 꺼내서 사용한다.



※데이터 레이크는 단순 스토리지이며, 그것만으로 데이터를 가공할 수 없다.

그래서 사용하는게 Map Reduce 이다.

└ 분석에 필요한 데이터를 가공, 정제하고, 이것을 데이터마트로

추출하는 것.

(분석용으로 일부를 추출한 Data)

데이터 소스 → Data lake → Data Mart.

Data Engineer.

수작업으로 데이터를 정제 ⇒ Ad hoc 분석
(일회성 데이터 분석)

Adhoc 분석은, Data Mart 만들려면 데이터베이스 와 웨어하우스에 직접 연결하는 경우가 많다.

☆ 데이터 파이프라인이 큰 흐름은 변하지 않는다!!!

☆① 처음부터 끝 일은 데이터 용량에 제한이 없음.

☆② 데이터를 효율적으로 추출할 수밖에 없음.

☆ DataFrame = 표형식의 데이터를 추상화한 객체.

스프링은 안에 안에서 가공과 정제를 할 수 있다.

JSON, 텍스트 파일들로 만들어야함

Q DataFrame은 쿼리를 실행한 결과로 얻을 수 있다.

Q 모니터링 : 커넥션으로 데이터의 연결을 유지해 나가는 것

BI 란 - 데이터의 움직임을 모니터링 하기위한 도구
└ 원인 파악위해서 데이터로 통계 계산을 할 수 있다.

① BI 도구에서 직접 데이터 소스에 접속!

② 데이터 마트를 통해 그것을 BI 도구로 옮겨오기

③ 웹 방식
BI 도구를 통해서 CSV 파일 업로드하기