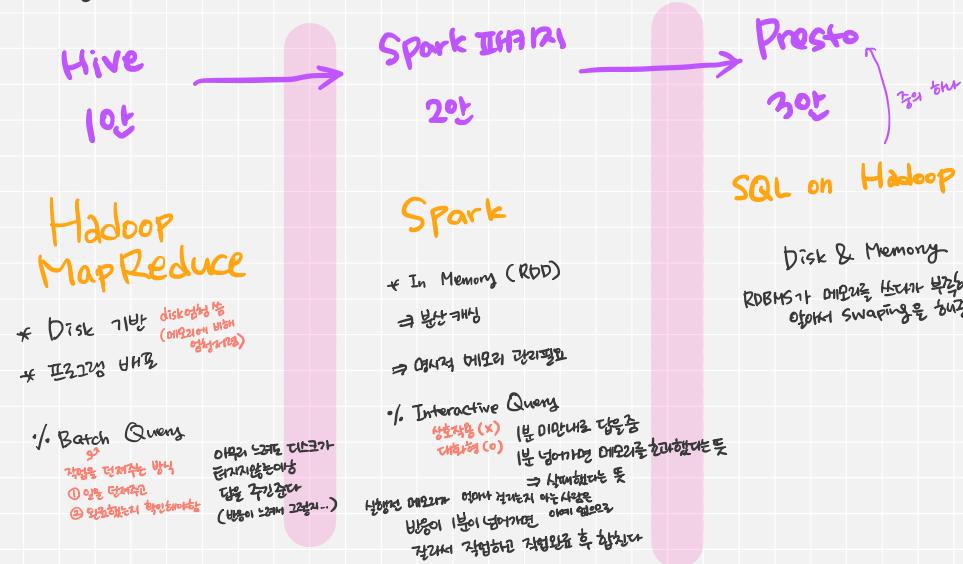


SQL on Hadoop

* MSSQL에서 Join을 쓰고 이것처럼 하면 당연히 느끼는 데에 압니다!
기본적으로 Client은, 서버는 Hadoop을 기본적으로 다 처리되어 있어서 약간의



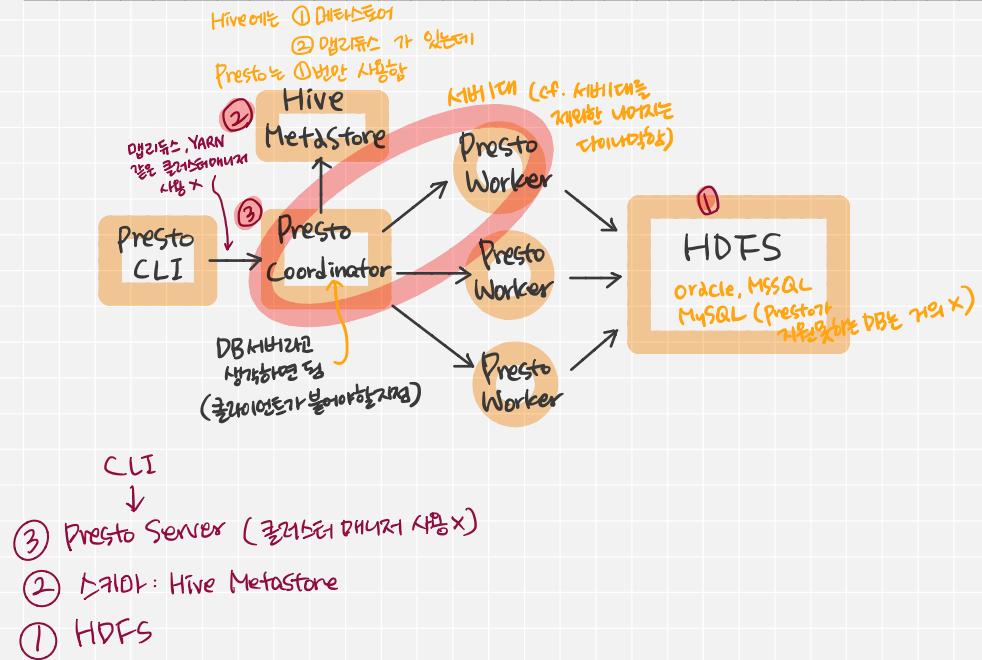
SQL on Hadoop 이란?

↳ HDFS에 저장된 데이터를 SQL / 유니SQL로 처리연산 / 분산처리

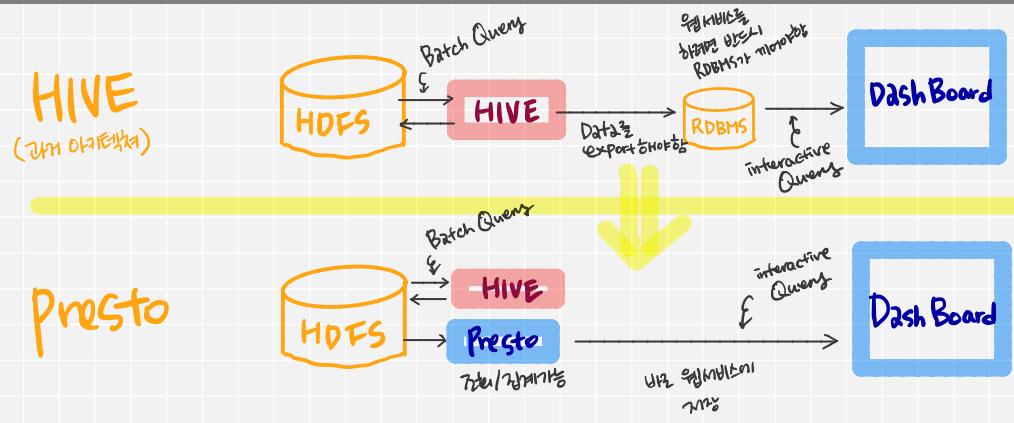
Software example

- Hive on Tez : Hive 2.0, Spark과 통합화된 양방향
• Drill : 구조화 데이터베이스... 데이터 분석
• Spark SQL : 스파크의 커널과 통합되었음. Spark의 쿼리구문입니다. 대체로 SQL과 유사
• Impala

Presto 마카테닉



HIVE vs Presto



[hadoop 시작]

- ① 네임노드가 있는 머신에 ssh (SSH namenode) HDFS 시작하기
`$ hadoop /sbin/start-dfs.sh`
- ② 클라이언트 머신에 Hive Metastore 실행하기
- ③ 네임노드가 있는 머신에 Presto Coordinator Server 실행하기
- ④ 워커노드에 Presto Worker 실행하고 시작하기
- ⑤ 클라이언트 머신에 Presto CLI 실행하기

Presto 시작

Prestodb.github.io 참고

Presto 개념

Presto의 개념

↳ 데이터가 있는 페더 (Federated)

Conn = execute (SQL)

row ← fetch (conn)

Service Query

fetch가 지원된다

Presto 커디네이터가

데이터를 가진다

나머지 데이터는

워커가 가진다

Interactive Query

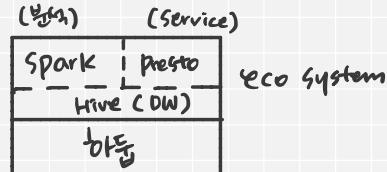
데이터를 가져온다

빅데이터 수집

Sqoop

Flume

Kafka



<빅데이터의 주요 수집 기술>

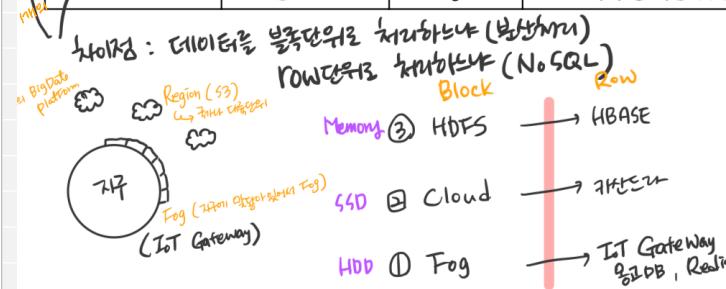
기술	개발	최초 공개	주요 기능 및 특징
Sqoop	아파치	2009년	RDBMS와 HDFS(NoSQL) 간의 데이터 연동
Flume	Cloudera	2010년	방대한 양의 이벤트 로그 수집
Kafka	LinkedIn	2010년	분산 시스템에서 메시지 전송 및 수집

- Hadoop과 DB 간 데이터 이동
- 방향 : Import & Export
- JDBC 드라이버를 통한 연동
- Sqoop Import : RDBMS → HDFS
- Sqoop Export : HDFS → RDBMS

빅데이터 저장

HDFS
Hive
NoSQL
Hadoop DW

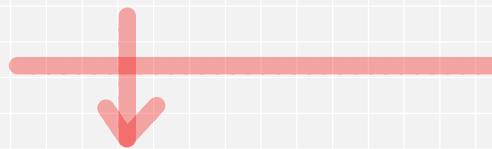
구분	기술	최초 개발	주요 기능 및 특징
분산파일시스템	HDFS	아파치	대표적인 오픈소스 분산파일시스템
	Hive	페이스북	HDFS기반의 DataWarehouse
	S3	아마존	아마존의 클라우드 기반 분산 스토리지 서비스
NoSQL	HBase	아파치	HDFS기반의 NoSQL
	Cassandra	A. Lakshman	ACID 속성을 유지한 분산 데이터베이스
	동고DB	10gen	DB의 수평 확장 및 범위 질의 지원, 자체 맵리듀스



불변형 \Rightarrow Block 타입 (o)
row 타입 (x)

Row 타입의
Update, Insert
delete 불가능

* Read Only



CRUD

1. Insert : Put (RowA, "B");
 2. Update : Put (RowA, "변경된값");
 이) Row Insert 후
 3. delete : Put (RowA, ""); \rightarrow NULL
 4. Select : value = get (RowA)

NoSQL

1. 전통DB : Row key (Unique) (Not Only)
→ sequence 가 반드시 필요! \Rightarrow NOSQL 등장원인
 2. 기능 : Put(), get() 2개의 헬퍼만 지원
 \hookrightarrow SQL을 지원하지 않는다 : NoSQL
o 데이터
 3. DB데이터구조
2개의 이중의 구조 < Rowkey
Value
 \hookrightarrow Put(rowkey, value), value = get(rowkey)

Value의 유형

1. (k, v) v는 단일 Data Type (문자열)
 2. (k, 문서) Data는 별도의 파일, 파일의 링크를 저장

(k, 결합의 집합) JSON 형태 \Rightarrow HBase

Rowkey 74

• 2020년 1월

→ 대표적인 NoSQL Value의 형식

Ակնարկներ

Map Reduce
SQL on Hadoop
HDFS
Spark Streaming

