= 하둡용 SQL 관련 기술 =

1. 아파치 하이브(Apache Hive)

- 하둡용 SQL 솔루션

- MySQL의 행위, 신택스(syntax), 인터페이스를 에뮬레이트 한 것이 특징

- **자바 API 및 JDBC 드라이버 포함**

- 느리고 읽기 전용이라는 단점

4. 스파크(Spark) SQL

- 하둡 데이터를 실시간으로 인메모리 병렬 처리하는데 목표

- 스파크 SQL은 데이터에 SQL 쿼리를 쓸 수 있는 것이 특징

- 아파치 스파크용 아파치 하이브(Apache Hive for Apache Spark)가 더 나은방식

-->

5. 아파치 피닉스(Apache Phoenix)

- SQL 같은 명령어로 HBase 쿼리를 처리하는 방법

- **내장된 JDBC 드라이버 사용**

- 높은 성능과 읽기/쓰기 작업 가능

6. 클라우데라 임팔라(Cloudera Impala)

- **HDFS나 HBase를 사용**

- SQL 신택스는 아파치 하이브와 동일하다

- HDFS나 HBase에 저장된 데이터를 대상으로 쿼리를 처리

2. 스팅거(Stinger)

3. 아파치 드릴(Apache Drill)

관련문서

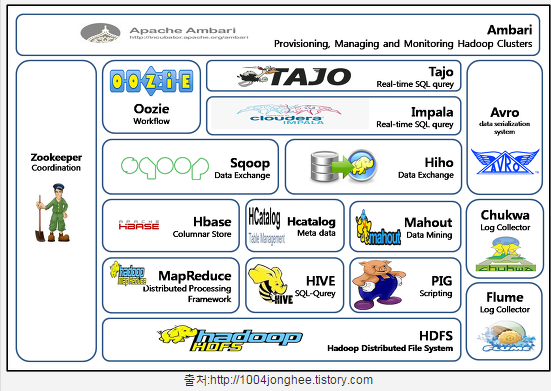
√ http://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/SingleCluster.html#Standlone\_Operation

√ SQL Hive CheatSheet 🡪 http://hortonworks.com/blog/hive-cheat-sheet-for-sql-users/

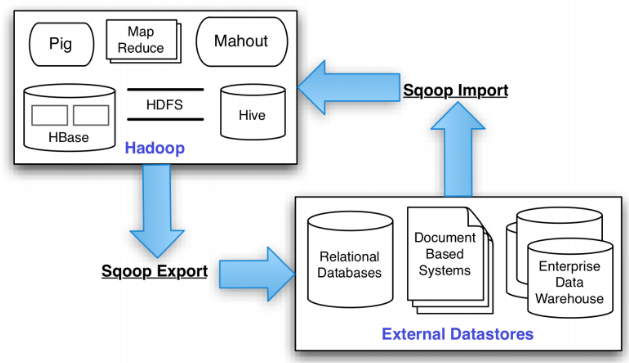
√ http://bigmark.tistory.com/35

★ Hive 설치시 2.xx 버전보다는 1.xxx 버전을 사용하는 것이 안정성 측면에서 낫다고 한다

★ Hadoop EcoSystem에 대하여



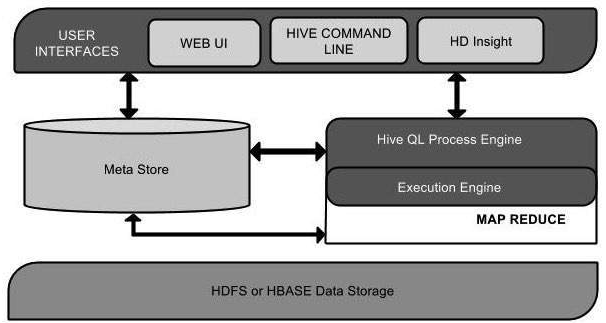
1. 코디네이터
   1. Zookeeper  
      - 분산환경에서 서버 간의 상호 조정을 하기 위하여 사용하는 시스템  
       √ 하나의 서버에만 서비스가 집중되지 않게 서비스를 분산  
       √ 하나의 서버에서 처리한 결과를 다른 서버와도 동기화하여 데이터의 안정성을 보장  
       √ 운영중인 서버에 문제가 생길경우, 대기중인 다른 서버로 전환시켜주는 역할을 함  
       √ 분산환경을 구성하는 서버의 환경설정을 통합적으로 관리
2. **리소스 관라**
   1. **YARN**  
      - 데이터 처리를 위한 클러스터 자원(CPU, 메모리, 디스크)과 스케줄링을 위한 프레임워크  
      - 맵리듀스, 하이브, 임팔라, 타조, 스파크 등의 Application은 YARN에서 리소스를 할당 받아서 처리
   2. Mesos  
      - YARN과 마찬가지로 자원관리를 하는 관리 프로젝트  
      - 동적으로 자원을 할당하고 격리시켜주는 메커니즘을 제공하며, 분산환경에 최적화  
      - 1만대 이상의 노드에도 대응이 가능하며, 여러 API도 제공  
      - 하둡, 스파크, 스톰, 엘라스틱 서치, 카산드라 등의 Application을 실행 가능
3. 데이터 저장
   1. HBase  
      - HDFS 기반의 컬럼 기반 Database  
      - 실시간 랜덤 조회 및 업데이트 가능   
      - 개인의 데이터를 비동기적으로 엡데이트 가능 (단, 맵리듀스는 일괄처리 방식)
   2. Kudu  
      - 컬럼 기반의 Storage / 특정컬럼에 대한 읽기를 고속화 가능  
      - 데이터 발생부터 분석까지의 시간을 HBase보다 단축시킬 수 있음
4. 데이터 수집
   1. 척와(Chukwa)  
      - -
   2. 플럼(Flume)
   3. Scribe
   4. **Sqoop**  
      - 데용량 데이터 전송 솔루션  
      - HDFS, RDBMS, DW, NoSQL 등 다양한 저장소에 대용량 데이터를 전송하는 방법 제공
   5. **Hiho**  
      - 스쿱과 같은 대용량 데이터 전송 솔루션 / Github에 공개되어 있는 상태  
      - 하둡에서 데이터를 가져오기 위한 SQL을 지정할 수 있음  
      - JDBC 인터페이스 지원 / 오라클과 MySQL의 데이터 전송만 지원
   6. Kafka



★ Sqoop 이란?

**Definition :** **Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases**.

🡪 Sqoop은 RDBMS와 Hadoop사이에서 대용량 데이터를 효율적으로 전송하게 고안된 Tool이다

1. 데이터 처리
   1. 피그(Pig)
   2. 머하웃(Mahout)
   3. 스파크(Spark)
   4. 임팔라(Impala)
   5. 프레스토(Presto)
   6. 하이브(Hive)  
      - 하둡 기반의 데이터웨어 하우징용 솔루션  
      - HiveQL이라는 SQL과 유사한 쿼리 언어를 제공  
        
      

< Hive Architecture >

* 1. 타조(Tajo)  
     - 고려대학교 박사과정 학생들이 주도해서 개발한 하둡기반의 데이터웨어 하우징용 솔루션  
     - 맵리듀스엔진이 아닌 자체 분산 처리 엔진을 사용 / **표준SQL을 지원**  
     - HDFS, AWS S3, HBase, DBMS 등에 저장되 데이터를 표준SQL로 조회 가능  
     - 이기종 저장소간의 데이터 조인처리도 가능  
     - 징의 유형에 따라서 하이브나 스파크보다 1.5~10배 빠른성능

1. 워크 플로우 관리
   1. Oozie
   2. Airflow
   3. Azkaban
   4. Nifi
2. 데이터 시각화
   1. Zeppelin
3. 데이터 직렬화
   1. Avro
   2. Thrift

참고 :::

<http://www.nextree.co.kr/p2865/> 🡪 check

<http://blrunner.com/48>

<http://ingest.tips/2015/05/20/a-roundtrip-from-mysql-to-hive/>

<https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_dataintegration/content/using_sqoop_to_move_data_into_hive.html> 🡪 check

<http://hochul.net/blog/datacollector_apache_sqoop_from_rdbms2/> 🡪 check

<http://hochul.net/blog/hive_sqoop_simple_stat_exampl/> 🡪

<http://www.w3ii.com/ko/hive/hive_data_types.html>

<http://blog.naver.com/PostView.nhn?blogId=gyrbsdl18&logNo=188056380>