Hbase

Hbase란 무엇인가?

- 아파치 HBase는 하둡 플랫폼을 위한 공개 비관계형 분산 데이터 베이스이다. 구글의 빅테이블을 본보기로 삼았으며 자바로 쓰여졌다. 아파치 소프트웨어 재단의 아파치 하둡 프로젝트 일부로서 개발되었으며 하둡의 분산 파일 시스템인 HDFS위에서 동작을 한다. 대량의 흩어져 있는 데이터 저장을 위한 무정지 방법을 제공하는 구글의 빅테이블과 비슷한 기능을 한다.

- HBase is a data model that is similar to Google's big table designed to provide quick random access to huge amounts of structured data.

Hbase란 무엇인가?

- 1970 년 이후 RDBMS는 데이터 저장 및 유지 보수 관련 문제에 대한 문제점을 해결할 수 있는 솔루션이었습니다.그러나 , 빅 데이터의 출현 이후 회사는 큰 데이터를 처리하는 이점을 인식하고 Hadoop과 같은 솔루션을 채택하기 시작했습니다.

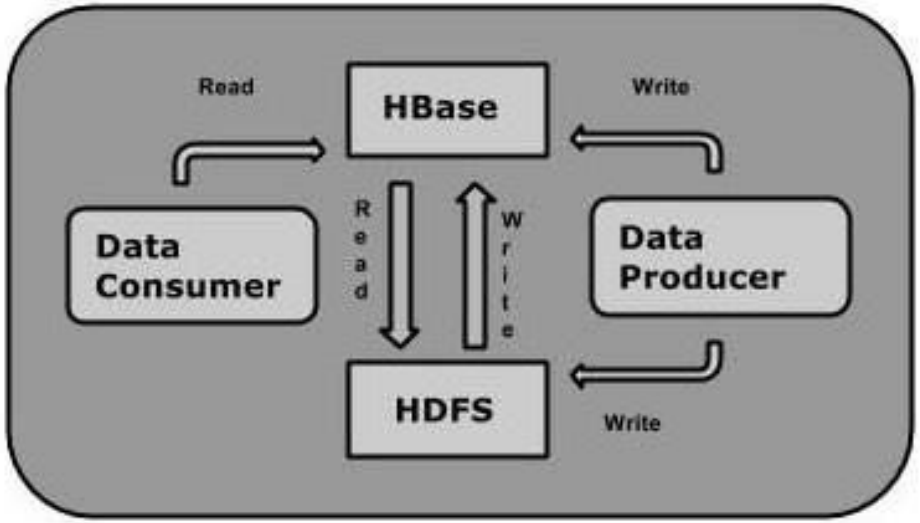
- Hadoop은 큰 데이터를 저장하기 위해 분산 파일 시스템을 사용하고 처리를 위해서는 MapReduce를 사용함. Hadoop은 임의적, 반 또는 비정형과 같은 다양한 형식의 거대한 데이터를 저장하고 처리하는 데 탁월합니다.

- Hadoop의 제한사항
  - Hadoop은 일괄 처리 만 수행 할 수 있으며 데이터는 순차적 방식으로만 접근 가능함. 즉, 가장 단순한 작업 일지라도 전체 데이터 세트를 검색 해야함.
  - Large 데이터 세트는 처리 될 때 또 다른 Large 데이터 세트를 생성하며, 이는 또한 순차적으로 처리됨. 원하는 시점 원하는 데이터를 엑세스 하는 것이 필요함(Random Access)

Hbase란 무엇인가?

- Hadoop Random Access Databases
  - HBase, Cassandra, couchDB, Dynamo 및 MongoDB와 같은 응용 프로그램은 Large 데이터를 저장하고 임의의 엑세스가 가능한 데이터 베이스임

- The Hbase is a distributed **column-oriented** database built on top of the Hadoop file system. It is an open-source project and is horizontally scalable



Hadoop vs HBase

HDFS	HBase
HDFS는 대용량 파일을 저장하기에 적합한 분산 파일 시스템임.	HBase는 HDFS 위에 구축 된 데이터베이스임.
HDFS는 빠른 개별 레코드 조회를 지원하지 않음.	HBase는 Large 테이블을 빠르게 검색함.
높은 대기 시간의 일괄 처리를 제공함. 일괄 처리의 개념이 없음.	수십억 개의 레코드 (무작위 액세스)에서 단일 행에 대한 낮은 대기 시간 액세스를 제공함.
데이터의 순차 액세스 만 제공함.	HBase는 내부적으로 해시 테이블을 사용하며 임의 액세스를 제공하며 빠른 조회를 위해 인덱싱 된 HDFS 파일에 데이터를 저장함.

Hbase 저장 구조

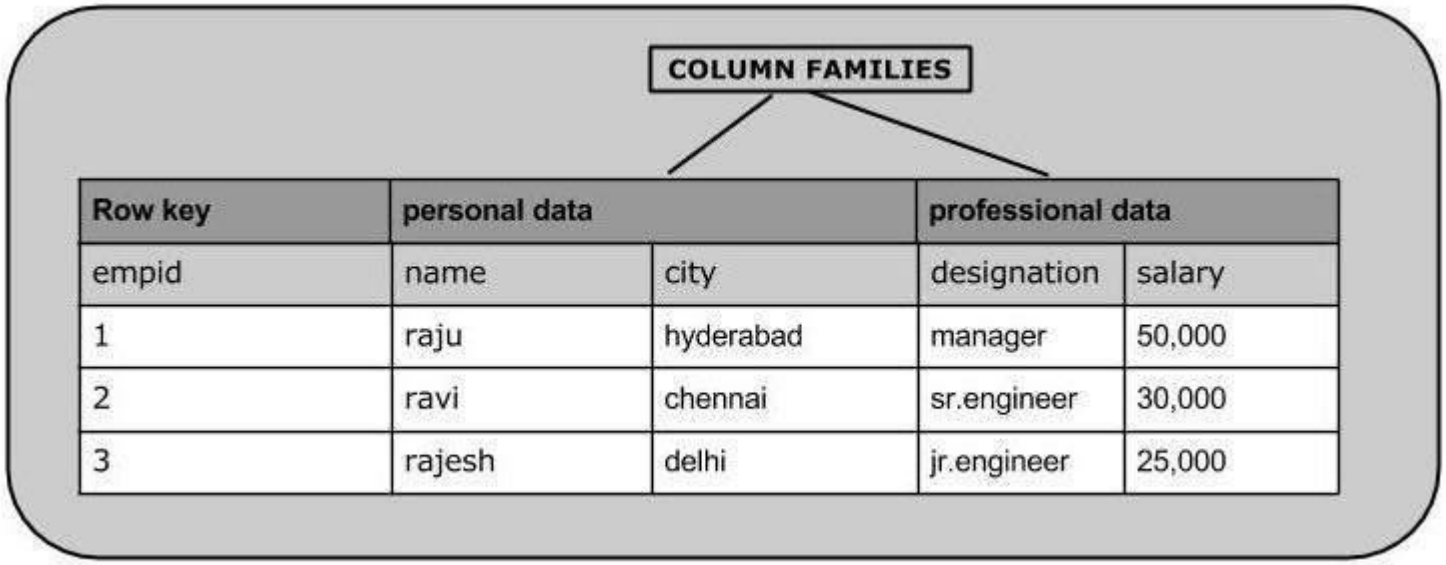
- HBase는 열 기반 데이터베이스(column-oriented database)이며 테이블은 행 별로 정렬됨.
- 테이블 스키마는 키 값 쌍인 열 패밀리만 정의함.
- 테이블에는 여러 열 패밀리가 있으며 각 열 패밀리에는 여러 열이 있을 수 임.
- 후속 열 값은 디스크에 연속적으로 저장됨.
- 테이블의 각 셀 값에는 timestamp가 있음.
  - Table is a collection of rows.
  - Row is a collection of column families.
  - Column family is a collection of columns.
  - Column is a collection of key value pairs

Rowid	Column Family			Column Family			Column Family			Column Family		
	col1	col2	col3	col1	col2	col3	col1	col2	col3	col1	col2	col3
1												
2												
3												

Column Oriented & Row Oriented

- 열 기반 데이터베이스는 데이터 행이 아닌 데이터 열의 섹션으로 데이터 테이블을 저장하는 데이터베이스

행 지향 데이터베이스	열 - 지향 데이터베이스
온라인 트랜잭션 프로세스 (OLTP)에 적합함	온라인 분석 처리 (OLAP)에 적합함
데이터베이스는 적은 수의 행과 열에 맞게 설계되었음.	열 기반 데이터베이스는 거대한 테이블을 위해 설계되었음.



Row key	personal data		professional data	
empid	name	city	designation	salary
1	raju	hyderabad	manager	50,000
2	ravi	chennai	sr.engineer	30,000
3	rajesh	delhi	jr.engineer	25,000

Hbase 및 RDBMS

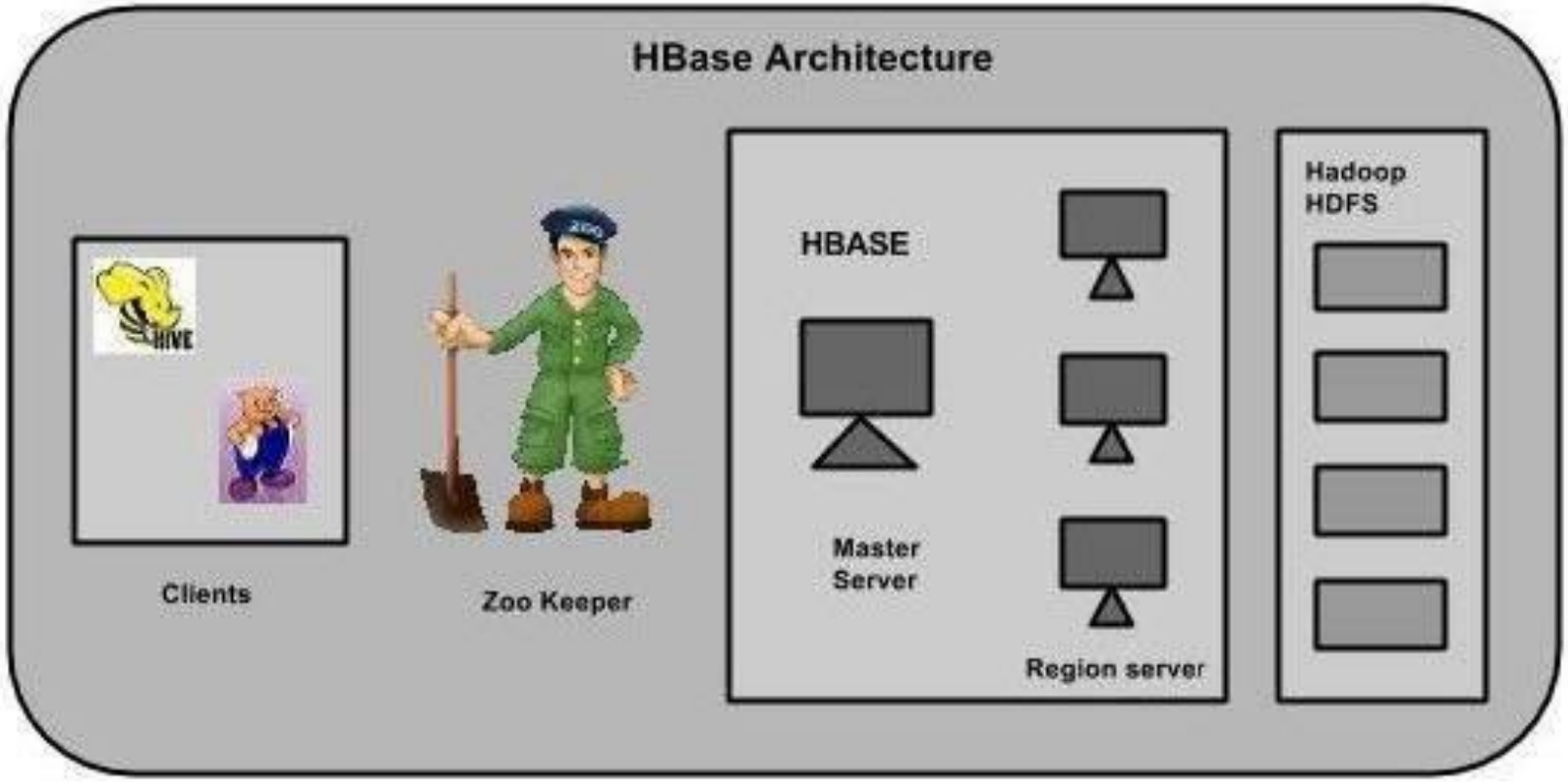
HBase	RDBMS
HBase는 스키마가 없으며 고정 열 스키마라는 개념이 없습니다. 열 패밀리만 정의함.	RDBMS는 테이블의 전체 구조를 설명하는 스키마에 의해 관리됨.
Large 테이블 용으로 설계되었음. HBase는 수평 확장이 가능함.	작은 테이블. 확장하기가 어렵움.
HBase에는 트랜잭션이 없음.	RDBMS는 트랜잭션이 존재함.
데이터를 비정규화.	정규화된 데이터.
반 구조화 된 데이터와 구조화 된 데이터에 유용함.	구조화 된 데이터에 유용함.

Hbase특징

- HBase는 선형 적으로 확장 가능함.
- 자동 장애 지원 기능이 있음.
- 일관된 읽기 및 쓰기 기능을 제공함.
- Hadoop과 소스 및 대상 모두를 통합함.
- 클라이언트를위한 쉬운 자바 API가 있음.
- 클러스터를 통한 데이터 복제를 제공함.

Hbase Architecture

- HBase에서 테이블은 영역으로 분할되며 영역 서버에서 제공됨.
- 영역은 열 패밀리에 의해 수직 나누어서 저장되고 저장은 HDFS에 저장됨



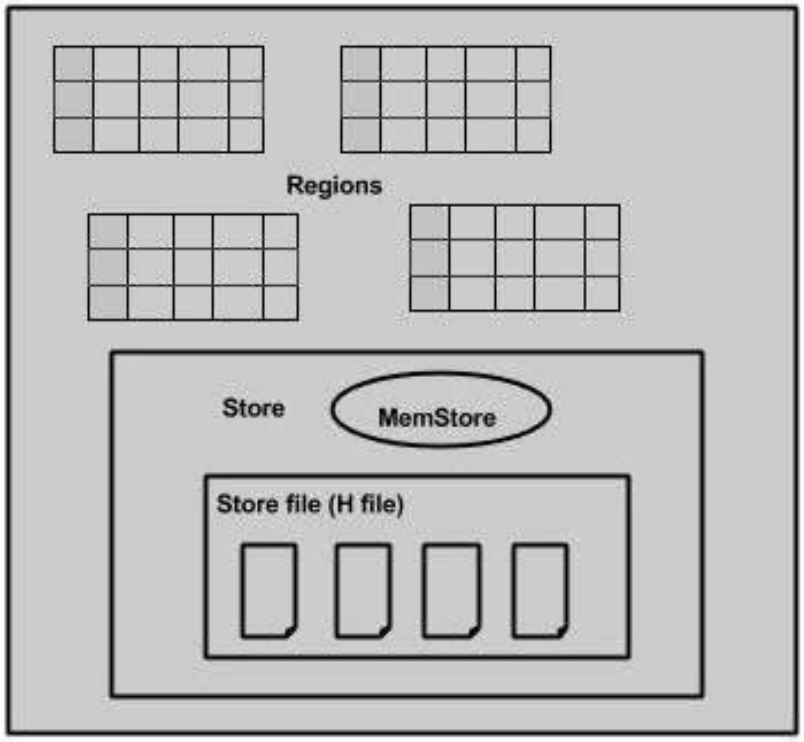
Hbase Architecture

- 주요한 요소는
  - Client library
  - A Master Server
  - Region Servers

- MasterServer
  - region을 region 서버에 할당하고 이 작업을 위해 Apache ZooKeeper의 도움을 받습니다.
  - region 서버에서 region의 로드 밸런싱을 처리합니다. 이 서버는 사용량이 많은 서버를 언로드 하고 영역을 덜 점유한 서버로 이동합니다.
  - 부하 분산을 협상하여 클러스터의 상태를 유지 관리합니다.
  - 스키마 변경 및 테이블 및 열 패밀리 작성과 같은 기타 메타 데이터 조작을 담당합니다.

Hbase Architecture

- Regions
  - Regions는 분할되어 지역 서버 전체에 퍼져있는 테이블들을 말함
  - 클라이언트와 통신하고 데이터 관련 작업을 처리합니다.
  - 그 아래의 모든 지역에 대한 읽기 및 쓰기 요청을 처리하십시오.
  - 영역 크기 임계 값을 따라 영역의 크기를 결정하십시오.



Zookeeper

- Zookeeper는 구성 정보 유지, 이름 지정, 분산 된 동기화 제공 등과 같은 서비스를 제공하는 오픈 소스 프로젝트입니다.
- Zookeeper 는 다른 지역 서버를 나타내는 임시 노드를 가지고 있습니다. 마스터 서버는 이 노드를 사용하여 사용 가능한 서버를 검색합니다.
- 가용성 이외에도 노드는 서버 장애 또는 네트워크 파티션을 추적하는데도 사용됩니다.
- 클라이언트는 Zookeeper 를 통해 지역 서버와 통신합니다.
- pseudo 및 standalone 모드에서 HBase 자체는 Zookeeper를 관리함