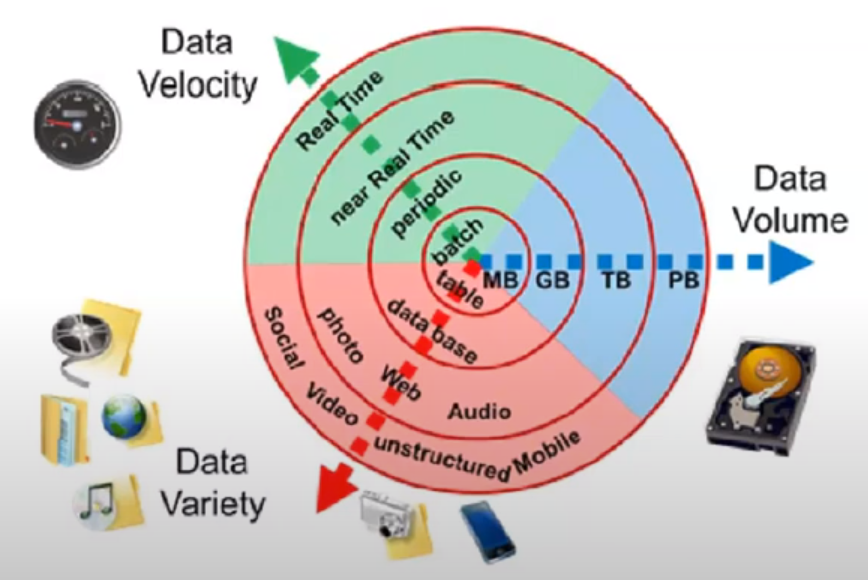
### **빅데이터란 : 대량의 비정형 데이터에 담긴 사람들의 "생각"과 "필요"를 읽어내는것**

* 대량의 정형 비정형 데이터 집합 및 데이터로 부터 가치를 축출하고 결과를 분석하는 기술
* 기존 데이터에 비행 양이나 종류가 너무 커서 기존방법으로는 도저히 수집,저장,분석이 어려운 데이터를 총칭

3V

* volume(규모) - GB -> PB
* variety(다양성) - 정형, 비정형 데이터 (db,sns,....)
* velocity(속도) - 저장,처리까지 실시간으로 데이터를 사용 ex)모빌리티 + Value



빅데이터 처리 프로세스





**Hadoop (High-Availability Distributed Object-Oriented Platform)**

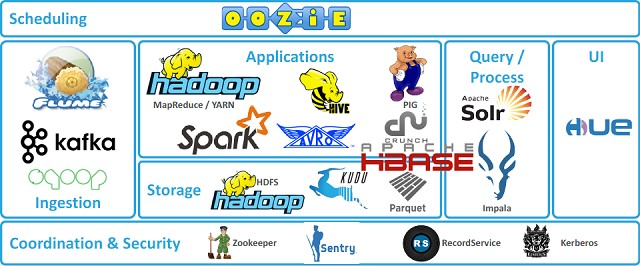
대량의 자료를 처리할 수 있는 큰 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 자바기반 프레임워크

**빅데이터 분석의 날개를 달아준 하둡(Hadoop)**

하둡은 대용량의 데이터를 적은 비용으로 더 빠르게 분석할 수 있는 소프트웨어이며, 빅데이터 처리와 분석을 위한 플랫폼 중 사실상 표준으로 자리잡고 있다.

여러대의 컴퓨터로 데이터를 분석하고 저장하는 방식으로 필요했던 많은 비용과 시간을 단축하게 해주었다

하둡 생태계를 통해 빅데이터를 보다 원할하고 효율적으로 분석할 수 있다.

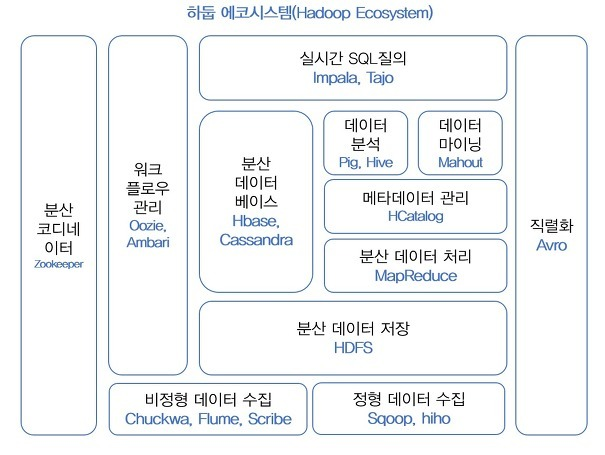


* 하둡 코어 프로젝트 : HDFS(분산 데이터 저장), MapReduce(분산 처리)
* 하둡 서브 프로젝트 : 데이터 마이닝, 수집, 분석 등

그래서 하둡 에코시스템(Hadoop EcoSystem)이 뭔데??

하둡의 코어 프로젝트는 HDFS와 MapReduce지만 그 외에도 다양한 서브 프로젝트들이 많이 있다.

하둡 에코시스템은 바로 그 프레임워크를 이루고 있는 다양한 서브 프로젝트들의 모임이라고 볼 수 있다. 아래 크림들은 하둡 에코시스템을 나타낸 것이다.



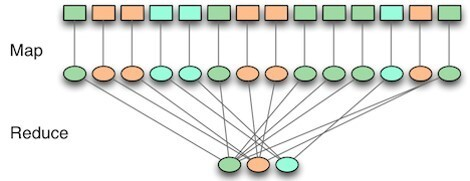
#### 1. 맵리듀스(MapReduce)

대용량의 데이터 처리를 위한 분산 프로그래밍 모델

맵 리듀스 프레임워크를 이용하면 대규모 분산 컴퓨팅 환경에서 대량의 데이터를 병렬로 분석 가능

프로그래머가 직접 작성하는 맵과 리듀스라는 두개의 메소드로 구성

흩어져있는 데이터를 수직화하여, 그 데이터를 각각의 종류별로 모으고(Map) -> 필터링과 sorting을 거쳐 데이터를 뽑아내는(Reduce) 분산처리 기술과 관련 프레임워크를 의미.



맵(Map)

흩어져 있는 데이터를 연관성 있는 데이터들로 분류하는 작업. (key, value의 형태)

리듀스(Reduce)

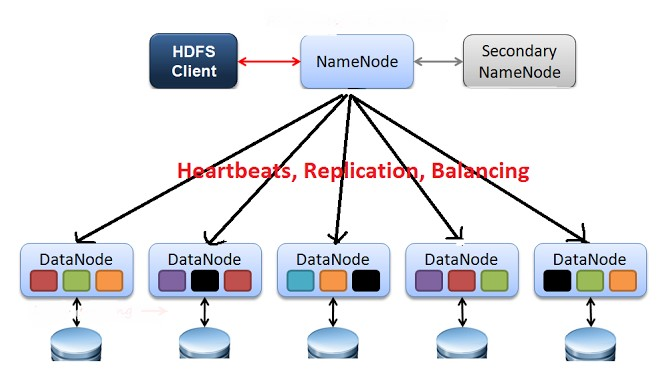
Map에서 출력된 데이터에서 중복 데이터를 제거하고 원하는 데이터를 추출하는 작업

#### 2. 하둡 분산형 파일 시스템 (Hadoop Distributed File System, HDFS)

하둡 분산 파일 시스템은 하둡 프레임워크를 위해 자바 언어로 작성된 분산 확장 파일 시스템이다.

특징

1. HDFS는 대용량의 데이터를 저장하면, 다수의 노드에 복제 데이터도 함께 저장해서 데이터 유실을 방지
2. HDFS에 파일을 저장하거나, 저장된 파일을 조회하려면 스트리밍 방식으로 데이터에 접근
3. 한번 저장한 데이터는 수정할 수 없고, 읽기만 가능해서 데이터 무결성을 유지
4. 데이터 수정은 불가능하지만 파일 이동, 삭제, 복사할 수 있는 인터페이스를 제공



네임노드는 파일과 디렉터리의 읽기(open), 닫기(close), 이름 바꾸기(rename) 기능을 수행하고 데이터 노드와 블록들의 맵핑을 결정한다.

데이터노드는 파일 시스템의 클라이언트가 요구하는 읽기(read), 쓰기(write) 기능을 담당하고 데이터 노드는 네임 노드에서의 생성, 삭제, 복제 등과 같은 기능도 수행

#### 3. Apache Saprk(인메모리기반)

분산 클러스터 컴퓨팅 프레임워크

RDD, Data Frame, Data Set의 3가지 API를 제공하는데, 이러한 데이터를 바탕으로 In-memory 연산을 가능하도록 하여 디스크 기반의 Hadoop에 비해 성능을 약 100배 정도 끌어올렸다.



Spark는 Cluster들을 관리하는 Cluster Manager와 데이터를 분산 저장하는 Distributed Storage System이 필요하다.

자주 사용되는 Cluster Manager로는 Hadoop의 YARN이나 Apache Mesos 등이 있다.

또한 Distributed Storage System으로는 HDFS(Hadoop Distributed File Syste), MapR-FS(MapR File System), Cassandra, OpenStack Swift, Amazon S3, Kudu, custom solution 등을 적용할 수 있다.

가장 많이 사용되는 Storage System는 Hadoop인데, zlib and Bzip2와 같은 압축 알고리즘을 지원하며, Spark와 같은 머신에서 구동가능하기 때문이다.

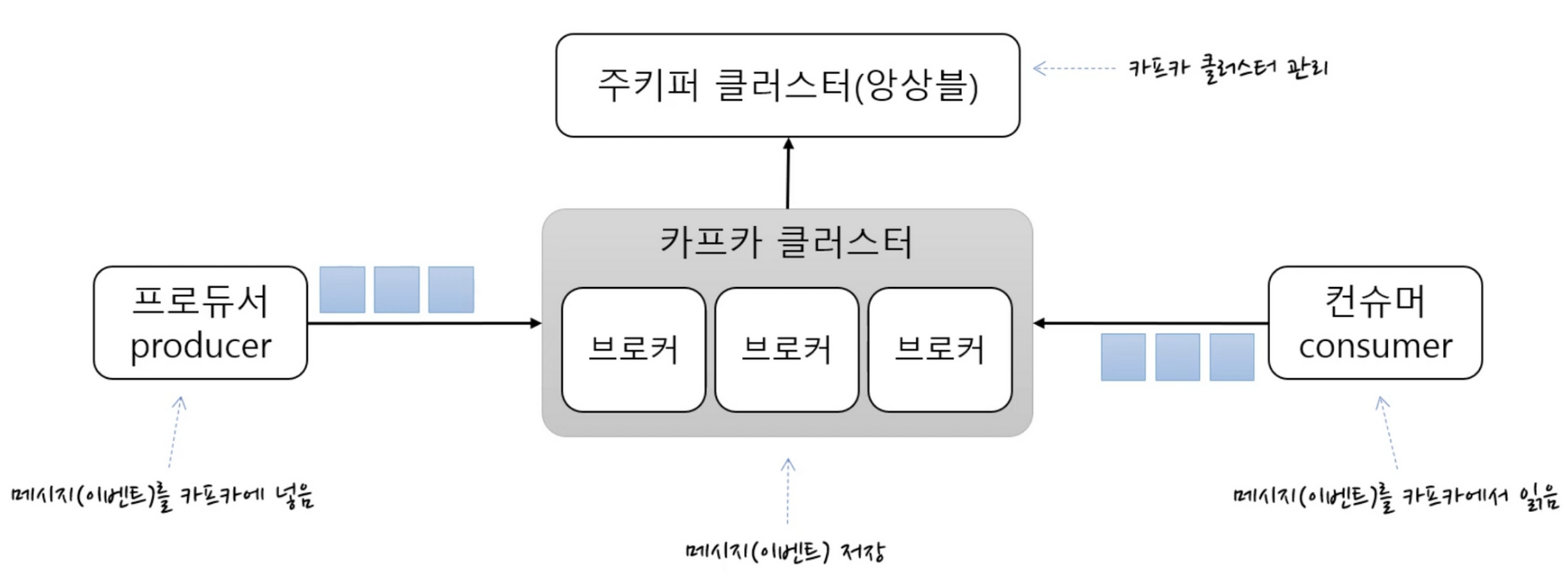
#### 4. Kafka(데이터 스트림 관리) - 분산 이벤트 스트리밍 플랫폼(distributed event streaming platform)

Kafka는 Linkedin에서 개발된 실시간 데이터 스트리밍 서버입니다.

Kafka를 경유하여 데이터를 입/출력하도록 하면 데이터 처리 패키지(Hbase나 spark등)과의 연동이 매우 쉽다.

심지어 HBASE등의 제품에 대해서는 CDC 연동도 가능합니다.

Oracle등의 RDBMS와도 연동됩니다.



#### 5. Zookeeper (분산코디네이터)

분산 환경에서 서버들간에 상호 조정이 필요한 다양한 서비스를 제공하는 시스템

하나의 서버에만 서비스가 집중되지 않도록 서비스를 알맞게 분산하여 동시에 처리하게 해줌

하나의 서버에서 처리한 결과를 다른 서버들과도 동기화 -> 데이터 안정성 보장

운영 서버에서 문제가 발생해 서비스를 제공할 수 없는 경우 다른 대기중인 서버를 운영 서버로 바꿔 서비스가 중지없이 제공되게 해줌

분산 환경을 구성하는 서버들의 환경설정을 통합적으로 관리

공식 사이트 : <http://zookeeper.apache.org/>

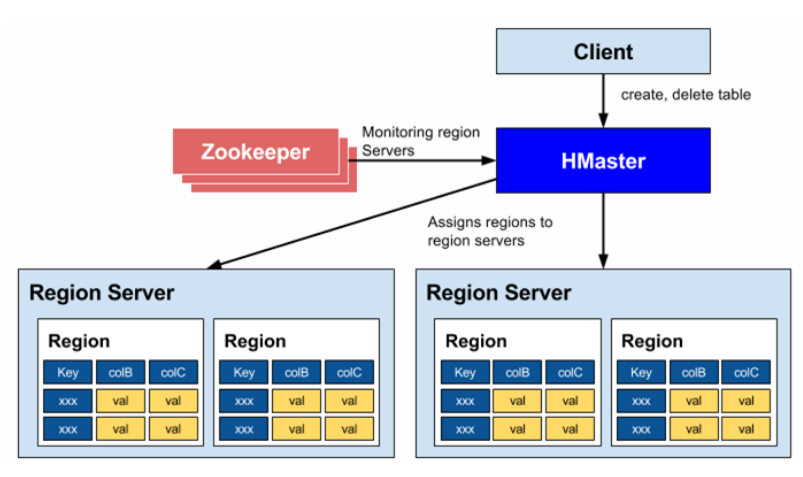
#### 

#### 6. HBase

HDFS의 칼럼 기반 데이터베이스

구글의 BigTable 논문을 기반으로 개발된 것으로, 실시간 랜덤 조회 및 업데이트가 가능하며, 각각의 프로세스들은 개인의 데이터를 비동기적으로 업데이트 할 수 있다. 단, MapReduce는 일괄처리 방식으로 수행

공식 사이트 : http://hbase.apache.org



#### 7. Hive

하둡 기반의 데이터웨어하우징용 솔루션

페이스북에서 만든 오픈소스로, SQL과 매우 유사한 HiveQL이라는 쿼리를 제공한다.

자바를 잘 모르는 데이터 분석가들도 쉽게 하둡 데이터를 분석할 수 있게 도와준다.

HiveQL은 내부적으로 MapReduce 잡으로 변환되어 실행된다.

공식 사이트 : <http://hive.apache.org>

#### 8. Ooozie(워크플로우관리)

하둡의 작업을 관리하는 워크플로우 및 코디네이터 시스템

자바 서블릿 컨테이너에서 실행되는 자바 웹어플리케이션 서버로, MapReduce 작업이나 Pig 작업 같은 특화된 액션들로 구성된 워크플로우를 제어

공식 사이트 : http://incubator.apache.org/oozie

#### 9. Pig

복잡한 MapReduce 프로그래밍을 대체할 Pig Latin이라는 자체 언어를 제공.

MapReduce API를 매우 단순화시키고 SQL과 유사한 형태로 설계되었다.

공식 사이트 : http://pig.apache.org

#### 10. Mahout

하둡 기반 데이터 마이닝 알고리즘을 구현한 오픈소스.

현재 분류(classification), 클러스터링(clustering), 추천 및 협업 필터링(Recommenders/Collavorative filtering), 패턴 마이닝(Pattern Mining), 회귀 분석(Regression), 차원 리덕션(Dimension Reduction), 진화 알고리즘(Evolutionary Algorithms) 등 주요한 알고리즘을 지원하고 있다.

Mahout을 그대로 사용할 수도 있지만, 자신의 비즈니스 환경에 맞게 최적화 해 사용하는 경우가 대부분.

공식 사이트 : http://mahout.apache.org

#### 11. HCatalog

하둡으로 생성한 데이터를 위한 테이블 및 스토리지 관리 서비스

HCatalog의 가장 큰 장점은 하둡 에코 시스템들간의 상호 운용성 향상이다. 예를 들어 Hive에서 생성한 테이블이나 데이터 모델을 Pig나 MapReduce에서 손쉽게 이용할 수가 있다.

공식 사이트: http://incubator.apache.org/hcatalog

#### 12. Avro

RPC(Remote Procedure Call)과 데이터 직렬화를 지원하는 프레임워크

JSON을 이용해 데이터 형식과 프로토콜을 정의하며, 작고 빠른 바이너리 포맷으로 데이터를 직렬화한다.

공식 사이트: http://avro.apache.org

#### 13. Chukwa

분산 환경에서 생성되는 데이터를 HDFS에 안정적으로 저장시키는 플랫폼

분산된 각 서버에서 에이전트(agent)를 실행하고, 콜랙터(collector)가 에이전트로부터 데이터를 받아 HDFS에 저장한다. 콜렉터는 100개의 에이전트당 하나씩 구동되며, 데이터 중복 제거 등의 작업은 MapReduce로 처리합니다.

공식 사이트: http://incubator.apache.org/chukwa

#### 14. Flume

Chukwa 처럼 분산된 서버에 에이전트가 설치되고, 에이전트로부터 데이터를 전달받는 콜랙터로 구성

차이점은 전체 데이터의 흐름을 관리하는 마스터 서버가 있어서, 데이터를 어디서 수집하고, 어떤 방식으로 전송하고, 어디에 저장할 지를 동적으로 변경할 수 있다.

공식 사이트: http://incubator.apache.org/projects/flume.html

#### 15. Scribe

페이스북에서 개발한 데이터 수집 플랫폼이며, Chukwa와는 다르게 데이터를 중앙 집중 서버로 전송하는 방식이다.

최종 데이터는 HDFS외에 다양한 저장소를 활용할 수 있으며, 설치와 구성이 쉽게 다양한 프로그램 언어를 지원함.

HDFS에 저장하기 위해서는 JNI(Java Native Interface)를 이용해야 한다.

공식 사이트: https://github.com/facebook/scribe

#### 16. Sqoop

대용량 데이터 전송 솔루션.

Sqoop은 HDFS, RDBMS, DW, NoSQL등 다양한 저장소에 대용량 데이터를 신속하게 전송할 수 있는 방법을 제공한다.

Oracle, MS-SQL, DB2 등과 같은 상용 RDBMS와 MySQL, PostgresSQL과 같은 오픈소스 RDBMS등을 지원합니다.

공식 사이트: http://sqoop.apache.org

#### 17. Hiho

Sqoop과 같은 대용량 데이터 전송 솔루션이며, 현재 github에서 공개되어 있습니다. 하둡에서 데이터를 가져오기 위한 SQL을 지정할 수 있으며, JDBC 인터페이스를 지원합니다.

공식 사이트: https://github.com/sonalgoyal/hiho

#### 12. Impala

클라우드데라에서 개발한 하둡 기반의 실시간 SQL 질의 시스템

맵리듀스를 사용하지 않고, 자체 개발한 엔진을 사용해 빠른 성능을 보여줍니다.

임팔라(Impala)는 데이터 조회를 위한 인터페이스로, HiveQL을 사용합니다. 수초 내에 SQL 질의 결과를 확인할 수 있으며, HBase와도 연동이 가능합니다.

공식 사이트 : https://github.com/cloudera/impala

#### 13. Tajo

고려대학교 정보통신대학 컴퓨터학과 DB연구실 박사 과정학생들이 주도해서 개발한 하둡 기반의 DW 시스템.

데이터 저장소는 HDFS를 사용하되, 표준DB언어인 SQL을 통해 실시간으로 데이터를 조회할 수 있습니다. Hive보다 2 ~ 3배 빠르며, 클라우드데라의 임팔라(Impala)와는 비슷한 속도를 보여줍니다. 임팔라가 클라우드데라의 하둡을 써야 하는 제약에 비해, 특정 업체 솔루션에 종속되지 않는 장점이 있습니다.

공식 사이트 : http://tajo.incubator.apache.org