빅데이터란?

> 빅데이터는 구조화 된 데이터와 구조화되지 않은 대량의 데이터를 기술하는 용어로 조직적 관점에서 더 나은 의사 결정과 전략적 비즈니스 이동으로 이어지는 통찰력을 분석할 수 있도록 해줍니다.

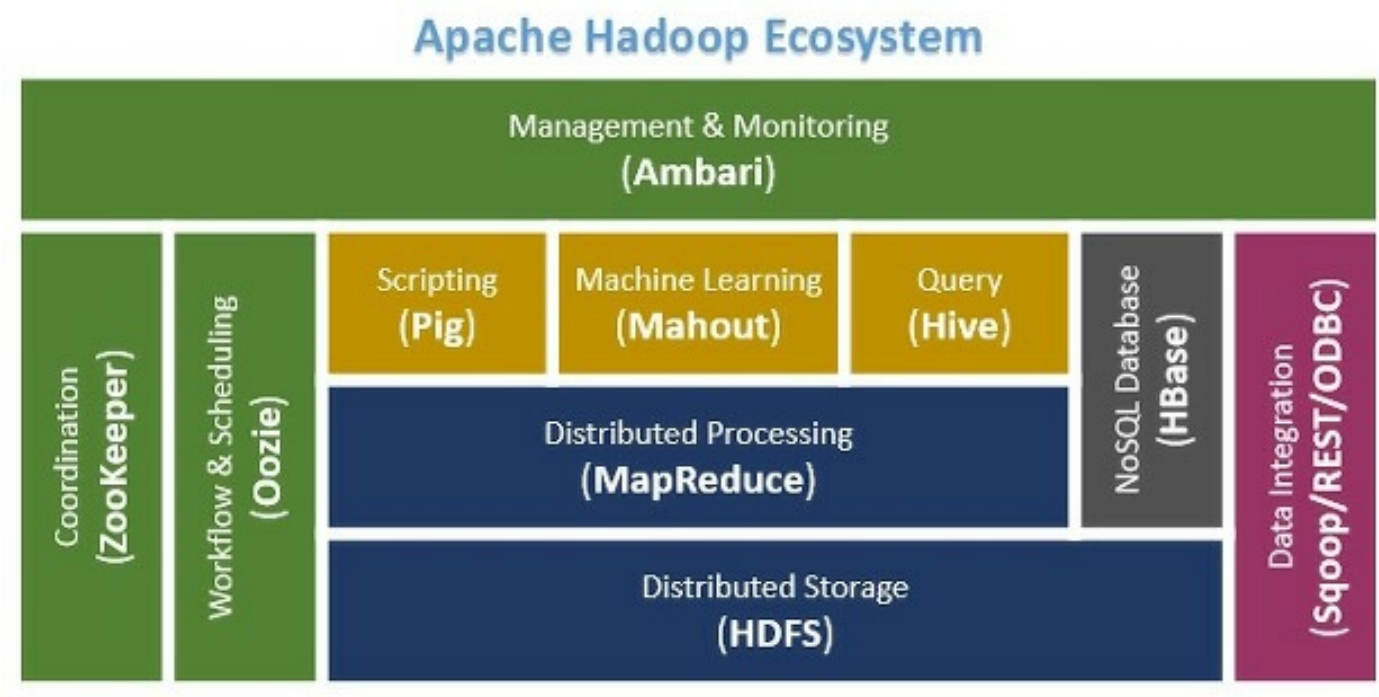
빅데이터의 특징 -3V + 5V

- **Volume** : 조직은 비즈니스 거래, 소셜 미디어 및 센서 또는 기계 대 기계 데이터의 정보를 비롯한 다양한 출처에서 데이터를 수집합니다. 과거에는 이를 저장하는 것이 문제 였을 것이지만 새로운 기술(Hadoop과 같은)이 부담을 덜어 주었습니다.
- **Velocity** : 전례없이 빠른 속도로 데이터가 스트리밍되므로 처리 또한 적절하게 되야 합니다. RFID 태그, 센서 및 스마트 미터링은 거의 실시간으로 급류를 처리 할 필요성을 불러 일으키고 있습니다.
- **Variety** : 데이터는 전통적인 데이터베이스의 구조화 된 숫자 데이터에서 구조화되지 않은 텍스트 문서, 전자 메일, 비디오, 오디오, 주식 시세 표시기 데이터 및 금융 거래와 같은 모든 유형의 형식으로 제공됩니다.
- **Variability** : 증가하는 속도 및 다양한 데이터 외에도 데이터 흐름은 주기적인 피크와 매우 일치하지 않을 수 있습니다. 구조화 되어있는 데이터도 변동성이 큰 상태인데 구조화되지 않은 데이터의 경우 더욱 그렇습니다.
- **Complexity** : 오늘날의 데이터는 여러 소스에서 제공되므로 시스템간에 데이터를 연결, 일치, 정리 및 변환하기가 어렵습니다. 그러나 관계, 계층 구조 및 여러 데이터 연계를 연결하고 상호 연관시켜야하거나 데이터가 신속하게 제어 할 수 없게 될 수 있습니다.

Hadoop이란?

Hadoop은 상용 하드웨어 클러스터에 데이터를 저장하고 응용 프로그램을 실행하기 위한 오픈 소스 소프트웨어 프레임 워크입니다. 모든 종류의 데이터, 엄청난 처리 능력 및 사실상 무제한의 동시 작업이나 작업을 처리 할 수 있는 능력을 제공합니다.

Hadoop Ecosystem이란? (그 중 HDFS, MAPREDUCE?)



Apache Hadoop은 저장기능인 HDFS와 처리기능인 MAPREDUCE 이렇게 기본만 제공합니다.

이를 보완하고 효율적으로 적용할 수 있도록 다양한 서브 프로젝트가 제공되는데, 이러한 것들로 구성된 것이 바로 HADOOP ECOSYSTEM입니다.

SPARK란?

Spark는 분석 응용 프로그램을 실행하기 위한 오픈 소스, 확장 가능하고 대규모 병렬 메모리 내 실행 환경입니다.

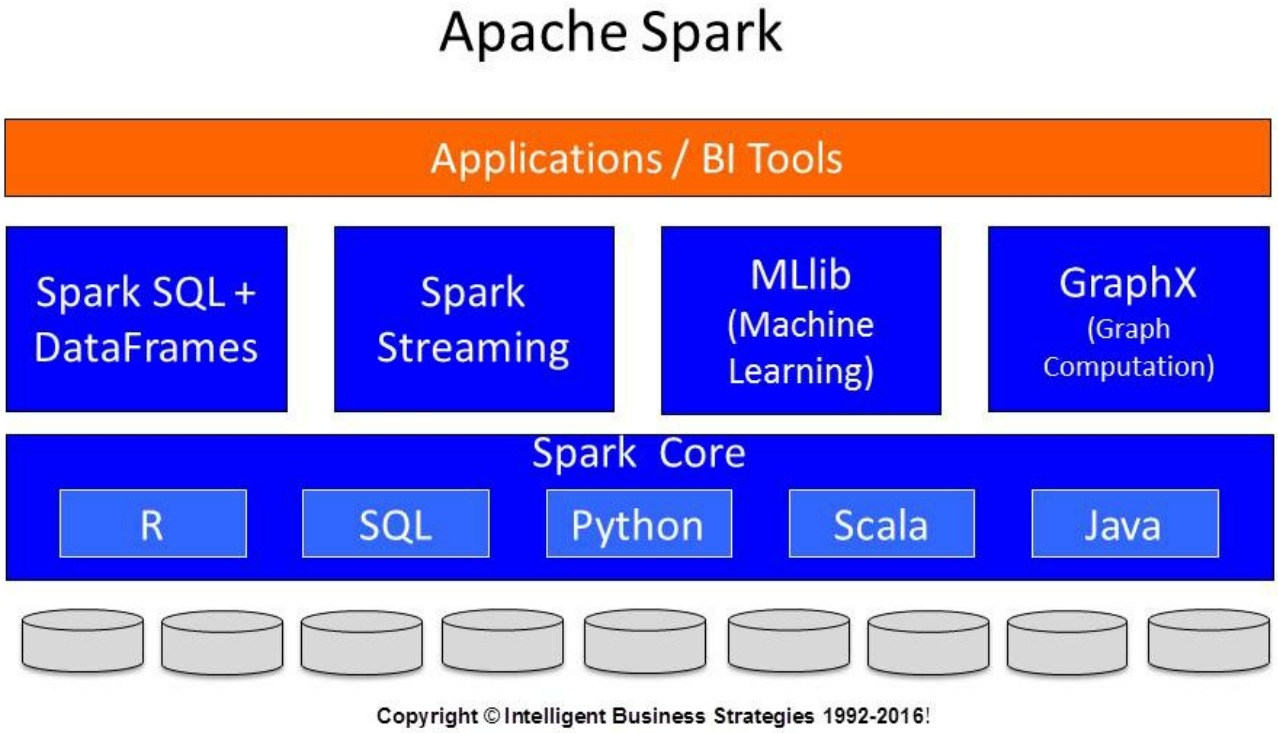
왜 Spark인가?

MapReduce와 마찬가지로 Spark은 클러스터를 통해 데이터를 배포하고 해당 데이터를 병렬로 처리합니다. 차이점은 MapReduce와 달리 Spark는 메모리에서 작동하므로 MapReduce보다 훨씬 빠른 데이터 처리가 가능하다는 점입니다.

RDD?

RDD는 Spark에서 데이터 세트로 수행하는 모든 작업에서 해당 기능으로 인해 생성되는 것입니다. RDD의 R은 장애에 대한 복원력을 의미합니다. RDD의 DD는 Distributed Dataset의 약자입니다. 따라서 RDD는 실패에 대한 내성을 가진 분산 데이터 세트를 의미합니다.

Spark 구조



CLOUDERA란?

Cloudera는 Big Data World와 관련된 소프트웨어를 개발, 배포, 구현 및 지원하는 Silicon Valley의 가장 큰 회사입니다.