빅데이터 처리 이해

몇가지 궁금한 점

01. 빅데이터는 무엇인가?

02. 빅데이터 처리 시스템은 무엇일까?

03. 그러면 빅데이터를 처리하는 기술은 어떤 것이 있을까?

1-1 빅데이터는 무엇인가?

▶ **기존 데이터 처리 응용 프로그램 소프트웨어가 적절하게 처리하기에는 너무 큰 데이터 세트를 나타내는 용어이다.**

▶ **빅데이터는 Volume(볼륨), Variety(다양성), Velocity(속도)와 관련이 있다.**

엄청난 데이터를 어떻게 해야 할까?

기존의 방식으로는 처리가 어렵구나~

1-2 빅데이터 처리 시스템은 무엇일까?

▶ **대규모 양의 데이터의 수집/관리/유통/분석을 처리하는 일련의 분산 병렬처리 프레임워크를 말한다.**

▶ **시스템의 꼭 필요한 요소는 다음과 같다.**

(가) 대규모 데이터 처리를 위한 **확장성**

(나) 데이터 생성 및 처리 속도를 지원하기 위한 **실시간 처리**

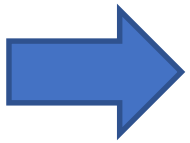
(다) **비정형 데이터 처리에 대한 지원**

1-3 빅데이터 처리하는 기술

▶ 왜 빅데이터 처리 기술이 필요한가?

(가) 전통적인 데이터 처리 기술은 현재의 폭발적인 데이터의 다양한 형태를 처리하기에 적합하지 않다.

(나) 다양한 형태의 포맷과 실시간 처리 속도에 대한 요구사항을 충족시키기에는 한계가 있다.



빅데이터 처리 기술은 생성되는 데이터에 대한 크기, 발생속도, 다양성에 따라 정의가 가능하다.

1-3 빅데이터 처리하는 기술

▶ 빅데이터 처리 기술(S/W)

HADOOP(하둡)

- 오픈소스
- HDFS와 맵리듀스를 구현한 빅데이터 처리 프레임 워크

맵리듀스(MapReduce)

- 맵 함수와 리듀스 함수 기반으로 구성되고, 데이터를 병렬처리하는 방식이다.

HDFS(하둡분산파일시스템)

그외 : AMAZON- S3(파일 분산 시스템 사용)

SPARK(스파크)

1-3 빅데이터 처리하는 기술

▶ 빅데이터 처리 기술(S/W)

프레스트(Presto)

- 페이스북에서 개발된 하둡을 위한 SQL 처리 엔진.
- SQL언어로 데이터를 빠르게 분석할 수 있다.
- 클라우데라 임팔라와 아파치 타조등과 유사

빅쿼리(BigQuery)

- 구글에서 개발한 대용량 데이터를 처리하는 엔진
- 빅쿼리 API를 사용하여 질의를 전송하는 방식
- 최대 2TB까지 데이터 업로드 후, 무료로 분석 가능.

1-3 빅데이터 처리하는 기술

- **빅데이터 처리 기술(S/W)**

  - 파이썬(Python)
  - 알(R)
  - Scala(스칼라)

빅데이터 처리 방식에 대해 알아보자.

1-4 빅데이터 처리하는 방식

- **대화형 처리**
- **배치 처리**
- **실시간 처리**

1-4 빅데이터 처리하는 방식

- **대화형 처리**
  - 질문을 던지고 이에 대한 답을 얻는 형태
  - 하이브 쉘(Hive), 임팔라(Impala), 피그(Pig)
  - Presto(프레스트), Impala(임팔라), 스파크(Spark), Hive(하이브), 피그(Pig)
- 배치 처리
- 실시간 처리

1-4 빅데이터 처리하는 방식

▶ 대화형 처리

▶ **배치 처리**

- 일일, 주간, 월간 보고서 작성 등 주기적으로 작업을 수행하는 형식
- 답을 얻기까지 일정시간이 소요된다.
- Mapreduce(맵리듀스), 하이브(Hive), 피그(Pig), 스파크(Spark)

▶ 실시간 처리

1-4 빅데이터 처리하는 방식

- 대화형 처리
- 배치 처리
- **실시간 처리**
  - 주로 이벤트성 응답이나 데이터 스트림의 준 실시간 처리를 위해 사용된다.
  - 결제나 비정상 카드 사용 등에 대한 데이터 분석에 사용된다.
  - Storm(스톰), KCL, 스파크 스트리밍(Spark Streaming)

빅데이터 처리 솔루션은 어떤 것이 있을까?

1-5 빅데이터 처리 솔루션

- 아파치 소프트웨어 파운데이션
- 클라우데라(Cloudera)
- 호튼웍스(Hortonworks)
- 맵알(MapR Technologies)
- 마이크로소프트 애저(Azure)
- 아마존 AWS(Amazon Web Service)

1-6 빅데이터 처리 방식

▶ 독립 모드(standalone mode)

- 데몬 프로세스가 없이 모든 프로그램이 하나의 JVM(Java Virtual Machine)에서 동작한다.
- 개발 테스트하는 동안에 사용하는 모드
- 실제 빅데이터 처리 환경으로 적합하지 않음.

▶ 의사분산 모드(pseudo-distributed mode)

▶ 완전분산 모드(fully distributed mode)

1-6 빅데이터 처리 방식

▶ 독립 모드(standalone mode)

▶ **의사분산 모드(pseudo-distributed mode)**

- 하둡 데몬 프로세스가 하나의 로컬 컴퓨터에 여러 개 동작하는 모드
- 작은 규모의 클러스터를 시뮬레이션하는 경우에 사용 가능

▶ 완전분산 모드(fully distributed mode)

1-6 빅데이터 처리 방식

- 독립 모드(standalone mode)
- 의사분산 모드(pseudo-distributed mode)
- **완전분산 모드(fully distributed mode)**
  - 하둡 데몬 프로세스가 클러스터로 구성된 여러 개의 컴퓨터에 나누어 동작
  - 마스터 노드(master node)와 슬레이브 노드(slave node)가 구분되어 있음.
  - 데이터들은 실제 데이터 노드(data node)에 분산 저장된다. 복제본을 여러 노드에 나누어 저장.