1️⃣

# Chapter 1

## 1-1 빅데이터의 정착

빅데이터라는 개념의 등장 이후로 두가지 문제에 당착

1. 데이터 분석 방법을 모른다
2. 데이터 처리에 수고와 시간이 걸린다

해당 책에서는 2번인 데이터 처리에 수고와 시간이 걸린다를 더 포커스 하여 해결하고자 함

데이터 처리에 수고와 시간이 걸린다 → 그렇다면 그것을 어떻게 효율적으로 실행할 것인가?

이에 따라 두가지 기술이 등장하였다:

**hadoop**

- 다수의 컴퓨터에서 대량의 데이터를 처리,관리하기 위함 시스템
- ex. 검색엔진을 만들기 위한 데이터를 저장해둘 스토리지와 처리할 수 있는 구조 필요

**NoSQL**

- RDB의 제약을 제거하는 것을 목표로 한 데이터베이스의 총칭
- RDB보다 고속의 읽기, 쓰기가 가능하고 분산처리에 뛰어남

모여진 데이터를 나중에 집계하는건 하둡, SQL은 어플리케이션에서 온라인으로 접속

기존에도 데이터 분석을 기반으로 하는 엔터프라이즈 데이터 웨어하우스라는 장치가 있었지만,

하드웨어와 소프트웨어의 결합으로 업데이트에 번거로움이 존재

따라서 사람들이 하둡으로 갈아타기 시작하고, 하둡을 베이스로 하는 비즈니스가 성립.

(중간에 하둡을 넣어줌으로써 데이터 웨어하우스의 부하를 줄여준다)

하둡과 SQL의 등장에 비슷한 시기에

개인이 직접 할 수 있는 데이터 분석의 폭이 다음과 같은 툴의 등장으로 인해 확대되기 시작

**클라우드 서비스**

- 여러 컴퓨터에 분산처리하는 빅데이터의 특징에 맞게 하드웨어를 준비하고 관리하는 일이 간단하지 않음 → 클라우드의 등장
- 시간 단위로 필요한 자원을 확보할 수 있어서 방법만 알면 언제든지 이용할 수 있음

**데이터 디스커버리**

- 셀프서비스용 BI도구: 데이터 웨어하우스와 조합되어 사용된 경영자용 시각화 시스템
- 대화형으로 데이터를 시각화 하여 가치 있는 정보를 찾으려고 하는 프로세스

# 1-2 빅데이터 시대의 데이터 분석 기반

→ 무엇이 빅데이터 기술을 기존의 기술보다 더 낫게 하는가?

## 빅데이터의 기술: 분산 시스템을 활용해서 데이터를 가공

**데이터 파이프라인**

- 차례대로 전달해나가는 데이터로 구성된 시스템
- 어디서 데이터를 수집/무엇을 실현 에 따라 변화

**수집**

- 데이터는 일반적으로 여러 장소에서 각각 다른형태로 존재
- 각각 서로 다른 기술로 데이터를 전송

**전송**

- **벌크:** 이미 어딘가에 존재하는 데이터를 정리해 추출, 정기적으로 수집
- **스트리밍:** 차례차례로 생성되는 데이터를 끊임없이 계속해서 보냄

**스트림 처리와 배치처리**

- 예전에는 벌크형이 많이 사용되었지만, 요즘은 스트리밍 방식이 더 많아짐
- **스트림 처리**: 스트리밍 형 방법으로 받은 데이터를 실시간으로 처리
- 이러한 스트림 처리는 장기적인 데이터분석에는 적합하지 않음 → 배치 처리의 등장

- **배치 처리**: 어느정도 정리된 데이터를 효율적으로 가공

**분산 스토리지**

- 여러 컴퓨터와 디스크로부터 구성된 스토리지 시스템
- **객체 스토리지**: 한 덩어리로 모인 데이터에 이름을 부여, 파일로 저장
- ex. Amazon S3, NoSQL

**분산 데이터 처리**

- 위에서 언급된 분산 스토리지에 저장된 데이터를 처리
- 나중에 분석하기 쉽도록 데이터를 가공해서 그 결과를 외부 데이터 베이스에 저장
- 대부분 SQL을 사용, 집계 방법에는 두가지 방법이 있음
  - 쿼리 엔진 사용: 대표적인 예로 Hive, 대화형 쿼리 엔진
  - 외부의 데이터 웨어 하우스 제품을 이용: 분산 스토리지에서 추출한 데이터를 데이터 웨어 하우스에 적합한 형식으로 변환 (call ETL 프로세스. 추출 → 가공 → 로드)

**워크플로 관리**

- 데이터 파이프라인의 동작을 관리
- 정해진 시간에 배치처리를 스케쥴대로 실행
- 오류 발생 시 관리자에게 통지하는 목적

빅데이터의 데이터 파이프라인을 실현하려면 많은 기술과 소프트웨어가 사용된다.
전부가 필요하진 않지만, 더욱 좋은 데이터 분석 환경을 구축하는 데는 각각의 특징을 이해해둘 필요가 있다.

# 데이터 웨어하우스와 데이터 마트

**데이터 마트**

데이터 웨어 하우스는 업무에 있어서 중요한 데이터 처리에 사용, 따라서 계속 접속하여 과부하를 불러 일으키는것은 곤란함.

이에 대한 해결책으로 데이터 웨어하우스에서 필요한 데이터만을 추출하여 데이터 마트를 구축

**데이터 레이크**

다양한 형데의 데이터를 원래의 형태로 축적

## 데이터 분석 기반을 단계적으로 발전시키기

**애드 훅 분석**

수작업으로 데이터를 집계

‘일회성 데이터 분석’

## 데이터를 수집하는 목적

1. 데이터 검색 : 대량의 데이터 중 조건에 맞는것을 찾기
2. 데이터 가공 : 업무 시스템의 일부로서 데이터 처리 결과를 이용하고 싶을 때
3. 데이터 시각화 : 데이터를 시각적으로 봄으로써 알고 싶은 정보가 있을때

# 1-3 스크립트 언어에 의한 특별 분석과 데이터 프레임

## 데이터 처리와 스크립트 언어

데이터를 수집할 시에는 다양한 case들이 존재하고, 이러한 범용성을 가지고 있는 스크립트 언어를 사용해야한다.

대표적인 언어로는 r과 python이 있다

## 데이터 프레임, 기초 중의 기초

데이터 프레임은 표 형식의 데이터를 추상화한 객체.

(사실 데이터 프레임 부분 좀 익숙해서...정리하면서 머리속에 넣을만한게 없음)

그 뒤의 부분은 약간 실습 느낌이고, 개념적인 부분이 거의 없어서 정리하는것이 의미가 없다고 생각.
따라서 skip

# 1-4 BI 도구와 모니터링

# 스프레드시트에 의한 모니터링

### 모니터링

데이터를 보다 계획적으로 데이터의 변화를 추적해 나가는것을 모니터링이라고 함

데이터는 현재 상황을 파악하기 위한 하나의 도구로 사용할 수 있기 때문에, 모니터링을 할 필요가 있음.

모니터링을 통해 알아낸 숫자의 의미를 제대로 이해하기 위해서는 사전지식이 필요함

데이터의 변화를 모니터링 하고 예상과 다른 움직임이 있다면, 행동으로 옮겨야 함

# 데이터에 근거한 의사결정

웹서비스의 KPI

| 약칭 | 정식 명칭 | 의미 |
| --- | --- | --- |
| DAU | Daily Active User | 서비스를 이용한 1일 유저 수 |
| 계속률 | Customer Retention | 서비스를 계속해서 이용하고 있는 유저의 비율 |
| ARPPU | Average Revenue Per Paid User | 유료 고객 1인당 평균 매출 |
| | | |

온라인 광고의 KPI

| 약칭 | 정식명칭 | 의미 |
| --- | --- | --- |
| CTR | Click Though Rate | 광고의 표시 횟수에 대한 클릭 비율 |
| CPC | Cost Per Click | 1회 클릭에 대해서 지불한 광고비 |
| CPA | Cost Per Acquisition | 1건의 고객 취득을 위해 지불된 광고비 |

### 데이터 기반 의사 결정

자신의 행동을 결정할 때 직감에 의지하는 것이 아니라 객관적인 데이터를 근거하여 판단하는것

# 수작업과 자동화해야 할 것의 경계를 판별하기

수작업으로 할 수 있는것은 수작업으로 하고, 자동화 하려는 경우에는 데이터 마트를 이용하기