## 문항 1

빅데이터와 전통적인 데이터를 구별하기 위해 몇 가지 특성이 사용되며, 그 특징들은 다음과 같습니다:

- 데이터의 크기
- 데이터 구성 방법
- 데이터 관리에 필요한 아키텍처
- 데이터가 파생되는 소스
- 데이터 분석에 사용되는 방법

**크기**

기존 데이터 세트는 일반적으로 기가바이트와 테라바이트 단위로 측정됩니다. 따라서, 서버 한 대에도 중앙 집중식 스토리지를 사용할 수 있습니다.

빅데이터는 크기뿐만 아니라 볼륨으로도 구분됩니다. 빅데이터는 일반적으로 페타바이트, 제타바이트 또는 엑사바이트 단위로 측정됩니다. 점점 더 커지는 빅데이터 세트의 규모는 현대적인 고용량 클라우드 기반 데이터 스토리지 솔루션에 대한 수요를 뒷받침하는 주요 요소 중 하나입니다.

**구성**

전통적인 데이터는 일반적으로 레코드, 파일 및 테이블로 구성된 정형 데이터입니다. 기존 데이터 세트의 필드는 관계형이므로 서로의 관계를 파악하고 그에 따라 데이터를 조작할 수 있습니다. SQL, Oracle DB 및 MySQL 과 같은 기존 데이터베이스는 사전 구성된 스태틱 스키마를 사용합니다.

빅데이터는 다이내믹 스키마를 사용합니다. 스토리지에서 빅데이터는 원시적(raw)이며 비정형입니다. 빅데이터에 접근하면 다이내믹 스키마가 원시 데이터에 적용됩니다. Cassandra 및 MongoDB 와 같은 최신 비관계형 또는 NoSQL 데이터베이스는 데이터를 파일에 저장하므로 비정형 데이터에 적합합니다.

**아키텍처**

전통적인 데이터는 일반적으로 중앙 집중식 아키텍처를 통해 관리되며, 이와 같은 아키텍처는 소규모의 정형화된 데이터 세트에 보다 비용 효율적이고 안전할 수 있습니다.

일반적으로, 중앙 집중식 시스템은 중앙 노드(예: 서버)에 연결된 하나 이상의 클라이언트 노드(예: 컴퓨터 또는 모바일 장치)로 구성됩니다. 중앙 서버는 네트워크를 제어하고 보안을 모니터링합니다.

빅데이터는 규모와 복잡성 때문에 중앙에서 관리할 수 없습니다. 따라서 분산 아키텍처를 필요로 합니다.

분산 시스템은 네트워크를 통해 여러 서버 또는 시스템을 연결하여 동일한 노드로 작동합니다. 아키텍처는 수평 확장이 가능하며(스케일 "아웃") 개별 노드에 장애가 발생하더라도 지속적으로 작동합니다. 분산 시스템은 상용 하드웨어를 활용하여 비용을 절감할 수 있습니다.

**출처**

전통적인 데이터는 일반적으로 ERP(전사적자원관리), CRM(고객관계관리), 온라인 트랜잭션 및 기타 엔터프라이즈 레벨 데이터에서 파생됩니다.

빅데이터는 소셜 미디어, 디바이스 및 센서 데이터, 시청각 데이터 등 다양한 엔터프라이즈 및 비엔터프라이즈 레벨 데이터에서 파생됩니다. 이러한 소스 유형은 동적이고 진화하며 매일매일 증가하고 있습니다.

비정형 데이터 소스에는 텍스트, 동영상, 이미지 및 오디오 파일도 포함될 수 있습니다. 전통적인 데이터베이스의 열과 행으로는 이러한 유형의 데이터를 활용할 수 없습니다. 점점 더 많은 양의 데이터가 비정형 구조를 띄고 있으며 여러 소스에서 제공되기 때문에 데이터에서 가치를 추출하려면 빅데이터 분석 방법이 필요합니다.

**분석**

전통적인 데이터 분석은 점진적으로 이뤄집니다. 이벤트가 발생하면 데이터가 생성되고, 이 데이터의 분석은 이벤트가 발생한 후에 수행됩니다. 전통적인 데이터 분석은 기업들이 정해진 기간 동안 특정 전략이나 변경 사항이 제한된 범위의 메트릭스에 미치는 영향을 이해하는 데 도움이 될 수 있습니다.

빅데이터 분석은 실시간으로 가능합니다. 빅데이터는 초 단위로 생성되므로 데이터가 수집되는 동안 분석할 수 있습니다. 빅데이터 분석은 기업의 요구사항과 전략에 대해 보다 동적이고 전체적인 이해를 제공합니다.

예를 들어, 기업이 직원을 위한 교육 프로그램에 투자했는데 그 효과를 측정하려고 한다고 가정해 보겠습니다.

전통적인 데이터 분석 모델에서는 기업이 세일즈와 같은 특정 운영 영역에 대한 교육 프로그램의 영향을 파악하고자 할 수 있습니다. 기업은 교육 전후의 판매량을 기록하고 관련 없는 요소는 제외합니다. 이론상으로는 교육의 결과로 매출이 얼마나 증가했는지 알 수 있습니다.

빅데이터 분석 모델을 활용하는 기업은 교육 프로그램이 특정 운영 영역에 어떤 영향을 주었는지에 대한 질문을 하지 않습니다. 대신, 전체 비즈니스에서 실시간으로 수집된 대량의 데이터를 분석하여 세일즈, 고객 서비스, 홍보 등 영향을 받은 특정 영역을 식별할 수 있습니다.

**문항 2 :**

데이터 정제

1. 데이터 전처리의 중요성

    1. 데이터 분석과정에서 데이터 전처리는 반드시 거쳐야 하는 과정
    2. 결과에 직접적인 영향을 끼치므로 전처리는 반복적으로 수행해야한다.
    3. 데이터 전처리는 데이터 정제 -> 결측값 처리 -> 이상값 처리 -> 분석 변수 처리

2. 데이터 정제의 개념 : 결측값을 채우거나 이상값을 제거하는 과정을 통해 데이터의 신뢰도를 높이는 작업.

3. 데이터 정제 절차 : 오류 원인 분석 -> 정제 대상 선정 -> 정제 방법 결정

    1. 오류 원인 원인
        1. 결측값 : 필수적인 데이터가 입력되지 않음
        2. 노이즈 : 실제는 입력되지 않았지만 입력된것으로 판단된 것.
        3. 이상값 : 데이터 범위에서 튀는값.(기준에서 많이 벗어난 값)
    2. 정제 대상 선정
        1. 모든 대상을 기준으로 선정하는 것이 기본
        2. 특히, 데이터의 품질을 떨어트리는 데이터를 더 많이 정제해야한다.
        3. 내부데이터 보단 외부데이터, 정형데이터 보단 비정형,반정형 데이터가 품질 저하 위협에 많이 노출
    3. 정제 방법 결정
        1. 삭제 : 무작위적인 삭제는 오류를 일으킬 수 있으므로 가급적으로 피한다.
        2. 대체 : 오류 데이터를 최빈값,중앙값,평균값 등으로 대체.
        3. 예측값삽입 : 회귀식등을 이용하여 예측값을 삽입

4. 데이터 정제 기술

    1. 일관성 유지를 위한 정제기술

1. 변환 : 다양한 형태로 표현된 값을 일관된 형태로 변환하는 작업
2. 파싱 : 정제 규칙을 적용하기 위해 유의미한 최소 단위로 분할하는 작업.
3. 보강 : 변환,파싱,수정,표준화 등을 통한 추가 정보를 반영하는 작업.

2. 데이터 정제 기술 : 분산 처리 기술을 기반으로 데이터를 정제 성능 보장을위해 인메모리 기반 컴퓨팅 기술사용
    1. ETL ; 수집 대상 데이터를 추출,가공,변환하여 DW,DM 에 저장하는 기술.
    2. MapReduce : 데이터를 추출하는 Map, 추출한 데이터를 중복이 없게 처리하는 Reduce 기술로 구성 (구글에서 대용량 데이터 세트를 처리하거나 생성하기 위한 목적으로 만든 SW 프레임워크)
    3. Spark.Storm : inMemory 방식의 데이터 처리 방식
        1. in memory : 디스크에 최적화된 데이터 베이스 보다 더 빠른 접근과 처리가 가능하도록 메인 메모리에 설치되어 운영되는 데이터 처리 방식을 말한다.
    4. CEP : 실시간으로 발생하는 이벤트 처리에 대한 결과값을 수집하고, 처리하는 기술
    5. Pig : 대용량 데이터 집합을 분석하기 위한 플랫폼
    6. Flumn : 로그 데이터를 수집하고 처리하는 기법, 실시간에 근접하게 데이터를 전처리,수집

5. 데이터 세분화
    1. 개념 : 데이터를 기준에 따라 나누고, 선택한 매개변수를 기반으로 유사한 데이터를 그룹화 효율적으로 사용함.
    2. 방법
        1. 계층적 방법 : 사전에 군집수를 정하지 않고 단계적으로 단계별 군집결과를 산출하는 방법
            1. 응집분석법 : 각 객체를 하나의 소집단으로 간주하고 단계적으로 유사한 소집단을 합쳐 새로운 소집단을 구성하는 방법
            2. 분할분석법 : 전체 집단으로 부터 시작하여 유사성이 떨어지는 객체들을 분리하는 방법.
        2. 비계층적방법 : 군집을 위한 소집단수를 정해놓고, 각 객체 중 하나의 소집단으로 배정하는 방법
            1. 인공신경망 : 생물학적 신경망에 영감을 얻어 통계학적 학습모델
            2. k-means : K 개의 소집단의 중심좌표를 정하여 각 객체와 중심간의 거리를 통해 군집을 나눔.

**데이터 결측값 처리**

1. 결측값의 종류
    1. 완전 무작위 결측 ( MCAR ) : 발생한 결측값이 다른 값들과 전혀 연관성이 없는 값
    2. 무작위 결측 (MAR) : 누락된 자료가 특정 변수와 관련되 일어나지만, 그 변수의 결과는 관계가 없는 경우. 누락이 전체 정보로 설명이 가능함.
    3. 비무작위결측 (MNAR) : 누락된 값이 다른 변수와 연관이 있는 경우
2. 결측값 처리 절차
    1. 결측값 식별 : 다양한 형태로 표현되어 있는 결측값의 현황을 파악한다.
    2. 결측값 부호화 : NA(기록되지 않은값), NULL(값이 없음), NaN(수학적으로 정의되지 않은 값), inf(무한)
    3. 결측값 대체 : 결측값의 자료형에 맞게 대체 알고리즘을 통해 결측값을 처리
3. 데이터 결측값 처리 방법
    1. 단순 대치법 : 결측값을 가진 자료 분석에 사용하기 쉽고, 통계적 추론에 사용된 통계량의 효율성 및 일치성 등의 문제를 부분적으로 보완해준다.
        1. 단순대치법의 종류
            1. 완전 분석법 : 불완전 자료는 모두 무시, 완전한 자료만 사용하여 분석, 분석은 쉽지만 효율성이 상실되고, 통계적 추론의 타당성 문제가 발생한다.
            2. 평균 대치법 : 자료의 평균값으로 결측값을 대치해서 불완전한 자료를 완전한 자료로 만드는 방법
            3. 단순 확률 대치법: 평균 대치법으로 추정된 통계량을 기반으로 확률적으로 값을 선정하여 대치
        2. 단순 확률 대치법의 종류
            1. 핫덱(Hot-Deck)대체 : 무응답 -> 비슷한 성향을 가진 응답자의 응답으로 대체, 표본조사에서 사용
            2. 콜드덱(Cold-Deck)대체 : 현재 진행중인 연구가 아닌 과거 혹은 외부의 연구에서 결과값을 참조
            3. 혼합방법
    2. 다중 대치법 : m 번의 대치를 통해 m 개의 가상적 완전한 자료를 만들어 분석하는 방법
        1. 다중 대치 적용 방식

1. 대치 : 결측자료의 예측분포 또는 사후분포에서 추출된 값으로 결측값을 대치, 베이지안 방법사용
2. 분석 : 같은 예측 분포로부터 D 개의 대치 표본을 구하여 원하는 분석을 각각 수행
3. 결합 : 모수 세타의 점추정과 표준오차의 추정치를 D 개 구한 후 이들을 결합하여 하나의 결과를 제시.

**데이터 이상값 처리**

1. 데이터 이상값 : 관측범위에서 많이 벗어난 아주 작은값 또는 아주 큰 값을 의미한다. 입력,데이터처리 오류 등의 이유로 특정 범위에서 벗어난 데이터값을 의미한다.
2. 데이터 이상값 발생원인
    1. 데이터 입력오류 : 전체 데이터 분포와 비교하여 쉽게 구분가능
    2. 측정오류
    3. 실험오류 : 실험조건이 동일하지 않아 오류
    4. 고의적인 이상값 : 자기 보고식에 의한 오류
    5. 표본추출에러 : 데이터 샘플링 과정에서 오류가 발생.
3. 데이터 이상값 검출 방법
    1. 개별 데이터 관찰 : 무작위 추출 혹은 전체 데이터 추이나 특이사항 관찰하여 이상값 검출.
    2. 통계값 : 통계(평균,중앙,최빈)값 과 데이터 분산도(범위,분산)을 활용한 이상값 검출
    3. 시각화 : 데이터 시각화를 통한 지표 확인으로 이상값 검출
    4. 머신 러닝 기법 : 데이터 군집화를 통한 이상값 검출
    5. 마할라노비스 거리 : 데이터 분포를 고려, 관측치가 평균으로 부터 벗어난 거리를 측정
    6. LOF : 관측치 주변의 밀도와 근접한 관측치 주변 밀도의 상대적인 비교를 통해 이상값을 탐색
    7. iForest : 관측치,관측치주변 밀도에 의존하지 않고 Decision Tree 를 이용하여 이상값 검출
4. 통계 기법을 이용한 이상값 검출
    1. ESD : 평균으로 부터 3 표준편차 떨어진 값을 이상값으로 판단한다.
    2. 기하평균을 활용 : 기하평균으로 부터 2.5 표준편차 떨어진 값을 이상값으로 판단.

3. 사분위 수를 활용 : 1,3 사분위를 기준으로 사분위간 범위(Q3-Q1)의 1.5 배 떨어진 값을 이상값으로 판단
4. 표준화점수(Z) : 평균 오메가와 표준편차가 알파인 정규 분포를 따르는 관측치들이 자료의 중심으로 부터 얼마나 떨어져 있는지를 기준으로 이상값을 검출
5. Q 검정 : 오름차순으로 정렬된 30 개 미만의 데이터의 범위에서 관측치간의 차이를 통해 이상값 여부 검증
6. T 검정 : 정규분포를 만족하는 단변량 자료에서 이상값을 검정하는 방법
7. 카이제곱검정 : 데이터가 정규분포를 만족하나, 자료의 수가 적을때 사용하는 방법

5. 시각화를 통한 이상값 검출
    1. 확률밀도함수
    2. 히스토그램
    3. 시계열차트
6. 머신러닝기법을 통한 이상값 검출
    1. 주어진 데이터를 K 개로 묶는 알고리즘, 각 클러스터와 거리 차이의 분산을 최소화 하는 방법
7. 마할라노비스 거리
    1. 데이터의 분포를 고려한 거리, 관측치가 평균으로 부터 얼마나 벗어났는가
    2. 이상값 탐색을 위해 고려되는 모든 변수간에 선형관계를 만족하고, 각 변수들이 정규분포를 따르는 경우 적용
8. LOF(Local Outlier Factor)
    1. LOF 는 관측치 주변의 밀도비교를 통해 이상값을 탐색
    2. 각 관측치에서 k 번째 근접이웃까지의 거리를 산출하여 해당 거리 안에 포함되는 관측치의 개수의 역수
9. iForest
    1. 관측치 사이의 거리 또는 밀도에 의존하지 않고, 데이터 마이닝 기법인 의사결정 나무를 이용하여 이상값 탐지.
    2. 데이터의 평균적인 관측치와 멀리 떨어진 관측치일수록 적은 횟수의 공간 분할을 통해 고립시킬 수 있다.
10. 데이터 이상값 처리 : 반드시 제거해야 하는 것이 아니므로 이상값을 처리할지는 분석의 목적에 따라 적절한 판단이 필요하다.
    1. 삭제 : 이상값으로 판단되는 결측값을 제외하고 분석, 이상값을 제거하기 위해 양극단값을 절단하기도 함 극단값 제거 방법보다 극단값 조절 방법을 활용하는 것이 적절하다

2. 대체법 : 하한,상한값을 결정한 후 이를 기준으로 더 아래,혹은 위일경우 하한,상한으로 대체하는 방법

3. 변환법 : 극단적인 값으로 인해 이상값이 발생했다면, 자연로그를 취해서 값을 감소

11. 박스플롯을 이용한 이상값 제거

1. 박스플롯 구성요소 : 1~3 사분위수, 최소,최대값, 하위경계(Q1-1.5IQR),최소값(하위경계 최근사값),최대값(상위경계의 최근사값),상위경계(Q#*1.5IQR), 수염(Q1,Q3 로부터 IQR 의 1.5 배 내에 있는 가장 멀리 떨어진 데이터까지 이은선) 이상값(수염보다 밖에 있는 값)