**데이터 분석과 빅데이터 분석의 차이점에 4가지에 대해서 설명하시오.**

빅데이터는 대규모 혹은 복잡한 데이터 세트와 이러한 유형의 데이터를 처리하는 데 사용되는 방법을 의미합니다. 빅데이터는 "V4"로 알려진 4가지 주요 특성이 있습니다.

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 MySQL과 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 MongoDB와 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.

## 항2. 데이터 전처리 기술 3가지가 무엇인지 명칭을 작성하고, 그 개념에 대해 설명하시오.(필요시, 사용되는 세부적인 기술을 설명하시오.)

- 데이터 정제 (cleansing): 결측 값(missing value; 빠진 데이터)들을 채워 넣고, 이상치를 식별 또는 제거하고, 잡음 섞인 데이터를 평활화 (smoothing)하여 데이터의 불일치성을 교정하는 기술
- 데이터 변환(transformation): 데이터 유형 변환 등 데이터 분석이 쉬운 형태로 변환하는 기술. 정규화 (normalization), 집합화(aggregation), 요약(summarization), 계층 생성 등의 방법을 활용한다.
- 데이터 필터링(filtering): 오류 발견, 보정, 삭제 및 중복성 확인 등의 과정을 통해 데이터의 품질을 향상하는 기술
- 데이터 통합(integration): 데이터 분석이 용이하도록 유사 데이터 및 연계가 필요한 데이터들을 통합하는 기술
- 데이터 축소(reduction): 분석 시간을 단축할 수 있도록 데이터 분석에 활용되지 않는 항목 등을 제거하는 기술