문제1.

1) 크기

기존 데이터 세트는 일반적으로 기가와 테라 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.

2) 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 스테틱 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.

3)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.

4) 출처

전통적인 데이터는 일반적으로 ERP, 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 자료들은 동적이고 변화하며 매일매일 증가하고 있습니다.

비정형 데이터 소스에는 텍스트, 동영상, 이미지 및 오디오 파일도 포함될 수 있습니다.

기존의 데이터베이스의 구조로는 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.

문제 2.

1. 데이터 유형변환 - 데이터 유형을 변환하거나 데이터 분석에 용이한 형태로 변환

2. 데이터 여과 - 오류발견, 보정, 삭제 및 중복성 확인 등 데이터 품질 향상

3. 데이터 정제 - 결측치 변환, 이상치 제거, 노이즈 데이터 교정 비정형 데이터를 수집할 때 반드시 수행