문항1. 데이터 분석과 빅데이터 분석의 차이점에 4가지에 대해서 설명하시오.

1. 사후처리와 사후처리
정보관리 시스템을 이용해 데이터 분석을 수행하기 위해서 데이터에 대해서 사전 처 리가 필요했습니다. 여기서 사전처리란 필요한 정보만 추려내고 필요 없는 정보는 제거해서 데이터를 축소하는 것을 말합니다. 오늘날에는 빅데이터 처리기술이 등장 함에 따라 사전처리보다 사후처리가 더욱 중요해졌습니다. 데이터를 저장하여 모아 놓고 데이터에서 직접 의미를 찾아내기 위해서 무엇을 해야 할까요? 데이터 유형에 따라 저장방법을 결정하는 저장관리, 신뢰성 있는 데이터를 확보하는 품질관리, 데 이터의 유출방지와 안전한 사용을 보장하는 보안관리 등의 역할의 수행이 중요해진 것입니다.
2. 표본조사와 전수조사
통계분석과 같은 전통적인 데이터 분석법도 표본 데이터에서 의미를 찾는 것에 초점 을 두고 발전해왔습니다. 왜냐하면 수집, 처리 비용, 분석 능력 등의 부담이 존재했 기 때문입니다. 하지만 이제 조사대상이 되는 자료 전체를 빠짐없이 조사하는, 전수 조사 가능해졌습니다. 전수조사의 장점은 표본조사가 주지 못하는 패턴과 같은 정보 를 제공한다는 점입니다.
3. 양과 질
세번째는 질과 양입니다. 데이터의 질보다 양이 더 강조되는 시대입니다. 빅데이터 란 말에도 나타나듯이 데이터의 양이 크지 않다면 빅데이터가 아닐 것입니다. 실시 간으로 분석할 수 있는 데이터의 양이 증가하면 사소한 몇몇 오류는 분석결과에 큰 영향을 미치지 않습니다.분석할 수 있는 데이터가 많으면 많을수록 결과의 정확성은 증가하게 됩니다.
4. 인과관계와 상관관계
기존의 데이터 분석은 이론적인 틀과 정해진 목적에 따라서만 진행되었습니다. 하지 만 비즈니스의 상황에서 중요한 것은 인과관계보다도 상관관계 파악입니다. 빅데이 터 분석을 통해 다양한 상관관계를 빠르게 저렴하게 도출할 수 있게 되었다고 말할 수 있습니다

문항2. 데이터 전처리 기술 3가지가 무엇인지 명칭을 작성하고, 그 개념에 대해 설명하시오.(필요시, 사용되는 세부적인 기술을 설명하시오.)

1. 데이터 저장방식

(1) 관계형 데이터베이스 관리시스템

데이터를 테이블 형식으로 저장하며 많은 데이터를 처리할 수 있습니다. 또한 5가지 성질은 원자성(Atomocity), 일관성(Consistency), 고립성(Isolation), 지속성(Durability)을 보장합니다. 관계형 데이터베이스 관리시스템의 한계점은 시스템 이용 불가시간이 발생한다는 점과 스케일 아웃의 한계가 발생한다는 점입니다.

(2) 비관계형 데이터베이스 관리시스템

비관계형 데이터베이스 관리시스템의 기술적 특징은 4가지입니다. 첫번째 노 스키마 입니다. 데이터를 모델링하는 고정된 데이터 스키마가 없이 키 값을 이용하여 다양한 형태의 데이터 저장 및 접근이 가능합니다. 다양한 데이터 저장 방식이란 열, 값, 문 서, 그래프 등의 이용하는 것입니다. 두번째 탄력성입니다. 탄력성이라는 시스템에 일부에 장애가 발생하더라도 클라이언트가 시스템에 접근이 가능합니다. 응용시스템 의 다운 타임이 없도록 동시에 대용량 데이터의 생성 및 갱신이 가능합니다. 또한 질 의에 대응할 수 있도록 시스템 규모와 성능 확장이 용이하며 입출력의 부하를 분산시 키는데 용이한 구조에 해당합니다.

세번째 질의입니다. 수십 대에서 수천 대 규모로 구성된 시스템에서도 데이터의 특 성에 맞게 효율적으로 데이터를 검색하고 처리할 수 있는 질의언어 관련 처리 기술 과 API를 제공합니다. 네번째 캐싱입니다. 대규모 질의에도 고성능 응답속도를 제 공할 수 있는 메모리 기반 캐싱 기술을 적용하는 것이 중요합니다.

(3) 분산파일 시스템

분산 파일 시스템은 막대한 양의 데이터를 저장하고 관리하기 위해 수많은 서버들에 데이터를 나누어 저장하고 관리하는 파일 시스템입니다. 빠른 처리 성능과 수백 페라바이트 이상의 데이터 저장을 지원하고 쉽게 시스템을 확장할 수 있습니다. 시스 템 장애에도 계속해서 안전하게 서비스를 제공할 신뢰성, 가용성을 확보합니다. 저 장소 성능 향상을 위한 여러 노드를 활용하여 용량과 속도를 늘리는 기능이 필요로 되어집니다. 분산 파일 시스템의 대표적인 예로는 구글 파일시스템과 하둡 분산 파 일시스템 등이 해당합니다.

2. 데이터 품질관리와 빅데이터 품질관리

➢ 정형데이터 품질 기준 5가지

① 완전성

개별 완전성과 조건 완전성이라는 세부기준을 포괄합니다. 완전성은 필수항목에 누락이 없다는 것을 의미합니다.

② 유일성

단독 또는 조건 유일성이라는 세부기준을 포괄합니다. 유일성은 데이터 항목은 유일 해야 하며 중복되어서는 안 된다는 것을 의미합니다.

③ 유효성

범위, 날짜, 형식의 유효성이라는 세부기준을 포괄합니다. 유효성은 데이터 항목은 정해진 데이터 유효범위 및 도메인을 충족해야 한다는 것을 의미합니다.

④ 일관성

기준코드 일관성, 참조 무결성, 데이터 흐름 일관성, 칼럼 일관성이라는 세부기준을 포괄합니다. 일관성은 데이터가 지켜야 할 구조, 값, 표현되는 형 태가 일관되게 정 의되고, 서로 일치해야 한다는 것을 의미합니다.

⑤ 정확성

선후 관계 정확성, 계산/ 집계 정확성, 최신성, 업무규칙 정확성이라는 세부기준을 포괄합니다. 실세계에 존재하는 객체의 표현 값이 정확히 반영되어야 한다는 것을 의미합니다.

➢ 비정형데이터 품질기준 5가지

① 기능성

적절성, 정확성, 상호 운 용성, 기능 순응성이라는 세부기준을 포괄합니다. 기능성 은 해당 콘텐츠가 특정 조건에서 사용될 때, 명시된 요구와 내재된 요구를 만족하는 기능을 제공하는 정도를 말합니다.

② 신뢰성

성숙성, 신뢰 순응성이라는 세부 기준을 포괄합니다. 신뢰성은 해당 콘텐츠가 규정 된 조건에서 사용될 때 규정 된 신뢰 수준을 유지하거나 사용자로 하여금 오 류를 방지할 수 있도록 하는 정도를 말합니다.

③ 사용성

이해성, 친밀성, 사용 순응성이라는 세부기준을 포괄합니다. 사용성은 해당 콘텐츠 가 규정된 조건에서 사용될 때, 사용자에 의해 이해되고, 선호될 수 있게 하는 정도 를 말합니다.

④ 효율성

시간 효율성, 자원 효율성, 효율 순응성이라는 세부기준을 포괄합니다. 효율성은 해 당 콘텐츠가 규정된 조건에서 사용되는 자원의 양에 따라 요구된 성능을 제공하는 정도입니다.

⑤ 이식성

적응성, 공존성, 이식 순응성이라는 세부기준을 포괄합니다. 이식성은 해당 콘텐츠 가 다양한 환경과 상황에서 실행될 가능성을 말합니다.

3. 데이터 보안관리 적용 기술

① 사용자 인증

누가 어떤 데이터에 어떤 조치를 취할 수 있는가를 미리 정한 바에 따라, 데이터나 데이터 관리 시스템에 접근하는 사람의 접근 자격을 확인하는 것입니다. 대표적으로 아이디/패스워드 방식, 일회용 패스워드 방식, 통합 사용자 인증 방식 등이 있습니 다.

② 접근 제어

어떤 주체가 어떤 객체를 읽거나, 객체에 기록하거나, 객체를 실행시키고자 할 때, 해당 주체가 객체에 대한 권한을 보유하고 있는지를 체크하고 통제하는 것입니다. 대표적으로 강제 접근제어, 임의 접근제어, 역할 기반 접근제어 등이 있습니다.

③ 암호화

평문을 해독 불가능한 형태로 변형하거나 또는 암호화된 암호문을 해독 가능한 형태 로 변형하기 위한 원리와 방법 등을 취급하는 기술입니다.

④ 개인 정보 비식별화

수집된 데이터에 포함된 개인 정보의 일부 또는 전부를 삭제하거나 다른 정보로 대 체 또는 다른 정보와 결합하여 특정 개인을 식별하기 어렵도록 하는 일련의 조치에 해당합니다.

⑤ 개인 정보 암호화

데이터베이스 전체에 대한 보호가 아닌 개인 정보가 포함된 특정 필드에 대한 보호 기술입니다. 개인 정보의 암호화 저장 후 데이터베이스에 저장된 개인 정보의 정상 적인 이용을 위해 데이터베이스를 안전하고 효율적으로 인덱싱하는 기술이라고 할 수 있습니다.