06_빅데이터 핵심 기술-생산, 수집, 관리, 분석

#1
1. 데이터의 생산
가. 데이터의 입력으로 생산하는 방식
데이터를 입력해 생산하는 방식에는 정형 데이터를 이용하는 방식과 비정형 데이터를 이용하는 방식이 있습니다.

- 정형 데이터: 조직 내에서 자체적으로 개발한 업무용 애플리케이션을 통해 생산합니다.
- 비정형 데이터: SNS나 블로그 등과 같이 개인이 생산합니다.

#2
나. 데이터의 자동 생산 방식
① 사물인터넷 플랫폼 서비스
사물인터넷 플랫폼은 서비스에 필요한 공통 요구 기능을 포함하고 있으며, 개별 사물과 서비스에서 독립적으로 동작할 수 있습니다. 서버나 클라우드 형태로 제공되거나 디바이스에 직접 위치할 수도 있습니다.

② 사물인터넷 플랫폼 기술
사물인터넷 플랫폼 기술은 센서나 액추에이터 또는 데이터 중심으로 인터넷 기반 서비스 도메인 수준의 규모를 지원해 왔습니다. 기술이 발전함에 따라 데이터, 프로세스, 지능 중심으로 바뀌어 수백억 개 이상의 글로벌 규모로 발전했습니다. 사물인터넷 플랫폼에 대한 표준은 사물인터넷 글로벌 협의체를 기반으로 하여 개방형으로 운영하고 있습니다.

#3
다. 환경에 따른 데이터 분류
① 조직 내에서 데이터가 생산되는 환경
데이터의 생산에서부터 완료에 이를 때까지 적합성 보장을 위한 트랜잭션 관리가 필요합니다. 특정 사용자의 정보를 수정할 때는 수정이 완료되기 전까지 과거의 정보가 일관되게 표현되어야 합니다. 작업 완료가 확정되면 새로운 정보가 나타나게 하는 것이 원칙입니다. 데이터가 업무 목적에 맞게 사용되어야 하므로 정형화된 형태를 띠게 됩니다. 이러한 환경에서는 관계형 데이터베이스

가 데이터 생산에 중요한 기술로 사용됩니다. 락(Lock 또는 Enqueue)과 같은 오브젝트로 트랜잭션을 보장할 수 있기 때문입니다.

② 개인이 생산하는 데이터
개인이 생산하는 데이터는 제약 없이 자유로운 형태를 띠고 있습니다. SNS나 블로그와 같은 웹을 통해 생산된 데이터는 여기저기 분산되어 있는 특징이 있습니다.

#4
2. 데이터의 수집
가. 데이터 수집 방법
데이터를 수집하는 이유는 업무에 활용하기 위함입니다. 데이터의 수집 방법은 외부 데이터를 수집하거나 내부에서 데이터를 직접 생산하는 방법이 있습니다.

① 전통적인 생산 방식
전통적인 데이터베이스 환경에서는 프론트엔드 애플리케이션을 활용하여 필요한 데이터를 직접 생산했다면, 점차 데이터를 외부에서 수집하는 방식으로 바뀌고 있습니다. 정형 데이터를 자체 생산하는 방식에서 외부의 비정형 데이터를 수집하는 방식으로 초점의 방향이 전환된 것입니다.

② 외부 데이터 생산
외부 데이터는 장비에서 기계적으로 생성되는 데이터가 많아 그 양이 방대합니다. 따라서 데이터의 수집 단계에서는 시스템의 성능이 중요합니다. 또 외부 데이터는 소스가 다양하기 때문에 다양한 소스의 데이터를 모두 수집하기 위해서는 여러 수집 방식을 사용해야 합니다. 정형 데이터뿐만 아니라 비정형 데이터를 수집하여 분석 목적에 적합하게 저장하는 기술이 필요합니다.

#5
나. 수집 방법 및 기술
데이터의 형태와 종류에 구애받지 않는 기술과 수집 방법이 필요합니다. 더불어 이와 같은 방식으로 수집한 데이터를 가공하고 재생산할 수 있어야 합니다. 수집 방법에는 데이터를 가공하고 재생산하는 과정이 포함됩니다. 서비스는 가시적인 형태로 제공할 수 있어야 하며, 데이터를 효율적으로 사용하기 위한 저장 방법이 필요합니다. 또 정형 데이터와 비정형 데이터의 형태에 따라 데이터

를 저장하는 기능이 제공되어야 합니다.

#6
① EAI(Enterprise Application Integration)
EAI는 기업 내·외부의 서로 다른 시스템을 통합하기 위해 사용하는 기법입니다. 여러 시스템에 분산되어 있는 정보들이 기업 내부의 의사결정이 필요할 때 빠르게 공유되어 합리적인 방안을 수립할 수 있습니다.

② ESB(Enterprise Service Bus)
ESB는 비즈니스 프로세스의 환경에 맞게 설계하고 또 전개할 수 있는 아키텍처를 제공합니다. 다시 말하면 애플리케이션 서비스들을 컴포넌트화된 논리적 집합으로 묶는 핵심 미들웨어입니다.

③ API(Application Programming Interface) 플랫폼
데이터 소유 주체가 웹 개발자나 사용자를 위해 정보와 데이터를 정해진 방식으로 공개하는 기술을 오픈 API라고 합니다. HTTP 프로토콜을 기반으로 하는 웹 서비스이며, SOAP(Simple Object Access Protocol)과 REST를 기반으로 하는 웹 서비스 기술을 제공합니다.

#7
3. 데이터의 관리
가. 지능화 및 자동화되는 데이터베이스 관리
① 데이터의 중요성
4차 산업 시대의 대표적인 키워드에는 모두 데이터가 포함되어 있습니다.
- ABCD(AI, Blockchain, Cloud, Data)
- ICBM(사물인터넷, Cloud, Big Data, Mobile)

데이터 기반의 실시간 의사결정은 불확실성이 커지는 비즈니스 환경에서 설득력이 있는 통찰력을 가질 수 있게 합니다. 하지만 기업이 데이터를 관리하는 환경은 전통적인 RDB(Relational DataBase)뿐만 아니라 하둡, NoSQL을 모두 사용하는 환경으로 복잡해지고 다양화되고 있습니다. 인프라 환경 또한 온프레미스와 클라우드를 혼용하는 경우가 많아지고 있습니다. 따라서 데이터베이스 관리자는 기존의 유지 관리 업무와 함께 정보의 유형, 중요성, 규제 대상의 여부, 가용성의 수준, 관리 비용 등을 고려할 수 있어야 합니다.

#8
② 데이터 관리의 매커니즘 변화
데이터 관리를 효과적으로 대응하기 위해 데이터베이스 시장도 지속적으로 변화하고 있는데, 대표적인 것이 바로 클라우드 환경으로의 전환입니다. AWS, MS, 오라클(Oracle)과 같은 클라우드 사업자는 클라우드상에서 안정적으로 운영될 수 있도록 확보된 데이터 서비스를 제공하기 위해 사용자의 개입을 최소화합니다. 또 데이터베이스 관리의 기능을 자동화하여 서비스를 제공합니다.

#9
나. 데이터 거버넌스의 확장
데이터 거버넌스란, 전사적으로 보유하고 있는 데이터에 대한 관리 체계를 의미합니다. 데이터에 대한 관리나 정책, 지침, 표준, 전략 및 방향 수립을 포함하며, 데이터를 관리할 수 있는 조직이나 서비스의 정의도 포함하는 의미입니다.

데이터 거버넌스는 고품질의 데이터를 확보하고 적극적인 활용을 통해 조직의 가치 창출에 지속적으로 기여하는 것을 목표로 합니다. 데이터를 통해 위험을 예측하고, 관리 비용을 최적화하며, 데이터의 활용이 촉진됨으로써 데이터의 가치가 향상됩니다. 이는 곧 비즈니스 목적에 부합하는 서비스가 지속될 수 있는 힘을 가지게 합니다.

#10
질문자: 조직 내에 거버넌스는 왜 중요한가요?
전문가: 조직 내에 거버넌스가 확립되지 못하면 품질이 낮은 데이터를 사용하게 되고 이에 따라 오류가 생성되거나 규제에 직면할 수 있습니다. 또 개인정보 관련 데이터가 유출되는 사태가 발생하여 한순간에 고객의 신뢰를 잃을 수도 있습니다.

#11
다. 데이터 거버넌스의 구현
효과적인 데이터 거버넌스의 구현을 위해서는 데이터의 생성부터 폐기까지 관리되어야 합니다. 대부분의 조직은 데이터 거버넌스의 구현을 위해 다양한 관리 시스템을 활용하지만 시스템 간의 프로세스가 미흡하거나 일부 영역이 통

합되지 않은 상태로 운영되고 있습니다. 단위 데이터 거버넌스 프로세스의 연계, 빅데이터 환경에 적합한 다양한 데이터 소스의 수집과 자동화는 이러한 제약을 극복하기 위한 진화의 방향입니다.

#12
라. 데이터 거버넌스의 기능
각각의 단위 거버넌스 프로세스와 관리 기능은 서로 연계될 때 더욱 효과적으로 사용할 수 있습니다. 예를 들어, 데이터의 품질 관리를 통해 오류가 예상되는 데이터 항목을 식별했다면 다음에는 메타데이터를 통해 해당 정보를 확인하고 오류 여부를 판단하는 것이 가능합니다.

또 오류 데이터가 발견되었다면 데이터 리니지를 확인하여 해당 데이터가 생성된 지점이나 다른 데이터와의 연관 관계를 파악할 수 있으며, 사용 프로그램 및 수정 현황 등을 추적할 수 있습니다. 이 과정에서 사용자는 연계된 프로세스와 기능을 통해 데이터를 관리할 수 있습니다.

#13
4. 데이터의 분석
가. 데이터의 분석 기술
데이터 분석 기술은 크게 전체 데이터 분석의 일부분인 단순 집계와, 전문적인 도구가 필요한 고급 분석으로 구분됩니다. 단순 집계는 중요성이 매우 높아 고급 분석에도 단순 집계가 전후 단계에서 함께 수행되면서 시너지 효과를 내고 있습니다. 많은 소프트웨어 도구들이 신축적인 단순 집계가 가능하도록 발전하고 있으며, 융합이 더욱 손쉽게 가능하도록 고급 분석을 지원하는 기능들이 점점 더 다양하게 제공되고 있습니다.

#14
① 단순 집계
단순 집계는 다양한 방식으로 가능합니다. SQL과 같은 전통적 수단이나 다차원 분석, 최근에 등장한 데이터 시각화(Data Visualization) 도구 등의 방식이 사용됩니다. 흔히 이 영역을 비즈니스 인텔리전스라고 부르며, 분석에 대한 별도의 전문 지식이 없더라도 일반 사용자들이 쉽게 수행할 수 있어 널리 사용되어 왔습니다. 또 결과를 직관적으로 이해할 수 있다는 장점이 있습니다.
② 고급 분석

고급 분석은 통계적 분석과 머신 러닝을 포함합니다. 전문적인 도구가 필요하고 보급 자체가 한정적이며 유용성과 활용 방법 부분에서도 한정적이어서 본격적인 활용이 지연되어 왔습니다. 빅데이터의 개념이 전파되면서 머신 러닝과 인공지능에 대한 일반인들의 이해가 높아졌고 오픈소스 소프트웨어가 보편화되기 시작했습니다. 이러한 발달의 과정은 고급 분석이 도입되고 실무 활용이 본격화되는 기반이 되었습니다.

#15
나. 데이터 분석 기술 관련 동향
데이터 분석 결과물을 만드는 것은 분석 외에도 데이터의 수집이나 저장, 전송, 변환 등과 같은 사전 작업이 필요합니다. 또 보고서 작성이나 중간 분석 결과를 저장하고 공유하며 사후 업무들도 있어 다양한 사항들이 요구됩니다. 그런 이유로 이를 종합적으로 지원할 수 있는 통합적 소프트웨어 도구와 방법론들이 발전하는 추세입니다.

데이터 분석 과정에서는 인력과 시간이 부족합니다. 이를 해결하기 위해서는 데이터 분석의 생산성 향상과 분석 결과의 품질을 높여야 합니다. 따라서 이러한 문제점을 해결하기 위한 데이터 분석 기술의 키워드는 크게 ‘통합’과 ‘자동화’로 요약될 수 있습니다.