LG전자 H&A본부 - 한양대

DX Intensive Course

데이터 베이스 설계와 운용

차 경 진 교수

LG전자

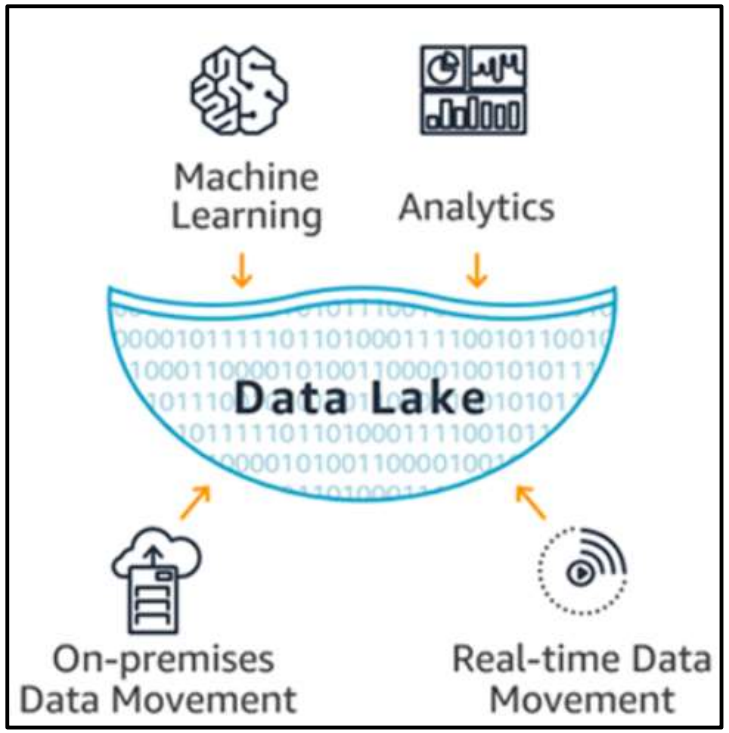
HANYANG UNIVERSITY

1. Data Lake

데이터 레이크 (Data Lake) 란?

✓ 데이터 레이크는 정형/비정형 데이터 종류와 모델에 상관없이 모든 유형의 데이터를 저장할 수 있는 중앙집중식 저장소.

✓ 전통적인 엔터프라이즈 IT 환경의 데이터 웨어 하우스는 구조적 정형 데이터만 저장했지만, 데이터 레이크는 비정형 데이터를 포함한 모든 데이터를 저장.



1. Data Lake

데이터 레이크 특징

✓ 데이터 분석을 위한 유연한 데이터 환경을 제공.

✓ 데이터 레이크는 여러 종류의 데이터를 담을 수 있는 저장소 역할 ➔ 인공지능과 머신러닝을 활용하기 위해서는 모든 데이터를 한 곳에 모으는 작업에 유용.

✓ 데이터 레이크를 구축한 조직은 데이터로 더 많은 가치를 창출할 수 있으며, 구축하지 않은 조직보다 매출 성장이 9% 앞선다고 함.

데이터 레이크 구축

✓ 데이터 레이크 구축은 어려움 ➔ 여러 개의 컴포넌트, 프레임 워크, 서로 다른 벤더의 여러 제품이 통합된 결과물이기에.

✓ 따라서, 철저한 사전 계획과 올바른 아키텍처를 가지고 데이터 레이크 설계 해야함.

✓ 그렇지 않을 경우, 1) 거버넌스와 보안이 취약 2) 데이터가 중복되고 퀄리티가 낮아짐 3) 데이터를 찾을 수 없게 됨.

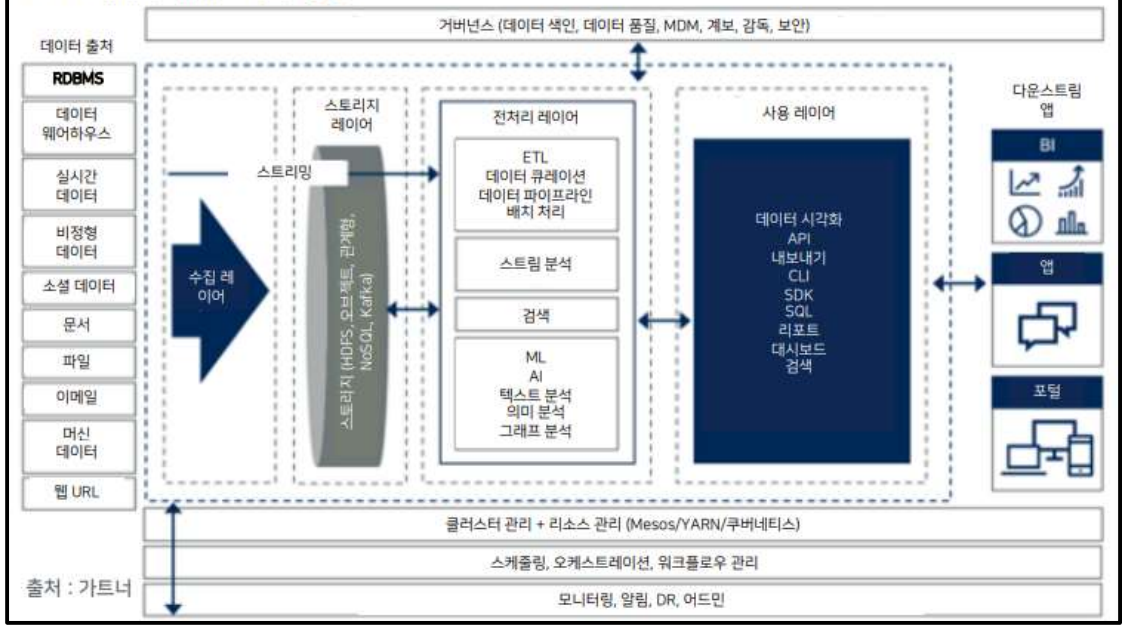
➔ 즉, 기업 내 중요 자산인 데이터를 충분히 활용할 수 없게 됨.

1-1. 아키텍처

✓ 하나의 데이터 레이크에 다 모을 것인가? 여러 개로 나눌 것인가?

✓ 데이터 레이크를 위해 꼭 필요한 핵심 기능과 컴포넌트를 포함한 전반적인 아키텍처는 표 1 과 같음.

표 1. 데이터레이크 아키텍처



데이터 레이크 설계 전에 기업이 고려해야 할 사항

■ 전체 조직을 위한 하나의 중앙화 된 데이터 레이크를 구축할 것인가?

■ 조직의 각 데이터 센터를 위한 세분화된 데이터 레이크를 여러 개 구축할 것인가?

■ 온프레미스 기반의 데이터 레이크를 구축할 것인가, 클라우드 기반의 데이터 레이크를 구축할 것인가, 또는 하이브리드 및 멀티 클라우드 기반의 데이터 레이크를 구축할 것인가?

1-1. 아키텍처

■ 중앙화 된 데이터 레이크의 장점.

✓ 하나의 중앙화 된 데이터 레이크는 관리하고 통제하기 쉬움.

✓ 모든 데이터가 한 장소에 모여 있으므로, 데이터 접근이 쉬우며 네트워크 대역폭 관련 제한을 다루기 용이.

✓ 하나의 통일된 데이터 소스를 확보할 수 있음.

✓ 비용을 통제할 수 있음.

표 2. 하나의 중앙화된 데이터레이크

■ 중앙화 된 데이터 레이크는 대규모 글로벌 기업에 적합하지 않을 수 있는 이유.

✓ 데이터 위치, 이동, 접근에 대한 법률적/주권적인 부분이 충돌할 수 있으며, 이 경우 사용성이 악화될 수 있음.

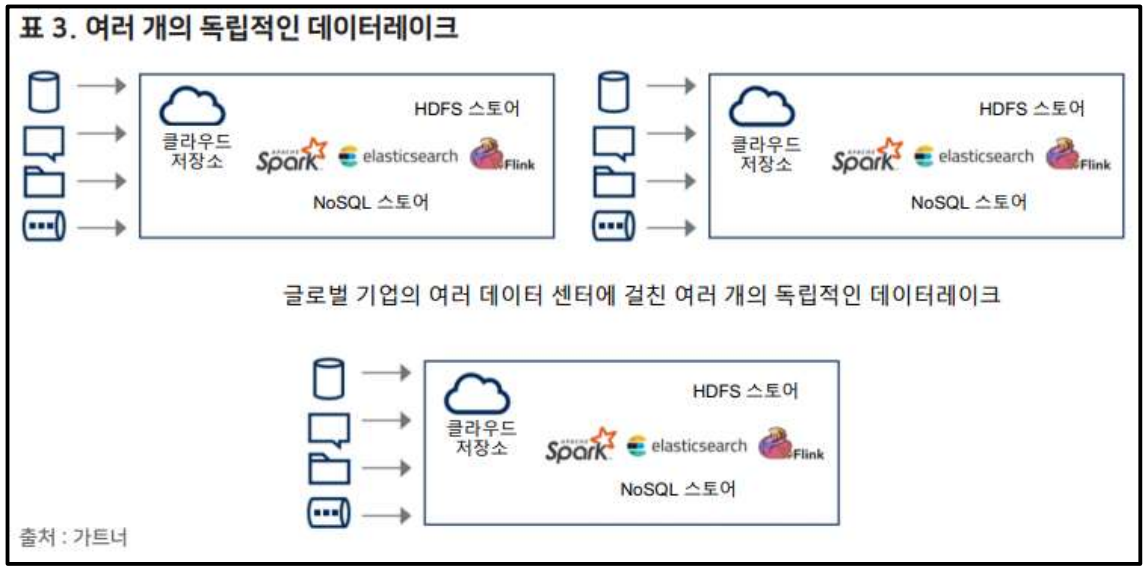
✓ 모든 것을 한 곳에 모으기 위해서는 매우 높은 네트워크 용량과 속도가 필요함.

✓ 한 곳에서 장애가 나면 전체 시스템에 영향을 줍니다. 즉 단일 장애 지점(Single point of failure)이 될 위험이 있음. 보안 리스크도 그만큼 높음.

1-1. 아키텍처

■ 대안 : 기업 내에 여러 개의 독립적인 데이터 레이크를 구축하는 방법.

표 3. 여러 개의 독립적인 데이터레이크



출처 : 가트너

해당 아키텍처의 장/단점

■ 장점

- ✓ 네트워크를 통해 하나의 위치로 이동하기 어려운 대용량 데이터의 경우, 로컬 데이터 레이크에 저장할 수 있음.
- ✓ 지역별, 국가별 규제에 맞게 요구사항을 맞출 수 있음.

■ 단점

- ✓ 이러한 접근법은 데이터 레이크 사일로(Silos)를 만듦.
- ✓ 데이터 거버넌스 정책과 툴에 일관성이 없어질 수 있음.
- ✓ 여러 개의 데이터 레이크에 저장된 공통 데이터를 동기화하기가 매우 어려움.
- ✓ 성격이 다른 여러 개의 기술이 각각의 데이터 레이크에 사용될 수 있음.
- ✓ 데이터 이동을 위한 네트워크 대역폭 요구사항 때문에 비용이 높아질 수 있음.
- ✓ 데이터 접근이 어려울 수 있음.

1-2. 수집

✓ 한 번에 불러올 것인가? 실시간으로 불러올 것인가?

✓ 데이터 레이크가 데이터 분석을 위해 쓰이기 위해서는, 여러 개의 데이터 출처로부터 데이터를 불러올 수 있어야 함.

✓ 데이터 레이크의 수집(Ingestion) 레이어가 이것을 가능하게 함.

✓ 수집 레이어는 출처가 되는 데이터 시스템에서 데이터를 추출하며, 데이터를 불러오는 서로 다른 전략들을 조율하는 플랫폼임.

✓ 표 5는 데이터를 데이터 레이크로 불러올 수 있는 여러가지 방법을 보여줌.

표 5. 데이터 불러오기 기술

출처 : 가트너

1-3. 스토리지

- ✓ 분산 파일 시스템과 오브젝트 스토리지, 각각의 장단점
- ✓ 데이터 수집이 끝나면 데이터를 저장해야 함. 데이터 스토리지는 데이터 레이크에서 가장 중요한 요소.
- ✓ 데이터 스토리지 종류와 아키텍처는 데이터의 성능, 확장성, 접근성을 결정함.
- ✓ 기업의 데이터 종류가 많고 복잡하며, 용량도 계속 늘어나기 때문에, 모든 스토리지 요구조건을 만족하는 하나의 스토리지 플랫폼은 존재하지 않음.

데이터 레이크가 가지는 일반적인 스토리지

- ✓ NoSQL 데이터 스토리지
- ✓ 관계형 데이터 스토리지
- ✓ 스트리밍 메세지/이벤트 스토리지
- ✓ New SQL 데이터 스토리지
- ✓ 인메모리 데이터 스토리지
- ✓ 분산 파일 시스템
- ✓ 오브젝트 데이터 스토리지

1-3. 스토리지

■ 스토리지 특징

- ✓ 사용 용도와 입력 및 출력 패턴에 따라, 기존의 스토리지를 사용할 수도 있고 데이터를 NoSQL 데이터 스토리지 같은 다른 형식의 스토리지로 이동할 수도 있음.
- ✓ 대부분의 조직은 수집과 전처리 이후의 데이터를 저장하기 위해 하둡 분산 파일 시스템(HDFS)을 사용함.
- ✓ 온프레미스 혹은 클라우드에서 사용하며, 클라우드 내 오브젝트 스토리지에서 사용하기도 함.
- ✓ 아마존 심플 스토리지 서비스(S3)나 구글 클라우드 스토리지와 같은 오브젝트 스토리지는 빠르게 HDFS를 대체하고 있음.
- ✓ 데이터 레이크 아키텍트는 파일 시스템을 오브젝트 스토리지 로 바꾸는 것의 함의를 이해해야 함.
- ✓ 오브젝트 스토리지는 파일 시스템과 다른 아키텍처를 가지고 있음.
- ✓ 초기 HDFS 위주로 구축된 패턴은 오브젝트 스토리지로 이동했을 때 일대일로 매칭되지 않을 수도 있음.

Thank you!

For More Information

차경진 교수 | 한양대학교 경영대학
kjcha7@hanyang.ac.kr/ 02-2220-1038