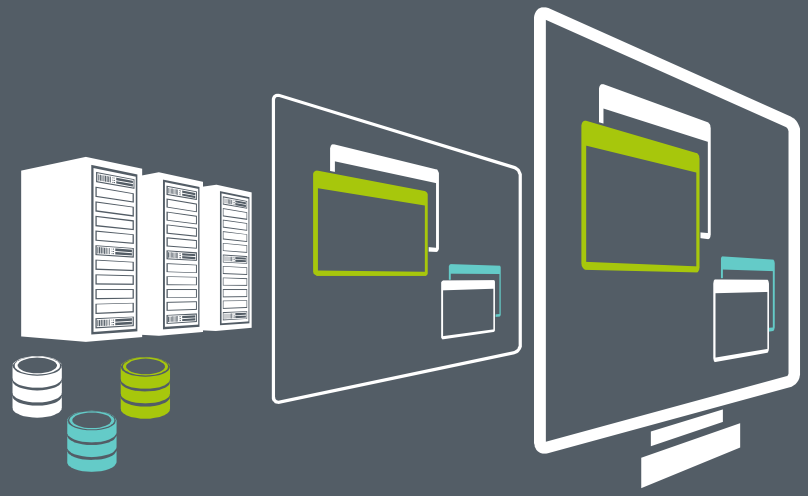
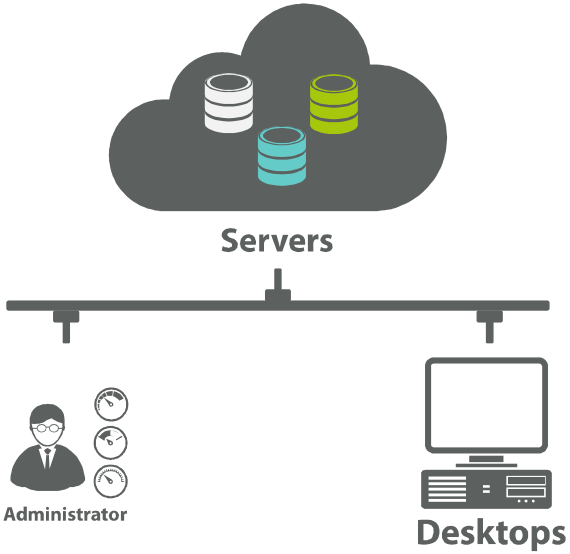


클라우드 기반 데이터레이크 및 분석

Data Lake란 무엇인가?

MEGAZONE CLOUD



Index

Data Lake의 개념

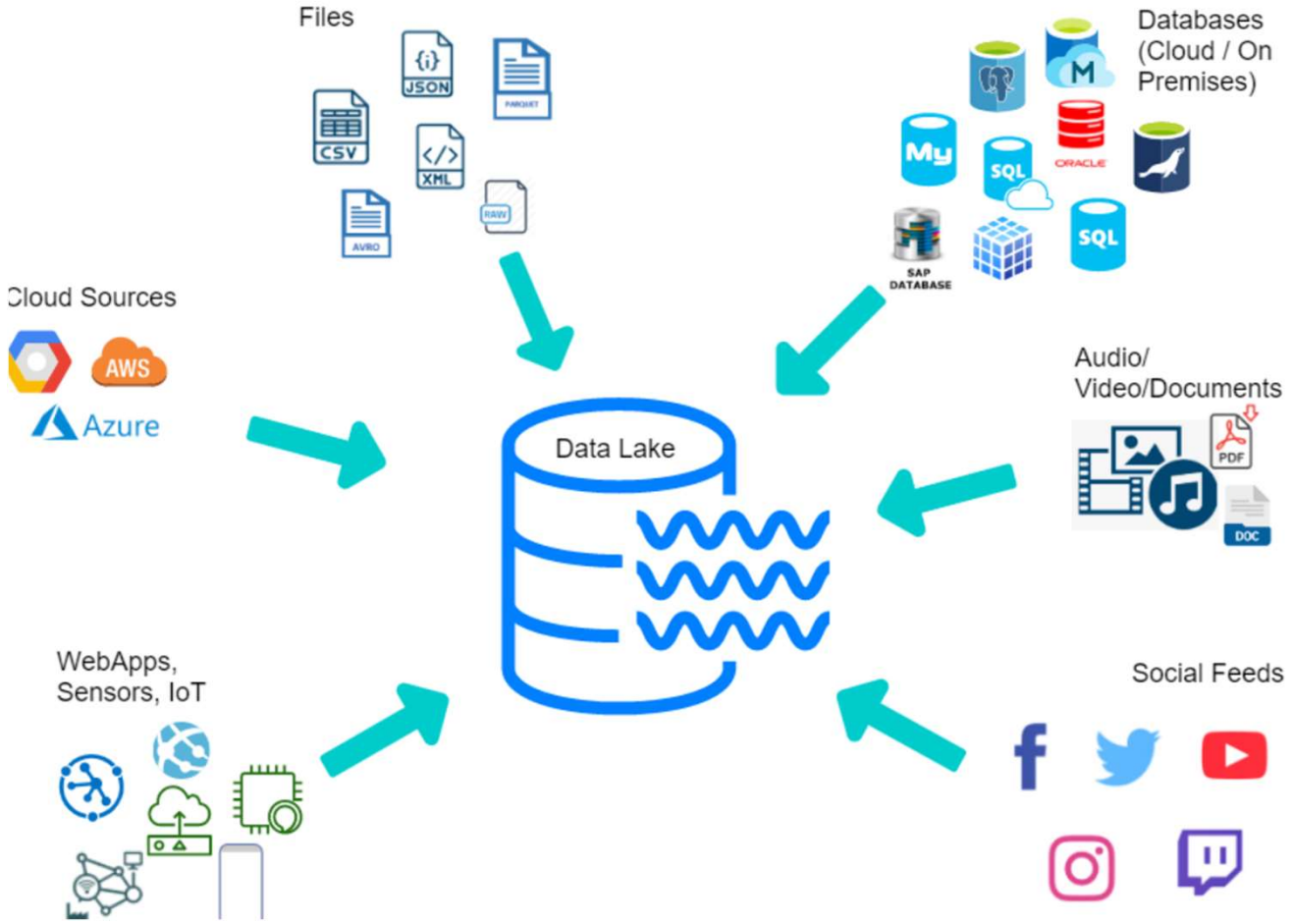
What's Data Lake?

- In Wikipedia

A data lake is a system or repository of data stored in its natural/raw format, usually object blobs or files. A data lake is usually a single store of data including raw copies of source system data, sensor data, social data etc., and transformed data used for tasks such as reporting, visualization, advanced analytics and machine learning. A data lake can include ***structured data*** from relational databases (rows and columns), ***semi-structured data*** (CSV, logs, XML, JSON), ***unstructured data*** (emails, documents, PDFs) and ***binary data*** (images, audio, video). A data lake can be established "on premises" (within an organization's data centers) or "in the cloud" (using cloud services from vendors such as Amazon, Microsoft, or Google).

https://en.wikipedia.org/wiki/Data_lake

What’s Data Lake? (Cont.)



https://vitalflux.com/data-lake-design-principles-best-practices/

What's Data Lake? (Cont.)

- In AWS

A data lake allows you to store all your structured and unstructured data, in one centralized repository, and at any scale. With a data lake, you can store your data as-is, without having to first structure the data, based on potential questions you may have in the future. Data lakes also allow you to run different types of analytics on your data like ***SQL queries***, ***big data analytics***, ***full text search***, ***real-time analytics***, and ***machine learning*** to guide better decisions.

https://aws.amazon.com/ko/big-data/what-is-a-data-lake/

What's Data Lake? (Cont.)

- In Text Book

여러 가지 다양한 유형의 데이터를 원본 포맷 형태로 저장하는 단일 저장소

Data Lake는 Hadoop 기반의 빅데이터 저장소로써, Raw Data 형태로 저장하여 Data Warehouse나 Business Application에 데이터 제공하는 역할을 수행

전사의 다양한 유형의 대용량 데이터를 Low Latency로 빅데이터를 수집하여, 사내의 모든 구성원들이 직접 필요한 데이터를 찾고, 이해하고, 확보하고, 분석할 수 있도록 해 주는 전사 데이터 플랫폼.

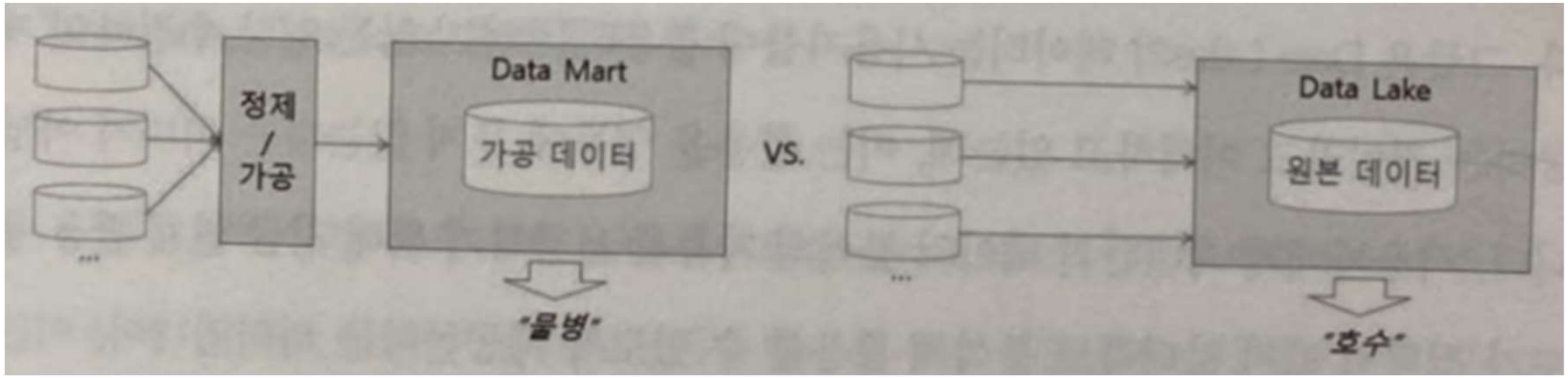
윤선웅, *차세대 빅데이터 플랫폼* ***Data Lake***, 좋은땅, 2021, p.41 ~ p.42.

What’s Data Lake? (Cont.)

- 전사
  - Enterprise
  - 전사에서 추진할 데이터 중심의 비즈니스 전환의 의미
- 모든 구성원
  - 모든 구성원에게 필요한 데이터 서비스 제공의 의미
- 활용
  - 사용자에게 최대한의 데이터 분석의 자유를 보장하기 위해
  - 현재 활용 용도가 정해져 있지 않더라도 분석에 활용할 수 있도록 제공의 의미
- 플랫폼
  - User Interface 기반으로 자동화하여 제공되는 데이터 서비스의 의미

Data Lake의 기원

- James Dixon(Pentaho - BI Solution Company CTO) in 2010
  - https://jamesdixon.wordpress.com/2010/10/
  - Data Mart와 대비되는 개념으로 설명
  - 원본 데이터 형태의 데이터 서비스로써 대용량의 물이 저장된 호수에 비유
  - 반면, Data Mart는 데이터를 편리하게 활용하기 위해 정제하고 가공하고 구조화하여 제공하는 데이터 서비스로써 소량의 정제수가 담긴 물병에 비유



  - https://jamesdixon.wordpress.com/2014/09/ → 재정리

윤선웅, *차세대 빅데이터 플랫폼 Data Lake*, 좋은땅, 2021, p.41 ~ p.44.

Data Lake vs. Data Warehouse

Data Lake vs. Data Warehouse

- Data Warehouse
  - Is a database optimized to analyze relational data coming from transactional systems and line of business applications.
  - The data structure, and schema are defined in advance to optimize for fast SQL queries
  - Where the results are typically used for operational reporting and analysis.
  - Data is cleaned, enriched, and transformed so it can act as the “single source of truth” that users can trust.

Data Lake vs. Data Warehouse (Cont.)

- Data Lake
  - Is different with Data warehouse.
  - Stores relational data from line of business applications, and non-relational data from mobile apps, IoT devices, and social media.
  - The structure of the data or schema is not defined when data is captured.
  - This means you can store all of your data without careful design or the need to know what questions you might need answers for in the future.
  - Different types of analytics on your data like SQL queries, big data analytics, full text search, real-time analytics, and machine learning can be used to uncover insights.

Data Lake vs. Data Warehouse (Cont.)

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

https://aws.amazon.com/ko/big-data/datalakes-and-analytics/what-is-a-data-lake/

Data Lake vs. Data Warehouse (Cont.)

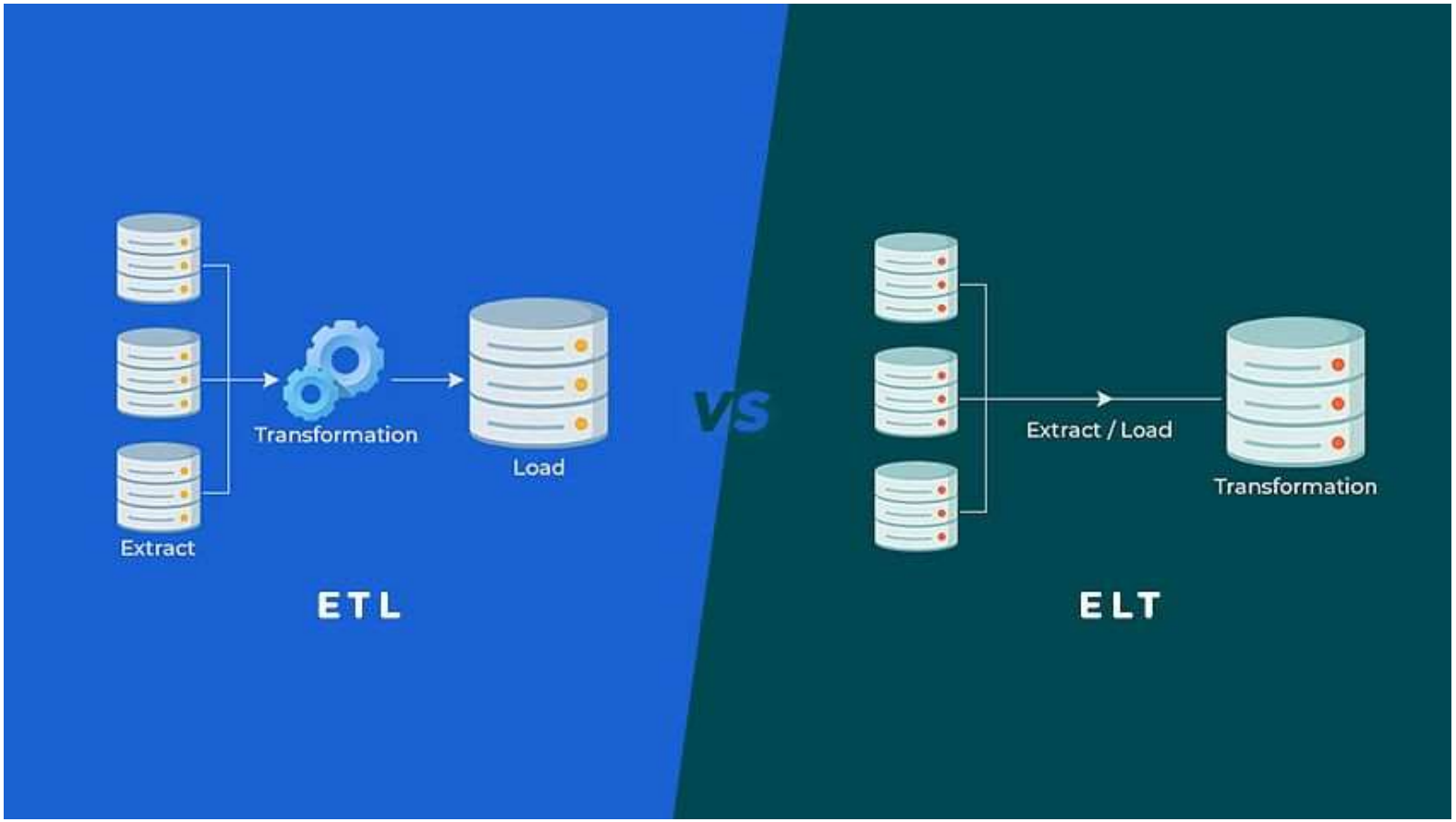
- Data Warehouse의 문제점을 보완하기 위한 방안으로 Data Lake 활용
  - 점점 더 늘어나는 데이터로 인한 성능 저하
  - 데이터 추출하고 변환하는 과정에서 ETL 부하 증가
  - 대체가 아닌 보완하기 위한 방안
- Data Lake는 수집과 유연한 분석이 가능
- Data Warehouse에 수용하기 어려운 데이터의 원천 데이터를 Data Lake에 적재
- But, Data Lake가 Data Warehouse를 대체하기 위해서는
  - 사용자의 데이터 분석에 대한 인식 전환과 역량 확보 필요
  - 기업 내 데이터 레이크 활용을 독려하는 분위기 필요
  - 사용자의 활용 데이터의 축적 필요

Data Lake vs. Data Warehouse (Cont.)

- Data Lake가 Data Warehouse를 대체하기 위한 2가지 방안

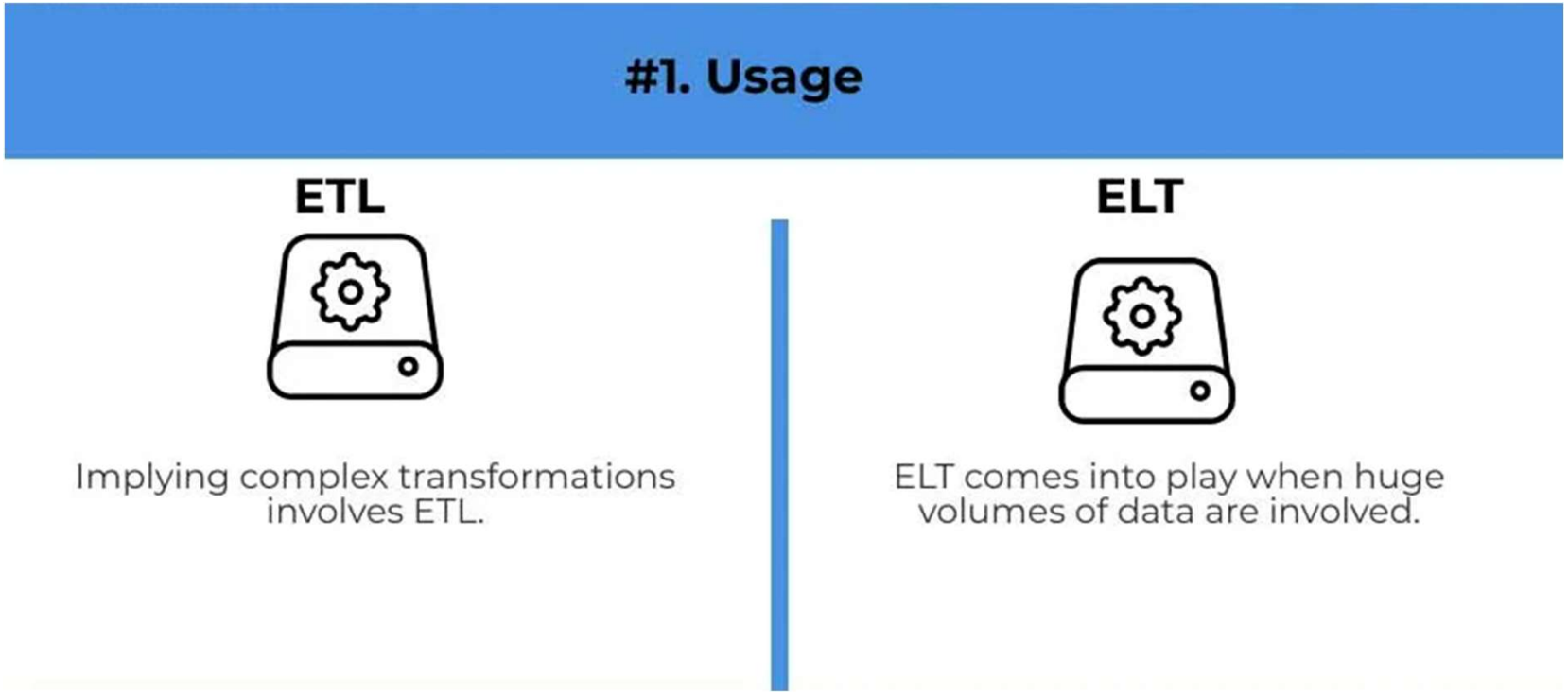
	1안)ETL 부하 경감하기 위한 방안	2안)Big Data를 Lake에 적재하기 위한 방안
장점	• ETL 부하 경감으로 ETL 비용 절감 및 DW 성능 강화 • Lake를 제한적 목적으로 활용하여 아키텍처 복잡성 감소	• DW는 일반 사용자, Lake는 Data Scientist로 명확한 사용자 구분 • 고급 데이터 분석가의 Lake 기반의 빅데이터 분석 가능
단점	• 기존 ETL을 모두 Hadoop 기반으로 변경 시 개발 부담 증가 • Lake는 ETL 역할만 수행하여, Lake 적재 데이터 활용 불가	• 기존 ETL 부하로 발생하는 문제점 미해결 • 데이터 관리 이원화로 사용자 혼란, 운영 부담 가중

ETL vs. ELT



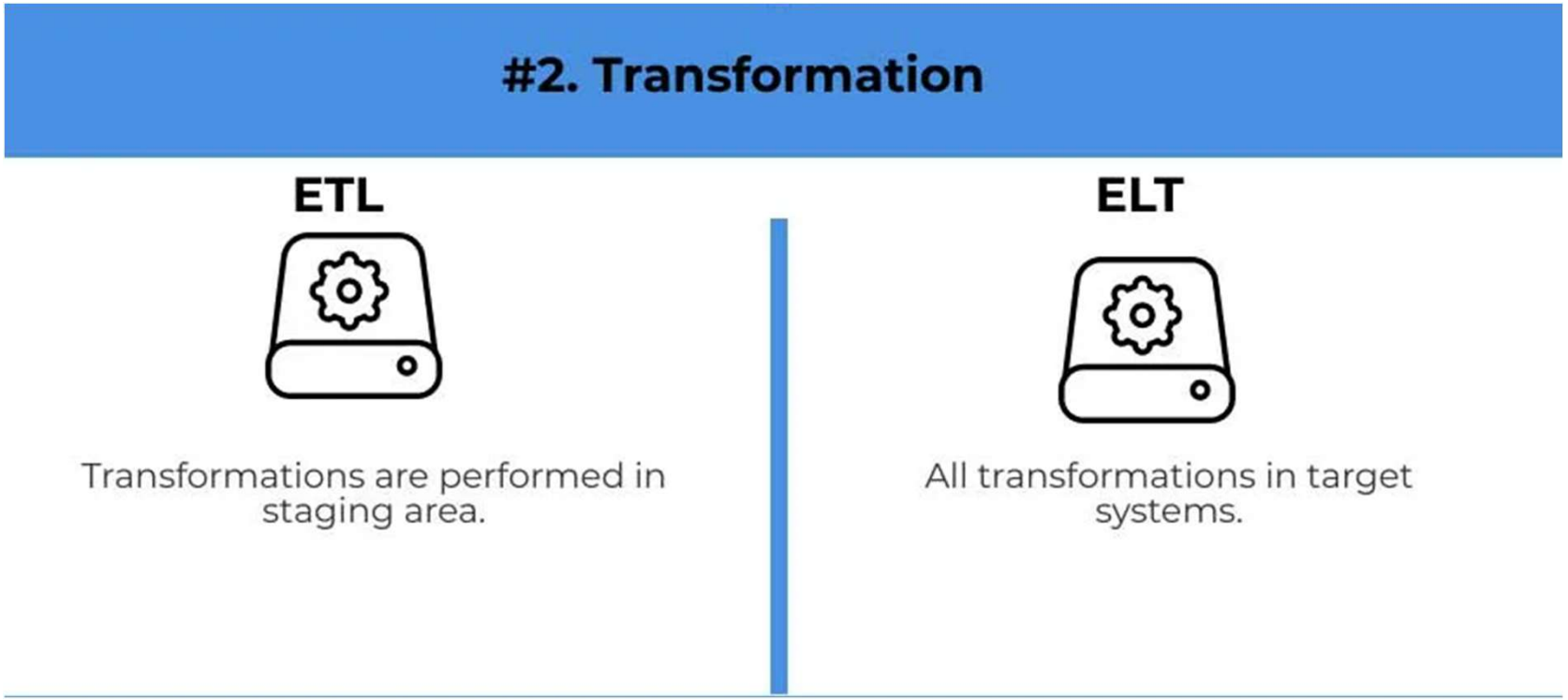
https://www.simplilearn.com/etl-vs-elt-article

ETL vs. ELT (Cont.)



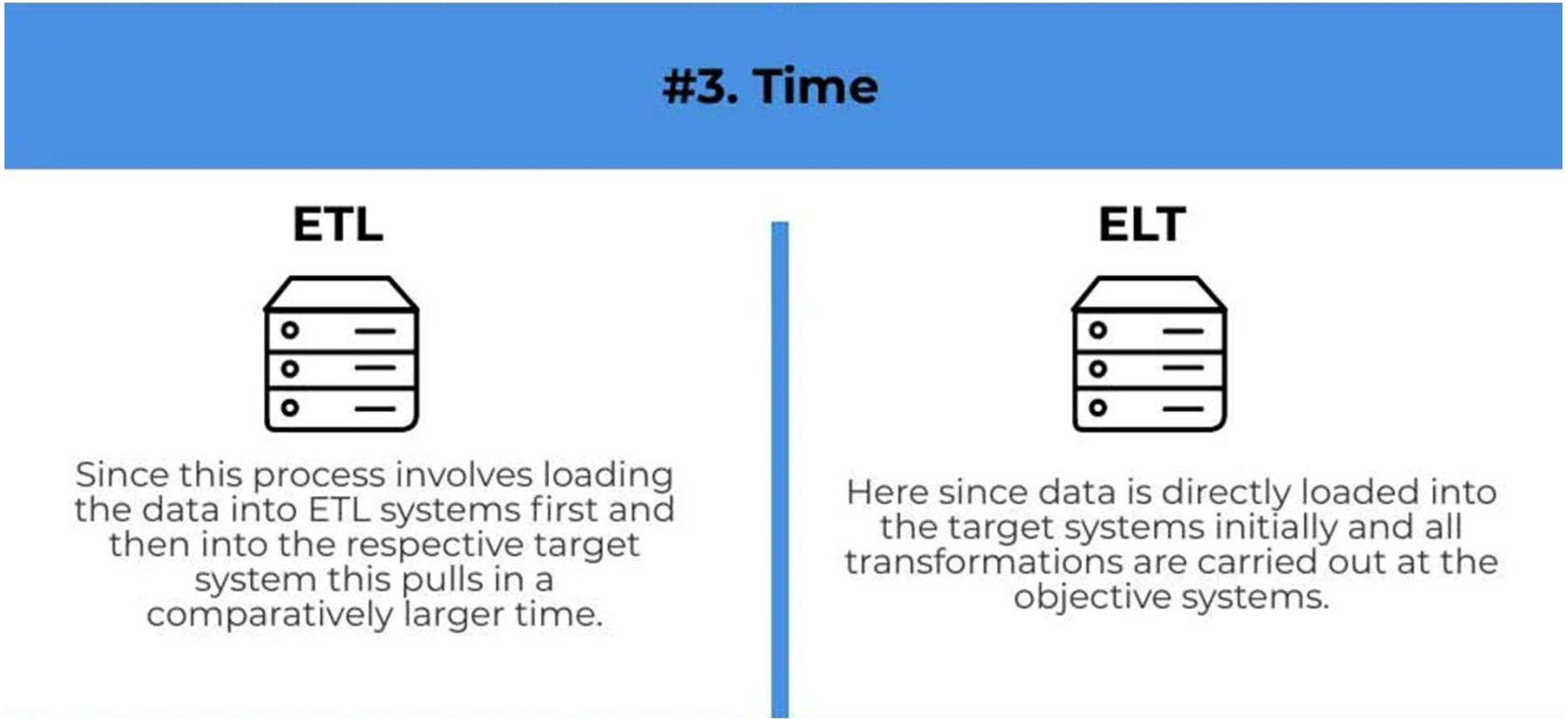
https://www.educba.com/etl-vs-elt/

ETL vs. ELT (Cont.)



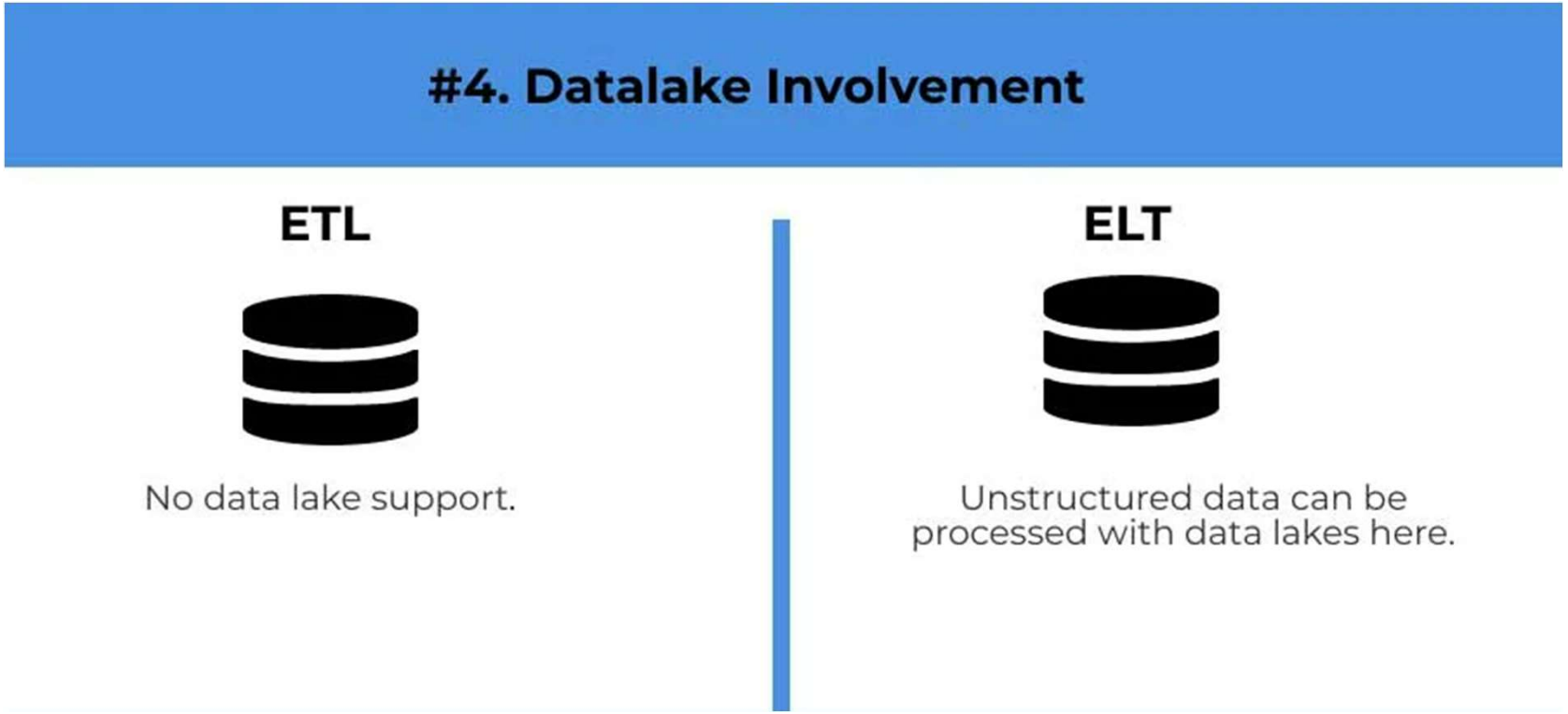
https://www.educba.com/etl-vs-elt/

ETL vs. ELT (Cont.)



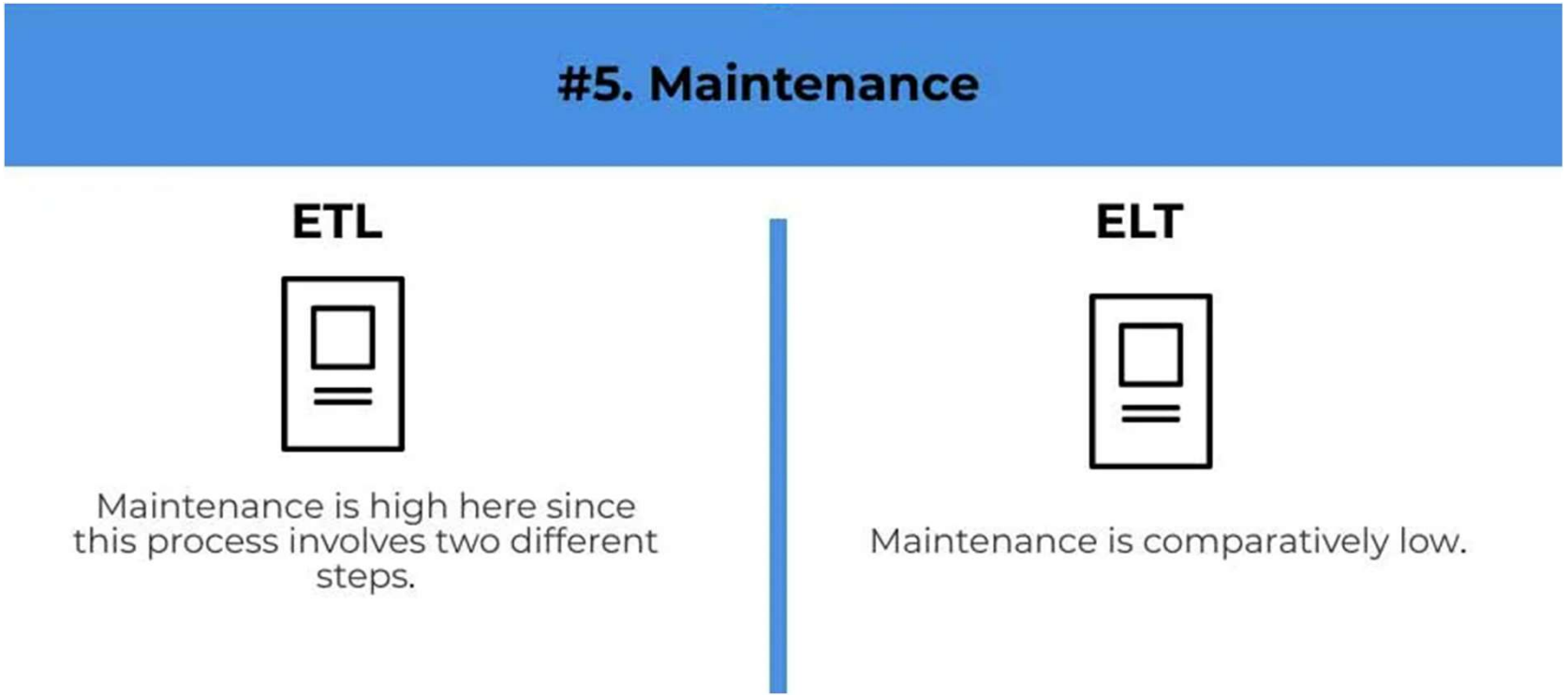
https://www.educba.com/etl-vs-elt/

ETL vs. ELT (Cont.)



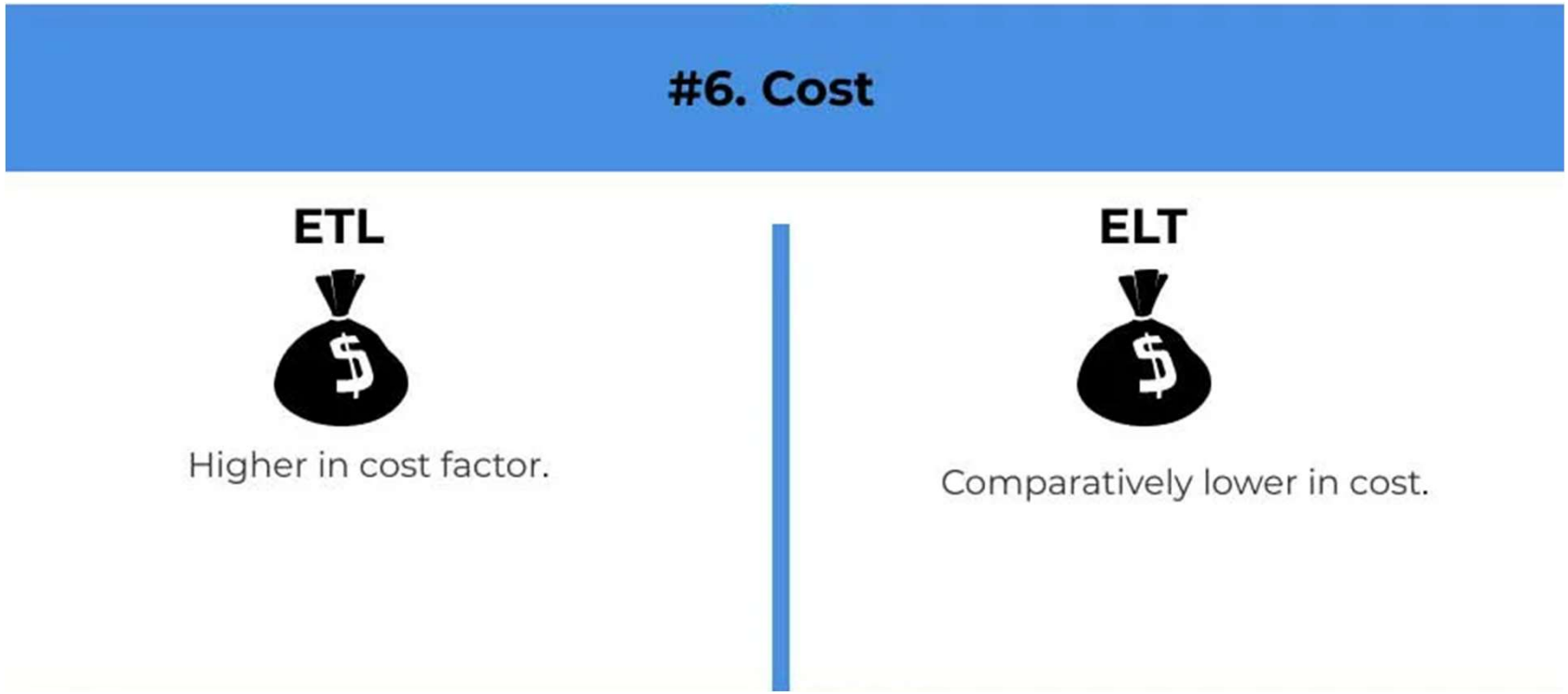
https://www.educba.com/etl-vs-elt/

ETL vs. ELT (Cont.)



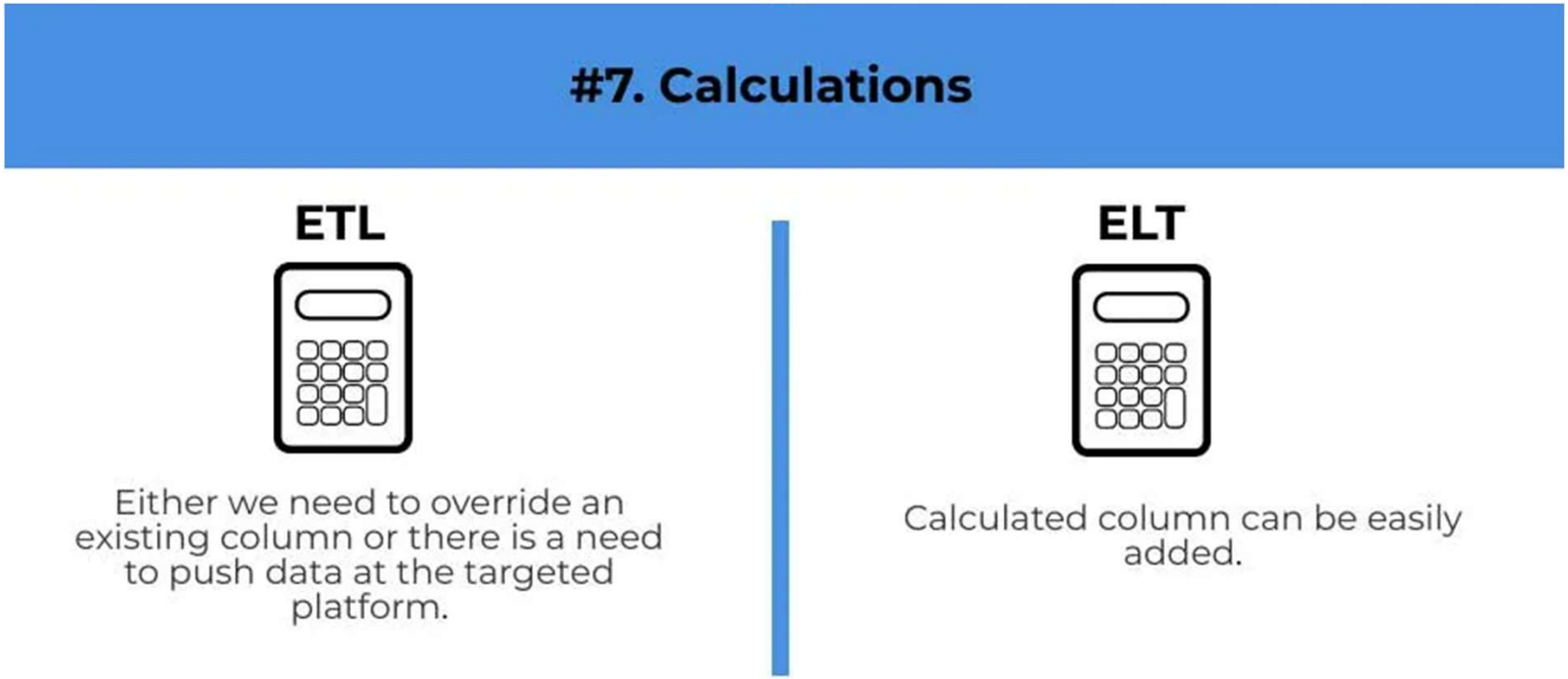
https://www.educba.com/etl-vs-elt/

ETL vs. ELT (Cont.)



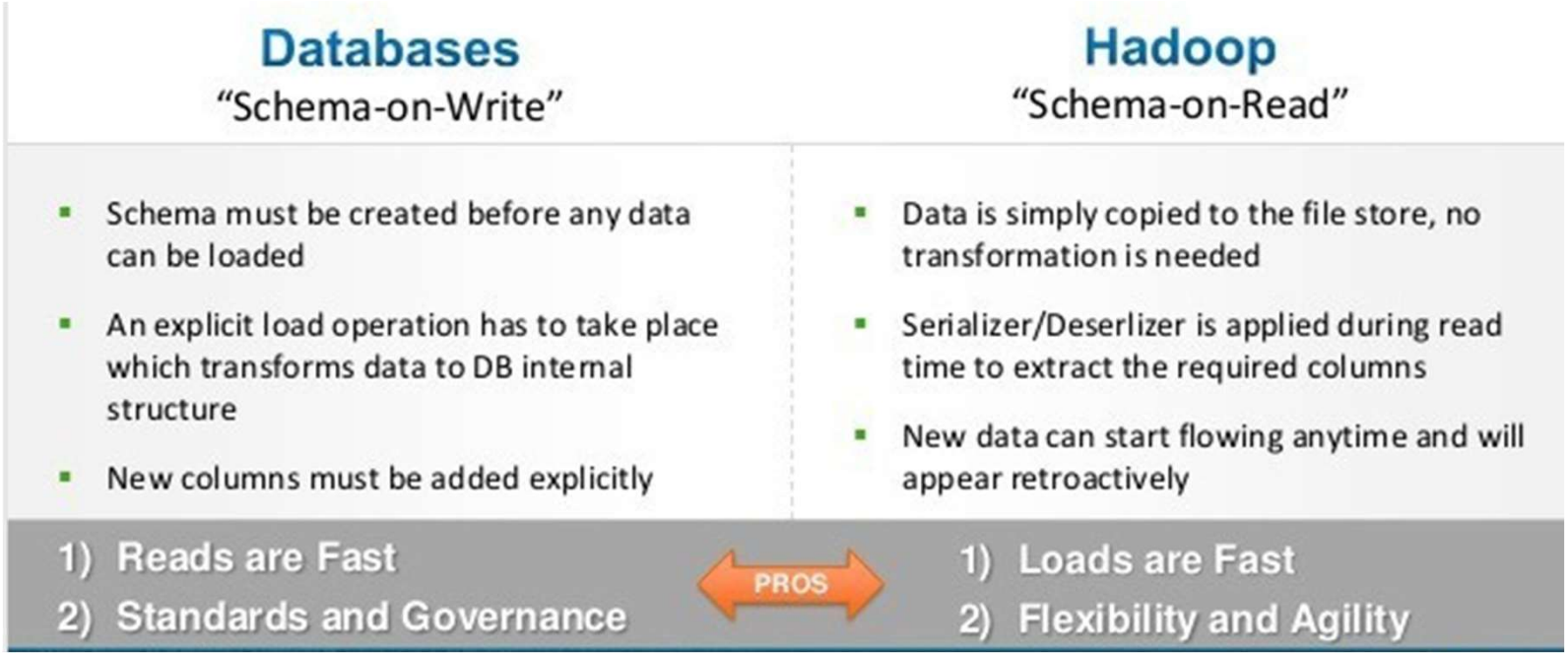
https://www.educba.com/etl-vs-elt/

ETL vs. ELT (Cont.)



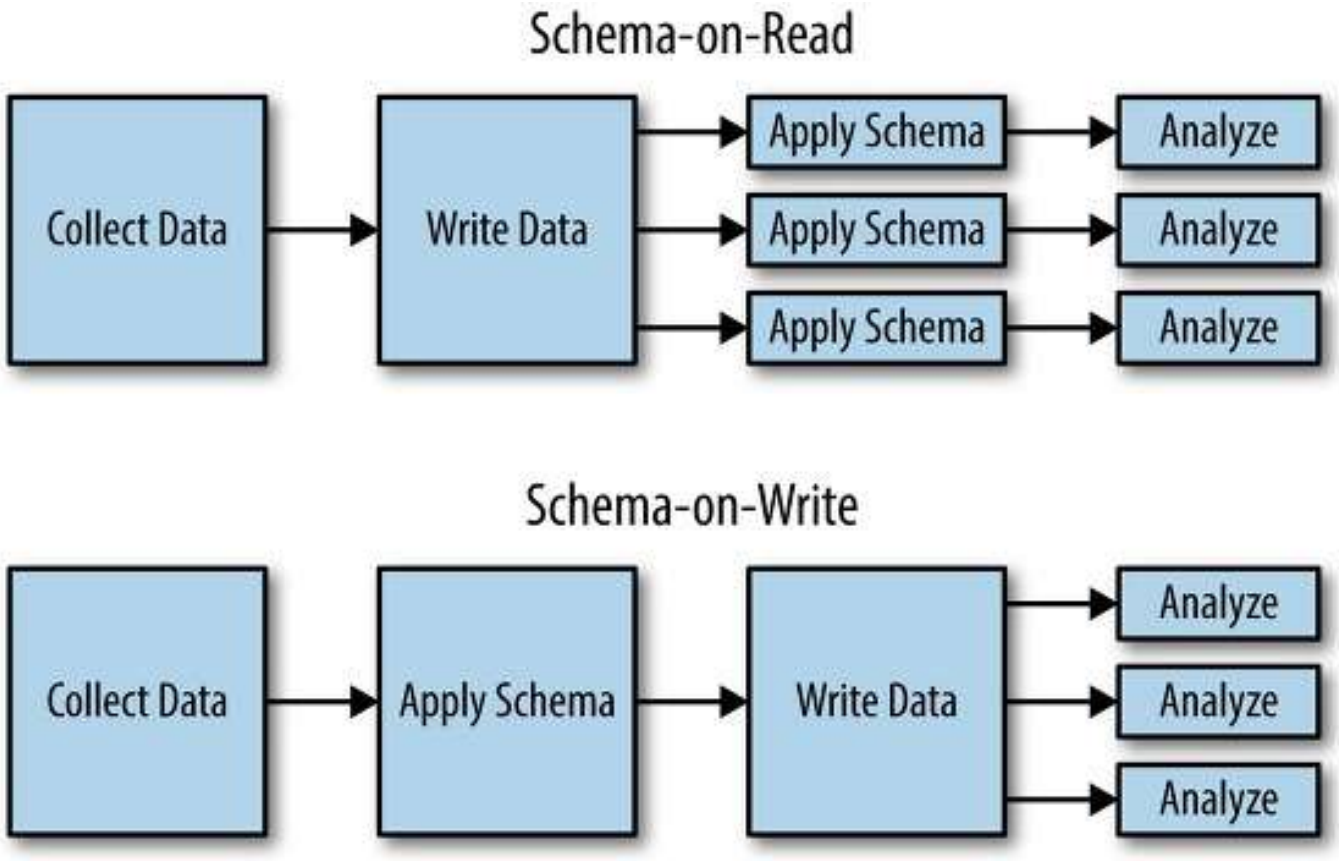
https://www.educba.com/etl-vs-elt/

Schema-on-Write vs. Schema-on-Read



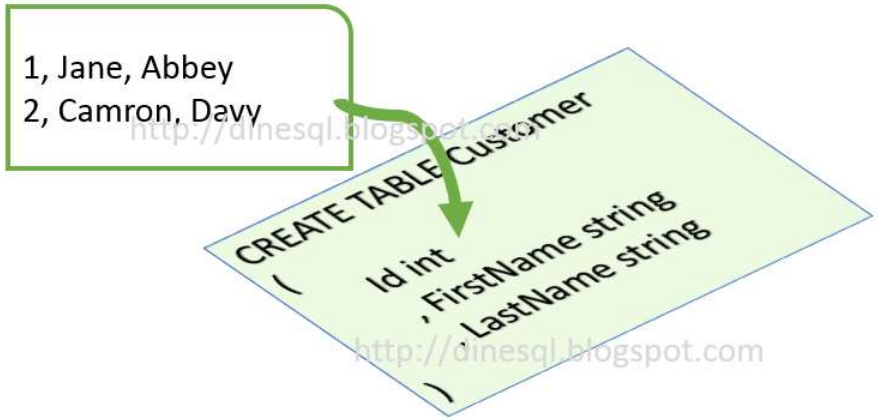
https://datacookbook.kr/90

Schema-on-Write vs. Schema-on-Read (Cont.)

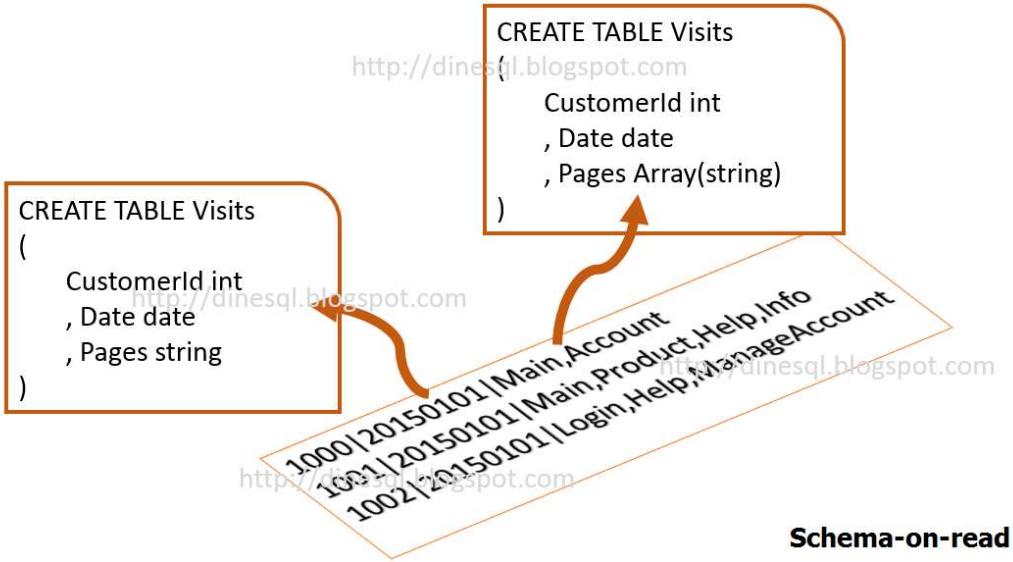


https://datacookbook.kr/90

Schema-on-Write vs. Schema-on-Read (Cont.)



Schema-on-write



Schema-on-read

http://dinesql.blogspot.com/2015/07/what-is-schema-on-write-and-schema-on.html