732A54/TDDE31

Big Data Analytics

Database Technologies for Data Analytics

Huanyu Li

Based on previous slides from Olaf Hartig

LINKÖPING UNIVERSITY

Topics

- NewSQL
- OLTP and OLAP, Data Warehouse
- Multidimensional Data Model and Operations
- Building a Data Warehouse
- Data Integration for Analytics
  - in the Age of Cloud Services

NewSQL

Characteristics of NewSQL Systems

- SQL and ACID guarantees for transactional read-write workloads (OLTP), just like traditional RDBMSs
- Performance and scalability comparable as NoSQL systems, through innovative software architecture
- Modern RDBMSs designed either to meet scalability requirements of distributed architectures or to improve performance so horizontal scalability is no longer a necessity

Categorization

- New architectures
  - New systems built from scratch to operate on shared-nothing resources, with components to support multi-node concurrency control, fault tolerance through replication, flow control, and distributed query processing
  - VoltDB, CockroachDB, MariaDB Xpand (formerly Clustrix), SingleStore (formerly MemSQL), SAP HANA, NuoDB
- Transparent sharding middleware
  - Centralized component that routes queries, coordinates transactions, and manages data placement, replication, and partitioning across a cluster of single-node DBMS instances
  - MariaDB MaxScale, ScaleArc
- Database as a service (DBaaS)
  - Amazon Aurora, ClearDB, Google Cloud Spanner

OLTP and OLAP

Online Transactional Processing (OLTP)

- The Most common use of relational DBs is for operational data
  - that is, data produced by the day-to-day operations of a business or an organization
  - e.g., students enrolling in courses, customers purchasing products, passengers purchasing airline tickets
- Workload characteristics:
  - simple queries (reads and writes)
  - many short transactions that make small changes
- Database systems that support the basic operations of a business are generally classified as OLTP systems
  - tuned to maximize throughput of concurrent transactions

Online Analytical Processing (OLAP)

- Enables analysts, managers, executives to gain insight into data as a basis for making decisions
- Primarily read-only workloads with complex queries
  - aggregations and grouping
  - touching large amounts of data
  - usually ad hoc

Data Warehouse

Data Warehouse

- Data warehouse: separate copy of the operational data, organized in a way that it can be used for executing decision support queries and/or data mining queries
  - usually a combination of data from multiple sources
  - data warehouses keep years' worth of data (in contrast, operational data in OLTP systems is short-lived and changes frequently)

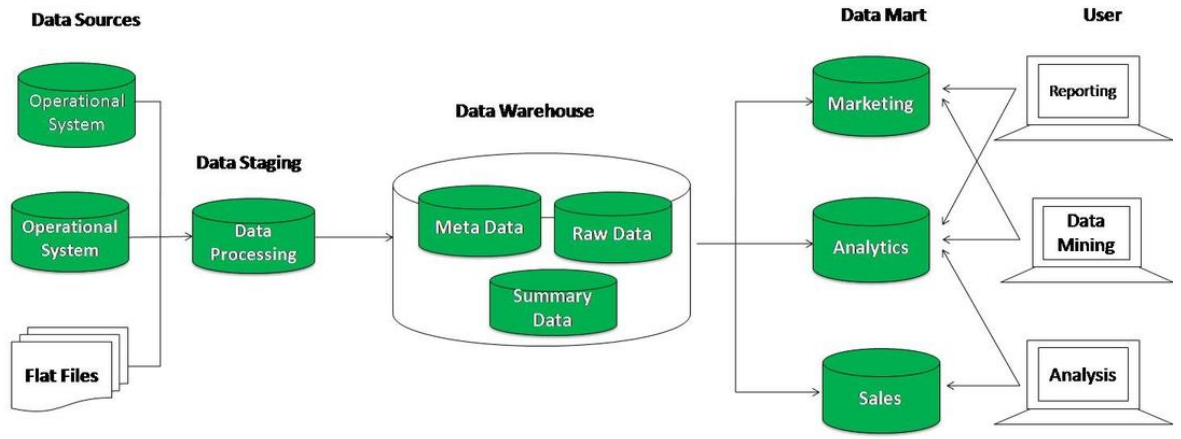
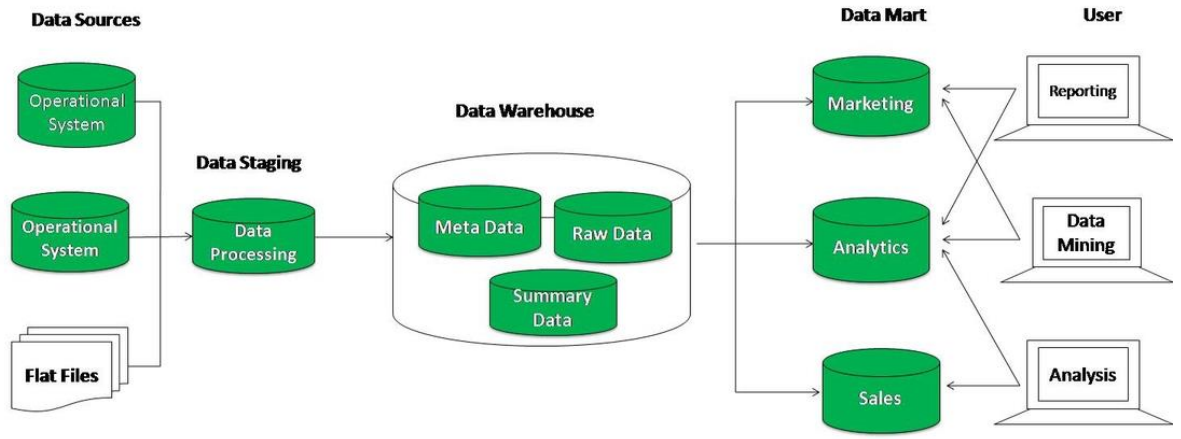


image source: https://www.geeksforgeeks.org/data-analysis/data-warehouse-development-life-cycle-model/

Why a separate system?

- Usually a combination of data from multiple sources
- Data organized differently, to better support OLAP queries
- Complexity of OLAP queries



- They take too much time to be executed in a transaction processing system with high throughput requirements
- They may lock the database for long periods of time and, thus, negatively affect all other OLTP transactions

image source: https://www.geeksforgeeks.org/data-analysis/data-warehouse-development-life-cycle-model/

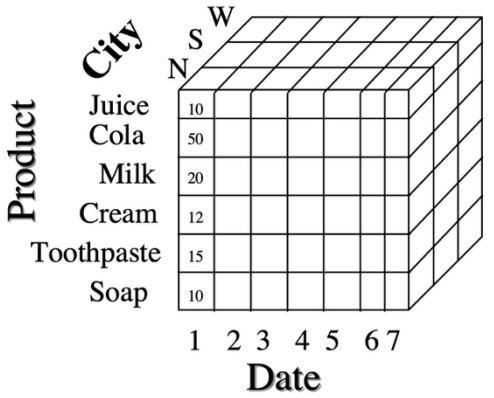
Categories of OLAP Systems

- Relational OLAP systems (“ROLAP”)
  - Store data in relations
  - Queries written in SQL
- Special-purpose OLAP systems
  - Represent and store data in a multi-dimensional array
  - OLAP-specific query language or spreadsheet-like UI

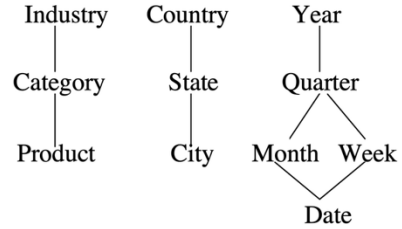
Multidimensional Data Model

Multidimensional Data Model

- Numeric measures that are the focus of the analysis
  - e.g., sales amount, budget, revenue, inventory counts
- Each such measure depends on a set of dimensions
  - e.g., dimensions of a sales amount may be product name, city, and date
- Each dimension is described by a set of attributes
  - e.g., product dimension may consist of product category, industry of the product, year of introduction, and average profit margin
- Some attributes may form a hierarchy of relationships



Dimensions: Product, City, Date
Hierarchical summarization paths



Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM Sigmod record*, *26*(1), 65-74.

Multidimensional Data in an RDBMS

- Dimension tables with the attributes of the dimensions
- Fact table with a column for each dimension (foreign keys to the dimension tables) and for the numeric measures
  - i.e., one tuple/row per cell of the multidimensional array
- Star schema: single dimension table for each dimension

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM Sigmod record*, *26*(1), 65-74.

Multidimensional Data in an RDBMS

- Snowflake schema: dimension tables normalized
  - hence, hierarchies represented explicitly
  - e.g., LOCATIONS(locid, city, state) and STATES(state, country)

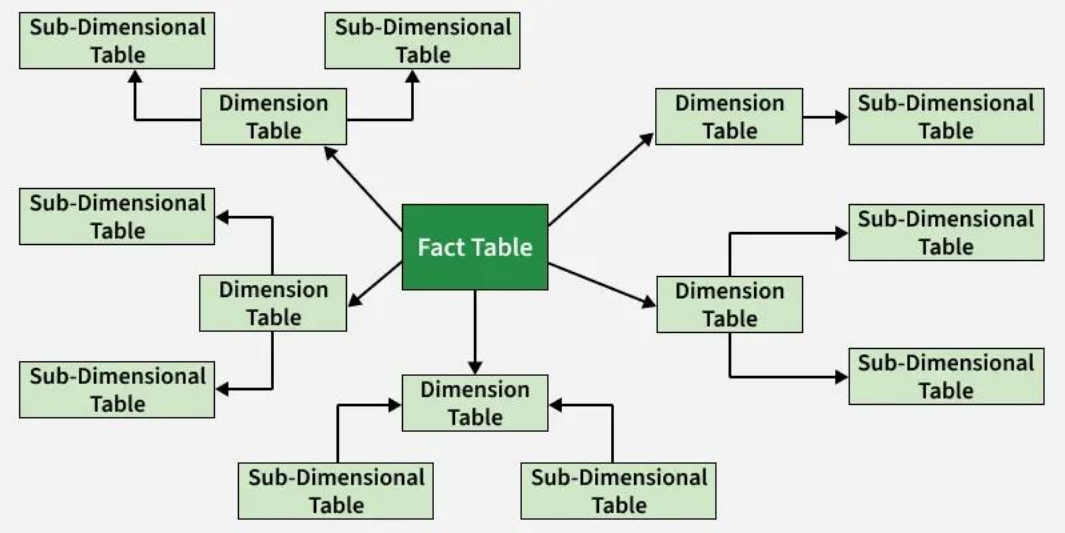


image source: https://www.geeksforgeeks.org/dbms/snowflake-schema-in-data-warehouse-model/

Operations over Multidimensional Data

Data Cube Operations

- Slicing
- Dicing
- Roll-up
- Drill-down
- Pivot

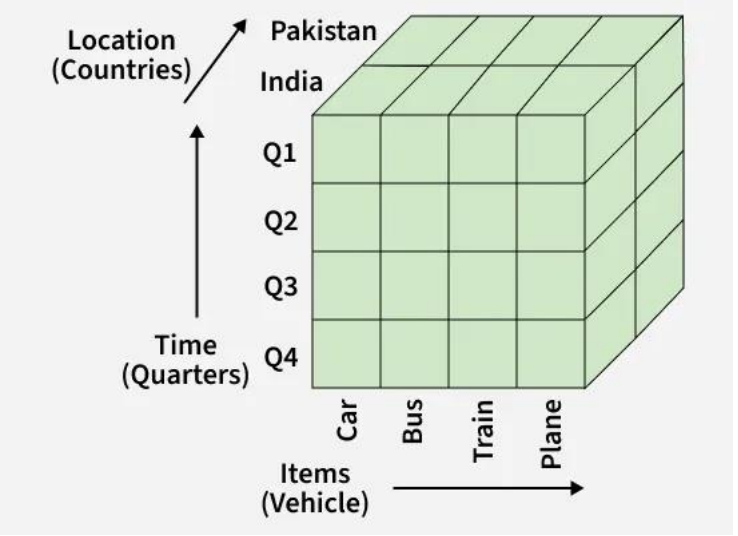
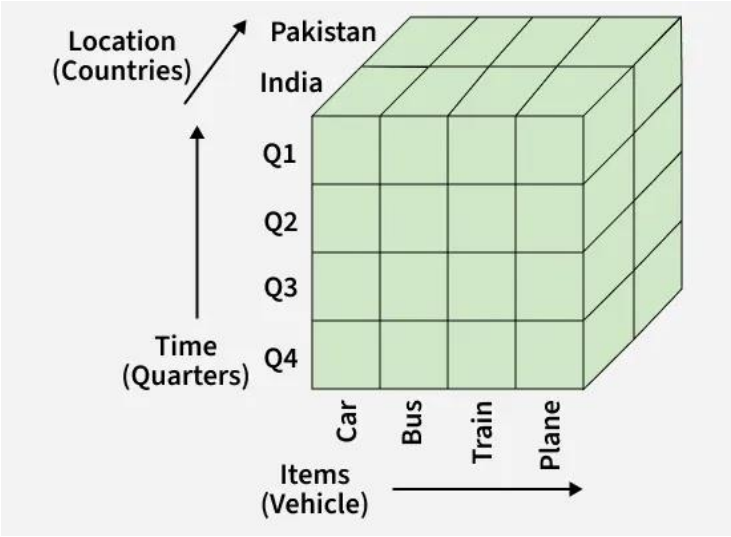


image source: https://www.geeksforgeeks.org/dbms/olap-operations-in-dbms/

Slicing

- An operation that filters the unnecessary portions, reducing the dimensions by selecting a single value for one of the dimensions
  - e.g., …WHERE dim_attr=xyz



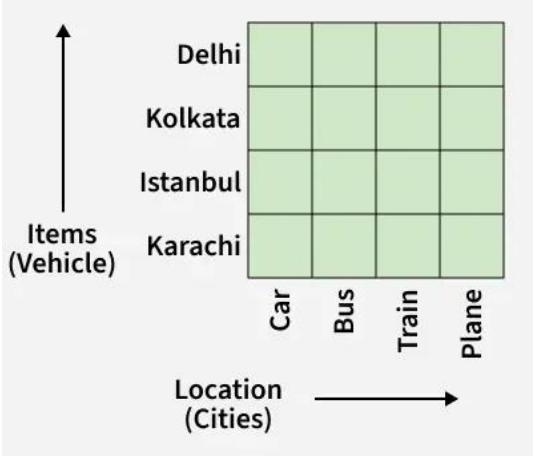
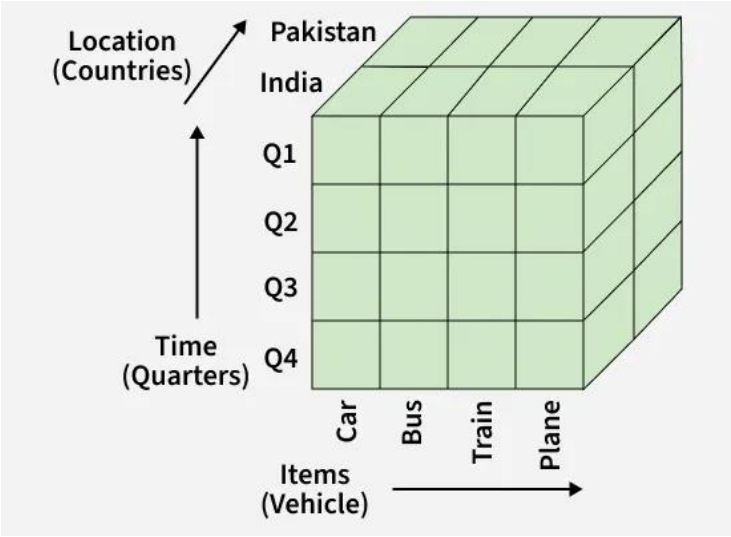


image source: https://www.geeksforgeeks.org/dbms/olap-operations-in-dbms/

Dicing

- An operation that does a multi-dimensional cutting, producing a sub-cube by selecting a range of values for one or more of the dimensions
  - e.g., …WHERE dim_attr>xyz
  - …WHERE dim_attr IN (x, y, z)



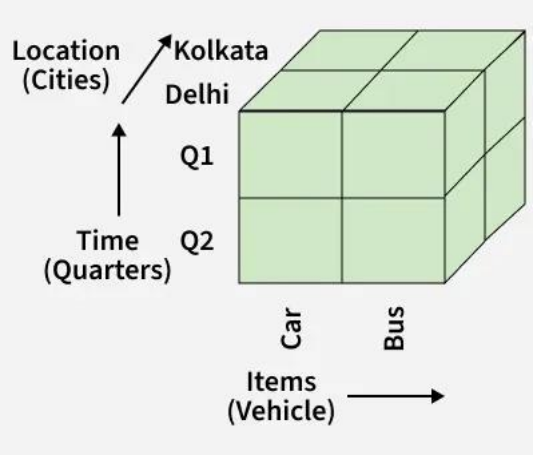
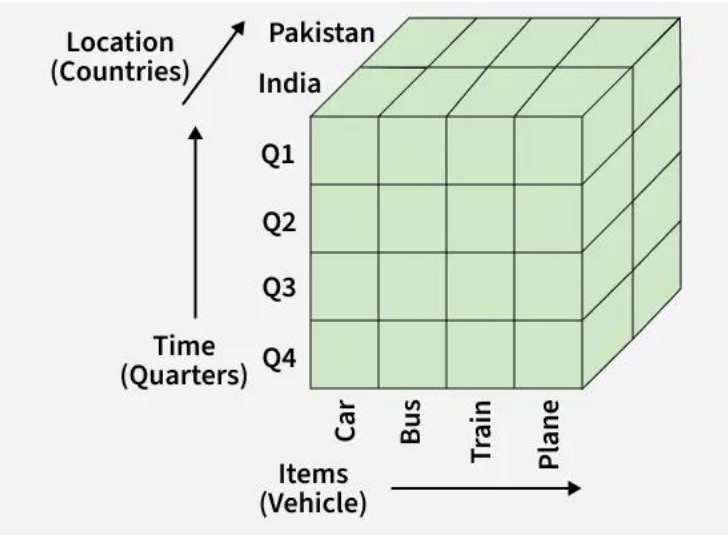


image source: https://www.geeksforgeeks.org/dbms/olap-operations-in-dbms/

Roll-up and Drill-down

- Roll-up: an operation that aggregates certain similar data attributes having the same dimension together;
  - e.g., sum up by months instead of days or by countries
- Drill-down: the reverse operation of roll-up



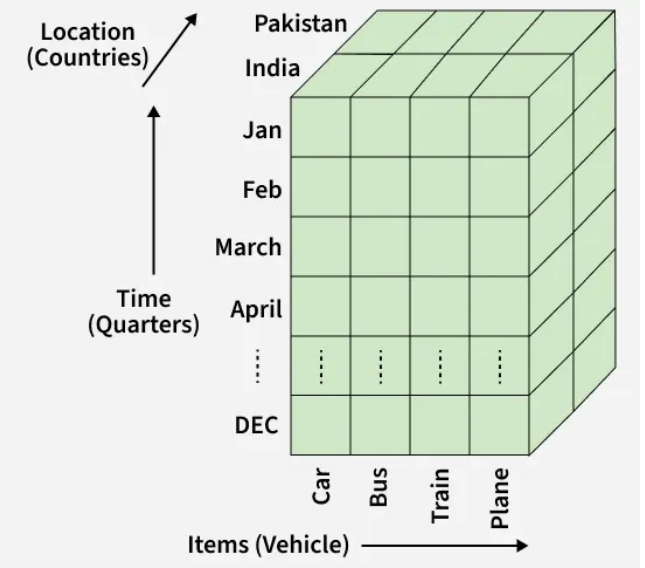
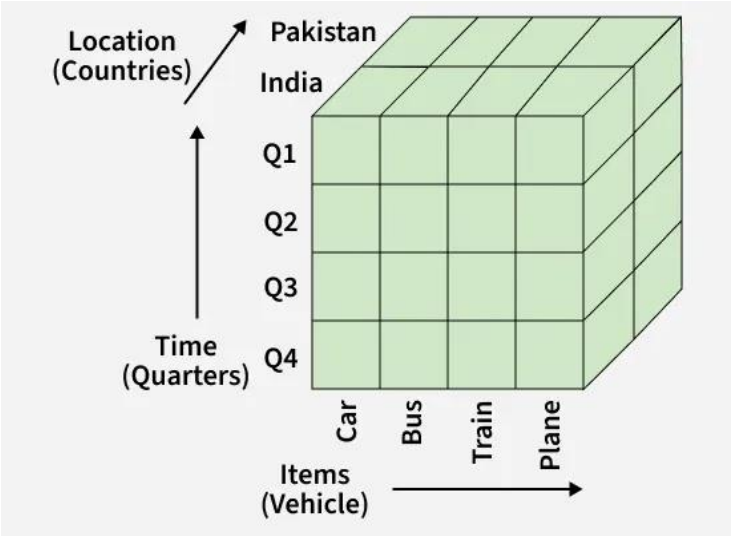


image source: https://www.geeksforgeeks.org/dbms/olap-operations-in-dbms/

Pivot (Rotation)

- An operation that transforms the data cube in terms of a view, rotating the cube to show a different orientation of the axes



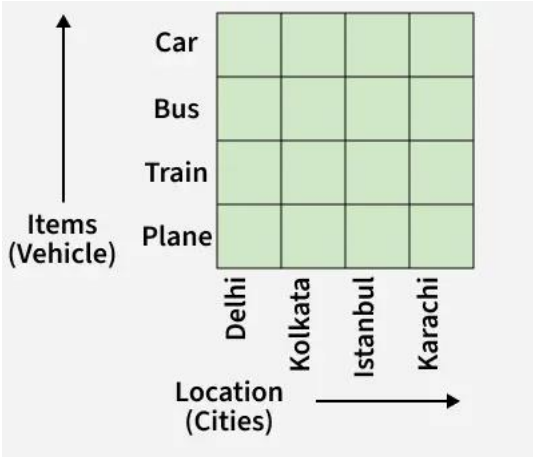


image source: https://www.geeksforgeeks.org/dbms/olap-operations-in-dbms/

Building a Data Warehouse

Building a Data Warehouse

- Identify desired data sources
- Scope the analytics needs that the project is meant to solve
- Define the data model/schema that the analysts and other end users need
- Build an extract-transform-load pipeline
- Conduct analytics work, extract insights

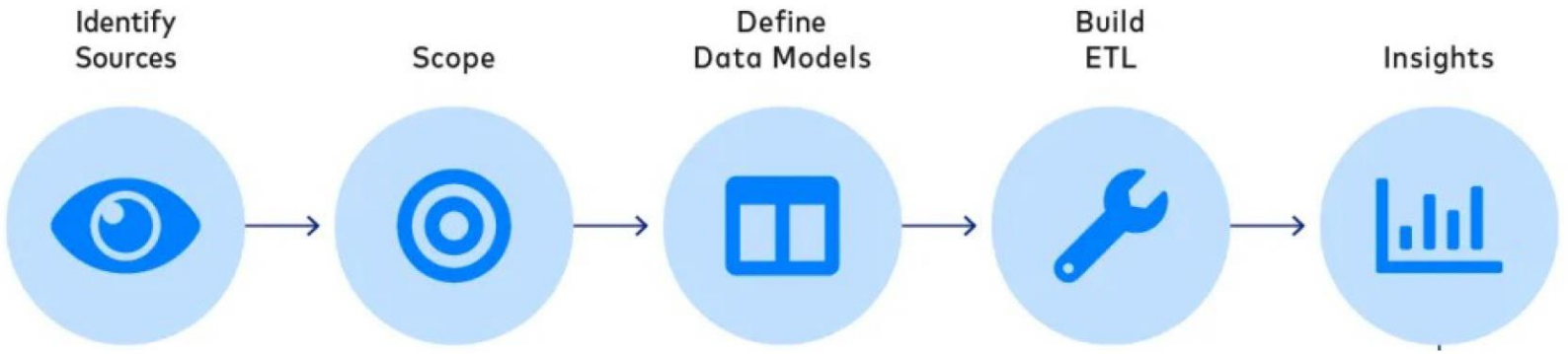


image source: https://www.fivetran.com/blog/etl-vs-elt

ETL

- Extract: query the operational databases to retrieve relevant data, and run scripts to extract from other types of sources
- Transform: clean the data (i.e., delete or repair tuples with missing or invalid information) and reorganize it to fit the schema of the warehouse
- Load: populate the warehouse with the data, build indexes

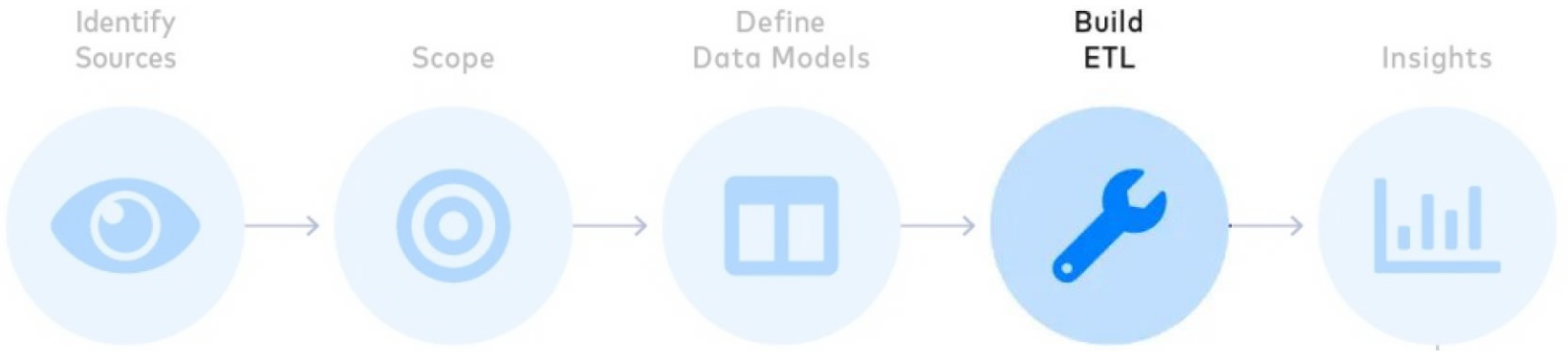


image source: https://www.fivetran.com/blog/etl-vs-elt

Challenges of Data Warehouses and ETL

- Data in the warehouse needs to be refreshed periodically
- Building and maintaining a data warehouse is a huge effort, may easily go into millions of dollars

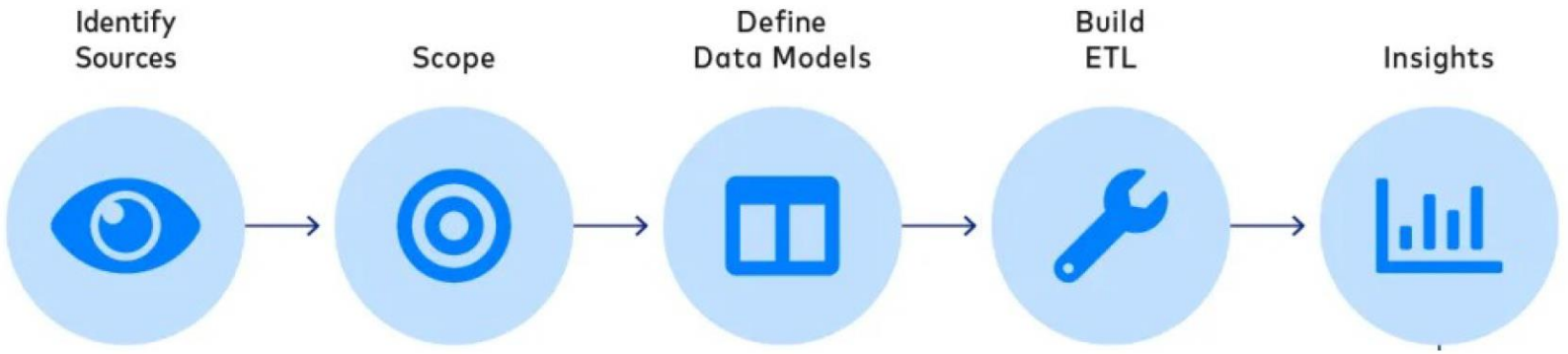


image source: https://www.fivetran.com/blog/etl-vs-elt

Challenges of ETL

- Data in the warehouse needs to be refreshed periodically
- Building and maintaining a data warehouse is a huge effort, may easily go into millions of

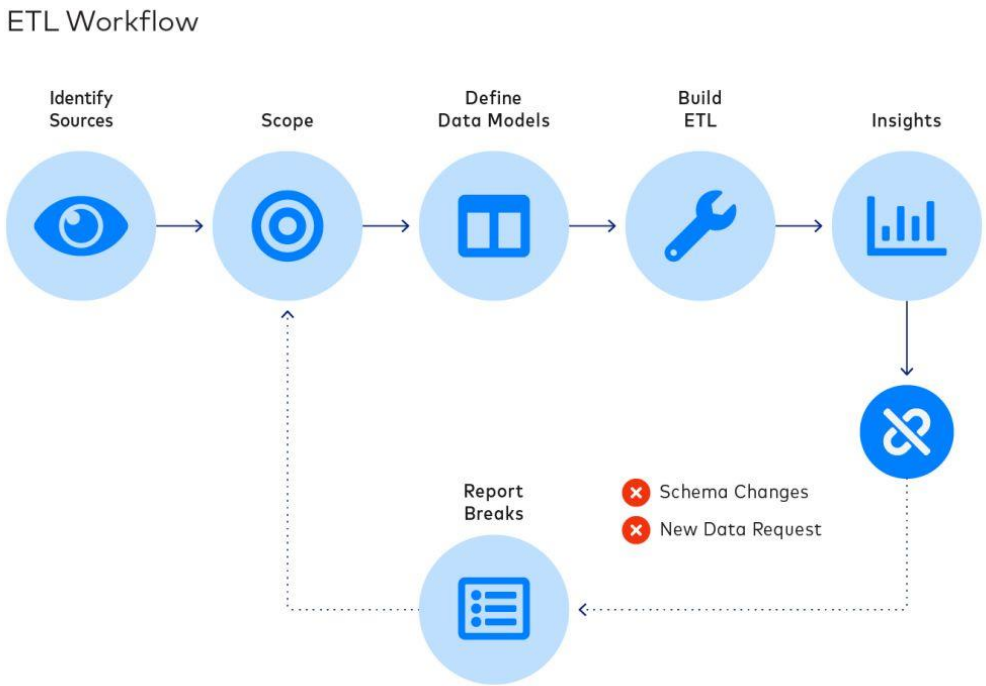


image source: https://www.fivetran.com/blog/etl-vs-elt

Building a Data Warehouse

Data Integration

- Data integration is the problem of combining data [from] different sources [into a single] unified view of these data*
  - schema mapping
  - record linkage (entity resolution)
  - inconsistent formats or units
- Modern technologies for data integration
  - Integration Platform as a Service (iPaaS)
  - ELT (Extract, Load, and Transform)
  - Reverse ETL

image source: https://www.fivetran.com/blog/etl-vs-elt

Integration Platform as a Service (iPaaS)

- Enable users to integrate applications with one another
  - in practice: an event in an application / system is transmitted to the iPaaS (via an API call or a Webhook) which then performs some predefined actions
- Data moves between applications directly through the iPaaS
- Little to no transformation takes place in the iPaaS
- Popular iPaaS
  - tray.io
  - workato
  - integromat
  - zapier
  - automate.io

image source: https://www.fivetran.com/blog/etl-vs-elt

ELT: Extract, Load, and Transform, and Load

- Cloud data warehouses have become extremely fast and reliable, which enables transformations to take place inside the warehouse itself
- ELT: Data moves directly from (cloud) applications to the data warehouse; afterwards, transformation in the data warehouse via SQL
  - No coding required!

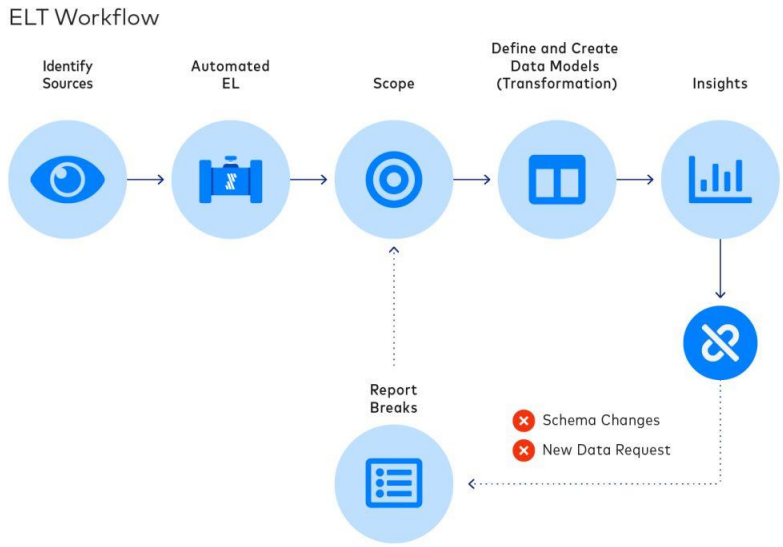


image source: https://www.fivetran.com/blog/etl-vs-elt

ELT Tools

- Modern ELT tools don't even offer in-built transformation capabilities
  - which was one of the major parts of traditional ETL tools
- Instead, to handle transformations in the data warehouse they integrate purpose-built solutions
  - e.g., dbt
- Leading companies:
  - Fivetran
  - Stitch
  - Matillion
  - Airbyte

Workflows for ETL versus ELT

- Identify desired data sources
- Automatically extract & load (can be outsourced, scaled up and down)
- Scope the analytics needs
- Define and create the data model needed for the analytics work

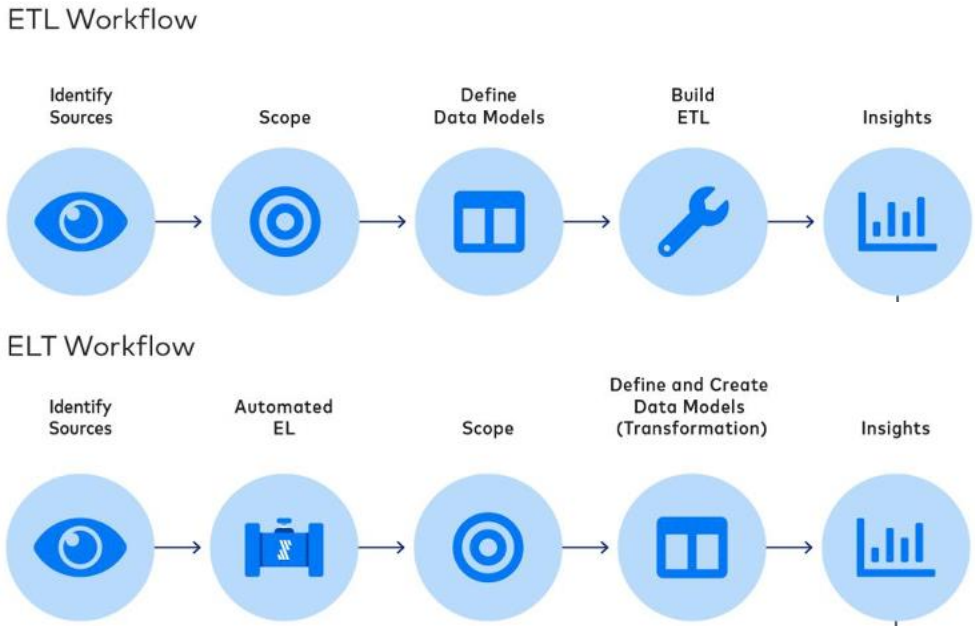
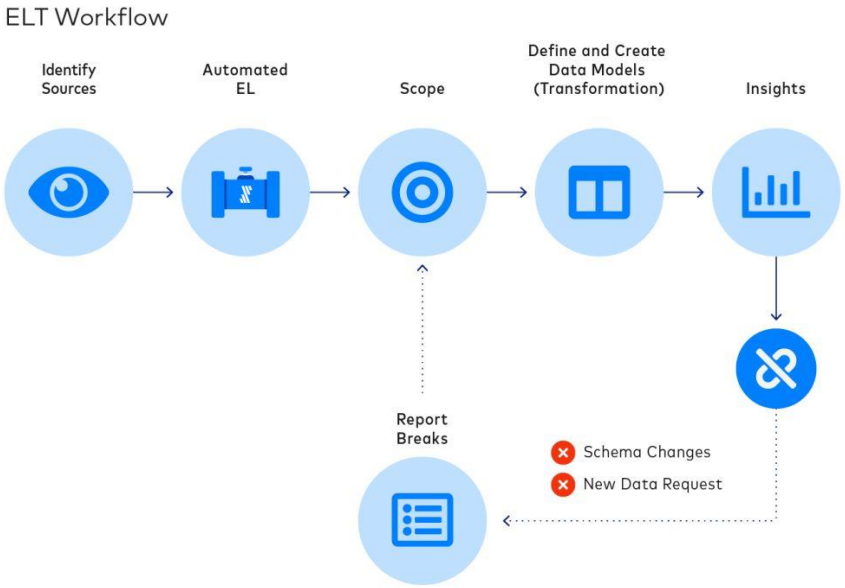


image source: https://www.fivetran.com/blog/etl-vs-elt

Workflows for ETL versus ELT

- Still, in ELT,
  - source schemas may change, and analytics needs may evolve and change
- But, transformation failures do not prevent data from being loaded



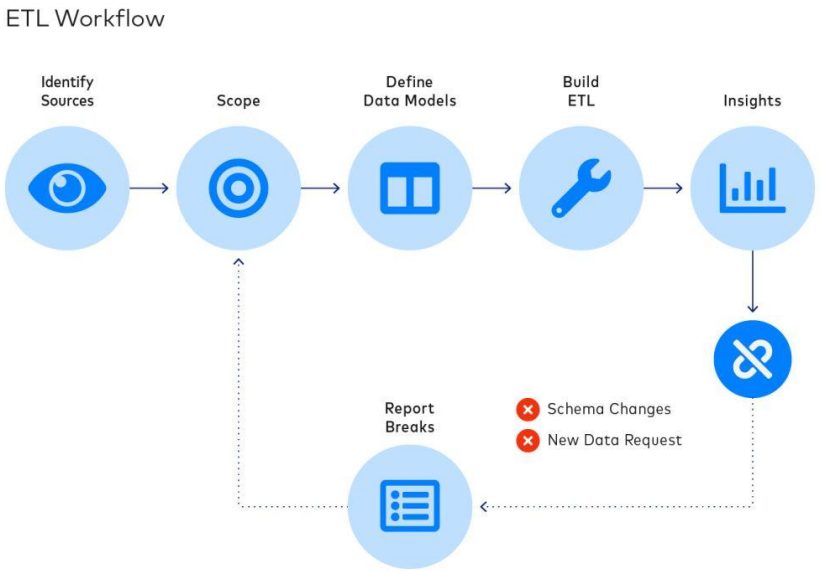


image source: https://www.fivetran.com/blog/etl-vs-elt

Reverse ETL

- Main use case: sync customer data from the data warehouse to sales, marketing and analytics tools
  - consistent view of the customer across all systems
  - enable operational analytics
- Main functionality of reverse ETL tools:
  - extract data from a data warehouse on a regular basis and load it into sales, marketing, and analytics tools
  - trigger a webhook or make an API call when data changes
  - move extracted data to a production database

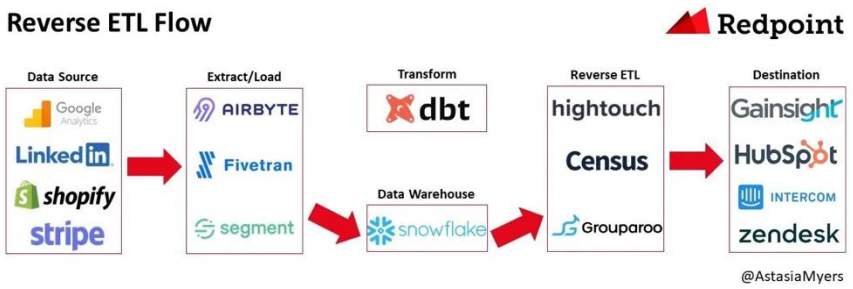


image source: https://medium.com/memory-leak/reverse-etl-a-primer-4e6694dcc7fb

Reverse ETL

- Main use case: sync customer data from the data warehouse to sales, marketing and analytics tools
  - consistent view of the customer across all systems
  - enable operational analytics
- Main functionality of reverse ETL tools:
  - extract data from a data warehouse on a regular basis and load it into sales, marketing, and analytics tools
  - trigger a webhook or make an API call when data changes
  - move extracted data to a production database
- Reverse ETL tools offer connectors for many cloud apps
- Startups that are building reverse ETL products:
  - Hightouch,Census, Grouparoo, Headsup, Polytomic, SeekWell

LINKÖPING UNIVERSITY