732A54 / TDDE31
Big Data Analytics

Topic:
Database Technologies for Data Analytics

Olaf Hartig

olaf.hartig@liu.se

LINKÖPING UNIVERSITY

Online Transactional Processing (OLTP)

- 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
  - touch large amounts of data
  - usually ad hoc

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 keeps years' worth of data (in contrast, operational data in OLTP systems is short-lived and changes frequently)

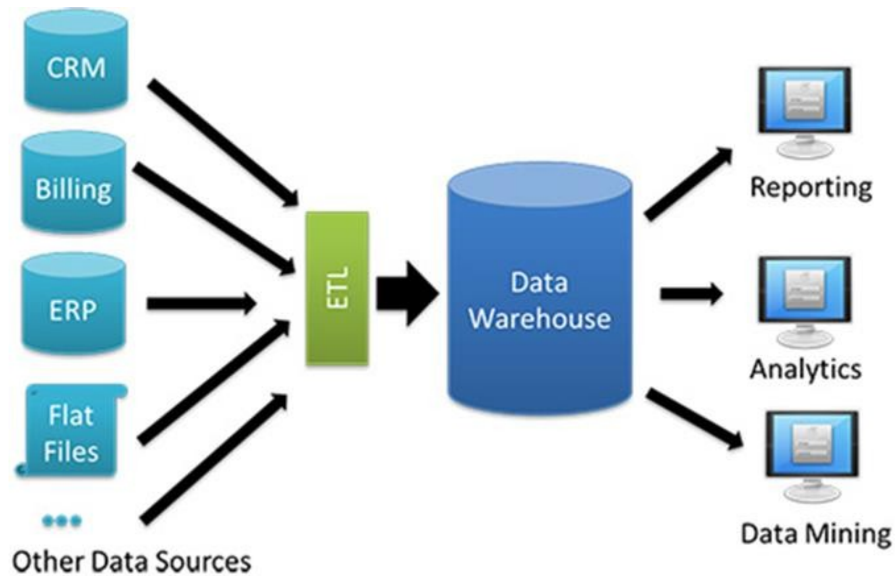


Figure from https://www.monitis.com/blog/top-5-data-warehouses-on-the-market-today/

Why a separate system?

- Usually a combination of data from multiple sources
- Data organized differently, to better support OLAP queries
- Complexity of OLAP queries
  - take too much time to be executed in a transaction processing system with high throughput requirements
  - may lock the database for long periods of time and, thus, negatively affect all other OLTP transactions

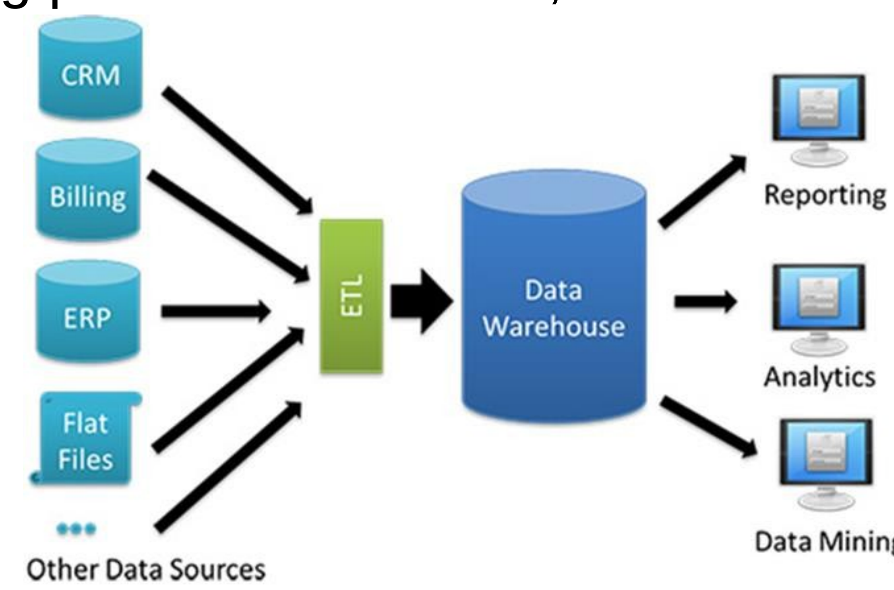


Figure from https://www.monitis.com/blog/top-5-data-warehouses-on-the-market-today/

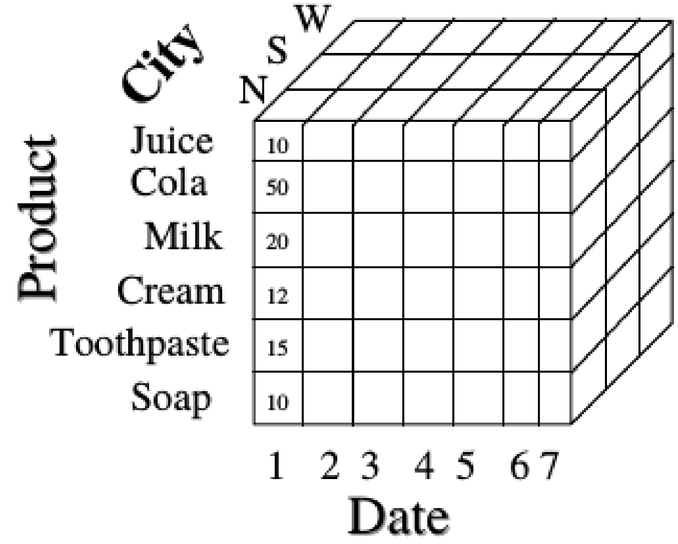
Categories of OLAP Systems

- 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

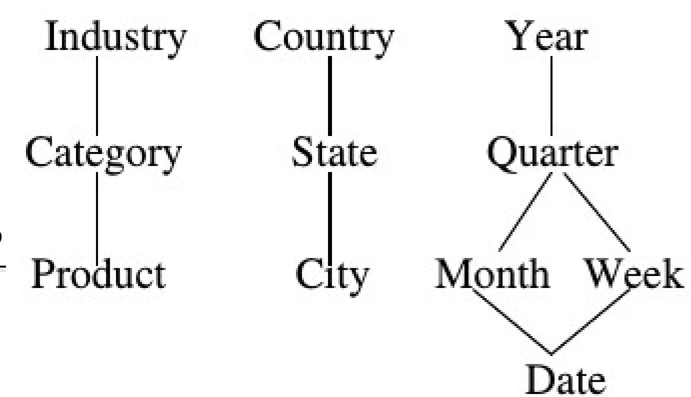
Multidimensional Data Model

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*



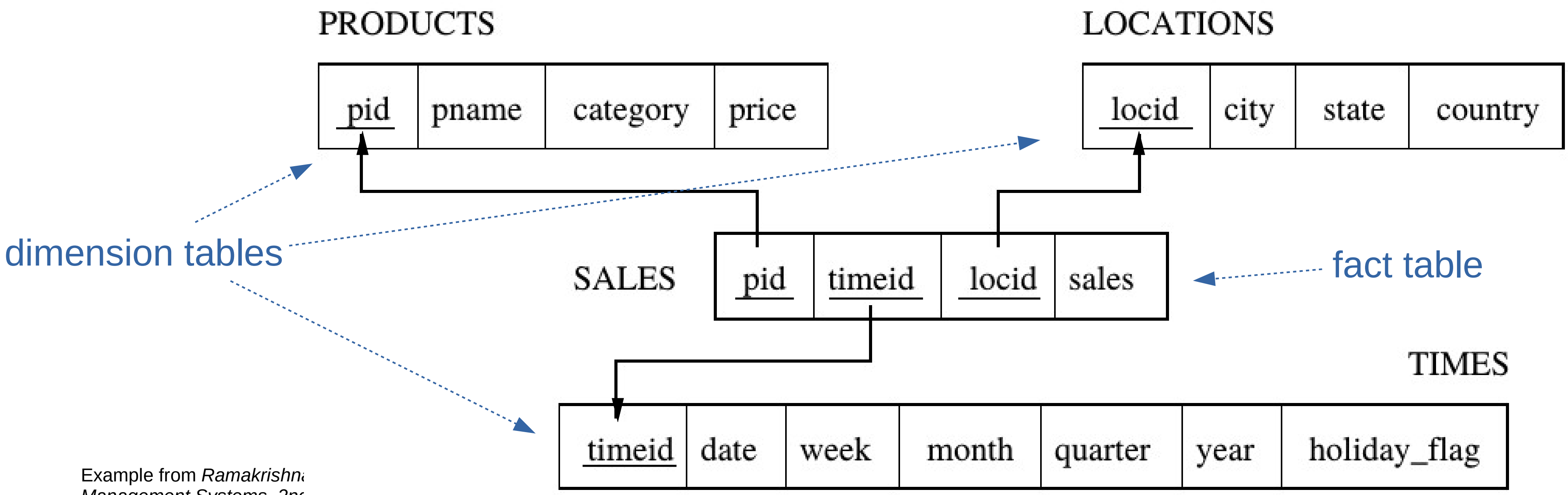
- Each dimension described by a set of attributes
  - e.g., *product* dimension may consist of *product category*, *industry of the product*, *year of introduction*, and *average profit margin*
- Some attributes may form a hierarchy of relationships



Example from *Chaudhuri and Dayal: An Overview of Data Warehousing and OLAP Technology. SIGMOD*

Multidimensional Model in an RDBMS

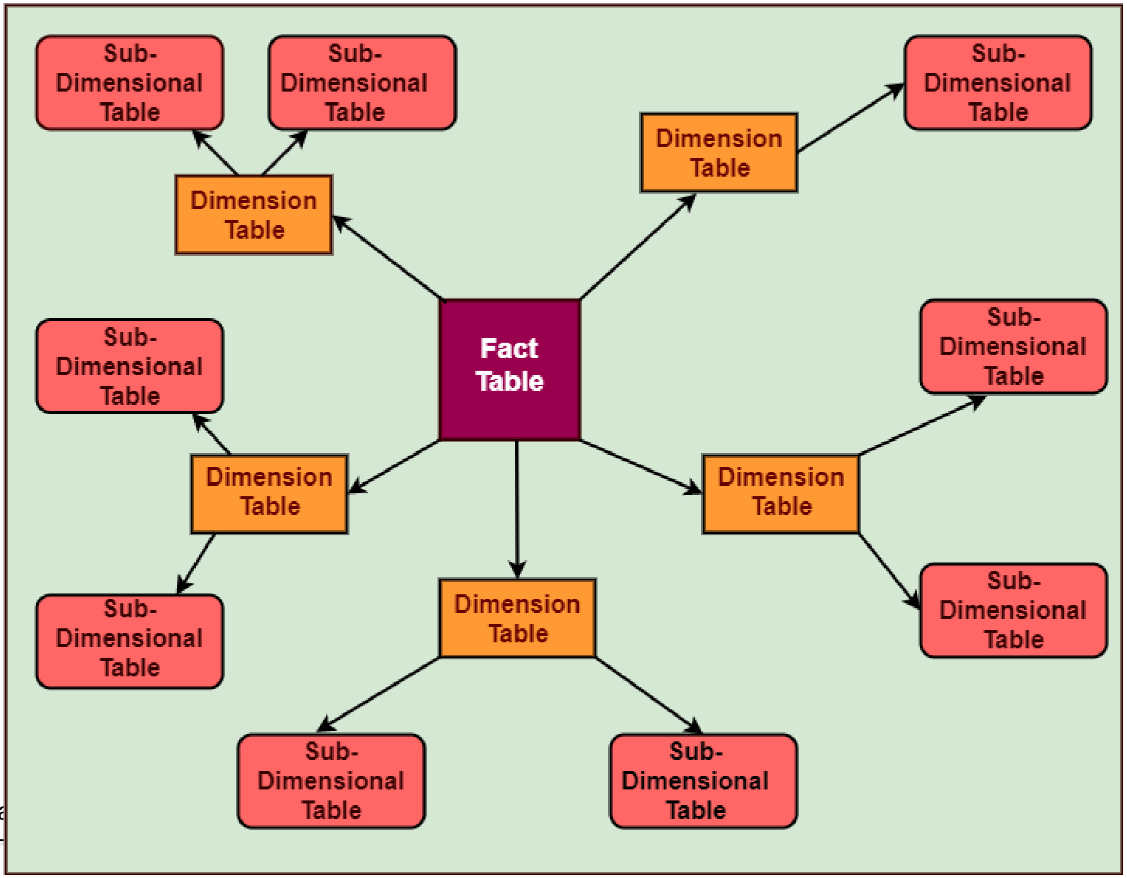
- Dimension tables with the attributes of the dimensions
- Fact table with a column for each dimension (foreign keys to the dimension tables) and for the numeric measures
  - i.e., one tuple/row per cell of the multidimensional array
- Star schema: single dimension table for each dimension



Example from *Ramakrishna* *Management Systems, 2nd*

Multidimensional Model in an RDBMS

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

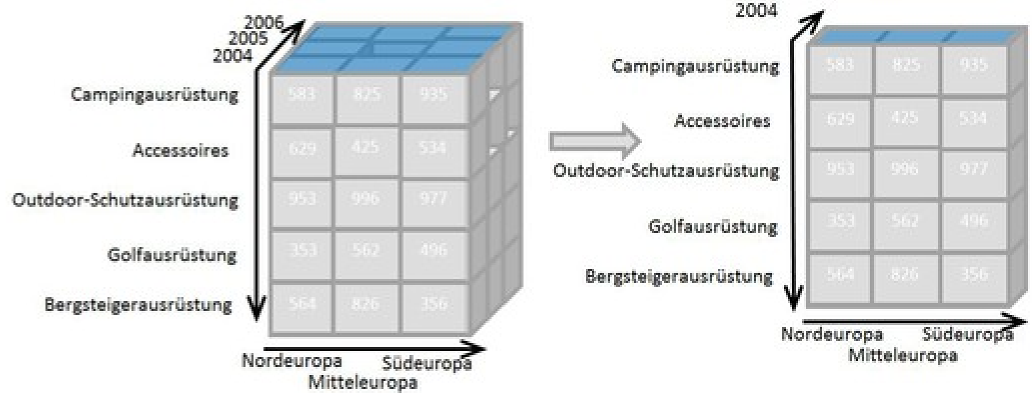


https://www.javatpoint.com/data-warehouse-star-schema-vs-snowfla

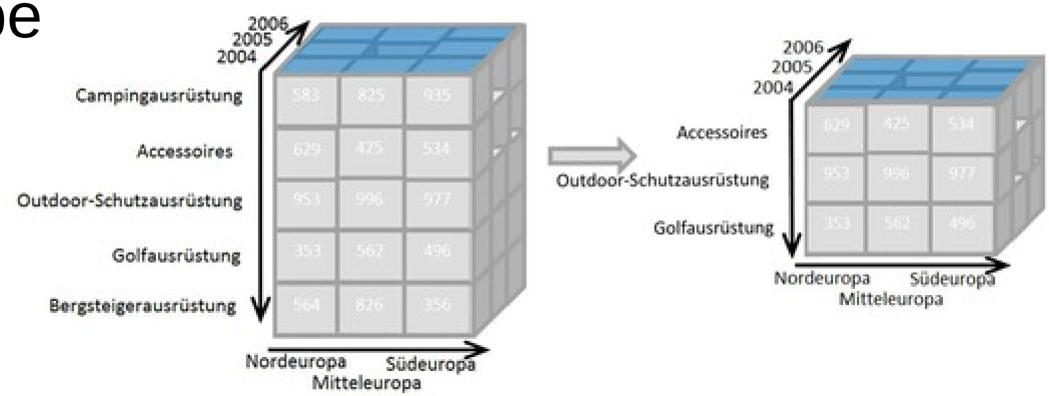
Operations over Multidimensional Data

Slicing and Dicing

- Slicing: reduce the dimensions by selecting a single value for one of the dimensions
  - like
    … WHERE *dimattr* = *xyz*



- Dicing: produce a sub-cube by selecting a range of values for one or more of the dimensions
  - like
    … WHERE *dimattr* > *xyz*
    … WHERE *dimattr* BETWEEN *x* AND *y*
    … WHERE *dimattr* IN (*x*,*y*,*z*)



Figures from Wikimedia Commons (https://en.wikipedia.org/wiki/File:OLAP_slicing.png *and* https://en.wikipedia.org/wiki/File:OLAP_dicing.png)

Roll-Up and Drill-Down

- Roll-up: aggregate the data along one or more dimensions (usually by moving up the hierarchy in these dimensions)
  - e.g., sum up by months instead of days,
    or by countries instead of cities
- Drill-down: opposite of roll-up
  - i.e., produce a more fine-grained view

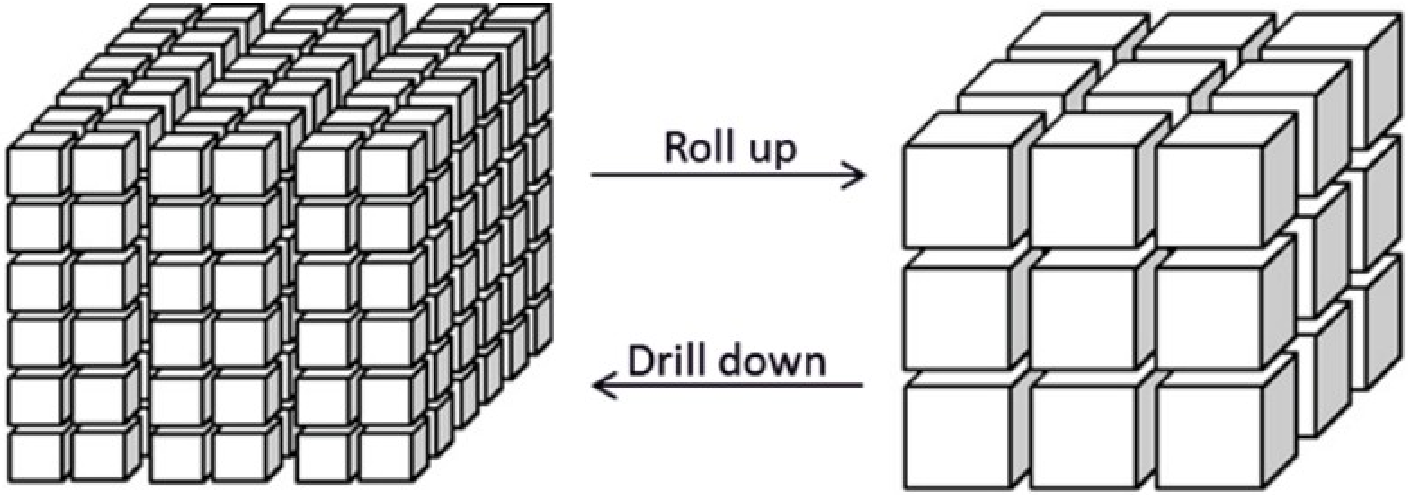


Figure from *Bolt and Van der Aalst: Multidimensional Process Mining Using Process Cubes. In BPMDS 2015.*

Pivoting

- Pivoting: rotate the cube to show a different orientation of the axes

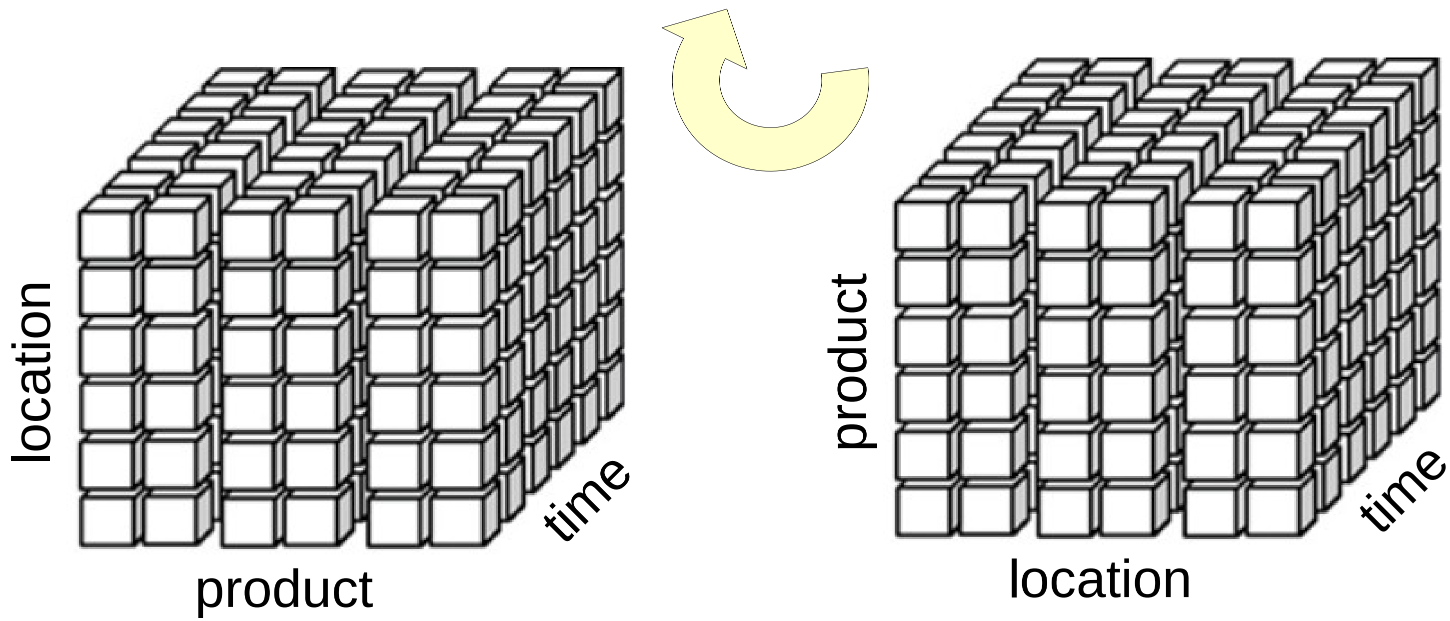
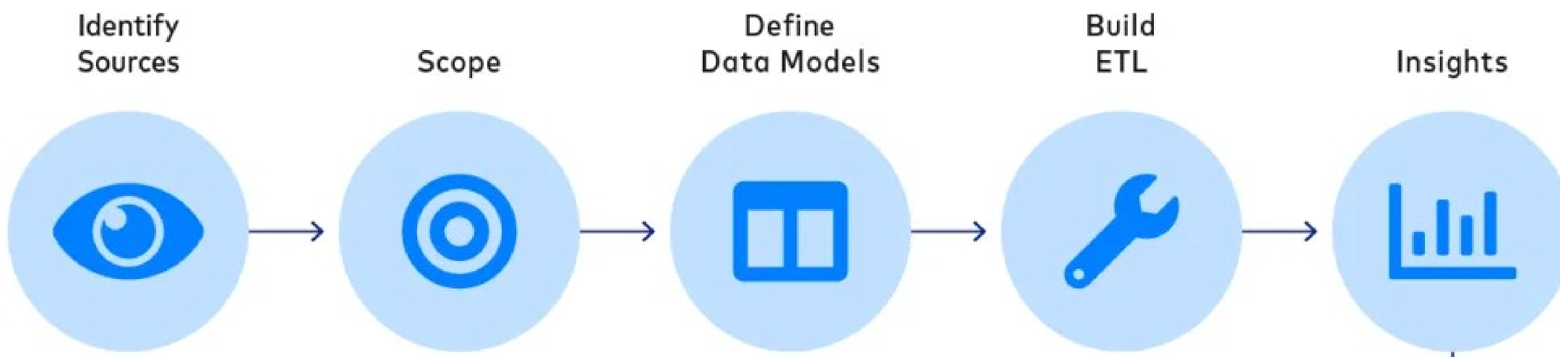


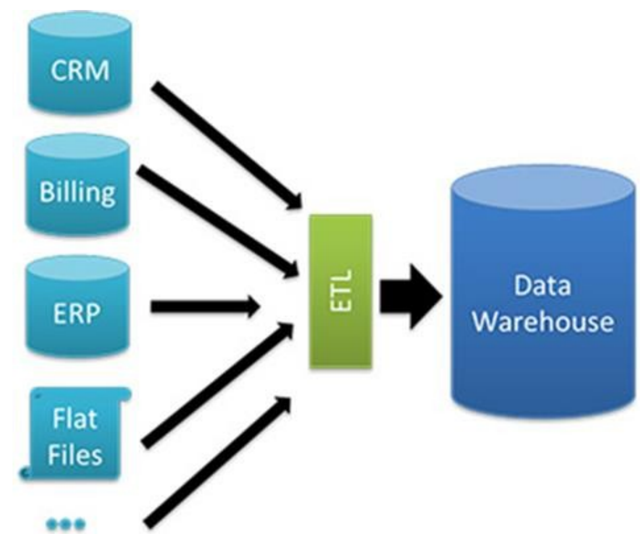
Figure from https://visibledata.wordpress.com/data/datacloud/datacube/

Building a Data Warehouse

Building a Data Warehouse

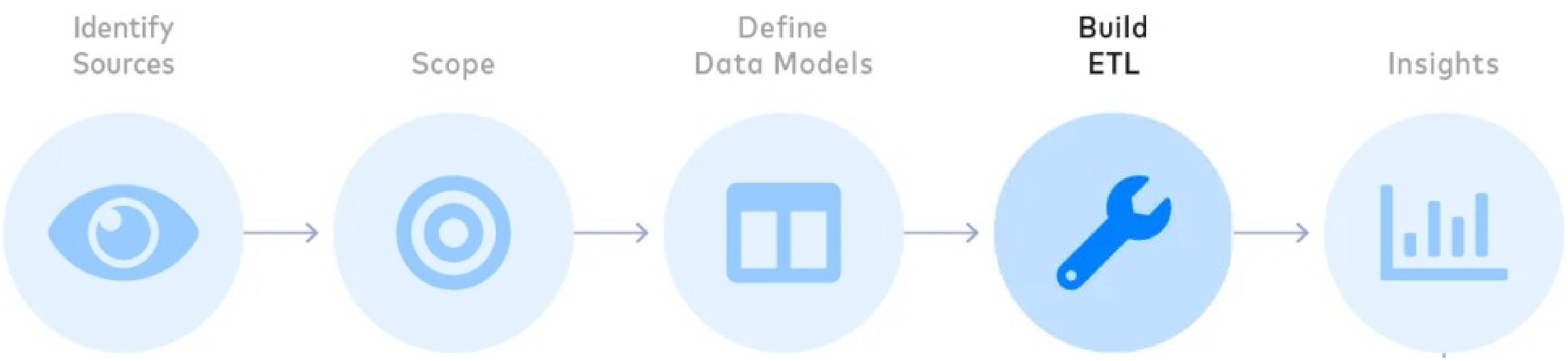


- Identify desired data sources
- Scope the analytics needs that the project is meant to solve
- Define the data model/schema that the analysts and other end users need
- Build an extract-transform-load pipeline
- Conduct analytics work, extract insights



Figures from https://fivetran.com/blog/etl-vs-elt and https://www.monitis.com/blog/top-5-data-warehouses-on-the-market-today/

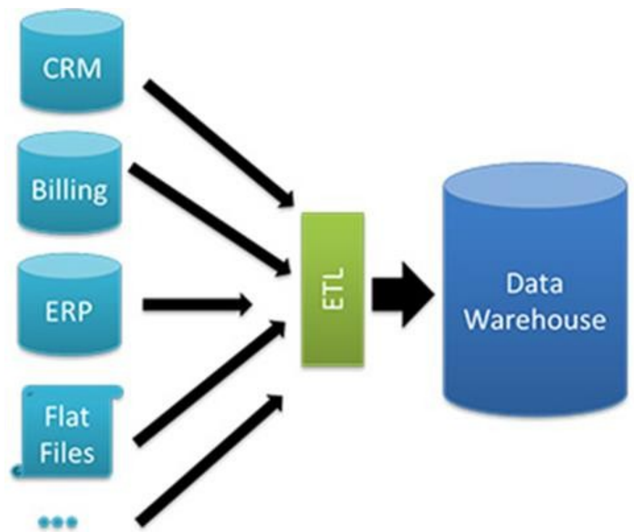
ETL



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

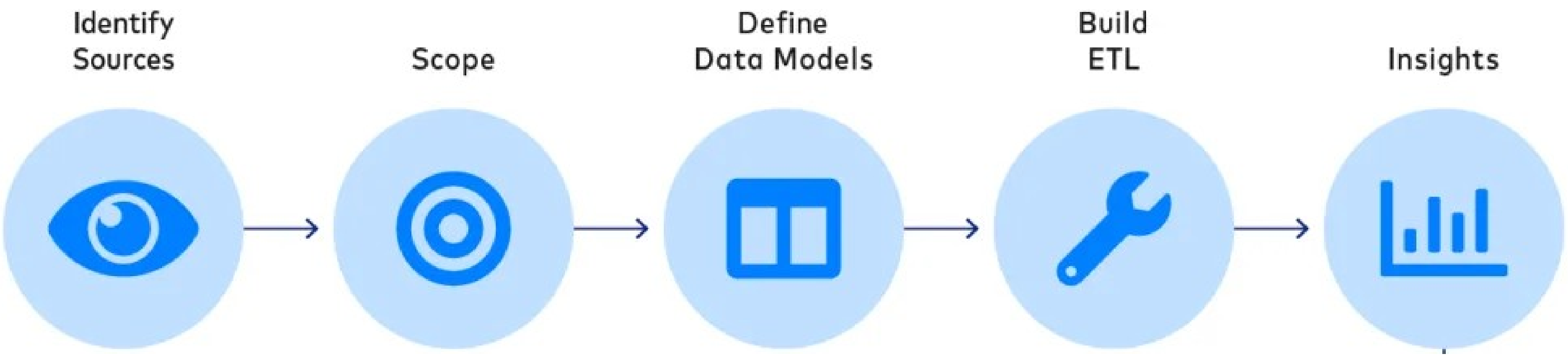


Figures from https://fivetran.com/blog/etl-vs-elt and https://www.monitis.com/blog/top-5-data-warehouses-on-the-market-today/

Data Warehouse Systems

- On premise (now also with cloud offerings)
  - Teradata
  - Oracle
  - Vertica
  - Netezza
  - Actian Vector (formerly Vectorwise)
  - SAP IQ (formerly Sybase)
  - etc.
- Cloud native
  - Snowflake
  - Amazon Redshift
  - Google BigQuery
  - Azure Synapse Analytics
  - etc.

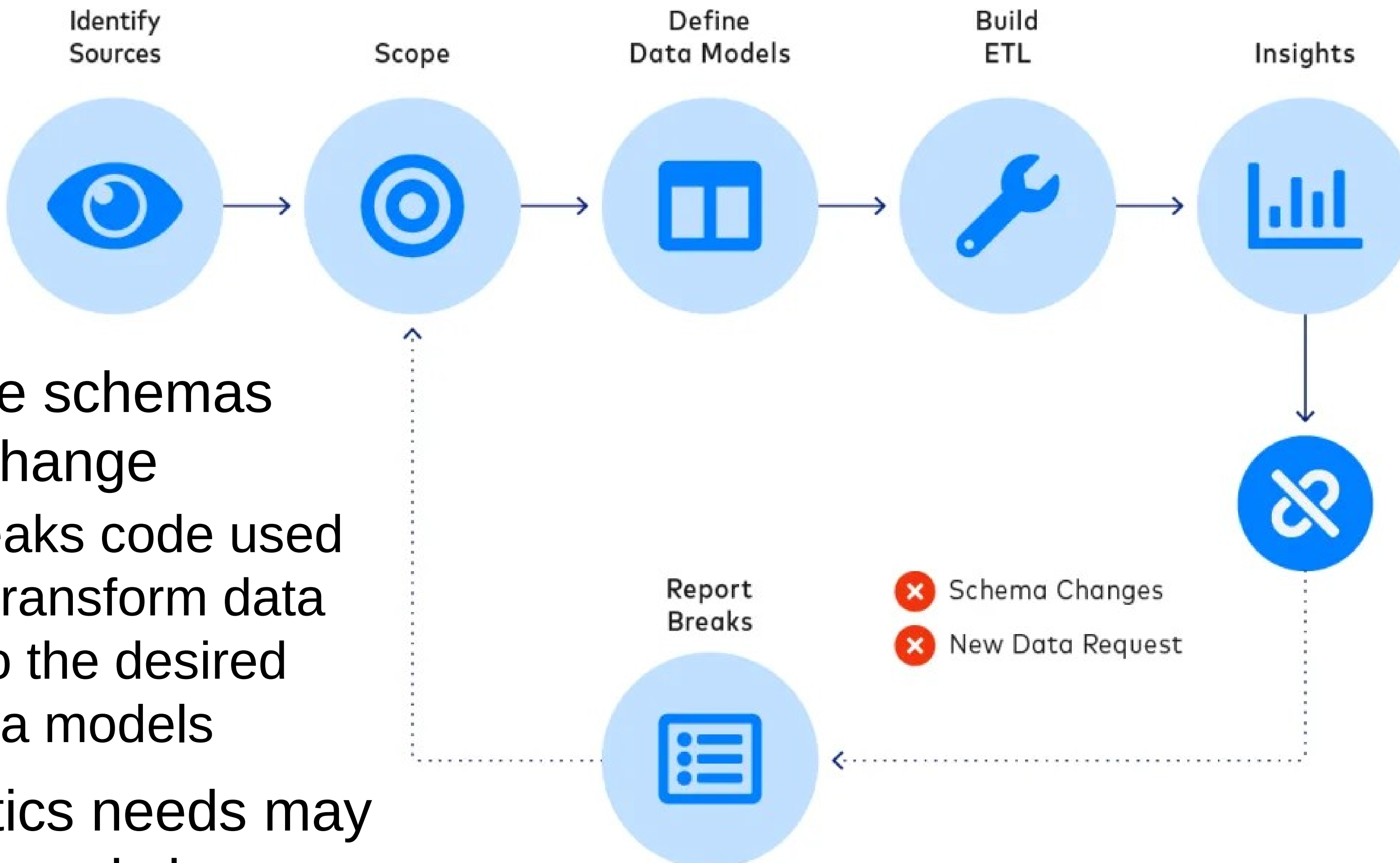
Challenges of Data Warehouses and ETL



- 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 $

Figure from https://fivetran.com/blog/etl-vs-elt

Challenges of ETL



- Source schemas may change
  - breaks code used to transform data into the desired data models
- Analytics needs may evolve and change
  - transformation code needs to be extended or rewritten

Figure from https://fivetran.com/blog/etl-vs-elt

Data Integration for Analytics in the Age of Cloud Services

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

*Quote from *Lenzerini: Data Integration: A Theoretical Perspective. PODS 2002*

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

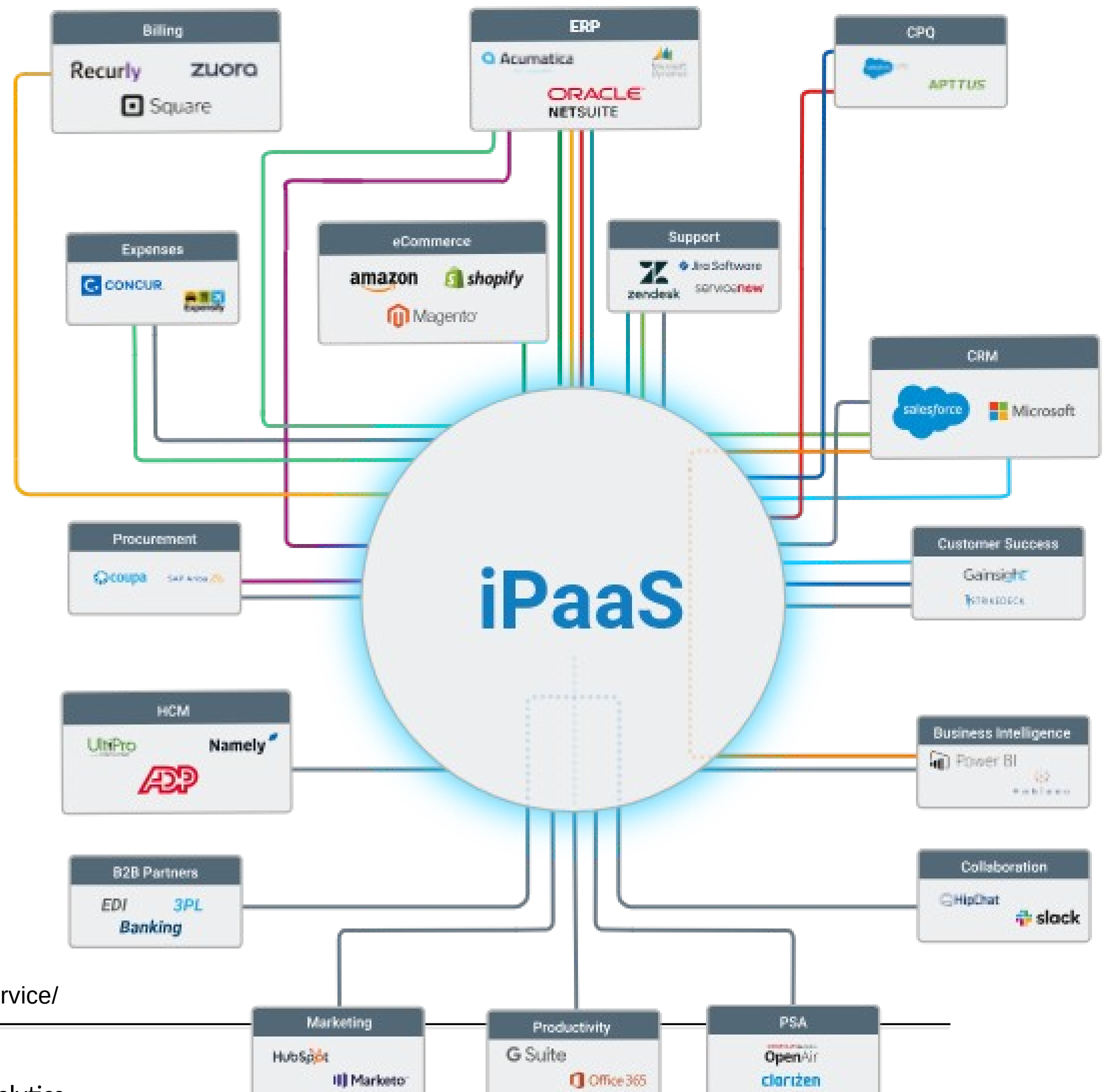


Figure from https://www.celigo.com/what-is-ipaas-integration-platform-as-a-service/

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

ELT: Extract, Load, and Transform

- 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!

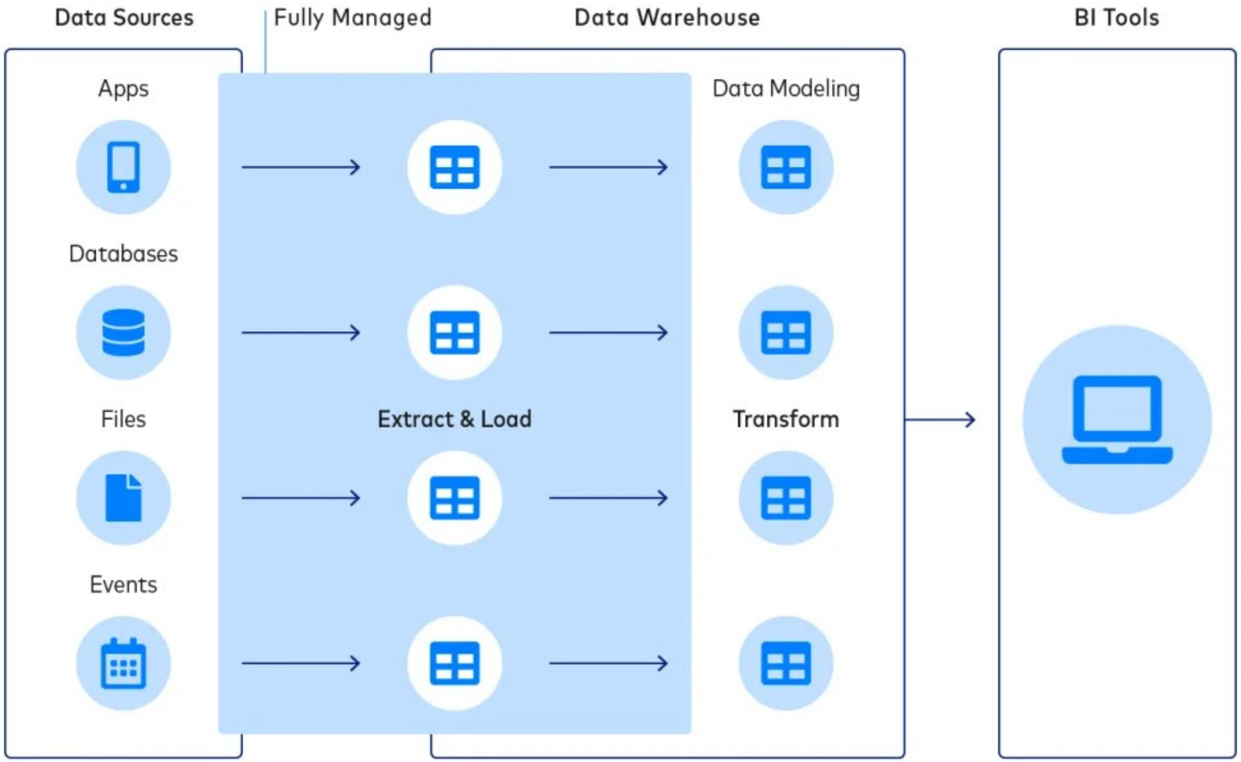
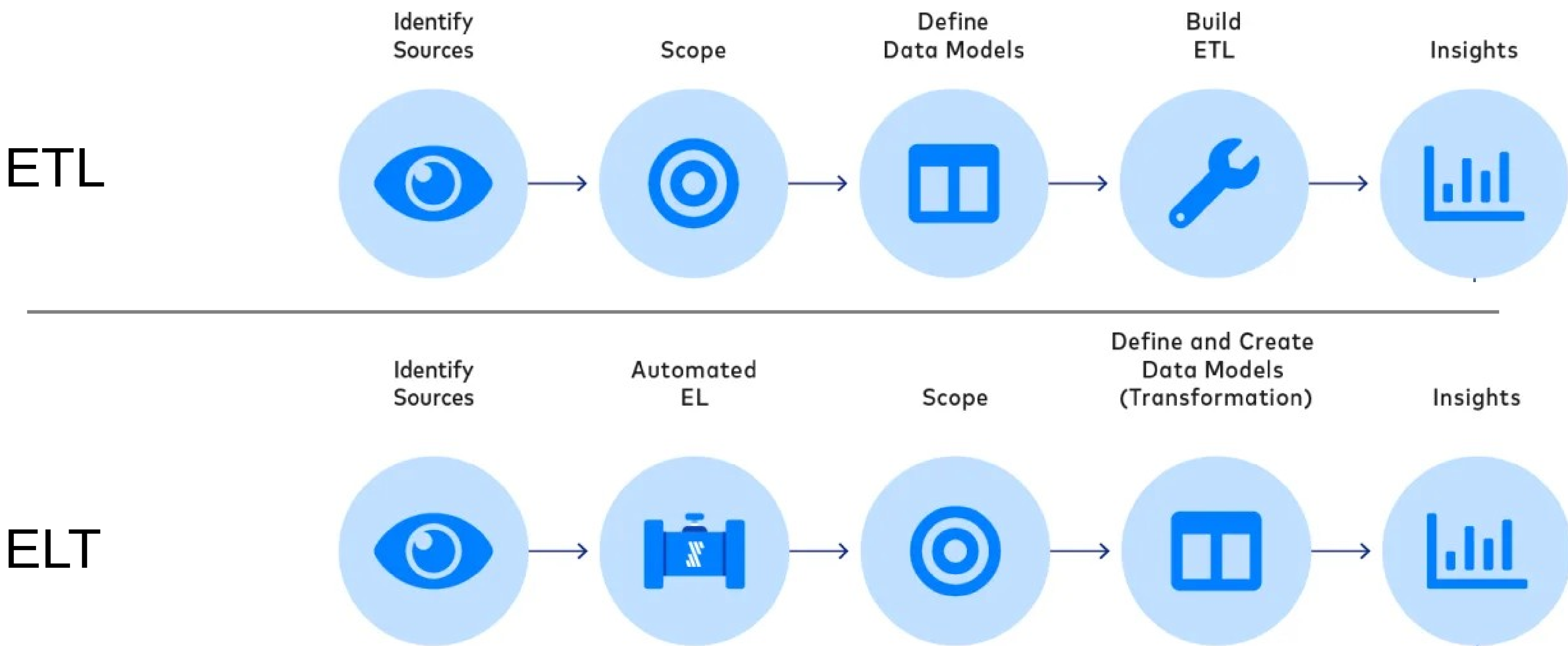


Figure from https://fivetran.com/blog/etl-vs-elt

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

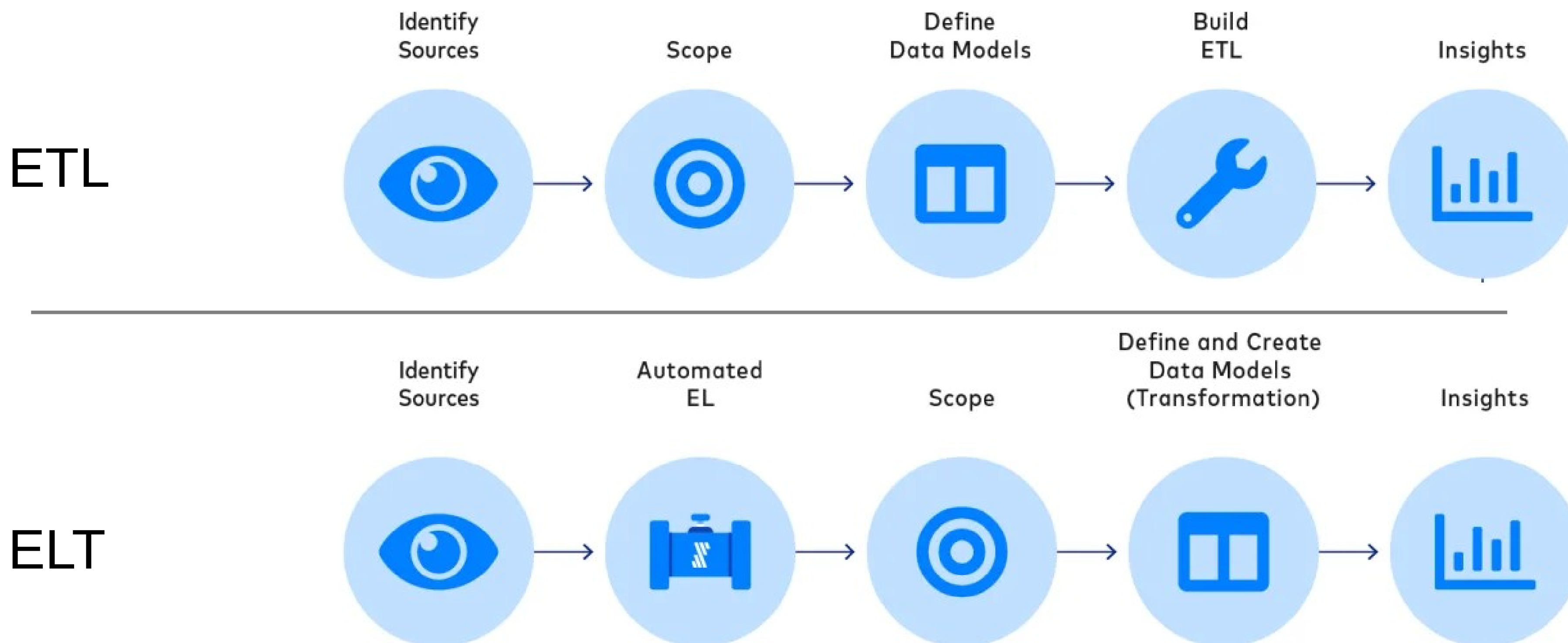
Workflows ETL versus ELT



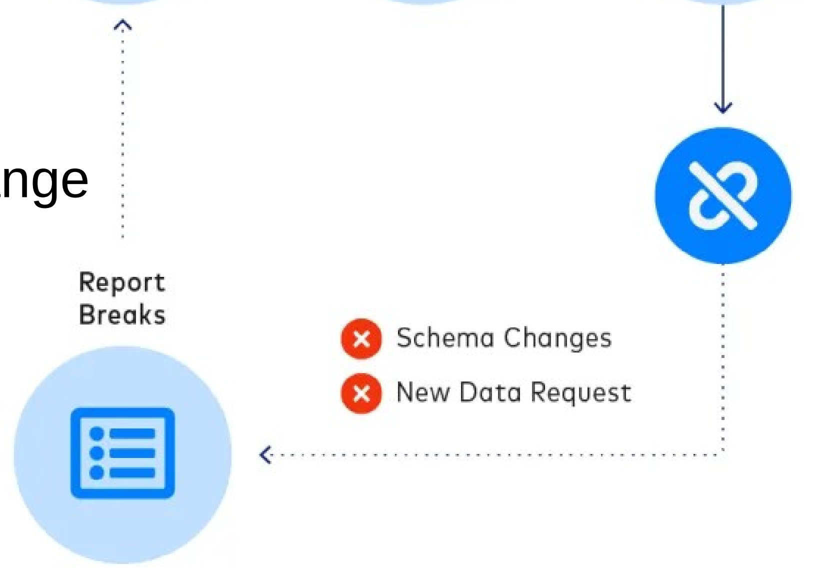
- Identify desired data sources
- Automatically extract & load (can be outsourced, scaled up and down)
- Scope the analytics needs
- Define *and create* the data model needed for the analytics work

Figures from https://fivetran.com/blog/etl-vs-elt

Workflows ETL versus ELT



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



Figures from https://fivetran.com/blog/etl-vs-elt

LINKÖPING UNIVERSITY

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

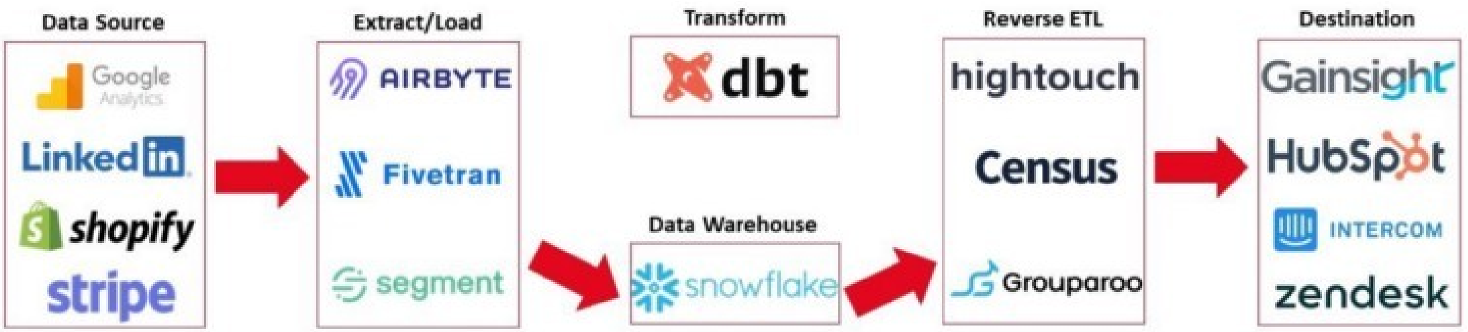


Figure from https://medium.com/memory-leak/reverse-etl-a-primer-4e6694dcc7fb

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

www.liu.se

LINKÖPING UNIVERSITY