University of Udine

Data Management for Big Data

Data Warehousing and Business Analytics

Andrea Brunello
andrea.brunello@uniud.it

Outline

Introduction

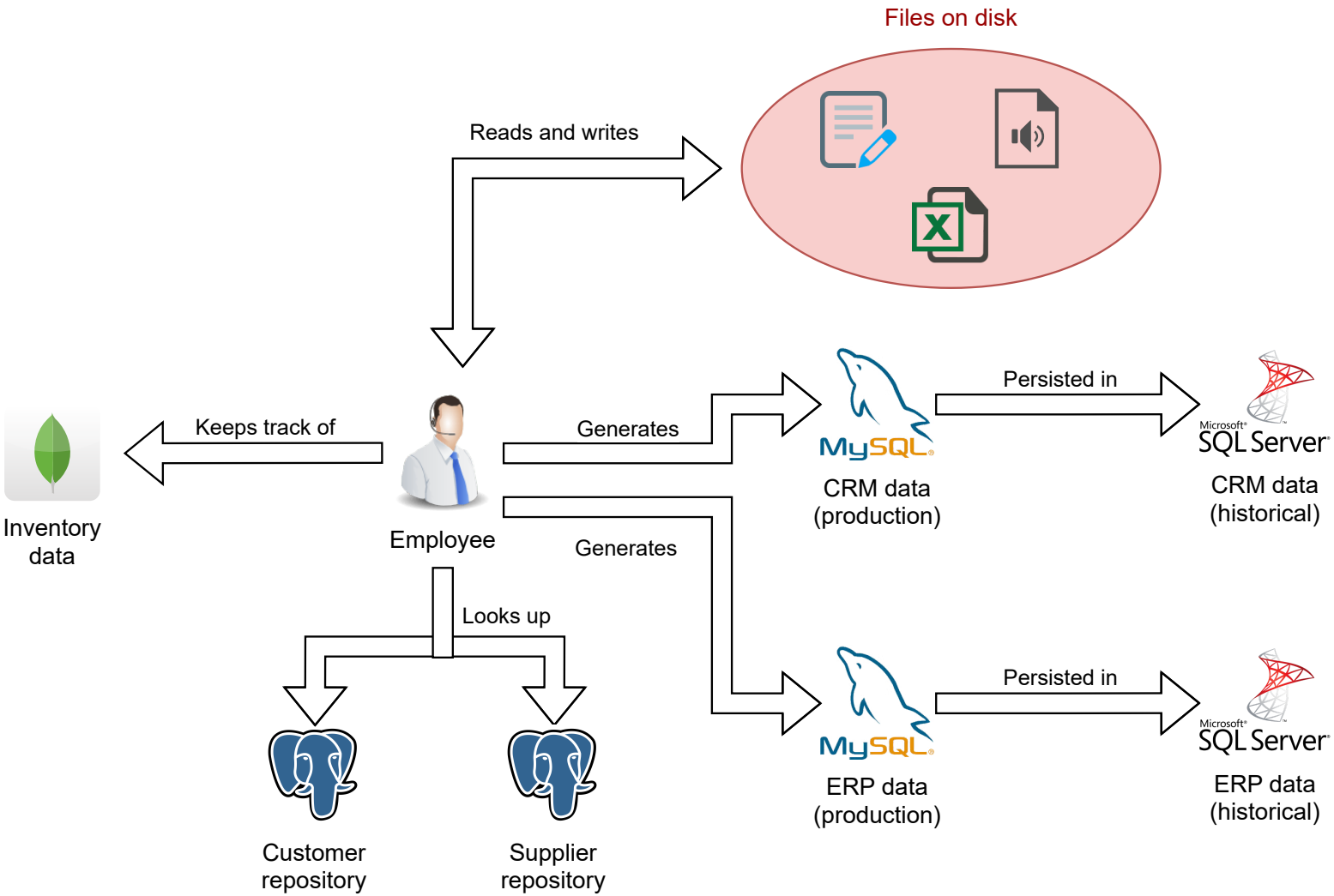
Nowadays, most of large and medium sized organizations are using information systems to implement their business processes.

As time goes by, these organizations produce a lot of (heterogeneous) data related to their business, but often these data are **not integrated**, being stored within one or more platforms.

Thus, they are hardly used for decision-making processes, though they could be a valuable aiding resource.

A **central repository** is needed; nevertheless, traditional databases are not designed to review, manage and store historical/strategic information, but deal with ever changing operational data, to support "daily transactions".

Introduction
Motivating example: A possible enterprise IT system



Introduction
What are the issues here?

Several problems:

- **Heterogeneous systems** require ad-hoc solutions for reading and writing data
- Different databases adopt **different conventions/formats** for storing the data
- Possibly (and probably) replicated and **inconsistent information**
- Difficult to perform queries and analyses involving **more than one repository**
- Some of the data are **not even considered** for analytics purposes
- Classical, operational systems are **not designed to perform complex analyses**, but to support applications and daily operations (OLTP, On Line Transaction Processing)

There is the necessity of having a **clear and uniform view** over all the company data.

This can be obtained by means of an **enterprise-wide repository**, in which information coming from different sources are brought together.

Such a repository should be explicitly designed to **support analytics tasks** (OLAP).

$\rightsquigarrow$ ***Data warehousing*** *provides solutions to these problems!*

Data Warehousing

Data warehousing is a technique for **collecting and managing data** from different sources to provide meaningful business insights.

It is a blend of components and processes which allows the strategic use of data:

- Electronic storage of a large amount of information which is designed for query and analysis instead of transaction processing
- The ultimate goal is that of transforming data into information and making it available to users in a timely manner to make a difference

Why Data Warehousing?

For example, a relational database for an inventory system has many tables related to each other through foreign keys.

A report on monthly sales information may include many joined conditions.

This can quickly slow down the response time of the query and report, especially with millions of records involved.

A data warehouse provides a new **multidimensional** design which reduces the response time and thus helps to enhance the performance of queries for reports and analytics.

Data Warehousing Fundamental Concepts

Data Warehouse

According to William Inmon, a data warehouse is a *subject-oriented, integrated, consistent, non-volatile, and time-variant collection of data* in support of management's decisions.

Thus, data warehousing is a technique for collecting and managing data originating from multiple sources, to provide meaningful business insights.

The data warehouse makes it easier to perform analysis tasks:

- Single, integrated, source of truth
- It typically poses at the heart of a **decision support system**, i.e., an information system that supports business or organizational decision-making activities

Subject-oriented

The data warehouse focuses on enterprise-specific **concepts**, as defined in the high-level corporate data model. Subject areas may include:

- Customer
- Product
- Order
- Claim
- Account

Conversely, operational databases hang on enterprise-specific *applications*, meaning that data in them is typically organized by business processes, around the workflows of the company.

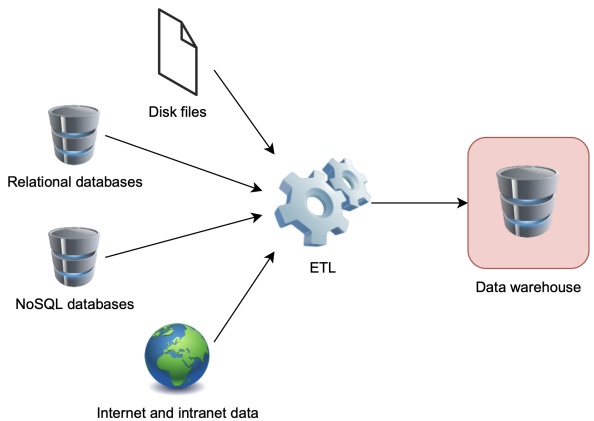
Integrated and Consistent

Data is fed from multiple, disparate sources into the data warehouse.

As the data is fed, it is converted, reformatted, resequenced, summarized, and so forth (**ETL** – Extract, Transform, Load).

Data is entered into the data warehouse in such a way that the many inconsistencies at the operational level are resolved.

Consistency applies to all application design issues, such as naming conventions, key structure, measurement of attributes, and physical characteristics of data.



Extract: data is gathered from multiple, heterogeneous sources

Transform:

- *Data cleansing*: removal of errors and inconsistencies
- *Data integration*: reconcile data on the same item coming from different sources
- *Data aggregation*: transform/aggregate data to match the data warehouse schema

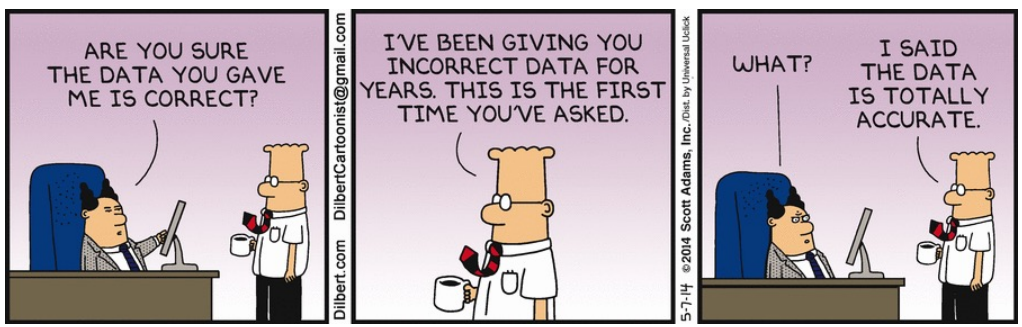
Load: initial bulk load, and subsequent continuous feed

Before processing all the dirty data, it is important to determine the cleansing and integration cost for every dirty data element (part of the **data quality process**).

Everyone would like to have all the data clean, but it is also a **matter of cost and time**.

Nevertheless, always remember:

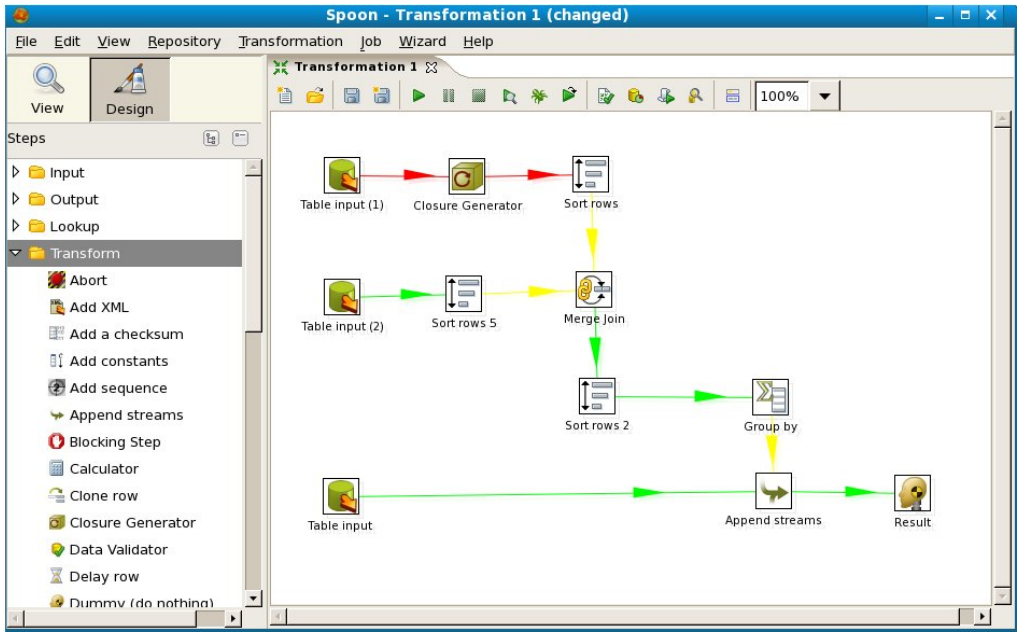
- Garbage in = garbage out
- Your analyses are as good as your data



QuerySurge is built specifically to automate the testing of Data Warehouses & ETL processes.

MarkLogic is a NoSQL data warehousing solution that includes a fully-fledged data integration and data management solution.

Pentaho Data Integration → *Talend Open Studio* → *Talend Stitch* support the creation of complex ETL workflows.



Non-volatile

After the data is inserted in the warehouse it is **neither changed nor removed**.

The only exceptions happen when false data is inserted or the capacity of the data warehouse is exceeded and **archiving** becomes necessary.

This means that data warehouses can be essentially viewed as **read-only databases**.

When subsequent changes occur, a new **snapshot** record is written. In doing so, a historical record of data is kept in the data warehouse.

Time-variant

Time variancy implies that the warehouse stores data representative as it existed at **many points in time in the past**.

A time horizon is the length of time data is represented in an environment; a **5-to-10-year** time horizon is normal for a data warehouse.

While operational databases contain current-value data, data warehouses contain **sophisticated series of snapshots**, each snapshot taken at a specific moment in time.

OLTP: On-Line Transaction Processing

OLTP queries are typical of operational, daily systems.

Such queries generally read or write a **small number of tuples**, executing transactions over **detailed data**.

A typical OLTP transaction in a banking environment may be the transfer of money from one account to another.

Always enforcing data consistency and handling concurrency aspects is essential for this kind of applications, because otherwise money may for example get lost or doubled.

"On-line" means that the analyst should obtain a response in almost real time.

OLAP: On-Line Analytical Processing

On the contrary, the type of query generally executed in data warehouses is OLAP.

Here, the user is interested in performing **read-only operations aggregating historical** data over large datasets.

E.g., calculate the average amount of money that customers under the age of 20 withdrew from ATMs in a certain region over the years 2018–2020.

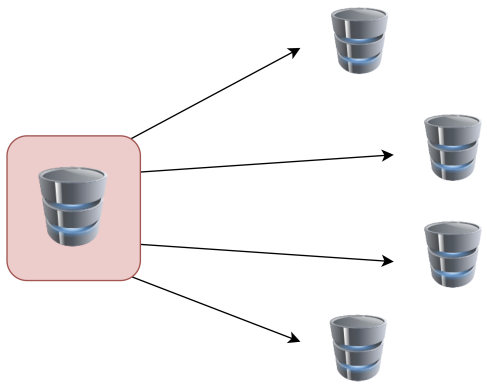
Such operations are typically complex and time consuming
$\rightsquigarrow$ **Multidimensional model**.

Data Mart

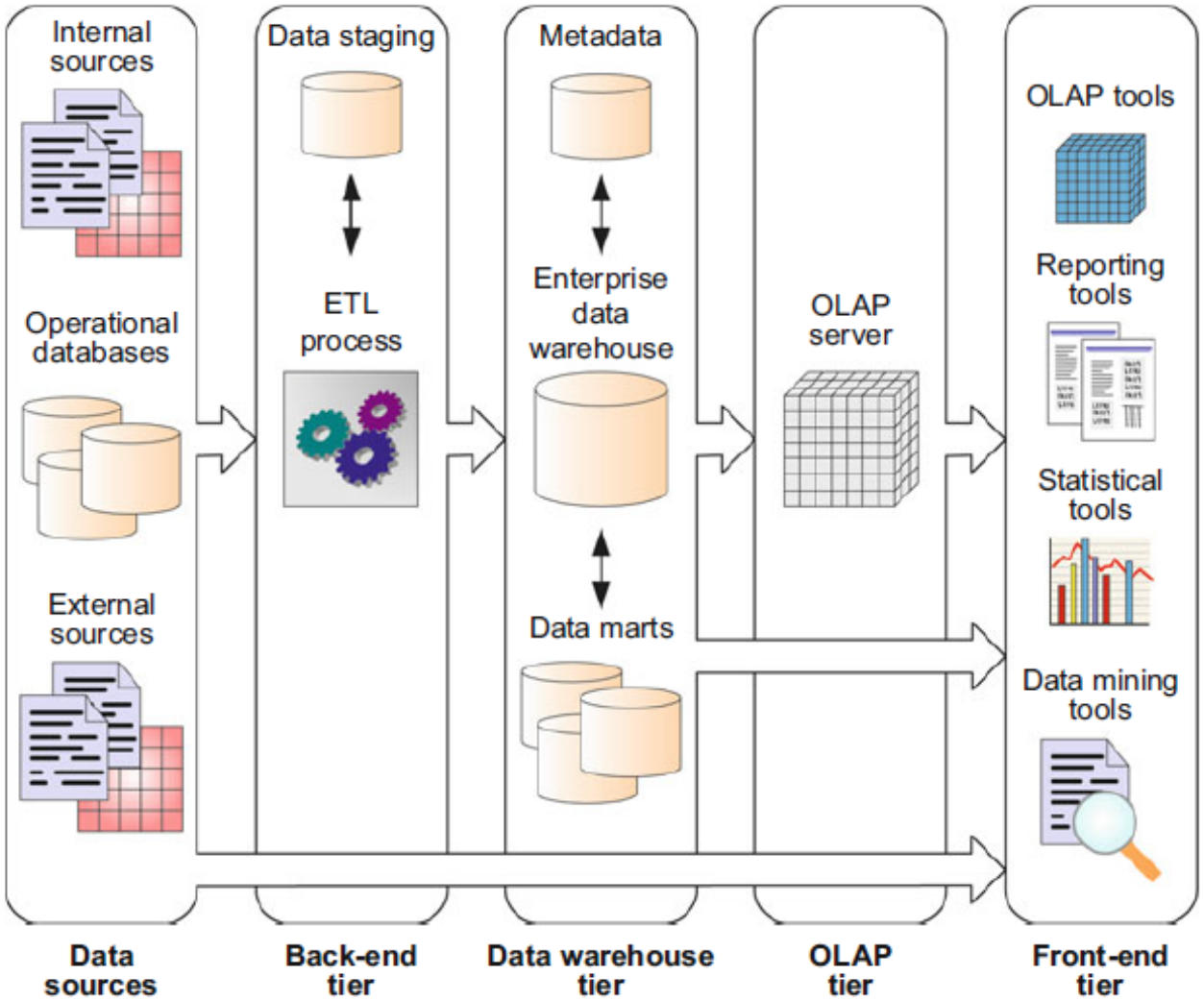
A data mart is focused on a **single functional area** of an organization and contains a subset of data stored in a Data Warehouse.

A data mart is a condensed version of Data Warehouse and is designed for use by a specific department, unit or set of users in an organization.

Data marts are **smaller in size and are more flexible** compared to a Data Warehouse.



Data Warehouse Architecture Schema



Data Warehouse Architecture Schema

A modern general data warehouse architecture typically consists of several tiers:

- **The back-end tier** includes extraction, transformation, and loading (ETL) tools and a data staging area
- **The data warehouse tier** is composed of an enterprise data warehouse and/or several data marts and a metadata repository (e.g., schema definitions, data lineage)
- **The OLAP tier** is composed of an OLAP server, which provides a multidimensional view of the data
- **The front-end tier** is used for data analysis and visualization. It contains client tools such as OLAP tools, reporting tools, statistical tools, and data mining tools

Data Lake

A Data Lake is a repository that can store large amounts of structured, semi-structured, and unstructured data.

- **Data is kept in its native format**, no limits on size or type
- It allows to access data before the ETL process, thus it retains all data coming from the sources
- Data is only transformed when the user is about to use it (**schema on read**, vs. schema on write in the warehouse)
- Storing information in a data lake is relatively inexpensive with respect to storing them in a data warehouse

A Data Lake is not a substitute for a Data Warehouse and, if not properly managed, can easily become a **data swamp**!

The Multidimensional Model

The Multidimensional Model

The distinctive features of OLAP applications suggest the adoption of a **multidimensional representation of data**, since running analytical queries against traditionally stored information would result in complex query specification and long response times.

The key idea is that of **pre-aggregating** some of the data.

The multidimensional model relies on the concepts of **fact**, **measure**, and **dimension**.

Facts and Measures

In a data warehouse context, a *fact* is the part of your data that indicates a **specific event or transaction** that has happened, like the sale of a product, or receiving a shipment.

A fact is composed of multiple numerical **measures**, that describe it.

As an example, a fact may be receiving an order for some shoes, detailed by the measures 'price' and 'quantity'.

Dimensions

Dimensions provide a way to **categorize/label/index facts**, e.g., considering spatial or temporal aspects.

Thus, they allow to filter and group facts.

The previous order may be detailed by the following 2 *measures* and 3 *dimension attributes*:

- total amount US$ 750
- quantity purchased is 10
- received yesterday at 2 pm
- served by our store in New York
- placed by customer #XAZ19

Data Hypercubes

In the multidimensional model, data is represented in an *n*-dimensional space, usually called **a data cube or hypercube**.

A data cube is defined by dimensions (cube edges) and facts (cube cells):

- Dimensions are perspectives used to analyse the data
- Facts have related numeric values, the measures

Data cubes can be sparse: there may not be a cell value for each combination of dimensions.

2–D data in a spreadsheet: Pivot tables

Bi-dimensional pivot table, that considers:

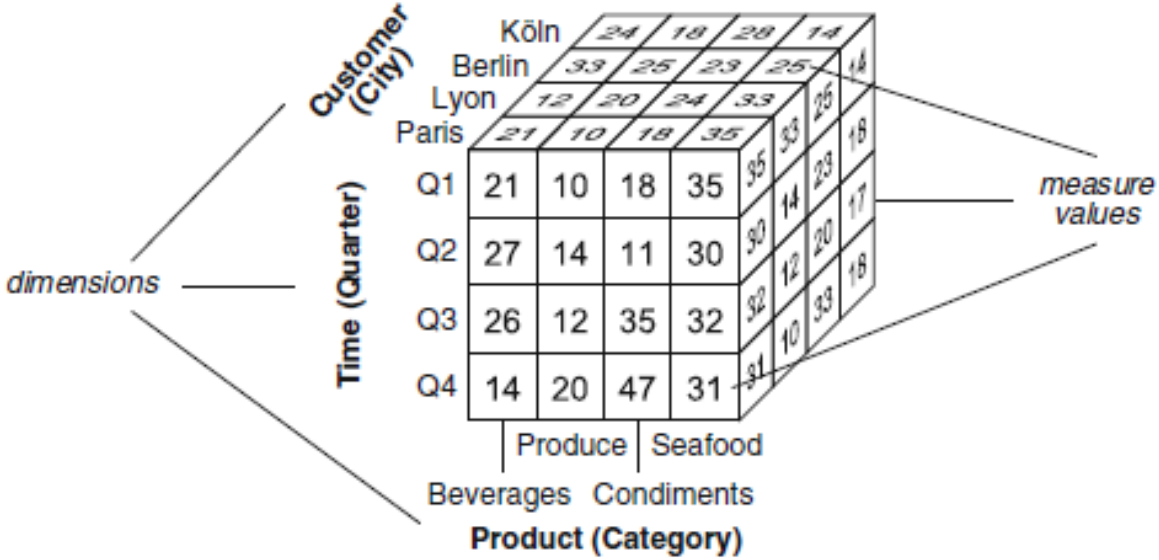
- Measure "Amount"
- Dimensions "Place" and "Product"
- Facts are the amounts of products sold in each country

	A	B	C	D	E	F	G	H	I	J
1	Category	(All)								
2										
3	**Sum of Amount**	**Column**								
4	**Row Labels**	**Apple**	**Banana**	**Beans**	**Broccoli**	**Carrots**	**Mango**	**Orange**	**Grand Total**	
5	Australia	20634	52721	14433	17953	8106	9186	8680	131713	
6	Canada	24867	33775		12407		3767	19929	94745	
7	France	80193	36094	680	5341	9104	7388	2256	141056	
8	Germany	9082	39686	29905	37197	21636	8775	8887	155168	
9	New Zealand	10332	40050		4390			12010	66782	
10	United Kingdom	17534	42908	5100	38436	41815	5600	21744	173137	
11	United States	28615	95061	7163	26715	56284	22363	30932	267133	
12	**Grand Total**	**191257**	**340295**	**57281**	**142439**	**136945**	**57079**	**104438**	**1029734**	
13										

The pivot table stores aggregated data; here, sums.

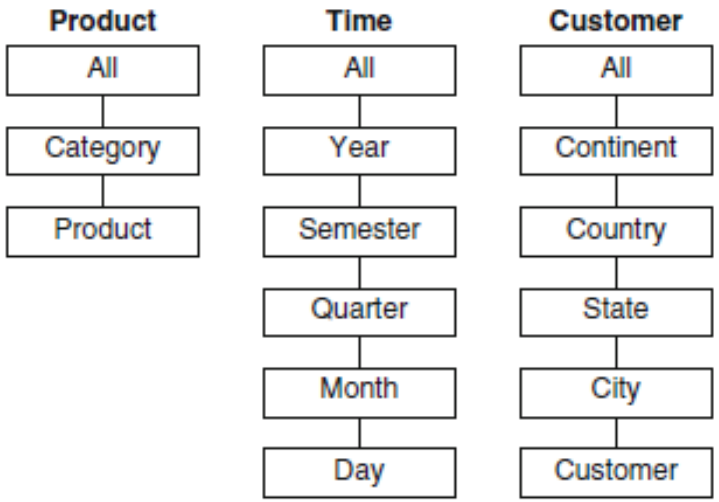
3-dimensional OLAP Cube Example

Amount of products sold in each quarter in different cities



Dimension Hierarchies

To extract strategic knowledge from a cube, it is necessary to view its data at **several levels of detail**.



In practice, OLAP cubes are implemented by **pre-aggregating** the data at the maximum level of detail allowed by the dimension hierarchies.

Operations over Multidimensional Data

OLAP operations

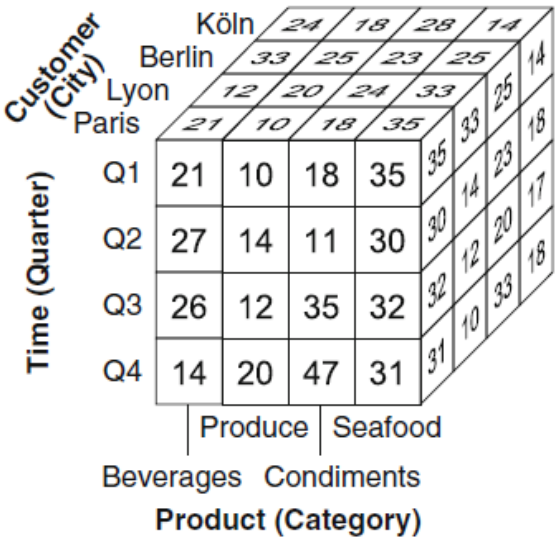
The four types of analytical operations performed on OLAP cubes are:

- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

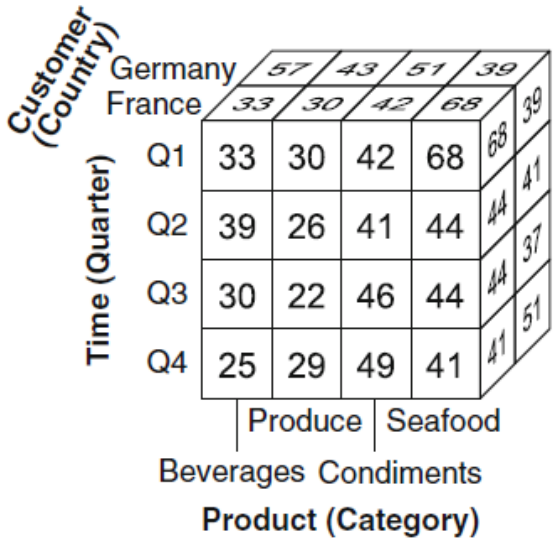
"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Roll-up

It involves summarizing the data along a chosen dimension (e.g., sum), navigating from a finer level of detail (down) to a coarser one (up).



(a) Original

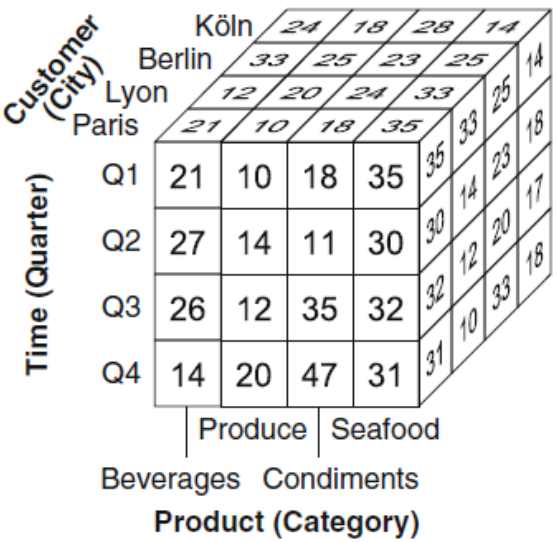


(b) Roll-up to the Country level

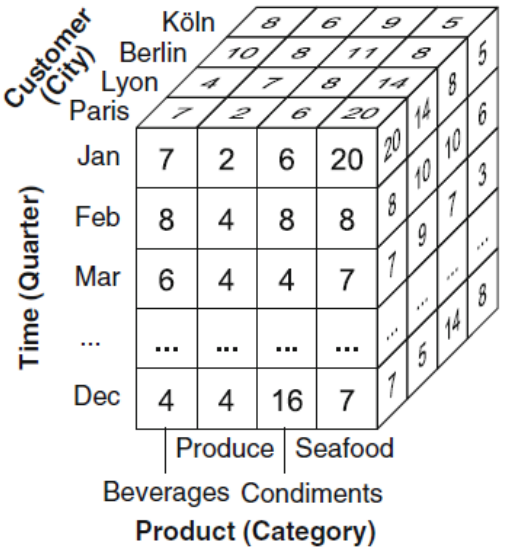
Drill-down

It allows the user to navigate among levels of data, ranging from the most summarized (up) to the most detailed (down), along a given hierarchy.

Looking at the detail beneath a summary number may be useful, especially where the summary number is surprising.



(c) Original



(d) Drill-down to the Month level

Roll-up and Drill-down
Aggregation of Measures

To allow roll-up and drill-down operations, each measure in a cube is associated with an **aggregation function**.

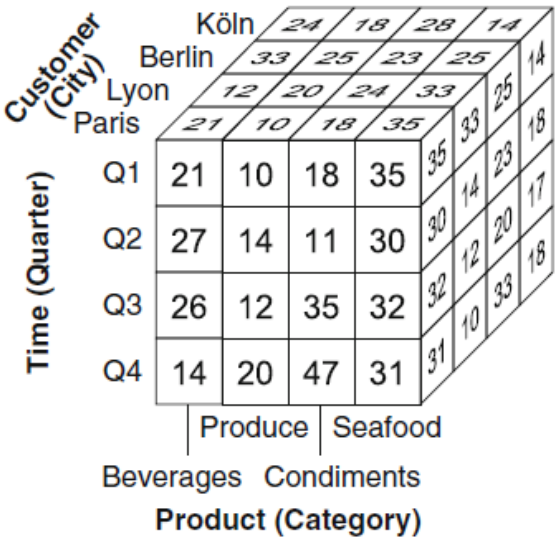
The aggregation function tells how to combine on-the-fly several measure values into a single one, or how to split a single value into multiple ones, when the hierarchies of the dimensions are being traversed.

As we mentioned, the cube holds the data at the finest level of granularity allowed by the dimensions.

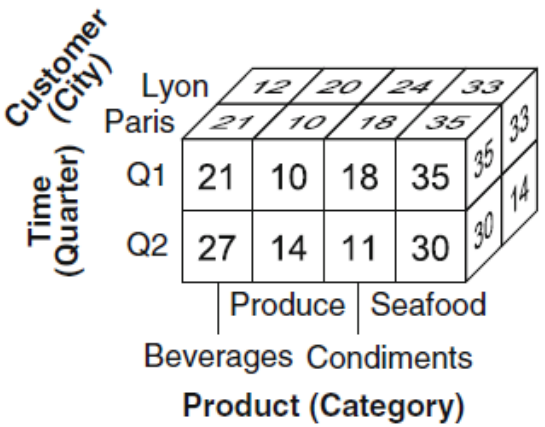
Pay attention to the aggregation functions that you apply to each measure while navigating the hierarchies (e.g., mind the difference between *sum* and *count*).

Slice and Dice

It is the act of picking a subset of a cube by fixing one or more values for one or more of its dimensions.



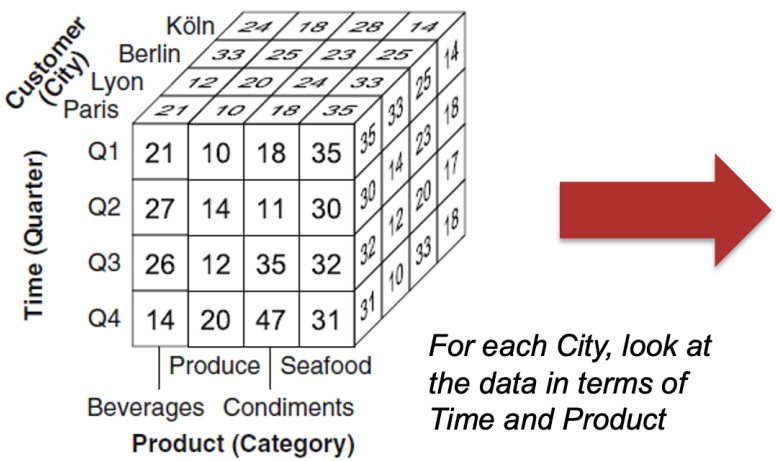
(e) Original



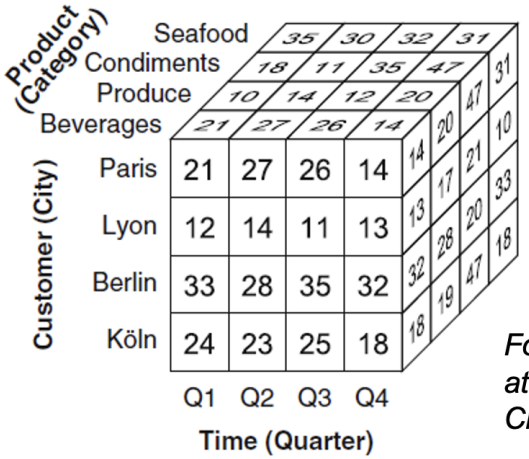
(f) Dice on City='Paris' or 'Lyon' and Quarter='Q1' or 'Q2'

Pivot

This operation allows an analyst to rotate the cube in space to see its various faces.



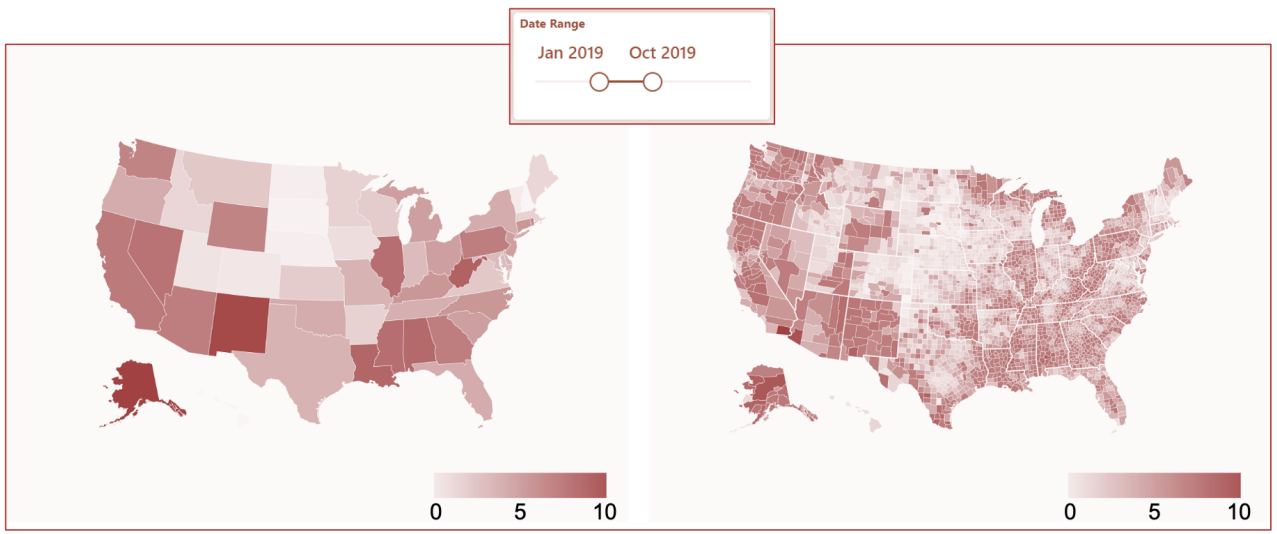
For each City, look at the data in terms of Time and Product



For each product, look at the data in terms of City and Time

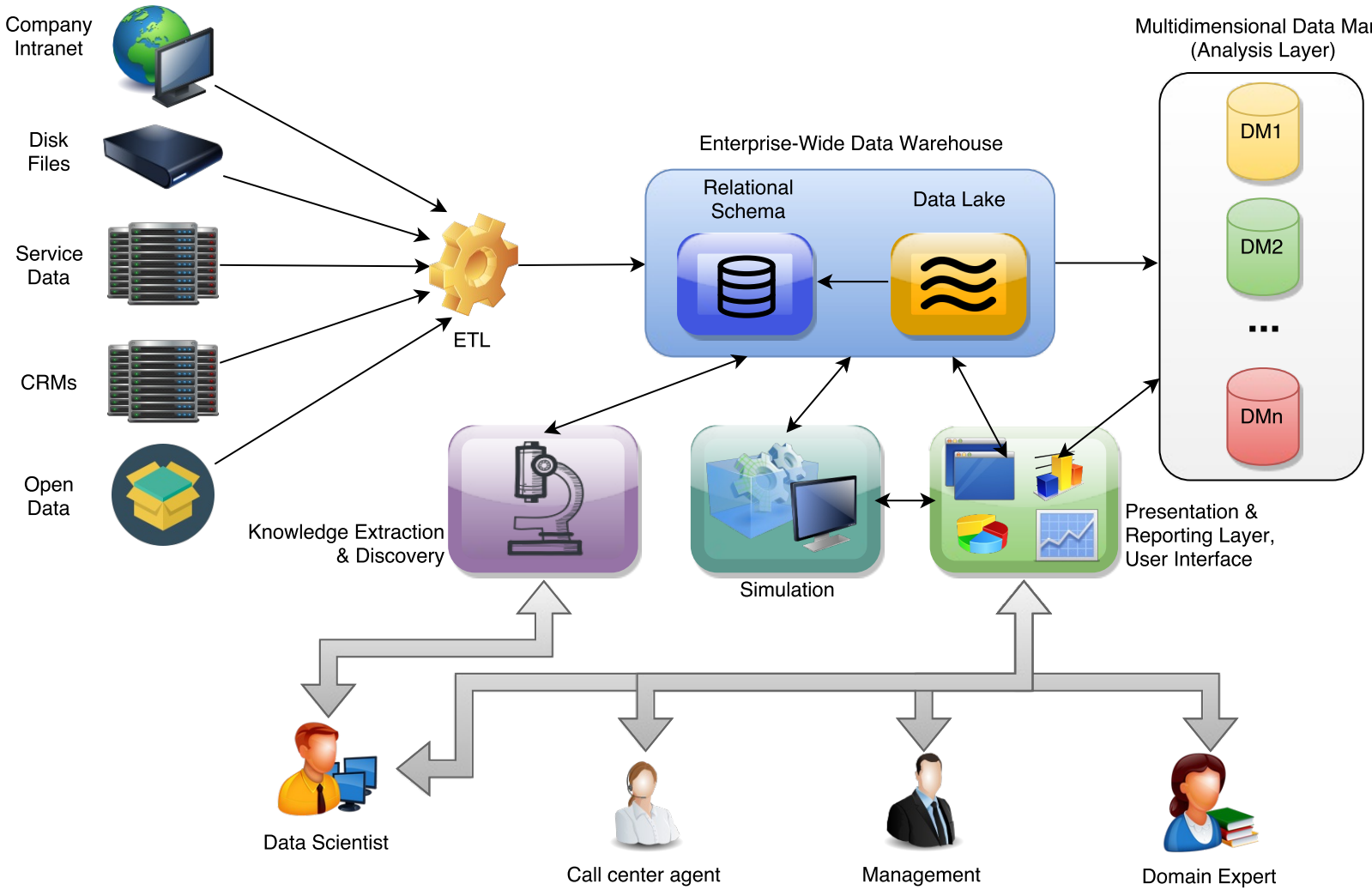
Hypercubes and representations

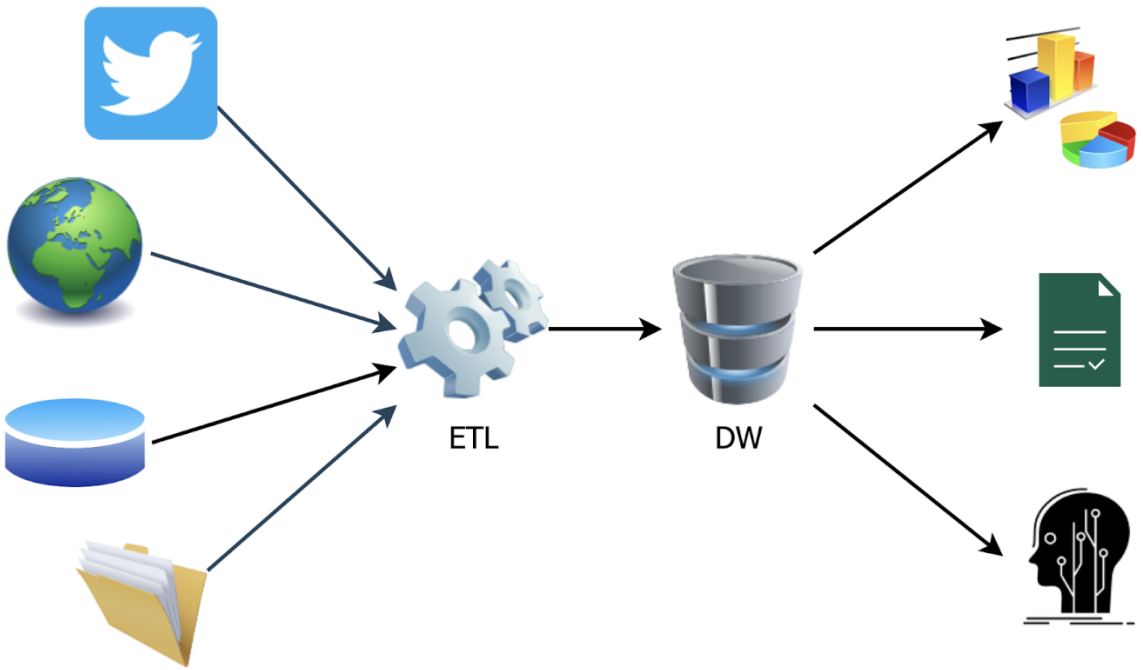
Note that hypercubes, like pivot tables, are just an **intuitive representation** of how the data are pre-aggregated and arranged within the system. The information can then be conveyed to the user in several manners, e.g., relying on different (interactive) graphs



Total number of sales in the US (tens of thousands), drill-down operation over the spatial dimension

A possible enterprise IT system
With Data Warehousing




ETL
DW

Data monetization

Collecting data is important but, without analyses, there is no value from them.

Creating value from data is also referred to as **data monetization**.

Data analyses may be performed by means of suitably designed *Business Intelligence* and *Business Analytics* systems.

There are three main analysis types to extract value from data: **descriptive analytics**, **predictive analytics** and **prescriptive analytics**.

Not only analyses. . . sometimes data monetization pertains just selling data (e.g., by social networks, companies).

Descriptive Analytics

It is used in almost every company (Business Intelligence).

It is focused on **historical or current data**, and makes use of OLAP and statistical techniques to analyse/summarise data.

E.g., drill down, roll up, slicing and dicing, and pivoting operations performed on data cubes.

Typical questions:

- How many customers were lost to the competition in the last year?
- What customers are likely to be committing a fraud regarding their currently opened claims (= which customers are currently deviating from their usual behaviour)?

Predictive Analytics

Predictive analytics aims at developing a **vision of the future** making use of past data (Business Analytics).

It relies on statistical analyses and machine learning, other than proper data modelling, pre-processing and querying.

Typical questions:

- What will the churn rate of customers be in the next three years?
- What customers are most likely to commit a fraud regarding their future claims?

Prescriptive Analytics

It has roots in predictive tasks, but it goes further.

It goes beyond predicting future outcomes, by also **suggesting** actions to benefit from the predictions and **showing the implications** of each decision option.

In the churn example:

- What are the factors that mostly influence the probability of churn?
- How should I act on such variables so to achieve a desired future churn rate?

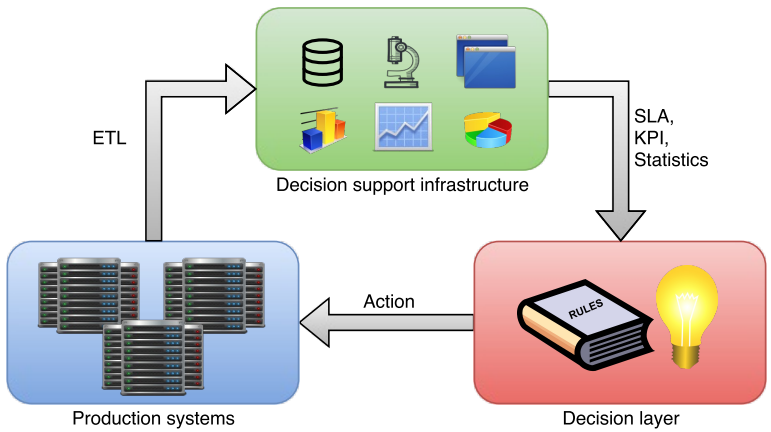
From DSSs to DMSs

A Decision Management System (DMS) is an **"action-oriented"** evolution of a Decision Support System (DSS)

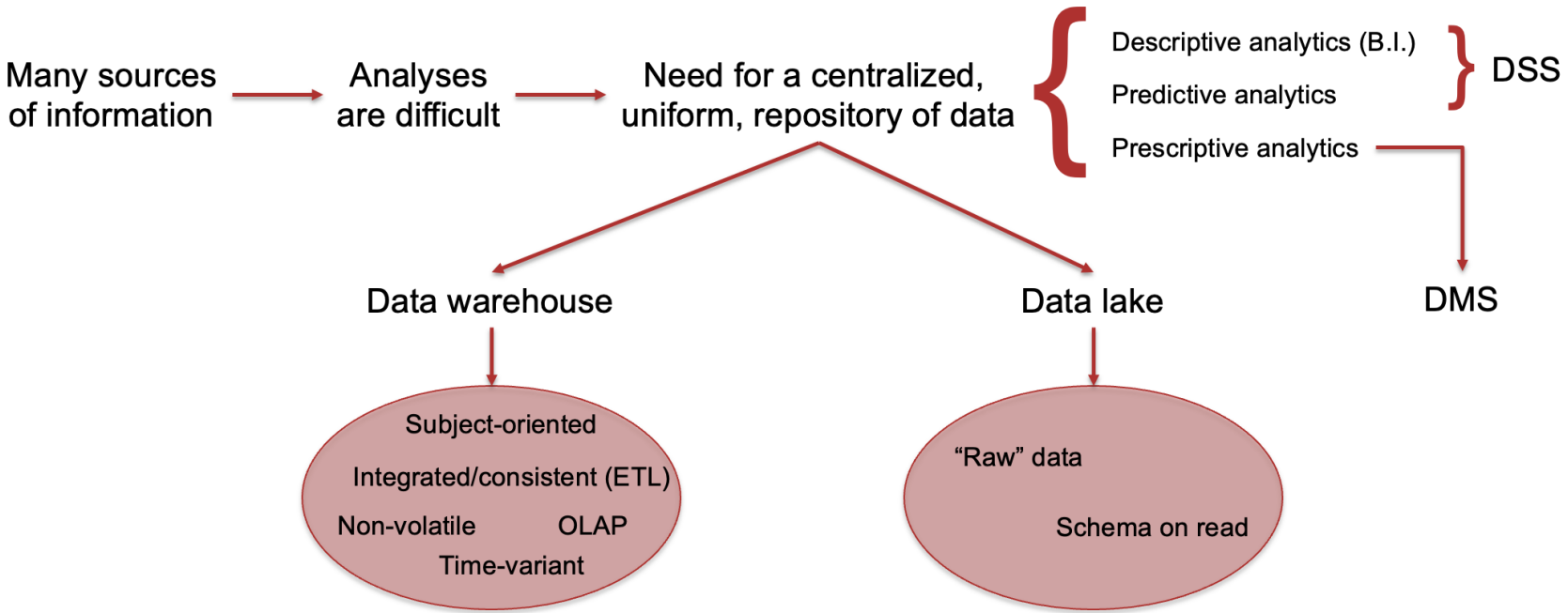
DSSs offer managers information upon which to come up with an idea and ultimately to make a choice

DMSs make one step more and take actions without human intervention based on known information and a set of coded business rules, or artificial intelligence models

Of course, this mainly involves **routinary decisions**



Recap


Many sources of information
Analyses are difficult
Need for a centralized, uniform, repository of data
Descriptive analytics (B.I.)
Predictive analytics
Prescriptive analytics
DSS
DMS
Data warehouse
Data lake
Subject-oriented
Integrated/consistent (ETL)
Non-volatile
OLAP
Time-variant
"Raw" data
Schema on read

References

A. Vaisman, E. Zimányi *Data Warehouse Systems - Design and Implementation*, 2014

A. Silberschatz, H. F. Korth, S. Sudarshan *Database system concepts*, 7th Edition, 2020

W. I. Immon *Building the Data Warehouse*, 4th Edition, 2005

The modern data stack.
`https://www.moderndatastack.xyz/categories`