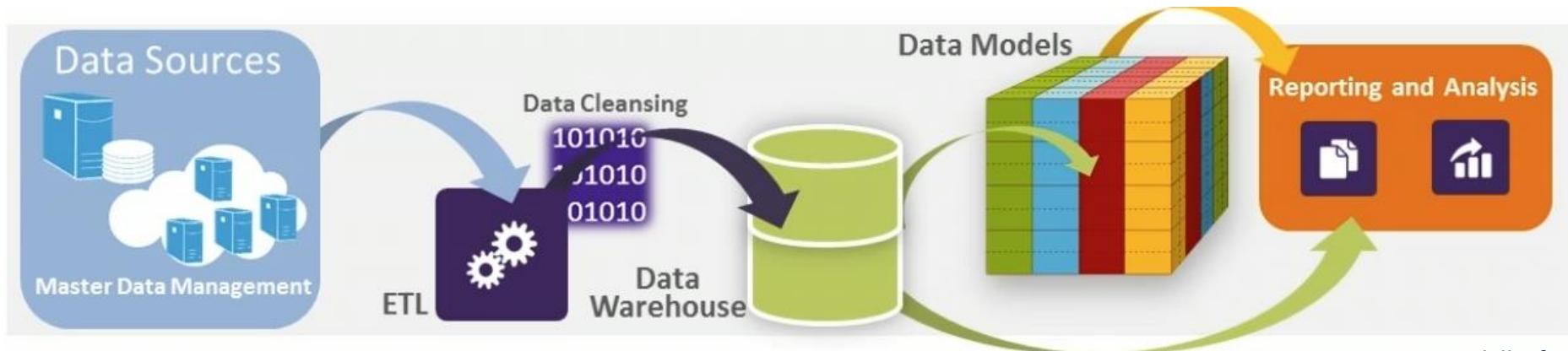


Business intelligence

Unit 1 – Introduction to BI and methodology

BI First class: An introduction

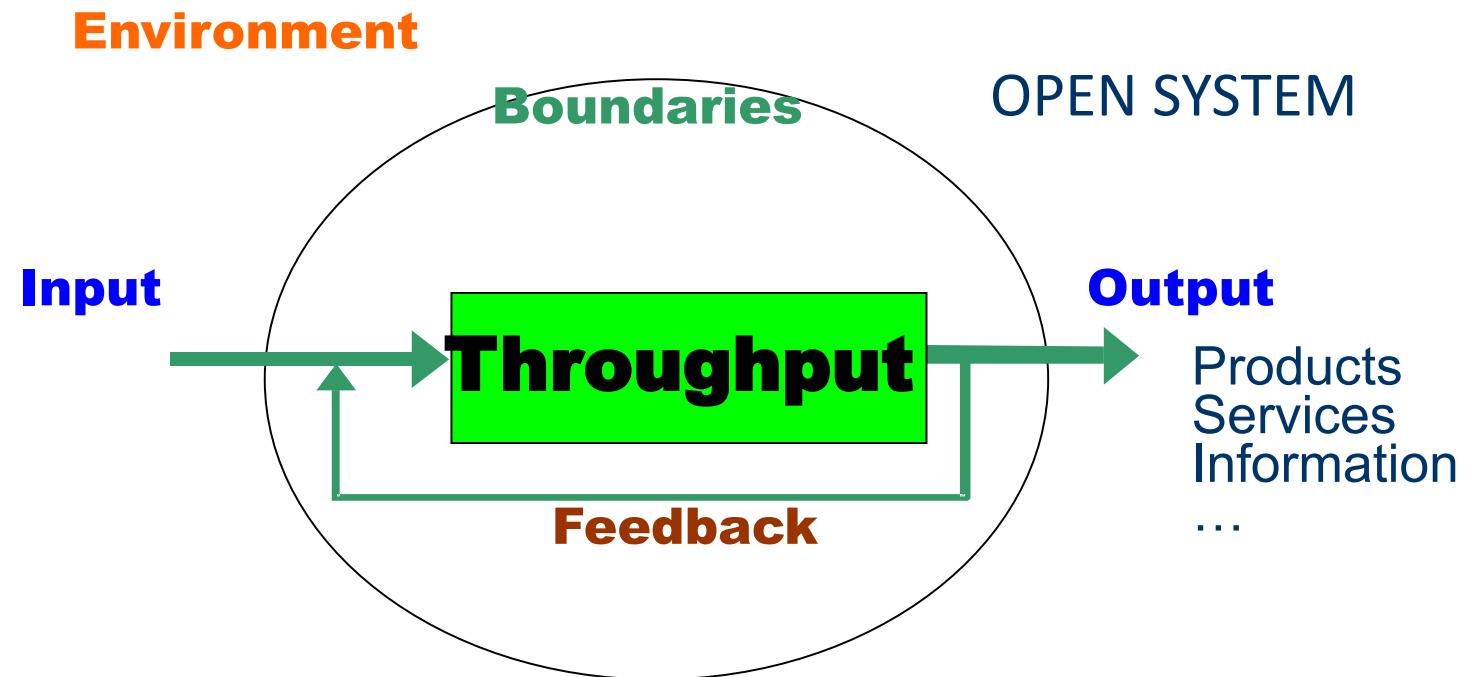
- **What is BI?**
- **Why is it important?**
- **How does it relate to Big Data?**
- **Where is it used?**
- **BI Tools**



Source: Skillsoft

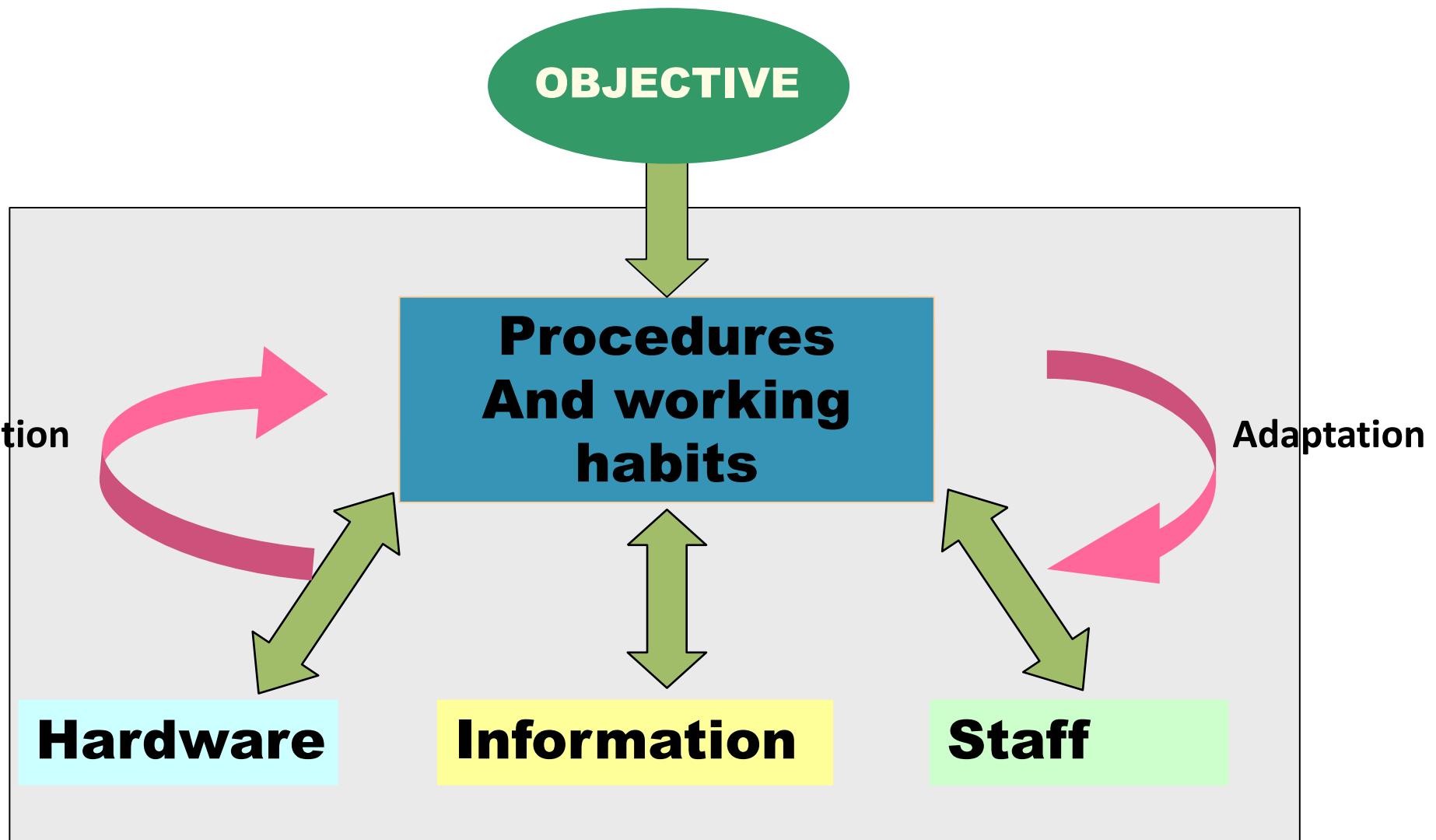
- **System:**
 - System is a set of interrelated things that contribute to a certain goal or perform a specific function.
 - A system consists of a set of interrelated elements, operating in a changing environment and with specific objectives.
- **Basic elements:**
 - System components.
 - The relations between them, which determine the structure of the system.
 - The objective of the system.

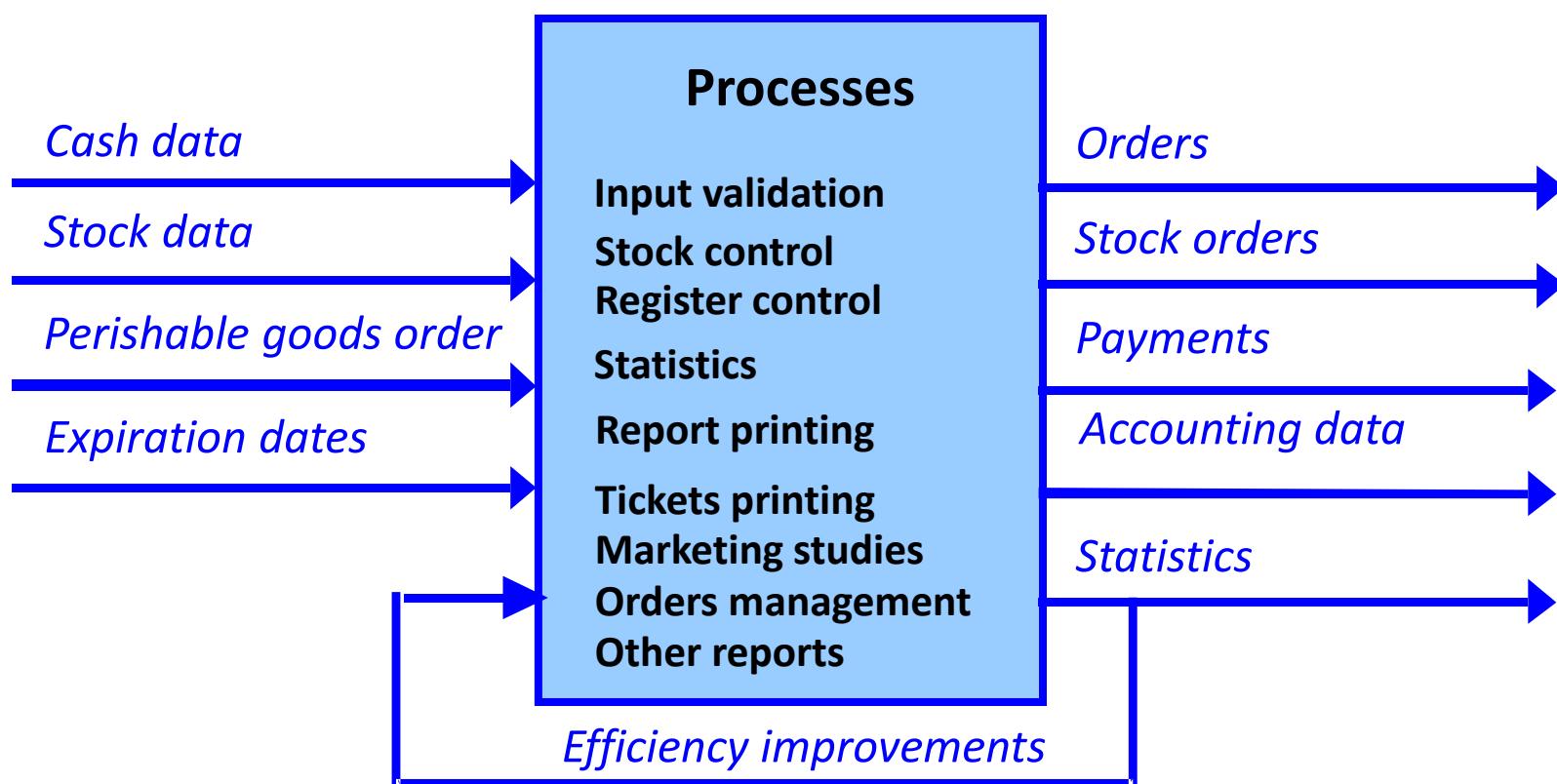
Customers
Providers
Law
Investors
Technology
Information
...

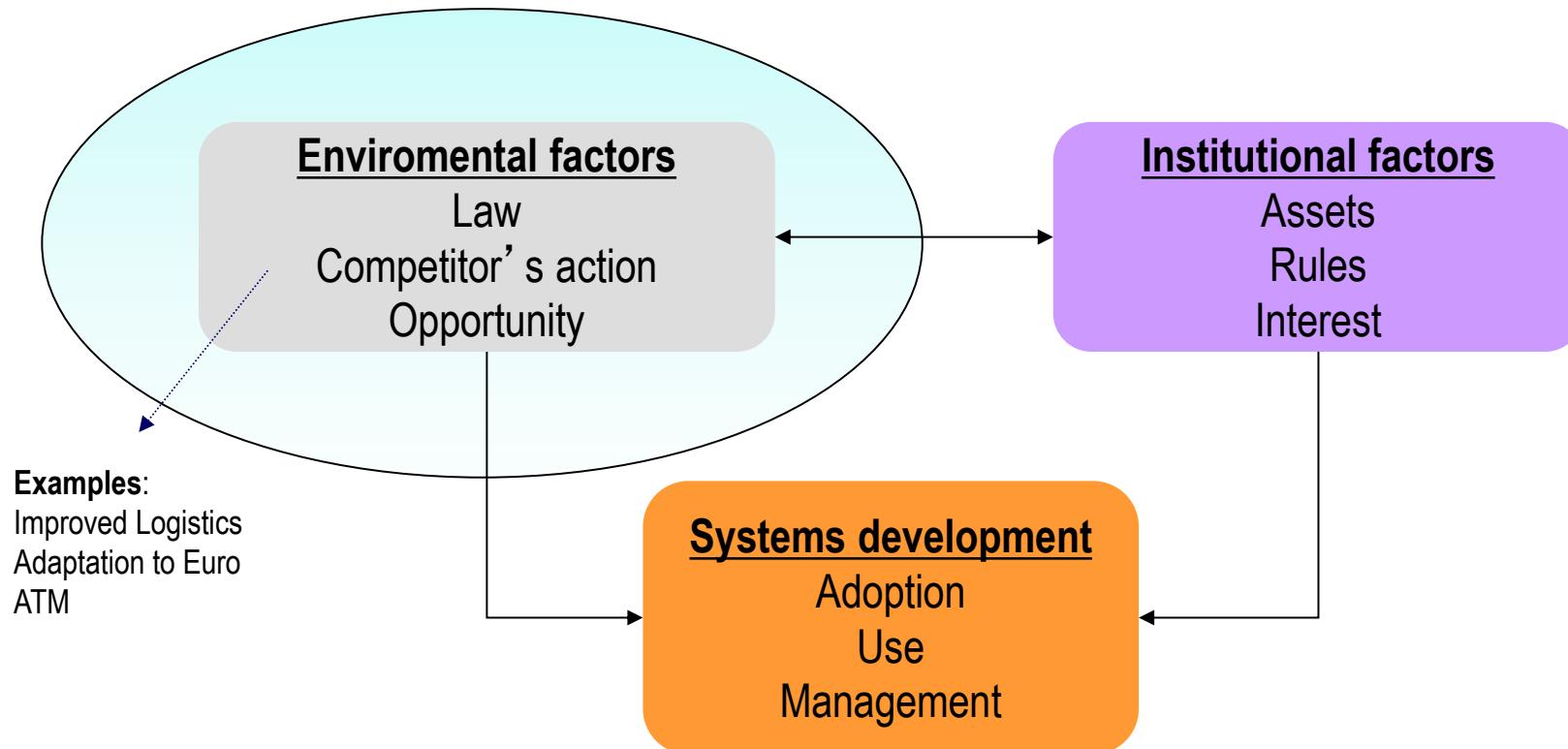


- The system **environment**: his surroundings, within which it is located.
- The system **boundaries**: the boundary between what the system is and what constitutes the environment.
- **Feedback**: In many systems the output influences the input of the system.
- **Throughput**: Transformation processes.

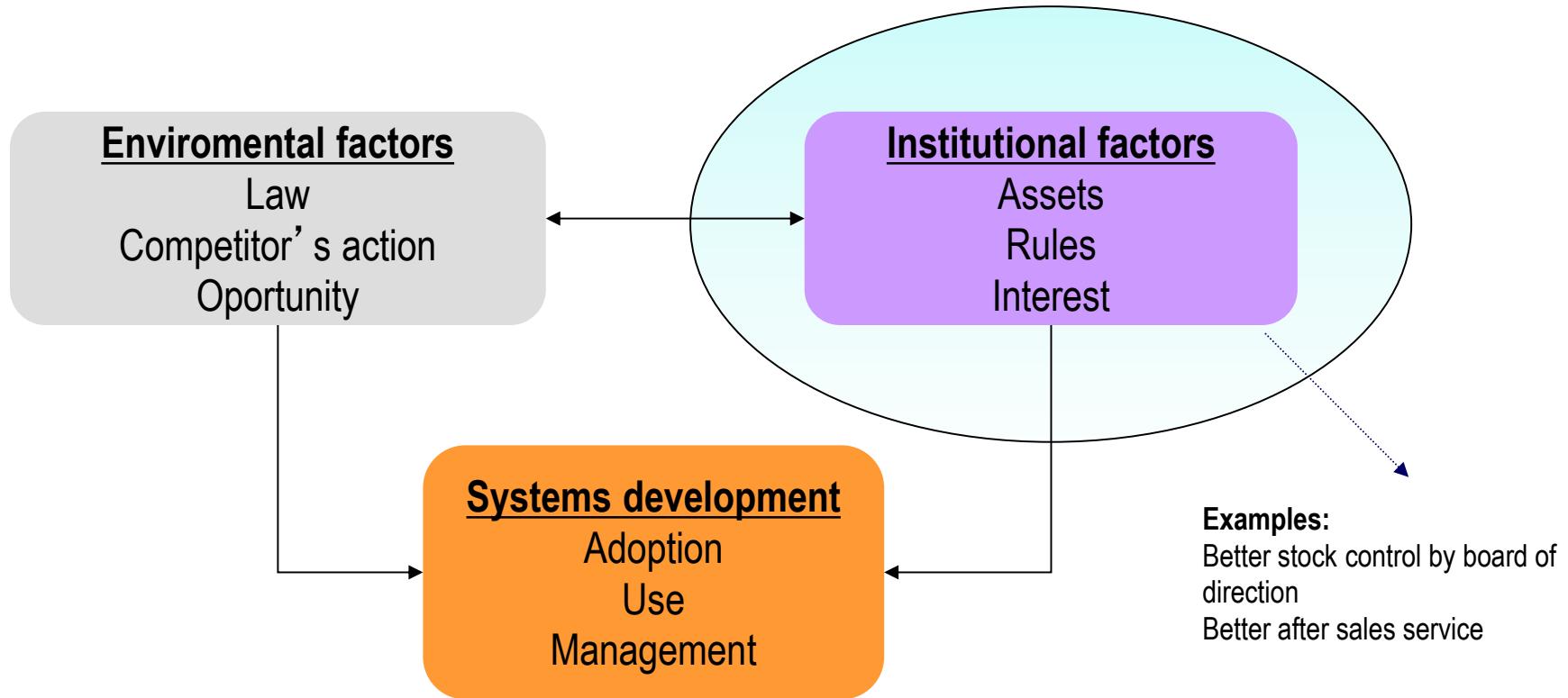
- Definition:
 - A formal set of processes that, operating on a collection of data structured according to the needs of the company, collect, process and distribute the information (or any part of it) necessary for the daily operations of the company and for the direction and control (decisions) to carry out their activities in accordance with its business strategy.
- Other definitions do emphasize that the goal is to provide quality information:
 - The goal of SI is to help the performance of activities at all levels of the organization, by providing the right information, with sufficient quality, to the right person at the right time and place, and with the most useful format to the receiver.



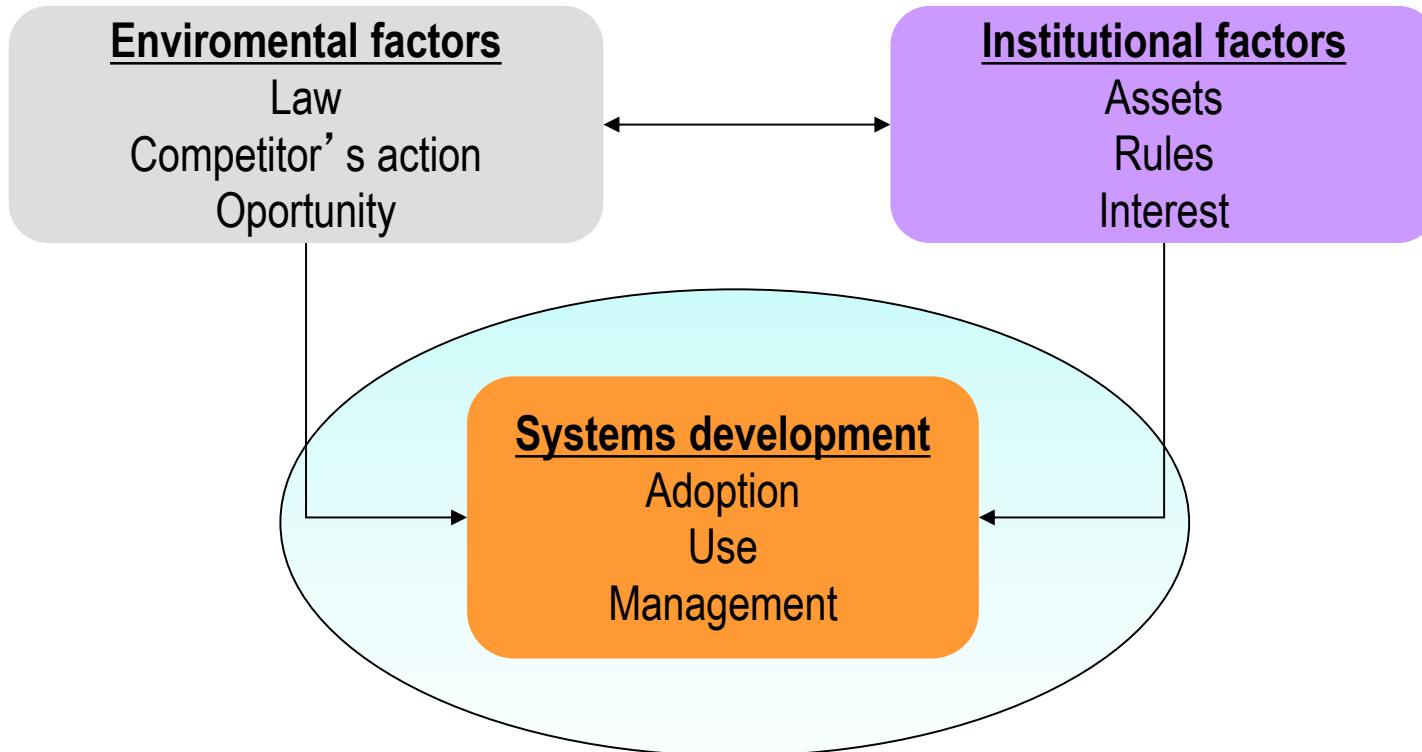




- From a business perspective, an information system is a solution for business organization and management based on information technology whose aim is to deal with an emerging challenge in the business context



- Strategic information systems: at any organizational level there are changes in goals, operations, products, procedures or relationships with the environment to gain a competitive advantage



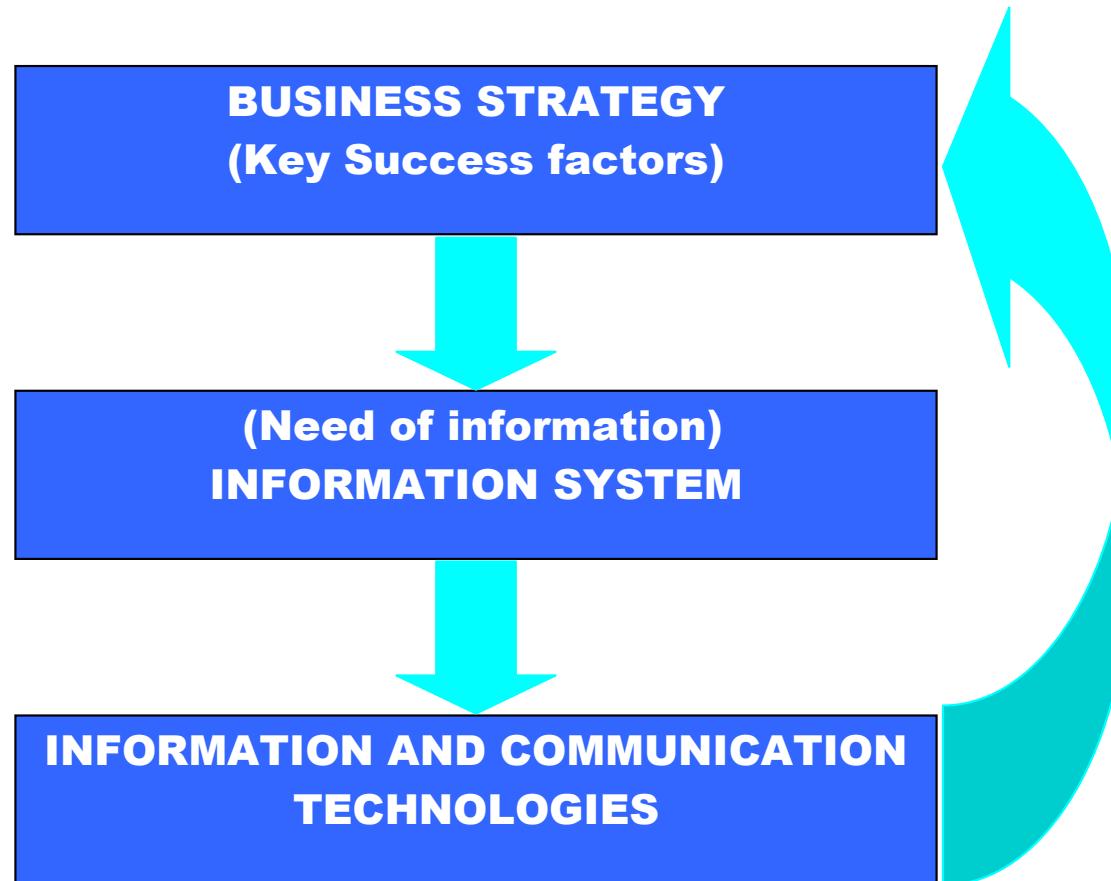
- The value of an IS depends on its effectiveness, its scope, its acceptance by those who use it, its cost, the quality of information that is produced, etc..

- Mussel farming IS
 - Some Environmental factors:

EU, national and regional regulations (food safety and hygiene)	Certifications
Climate change	Economic landscape
Technology improvements	Public administration support

- Some Institutional Factors:

Technology and Infrastructure	Human resources
Financial resources	Internal procedures
Interest in maximizing profits	Interest in expanding market



- “"The Enterprise IS coordinates the information flow and records necessary to carry out the functions of a company according to its business approach or strategy””
 - Business strategy is the key.
- Enterprise-resource planning ERP
- Workflow management systems
- Groupware systems
- E-commerce systems
- Electronic Data Interchange, EDI

Decade	Required Information	Type of Information Systems	Objectives
1950s	Basic Computational Data	Batch Processing Systems	Automate complex calculations and data processing tasks
1960s	Transactional Data	TPS (Transaction Processing Systems)	Automate routine business operations and record-keeping
1970s	Operational and Tactical Data	MIS (Management Information Systems)	Improve administrative control and reporting
1980s	Analytical Data	DSS (Decision Support Systems)	Support decision-making with data modeling and analysis tools
1990s	Strategic and Competitive Data	EIS (Executive Information Systems)	Enable strategic decision-making and real-time monitoring
2000s	Enterprise-Wide Data	ERP (Enterprise Resource Planning Systems)	Integrate business functions for efficiency and process automation
2010s	Big Data, Predictive Analytics	BI (Business Intelligence Systems)	Data-driven decision-making and predictive analytics
2020s	Real-Time and AI-Enhanced Data	AI Systems and Cloud-Based Analytics	Automate decisions with AI, scalable cloud infrastructure

Systems operational level: They monitor the activities, operations and basic transactions of the organization.

MIS and decision support : They support the monitoring, control and decision-making and administrative activities at management level.

Strategic level systems: They support long-term planning at management level in order to gain competitive advantage.

A fourth type: Knowledge level systems: support knowledge and information workers of the institution.

Example USC:

Operational- Student enrollments;

MIS-monthly reports on student performance

DSS-Analyzes trends in student enrollment to support decision making

Strategic- USC takes long-term decisions on new degrees and research lines

Knowledge- digital repository with theses, papers, ...

Its main features are:

Performs and records routine daily operations necessary for the operation of the company.

Designed to increase productivity

Investment in them is easy to justify to the directive board, as its benefits are visible and palpable.

They are often the first type of IS that is implanted in organizations. It starts supporting efforts at the operational level of the organization.

They are intensive in data input and output, their calculations and processes are usually simple and unsophisticated.

They provide administrators with reports and on-line access to historical and daily records

They are the main generators of information for other types of systems.

Examples: billing, plant scheduling, payroll, inventory, ...

- **Its main features are:**

They are always included after having implemented more relevant operational IS, since the latter are its information providers.

Calculations are usually intensive, while outputs are scarce.

The information generated provides support to middle and high management in the decision making process.

They combine changing information with sophisticated analytical models to support unstructured and semistructured decision-making.

They tend to be interactive, visual and friendly, and they are focused to the end user.

They do not intend to save work. As a result, the economic justification for investment in these systems is difficult because the direct benefits of the project are not known.

Involve analytical models, what-if analysis, simulations and forecasts.

- **Its main features are:**

They incorporate external information and get summarized information from operational and decision support systems.

They support the introduction of products and processes within the organization because they aim to gain advantage over competitors by innovating.

They function is to achieve advantages that competitors do not have, such as cost advantages and differentiated services to customers and suppliers. In this context, the Strategic System are creators of entry barriers to the business.

- For example, ATM, Internet banking, recommendation systems

They used to be developed ad hoc within the organization.

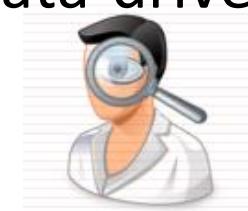
Aimed to achieve long-term strategic goals (e.g. expanding market share, create new markets).

◆ **Examples:**

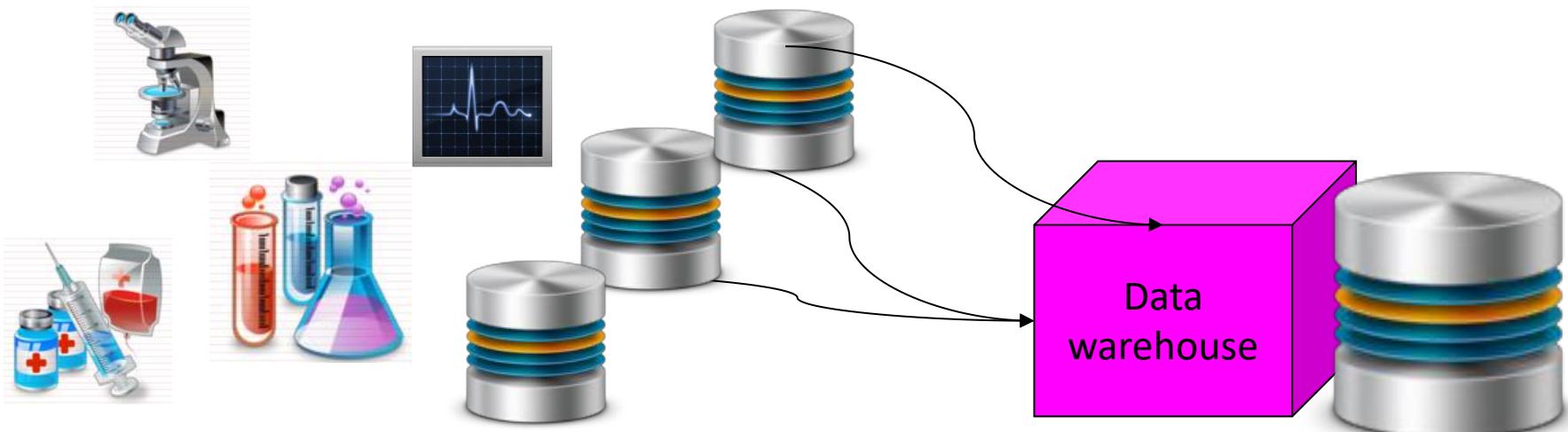
- Sales forecasts
 - Planning marketing campaign
- MRP (Manufacturing Planning Resource) focused on increasing productivity in a manufacturing process
- Product discovery and launching in banking: types of mortgages, types of accounts, ... with the purpose of achieving business goals:
 - Attracting new customers
 - Customer Loyalty
- Television schedule grid
 - Profile of viewers and appropriate advertising
 - Duration of advertising

System	Input	Processes	Output	User
Operational	Transactions,	Store, report, merge,	reports, summaries, transactions. Invoice, payroll	Operative staff, supervisor
DSS, tactical	Summarized operational information; high volume of data; Simple models	Model; simulations; analysis Optimization What-if analysis	Analytical reports, Decision Models	Technical staff, data analysts
Strategic systems	Aggregated information; External information;	Strategic analysis, scenario planning CI	Strategic plans, Forecasts	Directive

- This term refers to the management of information in a **specific** business or business area.
- It consists of a set of strategies and tools focused on knowledge management by **analyzing existing data** in an organization or company.
- In the business intelligence we focus on:
 - Setting business goals
 - Determine needs of data, information and knowledge to meet business objectives
 - Keywords: Integration, accessibility and data-driven decisions.



- Operative IS: gather and organize data



- **Tactical IS:** **Analyze**, summarize, transform, **visualize**

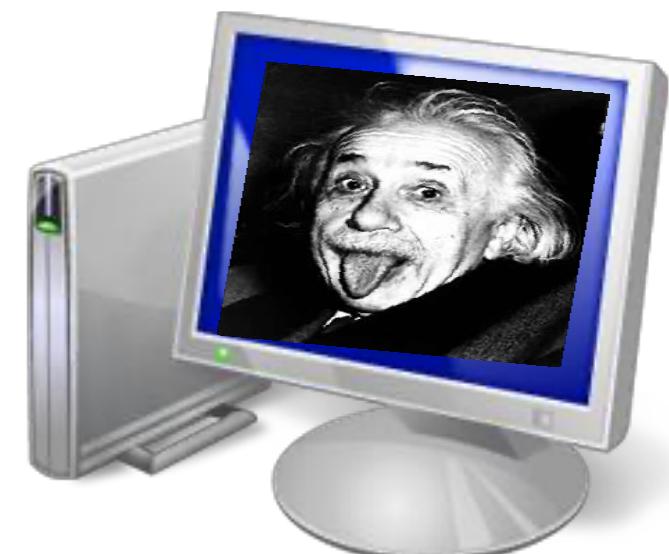


- Strategic IS: **Acquire, discover, evaluate and use knowledge**



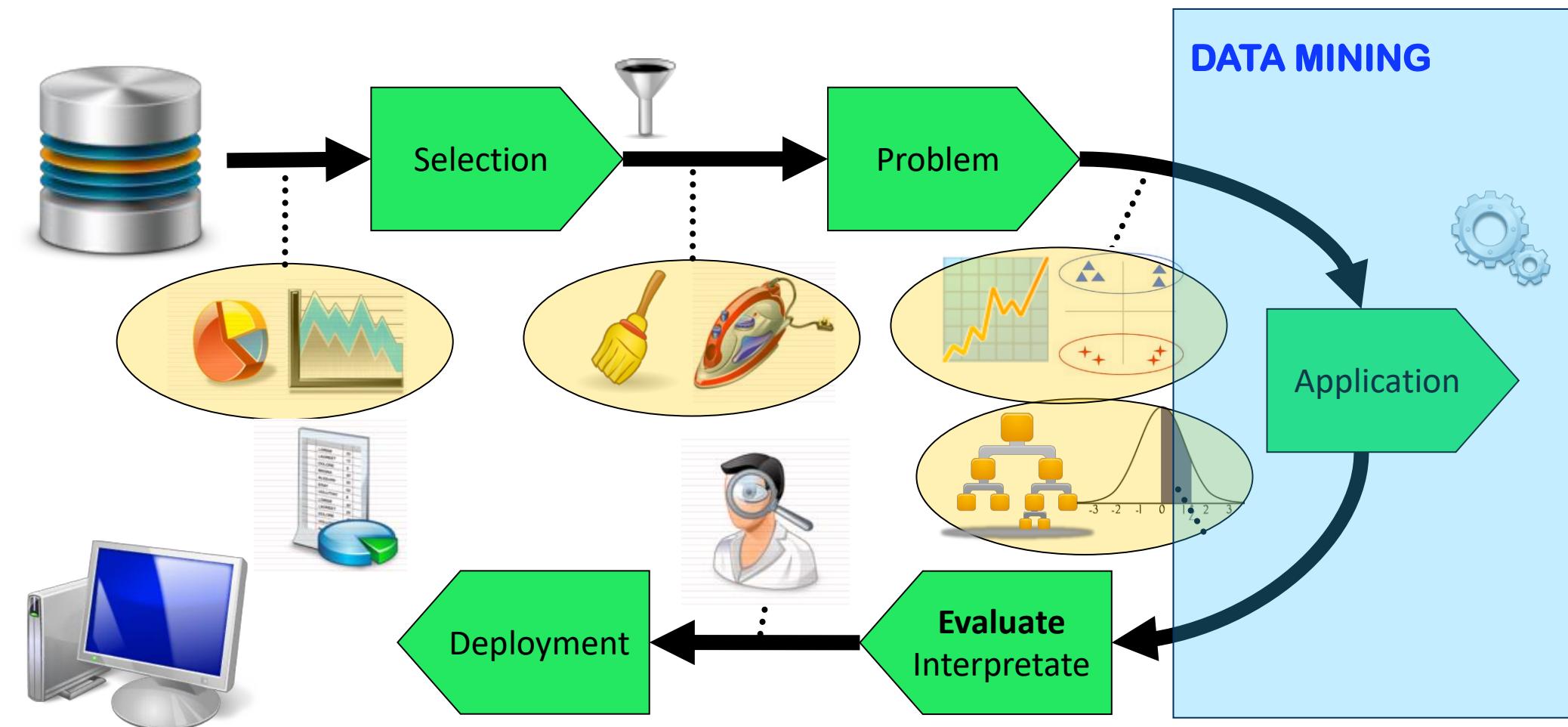
Knowledge acquistion

Knowledge discovery

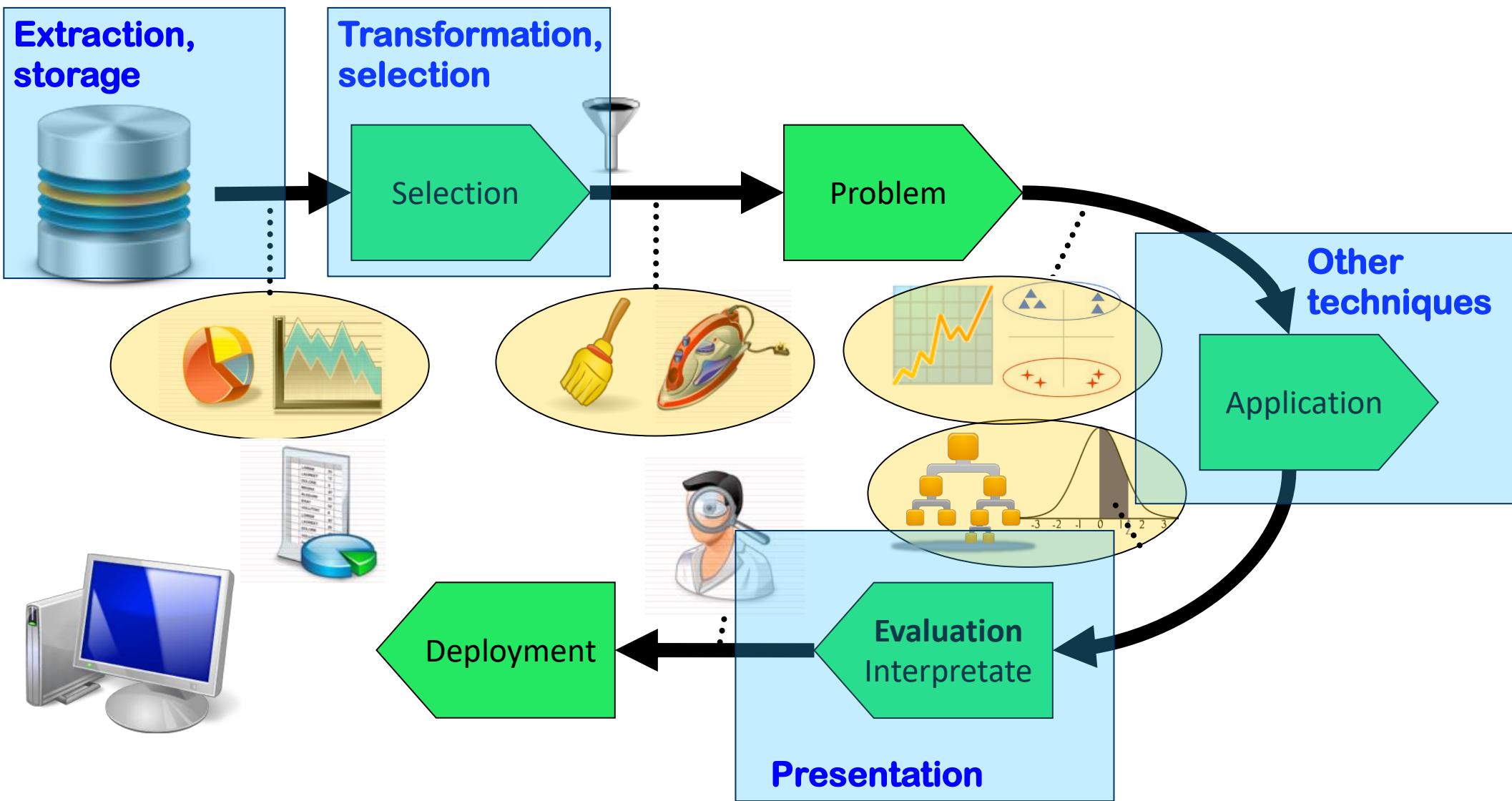


Expert system, dss, knowledge management, data mining, ...

- In the other courses you will focus on:



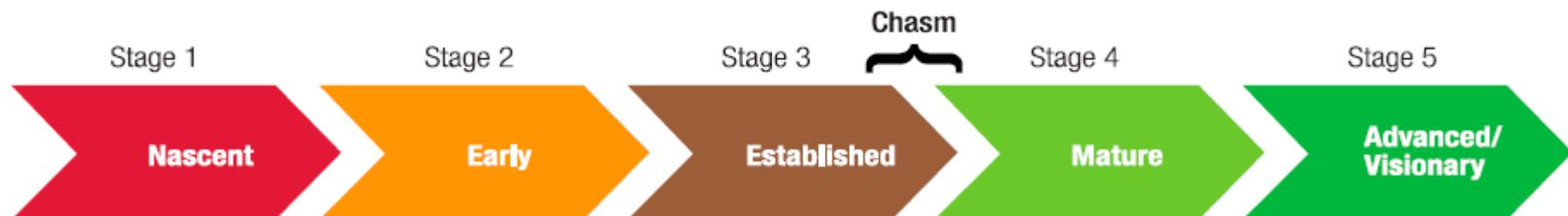
- Now we will focus on the complete process:



- Maturity models define levels of definition, efficiency, manageability and measurement of the monitored environment.
- The maturity model for Business Intelligence helps organizations understand where they are and how they can improve.
- Several MM:
 - TDWI MM
 - HP
 - Gartner
 - LOBI, Forrester's, ...
 - Each model has at least 5 stages of maturity
 - Each model starts with operational / one-off reporting and culminates in pervasive BI
 - The models do not focus on technology alone and hinge on the involvement of people and process as well.
 - Provide current capabilities, roadmap for improvement, benchmarking, alignment with business goals, improves data-driven culture, optimize resource allocation and support strategic planning.

Model Dimensions

Organizational Maturity	Resource Maturity	Data Infrastructure Maturity	Analytics Maturity	Governance Maturity
<ul style="list-style-type: none"> • Leadership • Culture • Impact • Strategy 	<ul style="list-style-type: none"> • Funding • Talent/skills • Roles/responsibilities • Training 	<ul style="list-style-type: none"> • Diversity, volume, and speed • Data access • Data integration and management • Data architecture 	<ul style="list-style-type: none"> • Scope of capabilities • Automation/ augmented • Deployment and delivery approaches • Innovation 	<ul style="list-style-type: none"> • Data governance processes and tooling • Model governance processes and tooling • Governance roles • Security/privacy



Stage 1: Nascent– Pre-analytical environment. Culture is not data-driven and no data-driven decisions. No formal BI infrastructure. Data scattered and in silos. Poor data quality. IT and business don't work together. Excel or manual reports.

Stage 2: Early- Basic BI tools. Starting to understand the value of analytics. The organization realizes some data infrastructure is needed to support data integration for analysis. Rudimentary analytics but advancing.

Stage 3: Established- A data warehouse has been implemented. There is a group that is responsible for analytics in IT. A data governance team has been formed. Data visualization tools are used.

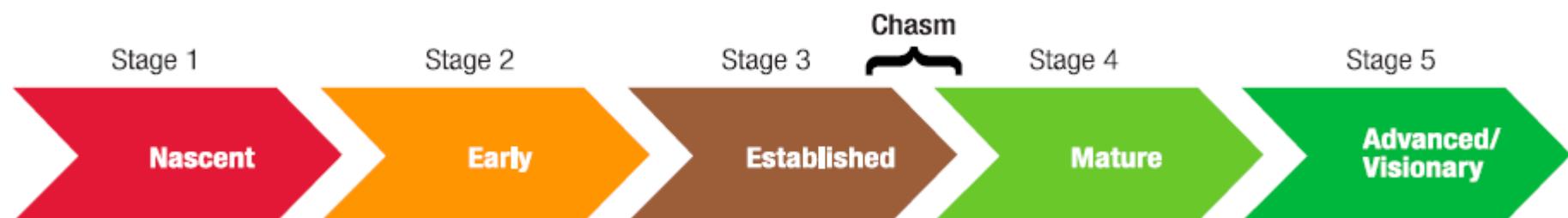
Stage 4: Mature- End users integrate analytics into decision-making and operations. Culture of innovation and collaboration between IT and business. Use of diverse sources, including semi and unstructured data. Workforce include data scientists, data engineers, MLOps and roles like Chief Analytics Officer.

Stage 5- Advanced/Visionary- Self service analytics. Organization considers analytics a critical, competitive weapon. Operational, flexible and scalable infrastructure (data lakes, DWH, cloud), ML, AI is used in NRT for decision-making.

Model Dimensions

Organizational Maturity	Resource Maturity	Data Infrastructure Maturity	Analytics Maturity	Governance Maturity
<ul style="list-style-type: none"> • Leadership • Culture • Impact • Strategy 	<ul style="list-style-type: none"> • Funding • Talent/skills • Roles/responsibilities • Training 	<ul style="list-style-type: none"> • Diversity, volume, and speed • Data access • Data integration and management • Data architecture 	<ul style="list-style-type: none"> • Scope of capabilities • Automation/ augmented • Deployment and delivery approaches • Innovation 	<ul style="list-style-type: none"> • Data governance processes and tooling • Model governance processes and tooling • Governance roles • Security/privacy

Chasm- Organizations need to cross the Chasm, ensuring the right governance, data architecture, skills, data culture, is in place.



Leadership

* Your leadership supports and evangelizes analytics across the company.

- Not at all
- They seem ambivalent about analytics, and they don't really evangelize it
- They support analytics efforts and are starting to evangelize it
- They firmly support analytics efforts, they use analytics to make decisions, and they evangelize it across the company

* Your company has a Chief Analytics Officer (CAO) who is in charge of your analytics efforts.

- We don't have anyone in charge of analytics in the organization
- Analytics is controlled by IT in my company
- We have a VP or Director of Analytics in my company, who is in charge of analytics
- Yes, have a Chief Analytics Officer

Strategy

* Your company has a strong strategy in place to support its data and analytics efforts.

- No and we have no plans to do so
- No, but we plan to do so in the next year
- Yes, we are in the process of putting a strategy together
- Yes, we have a solid strategy in place for analytics

* Analytics is an important part of your company's digital transformation strategy

- No, we do not have a digital transformation strategy
- Yes, we are in the process of putting our digital transformation strategy in place and analytics will play an important role
- Yes, analytics is an important part of my company's digital transformation strategy

Impact

* What % of business units in your company use analytics for day to day decision making

- Less than 25%
- 26-40%
- 41-55%
- 56-70%
- Greater than 70%

* Your organization has measured an impact with its analytics

- No
- No, but I think we've gained value
- Yes, we've measured a top or bottom line impact

Culture

* Your organization uses analytics to take action.

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

* There is a culture of trust in analytics across your company.

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

Interpretation

Once you complete the survey, you will receive your results. The breakdown of scores for each dimension is as follows:

SCORE PER DIMENSION	STAGE
≤ 5	Nascent
6-10	Early
11-15	Established
16-19	Mature
20	Visionary

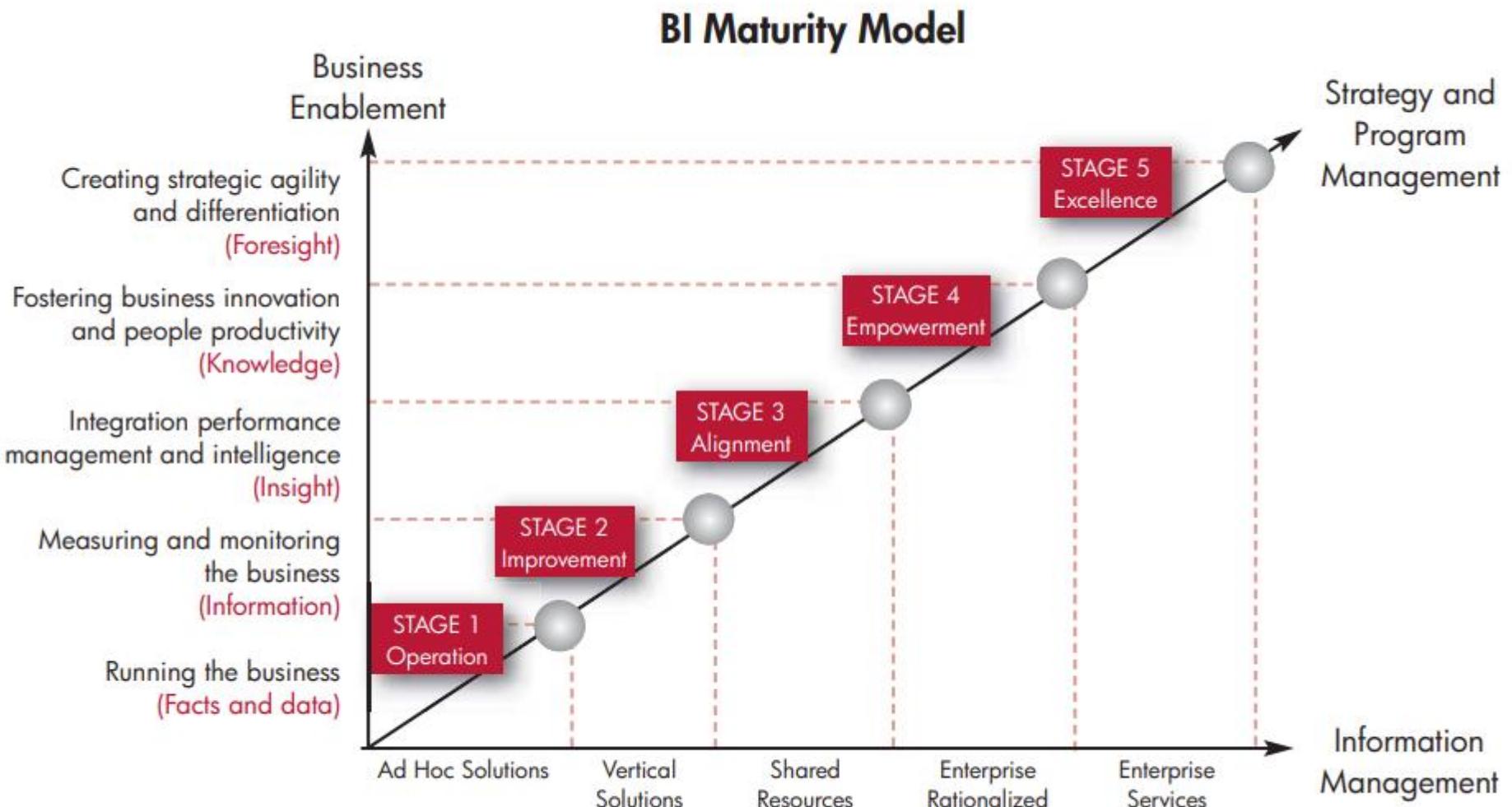
For instance, if you receive a score of 11 in the Organization dimension of the assessment, you are in the Established stage for that particular dimension. You should expect to see your scores vary for the different dimensions. Analytics programs don't necessarily evolve at the same rate across all of the dimensions. For example, your company might be more advanced in terms of bringing data sources together than it is in analyzing them or governing this data.

The Chasm can be overlaid between the Established and Mature stages.

When you complete the assessment, you might see scores like this:

DIMENSION	SCORE	STAGE
Organization	10	Early
Resources	7	Early
Data Infrastructure	11	Established
Analytics	4	Nascent
Governance	7	Early

Total Score: 39



Success = fn (business enablement, information management, strategy and program management)

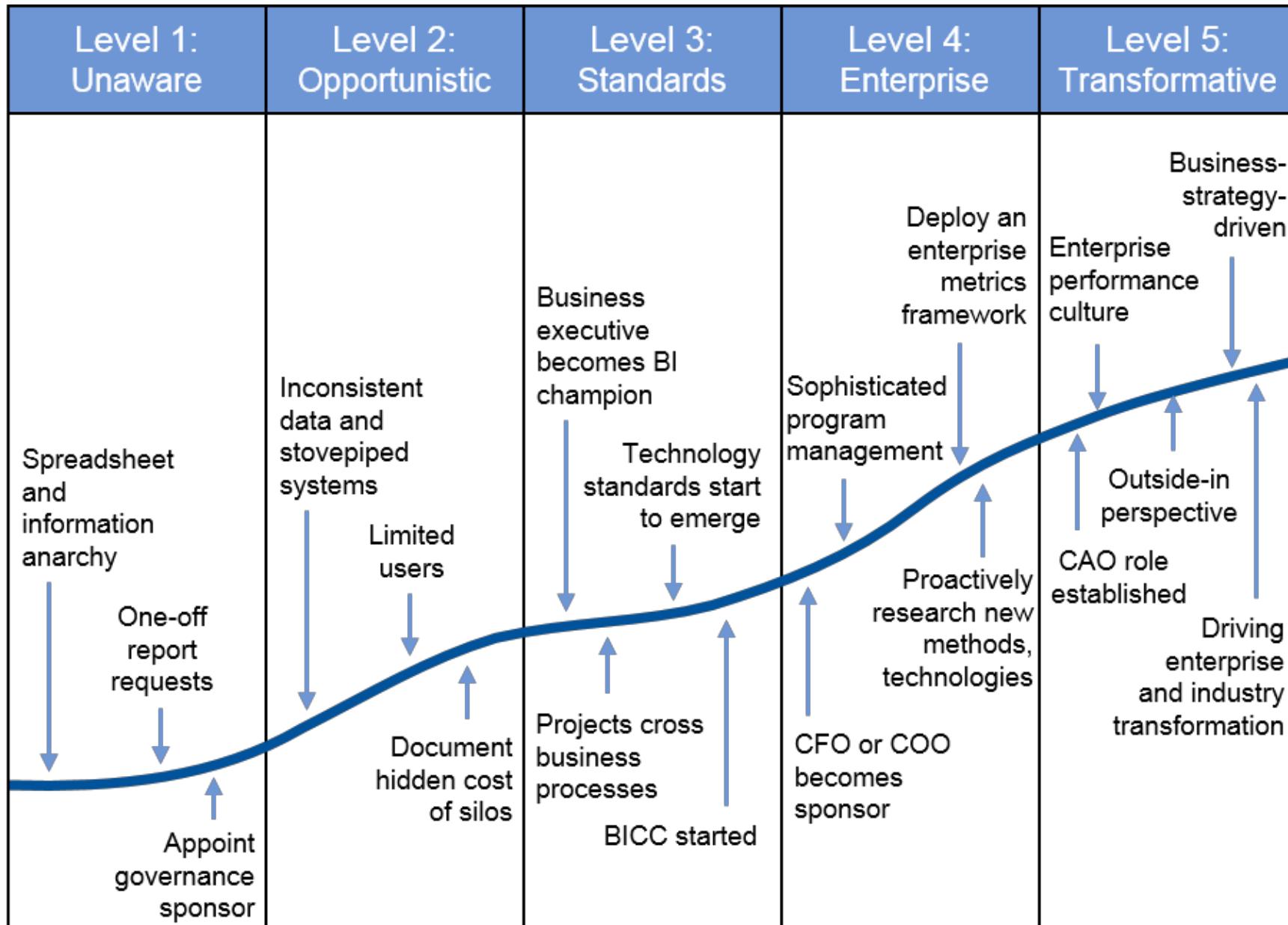
Stage 1 – Operation (Running the business) – involves ad-hoc solutions focused at local demand. No formal BI strategy or analytics.

Stage 2 – Improvement (Measuring and monitoring the business)- BI more structured. Analytics are being used. DW are still focused on specific business units.

Stage 3 – Alignment-BI aligned with strategic goals. Information integration across subject areas. Data quality and governance are increasing their importance. Organization evolved from BI project management to BI program management.

Stage 4 – Empowerment – Self-service tools being used at all levels. Advances analytics are implemented. Single version of truth across the organization.BI is critical.

Stage 5 – Excellence (Change the business) – predictive analytics for most business decisions. SOA for information delivery. Analytics to users are delivered quickly.



BI = Business intelligence

BICC = BI competency center

Level 1 – Unaware— ad-hoc BI and analytics. No formal decision-making process. No IT infrastructure.

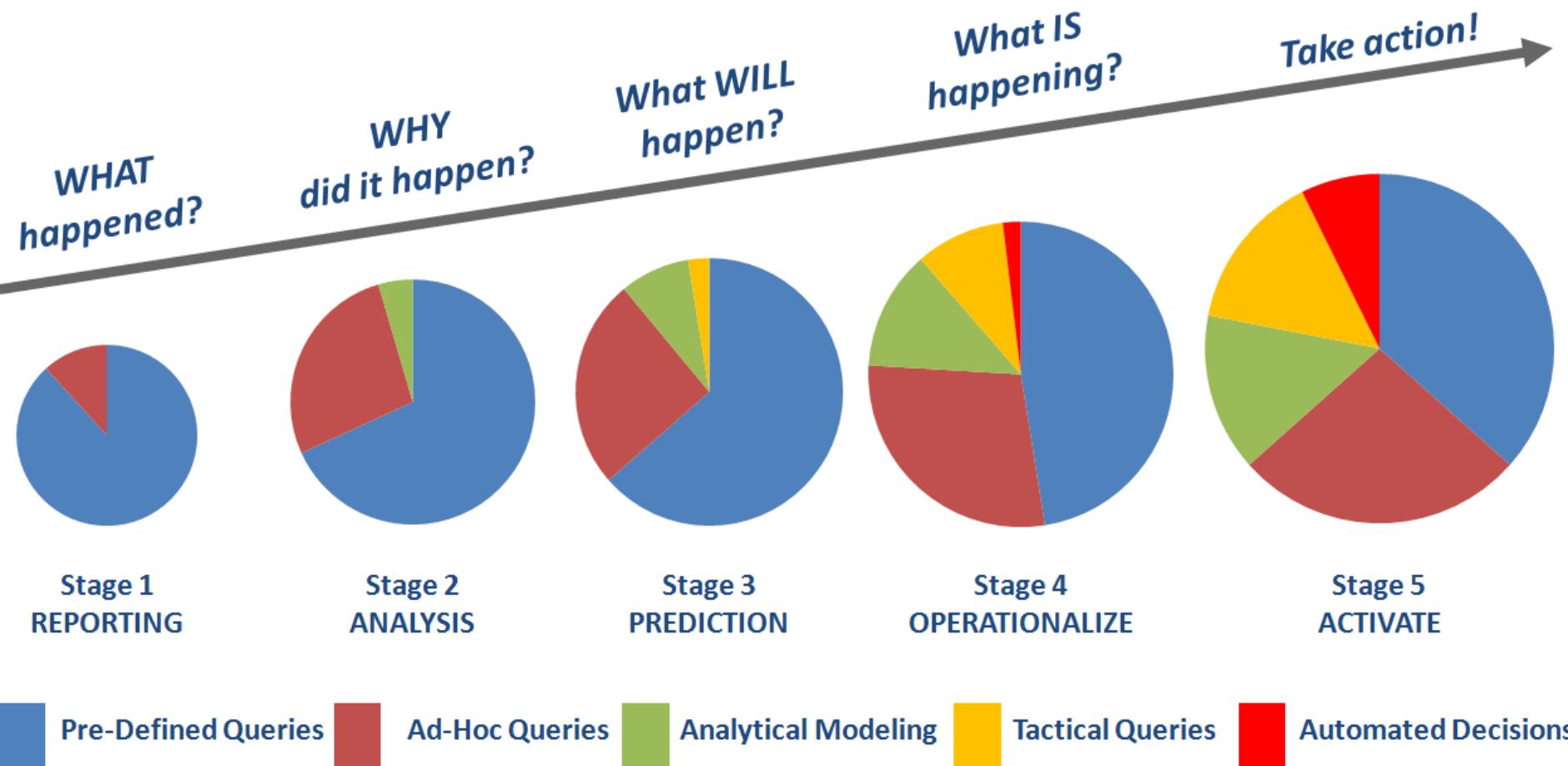
Level 2 – Opportunistic— business units use BI individually. Each project/unit has its own infrastructure.

Level 3 – Standards— Coordination improves. Projects across multiple business processes. A BI competence center is created. Technology standards start to emerge for infrastructure, DW and BI platforms.

Level 4 – Enterprise— Top executives sponsor BI. Performance metrics guiding strategy have been defined. BI supports companywide decision processes.

Level 5 – Transformative— BI and analytics are a strategic initiative run and supported by business and IT. They are used to generate revenue. Users across all levels, including customers and partners.

Evolution of BI capabilities



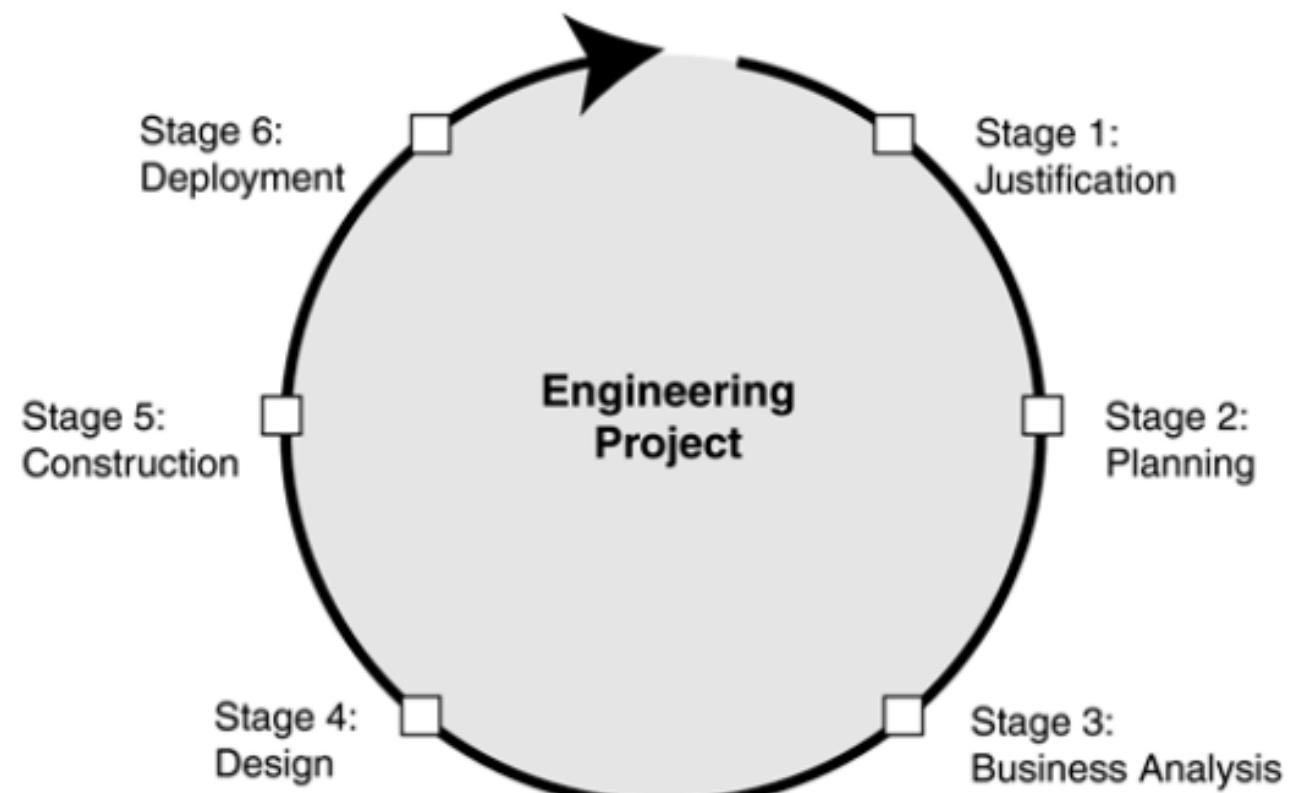
- **What is it and why?**

- Every process needs to be repeatable.
- It accumulates experience in a standard process.
- Created by experts.
- It helps in planning and process management.
- Reduces initial fear
 - There is a standardized process
 - Reduce dependency on specific staff

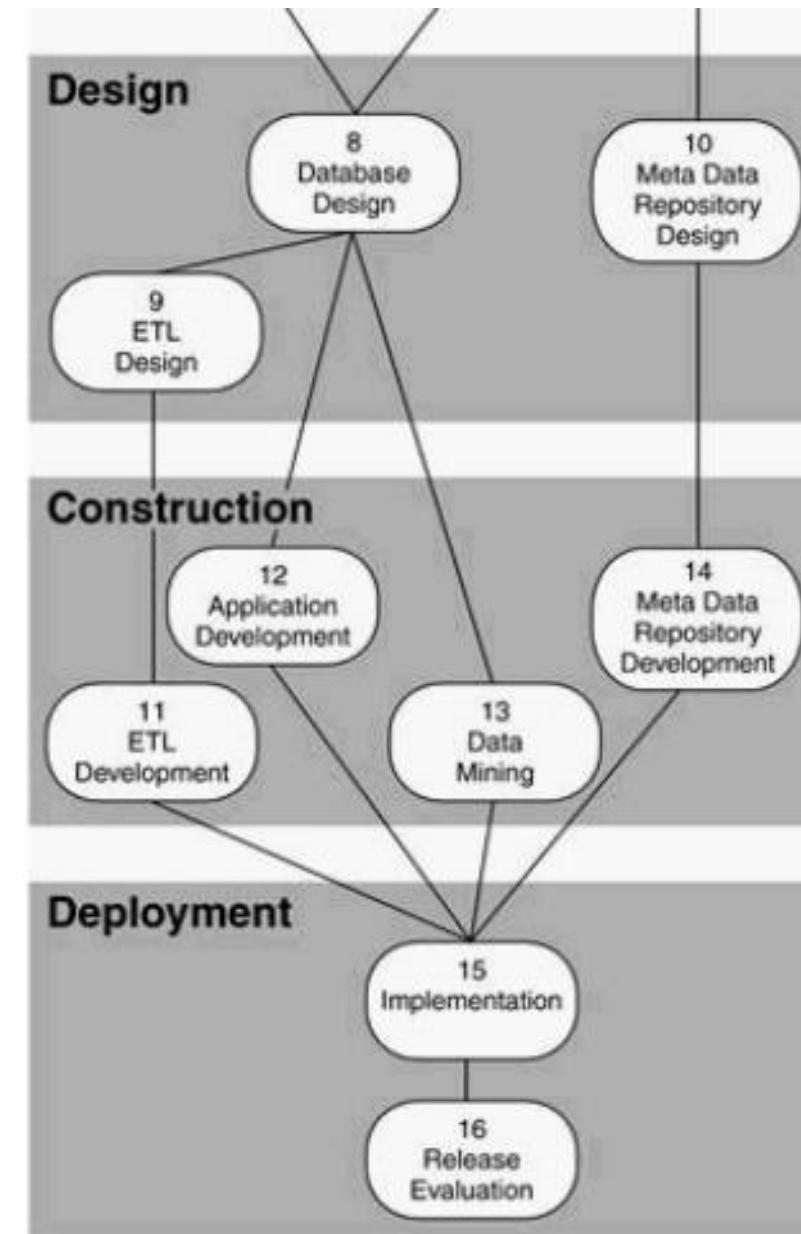
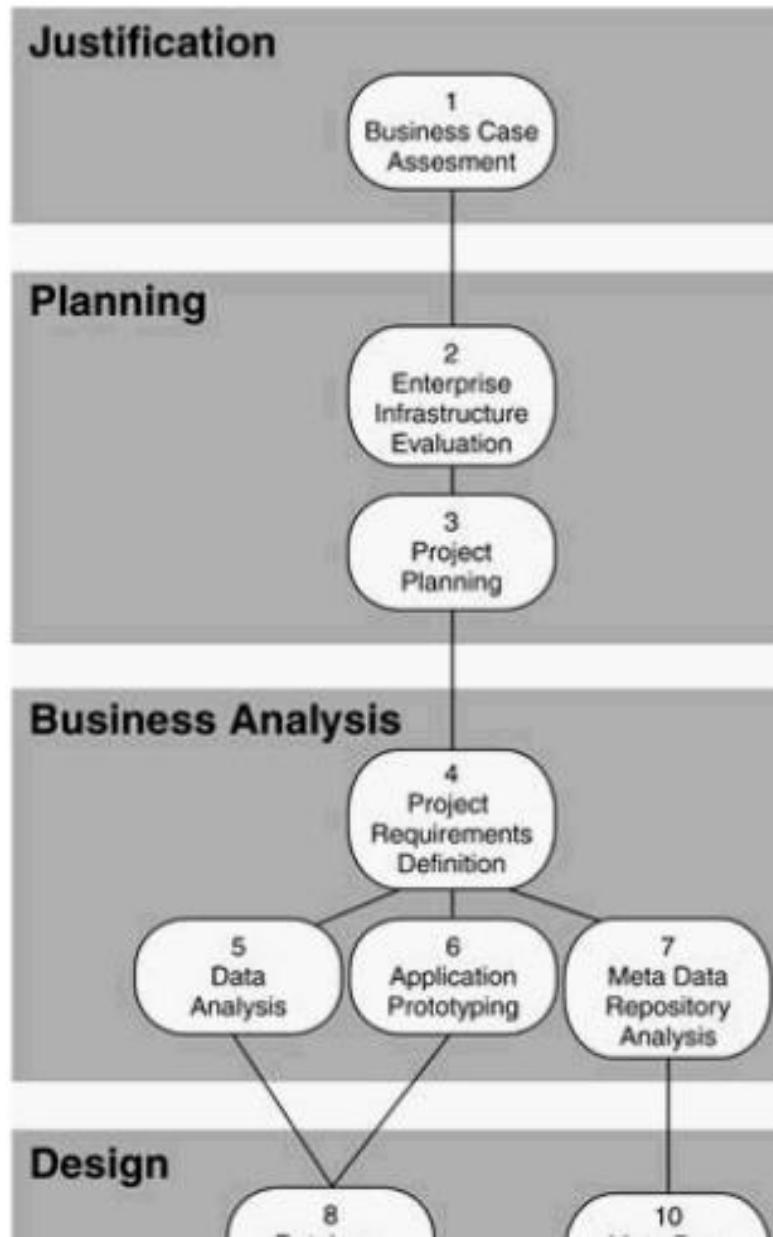
- **Focus on management and development**
 - Management: PMBok, Agile, ...
 - Management+Development: Larissa, Kimball, Inmon, SAFE, QPM (QlikView), ASAP (SAP), ... other tools
- **CRISP-DM: Generic tasks for data mining**
 - Cross Industry Standard Process for DM
 - Proposed by a consortium (SPSS, NCR, AG, OHRA)
 - Other focus on DM:
 - 5A'S, critikal, CAT (Clementine Application Template), SEMMA (SAS, Sample, Explore, Modify, Model, Assess) , associated to tools

- Defines stages, steps, roles, standards and deliverables.
- 6 iterative stages from inception to deployment.
- Agile and adaptive, promotes subprojects.

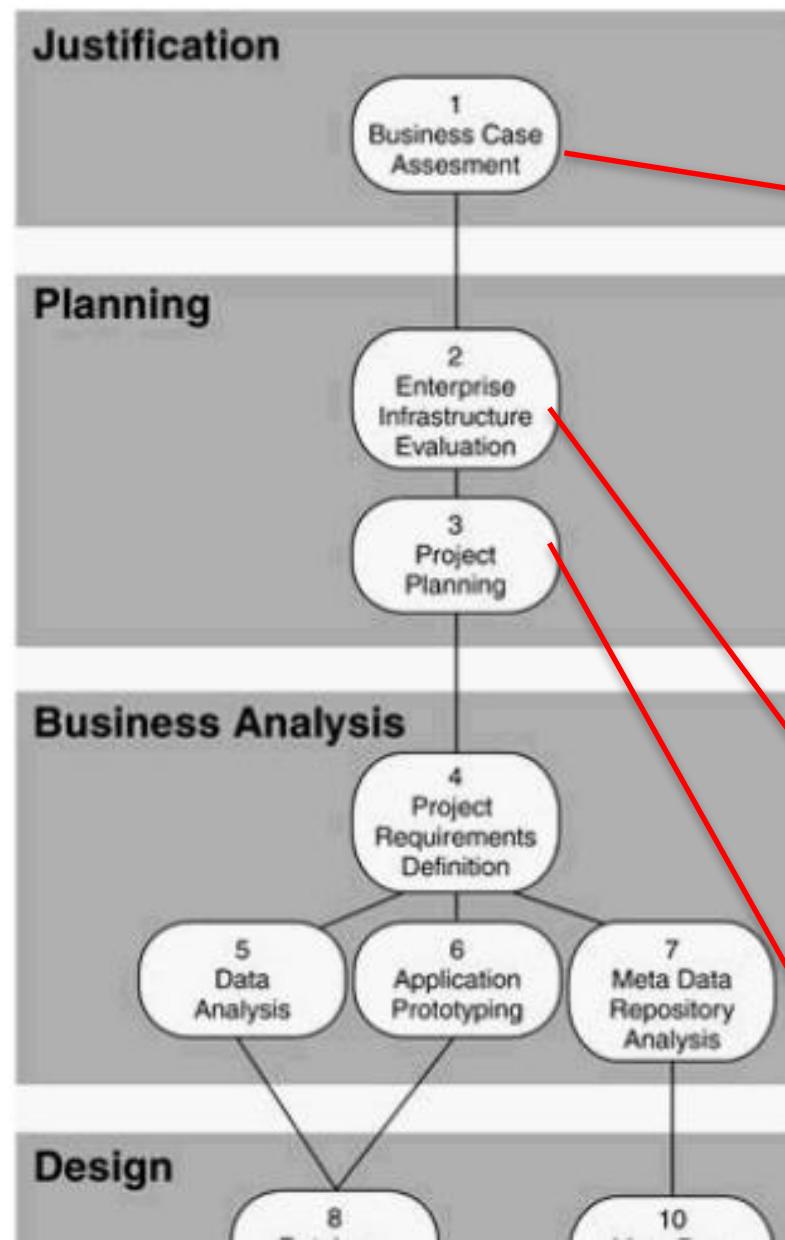
Figure 0.1. Engineering Stages



Larissa Moss: Roadmap to BI with 6 stages and 16 steps

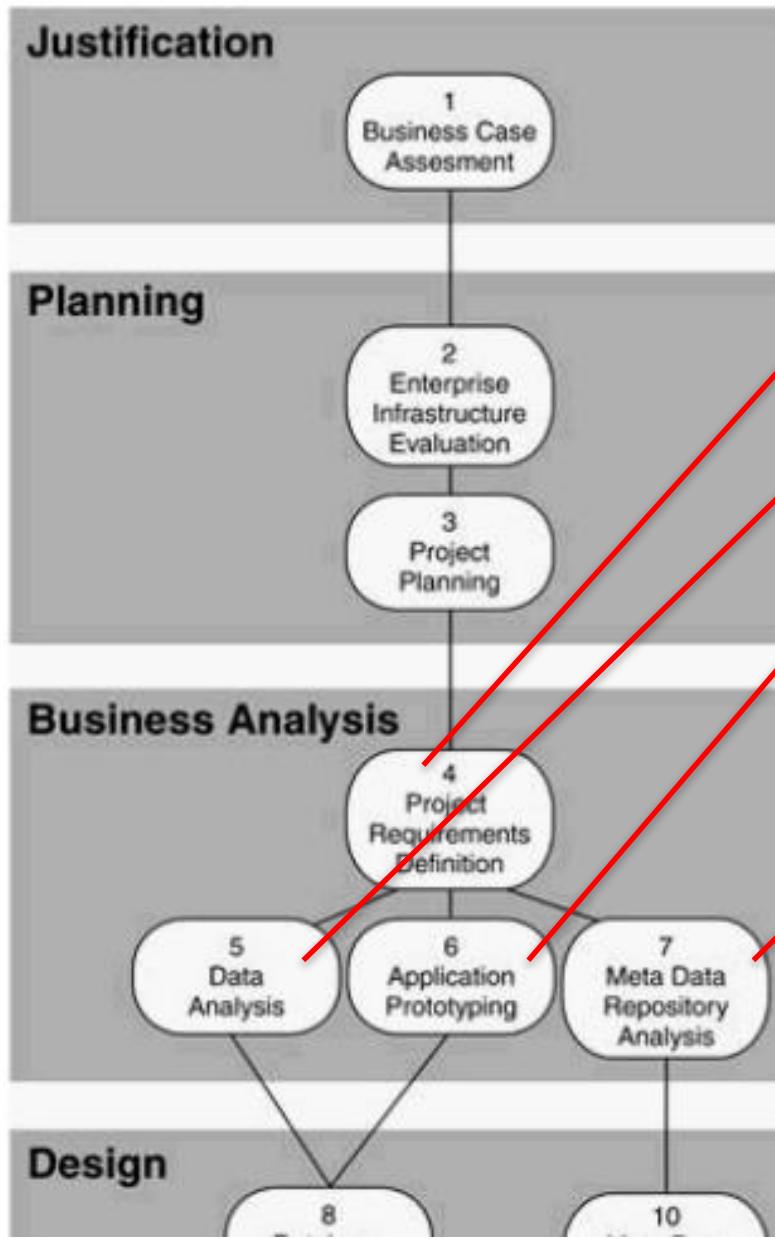


Larissa Moss: Roadmap to BI with 6 stages and 16 steps

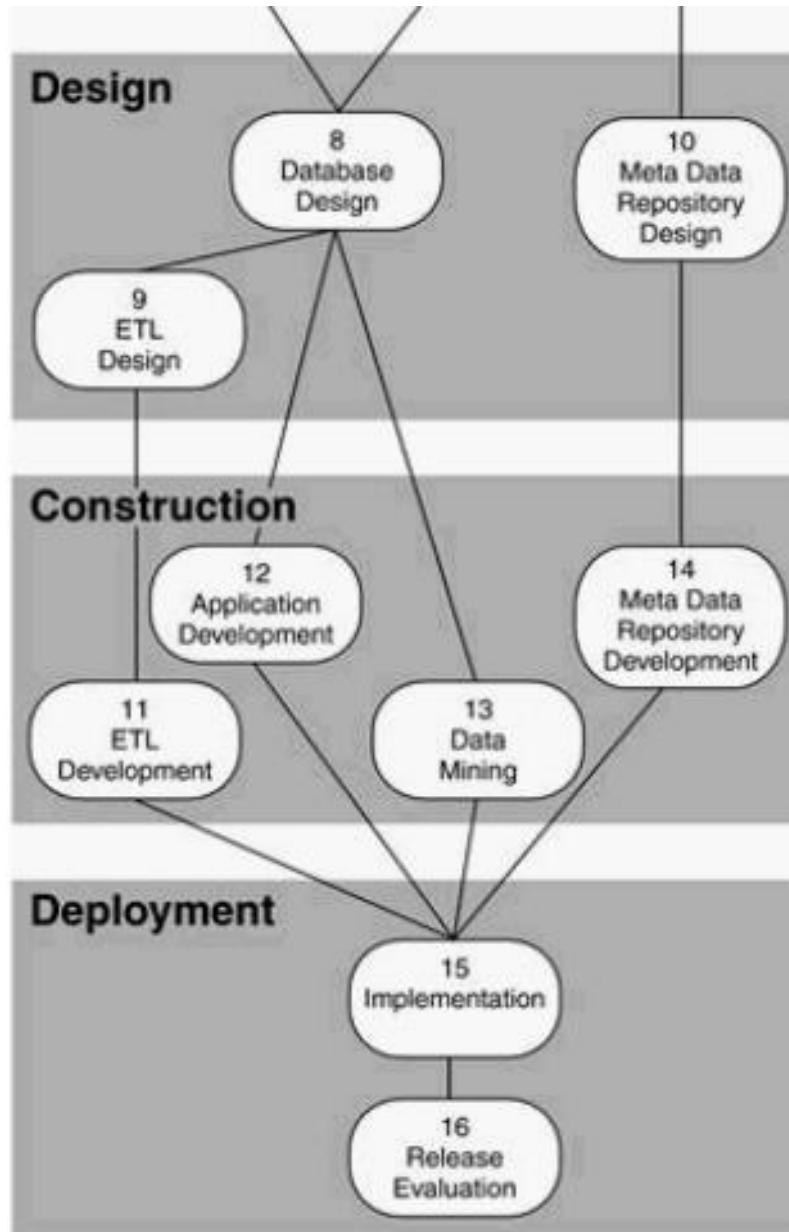


- **1. Justification:** Assess the business need that gives rise to the new engineering project.
- Step 1: Business Case Assessment
 - Defines business problem or business opportunity and proposes a BI solution.
 - Each BI application release should be cost-justified and should clearly define the benefits.
- **2. Planning:** Develop strategic and tactical plans, which lay out how the engineering project will be accomplished and deployed.
- Step 2: Enterprise Infrastructure Evaluation
 - Technical: hardware, software, networks, ...
 - Non-technical: procedures, methodologies, ...
- Step 3: Project planning: scope, staff, budget, technology, business representatives, ...

Larissa Moss: Roadmap to BI with 6 stages and 16 steps



- **3. Business analysis:** Perform detailed analysis of the business problem or business opportunity to gain a solid understanding of the business requirements for a potential solution (product).
- Step 4: Project Requirements Definition
 - Managing and specifying scope, user needs, ...
- Step 5: Data analysis, availability, quality, ...
- Step 6: Application prototyping
 - Helps in the requirements definition and avoid risks
- Step 7: Meta Data Repository analysis
 - Technical meta data needs to be mapped to the business meta data.
 - Meta data describes an organization in terms of its business activities, the business objects, and rules on which the business activities are performed.
 - Ex: definitions, units, relationships, sources....

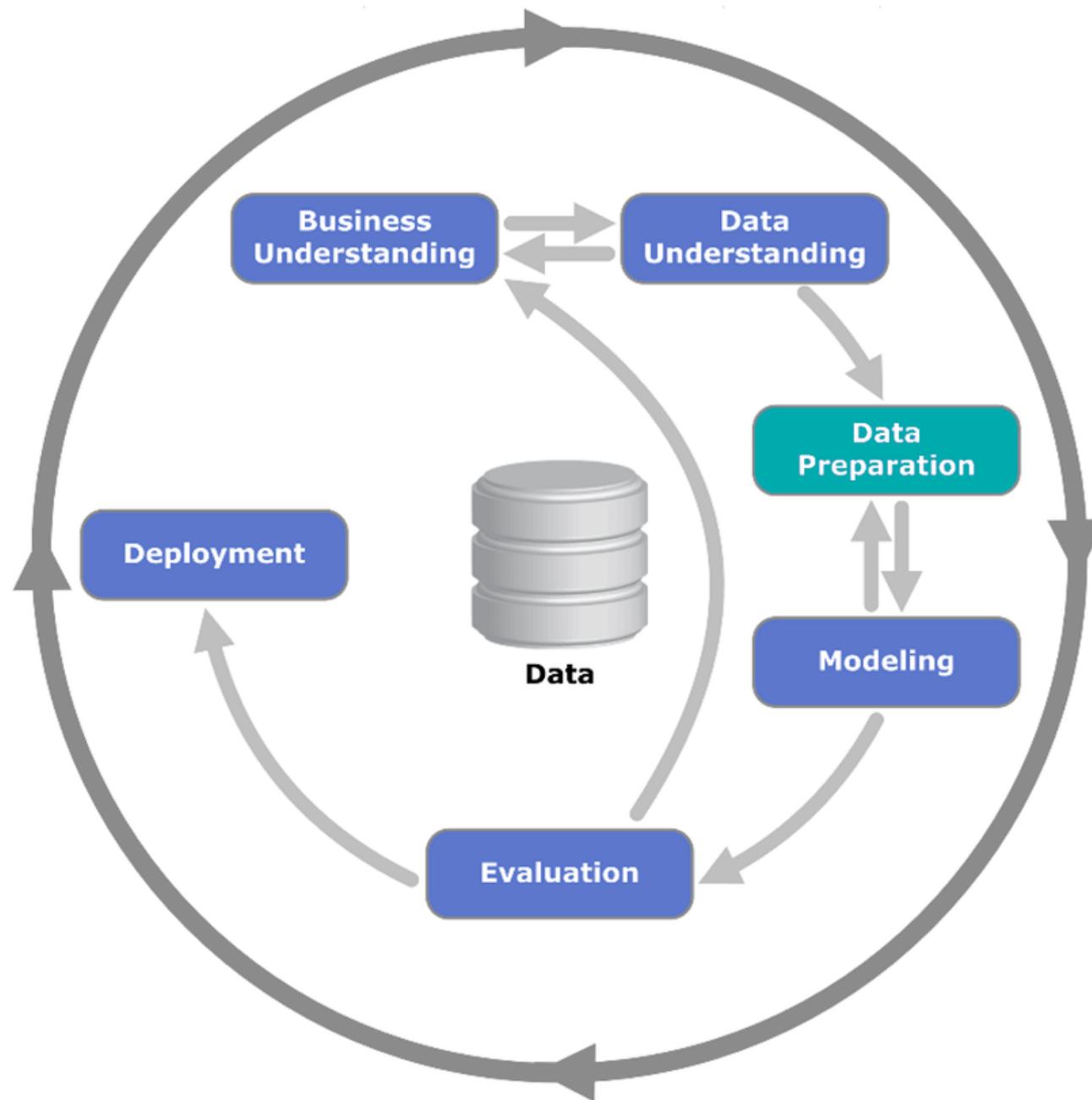


- **4. Design:** Conceive a product that solves the business problem or enables the business opportunity.
- **5. Construction:** Build the product, which should provide a return on investment within a predefined time frame.
- **6. Deployment:** Implement or sell the finished product, then measure its effectiveness to determine whether the solution meets, exceeds, or fails to meet the expected return on investment.

- Process without owner
 - Application-independent and
 - Context-independent
 - Tool-independent
 - It is a guide that consider the problem and techniques
 - Based on experience (developed in several workshops)
-
- 2 documents:
 - Reference model: describes phases, tasks and outputs.
 - User guide: practical application tips, check list



Phases



Executing Data Science Projects

BY JEFF SALTZ | LAST UPDATED APR 10, 2024 | LIFE CYCLE

Table of Contents

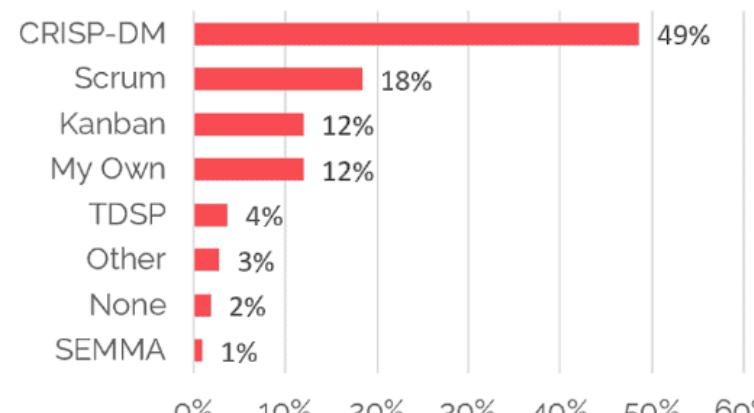
1. A quick review of the different alternatives
2. Previous Surveys
 - 2.1. Note on poll respondents
3. Google Searches
 - 3.1. CRISP-DM comes out as the most popular
4. Implications for teams
5. For additional information

During the past few months, we conducted a poll to see what project management framework teams used to help execute their data science projects.

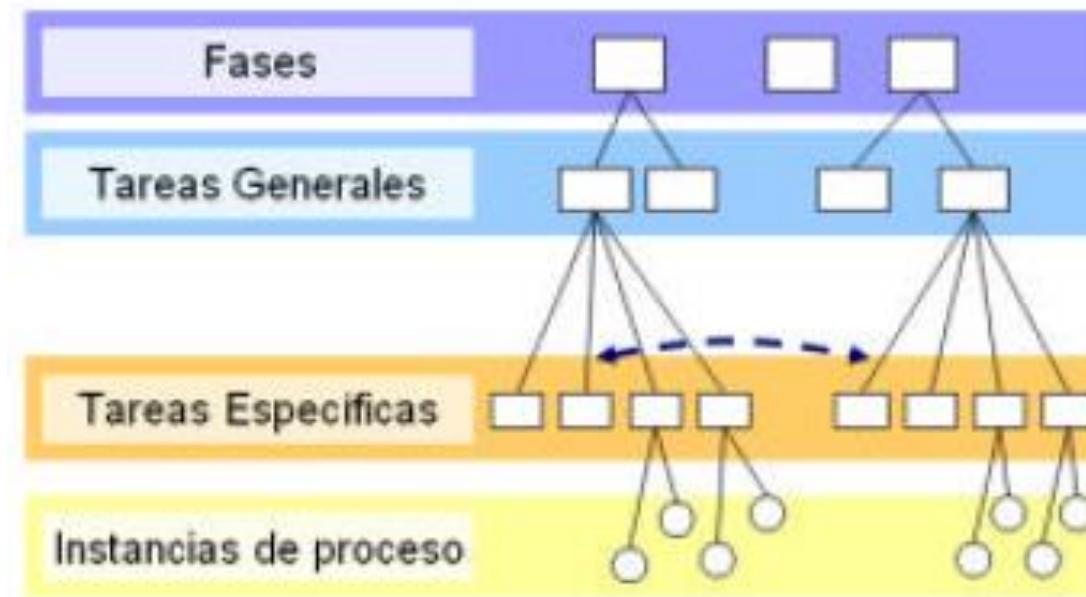
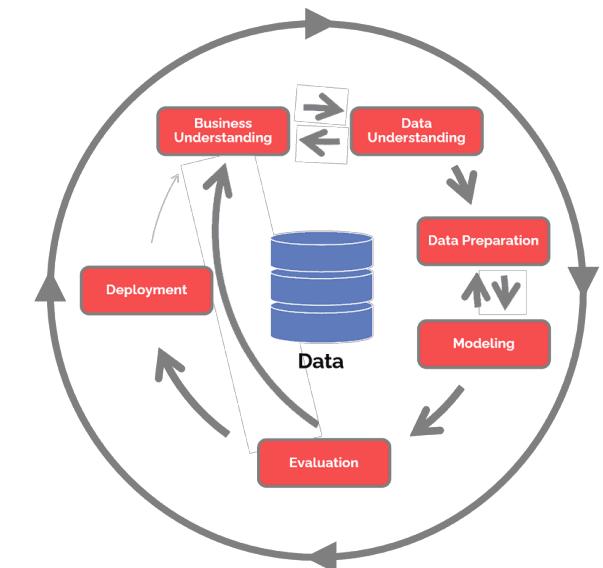
Based on our survey of 109 respondents, nearly half of the respondents most commonly use **CRISP-DM**. This was followed by **Scrum**, **Kanban** and "My Own". See results below.

datascience-pm.com Poll Results

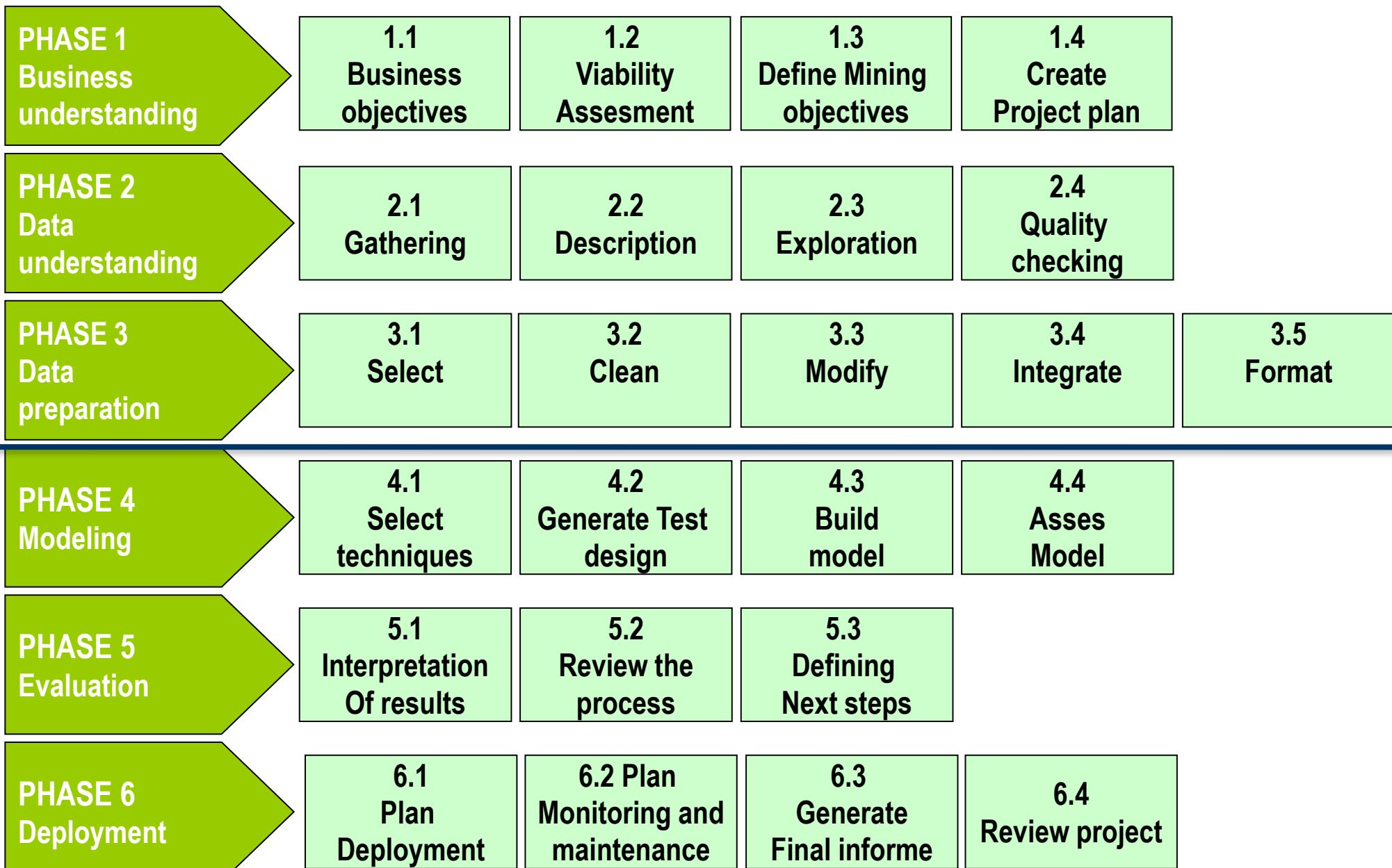
Which process do you most commonly use for data science projects?



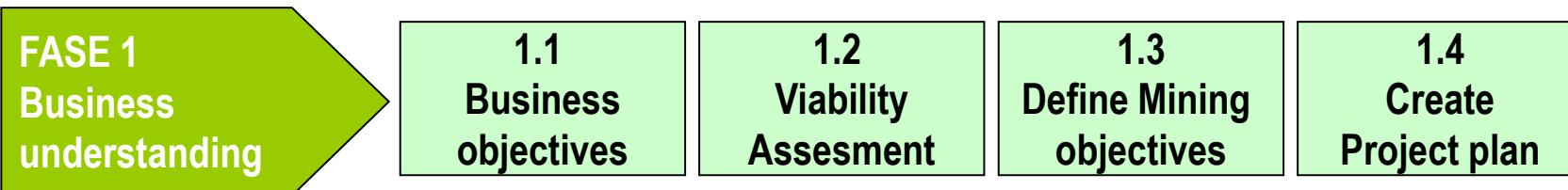
- **4 hierarchical levels**
 - General task; e.g.: cleansing
 - Specific task; e.g.: null cleansing



3 primeras fases representan el 80% del proyecto aprox, la tercera es la que mas tiempo lleva



un gran fallo que se puede tener es considerar que comprender el negocio es un objetivo secundario, cuando es de los aspectos más importantes



• 1.1 Defining business objectives

- Knowing what the customer wants from the business point of view
 - Is it to increase profitability campaign?
 - Is it to improve inventory management?
- Show factors affecting the project
- Establish success criteria
 - What does make the project successful?: Indicate objective measures

• 1.2 Viability Assesment

analisis dafo - debilidades, fortalezas, amenazas, oportunidades

- Indicate resources (people, data, software, ...), constraints, assumptions, requirements, and other factors: ARE WE ALLOWED TO USE DATA?
- Cost-benefit.
 - Equipment costs, human resources, etc.
 - ROI Benefits ...
- Risks: Data not available, not available knowledge, tools

FASE 1
Business understanding

**1.1
Business objectives**

**1.2
Viability Assessment**

**1.3
Define Mining objectives**

**1.4
Create Project plan**

•1.3 Define mining objectives

- Specific objectives of the problem:
 - Translate customer target into mining goals.
 - Eg segmentation, sequential patterns, and so on.
- Establish technical objectives
 - Translate goals into parameters of the output: rate of success, failure prediction, error cost, etc.

•1.4 Create project plan

- Establish steps:
 - Step: duration, resources required, inputs, outputs.
 - Scheduling and risks management
- Initial selection of techniques and tools



• 2.1 Gather

la parte de integracion suele ser muy complicada

- Define sources, integration, localization, problems and solutions, ...

• 2.2 Describe

- Formats, keys, amount...

• 2.3 Explore

- Query, visualize, report
- Value distributions, aggregations of data, statistics (properties, mean, std,...)

• 2.4 Quality check

- All type of data? All classes represented? Enough data? Complete data? Null or impossible values?

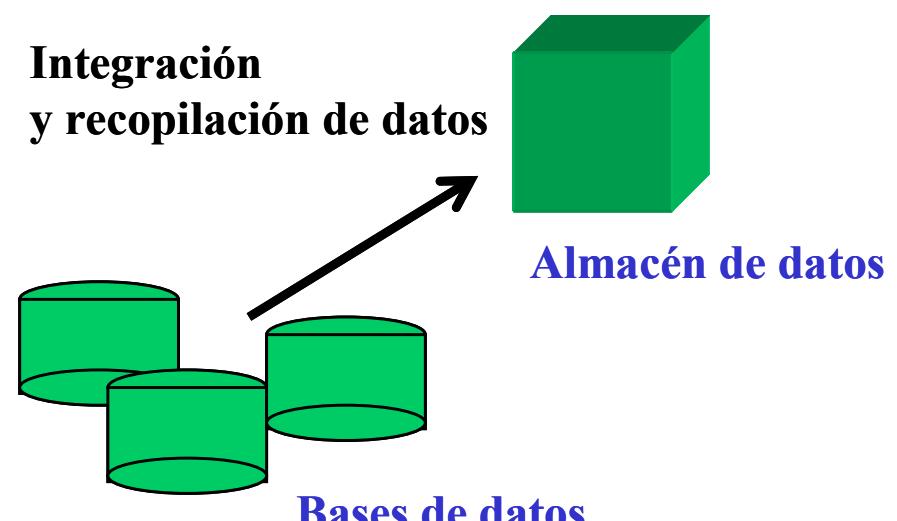


- It involves most of the total time

• 3.1 Selecting Data

- Selecting tables, attributes and rows
- Volumen suitable for tools
- selection:
 - Partitioning the data set:
 - Training data
 - Test Data
 - Validation Data
- Sampling Data

**Integración
y recopilación de datos**



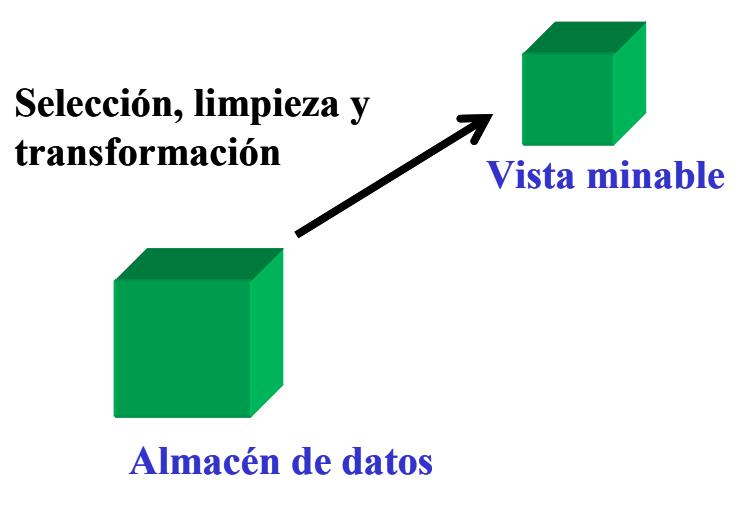
• 3.2 Cleaning: incomplete and erroneous data records:

- Improving data quality
- Select subset of data, replace null
- Complete, delete or ignore

esto es como un 50% del total del proyecto

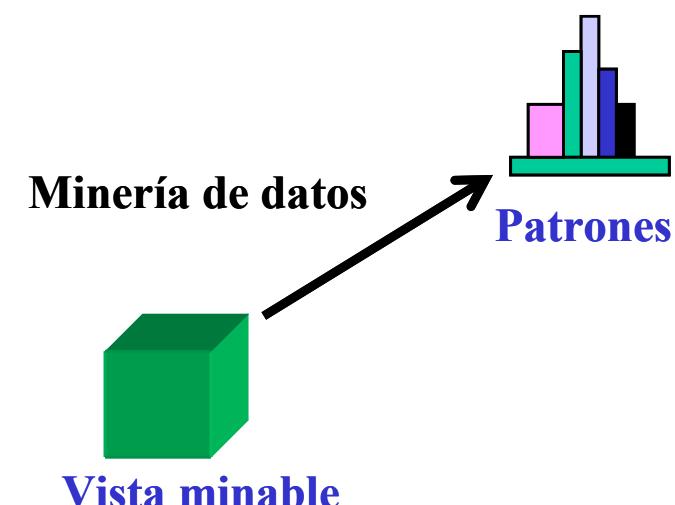


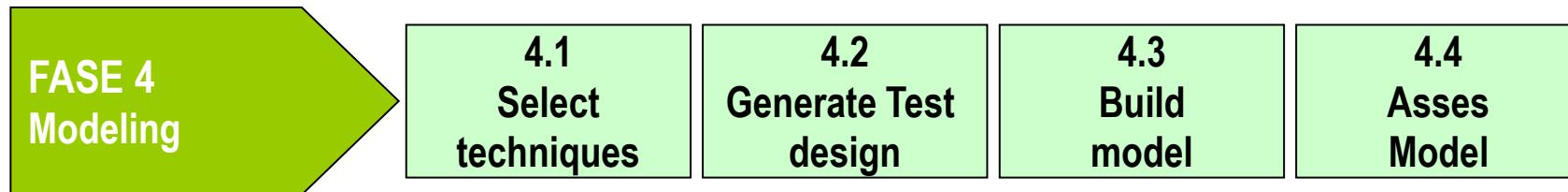
- 3.3 Transformation:
 - Choose more relevant attributes
 - Deriving most significant new features from the original
 - Discretize, map, normalize, ...
- 3.4 Integrate: **DATAWAREHOUSE**
- 3.5 Format: if needed by techniques
- Known as ETL
 - Extraction, Transformation, Load
 - Often uses a graphical model
 - Frequency of execution
 - It involves:
 - Execution monitoring
 - Recording
 - Exceptions and errors





- We assume a minable view: A table containing the relevant attributes, labeled as inputs or outputs
- 4.1 Select techniques: considering
 - Appropriateness to the problem
 - Classification, prediction, clustering, association, dependencies
 - With adequate data
 - Meeting the requirements of the problem
 - Executable in time
 - Knowledge of the technique





- 4.2 Design tests

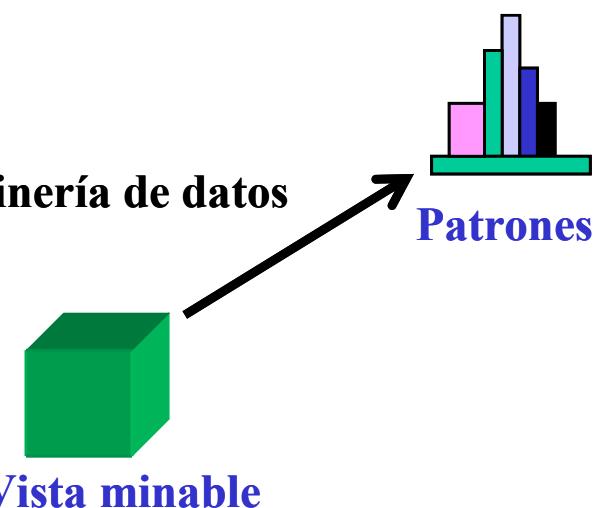
- Establish training, testing and validation
- Establish criteria for goodness of models

- 4.3 Creating model

- Set parameters
- run

- 4.4 Evaluate model

- Meets goodness?



FASE 5
Evaluation

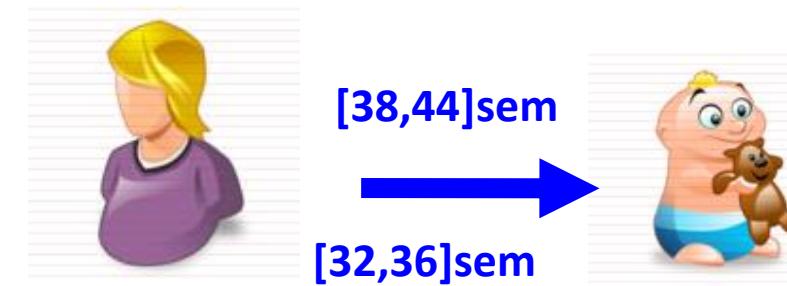
5.1
Interpretation
Of results

5.2
Review the
process

5.3
Defining
Next steps

- **5.1 Interpretate the results**

- Is the problem solved?
- Is the answer appropriate?
- Is it valid?
- Is the business objective well defined?
- Is the knowledge new?
- Is the model useful?
- Is it better than we had?
- Are there too many patterns?



(hair=blonde)
(profession=actress)

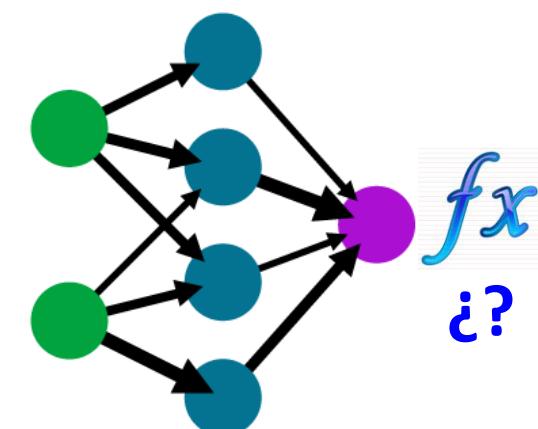
carrier=no

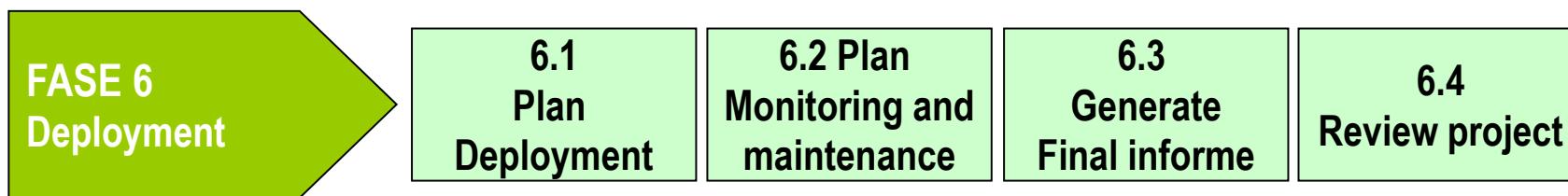
- **5.2 Review the process**

- Any error from the technical point of view?
- Have we overlooked anything?

- **5.3 Define next steps**

- Iterate? Deploy? Rebuild?





• 6.1 Plan deployment

- Who are the users?
- How and when the model will be used?
- How is it deployed? As a tool?
- A computer program is necessary? Paper?
- Strategic system for a doctor: hypothesis proposal
 - Screening, diagnose, prognosis

• 6.2 Plan monitoring and maintenance

- Is it being used? Is it properly used?
- Updatable models?

• 6.3 Generate final report

• 6.4 Review the project

- Strong and weak points, any aspect that can be improved?...

- **A Practical example: Research center focused on using remote sensing for assessing water quality in the Rías**
 - Business Understanding
 - Objective: Monitor and forecast water quality in the Galician estuaries using satellite data (business objective).
 - Viability Assessment:
 - Does this project align with the lab's strategy? + estimated costs & expected ROI + human and technical resources + estimated time + risks + expected outcomes.
 - Data mining objectives: what we want to obtain?: descriptions, predictions, classifications, detect anomalies, time-series,...
 - Create project plan:
 - Scope + define region of interest and timeframe + WBS & Tasks + Resources + milestones + internal and external communication, ...work breakdown structure

- **A Practical example: Research center focused on using remote sensing for assessing water quality in the Rías**
 - Data Understanding
 - Gathering:
 - Identify data sources: satellite, in-situ, external (GIS layers, climatologies), etc.
 - Obtain the data
 - Description: Formats, quality, we have to understand the data.
 - Exploration: statistical summaries, visual inspection, explore correlations, etc.
 - Data quality: outliers, missing data, does satellite and in-situ data correlate? Are the different datasets consistent?

- **A Practical example: Research center focused on using remote sensing for assessing water quality in the Rías**
 - Data Preparation (overlap with ETL):
 - Select: spatial, temporal + satellite bands
 - Clean: what we do with missing data and outliers? How to remove/reduce noise?
 - Modify: transform data (e.g. atmospheric correction to satellite data), normalize, feature engineering (e.g. to generate value added products), ...
 - Integrate: different data sources, spatial and temporal integration,...
 - Format: ensure we are using the same coordinate systems, raster vs vector, storage formats, ...

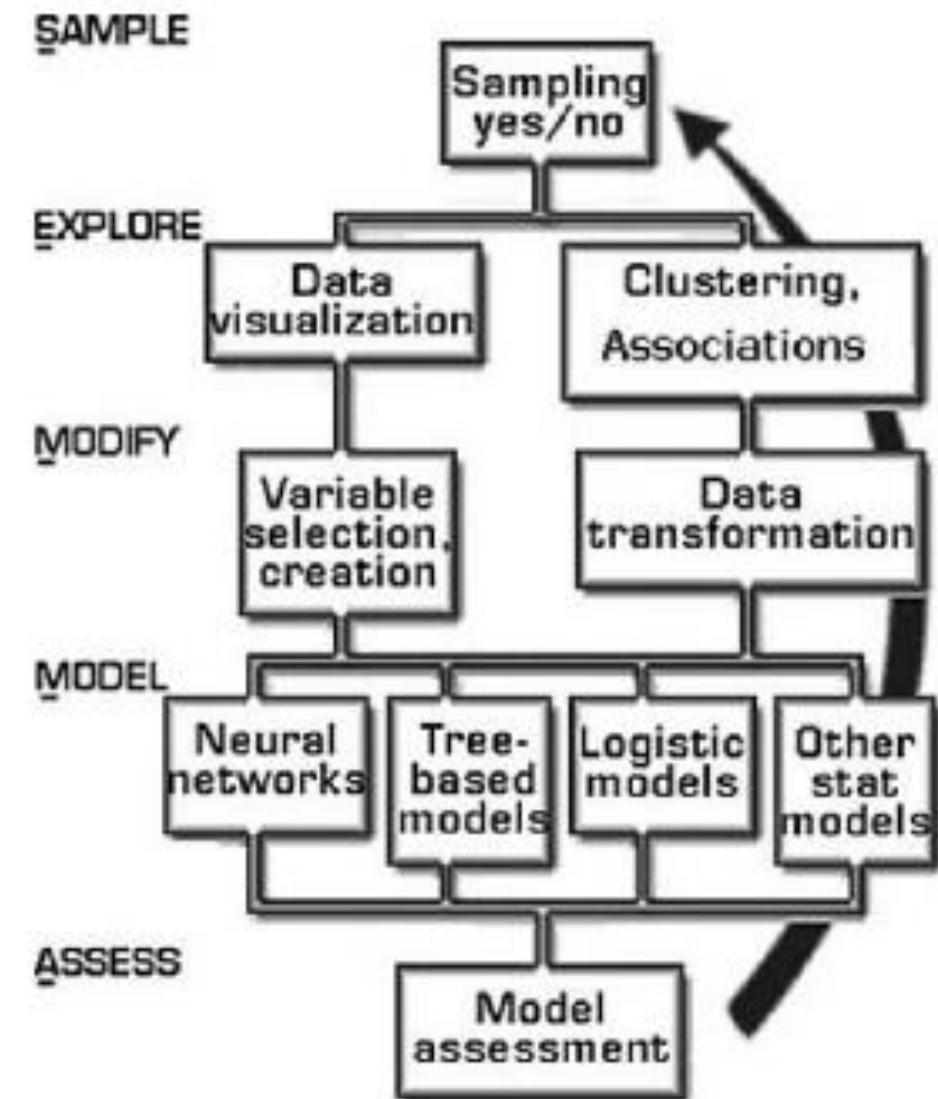
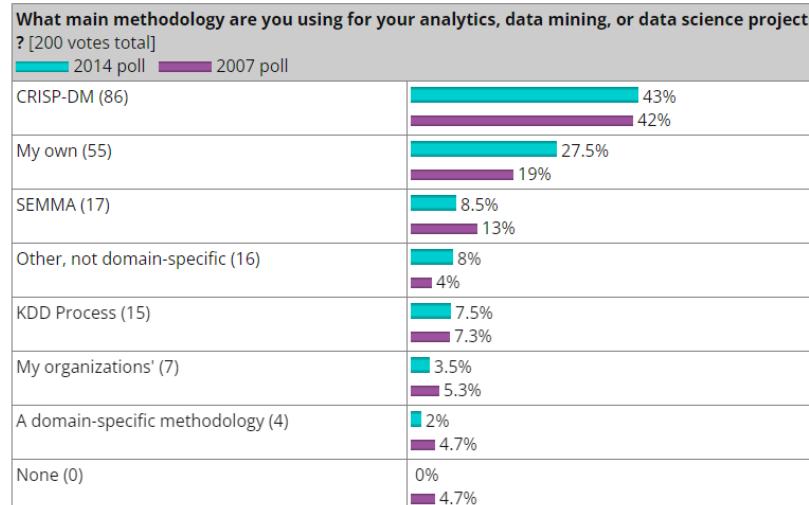
- **A Practical example: Research center focused on using remote sensing for assessing water quality in the Rías**
 - Modeling (usually very iterative):
 - Select techniques: based on the objectives, available data and the desired outputs: classification, regression, clustering, association, ...
 - Generate test design: training, validation and test datasets + what metrics are we going to use?
 - Build model: train the model and adjust hyperparameters
 - Assess model: assess performance + understand and interpret the model (e.g. when it performs poorly)

- **A Practical example: Research center focused on using remote sensing for assessing water quality in the Rías**
 - Evaluation:
 - Interpretation of the results: how good is the model for achieving the objectives of the project? Do we detect spatial/temporal patterns? Are there factors such as population/weather affecting water quality? Do the experts agree with the outcomes? How do the results compare to baselines?
 - Review of the process: we review the data, the techniques, the metrics, the feedback,...
 - Defining next steps: Does the model needs improvements? If it works well, deploy? Do we need to retrain the model with new data? Update the algorithm? Expand it to other regions? ...

- **A Practical example: Research center focused on using remote sensing for assessing water quality in the Rías**
 - Deployment:
 - Plan deployment: do we need to integrate the model into an existing system (e.g. in the cloud, GIS)? How do we plan to access the data to feed the model? What are the outputs and in which format?
 - Plan monitor and maintenance: how often we validate our system? We need to define metrics to assess model performance and trigger retraining. What happen in case of errors? Do we need a data backup?
 - Generate final report: describe project, data sources, procedures, results, recommendations, etc.
 - Review project: did we learn anything? Things that went well, challenges, bad things. We need to use this experience to improve future efforts.

- **CATs: Clementine Application Templates: [CATs]**
 - Specific libraries with best practices for specific applications: mapping to a project type
 - Following the CRISP-DM standard.
 - Each flow CAT is assigned to a phase of CRISP-DM.
- **Templates:**
ejemplos de plantillas
 - Telco CAT - loyalty and customer retention for telecoms
 - CRM CAT - understand and predict customer migration between segments,
 - Microarray CAT - specific functions for biological applications, finding genes for therapeutic purposes, predict genetic diseases
 - Fraud CAT - predict and detect fraudulent transactions, claims, taxes and so forth.
 - Web CAT - predict visitor behavior, access and merge web log data
 - Crime CAT- Identify and predict areas of high crime, offender type identification

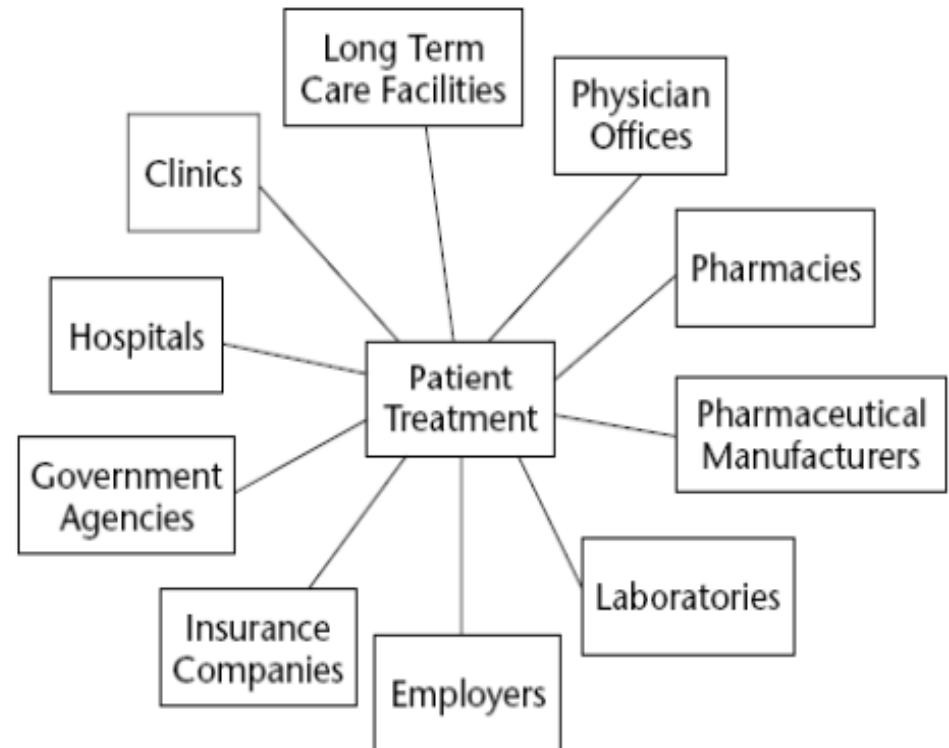
- SEMMA: SAS Enterprise Miner proposal for data mining projects
- Sampling:
 - A good sampling strategy almost guarantees good patterns
 - Rare patterns are statistically discovered
 - Subsets: training, validation, test
- Explore: simplify the problem



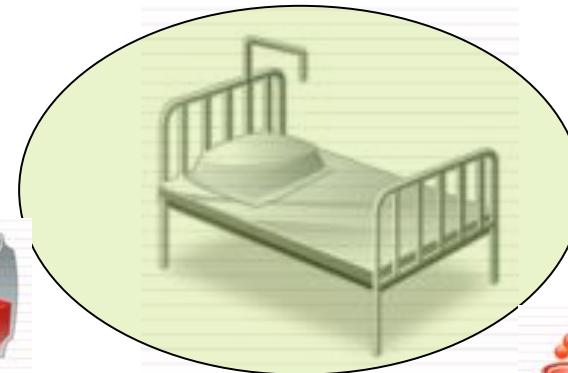
ejemplo:

◆ Application

- Definition of Health policies
- Detection of inefficient services
- Fraud detection by health providers



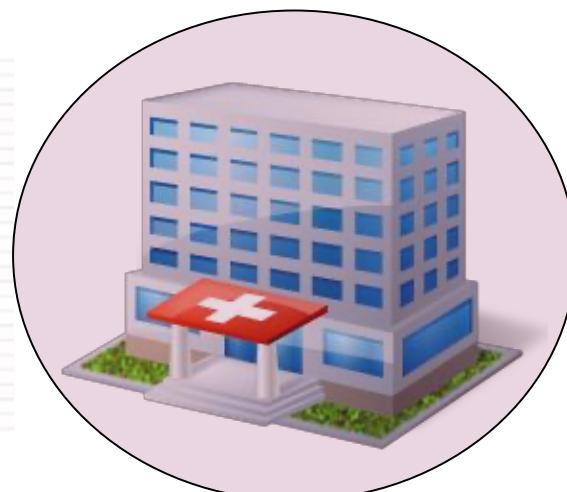
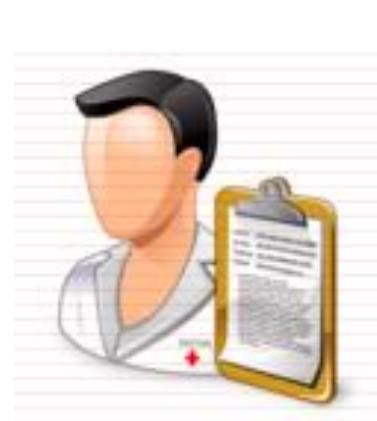
como mapa mental pero bien estructurado



**PATIENT
CARE**



MANAGEMENT



RESEARCH



FASE 3
Data preparation

**3.1
Select**

**3.2
Clean**

**3.3
Modify**

**3.4
Integrate**

**3.5
Format**

- Besides:

- Imprecise data, noise, incomplete data
- Subjective data due to human interaction
- High volume of data
- Complex and dynamic data
- Numeric, nominal, time series, images, video, 3d
- Text: reports, interpretation, ambiguous expressions, ...
- Difficult to characterize mathematically



- We find

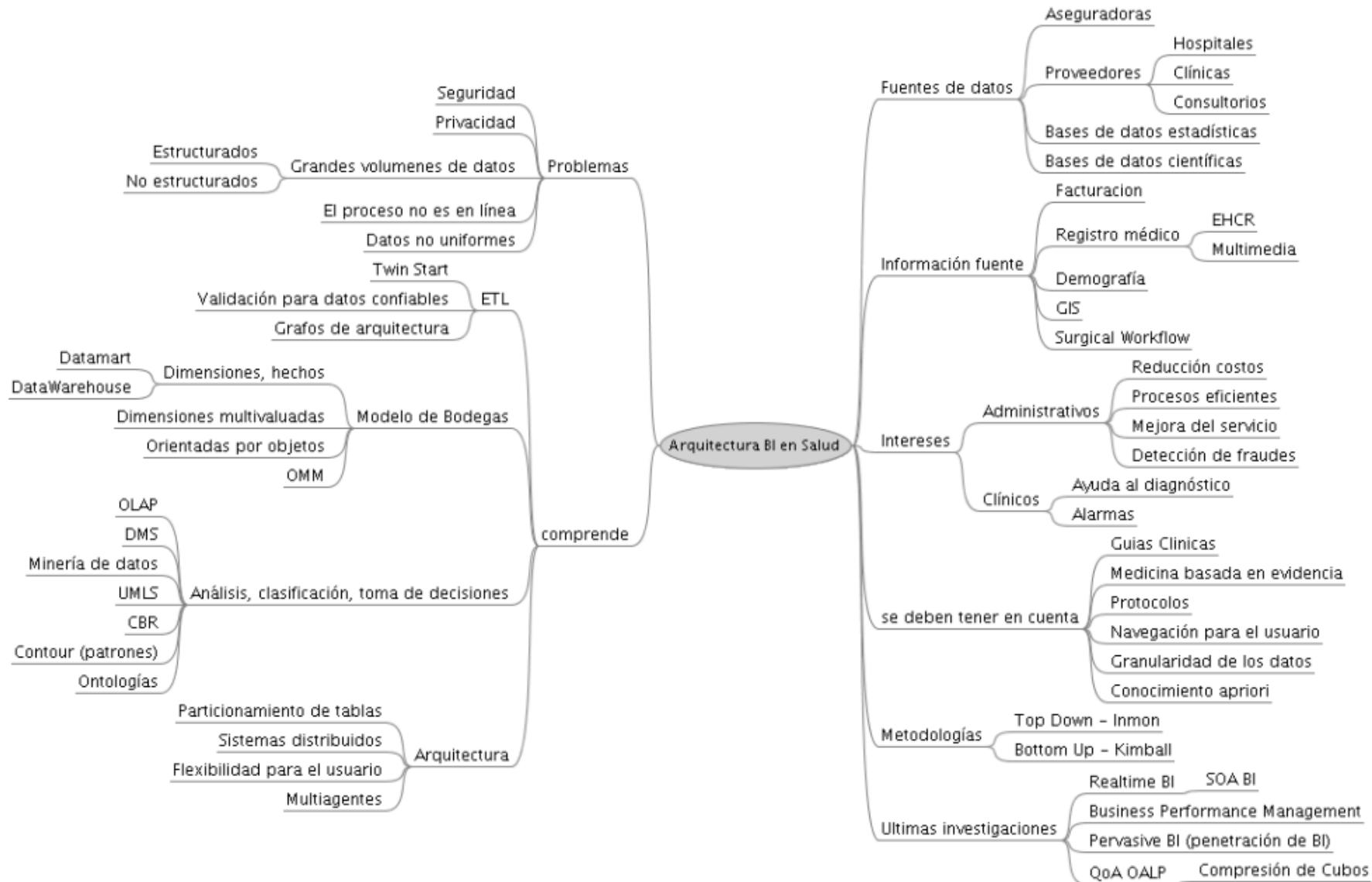
- Proprietary EHR
- Too many standards
- Low interoperability



- It implies:

- Big difficulty to acquire and consolidate changing data

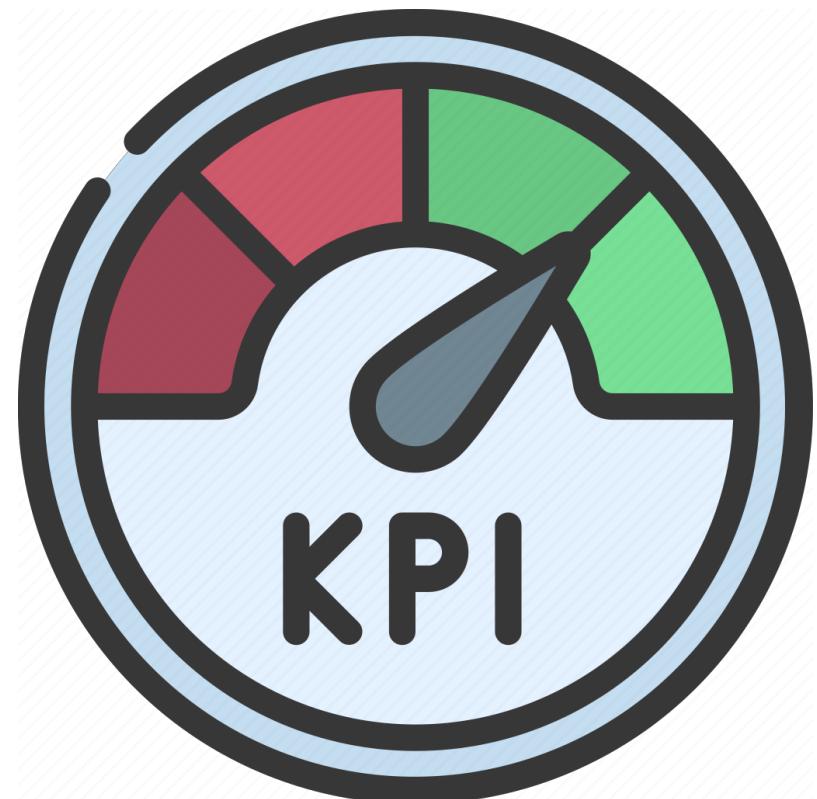




Business intelligence

Unit 1b – Key Performance Indicators

- A lot of words for Indicator, but mainly:
- KPIs represent a set of measures focusing on those aspects of organizational performance that are the most critical for the current and future success of the organization (Parmenter)
- Lets see this in detail



Goals, success factor, indicator

ejemplo: ganar la champions

ganar la champions

- **KGIs, Key Goal Indicators**, are defined, as measurable targets based on broad objectives such as increasing sales, improving customer satisfaction, raising availability, and raising employees' skill level. KGIs are defined as concrete goals.

buenas politicas de fichajes, buena cantera, gran numero de abonados, ...

- **CSFs, Critical Success Factor**, These are the essential elements needed to meet the KGIs. They identify the most important areas that need attention to ensure that goals are achieved. CSFs clarify what needs to be focused on to reach the KGIs.

define lo bien que van los CSF, que avancemos en ello, kpis asociados a cada csf

- **KPIs** to measure the current status of business processes related to CSFs. Usually a % or rate. Help assess progress towards meeting the goals and CSFs.

- Example:

- Objective: good health
- KGI: Improve results medical tests
- CSF1: Balanced diet CSF2:Regular physical activity
CSF3:Emotional well-being
- KPI for CSF1: Number days per week following diet
- KPI for CSF2: Number of days per week doing at least 30min exercise
- KPI for CSF3: Number of hours of sleep per day.

KGI: tener más beneficio
CSFs: publicidad, reducir costes y mejorar eficiencia,
KPIs: llevar desglose de costes, ver idíariamente defectos en las latas, ...



- **Example:**

- KGI: 10% increase in sales
- CSF: Increase net sales per customer
- KPI: the customers of accessories of more than \$ 10 >30%

- **Example:**

- KGI: Increase 10% customer satisfaction
- CSF: Provide positive customer experience
- KPI: Call Abandon rate
- KPI: Call Waiting rate
- KPI: Service Desk availability

just
another
example

Target, action

Once KPIs are defined, proceed to the definition of desired targets, decision about concrete actions, and so on. **Critical success factor:** get new customers

Indicator 1 New sales

Type:

Outcome

monitorizamos su estado pero no permite tomar decisiones para que haya mejora el mes que viene

Target:

5 per month



Action:

Follow-up

refrescarlo para ver como va

Indicator 2 New offers and proposals

Type:

Driver

se usa para estimar futuras metas

Target:

60 per month

Action:

Publish in social-media

ejemplo: he vendido 5 apps: eso es una medida
metrica: he vendido 5 apps por semana

- Indicator vs Measure: indicators are obtained from metrics, which are composed of measures. Indicators often aggregates or combines multiple metrics.
- [Kaplan&Norton] There are two fundamental types of KPIs: outcomes and drivers:
 - Outcome KPIs—sometimes known as lagging indicators—measure the output of past activity.
mas fáciles de definir
 - Driver KPIs—sometimes known as leading indicators or value drivers—measure activities that have a significant impact on outcome KPIs.



Indicators: Example

- **Objective**: Arrive at a meeting at 12.00.
- I'm in the car, what do I want to know? Which of the answers to these questions help me to act in order to fulfill the objective?

un leading para un objetivo puede ser lagging para otro !!!!!!



- Will I be late if I continue at the current speed? leading
- How many kilometers do I have left to reach the destination? leading
- How much time do I have left to reach the destination? leading
- How many kilometers have I traveled? lagging
- How long have I been traveling? lagging
- What is the current fuel consumption? lagging
- How many liters are left in the tank? leading
- How many kilometers can I travel without refueling? leading

Indicators: Example

- **Objective**: Arrive at a meeting at 12.00.
- What if there are more goals? For example:
Lower cost.
 - Can I throw the bags so that the car weighs less and spend less fuel to reduce cost while maintaining speed?
 - Can I bribe the cops so they don't fine me if I get up the speed?



- More examples:

- **Mortality rate:**

- Action is possible?
 - Can I “correct” it?
 - If not, It is a KRI not a KPI.

- **Length of stay:**

- Action is possible?
 - Can I just discharge the patients early to improve this?
 - It is possible, but other KRI such as readmissions will increase.



igual que KRIs y KPIs, pero sin ser KEYS

Parmenter - autor muy reconocido

- **RESULT indicators (RIs)** – are summary and retrospective indicators, that provide information about what has been done. They summarize **past** performance.
 - these can be measured daily, weekly, monthly, or quarterly.
- **PERFORMANCE indicators (PIs)** – track specific aspects of performance that are important but not critical for meeting strategic objectives. They are typically used for more focused areas of operation, providing information that can help teams improve day-to-day processes but are not as high-level or impactful as KPIs.
 - These are measured daily, weekly, or monthly and are not as important as the KPIs.
 - **Number:** 30-50 PIs and RIs.

10(KPIs max) - 80 (Pis + KIs maxs) - 10(KRIs max)

- **KEY RESULT indicators (KRIs)** – high-level summary indicators that provide broad insights into how well an organization or department has performed in achieving key goals. They help communicate to the board or senior leaders how effective management has been in executing the strategy, but they are not detailed enough to show how to improve performance in real-time.
 - Typically these are measured **monthly** or quarterly. **Number:** 6-10.
- **KEY PERFORMANCE indicators (KPIs)** –are the most critical indicators for driving strategic objectives. They give real-time, actionable insight into whether the organization is on track to meet its goals. dramatically.
 - These are measured either 24/7, daily, or weekly. **Number:** 6-

- KPIs characteristics [Parmenter] según el pibe, no es la biblia
 - Usually non financial measures (not expressed in €, \$,...)
 - Measured frequently e.g. 24 by 7, daily or weekly
 - Acted upon by the CEO and senior management team
 - All staff understand the measure and what corrective action is required
 - Responsibility can be tied down to a team
 - Significant impact e.g. it impacts on more than one of top CSFs and more than one balanced scorecard perspective
 - It is imperative that the importance and impact of these measures are well understood by all, and the corrective action required if the performance is not reaching targets.

- Responsibility can be tied down to a team
 - “Within 12 months, every employee will be able to check their email, calendar and their personal top 10 KPIs at the breakfast table using their smartphone or tablet.”
 - Performance management needs to become personal, simple and directly effective. We need to put business metrics in the hands of the people who do the work and drive performance and make sure that KPIs leave the ivory tower called the boardroom.



Indicators

- Other classification depending on what you are measuring

indicadores complementarios a lo que ya hemos visto

- Risk Indicators show the risk to exceed the defined risk appetite in the future and should be able to accurately predict losses. EJEMPLO: KPI de reducir un % costes medido como número de catéteres usados incrementa el KRI de tendencia reducida en número de infecciones de bacteriemia. Una tendencia sube y la otra baja, es una mirada al futuro.
- Control Indicators, set the desired internal control effectiveness of an organization. EJEMPLO: adherence to security standards to prevent data losses.
- Lead Indicators are being increasingly used to measure the achievement of strategy goals (for instance, in terms of customer satisfaction).
- DI: Diagnostic indicators help analyze and understand past performance

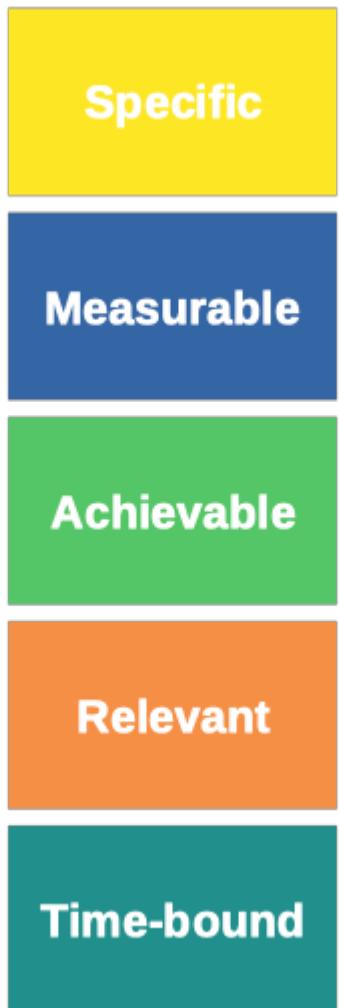
- Qualitative vs Quantitative.
- Quantitative data measures activity by counting, adding, or averaging numbers.
 - E.g.: inventory, purchasing, orders, accounting, employee injuries, number of training classes, ...
 - Quantitative data forms the backbone of most KPIs.
- The most common qualitative ones gather customer or employee satisfaction through surveys.
 - While the survey data itself is quantitative, the measures are based on a subjective interpretation of a customer's or employee's opinion on various issues.
 - These opinions can help explain why performance is dropping when all other indicators seem fine.

- **Specific:** it has to be clear what the KPI exactly describes and the context within which it is defined; Target an specific area rather than a general one

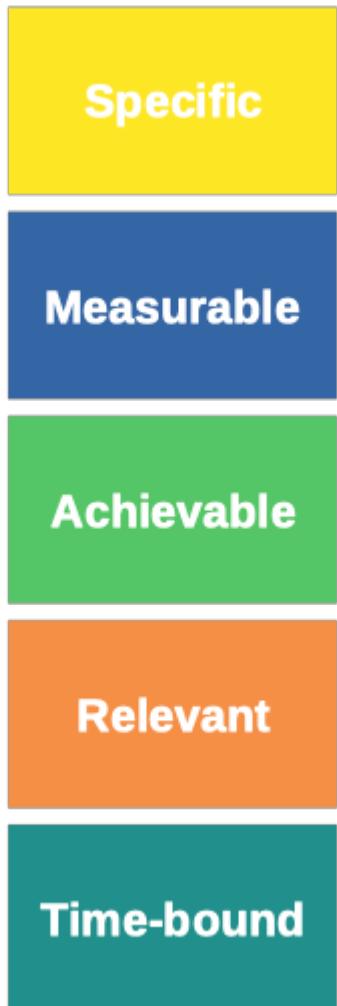
- What you want to achieve?
- Who is involved?
- Why is important?

- **Measurable:** Must be quantifiable

- Performance can be measured
- How much? How many? What is the baseline? What is the timeframe?



- **Achievable:** realistic and can be accomplished with the resources, constraints and time we have.
 - We need to assess resources
 - Take into account stakeholders, external factors and past performance
- **Relevant**, align with organization's goals.
 - It should matter to the organization.
- **Time-bound** – specify when the result(s) can be achieved
 - Measured each week? Time scale suitable.
 - Facilitates planning and tracking.

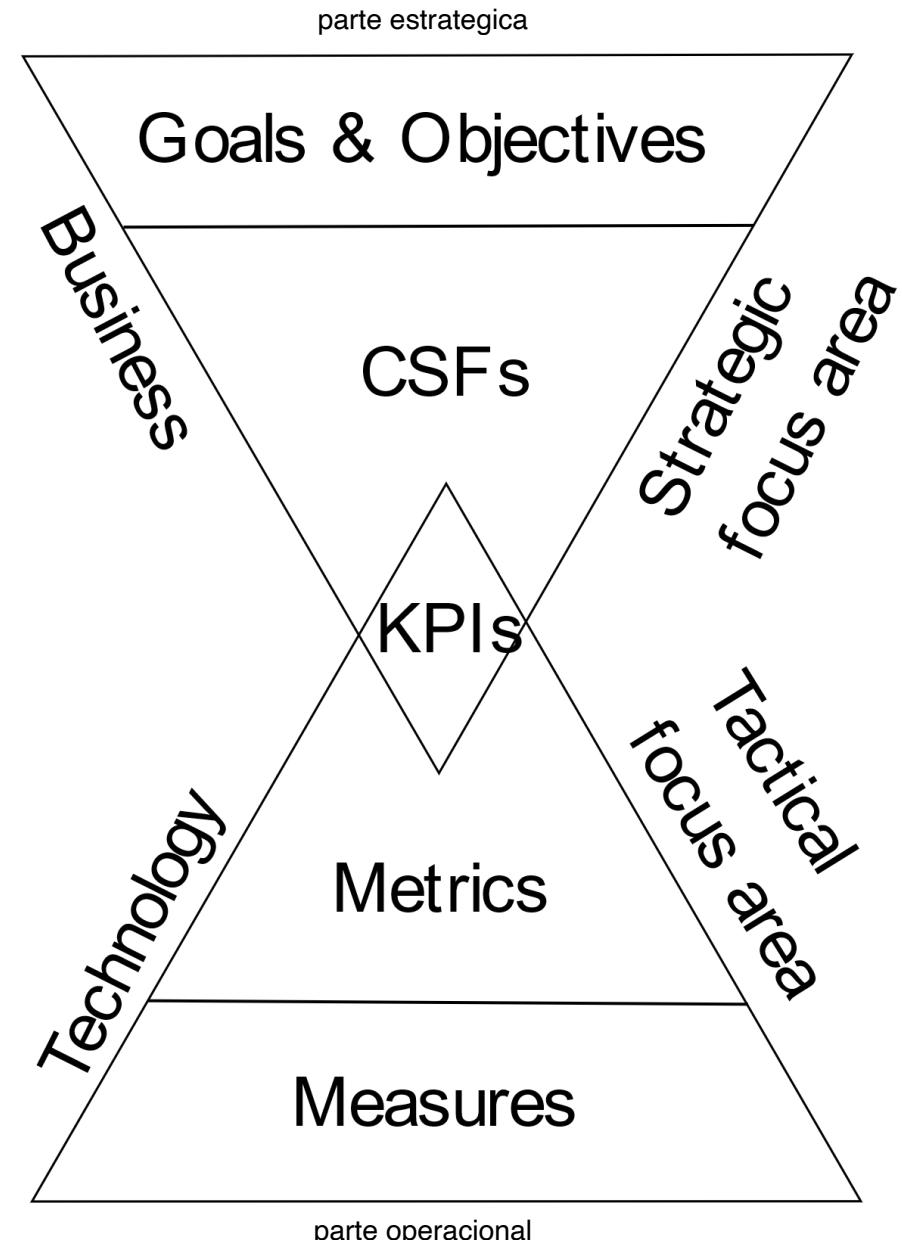


- **Others**
 - Comparable
 - Used to compare different organization and data available
 - Verifiable
 - For an effective data collection
 - Process repeatable with robust data
 - Cost-effective
 - The cost of measuring and collecting information must be balanced with benefits

KPIs: why we need them?

con los KPIs podemos:

- Evaluate status/progress
- Identify changes
- Show if we are meeting our strategic business objectives
- Drivers of improvement
- Save time and increase efficiency
- Improve decision making
- Support Benchmarking



- **KPI in HealthCare**

- Exitus rate
- Admission and readmission rates
 - ICU without walls
- Patient Wait Time: before check-in
- Patient Satisfaction
- What others?
 - **For next week: You have to search for national/local KPI for healthcare system or other field.**
 - **Eg. Vaccination rate.**



traer 3 KPIs apuntados para la semana que viene

fallos tipicos de los KPIs: fallos de comunicacion, de trabajo en equipo, de compromiso, ...

Just thinking a little bit

- KPI for SDG

- Example: SDG#6-Water and Sanitation

- Target 6.5: By 2030, implement integrated water resources management at all levels, including through transboundary cooperation as appropriate

- Indicators:

- 6.5.1- Degree of integrated water resources management implementation (0-100)

- 6.5.2- Proportion of transboundary basin area with an operational arrangement for water cooperation

- SDG#1- No Poverty

- Target 1.2: By 2030, reduce at least by half the proportion of men, women and children of all ages living in poverty in all its dimensions according to national definitions

- Ind 1.2.1- Proportion of population living below the national poverty line, by sex and age

- Ind 1.2.2- Proportion of men, women and children of all ages living in poverty in all its dimensions according to national definitions



Just thinking a little bit

KPI - Key Performance Indicators, global DU

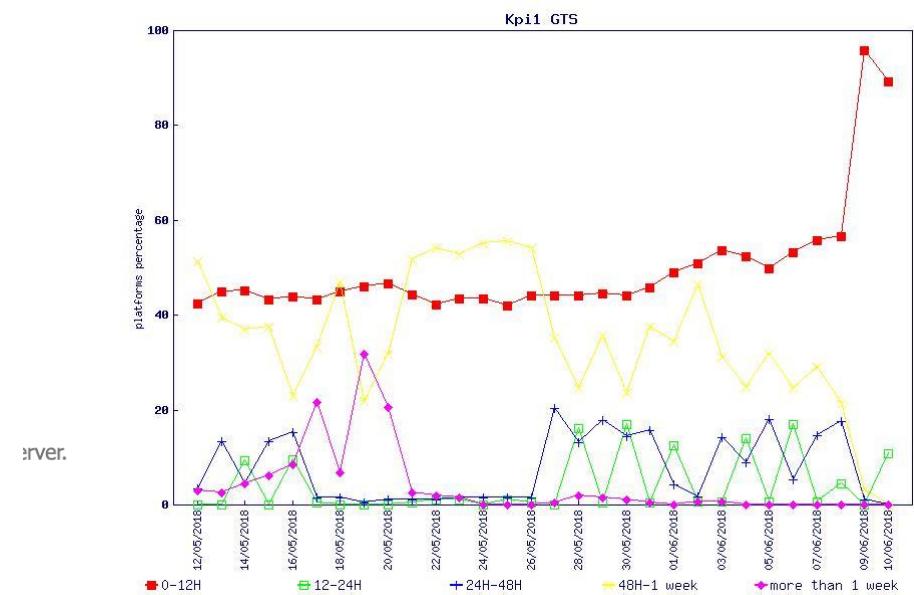
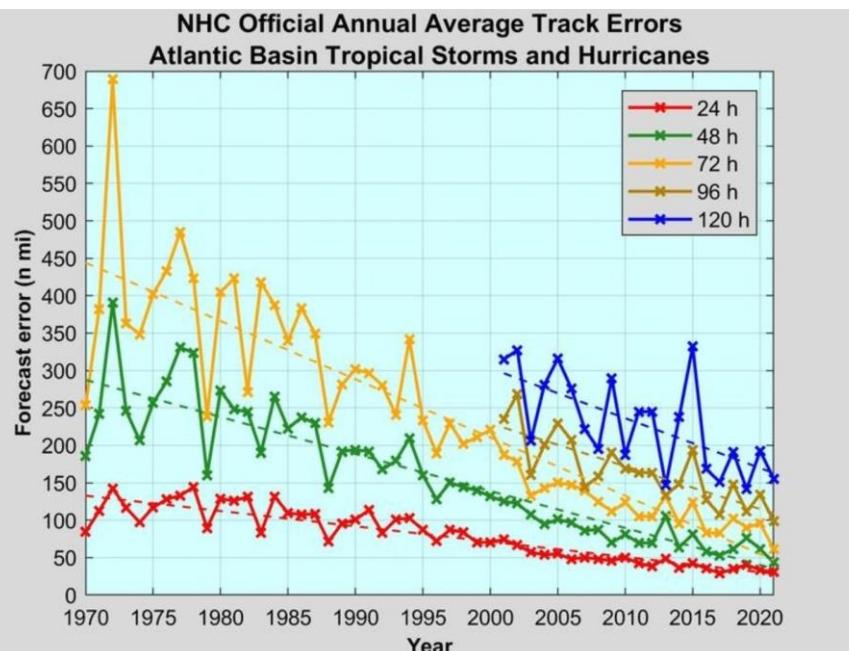
The Key Performance Indicators of the global in-situ TAC are daily calculated.

They provide indicators on availability and quality of in-situ data distributed by CMEMS.

KPI-1 : data availability

KPI-1 monitors the delay between the last date of observation and the first date of availability on CMEMS ftp server.

- > Drifting buoys
- > Gliders
- > GTS
- > Moorings
- > Profiling floats
- > Vessels



**NOAA Annual Average Track
Errors for Tropical Cyclones in the
Tropical Atlantic**

Business intelligence

conjuntos de datos a un determinado objetivo, son integrados, son volátiles, cambian en el tiempo, orientados a dar apoyo a la toma de decisiones

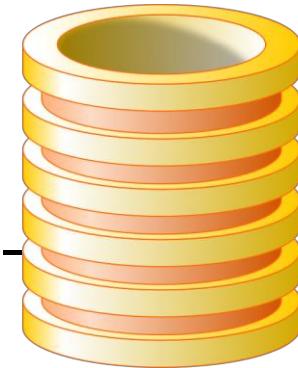
Unit 2 – Datawarehouse and OLAP
S2-1 – Datawarehouse

Introduction to datawarehouse

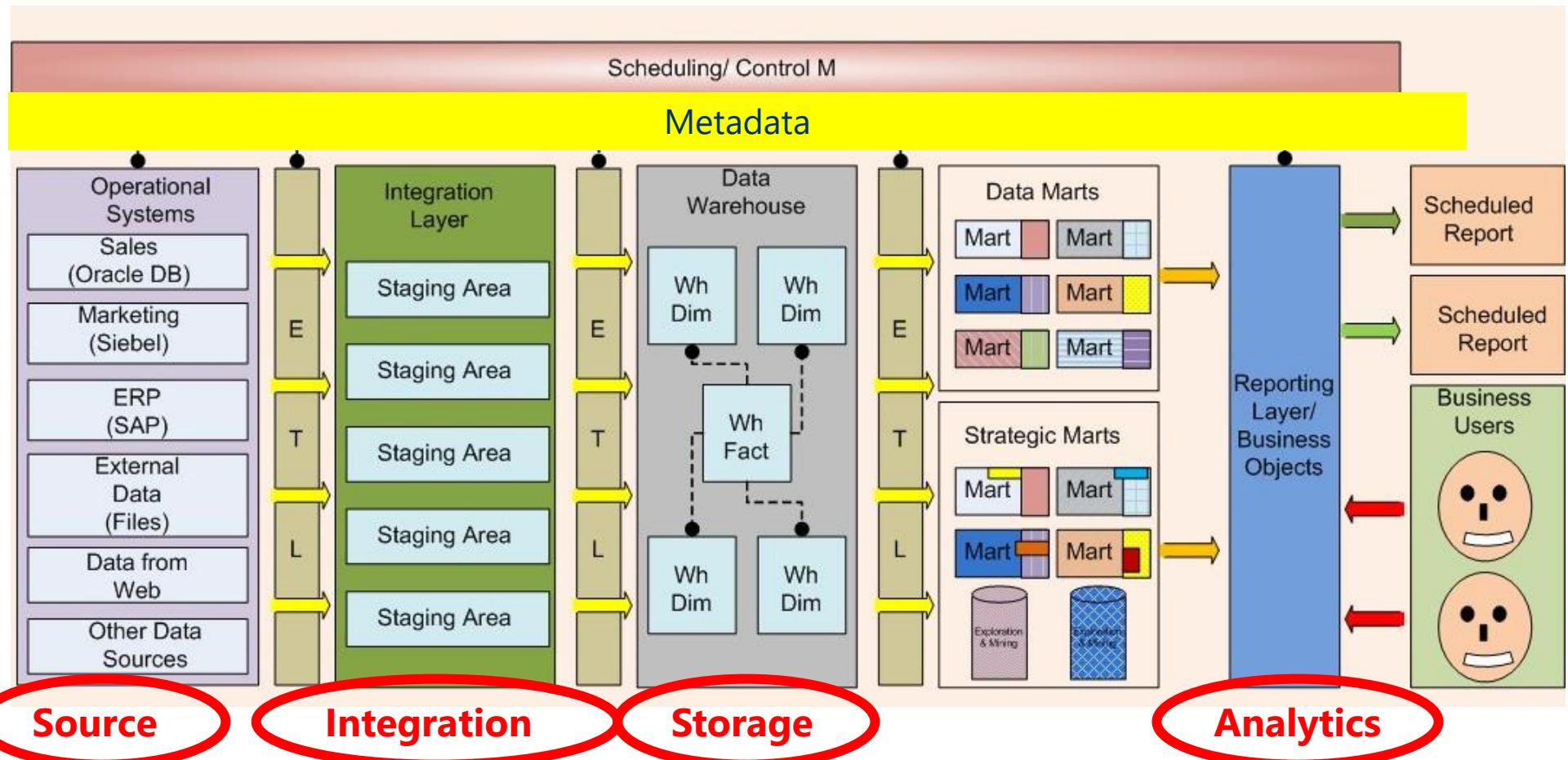
- **Objective:** to provide infrastructure that support **DSS (Decision Support System)** within an organization
 - Organizations begin **with basic** systems to manage **day-to-day operations**.
 - Powered by multiple **operational databases** (e.g. sales, inventory).
- **Requirements for DSS**
 - New requirements for **complex queries** and **data integration** of historical and operational data to generate knowledge.
 - Purposes:
 - Analysis of the organization**
 - Make predictions**
 - Define strategies**

Introduction to datawarehouse

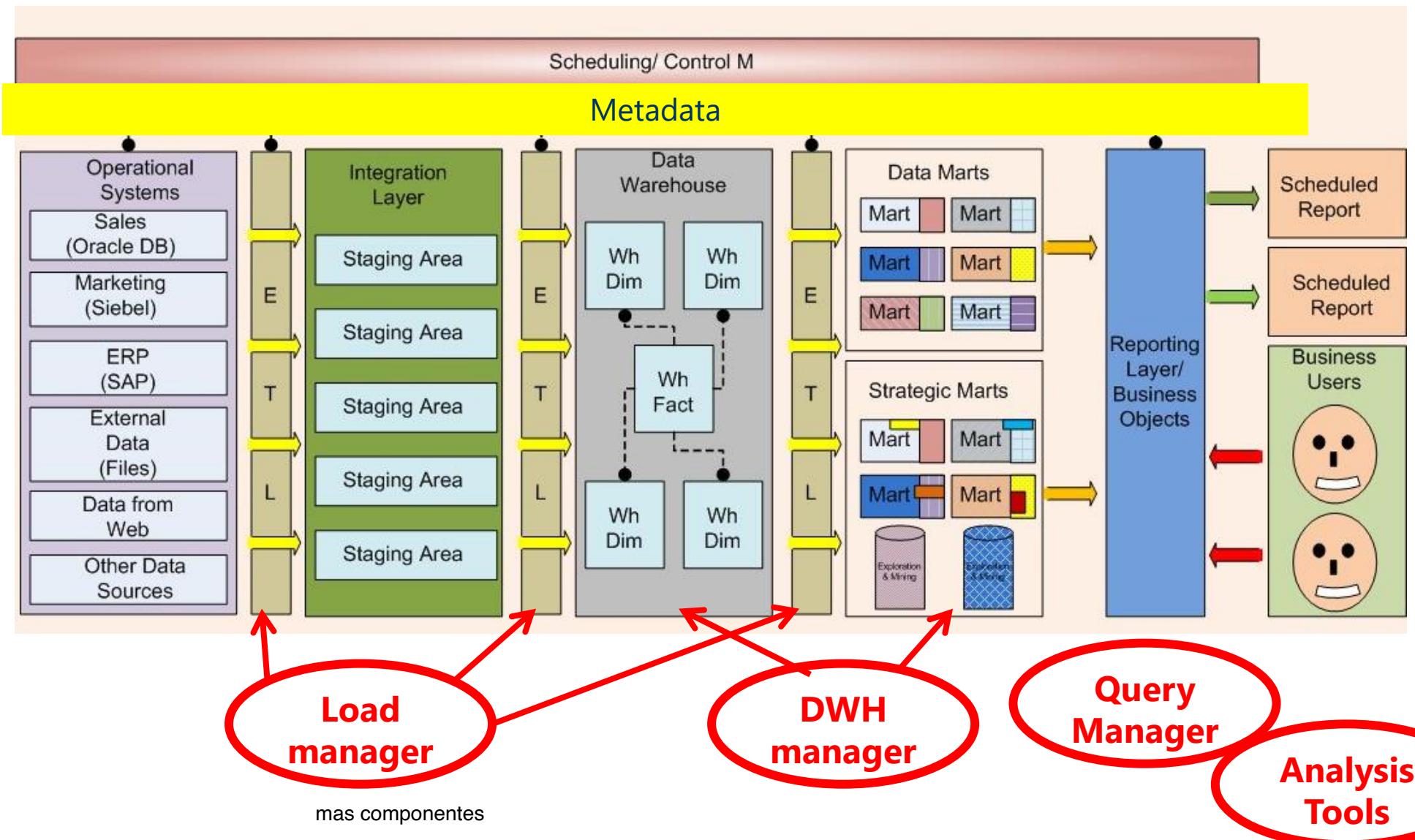
- You can still use the traditional system:
 - Maintaining **daily transactional work** in the original information systems (known as **OLTP**, On-Line Transactional Processing).
las operacionales no estan optimizadas para consultas analiticas
 - **Basic data analysis** is done in real time on the same database (known as **OLAP**, On-Line Analytical Processing).
para garantizar la integridad de los datos, hay que tomar la decisiones acerca de las transaccion
(lo que hablamos en BDGE)
- However, there are **limitations**:
 - Efficiency problems in daily work due to **complex queries** that are made when there is low load.
 - **No specific design** for analytical workloads.
cuellos de botella
 - **Performance bottlenecks** for in-depth analysis on operational data.



en un entorno de análisis los requisitos van cambiando !!

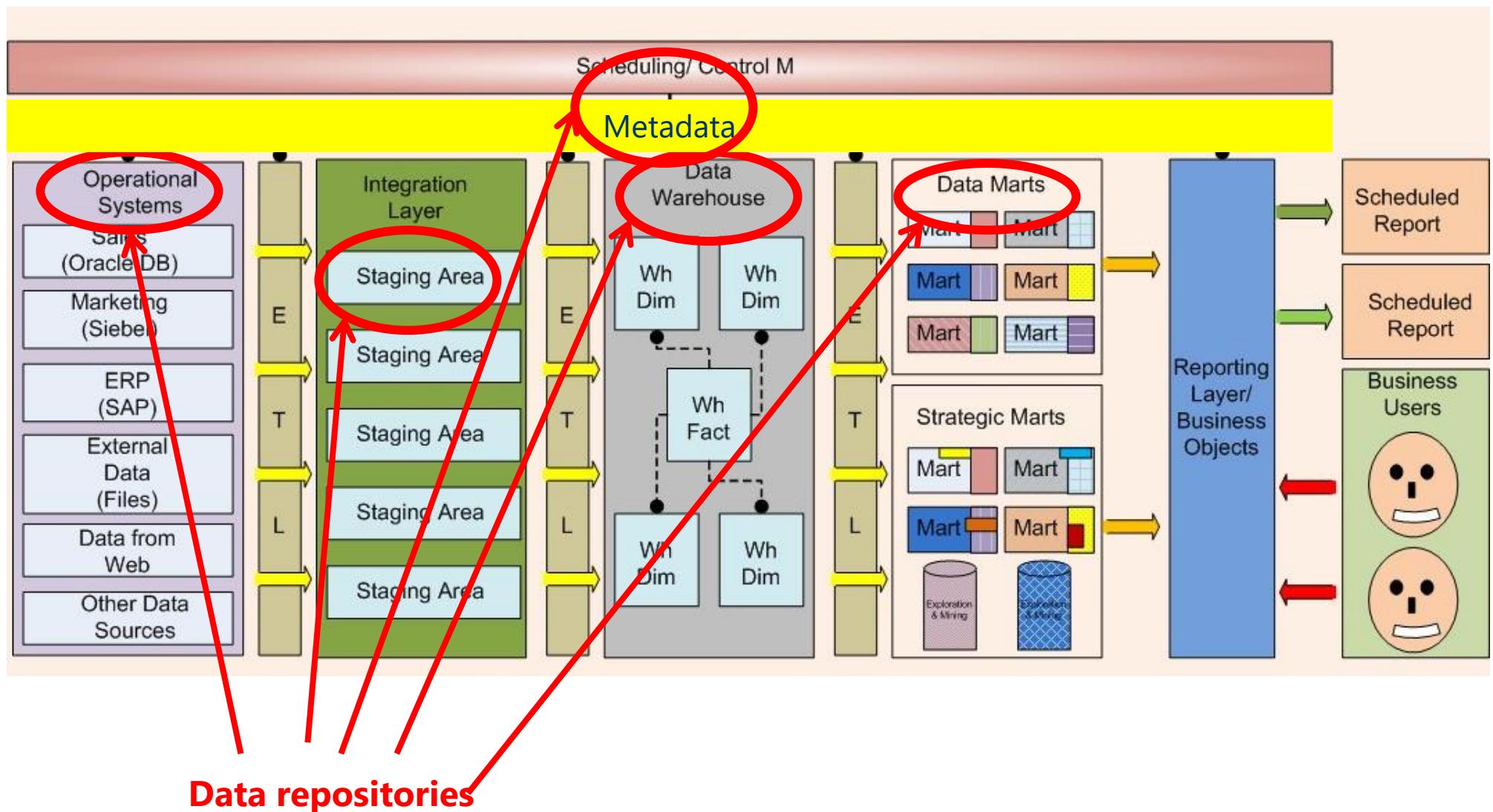


el almacen de datos alimenta los data marts



DWH Components

- **Load manager:** runs ETL tasks
 - **Extract:** Retrieves the data from various source systems.
 - **Transform:** applies business rules, clean data, format it for storage.
 - **Load:** inserts the transformed data into the DWH.
esquema de estrella siguen siendo valido en una base de datos columnar
- **DWH manager (server):** it allows to define and maintain the datawarehouse: data definition, aggregation, views, index, backup, etc..
- **Query manager:** Query execution, monitoring, ad-hoc forms, etc.
- **Access tools:** tools to design queries and reports, tools to develop end-user applications, OLAP tools, data mining tools, enterprise Information Systems (EIS)



- **Data sources:** files, web, xls, databases, ...
staging area procesa los datos, landing area es donde se guardan, no son lo mismo aunque se suelen mezclar
- **Staging area:** temporary storage where data is cleaned and transformed before going into the DWH.
 - E-R, Relational model. Not compulsory to have it implemented, but usually convenient.
- **Datawarehouse:** data collection for **decision making**
 - **Structure:** Multidimensional model
 - **Data mart:** departmental DWH un data mart para cada sección de la empresa
 - **Metadata:** describes business activities, the business objects, and rules that guide those activities.
 - Technical meta data needs to be mapped to the business meta data.
 - Includes documentation about data sources (origen, description, aggregation level, storage, ...)

- The core of a BI architecture is the **Data Warehouse**
- It provides a **consistent** and **unified** source of data for decision-making processes.
- Data Warehouse: A **central repository** of data designed to support the decision-making processes.
 - Information Oriented (not processes)
 - Integrated
 - Time-variant
 - Nonvolatile

unica fuente de verdad de la organizacion !!

Features of a DW

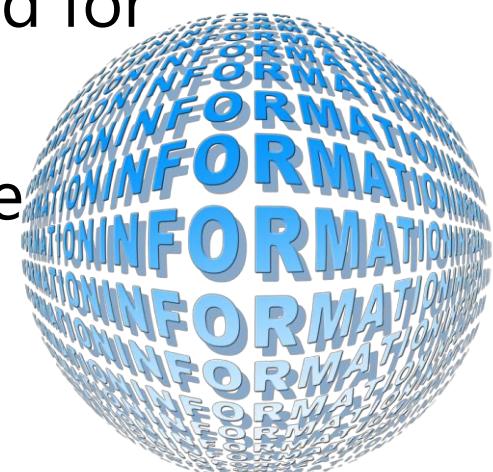
- **Information oriented** (not processes):

DWH is designed to **efficiently view information** on the basic activities (sales, purchasing, production, ...) of the organization.

It is not meant to support **operational processes** like **order management or billing**.

- Necessary information is extracted **from transactional systems** and efficiently stored for analysis.
- It leaves out irrelevant information. Only the **necessary data** is extracted and stored for analysis.

lagos de datos:
almacenes de datos
semi, no, o incluso
estructurados donde
almacenamos todo lo
que creemos que
puede tener valor para
la organización. Se
linkea el
datawarehouse con
ese lago

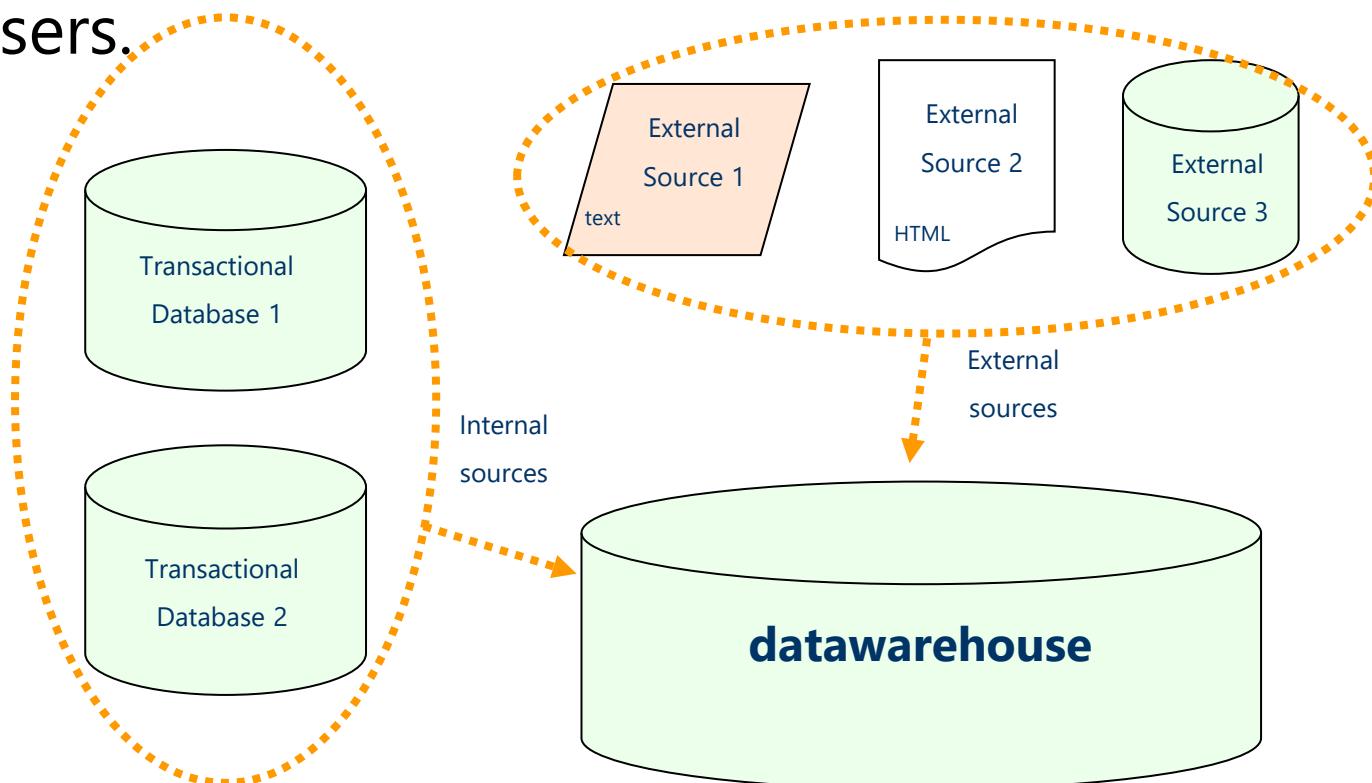


Features of a DW

- **Integrated:**

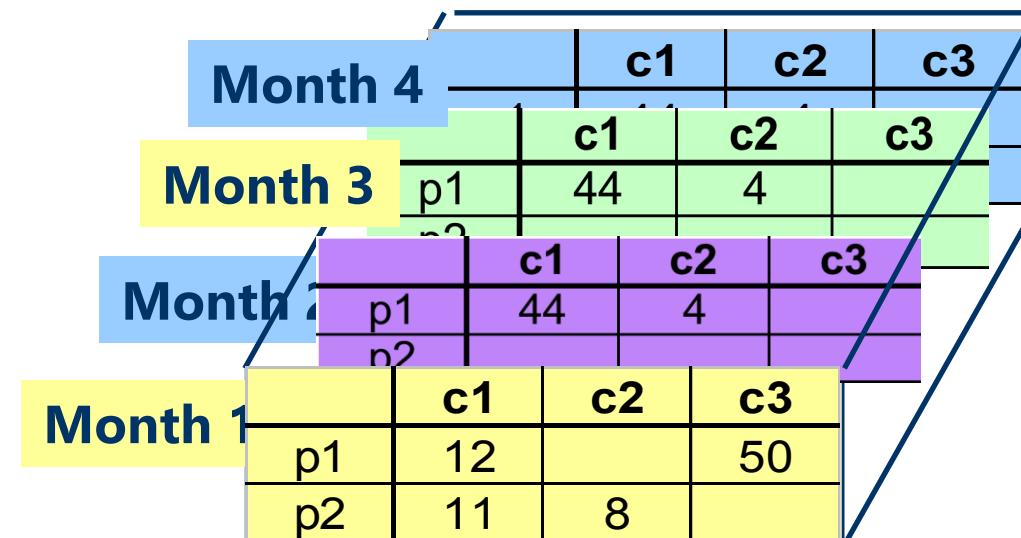
la integracion es muy dificil

- it collects data not only from the **internal transactional databases**, but it can also include **external sources**.
- It must be **consistent** to present a **unified view** of data to users.



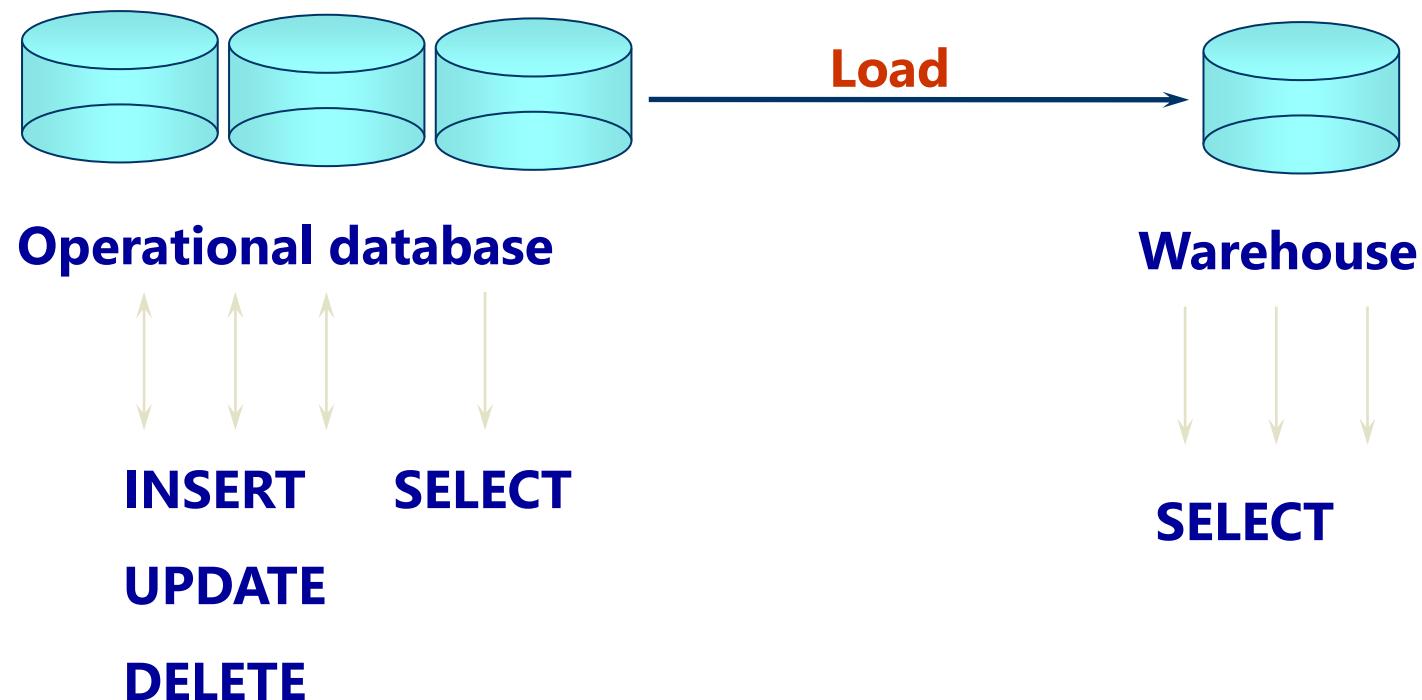
Features of a DW

- **Time-variant:**
 - Data are relative to **a period of time**.
 - It is **periodically updated** to include new information.
 - Detection of historical **changes, trends, patterns**, ...



Features of a DW

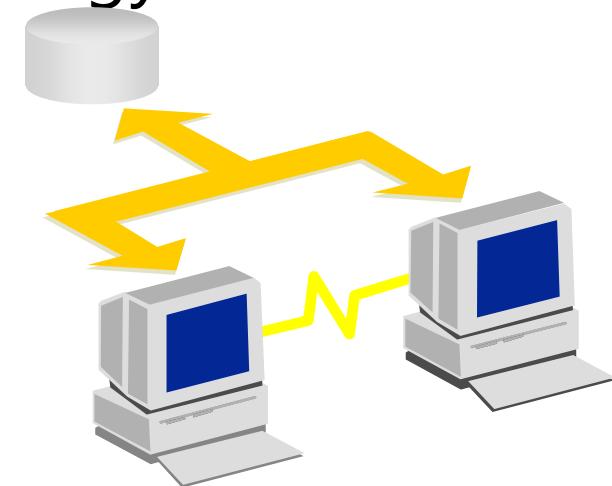
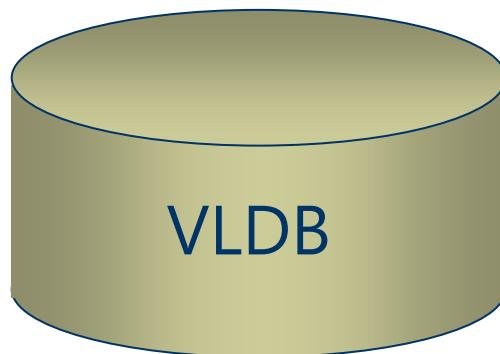
- **Non-volatile:**
 - the stored data are **not** (typically) **updated or deleted, only increased.**
 - Data in the DWH remains **stable** over time.



- Advances in technology have encouraged the development of data warehouse technology

- **Big data technology**

- Parallelism
- Hardware
- Distributed operating systems
- Database
- Query languages
 - VLDB
 - Big memory
 - Indexing techniques
 - Open systems (interoperability)
 - Specialized HW and SW for DWH
 - Tools for data analysis



resumen de diferencias

OLTP System	Data Warehouse
Stores current data	Stores historical data (including current data for trend analysis)
Stores detailed data	Stores summarized or aggregated data
Data are dynamic (updatable)	Data are static (non-volatile)
Repetitive processes	Ad-hoc, unforeseeable processes
Predictable usage pattern	Unpredictable usage pattern
High rate of transactions	Medium or low rate of transactions
Low response time (seconds)	Variable (usually longer) response time
Transaction-oriented	Data analysis-oriented
Application or process-oriented	Information-oriented
Supports daily operational decisions	Supports strategic decisions
High number of users (administrative)	Few users (managers, analysts)
Medium-sized databases	Large-sized databases

Advantages of using DW

- The **advantages** of using DW are, among others:

- **High ROI**

- **Competitive advantages:**

- Information non previously available, unknown or difficult to extract and incorporate.

- **Increased productivity:** decisiones más rápidas y más informadas

- More integrated and easy-to-access information.

- Make faster and more informed-decisions.

- **Better data quality and consistency**



Problems

Resources

Understatement of resources to load
High demand for resources
High cost of ownership

Data Integration

Complexity of integration
Hidden problems of source systems

Data Capture

Required data are not captured
Homogenization data

Usage

Increased demand from end users
Data ownership
Long-term projects

Multidimensional model

una dimension no deberia crecer al mismo tiempo que la de hechos

- **Multidimensional model:**

- A **fact** is the activity or event being analyzed.
- A **fact table** stores the quantitative data or measures, which are usually **aggregated** for analysis.
- **Dimensions** are descriptive attributes that provide context for the fact (e.g. time, location, customer, ...).
- The **dimension tables** store the **dimension attributes**. They are connected to the fact tables via **foreign keys**.

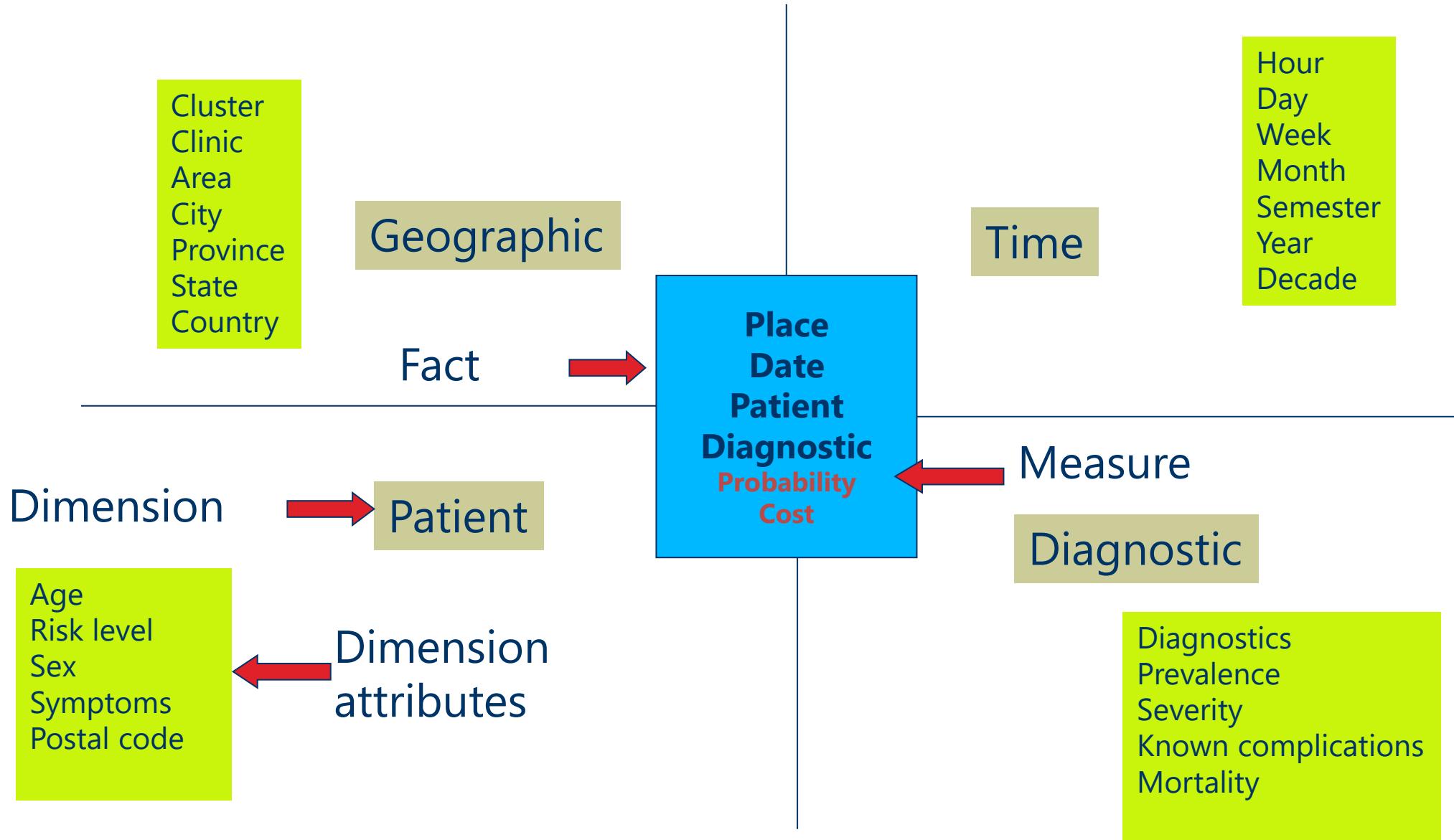


Multidimensional model

- **Activity analized:** Diagnostics.
- Information recorded about diagnostic:
“Avian influenza diagnostics has been realized in the **clinic “Morales”** on **Oct, 11th 2012** with a probability of **85%** to the **patient “Joseph”**. las dimensiones las vemos en la siguiente tabla
- The **geographic** and **temporal context** are important, not only the concrete diagnostic to a person.
- **Time** is hierarchical, sequential and many types of aggregations and calculations are made using it.



Multidimensional model



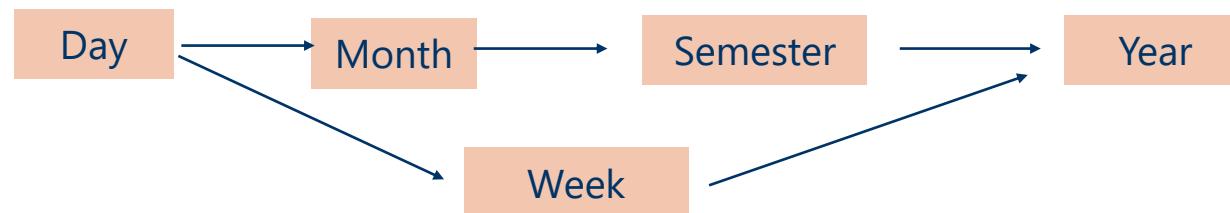
Multidimensional model

- **Hierarchy:** organization of dimension attributes in levels where data can be analyzed at different levels of detail.

Geography



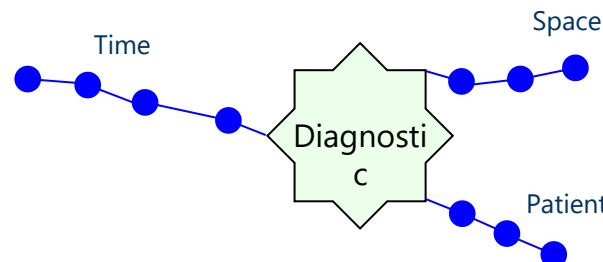
Time



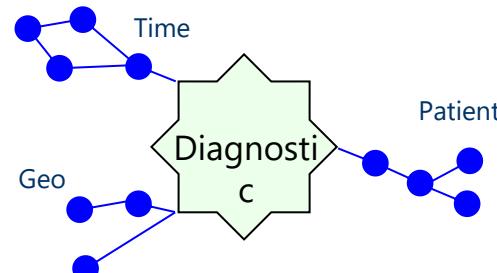
Multidimensional model

- **Basic structures**

- **Star schema:** Direct relationship between fact table and dimension tables. Dimensions are **denormalized**.



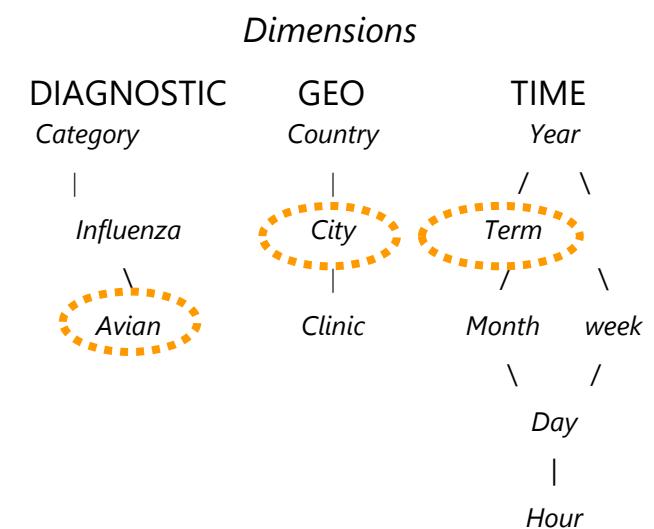
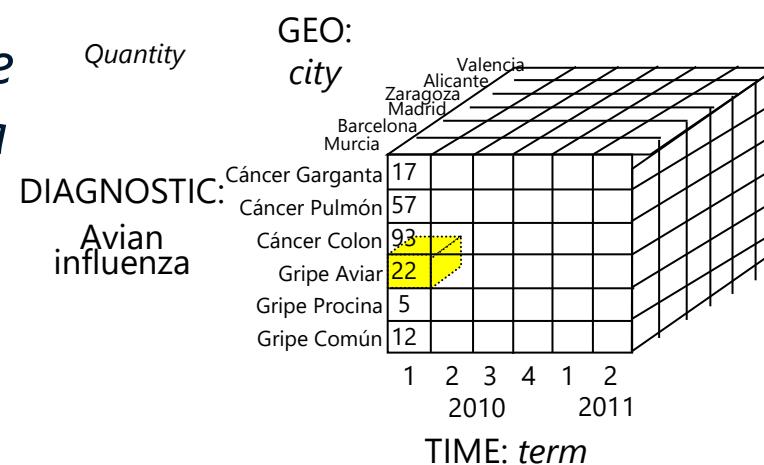
- **Snowflake schema:** Dimensions are **normalized**. Reduced redundancy but more joins during queries.



Multidimensional model

- **Facts** can be queried at **different aggregation levels**:
 - Query **measures** about the **facts** defined by **dimensions'** attributes and constrained by values of those **attributes**

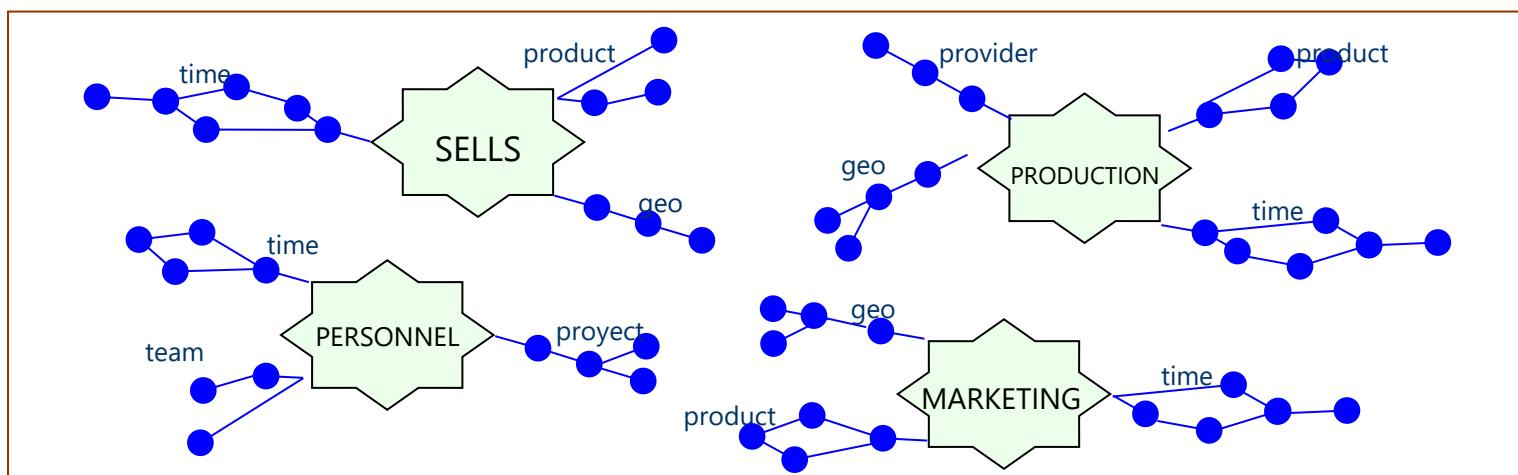
FACT: "The first term of 2010 22 cases of avian influenza were diagnosed in Murcia



- An aggregation level for a set of dimensions is called cube.

Multidimensional model

- Not all the information can be stored in a single schema
 - Some **departmental DWH** are needed
 - Each of these is called **datamart**. implementarlo si el DW esta sobrecargado
 - A data mart is a **subset** of the data warehouse that serves the **specific** needs of a department or business unit. They can be created from scratch, **independently** from the DWH.

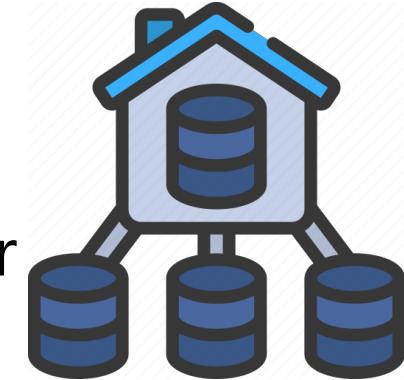


Design strategies

- **Datamart:**

- defined to meet the needs of a department or subdivision of the organization.
- contains less detail and more aggregated information.

como el DW es la unica fuente de verdad, tiene mas datos que un datamart, por lo que necesita tambien la maxima granularidad



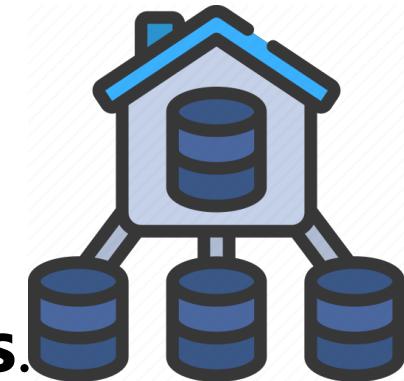
- **Top-down (Inmon): Data-centric**

- First define the data warehouse of the whole organization and then define data marts on it.

- **Bottom-up (Kimball): Business-centric**

- predefine departmental data marts and then integrate them into a data warehouse for the organization

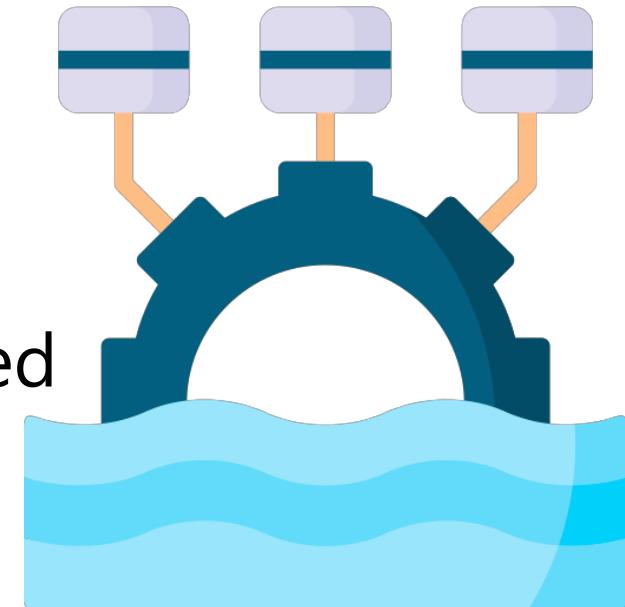
- **Datamart:** subset of the DWH.
- The **DWH** can be formed by **several datamarts** and, optionally, **additional tables**.
- Are defined to meet the needs of a department or subdivision of the organization.
- Contains **less detailed information** and **more aggregated** information.
- They are **easier to understand** and use.
- They may be the **intermediate step** between the data warehouse and transactional system



- **Why datamarts:**

- Provide users with access to the **most commonly** analyzed data.
- Smaller and focused data, with **better response time**.
- **Offloads** the primary DWH
- The data will already be **adapted** for OLAP queries or DM.
- **More control** over their data mart by department users.
- Being **simpler**, ETL processes are too.
- **Low cost** and **faster implementation**.
- Greater involvement of users in the overall data warehouse. **Better data quality, security, relevance**,...

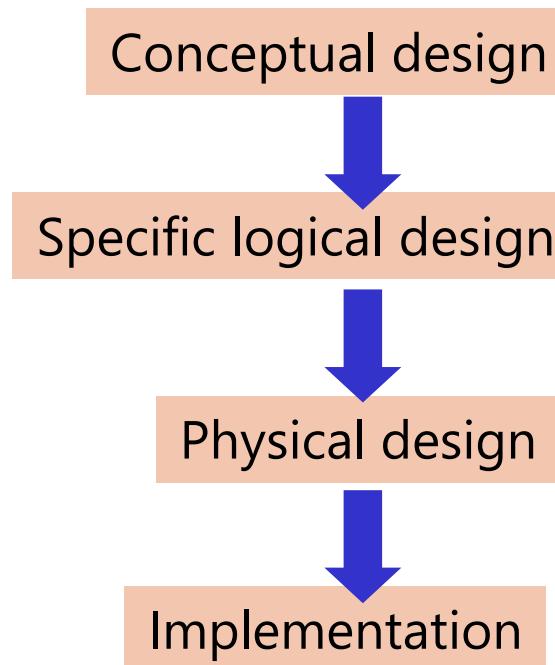
- In DWH data are transformed and structured, and we keep only the data needed for the analysis.
- Are we **losing** data that may be needed in the future?
- **Data lakes** store **ALL** source data in original format (may be not transformed)
- Easy to transform later for new DWH or for data analysis
 - But **schema-on-read**. Is it good?
 - Created for, not for business users**data scientist and analysts**

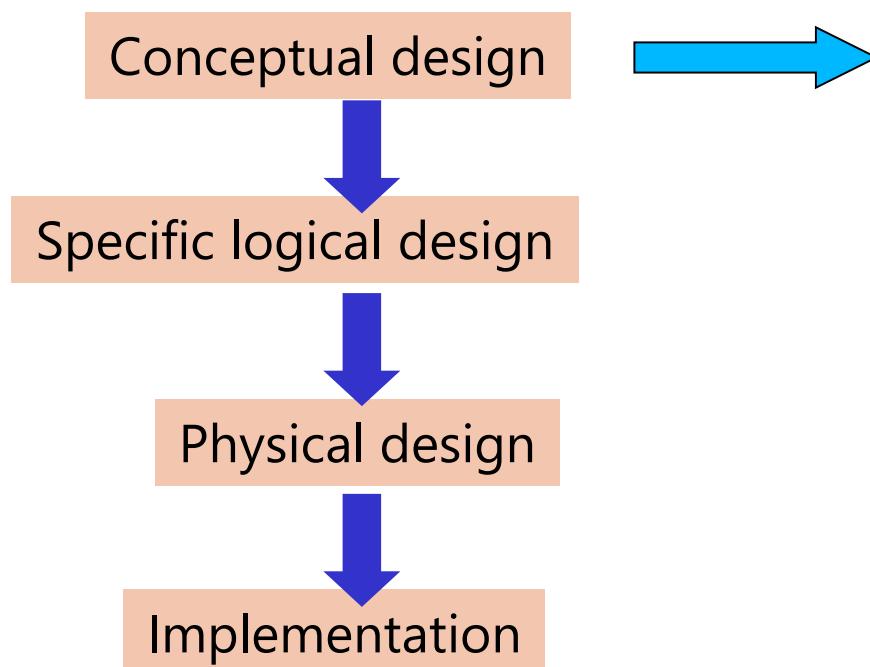


Business intelligence

Unit 2 – Datawarehouse and OLAP
S2-2 – Datawarehouse design

Does it ring a bell?





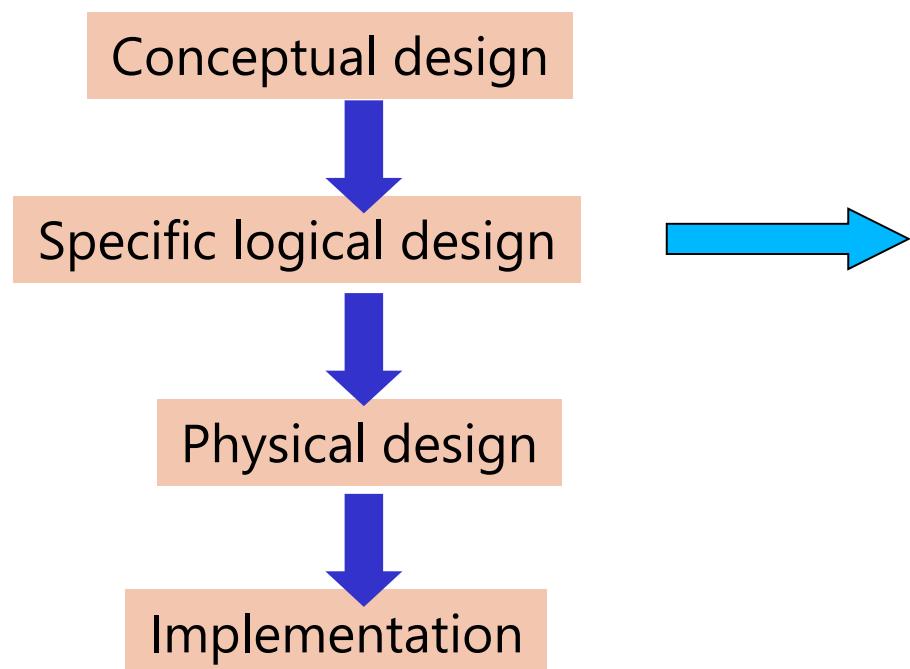
Requirement analysis

- Understand **business requirements**.
- Identify **business processes, KPIs**.
- Identify **data sources**.
- Identify **facts and measures**.

Conceptualization

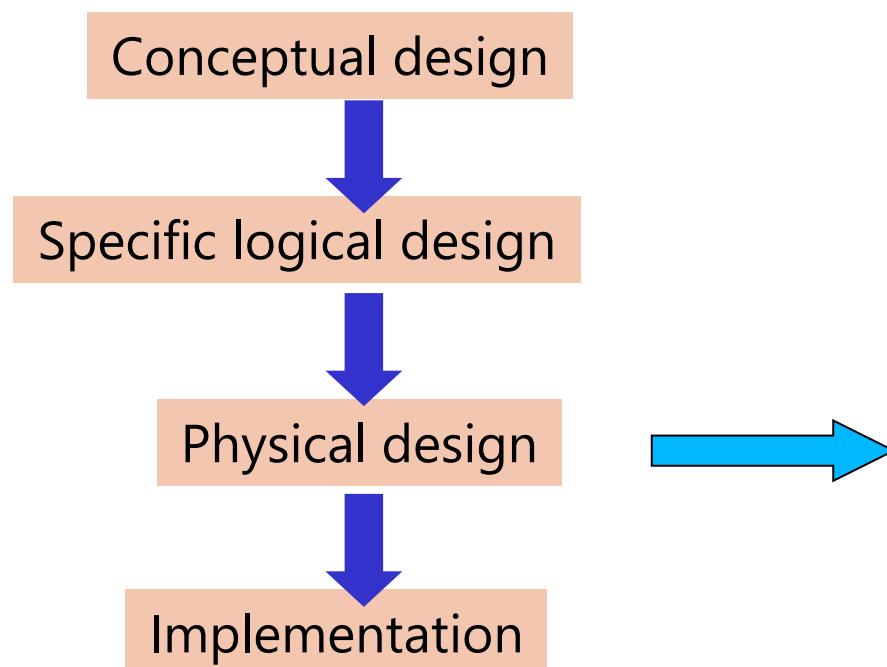
- Eg: Identify multidimensional model

Datawarehouse design



Multidimensional modeling

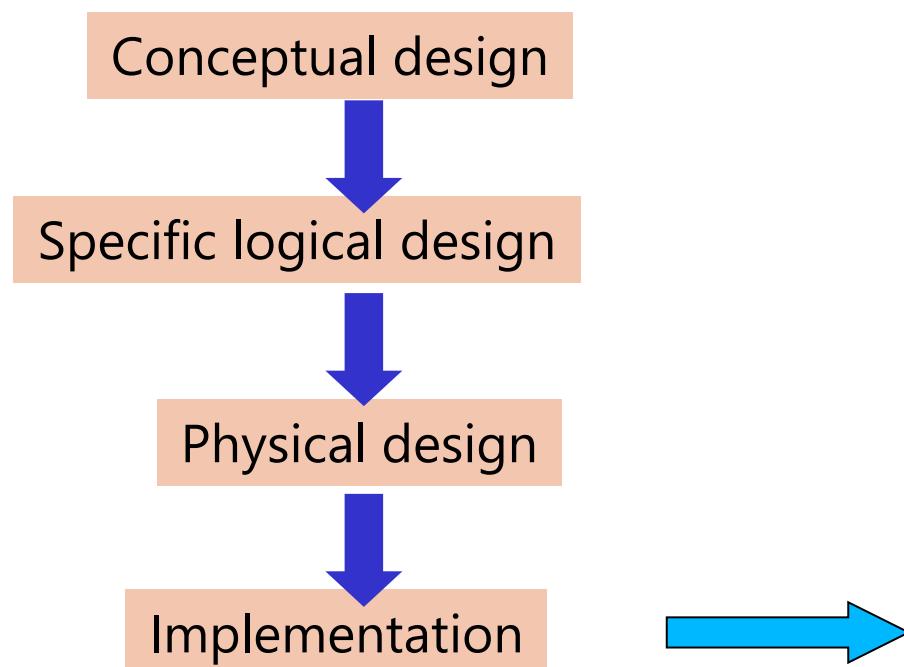
- Star and snowflake schemas
- Kimball's methodology



Storage management
(ROLAP, MOLAP, HOLAP)

Big data technologies

ETL Process design



**ETL and Data Integration
implementation
OLAP tools**

- **Multidimensional model:**
 - models an activity (**fact**) and **dimensions** that characterize the activity.
 - relevant information about the event (activity) is represented by **measures**.
 - Usually, the fact table uses a **composite key** created from the dimension's FK.
 - Descriptive information for each dimension is represented by **dimension attributes**.
 - Each dimension table uses a **simple key** to uniquely identify each record.

ER vs Multidimensional model

The Multidimensional Model and Entity Relationship have connections but are different.

- **application**

- ER is used for transaction systems (OLTP).
- MM is used for data analysis (OLAP).

- **structure**

- ER identifies and eliminates redundancy relations.
- MM usually includes denormalization for more efficient queries.

- **use**

- ER queries can be complex, with multiple joins.
- MM queries are simple and efficient.

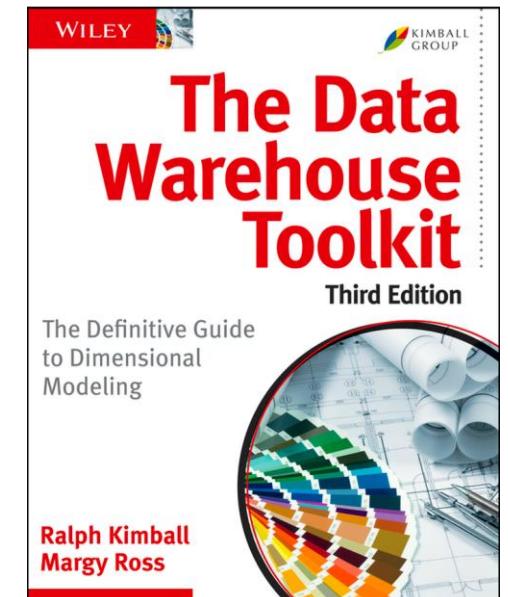
- **We start from:**
 - **Knowledge** about the domain (e.g. conceptual model)
 - **Data Sources**
 - **Indicators:** understand user queries
- **Objective:** resolve user queries efficiently.
 - **Methods** focused on **logical** and **physical design**
 - **[Kimball, 96]:** Methodology of **9 steps**



Methodology for multidimensional modeling

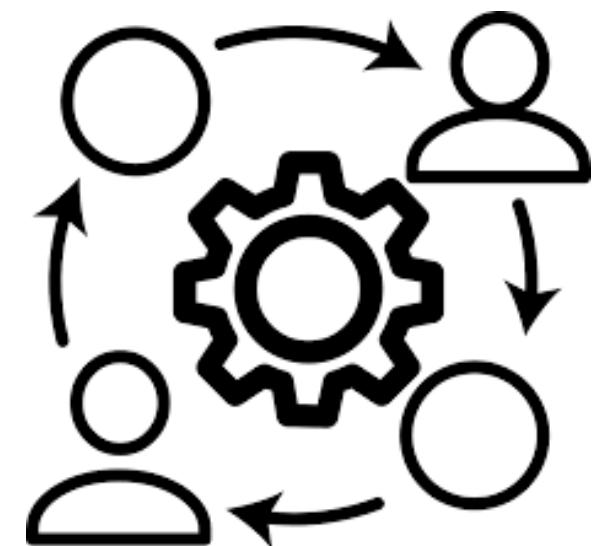
9-step methodology [Kimball, 96] :

1. Select the **process**.
2. Select the **granularity**.
3. Identification and conformation of the **dimensions**.
4. Selection of the **facts**.
5. Storing **derived** or **aggregated** values in fact tables.
6. **Complete** the dimension tables.
7. Select the **duration** of the database.
8. Manage **slowly changing dimensions**.
9. Select **priorities** and **query modes**.



Step 1: Select the process

- **Process:** business activity that the DWH will model.
- A process is supported by **OLTP** systems.
- Start with the **most important** to the organization.
- **Examples:** diagnoses, deceased, inventory, billing, ...



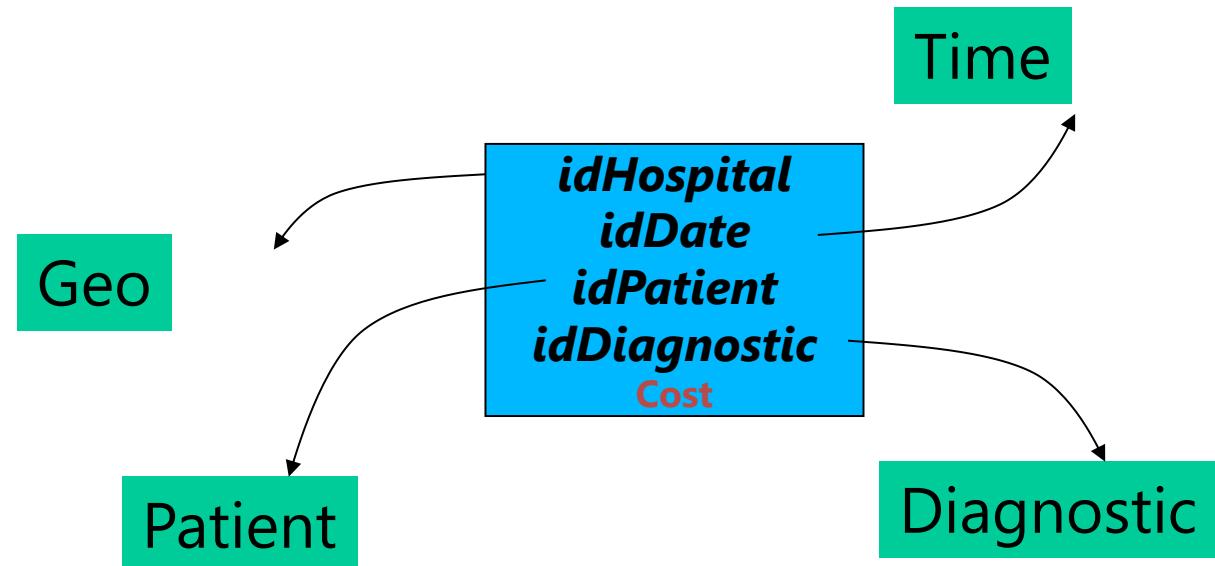
Methodology. Step 2

Step 2: Select the granularity

- **Granularity:** lowest level of detail in which information is stored in the fact table. Tell us how **detailed** our data is.
- It affects the **DWH size, analysis flexibility, query complexity**,...
- **Example:** weekly cost of diagnoses in health centers

¿Days? ¿Weeks?

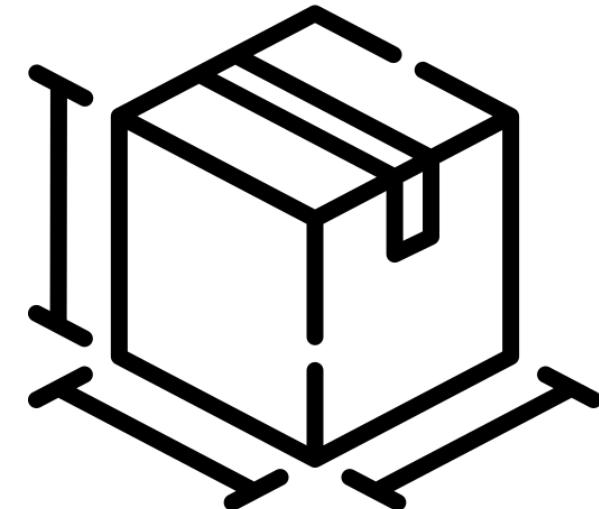
The level that allows the better analysis: Fine grain
¿Test performed or cost?



Methodology. Step 3

Step 3: Identification and conformation of the dimensions

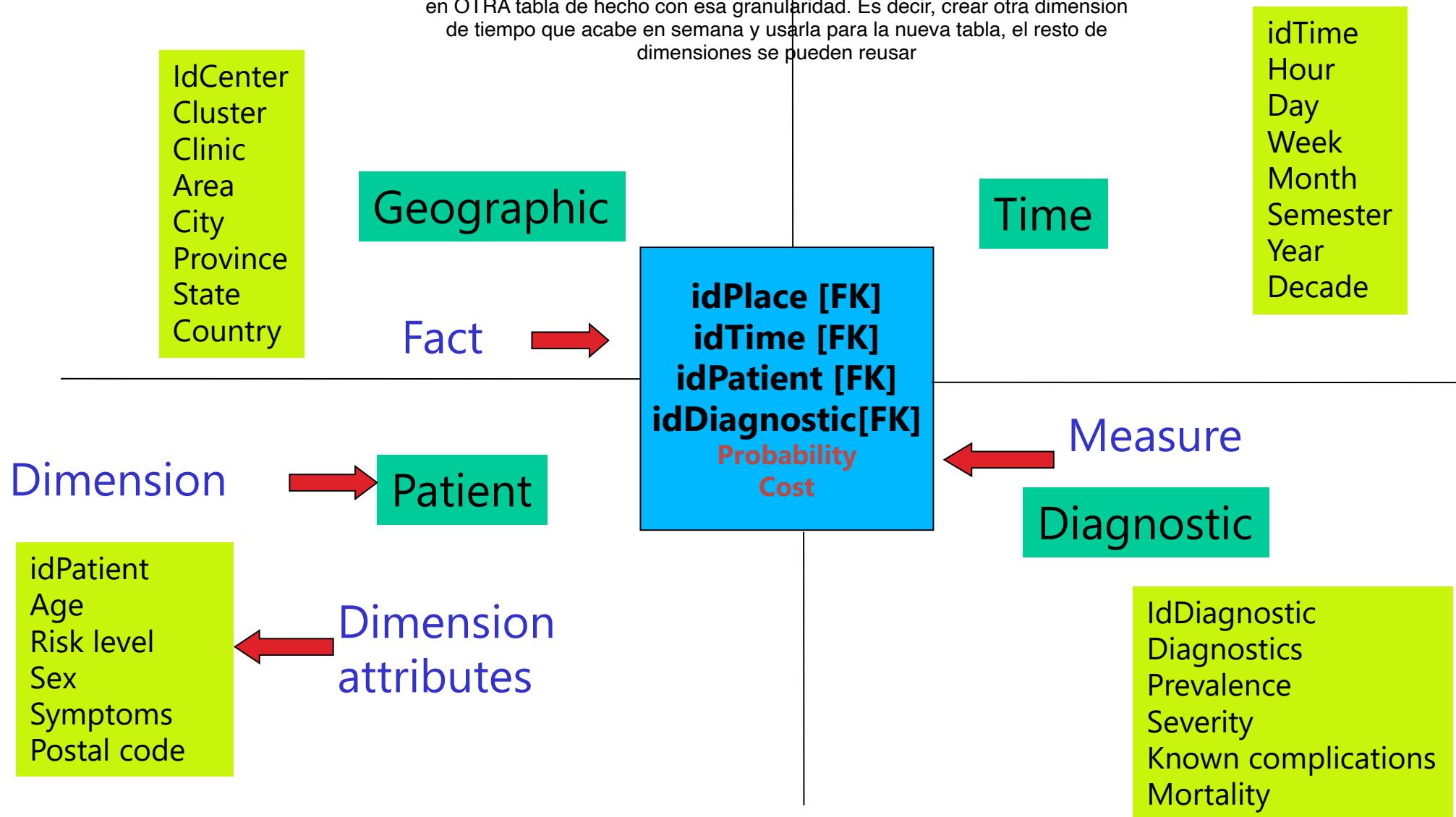
- After identifying the facts, the dimensions and their attributes **are defined**.
- **Dimensions:** characterization of facts at the selected level of detail.
- They are **descriptive** and are **query parameters**.
- **Hierarchies** between dimension attributes.



IdCenter
Cluster
Clinic
Area
City
Province
State
Country

Methodology. Step 3

is queremos granularidad semanal en la tabla de hechos, hay que guardarlos en OTRA tabla de hecho con esa granularidad. Es decir, crear otra dimensión de tiempo que acabe en semana y usarla para la nueva tabla, el resto de dimensiones se pueden reusar



Methodology. Step 3

- Dimensions define the **interest areas** for analyzing the facts:
 - Age
 - Age ranges
 - Sex
 - Risk level
 - Nacionality
- Special attention is paid to **time** and **space** (they apply to almost every type of business analysis)

idPatient
Age
Risk level
Sex
Symptoms
Postal code

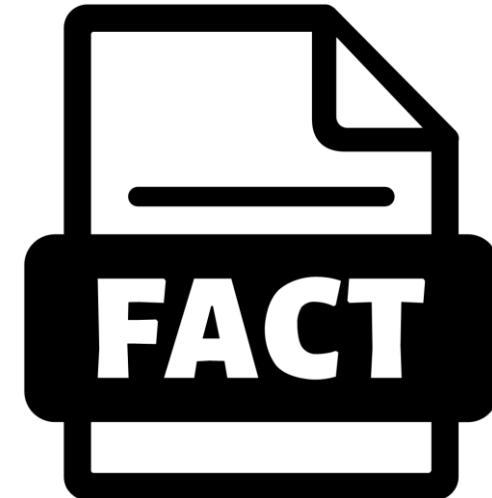
Methodology. Step 3

- Common attributes in the **time** dimension:
 - Day number, month number, year number, number of weeks.
 - day of the month (1 .. 31): allows comparisons on the same day across different months (e.g. sales on the 1st).
 - weekday (Monday ... Sunday).
 - month-end or Weekend indicators: comparisons of the last day of the month or weekend in different months.
 - quarter (1 .. 4): analysis of a specific quarter in different years.
 - holiday indicator: allows analysis on days adjacent to a holiday.
 - season (spring, summer, autumn, winter).
 - special event (football, elections, earthquake, ...)



Step 4: Select the facts.

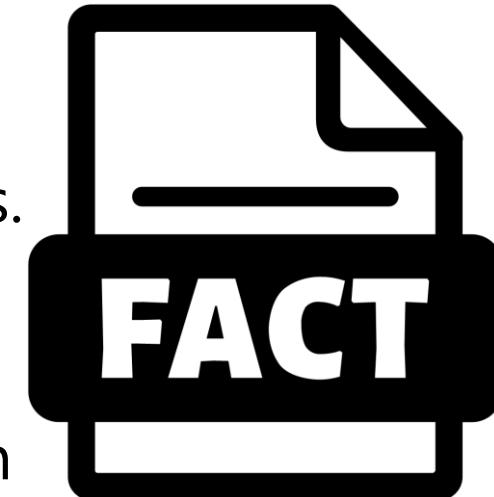
- **Fact:** analyzed information stored in the fact table.
- Useful facts are usually **numerical, additive**, or in general **facts that can be aggregated**.
- This is because large sets of the fact table are normally queried.
- **Select events** based on the **granularity of the information** chosen:
 - Cost of the treatment no se meten medidas cualitativas en una tabla de hechos
 - Cost of the tests
 - Deceased status
 - Total daily detected



Methodology. Step 4

- **Facts**

- **Additives:** can be aggregated in all dimensions.
 - **Activity data** are typically additives.
 - **Examples:** sales, units, money.
- **Semi-Additive:** they can be aggregated only in some dimensions.
 - **Intensity data** are not usually additives.
 - **Examples:** Stock, total existing stocks (can be added across some dimensions but not in time)
- **Non-Additives:** they can not be aggregated across any dimension.
 - **Examples:** temperature, unit price, percentage,
...
 - Can be aggregated by average values
 - **Dummy variables:** (0 or 1) to indicate occurrence.



Methodology. Step 4

- **Fact Tables:**
 - **Transactional:** represent detailed events in space time.
 - Provides the maximum level of exploration.
 - **Factless:** contain no measures, only the occurrence of certain events. tablas de hechos sin hechos - marcan la ocurrencia de un evento
 - Use to establish **relations between dimensions.**
 - **Examples:** students to class attendance, products that are not sold (negative analysis)
 - **Snapshot:** Each row is an instant of time.
 - Describe the state of the facts in a particular moment in time.
 - Normally includes semi-additive and non-additive facts.
 - They are often taken at predefined intervals and can be cumulative.

Step 5: Storing derived or aggregated values in fact tables.

- Include **derived attributes** that may be useful (e.g., non-additive attributes).
- **Examples:** differences, decomposed values (eg numerator and denominator) or calculated (amount= the price * units)

Step 6: Complete the dimension tables

- Add textual descriptions to the dimensions.
- Intuitive and understandable for users.

Step 7: Select of the duration of the database

- Set date from which to store data.
en muchos casos hay datos históricos
- It depends on the problem: data validity, business requirements, data availability, storage constraints, ...

por ejemplo guardar las facturas durante X años, etc

Step 8: Manage slowly changing dimensions (SCD)

- When an attribute changes value but **not the key**.
- **Example:** change in marital status, job category,
- Some **solutions** to slowly changing dimensions:
 - **Type 1:** overwrite a changed dimension attribute.
 - **Type 2:** create a new dimension record.
 - Current value (active, valid, ...) + validity date
 - **Note:** Business key are repeated. *Handle queries with care.*
 - **Type 3:** establish an alternate attribute, so that both the old and the new are accessible.
 - **Type 4:** mini-dimension (also historical table)
 - Also **combinations:** types 4, 5 (4+1), 6 (1+2+3), 7

Step 9: Select priorities and query modes (physical design)

Step 8: Control slowly changing dimensions

TABLA DE HECHOS

MEDIDAS	PROYECTO	INVESTIGADOR
1	P1	I1
2	P2	I1
3	P3	I1
4	P4	I1

TABLA DE HECHOS

MEDIDAS	PROYECTO	INVESTIGADOR
1	P1	I1
2	P2	I2
3	P3	I3
4	P4	I3

TABLA DE HECHOS

MEDIDAS	PROYECTO	INVESTIGADOR
1	P1	I1
2	P2	I1
3	P3	I1
4	P4	I1

TABLA DE HECHOS

MEDIDAS	PROYECTO	INVESTIGADOR	ESTADO
1	P1	I1	E1
2	P2	I1	E2
3	P3	I1	E3
4	P4	I1	E3

DIMENSION INVESTIGADOR

PK	BusinessKey	Nombre	Categoría	Facultad
I1	Manolo	M. Campos	AYD	Informática

TIPO 1: REESCRIBIR CATEGORÍA

TIPO 2: NUEVO REGISTRO. MANTENER BK

DIMENSION INVESTIGADOR

PK	BusinessKey	Nombre	Categoría	Facultad	FECHA_INI	FECHA_FIN	VIGENTE
I1	Manolo	M. Campos	AYD	Informática	2005	2009	N
I2	Manolo	M. Campos	CD	Informática	2010	2016	N
I3	Manolo	M. Campos	TU	Informática	2017	-	Y

TIPO 3: NUEVO ATRIBUTO

DIMENSION INVESTIGADOR

PK	BusinessKey	Nombre	Categoría AC	Categoría	Facultad	FECHA_INI	FECHA_FIN
I1	Manolo	M. Campos	TU	CD	Informática	2005	2009

TIPO 4: MINIDIMENSION

DIMENSION INVESTIGADOR

PK	BusinessKey	Nombre	Facultad	FECHA	ESTADO
I1	Manolo	M. Campos	Informática	2005	E1

MINIDIMENSION ESTADO

PK	Categoría ACTUAL
E1	AYD
E2	CD
E3	TU

TIPO 5: MINIDIMENSION +OUTRIGGER

Methodology: Errors / guidelines (Kimbball)

Error 1: Not sharing dimensions between different fact tables.

- Correct: Unify master files. Example: Gender {'M', 'F'}. 

Error 2: Not unifying facts from different fact tables

- Correct: Unify facts, even if they come from different departments or systems. Example: retail and enterprise sale.

Error 3: Ignoring aggregate tables

- Correct: instead of adding hardware to solve performance issues, use pre-aggregated tables to improve query performance.

Error 4: Forget the highest level of detail in the data model.

- Correct: Maximum detail in 3 areas: staging, relational and dimensional. area de procesamiento de ETL

Methodology: Errors / guidelines (Kimbball)

Error 5: Mix facts of different granularity in the same fact table.

- Correct: Avoid mixing different granularities in a fact table. It is better to create separate tables for different aggregation levels.



Error 6: Create a dimensional model to solve a particular report instead supporting multiple analytical needs.

Error 7: Adding dimensions to a fact table before setting its granularity.

- Correct: The fact table only contains FK and measures. Do not store dimensional data directly in the fact table.

Error 8: Create "smart keys" to relate a dimension table to a fact table.

- Correct: Surrogate keys should be auto-increment (even for the time dimension)



Why?

- Heterogeneous data sources keep their own business key.
- Changes in source applications should not affect the DWH.
- Performance (storage size and comparison speed).

TABLA DE HECHOS

MEDIDAS	PROYECTO	INVESTIGADOR
1	P1	Manolo
2	P2	Manolo
3	P3	Manolo
4	P4	Manolo

DIMENSION INVESTIGADOR

PK	BusinessKey	Nombre	Categoría	Facultad
I1	Manolo	M. Campos	AYD	Informática



Error 9: Not to address slowly changing dimensions.



Error 10: Splitting hierarchy levels into multiple dimensions.

“error suave”, no hacer normalmente hay que hacerlo de forma justificada

Error 11: Shorten the descriptions in the dimension tables with the intention of reducing the space required.

- The dimensions are the interface that users have to browse.
- They take up little space in relation to the facts.

Error 12: Include text attributes in a fact table, if done with the intention of filtering or grouping.

tipos de dimensión

- **Date and time**

- Minidimension
- Several time zones

- **(Shrunken) Rollup dimension**

- **Junk dimension**

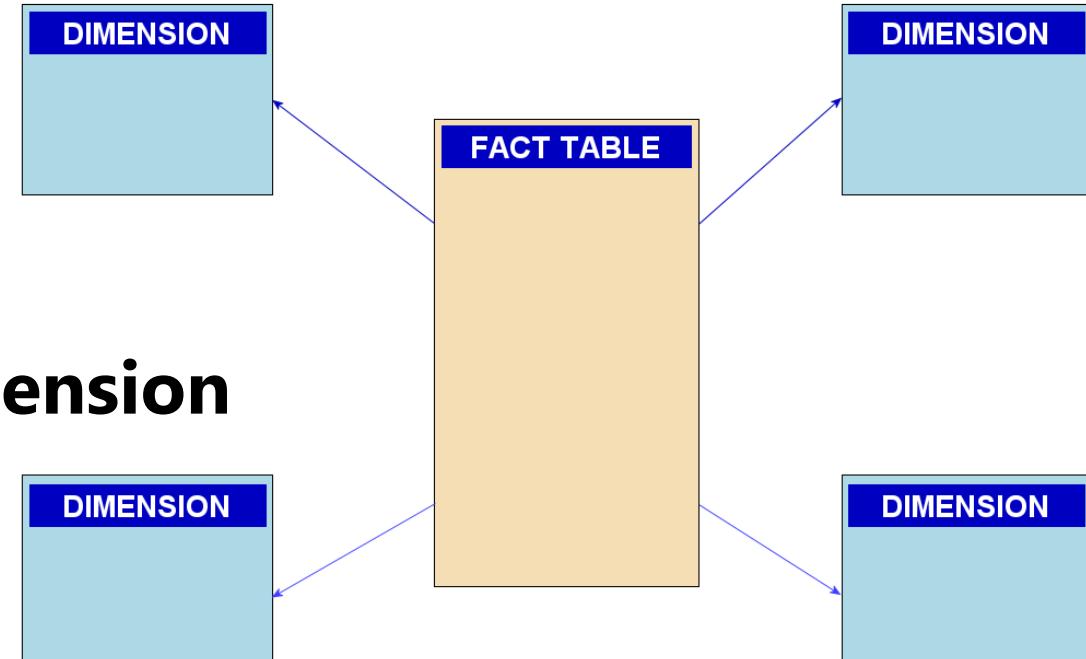
- **Degenerate dimension**

- **Bridge table**

- **Variable depth hierarchies**

- **Outrigger dimension**

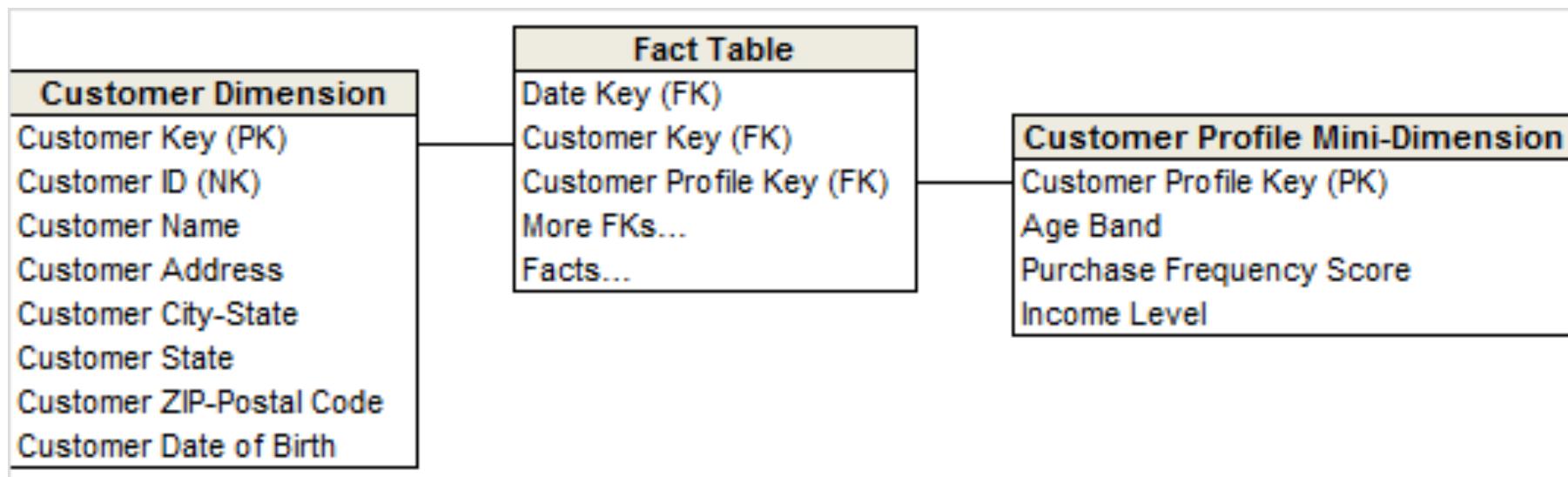
- **Snowflake dimension (normalization)**



DW Design techniques

Date and time

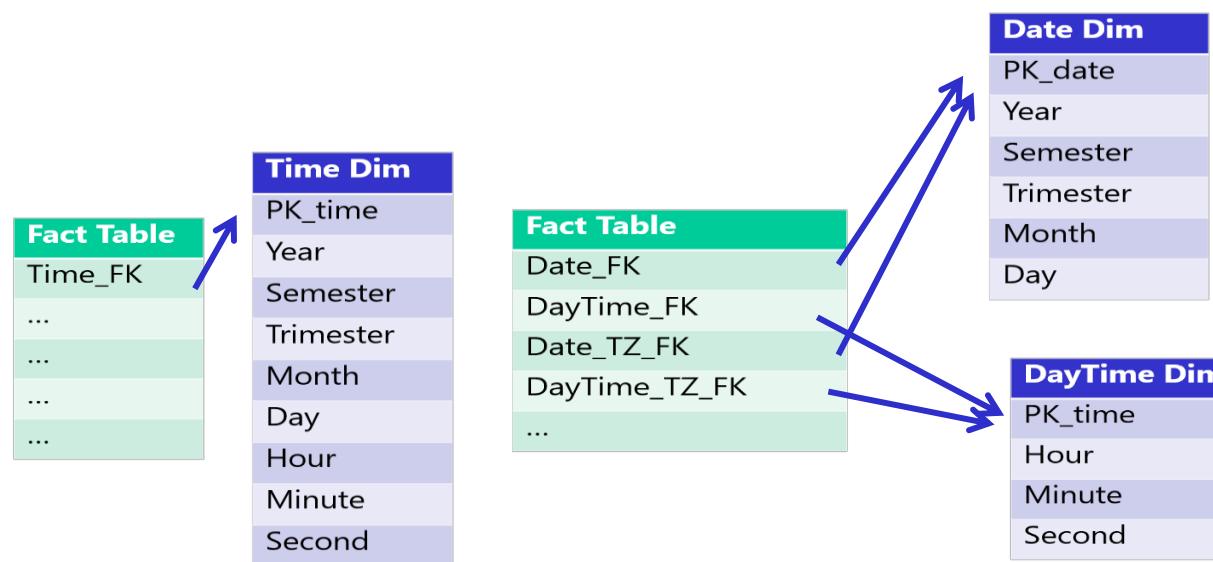
- Track time at different granularities (e.g. day, week, hour)
- Minidimension: smaller dimensions to store frequently changed attributes.



DW Design techniques

Date and time

- Track time at different granularities (e.g. day, week, hour)
- Minidimension: smaller dimensions to store frequently changed attributes.
- Several time zones: another FK to track other TZ.
- “Role playing”: several FK to same dim with different roles



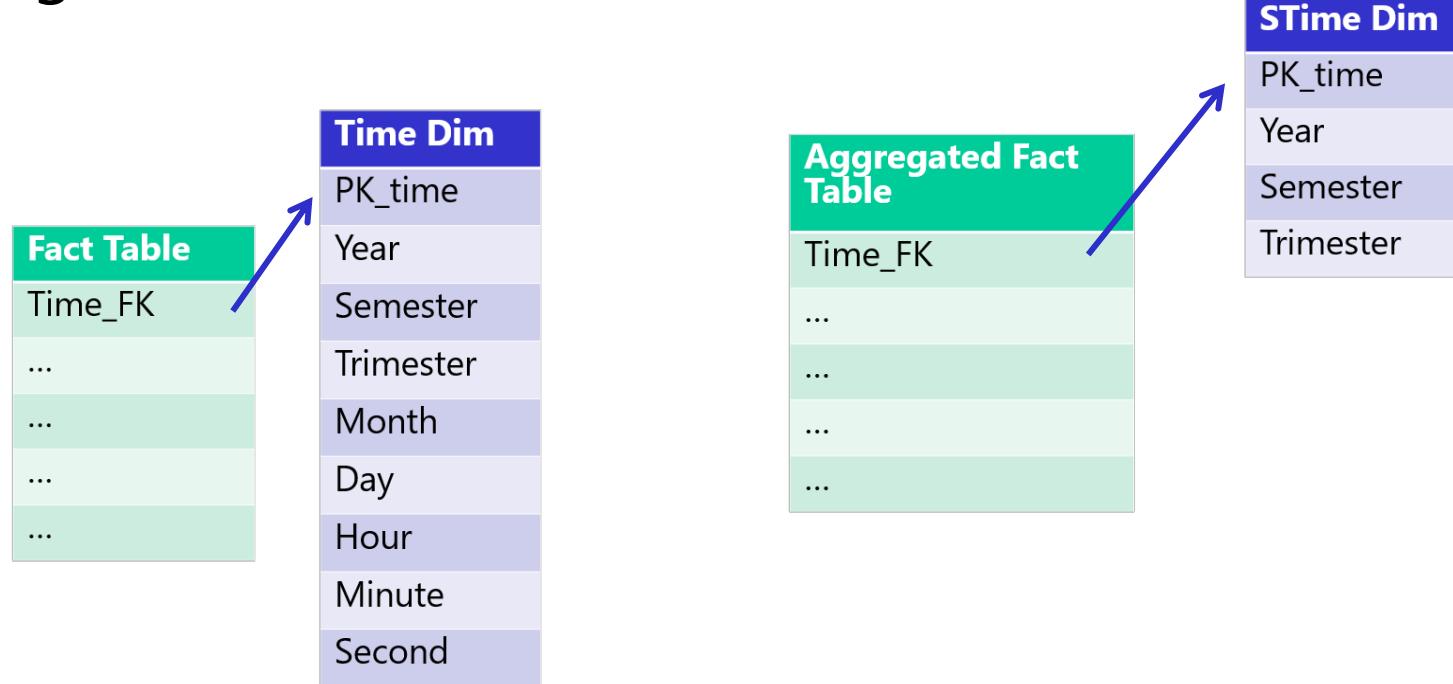
Not a
minidimension!
Dimension
splitting,
instead.

DW Design techniques

dimension agregada

(Shrunken) Rollup dimension

- For aggregated fact tables. A smaller version of an existing dimension.
- Smaller dimension -> better performance and storage for aggregated facts.

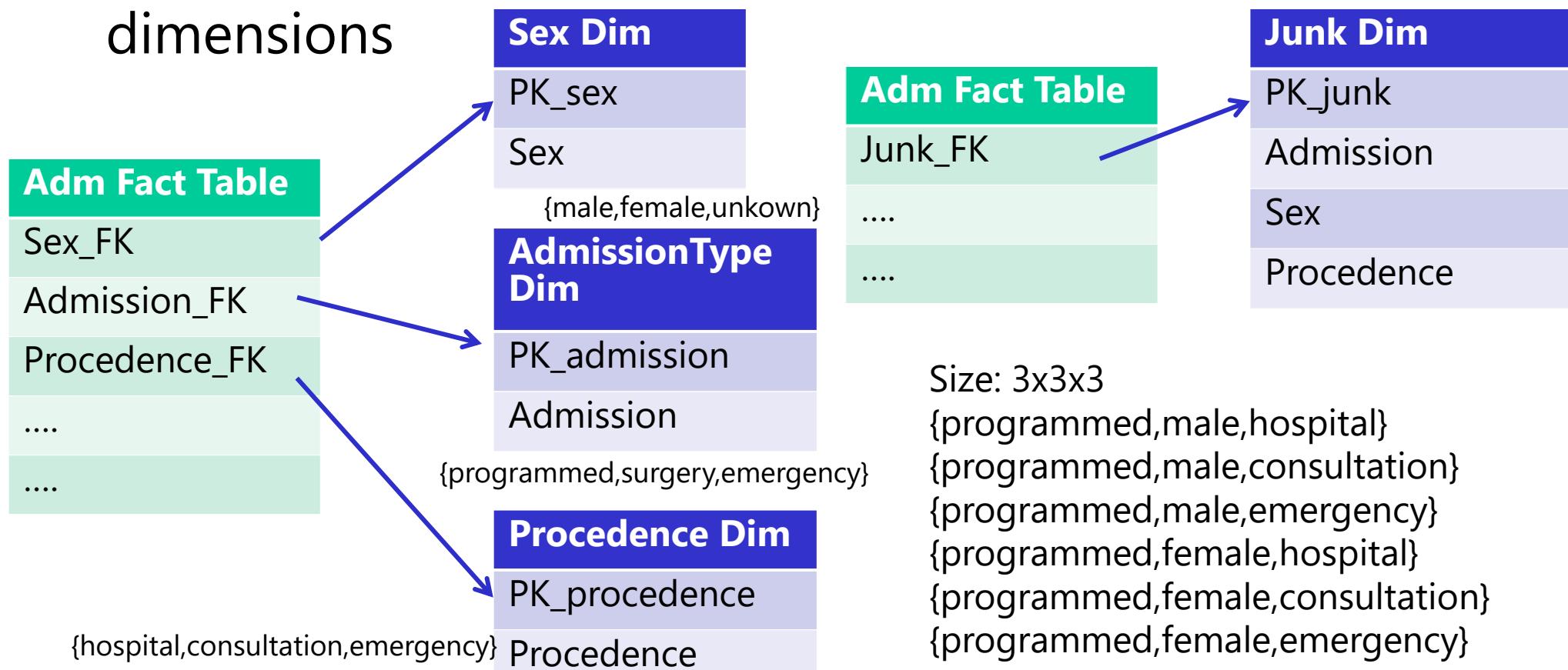


DW Design techniques

Junk dimension

- Fact tables with multiple dimensions with low cardinality
- Junk dimension holds cartesian product of small dimensions

condensa dimensiones pequeñas en otra



DW Design techniques

se usa para agrupar, para ver por ejemplo los productos asociados a una factura

Degenerate dimension

- Only contains the PK. Eg: order number, invoice number,... id, ...
- Store it in the fact table (number)
- Useful for grouping
- **Not for filtering**, as there are no descriptive attributes

Adm Fact Table
Sex_FK
Admission_FK
Procedence_FK
....
Episode_number:bigint
....

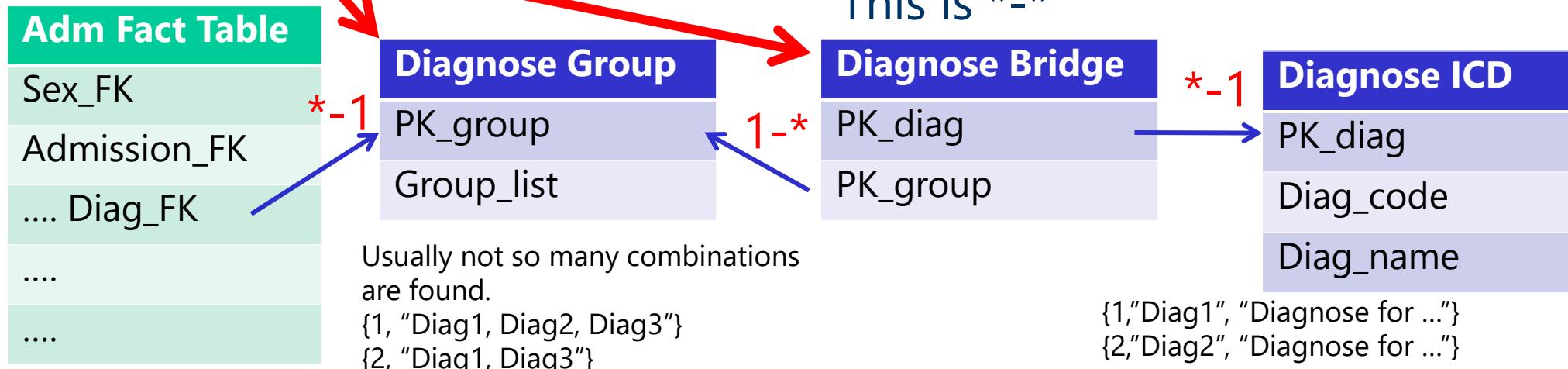
DW Design techniques

relacion entre tablas M-N

Problem M-N (*-*) associations between fact and dimensions. E.g.: admission has several diagnoses.

- Is this the right granularity? Solved in fact-tables!!!
 - Change granularity to diagnose? Another fact table?
- Standard solution: “**Bridge table**” with 2 additional tables
 - **Group table**: to keep association 1-* between fact and dimension. **Bridge table** between fact-table and dimension or dimension and values.

la mas complicada, se usa en BDs muy grandes



DW Design techniques

Problem M-N (*-*) associations between fact and dimensions. E.g.: admission has several diagnoses.

- Other alternatives? This is a simpler solution

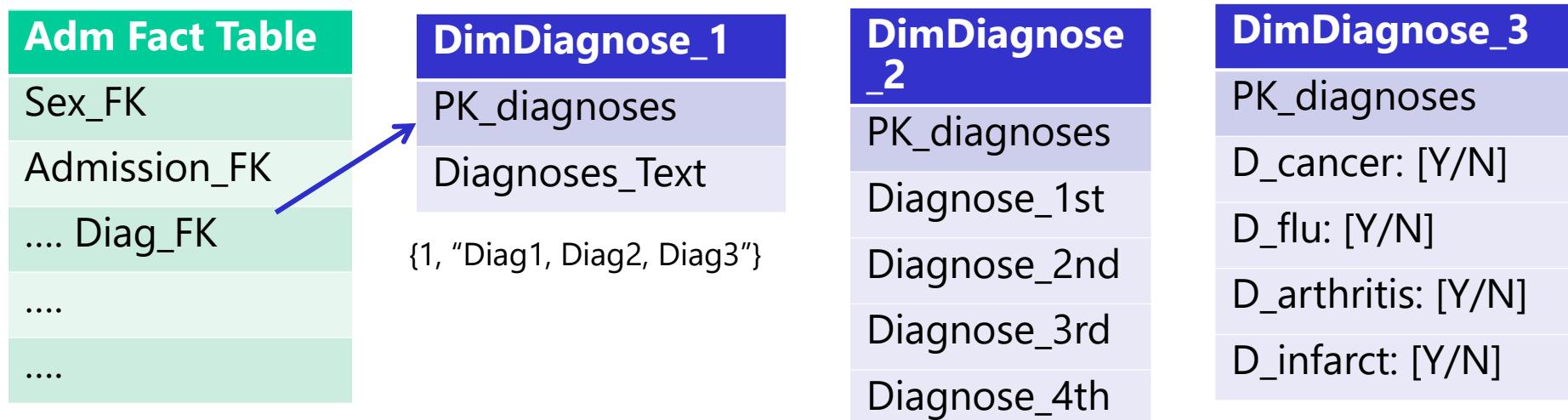


- Simple queries, fewer tables, good for small datasets. Difficult to group diagnoses (previous option is better), adds redundancy and doesn't scale well.

DW Design techniques

Alternatives to store and processing many-to-many relationships

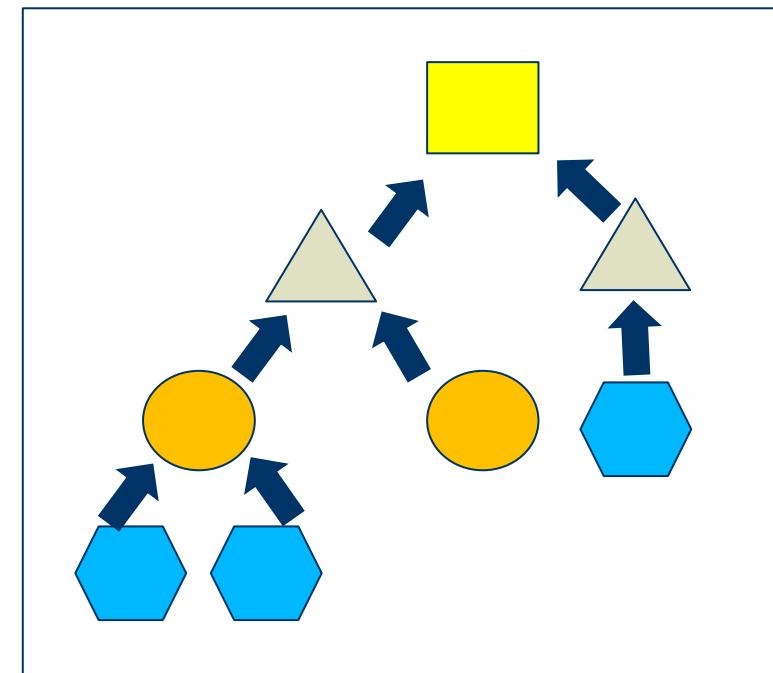
- String concatenation: "diag1 # diag 2" (Pathstring)
- Multiple attributes in the dimension. Eg. Diagnose1: diag1, diagnose2: diag2.
 - Are they sorted? Order is important? First value? Second value? Limited number of attributes? -> Multiple dummy attributes: Diagn



Variable depth hierarchies. The number of levels can change between records.

- Recursive queries in SQL and OLAP are limited.

Tratamiento	PK	ATC	descripcion	Padre
	1	J01AA01	Des J01AA01	J01AA
	2	J01AA02	Des J01AA02	J01AA
	3	J01AA	Des J01AA	J01A
	1	J01AB01	Des J01AB01	J01AB
	2	J01AB02	Des J01AB02	J01AB
	3	J01AB	Des J01AB	J01A
	3	J01A	Des J01A	J01
	3	J01B	Des J01B	J01
	4	J01	Desc J01	J
	5	J	Antibiotic	-



Solutions:

- Business decision: Not all levels apply: “Ceuta” and “Melilla” are not “Province”. Use business significative value
- Pathstring with complete path in hierarchy (same as with bridge tables but hierarchy flattened into a string).
- Slightly ragged: if range is small, force fix depth (same as with bridge tables)
- Bridge table with depth level (**closure table**)
 - Foreing key to dimension + depth attribute

ATC Treatment
PK_ATC
Description
Group_FK



Treatment Closure
PK_ATC
Group_ATC
Depth

ATC Treatment
PK_ATC
Description

DW Design techniques

Closure table (from previous slide)

- Bridge table: additional table.
- Foreign key to dimension + depth attribute
- Combined PK

**Fk_treat
m**

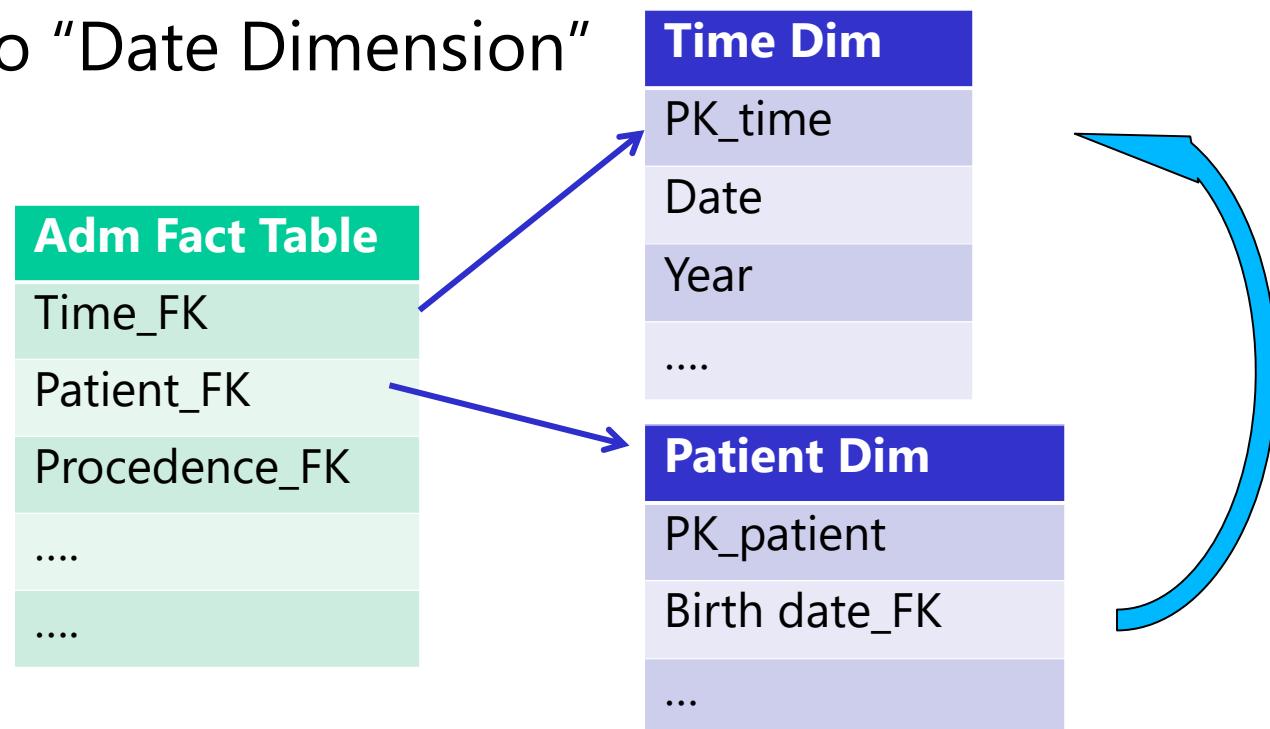
1-* →

Tratamiento		
PK	ATC	descripcion
1	J01AA01	Des J01AA01
2	J01AA02	Des J01AA02
4	J01AA	Des J01AA
5	J01AB01	Des J01AB01
6	J01AB02	Des J01AB02
7	J01AB	Des J01AB
8	J01A	Des J01A
9	J01B	Des J01B
10	J01	Desc J01
11	J	Antibiotic

Tratamiento Closure		
Hijo (PK, FK)	Padre (PK)	Profundidad
J01AA01	J01AA	4
J01AA01	J01A	3
J01AA01	J01	2
J01AA01	J	1
J01AA02	J01AA	4
J01AA02	J01A	3
J01AA02	J01	2
J01AA02	J	1
J01AB01	J01AB	4
J01AB01	J01A	3
J01AB01	J01	2
J01AB01	J	1
J01AB02	J01AB	4
J01AB02	J01A	3
J01AB02	J01	2
J01AB02	J	1

Outrigger dimension

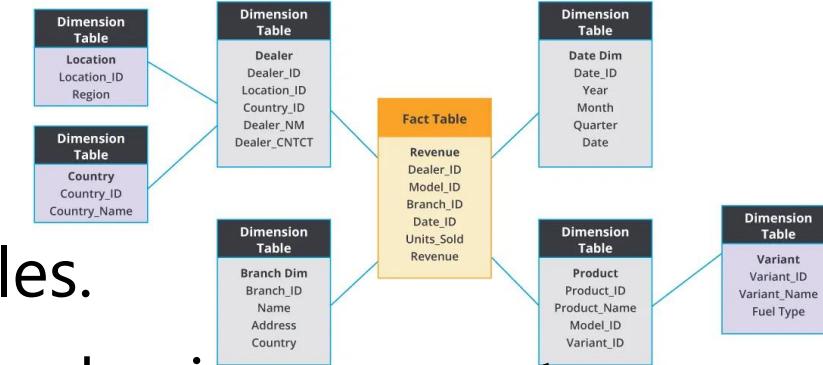
- Exception!!
- When a dimension has a FK to another dimension.
- Eg.: “Registered user” and “Unregistered user”
 - “Birth date” links to “Date Dimension”



Normalization

Snowflake dimensions

- A hierarchical relationship in a dimension table is normalized, low-cardinality attributes are stored in separate tables connected to the main dimension table through FK.
- Accurately models the hierarchy through normalization, yet a flattened (denormalized) dimension can store the same information in a single table.



- Typically used in big dimension tables.
- Snowflake schemas can be harder for business users to understand and navigate.
 - They can also negatively impact query performance, as they involve more table joins.

Normalization

DIMENSIÓN ESTRUCTURA

PK	BK	Nombre Área	Dept	Departamento	Dept ...	Facultad	Facultad Descripción	Fac ...
A1	LSI	Lenguajes y Sistemas	DIS	Informática y Sistemas	[varios]	FIUM	Facultad de Informática	[varios]
A2	ISA	Informática y Automática	DIS	Informática y Sistemas	[varios]	FIUM	Facultad de Informática	[varios]

DIMENSION ÁREA

PK	BK	Nombre Área	Dept
A1	LSI	Lenguajes y Sistemas	DIS
A2	ISA	Informática y Automática	DIS

FK

DIMENSION DEPTO

Dept	Departamento	Dept ...	Facultad
DIS	Informática y Sistemas	[varios]	FIUM

FK

DIMENSION FACULTAD

Facultad	Facultad Descripción	Fac ...
FIUM	Facultad de Informática	[varios]

No replicated, but join needed

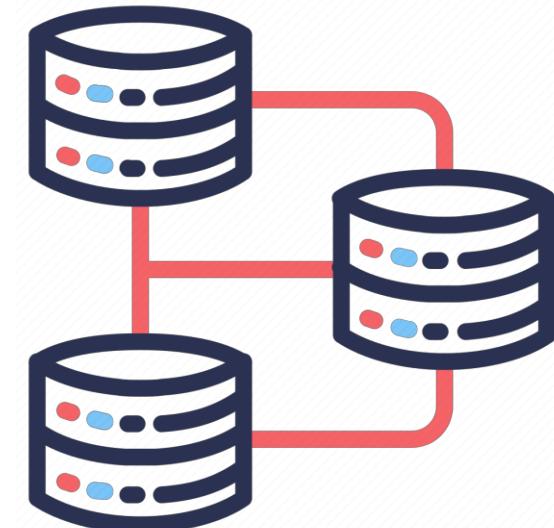
Replicated data

Normalization

Normalization is a process that organizes database tables by **removing anomalies and improving data consistency**. It helps prevent:

- **Update** anomalies
- **Inconsistent** Data
- **Addition** anomalies
- **Deletion** anomalies

ver ejemplos si eso



All **attributes depend on the key**, the whole key and nothing but the key.

- **1NF**: Keys exist and no repeating groups.
- **2NF**: No partial dependencies.
- **3NF**: No transitive dependencies.

1st Normal Form

distinguimos tres niveles

Table has a **primary key**

1NF

No **repeating groups or multivalued attributes**

A **multivalued attribute** is an attribute that may have several values for one record (e.g. list of courses for a teacher)

ID	Name	Courses
1	Pepito	BI, AMD, BDII

A **repeating group** is a set of one or more multivalued attributes that are related.

ID	Name	Course1	Course2	Course3
1	Pepito	BI	AMD	BDII

Already in 1NF. No partial dependencies

2NF

Every non-key attribute must **depend on the full concatenated key**, not just part of it.

T_ID	C_ID	T_Name	C_Name
T1	C1	Pepito	AMD
T1	C2	Pepito	BDII

T_name depends only on T_ID and C_Name on C_ID. How can we solve it?

no se encuentra en la segunda forma normal

C_ID	C_Name
C1	AMD
C2	BDII

T_ID	C_Name
T1	Pepito

T_ID	C_ID
T1	C1
T1	C2

3rd Normal Form

en relacional se suele usar esta

No transitive dependencies

3NF

A table is in **3NF** if it is **2NF** and all **non-key attributes** depend only on the **PK**.

Course	Area_Dpt	Teacher
AMD	LSI	Pepito
BDII	ARQ	Manolito

The candidate Key is Course, but Teacher depends on Area_Dpt.

Course	Teacher	Teacher	Area_Dpt
AMD	Pepito	Pepito	LSI
BDII	Manolito	Manolito	ARQ

Business intelligence

Unit 2 – Datawarehouse

S2-3 – Data integration - ETL

Data integration

Data come from different systems.

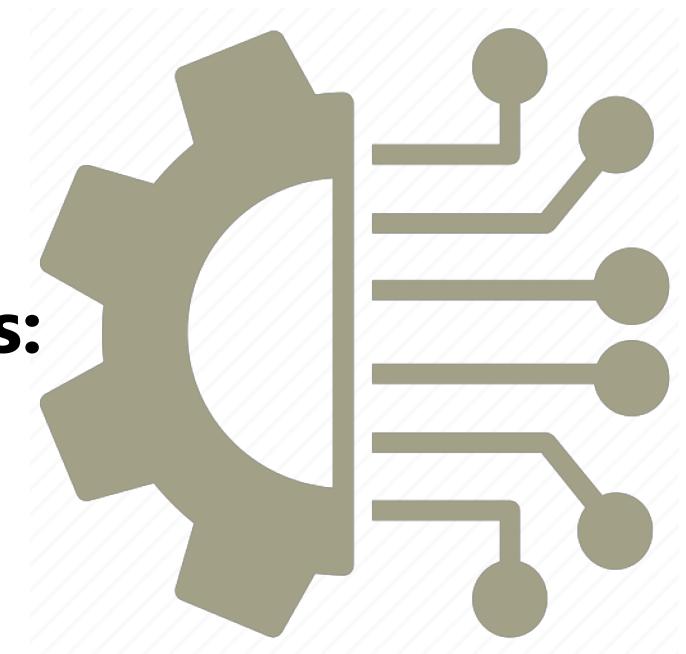
- Need for integration.
- Very different source systems

Data stored in at least in two places:

- The source system.
- The datawarehouse.

DWH features

- Exchange, integrate and consolidate data from many systems.
- It provides a new unified database.
- Volume tends to be very large.



Data integration - ETL

The maintenance of the data warehouse is done by means of the **ETL (Extraction - Transformation - Load)** process

- is **the responsibility of the team** developing the data warehouse.
- is **built specifically** for each DW.
- you can use **market tools** or **programs designed specifically**. **Example:** ETL tools: Oracle Data Integrator, Informatica Power Center, Talend, Stitch, Pentaho, AWS Glue, Alooma, Hevo Data, SSIS (SQL Server Integration Services)
- It is very **time consuming** when building our DW/BI solution.



Importance of ETL

General steps:

- Initial load
- Refreshment: daily,
weekly, monthly, ... periodic maintenance



It **adds value** to data

- **Metadata**: document data quality
- Captures **data flows** (processes)
- Increases **quality**: less mistakes
- **Conformed** data between departments

ETL description



- The data of one or more operating systems **needs to be accessible** in the DWH.
- ETL is the process of **extracting data from source systems** and **bringing them to DWH**
 - Load DWH **regularly**.
- It consists of several steps
 - **Extraction:** Pulling data from the sources.
 - **Cleaning:** Ensuring data is accurate, remove duplicates, ...
 - **Transformation:** Converting the data to match DWH.
 - **Load:** Inserting the transformed data into the DWH.

Business Requirements



- Definition: The business needs determine the selection of data sources and their transformation within ETL.
- Objective: Align ETL processes with business goals for optimal decision-making.

Compliance Requirements

- Regulatory Compliance: Ensure that legal and regulatory data requirements are met (e.g., telecommunications).
- Chain of Custody: Maintain data custody and adhere to legal frameworks for all data and reporting.

Data Integration Requirements



- Key Dimensions and Metrics: Identify common dimensions and attributes across processes to develop standardized KPIs and metrics.
- Objective: Ensure all systems can work together seamlessly with consistent data granularity and units.

Latency Requirements

- Data Availability: Determine the speed at which data must be made available to users.
- Impact: Latency directly influences ETL architecture design and performance.

Archiving and Lineage Requirements



- Archiving Policies: Define policies for data archiving, including the number of copies and retention strategy.
- Data Lineage: Use metadata to track the origin, changes, and quality of data, ensuring transparency in data handling.

Available Skills and Licensing Requirements

- Skills Assessment: Implement ETL based on the available expertise in the organization.
- Legacy Systems: Consider any restrictions imposed by legacy systems and their licensing agreements. It is good to differentiate between when the use of legacy systems is mandatory versus when it is merely recommended.

Data Quality Requirements



- Problem Identification: Identify elements with data quality issues and assess if operational systems can be modified for improvement.
- Monitoring: Include a system for ongoing quality tracking.

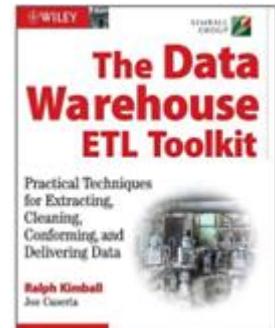
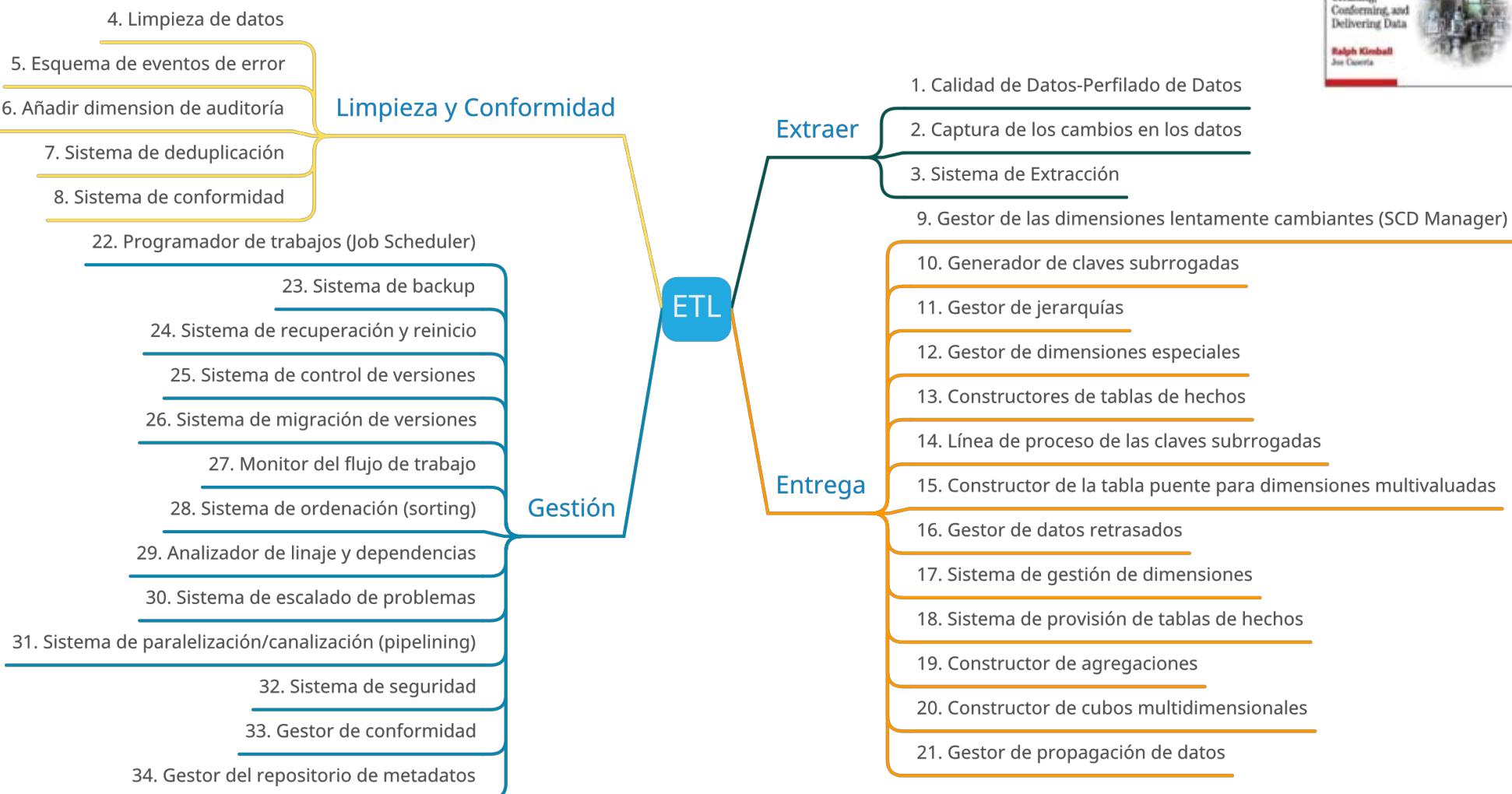
Security Requirements

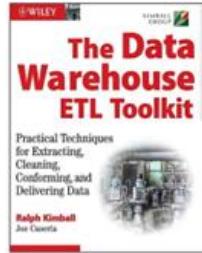
- Data Sensitivity: Assess the level of sensitivity for all data processed through ETL.
- Security Measures: Security extends to backups and must comply with compliance requirements.

Business Intelligence Interface Requirements



- Simple and Reliable Data Access: Ensure data reaches BI applications quickly, reliably, and simply.
- Collaboration: Identify facts and dimensions to expose to BI tools, working closely with data architects and application developers.

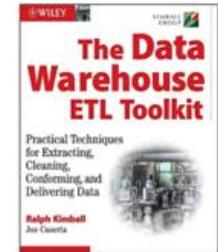




Comp 1: Extraer

- **Subsist 1: Perfilado de datos:** Analizamos los datos para conocer su contenido, consistencia y estructura. Nos permite hacernos una idea de los hechos y dimensiones, y del mapeado a implementar.
- **Subsist 2: Sist de captura de los cambios en los datos:** Podemos cargar todos los datos de las fuentes, pero con los volúmenes de datos de hoy en día eso en ocasiones no es factible. Debemos tener la capacidad de solo transferir aquellos cambios desde la última extracción. Técnicas comunes implican chequear los logs de transacciones, detectar cambios con CRC/Hash, por el timestamp un record ha sido creado/modificado, etc.
- **Subsist 3: Sistema de Extracción:** Relacionado con la obtención de datos de los sistemas fuente. Cada una de ellas podría tener estar en un entorno/BBDD diferente. Podemos obtener los datos en forma de ficheros (XML, csv, txt,...) o a través de un flujo de datos que podemos iniciar a través de una consulta, acceso a un servicio web, etc. Con un fichero no necesitamos re-consultar la fuente en caso de fallo. Aquí podemos decidir comprimir los datos para acelerar la transmisión. Debemos evaluar la necesidad de encriptar la información en tránsito.

Comp 2: Limpieza y conformidad de datos (la T de ETL)

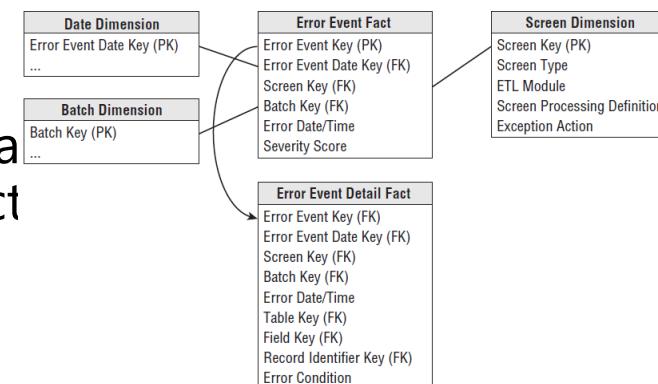


- **Subsist 4: Limpieza de datos:** Desarrollar una arquitectura para diagnosticar problemas, determinar métricas sobre la calidad de los datos, y responder a problemas (ej. resolviéndolo, reportándolo, abortando, continuando,...). Los procesos de filtrado se implementan en el pipeline de datos en la forma de chequeos de calidad, que pueden ser: de columnas (datos dentro de una misma columna), estructurales (relaciones entre los datos en distintas columnas) y de las reglas de negocio (ej. un cliente VIP ha realizado unas compras por encima de un umbral).
- **Subsist 5: Esquema de eventos de error:** Un esquema dimensional centralizado para almacenar los eventos de los chequeos de calidad.

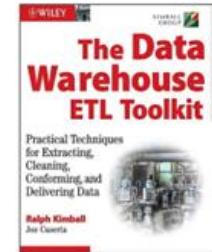
Cada error produce una entrada en Error Event Fact

Los procesos de filtrado son los responsables de añadir records a estas tablas.

Un error que añade una fila e Error Event Fact podría generar muchas entradas en la Error Event Detail Fact



Comp 2: Limpieza y conformidad de datos (la T de ETL)

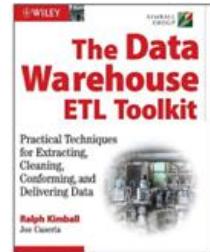


- **Subsist 6: Añadir dimensión de auditoría:** Dimensión añadida por el sistema ETL con el propósito de evaluar los resultados de los procesos de filtrado. Si todo va bien, solo se genera una fila en la tabla Audit Dimension, y la correspondiente FK será añadida a la tabla de hechos. Pero si tenemos condiciones de error, necesitaríamos más entradas en la tabla de dimensiones.

- **Subsist 7: Sistema de deduplicación**
En muchas ocasiones las dimensiones proceden de varias fuentes. Necesitamos combinarlas para crear una imagen que combine las columnas de más alta calidad para cada fila (proceso de supervivencia). Este proceso incluye reglas claras de negocio para decidir la prioridad de las diferentes fuentes a la hora de construir cada fila.

Shipments Facts	Audit Dimension
Ship Date Key (FK)	Audit Key (PK)
Customer Key (FK)	Overall Quality Rating
Product Key (FK)	Complete Flag
More FKs ...	Validation Flag
Audit Key (FK)	Out Of Bounds Flag
Order Number (DD)	Screen Failed Flag
Order Line Number (DD)	Record Modified Flag
Facts ...	ETL Master Version Number
	Allocation Version Number

Comp 2: Limpieza y conformidad de datos (la T de ETL)



- **Subsist 8: Sistema de conformidad:** Encargado de definir y crear dimensiones y hechos estructuralmente idénticos a partir de la combinación e integración de datos procedentes de diversos sistemas que se han deduplicado, filtrados para eliminar datos inválidos y estandarizados.



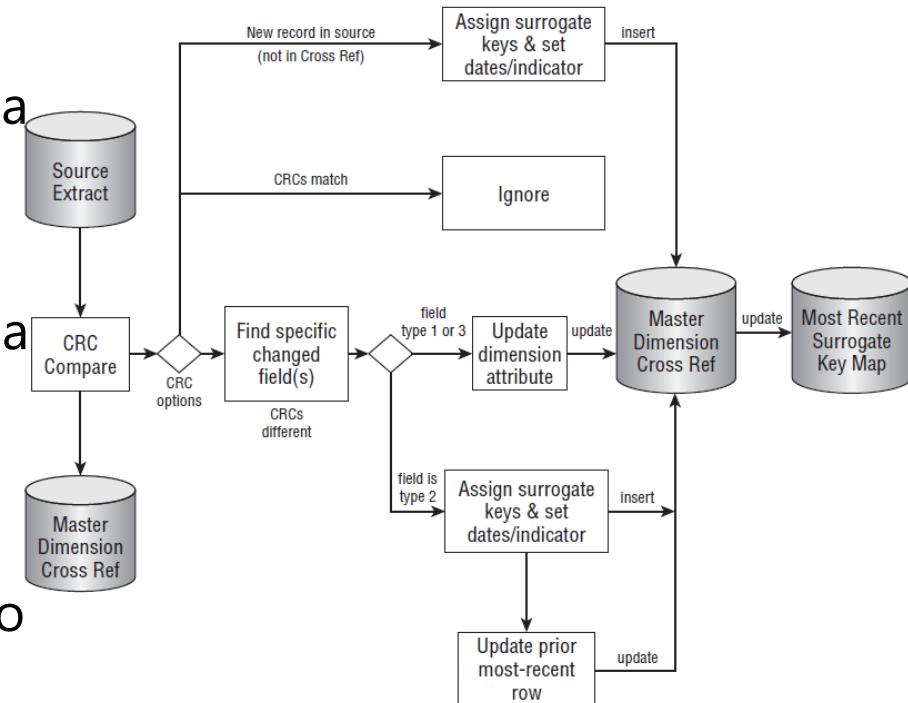
- **Comp 3: Entrega** (la L de ETL)

- **Subsist 9: Gestor de las dimensiones lentamente cambiantes** (SCD Manager): El sistema ETL tiene que gestionar el valor de un atributo que ha cambiado respecto al valor almacenado en el DW. Las respuestas típicas son: sobreescibir (tipo 1), añadir una nueva fila (tipo 2), y añadir una nueva columna (tipo 3)

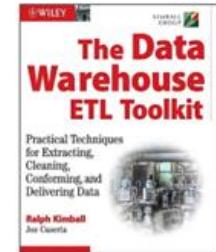
- **Tipo 1-** Cambiamos la dirección de un cliente. Solo actualizamos la fila sin tocar las claves de la tabla de dimensión ni la de la tabla de hechos.

- **Tipo 2-** Ahora imaginemos que queremos mantener el historial de cambios y, por tanto, al cambiar la dirección, necesitamos añadir una nueva fila. Seguimos sin tocar la tabla de hechos, pero usaremos ahora una nueva FK para ese cliente.

- **Tipo 3-** Añadimos una columna nueva para almacenar los cambios. Por ejemplo, en el caso anterior, añadimos una nueva columna para incluir la dirección vieja.



Comp 3: Entrega (la L de ETL)

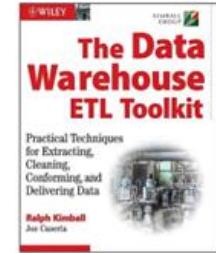
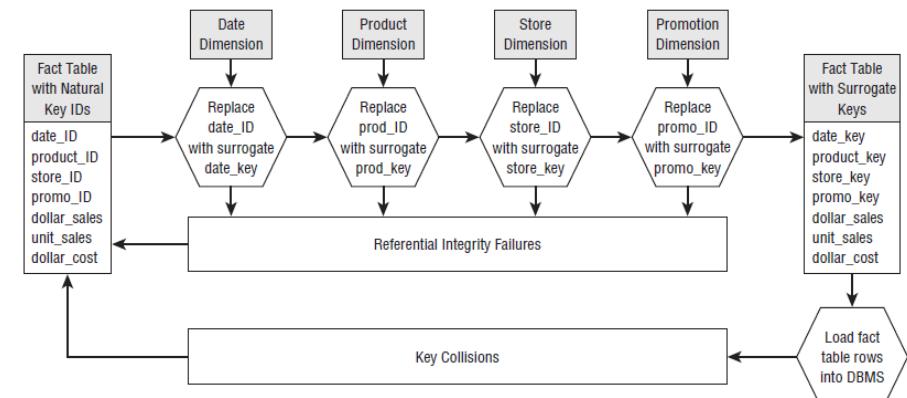


- **Subsist 10: Generador de claves subrrogadas:** Se recomienda usar este tipo de claves como PK de las tablas de dimensiones. Necesitamos un mecanismo robusto para generarlas en entornos ETL. En PostgreSQL, una secuencia valdría.
- **Subsist 11: Gestor de jerarquías:** Es normal encontrarnos con jerarquías en las dimensiones. Hay 2 tipos:
 - **Fijas:** número estable de niveles que se modelan como atributos de la dimensión. Ej: producto->fabricante
 - **Desiguales (ragged):** número variable de niveles. Podemos recurrir a un modelo copo de nieve para modelarlas. A veces , las direcciones postales son desiguales
- **Subsist 12: Gestor de dimensiones especiales:** Apoya las características de diseño de las dimensiones para la organización, pudiendo abarcar todas o algunas de:
 - Dimensiones de tiempo y fecha
 - Dimensiones basura (junk)
 - Dimensiones "encogidas" (shrunken)- subconjuntos de dimensiones mayores
 - Dimensiones pequeñas estáticas- normalmente tablas lookup sin fuente de referencia

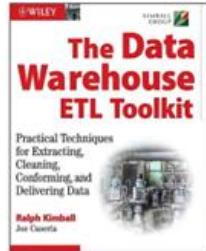
- **Comp 3: Entrega** (la L de ETL)

- **Subsist 13: Constructores de tablas de hechos**: Enfocado a la construcción de los diferentes tipos de tablas de hechos:
 - **Transaccionales**: Se cargan cuando la transacción ocurre o en intervalos a la máxima granularidad . Se cargan los datos y, a veces, si se requiere actualizar datos ya almacenados, primero se insertan y después de borran los viejos.
 - **Instantáneas (snapshot) periódicas**: Se cargan a intervalos regulares y representan períodos. Ej: balances mensuales de un banco.
 - **Instantáneas acumuladas**: para representar eventos finitos que evolucionan con un inicio y final bien definidos. Ej: procesar un pedido: tenemos primero la fecha de pedido, pero no sabemos la de envío, recogida,... A medida que las sabemos tenemos que ir actualizándolas.
- **Subsist 14: Línea de proceso de las claves subrogadas**:

Reemplazar las claves operacionales naturales del registro de la tabla de hechos con las subrogadas de la dimensión.



- **Comp 3: Entrega** (la L de ETL)
 - **Subsist 15: Constructor de la tabla puente para dimensiones multivaluadas:** Soporta relaciones Many-to-Many entre hechos y dimensiones. Ej: hecho: compra (cantidad, valor); dimensión: razones de la compra. Otras: pacientes/diagnósticos, clases/profesores,...
 - **Subsist 16: Gestor de datos retrasados:** En realidad, no todos los datos llegan al mismo tiempo. Ej: podemos tener datos de ventas diarios y datos de personal mensuales. El sistema ETL debe solucionar estas situaciones en las que hechos y dimensiones pueden llegar tarde para mantener la integridad referencial. Una posible solución podría ser asignar un valor por defecto para la dimensión hasta que se conoce ésta. Ej: Nombre cliente: TBD
 - **Subsist 17: Sistema de gestión de dimensiones:** Autoridad centralizada que prepara y publica las dimensiones conformadas (dimensiones cuya semántica, estructura y uso conviene a toda la empresa) a la comunidad del DW. Entre sus responsabilidades:
 - Implementar etiquetas descriptivas para los atributos
 - Gestionar los cambios en los atributos
 - Añadir nuevas filas a la dimensión conformada, generando nuevas claves subrrogadas
 - Distribuir las actualizaciones dimensionales

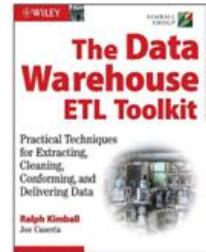


Comp 3: Entrega (la L de ETL)

- **Subsist 18: Sistema de provisión de tablas de hechos:** Administra una o varias tablas de hechos, siendo responsable de su creación, mantenimiento y uso. Responsabilidades:
 - Recibe las dimensiones replicadas del gestor de dimensiones.
 - Añade y actualiza las filas de las tablas de hechos
 - Asegura la calidad de los hechos
 - Recalcula los agregados
 - Notifica a los usuarios de cambios, actualizaciones y problemas.
- **Subsist 19: Constructor de agregaciones:** Creación y mantenimiento de estructuras físicas agregadas en la BBDD que ayudan a mejorar el rendimiento. Debemos controlar cuidadosamente cuales necesitamos, sin quedarnos cortos ni pasarnos. Los hechos/dimensiones sumario se crean a partir de los hechos/dimensiones base.
- **Subsist 20: Constructor de cubos multidimensionales:** Crear y mantener los esquemas ROLAP estrella que permiten crear los cubos. Los cubos MOLAP presentan datos multidimensionales fáciles de explorar. Los cubos deben actualizarse si los hechos o dimensiones cambian.

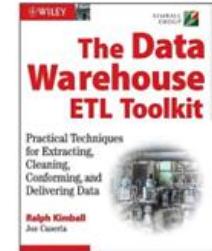
ETL: Comp 3: Entrega

- **Comp 3: Entrega** (la L de ETL)



- **Subsist 21: Gestor de propagación de datos:** Se encarga de mover datos del DW a otros entornos, como sistemas de minería de datos, o de auditoría.

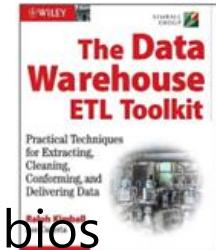




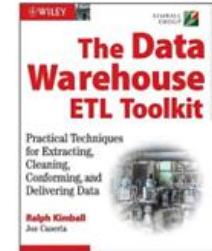
Comp 4: Gestionar el entorno ETL

- El entorno ETL debe ser fiable, disponible según los SLA, y adaptable según la evolución del negocio. Para ellos necesitamos los siguientes subsistemas:
 - **Subsist 22: Programador de trabajos** (Job Scheduler): Sistema para programar y lanzar los trabajos ETL. Debe conocer y controlar las dependencias entre trabajos ETL. Responsable de definir trabajos, programarlos, capturar metadatos y volcarlos en ficheros log, así como notificar problemas.
 - **Subsist 23: Sistema de backup**: Almacenar, archivar y extraer elementos del sistema ETL.
 - **Subsist 24: Sistema de recuperación y reinicio**: Los trabajos deben recuperarse en caso de errores y tener la capacidad de reiniciarse.
 - **Subsist 25: Sistema de control de versiones**: Almacenamos información sobre las diferentes versiones (scripts, SQL, Jobs,...) en el flujo de procesos de ETL: Git, SVN, ...

Comp 4: Gestionar el entorno ETL

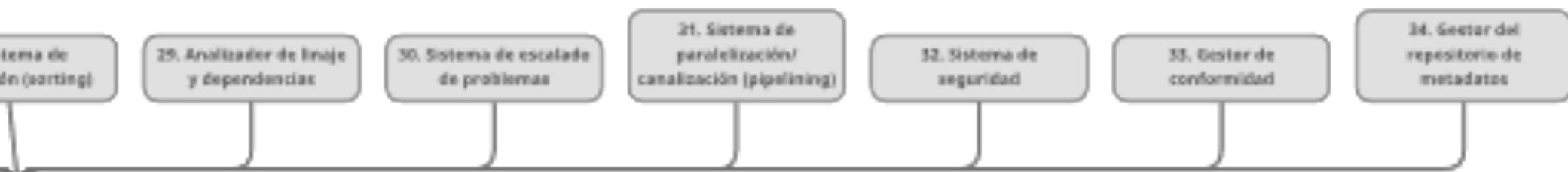


- **Subsist 26: Sistema de migración de versiones:** Transferir los cambios desde la etapa de desarrollo, a testeо, y después a producción
- **Subsist 27: Monitor del flujo de trabajo:** Panel de control y sistema de reporte de todos los trabajos iniciados por el programador de trabajos. Incluye un sistema de auditoría, los logs del ETL y la monitorización de la BBDD.
- **Subsist 28: Sistema de ordenación (sorting):** Ordenar es una capacidad muy importante en ETL, usada muy a menudo en agregaciones y joins. Es importante que se pueda aplicar de forma rápida y eficiente.
- **Subsist 29: Analizador de linaje y dependencias:** (linaje) -> Representa las fuentes físicas y las transformaciones de cualquier elemento de datos, en cualquier etapa del proceso ETL. (dependencia) -> Representar todos los elementos de datos "downstream" y reportes finales que se verían afectados por un cambio en un elemento "upstream". Ej dependencia: identificar los cubos y esquemas estrella que usan un elemento de la fuente.
- **Subsist 30: Sistema de escalado de problemas:** Sistema automático/manual donde una condición de error se eleva al nivel correspondiente para analizarlo y resolverlo. Va desde simples entradas en los logs, hasta a notificar a los operadores, supervisor o desarrollador.

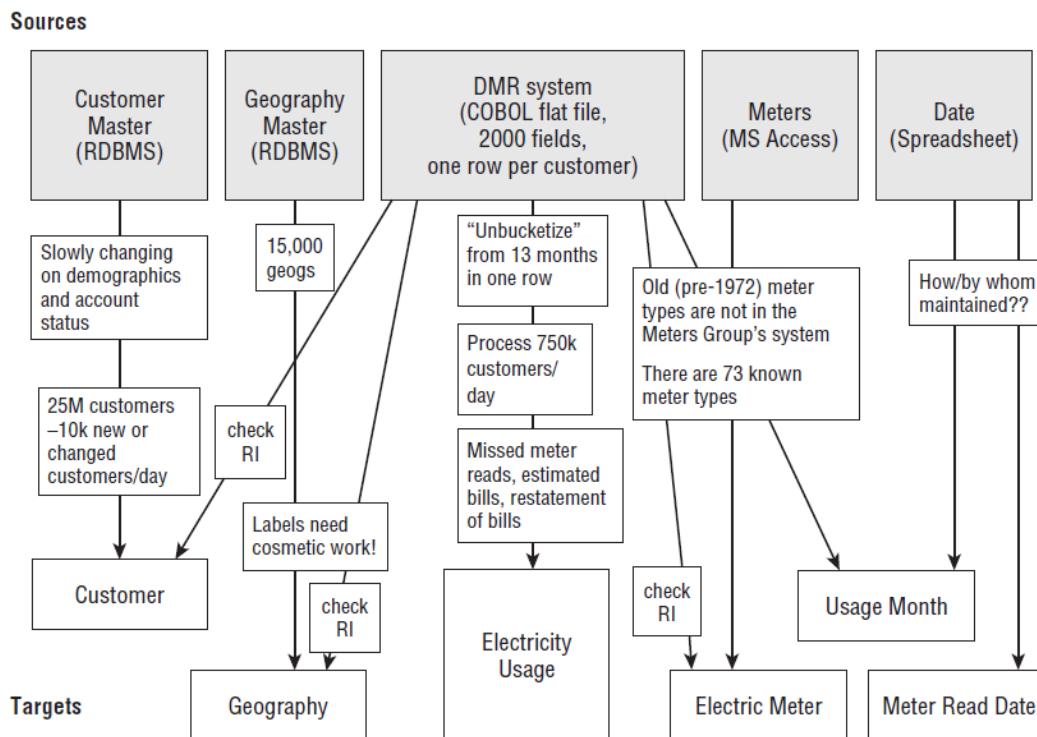


Comp 4: Gestionar el entorno ETL

- **Subsist 31: Sistema de paralelización/canalización** (pipelining): Normalmente las restricciones temporales nos obligan a aprovecharnos de la disponibilidad de recursos computacionales para aquellas tareas que se pueden aprovechar de ello.
- **Subsist 32: Sistema de seguridad**: Gestionar la seguridad a nivel individual y rol de todos los datos y metadatos en ETL. Esto también incluye el sistema de backup.
- **Subsist 33: Gestor de conformidad**: Mantener la cadena de custodia de los datos en entornos regulados (ej: bancos, salud,...) y registrar quién ha accedido de forma autorizada a los datos. Los cambios en los datos deben de reportarse.
- **Subsist 34: Gestor del repositorio de metadatos**: Administra todos los metadatos de ETL. Esto incluye la captura y mantenimiento, incluyendo toda la lógica de las transformaciones. Incluye metadatos de procesos, técnicos y de negocio.



- **Pasos preliminares (Kimball):**
 - Diseño lógico
 - Comprender la arquitectura DW/BI
- **Etapas planificación ETL:**
 1. **Esbozar un plan de alto nivel:** Identificar fuentes y destinos (dimensiones y tablas hechas)



ETL SUBSYSTEM			
EXTRACTING DATA	CLEANING AND CONFORMING	DELIVERING FOR PRESENTATION	MANAGING THE ETL ENVIRONMENT
ETL PROCESS STEP			
Plan			
Create a high level, one-page schematic of the source-to-target flow.	1		
Test, choose, and implement an ETL tool (Chapter 5).			
Develop default strategies for dimension management, error handling, and other processes.	3	4, 5, 6	10
Drill down by target table, graphically sketching any complex data restructuring or transformations, and develop preliminary job sequencing.	4, 5, 6	11	22
Develop One-Time Historic Load Process			
Build and test the historic dimension table loads.	3	4, 7, 8	9, 10, 11, 12, 15
Build and test the historic fact table loads, including surrogate key lookup and substitution.	3	4, 5, 8	13, 14
Develop Incremental Load Process			
Build and test the dimension table incremental load processes.	2, 3	4, 7, 8	9, 10, 11, 12, 15, 16, 17
Build and test the fact table incremental load processes	2, 3	4, 5, 8	13, 14, 16, 18
Build and test aggregate table loads and/or OLAP processing.			19, 20
Design, build, and test the ETL system automation.	6	17, 18, 21	22, 23, 24, 30



Etapas planificación ETL:

- 2. Seleccionar una herramienta ETL:** Existen muchas opciones que pueden leer de diversas fuentes. Podemos desarrollarla o bien emplear herramientas GUI.
- 3. Desarrollar estrategias** para las actividades más comunes en el sistema ETL. Incluye:
 - Extracción de datos y archivado
 - Evaluar calidad de dimensiones y hechos
 - Administrar cambios en atributos de dimensiones
 - Asegurar cumplimiento requisitos de disponibilidad del DW y ETL
 - Diseñar sistema de auditoría de datos
 - Organizar el almacenamiento de datos intermedio (staging area) para almacenamiento temporal.



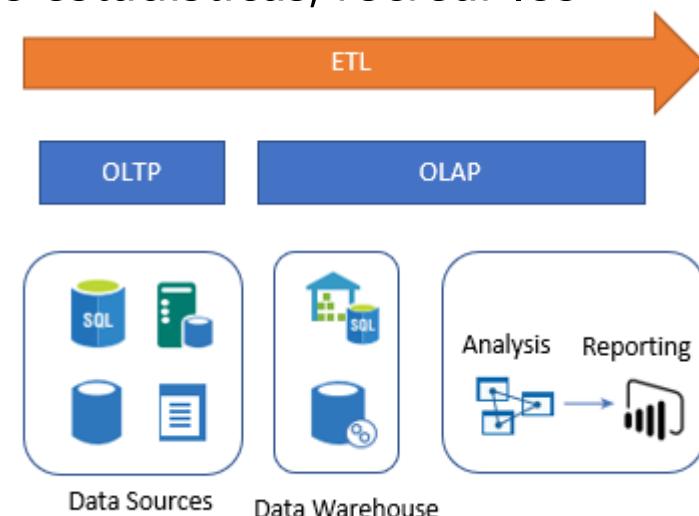
- **Etapas planificación ETL:**

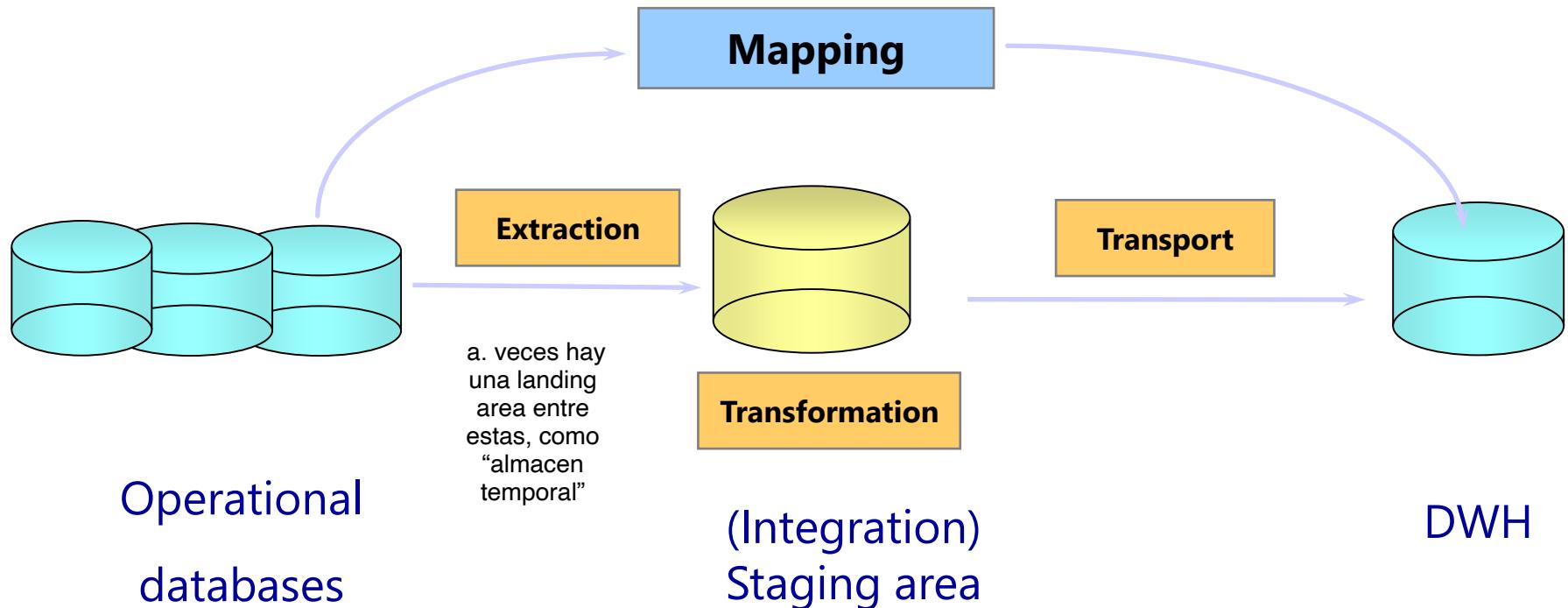
4. **Mapear de forma detallada** el flujo de datos desde las fuentes a las tablas de dimensiones y hechos. Podemos hacer uso de diagramas de flujo o esquemas que nos ayuden a realizar esta labor.
5. **Cargar las tablas de dimensiones con datos históricos**: Este es un proceso que hacemos una vez y que difiere de la carga incremental. Normalmente comenzamos con las tablas más simples (ej: tipo 1). Después podemos transformar los datos : convertir tipos, combinarlos de fuentes diferentes, asignarles claves subrrogadas, etc.
6. **Cargar la tabla de hechos con datos históricos**: Podemos realizar transformaciones: sustituir valores numéricos “especiales” por NULL en hechos aditivos o semiaditivos, calcular hechos complejos, añadir una clave de auditoría, etc.
7. **Procesado incremental de la tabla de dimensiones**: A veces podemos usar la misma lógica que en la carga de datos históricos. Un reto es identificar los datos nuevos/actualizados.

- **Etapas planificación ETL:**



8. **Procesado incremental de la tabla de hechos:** Esta fase normalmente requiere automatizar el proceso y puede requerir procesado en paralelo y chequeo de la calidad de datos.
9. **Agregar carga de tablas y OLAP:** A veces necesitamos procesado adicional más allá del esquema estrella. Necesitamos actualizar MOLAP, las vistas (incluyendo las materializadas) y las tablas sumario con agregaciones.
10. **Operación del sistema ETL y automatización:** en esta etapa necesitamos programar los trabajos, gestionar los errores/excepciones, operaciones sobre las BBDD(actualizar las estadísticas, recrear los índices, limpiar datos,...)

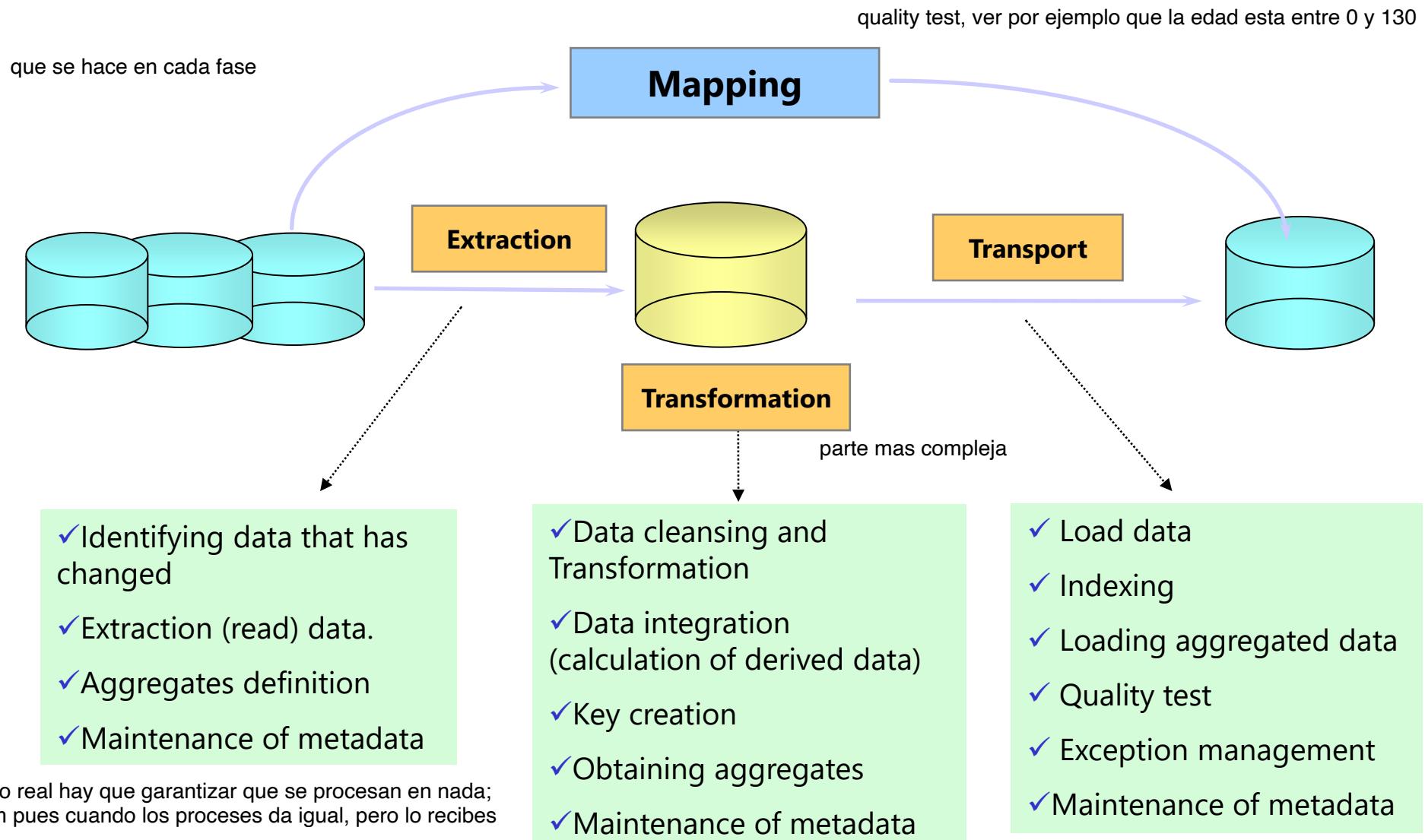




The **Staging** area:

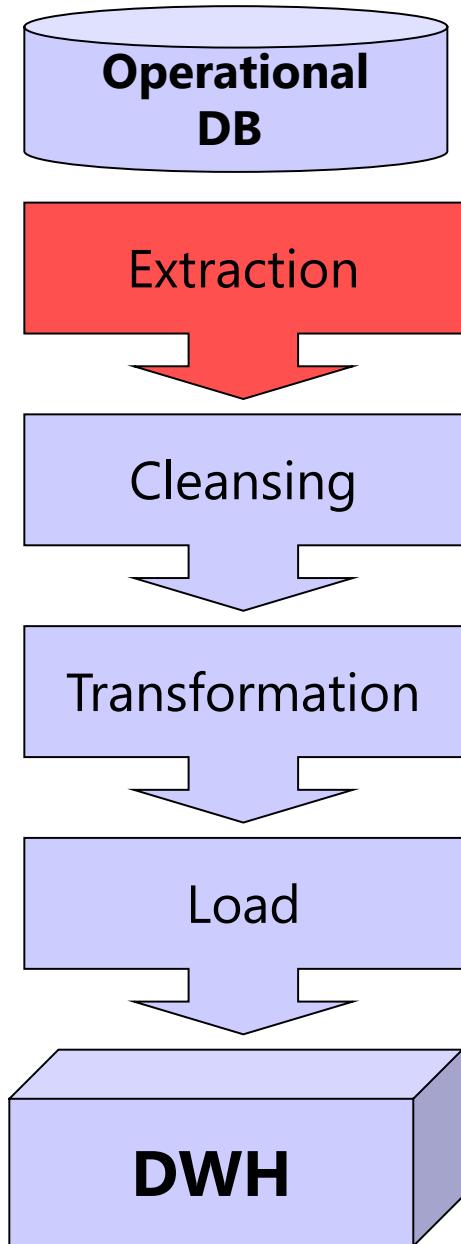
- allows transformations without paralyzing the operational databases or the datawarehouse.
- Allows storing metadata.
- Facilitates the integration of external sources.
- Not necessarily stores data in Relational model

ETL



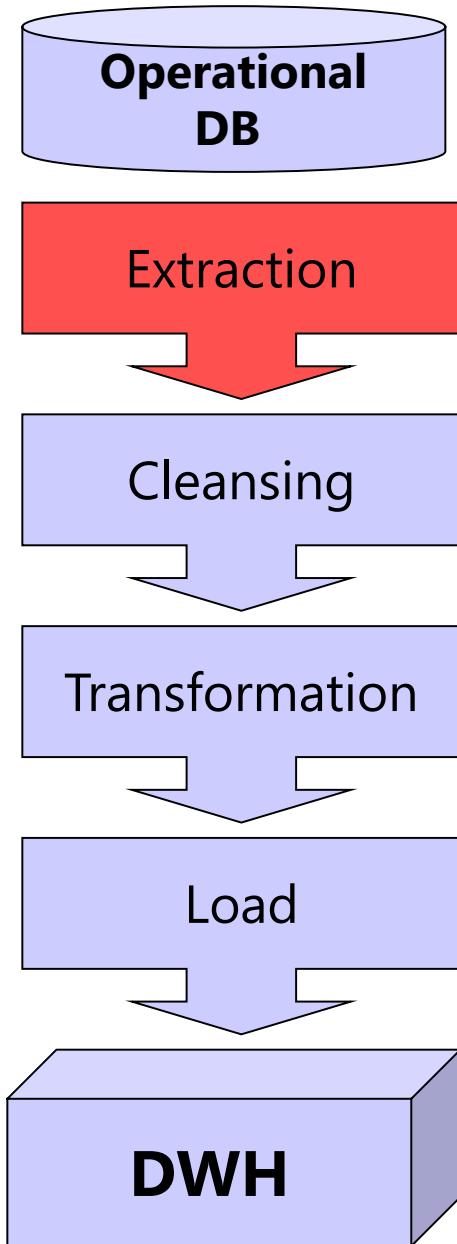
en tiempo real hay que garantizar que se procesan en nada;
en stream pues cuando los proceses da igual, pero lo recibes

los indices son muy importantes para el rendimiento (no pasarse !!). Un indice no es mas que otra tabla que ocupa espacio, asi que hay que ser cuidadoso



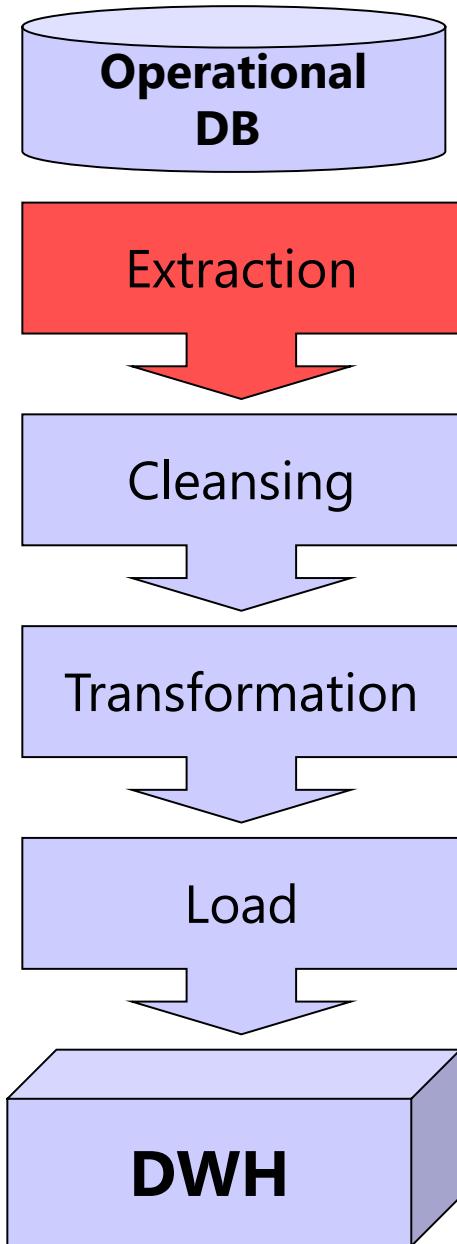
- **Extraction** = obtaining an image of a given subset of the source data to be loaded into the DWH
 - **Logical** and **physical** extraction.
 - Logical: Complete vs incremental
 - Physical: Online vs Offline
 - Also “**Ingestion**” in BigData





- **Complete extraction**

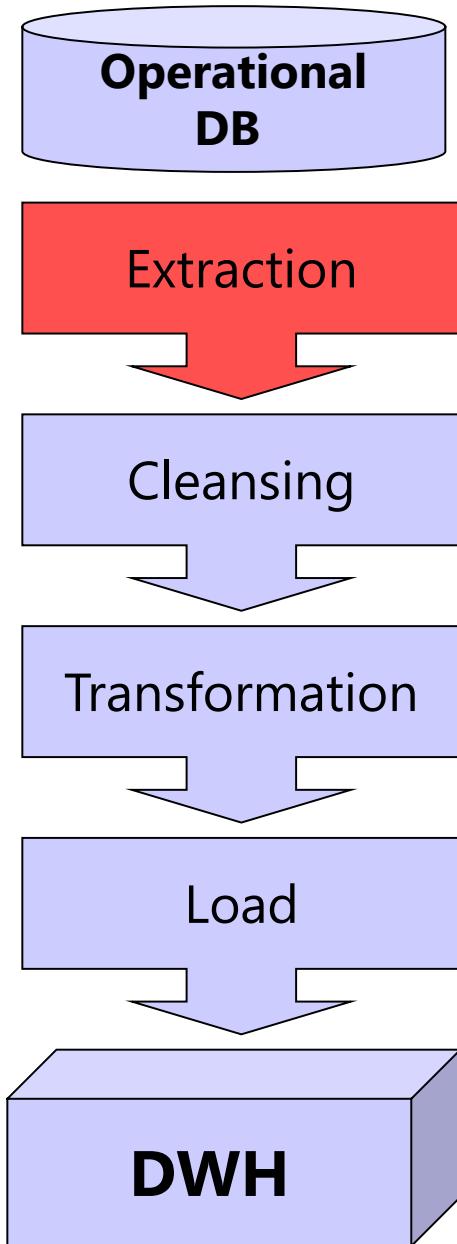
- Capture **all the source** data in an instant of time
- The extraction reflects **all available data** in the source system, it is not necessary to look at the source changes from the last extraction.
- The source data are taken "**as is**", and it does not require additional logic (e.g. timestamps).
- Example: dumping one table, or SQL that retrieves the entire table.
- Moving from **hadoop to relational and viceversa**: Apache Sqoop, Spark, NiFi, Flume, Kafka; Talend, ...



- **Incremental extraction**

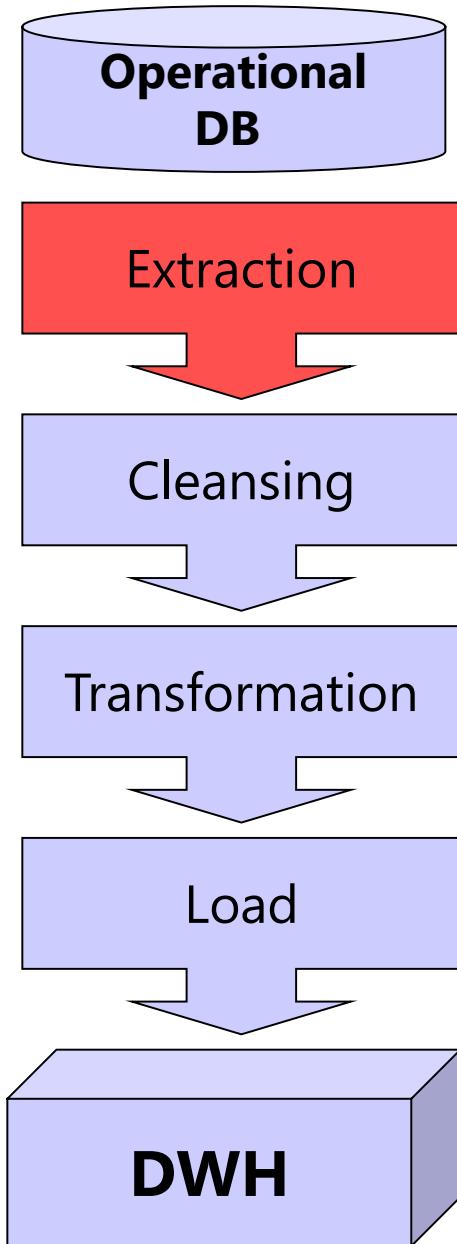
no necesitamos /queremos extraer todos los datos

- At a **specific point in time**. One event triggers it.
- Only **changed data** will be extracted.
- Identify the information from one point changing in time **CDC (Change Data Capture)**: posibles soluciones
 - **Timestamps** on row
 - **Version number** on row
 - **Status** on row
 - **Triggers** (on database or application)
 - **Logs** (on database or application)
 - Powerful tools for this choice
- Related to **Slowly Changing Dimension**
- Pull Vs Push
 - absicamente quien toma la iniciativa para coger los datos: pull hace peticion a la fuente. Push la fuente pone los datos a disposicion
- Real time or event oriented processing: eg. Flume



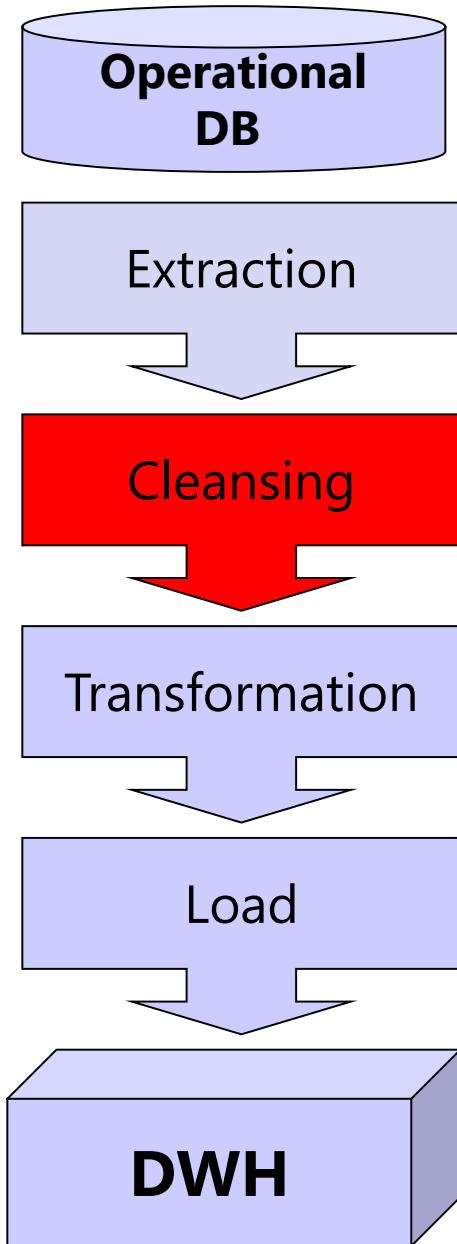
• **Online extraction**

- **Connect directly** to the source or an intermediate system that stores the required data (eg, snapshots or change tables).
- The **intermediate system** is not necessarily different from the source system. It can be a separate system designed for handling data extraction.
- In some cases, incremental extraction is handled by this intermediate system (e.g., CDC mechanisms, logs, snapshots), making it more efficient than directly querying the primary source.
- (See previous Incremental extraction)



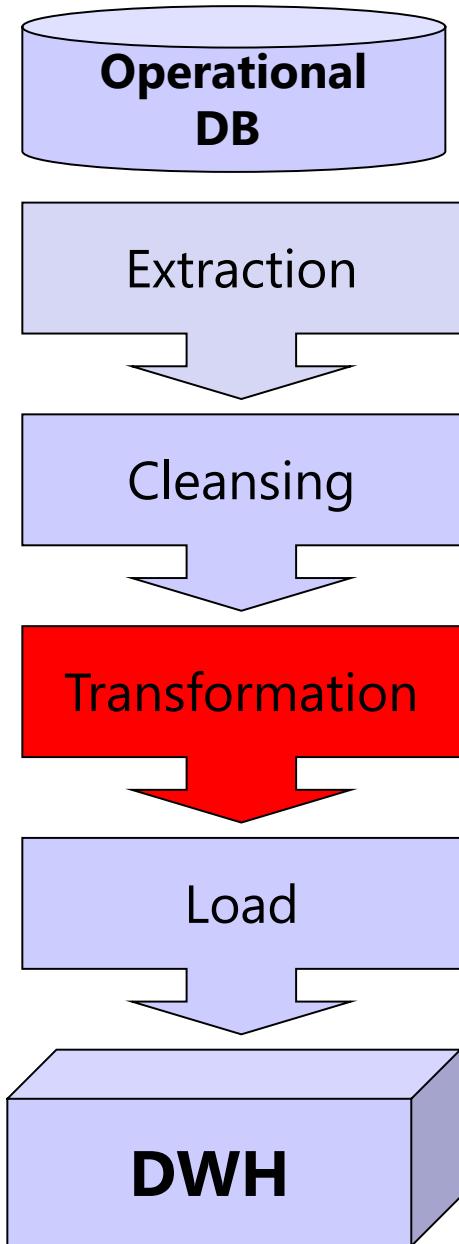
• Offline extraction

- Data not extracted directly from the source system. It runs outside the original source system. It is taken from external copies or backups of the data, which exist outside the original source system.
- The data have an existing structure or are created by a dumping routine.
 - **Flat files.** Data in a generic format, eg CSV. They may include extra info on format.
 - **Dump files.** DBMS specific format. Extra info on SQL + format.
 - **Redo logs and archive logs. Transportable tablespaces.**



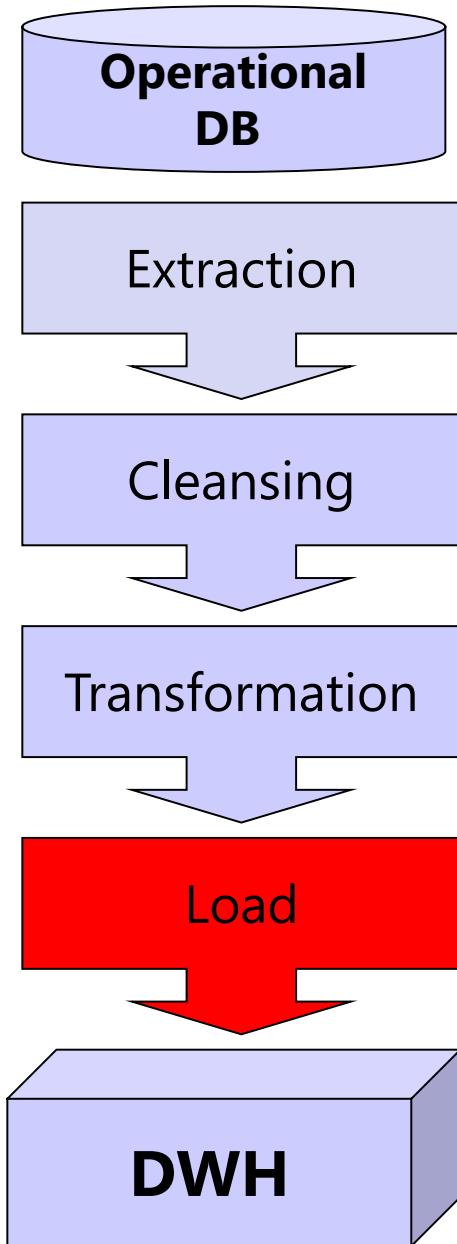
Cleansing = ensuring data quality

- **Correct mistakes:** format errors, wrong dates, misuse of fields, duplicate data, inconsistencies, restore referential integrity
- Also: decoding, reformatting, add timestamps, converting data, key generation, etc.
- **Alternative terms:** Wrangling, munging, tidying



Transform = convert data from the format of an operational system to the DWH format

- **Record Level:**
 - Selection - partitioning
 - Joining - combination
 - Aggregation - precalculation
- **Field level:**
 - Single field - from field to field
 - Multiple field - from multiple fields to a single field or vice versa
- **Types of data transformation**
 - Gradual: step by step transformation process.
 - Pipelined
 - Multiple steps grouped.
 - It is not necessary to use intermediate tables.



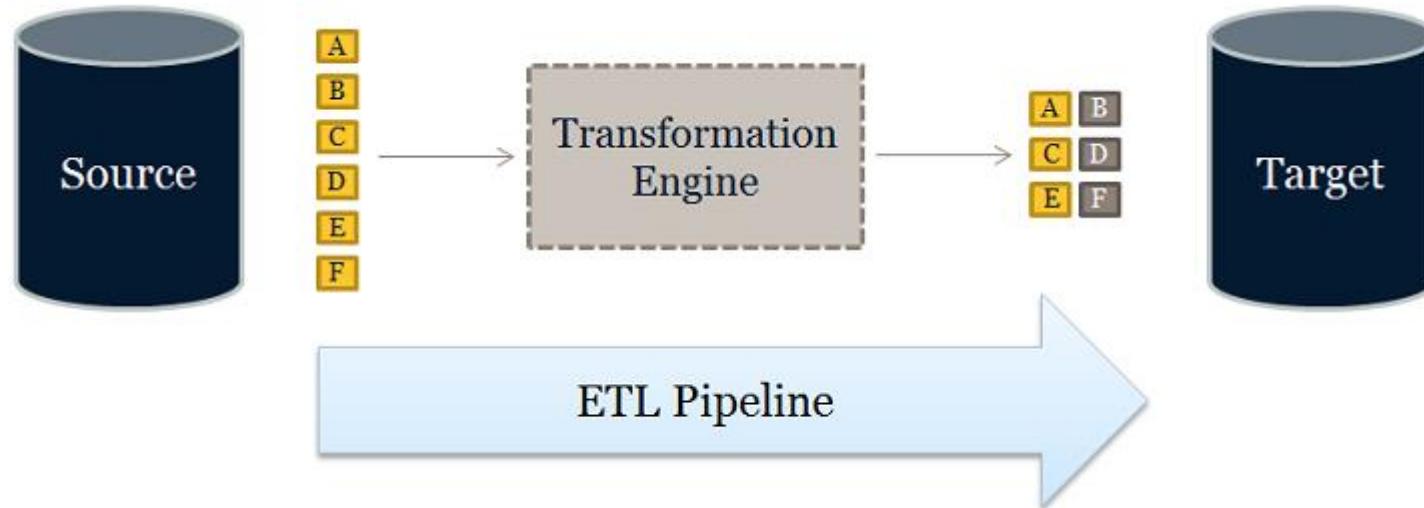
Load = enter data into the DW and create indexes

- Refresh method: determination of refresh intervals
 - Remove unused old data
- Update method: only changes in the sources are taken into account
- Loading mechanisms
 - SQL * Loader (pgload and COPY in Postgres)
 - External tables
 - OCI API and Direct-Path
 - Export / Import (pgdump in Postgres)
 - Sqoop

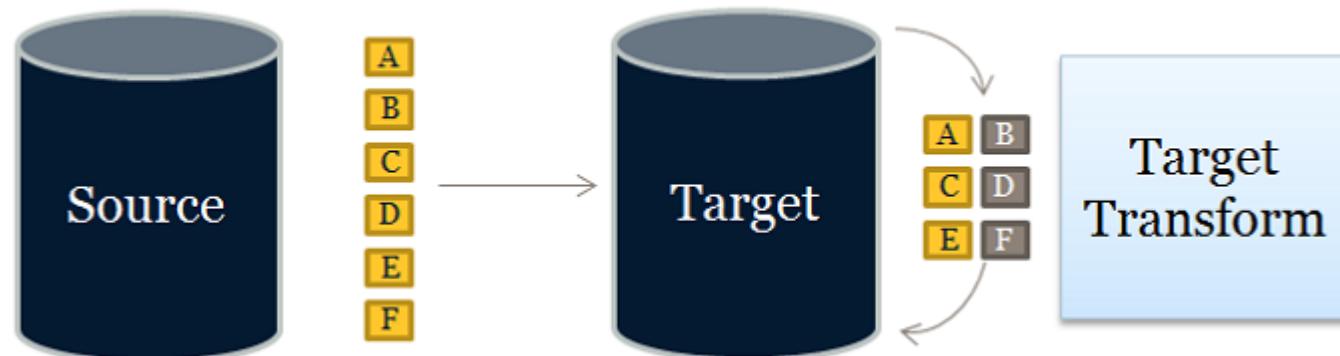
ETL vs ELT (note the order of letters)

- ETL Classic vs ELT “Big Data” approach

en elt los procesos de transformacion tienen lugar dentro del almacen.
Los datos no tienen que estar ajustados a un esquema determinado

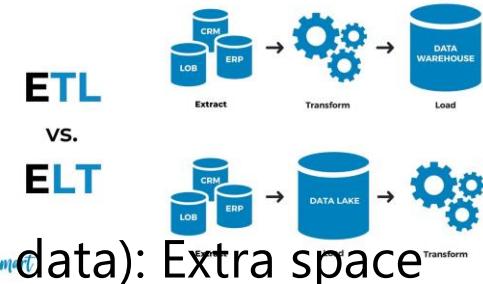


elt es mas costosa y se usa mejor en entornos regulares. Si tengo necesidad de ver datos crudos mejor elt o ingerir datos mas rapidos



ELT:

- **Data lakes**
 - **Raw data** in staging in DWH (minimal filter unneeded data): Extra space
 - **Faster**
 - For future needs, there are more data (not only the datamart)
- **Schemaless** / schema on read
- **Less processing on the source** since no transforming is being done
- **More processing when consuming** (need to transform)
- Languages with limited processing: hive, pig, ...
- **More scalability and better I/O** than ETL if classic ddbb is used
- Supports both batch or event oriented



Classic ETL

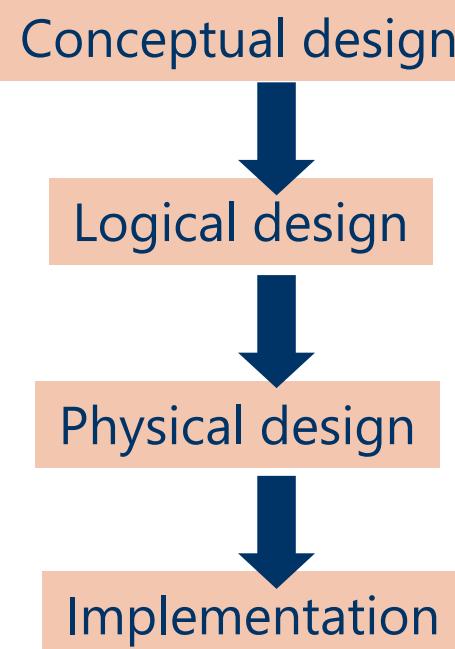
- Data marts: pre-aggregated data repositories optimized for queries. Cubes / **Schema on write**
- Only **clean, validated and quality** data
- Staging not needed: Less space
- More processing on extraction: a **pipeline**
 - Much slower
- GUI tools: abstraction independent from language and platform, very rich and powerful transformations

An ETL integration tool is essential for handling data movement, transformation, and quality assurance in data warehousing processes.

- An integration tool needs:
 - Many interfaces
 - Transformations
 - Quality: validation, filters, ...
 - Open Source:
 - Apache Airflow, Apache Kafka, Apache Nifi, Pentaho Data Integration, Talend OS, ...
 - Commercial:
 - Hevo Data, Informatica Power Center, SnapLogic, Jitterbit, Oracle Data Integrator, Mulesoft, IBM Datastage, Microsoft SQL Server Integration Services,

Business intelligence

Unit 2 – Datawarehouse and OLAP
S2-4 – Physical design



Efficiency in storage
Efficiency in queries

Storage model (ROLAP,
MOLAP, HOLAP)





Objective: improve performance

Focus on **improving query performance and scalability.**

Balance between storage efficiency and fast retrieval.

Architecture: **ROLAP, MOLAP, HOLAP**

Storage strategy and query optimization

Denormalization (ROLAP, HOLAP)

Partitioning (ALL)

Index (ALL)

Compression (ALL)

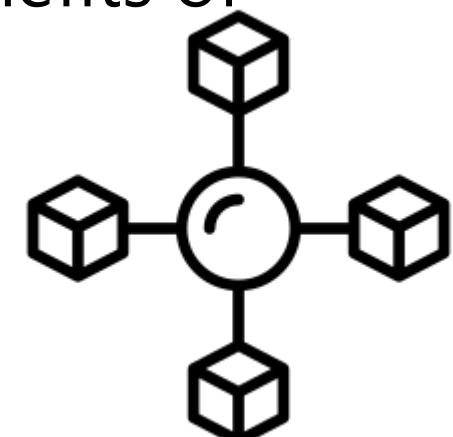
Materialized Views (ROLAP, HOLAP)

Big data challenges



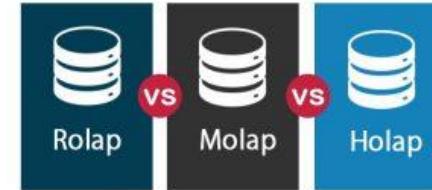
Different physical storage and query structures

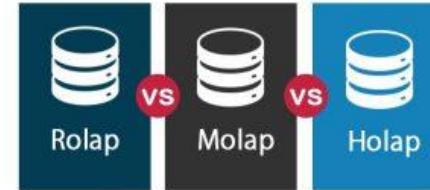
- Multidimensional OLAP (**MOLAP**): DBMS built specifically for data analysis. Data is stored in cubes to improve multidimensional queries.
- Relational OLAP (**ROLAP**) Relational DBMS, which can include both traditional row-based and modern columnar databases.
- Hybrid systems: **HOLAP**. Combines benefits of ROLAP and MOLAP for optimized performance.



ROLAP:

- The DWH is built **on top of relational DBMS**, often using columnar storage to improve performance with big datasets.
- RDBMS vendors provide OLAP-specific extensions, such as **window functions** (e.g., RANK(), SUM() OVER()), and **grouping sets** (GROUPING SETS, CUBE, ROLLUP) or integrated data mining tools: Oracle OLAP, Microsoft SSAS ROLAP, PostgreSQL with OLAP extensions.
- Typically uses a **star schema** or **snowflake** schema with fact and dimension tables.





ROLAP:

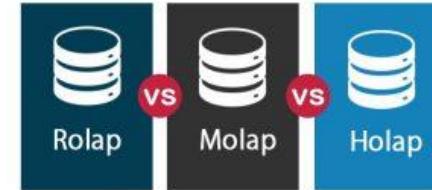
Advantages:

- Scalability: There is no limit on the amount of data.
- Proven technology: Functionality already available.
- Separation storage from OLAP processing.
- Cost-effective

Disadvantages:

- Performance overhead: queries on normalized db, usually with multiple joins. Complex multidimensional queries usually slower than in MOLAP systems.
- Limited functionality for SQL. For example, to perform complex calculations or hierarchical queries.
- No pre-aggregated data, slowing response times.
- Potential locking and concurrency problems in busy systems.

MOLAP



- Consist of physically storing data in multidimensional structures (**cubes**) so that the external representation *matches* the internal representation.
- It allows **fast retrieval** of pre-aggregated data.
- BD complexity is **hidden** from the users.
- Analysis is done on **aggregated data** and precalculated metrics or indicators.
- OLAP Engine: The **MOLAP engine** processes and responds to multidimensional queries with very low latency due to pre-aggregated data.
- Specific functionality: Array data structures (proprietary formats), query optimization , data compaction.

MOLAP:

Advantages:

- Fast Performance for small and medium size datasets: slide & dice specific operations.
- Hierarchical analysis: easy to navigate through dimensions.
- Complex calculations are pre-generated.
- Easier to use, even for inexperienced users.

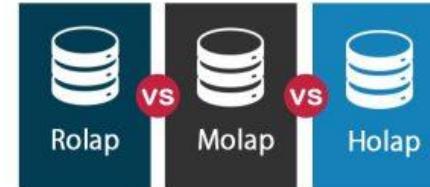


MOLAP:

Disadvantages:

- Limited size: Problems with large datasets and many dimensions.
- Latency when loading and pre-aggregate: it can take time.
- Storage overhead: pre-aggregated data can take a lot of storage, specially for large datasets.
- New investment: this technology is not usually present in enterprises. Also usually proprietary.
- Limited Flexibility: Less flexible when new dimensions or hierarchies are introduced, requiring a cube reprocessing.





MOLAP: Improving performance

- **Partial Cube Aggregation:** Selectively pre-aggregate dimensions or measures that are queried most frequently, reducing cube processing time.
- **Incremental Cube Refresh:** Use incremental refresh to update only the new data, minimizing downtime and latency.
- **Sparse Data Handling:** Manage sparse data more efficiently by compressing and storing only non-null values.
- **Advanced Partitioning:** Partitioning cubes by key dimensions (e.g., time) allows for faster access and easier management of data over time.

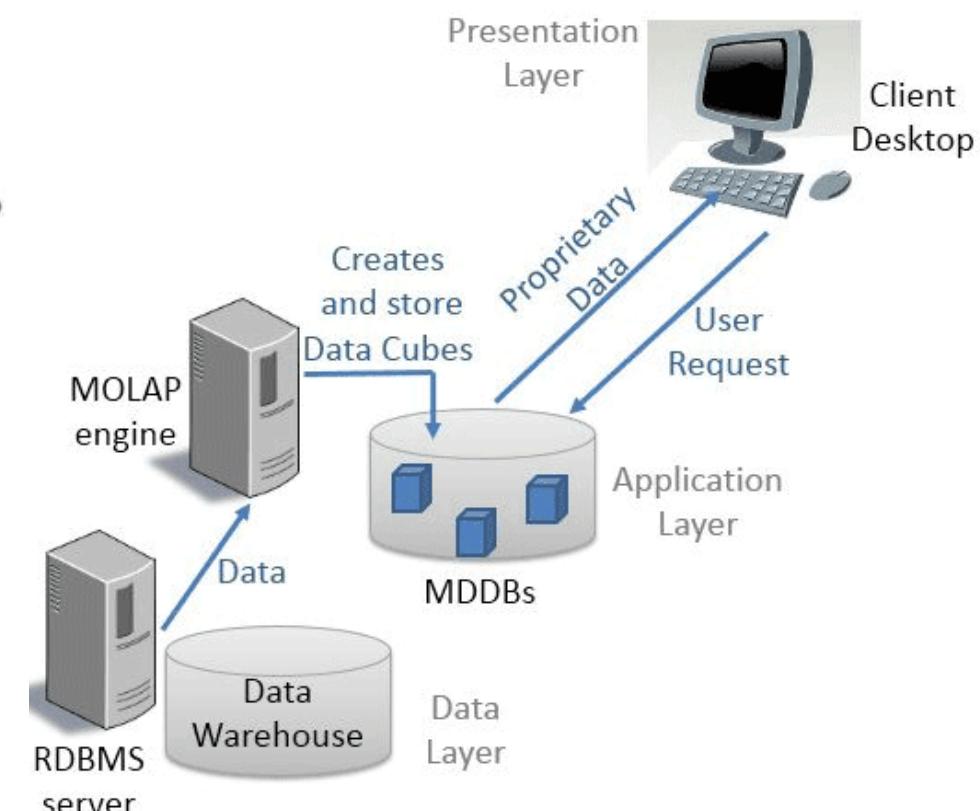
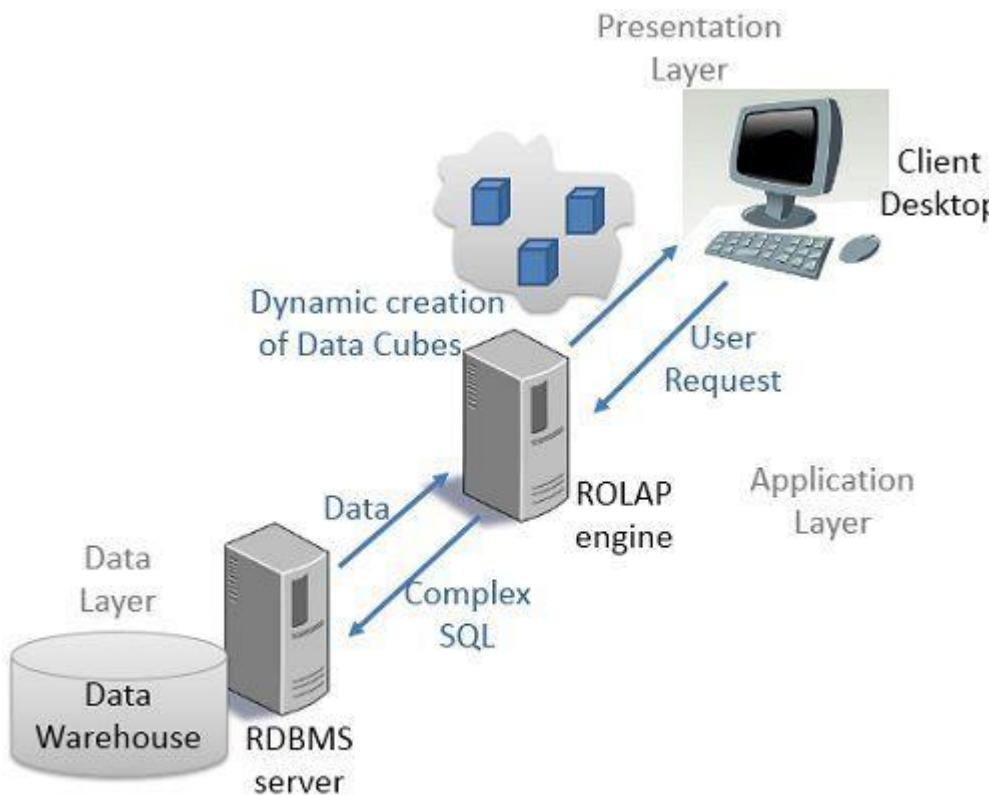


Image Source:EDUCBA

Database partitioning

Replace large tables with **more manageable** smaller subtables

- Allows **parallel and distributed** execution
- Partition **prunning** and **partition wise** joins
- **Partition data processing**: ETL, refreshing, backup, indexing.

Horizontal

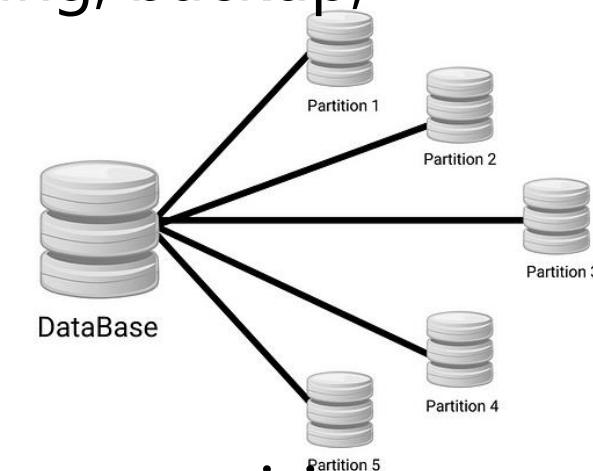
- Rows of a table are separated

Vertical

- Columns of a table are separated

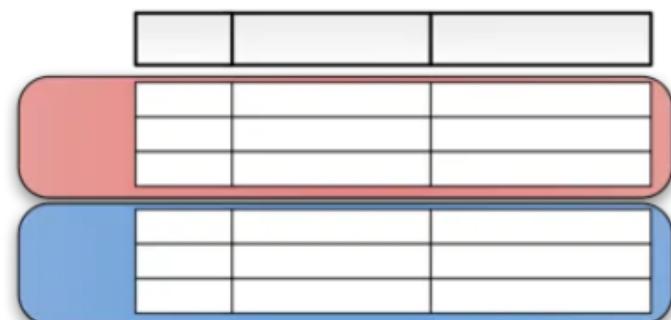
Dynamic Partitioning: automatically creates partitions based on incoming data, without the need for predefined ranges or categories.

Time-based Partitioning: Ideal for time series data. A flavor of horizontal partitioning.



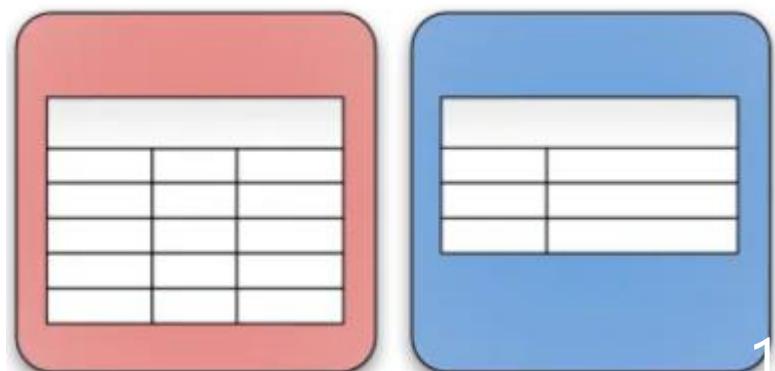
Horizontal partitioning

- **Rows of the same table** are stored in different locations (e.g. disks, nodes) according to some criteria.
- All partitions have the **same columns**.
- **Improves** query performance by scanning fewer rows.
- Remember that in DW the load is incremental (usually there are not updates).
 - New inserts do not produce table movements, can be just as simple as adding a partition.
 - Easy to delete by “time”.
- Criteria
 - Range, List, Hash, Partitioned Index
 - Composed: range + hash, list+hash



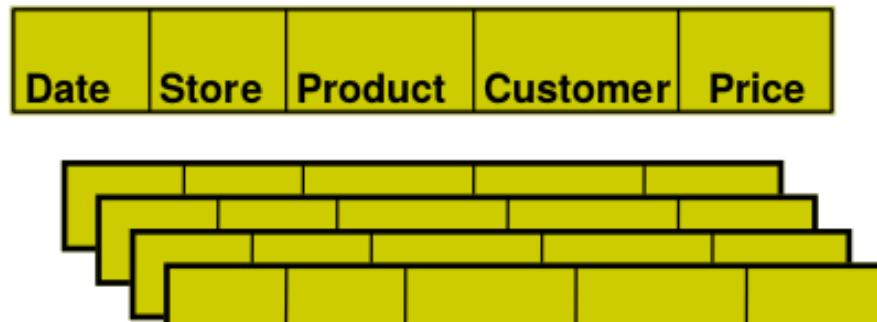
Vertical partitioning

- Divides a table into smaller tables with the **same rows** but a **subset of columns**.
- Not only for multidimensional db
- For **DWH**:
 - A number of operations are **aggregations by columns**
 - **Only columns affected** by operation are retrieved
 - **Compression** of columns is very efficient
 - Text columns without predefined size
 - Some columns may change more than others
 - **Improves performance** for OLAP queries by isolating frequently accessed columns..

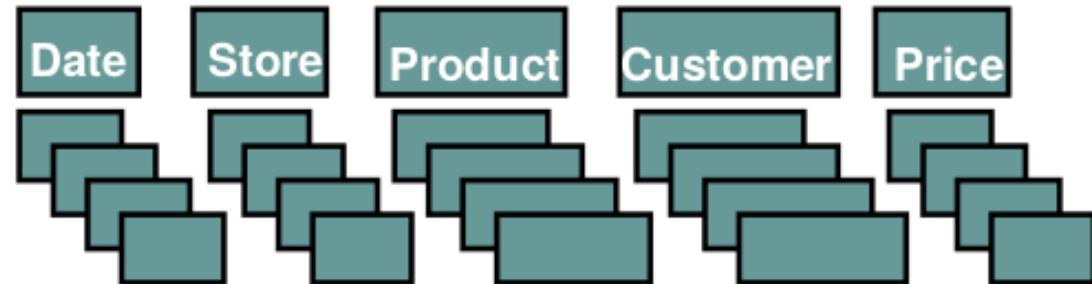


Columnar databases

row-store



column-store



Pros:

- easy to add/modify a record

Cons

- might read in unnecessary data

Pros:

- only need to read in relevant data
- Column compression easy and efficient (light algorithms)
- Aggregated queries very fast.

Cons:

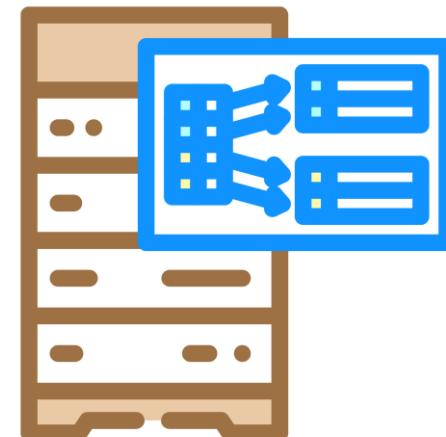
- tuple writes require multiple accesses

Note: We don't elaborate more here because you see them in other course.

Index: secondary data structure for arbitrary access and efficient

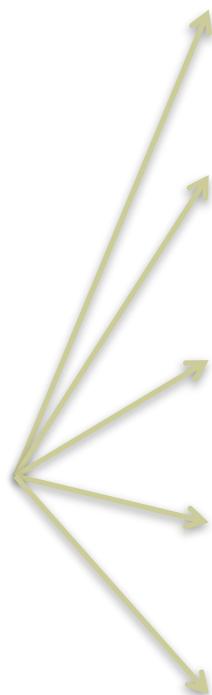
- **Characteristics:**

- **Content:**
 - Value of the attribute indexed
 - Address to the storage
- **Sorted entries** according to the value for efficient searches
- **Small size:** usually fits in one data block
- Some types: B-Trees, Hash
 - Specific for OLAP: Bitmap, Join Index
- Used only for retrieving **few values**



Index: General concept

cod A	Bloque
00	B4
11	B1
22	B2
33	B2
44	B4
55	B3
77	B3
88	B1
99	B5



Bloque	cod A	nombre	agencia	fechaNaci m
B1	88	Fele Martínez	Glam	22/02/75
	11	Najwa Nimri	Actors	14/02/72
B2	33	Nancho Novo	Rol	17/09/78
	22	Santiago Segura	Amiguete s	17/07/65
B3	77	Luis Tosar	BCN	13/10/71
	55	Candela Peña	Actors	14/07/73
B4	00	Maribel Verdú	Glam	02/10/70
	44	Penélope Cruz	BCN	28/04/74
B5	99	Javier Bardem	BCN	01/03/69

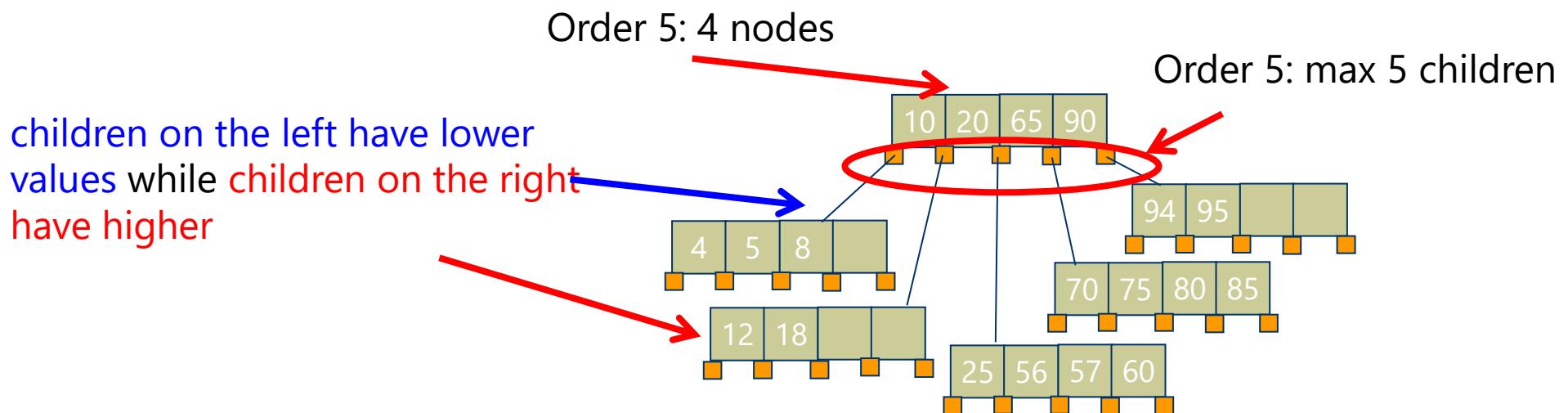
Just one
block

B-Tree Index

Deep balanced tree: all leaf nodes at the same depth. Each path from the root to a leaf has the same length.

N-order:

Each node can have up to n children and store n-1 keys.



Bitmap Index

Index on a particular column

Each value in the column has a bit vector: bit-op is fast

The **length** of the bit vector: **# of records** in the base table

The i-th bit is set if the i-th row of the base table has the value for the indexed column

Cust	Region	Type	RecID	Asia	Europe	America	RecID	Retail	Dealer
C1	Asia	Retail	1	1	0	0	1	1	0
C2	Europe	Dealer	2	0	1	0	2	0	1
C3	Asia	Dealer	3	1	0	0	3	0	1
C4	America	Retail	4	0	0	1	4	1	0
C5	Europe	Dealer	5	0	1	0	5	0	1

Bitmap Index

Reduced response time for large classes of ad hoc queries: Flexible WHERE.

Reduced storage requirements compared to other indexing.

Fully indexing a large table with a traditional B-tree index can lead to structures several times larger than the data in the table.

Bitmap indexes are typically **only a fraction of the size** of the indexed data in the table.

Efficient maintenance during parallel DML and loads.

The advantages of using bitmap indexes are greatest for columns in which the ratio of the number of distinct values to the number of rows in the table is small (1:100).

A table with one million rows, a column with 10,000 distinct values is a candidate

Comparison, union and aggregation use bit arithmetic.

Not suitable for high cardinality domains: too many columns (better B-Tree)

Often support compression

[Bitmap] Join Index

It materializes relational join in JI file and speeds up relational join

Join index: $\text{JI}(R\text{-id}, S\text{-id}) \text{ where } R(R\text{-id}, \dots) \text{ JOIN } S(S\text{-id}, \dots)$. R-fact table, S-dimension.

Better performance as it precomputes join results and stores them into bitmap format.

In DWH: join between facts and dimensions.

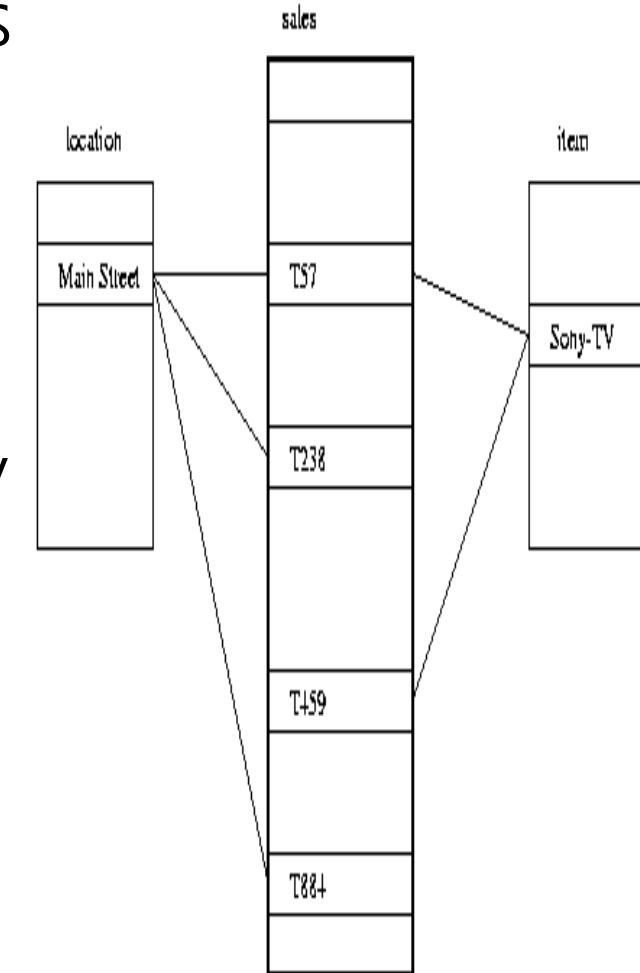
Equi-inner join between PK and FK

E.g. fact table: Sales and two dimensions city and product

A join index on city maintains for each distinct city a list of R-IDs of the tuples recording the Sales in the city

Join indices can span multiple dimensions

NOTE: Related to denormalization. They precompute join results and reduce the need for real-time joins.



Compression

Used in **multidimensional** DBMS.

Saving **disk space**, increasing **memory efficiency**, and improving **query performance** by reducing the amount of data read from disk (multidimensional databases are usually massive).

Overhead for updating, deleting, and processing in general. Data needs to be uncompressed, modified and compressed, leading to **overload during write operations**. In DWH environments, write operations are **less frequent**.

Hybrid Columnar Compression: groups of rows are stored in columnar format, with the values for a given column stored and compressed together.

Storing column data together, with the same data type and similar characteristics, drastically **increases the storage savings** achieved from compression.

Materialized Views

Materialized views are query results that have been **stored in advance** so long-running calculations are not necessary when you actually execute your SQL statements.

They act like tables and improve performance in complex queries, especially those involving joins and aggregations.

Useful for aggregates.

Problem:

- **Resources** consumption
- Refreshing policies: **on demand, on commit.**

Data cube can be viewed as a **lattice of cuboids**

- The bottom-most cuboid, with the most detailed data , is the **base cuboid**
- The top-most cuboid (**apex**) contains only one cell
- How many cuboids in an n-dimensional cube with L levels?

$$T = \prod_{i=1}^n (L_i + 1)$$

Materialization of data cube

- Materialize every cuboid (**full materialization**), none (**no materialization**, every query is computed from the base cuboid), or some (**partial materialization**)
- Selection of which cuboids to materialize
 - Based on size, sharing of cuboids between queries, access frequency, etc.

The “Compute Cube” Operator

Cube definition and computation in DMQL

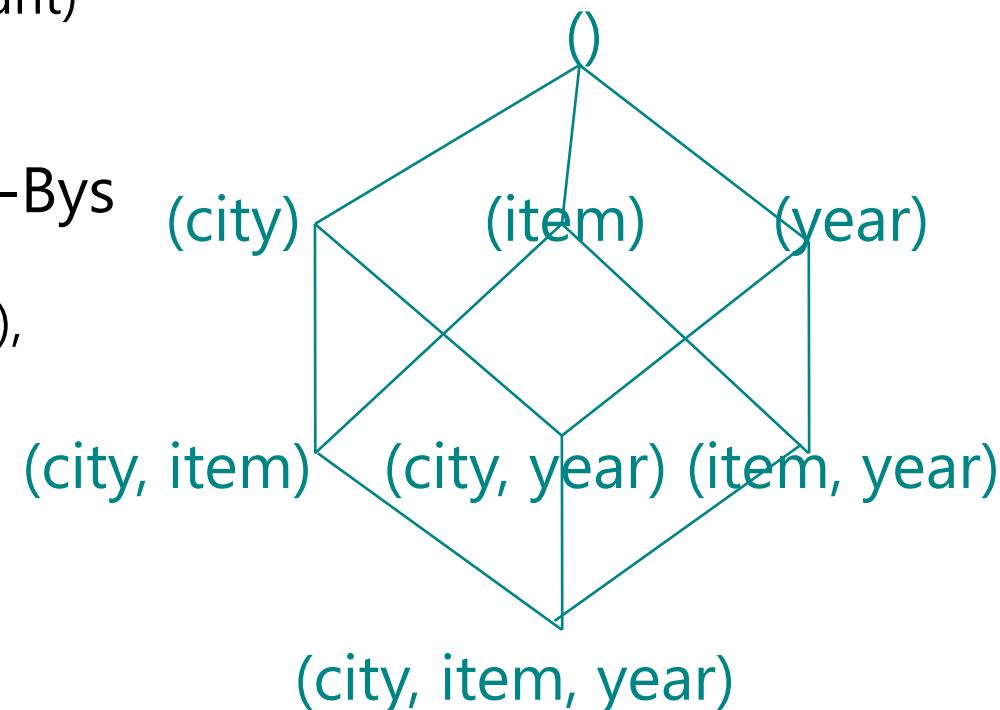
- define cube sales [item, city, year]: sum (sales_in_dollars)
- compute cube sales

Transform it into a SQL-like language (with a new operator cube by, introduced by Gray et al.'96)

```
SELECT item, city, year, SUM (amount)  
FROM SALES  
CUBE BY item, city, year
```

Need compute the following Group-Bys

- (item, city, year),
- (item,city),(item, year), (city, year),
- (item), (city), (year)
- ()



Determine which operations should be performed on the available cuboids

Transform drill, roll, etc. into corresponding SQL and/or OLAP operations, e.g., dice = selection + projection

Determine which materialized cuboid(s) should be selected for OLAP op.

Let the query to be processed be on {brand, province_or_state} with the condition "year = 2004", and there are 4 materialized cuboids available:

con country no puedes llegar a provincia o estado. Surven todos menos el 2

- 1) {year, item_name, city}
- 2) {year, brand, country}
- 3) {year, brand, province_or_state}
- 4) {item_name, province_or_state} where year = 2004

ROLAP MOLAP y OLAP SABER EN QUE CONSISTEN PREGUNTA BREVE
Particionamiento
Indexado importante
join bitmap y tal saber en consiste
cuboides

Which should be selected to process the query?

Explore **indexing structures** and **sparse vs. dense** array structs in MOLAP

tipo de examen:
alguna de tipo test
alguna de conceptos (uno o dos parrafos)
alguna un poco mas largo (2 o 3 parrafos)
Suele poner un alternativa en alguna pregunta.



A data lakehouse represents the synergy between a data warehouse and a data lake, offering the scalability and flexibility of data lakes with the ACID transactions and structure of data warehouses.

- **Storage Layers:**
 - **Raw Data Layer:** Stores unstructured or semi-structured data as it is ingested (e.g., log files, JSON, or Parquet).
 - **Curated Data Layer:** Stores cleaned and transformed data, often partitioned by time or domain to improve queries.
- **Query Optimization:** Indexing and data skipping techniques (e.g., Apache Parquet's predicate pushdown).
- **Columnar Formats:** storage formats like **Parquet** and **ORC** for efficient query performance.



- **Transaction Support:** Ensures ACID compliance for structured queries and updates, using technologies like **Delta Lake** (Databricks) and **Apache Iceberg**.
- **Partitioning and Compaction:** Strategies for partitioning data across file systems and regular data compaction to improve performance.
- **Serverless architectures** (scalable, on-demand processing of large datasets without infrastructure management). E.g. Google BigQuery, Amazon Redshift Spectrum.
- **Real-time analytics** (very fast query responses for low latency applications). E.g. Apache Pinot and ClickHouse.
- Combines **Data Lakes** (+ scalability for raw, unstructured data) and **Warehouses** (+ structure + ACID) for hybrid data management.

ROLAP:

- ClickHouse, Apache Pinot, Google BigQuery, Snowflake, Amazon Redshift

HOLAP:

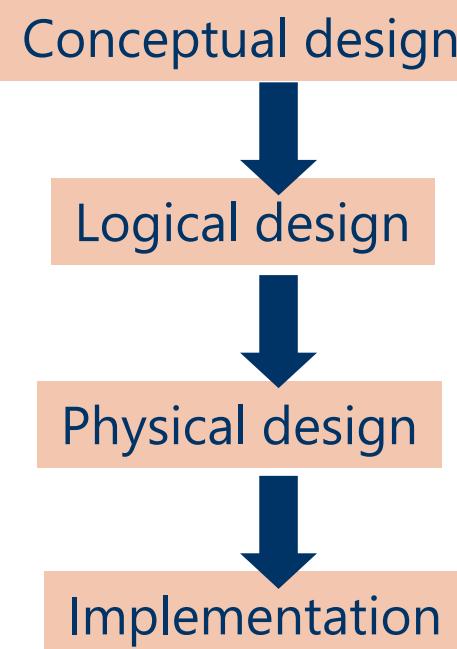
- Apache Druid, Apache Kylin, Atscale,

MOLAP:

Microsoft Azure Analysis Services, Kyvos Insights, Oracle Essbase,

Business intelligence

Unit 2 – Datawarehouse and OLAP
S2-4 – Physical design



Efficiency in storage
Efficiency in queries

Storage model (ROLAP,
MOLAP, HOLAP)





Objective: improve performance

Focus on **improving query performance and scalability.**

Balance between storage efficiency and fast retrieval.

Architecture: **ROLAP, MOLAP, HOLAP**

Storage strategy and query optimization

Denormalization (ROLAP, HOLAP)

Partitioning (ALL)

Index (ALL)

Compression (ALL)

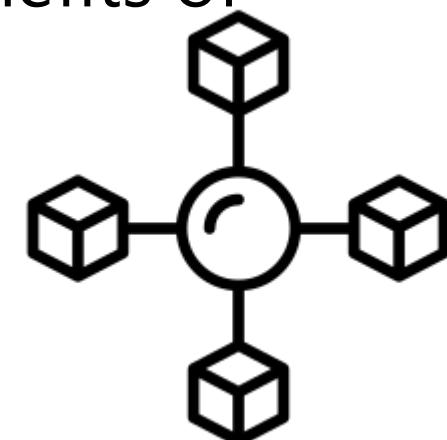
Materialized Views (ROLAP, HOLAP)

Big data challenges



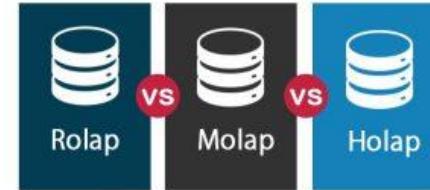
Different physical storage and query structures

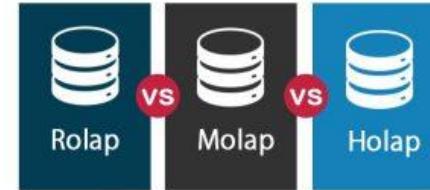
- Multidimensional OLAP (**MOLAP**): DBMS built specifically for data analysis. Data is stored in cubes to improve multidimensional queries.
- Relational OLAP (**ROLAP**) Relational DBMS, which can include both traditional row-based and modern columnar databases.
- Hybrid systems: **HOLAP**. Combines benefits of ROLAP and MOLAP for optimized performance.



ROLAP:

- The DWH is built **on top of relational DBMS**, often using columnar storage to improve performance with big datasets.
- RDBMS vendors provide OLAP-specific extensions, such as **window functions** (e.g., RANK(), SUM() OVER()), and **grouping sets** (GROUPING SETS, CUBE, ROLLUP) or integrated data mining tools: Oracle OLAP, Microsoft SSAS ROLAP, PostgreSQL with OLAP extensions.
- Typically uses a **star schema** or **snowflake** schema with fact and dimension tables.





ROLAP:

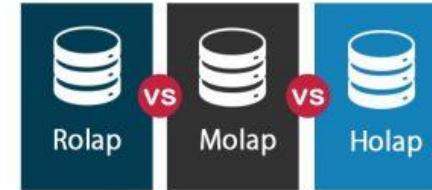
Advantages:

- Scalability: There is no limit on the amount of data.
- Proven technology: Functionality already available.
- Separation storage from OLAP processing.
- Cost-effective

Disadvantages:

- Performance overhead: queries on normalized db, usually with multiple joins. Complex multidimensional queries usually slower than in MOLAP systems.
- Limited functionality for SQL. For example, to perform complex calculations or hierarchical queries.
- No pre-aggregated data, slowing response times.
- Potential locking and concurrency problems in busy systems.

MOLAP



- Consist of physically storing data in multidimensional structures (**cubes**) so that the external representation *matches* the internal representation.
- It allows **fast retrieval** of pre-aggregated data.
- BD complexity is **hidden** from the users.
- Analysis is done on **aggregated data** and precalculated metrics or indicators.
- OLAP Engine: The **MOLAP engine** processes and responds to multidimensional queries with very low latency due to pre-aggregated data.
- Specific functionality: Array data structures (proprietary formats), query optimization , data compaction.

MOLAP:

Advantages:

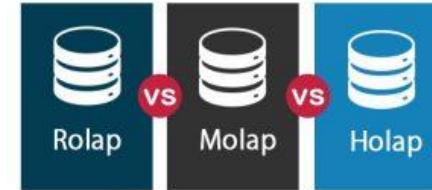
- Fast Performance for small and medium size datasets: slide & dice specific operations.
- Hierarchical analysis: easy to navigate through dimensions.
- Complex calculations are pre-generated.
- Easier to use, even for inexperienced users.



MOLAP:

Disadvantages:

- Limited size: Problems with large datasets and many dimensions.
- Latency when loading and pre-aggregate: it can take time.
- Storage overhead: pre-aggregated data can take a lot of storage, specially for large datasets.
- New investment: this technology is not usually present in enterprises. Also usually proprietary.
- Limited Flexibility: Less flexible when new dimensions or hierarchies are introduced, requiring a cube reprocessing.





MOLAP: Improving performance

- **Partial Cube Aggregation:** Selectively pre-aggregate dimensions or measures that are queried most frequently, reducing cube processing time.
- **Incremental Cube Refresh:** Use incremental refresh to update only the new data, minimizing downtime and latency.
- **Sparse Data Handling:** Manage sparse data more efficiently by compressing and storing only non-null values.
- **Advanced Partitioning:** Partitioning cubes by key dimensions (e.g., time) allows for faster access and easier management of data over time.

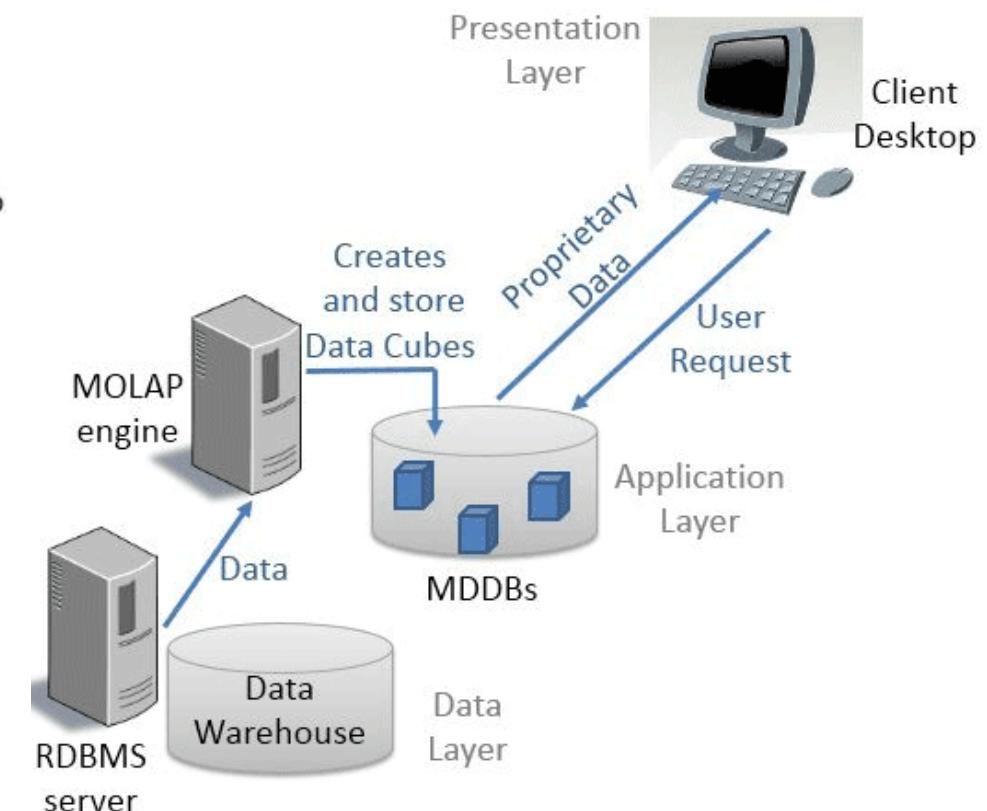
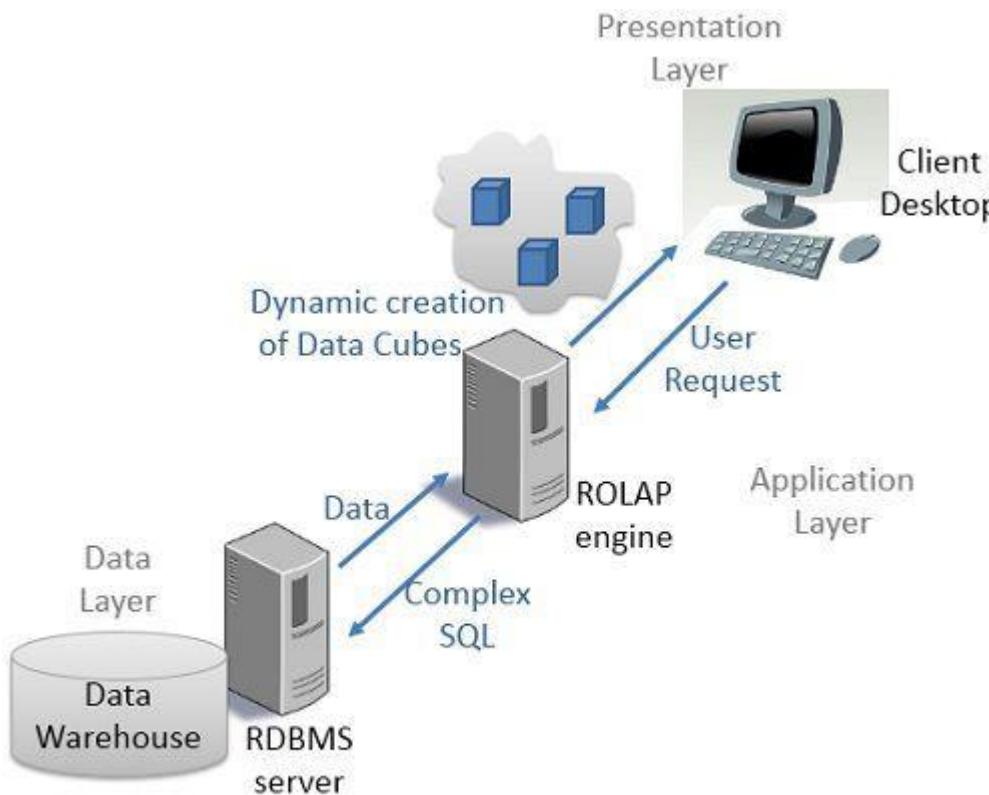


Image Source:EDUCBA

Database partitioning

Replace large tables with **more manageable** smaller subtables

- Allows **parallel and distributed** execution
- Partition **prunning** and **partition wise** joins
- **Partition data processing**: ETL, refreshing, backup, indexing.

Horizontal

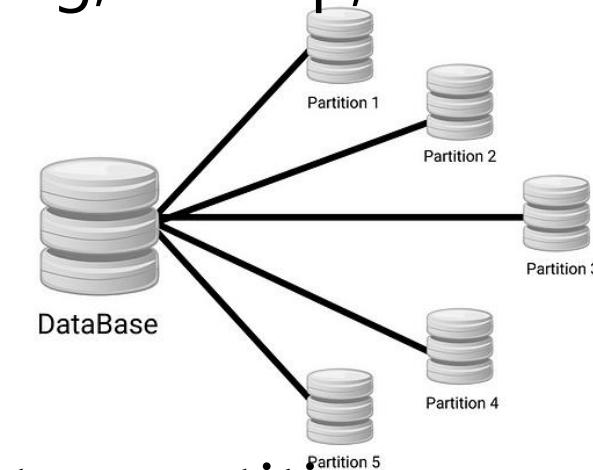
- Rows of a table are separated

Vertical

- Columns of a table are separated

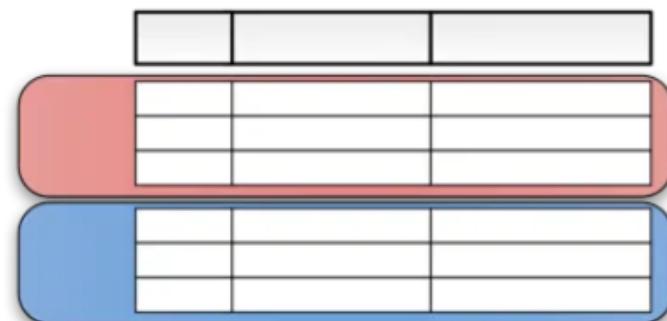
Dynamic Partitioning: automatically creates partitions based on incoming data, without the need for predefined ranges or categories.

Time-based Partitioning: Ideal for time series data. A flavor of horizontal partitioning.



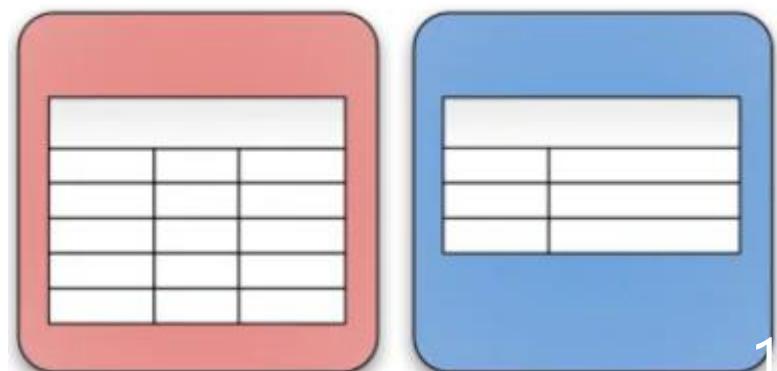
Horizontal partitioning

- **Rows of the same table** are stored in different locations (e.g. disks, nodes) according to some criteria.
- All partitions have the **same columns**.
- **Improves** query performance by scanning fewer rows.
- Remember that in DW the load is incremental (usually there are not updates).
 - New inserts do not produce table movements, can be just as simple as adding a partition.
 - Easy to delete by “time”.
- Criteria
 - Range, List, Hash, Partitioned Index
 - Composed: range + hash, list+hash



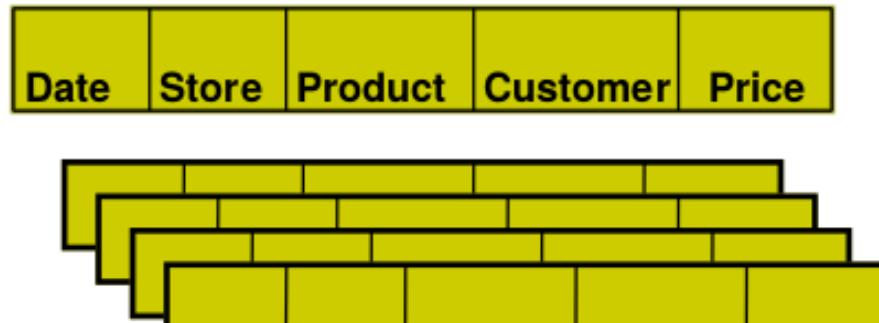
Vertical partitioning

- Divides a table into smaller tables with the **same rows** but a **subset of columns**.
- Not only for multidimensional db
- For **DWH**:
 - A number of operations are **aggregations by columns**
 - **Only columns affected** by operation are retrieved
 - **Compression** of columns is very efficient
 - Text columns without predefined size
 - Some columns may change more than others
 - **Improves performance** for OLAP queries by isolating frequently accessed columns..

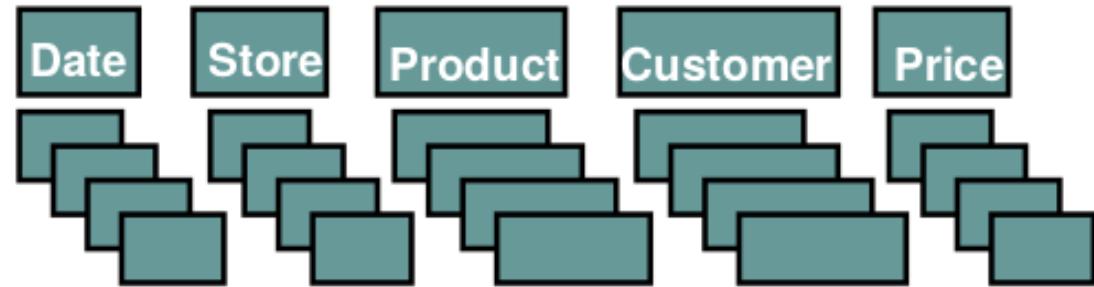


Columnar databases

row-store



column-store



Pros:

- easy to add/modify a record

Cons

- might read in unnecessary data

Pros:

- only need to read in relevant data
- Column compression easy and efficient (light algorithms)
- Aggregated queries very fast.

Cons:

- tuple writes require multiple accesses

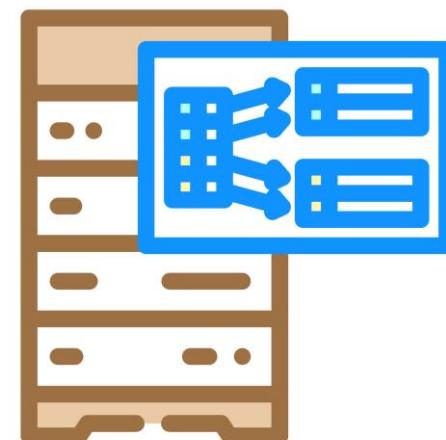
Note: We don't elaborate more here because you see them in other course.

http://cs-www.cs.yale.edu/homes/dna/talks/Column_Store_Tutorial_VLDB09.pdf

Index: secondary data structure for arbitrary access and efficient

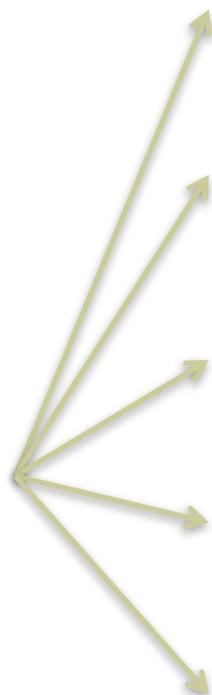
- **Characteristics:**

- **Content:**
 - Value of the attribute indexed
 - Address to the storage
- **Sorted entries** according to the value for efficient searches
- **Small size:** usually fits in one data block
- Some types: B-Trees, Hash
 - Specific for OLAP: Bitmap, Join Index
- Used only for retrieving **few values**



Index: General concept

cod A	Bloque
00	B4
11	B1
22	B2
33	B2
44	B4
55	B3
77	B3
88	B1
99	B5



Bloque	cod A	nombre	agencia	fechaNaci m
B1	88	Fele Martínez	Glam	22/02/75
	11	Najwa Nimri	Actors	14/02/72
B2	33	Nancho Novo	Rol	17/09/78
	22	Santiago Segura	Amiguete s	17/07/65
B3	77	Luis Tosar	BCN	13/10/71
	55	Candela Peña	Actors	14/07/73
B4	00	Maribel Verdú	Glam	02/10/70
	44	Penélope Cruz	BCN	28/04/74
B5	99	Javier Bardem	BCN	01/03/69

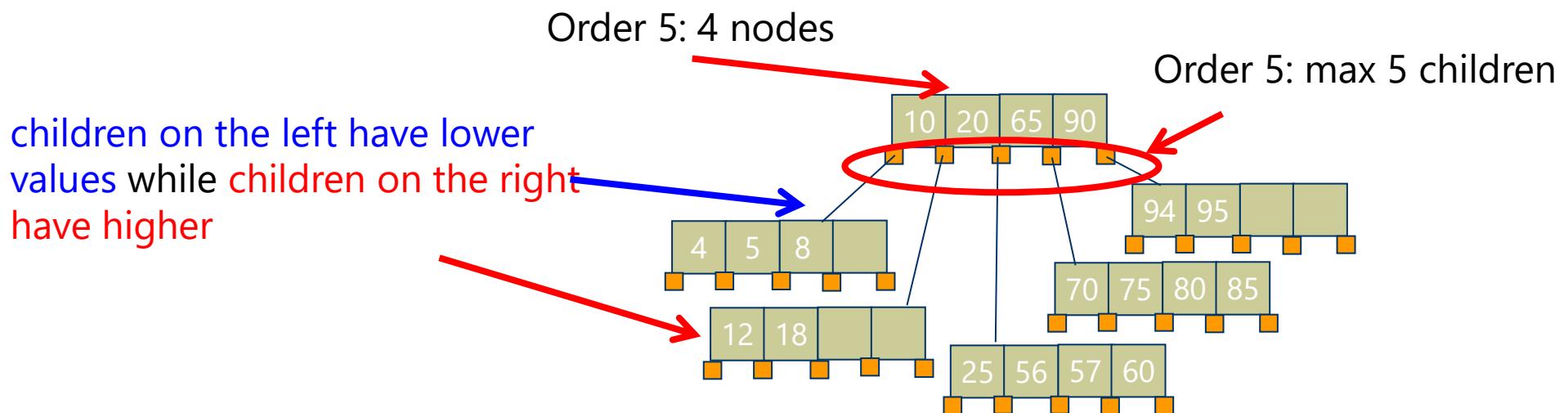
Just one
block

B-Tree Index

Deep balanced tree: all leaf nodes at the same depth. Each path from the root to a leaf has the same length.

N-order:

Each node can have up to n children and store n-1 keys.



Bitmap Index

Index on a particular column

Each value in the column has a bit vector: bit-op is fast

The **length** of the bit vector: **# of records** in the base table

The i-th bit is set if the i-th row of the base table has the value for the indexed column

Cust	Region	Type	RecID	Asia	Europe	America	RecID	Retail	Dealer
C1	Asia	Retail	1	1	0	0	1	1	0
C2	Europe	Dealer	2	0	1	0	2	0	1
C3	Asia	Dealer	3	1	0	0	3	0	1
C4	America	Retail	4	0	0	1	4	1	0
C5	Europe	Dealer	5	0	1	0	5	0	1

Reduced response time for large classes of ad hoc queries: Flexible WHERE.

Reduced storage requirements compared to other indexing.

Fully indexing a large table with a traditional B-tree index can lead to structures several times larger than the data in the table.

Bitmap indexes are typically **only a fraction of the size** of the indexed data in the table.

Efficient maintenance during parallel DML and loads.

The advantages of using bitmap indexes are greatest for columns in which the ratio of the number of distinct values to the number of rows in the table is small (1:100).

A table with one million rows, a column with 10,000 distinct values is a candidate

Comparison, union and aggregation use bit arithmetic.

Not suitable for high cardinality domains: too many columns (better B-Tree)

Often support compression

[Bitmap] Join Index

It materializes relational join in JI file and speeds up relational join

Join index: **JI(R-id, S-id)** where R (R-id, ...) JOIN S (S-id, ...). R-fact table, S-dimension.

Better performance as it **precomputes join results** and stores them into bitmap format.

In DWH: join between facts and dimensions.

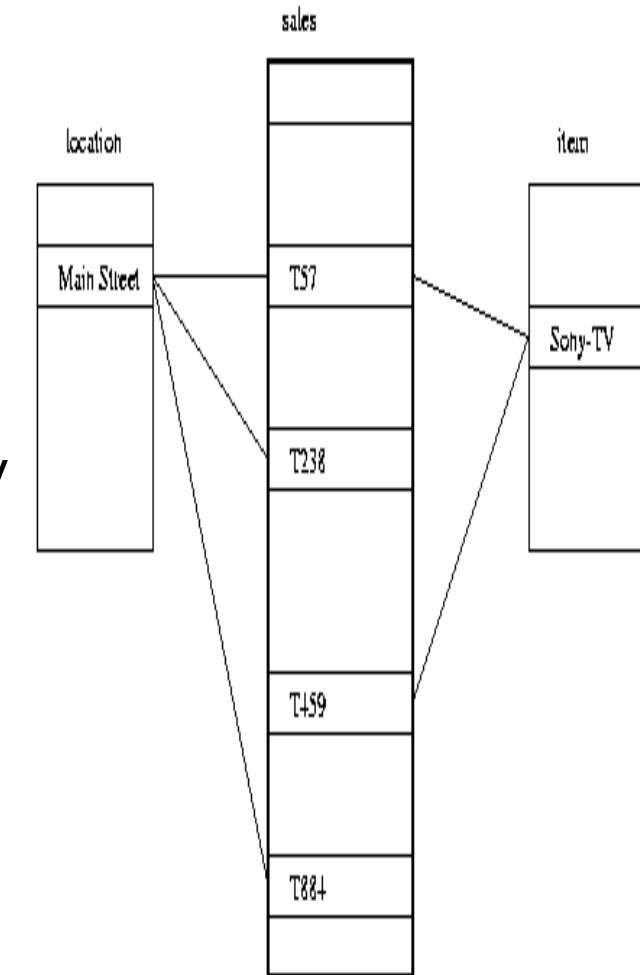
Equi-inner join between PK and FK

E.g. fact table: Sales and two dimensions city and product

A join index on city maintains for each distinct city a list of R-IDs of the tuples recording the Sales in the city

Join indices can span multiple dimensions

NOTE: Related to denormalization. They precompute join results and reduce the need for real-time joins.



Compression

Used in **multidimensional** DBMS.

Saving **disk space**, increasing **memory efficiency**, and improving **query performance** by reducing the amount of data read from disk (multidimensional databases are usually massive).

Overhead for updating, deleting, and processing in general. Data needs to be uncompressed, modified and compressed, leading to **overload during write operations**. In DWH environments, write operations are **less frequent**.

Hybrid Columnar Compression: groups of rows are stored in columnar format, with the values for a given column stored and compressed together.

Storing column data together, with the same data type and similar characteristics, drastically **increases the storage savings** achieved from compression.

Materialized Views

Materialized views are query results that have been **stored in advance** so long-running calculations are not necessary when you actually execute your SQL statements.

They act like tables and improve performance in complex queries, especially those involving joins and aggregations.

Useful for aggregates.

Problem:

- **Resources** consumption
- Refreshing policies: **on demand, on commit.**

Data cube can be viewed as a **lattice of cuboids**

- The bottom-most cuboid, with the most detailed data , is the **base cuboid**
- The top-most cuboid (**apex**) contains only one cell
- How many cuboids in an n-dimensional cube with L levels?

$$T = \prod_{i=1}^n (L_i + 1)$$

Materialization of data cube

- Materialize every cuboid (**full materialization**), none (**no materialization**, every query is computed from the base cuboid), or some (**partial materialization**)
- Selection of which cuboids to materialize
 - Based on size, sharing of cuboids between queries, access frequency, etc.

The “Compute Cube” Operator

Cube definition and computation in DMQL

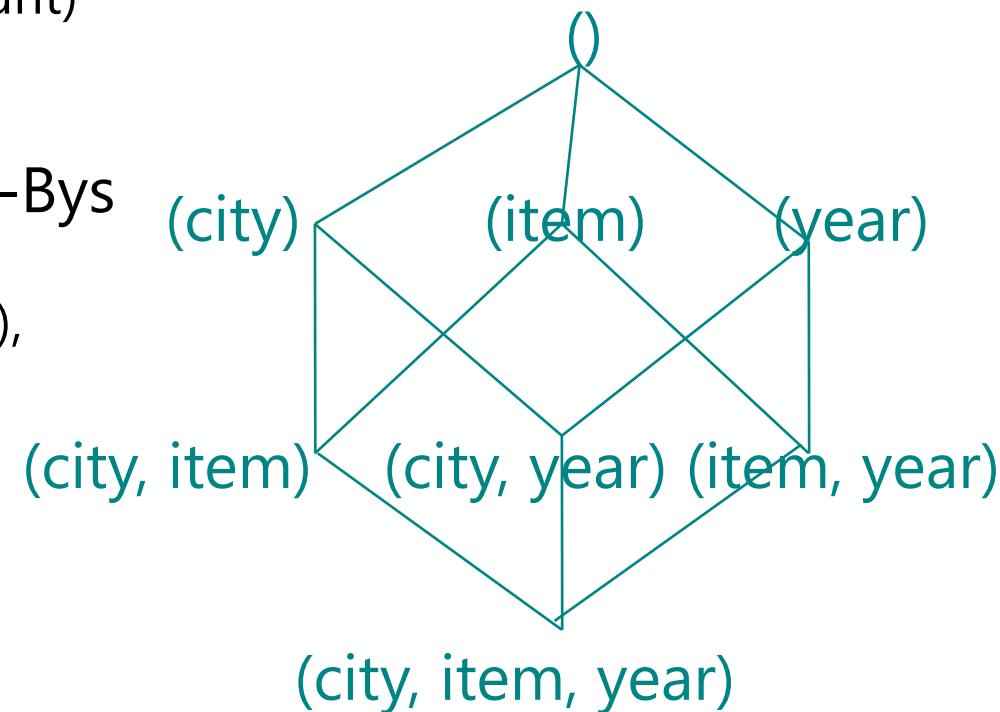
- define cube sales [item, city, year]: sum (sales_in_dollars)
- compute cube sales

Transform it into a SQL-like language (with a new operator cube by, introduced by Gray et al.'96)

```
SELECT item, city, year, SUM (amount)  
FROM SALES  
CUBE BY item, city, year
```

Need compute the following Group-Bys

- (item, city, year),
- (item,city),(item, year), (city, year),
- (item), (city), (year)
- ()



Determine which operations should be performed on the available cuboids

Transform drill, roll, etc. into corresponding SQL and/or OLAP operations, e.g., dice = selection + projection

Determine which materialized cuboid(s) should be selected for OLAP op.

Let the query to be processed be on {brand, province_or_state} with the condition "year = 2004", and there are 4 materialized cuboids available:

- 1) {year, item_name, city}
- 2) {year, brand, country}
- 3) {year, brand, province_or_state}
- 4) {item_name, province_or_state} where year = 2004

Which should be selected to process the query?

Explore **indexing structures** and **sparse vs. dense** array structs in MOLAP

DWH- Model for the flow of data from operational systems into decision support.

DWH+Big Data-> High maintenance costs, poor flexibility and difficult scalability.

DWH support **BI** and **Reporting**, but struggle with **ML**.

Data Lakes offer raw data repositories in multiple formats. Data processing frameworks: Hadoop (initial) -> Apache Spark (much more popular today).

DWH+Data Lakes: Multipurpose applications, such as SQL analytics, real-time monitoring, and ML. Cons: Additional complexity and multiple copies of data.

Scalable raw data repository fitted for Big Data (structured, semi- and unstructured data).

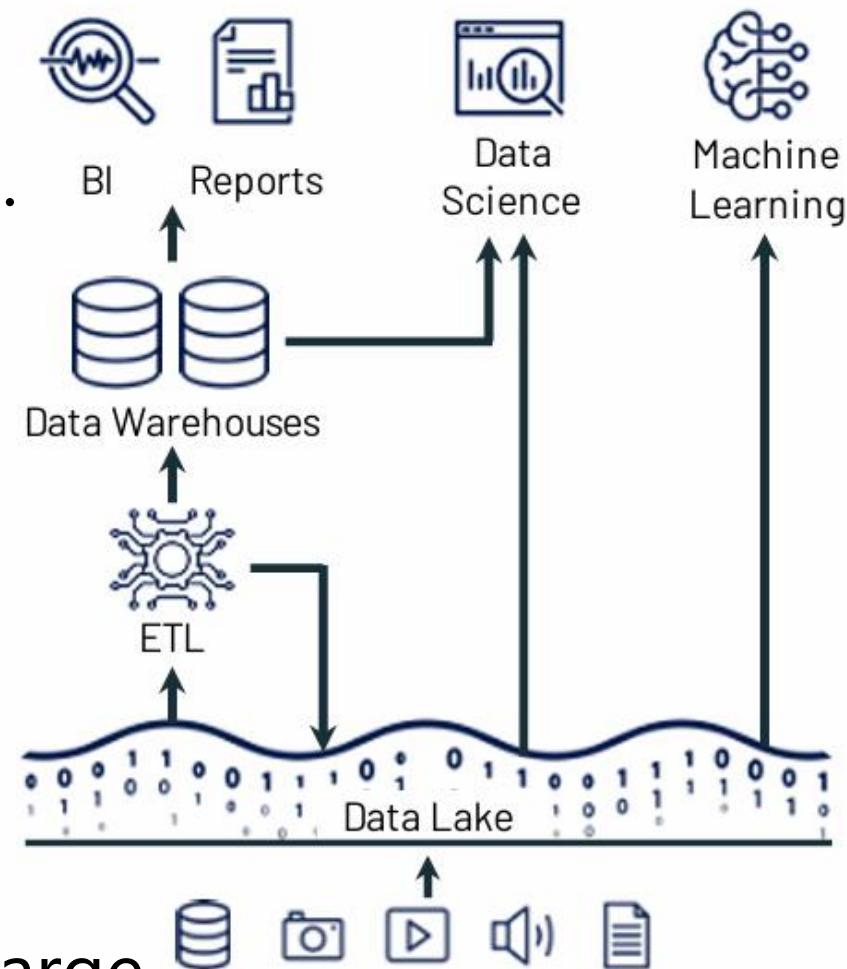
Raw data ingestion. All data types.

No ACID, **BASE** instead.

Data quality and governance:

Ideally-> Clean data repositories
But, often: **data swamps**.

Challenges: High cost, data redundancy, complex architecture, security, difficult managements of large metadata, ...

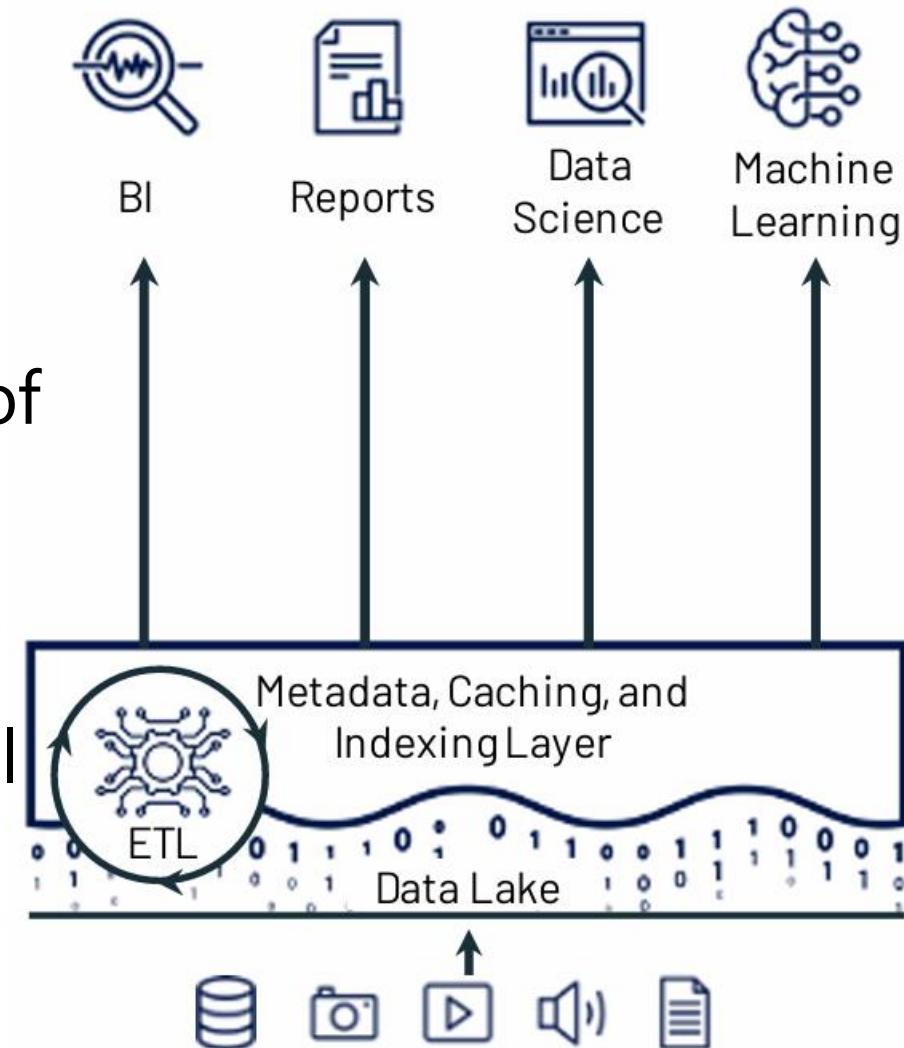


Data Lakehouses

A **data lakehouse** represents the synergy between a data warehouse and a data lake, offering the scalability and flexibility of data lakes with the ACID transactions and structure of data warehouses.

Simplifies data management.
Provides a **single platform** for all forms of data analysis.

Core OS technologies: Delta Lake, Apache Iceberg, Apache Hudi.



Data Lakehouses

ACID transactions.

Access to **early versions** of data

Improved **metadata** management.

An **unique architecture** for all applications.

Lower costs.

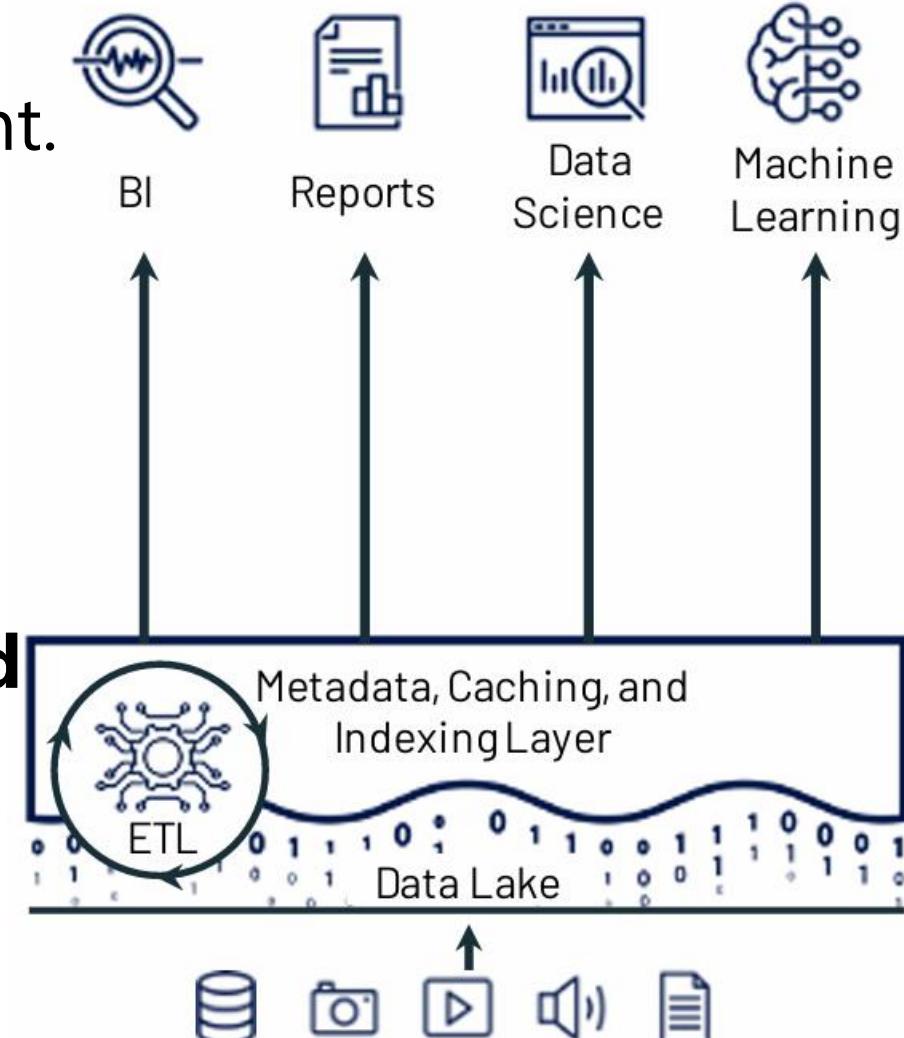
Reduced data redundancy.

Direct connection by BI Tools.

Uses mostly **open standards and open formats**, suitable for batch and streaming data.

Enables **BI in all the data**.

The query engine is **connected directly** to the data lake (reduced ETL).





DEMO

Lakehouse

<https://github.com/delta-io/delta-docker>

```
git clone https://github.com/delta-io/delta-docker.git
```

```
cd delta-docker
```

```
Start Docker desktop
```

```
In powershell, run bash.
```

```
docker build -t delta_quickstart -f Dockerfile_delta_quickstart .
```

```
docker run --name delta_quickstart --rm -it --entrypoint bash
```

```
delta_quickstart
```

Run a container from the image with a JupyterLab entrypoint:

```
docker run --name delta_quickstart --rm -it -p 8888-8889:8888-  
8889 delta_quickstart
```

Use the http to connect to Jupyter notebook:

<http://127.0.0.1:8888/lab>

Using python3 pykernel, select quickstart.ipynb

Delta-Lake

Write a Spark DataFrame to a Delta Lake table

```
data = spark.range(0, 5)

(data
    .write
    .format("delta")
    .save("/tmp/delta-table")
)
```

Read the above Delta Lake table to a Spark DataFrame and display the DataFrame

```
df = (spark
      .read
      .format("delta")
      .load("/tmp/delta-table")
      .orderBy("id")
    )
df.show()
```

```
+---+ part-00001-d39859ef-7163-47ac-b011-bacc9ccf04d0-c000.snapy.parquet  
| id | part-00001-d748a1c1-9a27-48b0-92cb-ac7e9b1d5468-c000.snapy.parquet  
+---+ part-00001-d78e6bbc-ce8f-4b0e-9d82-b690aa34c2cd-c000.snapy.parquet  
part-00001-d797bb15-3236-4346-9ba4-6cb050ecdd94-c000.snapy.parquet  
part-00001-d9e44ad4-8a6d-4658-bf10-88488c768f0ff-c000.snapy.parquet  
| 0 | part-00001-dac72292-a701-424a-8a8a-81592206c08e-c000.snapy.parquet  
| 1 | part-00001-db87183b-50df-4f73-a0cd-a7e616f03e99-c000.snapy.parquet  
| 2 | part-00001-dc625a90-7e1b-4736-84e6-cc069e1b3415-c000.snapy.parquet  
| 3 | part-00001-df79a539-8255-404c-bc86-4778b3b57558-c000.snapy.parquet  
| 4 | part-00001-e56019b5-22f0-4b06-8523-6584d6094773-c000.snapy.parquet  
part-00001-e87cf5df-7821-45ab-b9d7-f6fe6c369d0ff-c000.snapy.parquet  
part-00001-ebc17974-a298-4e66-ae86-5221a0197b2f-c000.snapy.parquet  
part-00001-ed0a96db-7ea4-451d-b635-3fcccb8c2b6e-c000.snapy.parquet  
part-00001-f08d8d55d-264e-42cc-95d4-56109857bf55-c000.snapy.parquet  
+---+ part-00001-f13ea236-0cd2-4cc4-adc8-37212977b40b-c000.snapy.parquet  
part-00001-f477c049-933f-4e03-9dd4-1fdceead3318-c000.snapy.parquet  
part-00001-f52d9aa-8de5-44ce-95c8-7f75d6aa357e-c000.snapy.parquet  
part-00001-f5a33748-f0d5-4673-bfc8-5b8b1760f81bb-c000.snapy.parquet  
part-00001-fb9e43e2-afa9-405d-9bf9-d5c769633593f-c000.snapy.parquet  
part-00002-29570c39-6caa-4cc8-88e6-bdb111e1629c-c000.snapy.parquet  
part-00002-6ad3429e-cfb5-4769-af72-b33a2013412c-c000.snapy.parquet  
part-00002-85144acf-782a-44b2-bde8-0d18462afeff-c000.snapy.parquet  
part-00002-8ea9cc0b-6640-465f-b6fc-c468208dc14f-c000.snapy.parquet  
part-00002-ba367c14-528c-4c66-904c-2143bd4ad441-c000.snapy.parquet  
part-00002-c77e47d2-8f7d-4afb-bd2a-6539a1a78d2b-c000.snapy.parquet  
part-00002-c89e9197-1d9a-4b58-84ef-4ae62266e264-c000.snapy.parquet  
part-00002-chf8e49b-810a-421c-b36b-a1bf2a216e24-c000.snapy.parquet
```

Overwrite a Delta Lake table

Overwrite the Delta Lake table written in the above step

```
data = spark.range(5, 10)

(data
    .write
    .format("delta")
    .mode("overwrite")
    .save("/tmp/delta-table")
)
```

Delta-Lake

Showcase `update` feature of Delta Lake and display the resulting DataFrame

```
from delta.tables import *
from pyspark.sql.functions import *

delta_table = DeltaTable.forPath(spark, "/tmp/delta-table")

# Update every even value by adding 100 to it
(delta_table
 .update(
   condition = expr("id % 2 == 0"),
   set = { "id": expr("id + 100") }
 )
)

(delta_table
 .toDF()
 .orderBy("id")
 .show()
)
```

Showcase `merge` feature of Delta Lake and display the resulting DataFrame

```
# Upsert (merge) new data
new_data = spark.range(0, 20)

(delta_table.alias("old_data")
 .merge(
   new_data.alias("new_data"),
   "old_data.id = new_data.id"
 )
 .whenMatchedUpdate(set = { "id": col("new_data.id") })
 .whenNotMatchedInsert(values = { "id": col("new_data.id") })
 .execute()
)

(delta_table
 .toDF()
 .orderBy("id")
 .show()
)
```

Showcase `delete` feature of Delta Lake and display the resulting DataFrame

```
# Delete every even value
(delta_table
 .delete(
   condition = expr("id % 2 == 0")
 )
)

(delta_table
 .toDF()
 .orderBy("id")
 .show()
)
```

Delta-Lake

Display the entire history of the above Delta Lake table

```
# get the full history of the table
delta_table_history = (DeltaTable
    .forPath(spark, "/tmp/delta-table")
    .history()
)

(delta_table_history
    .select("version", "timestamp", "operation", "operationParameters", "operationMetrics", "engineInfo")
    .show()
)

+-----+-----+-----+-----+-----+
|version| timestamp|operation| operationParameters| operationMetrics| engineInfo|
+-----+-----+-----+-----+-----+
| 4|2024-11-18 09:06:....| MERGE|{predicate -> ["(...|{numTargetRowsCop...|Apache-Spark/3.5....|
| 3|2024-11-18 09:06:....| DELETE|{predicate -> ["(...|{numRemovedFiles ...|Apache-Spark/3.5....|
| 2|2024-11-18 09:06:....| UPDATE|{predicate -> ["(...|{numRemovedFiles ...|Apache-Spark/3.5....|
| 1|2024-11-18 09:06:....| WRITE|{mode -> Overwrit...|{numFiles -> 6, n...|Apache-Spark/3.5....|
| 0|2024-11-18 09:05:....| WRITE|{mode -> ErrorIfE...|{numFiles -> 6, n...|Apache-Spark/3.5....|
+-----+-----+-----+-----+-----+
```

Time travel to the version 0 of the Delta Lake table using Delta Lake's history feature

```
df = (spark
    .read
    .format("delta")
    .option("versionAsOf", 0) # we pass an option `versionAsOf` with the required version number we are interested in
    .load("/tmp/delta-table")
    .orderBy("id")
)

df.show()

+---+
| id|
+---+
| 0|
| 1|
| 2|
| 3|
| 4|
+---+
```

Latest version of the Delta Lake table

```
df = (spark
    .read
    .format("delta")
    .load("/tmp/delta-table")
    .orderBy("id")
)

df.show()
```

id
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Delta-Lake

A little bit of Streaming

```
streaming_df = (spark
    .readStream
    .format("rate")
    .load()
)

stream = (streaming_df
    .selectExpr("value as id")
    .writeStream
    .format("delta")
    .option("checkpointLocation", "/tmp/checkpoint")
    .start("/tmp/delta-table")
)
```

```
stream2 = (spark
            .readStream
            .format("delta")
            .load("/tmp/delta-table")
            .writeStream
            .format("console")
            .start()
        )
```

```
00000000000000000000000000000001.json
:::::::::::
{"commitInfo":{"timestamp":1731920773177,"operation":"WRITE","operationParameters":{"mode":"Overwrite","partitionBy":[]},"readVersion":0,"isolationLevel":"Serializable","isBlindAppend":false,"operationMetrics":{"numFiles":6,"numOutputRows":5,"numOutputBytes":2686}, "engineInfo":"Apache-Spark/3.5.1 Delta-Lake/3.1.0","txId":"be8dc362-775c-4bafe9-d6de-a5bac8e4b7ed"}}

{"add":{"path": "/part-00001-d78ebbcc-ceef-4b0e-9d82-b690ae34c2cd-c000.snappy.parquet", "partitionValues":{}, "size":478, "modificationTime":1731920772801, "dataChange":true, "stats": {"("numRecords":1, "minValues": {"\\"id\\":5}, "maxValues": {"\\"id\\":5}, "nullCount": {"\\"id\\":0})"}}

{"add":{"path": "/part-0003-624fc4d-cb57-4595-b3ff-ac8f38a2be3d-c000.snappy.parquet", "partitionValues":{}, "size":478, "modificationTime":1731920772791, "dataChange":true, "stats": {"("numRecords":1, "minValues": {"\\"id\\":6}, "maxValues": {"\\"id\\":6}, "nullCount": {"\\"id\\":0})"}}

{"add":{"path": "/part-0004-86c3b683-3276-4e76-b567-1b9a86c1e4f-c000.snappy.parquet", "partitionValues":{}, "size":478, "modificationTime":1731920772791, "dataChange":true, "stats": {"("numRecords":1, "minValues": {"\\"id\\":7}, "maxValues": {"\\"id\\":7}, "nullCount": {"\\"id\\":0})"}}

{"add":{"path": "/part-0006-c763aa8-3d16-43dc-a8ca-95726c05acb6-c000.snappy.parquet", "partitionValues":{}, "size":478, "modificationTime":1731920772791, "dataChange":true, "stats": {"("numRecords":1, "minValues": {"\\"id\\":8}, "maxValues": {"\\"id\\":8}, "nullCount": {"\\"id\\":0})"}}

{"add":{"path": "/part-0007-48bb0cbd9-aae9-4b64-a0b5-e285d9e60b20-c000.snappy.parquet", "partitionValues":{}, "size":478, "modificationTime":1731920772801, "dataChange":true, "stats": {"("numRecords":1, "minValues": {"\\"id\\":9}, "maxValues": {"\\"id\\":9}, "nullCount": {"\\"id\\":0})"}}

{"remove":{"path": "/part-0004-7ebe48bc-2c83-4288-95ff-ae768968b371-c000.snappy.parquet", "deletionTimestamp":1731920773176, "dataChange":true, "extendedFileMetadata":true}}
```

```
Batch: 3
-----
+---+
| id|
+---+
| 11|
+---+



-----
Batch: 4
-----
+---+
| id|
+---+
| 12|
+---+



00000000000000001249.json      00000000000000001673.json
00000000000000001250.checkpoint.parquet 00000000000000001674.json
00000000000000001250.json      00000000000000001675.json
00000000000000001251.json      00000000000000001676.json
00000000000000001252.json      00000000000000001677.json
00000000000000001253.json      00000000000000001678.json
00000000000000001254.json      00000000000000001679.json
00000000000000001255.json      00000000000000001680.checkpoint.parquet
00000000000000001256.json      00000000000000001680.json
00000000000000001257.json      00000000000000001681.json
00000000000000001258.json      00000000000000001682.json
00000000000000001259.json      00000000000000001683.json
00000000000000001260.checkpoint.parquet 00000000000000001684.json
00000000000000001260.json      00000000000000001685.json
00000000000000001261.json      00000000000000001686.json
00000000000000001262.json      00000000000000001687.json
00000000000000001263.json      00000000000000001688.json
00000000000000001264.json      00000000000000001689.json
00000000000000001265.json      00000000000000001690.checkpoint.parquet
00000000000000001266.json      00000000000000001690.json
00000000000000001267.json      00000000000000001691.json
```

Delta-Lake

Import CSV

```
[3]: df = spark.read.format("csv").option("header", True).option("sep", ";").load("/opt/spark/examples/src/main/resources/people.csv")
df.show()
```

```
+-----+-----+
| name|age|    job|
+-----+-----+
|Jorge| 30|Developer|
| Bob| 32|Developer|
+-----+-----+
```

```
[4]: df.write.format("delta").save("tmp/people_delta")
!tree tmp/people_delta
```

```
24/11/18 11:54:57 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
[Stage 8:=====]          (41 + 8) / 50]
```

```
tmp/people_delta
└── delta_log
    └── 00000000000000000000.json
└── part-00000-ad0ac088-e4af-41f2-9013-83d18490580d-c000.snappy.parquet
```

```
1 directory, 2 files
```

```
[5]: spark.read.format("delta").load("tmp/people_delta").show()
```

```
+-----+-----+
| name|age|    job|
+-----+-----+
|Jorge| 30|Developer|
| Bob| 32|Developer|
+-----+-----+
```

Business intelligence

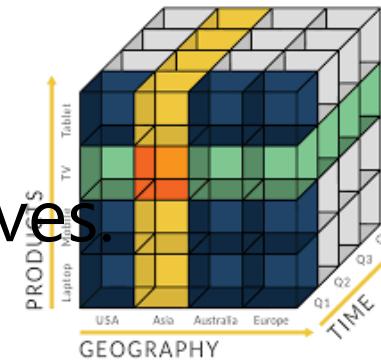
Unit 3 – Data exploitation. Query languages
and visualization
S3-1 – OLAP

OLAP tools

OLAP tools are used in BI for **analyzing multidimensional data** from multiple perspectives.

OLAP tools provide the user with a **multidimensional view** of data (multidimensional schema) for each activity that is being analyzed.
The user formulates queries to the OLAP tool selecting multidimensional attributes of this scheme **without knowing the internal structure** (physical schema) of the data warehouse.

The tool generates a corresponding **OLAP query** and sends it to the query management system (e.g. by means a SQL SELECT statement).



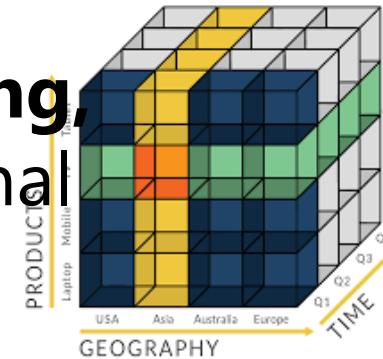
Multidimensional Analysis: OLAP enables **slicing, dicing, and drilling down** into data for additional insights.

Complex Calculations: OLAP tools support advanced calculations and metrics to evaluate business performance.

Fast Query Performance: MOLAP provides pre-aggregation, while HOLAP offers hybrid performance improvements.

Data Consistency: Ensures accurate, consistent analysis across BI applications.

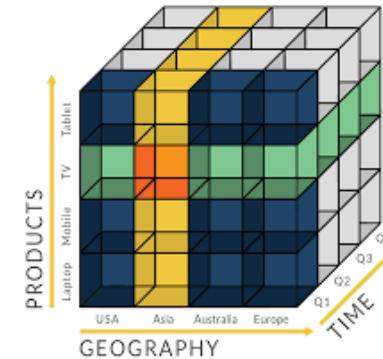
Query resolution procedure: Build the query → Extract → aggregated data → Visualize results → Analyze



An **OLAP query** consists of

- Retrieve **measures or indicators**
- About the **facts**
- **parametrized** by attributes in the dimensions
- Constrained by **conditions** imposed on the dimensions

Eg: What is the **total cost** per diagnostic with low mortality rate in the **last year** for each **province** and **sex**?



Problem

IdCenter
Cluster
Clinic
Area
City
Province
State
Country

Geographic

WHERE?

idPatient
Age
Risk level
Sex
Symptoms
Postal code

WHO?

Patient

idPlace [FK]
idTime [FK]
idPatient [FK]
idDiagnostic[FK]
Probability
Cost

Time

WHEN?

WHAT?

Diagnostic

idTime
Hour
Day
Week
Month
Semester
Year
Decade

IdDiagnostic
Diagnostics
Prevalence
Severity
Known complications
Mortality

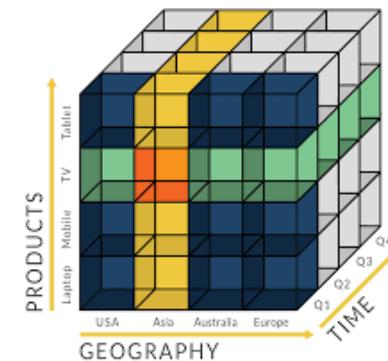
Fact tables

Diagnostic	Sex	Total
D1	M	100
D1	F	200
D2	M	150
D2	F	75



2D view

Diag\Sex	M	F
D1	100	200
D2	150	75



Fact table

Diagnostic	Sex	Province	Total
D1	M	P1	100
D1	F	P1	200
D1	M	P1	100
D1	F	P1	200
D2	M	P2	150
D2	F	P2	75
D2	M	P2	150
D2	F	P2	75

3D view

The diagram illustrates a 3D OLAP cube structure with three dimensions:

- Diagnostic:** D1, D2
- Sex:** M, F
- Province:** P1, P2

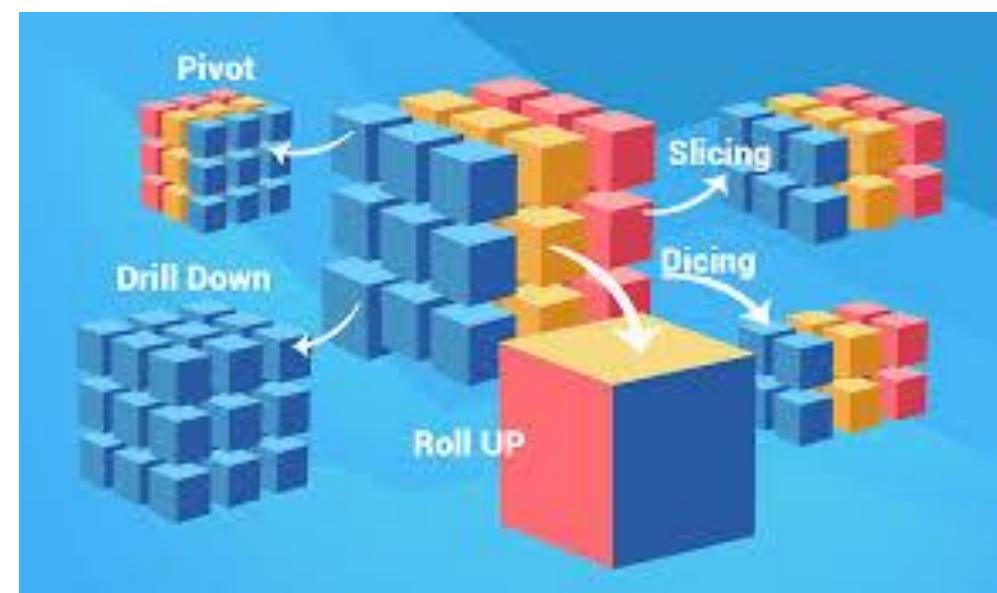
The cube is represented by two stacked tables:

Diag\Sex		M	F
D1		100	200
D2		150	75

Diag\Sex		M	F
D1		100	200
D2		150	75

The interesting thing is **NOT ONLY** to be able to query, in a way, something you can do with selections, projections, concatenation and traditional groupings.

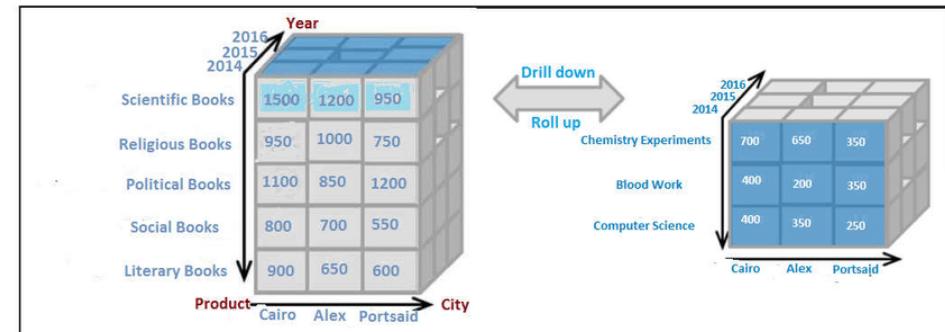
What is really interesting OLAP tools are its refinement operators for handling queries.

DRILL**ROLL****SLICE & DICE****PIVOT****ROLLUP****CUBE**

ROLL-DRILL

Diagnostic	Sex	Province	Total
D1	M	P1	100
D1	F	P1	200
D1	M	P1	100
D1	F	P1	200
D2	M	P2	150
D2	F	P2	75
D2	M	P2	150
D2	F	P2	75

————— **roll** —————
————— **drill** —————

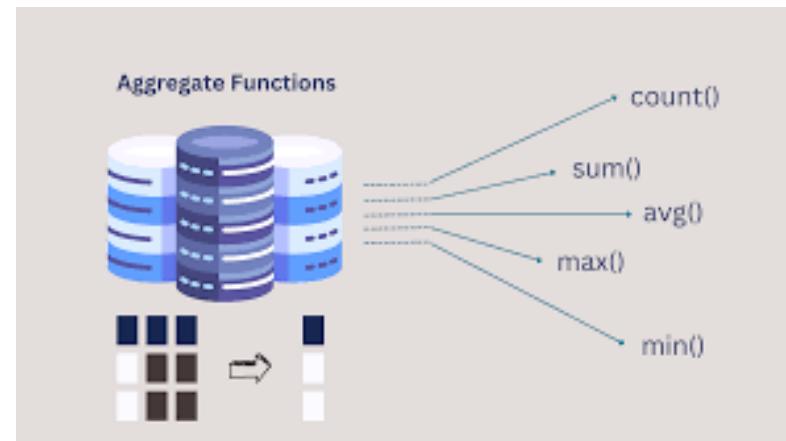


Diagnostic	Sex	Total
D1	M	200
D1	F	400
D2	M	300
D2	F	150

Aggregate (consolidate) and disintegrate (division):

- aggregation (**roll**): delete a grouping criterion in the analysis, aggregating the current groups. The granularity of one or more dimensions is aggregated.
- disintegrate (**drill**): enter a new grouping criterion in the analysis, breaking existing groups.

Aggregation in SQL: sum, count, max, min, average.



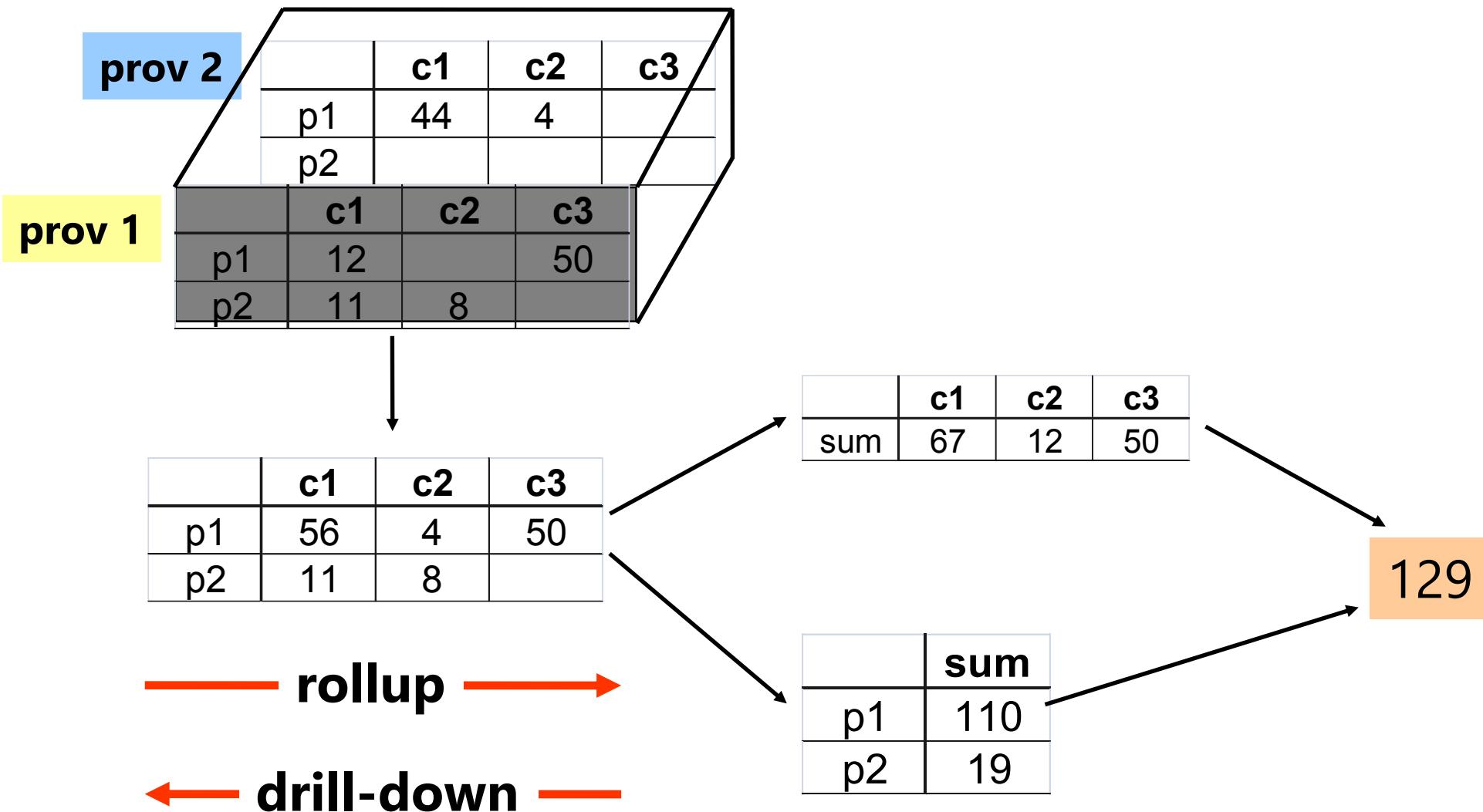
DRILL (ROLL) can be done on:

- **attributes** of one dimension on which a hierarchy has been defined:
 - DRILL-DOWN: upper to lower aggregation level
 - department – category - product (Product)
 - year – semester – month – day (Time)
 - ROLL-UP: lower to upper aggregation level.

Other “drill”:

- **DRILL-ACROSS**: join several fact tables.
- **DRILL-THROUGH**: Use SQL to explore up to the relational back-end tables.

OLAP. ROLL-DRILL



OLAP operators

SLICE & DICE: select and project

SLICE: Filter a specific dimension by selecting a single value.

DICE: Filter data by applying multiple conditions across dimensions.

PIVOT: Rotate, reorientate the data view to analyze different perspectives.



Store	Product	Sales
A	TV	2
A	TV	4
B	TV	6
B	DVD	8

Store	Avg(Sales) for TV	Avg(Sales) for DVD
A	3	(Empty)
B	6	8

Diagnóstico	Sexo	Provincia	Total	Núm
D1	H	P1	100	6
D1	M	P1	200	5
D1	H	P2	100	6
D1	M	P2	200	11
D2	H	P1	150	7
D2	M	P1	75	7
D2	H	P2	150	2
D2	M	P2	75	1

**Slice by removing
Prov dimension**



Diagnostic	Sex	Total
D1	M	100
D1	F	200
D2	M	150
D2	F	70

SELECT diagnostic, sex,
 SUM(total)
 FROM table
 WHERE province = 'P1'
 GROUP BY diagnostic, sex;

PIVOT

	Diagnóstico	Sexo	Total
P1	D1	H	100
	D1	M	200
	D2	H	150
	D2	M	75
P2	D1	H	100
	D1	M	200
	D2	H	150
	D2	M	75

Pivot
→

	Diagnóstico	Prov incia	Total
H	D1	P1	100
	D1	P2	100
	D2	P1	150
	D2	P2	150
M	D1	P1	200
	D1	P2	200
	D2	P1	75
	D2	P2	75

Example

```
CREATE EXTENSION IF NOT EXISTS tablefunc;
```

```
CREATE TABLE to_pivot (
    ID serial,
    Name TEXT, -- Name student
    Course TEXT, -- Course
    Grade INT,
    primary KEY(ID)
);
```

```
INSERT INTO to_pivot(Name,Course,Grade) VALUES
```

```
('Pepe', 'BDII', 9),
('Jose', 'BDII', 7),
('Pepe', 'BI', 8),
('Jose', 'BI', 5);
```

	123 id	A-Z name	A-Z course	123 grade
1	1	Pepe	BDII	9
2	2	Jose	BDII	7
3	3	Pepe	BI	8
4	4	Jose	BI	5

```
SELECT * FROM to_pivot;
```

```
SELECT * FROM crosstab ('Select Name,Course,Grade from to_pivot order by 1,2') as Pivoted (Name text, "BI" INT , "BDII" INT);
```

-

	A-Z name	123 BI	123 BDII
1	Jose	7	5
2	Pepe	9	8

SQL aggregation

- sum(), count(), avg(), min(), max()

Basic idea:

- Combine values in one column
- Into only one value

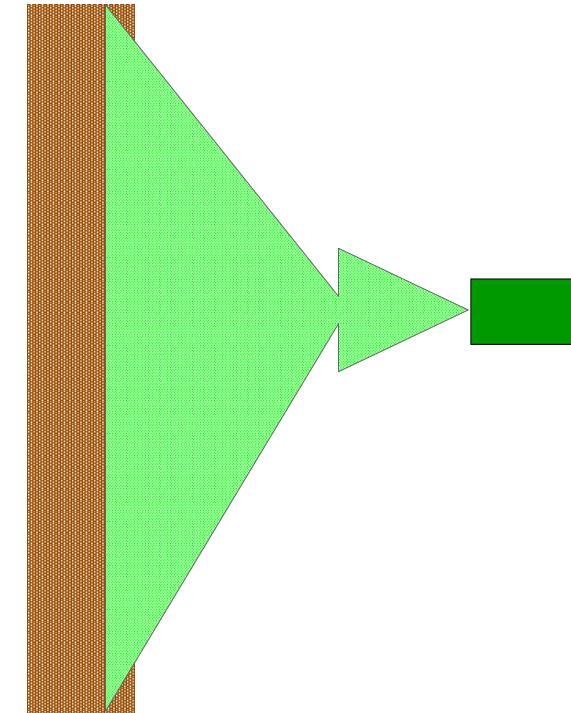
Syntax:

```
SELECT sum(cost) FROM diagnostic;
```

DISTINCT

- Allows the aggregation only of different values

```
SELECT COUNT(DISTINCT cost) FROM diagnostic;
```



OLAP extensions to SQL

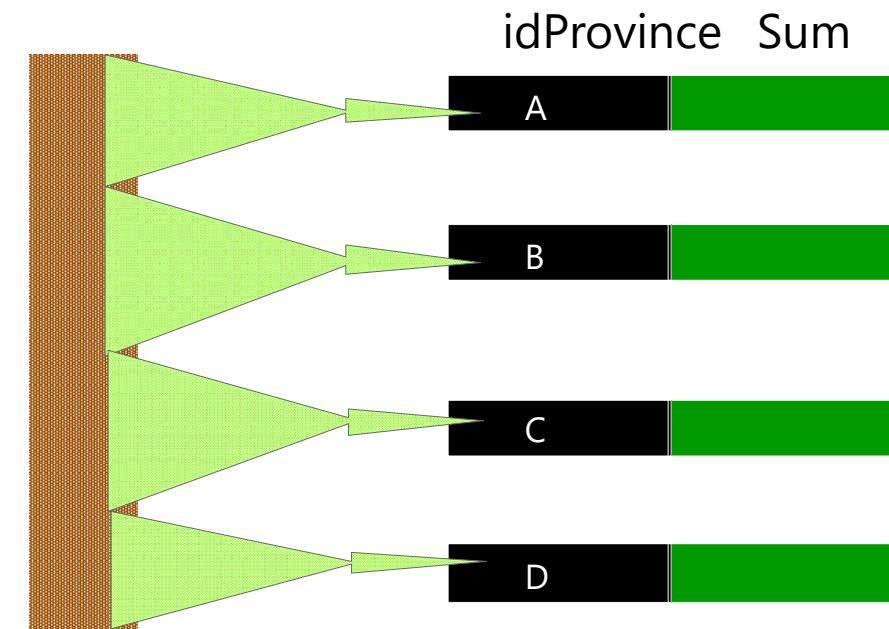
GROUP BY: Groups rows that have same values in specified columns. Aggregation functions are usually associated to it: `Select idprov, sum(b) from table group by idprov`

GROUP BY + HAVING

Aggregating in subgroups of the table that fulfill some condition applied after the grouping.

Syntax

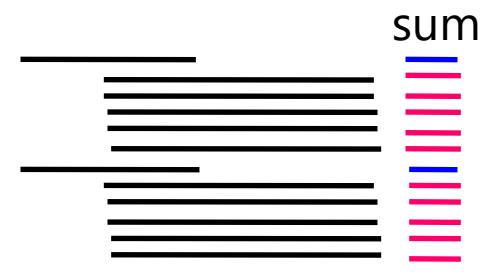
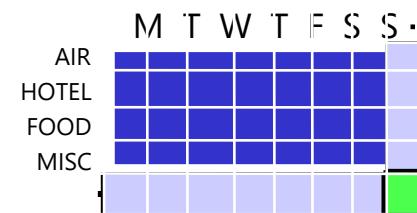
```
SELECT idProvinc, sum(cost)
FROM diagnostic
GROUP BY idProvinc
HAVING population > 2000;
```



OLAP extensions to SQL

Limitations without them

- Useful aggregations are difficult to calculate
 - Data cube
 - Complex: median, variance
 - Moving average
 - Rankings
- Marginals or crosstabs
 - SQL requires additional functionality
- Include sum and partial sums
 - drill-down & roll-up



ROLLUP

ROLLUP: performs the aggregation for the set of prefix of the attributes given

Example:

```
SELECT item-name, color, size, SUM(number)  
FROM sales
```

```
GROUP BY ROLLUP(item-name, color, size)
```

Calculates **SUM for the n+1 prefixes:**

{ (item-name, color, size), (item-name, color), (item-name), () }

Very useful for aggregating in hierarchies defined on dimensions

It can be done in SQL without OLAP extensions, but very inefficiently (multiple GROUP BY and UNION operations).

To improve efficiency: calculate the higher level aggregations using partial results of the more detailed levels

CUBE: generalization of GROUP BY to n-dimensions.

- Calculates the aggregation function for all the subsets of the attributes given instead for only the prefixes (ROLLUP)
- Example:

```
SELECT item-name, color, size, SUM(number)  
FROM sales
```

```
GROUP BY CUBE (item-name, color, size)
```

- Calculates the aggregate for the set of 2^n combinations:
 - $\{(\text{item-name}, \text{color}, \text{size}),$
 $(\text{item-name}, \text{color}), (\text{item-name}, \text{size}), (\text{color}, \text{size}),$
 $(\text{item-name}), (\text{color}), (\text{size}),$
 $\emptyset\}$
 - For each combination, the result is null for attributes that are not present in the combination.

SQL:1999 uses NULL for representing both aggregated rows (ALL) and “usual” null (missing values).

When we have an OLAP query, how to know?

- In order to distinguish them we can use the **GROUPING** function that applied to an attribute
 - Returns 1 if NULL represents ALL
 - Returns 0 otherwise
 - Combined with DECODE (or CASE) we can return the desired value
- ```
SELECT DECODE(GROUPING(Year), 1, 'Total',
Year) AS Year, DECODE(GROUPING(Region), 1, 'Total', Region) AS
Region, SUM(SalesAmount) AS TotalSales FROM Sales GROUP BY CUBE
(Year, Region);
```

# GROUPING SETS

**GROUPING SETS** allows us to specify multiple groupings in a single query.

We can define the subsets of columns for grouping.

No need to have separate queries or UNION ALL.

Provides better efficiency.

```
select Name, Course, AVG(Grade) from
to pivot group by
grouping sets
((Name,Course),(Course),());
```

| O | A-Z name | A-Z course | 123 avg |
|---|----------|------------|---------|
| 1 | [NULL]   | [NULL]     | 7.25    |
| 2 | Pepe     | BI         | 8       |
| 3 | Jose     | BI         | 5       |
| 4 | Jose     | BDII       | 7       |
| 5 | Pepe     | BDII       | 9       |
| 6 | [NULL]   | BDII       | 8       |
| 7 | [NULL]   | BI         | 6.5     |

# WINDOW FUNCTIONS

**WINDOW** clause defines **ordered** and **overlapping** groups of rows to calculate aggregates based on a defined "window", while retaining the original rows.

**GROUP BY** clause defines disjoint partitions of tuples in a sorted table, then calculates aggregates on those partitions, and generates a tuple with the result of the aggregate for each partition. It eliminates rows-level granularity

- Example: "For each day, we want the average cost of obtaining diagnoses from the previous day, the current and the next, and cumulatively in the last 7 days":

```
SELECT date, sum(cost) OVER (order by date ROWS BETWEEN 1 preceding and 1 following), sum(cost) OVER (order by date ROWS BETWEEN 7 preceding and CURRENT ROW)) FROM diagnostics;
```

## Syntax:

- **SELECT** attribute\_list\_1, + Aggregated\_function **OVER** W as windowName
- **FROM** table\_list
- **WHERE** constraints

**WINDOW** W AS (

- **PARTITION BY** attribute\_list\_2
- **ORDER BY** attribute\_list\_3
- **frame declaration**)

Frame declaration is optional. By default, RANGE BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW

## Execution:

- FROM, WHERE, GROUP and HAVING generate an **intermediate** table.
- PARTITION: each partition contains tuples with the same values in the attributes given in attribute\_list\_2
- ORDER BY: rows in each partition are sorted according to the values of the attributes in attribute\_list\_3
- SELECT the tuples under the constraints established in the frame declaration
  - RANGE: logical conditions (ie: 5 days)
  - ROWS: in rows (ie: 5 preceding rows)

## Frame examples:

- between rows unbounded preceding and current row
- rows unbounded preceding
- range between 10 preceding and current row
- range interval 10 day preceding
- range between interval 1 month preceding and interval 1 month following

**Default frame:** If the frame is not specified, all preceding and current rows are considered in the partition

- RANGE BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW

**RANK assigns** to every tuple a rank based in some sorting of some attribute

- Example: given a cost-province relation rank each province by its cost.

```
SELECT province, rank() over (order by coste desc) as
provrank FROM diagnostic
```

- Afterwards, the result can be sorted by that field

```
SELECT province, rank() over (order by coste desc) as
provrank FROM diagnostic order by provrank
```

**RANK** allow gaps if there are 2 values with the same ranking.

- Example: if the 1<sup>rst</sup> and 2<sup>nd</sup> classified have the same cost, then both will be assigned rank 1, and the next row will have rank 3
- DENSE\_RANK does not allow gaps, so the next row will have rank 2

- RANK over partitions:
  - “Rank the community and provinces by their cost”

```
SELECT province, community,
rank () over (partition by community order by cost
desc) as prov-community-rank
FROM diagnostic
ORDER BY community, prov-community-rank
```

si particionamos por provincia, daria que todas las provincias tendrian ranking 1

- Several RANK can be included in the same query.

# Other functions

## Other rank functions

- **percent\_rank**: it displays each row as a percentage of all the other rows up to 100% in a rank.  $= (\text{rank}-1)/(\text{Number\_rows\_partition}-1)$ .

```
SELECT Student, Score,
PERCENT_RANK() OVER
(ORDER BY Score DESC) AS
Percent_Rank
FROM Students;
```

| O | A-Z student | 123 score | 123 percent_rank |
|---|-------------|-----------|------------------|
| 1 | Elena       | 92        | 0                |
| 2 | Jose        | 90        | 0.25             |
| 3 | Pepe        | 85        | 0.5              |
| 4 | Marco       | 85        | 0.5              |
| 5 | Manolo      | 78        | 1                |

- **cume\_dist**: cumulative distribution: It displays the number of values in the set preceding and including in the specified order divided by the number of rows.

```
SELECT Student, Score,
CUME_DIST() OVER (ORDER BY Score
DESC) AS Cume_Dist
FROM Students;
```

| O | A-Z student | 123 score | 123 cume_dist |
|---|-------------|-----------|---------------|
| 1 | Elena       | 92        | 0.2           |
| 2 | Jose        | 90        | 0.4           |
| 3 | Pepe        | 85        | 0.8           |
| 4 | Marco       | 85        | 0.8           |
| 5 | Manolo      | 78        | 1             |

# Other functions

## Other rank functions

- **row\_number:**

```
SELECT Student, Score,

ROW_NUMBER() OVER (ORDER BY Score DESC) AS Row_Number

FROM Students;
```

| O | A-Z student | 123 score | 123 row_number |
|---|-------------|-----------|----------------|
| 1 | Elena       | 92        | 1              |
| 2 | Jose        | 90        | 2              |
| 3 | Pepe        | 85        | 3              |
| 4 | Marco       | 85        | 4              |
| 5 | Manolo      | 78        | 5              |

- **ntile(x): cuantile**
  - Divides the rows in the partition in x buckets with the same number of rows

```
SELECT Student, Score,

NTILE(2) OVER (ORDER BY Score

DESC) AS ntile

FROM Students;
```

| O | A-Z student | 123 score | 123 ntile |
|---|-------------|-----------|-----------|
| 1 | Elena       | 92        | 1         |
| 2 | Jose        | 90        | 1         |
| 3 | Pepe        | 85        | 1         |
| 4 | Marco       | 85        | 2         |
| 5 | Manolo      | 78        | 2         |

# Other functions

- Numeric functions (exp, cos, ln, ...) `SELECT EXP(2);`
- Aggregated (std, var, corr, regr, ...) `SELECT STDDEV(Score) FROM Students;`
- Frame functions: lag, lead, ...

|   | A-Z student | 123 score | 123 previous_score | 123 next_score |
|---|-------------|-----------|--------------------|----------------|
| 1 | Elena       | 92        | [NULL]             | 90             |
| 2 | Jose        | 90        | 92                 | 85             |
| 3 | Pepe        | 85        | 90                 | 85             |
| 4 | Marco       | 85        | 85                 | 78             |
| 5 | Manolo      | 78        | 85                 | [NULL]         |

```
SELECT Student, Score,
LAG(Score) OVER (ORDER BY Score DESC) AS Previous_Score,
LEAD(Score) OVER (ORDER BY Score DESC) AS Next_Score
FROM Students;
```

SQL:1999 allows the use of **nulls first** and **nulls last**. It serves to define if nulls appear before or after non-null values in the sort ordering. By default, NULLS FIRST for DESC order and NULLS LAST for ASC order.

```
SELECT Student, Score, ROW_NUMBER() OVER (ORDER
BY Score ASC) AS Default_Order, ROW_NUMBER()
OVER (ORDER BY Score ASC NULLS FIRST) AS
Nulls_First_Order, ROW_NUMBER() OVER (ORDER BY
Score ASC NULLS LAST) AS Nulls_Last_Order
FROM Students;
```

|   | A-Z student | 123 score | 123 default_order | 123 nulls_first_order | 123 nulls_last_order |
|---|-------------|-----------|-------------------|-----------------------|----------------------|
| 1 | Manolo      | 78        | 1                 | 2                     | 1                    |
| 2 | Pepe        | 85        | 2                 | 3                     | 2                    |
| 3 | Marco       | 85        | 3                 | 4                     | 3                    |
| 4 | Jose        | 90        | 4                 | 5                     | 4                    |
| 5 | Elena       | 92        | 5                 | 6                     | 5                    |
| 6 | Pat         | [NULL]    | 6                 | 1                     | 6                    |

# Codd rules

**1 Multidimensional view of data**

**2 Transparency to support (ROLAP, MOLAP)**

interfaz simple para el usuario, le da igual lo que hay detrás

**3 Accessibility**

el usuario debe acceder de forma simple a los datos, da igual donde estén

**4 Coherent performance in reporting**

resultados rápidos y consistentes

**5 Client-Server Architecture**

**6 Generic operations regarding the number of dimensions**

**7 Dynamic sparse matrix**

**8 Multiuser support**

**9 Flexibility in the definition of the dimensions: constraints, aggregations and hierarchies among them.**

**10 Intuitive handling of operators: drill, roll, slice-&-dice, pivot.**

**11 flexible report generation**

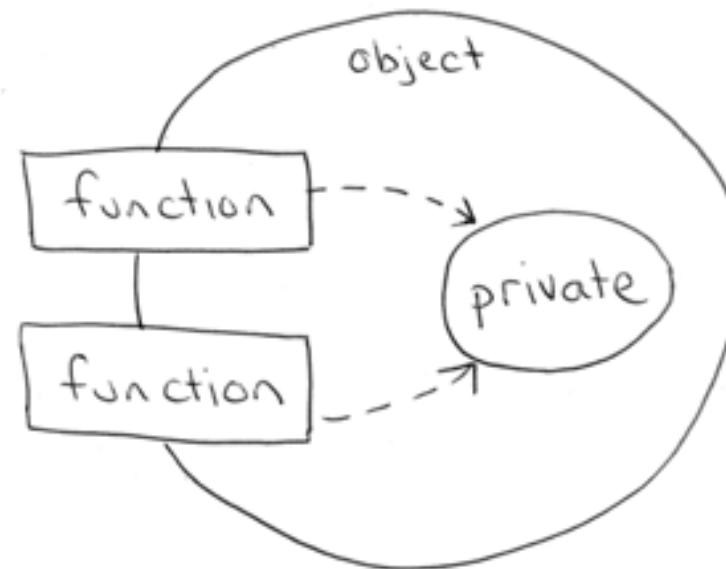
**12 No limit dimensions**

# REPORTING

Unit 3 – Data exploitation. Query languages and  
visualization  
S3 –2 – MDX

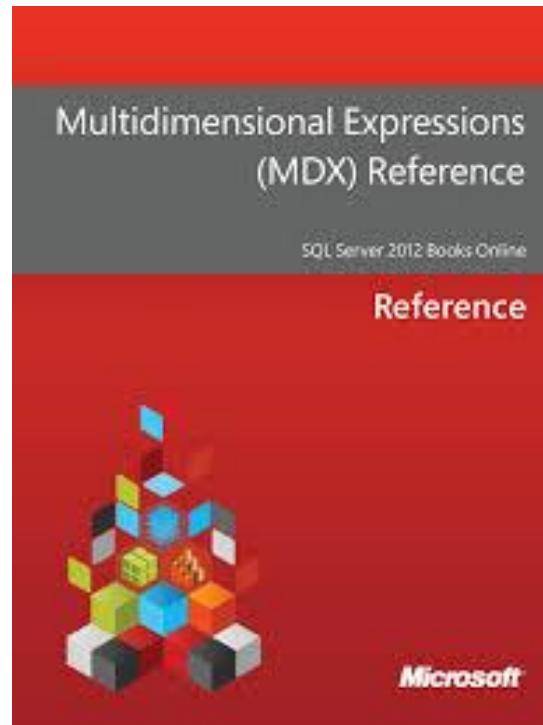
- **CUBES**
- **LANGUAGE**
- **MultiDimensional eXpression.**
  - Microsoft in 1997

- Do you remember studying Object-Oriented Programming?



# Outline

1. Introduction
2. Tuples, Sets & Cells
3. MDX Spells
4. MDX Query Syntax



**What is MDX?** query language for OLAP cubes and multidimensional databases, designed to retrieve and analyze data stored in cubes.

Primarily used in MOLAP systems like Microsoft SQL Server Analysis Services (SSAS), Oracle Essbase, and IBM Cognos.

## **Key Features of MDX:**

- Retrieval of data from multidimensional structures.
- Definition of calculated members and custom measures.
- Aggregation and slicing of data for analysis.

## **Why Use it?**

- Advanced capabilities for querying hierarchical and multidimensional datasets, making it ideal for complex analytics in BI systems.

- **Fact Tables:** Contain measures, which are numerical values or metrics used for analysis.
- **Dimension tables:** Contain Members. Members are values in a dimension that describe or categorize the data in the fact table. These values can be qualitative or quantitative.
- E.g.: 2 dimension cube
  - **1 measure:** discharged patients.
  - Time Dimension with 4 members: Jan to April.
  - Hospital Dimension with 4 members: H1,H2,H3,H4.

| Discharged | H1 | H2 | H3 | H4 |
|------------|----|----|----|----|
| January    | 20 | 44 | 81 | 44 |
| February   | 15 | 32 | 78 | 32 |
| March      | 23 | 65 | 88 | 65 |
| April      | 19 | 67 | 67 | 67 |

- **Fact Tables:** Contain measures, which are numerical values or metrics used for analysis.
- **Dimension tables:** Contain Members. Members are values in a dimension that describe or categorize the data in the fact table. These values can be qualitative or quantitative.
- E.g.: 2 dimension cube
  - **2 measure:** no. discharged patients, total cost (M€).
  - Time Dimension with 4 members: Jan to April.
  - Hospital Dimension with 4 members: H1,H2,H3, H4.

| Discharged | H1       | H2       | H3        | H4       |
|------------|----------|----------|-----------|----------|
| January    | 20 1.5M€ | 44 4.1M€ | 81 10.5M€ | 44 4.1M€ |
| February   | 15 1.1M€ | 32 3.9M€ | 78 10.4M€ | 32 3.9M€ |
| March      | 23 1.6M€ | 65 5.4M€ | 88 10.7M€ | 65 5.4M€ |
| April      | 19 1.5M€ | 67 5.6M€ | 67 9.5M€  | 67 5.6M€ |

CUBE: 3 DIMENSIONS  
TIME, DX, HOSPITAL

JANUARY

FEBRUARY

MARCH

APRIL

ENDOCRINE  
E00-E90

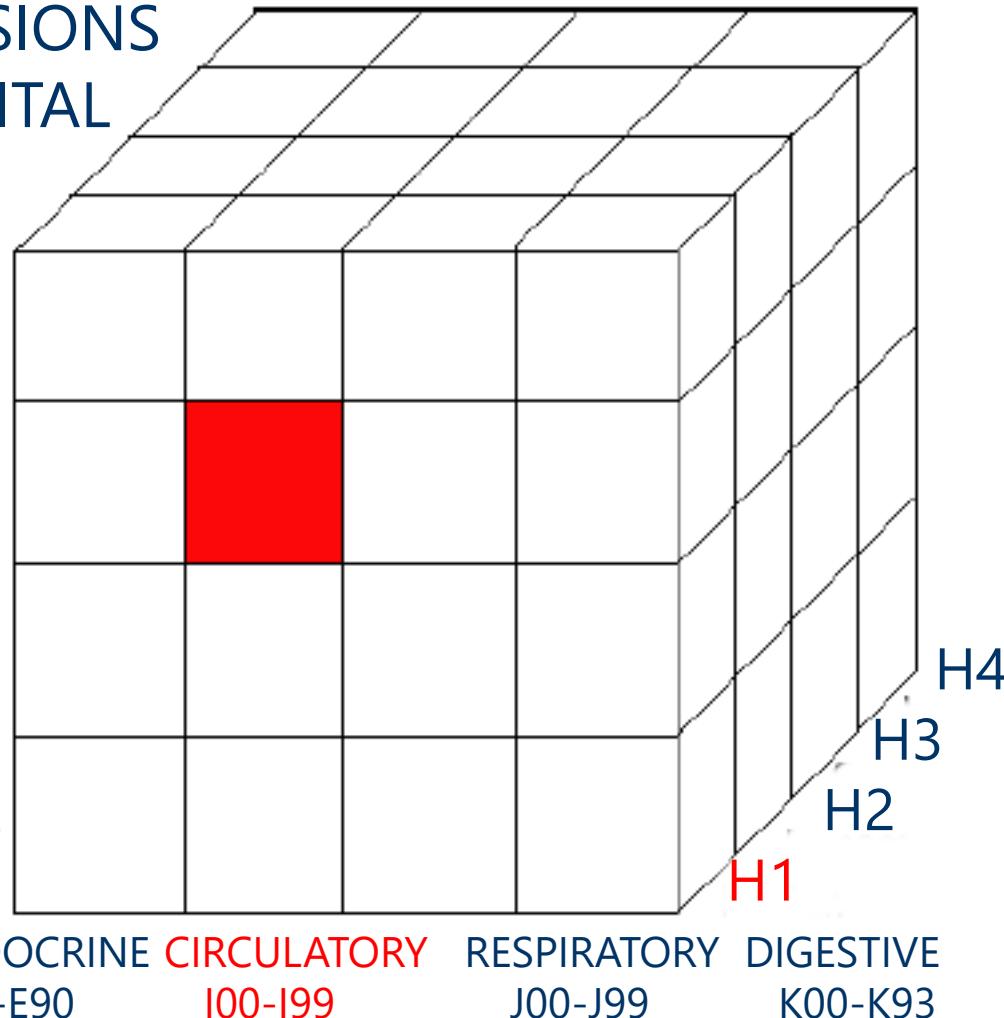
CIRCULATORY  
I00-I99

RESPIRATORY  
J00-J99

DIGESTIVE  
K00-K93

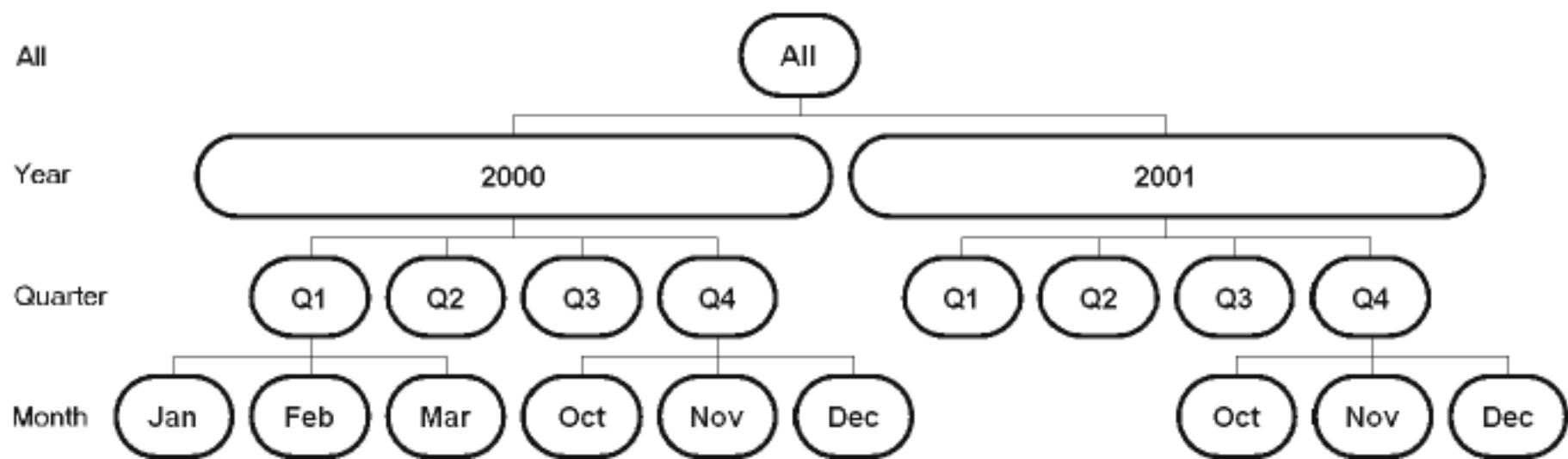
H1

H2  
H3  
H4



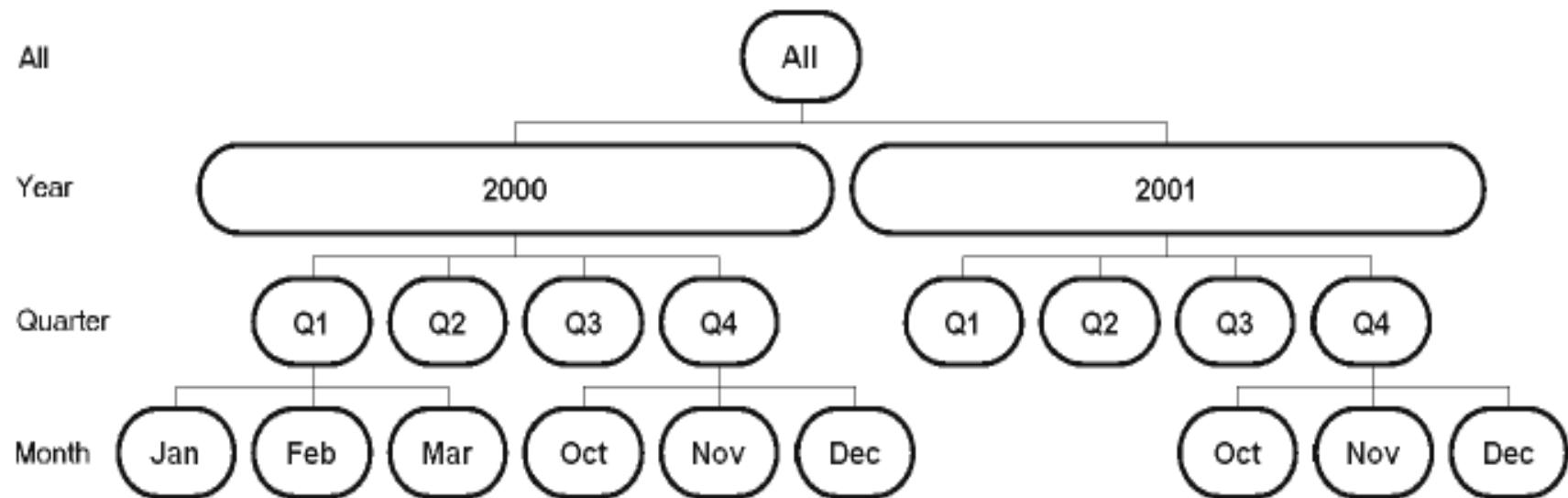
# Tuples, Sets & Cells

- Dimension has **hierarchies**.
- Hierarchy has **levels**: All, Year, Quarter, Month.
- A hierarchy organizes members of a dimension into multiple levels for drill-down or roll-up operations.
- All → Highest aggregation level.



- Naming Conventions

[Time].[All].[2000].[Q4].[Oct] = [Time].[2000].[Q4].[Oct]



Naming Conventions:  **Tuple**

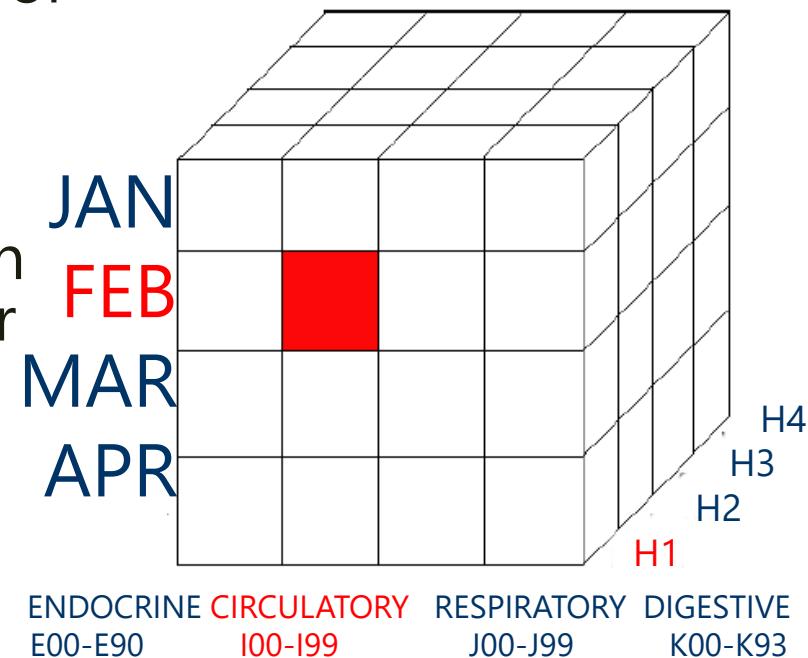
Tuple in pseudo-MDX:

([Time].[Feb],[Dx].[Circ],[Hosp].[H1])

Def1: “ **Tuple** is the intersection of  
**one member**  
**from each dimension**”

Def2: “A tuple is the intersection  
of **one (and only one)** member  
taken from one or several of  
the dimensions in the cube.”

(tuple=single cell in the cube ??)



- Naming Conventions: **Tuple**

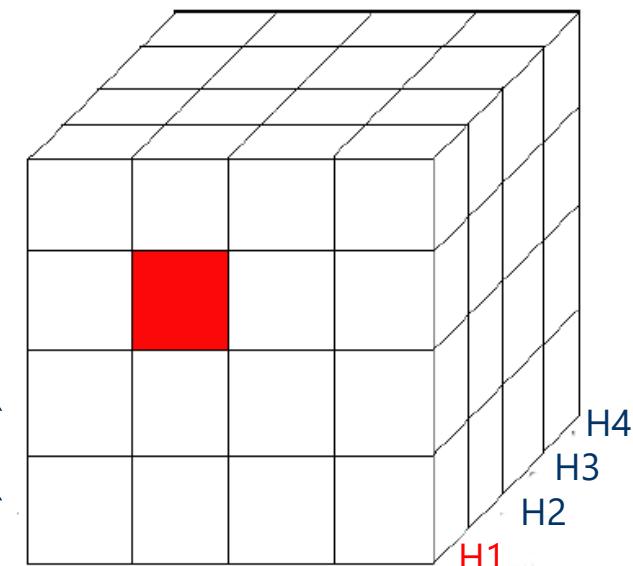
Tuple in pseudo-MDX:  $(x,y,z)=(y,z,x)$   
([Time].[Feb],[Dx].[Circ],[Hosp].[H1])

Def1: "**Tuple** is the intersection  
choosing **one member**  
**of each dimension**"

Def2: "A tuple is the intersection  
of **one (and only one)** member  
taken from one or several of  
the dimensions in the cube."

(tuple=single cell in the cube ??)

JAN  
FEB  
MAR  
APR



ENDOCRINE CIRCULATORY RESPIRATORY DIGESTIVE  
E00-E90 I00-I99 J00-J99 K00-K93

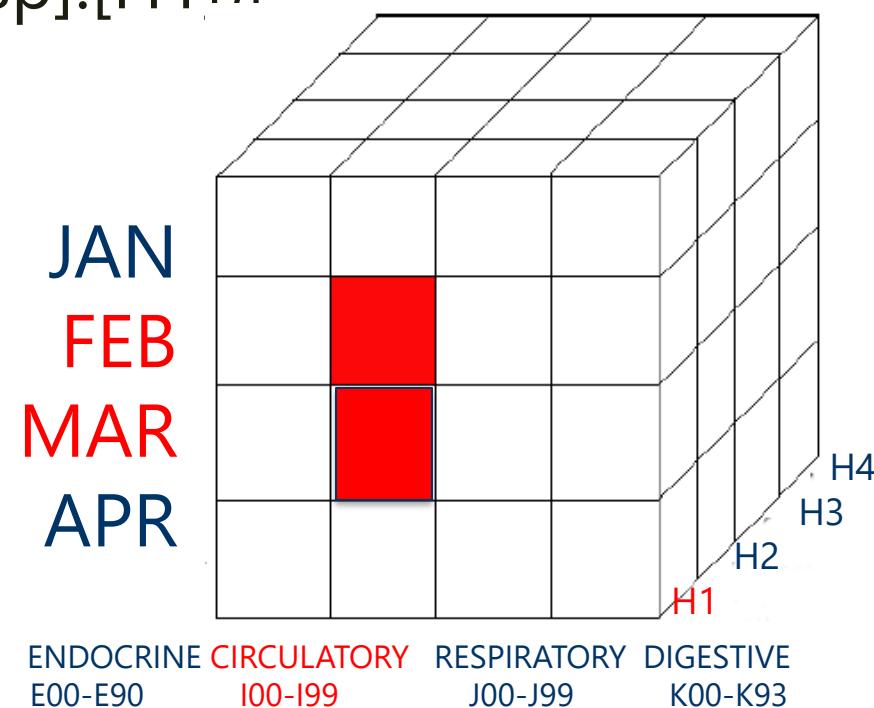
- Naming Conventions: **Set**

Set in pseudo-MDX:  $\{(x_1, y_1, z_1), \dots, (x_n, y_n, z_n)\}$

$\{([Time].[Feb].[Dx].[Circ].[Hosp].[H1]),$   
 $([Time].[Mar].[Dx].[Circ].[Hosp].[H1])\}$

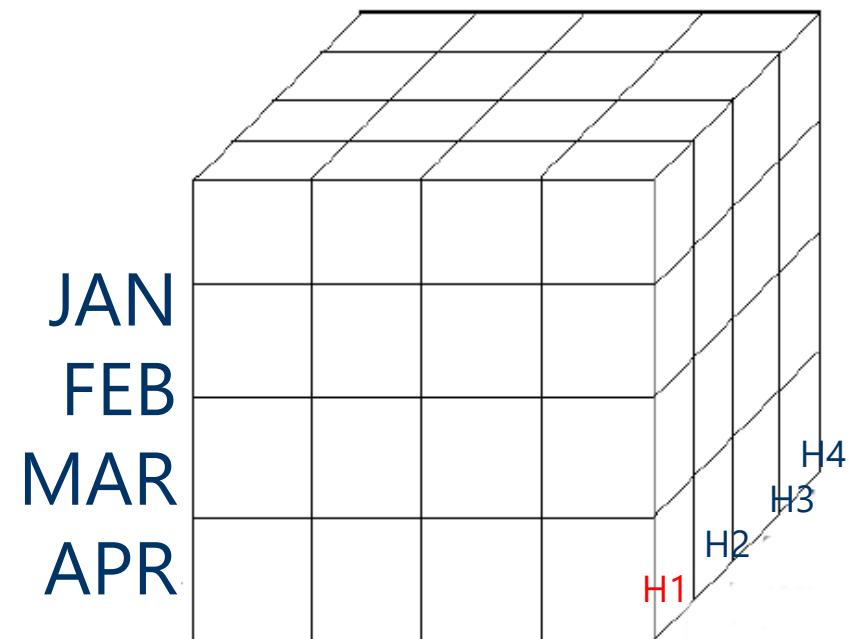
**"Set** is a set of tuples with  
the same dimensionality"  
(set of cells in the cube)

A set can be aggregated in MDX:  
 $\text{AVG}(\text{SET}) \rightarrow \text{FLOAT}$



- Question: Tuple or Set?

[DX].[Circulatory],[Hosp].[H1]

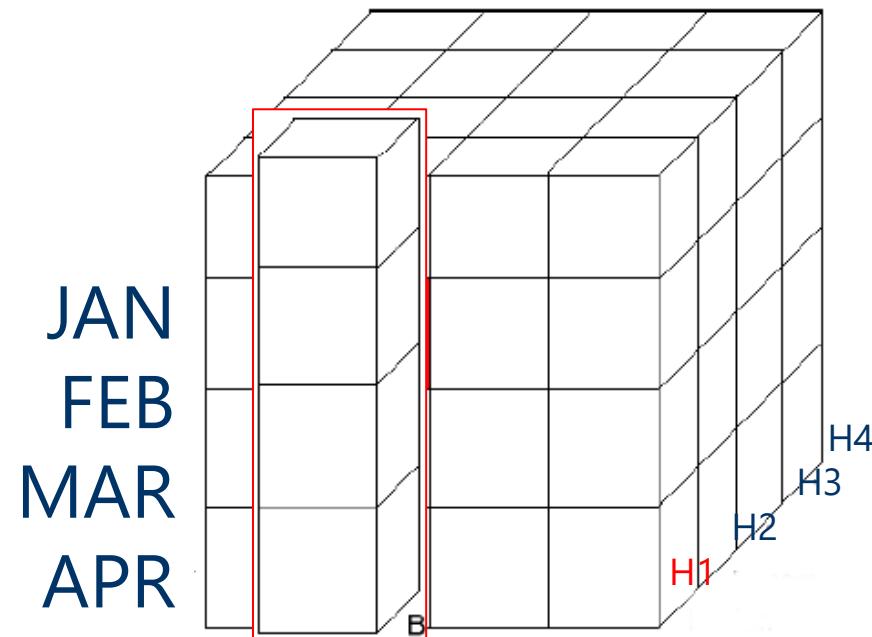


ENDOCRINE CIRCULATORY RESPIRATORY DIGESTIVE  
E00-E90 I00-I99 J00-J99 K00-K93

- Question: Tuple or Set?

[DX].[Circulatory],[Hosp].[H1]

Is a TUPLE!  
(but MANY CELLS!)



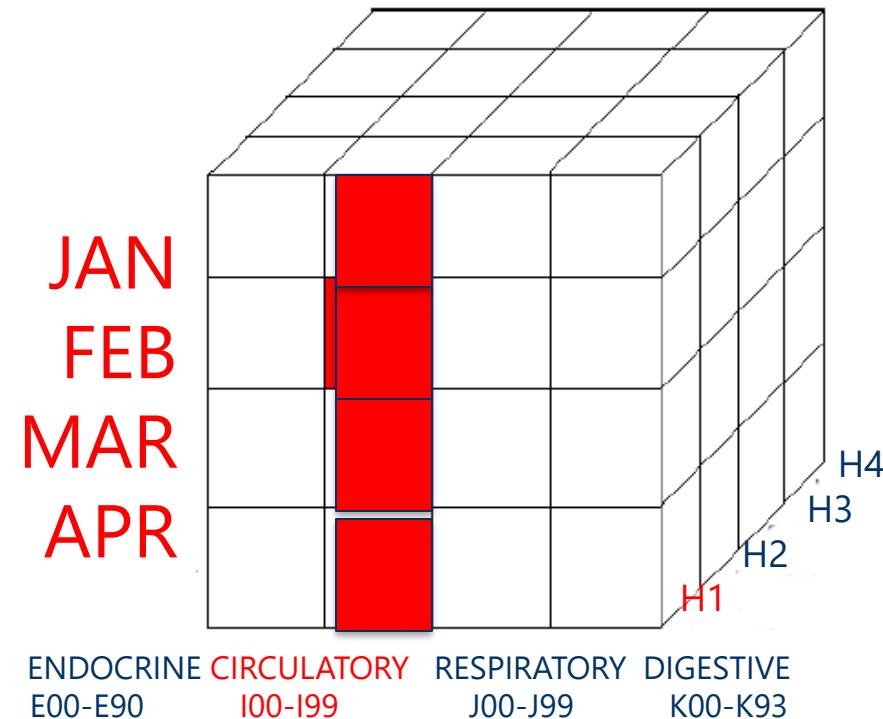
ENDOCRINE CIRCULATORY RESPIRATORY DIGESTIVE  
E00-E90 I00-I99 J00-J99 K00-K93

- Question: Difference between?

- a)  $([Dx].[Circ],[Hosp].[H1])$

- b)

- $\{([Dx].[Circ],[Hosp].[H1],[Time].[Jan]),$   
 $([Dx].[Circ],[Hosp].[H1],[Time].[Feb]),$   
 $([Dx].[Circ],[Hosp].[H1],[Time].[Mar]),$   
 $([Dx].[Circ],[Hosp].[H1],[Time].[Apr])\}$

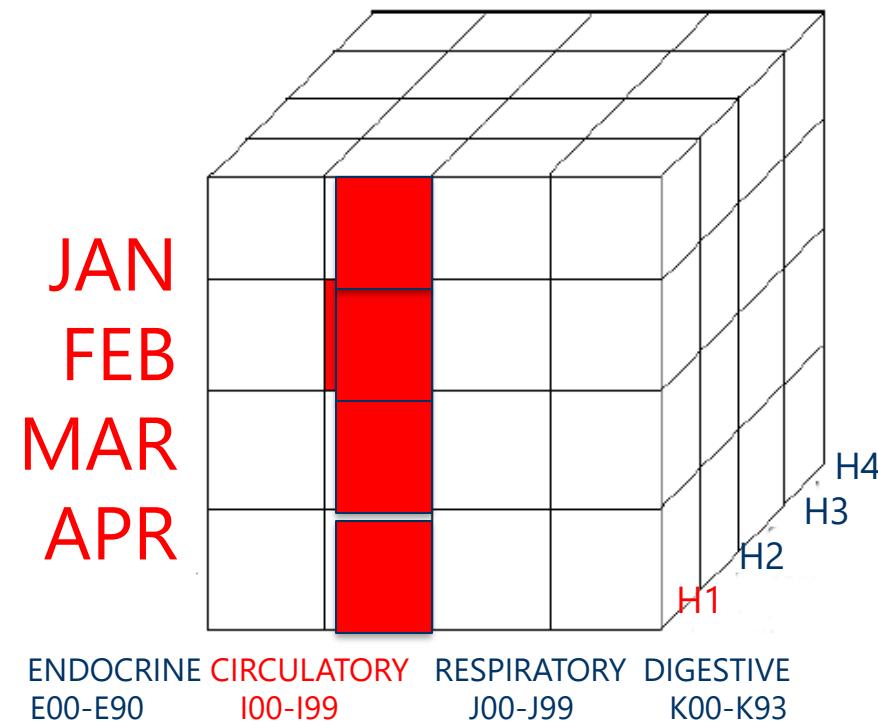


- Question: Difference between?

- a)  $([Dx].[Circ],[Hosp].[H1])$  **IS A TUPLE** (SEE DEFINITION 1)
- b)

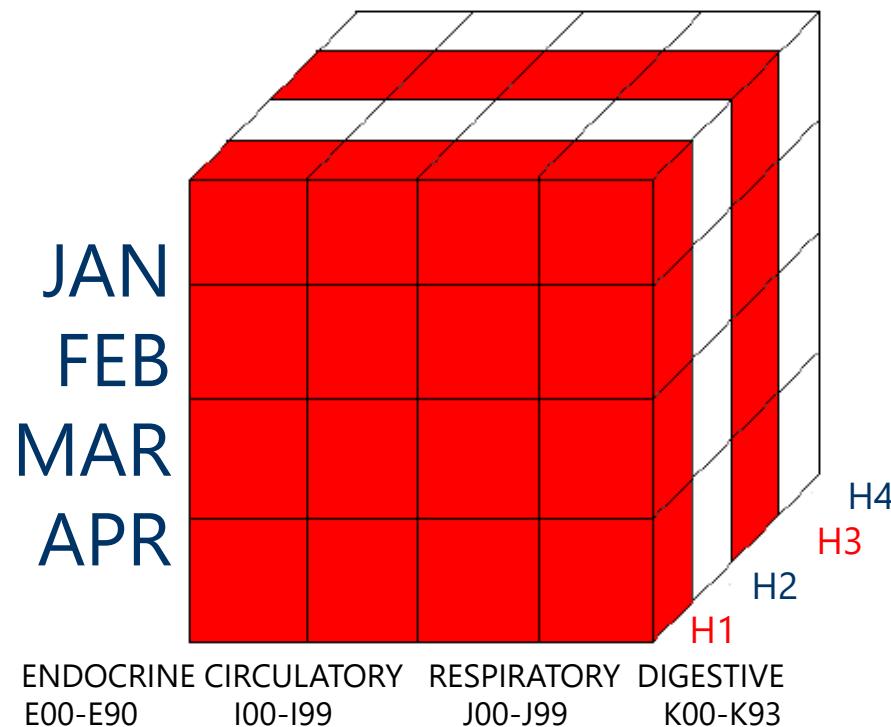
- $\{([Dx].[Circ],[Hosp].[H1],[Time].[Jan]),$   
 $([Dx].[Circ],[Hosp].[H1],[Time].[Feb]),$   
 $([Dx].[Circ],[Hosp].[H1],[Time].[Mar]),$   
 $([Dx].[Circ],[Hosp].[H1],[Time].[Apr])\}$

**IS A SET**



- Question: Tuple or Set?

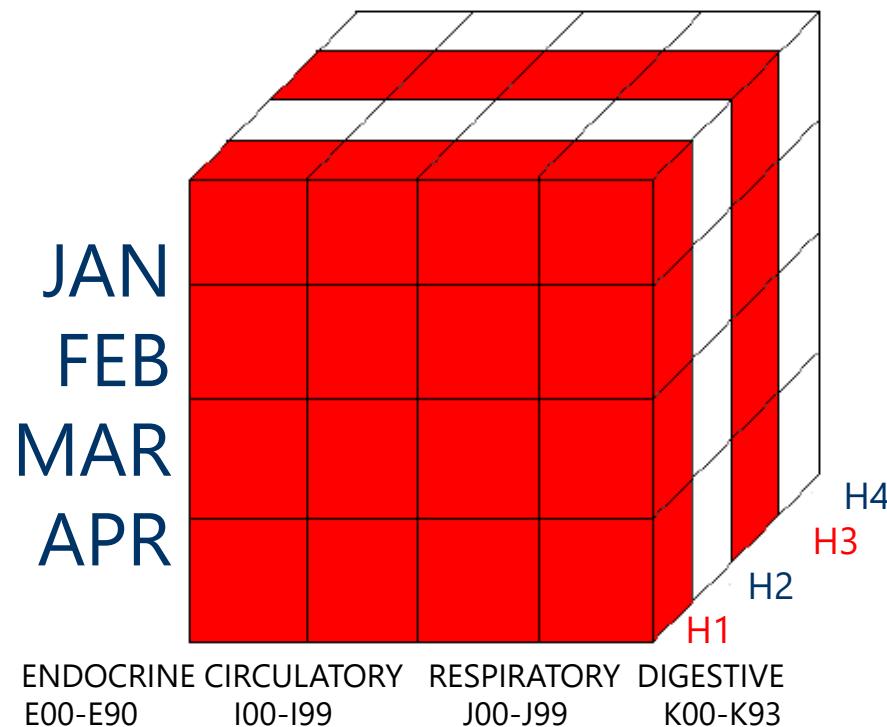
[Hosp].[H1], [Hosp].[H3]



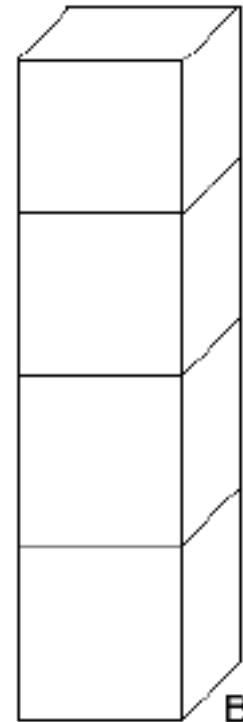
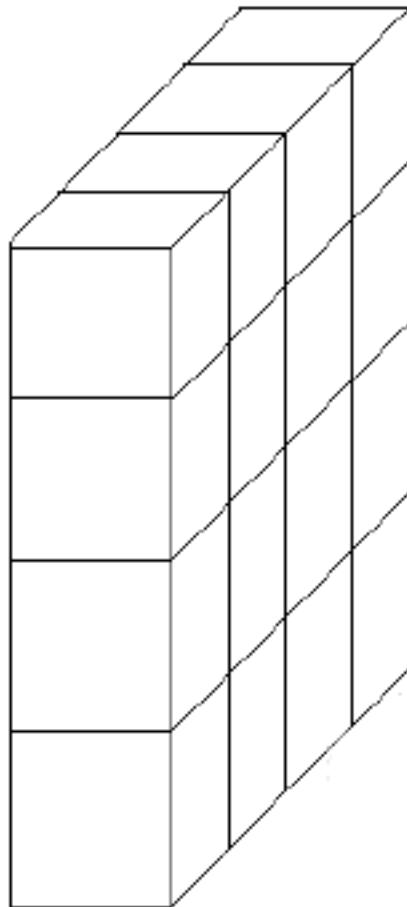
- Question: Tuple or Set?

[Hosp].[H1], [Hosp].[H3]

is a Set  
(see Def2!)



[DX].[Circulatory] [DX].[Circulatory],[Hosp].[H1]



ALL THESE ARE TUPLES  
SINCE THEY HAVE THE  
“CAPACITY TO POINT TO  
A SINGLE CELL”  
(actually they don't)

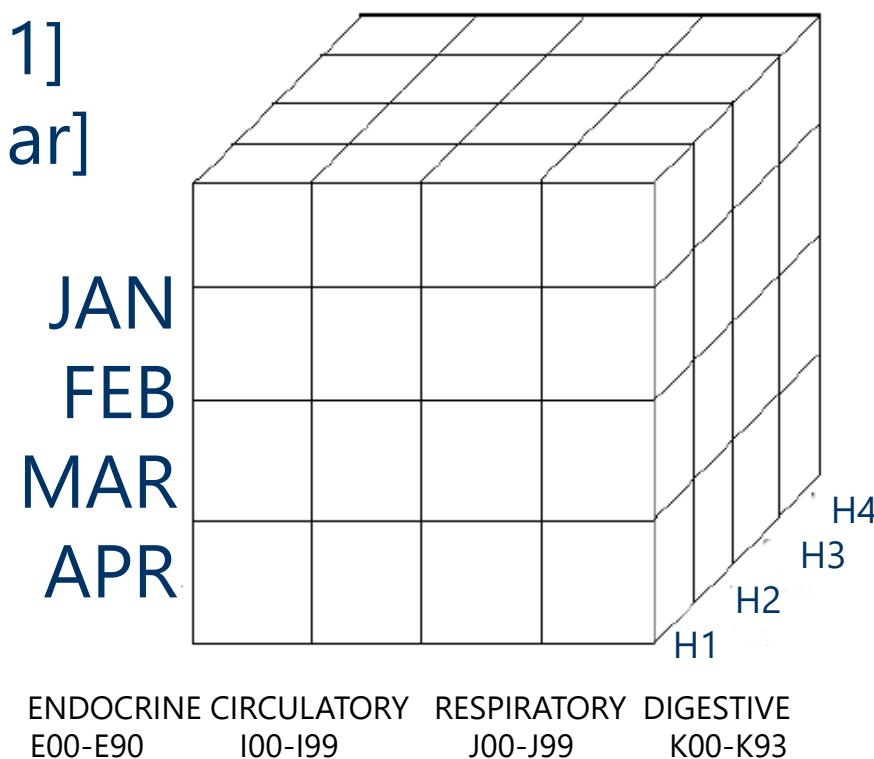
- Question: Do these tuples point to a single cell?

[DX].[Circulatory],[Hosp].[H1], [Time].[Mar]

[DX].[Circulatory],[Hosp].[H1]

[DX].[Circulatory],[Time].[Mar]

[Hosp].[H1]



- Question: Do these tuples point to a single cell?

[DX].[Circulatory],[Hosp].[H1], [Time].[Mar]

[DX].[Circulatory],[Hosp].[H1]

[DX].[Circulatory],[Time].[Mar]

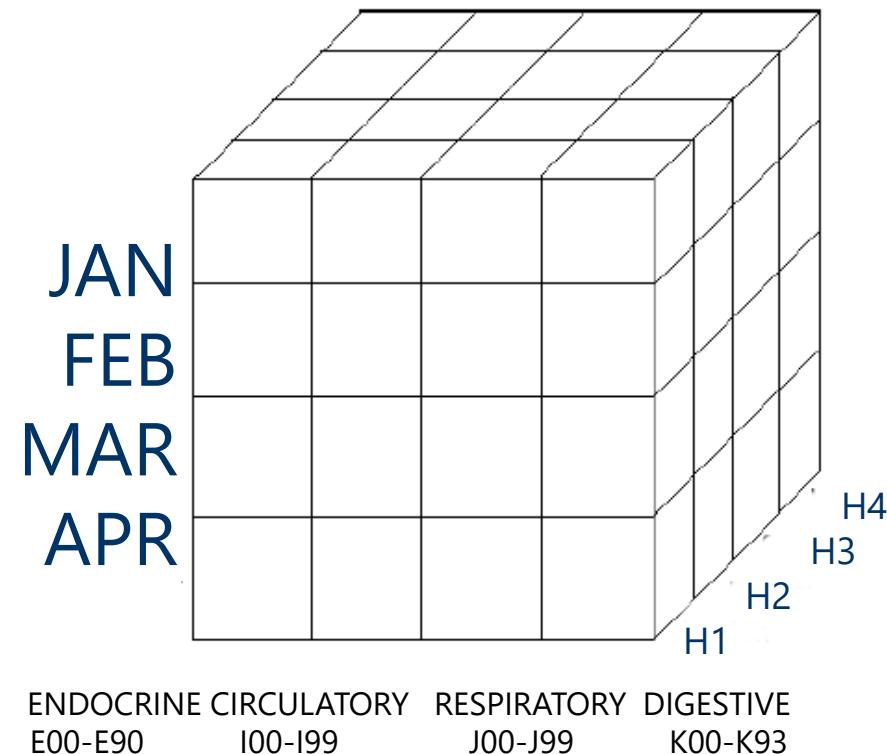
[Hosp].[H1]

(or an aggregated cell)...

YES

If we consider that all dimensions have a  
*'DEFAULT MEMBER'*

In MDX if you don't specify a member of a  
dimension the default member is implied



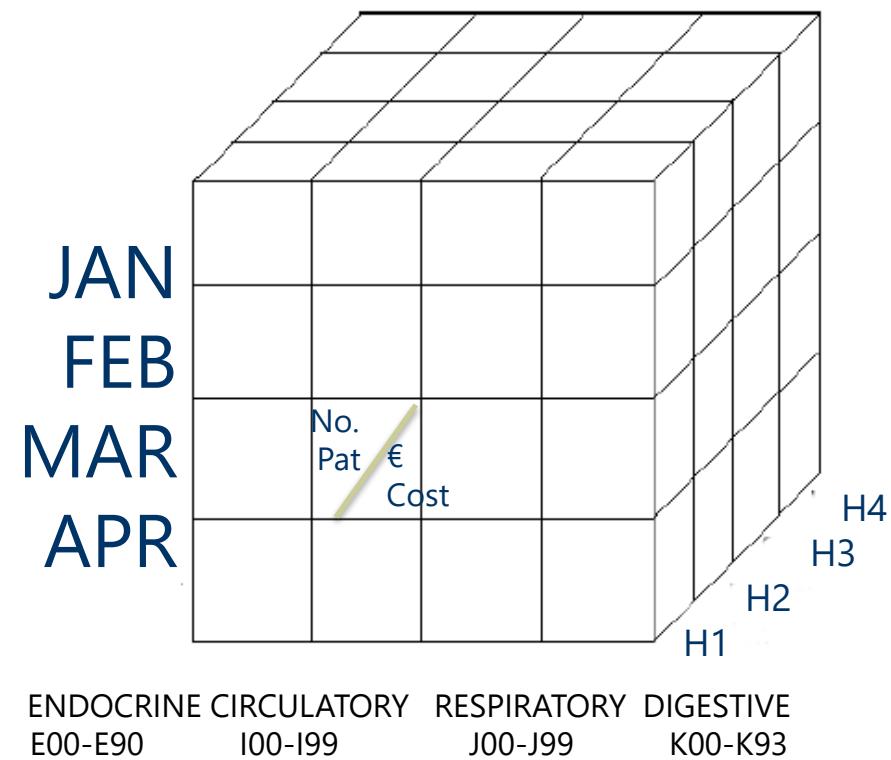
## WHAT IF...

- Measures like dimensions

Suppose a cube **with 2 measures**:  
No.Patients and €Cost.

[Hos].[H1],[TIME].[Mar],  
[Dx].[Car],**[Measures].[NoPat]**

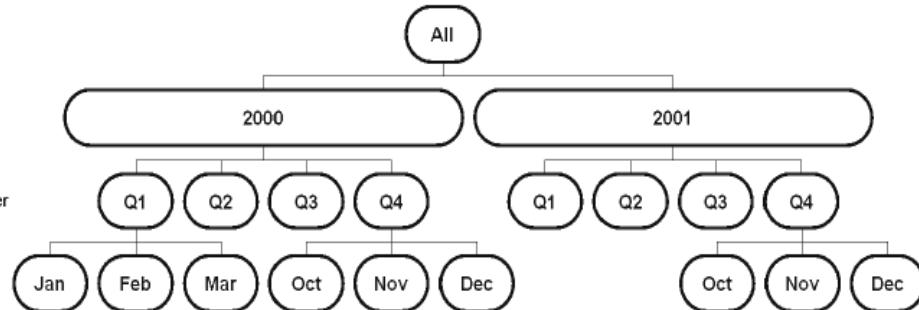
Measure behaves like member  
of a dimension



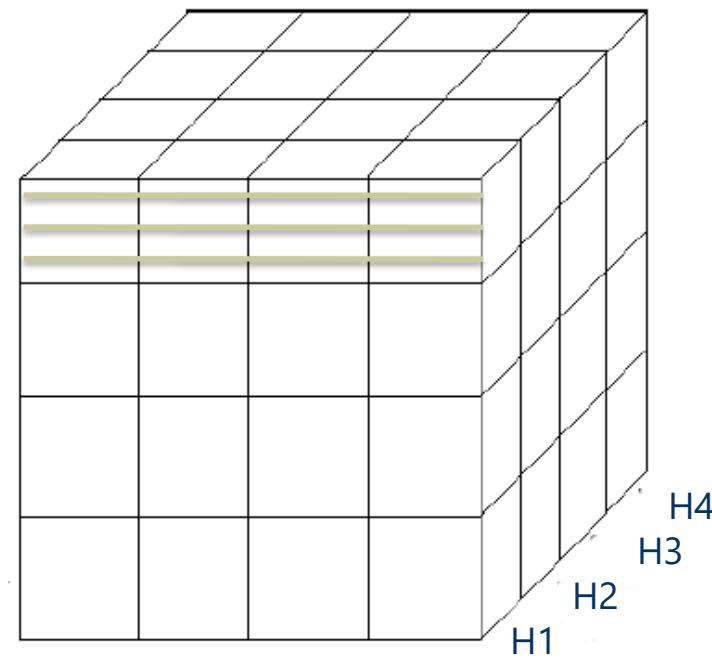
- Measures & Hierarchies

[Hos].[H1],**[TIME].[Q1]**,[Dx].[Cir]

**SET?**  
**TUPLE?**  
**CELL?**



2000  
Q1 JAN  
FEB  
MAR  
Q2 APR  
MAY  
JUN  
JULY ...  
Q3  
Q4

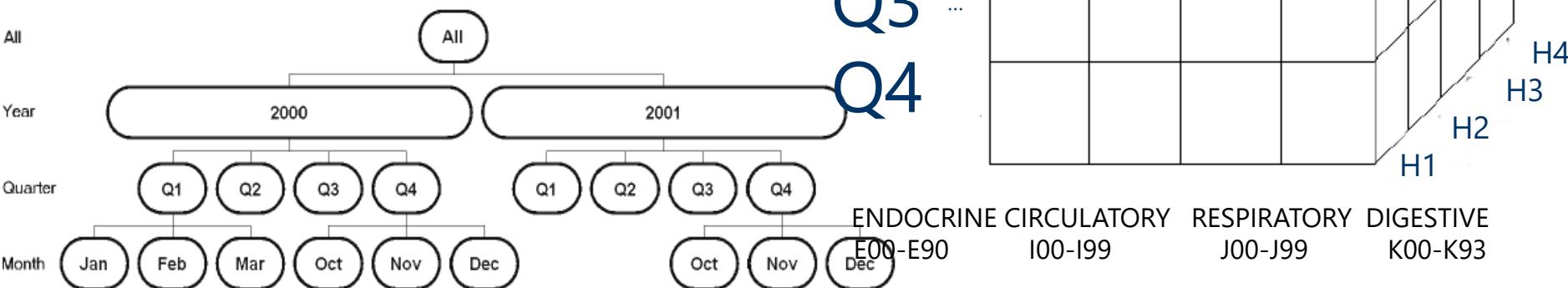


|           |             |             |           |
|-----------|-------------|-------------|-----------|
| ENDOCRINE | CIRCULATORY | RESPIRATORY | DIGESTIVE |
| E00-E90   | I00-I99     | J00-J99     | K00-K93   |

- Measures & Hierarchies

[Hos].[H1],**[TIME]**.[Q1],[Dx].[Cir]

**TUPLE!**  
**CELL! (AGGREGATION)**

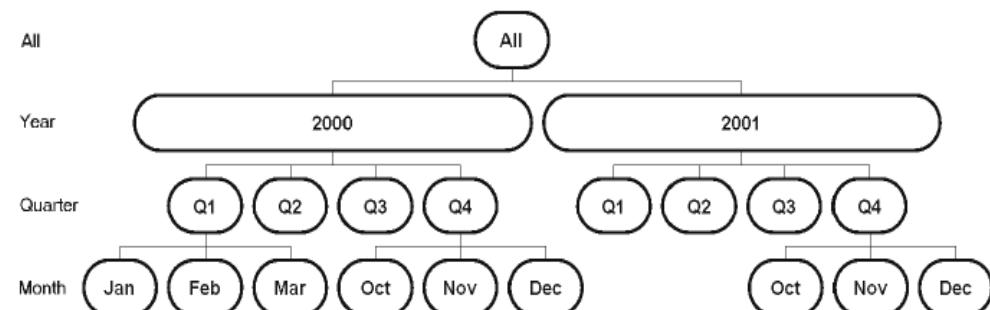


# MDX

- Designed for multidimensional data retrieval and analysis.
- Supports hierarchical navigation through dimensions.
- Enables slicing, dicing, drilling down/up, and pivoting data.
- SQL returns a subset of a 2D table. MDX returns a multidimensional subset from a cube.
- Case-insensitive but conventionally capitalizes reserved keywords.
- Square brackets [ ] are used to reference members and dimensions.
- Curly braces { } are used for defining sets.
- A Typical SQL query has SELECT, FROM and WHERE.

MDX ≠ SQL ?   SQL: Country="Spain"   MDX: Location.[Spain]

|             |                   |
|-------------|-------------------|
|             | <b>ALL (TIME)</b> |
| <b>COST</b> | 45,300,000 €      |



- **SELECT**  
  {[TIME].[ALL]} **ON COLUMNS**  
  {[Measure].[Cost]} **ON ROWS**  
**FROM** [MyCube]  
(shows costs of [HOSP].[H1], default member of HOSP)  
(also for [Dx].[Circulatory])

*(COL dimension)*  
*(ROW dimension)*

**SELECT**-> Axis or layout of the results  
**FROM**-> Refers to the Cube we query

|       | COL 1 | COL 2 | COL 3 |
|-------|-------|-------|-------|
| ROW A |       |       |       |
| ROW B |       |       |       |
| ROW C |       |       |       |

SELECT

{column headers} ON COLUMNS → SET  
{row headers} ON ROWS → SET  
FROM [cube] → name

SELECT defines the resulting set with the subset of multidimensional data from the cube.

We define the number of axes and the members from each dimension to include in each axis

```
SELECT
{[Measure].[Patient]} ON COLUMNS
{[Hospital].[Hosp1],
[Hospital]. [Hosp2],
[Hospital]. [Hosp3],
[Hospital]. [Hosp4]} ON ROWS
FROM [MyCube]
(shows a default member of TIME)
```

|               | PATIENT |
|---------------|---------|
| <b>HOSP 1</b> | 23      |
| <b>HOSP 2</b> | 65      |
| <b>HOSP 3</b> | 88      |
| <b>HOSP 4</b> | 65      |

FROM-> What is the source of the multidimensional data? A cube (restricted to 1 cube).

With LookupCube() we can bypass that restriction.

SELECT

{[Measure].[Patient]} ON COLUMNS

{[Hospital].Children} ON ROWS

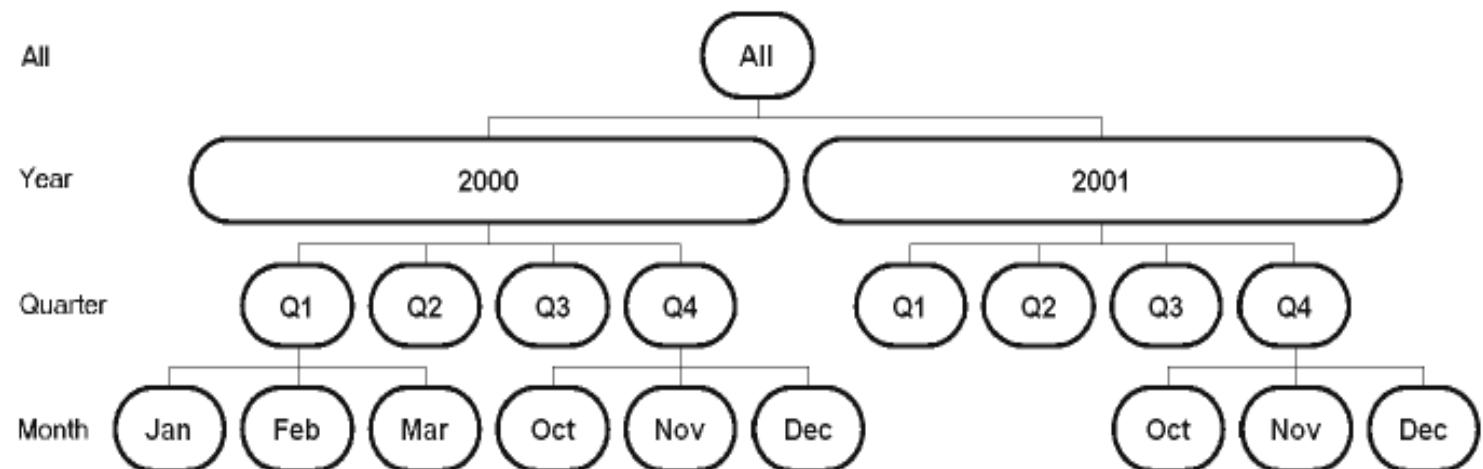
FROM [MyCube]

|               | <b>PATIENT</b> |
|---------------|----------------|
| <b>HOSP 1</b> | 23             |
| <b>HOSP 2</b> | 65             |
| <b>HOSP 3</b> | 88             |
| <b>HOSP 4</b> | 65             |

# MDX Basics

- QUESTION:  
Cost in H1,H2  
during 2000 (by Q), for  
circulatory diseases.

| Y2000 | HOSP1 | HOSP2 |
|-------|-------|-------|
| Q1    | 2M€   | 0.3M€ |
| Q2    | 3.2M€ | 0.7M€ |
| Q3    | 1.5M€ | 0.6M€ |
| Q4    | 0.4M€ | 0.5M€ |



Hint:  
cost/circulatory are default members

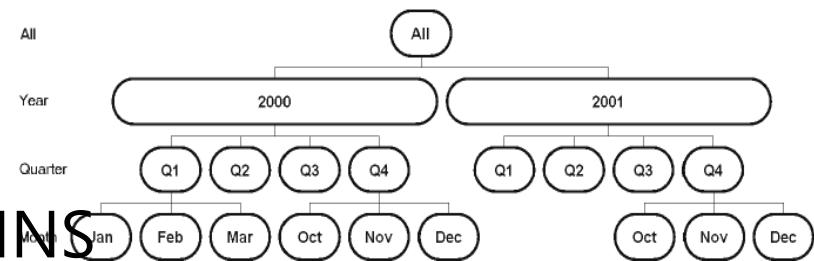
## QUESTION:

Cost in H1,H2 in 2000 (by Q), for circulatory diseases.

SELECT

```
{[Hospital].[Hosp1],
 [Hospital].[Hosp2]} ON COLUMNS
 {[Time].[All].[2000].Children} ON ROW
 FROM [MyCube]
```

| Y2000 | HOSP1 | HOSP2 |
|-------|-------|-------|
| Q1    | 2M€   | 0.3M€ |
| Q2    | 3.2M€ | 0.7M€ |
| Q3    | 1.5M€ | 0.6M€ |
| Q4    | 0.4M€ | 0.5M€ |



- **QUESTION:**  
Patients no. in H1,H2  
during 2000 (by Q), for  
circulatory diseases.

| Y2000     | HOSP1   | HOSP2  |
|-----------|---------|--------|
| <b>Q1</b> | 121 pat | 78 pat |
| <b>Q2</b> | 165 pat | 61 pat |
| <b>Q3</b> | 115 pat | 41 pat |
| <b>Q4</b> | 120 pat | 76 pat |

Hint:

Number of patients is NOT a default member

QUESTION:

Patients in H1,H2  
during 2000 (by Q), for  
circulatory diseases.

SELECT

```
{[Hospital].[Hosp1],
 [Hospital].[Hosp2]} ON COLUMNS
 {[Time].[All].[2000].Children} ON ROWS
 FROM [MyCube]
 WHERE ([Measures].[NoPat])
```

| Y2000 | HOSP1   | HOSP2  |
|-------|---------|--------|
| Q1    | 121 pat | 78 pat |
| Q2    | 165 pat | 61 pat |
| Q3    | 115 pat | 41 pat |
| Q4    | 120 pat | 76 pat |

Hint:

Number of patients is NOT a default member

## WHERE clause

Not restricted to measures.

Not restricted to 1 dimension.

It is a *SLICER/DICER*.

WHERE clause

**Not restricted to measures.**

SELECT

```
{[Hospital].[Hosp1],
 [Hospital].[Hosp2]} ON COLUMNS
{[Time].[All].[2000].Children} ON ROWS
FROM [MyCube]
WHERE ([Dx].[Respiratory])
```

| Y2000 | HOSP1 | HOSP2 |
|-------|-------|-------|
| Q1    | 1M€   | 0.4M€ |
| Q2    | 1.2M€ | 0.1M€ |
| Q3    | 0.5M€ | 0.5M€ |
| Q4    | 0.4M€ | 0.3M€ |

WHERE-> optional. Defines the slicer dimension, that we use to filter the multidimensional data.

WHERE clause

**Not restricted to 1 dimension.**

SELECT

{[Hospital].[Hosp1],  
[Hospital].[Hosp2]} ON COLUMNS

{[Time].[All].[2000].Children} ON ROWS

FROM [MyCube]

**WHERE ([Dx].[Respiratory],[Measures].[NoPat])**

| Y2000 | HOSP1   | HOSP2  |
|-------|---------|--------|
| Q1    | 61 pat  | 28 pat |
| Q2    | 75 pat  | 41 pat |
| Q3    | 105 pat | 11 pat |
| Q4    | 112 pat | 56 pat |

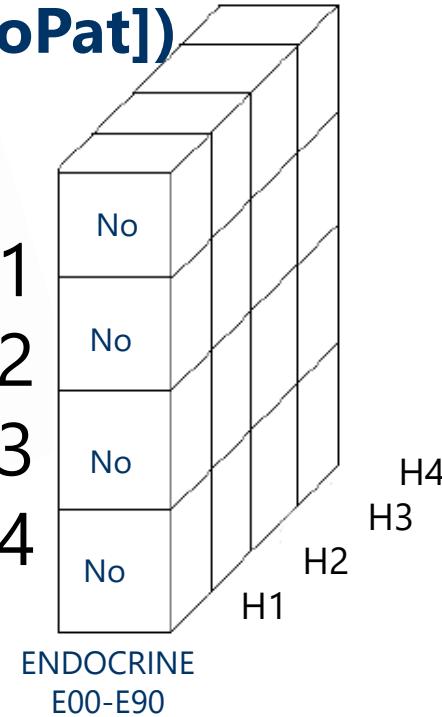
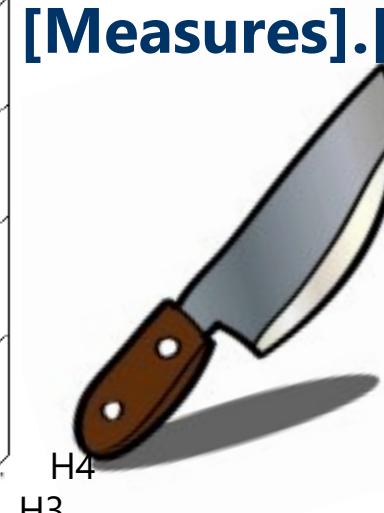
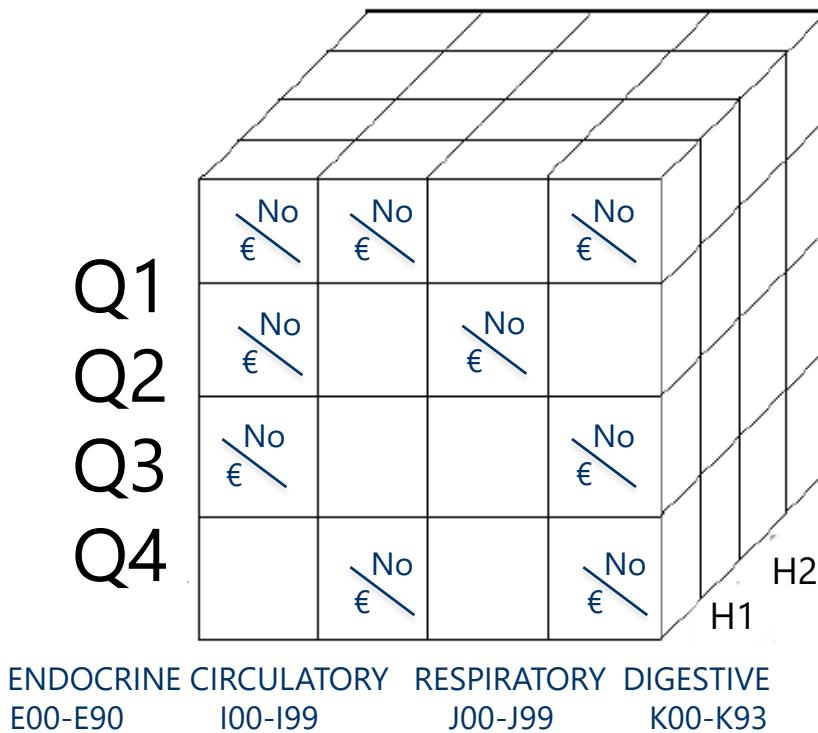
ORDERED BY Number of PATIENTS?

SELECT

```
{[Hospital].[Hosp1],
 [Hospital].[Hosp2]} ON COLUMNS
 ORDER({[Time].[All].[2000].Children},
 ([Dx].[Respiratory],[Measures].[NoPat]),BDESC) ON
 ROWS
 FROM [MyCube]
WHERE ([Dx].[Respiratory],[Measures].[NoPat])
```

- WHERE clause
  - It is a *SLICER/DICER*.

**WHERE**  
 $([Dx].[ENDOCRIN],$   
 $[Measures].[NoPat])$



# 3. MDX syntax

( ) [ ] { } .

- Brackets []
  - Dimensions: [Time]
  - Members: [2000]
- Dots .
  - Separators: [Time].[2000].[Q3]
- Parentheses ()
  - Tuples: ([DX].[Circulatory],[Hosp].[H1])

- Braces {}
  - Sets: {[Hosp].[H1], [Hosp].[H3]}  
{[Dx].Children}  
{ ([Dx].[Circ],[Hosp].[H1],[Time].[Jan]),  
([Dx].[Circ],[Hosp].[H1],[Time].[Feb]),  
([Dx].[Circ],[Hosp].[H1],[Time].[Mar]),  
([Dx].[Circ],[Hosp].[H1],[Time].[Apr]) }

**SELECT**  
  { SET } **ON COLUMNS**  
  { SET } **ON ROWS**  
**FROM** [cube]  
**WHERE** (TUPLE)

- QUESTION: Correct? Why?

```
SELECT
([Measures].[NoPatients]) ON COLUMNS,
{[Time].[2000].Children} ON ROWS
FROM [MyCube]
```

- QUESTION: Correct? Why?

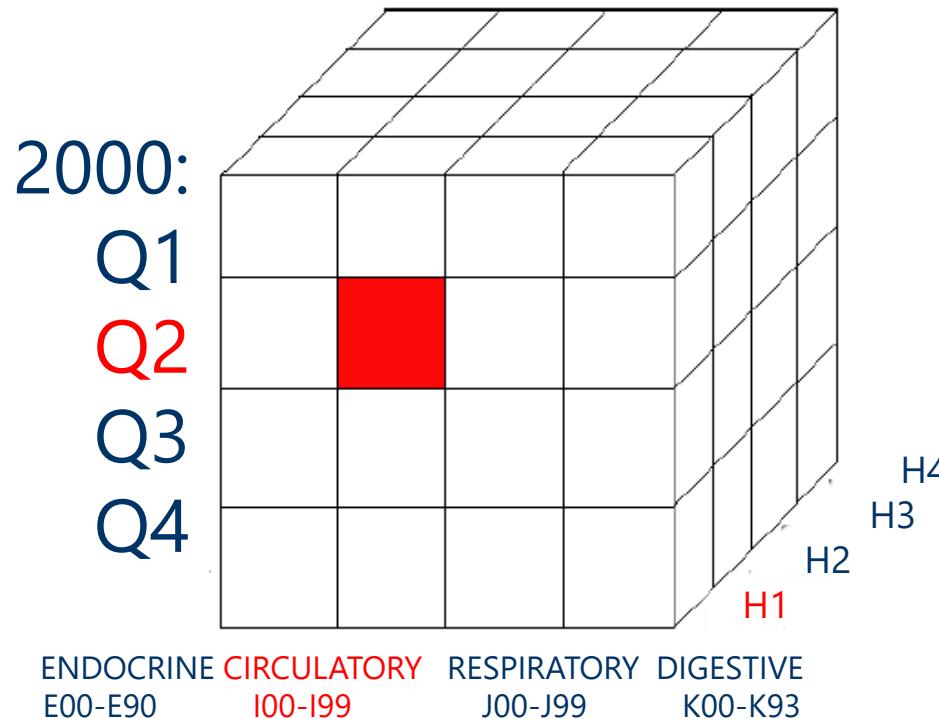
```
SELECT
{[Dx].Children} ON COLUMNS,
{[Time]. [2000].[Q1].[May].Children} ON ROWS
FROM [MyCube]
WHERE {[Measure].[cost],[Hosp].[H2]}
```

# MDX syntax

- Name of a CELL.
  - In a cube, each cell has a name.

The name of this cell is:

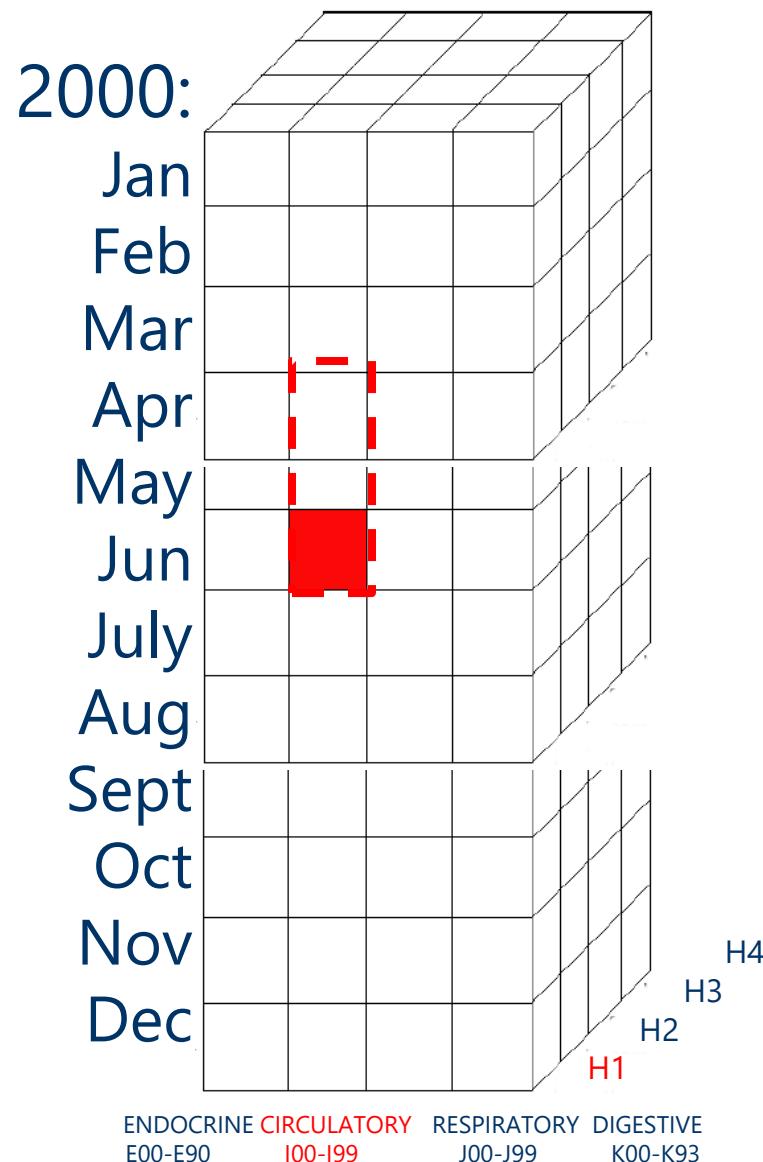
([Time].[2000].[Q2],  
[Dx].[Circulatory],  
[Hospital].[H1])



# MDX syntax

- Name of a CELL.

The name of this cell is:  
([Time].[2000].[Q2].[Jun],  
[Dx].[Circulatory],  
[Hospital].[H1])

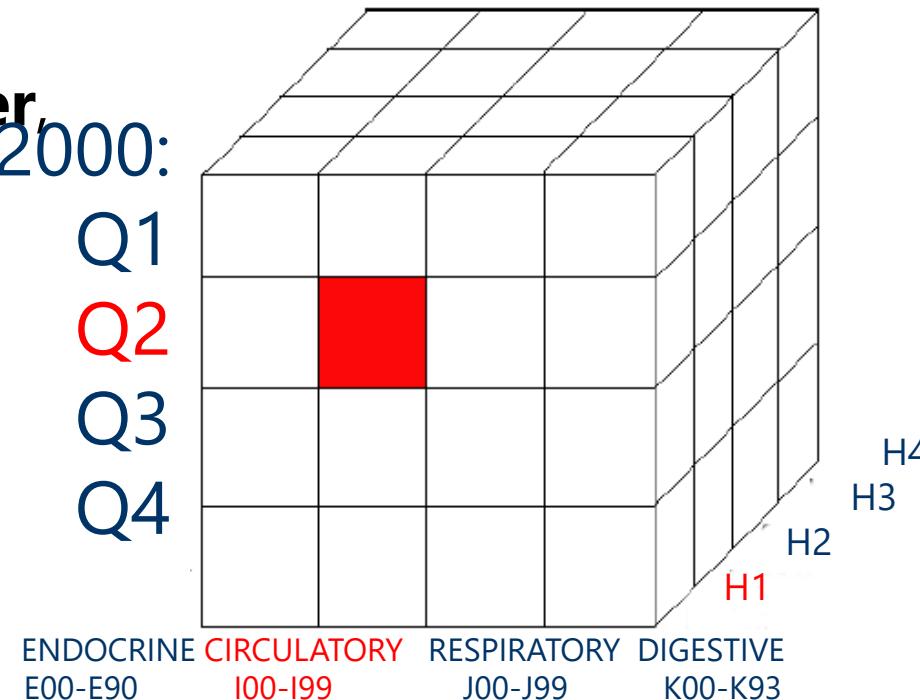


# MDX syntax

- Relative Cell Referencing:
  - CurrentMember, PrevMember, NextMember.

The name of this cell is:

**([Time].[2000].[Q3].PrevMember,**  
[Dx].[Circulatory],  
[Hospital].[H1])



- Calculated Members: +-\*/ %

"Attention improvement on circulatory patients of the 1<sup>st</sup> quarter of years 1999 and 2000".

Calculus:

([Hosp].[H1].[Dx].[Circ].[Time].[2000].[Q1],[Measures].[NoPatient])

-

([Hosp].[H1].[Dx].[Circ].[Time].[1999].[Q1],[Measures].[NoPatient])

Calculated elements bring flexibility to MDX to generate complex queries. MDX supports multiple logical and arithmetic clauses.

- Calculated Members: +-\*/ %

“Growth cost throughout year 2000 in H1 for Circulatory patients”.

```
WITH MEMBER [Measures].[Increment] AS
([Measures].[Inputs] - [Measures].[Outputs]) SELECT
{[Measures].[Inputs], [Measures].[Outputs],
[Measures].[Increment]} ON COLUMNS,
[Time].[Year].Members ON ROWS FROM [MyCube]
```

- Calculated Members: +-\*/ %

**"Growth cost throughout year 2000 in H1 for Circulatory patients".**

- Growth cost:  $\text{cost}(t) - \text{cost}(t-1)$  (increment/derivate)
- Obviate: H1 for Circulatory patients

Growth cost throughout year 2000".

| Year | Quarter   | Month    | Cost M€   | No Patient |
|------|-----------|----------|-----------|------------|
| 2000 |           |          | 57        | 280        |
|      | <b>Q1</b> |          | <b>15</b> | <b>90</b>  |
|      |           | January  | 5         | 20         |
|      |           | February | 5         | 30         |
|      |           | Mach     | 5         | 40         |
|      | <b>Q2</b> |          | <b>10</b> | <b>60</b>  |
|      |           | April    | 5         | 15         |
|      |           | Jun      | 3         | 15         |
|      |           | July     | 2         | 30         |
|      | <b>Q3</b> |          | <b>12</b> | <b>50</b>  |
|      |           | April    | 5         | 10         |
|      |           | Jun      | 5         | 10         |
|      |           | July     | 2         | 30         |
|      | <b>Q4</b> |          | <b>20</b> | <b>80</b>  |
|      |           | April    | 5         | 15         |
|      |           | Jun      | 5         | 15         |
|      |           | July     | 10        | 50         |

|      |           |          |    |           |           |
|------|-----------|----------|----|-----------|-----------|
| 2000 |           |          |    | 57        | 280       |
|      | <b>Q1</b> |          |    | <b>15</b> | <b>90</b> |
|      |           | January  | 5  | 20        |           |
|      |           | February | 5  | 30        |           |
|      |           | March    | 5  | 40        |           |
|      | <b>Q2</b> | 0        | 10 | 60        |           |
|      |           | April    | 5  | 15        |           |
|      |           | Jun      | 3  | 15        |           |
|      |           | -1       | 2  | 30        |           |
|      | <b>Q3</b> |          | 12 | 30        |           |
|      |           | April    | 5  | 10        |           |
|      |           | Jun      | 5  | 10        |           |
|      |           | July     | 2  | 30        |           |
|      | <b>Q4</b> |          | 20 | 80        |           |
|      |           | April    | 5  | 15        |           |
|      |           | Jun      | 5  | 15        |           |
|      |           | July     | 10 | 50        |           |

"Growth cost throughout year 2000".

Year      Quarter      Month      Cost M€      No Patient

|      |           |          |           |           |
|------|-----------|----------|-----------|-----------|
| 2000 |           |          | 57        | 280       |
|      | <b>Q1</b> |          | <b>15</b> | <b>90</b> |
|      |           | January  | 5         | 20        |
|      |           | February | 5         | 30        |
|      |           | Mach     | 5         | 40        |
|      | <b>Q2</b> |          | <b>10</b> | <b>60</b> |
|      |           | April    | 5         | 15        |
|      |           | Jun      | 3         | 15        |

Growth cost throughout year 2000".

[Time].CurrentMember,[Measures].[Cost]

-

[Time].CurrentMember.PrevMember,[Measures].[Cost]

|  |           |       |           |           |
|--|-----------|-------|-----------|-----------|
|  | <b>Q4</b> |       | <b>20</b> | <b>80</b> |
|  |           | April | 5         | 15        |
|  |           | Jun   | 5         | 15        |
|  |           | July  | 10        | 50        |

- Calculated Members: +-\*/ %

"**Growth** cost **throughout** year 2000 in H1 for Circulatory patients".

- ([Time].**CurrentMember**,[Measures].[Cost])  
- [Time].**CurrentMember.PrevMember**,[Measures].[Cost])

YES... WHAT IF WE FOCUS ON **THE QUARTER GROWTH** ?????

|      |    |          |    |    |     |
|------|----|----------|----|----|-----|
| 2000 |    |          |    | 57 | 280 |
|      | Q1 |          |    | 15 | 90  |
|      |    | January  | 5  | 5  | 20  |
|      |    | February | -5 | 5  | 30  |
|      |    | Mac      | 5  | 5  | 40  |
|      | Q2 |          | 10 | 10 | 60  |
|      |    | April    | 5  | 5  | 15  |
|      |    | Jun      | 3  | 3  | 15  |
|      |    | July     | 2  | 2  | 30  |
|      | Q3 |          | 12 | 12 | 50  |
|      |    | Apr      | 5  | 5  | 10  |
|      |    | Jun      | 5  | 5  | 10  |
|      | Q4 | 8        | 2  | 2  | 30  |
|      |    | July     | 20 | 20 | 80  |
|      |    | April    | 5  | 5  | 15  |
|      |    | Jun      | 5  | 5  | 15  |
|      |    | July     | 10 | 10 | 50  |

- Calculated Members: +-\*/ % AVG SUM ...  
**"Growth cost throughout year 2000 in H1 for Circulatory patients".**
- ([Time].**CurrentMember**,[Measures].[Cost])  
- [Time].**CurrentMember.PrevMember**,[Measures].[Cost])

YES... WHAT IF WE FOCUS ON **THE SEMESTER GROWTH !!!**

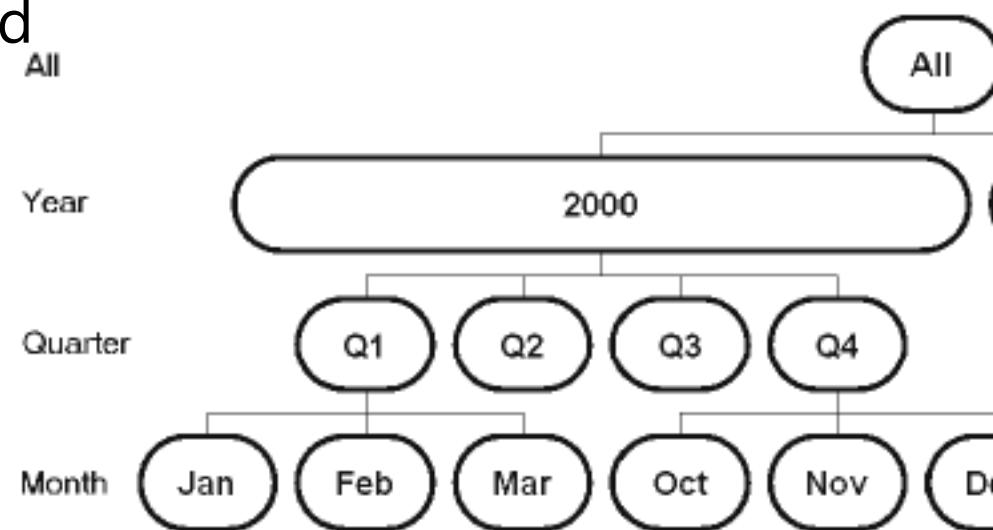
... **SAME EXPRESSION. THAT'S THE COOL THING. DEPENDS ON THE TIME DIMENSION, WHICH DEFINES THE CURRENT MEMBER PROPERTY**

- OTHER FUNCTIONS:
  - Sum (X)→Number : Sums all members of X
  - X.Lag(N) : N positions back from X.
  - X.Lead(M) : M position forward from X.
  - YTD(X)→Set : YearToDate: Members of the Year until member X.  
e.g. YTD(March) → {Jan, Feb, March}

ClosingPeriod, OpeningPeriod, ParallelPeriod,

# MDX syntax

- Hierarchy navigation:
  - *Member.Children*
  - *Member.Parent*
  - *Member.FirstChild / LastChild*
  - *Descendants(X,n)*
  - *Ancestors(X,n)*
  - *Siblings / Cousins*
  - ~~Aunt / Uncle~~



## Security

As with SQL, MDX can be vulnerable to MDX Injection

Do we want to prevent unauthorized access to the cube?

Do we want to prevent unauthorized modifications?

If we can inject in the first dimension of SELECT we can write a full custom query... very dangerous.

If the injection is in WHERE, blind injection can be used.

We have to sanitize the inputs, implement least privilege, testing for injection vulnerabilities, audit, etc.

- Bibliography and Resources:
- Mark Whitehorn et al. Fast Track to MDX (2<sup>nd</sup> Ed). Springer. 2004.
- Microsoft, "Key Concepts in MDX (Analysis Services)",  
[https://docs.microsoft.com/en-us/analysis-  
services/multidimensional-models/mdx/key-concepts-  
in-mdx-analysis-services?view=asallproducts-  
allversions](https://docs.microsoft.com/en-us/analysis-services/multidimensional-models/mdx/key-concepts-in-mdx-analysis-services?view=asallproducts-allversions)
- InterSystems, "Introduction to MDX Queries",  
[https://docs.intersystems.com/irislatest/csp/docbook/DocBook.UI.Page.cls?KEY=D2GMDX\\_CH\\_MDX\\_INTRO](https://docs.intersystems.com/irislatest/csp/docbook/DocBook.UI.Page.cls?KEY=D2GMDX_CH_MDX_INTRO)

# SCORECARDS & DASHBOARDS

un dashboard es menos transversal, mas focalizado !!!!

Unit 3 – Data exploitation. Query languages  
and visualization

S3-3 – SCORECARDS & DASHBOARDS



V S  
O T



## KEY CONCEPTS:

importante para el examen !!!!!

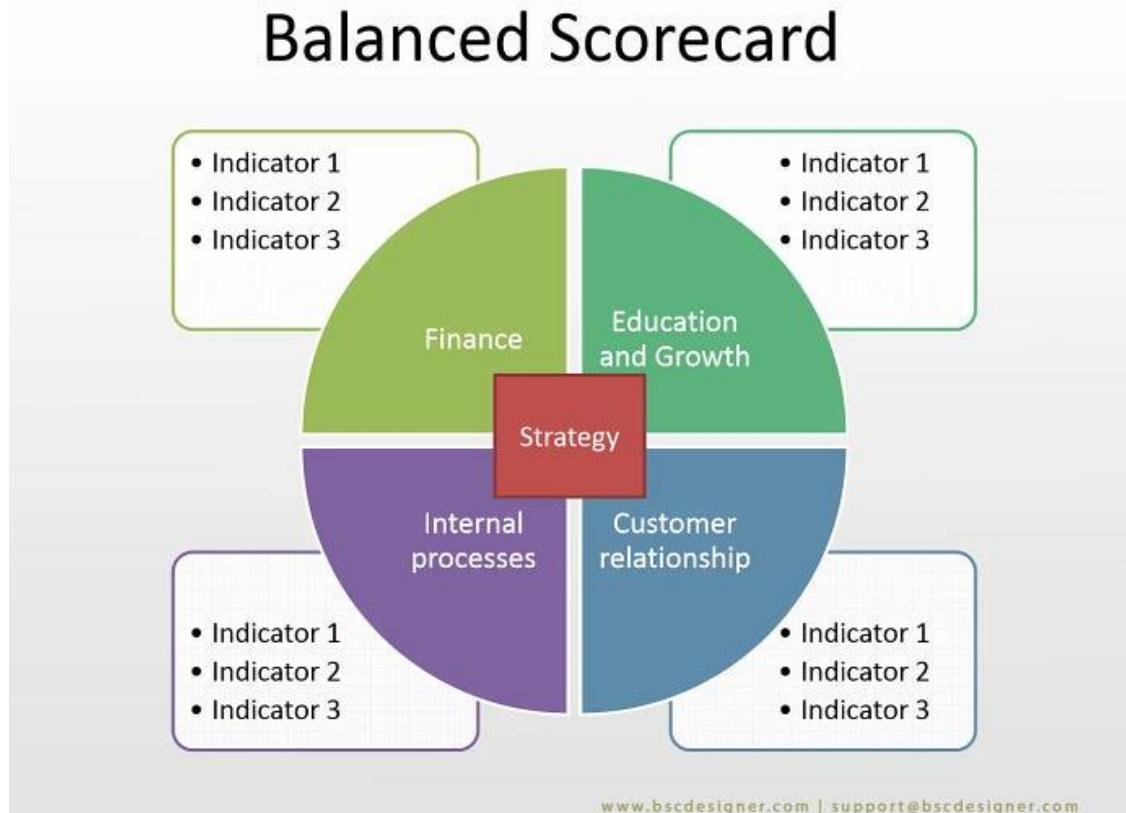
- **Vision:** Vision provides a shared mental framework for everyone in the organization, defining a clear picture of the desired future state and long-term aspirations.
- **Strategy:** the set of priorities and plans an organization adopts to pursue its mission, considering the operating environment.
- **Objective:** a concise statement describing the specific things organization must do well in order to execute a strategy.  
especificos pero no mucho, eso son los target
- **Target:** A target is the specific, desired value or outcome associated with a performance measure, indicating success in achieving objectives.

## OUTLINE:

1. Balanced Scorecards (BSC)
2. Strategy Maps
3. Example: BSC in Healthcare
4. Dashboards Essentials
5. Scorecards vs. Dashboards
6. Examples of Dashboards: IT company and ICU.

perspectivas que implementaremos en nuestro cuadro de mando integral

# 1. BALANCED SCORECARDS (BSC)



- BALANCED SCORECARDS (BSC)

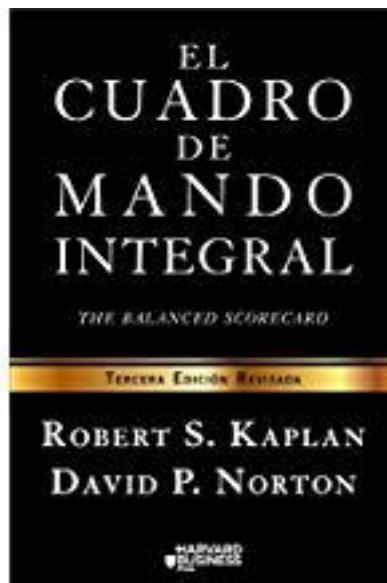
- TOP-DOWN methodology in organizations to align organizational objectives and activities with strategy.
- **Tool to manage and monitor long-term strategy.**
- It is NOT: tool to design a strategy.
- It is: tool to move strategy into actionable objectives.



- **BALANCED SCORECARDS (BSC)**    englobar todos los elementos en uno

“A carefully selected set of quantifiable measures derived from an organization’s strategy”.

By Robert Kaplan (Harvard Univ.) & David Norton (consultant Boston area)  
BOOK: The Balanced Scorecard (in 1996)



- BALANCED SCORECARDS (BSC)
- Scorecard can be used as:
  - measurement system. Monitoring KPIs aligned with strategic goals.
  - strategic management system. Aligning organization vision and strategy and operations.
  - communication tool across all levels of the organization.

## HOW TO DESIGN A BALANCED SCORECARD:

Know **guiding principles**: objectives & strategy.

Split the strategy objectives into **actions**.

Strategy Objective Actions from **4 perspectives**:



Financial



Customer



Internal-Business-Processes



Learning & Growth.

## HOW TO DESIGN A BALANCED SCORECARD:

- Strategy **Scores**: measure progress towards achieving each strategy. *"If you cannot measure something, you cannot manage it"*.
- Target: the score value that we expect to reach.
- Program actions: specific short-term actions that must be done **to reach a target**. Tactics.

## HOW TO DESIGN A BALANCED SCORECARD:

- Strategy **Scores**, Target, Program Actions.



## HOW TO DESIGN A BALANCED SCORECARD:

- Strategy **Scores**, Target, Program Actions.



## HOW TO DESIGN A BALANCED SCORECARD:

- Strategy **Scores**, Target, Program Actions.



## HOW TO DESIGN A BALANCED SCORECARD:

- Strategy **Scores**, Target, Program Actions.



## HOW TO DESIGN A BALANCED SCORECARD:

1. Know **guiding principles**: objectives & strategy.
2. Split the strategy objectives into **actions**.
3. Strategy Objective Actions from **4 perspectives**:



Financial



Customer



Internal-Business-Processes



Learning & Growth.



## HOW TO DESIGN A BALANCED SCORECARD:

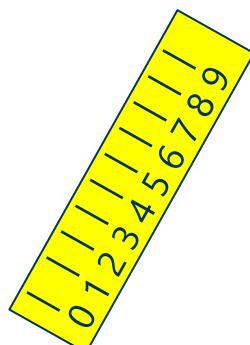
- Strategy Objective Actions from **4 perspectives**:
  - **FINANCIAL DIMENSION**:



- A scorecard must encourage including the financial goal in the organization strategy.
- The financial goals matches to the phases of the organization life-cycle:
  - Grow, Maintenance, Production.

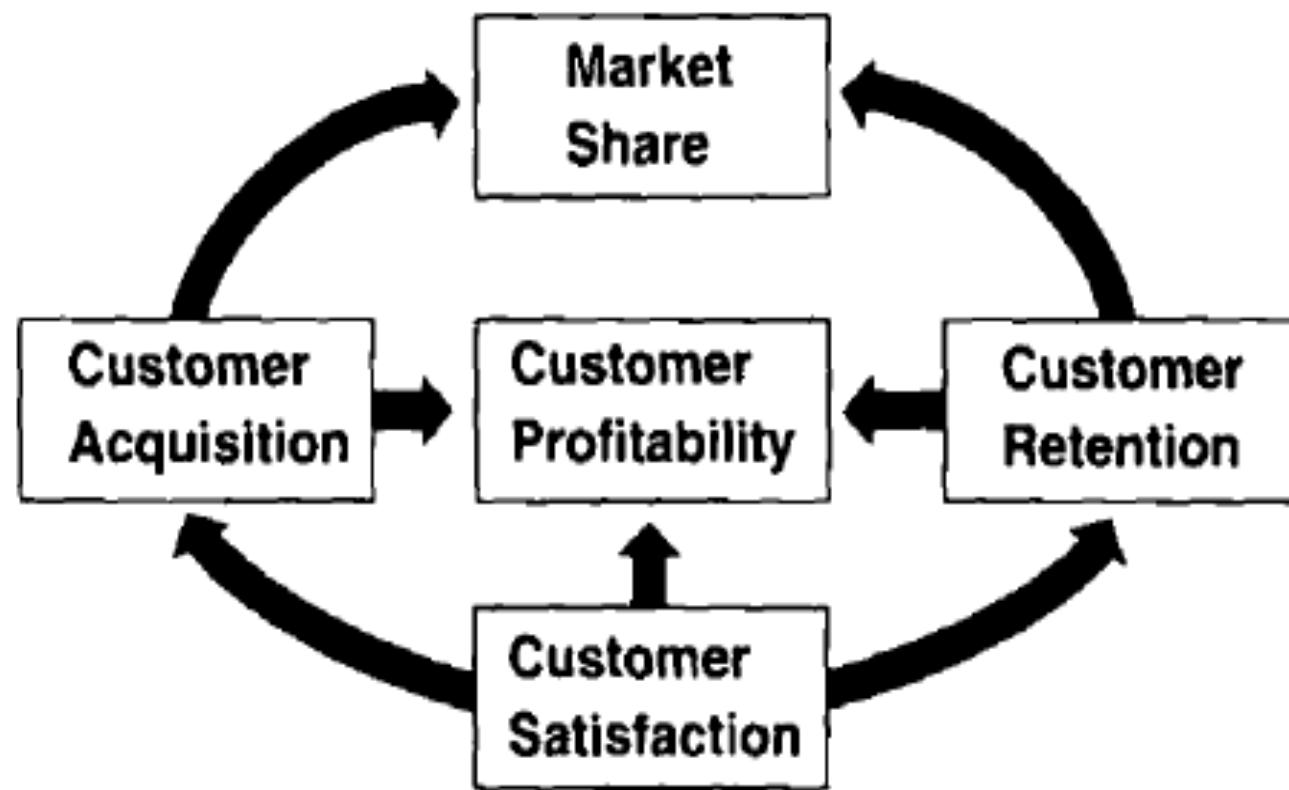
## HOW TO DESIGN A BALANCED SCORECARD:

- Strategy Objective Actions from **4 perspectives**
  - **CUSTOMER DIMENSION:**
    - Identify those **segments of clients and markets** they want to compete.
    - **Key scores** of clients:
      - Market and Account Share (ES: *Cuota Mercado*).
      - Customer Retention: maintain and increase the market based on customer segments.
      - Customer Acquisition: increase customer base in targeted segments.
      - **Customer satisfaction.**
      - Customer profitability (ES: *rentabilidad*)



## HOW TO DESIGN A BALANCED SCORECARD:

- Strategy Objective Actions from **4 perspectives**:
  - **CUSTOMER DIMENSION:**



## HOW TO DESIGN A BALANCED SCORECARD:

- Strategy Objective Actions from **4 perspectives**:
    - **CUSTOMER DIMENSION**: Customer Satisfaction:
    - Time: major competitive weapon: respond rapidly and reliably to customer requests (TOYOTA Japanese manufacturers).
    - Quality.
    - Price: whether customers care about current price.
- 

## HOW TO DESIGN A BALANCED SCORECARD:

- Strategy Objective Actions from **4 perspectives:**

**INTERNAL BUSINESS-PROCESS DIMENSION:**

- Once the financial and customer goals have been developed
- The main internal processes are:
  1. Innovation processes
  2. Operations processes
  3. Post-sale process



## • INTERNAL BUSINESS-PROCESS DIMENSION



Jpatrick McCann [Follow](#)

Director of Regional Events at C2Eventz



25



1



6

We offer three kinds of service  
**GOOD-CHEAP-FAST**



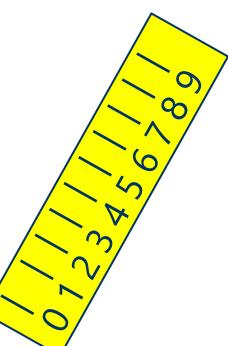
## HOW TO DESIGN A BALANCED SCORECARD:

- Strategy Objective Actions from **4 perspectives**:

**Learning & Growth DIMENSION:**

Goals to follow in order to induce new infrastructures to improve aspects of previous dimensions.

- Invest in the future.
- Key scores:
  1. Employee capacities.
  2. Information system capacities.
  3. Motivation, power, delegation and objective coherency.



## HOW TO DESIGN A BALANCED SCORECARD:

- EXAMPLE 1: 'Kenyon Stores'
  - **Financial**: aggressive growth , maintain overall margins.
  - **Customer**: customer loyalty , complete product-line offering.
  - **Internal Business Process**: Build the brand, fashion leader, quality product, superior shopping experience.
  - **Learning and Growth**: strategic skills, personal growth.

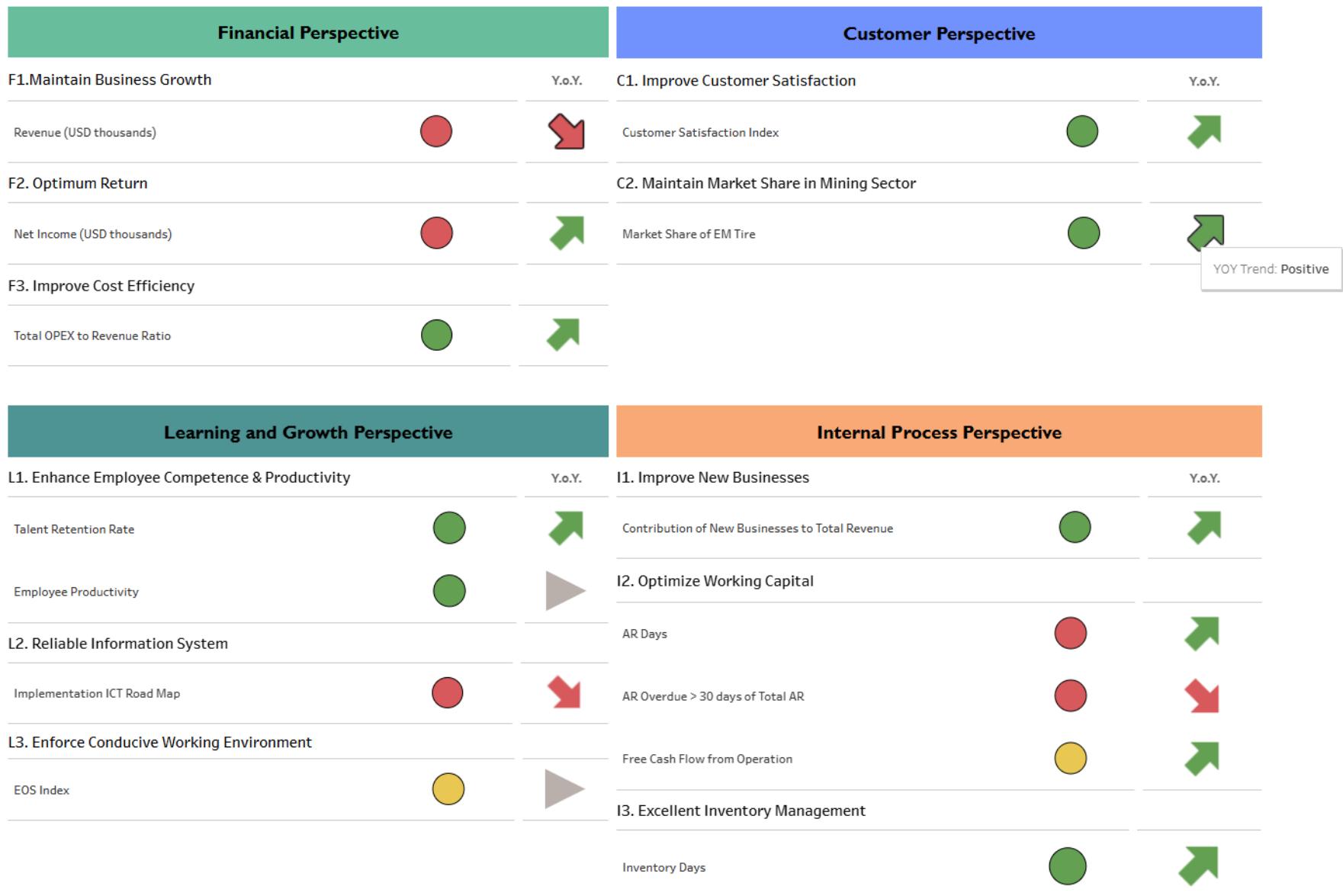
| Dimensions     | Guiding principles | Scores | Target | Program actions |
|----------------|--------------------|--------|--------|-----------------|
| Financial      |                    |        |        |                 |
| Client         |                    |        |        |                 |
| Internal Proc. |                    |        |        |                 |
| Learn&Grow     |                    |        |        |                 |

# SCORECARDS & DASHBOARDS



## Scorecard

Click on any item of interest to reveal the menu



## 2. STRATEGY MAPS

The Balanced Scorecard is a step in a continuum that describes what value is and how it is created

*Translating Mission into Strategic Outcomes*



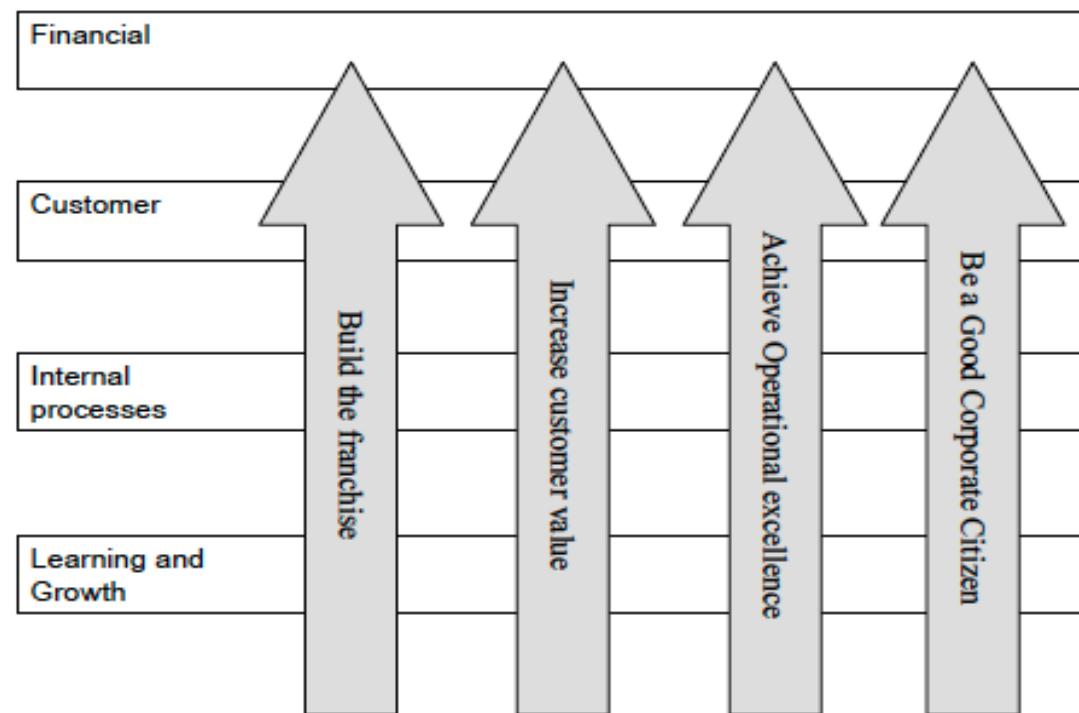
Source: Strategy Maps, Kaplan and Norton, 2004

## STRATEGY MAPS:

- **Measures** must be chosen **more “strategically”**.
- **Relate** specific high level objectives
- Causal relationship between objectives indicated in the “Strategy Map”.

# SCORECARDS & DASHBOARDS

# STRATEGY MAPS:



**Figure 2.1: Architecture of a Strategy Map**  
(Source: Kaplan and Norton 2001, p. 79)

# SCORECARDS & DASHBOARDS

## STRATEGY MAPS:

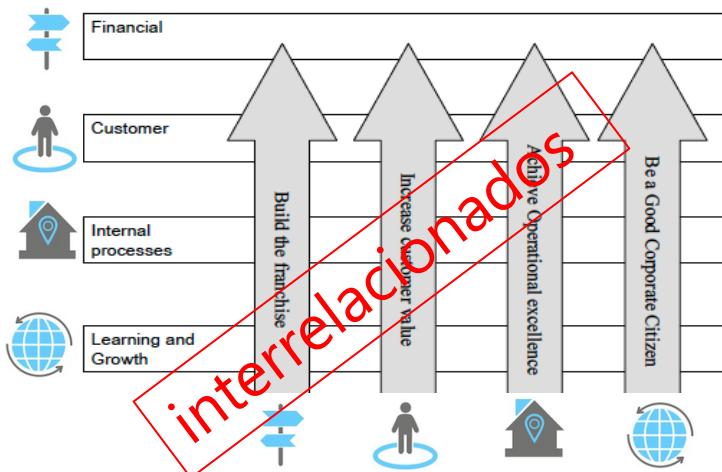
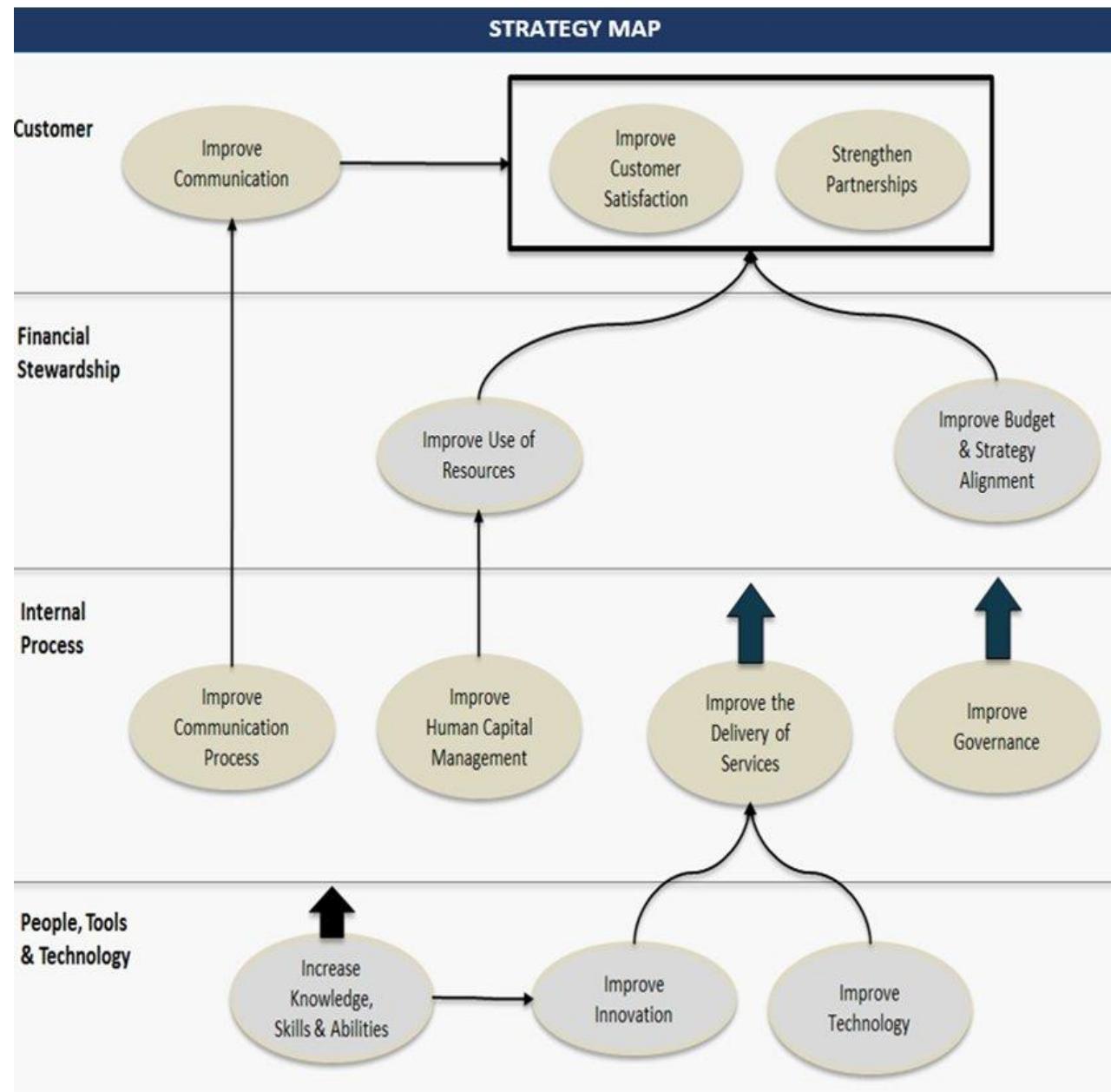
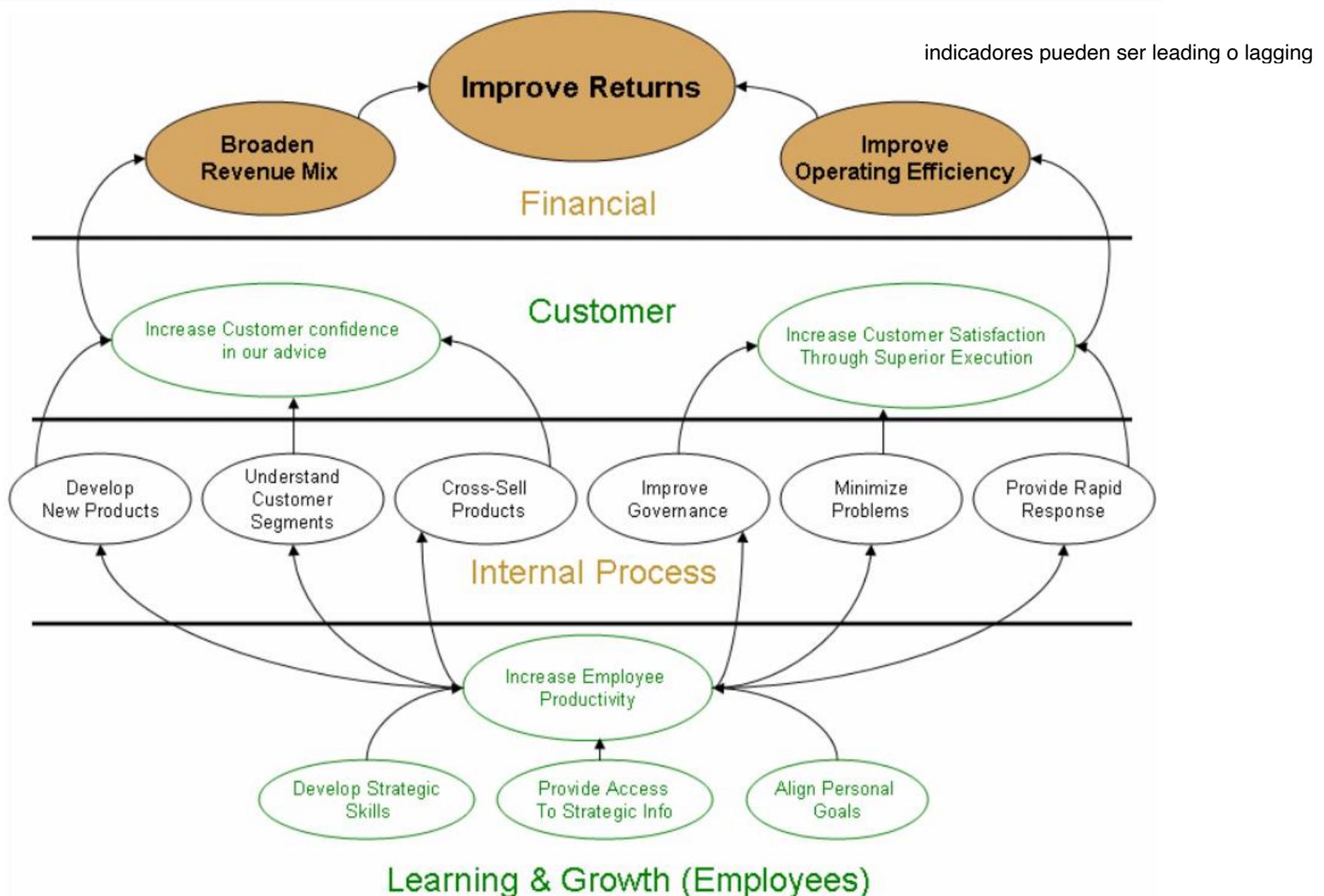


Figure 2.1: Architecture of a Strategy Map  
(Source: Kaplan and Norton 2001, p. 79)



Source: fdic.gov

# SCORECARDS & DASHBOARDS



## 3. EXAMPLE: BSC IN HEALTHCARE

- BSC in Healthcare

**To what extent has the BSC been introduced to healthcare:**

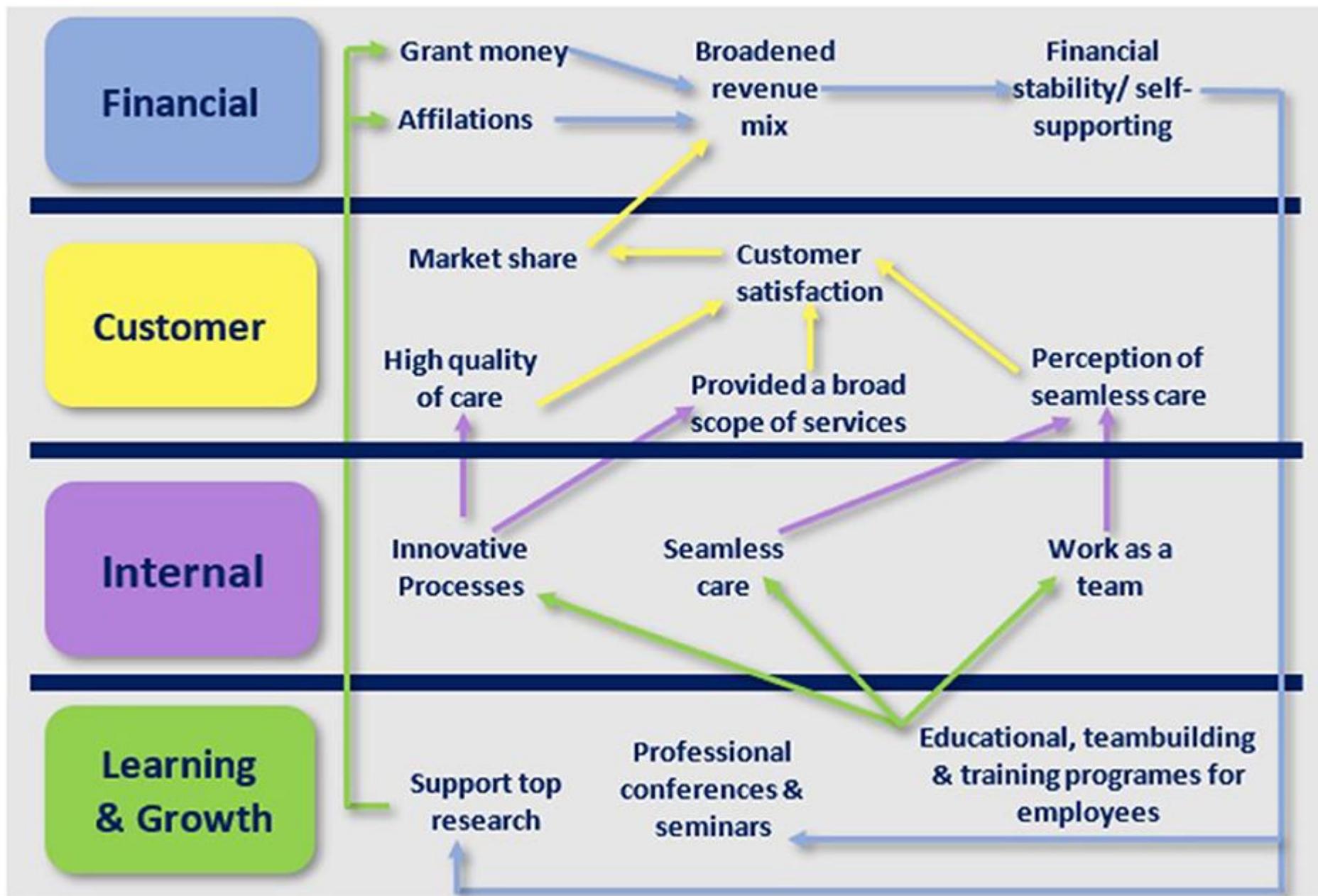
Hospitals, Healthcare Systems, University medical depts., long-term care, mental health centers, pharmaceutical care, health insurance companies.

BSC is the most adopted model to measure performance.

- BSC in Healthcare

## Reasons to use BSC in healthcare (diversity):

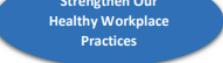
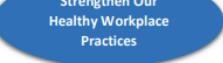
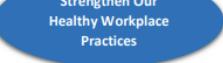
1. To ensure to be a **high performing** healthcare provider (Northumbria Healthcare Foundation, United Kingdom).
2. How can we demonstrate to the community that they are getting **value for our tax payer** funded services? (>AUS\$ 1 billion). (Hunter New England Health District, Australia).
3. To take a more strategic approach to differentiate their services and **attract more business** (Mackay Memorial Hospital, Taiwan).
4. To achieve **better outcomes** for patients and staff (St Vincent's Private Hospital, Australia).
5. To combine **financial control** with quality improvement (Högland Hospital, Sweden).



# SCORECARDS & DASHBOARDS

## ejemplo de BSC

### Waypoint Strategic Balanced Scorecard 2020-25 (Year 4 - 2023-24)

| MISSION                                                                                                                                                                                                                                                                                               | We are a Catholic hospital committed to providing excellence in specialized mental health and addictions services grounded in research and education and guided by faith-based values.    |                                                                                     |                                                                                     |                                                                                       |                                                                                                                                                      |                                                                                                                                                               |            |           |                   |                                                                                                                                                                                                                             |             |  |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|-----------|-------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|--|
| VISION                                                                                                                                                                                                                                                                                                | As an inspired organization, we will change lives by leading the advancement and delivery of compassionate care.                                                                          |                                                                                     |                                                                                     |                                                                                       |                                                                                                                                                      |                                                                                                                                                               |            |           |                   |                                                                                                                                                                                                                             |             |  |
| STRATEGIC DIRECTIONS                                                                                                                                                                                                                                                                                  | SERVE                                                                                                                                                                                     | DISCOVER                                                                            |                                                                                     |                                                                                       |                                                                                                                                                      | LEAD                                                                                                                                                          |            |           |                   |                                                                                                                                                                                                                             |             |  |
| STRATEGIC RESULTS                                                                                                                                                                                                                                                                                     | We will include patients and families as partners in all we do, fostering a healing culture where staff, physicians, and volunteers are inspired to provide exceptional service and care. |                                                                                     |                                                                                     |                                                                                       |                                                                                                                                                      | We will embrace education, advance research, and seek, generate, and apply best practice and new knowledge to create the best possible outcomes for patients. |            |           |                   |                                                                                                                                                                                                                             |             |  |
| OBJECTIVES & STRATEGY MAP<br>(read from bottom to top)                                                                                                                                                                                                                                                |                                                                                                                                                                                           | MEASURE<br><small>*Quality Improvement Plan indicator</small>                       | BASELINE<br>Q3 2022-23                                                              | TARGET<br>2023-24                                                                     | Q1<br>YTD unless indicated with ^                                                                                                                    | Q2                                                                                                                                                            | Q3         | Q4        | TARGET<br>2020-25 | 2023-24 INITIATIVES**<br><small>Initiatives not directly responsible for the measure listed to the left</small>                                                                                                             |             |  |
| <b>FIDUCIARY PERSPECTIVE:</b> If we succeed, how will we look to funders or donors?                                                                                                                                                                                                                   |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       |                                                                                                                                                      |                                                                                                                                                               |            |           |                   | Develop regional integrated care pathway through<br>■ the Central Ontario Specialized Health Networks for adult depression and anxiety                                                                                      |             |  |
|                                                                                                                                     |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Increase % eligible programs demonstrating improvements in patient health outcomes through the use of standardized measures (e.g. Composite Index) ^ | ■ 55%                                                                                                                                                         | ■ 63-65%   | ■ 64%     | ■ 64%             | ■ 18%                                                                                                                                                                                                                       | ■ 80-85%    |  |
|                                                                                                                                     |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Decrease repeat Emergency Department visits (30 days return visit) for mental health and addictions ^                                                | ■ 21.4%<br>(Q2 2022-23)                                                                                                                                       | ■ 19.8%    | ■ 22.0%   | ■ 22.5%           | ■ 22.1%                                                                                                                                                                                                                     | ■ 18.5%     |  |
|                                                                                                                                 |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Decrease Alternate Level of Care (ALC) Days for regional programs (NEW)                                                                              | ■ 27.8%                                                                                                                                                       | ■ 25.1%    | ■ 18.0%   | ■ 18.0%           | ■ 18.8%                                                                                                                                                                                                                     | ■ 22.4%     |  |
|                                              |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | *Maintain total margin ~                                                                                                                             | ■ 2.99%                                                                                                                                                       | ■ > 0      | ■ (1.87%) | ■ (0.20%)         | ■ (3.38%)                                                                                                                                                                                                                   | ■ > 0       |  |
| <b>PATIENTS, FAMILIES, PARTNERS PERSPECTIVE:</b> To achieve our vision, how must we look to our patient, families, and partners? What do they want? How will we satisfy them? How will we serve them?                                                                                                 |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       |                                                                                                                                                      |                                                                                                                                                               |            |           |                   | Implement coordinated access: Central Waitlist Management Service<br>Continue work on new 20 bed acute mental health unit on Toanche Level 3<br>Advance urgent and emergent mental health services regionally               |             |  |
|                                                                                                                                     |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | *Increase overall inpatient satisfaction                                                                                                             | ■ 70%                                                                                                                                                         | ■ 75%      | ■ n/a     | ■ n/a             | ■ n/a                                                                                                                                                                                                                       | ■ 84%       |  |
|                                                                                                                                 |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Decrease reported patient incidents per 1000 patient days (Severity 2-4)                                                                             | ■ 12.54                                                                                                                                                       | ■ 9.87     | ■ 11.10   | ■ 12.23           | ■ 12.46                                                                                                                                                                                                                     | ■ 8.98      |  |
| <b>INTERNAL PROCESSES PERSPECTIVE:</b> To satisfy our patients, families, partners, funders, donors, and our mission, what processes must we excel at? What are the few things we need to do better, from amongst our many processes, that will make the biggest difference?                          |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       |                                                                                                                                                      |                                                                                                                                                               |            |           |                   | Implement Model of Care<br>Implement Six Core Strategies to prevent restraint & seclusion                                                                                                                                   |             |  |
|                                                                                                                                 |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Number of clients enrolled in Ontario Structured Psychotherapy (@Waypoint) (NEW)                                                                     | ■ 1184<br>(93% YTD)                                                                                                                                           | ■ 2537     | ■ 440     | ■ 1013            | ■ 1583                                                                                                                                                                                                                      | ■ 2537      |  |
| <b>LEARNING &amp; GROWTH PERSPECTIVE:</b> To achieve our vision, how will we build capability for our people to learn and grow, communicate and work together? What skills, knowledge, culture, behaviours, values technology, capability or capacity do we have to grow or learn as an organization? |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       |                                                                                                                                                      |                                                                                                                                                               |            |           |                   | Participate in Pursuing Equity Learning Network (Institute for Healthcare Improvement)<br>Develop Human Capital Management System (phase 1)                                                                                 |             |  |
|                                              |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Reduce levels of medium to high staff burnout                                                                                                        | ■ 89%                                                                                                                                                         | ■ 74%      | ■ n/a     | ■ n/a             | ■ n/a                                                                                                                                                                                                                       | ■ 70%       |  |
|                                              |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | *Decrease workplace violence frequency (lost time claims per 100 full time equivalents)                                                              | ■ 2.8                                                                                                                                                         | ■ 1.5      | ■ 2.7     | ■ 3.6             | ■ 3.9                                                                                                                                                                                                                       | ■ 1.2       |  |
|                                              |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | *Decrease workplace violence severity (lost time claims per 100 full time equivalents)                                                               | ■ 19.2                                                                                                                                                        | ■ 25       | ■ 8.4     | ■ 56.7            | ■ 59.2                                                                                                                                                                                                                      | ■ 22        |  |
|                                              |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Increase research projects with patient involvement^ (cumulative since 2019-20)                                                                      | ■ 3                                                                                                                                                           | ■ 3        | ■ 5       | ■ 5               | ■ 5                                                                                                                                                                                                                         | ■ 5         |  |
|                                              |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       |                                                                                                                                                      |                                                                                                                                                               |            |           |                   | Develop regional integrated care pathway for schizophrenia: Health Quality Ontario quality standards in the hospital and community<br>Expand research training<br>Refresh Long Term Master Plan<br>Implement patient portal |             |  |
|                                              |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Increase annual peer reviewed publications (cumulative) (NEW)                                                                                        | ■ 76                                                                                                                                                          | ■ 98 - 101 | ■ 92      | ■ 98              | ■ 112                                                                                                                                                                                                                       | ■ 120 - 126 |  |
|                                              |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Increase number of quality statements implemented (cumulative) (NEW)                                                                                 | ■ 5                                                                                                                                                           | ■ 14       | ■ 5       | ■ 9               | ■ 9                                                                                                                                                                                                                         | ■ 30        |  |
|                                              |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Increase % of Electronic Medical Record Analytics Maturity (EMRAM) standards met                                                                     | ■ 86%                                                                                                                                                         | ■ 100%     | ■ 99%     | ■ 99%             | ■ 99%                                                                                                                                                                                                                       | ■ 100%      |  |
|                                              |                                                                                                                                                                                           |                                                                                     |                                                                                     |                                                                                       | Measures relate to Strategic Plan, Service Accountability Agreements, Quality Improvement Plan                                                       |                                                                                                                                                               |            |           |                   | ~ Total Margin target parameters differ                                                                                                                                                                                     |             |  |
| <b>VALUES</b>                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                           |  |  |  |                                                                 | Within 5% of Target      Between 5 & 10%      >10% from Target                                                                                                |            |           |                   |                                                                                                                                                                                                                             |             |  |

## 4. DASHBOARDS ESSENTIALS

- DASHBOARDS



# SCORECARDS & DASHBOARDS

- DASHBOARDS



FROM: ProjectManager.com

- DASHBOARDS
  - WHAT IS A DASHBOARD?
  - Something called a dashboard :
    - Includes **graphical display mechanisms**: traffic lights, gauges, meters...etc.
    - **Overviews something** going on in the business.

- DASHBOARDS

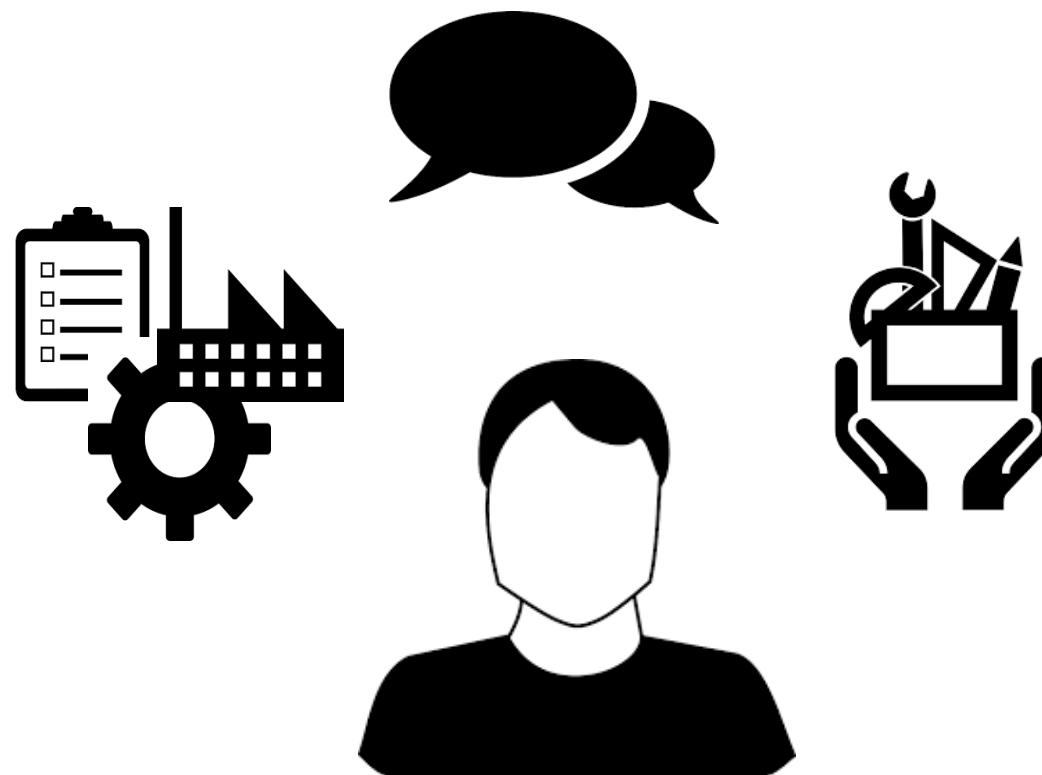
- WHAT IS A DASHBOARD?
- DEFINITION:

'It is a **visual display** of the most **important information** needed to achieve one or more **objectives**; arranged on a single screen so the information can be **monitored at a glance**'

*Stephen Few, "Dashboard Confusion,"  
Intelligent Enterprise, March 20, 2004*

- DASHBOARDS
  - According to this definition:
    1. Dashboards display the information **to achieve an objective**.
    2. Dashboards **fits** on a single computer **screen**.
    3. Dashboards used to **monitor** information **at a glance**.

- DASHBOARDS
  - So... a Dashboard:  
communication tool, good understanding, good design.



- DASHBOARDS
  - Some (obvious) Dashboard principles:
    1. Display right data to the right audience
    2. Right dashboard: adapt the dashboard model to the context
    3. Easy to find: keep the dashboard tidy
    4. Only essentials
    5. Perfection in DB design can never be achieved

- DASHBOARDS
  - Categorization of dashboards:
    - By ROLE: Strategic, Analytical or Operational.
    - By TYPE OF DATA: Quantitative, Qualitative
    - By DOMAIN: Sales, Finance, Marketing, Manufacturing
    - By TYPE OF MEASURE: **BSC**, *Six Sigma*, ...
    - By MECHANISMS OF DISPLAY: Graphical, text, integrated
    - ...etc.

- DASHBOARDS
  - Aspects to consider
    - More KPIs don't mean a better dashboard.
    - Functionality has higher priority than aesthetics.
    - Data Quality is usually lower than expected.
    - Align it with business processes.
    - Take into account the data context.
    - Update!
    - Train and communicate with your intended users.
    - Now, it is easy to incorporate interactivity.
    - Always define the purpose of the dashboard.

## 5. SCORECARDS vs. DASHBOARDS

- SCORECARDS vs. DASHBOARDS
- BOTH:
  - Tools for **supporting management** in companies
  - Support **data-driven decision making**
  - Can display **KPI** (key performance indicators)
- KEY DIFFERENCES:
  - Methodology, users, level of details, timing

- KEY DIFFERENCES:
  - CAR METAPHOR

el gps seria el bsc (a largo plazo, me permite llegar a mi destino) y el caudro de mando del salpicadero seria el dashboard

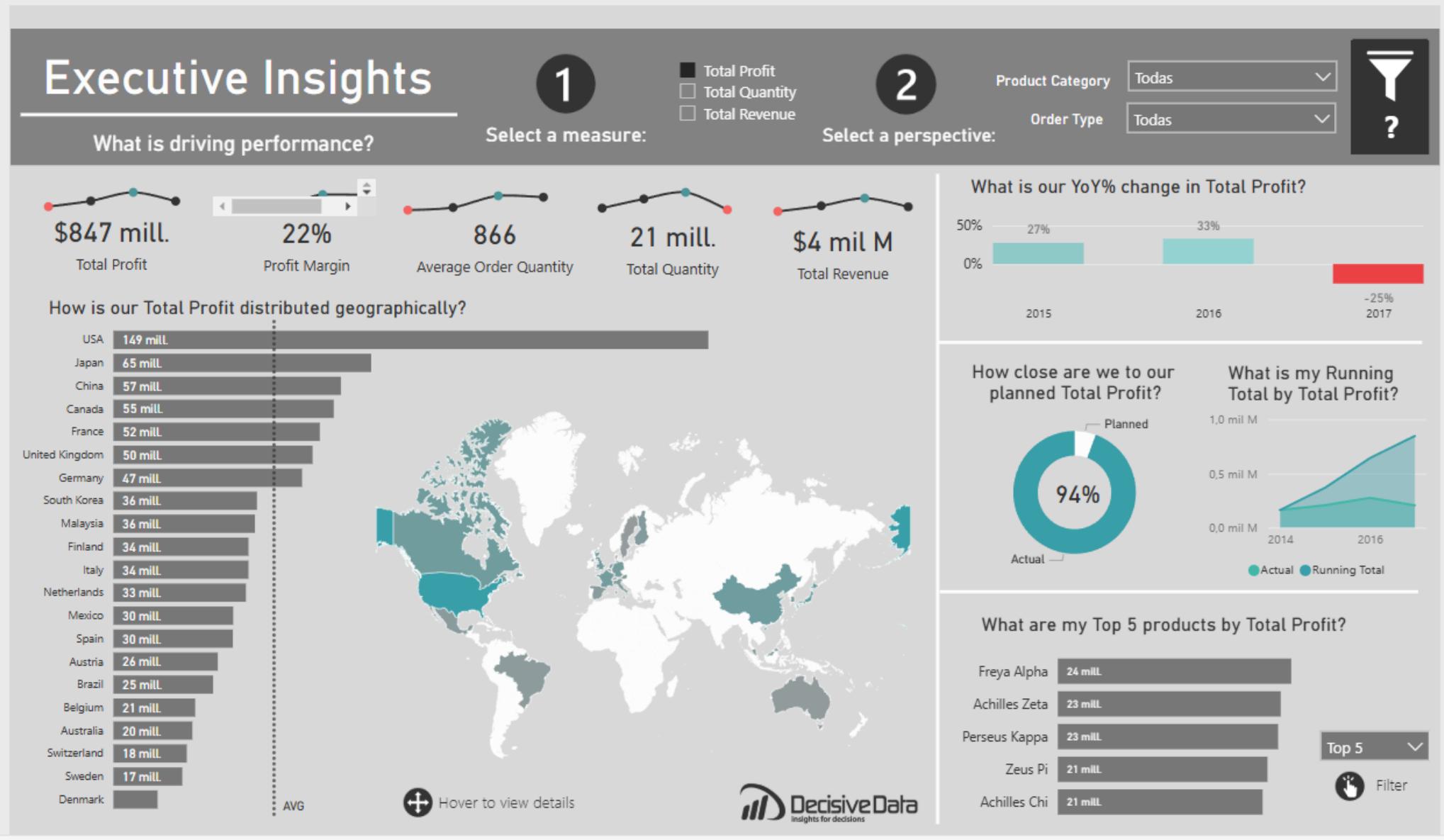


- SCORECARDS vs. DASHBOARDS
- SCORECARDS:
  - Controlling the progress towards the strategy
  - Formal business methodology (holistic approach)
  - Focus on long-term strategic performance.
    - Periodic snapshots (e.g. quarterly, annually) to evaluate progress
  - Present summaries, not particular data
  - Can be graphically displayed

- SCORECARDS vs. DASHBOARDS
- DASHBOARDS:
  - No specific for business environments.
  - Usually focus on an **specific problem** (no holistic approach)
  - Measuring general performance AND its specific aspects.
  - Used by executives / specialists and other employees
  - Provide updates in real time or **in right-time**.
  - Concerning data, including summaries.
  - Data visualized nicely (raw data also available through drill-down features)

## 6. EXAMPLES OF DASHBOARDS

# SCORECARDS & DASHBOARDS



Source: Microsoft

# SCORECARDS & DASHBOARDS

## SUMMARY

CAGR

4,4%

TRADE VOLUME

41.997.925

EXPORT

22.078.619

IMPORT

19.919.306

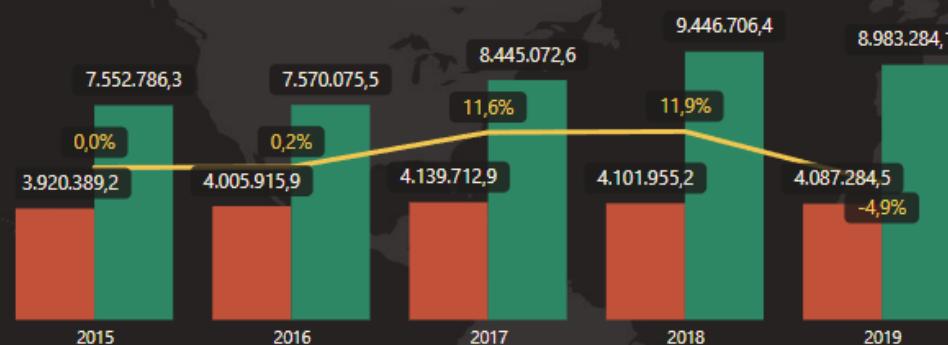
2015 2019

EXP / IMP BALANCE

2.159.313

## GLOBAL TRADE DYNAMICS

● Trade volume, Ton, K ● Trade volume, Euro, M ● CAGR (Year to Year)



## TRADE BY COUNTRIES

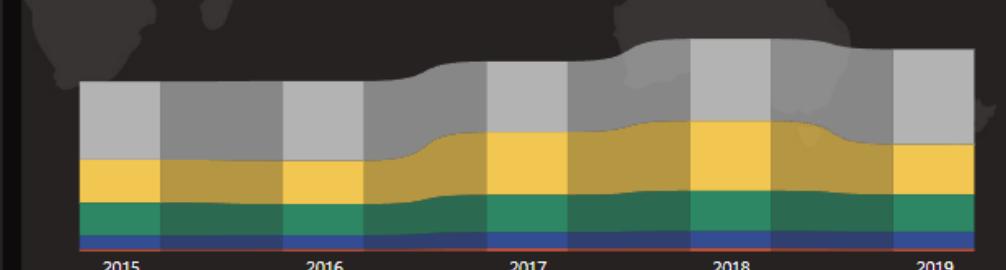
| COUNTRY     | CAGR, T | Ton, K    | Euro, M    | CAGR Euro |
|-------------|---------|-----------|------------|-----------|
| Germany     | -0,2%   | 3.985.724 | 12.518.084 | 3,5%      |
| France      | 0,5%    | 2.460.968 | 5.586.650  | 5,2%      |
| Netherlands | 0,5%    | 3.573.633 | 5.431.377  | 6,6%      |
| Italy       | 0,2%    | 2.266.334 | 4.835.215  | 3,7%      |
| Belgium     | 3,3%    | 2.678.756 | 4.370.588  | 3,4%      |
| Spain       | 1,3%    | 2.068.190 | 3.181.991  | 4,2%      |
| Poland      | 3,1%    | 1.238.855 | 2.238.560  | 7,0%      |
| Austria     | 1,8%    | 658.577   | 1.632.857  | 4,5%      |
| Sweden      | 1,0%    | 875.428   | 1.567.313  | 3,4%      |
| Finland     | 3,0%    | 448.791   | 635.289    | 5,3%      |

## CARGO CATEGORIES



## DYNAMICS BY MODES OF TRANSPORT

● Air ● Rail ● Road ● Sea ● All others or Unknown



- + 99 %

Source: Microsoft

- DASHBOARDS IN HEALTHCARE
  - Focused on clinical performance indicators.
  - Performance Indicator: *“a statistic or other unit of information which reflects, directly or indirectly, the performance of a health or welfare intervention, facility, service or system in maintaining or increasing the well being of its target population”*. (Armstrong, 1994)

- DASHBOARDS IN HEALTHCARE
  - Focused on clinical performance indicators.
  - Performance Indicator Properties:
    - Definable
    - Clear intent
    - Accessible
    - Reliable
    - Useful

- DASHBOARDS IN HEALTHCARE
  - Types of performance indicators.  
(by ACHS Performance Indicators)
    - Continuity of care: medical record, assessment system, consent, care evaluation, discharge, etc.
    - Access: information about services, access to the needs.
    - Effectiveness: evidence-based , process effective.
    - Patient safety: medication, infection control, blood management, etc.

- **DASHBOARDS IN HEALTHCARE**
  - Examples of performance indicators.
    - Regarding Medical Records:
      - Percentage of medical records where illegible writing resulted in an adverse event.
      - Percentage of medical records where care plans are not documented.
    - Regarding Blood Management:
      - Percentage of inappropriate storage
      - No. of patients transfused with Hb>100g/L

# SCORECARDS & DASHBOARDS

## Patients Discharged Receiving Home Health (HH) Services

Overview Hospital 248, Years: Todo, Service: Todo



| Total Discharges | Discharges to Home Health | Home Health CMI | Number of Home Health Agencies | Avg # of Patient Referrals per HH Agency |
|------------------|---------------------------|-----------------|--------------------------------|------------------------------------------|
| <b>56.688</b>    | <b>9,9% (n=5.640)</b>     | <b>1,24</b>     | <b>211</b>                     | <b>27</b>                                |

**Home Health discharges by service line**

| Service Line                             | Discharges | Overall CMI | Discharges CMI | % of Discharges to HH |
|------------------------------------------|------------|-------------|----------------|-----------------------|
| Hospital Overall                         |            |             |                | 9,9%                  |
| <a href="#">General Medicine</a>         | 10.536     | 1,17        | 1.29           | 10,0%                 |
| <a href="#">Obstetrics</a>               | 9.432      | 0,70        | 0,68           | 9,4%                  |
| <a href="#">Normal Newborns</a>          | 6.720      | 0,17        | 0,17           | 8,9%                  |
| <a href="#">Cardiovascular Diseases</a>  | 6.144      | 1,06        | 1,03           | 11,3%                 |
| <a href="#">Pulmonary Medical</a>        | 4.632      | 1,25        | 1,08           | 14,0%                 |
| <a href="#">General Surgery</a>          | 3.912      | 2,69        | 3,39           | 11,0%                 |
| <a href="#">Nephrology/Urology</a>       | 3.288      | 1,03        | 0,97           | 7,3%                  |
| <a href="#">Orthopedics</a>              | 3.072      | 2,03        | 2,72           | 7,8%                  |
| <a href="#">Neuro Sciences</a>           | 2.448      | 1,24        | 1,14           | 13,7%                 |
| <a href="#">Cardio\Vasc\Thor Surgery</a> | 2.352      | 3,69        | 2,72           | 4,1%                  |
| <a href="#">Neonatology</a>              | 2.352      | 1,85        | 1,55           | 7,1%                  |
| <a href="#">Oncology</a>                 | 840        | 1,71        | 1,40           | 11,4%                 |
| <a href="#">Gynecology</a>               | 456        | 1,11        | 1,00           | 26,3%                 |
| <a href="#">Alcohol &amp; Drug Abuse</a> | 216        | 0,72        | 0,65           | 11,1%                 |
| <a href="#">ENT</a>                      | 192        | 0,86        | 0,0%           |                       |
| <a href="#">Ophthalmology</a>            | 72         | 0,90        | 0,0%           |                       |
| <a href="#">Psychiatry</a>               | 24         | 0,99        | 0,0%           |                       |

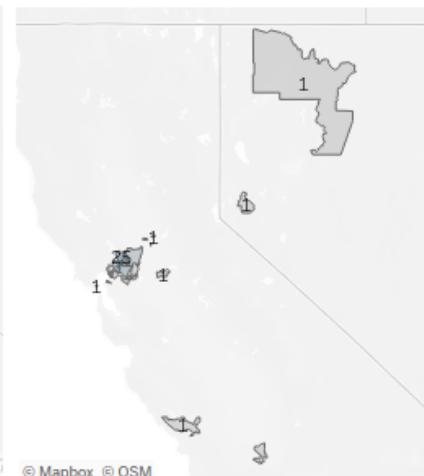
▲ ▲ Hover for trend ▲ ▲

**Home Health agency locations**  
Click a dot to see specific agencies and the locations of patients they serve



© Mapbox © OSM

**Number of patients discharged to Home Health by ZIP Code** (Only patients with ZIP Code data)



© Mapbox © OSM

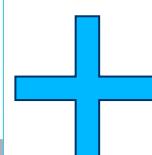
**Home Health agency referrals** Click to see ratings

| ZIP Code | HH Agency Name               | Rating |
|----------|------------------------------|--------|
| 90006    | <a href="#">Agency 8919</a>  | 1      |
| 90008    | <a href="#">Agency 33998</a> | 1      |
| 90010    | <a href="#">Agency 11090</a> | 1      |
|          | <a href="#">Agency 13328</a> | 1      |
|          | <a href="#">Agency 18960</a> | 1      |
|          | <a href="#">Agency 19316</a> | 1      |
|          | <a href="#">Agency 19932</a> | 1      |
|          | <a href="#">Agency 23810</a> | 1      |
|          | <a href="#">Agency 25278</a> | 1      |

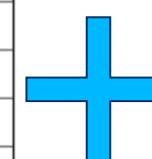
# SCORECARDS & DASHBOARDS

- DASHBOARDS IN HEALTHCARE
  - ICU Dashboard Example.
  - Fluid balance
  - Severity Score:
    - APACHE III :
    - [0..299]
    - Initial risk classification of severely ill hospitalized patients.
    - 20 physiologic variables

| Variables fisiológicas                                                                   | Límites altos anormales |          |         |            | Norma               |                       |           |                       | Límites bajos anormales |     |     |     |
|------------------------------------------------------------------------------------------|-------------------------|----------|---------|------------|---------------------|-----------------------|-----------|-----------------------|-------------------------|-----|-----|-----|
|                                                                                          | -4                      | -3       | -2      | -1         | 0                   | +1                    | +2        | +3                    | +4                      | +5  | +6  | +7  |
| Temperatura, rectal (°C)                                                                 | ≥41°                    | 39-40,9° | ---     | 38,5-38,9° | 36-38,4°            | 34-34,5°              | 32-33,9°  | 30-31,9°              | ≤29,9°                  | --- | --- | --- |
| Presión arterial media (mmHg)                                                            | ≥160                    | 130-159  | 110-129 | ---        | 70-109              | ---                   | 50-69     | ---                   | ≤49                     | --- | --- | --- |
| Frecuencia cardíaca (latidos/minuto)                                                     | ≥180                    | 140-179  | 110-139 | ---        | 70-109              | ---                   | 55-69     | 40-54                 | ≤39                     | --- | --- | --- |
| Frecuencia respiratoria (resp/minuto)                                                    | ≥50                     | 33-49    | ---     | 25-34      | 12-24               | 10-11                 | 6-9       | ---                   | ≤5                      | --- | --- | --- |
| Oxigenación: AaDO <sub>2</sub> o PaO <sub>2</sub> (mmHg)                                 | ---                     | ---      | ---     | ---        | ---                 | ---                   | ---       | ---                   | ---                     | --- | --- | --- |
| a. FIO <sub>2</sub> ≥0,5, registrar AaDO <sub>2</sub>                                    | ≥500                    | 350-499  | 200-349 | ---        | <200                | ---                   | ---       | ---                   | ---                     | --- | --- | --- |
| b. FIO <sub>2</sub> <0,5, registrar sólo PaO <sub>2</sub>                                | ---                     | ---      | ---     | ---        | PO <sub>2</sub> >70 | PO <sub>2</sub> 61-70 | ---       | PO <sub>2</sub> 55-60 | PO <sub>2</sub> <55     | --- | --- | --- |
| pH arterial                                                                              | ≥7,7                    | 7,6-7,69 | ---     | 7,5-7,59   | 7,33-7,49           | ---                   | 7,25-7,32 | 7,15-7,24             | <7,15                   | --- | --- | --- |
| Sodio sérico (mmol/L)                                                                    | ≥160                    | 160-179  | 155-159 | 150-154    | 130-149             | ---                   | 120-129   | 111-119               | ≤110                    | --- | --- | --- |
| Potasio sérico (mmol/L)                                                                  | ≥7                      | 6-6,9    | ---     | 5,5-5,9    | 3,5-5,4             | 3-3,4                 | 2,5-2,9   | ---                   | <2,5                    | --- | --- | --- |
| Creatinina sérica (mg/dl) (puntuación doble para la Renal Aguda)                         | ≥3,5                    | 2-3,4    | 1,5-1,9 | ---        | 0,8-1,4             | ---                   | <0,6      | ---                   | <20                     | --- | --- | --- |
| Hematocrito (%)                                                                          | ≥60                     | ---      | 50-59,9 | 46-49,9    | 30-45,9             | ---                   | 20-20,9   | ---                   | <20                     | --- | --- | --- |
| Recuento de leucocitos (total/mm <sup>3</sup> )                                          | ≥40                     | ---      | 20-39,9 | 15-19,9    | 3-14,9              | ---                   | 1-2,9     | ---                   | <1                      | --- | --- | --- |
| Puntuación GLASGOW COMA SCORE = 15 - Puntuación GCS real                                 | ---                     | ---      | ---     | ---        | ---                 | ---                   | ---       | ---                   | ---                     | --- | --- | --- |
| <b>A PUNTUACIÓN FISIOLÓGICA AGUDA (PFA) total = Sumar los puntos de las 12 variables</b> | ---                     | ---      | ---     | ---        | ---                 | ---                   | ---       | ---                   | ---                     | --- | --- | --- |
| HCO <sub>3</sub> sérico (venoso, mmol/L) (no es de elección, usar si no hay GSA)         | ≥52                     | 41-51,9  | ---     | 32-40,9    | 22-31,9             | ---                   | 18-21,9   | 15-17,9               | <15                     | --- | --- | --- |



| Edad          | Puntos |
|---------------|--------|
| < 44 años     | 0      |
| 45 - 54 años  | 2      |
| 55 - 64 años  | 3      |
| 65 - 74 años  | 5      |
| > o = 75 años | 6      |



Liver  
Cardiovascular  
Respiratory  
Renal  
Immune depressed

# SCORECARDS & DASHBOARDS

- DASHBOARD IN HEALTHCARE



- REFERENCES & RESOURCES:
  - Robert S. Kaplan, David P. Norton The Balanced Scorecard Translating Strategy into Action.1996.
  - Stephen Few. Information Dashboard Design: the effective visual communication of data. O'Reilly. 2006.
  - Harold Kerzner. Project Management Metrics, KPIs, and Dashboards: a guide to measuring and monitoring project performance. Wiley.2 ed. 2013
- ADDITIONAL READINGS:
  - Beata Kollberg. Exploring the use of balanced scorecards in a Swedish health care organization
  - Casos de éxito con BSC:  
<https://balancedscorecard.org/Resources/Examples-Success-Stories>

# REPORTING

Unit 3 – Query languages and visualization  
S3 –4 – REPORTING

## OUTLINE

1. REPORTING
2. REPORTING TOOLS REVIEW
3. PENTaho REPORT
4. MY FIRST REPORT
5. BI: FUTURE

mas detallado que un dashboard !!!!!

## REPORTING

- Involves **preparing, analyzing, and displaying business metrics**.
- Historically, **technical employees** prepared the data to generate managed reports.
- But now, **non-technical users can generate ad-hoc reports** autonomously, often using **drag- and drop interfaces**.
- In many cases, they are generated on a defined **schedule**.
- Some tools (e.g. Tableau, PowerBI, SAP Analytics) support **Natural Language processing (NLP)** to generate the reports.
- Reports can take many forms: **tables, PDFs, spreadsheets, or webpages**.
- They rely on data from the company data sources, using them as a **single source of truth**.

## REPORTING

- Wide experience in organizations
  - Essential for decision-making and tracking performance.
- Evolution of Reporting Systems:
  - Traditional: Document Oriented (e.g. manual reports, spreadsheets)
  - Modern:
    - Document Management Systems (e.g. Sharepoint, Docuware, Google Workspace).
    - Relational DB Systems & Reporting Tools (e.g. Tableau, Power BI)
- Mature technology
  - Integration between databases and reporting tools for automated analysis.
  - Big data friendly, scalable and real-time analytics.

## REPORTING TOOLS: why?

- Simplifies data understanding: Easier for technical and non-technical users to use the data and better understand business metrics.
- Enable data-driven decisions: Teams can use them to make informed decisions and share the data.
- Can aggregate data from multiple departments, allowing users to interact with the data and provide an integral view of the company.
- Speed up reporting: Can generate reports in short time.
- Automates custom reporting: Can publish customized reports automatically.
- In today's world and besides analytics, data must be communicated to promote insight, action and understanding.

## REPORTING TOOLS: which ones?

- User-friendly: Users should easily query and generate responses.
- Customized reports, with text and graphics
- Scalable for growing datasets and user needs.
- Can easily integrate within our current infrastructure.
- This also applies to dashboards.

Dashboards and reports both provide visual information about the performance of the organization, enabling decision making.

Dashboards are more high-level, and provide a quick view of performance. Reports focus on detailed analysis and historical data, usually a snapshot.

## REPORTING TOOLS: Design rules

- User centered design: Focused on user needs and preferences.
- Organize information: Structure data logically to enhance clarity and usability. Consistent interface.
- Keep it simple: Avoid unnecessary stuff, focus on the core message. Minimalism design principles.
- Provide visual clues: Use icons, charts, and other visual elements to guide users effectively.
- Right color scheme: Choose the right color palette that improves readability and understanding.
- Other aspects to consider: accessibility, responsiveness (if interactive), ...

# IBCS

## International Business Comunication Standard

### Proposals for business communication

Source: IBCS®

We know how important are **standards**

#### Traffic



## IBCS

- Universal set of rules for reporting, visualization and presentation of business information.
- Promotes common structure and design, improving comparison, benchmarking.
- Improves the quality of business communication (46% faster understanding, 61% less errors answering)
- As the data is easier to understand, it supports faster and more informed decision making.
- It provides a notation to present data in a consistent way.
- BI tools (e.g. PowerBI) integrate IBCS principles.
- IBCS is the basis of ISO/AWI 24896 Standard notation for business reports, started in July 2024.

## IBCS

### Key characteristics of Business communication

- Purpose: Making critical decisions.
- Challenges:
  - Operating under constraints of time and efficiency.
  - Managing high volumes of reports and vast amounts of data.
- Impact:
  - Reach company-wide users who require consistency.
  - Mitigating the high risk on misinterpretation.

## **IBCS- Based on scientific, experimental and practical experience.**

- **Conceptual Rules:** How the content is structured and organized to transmit the intended message clearly.
- **Perceptual Rules:** How information is **visually presented** to make it easy to perceive and understand.
- **Semantic Rules:** Ensure consistency and standardization in how similar content is represented.

## IBCS: SEVEN AREAS ACRONYM “SUCCESS”

- **SAY:** Transmit a clear and concise message (Conceptual Rule)
- **UNIFY:** Similar things have a consistent appearance. (Semantic Rule)
- **CONDENSE:** Increase information density while maintaining clarity. (Perceptual Rule)
- **CHECK:** Present information in a way that it is easy to understand and it doesn't mislead. (Perceptual Rule)

Communications products: reports, presentations, statistics, dashboards and BCSs.

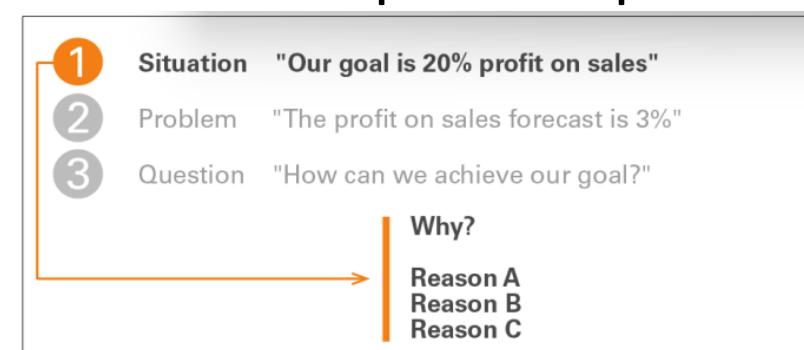
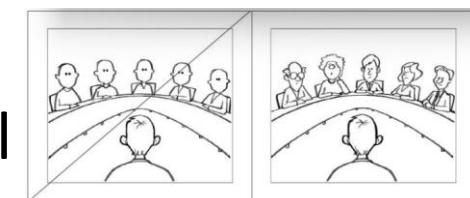
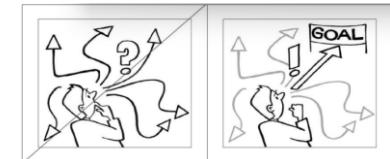
## IBCS: SEVEN AREAS ACRONYM “SUCCESS”

- **EXPRESS:** Ensure the message and facts are communicated clearly and intuitively. (Perceptual Rule)
- **SIMPLIFY:** Remove components that do not add value. (Perceptual Rule)
- **STRUCTURE:** Organize content logically to guide audience effectively. (Conceptual Rule)

## IBCS- SAY: Convey a message

Reports intend to say something to the recipients.

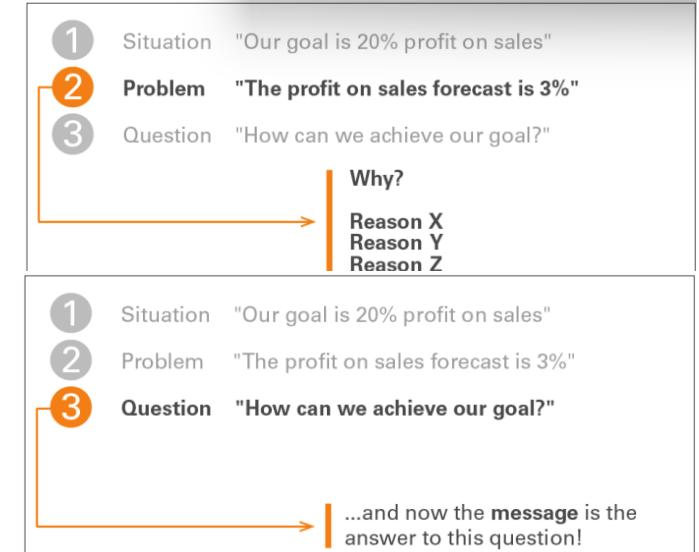
- SA1 Know your objectives:
  - SA1.1 Know own goals: First, we need a clear vision of what we want to achieve.
  - SA1.2 Know your target audience: function, position, knowledge, experience, attitude, cultural background,...
- SA2 Introduce your message: Context and background
  - SA2.1 Map situation: Compile and present related facts.



## IBCS- SAY: Convey a message

Reports intend to say something to the recipients.

- SA2.2 Explain the problem: What is the challenge?
- SA2.3 Raise the question:  
What is the relevant questions from receiver's perspective?
- SA3 Deliver the message: Answer it.
  - SA3.1 Detect, explain, or suggest



## IBCS- SAY: Convey a message

Reports intend to say something to the recipients.

- SA3.1 Detect, explain, or suggest:
- SA3.2 Say the message first:

Summed up clear overall message:

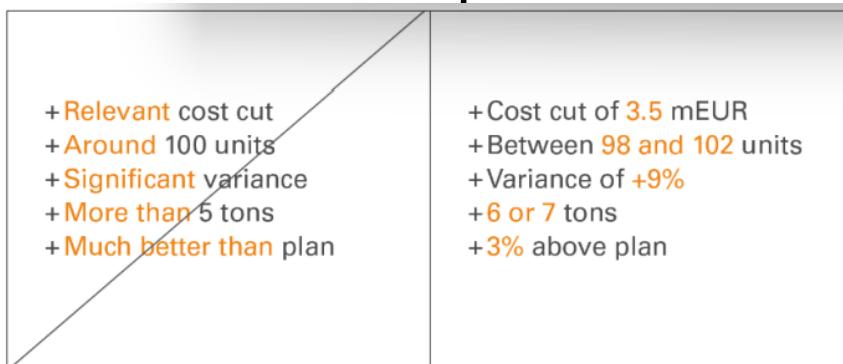
|                                                                                                                                                                                                     |                                                                                                                                                                                                                       |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>The market for Project B is too small and we expect high competition. In addition, costs are above budget and desired quality is unattainable. <b>Therefore, we should stop the project.</b></p> | <p><b>We should stop Project B</b></p> <ul style="list-style-type: none"> <li>+ Market is too small</li> <li>+ Competition is very high</li> <li>+ Costs are above plan</li> <li>+ Quality is unattainable</li> </ul> |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

| Detections                                                                         | Explanations                                                        | Suggestions                                                         |
|------------------------------------------------------------------------------------|---------------------------------------------------------------------|---------------------------------------------------------------------|
| ① Export share of PSI sank in Q1 from 45% to 40%                                   | ...because product approvals are lacking in Asia and America .      | Therefore, we should move up authorization for the USA to August.   |
| ② Inventory is replaced 5 times, which is below industry average of 6.5            | ... because suppliers are not particularly reliable.                | Therefore, we should re-examine the selection of suppliers.         |
| ③ Sales in spare parts fell by 12%                                                 | ...because competitors are increasingly imitating our products.     | Therefore, we should redesign our spare parts business.             |
| ④ Personnel costs in Berlin amount to 35% of sales, which is 6app above the target | ...because the personnel situation has not been adjusted.           | Therefore, we should utilize logistics through cooperations.        |
| ⑤ Results for kitchens in Q1 were 3 million under budget for the first time        | ...because we had a lot of down time at our new plant C.            | Therefore, we should re-examine the new management concept.         |
| ⑥ Production costs at C are 11% higher than the average                            | ...because lot sizes here are smaller than at all the other plants. | Therefore, we should build the planned warehouse already this year. |
| ⑦ Result has deteriorated despite a 6% increase in sales                           | ...because sales often gave discounts that were too large           | Therefore, we should reorganize the management of sales.            |
| ⑧ Margins in films are 5% lower than in the previous year                          | ...because excess capacity still exists in Eastern Europe-          | Therefore, we should not begin production in China as planned.      |
| ⑨ Consulting time for small customers takes up 50%                                 | ...because consulting was provided without concrete specifications. | Therefore, we should focus consultations on customer potential.     |

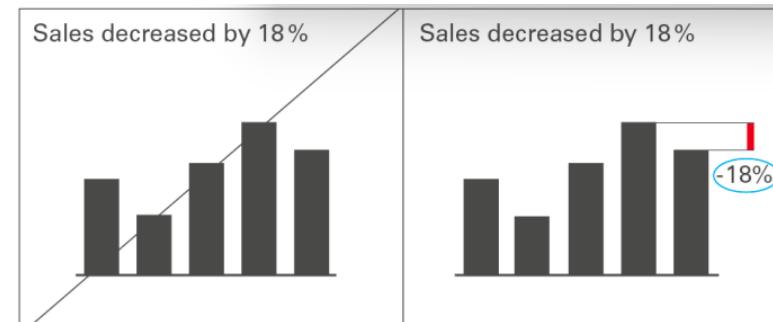
## IBCS- SAY: Convey a message

Reports intend to say something to the recipients.

- SA4 Support message
  - SA4.1 Provide evidence: Prove the message with facts and figures.
  - SA4.2 Use precise words:



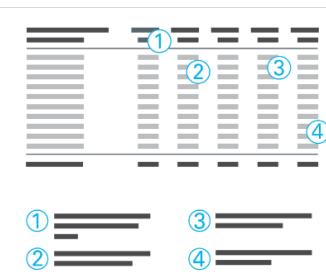
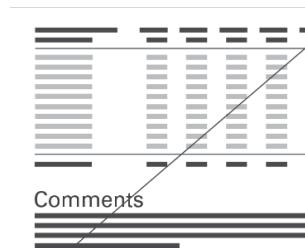
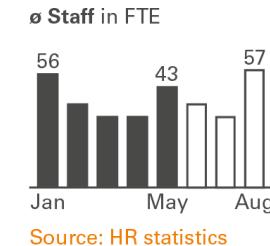
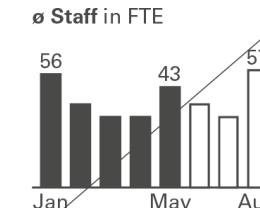
- SA4.3 Highlight message:



## IBCS- SAY: Convey a message

Reports intend to say something to the recipients.

- SA4.4 Name the sources
- SA4.5 Link comments:
- SA5- Summarize message
  - SA5.1 Repeat the message
  - SA5.2 Explain the consequences



"...and I hope that  
I could get across,  
that we have to..."

...and then the audience  
says "Thank you".

...and now we should decide on the next steps:

- a Staff** Additional resources are needed...
- b Machines** Two more assembly lines must be...
- c Marketing** Our web shop features should...

## IBCS- UNIFY: Apply Semantic Notation

Similar content should be visualized in the same way.

- UN1 Unify terminology
  - UN1.1 Unify terms and abbreviations
  - UN1.2 Unify numbers, units and dates

| Term                   | Abbreviations | Definition    |
|------------------------|---------------|---------------|
|                        | short long    |               |
| + Return on investment | ROI           | Ret. on inv.  |
| + Accounts receivable  | AR            | Acc. receiv.  |
| + Profit before tax    | PBT           | Profit b. tax |
| + Profit and loss      | P&L           | Profit & loss |
| + Human resources      | HR            | Human res.    |
| + Net sales per capita | NS/c          | NS per cap.   |

## ISO80000-1 Numbers

ISO4217 Currencies

SI Physical Units

ISO8601 Dates

|          |             |                |             |
|----------|-------------|----------------|-------------|
| 23 mtr.  | 100.000.000 | 23 m           | 100 000 000 |
| 34 kg.   | 123456      | 34 kg          | 123 456     |
| 20 sec.  | 1234567 CHF | 20 s           | 1.23 mCHF   |
| 22 tons  | €           | 22 t           | EUR         |
| [kg]     | US\$        | kg             | USD         |
| sqm      | £           | m <sup>2</sup> | GBP         |
| 1.5.2021 | II/2021     | 2021-05-01     | 2021-Q2     |
| 01/05/21 | W17-2021    | 2021-05-01     | 2021-W17    |
| 05/01/21 | Jun/2021    | 2021-05-01     | 2021-06     |

## IBCS- UNIFY: Apply Semantic Notation

Similar content should be visualized in the same way.

- UN2 Unify descriptors
    - UN2.1 Unify messages
    - UN2.2 Unify titles and subtitles
- Who, what, when



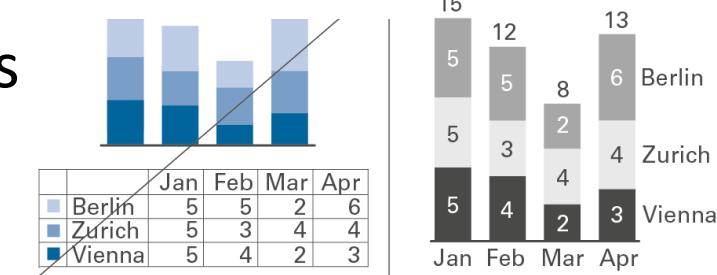
Net Sales Development from  
January to July 2021  
Alpha Corporation [mEUR]

Profit ratio: In thousand Euros  
per Employee in Division D  
Actual and Budget in 2021

Alpha Corporation  
**Net sales** in mEUR  
Jan..Jul 2021

ABC Corporation, Division D  
**Profit per employee** in kEUR  
2021 AC, BU

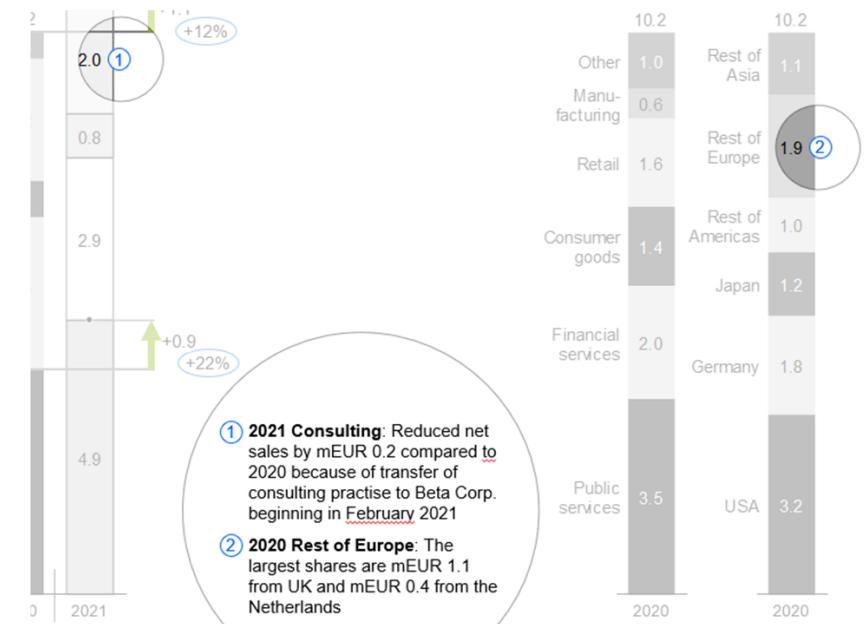
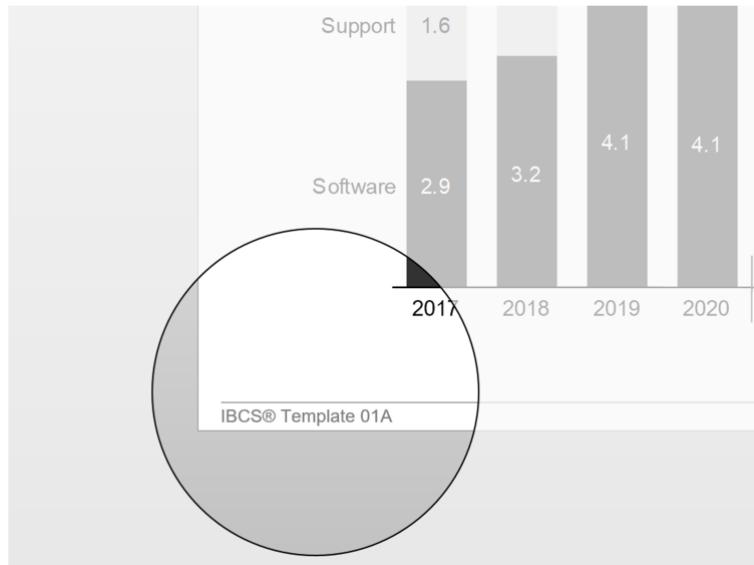
- UN2.3 Unify position legends and labels



## IBCS- UNIFY: Apply Semantic Notation

Similar content should be visualized in the same way.

- UN2.4 Unify comments
- UN2.5 Unify footnotes



## IBCS- UNIFY: Apply Semantic Notation

Similar content should be visualized in the same way.

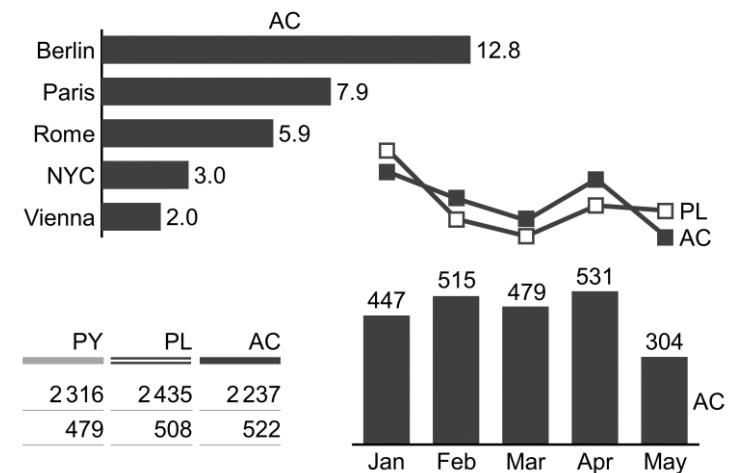
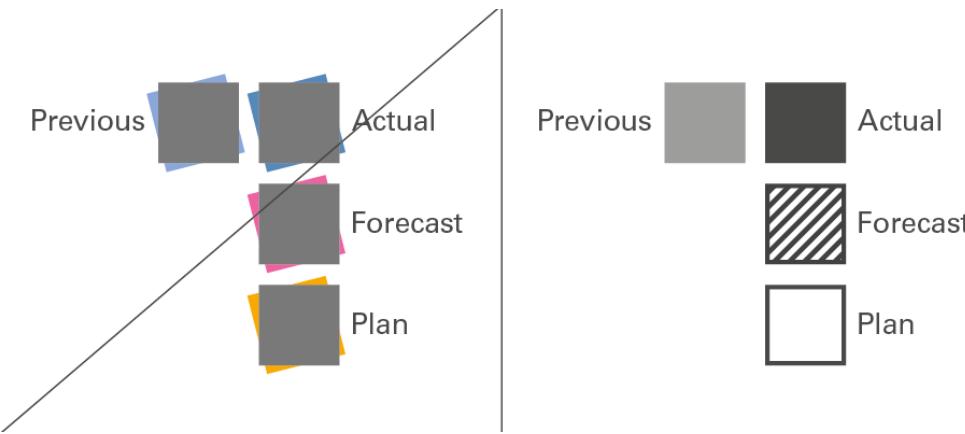
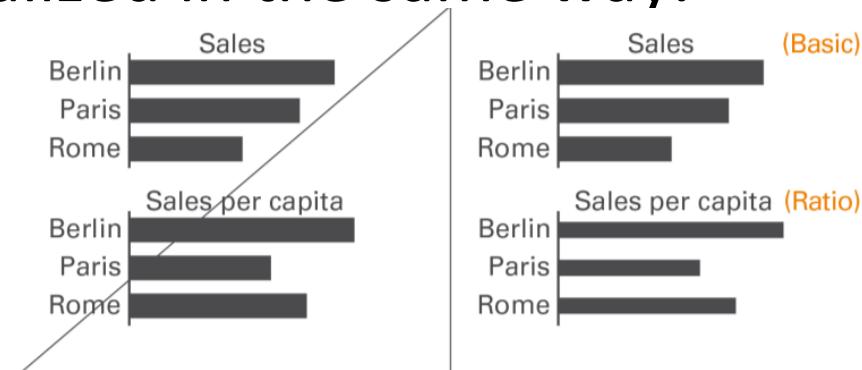
### UN3 Unify dimensions

- UN3.1 Unify measures

Basic-2/3 category width

Ratio- 1/3 category width

- UN3.2 Unify scenarios



## IBCS- UNIFY: Apply Semantic Notation

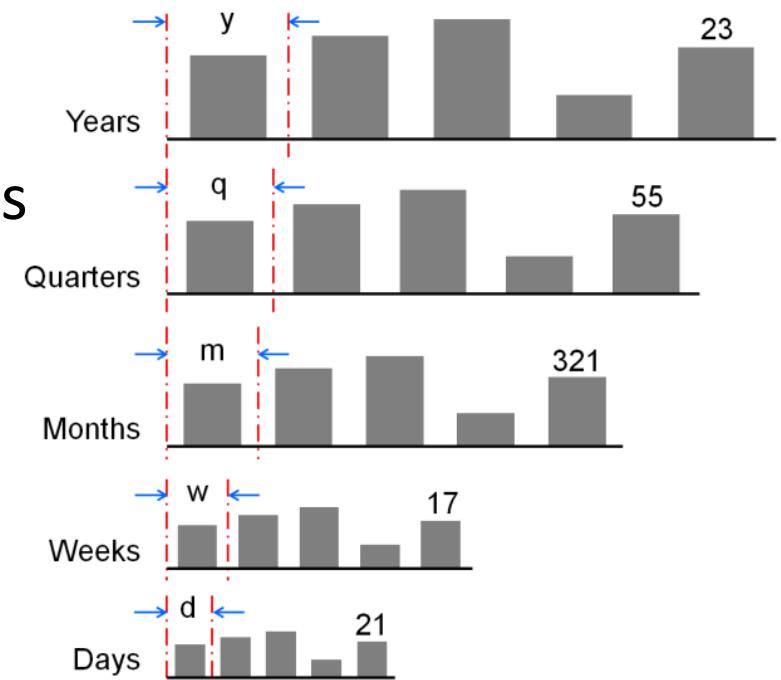
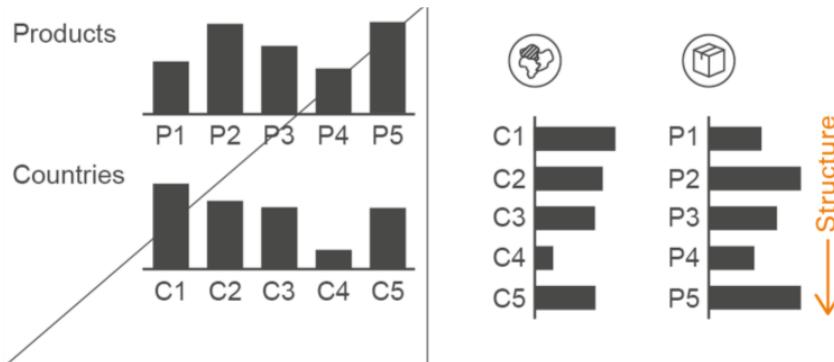
Similar content should be visualized in the same way.

- UN3.3 Unify time periods, use horizontal axes

Visual direction of time

Category widths

- UN3.4 Unify structure dimensions



## IBCS- UNIFY: Apply Semantic Notation

Similar content should be visualized in the same way.

### UN4 Unify analysis

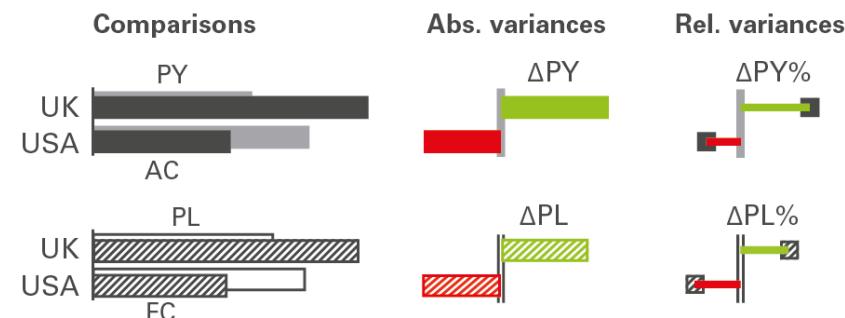
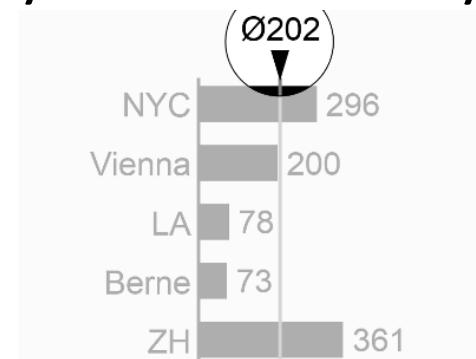
- UN4.1 Unify scenario analysis

Scenario comparison

Absolute variances

Relative variances

- UN4.2 Unify time series analysis
- UN4.3 Unify structure analysis



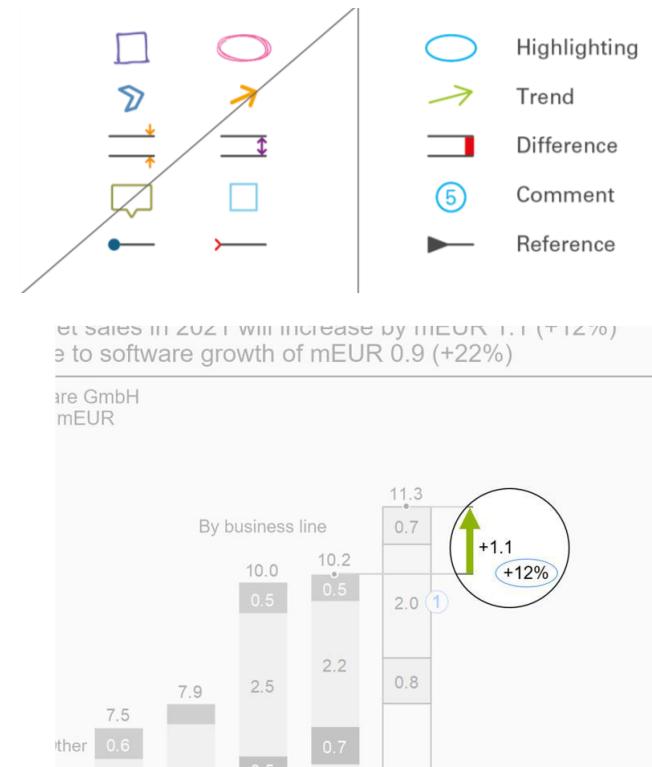
| Analyses     | Symbol | Example       | Application    |
|--------------|--------|---------------|----------------|
| Difference   | x - y  | AC'20 - AC'19 | "Year to date" |
| Time span    | a..b   | Feb..Jun'21   | Jan Feb Mar    |
| Year to date | -x     | -Jun'21       | 123 234 546    |
| Year to go   | x-     | Jun'21-       | 22 46 86       |
| Rolling      | ~x     | ~Jun'21       |                |
| Average      | ø      | ø'21          |                |
| First day    | .x     | .Aug'21       |                |
| Last day     | x.     | Aug.'21       |                |



## IBCS- UNIFY: Apply Semantic Notation

Similar content should be visualized in the same way.

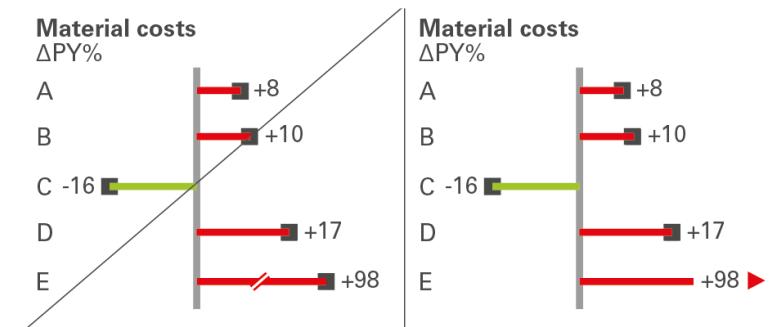
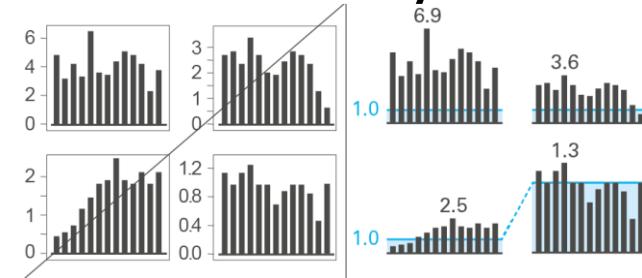
- UN4.4 Unify adjustment analyses
  - UN5 Unify indicators
    - UN5.1 Unify Highlighting indicators
- Assisting lines and areas  
 Difference markers  
 Trend arrows  
 Highlighting ellipses  
 Reference arrowheads  
 Comment references



## IBCS- UNIFY: Apply Semantic Notation

Similar content should be visualized in the same way.

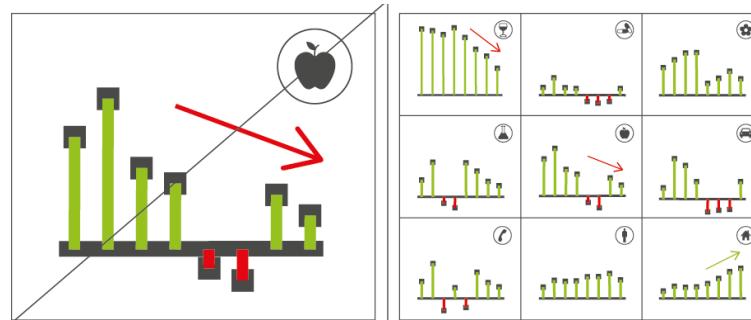
- UN5.2 Unify scaling indicators
- UN5.3 Unify outlier indicators



## IBCS- CONDENSE: Increase Information Density

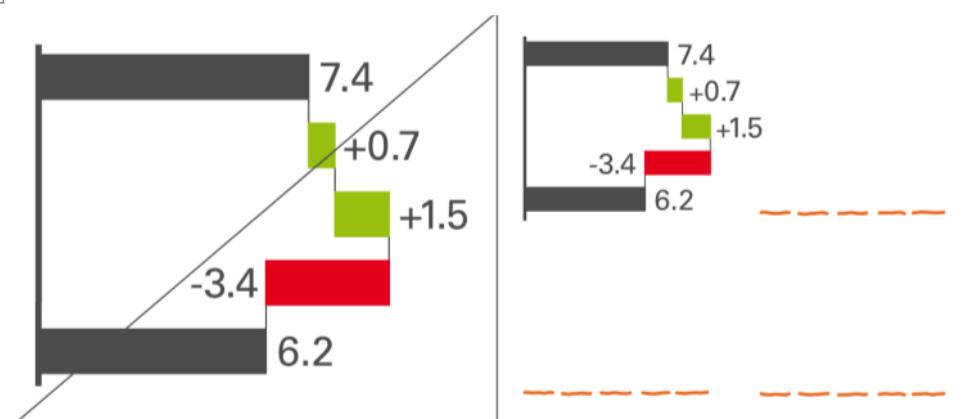
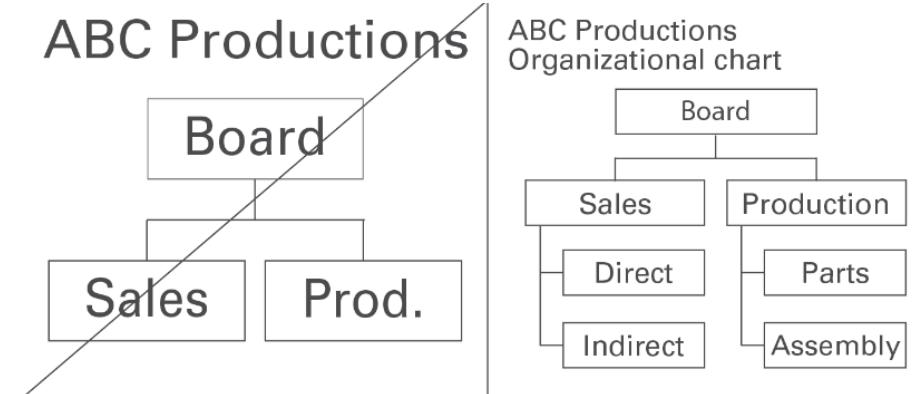
All reports include all needed information on one page

- CO1 Use small components
  - CO1.1 Use small fonts
  - CO1.2 Use small elements



- CO1.3 Use small objects

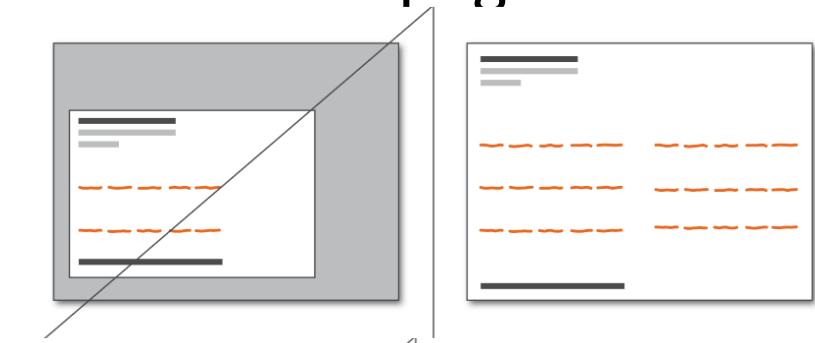
ABC Productions



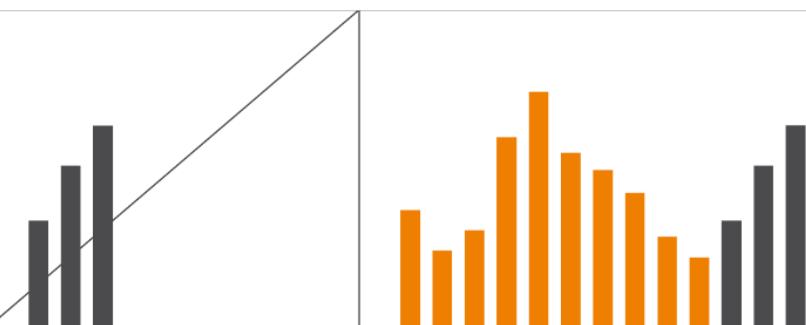
## IBCS- CONDENSE: Increase Information Density

All reports include all needed information on one page

- CO2 Maximize use of space
  - CO2.1 Use narrow page margins
  - CO2.2 Reduce empty space
- CO3 Add data
  - CO3.1 Add data points



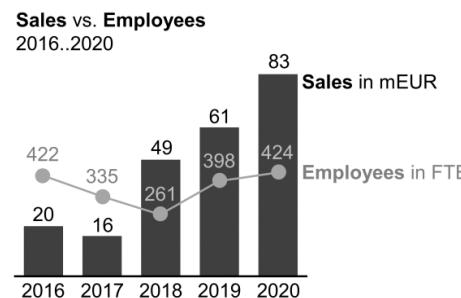
|         | Sales in EUR |     |     |
|---------|--------------|-----|-----|
|         | Jan          | Feb | Mar |
| Germany | 502          | 456 | 243 |
| Austria | 49           | 34  | 44  |
| France  | 89           | 83  | 89  |
| Italy   | 123          | 101 | 117 |
| Sweden  | 77           | 88  | 8   |
| Denmark | 34           | 37  | 45  |



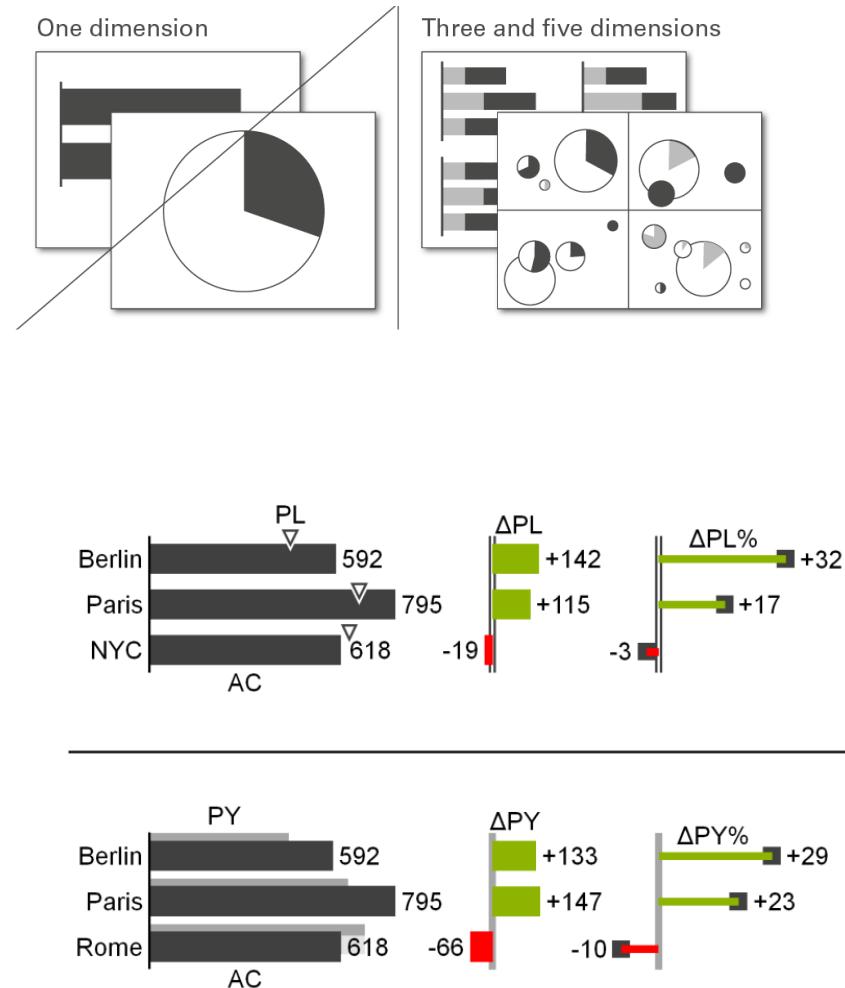
## IBCS- CONDENSE: Increase Information Density

All reports include all needed information on one page

- CO3.2 Add dimensions
- CO4 Add elements
  - CO4.1 Show overlay charts



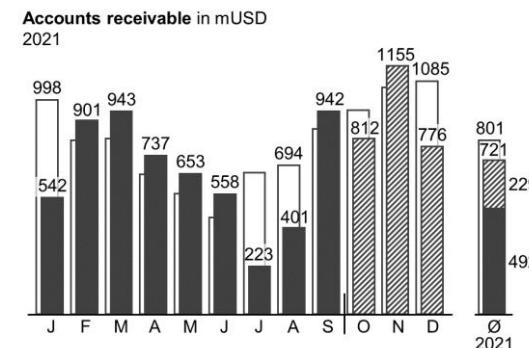
- CO4.2 Show multi-tier charts



## IBCS- CONDENSE: Increase Information Density

All reports include all needed information on one page

- CO4.3 Show extended charts



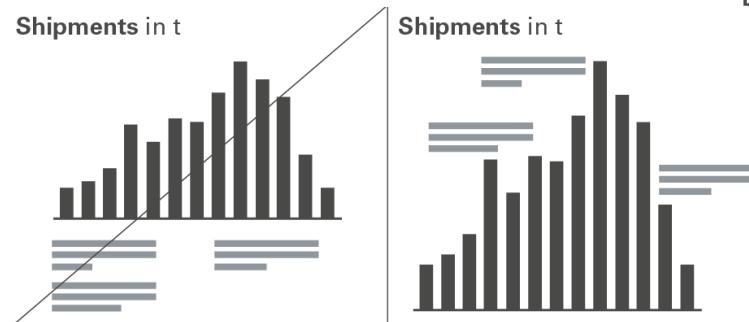
- CO4.4 Embed chart elements in tables
- CO4.5 Embed explanations

Sales in kEUR  
2020

|         | PY  | AC  | ΔPY |
|---------|-----|-----|-----|
| Germany | 84  | 87  | +3  |
| Austria | 19  | 17  | -2  |
| France  | 28  | 27  | -1  |
| Rest    | 36  | 39  | +3  |
| Europe  | 167 | 170 | +3  |

Sales in kEUR  
2020

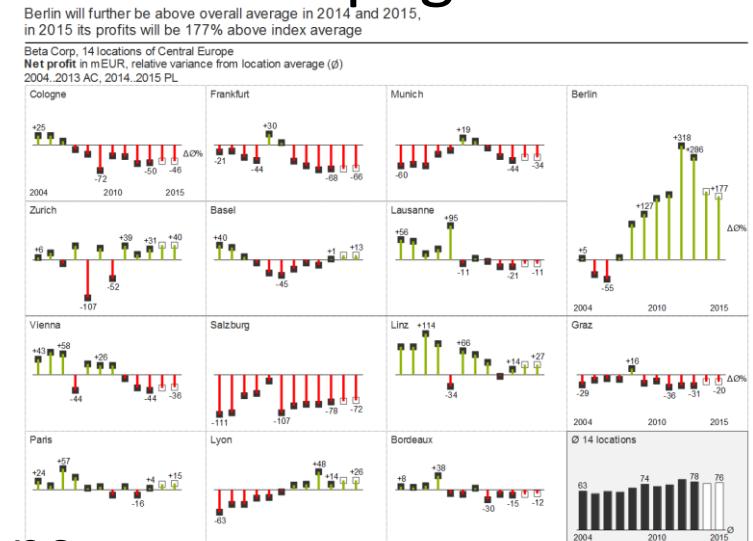
|         | PY  | AC  | ΔPY |
|---------|-----|-----|-----|
| Germany | 84  | 87  | +3  |
| Austria | 19  | 17  | -2  |
| France  | 28  | 27  | -1  |
| Rest    | 36  | 39  | +3  |
| Europe  | 167 | 170 | +3  |



## IBCS- CONDENSE: Increase Information Density

All reports include all needed information on one page

- CO5 Add objects
  - CO5.1 Show small multiples
  - CO5.2 Show related charts on one page
  - CO5.3 Show chart-table combinations
  - CO5.4 Show charts and tables in text pages



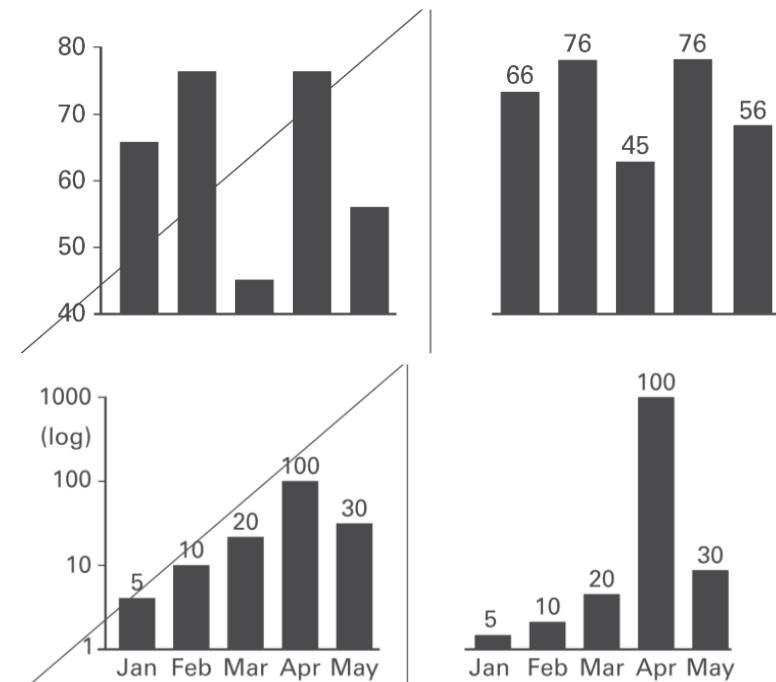
## IBCS- Check: Ensure Visual Integrity

Information in reports is truthful and understandable

- CH1 Avoid manipulated axes
  - CH1.1 Avoid truncated axes (exception: variances)
  - CH1.2 Avoid logarithm axes
  - CH1.3 Avoid different class sizes



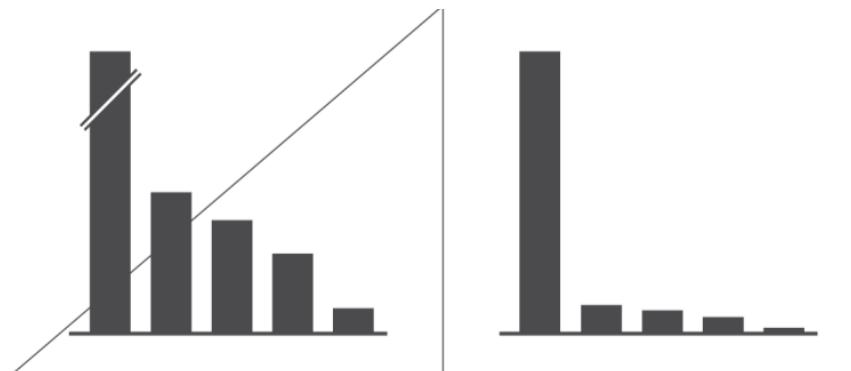
|       |    |
|-------|----|
| 56_60 | 0  |
| 51_55 | 2  |
| 46_50 | 5  |
| 41_45 | 7  |
| 36_40 | 10 |
| 31_35 | 13 |
| 26_30 | 6  |
| 21_25 | 4  |
| 16_20 | 2  |



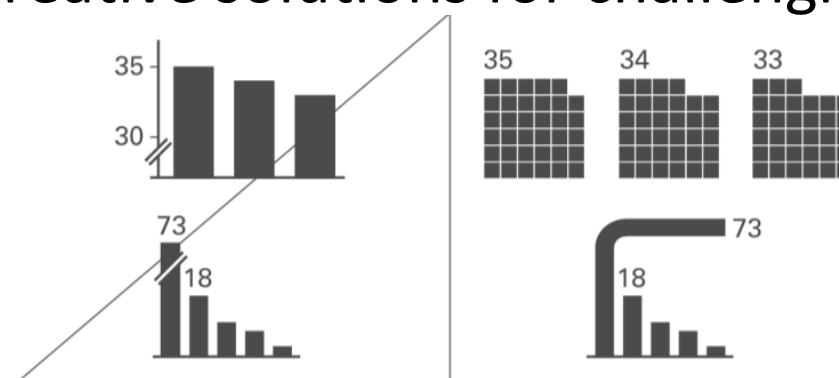
## IBCS- Check: Ensure Visual Integrity

Information in reports is truthful and understandable

- CH2 Avoid manipulated visualization elements
  - CH2.1 Avoid clipped visualization elements



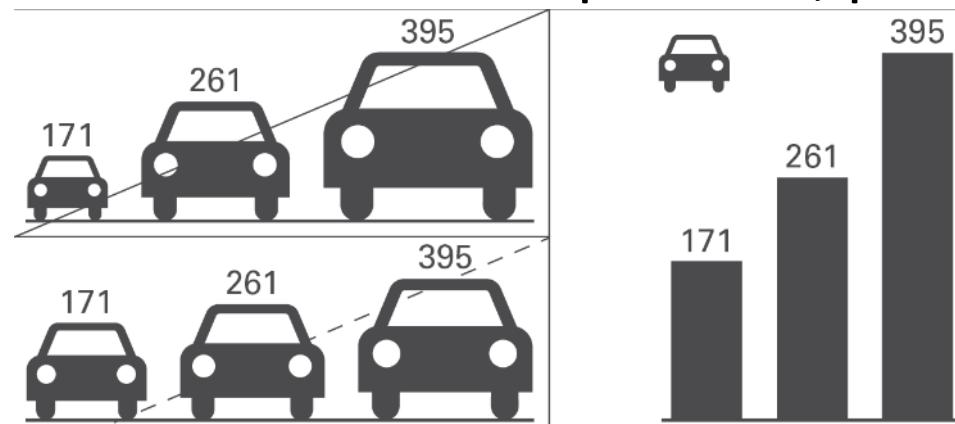
- CH2.2 Use creative solutions for challenging scaling issues



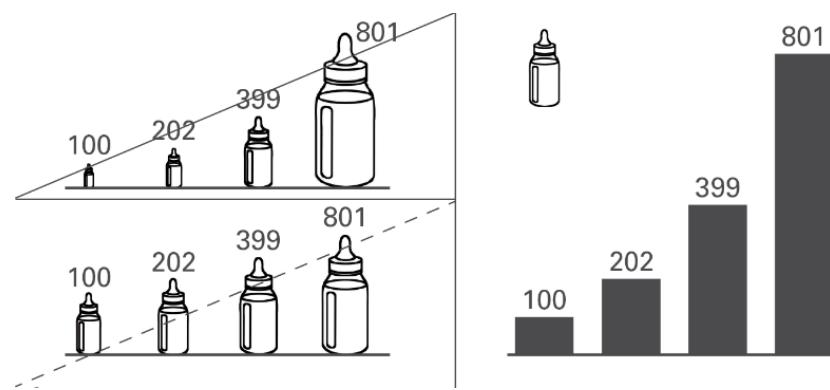
## IBCS- Check: Ensure Visual Integrity

Information in reports is truthful and understandable

- CH3 Avoid misleading representations
  - CH3.1 Use correct area comparisons, prefer linear ones



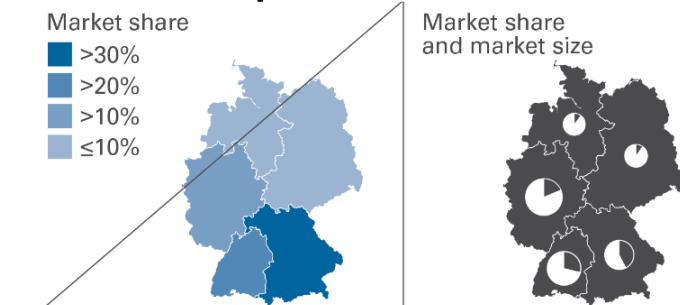
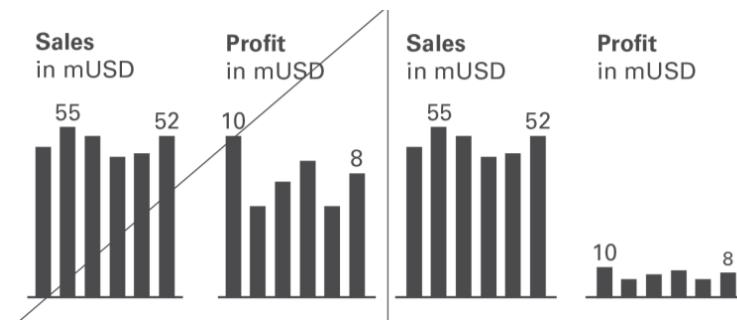
- CH3.2 Use correct volume visualizations, prefer linear ones



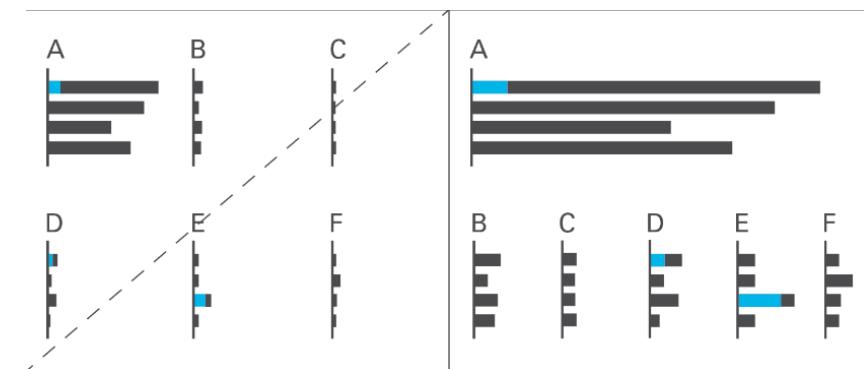
## IBCS- Check: Ensure Visual Integrity

Information in reports is truthful and understandable

- CH3.3 Avoid misleading colored areas in maps
- CH4 Use the same scales
  - CH4.1 Use identical scale for the same unit



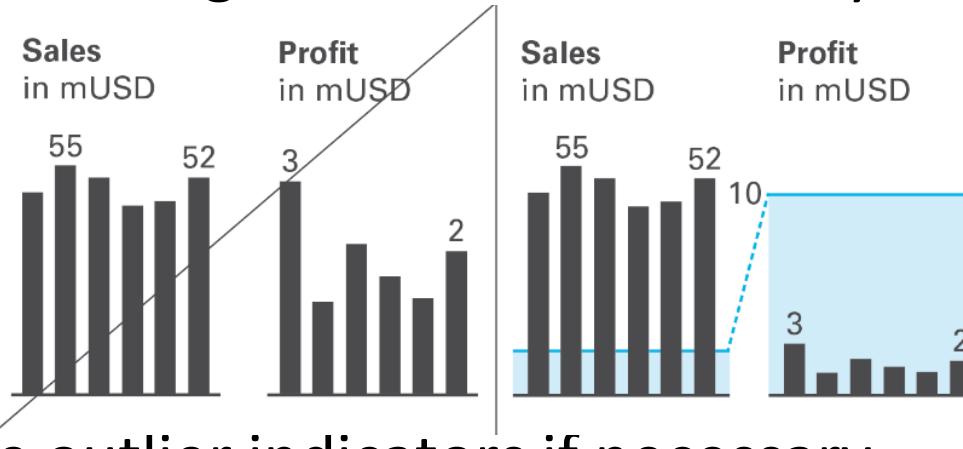
- CH4.1 Size charts to given data



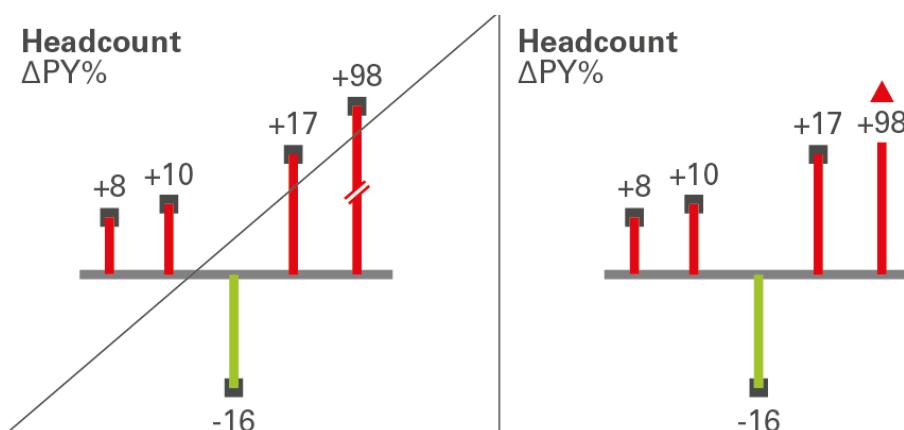
## IBCS- Check: Ensure Visual Integrity

Information in reports is truthful and understandable

- CH4.3 Use scaling indicators if necessary



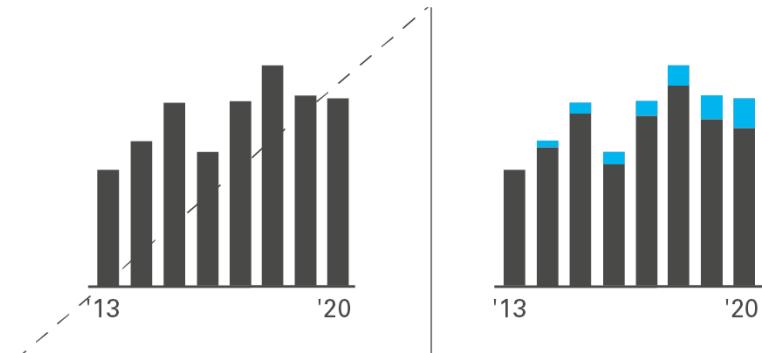
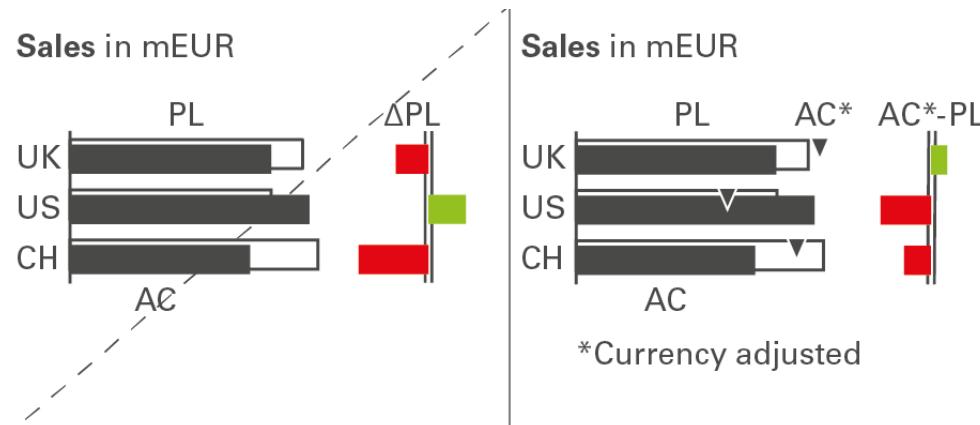
- CH4.4 Use outlier indicators if necessary



## IBCS- Check: Ensure Visual Integrity

Information in reports is truthful and understandable

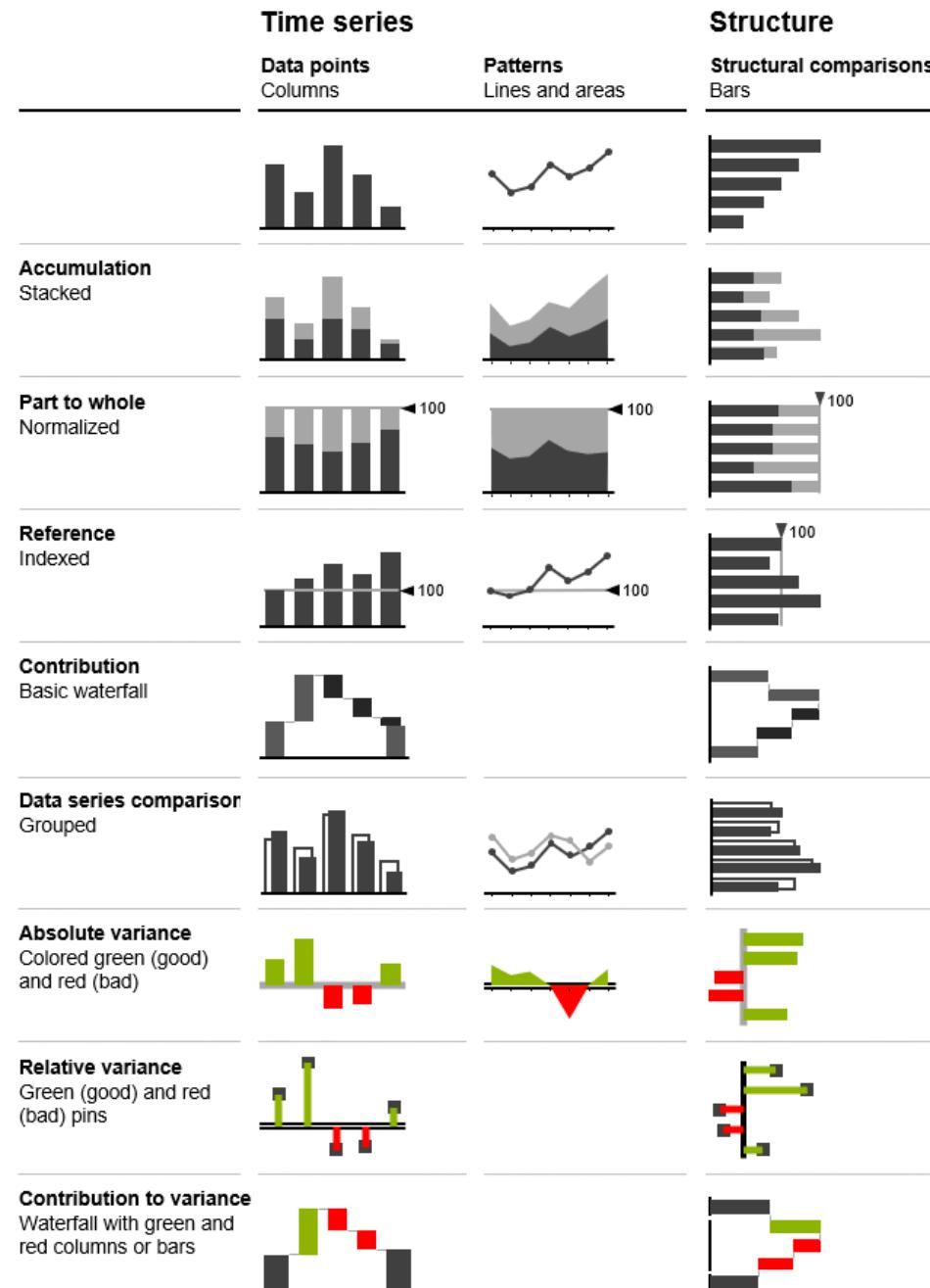
- CH4.5 Use magnifying glasses
- CH5 Show data adjustments
  - Ch5.1 Show the impact of inflation
  - CH5.2 Show the currency impact



## IBCS- Express: Choose Proper Visualization

Charts and tables in reports help to identify interesting facts and communicate the desired message

- EX1 Use appropriate object types
    - EX1.1 Use appropriate chart types
- Bar-comparisons.  
Line- trends over time.



## IBCS- Express: Choose Proper Visualization

Charts and tables in reports help to identify interesting facts and communicate the desired message

- EX1.2 Use appropriate table types

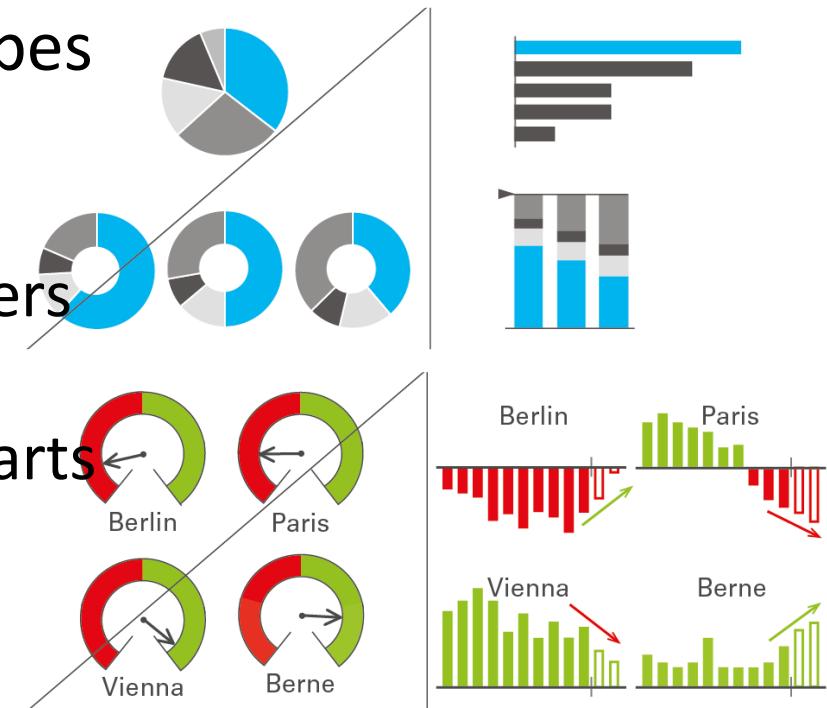
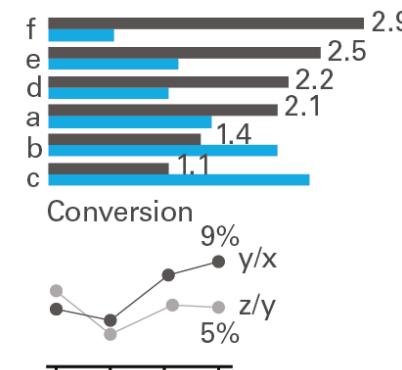
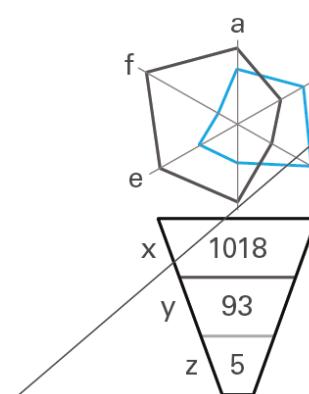
Electronic Inc.  
Profit after tax in kEUR  
Nov 2020

| Nov          |              |              |             |            |             |             |                      |  |  | Jan_Nov       |               |               |             |            |             |            |
|--------------|--------------|--------------|-------------|------------|-------------|-------------|----------------------|--|--|---------------|---------------|---------------|-------------|------------|-------------|------------|
| PY           | PL           | AC           | ΔPY         | ΔPY%       | ΔPL         | ΔPL%        |                      |  |  | PY            | PL            | AC            | ΔPY         | ΔPY%       | ΔPL         | ΔPL%       |
| 560          | 590          | 559          | -1          | -0%        | -31         | -5%         | Austria              |  |  | 5 078         | 5 611         | 5 509         | +431        | +8%        | -102        | -2%        |
| 56           | 72           | 58           | +2          | +4%        | -14         | -19%        | Belgium              |  |  | 531           | 529           | 484           | -47         | -9%        | -45         | -9%        |
| 140          | 149          | 134          | -6          | -4%        | -15         | -10%        | France               |  |  | 1 290         | 1 488         | 1 354         | +64         | +5%        | -134        | -9%        |
| 345          | 279          | 260          | -85         | -25%       | -19         | -7%         | Germany              |  |  | 3 124         | 2 815         | 2 850         | -274        | -9%        | +35         | +1%        |
| 78           | 91           | 86           | +8          | +10%       | -5          | -5%         | Poland               |  |  | 816           | 818           | 854           | +38         | +5%        | +36         | +4%        |
| 77           | 81           | 86           | +9          | +12%       | +5          | +6%         | Sweden               |  |  | 809           | 722           | 764           | -45         | -6%        | +42         | +6%        |
| 61           | 70           | 66           | +5          | +8%        | -4          | -6%         | Switzerland          |  |  | 604           | 582           | 678           | +74         | +12%       | +96         | +16%       |
| 502          | 498          | 545          | +43         | +9%        | +47         | +9%         | Other                |  |  | 5 602         | 6 022         | 5 441         | -161        | -3%        | -581        | -10%       |
| <b>1 819</b> | <b>1 830</b> | <b>1 794</b> | <b>-25</b>  | <b>-1%</b> | <b>-36</b>  | <b>-2%</b>  | <b>Europe</b>        |  |  | <b>17 854</b> | <b>18 587</b> | <b>17 934</b> | <b>+80</b>  | <b>+0%</b> | <b>-653</b> | <b>-4%</b> |
| 119          | 109          | 121          | +2          | +2%        | +12         | +11%        | Brazil               |  |  | 1 205         | 1 254         | 1 314         | +109        | +9%        | +60         | +5%        |
| 65           | 71           | 59           | -6          | -9%        | -12         | -17%        | Canada               |  |  | 629           | 656           | 718           | +89         | +14%       | +62         | +9%        |
| 346          | 326          | 311          | -35         | -10%       | -15         | -5%         | USA                  |  |  | 3 406         | 3 124         | 3 239         | -167        | -5%        | +115        | +4%        |
| 438          | 401          | 399          | -39         | -9%        | -2          | -0%         | Other                |  |  | 4 166         | 4 219         | 4 008         | -158        | -4%        | -211        | -5%        |
| <b>968</b>   | <b>907</b>   | <b>890</b>   | <b>-78</b>  | <b>-8%</b> | <b>-17</b>  | <b>-2%</b>  | <b>Americas</b>      |  |  | <b>9 406</b>  | <b>9 253</b>  | <b>9 279</b>  | <b>-127</b> | <b>-1%</b> | <b>+26</b>  | <b>+0%</b> |
| 54           | 66           | 62           | +8          | +15%       | -4          | -6%         | Australia            |  |  | 517           | 609           | 588           | +71         | +14%       | -21         | -3%        |
| 266          | 204          | 231          | -35         | -13%       | +27         | +13%        | China                |  |  | 2 107         | 1 925         | 2 399         | +292        | +14%       | +474        | +25%       |
| 9            | 70           | 11           | +2          | +22%       | -59         | -84%        | Japan                |  |  | 67            | 855           | 144           | +77         | +115%      | -711        | -83%       |
| 234          | 311          | 255          | +21         | +9%        | -56         | -18%        | Other                |  |  | 2 351         | 2 099         | 2 145         | -206        | -9%        | +46         | +2%        |
| <b>563</b>   | <b>651</b>   | <b>559</b>   | <b>-4</b>   | <b>-1%</b> | <b>-92</b>  | <b>-14%</b> | <b>Rest of World</b> |  |  | <b>5 042</b>  | <b>5 488</b>  | <b>5 276</b>  | <b>+234</b> | <b>+5%</b> | <b>-212</b> | <b>-4%</b> |
| <b>3 350</b> | <b>3 388</b> | <b>3 243</b> | <b>-107</b> | <b>-3%</b> | <b>-145</b> | <b>-4%</b>  | <b>World</b>         |  |  | <b>32 302</b> | <b>33 328</b> | <b>32 489</b> | <b>+187</b> | <b>+1%</b> | <b>-839</b> | <b>-3%</b> |

## IBCS- Express: Choose Proper Visualization

Charts and tables in reports help to identify interesting facts and communicate the desired message

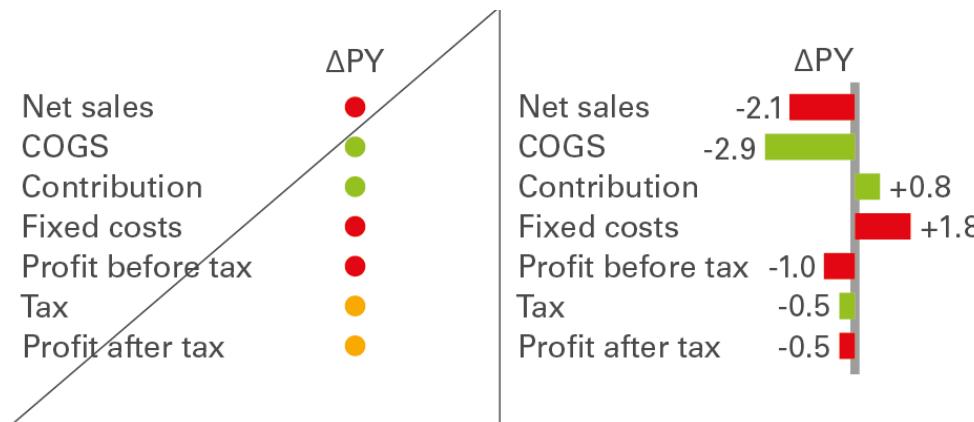
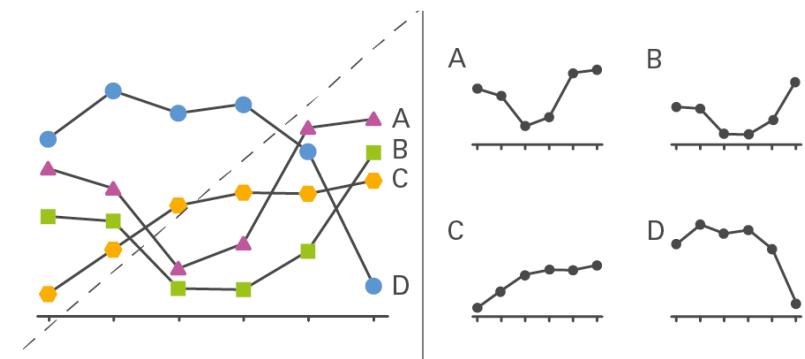
- EX2 Replace inappropriate chart types
  - EX2.1 Replace pie and ring charts
  - EX2.2 Replace gauges, speedometers
  - EX2.3 Replace radar and funnel charts



## IBCS- Express: Choose Proper Visualization

Charts and tables in reports help to identify interesting facts and communicate the desired message

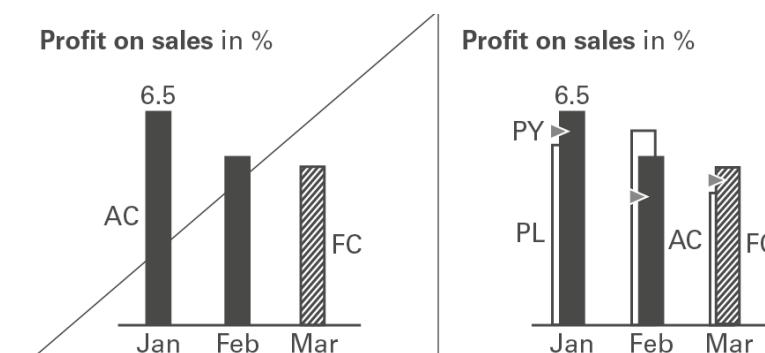
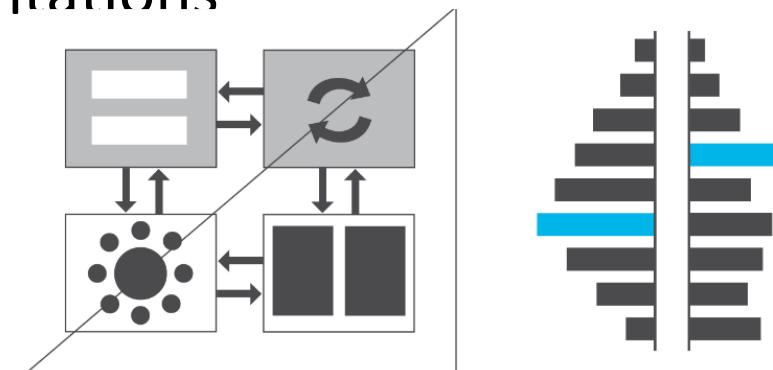
- EX2.4 Replace spaghetti charts
- EX2.5 Replace traffic lights



## IBCS- Express: Choose Proper Visualization

Charts and tables in reports help to identify interesting facts and communicate the desired message

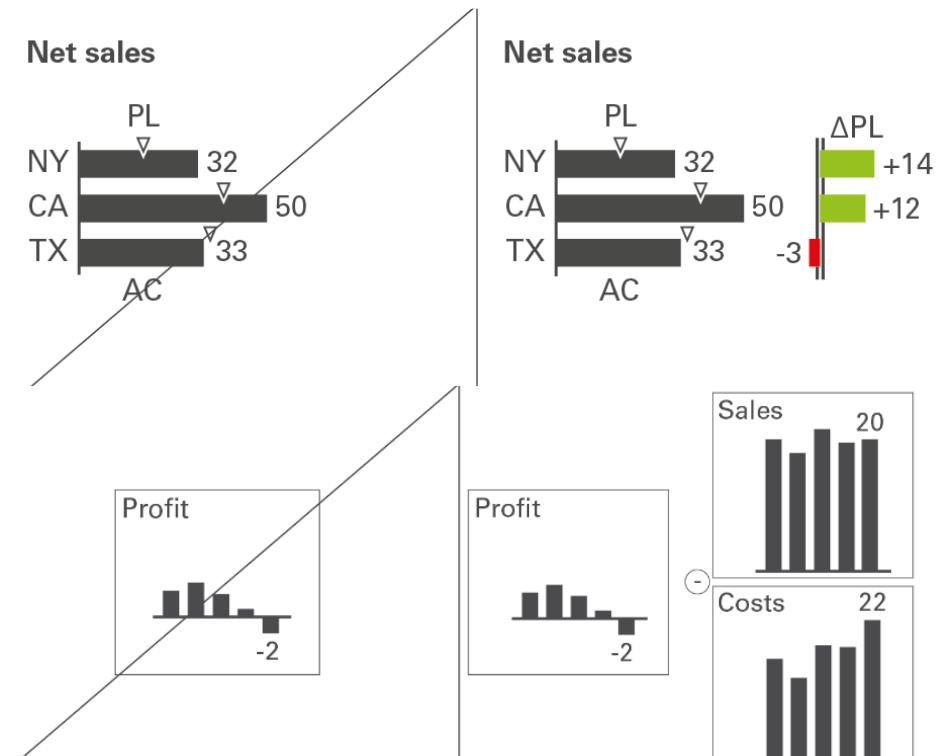
- EX3 Replace inappropriate representations
  - EX3.1 Prefer quantitative representations
  - EX3.2 Avoid text slides in presentations
- EX4 Add comparisons
  - EX4.1 Add scenarios



## IBCS- Express: Choose Proper Visualization

Charts and tables in reports help to identify interesting facts and communicate the desired message

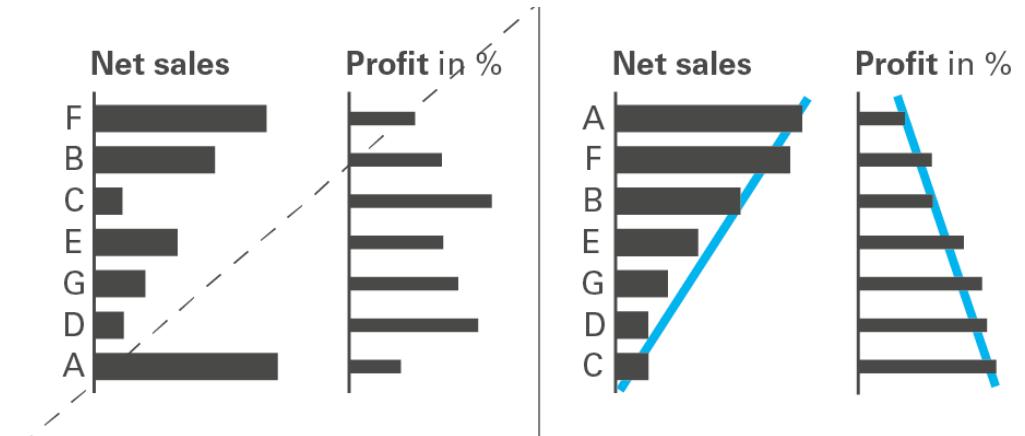
- EX4.2 Add variances
- EX5 Explain causes
  - EX5.1 Show tree structures
  - EX5.2 Show clusters



## IBCS- Express: Choose Proper Visualization

Charts and tables in reports help to identify interesting facts and communicate the desired message

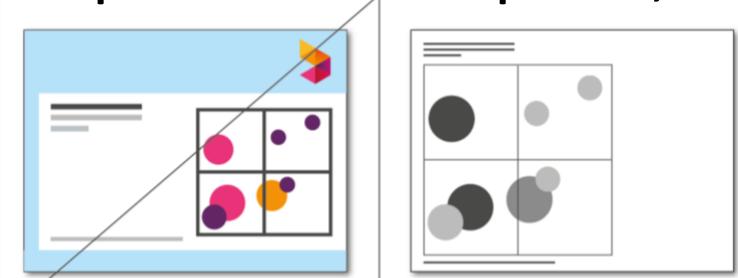
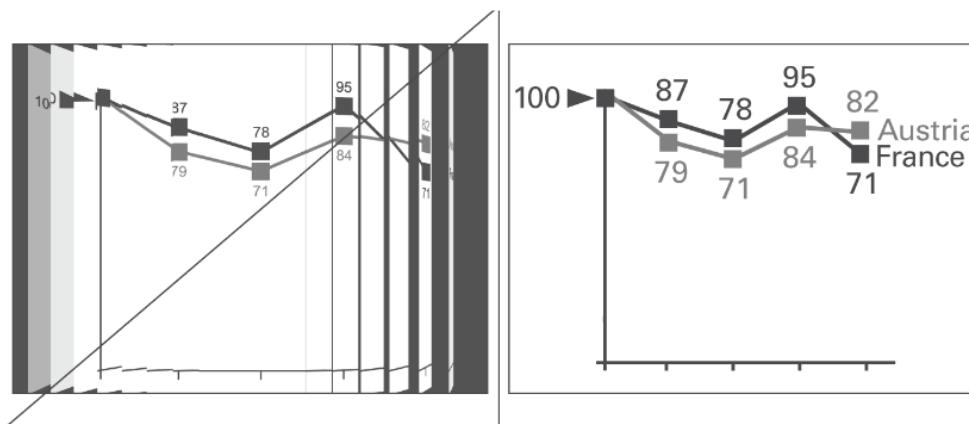
- EX5.3 Show correlations



## IBCS- Simplify: Avoid clutter

Avoid all unnecessary and decorative components in reports, presentations and dashboards.

- SI1 Avoid unnecessary components
  - SI1.1 Avoid cluttered layouts
  - SI1.2 Avoid colored or filled backgrounds
  - SI1.3 Avoid animations and transition effects



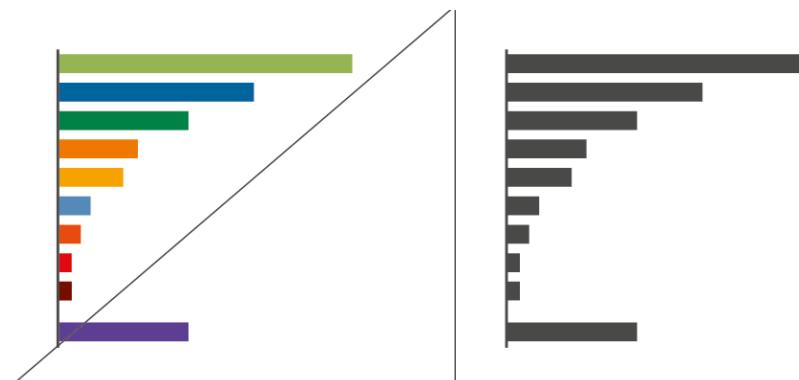
|        | AC     | ΔPL |
|--------|--------|-----|
| 34 567 | 321    |     |
| 22 343 | -1 122 |     |
| 1 231  | 34     |     |
| 32 557 | -234   |     |
| 8 990  | 2 289  |     |
| 11 887 | 199    |     |

|        | AC     | ΔPL |
|--------|--------|-----|
| 34 567 | +321   |     |
| 22 343 | -1 122 |     |
| 1 231  | +34    |     |
| 32 557 | -234   |     |
| 8 990  | +2 289 |     |
| 11 887 | +199   |     |

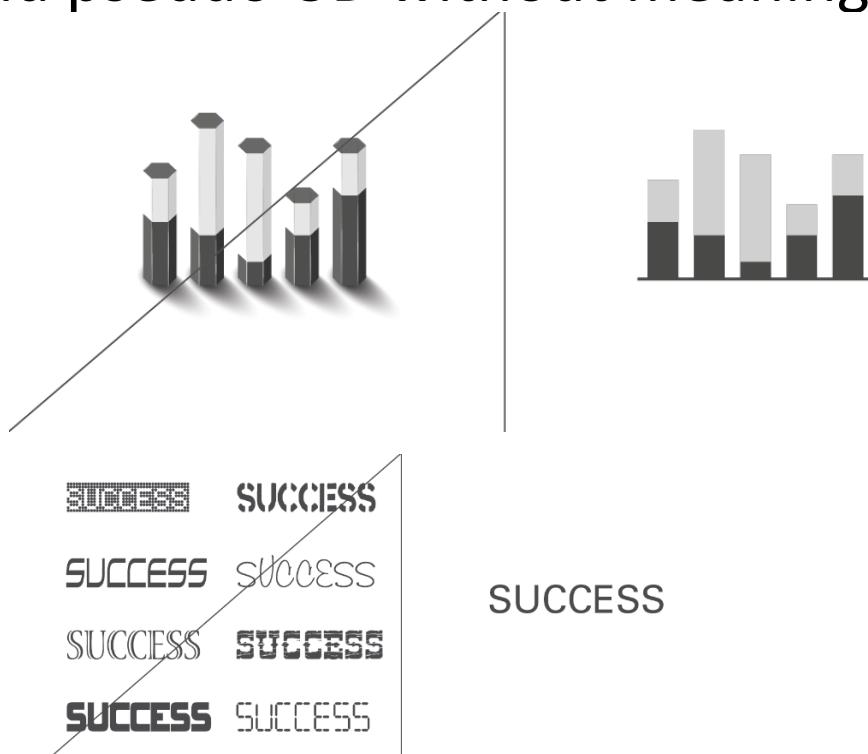
## IBCS- Simplify: Avoid clutter

Avoid all unnecessary and decorative components in reports, presentations and dashboards.

- SI2 Avoid decorative styles
  - SI2.1 Avoid frames, shades, and pseudo-3D without meaning
  - SI2.2 Avoid decorative colors



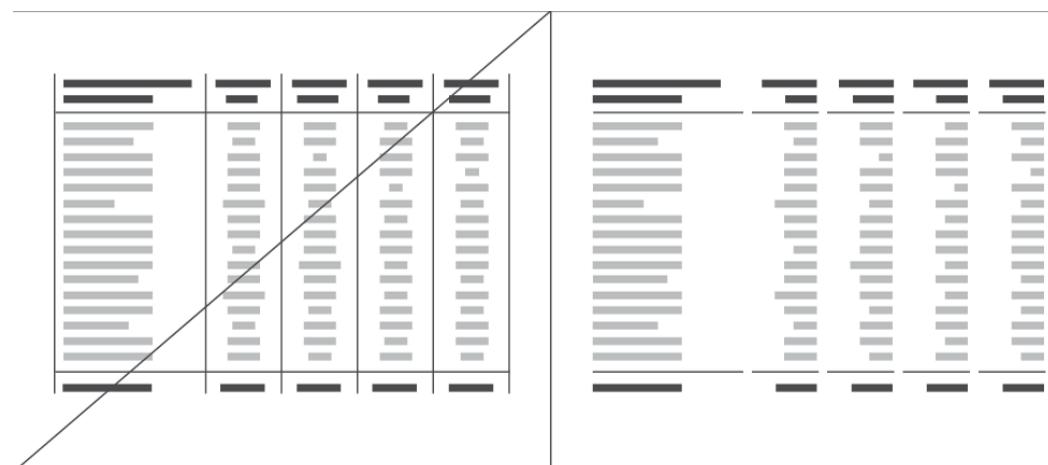
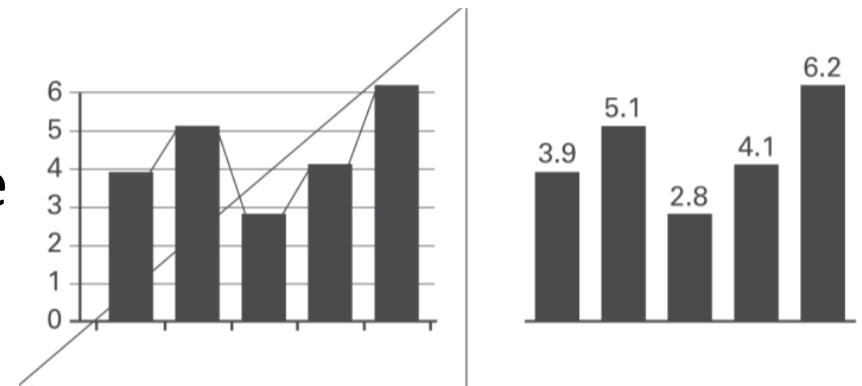
- SI2.3 Avoid decorative fonts



## IBCS- Simplify: Avoid clutter

Avoid all unnecessary and decorative components in reports, presentations and dashboards.

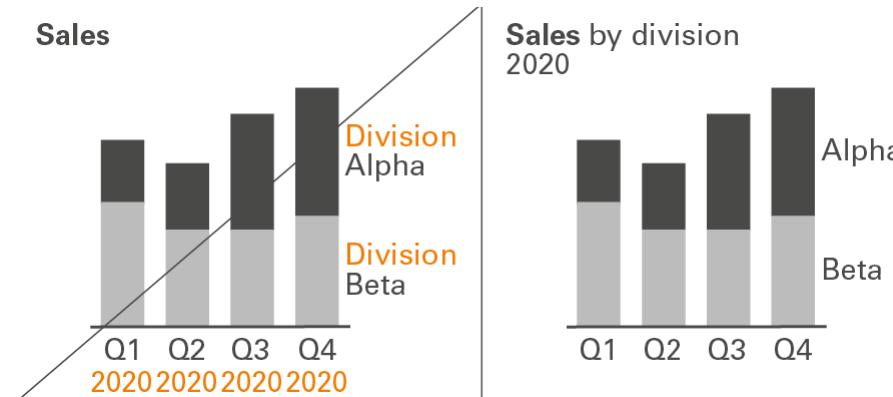
- SI3 Replace with cleaner layout
  - SI3.1 Replace grid lines and value axes with data labels
  - SI3.2 Avoid vertical lines by right-align data



## IBCS- Simplify: Avoid clutter

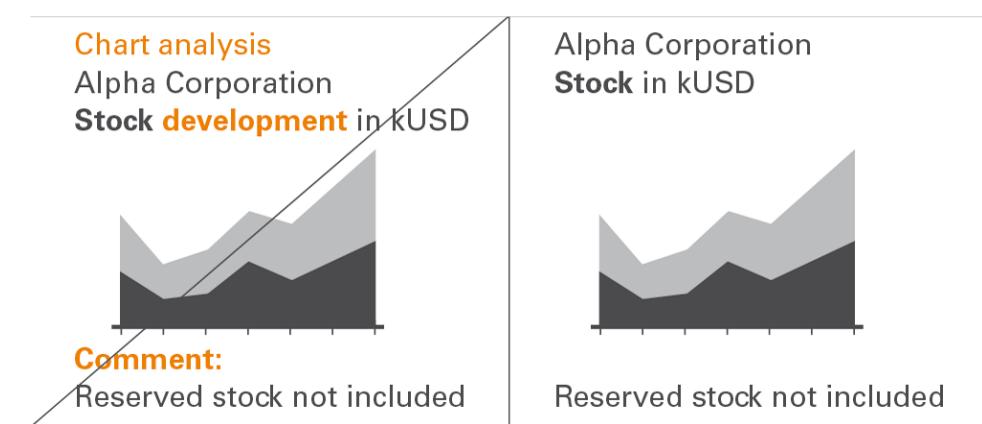
Avoid all unnecessary and decorative components in reports, presentations and dashboards.

- SI4 Avoid redundancies
  - SI4.1 Avoid superfluous extra words
  - SI4.2 Avoid obvious terms
  - SI4.3 Avoid repeated words



The table shows sales data for Europe and its constituent countries in 2020. The columns are 'Total sales in kEUR' and 'Sales in kEUR'. The table includes a 'Full year 2020' summary row and a 'Sum of Europe' row. The data is identical in both columns.

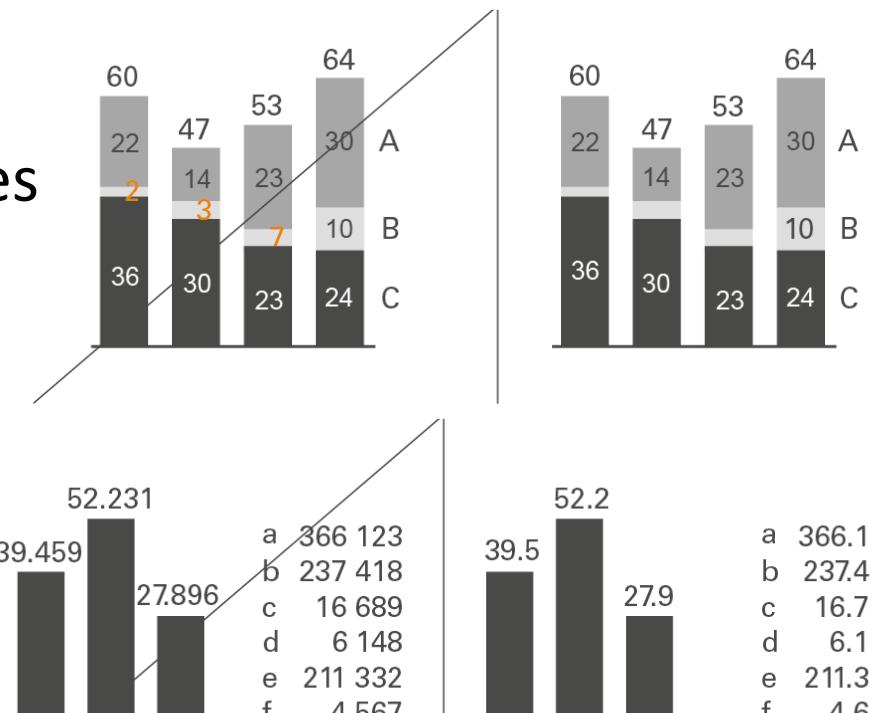
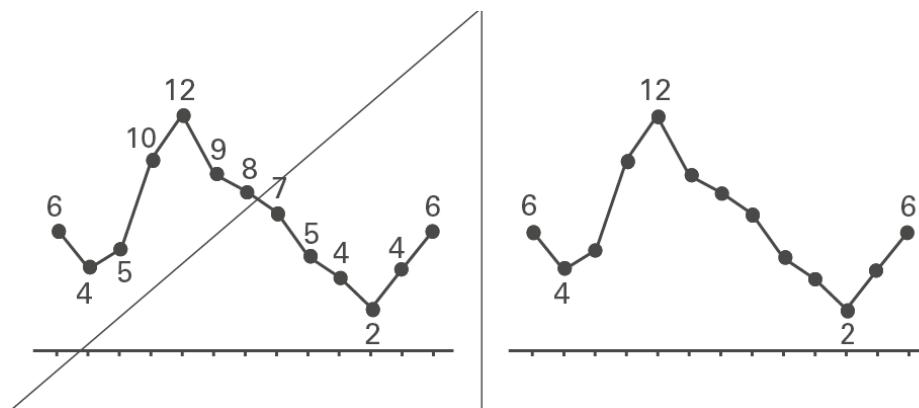
|                      | Total sales in kEUR | Sales in kEUR |
|----------------------|---------------------|---------------|
| Germany              | 788                 | 788           |
| France               | 34                  | 34            |
| Italy                | 122                 | 122           |
| Other countries      | 345                 | 345           |
| <b>Sum of Europe</b> | <b>1 289</b>        | <b>1 289</b>  |



## IBCS- Simplify: Avoid clutter

Avoid all unnecessary and decorative components in reports, presentations and dashboards.

- SI5 Avoid distracting details
  - SI5.1 Avoid labels for small values
  - SI5.2 Avoid long numbers
  - SI5.3 Avoid unnecessary labels



## IBCS- Structure: Organize Content

Reports, presentations and dashboards follow a logical structure forming a convincing storyline.

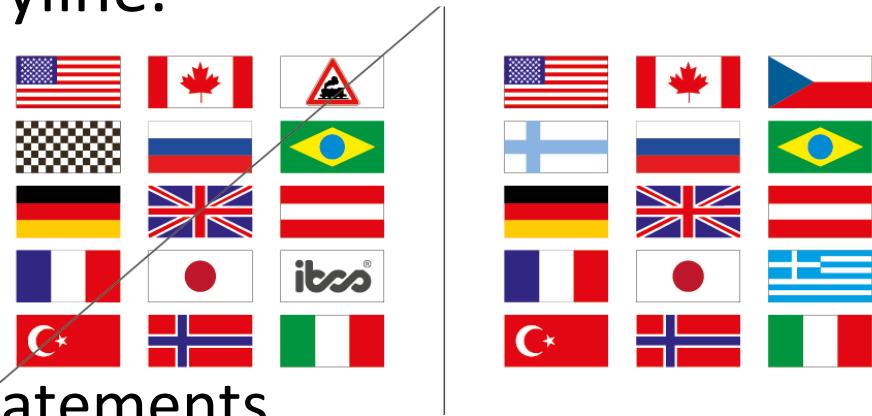
- ST1 Use consistent elements
  - ST1.1 Use consistent items
  - ST1.2 Use consistent types of statements
  - ST1.3 Use consistent wording

**Activities**

- + Sell division A
- + **Restructuring of** division B
- + Split division C
- + Improve division D

**Activities**

- + Sell division A
- + **Restructure** division B
- + Split division C
- + Improve division D



**Objectives**

- + Improve quality
- + Reduce costs
- + **We have** delays
- + Reduce price

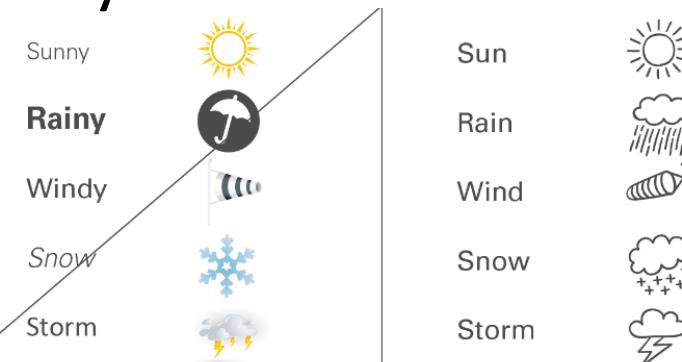
**Objectives**

- + Improve quality
- + Reduce costs
- + **Avoid** delays
- + Reduce price

## IBCS- Structure: Organize Content

Reports, presentations and dashboards follow a logical structure forming a convincing storyline.

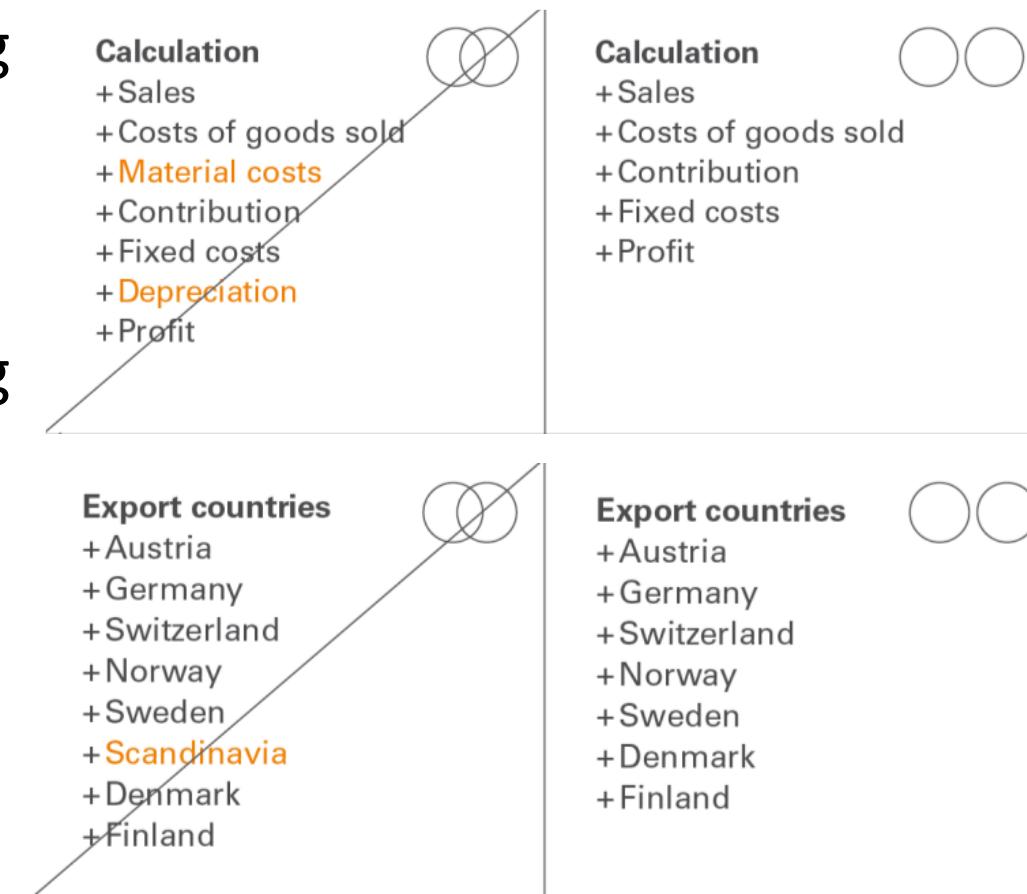
- ST1.4 Use consistent visualizations
- ST2 Build non-overlapping elements
  - ST2.1 Build non-overlapping report structures



## IBCS- Structure: Organize Content

Reports, presentations and dashboards follow a logical structure forming a convincing storyline.

- ST2.2 Build non-overlapping business measures



- ST2.3 Build non-overlapping Structure dimensions

## IBCS- Structure: Organize Content

Reports, presentations and dashboards follow a logical structure forming a convincing storyline.

- ST3 Build collectively exhaustive elements

- ST3.1 Build exhaustive arguments

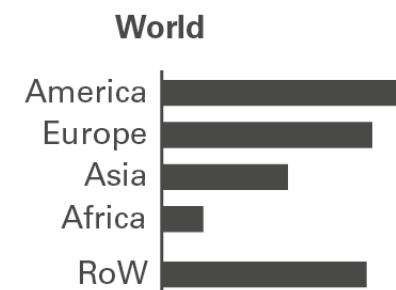
We think of all options

- + New products, old location
- + Old products, old location
- + New products, new location

We think of all options

- + New products, old location
- + Old products, old location
- + New products, new location
- + Old products, new location

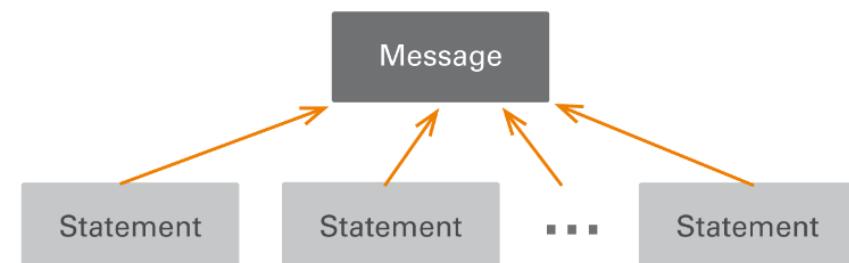
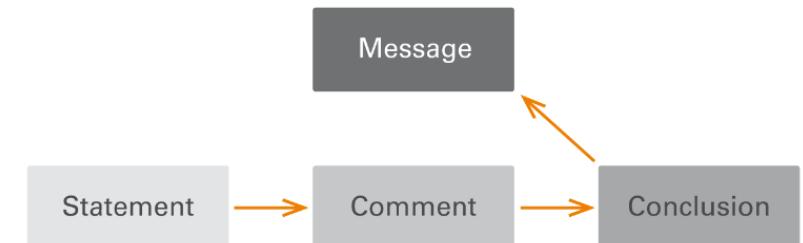
- ST3.2 Build exhaustive structures



## IBCS- Structure: Organize Content

Reports, presentations and dashboards follow a logical structure forming a convincing storyline.

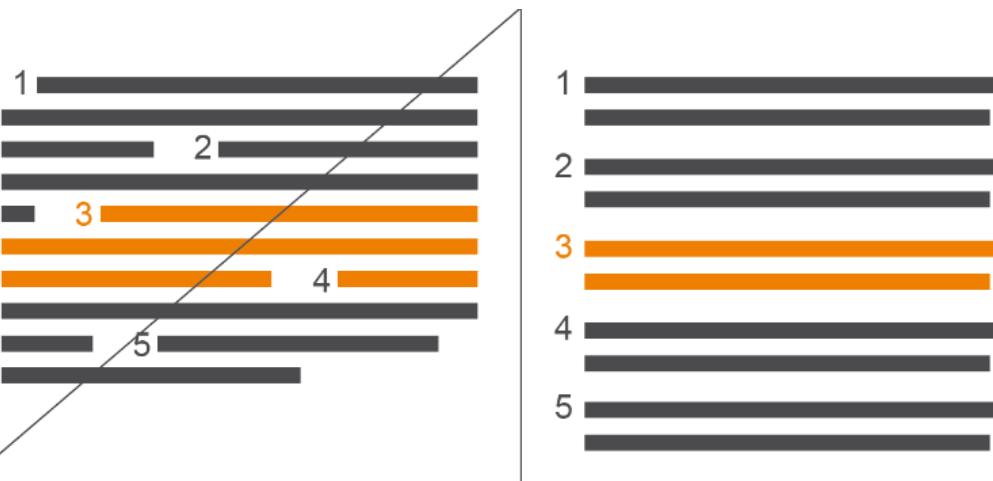
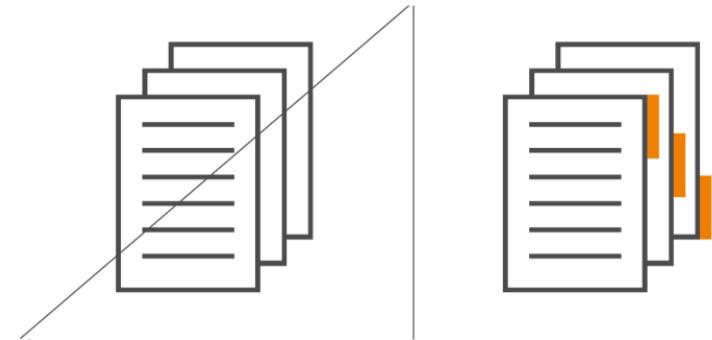
- ST4 Build hierarchical structures
  - ST4.1 Use deductive reasoning
  - ST4.2 use inductive reasoning



## IBCS- Structure: Organize Content

Reports, presentations and dashboards follow a logical structure forming a convincing storyline.

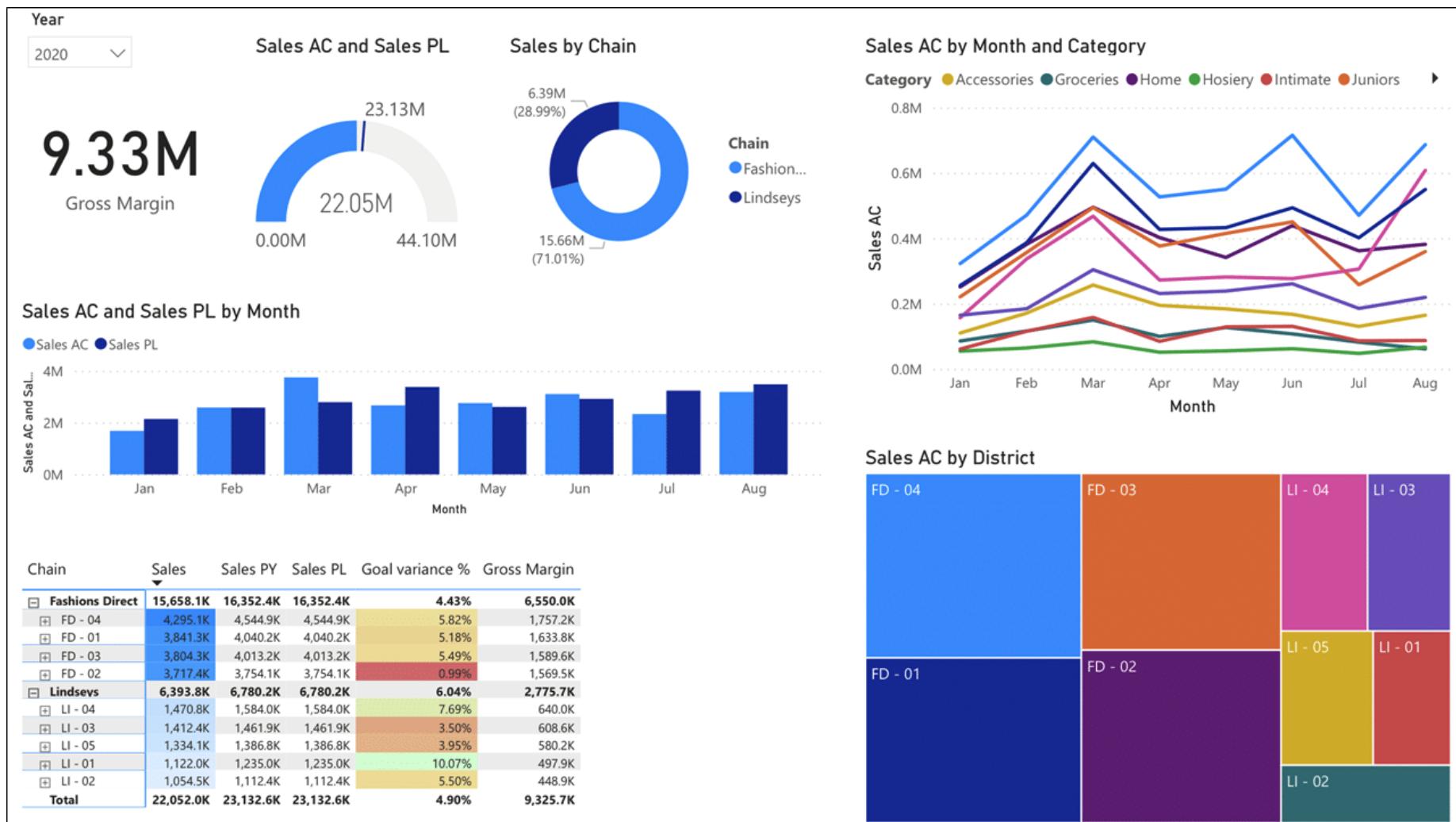
- ST5 Visualize structure
  - ST5.1 Visualize structure in reports
  - ST5.2 Visualize structure in tables
  - ST5.3 Visualize structure in notes



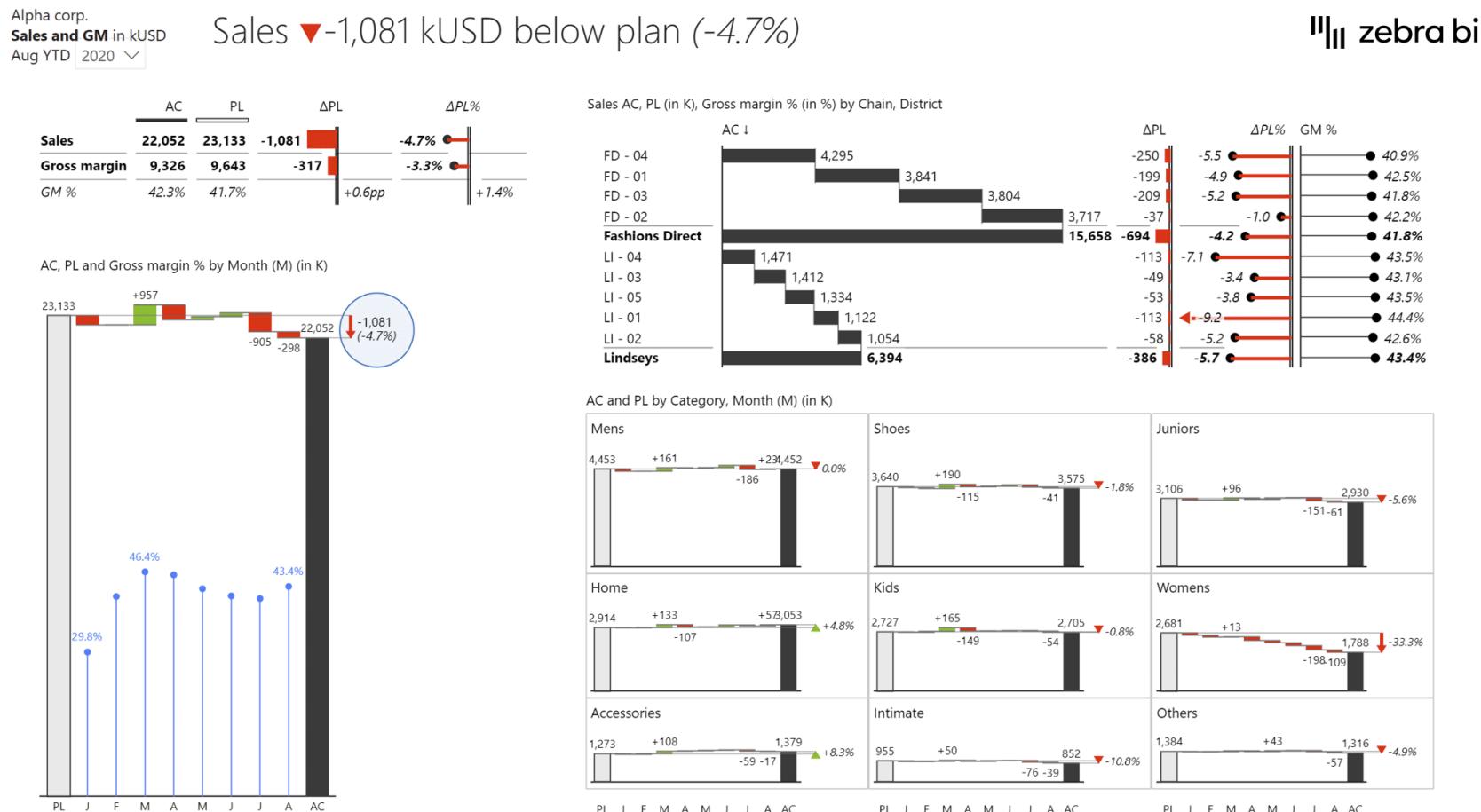
|           | J  | F  | M  | Q1  |
|-----------|----|----|----|-----|
| Hamburg   | 12 | 11 | 9  | 32  |
| Berlin    | 19 | 16 | 14 | 49  |
| North     | 31 | 27 | 23 | 81  |
| Munich    | 16 | 14 | 15 | 45  |
| Stuttgart | 23 | 20 | 21 | 64  |
| South     | 39 | 34 | 36 | 109 |
| Germany   | 70 | 61 | 59 | 190 |

|           | J  | F  | M  | Q1  |
|-----------|----|----|----|-----|
| Hamburg   | 12 | 11 | 9  | 32  |
| Berlin    | 19 | 16 | 14 | 49  |
| North     | 31 | 27 | 23 | 81  |
| Munich    | 16 | 14 | 15 | 45  |
| Stuttgart | 23 | 20 | 21 | 64  |
| South     | 39 | 34 | 36 | 109 |
| Germany   | 70 | 61 | 59 | 190 |

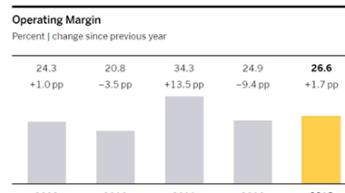
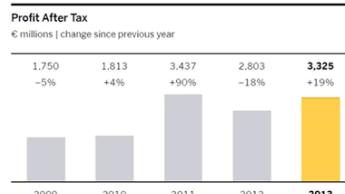
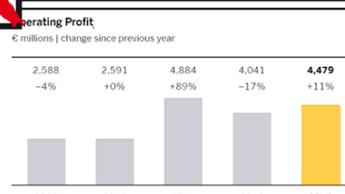
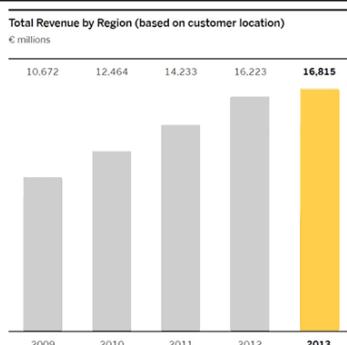
## IBCS- How we do apply it? An example (Source: ZebraBI)



## IBCS- How we do apply it? An example (Source: ZebraBI)



## IBCS- How we do apply it? An example (Source: IBCS)



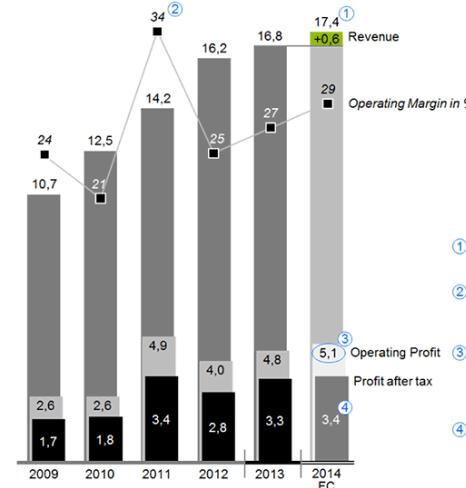
SAP AR 2013 pages 94ff

We expect a revenue increase of bEUR 0,6 in 2014 mainly due to the acquisition of Alpha. This will lead to an operating profit of bEUR 5,1



acquisition

SAP AG  
Revenue, Operating Profit and Profit after Tax in bEUR  
Operating Margin in %  
2009..2013



① Revenue 2014: The growth of bEUR 0,6 is due to the acquisition of Alpha.

② Operating Margin 2011: Again, aenean commodo ligula eget dolor. Aenean massa. Cum sociis bEUR 0,1 natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

③ Operating Profit 2014: Compared to donec quam felis, ultricies nec, pellentesque eu, pretium bEUR 0,3 quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. lo af

④ Profit after tax 2014: Because of bEUR 2,3 birur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget cond

Caution: Data of 2014 are made up for demonstration purposes

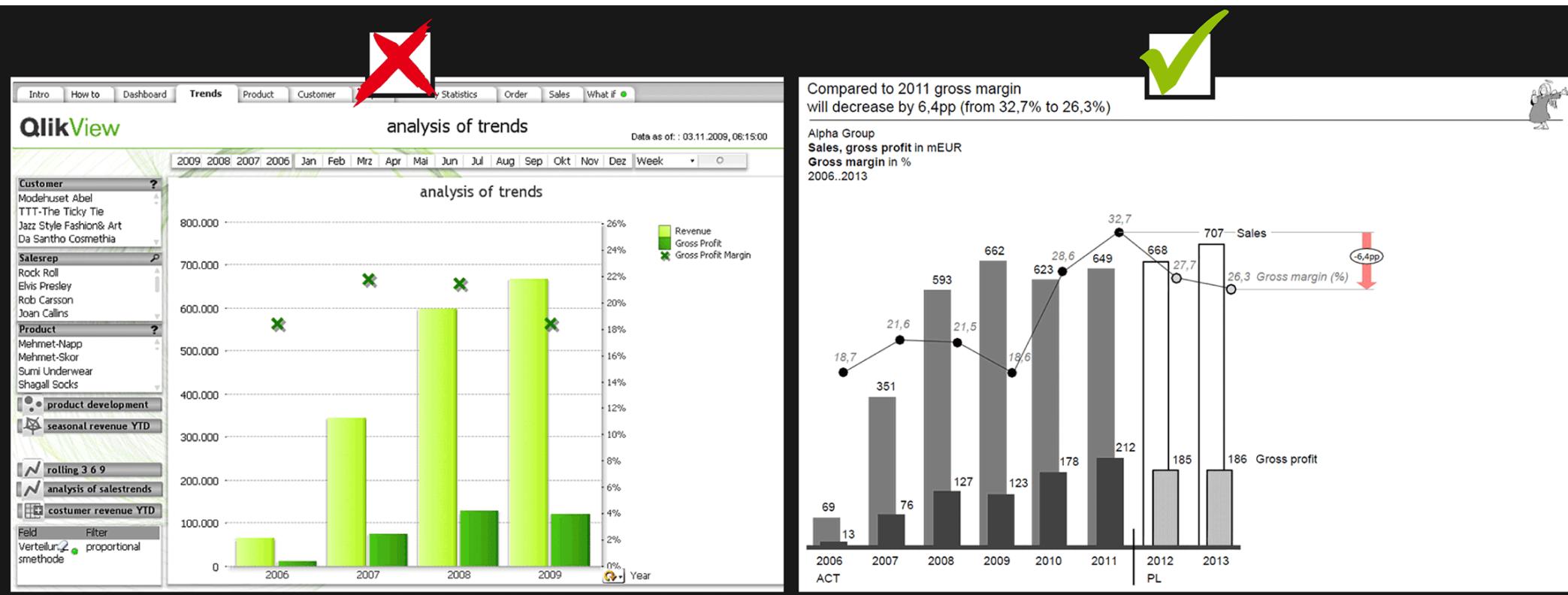
## IBCS- How we do apply it? An example (Source: IBCS)



| <b>Total Result</b> |                | <b>6,738,734.78 €</b> | <b>6,665,225.07 €</b> |
|---------------------|----------------|-----------------------|-----------------------|
| ▼ 10                | Americas       | 1,902,453.40 €        | 1,854,604.14 €        |
| ► 1                 | United States  | 1,010,268.15 €        | 1,051,664.96 €        |
| ► 2                 | Canada         | 447,983.74 €          | 437,579.29 €          |
| ► 3                 | Mexico         | 77,521.00 €           |                       |
| ► 4                 | Columbia       | 297,021.84 €          | 301,593.07 €          |
| ► 5                 | Argentina      | 69,658.67 €           | 63,766.82 €           |
| ▼ 11                | Europe         | 3,177,731.45 €        | 3,113,743.11 €        |
| ► 1                 | United Kingdom | 627,810.75 €          | 597,254.99 €          |
| ► 2                 | Ireland        | -10,879.34 €          | 43,885.82 €           |
| ► 11                | Germany        | 1,619,139.28 €        | 1,526,285.05 €        |
| ► 20                | Spain          | 62,801.15 €           | 59,425.35 €           |
| ► 21                | Poland         | 784,520.19 €          | 842,321.40 €          |
| ► 24                | Switzerland    | 72,700.00 €           | 69,506.96 €           |
| ► 25                | Netherlands    | 21,639.42 €           | -24,936.46 €          |
| ▼ 16                | Asia-Pacific   | 1,446,992.86 €        | 1,425,083.69 €        |

|                 | AC           | BU           | ΔBU        | ΔBU%       |
|-----------------|--------------|--------------|------------|------------|
| United States   | 1.010        | 1.052        | -41        | -4%        |
| Canada          | 448          | 438          | +10        | +2%        |
| Mexico          | 78           | 0            | +78        | ∞          |
| Columbia        | 297          | 302          | -5         | -2%        |
| Argentina       | 70           | 64           | +6         | +9%        |
| <b>Americas</b> | <b>1.902</b> | <b>1.855</b> | <b>+48</b> | <b>+3%</b> |
| United Kingdom  | 628          | 597          | +31        | +5%        |
| Ireland         | -11          | 44           | -55        | -125%      |
| Germany         | 1.619        | 1.526        | +93        | +6%        |
| Spain           | 63           | 59           | +3         | +6%        |
| Poland          | 785          | 842          | -58        | -7%        |
| Switzerland     | 73           | 70           | +3         | +5%        |
| Netherlands     | 22           | -25          | +47        | +187%      |
| <b>Europe</b>   | <b>3.178</b> | <b>3.114</b> | <b>+64</b> | <b>+2%</b> |

## IBCS- How we do apply it? An example (Source: IBCS)



## Summary

**IBCS**-standards for clear, consistent and actionable business communication using charts, tables and reports.  
Enhances clarity, comparability and efficiency.

## Why?

Better decision-making through standard and interpretable visuals.  
Minimizes misinterpretation through consistent interfaces.

## Applications

Financial reports, dashboards, operational KPIs, etc.  
Compatible with tools like PowerBI, Excel, Qlik, SAP Analytics,...

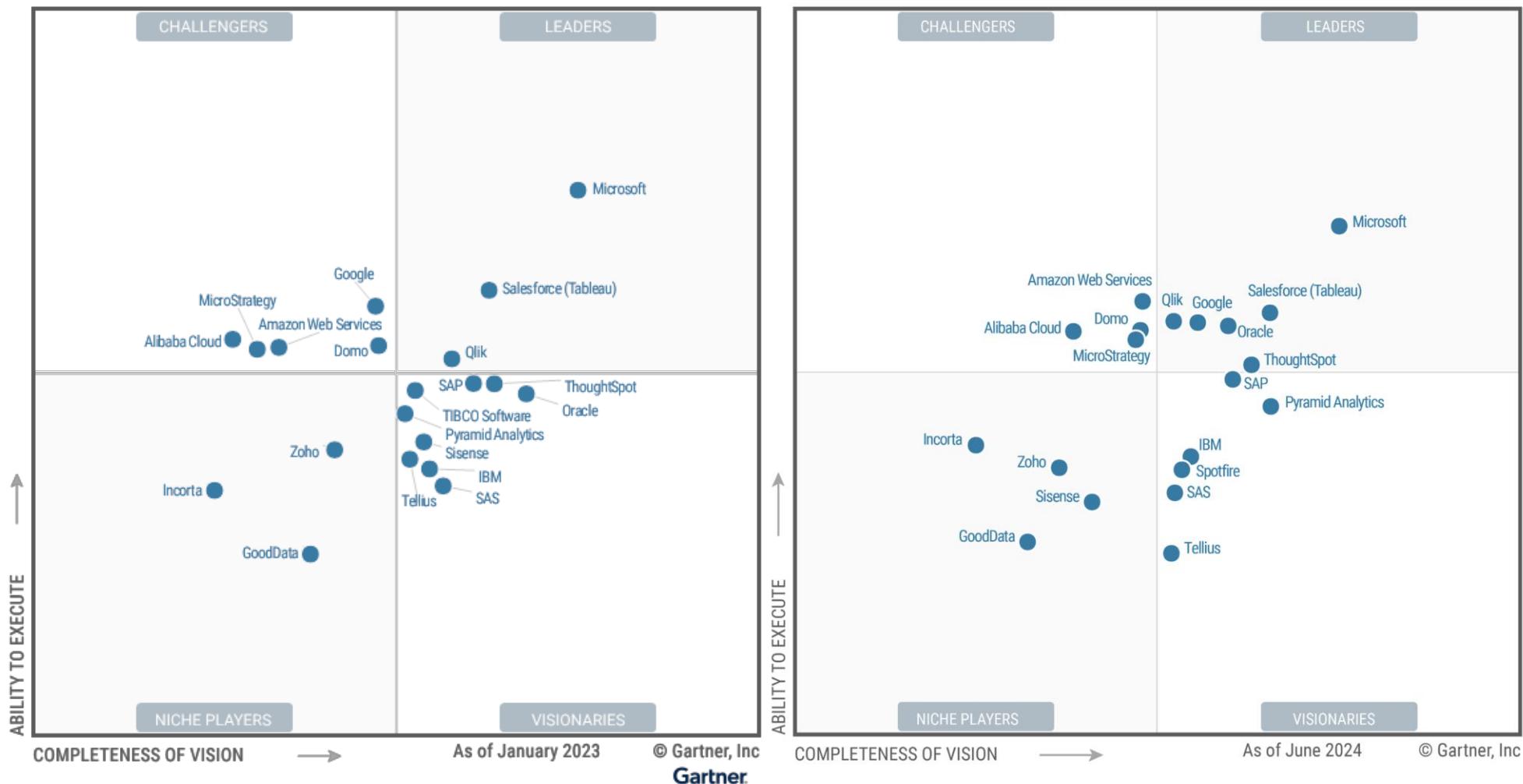
# BI: FUTURE

NLP, limpieza de datos automatica, ...

# BI:Future

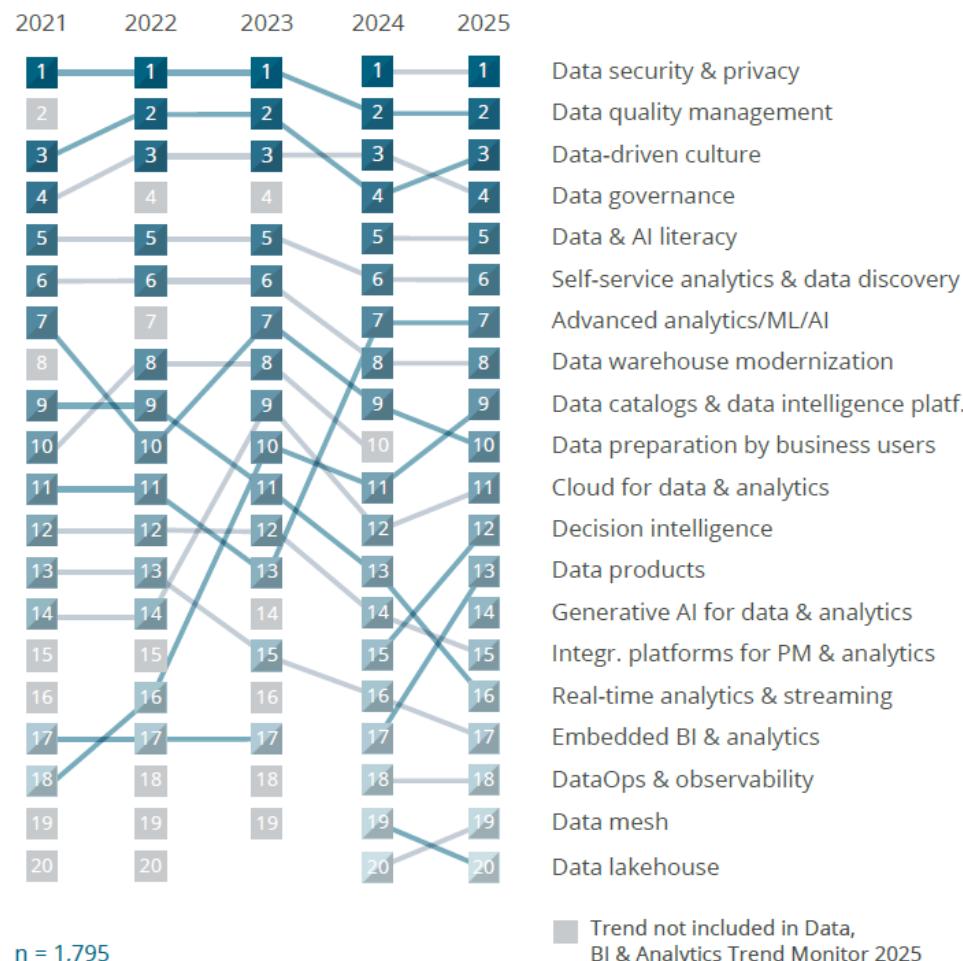
gobernanza de datos

## Magic Quadrant for Analytics and Business Intelligence Platforms

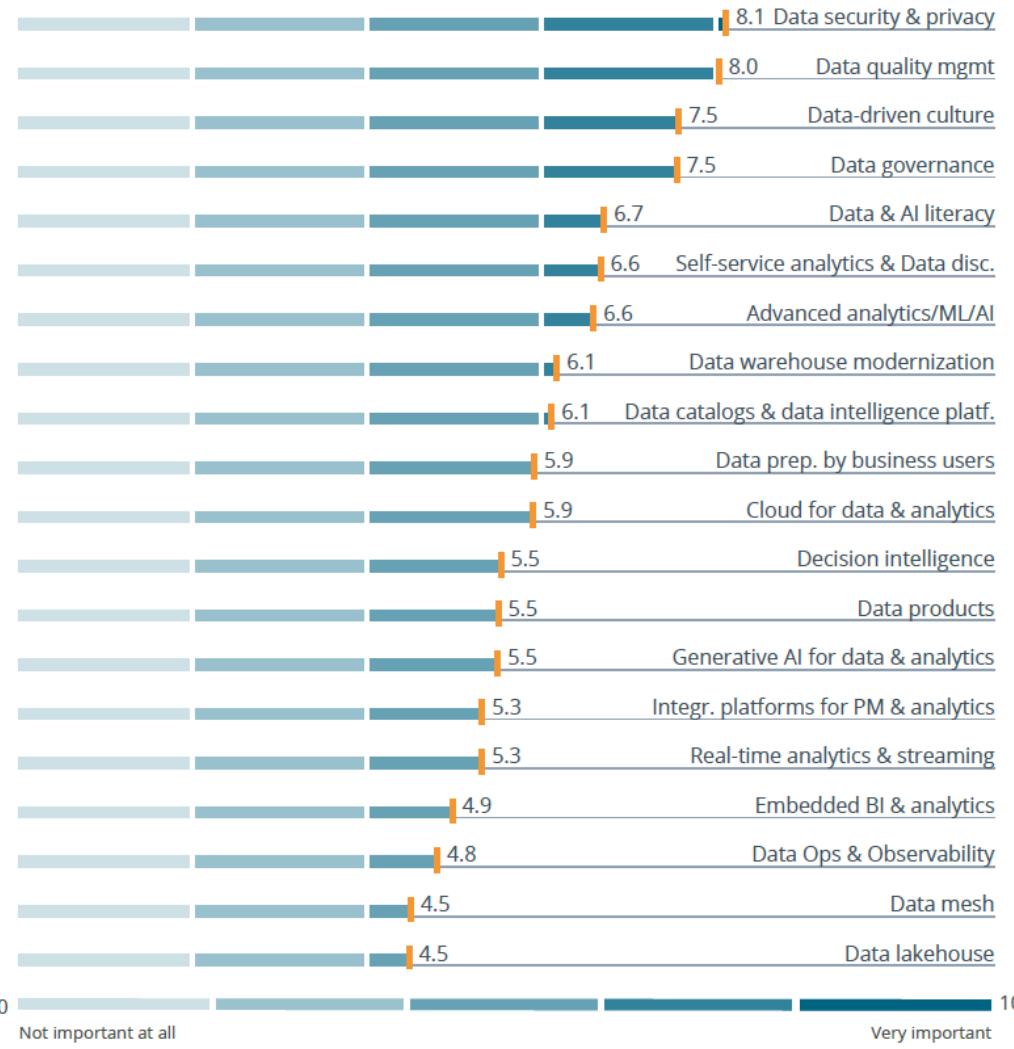


# BI:Future

- BI continues to be a top priority for organizations. Very competitive market.
- Fastest growing technologies in IT (also # of professionals)

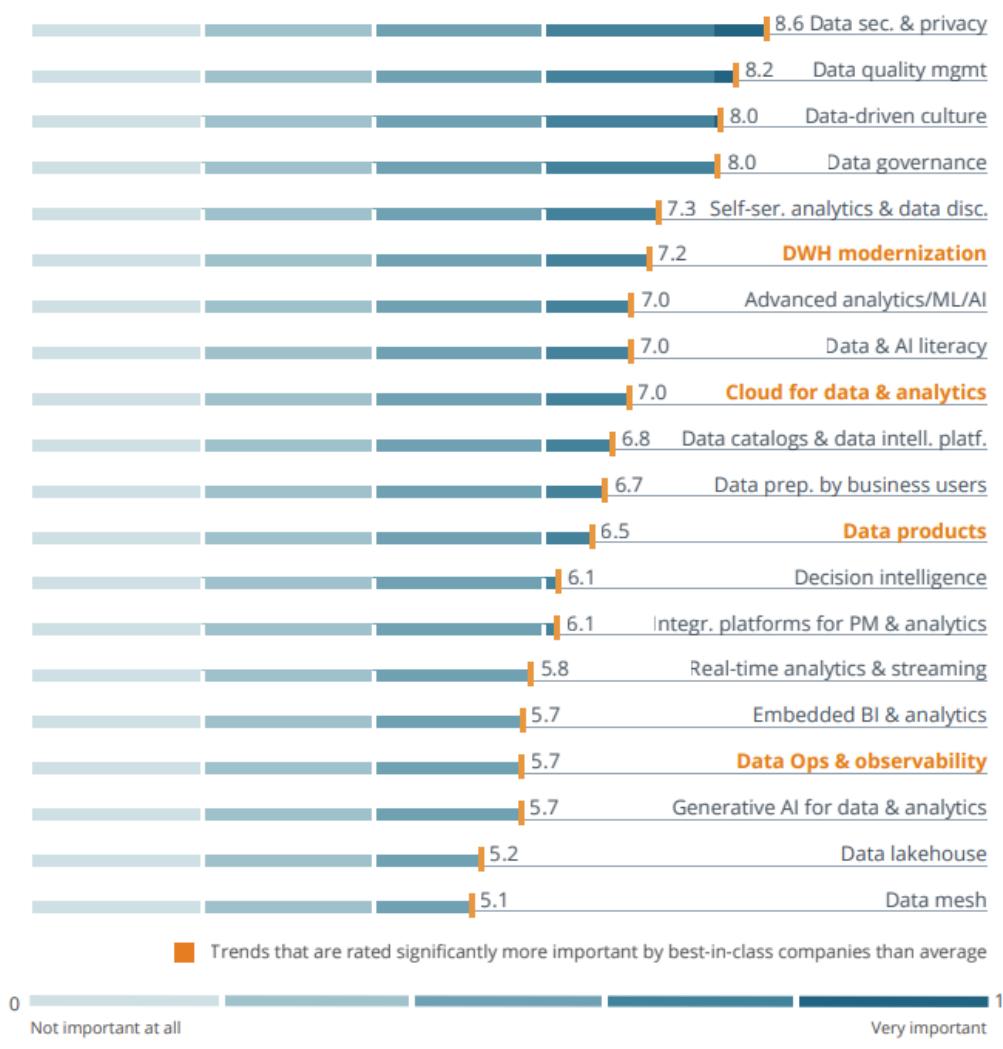


Source: Data, BI and Analytics Trend Monitor 2025. BARC Research Study



n = 1,795

Overall Results



n = 161

Best-in-class Companies

■ Trends that are rated significantly more important by best-in-class companies than average

# BI:Future. Recommendations

1

## Venture into trending topics

The focus on **decision intelligence** and **predictive planning** has grown, especially as companies are now scaling their use of **machine learning** to automate decision-making processes. We can confidently suggest starting pilot projects in these areas, which are delivering measurable benefits in operational decision-making, **embedded BI** and **AI-driven automation**. These initiatives, alongside efforts in **data culture**, **security** and **quality**, remain vital.

3

## Pay attention to data governance

Data governance is becoming increasingly vital as businesses adopt decentralized data ownership and self-service analytics. With more users accessing and analyzing data, governance frameworks are needed to ensure integrity, consistency and security. As decision intelligence and AI play a larger role, high-quality, governed data is crucial for reliable automated decision-making. This shift is driving the rise of federated governance models, which balance flexible data access with robust oversight to maintain trust in data-driven decisions.

5

## Playtime is over! Is it?

As companies move beyond prototypes, the focus is shifting toward **operationalizing AI**, particularly in areas like **decision intelligence** and **predictive planning**. Businesses are scaling AI implementations to automate operational processes in areas such as **fraud detection** and **dynamic pricing**.

2

## Enable your staff

With the growing importance of **data and AI literacy**, empowering employees across all roles to engage with **AI-driven tools** and **decision intelligence** is crucial for operational success. This ensures that not only specialized data professionals, but also employees at every level, can confidently interpret and use data to drive informed decisions. A lack of skills remains one of the most significant barriers to progress, not only in AI adoption but across the broader data and analytics landscape.

4

## Modernize your information architecture

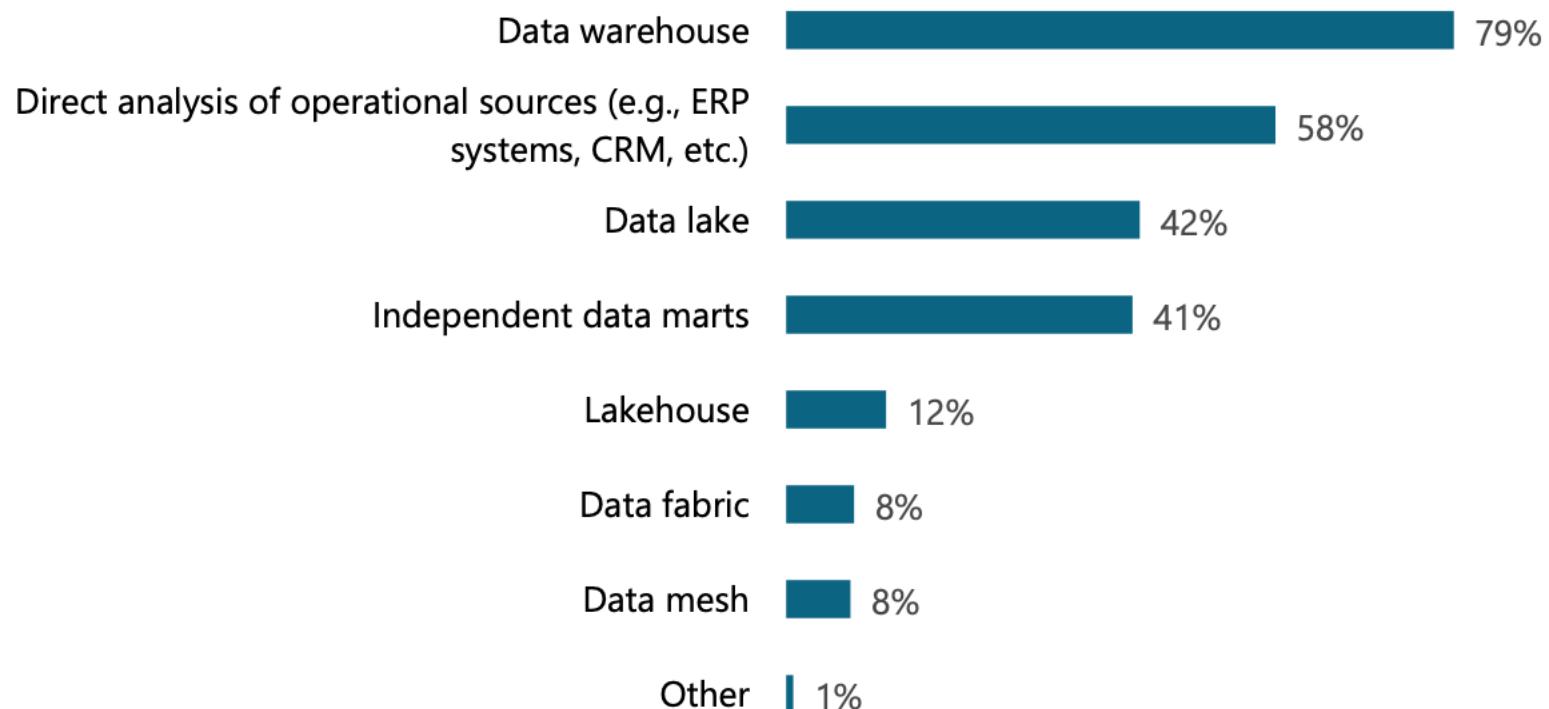
Review your existing information architecture to ensure it can handle the growing demand for cloud-based analytics, real-time data and AI-driven insights. The need for data warehouse modernization remains critical, as companies aim to integrate cloud-first strategies and support poly-structured data. Embedded BI is increasingly part of this architecture, enabling real-time decision-making within operational workflows. A modern, flexible architecture will support advanced analytics, enabling businesses to respond to the ever-growing complexity of data environments.

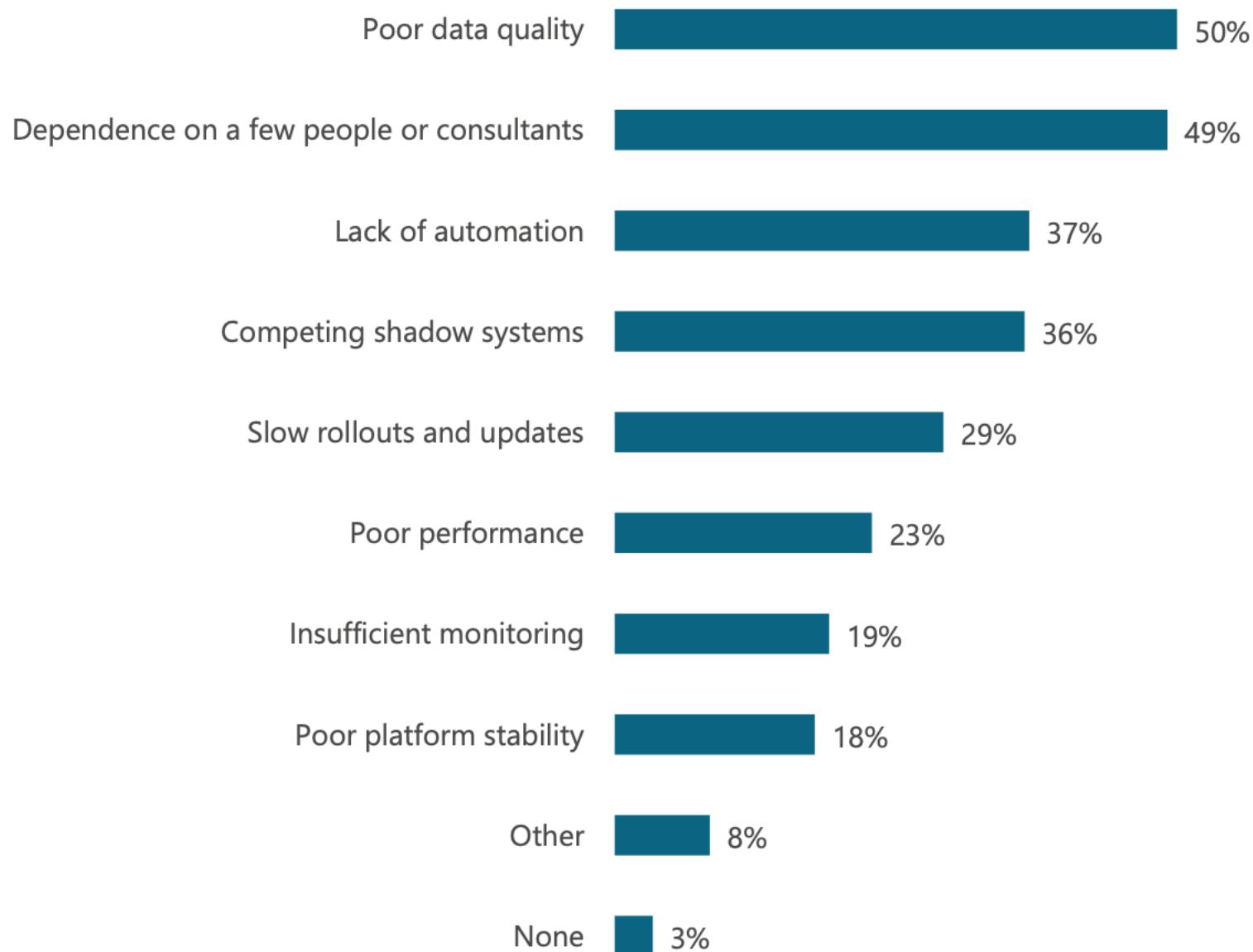
6

## Get ready for a data-driven culture

The foundational pillars from the **BARC Data Culture Framework** continue to be critical, with both **data literacy** and **AI literacy** becoming increasingly important in 2025. Companies must foster a culture that not only supports data-driven decision-making but also incorporates AI as a fundamental part of their operations. Data literacy and AI literacy are essential for building a culture that embraces the opportunities and challenges of the future.

**Figure 2a. Which of the following types of architecture do you have in your environment? (n=236)**

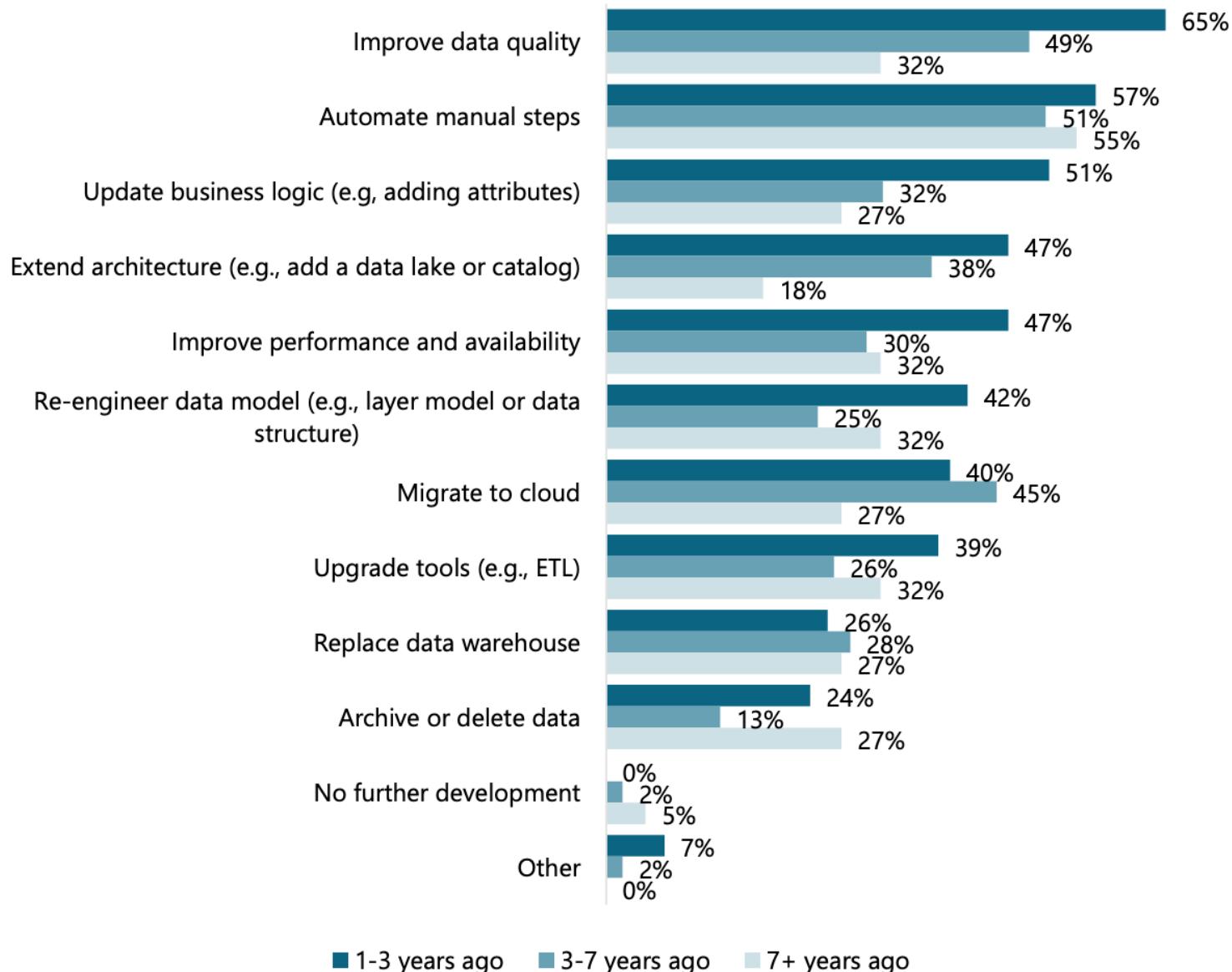


**Figure 3. What are the biggest challenges in your current analytics environment? (n=238)**

Source: Data Warehouse and data vault adoption trends. BARC Research Study

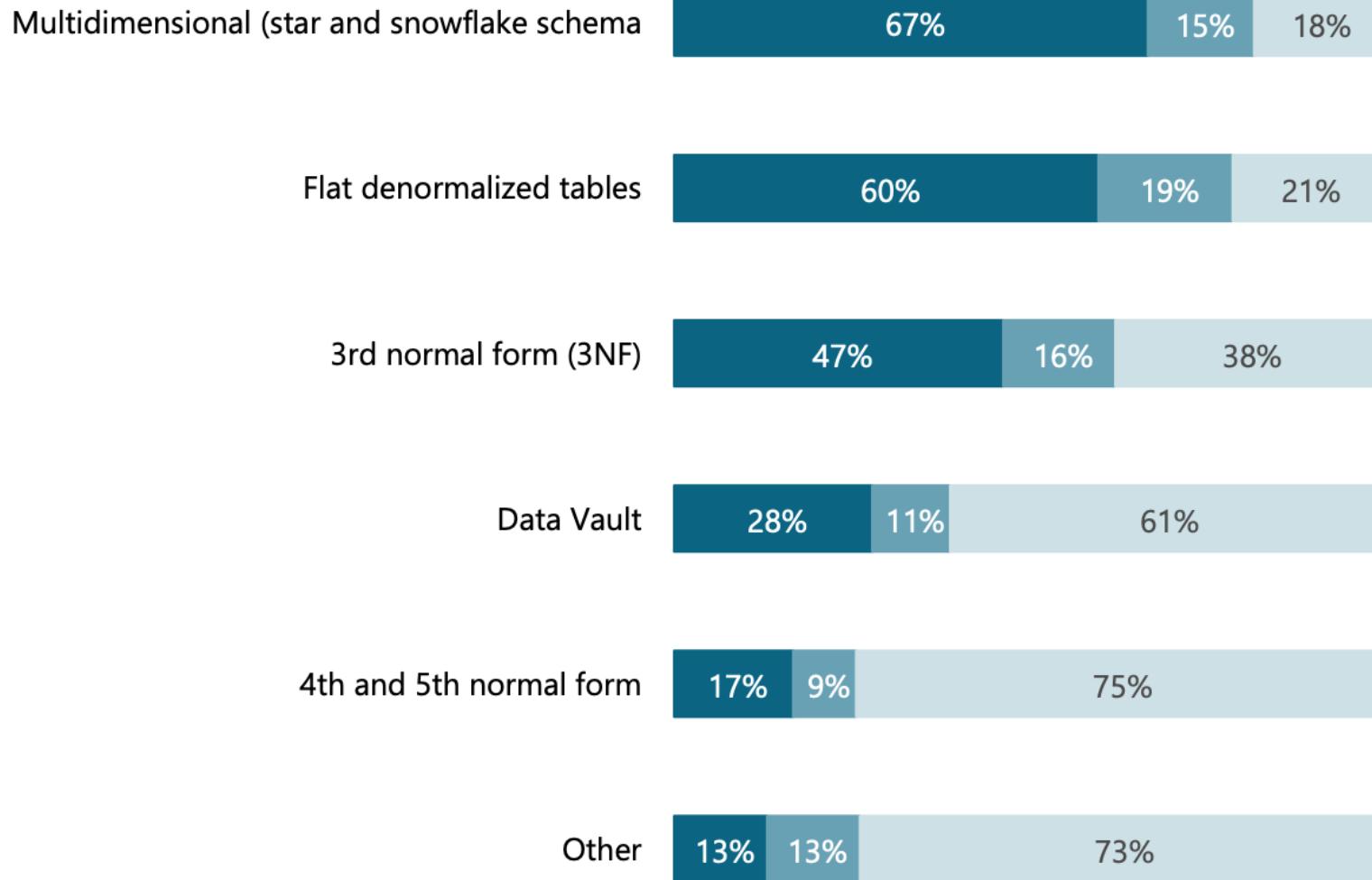
# BI:Future

**Figure 6. What environment updates and modernization steps do you plan in the next 3 years?**  
(n=237)



Source: Data Warehouse and data vault adoption trends. BARC Research Study

**Figure 9. What data modeling techniques does your company use, or has it used in the past?**  
(n=238)



■ Currently Using   ■ Previously Used   ■ Not at all

Source: Data Warehouse and data vault adoption trends. BARC Research Study

# BI:Future. Big Data and Key Technologies

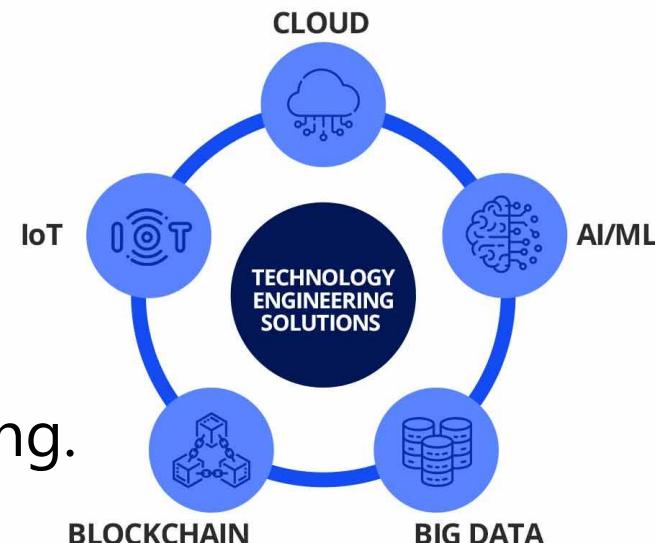
**Big Data** changes how enterprises manipulate data.

Shift to **predictive**, **real-time**, and **user-friendly** analytics.

**No more data silos:** Lakes, Warehouses, and Lakehouses complement each other.

## Key Technologies in BI Future:

- **AI:** Automate data analysis and decisions.
- **ML:** Enhance predictive analysis and modeling.
- **IoT:** Vast real-time data streams.
- **Blockchain:** Secure, transparent storage.
- **Cloud Computing:** Scalable, on-demand processing.

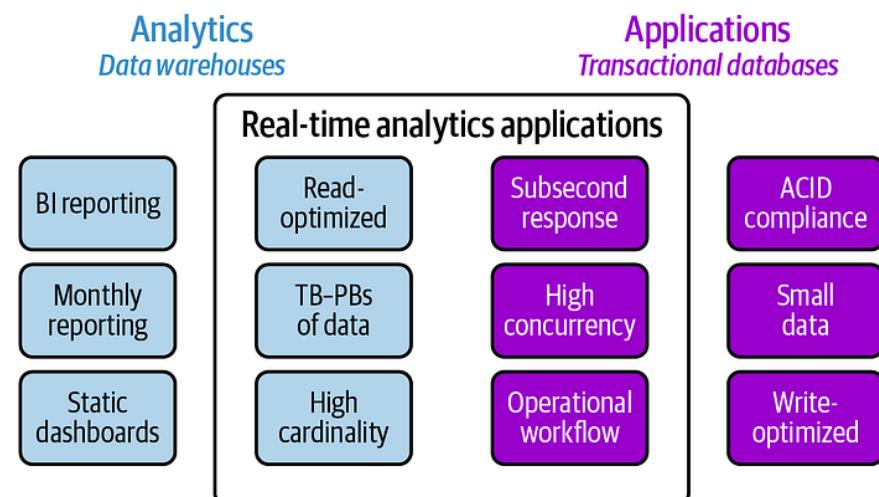
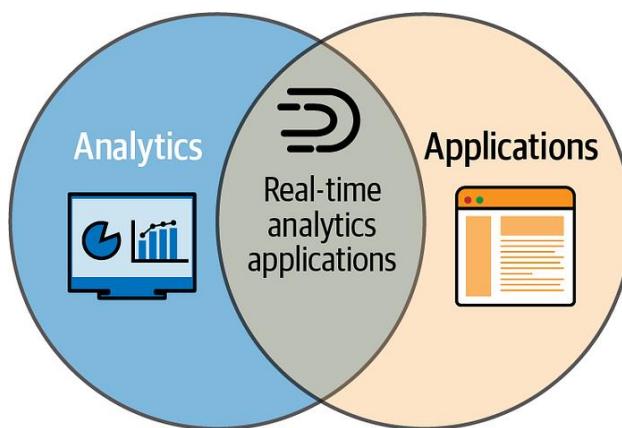


## Real-time Analytics:

- **Immediate decision-making:** trading, social media, logistics.
- Data accuracy, privacy concerns, and infrastructure requirements.

## Infrastructure Perspective:

- **Hybrid:** MultiCloud and on-premise solutions.
- **Elastic:** Adapts to changing needs with data quality a must.



## What we want to improve?

---

Business/User satisfaction

---

Optimize resource usage

---

Enhance productivity

---

Adaptation to changes

---

Better (timely) information

---

ROI

---

Integration

## Challenges:

Find BI professionals understanding technologies and business  
DWH needs to become agile

**Data privacy and Ethics:** Encryption, anonymization, secure data storage.

- **Compliance** with GDPR, and similar legal frameworks for user data protection.
- **Fair use of data**, avoiding bias in data analysis, ensuring transparency in data handling.

**Edge Computing:** Processing data near the source

Less latency, faster decision making and massive data processing

**Areas:**

Healthcare- predictive patient care, customized medicine

Finance-Real time fraud detection, trading, customized banking

Retail- Optimize inventory, predict customer behaviour

Manufacturing- predictive maintenance, quality control, process optimization.

## DataOps:

- Inspired by DevOps, decreases the time from data to value
- Users: Analysts and scientists looking for creating and deploying models and visualizations
- Improves data and analytic pipelines, automating data ingestion, transformation and “orchestrating” of data workflows.

## Decision Intelligence:

Use AI techniques to improve decision making. Use ML, statistics ands analytics to solve business needs.

## Edge Analytics:

Descentralized & near-sensor analysis (usually IoT devices)

# BI:Future. Market Trends

## Market

- \$14.3 billion (2018), \$27.11 billion (2022) , \$29.42 billion (2023), \$54.27 billion (2030), \$63.76 billion (2032). Annual growth rate ~9%
- BI becomes a core component of operations
- By 2023, ~ 33% of large companies implement decision intelligence
- Self-service BI essential for 60% R&D departments

Adoption Rates: 26% (global), 80% (#staff > 5000)

- Cloud-based BI market fastest growing BI segment
- Cloud-based BI being adopted by manufacturing (~58%), and business and financial services (~40% each)

## Job Market

- Growing job market (55% business have dark data)
- The US Bureau of Labor Statistics predicts the creation of around 11.5 million data scientist jobs by 2026.
- Demand for data engineers is projected to grow at a rate of 50% annually.



## Job Market (Cont)

- **Demand for BI Professionals:** The BI field is expected to grow due to the increasing volume of data and the need for data-driven decision-making. Roles such as BI Analysts, Data Scientists, and BI Developers are in high demand.
- **Skills in Demand:**
  - **AI and Machine Learning:** Understanding how to apply these technologies for predictive analysis and automation in BI.
  - **Data Visualization:** Proficiency in tools like Tableau, Power BI, or Qlik for creating interactive dashboards.
  - **Data Management and Governance:** Skills in managing data quality, security, and compliance.
  - **Cloud Proficiency:** Knowledge of cloud BI solutions like AWS, Google Cloud, or Azure for scalability and flexibility.
- **Emerging Roles:**
  - **Decision Intelligence Analysts:** Roles that focus on using BI to directly influence business decisions.
  - **Data Storytellers:** Professionals who can translate complex data into compelling narratives for stakeholders.