Master en
Big Data e Inteligencia Artificial

Module 5

Data Analytics

Lesson 3: Building a data warehouse solution

Juan Carlos Trujillo – Lucentia (jtrujillo@dlsi.ua.es)
Alejandro Maté – Lucentia (amate@dlsi.ua.es)

Universitat d'Alacant
Universidad de Alicante

1

Master en
Big Data e Inteligencia Artificial

Table of contents

- User tools
  - OLAP & MDX
  - Dashboards
  - Scorecards
  - KPIs
- Annex: Analytic solutions
  - Pentaho

Universitat d'Alacant
Universidad de Alicante

2

Introduction

- Users require information to make complex decisions
- Traditionally, reports and analysis are created on demand by IT experts (that is you!) to satisfy decision making needs
  - Reports, Descriptive, Predictive, Prescriptive analytics…
  - Many questions and information needs to be answered!

3

Introduction

- Concrete answers are not always adequate
- Operational decisions: Require quick, specific information. Low level of abstraction.
  - When should we restock our products?
- Strategic decisions: Require information from various sources, validate multiple hypotheses. High level of abstraction.
  - Should we start an expansion in a new county or consolidate our position?
    - How much profit will we be able to achieve in the new country?
    - What is the cost of opening new stores in the main selling points?
    - How much room to grow do we have if we do not expand?

4

OLAP

- In order to satisfy all the information needs we need several tools
- The first one is OLAP: On-line Analytical Processing
- OLAP…
  - Allows decision makers to validate their own hypotheses
  - Facilitates multiple crossing data without looking at multiple data pipelines and sources
  - Improves the understandability of the results

OLAP

- **Self-Service:** IT does no longer provide direct answers for questions
- Instead, the user should satisfy his own decision making needs ***himself***
  - Find answers for his initial questions
  - Formulate new hypotheses based on his findings
  - Validate these new hypotheses on its own
  - Make a decision

Traditional Results vs OLAP Results

- Traditional
  - Results are presented as-is
  - Bi-dimensional (Tables)
  - Static
  - Focused on presenting the results of the analysis
  - Contains a limited number of records
  - Aims to provide an answer for a single or small set of questions

- OLAP
  - Results are hierarchically structured
  - N-Dimensional
  - Dynamic
  - Focused on facilitating information browsing
  - May contain hundreds of thousands of records
  - Aims to provide answers for as many questions as possible

7

Advantages of OLAP Analysis

- More intuitive for the user (levels of abstraction)
- Context-aware analysis
- Analyze large volumes of data with ease
- Easy comparison of values
- Sandbox-like question answering

8

OLAP: Concepts and Operations

- Building an OLAP solution requires careful design around a specific concept:
  - The Analysis Cube
- An analysis cube is an abstraction of the data that determines (i) what information is included and (ii) how we can operate with it
- Every possible cube is defined by two aspects:
  - Its static structure
  - A dynamic set of operations performed

OLAP: Static Structure

- How to structure the information to meet these needs?
- Multidimensional modeling!
  - Facts
  - Dimensions
  - Levels

OLAP: Static Structure

- Facts and measures
  - A fact is an activity or process of the company
  - A fact should have at least a measure that determines its performance
  - Fact measures should always be numeric
  - Facts are not constrained to financial activities only
- Sales, Tickets, Support center calls, Football matches, Classroom lessons….

11

OLAP: Static Structure

- Dimensions
  - Dimensions are concepts or entities that participate in one or more activities (Facts)
  - Dimensions may include multiple concepts organized hierarchically (Levels)
  - Dimensions should have 1:N relationships with the Fact
  - Dimensions can have multiple relationships with the Fact (Roles)
  - Each dimension should be related with the fact through one of its levels only
- Location (City, Province, Country), Product (Product, Family, Type), Customer (Customer, City, Province, Country)…

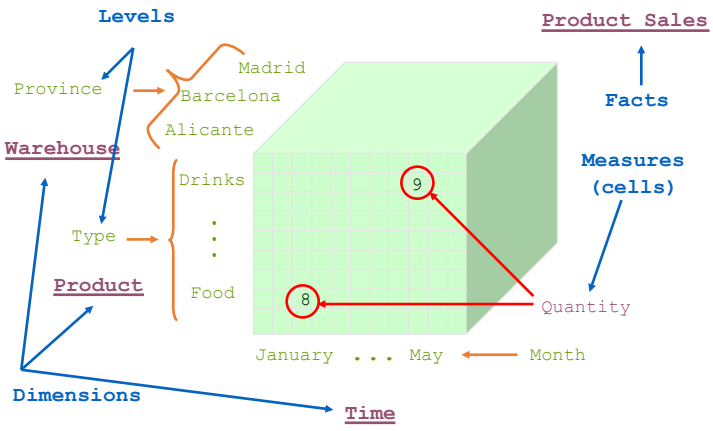
12

OLAP: Static Structure

- Levels
  - Levels represent different abstraction levels of a dimension
  - Each level should correspond with an identifiable concept
  - Levels may have attributes or properties, which provide additional information about each instance of the level
  - If there are two or more levels in a dimension, each level should be related to at least one other level
- Model, Type of Vehicle, Brand; Football Player, Team; Day, Month, Year….

OLAP: Static Structure

- Result:



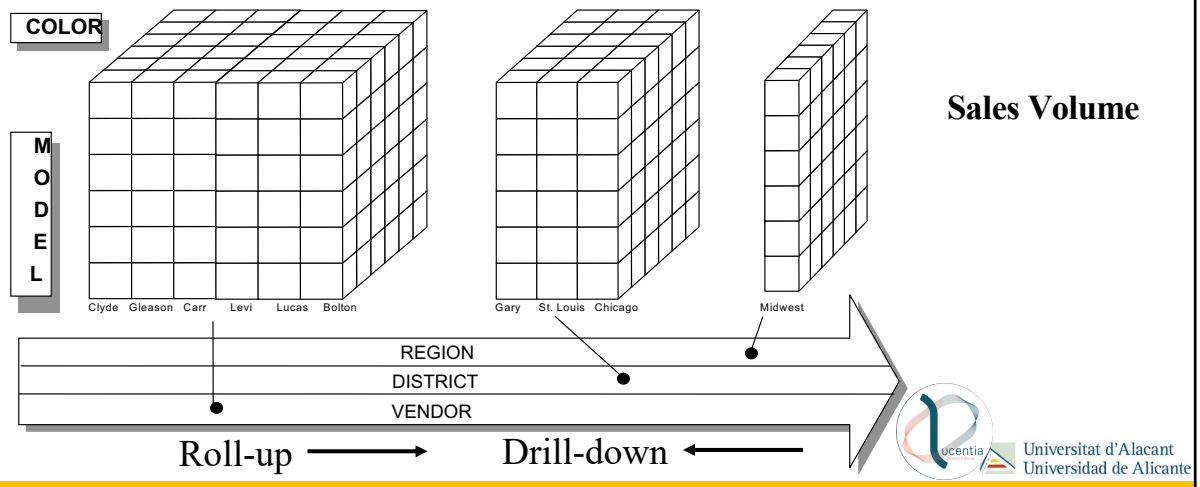
OLAP: Concepts and Operations

Once we have covered the static aspect of an OLAP Cube we can proceed to the dynamic aspect:

- Provides interactivity to the cube
- Is comprised by a set of operations that can be performed:
  - An OLAP operation over a cube always returns a cube
  - The set of operations is implemented in almost every OLAP server
  - The set of operations is almost always written in MDX

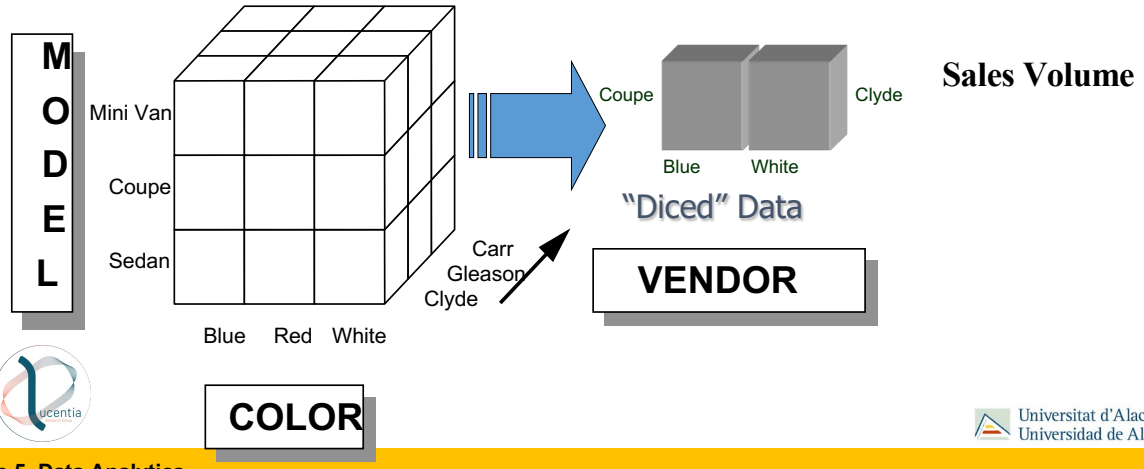
15

OLAP Operations: Roll-up/Drill-down



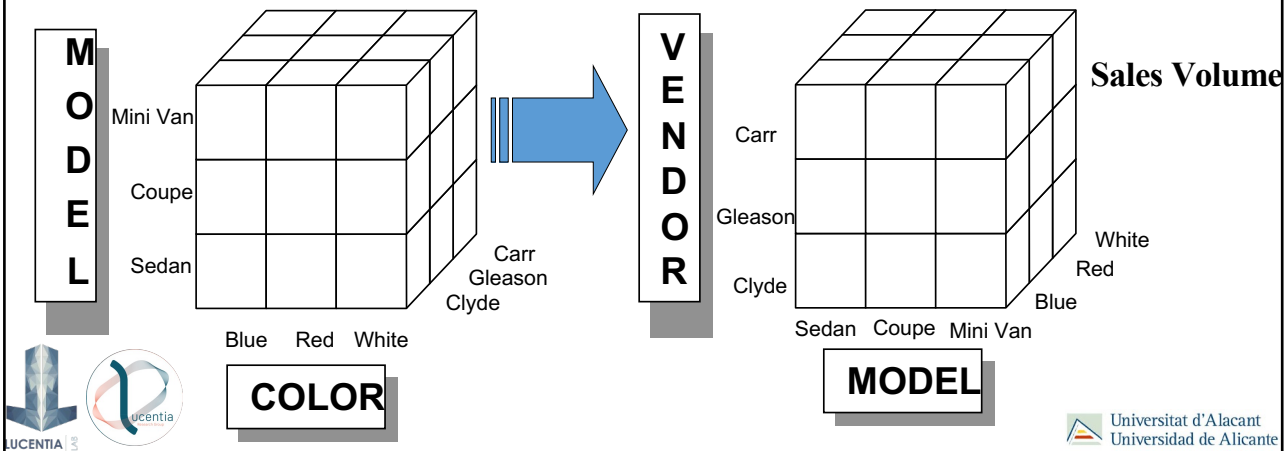
16

OLAP Operations: Slice & Dice



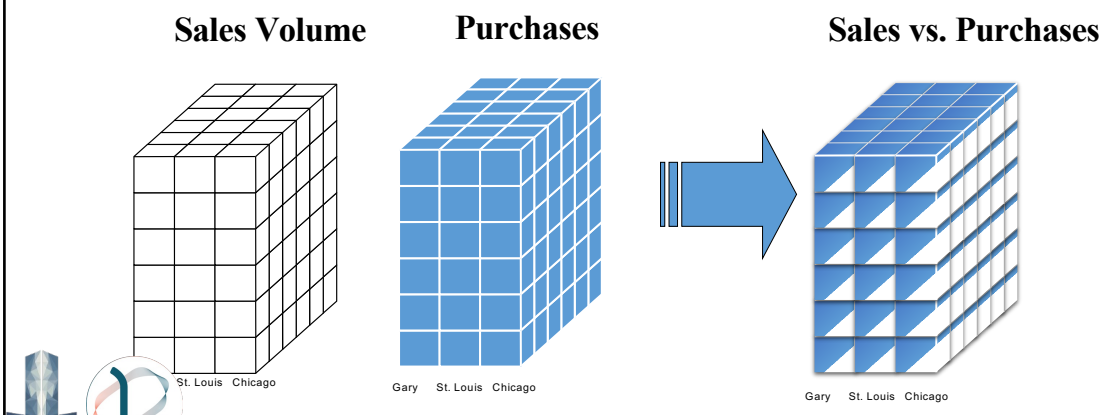
17

OLAP Operations: Pivoting



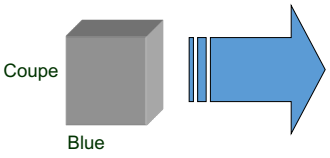
18

OLAP Operations: Drill-Across



19

OLAP Operations: Drill Through



Model	Vendor	Color	Date	Sales
Coupe	Carr	Blue	01/02/2016	3
Coupe	Carr	Blue	07/05/2016	2
Coupe	Gleason	Blue	02/02/2016	5
Coupe	Clyde	Blue	01/06/2016	2
Coupe	Gleason	Blue	17/09/2016	3
Coupe	Carr	Blue	11/10/2016	1

20

Frequently Asked Questions

- Think for a moment and try to answer the following:
  - Must the data be physically/logically structured according to the multidimensional principles?
  - Must we specify at some point a multidimensional structure if we are going to make use of OLAP?
  - Are dimensions always related to the fact through the most detailed level?

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

Frequently Asked Questions

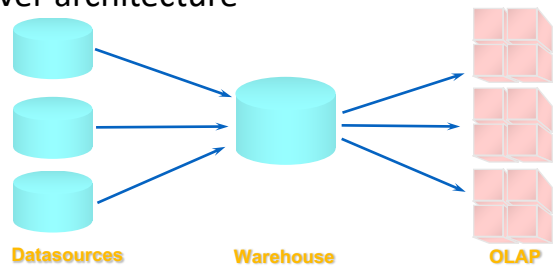
- Should the data be physically/logically structured according to the multidimensional principles?
  - NO, data can be formatted on the fly to accommodate a multidimensional structure
  - It is recommended though
- Should we specify at some point a multidimensional structure if we are going to make use of OLAP?
  - YES, otherwise you will not be able to perform OLAP operations
- Are dimensions always related to the fact through the most detailed level?
  - NO, there can be multiple cubes each relating to its own level in the dimension

OLAP Server Implementation

- An OLAP server acts as a mediator between the data and the user
- The OLAP server reads the data and builds a cube in memory according to the multidimensional schema specified
- Then, it allows the user to apply OLAP operations over the data

OLAP Server Implementation

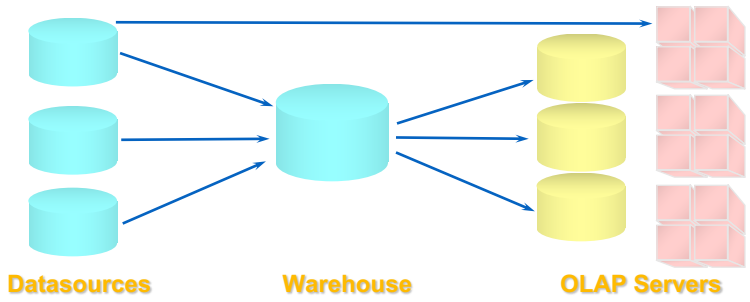
- Traditional OLAP server architecture



- Easy and efficient data access
- Multidimensional view
- Client/Server configuration
- Complex data navigation

OLAP Server Implementation

- Alternative OLAP architecture with direct data feed



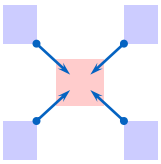
25

OLAP Server Implementation

- Depending on how the server manages and stores the data we can differentiate three groups of OLAP Servers:
  - ROLAP
  - MOLAP
  - HOLAP

26

OLAP Server Implementation

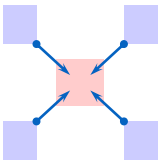


- ROLAP: Relational OLAP Servers
  - Works on top of RDBMS
  - Base data and dimension tables stored as relational tables
- Advantages:
  - Good management of space for sparse data warehouses
  - Data can be accessed by non-OLAP tools
  - Data loading times are usually shorter
  - Does not require pre-computation

27

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

OLAP Server Implementation

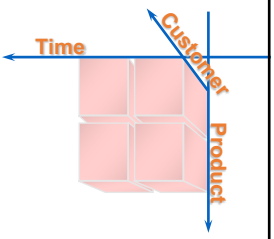


- Disadvantages of ROLAP Servers:
  - Slower performance than MOLAP servers
  - Aggregated tables must be created and managed by the user
  - Depends on the database for querying and caching
- Examples:
  - Oracle
  - Microsoft Analysis Services
  - Microstrategy

LUCENTIA ucentia
Universitat d'Alacant Universidad de Alicante

28

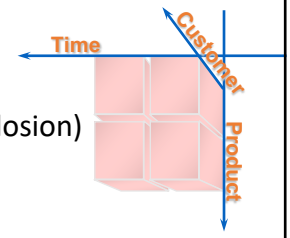
OLAP Server Implementation



- MOLAP: Multidimensional OLAP
  - Precomputes a cube from the Base data
  - Represents in-memory the multidimensional structure of the cube
  - Includes higher level aggregations
- Advantages:
  - Very fast querying speed
  - Data can be compressed in order to reduce space required
  - Easily implemented through array models

29

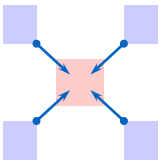
OLAP Server Implementation



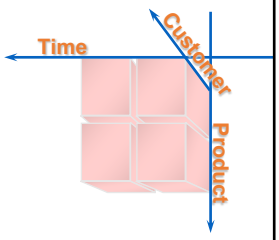
- Disadvantages of MOLAP Servers
  - Poor management of high dimensionality and sparse cubes (data explosion)
  - Long pre-computation times make it unusable for real-time data
  - Information cannot be queried outside the server
  - Cannot drill everywhere on the data
- Examples:
  - Kylin
  - Microsoft Analysis Services

30

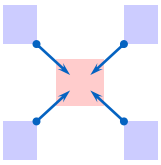
OLAP Server Implementation



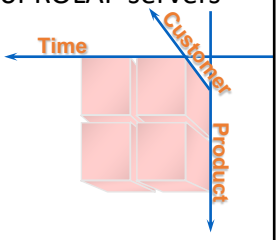
- HOLAP: Hybrid OLAP
  - Base data is stored within a database, like in ROLAP solutions
  - However, the OLAP Engine itself includes in-memory caching and aggregation
  - Base data is only re-accessed when needed
  - Tries to bring the best of both worlds
- Advantages
  - Utilizes both pre-calculated cubes as well as relational data
  - Scales better than ROLAP
  - Computes faster than MOLAP



OLAP Server Implementation



- Disadvantages of HOLAP Servers
  - A HOLAP implementation varies from server to server
  - May require manual configuration and tuning of the aggregated values
  - Cannot achieve the performance of MOLAP servers nor the scalability of ROLAP servers
- Examples:
  - Mondrian (Pentaho)
  - SAP



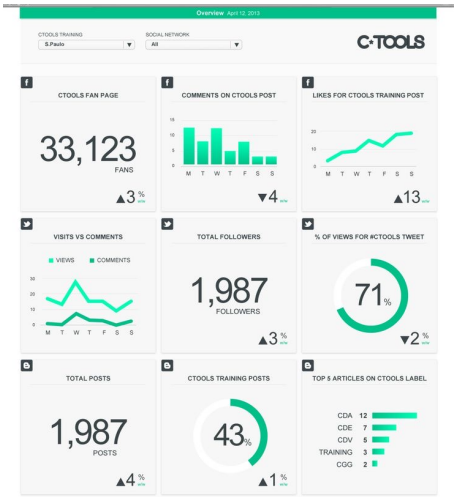
Dashboards

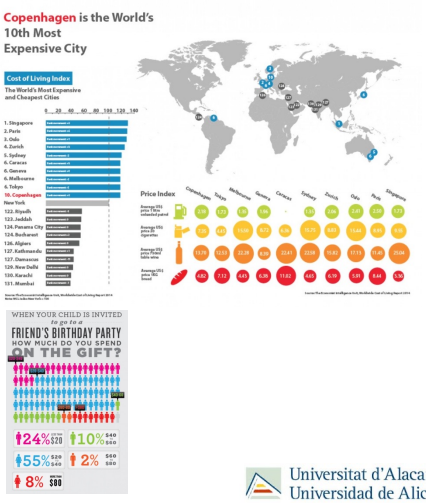
- While OLAP is useful for testing hypotheses, sometimes users wish to **monitor** a certain **process** or the **status of the company**
  - The trend of our sales
  - The rate of manufacture at the car plant
  - The results of the last election poll per state
  - …
- To understand these results, users need more than just data, they need **visualizations** designed to **answer their questions** at a glance

Dashboards

- "A dashboard is a **visual display** of the most **important information** needed to achieve one or more **objectives**; consolidated and arranged on a single screen so the information can be **monitored at a glance.**"

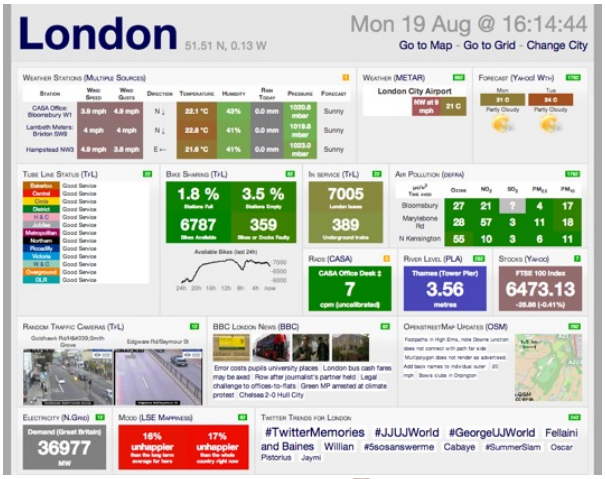
Dashboards





Dashboards

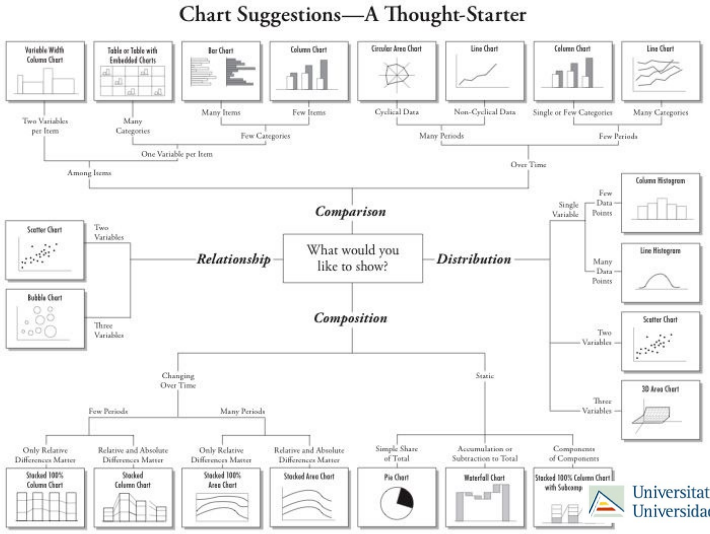
- Types of dashboard:
  - By Area
    - Sales, Marketing, Finances, HR
  - By Role
    - Strategic, Analytic, Operational
  - By Scope
    - Organization, Department, Individual
  - By Interaction
    - Static, Dynamic
  - By refresh rate
    - Continuous, Daily, Weekly, Monthly



Dashboards

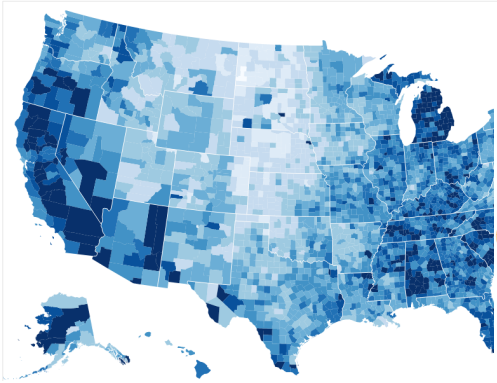
- Choosing the **right visualizations**:
  - Depend on the **user preferences**
  - And the **data at hand**

Source: http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html



37

Dashboards

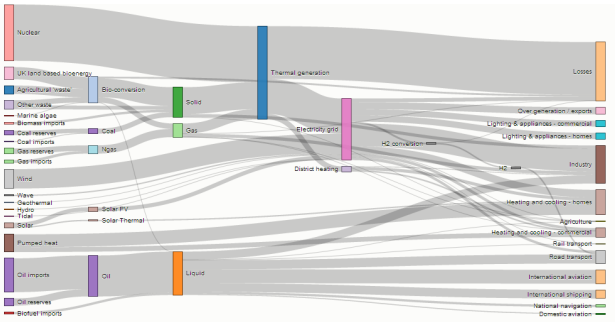


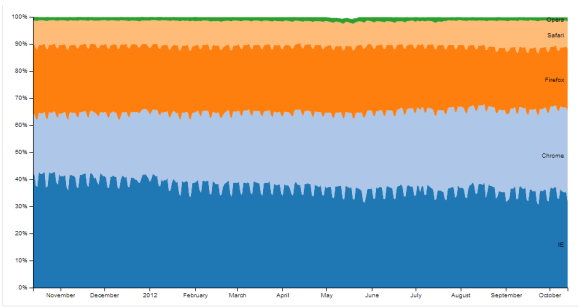
- Choropleth

- Bubble chart

38

Dashboards





- Sankey Diagram
- Stacked Area Chart

Scorecards

- What is a scorecard? A special kind of dashboard designed to **objectively measure the success** of **company objectives**
- Proposed by Kaplan & Norton in 1992 following the motto:

"You cannot control what you cannot measure"

LUCENTIA
ucentia
Universitat d'Alacant
Universidad de Alicante

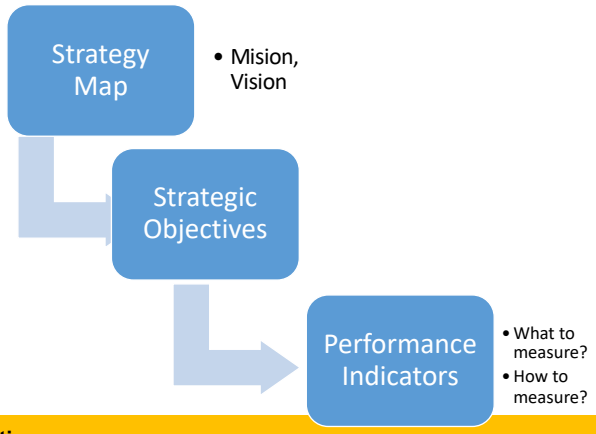
Máster de Formación Permanente en BIG DATA e Inteligencia Artificial

Master en
Big Data e Inteligencia Artificial

khaos RESEARCH

Scorecards

- How to design a Scorecard?



LUCENTIA | ucentia

Universitat d'Alacant
Universidad de Alicante

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial

Master en
Big Data e Inteligencia Artificial

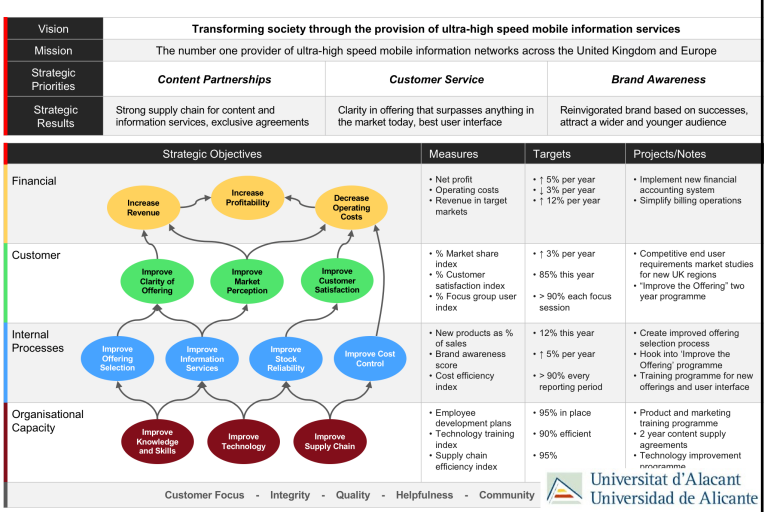
khaos RESEARCH

Scorecards

- Strategy maps:
  - Organized into 4 perspectives: Finance, Customer, Process, Learning

Source: https://www.intrafocus.com/balanced-scorecard/

Vision	Transforming society through the provision of ultra-high speed mobile information services		
Mission	The number one provider of ultra-high speed mobile information networks across the United Kingdom and Europe		
Strategic Priorities	***Content Partnerships***	***Customer Service***	***Brand Awareness***
Strategic Results	Strong supply chain for content and information services, exclusive agreements	Clarity in offering that surpasses anything in the market today, best user interface	Reinvigorated brand based on successes, attract a wider and younger audience



LUCENTIA | ucentia

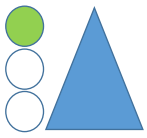
Universitat d'Alacant
Universidad de Alicante

KPIs

- Key Performance Indicators (KPIs) are the **basic building blocks** of scorecards and many dashboards
- A KPI measures a **business objective** and provides a **threshold** that it must meet to be considered **"satisfied"**
  - ***Objectivizes*** the degree of achievement of business objectives

KPIs

- Characteristics:
  - Quantitative and measurable
  - Measures a critical success factor
  - Represents both the business objective as well as its target goal
  - Limited in number
  - Are applied throughout the whole organization



- A KPI **is not** a report, a matter of discussion, nor an unclear metric

KPIs

- While KPIs are **extremely widespread**…
  - Satisfaction of students on courses
  - Number of visitors to our webstore
  - Number of patient deaths in a hospital
- … there can be serious consequences for their **misuse and misinterpretation**
  - A course where no one can fail to avoid dissatisfaction
  - A web where everyone enters but no one buys
  - A hospital that does not treat any high risk patient
- And whoever is responsible for the KPI **will reject changing it**

45

Tips for designing dashboards and scorecards

- How to design **effective** dashboards and scorecards:
  - Establish **clear objectives** to be translated into measures and visualizations
  - **Prototype the dashboards** and show them to the users → They MUST **understand** them
  - Try to imagine **alternative scenarios** where you get the right results but **the situation is completely wrong**
  - Provide **additional information** to identify these scenarios

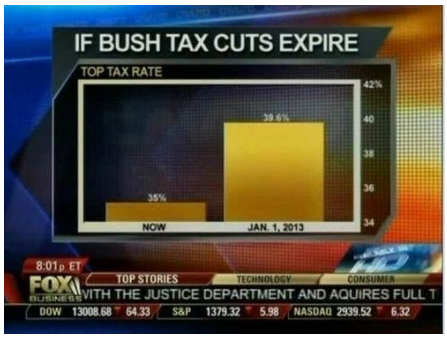
46

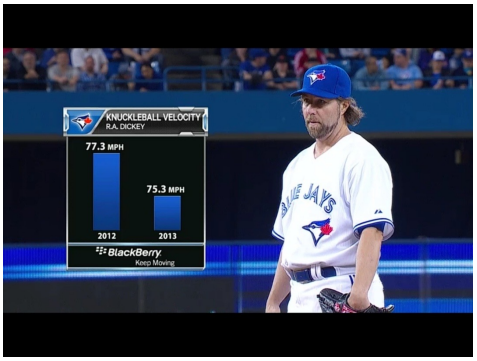
Máster de Formación Permanente en BIG DATA e Inteligencia Artificial

Master en
Big Data e Inteligencia Artificial

khaos RESEARCH

Errors to Avoid





Source:
http://www.huffingtonpost.com/raviparikh/lie-with-data-visualization_b_5169715.html

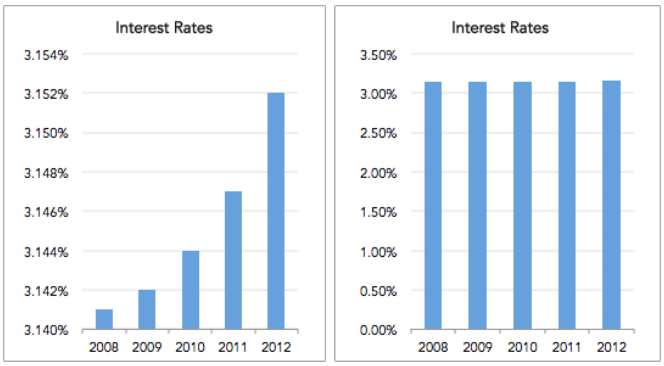
Máster de Formación Permanente en BIG DATA e Inteligencia Artificial

Master en
Big Data e Inteligencia Artificial

khaos RESEARCH

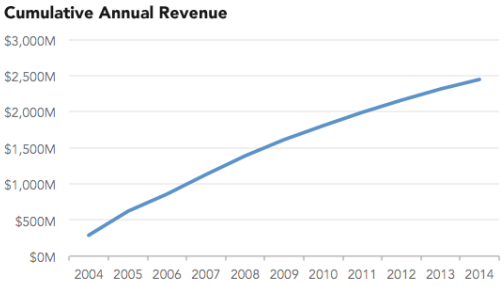
Errors to Avoid

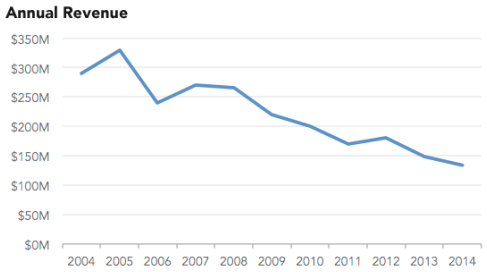
Same Data, Different Y-Axis



Source:
http://www.huffingtonpost.com/raviparikh/lie-with-data-visualization_b_5169715.html

Errors to Avoid





Source: http://www.huffingtonpost.com/raviparikh/lie-with-data-visualization_b_5169715.html

Analytic solutions

- In order to facilitate the analytic results to users we need a **complete** analytic solution
- Ideally, our analytic solution should provide tools for:
  - Data Manipulation (Extraction, Transformation, Load)
  - Data Storage
  - Data Analysis
  - Data Presentation

Analytic solutions

- There are several solutions in the market

Power BI, pentaho, SAP, MicroStrategy, Qlik, JASPERSOFT THE INTELLIGENCE INSIDE, ORACLE DATA INTEGRATOR, +ableau

- We use Pentaho as it includes:
  - One of the most powerful **data manipulation tools** available
  - It is **open source** and free
  - Includes modules for most of the data analytics tasks

Module 5

Data Analytics

Lesson 3: Building a data warehouse solution

Juan Carlos Trujillo – Lucentia (jtrujillo@dlsi.ua.es)
Alejandro Maté – Lucentia (amate@dlsi.ua.es)