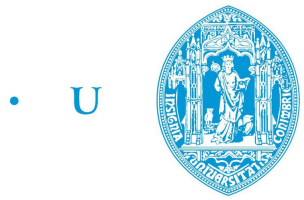


FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Business Intelligence 2017
DEI, FCT, University of Coimbra

*BI Project*

# Data Warehouse and OLAP

## Deliverables and deadlines

1. **Parts 1, 2 and 3**: 8th of March 2017
2. **Parts 4 and 5**: 5th of April 2017
3. Submit a .pdf file into *Inforestudante* with your presentation slides.
4. Deliver a 10 minutes presentation in class

## Part 1

Ralph Kimball identified 3 main success factors for a BI project:

1. The level of commitment and sponsorship of the project from senior management
2. The level of business need for creating a BI implementation
3. The amount and quality of business data available

**To begin this project, you must start by gaining support from the senior management** (in this case, the teacher) for your project. Thus you must convince management of the importance and relevance of your BI project.

**You must identify the team, context and objectives of the project.**

**List of relevant information you should provide:**

1. **Identify the team and the title of the project**
2. **Describe the context for your work**
    a. Present the company/institution (or companies)
    b. Present the "business"

... i. *Products or services, location, clients, etc*

c. Identify the reason for its success
   i. *In your own words, describe why the company is successful*

d. Identify the main challenges the company faces in order to grow
   i. *You don't need to be very thorough! Use your insight and common knowledge to talk about: Competitors, new products, new tendencies, new markets, distribution/transportation issues, cheaper services, royalties, human resources, etc*

**3. Describe your data source**

a. Identify the location of the data that you will use in the project
b. Explain (gross grain) how you will access the data
c. Describe the amount and the quality of data that you will have at your disposal
   i. *Estimate how many years of data can be accessed, how many "records", how many*
... *megabytes (if you can), etc*
   ii. *Estimate how often the data should be refreshed to get the most up-to-date results*
d. Describe potential issues with the quality of data
   i. *Bad introduction of data by the users of the system, lack of compatibility between data in*
... *different locations, missing information, etc*

**4. Explain the objective of your BI solution**

a. Explain how your BI solution can help the company thrive
b. Identify the business processes for which you can provide new and relevant information based on the data available
c. Explain why is this information useful
   i. *How can this information be used to support business decisions*
   ii. *Who, in the organization, would be interested in seeing your results*
d. Identify a set of the most relevant questions you believe you can find answers for in your data
   i. *E.g., "What are our best selling products?", "What is the average age of our best clients?", "What has been our profit in the last 5 years?", etc*
   ii. *These questions are not definitive, they can change, but they are useful to better understand the problem that you are proposing to solve*
   iii. *On the final report, there should be a much larger set of questions and answers*

## Guidelines

To identify the problem that you will tackle in this project you can:

1. Just use a dataset from:
   a. http://www.kdnuggets.com/datasets/index.html
   b. http://archive.ics.uci.edu/ml/
   c. http://www.rdatamining.com/resources/data
   d. http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html
2. Use a dataset which you have access to and license to use
3. Use the list of Web Services available at: http://www.programmableweb.com/apis/directory/1?apicat=Music&protocol=REST&format=XML

# Part 2

In the last phase you have completed the selection of the domain for you project and gained the full support from senior management, now it is time to define the **requirements** for your project. And, since this project is all about data, we will focus on a subset of the full range of requirements usually associated with a software project. We will be leaving out aspects related with technology, non-functional requirements (quality attributes), and with the way the overall system should function. In this stage, we will be focusing on the how data can be combined and exhibited on your BI solution: **User Interface and Reports**. Thus, this is not a full requirements elicitation process, which would take much more time than we have available, it is a process specifically tailored for the needs of this project.

Your objective for this phase is to understand the raw data that you have available to work with, and decide which is the best way to expose the information that you will be extracting from this data to non-technical users. Thus you convince management that you will be delivering a BI solution that fits the needs of the organization. Be sure to include drill-down and roll-up actions, and lots of slice-and-dice capabilities.

## Important Notes

This phase is one of the most important ones that you will have in the project, because it will give you a target, an objective for which you will be working. But, the result of this phase will not be complete because:

1. You don't know the data very well;
2. You don't know everything you can do with the data (e.g., data mining, predictive analytics);
3. You have no experience in the BI field.

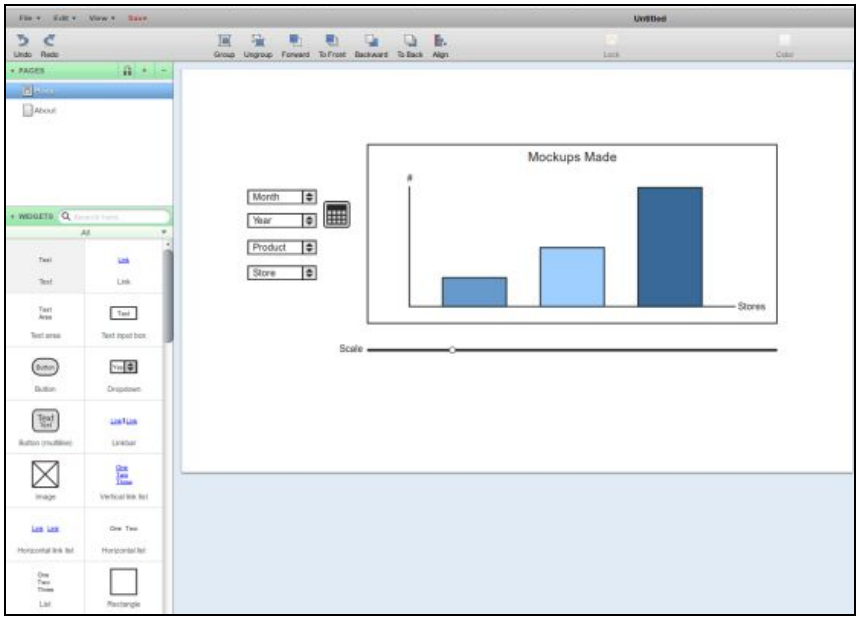
On the other hand, the results from this iteration will represent the greater part of the objectives of this work. Thus, it must be as complete and correct as possible. Keep in mind that the data, which you will be using, is still raw and that it will be your job to prepare, clean, and transform this data into a form that might suite your interests better. To give you a hint on what operations usually are performed over raw data, here is a small list:

1. Deletion: elimination of records and fields which are not necessary or have low quality data;
2. Value modification: e.g., change a numeric value into a category – salary to "pay level", change a date into a season;
3. Value aggregation: count the number of records for the same characteristic (e.g., number of cars sold per month);
4. Value sum: e.g., sum of all sales per day/week/month:
5. Adding new values: add missing information or even new fields to complement existing information;
6. Value correction: removing errors in data (e.g., NULL values to something, misspelled words).

## Objective and guidelines

The objective of this work is to create the mockups for the screens and documents that you will be providing with your BI solution.

If for the documents you can simply use MS Word or something similar, for the screens it is much simpler to use a mockup tool such as https://gomockingbird.com or https://balsamiq.com/.




Use the mockup tool to describe how users will be interacting with the data and how information will be displayed.

# Part 3

The next step of the BI project is the creation of a Data Warehouse. Creating the DW involves knowing the source data, designing the data model, selecting the necessary technology, planning and performing ETL, and deployment. For now, we will focus on data modeling.

By now you already have a good understanding about the operational/raw data available and also about the scope and objectives of the project. **The objective in this phase is to design the Multidimensional Data Model for the Data Warehouse**. The DW will collect all the necessary data to answer the questions which you have previously identified and derive the information present in the mockups. Thus, you must describe the data model for the DW.

## Objective and guidelines

The objective of this phase is the definition of the data model for the DW. Thus, it is essential to:

1. Identify the “stars” in the model;

2. Define the fact tables and the facts;
3. Identify the dimensions and their attributes;
4. Define the granularity of the facts.

# Part 4

The next step of the BI project is the creation of a Data Warehouse. Creating the DW involves knowing the source data, designing the data model, selecting the necessary technology, planning and performing ETL, and deployment. In this phase we will focus on the selection of technology.

In the last phase you created a multidimensional data model for the DW. Now, you will probably use a relational database to implement this data model (but that is your decision). Regardless of the technology that you will use to store the data, you will also need to select the tools for preparing and loading data into the DW, and to perform analytical processing over this data. **The objective is to select the software for a) storing data (e.g., relational database), b) perform the ETL process, and c) do OLAP.** Thus, you must identify the software that you selected for your project, the alternatives that you considered and the reason of your choices.

## Objective and guidelines

The objective is the selection of the tools to use for storing data, ETL and OLAP. Thus, it is essential to:

1. Identify the "must-have" requirements for your tools. For instance:
    a. What features should the ETL tool have? Visual modeling, have process flow definition, etc
    b. What it the size of the DW for the server to handle?
    c. Should the OLAP tool be web-based?
    d. Open-source? Freeware? Can be trial? Most used? Most recent?
    e. Among others.
2. Describe the criteria used for comparing the different solutions based on the elicited requirements;
3. Describe the approach for searching and comparing the multiple software solutions;
4. Present the results and identify the list of software that you will be using.

## Where to start…

Here is an unstructured list of software where for you to start your search. Beware that this is not a complete list (in every aspect) and the order in which names appear in the list does represents any specific sorting preference.

**Oracle (available in house)**
OWB - Oracle Data Warehouse Builder
Oracle Discoverer

**Trial Software**
Pentaho
IBM InfoSphere Warehouse
SQL Power Architect
Visual Importer Enterprise
Warehouse Workbench
Business Intelligence
Datawarehouse
PowerOLAP
IBM Cognos
IntelliView

**Free Software**
InstantOLAP
Talend Open Studio
CloverETL
KETL
DataCleaner
Apatar
Cubulus
icCube
Talend

**Database servers**
Oracle
MySQL
PostGres
SQL Server
LucidDB
...

# Part 5

With the data model and the DW software selected it is time to get the "job" done. In this phase you will focus on planning the ETL process, executing the ETL, and creating the tools for you user OLAP activities.

You must explain your ETL process and your OLAP solutions.

## Objective and guidelines

The objective is to present the complete ETL plan for the first and all the subsequent loads of the DW. Thus, it is essential to:

1. Identify and describe the sources of data;
2. Present the overall ETL plan for your solution;
3. Describe the staging area;
4. For each major action in the plan explain why it is necessary and how it is implemented;
5. Present the major challenges to the implementation;
6. Present the following metrics: size of the source data, size of data on the 1st load, size of data on the subsequent loads, time elapsed on the first load, time used for each update;
7. Identify problems with the source and DW data that were not dealt with (if they exist).
8. Explain how the ETL process is automated in order to allow the future updates of the data in the DW, what is the update strategy for the dimensions and the facts tables and its frequency.
9. Describe how the OLAP data is presented to final users, how they can access it and modify the search parameters;
10. Describe the analyses being performed;
11. Present and discuss the initial findings;
12. Discuss results from the business perspective and explain how the information can be used for Decision Support;

Explain the strategies and techniques used for optimizing the DW and the OLAP queries performance (views, indexes, partitioning, etc).