· U C ·

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Business Intelligence 2023-24
DEI, FCT, University of Coimbra

BI Project

Data Warehouse, OLAP and Data Mining

Deliverables and deadlines

1. **Project idea:** 20th of February 2023
2. **Part 1**: 15th of March 2024
    a. Submit a .pdf file to *Inforestudante* with your presentation slides.
    b. Deliver a 10-minute presentation in next week's class, make a demonstration of your project
3. **Part 2**: 10th of May 2024
    a. Submit a .pdf file to *Inforestudante* with your presentation slides.
    b. Deliver a 10-minute presentation in next week's class, make a demonstration of your project

Part 1

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

1. The level of commitment and sponsorship of the project from senior management
2. The level of the 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 instructors) for your project. Thus, you must convince management of the importance and relevance of your BI project. **To do so, you must submit a document explaining your project idea. The subject you select must have data that allows you to implement both OLAP and Data Mining approaches.**

You must **identify the team and specific objectives** of your 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 customer (the one paying for your work)
    b. Present the "business."
        i. *Products or services, location, clients, etc.*
    c. Identify the Key Performance Indicators (KPI) of the business
        i. *In your own words, describe what influences these KPI*
    d. Identify the main challenges that the company faces to grow
        i. *You don't need to be very thorough! Use your insight and shared knowledge to discuss: 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. *Wrong data input by the system users, 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 business thrive
    b. Identify the business processes for which you can provide new and relevant information based on the data available
    c. Explain why this information is useful
        i. How can this information be used to support business decisions
        ii. Who in the customer organization would be interested in seeing your results
    d. Identify a set of the most relevant questions you believe you can find answers to 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 five years?" etc.
        ii. Clearly state what problems would be handled using Machine Learning approaches, such as prediction, classification, etc.
        iii. These questions are not definitive, they can change, but they are helpful to understand better the problem that you are proposing to solve
        iv. In the final presentation, 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. Use the dataset from the UC project Greenhub:
   a. https://greenhubproject.org/
   b. https://play.google.com/store/apps/details?id=com.hmatalonga.greenhub
2. Just use a dataset from:
   a. http://www.kdnuggets.com/datasets/index.html
   b. http://archive.ics.uci.edu/ml/
   c. https://www.rdatamining.com/datasets
   d. http://kaggle.com
   e. http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html
3. Use a dataset that you have access to and a license to use
4. Use the list of Web Services available at:
   a. https://free-web-services.com/

Part 1

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 leave out technological aspects, non-functional requirements (quality attributes), and how the overall system should function. In this stage, we will focus on how data can be combined and exhibited on your BI solution: **User Interface and Reports**. Thus, this is not a complete requirements elicitation process, which would take much more time than we have available. It is a process specifically tailored to the needs of this project.

Your objective for this phase is to understand the raw data you have available to work with and decide the best way to expose the information you will extract from this data to non-technical users. Thus, you convince management that you will deliver a BI solution that fits the project's needs. **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 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:

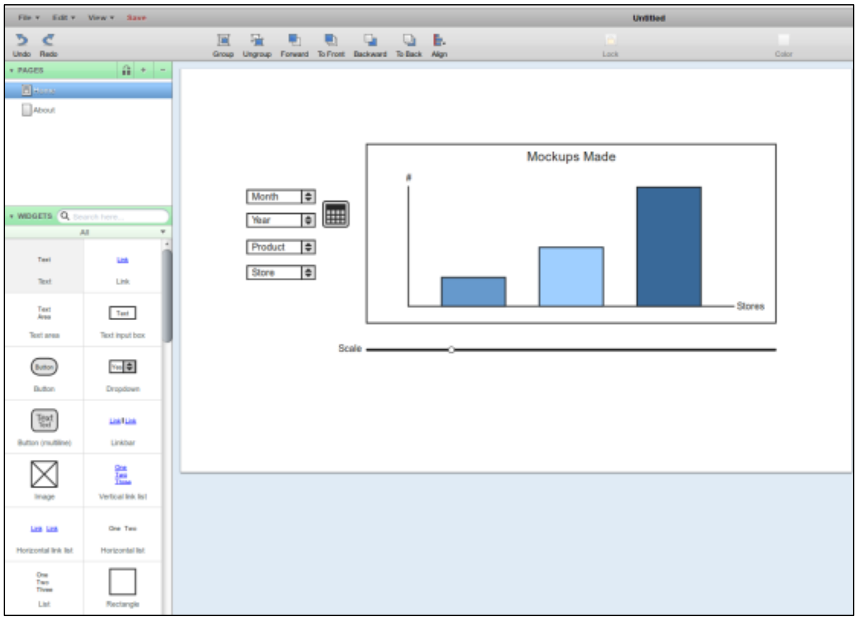
1. You don't know the data very well.
2. You don't know everything you can do with the data (e.g., data mining, predictive analytics).
3. You have no experience in the BI field.

On the other hand, the results from this iteration will represent the more significant part of the objectives of this work. Thus, it must be as complete and correct as possible. Keep in mind that the data 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 suit your interests better. To give you a hint on what operations usually are performed over raw data, here is a small list:

1. Deletion: eliminating records and fields that are unnecessary or have low-quality data.
2. Value modification: e.g., change a numeric value into a category – salary to "pay level," transform 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., the 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).

This work aims to create mockups for the screens and documents you will be provided 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 interact with the data and how the information will be displayed.

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 of the operational/raw data available and the scope and objectives of the project. **This phase aims to design the Multidimensional Data Model for the Data Warehouse**. Thus, you must describe the data model for the DW. The DW will collect all the necessary data to answer the questions you have identified and derive the information in the mockups.

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.

Next, we will focus on the selection of technology.

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

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, process flow definition, etc.
    b. What is the size of the DW for the server to handle?
    c. Should the OLAP tool be web-based?
    d. Freeware? Open source? Can it be a trial? Most used? Most recent?
    e. Among others.
2. Describe the criteria 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 represents any specific sorting preference.

Oracle OWB - Oracle Data Warehouse Builder
Oracle Discoverer
Microsoft Power BI

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
...

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 your user OLAP activities.

You must explain your ETL process and your OLAP solutions. The objective is to present the complete ETL plan for the first and all the subsequent loads of the DW and the OLAP/Reporting tools you developed. 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 significant 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: the size of the source data, the size of data on the 1st load, the size of data on the subsequent loads, the 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 to allow future data updates in the DW and the update strategy for the dimensions, facts tables, and 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 for optimizing the DW and the OLAP queries performance (views, indexes, partitioning, etc.)

Next, it is time to build the front end of the project. Now that you have all the data ready in the DW, using the mockups as guidelines, make the user interfaces for your BI application using the software you selected for this effect. Create the necessary UI to provide your users with the means to answer their questions with the data you prepared. Make sure to include exciting visualizations, such as pivot tables, graphs, etc. And have all the necessary controls for allowing the "slice and dice" of information, the "drill-down" and "roll-up" of the data, etc.

Part 2

Data mining tools complement OLAP data analyses with descriptive and predictive capabilities. In this assignment, you will prepare your dataset and use it to explore two knowledge extraction techniques of your choice. But one of the two choices must be a classification study and the other a different study like clustering, association, time series analysis, etc.

This phase aims to describe the data mining approach you used and show how it is integrated into your product and made available to end users. Thus, it is essential to:

1. Describe the main goals of the Data mining project, and do not forget to detail the two proposed techniques.
    a. For each technique, detail the specific goals
    b. Justify the algorithm choice for each analysis
2. Identify the software being used.
3. Present the data source and identify the attributes being used.
4. Describe each step of the data mining process:
    a. Get data
    b. Clean, prepare and manipulate data, e.g., feature Selection and Reduction
    c. Detail the design of experiences, namely the hyperparameter search/choice
    d. Train and test the models
    e. Test model
    f. Improve

5. Discuss the results of the study. Present and discuss the results obtained in your data mining assignment. Compare your results with the results from other sources. In this problem, one important aspect is to evaluate among the data available the more appropriate for the different scenarios.
6. Describe how your models are made available to end users and integrated into your BI product.