Business Intelligence project

Phase 2 report

Johanna Chapman, Tracy Chapman, Antoine Chédin Clément Chaussade, Léo Fourrier, Valentin Maréchal

Table of contents

[Introduction 2](#_Toc500938648)

[Work Organization 3](#_Toc500938649)

[Review of the model and additional work 3](#_Toc500938650)

[Document description 4](#_Toc500938651)

[Final DF-model 4](#_Toc500938652)

[Architecture schema 4](#_Toc500938653)

[Star-cluster schema 4](#_Toc500938654)

[Talend jobs 4](#_Toc500938655)

[Hive implementation 4](#_Toc500938656)

[Conclusion 4](#_Toc500938657)

# Introduction

This report describes our work for the second phase of the Business Intelligence project.

This project was proposed by a company specialized in predictive analysis, the LIG Computer Sciences Lab of Grenoble, at the french conference EGC 2017.

The objective of this project is to implement a business intelligence system supporting the analysis of tree disease. That involves the integration of several data sources in a data warehouse, the processing of these data and some analyses.

To do this work, three open data sets related to the diseases of the trees around the city of Grenoble have been provided: they describe various attributes of those trees; and while two of them describe their diseases, the last one doesn't. Our purpose is to study the first two data sets in order to predict the diseases of the trees in the last one.

During this phase, we fixed several mistakes from phase 1, created logical schemas related to our data warehouse model, and implemented a prototype of our data warehouse.

# Work Organization

For this phase, Clément was the group leader. Regular meetings were organized in order to share our progress and distribute tasks.

The first objective of this phase was to fixed some mistakes from the first phase of the project. Indeed, we needed a concept more challenging so we decided to add more external data sources to our model. Tracy was in charge to find these data and found pollution information. Clément included them in the Data-Facts model. Antoine was a new member in our team and was responsible for find other data sources and integrate them in the actual D-F model. All the other technical documents were update.

Also, we needed to transform our DF-model into a logical schema. Therefore, Johanna created a star-cluster schema. Valentin designed the schema of our system architecture.

The entire group made Extract-Transform-Load (ETL) jobs on the furnished and the collected data. We used the open source distribution of Talend (Open Talend Studio). Before to do this, Clément made a summary diagram which shows all envisaged jobs and the global workflow.

Léo installed the infrastructure Hive on virtual machine and realised many tests of data loading.

# Review of the model and additional work

(pollution + travail d’Antoine)

# Document description

This section will describe each document produced during this phase. They can be found attached in the provided zip file. The documents from phase 1 have been updated to match our current structure but are not presented here. Refer to the report of the first phase for a description of those documents.

## Final DF-model

## Architecture schema

## Star-cluster schema

# Talend jobs

# Hive implementation

# Conclusion