|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Картинки по запросу UNS logo | Картинки по запросу hospital pasteur |  |  |  |

|  |  |
| --- | --- |
| Big Bridge – SE :  Projet **B**ig **D**ata **S**anté et **E**nvironnement dans la ville de Nice en partenariat avec l’IMREDD  **MDBS France** | |
| Novembre 2016 – Mars 2017  Tuteur MBDS : Gabriel MOPOLO-MOKE  Responsable scientifique : Serge MIRANDA | |
| **Etudiants :**  Mlle Irina GRIGOREVA  M. Sergey GORIANIN  M. Oualid MANAI  M. Walid RHAZADI | **Membres du jury:**  M. Gabriel MOPOLO  M. Serge MIRANDA  M. Pascal VIGNON |

**Résumé**

Ce document est le rapport de notre projet Big Bridge – SE : Big Data Santé et Environnement, réalisé au sein du MBDS en partenariat avec l'IMREDD.

L’objectif de ce projet est de développer un outil logiciel utilisant les approches Big Data permettant de trouver des corrélations entre des données environnementales et les données sur la santé. Afin de vérifier cette corrélation entre environnement et santé, nous avons sont utilisé des outils d’analyse de données issue de plateformes Open Source.

Le but de ce projet est donc de permettre de vérifier si la pollution de l'air influe ou non sur le nombre de patients souffrant de dyspnée à Nice.

Les données utilisées pour la recherche ont été collectées par le service AirPaca et l'Hôpital Pasteur de Nice entre janvier 2014 et décembre 2016.

La réalisation de ce projet a été faite avec les outils suivants : Java / JEE, R, Oracle SQL et NoSQL (Hadoop HDFS, Oracle NoSQL).

Mots clés : Big Data, base de données, Oracle, Apache, Hadoop, Hive, HDFS, NoSQL, DWH, SQL, Web, Java, JSF, JEE, EJB, R, IMREDD.

**Abstract**

This document is the report of our project Big Bridge - SE: Big Data Santé et Environnement, realized within the MBDS in partnership with the IMREDD.

The objective of this project is to develop a software tool using Big Data approaches to find correlations between environmental data and health data. In order to verify this correlation between environment and health are used data analysis tools derived from Open Source platform.

The aim of this project is to make it possible to check whether or not the air pollution influences the number of patients suffering from dyspnea in Nice.

The data used for the research were collected by the AirPaca service and the Hospital Pasteur between January 2014 and December 2016.

The realization of this tool has been made using Java/JEE, R language, Oracle SQL and NoSQL databases (Hadoop HDFS, Oracle NoSQL).

Key words: Big Data, database, Oracle, Apache, Hadoop, HDFS, Hive, NoSQL, DWH, SQL, Web, Java, JSF, JEE, EJB, R, IMREDD

**List of figures and tables**

[Table 1 Comparing the solutions 12](#_Toc478324291)

[Table 2 Risk plan 13](#_Toc478324292)

[Table 3 List of tasks of Sprint 1 14](#_Toc478324293)

[Table 4 List of tasks of Sprint 2 15](#_Toc478324294)

[Table 5 List of tasks of the Sprint 3 15](#_Toc478324295)

[Table 6 List of tasks of the Sprint 4 16](#_Toc478324296)

[Table 7 The Structure of The Health Data 17](#_Toc478324297)

[Table 8 The Structure of The Environmental Data 17](#_Toc478324298)

[Table 9 Result of testing the databases 21](#_Toc478324299)

[Table 10 Result of the searching correlation. Case 1 29](#_Toc478324300)

[Table 11 Result of the searching correlation. Case 2 30](#_Toc478324301)

[Table 12 Result of the searching correlation. Case 3 30](#_Toc478324302)

[Figure 1 Architecture the project (common scheme) 18](#_Toc478116835)

[Figure 2 Options to store the pollution data 19](#_Toc478116836)

[Figure 3 Case of using Oracle NoSQL Database to store the pollution data 20](#_Toc478116837)

[Figure 4 Case of using Mongo DB to store the pollution data 20](#_Toc478116838)

[Figure 5 Case of using HDFS to store the pollution data 21](#_Toc478116839)

[Figure 6 Current architecture of the project 22](#_Toc478116840)

[Figure 7 Residuals vs Fitted chart 31](#_Toc478116841)

[Figure 8 Chart of the extrapolated environmental data 32](#_Toc478116842)

[Figure 9 Chart of the extrapolated health data 33](#_Toc478116843)

[Figure 10 Number of patients – Value of NO Chart 33](#_Toc478116844)

[Figure 11 Number of patients – Value of NO2 Chart 34](#_Toc478116845)

[Figure 12 Number of patients – Value of NOX Chart; 34](#_Toc478116846)

[Figure 13 Dyspnea data 35](#_Toc478116847)

[Figure 14 The linear correlation coefficient of filtered data by diagnosis 35](#_Toc478116848)

[Figure 15 Chart of Use Case 1 36](#_Toc478116849)

Table of Contents

[1 General Introduction 6](#_Toc478116850)

[2 Project Presentation 7](#_Toc478116851)

[2.1 Presentation of the Master MBDS and the Big bridge project 7](#_Toc478116852)

[2.2 Presentation of the project partner IMREDD 8](#_Toc478116853)

[2.3 Presentation of the subject and project goals 8](#_Toc478116854)

[3 State of Art 9](#_Toc478116855)

[3.1 Criteria for comparing 9](#_Toc478116856)

[3.2 Existing projects 9](#_Toc478116857)

[3.3 Comparing projects 11](#_Toc478116858)

[4 Envisaged solution 12](#_Toc478116859)

[4.1 Deliverables 12](#_Toc478116860)

[5 Project Organization 12](#_Toc478116861)

[5.1 Used Project method 12](#_Toc478116862)

[5.2 Project team and member’s role 12](#_Toc478116863)

[5.3 Used tools in the project 13](#_Toc478116864)

[5.4 Configuration management 13](#_Toc478116865)

[5.5 Risk Plan 13](#_Toc478116866)

[5.6 Project planning 14](#_Toc478116867)

[5.7 Project Budget 16](#_Toc478116868)

[6 Environment and health data description 16](#_Toc478116869)

[6.1 Health data Description 16](#_Toc478116870)

[6.2 Environmental Data description 17](#_Toc478116871)

[7 Big Bridge project Architecture and Data Management 18](#_Toc478116872)

[7.1 Objective 18](#_Toc478116873)

[7.2 Architecture of Big Data and Oracle DWH implementation used in the application 18](#_Toc478116874)

[7.3 Managing Health data around Oracle SQL Database 12c in the Oracle Big Data Environment 22](#_Toc478116875)

[7.3.1 Importing Health Data to Oracle SQL Database 12c 22](#_Toc478116876)

[7.4 Managing Pollution data around Oracle NoSQL Database in the Oracle Big Data Environment 22](#_Toc478116877)

[7.4.1 Managing Pollution data around MongoDB 23](#_Toc478116878)

[7.4.2 Managing Pollution data around HDFS 23](#_Toc478116879)

[7.4.3 Creating Oracle NoSQL, MongoDB and HDFS external tables on Hive 23](#_Toc478116880)

[7.4.4 Creating Hive external tables on Oracle SQL Database 24](#_Toc478116881)

[8 Data Analysis with OLAP Query 25](#_Toc478116882)

[9 Data Analysis with the R Language 27](#_Toc478116883)

[9.1 Result of searching the correlation between data with the R Language 28](#_Toc478116884)

[9.1.1 The linear correlation coefficient of unfiltered data 28](#_Toc478116885)

[9.1.2 The linear correlation coefficient of filtered data by diagnosis 29](#_Toc478116886)

[9.1.3 The linear correlation coefficient of interval data 30](#_Toc478116887)

[9.1.4 Linear Regression 30](#_Toc478116888)

[9.2 Prediction Data with R Language 31](#_Toc478116889)

[9.2.1 Prediction of environmental data 31](#_Toc478116890)

[9.2.2 Prediction of health data 32](#_Toc478116891)

[9.3 Different Charts in R 33](#_Toc478116892)

[10 Big Bridge - SE Web Application 34](#_Toc478116893)

[10.1 Displaying the data on Data Tables 35](#_Toc478116894)

[10.2 Displaying OLAP Query result on the Map 35](#_Toc478116895)

[10.3 Displaying R charts 36](#_Toc478116896)

[11 General Conclusion 38](#_Toc478116897)

[Reference and bibliography 38](#_Toc478116898)

[Annexes 41](#_Toc478116899)

1 General Introduction

The most important mechanism for management decisions aimed at improving air quality and reducing the negative impact of environmental factors on the human body is to carry out a health risk assessment of the population. Methodology for assessing the health risk is an element of mathematical modeling of the causal relationships between environmental factors and health, under their influence in the specific conditions of time and area.

At the present time, there is a large number of studies on the impact of air pollution on human health. Studies in various geographical areas have shown associations of respiratory symptoms and conditions with long-term exposure to total suspended particulates (TSP) and SO2 ([1-7](#_Reference_and_bibliography)), to particulate matter ([8-10](#_Reference_and_bibliography)), to black smoke ([11](#_Reference_and_bibliography)), and to NO2 ([7](#_Reference_and_bibliography)). Furthermore, studies of hospital admissions and mortality studies point to an association of short- and long-term exposure to air pollution with symptoms that are related both to pulmonary and to cardiac diseases ([12-19](#_Reference_and_bibliography)).

Thus, it is important to investigate a correlation between the level of air pollution and human health in the city of Nice.

Our mission is to develop a software product using the Big Data approach to improve the efficiency of process and analyze the data, using to make this research.

Plan of report:

[Chapter 1: General Introduction](#_1_General_Introduction) — the reasons of project implementation, a short summary of the sphere, our mission

[Chapter 2: Project Presentation](#_2_Project_Presentation) — Big Bridge project for master MBDS, project partners, project’s objective and goals

[Chapter 3: State of Art](#_4_Study_of) — current state of the projects, studying the correlation of environment and health

[Chapter 4: Envisaged solution](#_4_Envisaged_solution) — deliverables of the project

[Chapter 5: Project Organization](#_5_Project_Organization) — information about team, planning of work and risk plan

[Chapter 6: Environment and health data description](#_6_Environment_and) — description of the content and volume of the source data

[Chapter 7: Project Architecture and Data Management](#_7_BDSE_project) — detailed description of the architecture of Big Data project

[Chapter 8: Data Analysis with OLAP Query](#_8_Data_Analysis) — description of the used data analysis methods with OLAP queries and the results

[Chapter 9: Data Analysis with the R Language](#_9_Data_Analysis) — description of the used data analysis methods with R language and the results

[Chapter 10: Big Bridge – SE Web Application](#_10_Big_Bridge) — description of the web application

[Chapter 11: General Conclusion](#_11_General_Conclusion) — current results of the project, plan of the continuation of research

# 2 Project Presentation

## 2.1 Presentation of the Master MBDS and the Big bridge project

Since more than 26 years Master MBDS form project managers to the information services of the future in symbiosis with the industrial world. Strong industrial links signed in 1990 with all major global technology information (Oracle, Sun, IBM/Informix, Computer Associates, Microsoft, Sybase, CNAF, Amadeus, Unilog, etc.) and then from 1999 telecommunications (Cegetel, Siemens, Lucent, Intel, 3Com, Nokia, etc.) and large users (Amadeus, SFM, Crédit Agricole, GMF, etc.) to prototype future wireless online services. Training of students by practice in the future of information technologies. Construction of a single strategic vision of the information technology market in Europe and around the world with the MBDS inscription in a clear vision of development in future.

The Nice Sophia Antipolis University is a university located in Nice, France and neighboring areas. It was founded in 1965 and is organized in eight faculties, two autonomous institutes and an engineering school.

The project Big Bridge is a generic Big Data project in the MBDS, which will enable students involved experience of practical approaches Big Data market and including bridges between the management of structured and unstructured data as well as data analysis tools using the Open Source platforms. The main databases publishers (IBM, Microsoft, Oracle, etc.), major IT companies (ATOS, CAPGEMINI, Sopra, etc.) and major key accounts (AIR France, Amadeus, HP, etc.) offer Big Data solutions for manage unstructured and structured production data by DBMS. The goal of the MBDS Big Bridge project is propose Big Data projects to demonstrators with publishers, IT services companies and / or large accounts that wish. These projects are spread out on concrete issues such as healthcare, government data, consumer, security, insurance, environment, sports, etc. The different solutions are generally structured around open source ecosystem Hadoop, NoSQL databases (such as MongoDB), SQL databases and various strategies of data analysis (data analytics / data science with the use of the language open source R).

In this document, is presented a project, called Big Bridge, around the Big Data of Oracle solution and the Open Source Hadoop/Map Reduce platforms and R language.

## 2.2 Presentation of the project partner IMREDD

IMREDD is the Mediterranean Institute for Risk, Environment and Sustainable Development. The IMREDD is a new form of cooperation between research, business and the territory in the areas of green technology and intelligent city (Smart City).

IMREDD's mission is to stimulate research, create initial and continuing training courses on environment and sustainable development, and promote expertise and innovation in these fields. It pursues a triple mission:

* conduct and promote the scientific, technological, economic, social and human scientific and technological research and training activities of sustainable development;
* impulse the logics of platforms necessary for mutualized research and development activities;
* promote the valorization of previous activities by helping to identify and support innovative new entrepreneurs and by facilitating collaborative networks between existing actors, such as laboratories, companies, local authorities and associations.

The role of IMREDD in this project is to assist in preparing specifications and obtaining data from the Central Hospital of Nice and the pollution data.

## 2.3 Presentation of the subject and project goals

The objective of this project is to design and develop software tool, using experience of Big Data market approaches including bridges between the management of data and data analysis tools using open source platforms, to search the correlations between environmental and health data.

The aim of the research is to investigate whether or not the air pollution affect the growth of the number of patients with Dyspnea in Nice. The data used for the research were collected by the AirPaca service and the Nice Central Hospital between January 2014 and December 2016.

# 3 State of Art

## 3.1 Criteria for comparing

* Analysis of medical data
* Analysis of environment data
* Open-source solution
* Big Data
* Real time analysis

## 3.2 Existing projects

Today, there are not a lot of solutions in both spheres at the same time: medicine and pollution analysis. But there are projects, which can analyze one of this sphere :

* *Air Paca*

Air PACA is the Association Approved by the Ministry for the Environment for Monitoring Air Quality in the region Provence-Alpes-Côte d'Azur ( AASQA ).

Link: <http://www.airpaca.org/>

* *Open Air*

The Open Air project is a Natural Environment Research Council (NERC) knowledge exchange project that aims to provide a collection of open-source tools for the analysis of air pollution data. These pages provide some background information to the project. The project is also supported by Defra. The project is led by the Environmental Research Group at King's College London, supported by the University of Leeds.

Technologies: R language

Link: <http://www.openair-project.org/>

* *IBM Analytics (Watson)*

Watson Analytics offers you the benefits of advanced analytics without the complexity. A smart data discovery service available on the cloud, it guides data exploration, automates predictive analytics and enables effortless dashboard and infographic creation. You can get answers and new insights to make confident decisions in minutes—all on your own.

Technologies: Cloud based service

Link: <https://www.ibm.com/analytics/watson-analytics/us-en/>

* *OHDSI*

The Observational Health Data Sciences and Informatics (or OHDSI, pronounced "Odyssey") program is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics. All our solutions are open-source. OHDSI has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University.

Technologies:

* Atlas (a web-based integrated platform for database exploration, standardized vocabulary browing, cohort definition, and population-level analysis)
* Achilles (a standardized database profiling tool for database characterization and data quality assessment)
* Calypso (an analytical component for clinical study feasibility assessment)
* KnowledgeWebBase (an experimental user interface for exploration of data present in the LAERTES evidence base)

Link: <http://www.ohdsi.org/>

* *Easy Med Stat*

Medical Statistic Analysis

Technologies: PHP

Link: <http://www.easymedstat.com/>

* *FreeMED*

FreeMED is an open-source, old-as-dirt EMR. Founded in 1999, it’s one of the longest-running open source EMRs out there. It boasts over 81,000 downloads and implementation in everything from small private practices to large government hospitals.

Technologies: Web service

Link: <http://freemedsoftware.org/>

* *REMITT*

REMITT is a revolutionary medical information translation and transmission system, which is primarily used for preparing and submitting medical billing data.

Technologies:

* Written using Java 1.6 / J2EE application standard.
* MySQL-database backed operation.
* Full REST/SOAP interface with WSDL.
* Supports processing X12 835 remittance information and pushing it back to an EMR/PM system via SOAP callbacks.
* Web interface to allow configuration per user, testing of individual plugins, etc.
* JUnit testing using JUnitEE with web interface for full regression and functionality testing.
* File scooper support for pulling remittance and other claim data from clearinghouses.
* Scriptable claim submission using Javascript scripting for clearinghouses.
* Fully database-backed filestore for claims, remittance and processing data with audit/processing trail.

Link: <http://remitt.org/>

Existing solutions do not make a complex analysis of medical and pollution data at the same time. They just analyze one sphere: or medicine, or pollution data. Also, some of existing solutions are commercial and not open source.

## 3.3 Comparing projects

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Criteria  Project | Analysis of medical data | Analysis of environment data | Open-source solution | Big Data | Real time analysis |
| Air Paca |  | **✓** | **✓** | **✓** | **✓** |
| Open Air |  | **✓** | **✓** | **✓** |  |
| OHDSI | **✓** |  |  | **✓** |  |
| Easy Med Stats | **✓** |  |  | **✓** |  |
| FreeMED | **✓** |  | **✓** | **✓** |  |
| REMITT | **✓** |  | **✓** | **✓** |  |
| Big Bridge - SE | **✓** | **✓** | **✓** | **✓** | **✓**  **(in future)** |

Table 1 Comparing the solutions

# 4 Envisaged solution

Big Bridge SE is the software tool, which use Big Data approach and Open-source solutions to make the effective management and analysis of health and environment data.

## 4.1 Deliverables

1. Databases: loading and managing data;

2. Analysis script on R language for prediction, linear correlation and charts;

3. OLAP query script on the project’s data;

4. Web application;

5. Installation guide.

# 5 Project Organization

## 5.1 Used Project method

For managing the project, the Scrum method was used. The description of this methodology is presented in [Annexes](#_Annexes_1).

The team meetings were held every Monday and Tuesday for a duration of 30 minutes. Each member of the team was presenting his progress and his difficulties, and current tasks were identified.

## 5.2 Project team and member’s role

The Product Owner: the mentor of the Master 2 MBDS Mr. Mopolo

The Scrum Master: Irina Grigoreva

The Development Team: Irina Grigoreva, Sergey Gorianin, Oualid MANAI, Walid RHAZADI

## 5.3 Used tools in the project

* *IDE:* NetBeans IDE 8.0.2
* *Web Application Server:* GlassFish 4.1
* *Project management tool:* Microsoft Project 2016
* *Modeling of schemes:* Creately
* *Environment:* Oracle DWH 12g
* *Databases:* Oracle NOSQL, Hadoop HDFS, HBASE
* *Programming languages:* Java/JEE, R
* *Version Control System:* Git (Bitbucket)

## 5.4 Configuration management

Configuration management is done using Dropbox. Project folder directory, named “ProjetBigdataImredd2017” has been created containing the subdirectories for each type of documents. Each member deposits its files in its directory. Source code is available in the Dropbox. Folder description:

BigDataWebApp —NetBeans workspace with the source code of the web application;

StatisticR — RStudio workspace with the source code of the program performs the calculating the correlation and prediction methods;

Report — this report with annexes, report for the partners and the presentation.

The CamelCase notation was used to denote the variable names. The name of the variable is the purpose to which data they refer. Variable from the example refers to the type of style for a block with personalized recommendations.

## 5.5 Risk Plan

|  |  |
| --- | --- |
| Risk | Solution |
| Lack of experience in the field of statistics and mathematical analysis | Consultation with an expert, involving a mathematician to the project |
| The problem with obtaining data from project partners | In the case of a lack of data - to change the theme of the study |
| Loss of time waiting for data and specifications from partners | To use this time for study Big Data approaches and similar projects |
| Server of MBDS cluster crash | Calling a system administrator, working on local virtual machines |
| No external connections to the cluster | Calling a system administrator to make the external connections |

Table 2 Risk plan

## 5.6 Project planning

*Release "Developing the Big Data software to search the correlation between the data"*

*Sprint 1*

Objective of the sprint: define the most effective architecture, configure the cluster.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task | Description | Duration | Start Date | Finish Date |
| 1.1 | Installation and configuration the Oracle Server (Cluster Big Data du MBDS) | 3 days | 21.11.16 | 23.11.16 |
| 1.2 | Determine several software architectures / tools available from the Big Data tools for storing and analyzing this data | 2 days | 21.11.16 | 22.11.16 |
| 1.3 | Test and confirm the most effective architecture | 5 days | 22.11.16 | 28.11.16 |

Table 3 List of tasks of Sprint 1

*Sprint 2 (The duration of this sprint was increased because of the expectation of data from the partner)*

Objective of the sprint: preparing data with project partners, analysis of the existing Big Data projects.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task | Description | Duration | Start Date | Finish Date |
| 2.1 | Preparing state of the art on projects and / or tools big data for cross-cutting environmental and health data | 7 days | 29.11.16 | 07.12.16 |
| 2.2 | Analysis of the existing Big Data projects, learning the approach | 39 days | 08.12.16 | 31.01.17 |
| 2.3 | Preparing data with the IMREDD | 39 days | 08.12.16 | 31.01.17 |
| 2.4 | Preparing the specifications with IMREDD and CHU | 39 days | 08.12.16 | 31.01.17 |

Table 4 List of tasks of Sprint 2

*Sprint 3*

Objective of the sprint: to create and to fill the databases with Health and Environmental data, using Big Data approach.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task | Description | Duration | Start Date | Finish Date |
| 3.1 | Import csv files with environmental data into Hadoop HDFS and into Oracle NOSQL Database | 6 days | 01.02.17 | 08.02.17 |
| 3.2 | Creating and filling tables in Apache Hive. Creating external tables from Apache HDFS and Oracle NoSQL Database | 3 days | 09.02.17 | 13.02.17 |
| 3.3 | Compare the speed of work with both solutions for environmental data, choose the optimal one | 1 day | 14.02.17 | 14.02.17 |
| 3.4 | Creating and filling tables of Health Data in Oracle SQL Database | 1 day | 15.02.17 | 15.02.17 |

Table 5 List of tasks of the Sprint 3

*Sprint 4*

Objective of the release: to investigate the correlation between environmental and health data, using R language, and to display the results in the web application.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task | Description | Duration | Start Date | Finish Date |
| 4.1 | Configuration of the R Studio for access the databases | 1 day | 16.02.17 | 16.02.17 |
| 4.2 | Preparation the data for analysis in R, making OLAP queries | 2 days | 17.02.17 | 20.02.17 |
| 4.3 | Investigation of the data correlation using tools of the R language | 5 days | 21.02.17 | 27.02.17 |
| 4.4 | Creation the NetBeans Project, configuration the connection to database | 3 days | 16.02.17 | 20.02.17 |
| 4.5 | Making the prediction methods for data in R | 1 day | 28.02.17 | 28.02.17 |
| 4.6 | Adjusting the display of the analysis results in the application | 8 days | 01.03.17 | 10.03.17 |

Table 6 List of tasks of the Sprint 4

The full project plan is presented in [Annexes](#_Annexes_1).

## 5.7 Project Budget

For this moment, this project has no budget.

# 6 Environment and health data description

## 6.1 Health data Description

For the health part of the research was used the data about patients with Dyspnea of the Central Hospital of Nice, in the period from January 2014 to December 2016. The data structure is presented in Table 6.1. Data is presented in csv format.

Volume: 7854 rows, 25 columns.

|  |  |
| --- | --- |
| Field | Description |
| Gender | *Gender* |
| Age | *Age* |
| Address | *Address* |
| Postal code | *Postal code* |
| Ville | *Ville* |
| Admission | *The date of the admission* |
| Sortie | *The date of the exit* |
| Examen | *The date of the examination* |
| Categorie de Recours | *The category of remedy – group of medicines that patient got* |
| Libelle de Recours | *Diagnosis and value of DEP (peak expiratory flow) and suspicion and/or theoretical diagnosis* |
| Code de Recours | *Code of ICD-10 (International Statistical Classification of Diseases and Related Health Problems) by World health organization* |
| Libelle gravite | *Level of seriousness* |
| Libelle CCMU | *Wording of Classification Clinique des maladies aux urgences)* |
| Destination Confirmee | *Сonfirmed patient organization* |
| Type de sortie | *Format of the treatment* |
| Diag1 – diag10 | *Final diagnosis - code of ICD-10 (International Statistical Classification of Diseases and Related Health Problems) by World health organization* |

Table 7 The Structure of The Health Data

## 6.2 Environmental Data description

For the environmental part of this research was used the data about the air pollution in Nice from the service AirPaca ([20](#_Reference_and_bibliography)), in the period from January 2014 to December 2016. The data structure is presented in Table 6.2. Data is presented in csv format.

Volume: 23016 rows, 6 columns.

|  |  |
| --- | --- |
| Field | Description |
| Station | *Station* |
| Pollutant | *Pollutant* |
| Description | *Full name of the pollutant* |
| Unité | *µg/m3* |
| Date | *Date* |
| Value | Value |

Table 8 The Structure of The Environmental Data

# 7 Big Bridge project Architecture and Data Management

## 7.1 Objective

The objective of Big Bridge Project Architecture and Data Management is:

* understanding the Oracle Linux Server architecture;
* understanding the Oracle Big Data SQL solution;
* understanding different types of data storing;
* understanding the “Big Bridge mechanism”;
* testing different types of data storing and choosing the best one for our solution;
* acquisition of data (Oracle NoSQL, Apache Hive, HDFS) for our reference Big Bridge project;
* access to data from the database DWH through access drivers (Hadoop Big Data SQL access driver, Hive Big Data SQL access driver);
* analysis with tools beyond data analytics tools (such as open language source R).

## 7.2 Architecture of Big Data and Oracle DWH implementation used in the application

The scheme below shows the architecture of data streams and data storing in our project with our vision of next developing Big Bridge project:

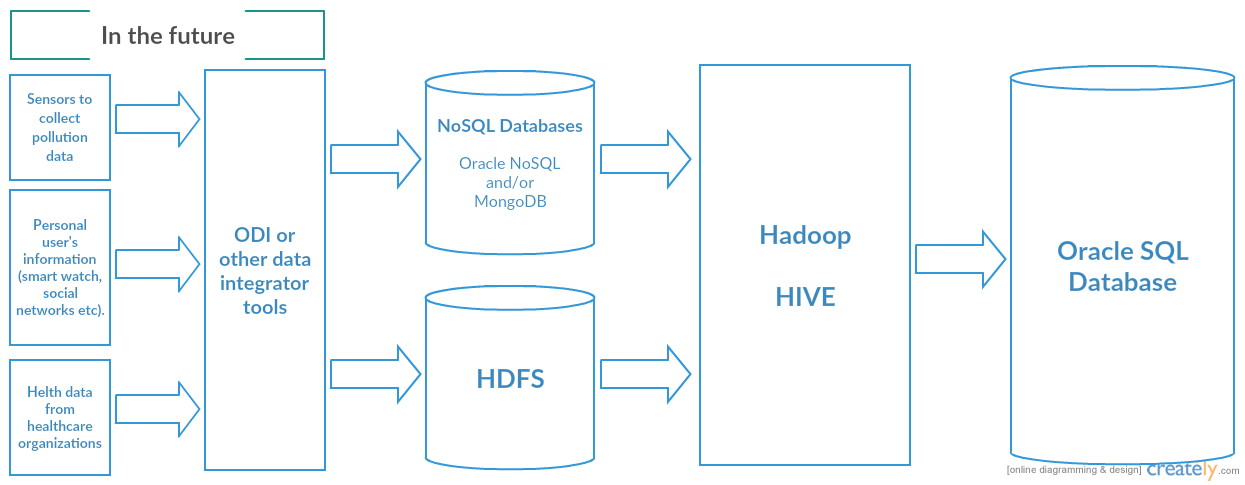


Figure 1 Architecture the project (common scheme)

As we can see, to continue the project the additional data sources may be added, like sensors in Nice to monitor pollution; user’s personal information, like age, gender, living place, heart rate, using smart watches and social network’s profile information; health data from healthcare organizations, like hospitals. The next step is to integrate the data into Oracle Big Data Approach. It’s possible using ODI Tool. As database, it can be used MongoDB, Oracle NoSQL Database, HDFS (for non-structured and more dynamical updated data) or Oracle SQL Database (for structured and more static data). We use Hadoop Hive as a bridge or a “universal” gates between NoSQL Databases and SQL Databases. It’s a relevant solution for this project.

At the present time, we have two data sources: pollution data from Air PACA service and health data from Pasteur Hospital of Nice. All data are presented in .csv files. To store the health data, we use Oracle SQL Server, because, in the future, health data from healthcare organizations should be more static, than pollution or user’s data. For pollution data, we could use one of such databases: Oracle NoSQL Database, MongoDB or HDFS.

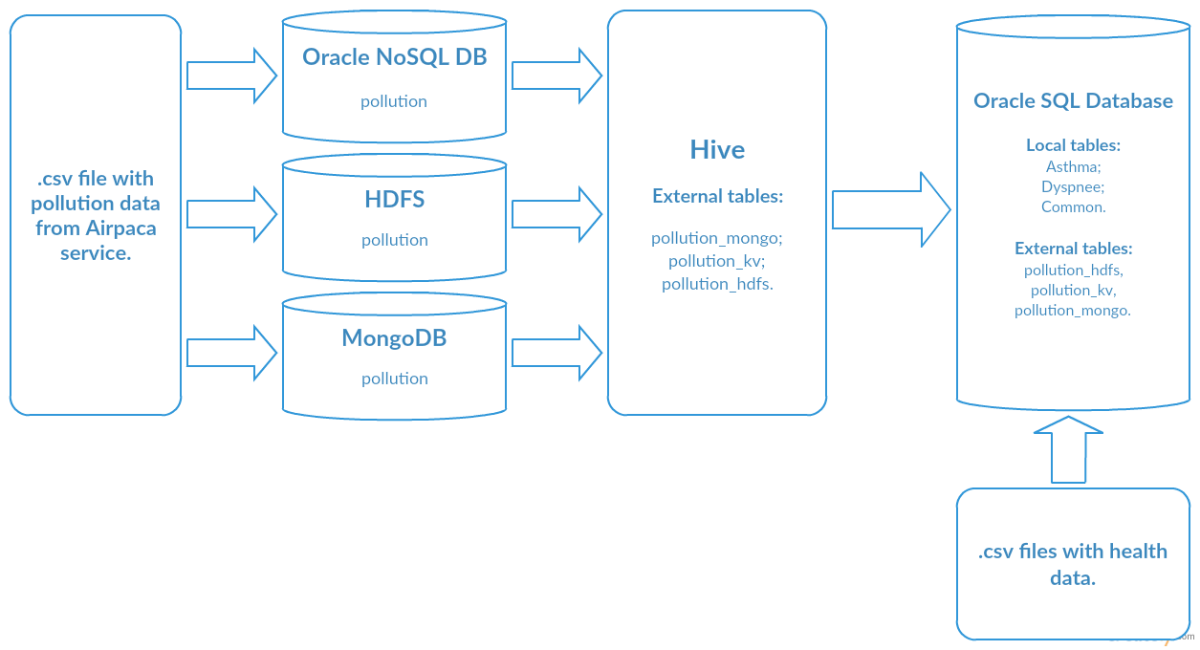


Figure 2 Options to store the pollution data

As we can see on this diagram, we’ve imported .csv pollution data file to Oracle NoSQL DB, HDFS and MongoDB, then we’ve made external tables in Hadoop Hive. Finally, we’ve created external tables with pollution data in Oracle SQL Server, making the “bridge” between NoSQL Databases and Oracle SQL Database, using Hadoop Hive. More detailed information can be seen on these schemes, also information about importing .csv data files into NoSQL Databases, HDFS, Oracle SQL Database and creating external tables is presented in [Installation Guide](#_Annexes_1).

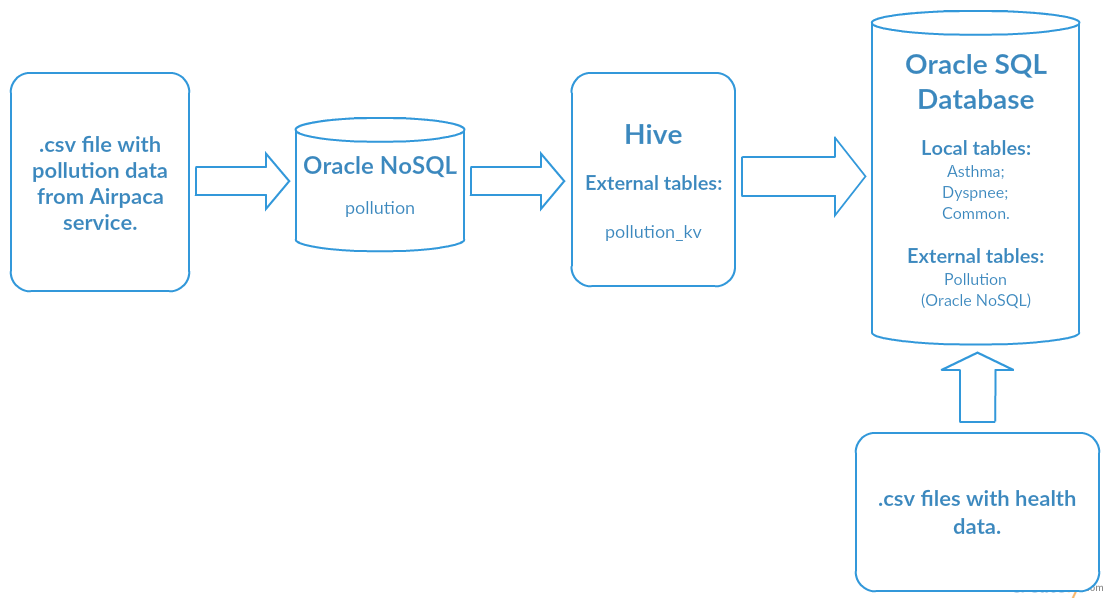


Figure 3 Case of using Oracle NoSQL Database to store the pollution data

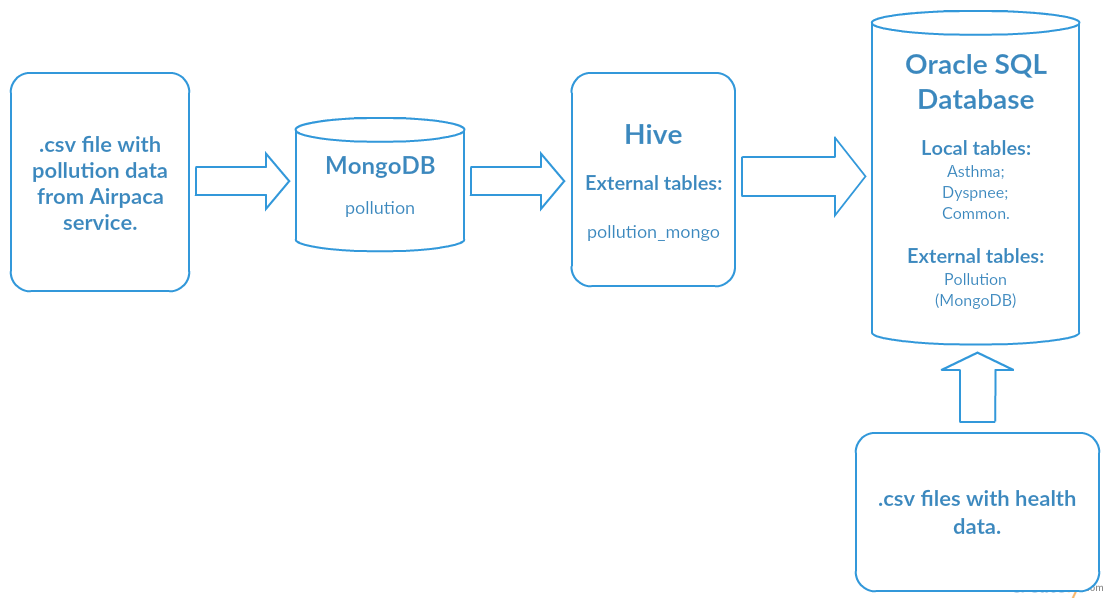


Figure 4 Case of using Mongo DB to store the pollution data

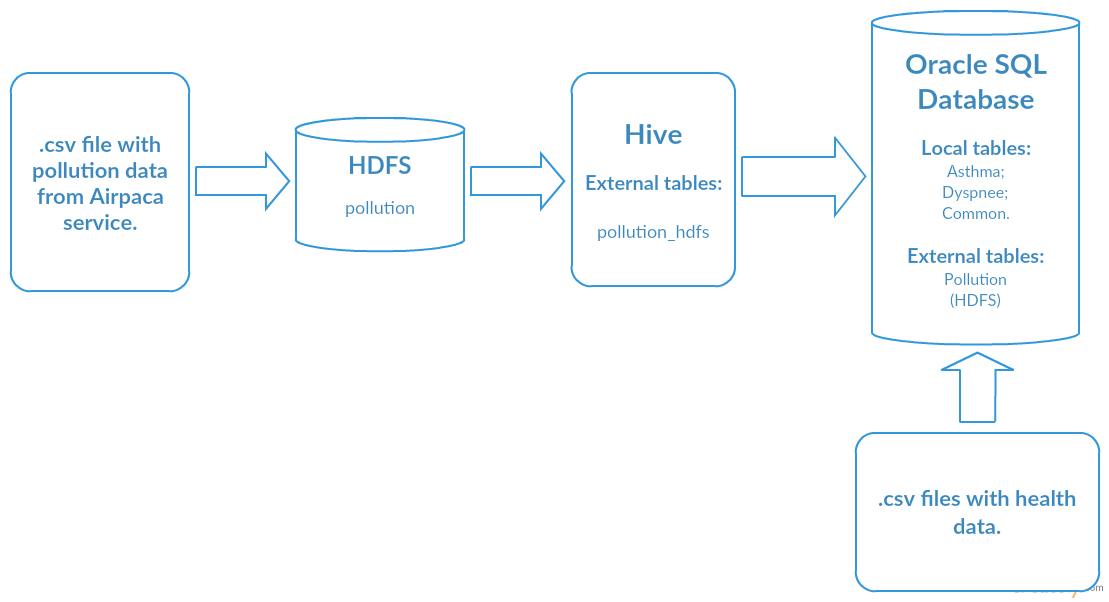
****

Figure 5 Case of using HDFS to store the pollution data

The next step is testing, which solution is better; which database queries to pollution data will execute faster. The result of this test is in the table below, also the screenshots of this testing are presented in [Installation Guide](#_Annexes_1).

|  |  |  |  |
| --- | --- | --- | --- |
| Database  Executing  time | Oracle NoSQL DB | HDFS | MongoDB |
| Time (secs.) | **0.83** | **0.257** | **0.57** |

Table 9 Result of testing the databases

As we can see, HDFS shows the best result of speed of the executing queries. So, in our project we’ve decided to use HDFS to store pollution data and the final architecture of our project is on the scheme below:

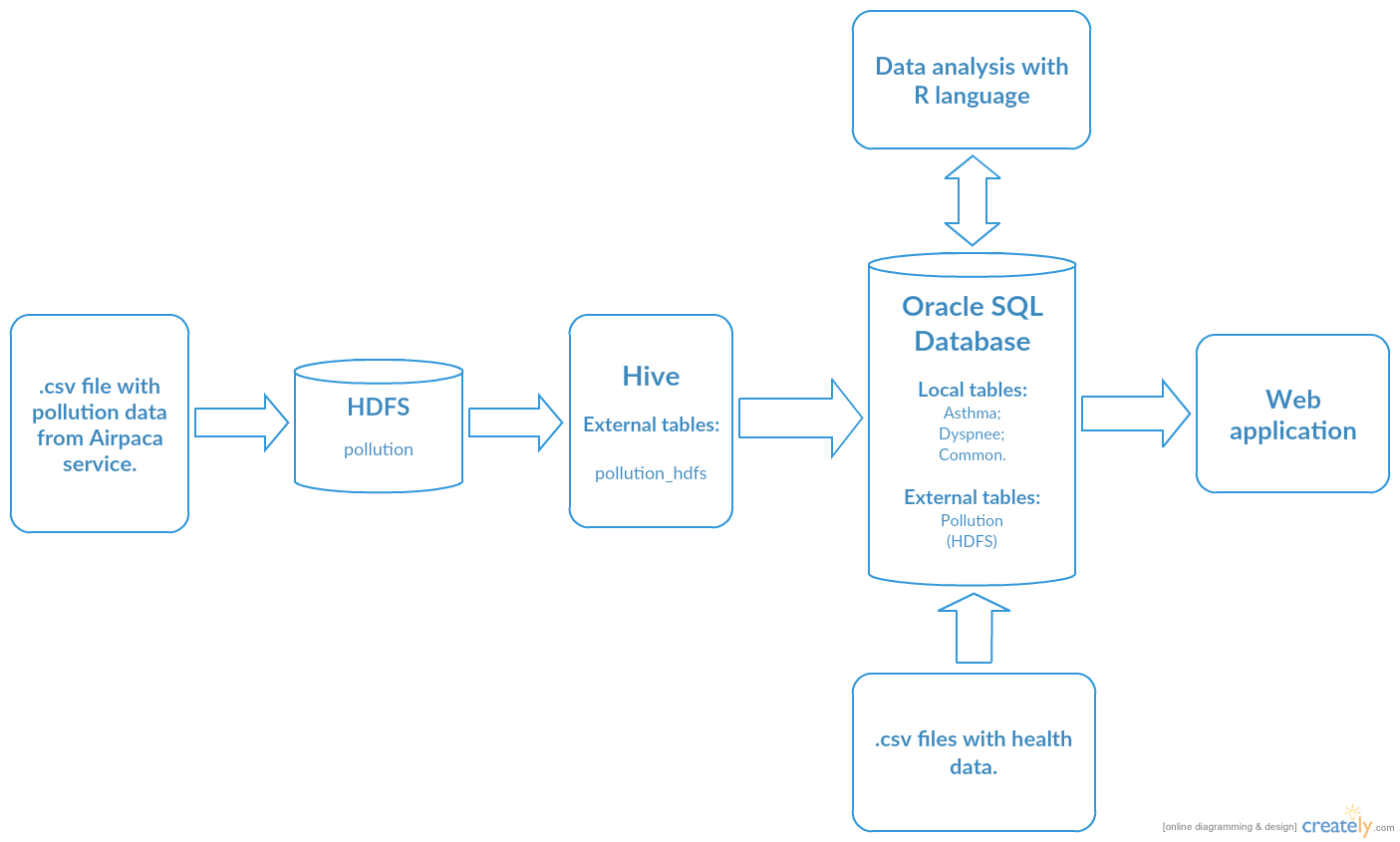


Figure 6 Current architecture of the project

## 7.3 Managing Health data around Oracle SQL Database 12c in the Oracle Big Data Environment

### 7.3.1 Importing Health Data to Oracle SQL Database 12c

To import .csv data files with health information use SQL Developer, right click on “Tables” field and select “Import data…” to open “Data Import Wizard”. Next, follow steps in “Data Wizard” to import data as local tables. More detailed instruction with SQL Developer configurations and importing data are presented in [Installation Guide](#_Annexes_1).

## 7.4 Managing Pollution data around Oracle NoSQL Database in the Oracle Big Data Environment

To connect to Oracle NoSQL Database, use the command below:

**java -jar $KVHOME/lib/kvstore.jar runadmin -port 5000 -host bigdatalite.localdomain**

Then, connect to store:

**connect store -name kvstore**

To create table “pollution”, print:

**execute 'create table pollution (station string, pollutant string, description string, units string, date\_pol string, value\_pol integer, primary key(shard(date\_pol), station, pollutant, value\_pol))'**

And then, to import data from data file, use the command:

**put table -name pollution -file /home/oracle/bigdataProject/json/pollution.json**

### 7.4.1 Managing Pollution data around MongoDB

Information about installation and configuration Oracle Server to work with MongoDB can be found in [Installation Guide](#_Annexes_1). Here is described importing data mechanism from .csv data files to MongoDB.

To import .csv data file to MongoDB use the command below:

**mongoimport -d bigdataProjectDB -c pollution --type csv --file /home/oracle/bigdataProject/csv/pollution/pollution.csv --headerline**

### 7.4.2 Managing Pollution data around HDFS

To import .csv data files to HDFS, first, create the directory, using the command:

**hdfs dfs -mkdir /pollution**

And then put data file to this directory:

**hdfs dfs -put /home/oracle/bigdataProject/csv/hdfs/pollution.csv /pollution**

### 7.4.3 Creating Oracle NoSQL, MongoDB and HDFS external tables on Hive

How to connect to Hadoop Hive is described in [Installation Guide](#_Annexes_1).

To create Oracle NoSQL external table on Hive, print:

**create external table pollution\_kv (station string, pollutant string, description string, units string, date\_pol string, value\_pol int)**

**stored by 'oracle.kv.hadoop.hive.table.TableStorageHandler'**

**tblproperties**

**('oracle.kv.kvstore'='kvstore',**

**'oracle.kv.host'='bigdatalite.localdomain:5000',**

**'oracle.kv.hadoop.host'='bigdatalite.localdomain/127.0.0.1',**

**'oracle.kv.tableName'='pollution');**

To create MongoDB external table on Hive, use the command below:

**CREATE EXTERNAL TABLE pollution\_mongo(**

**station string, pollutant string, description string, units string, date\_pol string, value\_pol int)**

**ROW FORMAT SERDE**

**'com.mongodb.hadoop.hive.BSONSerDe'**

**STORED BY**

**'com.mongodb.hadoop.hive.MongoStorageHandler'**

**WITH SERDEPROPERTIES (**

**'mongo.columns.mapping'='{"station":"station",**

**"pollutant":"pollutant", "description":"description", "units":"units",**

**"date\_pol":"date\_pol", "value\_pol":"value\_pol"}')**

**TBLPROPERTIES (**

**'mongo.uri'='mongodb://localhost:27017/bigdataProjectDB.pollution');**

To create HDFS external table, use the command:

**create external table pollution\_hdfs (station string, pollutant string,**

**description string, units string, date\_pol string, value\_pol int)**

**row format delimited fields terminated by ','**

**stored as textfile location 'hdfs:/pollution/';**

### 7.4.4 Creating Hive external tables on Oracle SQL Database

To create Hive external tables on Oracle SQL Database, use the commands below:

**create table pollution\_kv (station varchar2(55), pollutant varchar2(10),**

**description varchar2(55), units varchar2(12), date\_pol varchar2(32),**

**value\_pol numeric(8))**

**organization external (type oracle\_hive**

**default directory oracle\_bigdata\_config**

**access parameters (**

**com.oracle.bigdata.tablename = bigdataprojectdb.pollution\_kv)) reject limit unlimited;**

And :

**create table pollution\_hdfs (station varchar2(55), pollutant varchar2(10),**

**description varchar2(55), units varchar2(12), date\_pol varchar2(32),**

**value\_pol numeric(8))**

**organization external (**

**type oracle\_hive**

**default directory oracle\_bigdata\_config**

**access parameters (**

**com.oracle.bigdata.tablename = bigdataprojectdb.pollution\_hdfs)) reject limit unlimited;**

Then two external tables has been created in Oracle SQL Database: pollution\_kv (Oracle NoSQL DB) and pollution\_hdfs (HDFS).

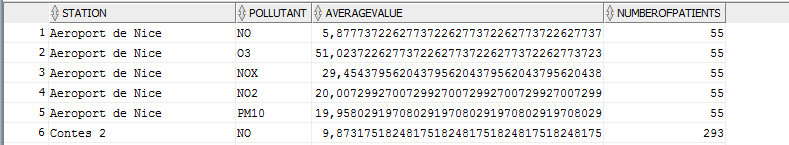
The next step is testing these tables – with what table queries will execute faster. The details of this test are described in [Installation Guide](#_Annexes_1).

Simple select-query executes for HDFS in 0.257 seconds. For Oracle NoSQL the result is 0.83 s. Thus, it will be rational to use HDFS in this project.

# 8 Data Analysis with OLAP Query

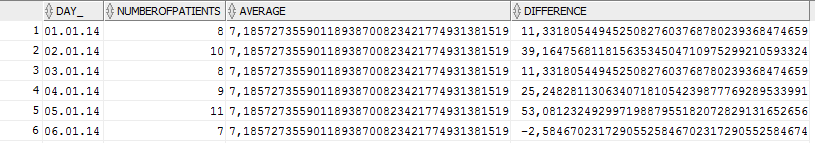
To display the average value of each pollutant on the map, as well as the total number of patients for each station, the following query was used:

|  |
| --- |
| SELECT Pollution.station, Pollution.pollutant, AVG(Pollution.value\_pol) AS averageValue,  (SELECT COUNT(\*) FROM dyspnee WHERE station=Pollution.station GROUP BY station) AS numberOfPatients  FROM Pollution  GROUP BY Pollution.station, Pollution.pollutant  ORDER BY Pollution.station; |



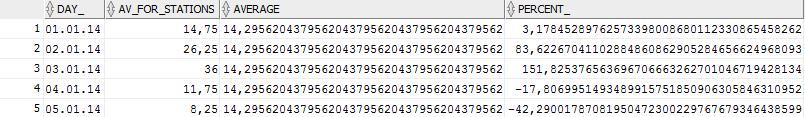
The queries for preparing health data for analysis in R:

|  |
| --- |
| SELECT TO\_DATE(SUBSTR(admission, 0, 10),'MM/DD/YYYY') AS day\_,  COUNT(\*) AS numberOfPatients, ((SELECT AVG(numberOfPatients) FROM  (SELECT TO\_DATE(SUBSTR(admission, 0, 10),'MM/DD/YYYY') AS day\_, COUNT (\*) as numberOfPatients FROM dyspnee GROUP BY TO\_DATE(SUBSTR(admission, 0, 10),'MM/DD/YYYY')))) AS Average,  (((COUNT(\*) - ((SELECT AVG(numberOfPatients) FROM (SELECT TO\_DATE(SUBSTR(admission, 0, 10),'MM/DD/YYYY') as day\_, COUNT(\*) as numberOfPatients FROM dyspnee GROUP BY TO\_DATE(SUBSTR(admission, 0, 10),'MM/DD/YYYY')))))/((SELECT AVG(numberOfPatients) FROM (SELECT TO\_DATE(SUBSTR(admission, 0, 10),'MM/DD/YYYY') as day\_, COUNT(\*) as numberOfPatients FROM dyspnee GROUP BY TO\_DATE(SUBSTR(admission, 0, 10),'MM/DD/YYYY')))))\*100) AS Difference  FROM dyspnee  GROUP BY TO\_DATE(SUBSTR(admission, 0, 10),'MM/DD/YYYY')  ORDER BY TO\_DATE(SUBSTR(admission, 0, 10),'MM/DD/YYYY'); |



The queries for preparing environmental data for analysis in R:

|  |
| --- |
| SELECT TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss') AS day\_, AVG(value\_pol) AS AV\_FOR\_STATIONS, (SELECT AVG(value\_) FROM (SELECT AVG(value\_pol) AS value\_, TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss')  FROM Pollution WHERE POLLUTANT=**'Name Of Pollutant'**  GROUP BY TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss')  ORDER BY TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss'))) AS Average,  (((AVG(value\_pol) - (SELECT AVG(value\_) FROM  (SELECT AVG(value\_pol) AS value\_, TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss')  FROM Pollution  WHERE POLLUTANT =**'Name Of Pollutant'**  GROUP BY TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss')  ORDER BY TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss'))))/(SELECT AVG(value\_)  FROM (SELECT AVG(value\_pol) AS value\_, TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss') FROM Pollution WHERE POLLUTANT=**'Name Of Pollutant'**  GROUP BY TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss') ORDER BY TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss'))))\*100) AS Percent\_  FROM Pollution  WHERE POLLUTANT=**'Name Of Pollutant'**  GROUP BY TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss')  ORDER BY TO\_DATE(SUBSTR(date\_pol, 0, 10),'MM/DD/YYYY hh24:mi:ss'); |



# 9 Data Analysis with the R Language

The statistical analysis of the data from 7854 patients and 6 pollutants (day-by-day) was processed with the R language. R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis ([21](#_Reference_and_bibliography)).

R Language allows to group and filter the data, automatically calculate a correlation coefficients, as well as to build the necessary charts for visualization.

For the calculate measure of dependence between quantities was used the Pearson product-moment correlation coefficient, or "Pearson's correlation coefficient", commonly called simply "the correlation coefficient". It is obtained by dividing the covariance of the two variables by the product of their standard deviations.

The population correlation coefficient ρX,Y between two random variables X and Y with expected values μX and μY and standard deviations σX and σY is defined as:

where E is the expected value operator, cov means covariance, and cor is a widely used alternative notation for the correlation coefficient.

The Pearson correlation is +1 in the case of a perfect direct (increasing) linear relationship (correlation), −1 in the case of a perfect decreasing (inverse) linear relationship (anticorrelation), and some value in the open interval (−1, 1) in all other cases, indicating the degree of linear dependence between the variables. As it approaches zero there is less of a relationship (closer to uncorrelated). The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables (22).

To calculate the linear correlation coefficient in R, it can be used the function *cor(firstDataSet, secondDataSet)*, for example:

|  |
| --- |
| > cor(environmentDataNO$PERCENT\_, patientsData$DIFFERENCE, method="pearson")  [1] 0.1804087 |

To find the linear correlation were used the calculated values of the deviation from the mean value for the entire set of data:

where x – the value of the data set, xav - the mean value.

## 9.1 Result of searching the correlation between data with the R Language

### **9.1.1 The linear correlation coefficient of unfiltered data**

The data sets for analysis in this case:

* a set of average values (for all stations) for each pollutant: NO, NO2, NOX, O3, PM10, PM2,5;
* a set of the total number of patients the day-by-day.

The description of R data frames is presented in [Annexes](#_Annexes_1).

. The result is presented in the table:

|  |  |  |
| --- | --- | --- |
| Data set 1 | Data set 2 | The correlation coefficient |
| ∆NO | ∆Patients | 0.1804087 |
| ∆NO2 | ∆Patients | 0.1012164 |
| ∆NOX | ∆Patients | 0.1616413 |
| ∆O3 | ∆Patients | -0.2199504 |
| ∆PM10 | ∆Patients | 0.05497636 |
| ∆PM2,5 | ∆Patients | 0.0786113 |

Table 10 Result of the searching correlation. Case 1

As we can see from the results, the largest correlation coefficient with the set of data with numbers of patients has data of pollutant NOX. But this value can’t be considered sufficient for the approval of the existence correlation.

### 9.1.2 The linear correlation coefficient of filtered data by diagnosis

The data sets for analysis in this case:

* a set of average values (for all stations) for each pollutant: NO, NO2, NOX, O3, PM10, PM2,5;
* a set of the total number of patients the day-by-day, filtered by diagnosis to 2 data sets: patients with lung diseases and patients with heart diseases.

The description of R data frames is presented in [Annexes](#_Annexes_1).

The result is presented in the table:

|  |  |  |
| --- | --- | --- |
| Data set 1 | Data set 2 | The correlation coefficient |
| ∆NO | ∆Patients(lung) | 0.1719213 |
| ∆NO2 | ∆Patients(lung) | 0.08888615 |
| ∆NOX | ∆Patients(lung) | 0.1535681 |
| ∆O3 | ∆Patients(lung) | -0.2045952 |
| ∆PM10 | ∆Patients(lung) | 0.06846391 |
| ∆PM2,5 | ∆Patients(lung) | 0.0973422 |
| ∆NO | ∆Patients(heart) | 0.1060834 |
| ∆NO2 | ∆Patients(heart) | 0.03138159 |
| ∆NOX | ∆Patients(heart) | 0.08371402 |
| ∆O3 | ∆Patients(heart) | -0.1386231 |
| ∆PM10 | ∆Patients(heart) | -0.005957092 |
| ∆PM2,5 | ∆Patients(heart) | 0.002553244 |

Table 11 Result of the searching correlation. Case 2

In the results, there is no such coefficient to affirm the correlation between the data.

### 9.1.3 The linear correlation coefficient of interval data

To search the correlation considering the factors such as the presence of disease incubation period, and other anthropological factors, it was decided to group the data of air pollution and the total number of patients shown in paragraph 2.1, by 5 days. For each pollutant - the average value of 5 days, and for health data - the total number of patients of the 5 days.

The description of R data frames is presented in [Annexes](#_Annexes_1).

The result is presented in the table:

|  |  |  |
| --- | --- | --- |
| Data set 1 | Data set 2 | The correlation coefficient |
| ∆NO | ∆Patients | 0.36395 |
| ∆NO2 | ∆Patients | 0.2697953 |
| ∆NOX | ∆Patients | 0.3613143 |
| ∆O3 | ∆Patients | -0.371573 |
| ∆PM10 | ∆Patients | 0.1243245 |
| ∆PM2,5 | ∆Patients | 0.1249298 |

Table 12 Result of the searching correlation. Case 3

These coefficients are not sufficient to confirm the existence of a correlationю

### 9.1.4 Linear Regression

For the NO pollutant (interval data set), which has the highest correlation coefficient with patient data, was constructed a linear regression model for visual verification of the correlation of data. In R language for this reason is used function *lm*:

|  |
| --- |
| > model <- lm(formula = fiveDaysNO$av\_value ~ fiveDaysPatients$DIFFERENCE)  > summary(model)  Call:  lm(formula = fiveDaysNO$av\_value ~ fiveDaysPatients$DIFFERENCE)  Residuals:  Min 1Q Median 3Q Max  -10.537 -3.857 -1.577 2.705 20.811  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 14.30545 0.39801 35.942 < 2e-16 \*\*\*  fiveDaysPatients$DIFFERENCE 0.08759 0.01518 5.769 2.71e-08 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 5.904 on 218 degrees of freedom  Multiple R-squared: 0.1325, Adjusted R-squared: 0.1285  F-statistic: 33.29 on 1 and 218 DF, p-value: 2.708e-08 |

“Residuals vs Fitted” chart:

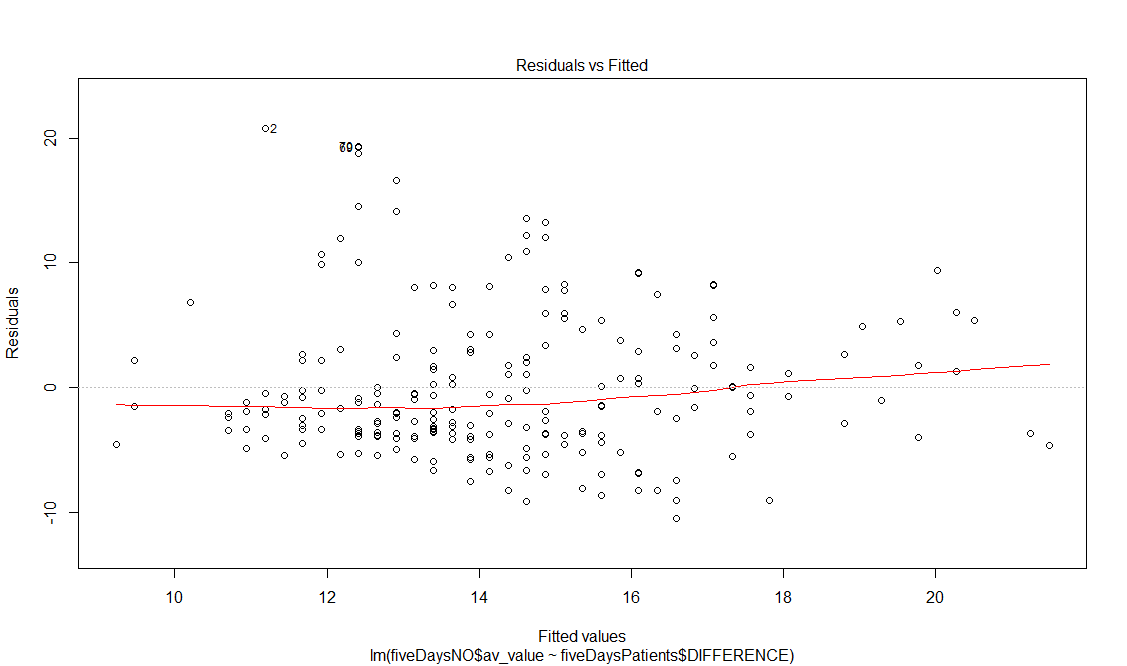


Figure 7 Residuals vs Fitted chart

## 9.2 Prediction Data with R Language

### 9.2.1 Prediction of environmental data

In the R language, there are tools that allow to extrapolate and interpolate data. Thus, we can extrapolate air pollution data for 2017 using the prediction function:

|  |
| --- |
| #creating the sequence of dates for the year of 2017  > pred <- seq(as.Date("2017/1/1"), as.Date("2017/12/31"), by = "day")  > daysDataframe <- data.frame(pred)  > daysDataframe$pred <- strftime(daysDataframe$pred,"%Y-%m-%d %H:%M:%S.0")  > names(daysDataframe) <- c("DAY\_")  #creating the linear regression model of the current data  > model <- lm(AV\_VALUE ~ as.Date(DAY\_), data=environmentDataNO)  #adding predicted value to the dataframe  > daysDataframe$predValueNO = predict(model,daysDataframe) |

As a result, we have a set of data transformed into a linear function and extrapolated to the future. The result is shown in the graph below:

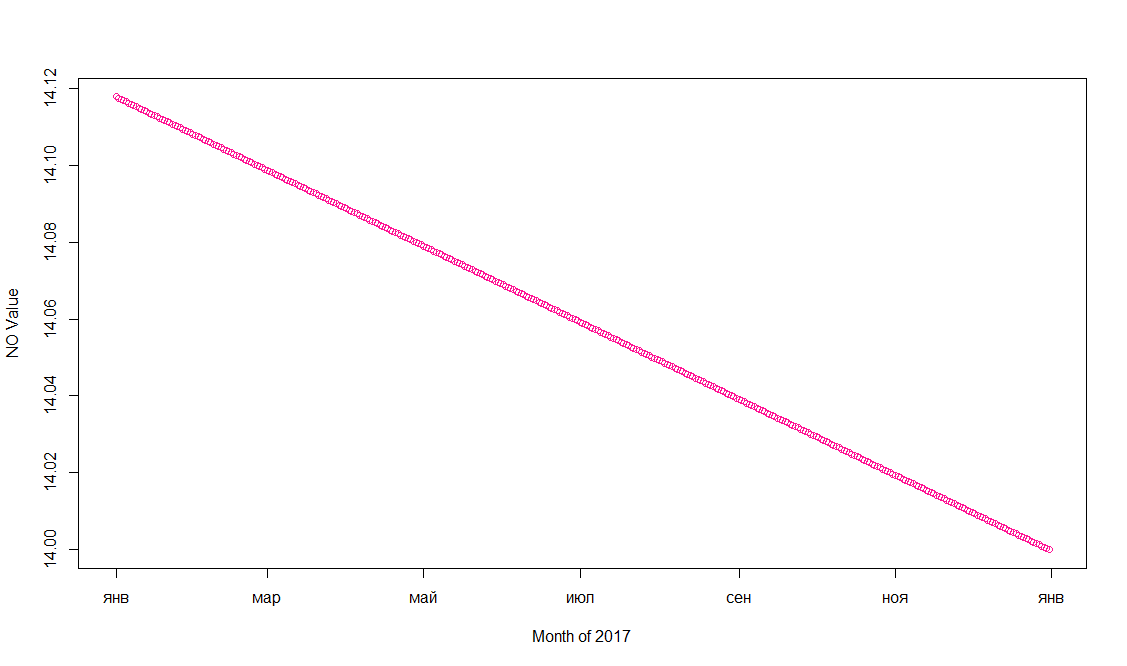


Figure 8 Chart of the extrapolated environmental data

### 9.2.2 Prediction of health data

The actions described above can be repeated for the data on the number of patients:

|  |
| --- |
| #creating the linear regression model of the current data  > model <- lm(NUMBEROFPATIENTS ~ as.Date(DAY\_), data=patientsData)  #adding predicted value to the dataframe  > daysDataframe$predNumberOfPatients = predict(model,daysDataframe) |

The result is shown in the graph below:

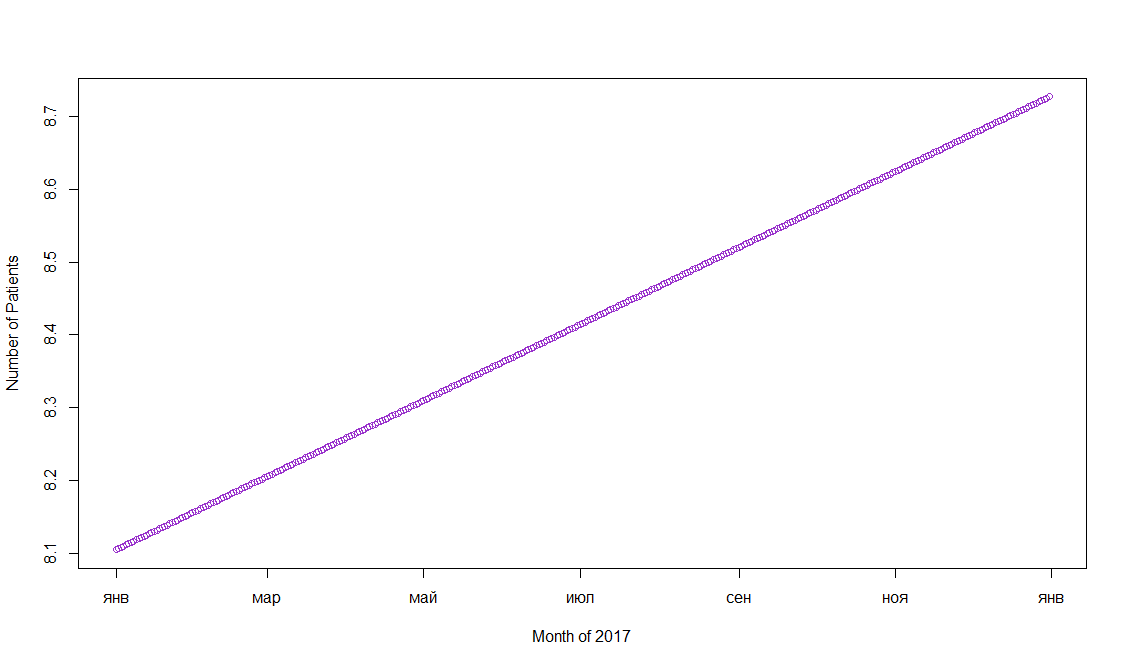


Figure 9 Chart of the extrapolated health data

## 9.3 Different Charts in R

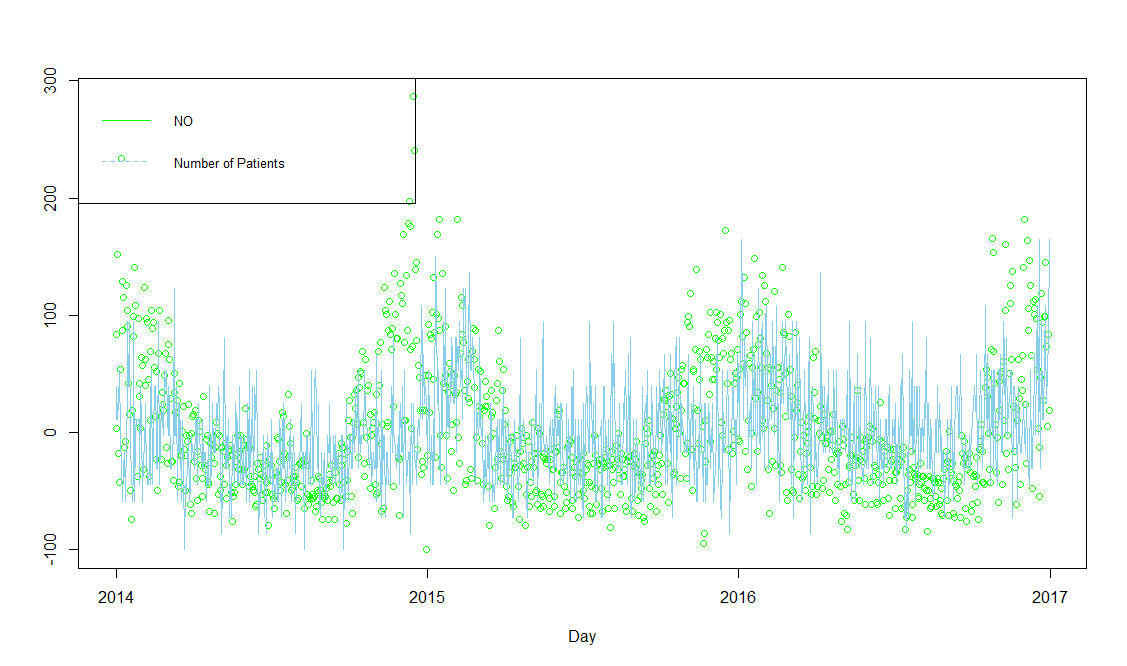


Figure 10 Number of patients – Value of NO Chart

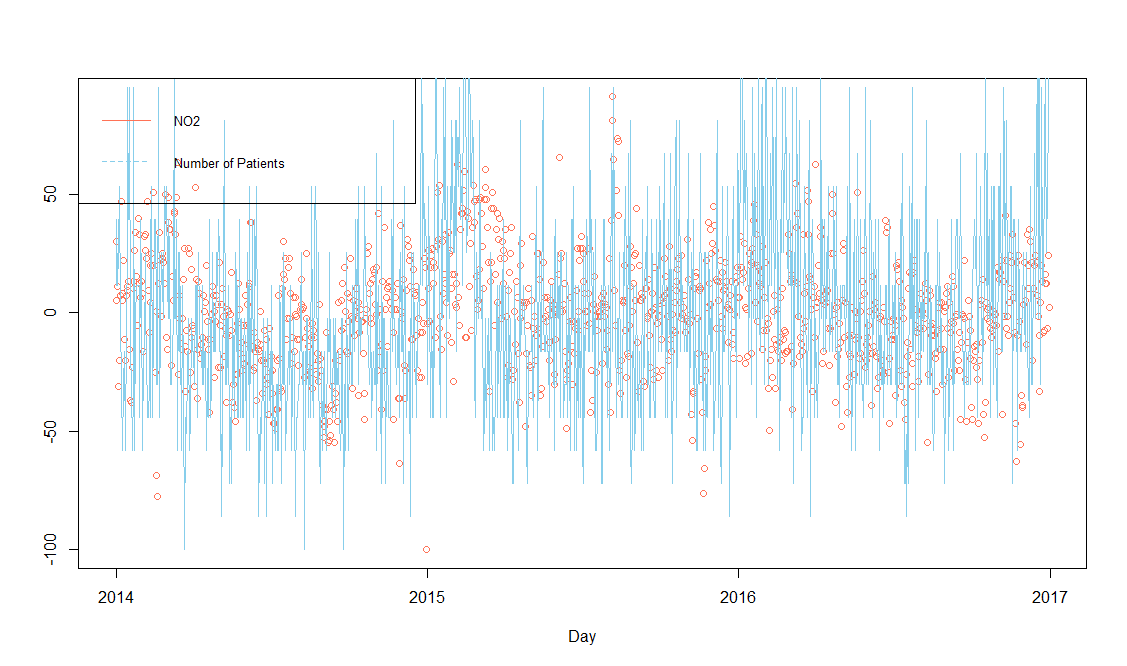


Figure 11 Number of patients – Value of NO2 Chart

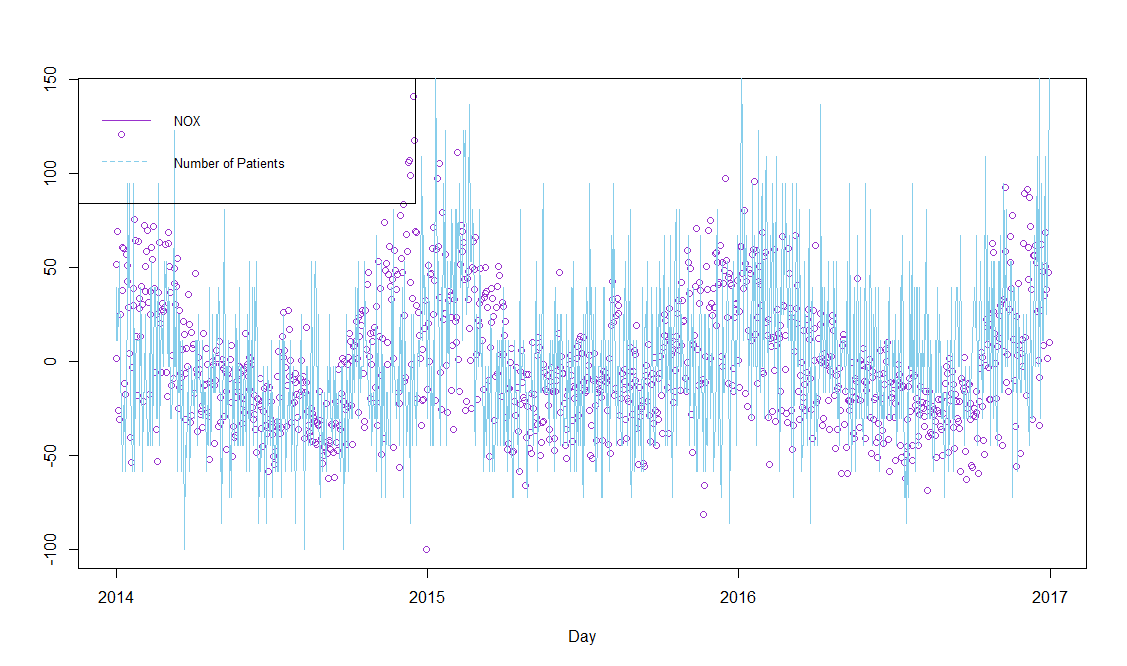


Figure 12 Number of patients – Value of NOX Chart;

# 10 Big Bridge - SE Web Application

This part describes the web application we implemented to display the different results we have processed.

When the application is launched, a connection is made to a cluster of MBDS which provides a list of Oracle Big Data platform components we used for this project. It's also in this server we create ant put the data we use throughout the Big Bridge project.

## 10.1 Displaying the data on Data Tables

We create views in order to display the data of the tables we implemented, for example this view displays the data from the SQL table «Dyspnea»:

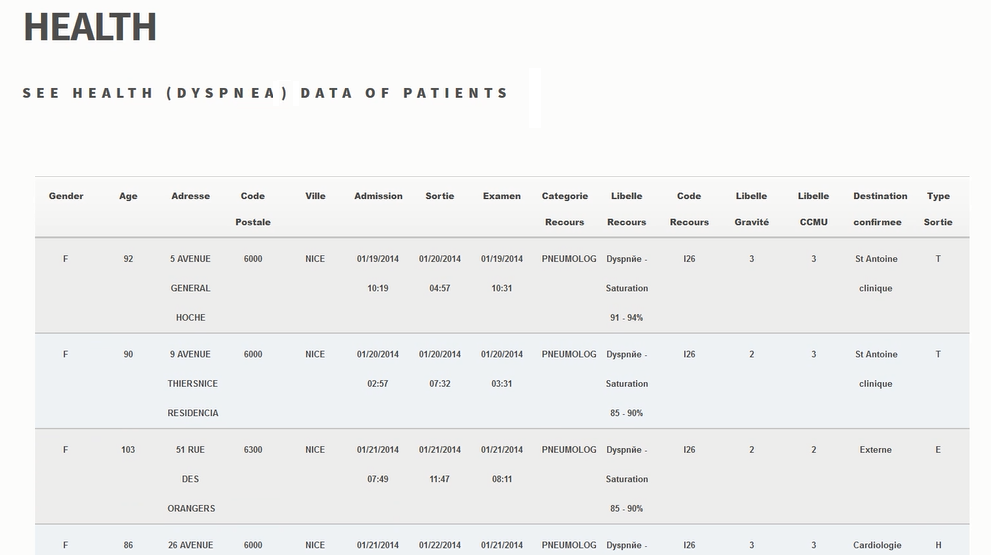


Figure 13 Dyspnea data

We also display the different use cases obtains from R results.

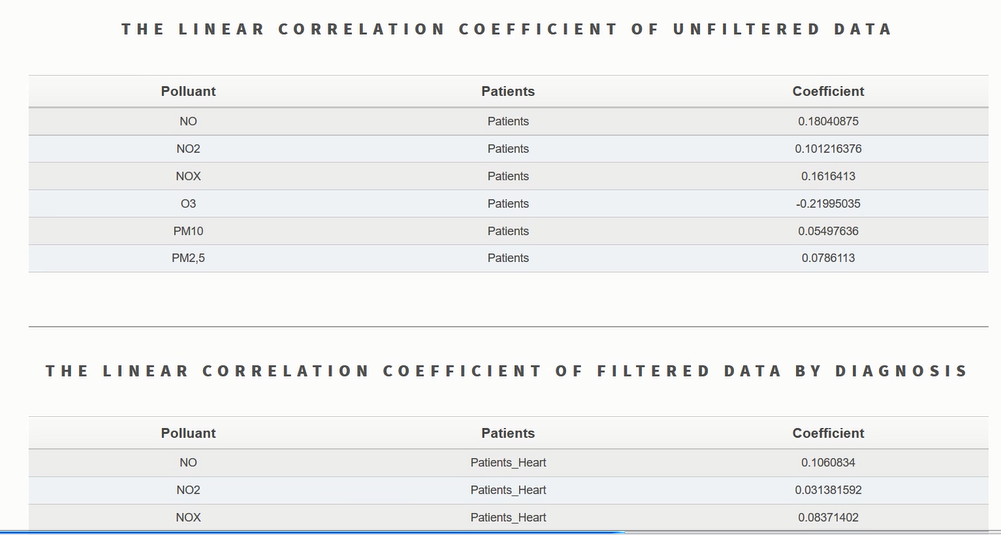


Figure 14 The linear correlation coefficient of filtered data by diagnosis

## 10.2 Displaying OLAP Query result on the Map

We used the Google Map API to display the results for each pollution station of the number of patients concerned by dyspnea and the average of each pollutant measured.



Figure 15 Displaying OLAP Query result on the Map

## 10.3 Displaying R charts

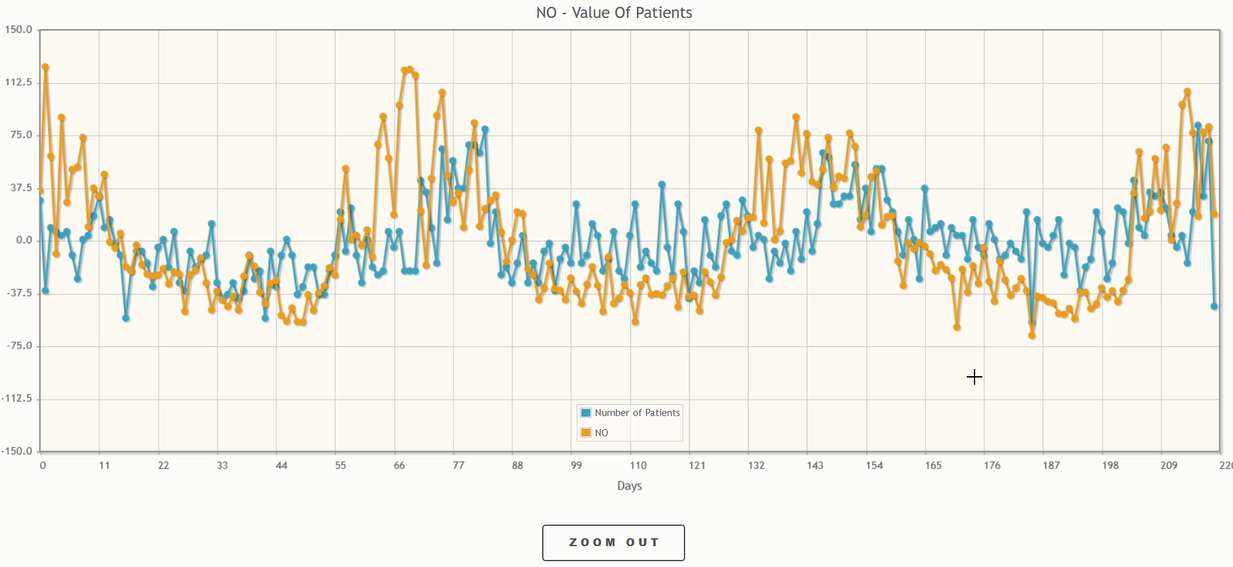


Figure 16 Chart of Use Case 1

Those following charts are from the R results of the previously use cases. But in order to use the data we had to store the R results in SQL tables and then make a SQL Query to get the data and fill our charts.

Sample of the code to get the data:

|  |
| --- |
| public static List<Double> getNOValuesRchartPol() {          List<Double> RchartList = new ArrayList<>();          try {              String sql = "select value from rchart\_pol where pollutant='NO'";              PreparedStatement statement = OracleConnection.getConnection().prepareStatement(sql);              ResultSet result = statement.executeQuery(sql);              while (result.next()) {                  RchartList.add(result.getDouble(1));              }              result.close();          } catch (SQLException e) {              e.printStackTrace();          }          return RchartList;      } |

Sample of code to fill the chart :

|  |
| --- |
| private LineChartModel initLinearModel2() {          LineChartModel model = new LineChartModel();          LineChartSeries series1 = new LineChartSeries();          series1.setLabel("Number of Patients");          allRPatientsList = RequestManager.getAllValuesRchartPatients();          NO2ValuesList = RequestManager.getNO2ValuesRchartPol();            for (int i = 0; i < allRPatientsList.size(); i++) {              series1.set(i,allRPatientsList.get(i));          }          LineChartSeries series2 = new LineChartSeries();          series2.setLabel("NO2");         for (int i = 0; i < NO2ValuesList.size(); i++) {              series2.set(i,NO2ValuesList.get(i));          }          model.addSeries(series1);          model.addSeries(series2);          return model;      } |

# 11 General Conclusion

The result of this project is an implemented software, using the Big Data approach, which allows to manage and analyze data, using Open Sours tools.

An analysis of the presented data for the existence of a linear correlation was also carried out. The linear correlation between data of air pollution and the number of patients with Dyspnea was not established in this study.

As can be seen from the results, there is a sense to continue experiments of finding correlation with such pollutants: NO, NО2 and NОХ, which show the greatest possibility of correlation.

Also, it is planned to enrich the project with unstructured data coming both from social networks and captors in connected watches (monitoring heart beats in real time) to enable real-time cartography for monitored individuals.

For further analysis of the existing data, the following approaches are proposed, which can be implemented with the language R:

1) Clustering: to identify the Risk Factors (such as age, for example, or place of living) for patients with Dyspnea in Nice for existing health data

2) Classification: a method that allows a person to be assigned to a group defined in point 1, according to his description. It allows to determine the risk of Dyspnea for a person by means of predictive functions. It can be used in the future when including to the system the individual user profiles.

Then - to correlate personal Risk Factors and the environment data for constructing recommendations for the user (for example, caution about visiting heavily polluted areas if a person has high blood pressure).

# Reference and bibliography

1. Holland W. W., Reid D. D. The urban factor in chronic bronchitis. Lancet
2. PAARC: Groupe Cooperative /Lelouche J Pollution atmosphérique et affections respiratoires chroniques ou à répétition. Bull. Eur. Physiopathol. Respir
3. Schenker M. B., Samet J. M., Speizer F. E., Gruhl J., Batterman S.Health effects of air pollution due to coal combustion in the Chestnut Ridge region of Pennsylvania: results of cross-sectional analysis in adults. Arch. Environ. Health
4. Euler G. L., Abbey D. E., Magie A. R., Hodlkin J. E.Chronic obstructive pulmonary disease symptom effects of long term cumulative exposure to ambient levels of total suspended particulates and sulfur dioxide in California Seventh-Day Adventist residents. Arch. Environ. Health
5. Portney P., Mullahy J.Urban air quality and respiratory disease. Reg. Sci. Urban Econ.
6. Schwartz J. Particulate air pollution and chronic respiratory disease. Environ. Res.
7. Forsberg, B., N. Stjernberg, and S. Wall. 1997. Prevalence of respiratory and hyperreactivity symptoms in relation to levels of criteria air pollutants in Sweden. Eur. J. Public Health 7/3:291–296.
8. Abbey D. E., Lebowitz M. D., Mills P. K., Petersen F. F., Beeson W. L., Burchette R. J.Long-term ambient concentrations of particulates and oxidants and development of chronic disease in a cohort of nonsmoking California residents. Inhal. Toxicol.
9. Abbey, D. E., B. E. Ostro, F. Petersen, and R. J. Burchette. 1995. Chronic respiratory symptoms associated with estimated long-term ambient concentrations of fine particulates less than 2.5 microns in aerodynamic diameter (PM2.5) and other air pollutants. J. Exp. Anal. Environ. Epidemiol. 5/2:137–159.
10. Abbey D. E., Hwang B. L., Burchette R. J.Estimated long term ambient concentrations of PM10 and development of respiratory symptoms in a nonsmoking population. Arch. Environ. Health
11. Scarlett J. F., Griffiths J. M., Strachan D. P., Anderson H. R.Effect of ambient levels of smoke and sulphur dioxide on the health of a national sample of 23-year-old subjects in 1981. Thorax
12. Schwartz J., Dockery D. W.Increased mortality in Philadelphia associated with daily air pollution concentrations. Am. Rev. Respir. Dis.
13. Spix C., Heinrich J., Dockery D., Schwartz J., Volksch G., Schwinkowski K., Collen C., Wichmann H. E.Air pollution and daily mortality in Erfurt, East Germany, 1980–1989. Environ. Health Perspect.
14. Dockery D., Pope A., Xu X., Spengler J. D., Ware J. D., Fay M. E., Ferris B. J., Speizer F. E.An association between air pollution and mortality in six U.S. cities. N. Engl. J. Med.
15. Touloumi G., Pocock S. J., Katsouyanni K., Trichopoulos D.Short-term effects of air pollution on daily mortality in Athens—a time-series analysis. Int. J. Epidemiol.
16. Schwartz J.Air pollution and daily mortality: a review and meta-analysis. Environ. Res.
17. Pope A., Thun M., Namboodiri M., Dockery H. D. W., Evans J. S., Speizer F. E., Heath C. W.Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am. J. Respir. Crit. Care Med.
18. Schwartz J., Morris R.Air pollution and hospital admissions for cardiovascular disease in Detroit, Michigan. Am. J. Epidemiol.
19. Burnett R., Dales R., Krewski D., Vincent R., Dann T., Brook J. Associations between ambient particulate sulfate and admissions to Ontario Hospitals for cardiac and respiratory diseases. Am. J. Epidemiol.
20. AirPaca. Association de surveillance de la qualité de l'air agréée par le ministère de l'environnement – URL – <http://www.airpaca.org/>
21. Wikipedia. R (programming language) – URL – <https://en.wikipedia.org/wiki/R_(programming_language)>
22. Wikipedia. Correlation and dependence – URL – <https://en.wikipedia.org/wiki/Correlation_and_dependence>
23. Wikipedia. Scrum (software development) – URL – <https://en.wikipedia.org/wiki/Scrum_(software_development)>
24. G. Mopolo-Moké, course Big Data et les SGBDs NoSQL, 2016
25. N. Pasquier, Data Analytics & Mining course, 2016
26. B. Renaut, Introduction à Hadoop & MapReduce course, 2015
27. S. Miranda, course Introduction stratégique des Bases de Données à Big Data, 2015

# Annexes

*Annex 1. Scrum methodology of project organization*

It is the methodology of agile development software. It focuses on the quality control of the development process. The project is divided into releases, which are divided into iterations, called "Sprint" ([23](#_Reference_and_bibliography)). The release is a major objective of the project, and the sprint is for a minor (or intermediate) objective. The duration of the sprint should be on average 2 to 3 weeks.

The roles and responsibilities are divided into three categories, according to the Scrum method: The Product Owners, The Scrum Master and The Development Team.

* The Product Owners

The Product Owner represents the product's stakeholders and the voice of the customer.

* The Scrum Master

Scrum Master is accountable for removing impediments to the ability of the team to deliver the product goals and deliverables. The Scrum Master ensures that the Scrum framework is followed. The Scrum Master helps to ensure the team follows the agreed processes in the Scrum framework, often facilitates key sessions, and encourages the team to improve.

* The Development Team

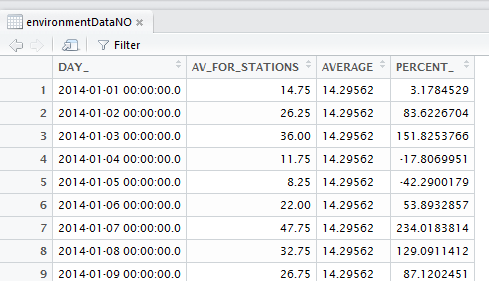
The Development Team is responsible for delivering potentially shippable increments of product at the end of each Sprint (the Sprint goal). A team do the actual work (analyse, design, develop, test, technical communication, document, etc.).

*Annex 2: Project Plan*

*Annex 3: Installation guide*

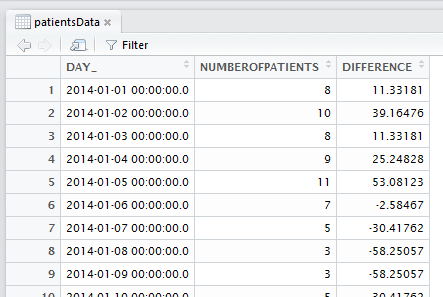
*Annex 4: The description of R data frames*

Example of environmental data frame:

**

DAY\_ - date, AV\_FOR\_STATIONS – the average value of this pollutant for all the stations, AVERAGE – the average value of the 2nd column, PERCENT\_ - delta according to the mean value to calculate the correlation (in percentages)

Example of the patient’s data frame:



DAY\_ - date, NUMBEROFPATIENTS – total number of patients for this date, DIFFERENCE - delta according to the mean value to calculate the correlation (in percentages)

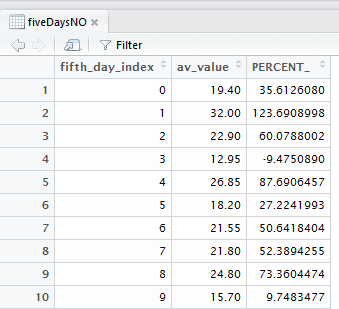
Grouping patients by diagnosis:

|  |
| --- |
| heartPatiensData <- dbGetQuery(conn\_oracle, "Select count (\*) as numberOfPatients, day\_ from  (Select diag1, to\_date(substr(admission, 0, 10),'MM/DD/YYYY') as day\_ from dyspnee where substr(diag1,0,1)='I') group by day\_ order by day\_")  lungPatiensData <- dbGetQuery(conn\_oracle, "Select count (\*) as numberOfPatients, day\_ from  + (Select diag1, to\_date(substr(admission, 0, 10),'MM/DD/YYYY') as day\_ from dyspnee where substr(diag1,0,1)='J') group by day\_ order by day\_") |

Grouping data by 5 days (example for the environmental data):

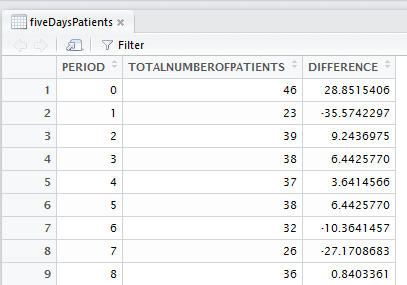
|  |
| --- |
| testEnvironmentDataNO <- environmentDataNO  testEnvironmentDataNO$fifth\_day\_index <- c(0, rep(1:(nrow(testEnvironmentDataNO)-1)%/%5))  fiveDaysNO <- group\_by(testEnvironmentDataNO, fifth\_day\_index) %>% summarise(av\_value = mean(AV\_FOR\_STATIONS)) |

Example of the grouping environmental data frame (by 5 days):



fifth\_day\_index – the number of the period, av\_value – the average value for this 5 days, PERCENT\_ - delta according to the mean value to calculate the correlation (in percentages)

Example of the grouping patient’s data frame (by 5 days):



fifth\_day\_index – the number of the period, TOTALNUMBEROFPATIENTS – the total number of patients for this 5 days, PERCENT\_ - delta according to the mean value to calculate the correlation (in percentages)