

MINISTRY OF HIGHER EDUCATION
FOR SCIENTIFIC RESEARCH AND TECHNOLOGY
UNIVERSITY OF MANOUBA



NATIONAL SCHOOL OF COMPUTER SCIENCES

SOFTWARE REQUIREMENTS SPECIFICATION
OF ENGINEER DIPLOMA THESIS

Big Data Management for Manufacturing Intelligence Application

Performed By:
Mèhdi Ben Hamida

Supervised By:
Anis Mlaouah

HOSTING COMPANY:



LAKE GARDENS, DOLLAR STREET LAKE CITY CENTER BUILDING BLOC A
1ST FLOOR, TUNIS 1053

WEBSITE: WWW.INTEGRATIONOBJECTS.COM
PHONE NUMBER: +216 71 195 360

Academic Year:
2016-2017

Signature & stamp:

A large, empty rectangular box with a thin black border, intended for a signature and stamp. It occupies the lower half of the page.

Contents

1	Introduction	1
1.1	Purpose	1
1.2	Scope	1
1.3	Hosting Organism	1
1.4	General Context	1
1.5	Project Description	2
2	Requirements Specification	3
2.1	Functional requirements	3
2.2	Non-functional requirements	3
3	Constraints	5
3.1	Environment	5
3.1.1	Hardware environment	5
3.1.2	Software environment	5
3.2	Plan	6
4	Conclusion	7

Chapter 1

Introduction

1.1 Purpose

This document is the software requirements specification for a design and implementation of a Big Data Management for Manufacturing Intelligence Application. Throughout this document we will define the project's context and present it. This document will specify the project's requirements, the working environment and the estimated plan.

1.2 Scope

This project is a requirement for the final year in the **National School of Computer Science** in order to obtain the National Diploma of a Software Engineer. It will be achieved during an internship at **Integration Objects** that will last 4 months starting from February 1st, 2017 to May 31th, 2017.

1.3 Hosting Organism

Integration Objects is a world-leading manufacturing tech solution provider for knowledge management, automation, plant performance management, decision support, and system connectivity for the process, power, and utilities industries around the globe. The company provides advanced software solution software for multinational companies in the oil and gas industry to empower manufactories operations and increase its efficiency. **Integration Objects** is **Microsoft** gold partner and a member of the **OPC Foundation**, **MIMOSA** and **ISA**.

1.4 General Context

KnowledgeNet Analytics is an out of the box data analysis software product designed to meet operations needs in terms of knowledge mining in order to

unlock hidden knowledge and profits within plant historical data. **KnowledgeNet Analytics** helps analyzing and investigating the plant process behavior and states, identifying opportunities to increase operational efficiency and reliability, and predicting abnormal conditions. One of the known issues in this field is how to manage huge volumes of data in term of storage, algorithms optimizations and presentation. The purpose of this project is to study and provide the best way for big data management.

1.5 Project Description

The main objective of this project is to:

- Perform a literature review to understand the project scope
- Engineer the best technical solution to load and manage huge volumes of data into the platform
- Apply standard operations such as matrix operations and complex algorithms (such as PCA (Principal Components Analysis) and PLS (Partial Least Square), etc.) using the best solution for big data
- Compare results with other products

Chapter 2

Requirements Specification

2.1 Functional requirements

The delivered application consists on designing and implementing Client / Server application that should satisfy the specific functionalities listed below:

- The system must offer to the user the possibility to load big data from different sources into HDFS. The user is not coming from An IT background; all command lines should be hidden behind a Graphical User Interface.
- The system allows user to apply on large volume of data several data mining algorithms such as:
 - Principal components analysis
 - Linear Model
 - K-Means
- The User should be able to visualize algorithm results in different interactive forms (Charts, Graphics, data visualizations, grids ...)

2.2 Non-functional requirements

While achieving the functional requirements, the solution should satisfy the following non functional requirements:

- **Scalability:** The modularity and extensibility of the application architecture in case of the addition of new services. In our case, when the size or volume of data becomes larger or the number of servers increase, the application should guarantee a regular behaviour.
- **Ease of use:** The application must have a user-friendly and ergonomic interface.

- **Security:** The data storage should be far from being accessed by an unwanted user. Every operation of access should be preceded by an authentication. Data storage servers should be able to provide disponibility of the data when required.
- **Documentation:** The solution should be well-documented in order to provide the best use for the costumer.

Chapter 3

Constraints

3.1 Environment

3.1.1 Hardware environment

To achieve this work we used a personal computer with the following characteristics:

Workstation :

Band: Laptop HP Probook 450 g3

Processor: Intel Core i7 CPU

Memory: 8,00GO of RAM

Operating system: Windows 10.

3.1.2 Software environment

To achieve this work we have started by installing the following software:

- **Microsoft Visual Studio:** Microsoft Visual Studio is an integrated development environment provided by Microsoft. This IDE has a rich set of tools that facilitates web applications, web sites, web services and computer programs development. Moreover, Microsoft visual studio has very
- **Spark:** Apache Spark is a fast and general engine for large-scale data processing. Spark powers a stack of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and spark Streaming. Spark extends its predecessors with in-memory processing. Its Resilient Distributed Dataset (RDD) abstraction enables developers to materialize any point in a processing pipeline into memory across the cluster, meaning that future steps that want to deal with the same data set need not recompute it or reload it from disk. Sparks runs on Hadoop, Mesos, standalone, or in the cloud.
- **Hadoop Distributed File System (HDFS:)** HDFS is a distributed file system that provides high-performance access to data across Hadoop clus-

ters. Like other Hadoop-related technologies, HDFS has become a key tool for managing pools of big data and supporting big data analytics applications.

3.2 Plan

During these four months, starting from February 1st, 2017 and ending in May 31th, 2017, the project will follow the plan illustrated in the figure below. This plan represents the important tasks distribution during the next sixteen weeks of work. This plan is not final, depending on the work evolution, some changes may happen.

Table 3.1: The Project Plan

	February				Mars				April				May			
Documentation																
Specification																
Prototyping																
Design																
Implementation																
Tests																
Report redaction																

Chapter 4

Conclusion

Through this software requirements specification we tried to present a reference for the next steps during the project's development process. Considering the work progress, we will do our best to be faithful to the specification characterized above.