Data Mining
Open Data

Lab Semantic Data Web

Outlier Detection on Financial RDF Data

Mentor

Christiane Engels

Group students

Berivan Ekmez, Zuhair Almhithawi, Nayef Roqaya

Requirement specification Document

V1

2016

Table of Contents

1 Introduction:

1.1 Motivation:

The World Wide Web has enabled the creation of a global information space comprising linked documents. Linked Data provides a publishing paradigm in which not only documents, but also data, can be a first class citizen of the Web, thereby enabling the extension of the Web with a global data space based on open standards (Heath and Bizer, 2011).

Growing of Semantic data Web and Linked Data technologies additional to development data mining methods in data sciences field, contribute in growing the role semantic data web in developing enterprise systems for many goals. The idea is to use the Linked Open Data Cloud for retrieving additional information on e.g. demographics or economics to enrich the data sets at hand before analyzing them. The project concentrates on financial data and detects outlier then compares the results. Certain data mining methods can be applied to these separate datasets (Financial data) to efficiently detect anomalies/outliers, and by the use of Semantic Web concepts, we can have even more accurate findings.

1.2 Project Purpose:

Applying different methods to detect outliers and anomalies in financial RDF data and compare the results.

1.3 Definitions:

1.3.1 What is an outlier/anomaly?

An anomaly is something that deviates from what is standard, normal, or expected

1.3.2 Anomaly detection in data-mining

In data mining, anomaly detection (outlier detection) is the identification of items, events or observations, which do not conform to an expected pattern or other items in a dataset.

1.3.3 Outliers in financial data:

An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).

2 Project Time plan:

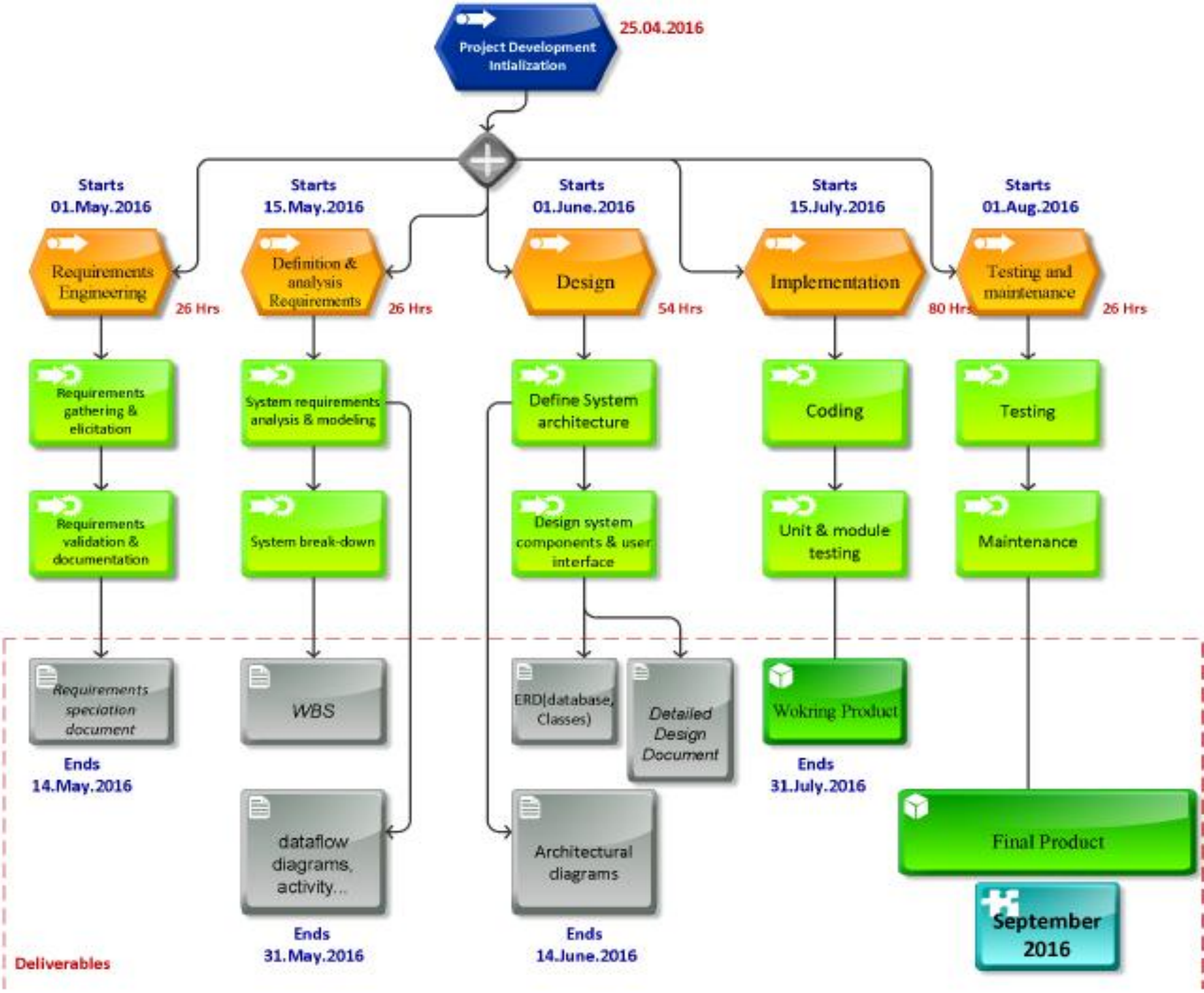

Figure 1 Time plan

3 System Requirements:

3.1 Functional requirements:

ID	Name	Comments
FRQ - 1	The user shall import his/her RDF datasets	Datasets may be in XML/RDF format
FRQ - 2	The system shall Link local datasets to external dataset cloud	External dataset will be chose manually in financial field
FRQ - 3	The system shall apply and compare different data mining methods to detect outliers	Comparing aspects like accuracy and performance of the applied DM methods
FRQ – 4	The system shall provide reports on analyzed data	Reports may include: • Comparison for DM methods • Results visualization • Explanations

3.2 Technical Requirements:

The system shall be implemented using Java SE environment

- User interface: cross browser interface that runs on most web browsers (e.g. IE, Chrome, Firefox, Safari)
- Front end languages and technologies: HTML5/CSS3, JavaScript, JQuery library
- Restful API for web services
- Tools to be used: Fusike, Rapid miner and Apache Jena

3.2.1 Fusike:

Expose your triples as a SPARQL end-point accessible over HTTP. Fuseki provides REST-style interaction with your RDF data (Jena.apache.org, 2016).

3.2.2 Apache Jena:

A free and open source Java framework for building Semantic Web and Linked Data applications (Jena.apache.org, 2016).

3.2.3 Rapid miner:

Open Source Predictive Analytics Prep data, create models, and embed in business processes Faster than ever before (Platform, 2016).

3.3 Use case:

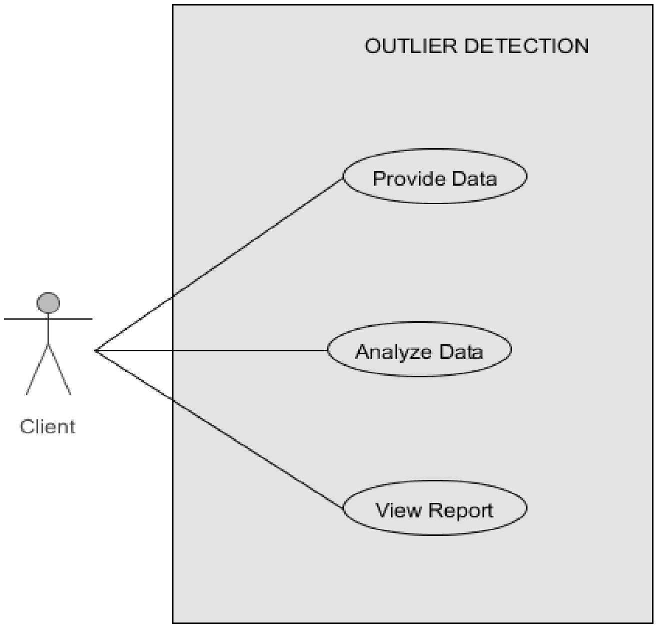

Figure 2 Use case Diagram Level 0

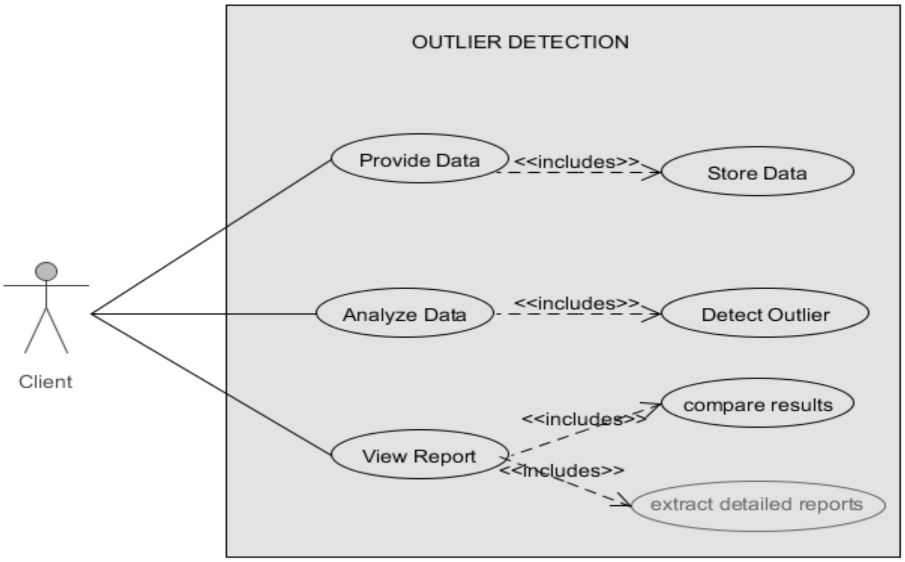

Figure 3Use case Level 1

3.4 Non- functional requirements:

ID	Name	Comments
NRQ - 1	Reliability	The system should be able to perform the required functions under stated conditions for specific period.
NRQ - 2	Usability	The system has to give ease with which a user can learn to operate, Prepare inputs for, and interpret outputs of system. Additionally easy-to-use interfaces.
NRQ - 3	Capacity	The ability of the system to handle transactional volumes is a very important characteristic for a system
NRQ – 4	Performance	The system should consider the Processed transactions per second and Response time to user input and performance of the implemented DM methods
NRQ – 5	Open source	Source code made available with a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose.
NRQ – 6	Quality	Quality of the outlier and detection methods for giving precise result.

4 reference:

[1]. Jena.apache.org. (2016). Apache Jena -. [online] Available at: https://jena.apache.org/index.html [Accessed 5 Jun. 2016].

[2]. Platform, R. (2016). RapidMiner | #1 Open Source Predictive Analytics Platform. [online] RapidMiner. Available at: https://rapidminer.com/ [Accessed 5 Jun. 2016].

[3]. Heath, T. and Bizer, C. (2011) ‘Linked data: Evolving the web into a global data space’, Linked Data Evolving the Web into a Global Data Space, 1(1), pp. 1–136, page 7.