# DIGIT.B4 – Big Data PoC

## GROW – Transpositions

## D02.02 Technological Architecture

everis Spain S.L.U

# Table of contents

## Tables

## Figures

# 1 INTRODUCTION

The construction of this proof of concept is based on the definition of a technical architecture. which includes the following section:

- Definition of the technical architecture.

# 2 TECHNICAL ARCHITECTURE

The system architecture of the DIGIT.B4 – Big Data PoC for the purpose of supporting policy DGs of the European Commission in the domain of Member States transposition of EU legislation will be based on the NTT Data – everis Big Data Reference Architecture.

The reference architecture is a collection of building blocks which consist of a series of technologies and methodologies.

NTT DATA and everis Big Data Reference Architecture is composed of two levels:

1. **Layer:** A group of building blocks which have the same technology/methodology. There are eight layers that are categorised for all technologies/methodologies important for the architecture.
2. **Building blocks:** Each one represents a technology/methodology. There are 38 building blocks in the architecture.

The building blocks that will be used in the PoC for each layer have been highlighted.

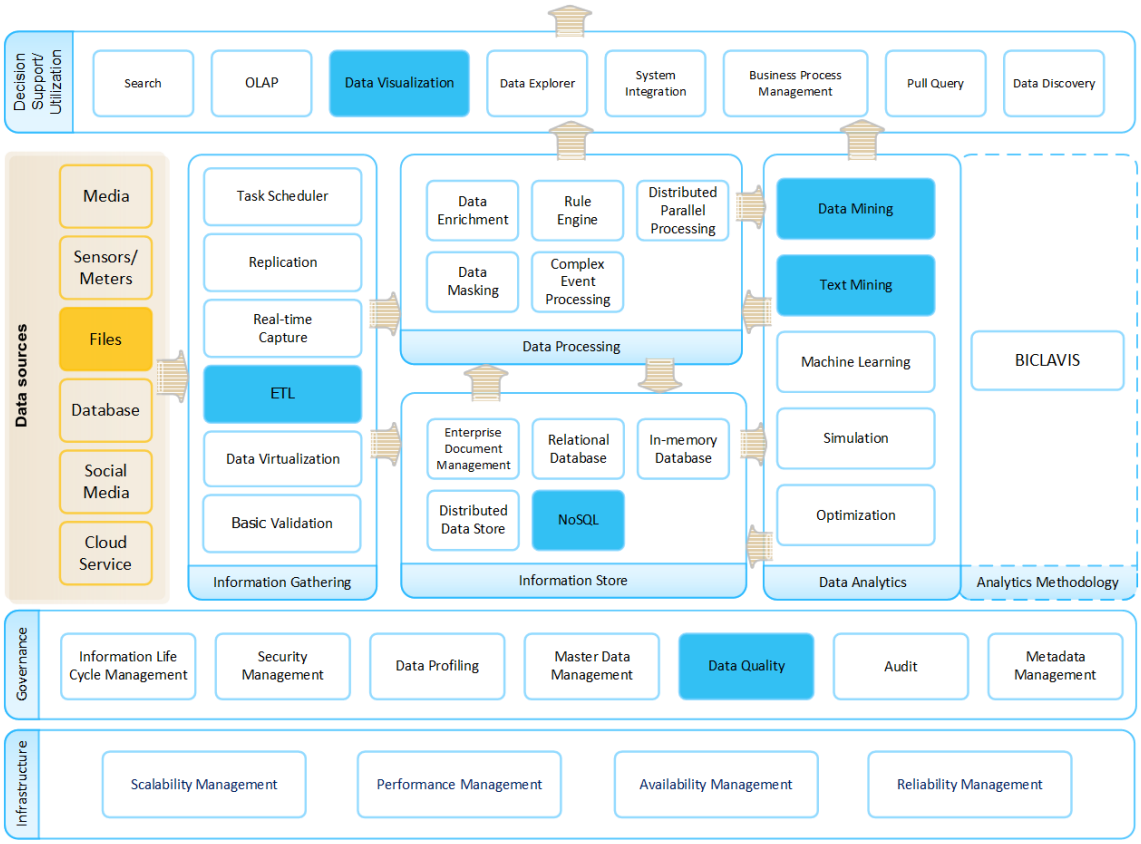


**Figure 1 - System architecture**

The technologies chosen for each building block are listed, and the descriptions for the use that the PoC will make of each highlighted block are listed below:

## 2.1 Information gathering

| ETL (Extract, Transform and Load) | |
|---|---|
| **Building block description** | Extract, Transformation and Load (ETL) is an industry standard term used to represent the data movement and transformation processes. |
| **Building block PoC use** | The files obtained from the data sources have to be transformed to other formats in order to be readable by the algorithms. |
| **Building block technologies** | R, python |

**Table 1- Information gathering**

## 2.2 Information storage

| Relational Data Base | |
|---|---|
| **Building block description** | NoSQL (Generally, called "Not only SQL") is a word categorising database management systems different from a relational database management system (RDBMS). |
| **Building block PoC use** | The information will be stored without a predefined scheme. |
| **Building block technologies** | Vertica, Cassandra, Apache Accumulo, Apache Hbase, MarkLogic Server, MongoDB, Couchbas, Neo4j, AllegroGraph, InfiniteGraph, FlockDB, Sqrrl Enterprise, Oracle NoSQL Database, Riak, Redis. |

**Table 2 - Information storage**

## 2.3 Data analytics

| Data Mining, Text Mining, Machine Learning | |
|---|---|
| **Building block description** | **Data Mining**: Data mining refers to the statistical techniques and rule-based models that aim at discovering new knowledge and trends from a large data set. In contrast to machine learning, data mining emphasises the process of analysing data rather than the accuracy of the used models.<br>**Text Mining**: Text mining can be defined as an intensive process in which a user interacts with a document collection over time by using a suite of analysis tools to extract knowledge. |
| **Building block PoC use** | The algorithms used for the PoC will use techniques from all of these blocks. |
| **Building block technologies** | R, python |

**Table 3 - Data analytics**

## 2.4 Governance

| Quality of Data | |
|---|---|
| **Building block description** | The goal of a data quality exercise is to establish high quality data and maintain it. The main tools/techniques that aid success in data health are related to Master Data Management (MDM). |
| **Building block PoC use** | The gathered data has to be treated to ensure its quality before sending it to classification |

| | |
|---|---|
| | model. |
| **Building block technologies** | R, python |

**Table 4 - Governance**

## 2.5 Decision support/utilisation

| **Visualisation** | |
|---|---|
| **Building block description** | Data visualisation means showing data in a graphical format. It involves the creation and study of the visual representation of data. The primary goal of data visualisation is to communicate information clearly and efficiently to users via graphics, such as tables, charts and maps. |
| **Building block PoC use** | The PoC results will be published in a web that will illustrate the outcome of the algorithms. |
| **Building block technologies** | R, JavaScript, HTML |

**Table 5 - Decision support/utilisation**