A Multi-Agent Framework for Multi-Criteria Business Intelligence driven Smart Data in a Big Data Environment

Zakarya Elaggoune*, Ramdane Maamri† and Imane Boussebough‡
LIRE Laboratory, University of Constantine 2-Abdelhamid Mehri, Constantine- Algeria
Email: *zakarya.elaggoune@univ-constantine2.dz, †ramdane.maamri@univ-constantine2.dz, ‡iboussebough@gmail.com

***Abstract*—Business Intelligence (BI) is a very important process in an entreprise because it provides decision support and enables business strategy managers to have an overview of the activity being treated. The application of new technologies and the era of Big Data pose new challenges for BI, a common problem affecting data quality is the presence of noise and irrelevant information wich can lead decision makers to a wrong decision. In this paper, a Multi-Agent Framework for BI driven Smart Data in a Big Data Environment is presented. For the unstructured and semi-structured data collection and preprocessing we use the framework Hadoop; Apache Flume, Apache Sqoop and ODBC/JDBC Connectors for data extraction and intagration, the Hadoop Distributed File System (HDFS) for data storage and MapReduce for preprocessing. For the structured data, an Extraction,Transformation and Loading (ETL) process based agents is used. Agents perform specific task assigned to them for treating the noise in Big Data problems by applying Analytic Hierarchy Process (AHP) that is one method of Multi-Criteria Decision Making (MCDM), providing high quality and relevant information, also known as Smart Data.**

***Index Terms*—Big Data, Business Intelligence, Hadoop, Multi-Agent System, Smart Data, Decision Making, Analytic Hierarchy Process, Multi-Criteria Decision Making**

I. Introduction

The current world is the one of data, commonly used applications such as social networks (Facebook, Twitter, etc.), forums, messaging systems, research articles, online transactions and corporate data produce heterogeneous data that is enormous in volume and generated exponentially [1]. These data can be very effective in developing business strategies and planning effective business decisions, but the era of big data and its characteristics put many challenges on the storage and analysis of these data that require intelligent mechanisms and tools to manage these data sets.

Big Data, as a concept, is defined around five aspects: data volume, data velocity, data variety, data veracity and data value [2], [3]. Although aspects of volume, variety and velocity refer to the process of data generation, data capture and storage. Veracity and value aspects deal with the quality and the relevance of the data. These two last aspects become crucial in any decision making process, where the quality of decision making is strongly influenced by the quality of the used data.

Smart Data (focusing on veracity and value) has been introduced, aiming to filter out the noise and to highlight the relevant data, which can be effectively used by companies and governments for planning, operation, monitoring, control, and intelligent decision making.

The aim of this paper is to propose a BI solution in a Big Data environment that can filter the noise and extract the relevant information for an intelligent decision making. Multi-agent systems (MAS) can manage complex and distributed computer scenarios, according to that, a new BI framework based MAS is suggested.

II. Problem Statement and Definition

The central research question will be " How can we extract Smart Data to improve the quality of decision making in a Big Data environment ? " This study intends is to find answers to the following questions:

- How to cover the problems of heterogeneous and incompleteness of data ? and how to deal with unstructured data ?
- How to ensure the veracity of data that is extracted from external sources ?
- How to extract Smart Data ?

III. Extracting Smart Data within AHP

The term Smart Data is utilized to denote the challenge of transforming raw data into data that can be processed later to obtain valuable information [4]. Smart data discovery involves filtering big data holding useful information, becoming a subset of data that is important for companies and researchers [5]. Obtaining a reduced / filtered amount of data may involve a large reduction in data storage costs and it influences decision makers to make the right decision.

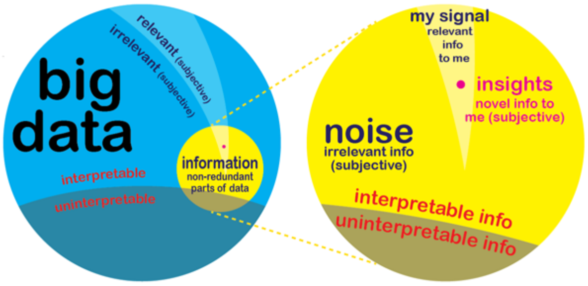

Fig. 1. Extracting Smart Data from Big Data [6]

Figure 1 illustrates an overview of filtering noisy data and extracting Smart Data (relevant information). The different steps of the process are listed in the following paragraphs:

Step 1 (Extract information from Data). "Although data and information are different, many people speak of them as if they are synonymous, which is almost never true" explains Michael Wu in his web article [5]. Data is simply a record of the events that took place. It is the raw data that describes what happend, when, where, how, who is involved, etc. Data does give us information, they are not the same. Information is only the non-redundant parts of the data [7], the maximum amount of extractable information can be measured through lossless compression algorithms.

Step 2 (Extract relevant information). Information must satisfy some criteria to provide relevant information that are valuable. Autohrs in [8] define three key attributes for data to be smart, it must be accurate, actionable and agile:

- **Accurate**: data must be what it says it is with enough precision to drive value. Data quality matters.
- **Actionable**: data must drive an immediate scalable action in a way that maximizes a business objective like media reached across platforms. Scalable action matters.
- **Agile**: data must be available in real-time and ready to adapt to the changing business environment. Flexibility matters.

The Analytic Hierarchy Process (AHP), presented by Thomas Saaty (1980), is an effective tool for dealing with complex decision-making and can help the decision-maker prioritize and make the best decisions. By reducing complex decisions to a series of paired comparisons and then synthesizing the results, AHP helps to understand the subjective and objective aspects of a decision. In addition, AHP incorporates a useful technique for checking the consistency of evaluations of decision-makers, which reduces bias in the decision-making process.

In our case we use the AHP process to filter the noise and selected only the Smart Data According to the following steps:

1) Define the problem, which is in our case "How to extract Smart Data from Big Data?".
2) Structure the top decision hierarchy with the objective of the decision which is "Extracting Smart Data", and then the objectives from a broad perspective, through intermediate levels (The three factors already cited: data must be accurate, actionable and agile) to the lowest level (usually a set of Alternatives which is in our case Big Data).
3) Construct a set of pairwise comparison matrices.
4) Use the priorities obtained from the comparisons to weigh the priorities in the level immediately below. In our case the weight represents the level of relevance of the information, Therefore the selected Smart Data are the data which have a high weight.

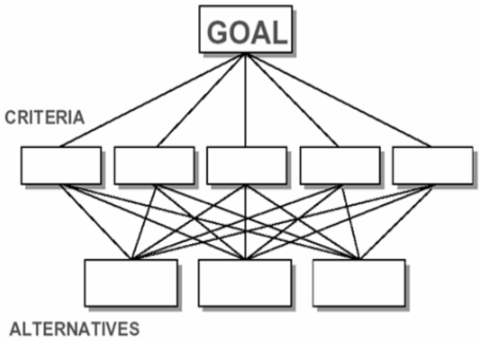

Fig. 2. AHP flow method [9]

IV. An Overview of the Decision-Support Cycle

The decision support cycle in the proposed framework includes four sub-processes (Figure 3) :

• **Data Extraction**. External sources (social networks, forums..etc.) generate heterogeneous data that have an unstructured and semi-structured formats, for that we propose the use of some Apache Hadoop components (Apache Flume,Apache Sqoop and ODBC/JDBC Connectors) that are very useful for ETL, and we use the Hadoop Distributed File System (HDFS) for data storage and MapReduce for preprocessing. Hadoop brings at least two major advantages to traditional ETL [10]:

- Ingest massive amounts of data without specifying a schema on write.
- Offload the transformation of raw data by parallel processing at scale.

On the other hand, we use agents to perform ETL process to the structured data generated by the existing systems like the OLTP (OnLine Transaction Processing).

• **Data Preprocessing**. Data preprocessing is the process of extracting Smart Data. In addition to data parsing, the preprocessing stage is composed of two main sub-processes: redundancy elimination and relevant information extraction in which we use the AHP to extract the Smart Data.

Structured data will be preprocessed in the staging area while semi-structured and unstructured data will be preprocessed with MapReduce in the HDFS. After the preprocessing step, data will be loaded into the Data Warehouse (DW) for analysis.

• **Data Analysis and Visualization**. The last step in the Decision-Support Cycle is the data analysis and visualization. In this step agents perform specific task assigned to them like Data Mining (DM) and Online Analytical Processing (OLAP) in order to visualize the results to support the decision making.

The aimed system is composed of the following set of components:

- External data integration component
- Hadoop Distributed File System (HDFS)
- Staging Area
- Data warehouse
- User-interface component

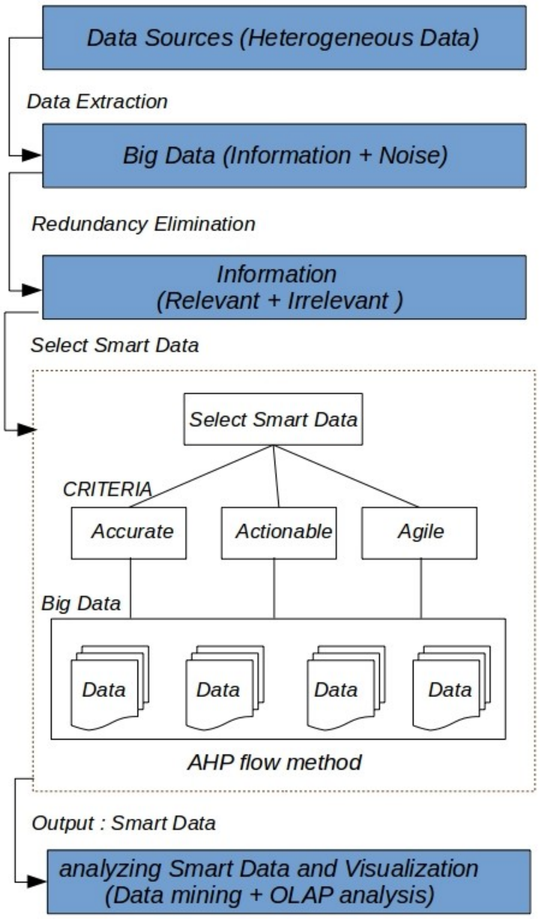

Fig. 3. An Overview of the Decision-Support Cycle

V. The Proposed Framework

A. Description of the Framework

Intelligent agents are utilized nowadays in each field of life to take care of complex issues by distributing the work. Agents are software programs that take the autonomous action in different states to achieve design goals. As indicated by [11], responsive, proactive, autonomous and social are essential attributes of agents. In a multi-agent based system, agents work collectively and each agent performs specific tasks according to the role assigned [12].

Figure 4 illustrates an overview of the proposed framework. The different agents and their roles are listed in the following paragraphs.

- *Extractor and Integrator Multi-Agent Group*
  - **Source Identification Agent (SIA):** SIA identifies the external data extraction sources and evaluates the quality of data. This agent ensures that the extracted data complies with these three main factors [13]:
    1) Provenance factors, refer to the source of information.
    2) Quality factors, concentrate on factors that reflect how an information object fits for use.
    3) Trustworthiness factors, influence how end-users make decisions regarding the trust of information.
  - **Extractor Agent (EA):** EA builds up a connection with the sources system and extracts data.
  - **Loader Agent (LA):** The role of LA is to guarantee effectiveness and consistency to enhance the performance of data warehouse operations and decrease the stacking time [12].

- *Preprocessor Multi-Agent Group*
  Redundancy Elimination Agent and Filtering Irrelevant Information Agent are situated in the staging area and they are responsable for structured data preprocessing, while Driver Agent and Workers Agents are responsable for unstructured and semi-structured data preprocessing in the HDFS.
  - **Redundancy Elimination Agent (REA):** REA is concerned with identifying and eliminating contradictions and inconsistencies. REA removes duplicate, missing and redundancy from data.
  - **Filtering Irrelevant Information Agent (FIIA):** His role is to identify irrelevant information and ignore them, FIIA filter all the noise and only relevant information that are actionable, accurate and agile will be loaded into the Data Warehouse. FIIA applies the AHP method and other techniques like Feature Extraction (FE) and Feature Selection (FS) [14] to filter the irrelevant information.
  - **Driver Agent and Workers Agents (DA and WA):** These agents apply the preprocessing process to unstructured and semi-structured data stored in the HDFS. They have the same role as Redundancy Elimination Agent and Filtering Irrelevant Information Agent. A MapReduce model is implemented, from where the Driver Agent is located in the master node, it is created to execute specific tasks, while the real work is carried out by the workers agents who are located in the slave nodes as illustrated in figure 5.

- **Analyzer Agents (AA):** The purpose of these agents is to convert the amount of data stored in the Data Warehouse into valuable information by applying a fast and efficient analysis and creating various views and representations of that data. The objective is to carry out OLAP and Data mining on behalf of an agent or a user and to report the results to the requesting entity and to all other entities that should be informed.

- **User-Interface Agent (UIA)** The UIA enhances the ability of the system user to use and entirely benefit from the Decision-Support System. It is responsible for all communications between the user of the system and the other agents in order to transmit reports and results to the end user.

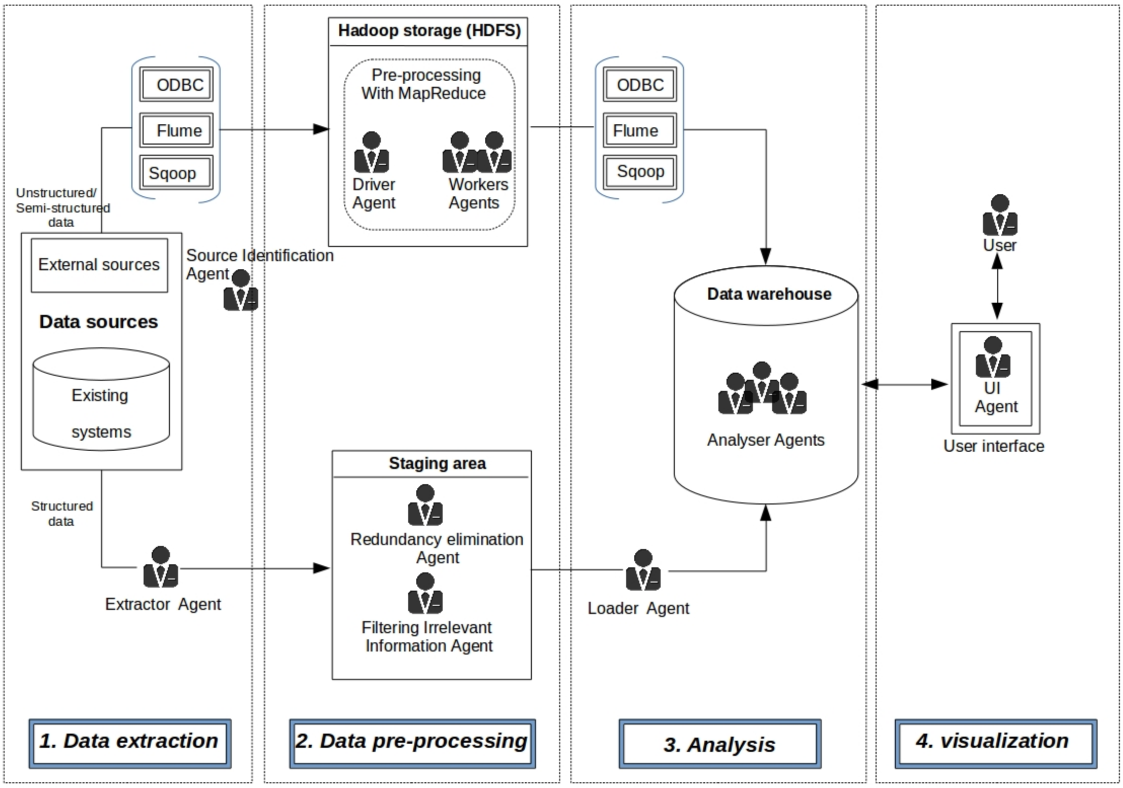

Fig. 4. The proposed framework

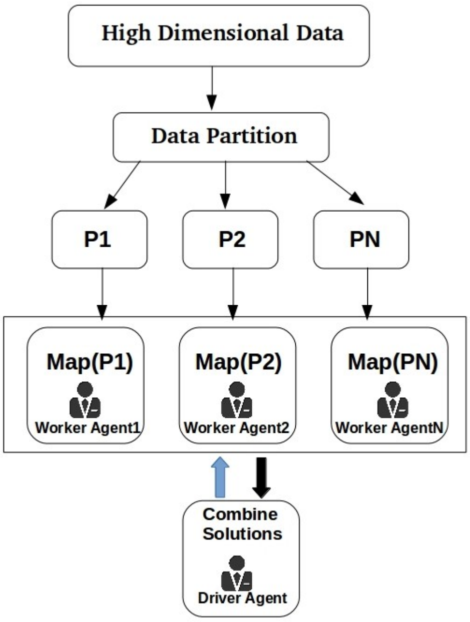

Fig. 5. Driver Agent and Workers Agents

B. Inter-Agent Communication

The sequence diagram based on the AUML language [15] shown in Figure 6 describes the interaction between the agents in the decision support system:

- *Step 1 (Data extraction)*
  1) The user requests data through user interface.
  2) The User-Interface Agent (UIA) receives and redirects the requet to both Source Identification Agent (SIA) for external data extraction (social media data) and Extractor Agent (EA) for internal data extraction.
  3) The Source Identification Agent (SIA) checks for the evaluation methods and techniques to mesure the external data quality. SIA allows data extraction only from sources that meet the quality attributes defined by the company.
  4) The Extractor Agent (EA) extracts data from existing system to the staging area.

- *Step 2 (Data preprocessing)*
  1) Unstructured and semi-structured data will be preprocessed (Data parsing, Redundancy Elimination and Filtering Irrelevant Information) in the HDFS, Driver Agent (DA) and Workers Agents (WA) implement a MapReduce model for preprocessing the data. The resulting Smart Data will be loaded into the Data Warehouse by the use of Apache Hadoop components (Apache Flume,Apache Sqoop and ODBC/JDBC Connectors).
  2) Redundancy Elimination Agent (REA) parse, eliminates redundacy from structured data and notifies the Filtering Irrelevant Information Agent (FIIA) to filter the irrelevant information.

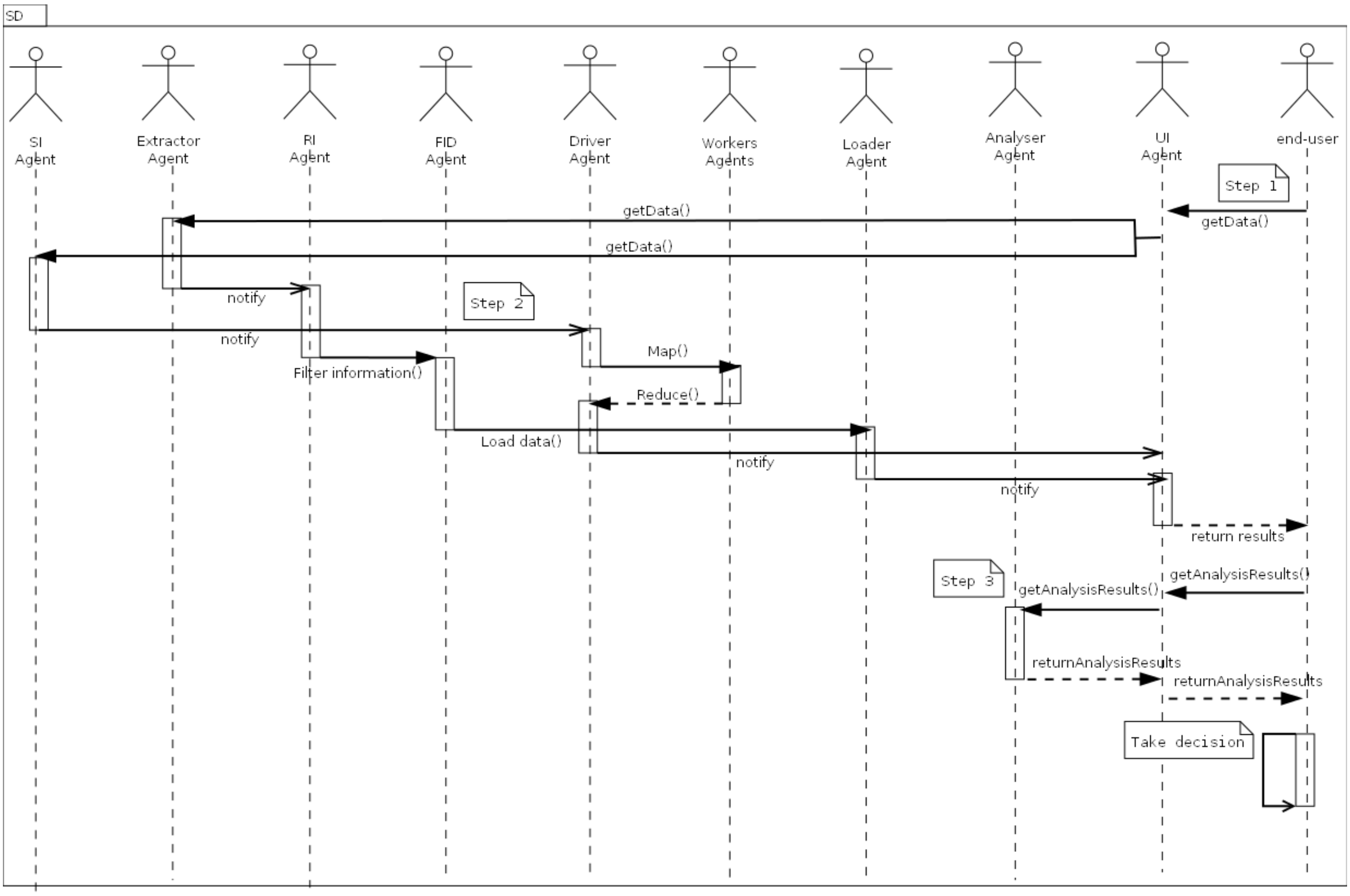

Fig. 6. The communication between agents

3) The Loader Agent (LA) loads the data into the Data Warehouse for analysis.
4) After storing all the data in the Data Warehouse, the User-Interface Agent displays the results to the User.

- *Step 3 (Analysis and Visualisation)*

1) The user request data analysis through user interface.
2) The User-Interface Agent (UIA) receives and redirects the requet to Analyzer Agents Group (OLAP Agents and Data Mining Agents).
3) Analyzer Agents Group (AA) anlyses the data and returns results to the User-Interface Agent (UIA) which displays the results to the user.

- *Step 4 (Decision Making)*

1) The user reviews the results of the analysis and makes decisions.

VI. Discussion

There has already been a lot of work in the area of Big Data and Business Intelligence, in particular with MAS. The authors in [11], [12] propose a conventional BI solution based agents, but the disadvantage in conventional BI solutions is that they use only structured data, in contrast to modern solutions that use Hadoop for processing the unstructured and semi-structured data.

There exist other architectures and systems proposed based agents [16]–[19], which supports the use of Big Data (structured, semi-structured and unstructured data), but most of these works focuse only on the three challenges: 'volume', 'velocity' and 'variety', and neglects the other challenges that are 'value' and 'veracity'.

In this paper, we have tried to propose a BI framework based agents in a Big Data Environment that can handle the 5Vs challenges:

- We have proposed the use of Hadoop for the preprocessing of semi-structured and unstructured data. With the distributed storage and the MapReduce programming model, Hadoop can handle the first 3Vs: volume, velocity and variety.
- For the fourth challenge, which is the veracity of the data, we have proposed a Source Identification Agent whose main role is to ensure some factors [13] like: provenance factors, quality factors and trustworthiness factors.
- The fifth and last challenge is the value. This is the main challenge that we have based on, where we described

how to transform Big Data into Smart Valuable Data.

VII. Conclusion

In this paper, we have tackled the problem of noise in Big Data, which is a crucial step in transforming such raw data into Smart Data. We have proposed a Multi-Agent Framework for Business Intelligence driven Smart Data in a Big Data Environment. In the proposed framework, agents work collectively to perform tasks according to the roles assigned. The system contains different groups of agents to reduce the extraction time, filter the noise and optimize the performance of decision making. Hadoop is used to ensure fast data loading, fast query processing and efficient storage. The highly tolerant nature of Hadoop's failures, flexibility, extensibility, efficient load balancing and platform independence are also useful features for the development of any distributed process.

The solution may be adapted to different contexts, enabling the user to select the relevant data attributes and apply them in a suitable way into a certain situation. Moreover, the implementation of the agent based scenario for analysis purposes in different fields of life can be done.

References

[1] B. Prakash and M. Hanumanthappa, "Issues and challenges in the era of big data insight," *International Journal of Emerging Trends and Technology in Computer Science*, vol. 3, no. 4, pp. 321–325, 2014.

[2] M. Kendrick, *Big Data, Big Challenges, Big Opportunities : 2012 ioug Big Data Strategies Survey*, Oracle, 2012.

[3] N. Wallis, *Big Data in Canada : Challenging Complacency for Competitive Advantage*, 3rd ed., Yahoo press, 2012.

[4] A. Lenk, L. Bonorden, A. Hellmanns, N. Roedder, and S. Jaehnichen, "Towards a taxonomy of standards in smart data," in *2015 IEEE International Conference on Big Data (Big Data)*, 2015.

[5] I. Triguero, J. Maillo, J. Luengo, S. Garcia, and F. Herrera, "From big data to smart data with the k-nearest neighbours algorithm," in *IEEE International Conference on Smart Data (Smart Data 2016)*, 2016.

[6] M. Wu. (2013) The key to insight discovery: Where to look in big data to find insights. [Online]. Available: https://community.lithium.com/t5/Science-of-Social-Blog/The-Key-to-Insight-Discovery-Where-to-Look-in-Big-Data-to-Find/ba-p/70116

[7] ——. (2012) The big data fallacy. [Online]. Available: https://community.lithium.com/t5/Science-of-Social-Blog/The-Big-Data-Fallacy-Data-Information/ba-p/59250

[8] D.Garcia-Gil, J.Luengo, S.Garcia, and F.Herrera, "Enabling smart data: Noise filtering in big data classification," *Computer science*, 2017, arXiv:1704.01770 [cs.DB].

[9] P. Vincke, J. Wiley, and Sons, "Multi criteria decision-aid," *Mathematics Applied in Business nad industry*, 1992.

[10] *Extract, Transform, and Load Big Data with Apache Hadoop**, Intel, 2013, white paper "Big Data Analytics".

[11] A.Bologa and R.Bologa, "Business intelligence using software agent," *Database Systems Journal*, vol. 2, no. 4, pp. 31–42, 2011.

[12] R.Talib, M.K.Hanif, F.Fatima, and S.Ayesha, "A multi-agent framework for data extraction, transformation and loading in data warehouse," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, 2016.

[13] J. Nurse, S.Rahman, S.Creese, M.Goldsmith, and K.Lambert, "Information quality and trustworthiness: A topical state-of-the-art review," in *International Conference on Computer Applications and Network Security*. ieee, 2011, pp. 492–500.

[14] S. Ramirez-Gallego, B. Krawczyk, S. Garcia, M. W. zniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, 2017.

[15] J. Odell, D. Parunak, and B. Bauer, "Representing agent interaction protocols in uml," *Agent-Oriented Software Engeneering*, 2001.

[16] E. Belghache, J. Georgé, and M. Gleizes, "Towards an adaptive multi-agent system for dynamic big data analytics," in *Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress*, 2016.

[17] B. Twardowski and D. Ryzko, "Multi-agent architecture for real-time big data processing," in *International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, 2014.

[18] A. Fazziki, A. Sadiq, J. Ouarzazi, and M. Sadgal, "A multi-agent framework for a hadoop based air quality decision support system," in *Advenced Information Systems Engineering*, 2015.

[19] K. Qayumi and A. Norta, "Business-intelligence mining of large decentralized multimedia datasets with a distributed multi-agent system," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 10, no. 06, 2016.