A Machine Learning and Multi-Agent Model to Automate Big Data Analytics in Smart Cities

Fouad SASSITE, Malika ADDOU, Fatimazahra BARRAMOU
Team (ASYR) - Laboratory of Systems Engineering
Hassania School of Public Works (EHTP)
Casablanca, Morocco

***Abstract*—The objective of this paper is to present an architecture to improve the process of automating big data analytics using a multi-agent system and machine learning techniques, to support the processing of real time big data streams and to enhance the process of decision-making for urban planning and management. With the rapidly evolving information technologies, and their utilization in many areas such as smart cities, social networks, urban management and planning, massive data streams are generated and need an efficient approach to deal with. The proposition in this paper adopts the concept of smart data which focuses on the value aspect from big data. The proposed architecture is composed of three layers: data acquisition and storage, data management and processing and the service layer, based on a multi-agent system to automate the big data analytics; the proposed model describe the functionalities of the system and the collaboration between agents, these autonomous entities receive data streams in real time, they perform operations of preprocessing, big data analytics and storage into a Hadoop cluster. The techniques of machine learning are also used to enhance the process of decision making, such the use of classification algorithms to predict habitat type based on the characteristics of a population to help making efficient urban planning decisions. The proposed system can serve as a platform to support data management and to conduct effective decision-making in smart cities.**

Keywords*—Big data analytics; machine learning; smart data; multi-agent system; automation; decision-making; urban planning

I. INTRODUCTION

With the increasingly rapid evolution of information technologies, and its use in many areas such as smart cities, social networks, online business applications, transportation data, urban management, and planning. These massive data are generated with high velocity and require real-time processing, the emergence of the concept of Big Data in many areas has become a reality that imposes itself on systems based on traditional data management and processing technologies such as relational databases. The amounts of generated data are increasingly unstructured due to the diversity and the heterogeneity of the data sources, some studies [1] are claiming that the use of connected object can reach more than 25 billion in 2020. This rapid growth of generated big data imposes new challenges in terms of storage, data analytics and processing, it requires new approaches to provide more reliable solution to deal with the big data.

Useful information, actionable knowledge and valuable data is normally incorporated into these voluminous raw data, Smart Data is the approach that focuses on the value aspect of the big data and try to exploit these huge masses and discover knowledge from these data [2].

Smart data with a focus on veracity and value has been introduced, The goal is to effectively clean or rectify imperfections in the raw data and put the action on the valuable data, which can be effectively used by businesses and governments for planning, evaluation, control and intelligent decision making [3], [4], in order to turn big data into smart data, some researches [5] focuses on the importance of the preprocessing steps and others [6] proposes to highlight the importance of improving four fields of interest:

- Reliable infrastructures,
- Data Organization and management,
- analyze and prediction,
- Decision support and automatization.

This research is in line with this theme of research, and it tries to give a contribution in this area by proposing a multi-agent system in order to automate the processing and analysis of Big Data and to improve the decision-making process, to better use the collected Big Data and to give a sense of these data.

The use of this paradigm in smart cities constitutes an added value in the process of data management and the exploitation of the most of these data to refine the decision making in real time, for problems that require actions in real time and in an automatic way without human intervention based on autonomous agents dotted with artificial intelligence to solve problems related to smart cities such as the congestion of traffic flows, or for the decision making in the long and medium term such as to predict and anticipate the equipment and the necessary infrastructures for the urban planning.

The applications of this proposition can be in several domains notably in the field of urban management and planning, the use of this system can improve the handling of big data issued from the management of smart cities and help to exploit the amount of data generated and analyze it in an intelligent way to create machine learning models that can guide efficient decision making in these cities.

This paper is structured as follows: Section II will discuss some paradigms related to this research topic. Section III will highlight and discuss some related works. Section IV describes

the proposed approach. Section V will study the conducted experimentations in this work and the final section will conclude and gives some perspectives.

II. Background of the Study

This section will highlight some of the main concepts related to this research work, First, we will highlight the complexity and diversity of the nature of big data from smart cities, and then show the value of smart data which focuses on the value aspect of the big data and on intelligent processing. Next, we will expose the importance of multi-agent systems in the process of big data processing automation in order to automate the process of real time decision making. Finally, we will explain the importance of machine learning techniques and algorithms, to strengthen the cognitive part of the proposed system.

A. Big Data to Smart Data

Big data usually refers to data with large volume of datasets characterized by the complexity and challenges to handle it due the nature of those data and the actual technologies to store and process this type of data [7], [8] the main characteristics in the literature to classify the data as big data:

- Volume, Variety velocity shows the manner in which the data has been generated and the process of storing and processing it.
- Value and Veracity focuses on value aspect, the usefulness, and the quality of data.

In order to automate the process of big data analytics a study of the nature of this data should be done to understand it, basically the big data sources generates three main formats [9] as described in the Fig. 1.

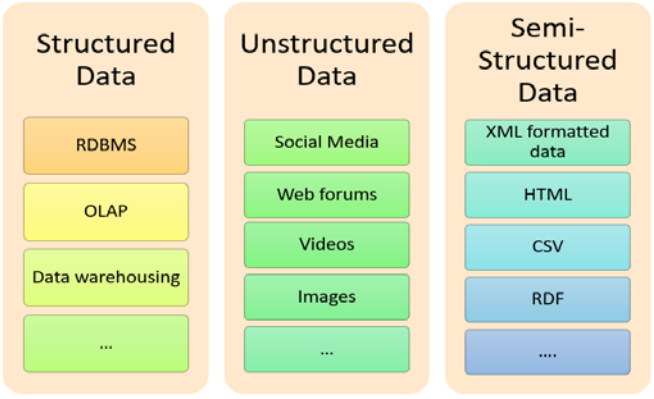

Fig. 1. Taxonomy of Big Data Types.

Smart data (value-based) has been introduced, to highlight valuable data, which can be effectively used by companies and governments for effective planning, monitoring, evaluation, reporting and intelligent decision-making. Three core attributes are necessary for data to be intelligent: it has to be accurate, actionable and agile [10], [11].

A. Multi-agent System

Multi-agent System can be defined as a collection of autonomous entities know as agents [12]. The agents can in a collaborative way solve complex problems [13], [14] they can interact with the environment and the other agents to achieve goals or the complete tasks the agents are characterized with:

- Sociability: To reach their goals they can share their knowledge or request useful information other agents.
- Autonomy: The agents can perform some appropriate actions and executing the process of decision-making.
- Proactivity: The agents can perform effective actions by using their historical data, the data from the sensors or from other agents or their environment.

The real value added of agents can be reached through the use of the collaborative work of each agent to solve complex problems.

In this work the framework JADE is used, [15] which is a multi-agent system platform that provides the infrastructure to deploy agents, this platform assures this functionality with the help of other components like:

- Directory Facilitator: DF is the component responsible for providing yellow pages services in the platform, it allows agents to publish their services by sending requests of registration, and this operation allows other agents to collaborate through the published services.
- Agent Management System: AMS is a necessary component in an agent platform; this module is responsible for managing the agents, and assuring the operations of creation, destruction, migration, etc. of the agents in the platform.
- Message Transport Service: this service is responsible of transporting messages between agents in the platform; these messages meet the standards described in the FIPA-ACL.

B. Machine Learning

Machine learning is field of Artificial intelligence that relates to a wide range of algorithms for making intelligent predictions based on a data set. These data sets are often characterized with large volume [16], Recent advances in machine learning field have achieved what seems like a human standard of semantic understanding and information extraction and some ability to sense abstract patterns with higher accuracy than human experts [17], [18].

The machine learning techniques are nowadays widely used in different domains like the computer vision, prediction, classification, recommendation, semantic analysis, natural language processing and information retrieval [19] [16], [20].

III. Related Work

This section will highlight some research works with relation of this research question.

In the literature, several researches are done to study the processing and management of big data.

With the technological advancements we are experiencing today, the large amounts of data generated are multiplying rapidly, and the process of manipulating these data is

complicated, hence the need to automate the process is justified in several areas like the in industry [21], [22] or in the urban planning and management domain [23].

The use of big data analytics[24] is widely demanded in multiple areas such as security and intrusion detection [25], healthcare [26], The proposition of [27] aims to provide a process composed of the two sub-processes the first for big data management and the second for the big data analytics each sub-process is composed of set of operations described as bellow:

- Big Data Management:
    - Acquisition and recording,
    - Extraction, cleaning, and annotation,
    - Integration, Aggregation and Representation.
- Big data analytics:
    - Modeling and Analysis,
    - Interpretation.

The authors [28], [29] proposed a method based on the generation of workflows of data processing based on a service oriented approach[30], the idea is to divide the totality of the functionalities of the system into a set of services, these services are organized according to four main phases:

- Planning stage,
- Discovery stage,
- Selection stage,
- Execution stage.

The system selects according to the request received a set of services in each phase and then generates a workflow based on the selected services

Other authors [31] proposed a the main steps for extracting value form big data through the definition of four steps:

- Gather data,
- Load data,
- Transform data
- Extract data.

Several works presented in the literature try to give the necessary steps for the system to realize the tasks of analysis and big data processing, but they don't put the action on giving the possibility to the system to have an adaptive behavior according to the nature of the processing and the requests received by the system in an autonomous way, and they need a human intervention to program and manage tasks. The use of the multi-agent paradigm in this sense can enhance the cognitive capabilities of the system of learning efficient ways to process and analyze data and to help in improving the processes of decision-making based on a knowledge base.

IV. Proposed Approach

In order to increase the efficiency of the big data management and analytics process and to implement a new strategy to improve the analysis and processing operations through automation, a model based on three layers architecture and a multi-agent system is proposed.

A. The Proposed Model

The proposed model is mainly based on a multi-agent system and adopts the Smart Data approach that focuses on the value aspect of data in order to optimize the exploitation of the large masses of data stored on big data storage clusters. This model has the objective of automating the processing of big data and improving the efficiency of decision making in this context using machine learning models.

As presented in the Fig. 2, this architecture is based on a collection of autonomous entities to constitute a multi-agent system. The multi-agents are organized according to three-layer architecture:

- The data acquisition and storage layer: This layer is responsible for interacting with the heterogeneous data sources that usually a system handling big data can interact with, for example in the field of smart cities management often the system interacts with data coming from different sources whether it is from sensors or external data sources, these data being of various types and generated at different rates, in batches or in real time streams.

This layer also allows to manage the storage of the generated or collected data platform through the use of an HDFS storage cluster which is distributed over several nodes in order to guarantee high availability and also to distribute the data processing operations.

There is also a knowledge base that can help the agents in making decisions and choosing the necessary processing step for each request received by the system.

- The data management and processing layer: In this layer, there are principally the modules dedicated to data processing and analysis, such as components for pre-processing operations like cleaning, filtering, normalization, transformation, and data reduction, as well as the components for data processing like the entities for data processing and analysis in batches specially for historical data and stream processing in the case when the system deals with real time streams of data.
- The data services layer: The main role of this layer is to communicate the results of the processing performed by the system as a result of a query to the applications and services, the monitoring of dashboards and the production of reports, as well as to provide interfaces that facilitate the interaction of the applications with the proposed system.

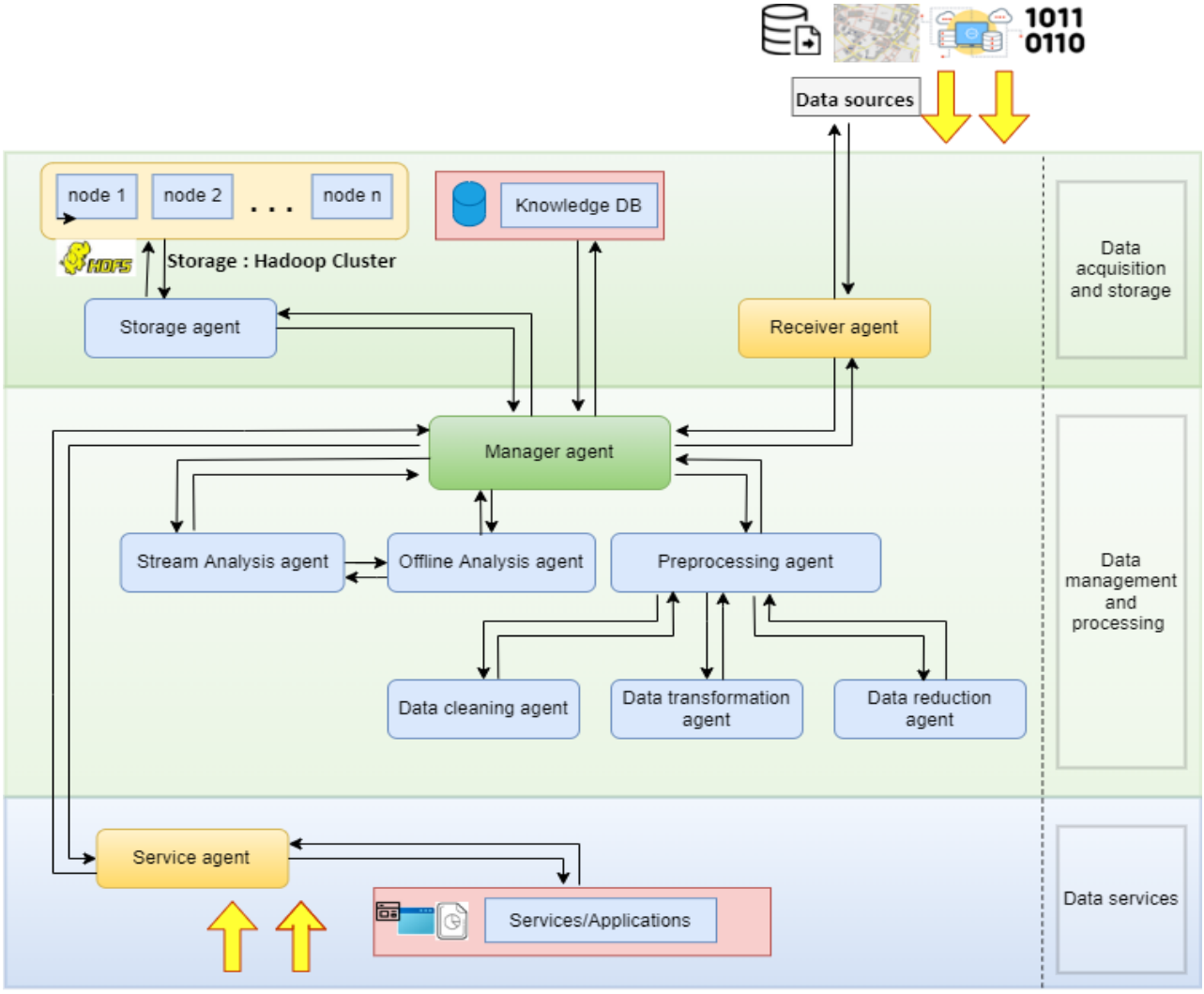

Fig. 2. The Proposed Multi-layer Architecture for the MAS.

Such functionality can help improve and automate the decision-making process, as well as the management and processing of real-time data. The proposed system can serve as a platform to support data management and to conduct effective decision-making in smart cities.

B. Components of the Proposed Multi-agent System

The proposed multi-agent system, through the cooperation of the agents, can automate the processing and the management of the data, in accordance with the three layers of the system: data acquisition, data management and processing, and the services and communication layer.

The system is composed of multiple agents:

- Receiver agent: This agent represents the system's interface with the data sources. The receiving agent handles data of multiple types and from different sources, such as sensors, external databases, web services, etc. and at different rates, streaming or batch.
- Storage Agent: this agent is in charge of all operations related to storage management and data reading from the Hadoop cluster.
- Service Agent: This agent is designed to interact with applications and services and to communicate the processing results through user friendly interfaces and applications. With this agent, the end-users can also schedule or send requests or execute certain tasks.
- Offline Analysis Agent: This agent allows the system to perform operations and tasks related to data batch processing, it can handle the requests of data analysis using voluminous historical data, and usually those data are collected and stored into the cluster storage. Those tasks of processing can help to automate the process of scheduled or per period data analysis tasks.
- Stream Analysis Agent: the main function of this agent is to process data streams that are usually in real time. The use of technologies that are generating data in real time such as sensors and devices that are transmitting requests and data streams in real time. These queries expect responses from the system in real time.

The collaboration between the two agents responsible of processing: the stream Analysis Agent and the Offline Analysis

Agent can resolve complex problems of data processing for example the requests they require to handle data streams in real time and needs to perform operations on the historical data already stored on the system, this collaboration is carried by the Manager Agent.

- Knowledge Base Component: is a database used by the multi-agent system to store rules to facilitate the selection process of the adaptive behavior of the agents based on the type of the request and the data to process.
- HDFS storage Cluster: is the component of the system that supports distributed data storage through the use of many data blocks or partitions. This distributed file system stores the data using several nodes, called data nodes, and managed by the name node which try to balance the load between theses nodes.

The HDFS also assure the replication of the data to guarantee high availability. adopting this type of storage is more appropriate due to requirement of data processing and to the nature of the big data constraints [32], especially when the system deals with fast processing queries with low latency.

- Preprocessing Agent: The preprocessing stage is considered as a primordial phase that help extract to value aspect from the raw big data. The smart Data approach focuses and consider it like a key to prepare the data, since the data generated from real world applications are imperfect and may contains some redundancies, inconsistencies, and noisy values.
- Data Cleaning Agent: This is an agent that checks the received data and imposes a strategy to clean the received dataset. The Smart data aspect focuses on the preprocessing steps and the cleaning phase is an important one with the aim to remove noisy and redundant instances [33] to allow the system to operate automatic learning algorithms on reliable datasets.
- Data Transformation Agent: This agent is designed to change the data format according to the processing needs, to smooth out the training phase by applying data normalization, aggregation, and filtering operations.
- Data Reduction Agent: This agent has an essential contribution in the work of the system by reducing the time of data processing which will impact the positively the process by reducing the time latency, deleting the faulty data and reducing the volume, in this phase there are several techniques in the literature [34] that can be applied with the aim to simplify the datasets dimensionality reduction, discretization and instance reduction.
- Manager Agent: The manager agent has in important role in the proposed system it represents the main component responsible of coordination of the tasks sharing between the other agents.

This entity can take an adoptive behavior based on the nature of the requests received and a knowledge base.

C. Operating Principle

The proposed multi-agent system described previously shows the necessary steps to assure the functioning of the system in different situations.

The system can interact with external applications and data sources and subsequently initiate or execute data analysis operations mainly by triggering two agents, the one responsible for data acquisition or services:

- Receiving a request from a sensor generating a data flow or from an external api or service applies the collaboration of the agents of the system.

Once the request is received by the receiving agent and sends it to the Manager, the latter distributes the tasks to store and process this request. The tasks of pre-processing, online analysis and offline analysis are performed to respond to this request in real time, after that the service agent can transfer the result.

- Receiving a request from the application layer as in the case of scheduling a task or generating a report or a specific request to execute.

The Service agent sends the request to The Manager agent which decides to collect the necessary data from the data cluster or to retrieve it from the data sources; and according to the type of this request, it makes the decision to send this data to one of the data processing agents, either in stream or in offline analysis and returns the result through the service agents.

V. Experimentation and Tests

The proposed approach consists in developing a multi-agent model for the automation of big data processing and analysis by applying the smart data approach.

This section attempts to highlight the adaptive behavior of the proposed system by the study of three cases of use of the system.

The system is composed a set of agents linked to stored algorithms that are defined by the organization's business rules. These agents will be distributed across multiple nodes in order to promote task distribution and load balancing across nodes.

For the implementation of this approach the JADE [15] Framework was adopted.

The communication and the interactions between agents is principally done through the exchange of messages [35] each message is labeled with an act of communication such as: request, inform, notify, propose, etc.

The Fig. 3 represents a snapshot of the behavior of the multi-agent agent system at a time t, where the different communications between the agents of the proposed system can be observed, this communication can be done by sending ACL messages or by sharing information in the environment of the agents, this communication can be in the context of a collaboration, coordination, or negotiation, the agents adopt an adaptive behavior according to the nature of the data and the requests received by the system.

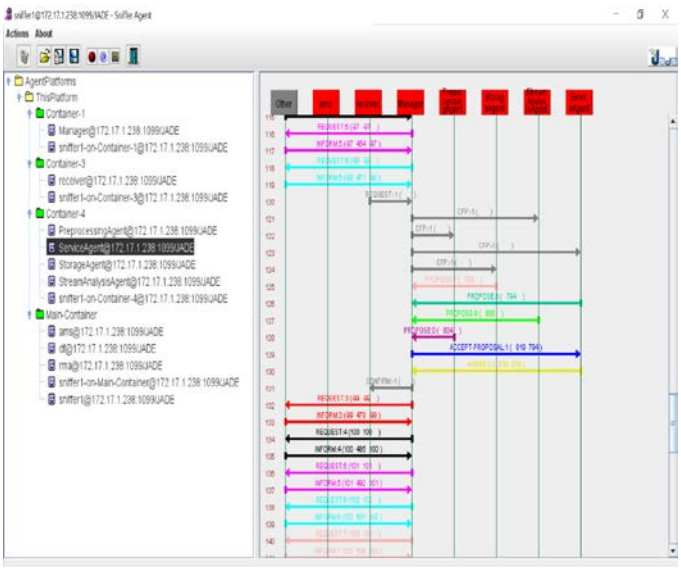

Fig. 3. The Communication between the Agents of the Proposed System.

This part will study the adaptive behavior of the proposed multi-agent system by illustrating three case studies where an explanation of the operating principle of the system and the flow of processing that changes depending on the nature of the request received by the system will be presented, the three scenarios are related to the field of management and planning of smart cities.

The first is to study the behavior of the system using a large dataset that contains data from the last census in Morocco for the city of Casablanca. The used data contain information and characteristics about the habitat and individuals in order to build a machine-learning model to predict the type of habitat according to the evolution of individuals. Thus, predict the equipment and infrastructure necessary for the urban planning of this city.

The second scenario is to show the behavior of the system and its use for listening to changes and storing data from different data sources.

And the third scenario is to use the system for real-time prediction tasks using the machine learning model already created in the first scenario.

A. *Case 1: Training a Machine Learning Model from a DataSet Stored in HDFS*

In this case the request issuing from the services and applications layer which consists in asking the system to launch a learning operation from a dataset stored in the Hadoop storage cluster, the Fig. 4 explain the explains the information exchange between the different components of the system in this case.

An explanation of the steps will be attempted using a sequence diagram illustrated in Fig. 5.

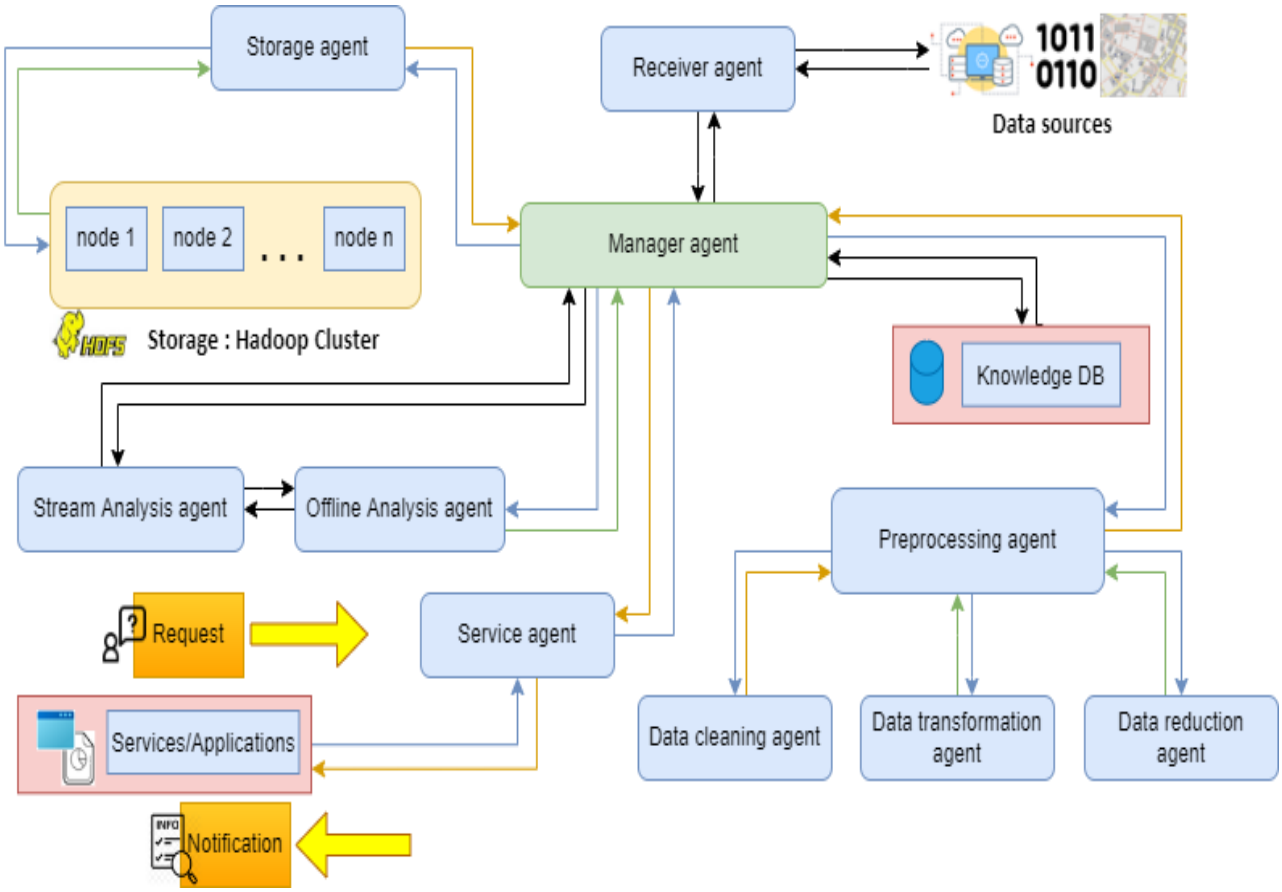

Fig. 4. The Flowchart of Training a ML Model using the Proposed MAS.

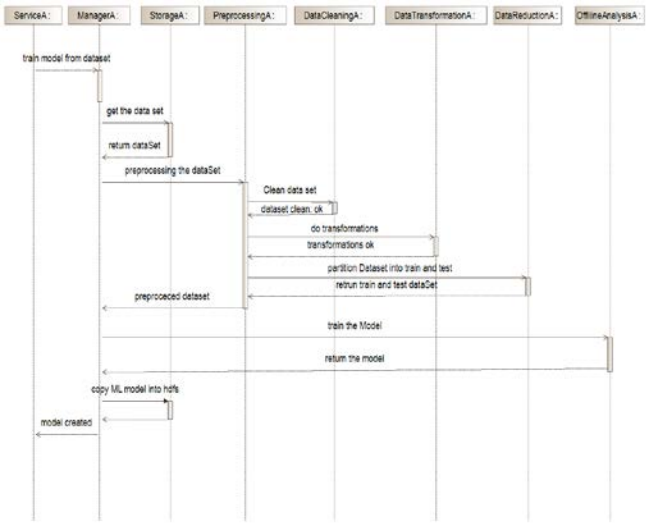

Fig. 5. Sequence Diagram of Training a ML Model using the Proposed MAS.

The service agent receives the request and sends it to the agent manager who is in charge of distributing the operations between agents, it sends the request to retrieve the Dataset to the storage agent, then it sends the Dataset for preprocessing operations by the Preprocessing Agent, which delegates tasks as necessary to the transformation, reduction and data cleaning agents.

Then the partitioned datasets are ready for the phases of training and testing by the Offline Analysis agent, once the model is valid the Manager agent sends the request to the storage agent to copy it on the storage cluster and notifies the service agent of the creation of the requested model.

This generated model will be used to serve prediction requests that the system can receive and respond to in real time.

B. *Case 2: Collecting Data from a Data Source and Storing it in HDFS*

The proposed system can interact with several data sources such as databases, sensors, connected objects-IOT and external applications.

In this case the system is used to receive data from an external source which is an API[36] that provides weather data, the system opted to collect and store its data for the city of Casablanca in the HDFS file system.

In this example the agent will send request to the API in every specified lapse of time to get and store the last data of the weather about the city of Casablanca, the Fig. 6 represents a snapshot of a request of an example of a request sent by the system to get weather data.

```
07:01:09.568 [IPC Client (1827453825) connection to localhost/127.0.0.1:9000 from user] DEBUG org.apache.hadoop.ipc.Client - IPC Cli
{"consolidated_weather":[{"id":5601851798978560,"weather_state_name":"Light Rain","weather_state_abbr":"lr","wind_direction_compass"
>>>>> Get data from datasources
_____2021/12/24 07:01:17
07:01:17.512 [receiver] DEBUG org.apache.hadoop.fs.FileSystem - Looking for FS supporting hdfs
```

Fig. 6. Agent's request to the API.

{"consolidated_weather":[{"id":4697511437533184,"weather_state_name":"Clear","weather_state_abbr":"c","wind_direction_compass":"SE"
,"created":"2021-12-29T22:47:01.9628522","applicable_date":"2021-12-30","min_temp":13.155000000000001,"max_temp":23.865000000000002
,"the_temp":24.025,"wind_speed":3.894206296211459,"wind_direction":129.15947876848557,"air_pressure":1022,"humidity":56,"visibility"
:14.422646743020758,"predictability":68},{"id":6511479063838720,"weather_state_name":"Clear","weather_state_abbr":"c"
,"wind_direction_compass":"ENE","created":"2021-12-29T22:47:04.458998Z","applicable_date":"2021-12-31","min_temp":14.275,"max_temp":21
.915,"the_temp":22.22,"wind_speed":3.203928784289085,"wind_direction":70.16246662280334,"air_pressure":1023.5,"humidity":53
,"visibility":14.444705420345183,"predictability":68},{"id":6165433112592384,"weather_state_name":"Light Cloud","weather_state_abbr"
:"lc","wind_direction_compass":"NE","created":"2021-12-29T22:47:07.5564782","applicable_date":"2022-01-01","min_temp":13.885
,"max_temp":21.935,"the_temp":21.880000000000003,"wind_speed":3.0848885396802674,"wind_direction":38.90329055262177,"air_pressure"
:1026,"humidity":55,"visibility":14.088892587290224,"predictability":70},{"id":6339998865360896,"weather_state_name":"Clear"
,"weather_state_abbr":"c","wind_direction_compass":"NE","created":"2021-12-29T22:47:10.4784872","applicable_date":"2022-01-02"
,"min_temp":13.245000000000001,"max_temp":20.72,"the_temp":19.795,"wind_speed":3.439518859979745,"wind_direction":47.80552040716503
,"air_pressure":1027,"humidity":65,"visibility":13.141379344627376,"predictability":68},{"id":6676595793199104,"weather_state_name"
:"Clear","weather_state_abbr":"c","wind_direction_compass":"ENE","created":"2021-12-29T22:47:13.4515002","applicable_date":"2022-01
-03","min_temp":12.05,"max_temp":20.39,"the_temp":20.16,"wind_speed":3.2158894058697207,"wind_direction":62,"air_pressure":1025
,"humidity":62,"visibility":9.999726596675416,"predictability":68},{"id":5358119593967616,"weather_state_name":"Light Cloud"
,"weather_state_abbr":"lc","wind_direction_compass":"WSW","created":"2021-12-29T22:47:16.4708572","applicable_date":"2022-01-04"
,"min_temp":11.5,"max_temp":20.689999999999998,"the_temp":20.11,"wind_speed":5.363079018531774,"wind_direction":255,"air_pressure"
:1021,"humidity":59,"visibility":9.999726596675416,"predictability":70}],"time":"2021-12-30T01:06:30.0126112","sun_rise":"2021-12
-30T07:34:27.3686192","sun_set":"2021-12-30T17:31:49.8254642","timezone_name":"UTC","parent":{"title":"Morocco","location_type"
:"Country","woeid":23424893,"latt_long":"31.434200,-6.402450"},"sources":[{"title":"BBC","slug":"bbc","url":"http://www.bbc.co.uk
/weather/","crawl_rate":360},{"title":"Forecast.io","slug":"forecast-io","url":"http://forecast.io/","crawl_rate":480},{"title":"Met
Office","slug":"met-office","url":"http://www.metoffice.gov.uk/","crawl_rate":180},{"title":"OpenWeatherMap","slug":"openweathermap"
,"url":"http://openweathermap.org/","crawl_rate":360},{"title":"Weather Underground","slug":"wunderground","url":"https://www
.wunderground.com/?apiref=fc30dc3cd224e19b","crawl_rate":720},{"title":"World Weather Online","slug":"world-weather-online","url"
:"http://www.worldweatheronline.com/","crawl_rate":360}],"title":"Casablanca","location_type":"City","woeid":1532755,"latt_long":"33

Fig. 7. The Response of the API formatted as Json Object.

The Receiver agent can listen to changes in a data source or collect values from a data source according to the desired behavior e.g. in this case the Receiver agent adopts a TickBihaviour for each time lapse it retrieves the data and sends the request to the Manager agent to initiate the request for preprocessing to the Preprocessing agent, once the data is cleaned up and ready to be stored the Manager agent asks the Storage agent to copy it to the storage cluster. Fig. 7 shows an example of stored data: Name of the location, time of the location, name of the time zone in which the location is located, internal identifier of the forecast, date to which the forecast or observation refers, compass point of the wind direction, one or two letter abbreviation of the weather condition.

Fig. 8 tries to illustrate the functioning of the system using a diagram sequence.

C. *Case 3: The use of a Machine Learning Model for Real-Time Prediction*

In this case of study, the behavior of the system will be analyzed in the case where it must process prediction requests in real time.

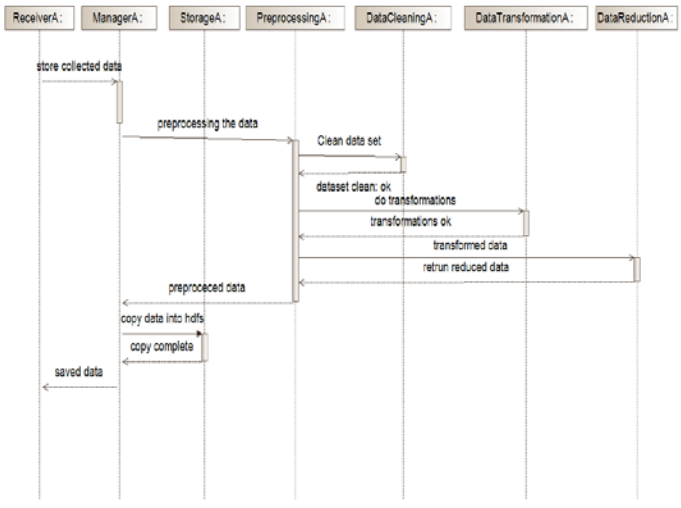

Fig. 8. Sequence Diagram of Collecting Data from a Data Source and Storing it in HDFS.

The first step is to build the model machine learning that will be used for future predictions, and to store it in the cluster, in this step a dataset will be used which contains data about households and individuals in Casablanca from the last general census of population and housing from HCP[37] this data set is made public in 2019, we conducted based on the analyze of this data a previous work to help in enhancing urban planning in the region of Casablanca by predicting the type of habitat and estimate the necessary equipment [23].

1) Training and test machine learning model: In the beginning of this learning phase the dataset used is already prepared cleaned and preprocessed, the start will be splitting the main dataset into two datasets usually in the literature the split is made into two part a larger one with 80% used to train the model and the part with 20% used to test the model, the Fig. 9 illustrate the steps of training the machine learning model used in this study.

The step of choosing the adequate algorithm is data-centric step based on the type of the data and the nature of the problem, in this study this dataset will be used to predict the type of habitat based on the characteristics of the studied population, it's a classification problem, in the next part a test of the main machine learning algorithms for supervised learning on the dataset will be performed.

We will present in this part the metrics and the obtained results for each used machine learning algorithm on the same dataset:

To evaluate the relative performance of different classification models, it is necessary to specify the appropriate quantitative criteria for the evaluation. The performance metrics are specified for the classifier:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Nagative} \quad (2)$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

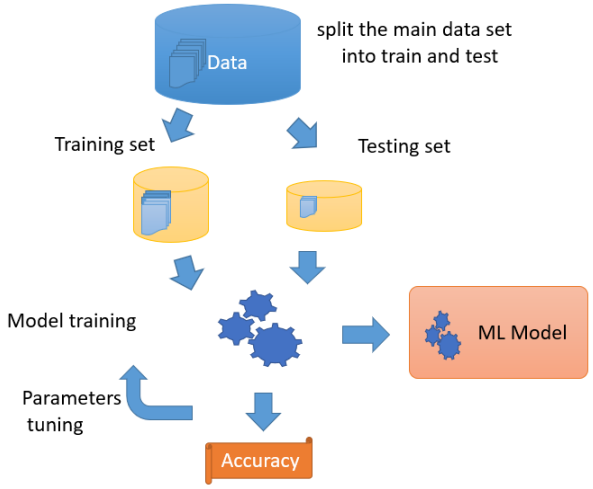

Fig. 9. Steps of Training ML Model using Classification Algorithms.

This section highlights the evaluation of the proposed model on the dataset for each classification algorithm.

a) Naïve bayes: A Naive Bayes classifier is a probabilistic machine learning model used in classification cases to discriminate different objects and labels based on a set of features, this classifier is based on the bayes theorem [38], with the naïve hypothesis of a conditional independence of each combination of features given the value of the class variable, the Table I presents the evaluation metric for the naïve bayes algorithm:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1,\dots,x_n|y)}{P(y|x_1,\dots,x_n)} \quad (4)$$

TABLE I. CLASSIFICATION REPORT FOR THE ML ALGORITHM: NAÏVE BAYES

METRIC	SCORE (%)
precision	64
recall	61
f1-score	61

b) Multilayer perceptron: The MLP [39] multilayer perceptron is a type of artificial neural network organized as a set of multilayers, the process of information flowing from the input layer to the output layer only. It is therefore a feedforward network, each layer consists of a variable number of neurons, the neurons of the last layer being the outputs of the global system, The MLP are widely used in classification, prediction and recognition use cases, the Table II presents the metrics for the use of the multilayer perceptron classifier:

TABLE II. CLASSIFICATION REPORT FOR THE ML ALGORITHM MULTI LAYER PERCEPTRON

METRIC	SCORE (%)
precision	73
recall	73
f1-score	71

c) Random Forest: Random forest is a classifier that can be defined as a meta estimator which corresponds to a number of decision tree classifiers, this algorithm performs a splitting of the training data into a specific number of data subsets used to perform training on multiple decision trees, the subsets used for training the model are slightly different, and then a voting method is used to select the best model [40], the Table III shows the metrics for the use of the Random forest algorithm:

TABLE III. CLASSIFICATION REPORT FOR THE ML ALGORITHM RANDOM FOREST

METRIC	SCORE (%)
precision	85
recall	85
f1-score	84

d) Summary: The Table IV summarizes the evaluation metrics and give a comparison for the used classification algorithms.

TABLE IV. SUMMARY OF THE METRICS USING SUPERVISED LEARNING

METRIC	Precision	Recall	F1-Score	Accuracy
Naïve Bayes	0.64	0.61	0.61	0.61
MLP	0.73	0.73	0.71	0.73
Random Forest	0.85	0.85	0.84	0.85

To select the model to be deployed and used for future prediction requests some metrics are used like precision, recall, f1-score, and the accuracy.

Based on the results obtained, the most accurate model can be choose, in this case the model trained with Random Forest Algorithm, the next step will be to the step of serializing this model into the storage of the proposed system.

Serialization of the trained ML Model: In order to use the trained ML models in the different applications, the step of serialization still necessary, especially when the trained model is a result of huge dataset and machine learning algorithm that consumed a lot of resources to get to the validation phase of the model.

The reuse of the stored model still a good strategy that allows the system to use immediately the model, as explained in the Fig. 10.

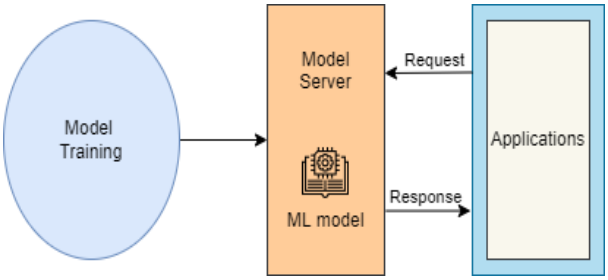

Fig. 10. Serialization and use of the Machine Learning Model.

The use of the trained model by the MAS system:

The proposed system can handle real-time prediction requests using the principle of collaboration between different agents, as shown in the Fig. 11 first the system get the request of prediction from the application and service layer this request is received with the service agent which will transmit it to the agent manager, the latter will call the storage agent to retrieve the serialized model in order to use it for this request and send the request for preprocessing to format it and to get the parameters or the datasets in this request, once done the operation of prediction by the stream analysis agent, this result will be sent to the service agent to bring up the response to the demanding application.

The system can handle the communication with different applications by defining an interface for communication with the proposed system,

And defining multiple functionalities to download, upload, store or analyze data by defining several functions.

For example, the receiver agent can receive requests of type:

$Function_i(Arg_1, Arg_2, \ldots, Arg_n) => D_i$

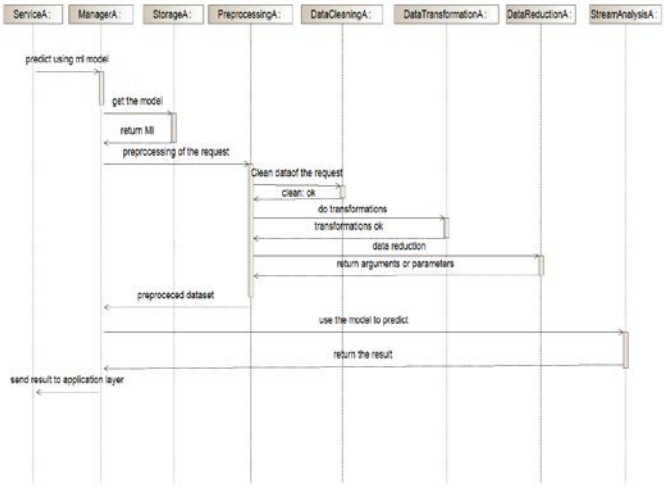

Fig. 11. Sequence Diagram of using a Machine Learning Model for Real-time Prediction.

These functions and services could be used in other applications like the one in the Fig. 12, which will interface with the proposed system and use these features.

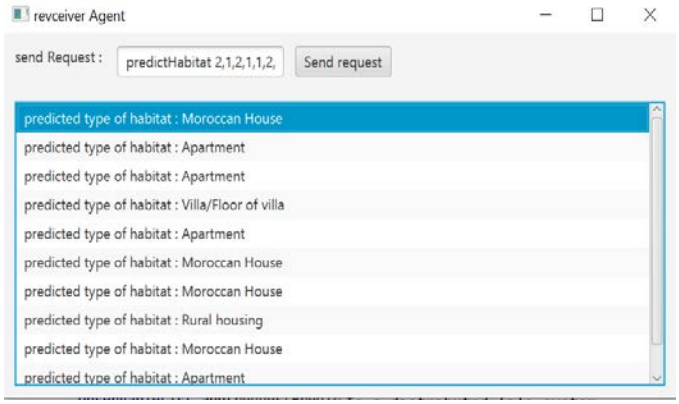

Fig. 12. Example of an Application Consuming the Prediction Service of the Proposed System.

VI. CONCLUSION AND PERSPECTIVES

In this paper, a multi-agent system to automate big data analytics in real time based on the smart data approach was proposed, a model with a multilayer architecture was presented, and the adaptive behavior of this system was illustrated through some case studies, especially in the case of training and the using machine learning models.

The use of intelligent processing, automating big data analytics through the use of the multi-agent system and machine learning techniques can help to solve complex problems such in the field of Urban planning and management in smart cities, by providing actionable data and efficient decisions.

To summarize, the paper tries to highlight the challenges following the processing of big data specifically issued from smart cities, these data are in different formats (structured, semi-structured, unstructured). In addition, they are generated at different rates by batch or in real time in streams. These voluminous amounts of data present many challenges including

the management of these data, the latency of processing time and mainly the effective exploitation of these data to extract useful information.

To overcome these problematic, the paper proposes a model based on a multi-agent system that adopts the smart data approach which focuses on the value aspect of data and the application of intelligent processing in order to extract useful information from these voluminous masses of big data and to drive an efficient decision making, the proposed system automates its data processing processes, and thanks to the autonomous and cognitive capabilities of the agents they can manage and make decisions in real time, And to illustrate the adaptive behavior of the system through the autonomous behavior of the agents, we proposed 3 scenarios:

- The system builds a machine learning model of the classification from a large dataset that contains the demographic data of the city of Casablanca from the last census in Morocco in 2019, to predict the type of housing adequate to a profile of household or citizen for this several tasks were performed by the agents including the preprocessing phase and the choice of appropriate classification algorithm according to the study of metrics and then the serialization of the model, in order to use it for future requests for prediction in real time.
- The system listens to the various data sources and stores the data in the Hadoop storage cluster in real time.
- The system responds to real-time prediction demands by using machine learning models already built and stored in the system to improve the real-time decision-making process.

Such a proposition can find applications in several areas including the management and planning of smart cities known with the generation of huge amounts of data, this proposition can contribute in the effective use of these data and the extraction of valuable information that can guide effective decision-making to improve for example the management of traffic flows and road traffic, the anticipation of needs in terms of equipment in smart cities, reduce energy consumption, the monitoring of KPIs, the reporting and monitoring of smart cities.

In perspective, we will work on enhancing the system with other functionalities, focusing on the improvement of agent behaviors and the machine learning models. To test this approach, solve several problems in urban planning and management in smart cities.

The proposed approach in this paper represents a real opportunity to enhance the process of big data analytics by the exploit of the cognitive functionalities of the agents and to empower their capabilities using a knowledge base which will facilitates the process of decision- making made by agents to take actions in a autonomous way, such adaptive behavior can solve serious problems in smart cities like energy consumption and traffic congestion.

REFERENCES

[1] Lee, “Big data: Dimensions, evolution, impacts, and challenges,” Business Horizons, vol. 60, no. 3, pp. 293–303, May 2017, doi: 10.1016/j.bushor.2017.01.004.

[2] F. Iafrate, “A Journey from Big Data to Smart Data,” in Digital Enterprise Design & Management, 2014, pp. 25–33.

[3] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, “Smart Data,” in Big Data Preprocessing, Cham: Springer International Publishing, 2020, pp. 45–51.

[4] F. Sassite, M. Addou, and F. Barramou, “A smart data approach for Spatial Big Data analytics,” in 2020 IEEE International conference of Moroccan Geomatics (Morgeo), May 2020, pp. 1–6, doi: 10.1109/Morgeo49228.2020.9121920.

[5] S. García, J. Luengo, and F. Herrera, “Tutorial on practical tips of the most influential data preprocessing algorithms in data mining,” Knowledge-Based Systems, vol. 98, pp. 1–29, Apr. 2016, doi: 10.1016/j.knosys.2015.12.006.

[6] A. Lenk, “Smart Data: Von Technologien zu Standards,” 2015.

[7] D. Agrawal, S. Das, and A. El Abbadi, “Big data and cloud computing: current state and future opportunities,” in Proceedings of the 14th International Conference on Extending Database Technology - EDBT/ICDT ’11, Uppsala, Sweden, 2011, p. 530, doi: 10.1145/1951365.1951432.

[8] B. R. Prasad and S. Agarwal, “Comparative Study of Big Data Computing and Storage Tools: A Review,” International Journal of Database Theory and Application, vol. 9, no. 1, pp. 45–66, Jan. 2016, doi: 10.14257/ijdta.2016.9.1.05.

[9] Y. K. Gupta and C. Jha, “A review on the study of big data with comparison of various storage and computing tools and their relative capabilities,” International Journal of Invocation in engineering & technology (IJIET), vol. 7, no. 1, pp. 470–477, 2016.

[10] D. García-Gil, J. Luengo, S. García, and F. Herrera, “Enabling Smart Data: Noise filtering in Big Data classification,” Information Sciences, vol. 479, pp. 135–152, Apr. 2019, doi: 10.1016/j.ins.2018.12.002.

[11] F. SASSITE, M. ADDOU, and F. BARRAMOU, “A Smart Data Approach for Automatic Data Analysis,” in ESAI’19: 1St international conference on embedded systems and artificial intelligence, 2019.

[12] A. Dorri, S. S. Kanhere, and R. Jurdak, “Multi-Agent Systems: A Survey,” IEEE Access, vol. 6, pp. 28573–28593, 2018, doi: 10.1109/ACCESS.2018.2831228.

[13] N. Benmoussa, M. Fakhouri Amr, S. Ahriz, K. Mansouri, and E. Illoussamen, “Outlining a Model of an Intelligent Decision Support System Based on Multi Agents,” Engineering, Technology & Applied Science Research, vol. 8, no. 3, pp. 2937–2942, Jun. 2018, doi: 10.48084/etasr.1936.

[14] M. Naserian, A. Ramazani, A. Khaki-Sedigh, and A. Moarefianpour, “Fast terminal sliding mode control for a nonlinear multi-agent robot system with disturbance,” Systems Science & Control Engineering, vol. 8, no. 1, pp. 328–338, Jan. 2020, doi: 10.1080/21642583.2020.1764408.

[15] F. L. Bellifemine, G. Caire, and D. Greenwood, Developing multi-agent systems with JADE. 2010.

[16] P. P. Shinde and S. Shah, “A Review of Machine Learning and Deep Learning Applications,” in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Aug. 2018, pp. 1–6, doi: 10.1109/ICCUBEA.2018.8697857.

[17] J. A. Nichols, H. W. Herbert Chan, and M. A. B. Baker, “Machine learning: applications of artificial intelligence to imaging and diagnosis,” Biophysical Reviews, vol. 11, no. 1, pp. 111–118, Feb. 2019, doi: 10.1007/s12551-018-0449-9.

[18] N. C. Eli-Chukwu, “Applications of Artificial Intelligence in Agriculture: A Review,” Engineering, Technology & Applied Science Research, vol. 9, no. 4, pp. 4377–4383, Aug. 2019, doi: 10.48084/etasr.2756.

[19] M. Anwer, S. M. Khan, M. U. Farooq, and W. Waseemullah, “Attack Detection in IoT using Machine Learning,” Engineering, Technology & Applied Science Research, vol. 11, no. 3, pp. 7273–7278, Jun. 2021, doi: 10.48084/etasr.4202.

[20] D. Xia, P. Chen, B. Wang, J. Zhang, and C. Xie, "Insect Detection and Classification Based on an Improved Convolutional Neural Network," Sensors, vol. 18, no. 12, p. 4169, Nov. 2018, doi: 10.3390/s18124169.

[21] P. Ghadimi, C. Wang, M. K. Lim, and C. Heavey, "Intelligent sustainable supplier selection using multi-agent technology: Theory and application for Industry 4.0 supply chains," Computers & Industrial Engineering, vol. 127, pp. 588–600, Jan. 2019, doi: 10.1016/j.cie.2018.10.050.

[22] E. Erturk and K. Jyoti, "Perspectives on a Big Data Application: What Database Engineers and IT Students Need to Know," Engineering, Technology & Applied Science Research, vol. 5, no. 5, pp. 850–853, Oct. 2015, doi: 10.48084/etasr.592.

[23] F. Sassite, M. Addou, and F. Barramou, "Towards a Multi-agents Model for Automatic Big Data Processing to Support Urban Planning," in Geospatial Intelligence, F. Barramou, E. H. El Brirchi, K. Mansouri, and Y. Dehbi, Eds. Cham: Springer International Publishing, 2022, pp. 3–17.

[24] G. Lombardo, P. Fornacciari, M. Mordonini, M. Tomaiuolo, and A. Poggi, "A Multi-Agent Architecture for Data Analysis," Future Internet, vol. 11, no. 2, p. 49, Feb. 2019, doi: 10.3390/fi11020049.

[25] K. Dounya, K. Okba, S. Hamza, and B. Omar, "Design and Implementation of a New Approach using Multi-Agent System for Security in Big Data," International Journal of Software Engineering and Its Applications, vol. 11, no. 9, pp. 1–14, Sep. 2017, doi: 10.14257/ijseia.2017.11.9.01.

[26] C. Yao, S. Wu, Z. Liu, and P. Li, "A deep learning model for predicting chemical composition of gallstones with big data in medical Internet of Things," Future Generation Computer Systems, vol. 94, pp. 140–147, May 2019, doi: 10.1016/j.future.2018.11.011.

[27] A. Shashwat and D. Kumar, "A service identification model for service oriented architecture," in 2017 3rd International Conference on Computational Intelligence Communication Technology (CICT), Feb. 2017, pp. 1–5, doi: 10.1109/CIACT.2017.7977299.

[28] T. H. Akila, S. Siriweera, I. Paik, and B. T. G. S. Kumara, "Onotology-based service discovery for intelligent Big Data analytics," in 2015 IEEE 7th International Conference on Awareness Science and Technology (iCAST), Sep. 2015, pp. 66–71, doi: 10.1109/ICAwST.2015.7314022.

[29] B. T. G. S. Kumara, I. Paik, J. Zhang, T. H. A. S. Siriweera, and K. R. C. Koswatte, "Ontology-Based Workflow Generation for Intelligent Big Data Analytics," in 2015 IEEE International Conference on Web Services, New York, NY, USA, Jun. 2015, pp. 495–502, doi: 10.1109/ICWS.2015.72.

[30] I. Paik, W. Chen, and M. N. Huhns, "A Scalable Architecture for Automatic Service Composition," IEEE Transactions on Services Computing, vol. 7, no. 1, pp. 82–95, Jan. 2014, doi: 10.1109/TSC.2012.33.

[31] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," International Journal of Information Management, vol. 35, no. 2, pp. 137–144, Apr. 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[32] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, "Big Data: Technologies and Tools," in Big Data Preprocessing, Cham: Springer International Publishing, 2020, pp. 15–43.

[33] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," Big Data Analytics, vol. 1, no. 1, p. 9, Dec. 2016, doi: 10.1186/s41044-016-0014-0.

[34] D. Pyle, Data preparation for data mining. morgan kaufmann, 1999.

[35] M. K. Eddy, A. Ahmad, and A. Y. C. Tang, "Agents of Things (AOT): Utilizing JADE Agent Technology as Communication Middleware for Vehicle Monitoring System," International Journal of Future Generation Communication and Networking, vol. 11, no. 1, pp. 47–55, Jan. 2018, doi: 10.14257/ijfgcn.2018.11.1.05.

[36] "API - MetaWeather." https://www.metaweather.com/api/location/1532755/ (accessed Dec. 30, 2021).

[37] "RGPH 2014 | Téléchargements | Site institutionnel du Haut-Commissariat au Plan du Royaume du Maroc," Nov. 28, 2020. https://www.hcp.ma/downloads/RGPH-2014_t17441.html (accessed Nov. 28, 2020).

[38] H. Zhang, "The Optimality of Naïve Bayes," in In FLAIRS2004 conference, 2004.D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv:1412.6980 [cs], Jan. 2017, Accessed: Mar. 21, 2022. [Online]. Available: http://arxiv.org/abs/1412.6980.

[39] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review," Frontiers in Aging Neuroscience, vol. 9, 2017, doi: 10.3389/fnagi.2017.00329.