# Semantic Integration in Big Data: State-of-the-Art

Zaoui Sayah[1,*], Okba Kazar[2] and Ahmed Ghenabzia[1]

[1]*LINATI Laboratory, Department of Computer Science and New Technologies, KASDI Merbah University of Ouargla, Algeria*
[2]*Intelligent Computer Science Laboratory, Computer Science Department, University Mohamed Khider of Biskra, Algeria*
*E-mail: sayahzao@gmail.com; kazarokba@gmail.com; a_ghenabzia@esi.dz*
[*]*Corresponding Author*

Received 26 November 2018; Accepted 04 February 2020;
Publication 29 February 2020

## Abstract

Nowadays, web users and systems continually overload the web with an exponential generation of a massive amount of data. This leads to making big data more important in several domains such as social networks, internet of things, health care, E-commerce, aviation safety, etc. The use of big data has become increasingly crucial for companies due to the significant evolution of information providers and users on the web. However, big data remain meaningless without semantics. In order to get a good comprehension of big data, we raise questions about how big data and semantic are related to each other and how semantic may help. To overcome this problem, researchers devote considerable time to the integration of ontology in big data to ensure reliable interoperability between systems in order to make big data more useful, readable and exploitable. This technology can hide the heterogeneity of different data resources. Moreover, in given domains, users can exchange knowledge without caring to choose the suitable semantic that makes their content more expressive. This paper aims to provide a comprehensive overview for readers about big data and the appropriate tools to manipulate and analyse them such as Hadoop. Afterwards, we talk about ontology and how it can be used to improve big data management and analyses for decision makers.

*Journal of Mobile Multimedia, Vol. 15_3,* 191–238.
doi: 10.13052/jmm1550-4646.1533

Finally, different semantic integration approaches are seen in a comparative study. This survey is concluded with a discussion and some perspectives.

**Keywords:** Big data, Hadoop, interoperability, ontology, semantic integration.

## 1 Introduction

Over the past two decades, people and systems have been generating a colossal amount of data from heterogeneous sources and overloading the web with a massive exponential volume of data on a daily basis. The digital data are produced continuously from millions of devices and applications (social networks, smart-phones, sensors, logs...). For instance, Google processes data of hundreds of Petabytes (PB), Facebook also generates more than 10 PB's of data per month [1]. In 2013, (IDC) reports that the total of digital data created was estimated at 4.4 zettabytes (ZB) [2] and it will, in 2020, reach around 40 ZB's [3]. The notion of big data appeared for the first time at the Association for Computing Machinery (ACM) in 1997 [4]. Gartner defines big data as: "big data is a large information resource, high-velocity and/or high-quality that requires new forms of processing to improve decision-making, discovery and the optimisation process" [5]. In recent years, the majority of companies are found emerging in the big data paradigm such as EMC, Oracle, IBM, Microsoft, Google, Amazon, and Facebook, etc.

The aspect of large data volume appears as a challenge to be manipulated by conventional data processing tools and with current means, which has led to a search for new methods and technologies capable of handling this large amount of data in a reasonable time period. According to [1], it is not only the volume aspect that characterizes big data [6]. But he gave a well-known definition (called 3Vs) to clarify the meaning of big data: volume, velocity and variety. Other studies [4] add veracity and value, which give the 5Vs definition. Besides, validity, variability, location, vocabulary, and vagueness (inaccuracy) were added later to the 5Vs to fully explain the term big data.

Regarding the Big Data market, it reached about $ 16.1 billion in 2014 and is predicted to attain $ 114 billion by 2018 (IDC) [4]. These forecasts undoubtedly confirm that the future of big data is very promising. In addition, the importance is seen in other vital areas such as the control and prevention of epidemic diseases [1, 2], electronic commerce, smart cities and air traffic management. Thanks to big data, business leaders can really measure and improve the basic knowledge of their activities and directly translate this knowledge into safer decisions. This allows managers to make decisions

based on evidence rather than intuition, which leads to a considerable reduction in the waste of time, effort and resources and a significant gain in earnings. It is very apparent that the big data is well positioned to be a very active research area to fulfil the exorbitant needs of companies.

With the exponential progress of big data and the inevitable need for proper control of these domains whether for commercial, strategic or political and social reasons, traditional tools show a big fiasco to manipulate the huge amount of data in research, analysis and knowledge extraction that can be used by decision makers [7, 8]. To overcome this deficit, several techniques and tools are developed to further provide storage capacity, parallel processing and real-time analysis of different heterogeneous sources and complex data [2]. The advantage of these solutions is to offer more reliability, flexibility, scalability [6] and performance with a continuously decreasing cost. These tools are mainly characterized by the Hadoop ecosystem which gave an excellent opportunity for high storage capacity and fast data processing thanks to these two main components HDFS and MapReduce [9, 10]. Nevertheless, operating big data is not only a storage and management challenge. However, understanding and extracting coherent knowledge from these data [1] become a primordial requirement to reply to the excessive needs of customers and companies in particular decision-makers [11].

This process serves to endow big data with semantics for data derived from independent and heterogeneous sources. In the artificial intelligence field, ontology refers to the combination of data in a way that users can get a standard view [12]. Ontologies offer a solution for the heterogeneity issue and guarantee interoperability between applications exploiting the big data environment [13]. It makes the data readable and comprehensive and allows sharing and exchanging data between individuals, systems and organisations without any particular effort [5, 11]. This work aims to provide a brief overview of semantic integration approaches in big data. The rest of paper is organised as follows. Section 2 consists of motivation and related works. Section 3 presents a general overview of big data. In Section 4, we focus on semantic usage in big data. Section 5 is devoted to semantic integration in big data with several illustrating examples of integration. Finally, in Section 6, a brief conclusion and future work are mentioned.

## 2 Motivation and Related Works

With the rapid development of information and network technology, people's ability to search, store and share data are increasing. The data have exploded dramatically. In sharp contrast, the ability to get valuable data for decision

makers remain very poor. To extract the complete knowledge that is usually hidden in the raw data one needs to apply analysis and mining. However, to share and reuse knowledge remains difficult issues in big data applications. To overcome this challenge, semantic integration in big data seems to be the right solution [14–17]. The remark of bibliographic research reveals that this type of survey on the integration of ontologies in the field of big data is not sufficiently widespread. However, the number of recent efforts regarding the application of ontology integration in various aspects of big data is quite perceptible. In our research, the related work we have seen deals with the topic of integrating with different perspectives. Thus, the efforts in [11, 15, 16, 18–24] seems to be quite limited regarding in scope, comprehensiveness and content to provide a clear understanding of how ontologies are actually applied in different aspects of big data. The latter facts prompted us to work on a comprehensive study of the current state-of-the-art approaches, techniques and practices in the ontology integration literature in the big data domains.

The main purpose of this paper is to explore the major ontology landscapes in this domain and surveys what and how big data technologies are being used in the recent domains such as healthcare, social media, IoT and business. In this article, we try to discuss semantic role to improve data value and clarify how ontology helps to structure data and facilitates data understanding and shared, which lead to allow easy exploration, reliable and efficient analysis of multidimensional datasets. It also shows clearly how Big Data and ontologies are related to each other and how Semantic integration may help. We expect that this review will potentially promote new research perspectives and provide a new opportunity for both ontology and big data communities.

## 3 Big Data

For well positioning in big data domain, we devote this section to focus primarily on big data development, aspects, technologies and different analysis applied in their fields. In addition, this section offers added value through a comprehensive overview of big data.

### 3.1 Understanding Big Data

In the literature, big data have several definitions, according to Davis and Patterson “big data are data too big to be manipulated and analysed by

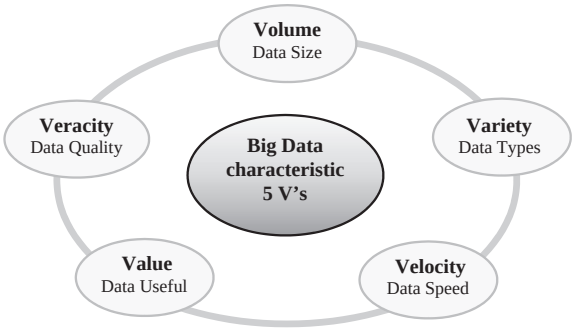


**Figure 1** Characteristics of Big Data [33].

traditional database protocols such as SQL" [9, 25] and others share this point of view. Similarly, Manyika et al. [26] defined big data as "a data set that exceeds the capacity of typical database software tools to capture". These two groups of authors agree that only data size is the factor that characterizes big data. Edd Dumbill showed the multidimensional aspect of big data when he added that the data are too big, move very fast, or do not match the restrictions of database architectures [9]. Through all of this, we clearly understand that we should add other features so that a significant amount of data is considered as big data.

The authors [9, 25] use 3Vs (Volume, Variety and Velocity) to characterize big data. Moreover, Lomotey in [27] described the model 5Vs (Volume, Variety, Velocity, Value and Veracity), illustrated in Figure 1, that is an extension of the previous model of 3 Vs. Other authors [9, 26] and institutions like IEEE rely upon visualisation more than value and veracity [28]. The visualisation shows their importance by the new tools which are used to understand the data and analyse the results [29]. In order to understand these Vs, we explain the characteristics of big data:

### Volume (data in storage support)

It refers to a significant amount of data generated and collected by individuals and businesses at a scale of petabytes and even terabytes, these data of different types continuously flood the databases of several companies such as Facebook, Twitter, YouTube, etc. [30, 31]. This challenge obliges researchers and companies to find convincing ways and tools to store and manage these massive data in a reliable way. Figure 2 shows data growth between 2005–2020 [26].

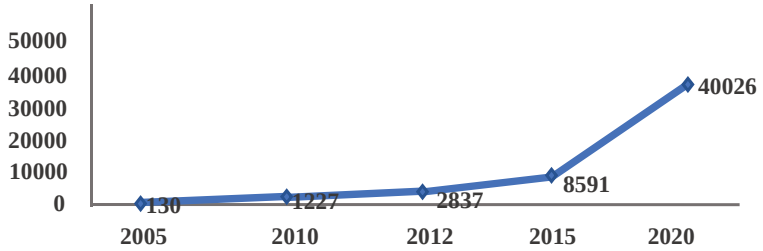


**Figure 2** Data growth by year (Exabytes) [32].

### Variety (data in many forms)

The different sources generate different heterogeneous data types that are classified into three formats: structured (relational data databases), semi-structured (weblogs, email, social media feeds, etc.) and unstructured (photo, audio, video, sensors, etc.) [32]. This requires new techniques able to handle and analyse the data effectively [2]. Therefore, it is rarely presented in a standard format and it is necessary to pass through a standardization or integration step before processing the data [11].

### Velocity (moving data)

Meaning the speed of real-time data generated from billions of connected machines [7]. Gartner (2015) estimated that connected devices will reach 20.8 billion by 2020 [33]. The data produced by various sources are continuously changing and evolving, which poses a significant challenge for real-time retrieval, processing, sharing and analysis by traditional devices [1]. In reality, some activities are critical and require immediate answers to maximize the value of information and the effectiveness of the application role [2]. This gives the motivation to build infrastructure able to promptly react and agilely respond [34].

### Veracity (quality or reliability of the data)

It is suggested by IBM and Microsoft as another dimension to measure the reliability of data from different sources [35]. Hence, veracity refers to the degree of trust given to a data leader treated and analysed to make a decision [9, 30]. Tools and statistical techniques have been developed to address the unreliability of data with specified confidence levels or confidence intervals [33].

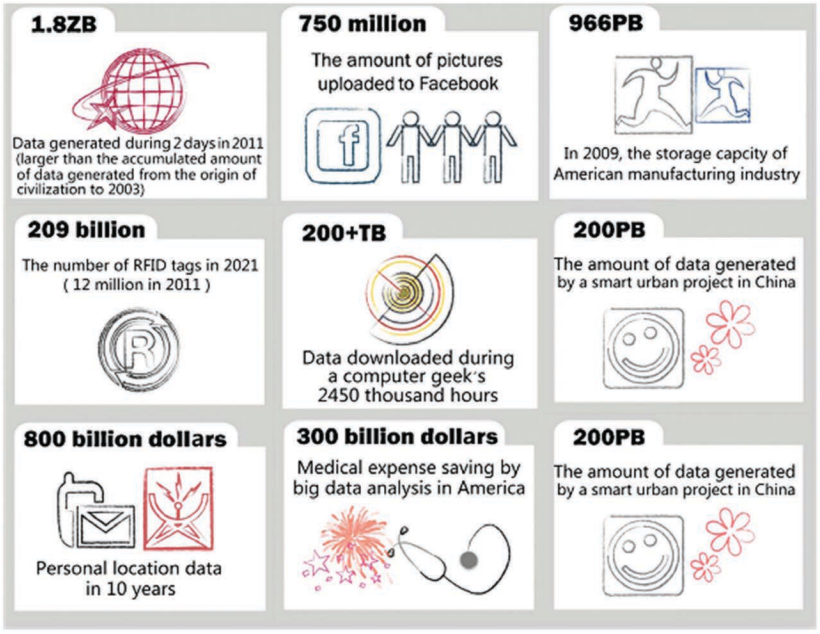


**Figure 3** Big Data in constant evolution [1].

### Value (value of the extracted data)

Oracle introduced this feature as an extra dimension to the big data 3Vs [33]. The data will have a significant value after the processing and analysis phase, which is considered a goal of big data technology [36]. For this reason, the big data architectures are designed to economically extract the relevant data from huge volumes of a wide variety of data [32]. This allows companies to increase returns, reduce operating costs and better serve customers, with reasonable amounts of investment. Figure 3 shows the volume of big data and their financial impact [1].

Other aspects, such as the variability and the visualisation [37], of big data are cited in the literature and found to be necessary. The variability of the data stands for the change of meaning according to the context and the visualisation refers to the way of representing the processed data in an easily readable and interpretable form [38] such as a graphical representation in the form of a table, images and diagrams [36]. The security aspect and the confidentiality of private and strategic data are also very demanding objects of research.

For an efficient and reliable processing of big data, researchers must evaluate the value criterion regarding the volume, capturing and manipulating the variety and veracity of the data while they are still moving (speed). Aruna et al. [39] recommends that scientists must attack all the criteria and features of big data jointly without leaning towards a criterion to the detriment of others, otherwise big data no longer meets the expected ambitions [35].

Big data characteristics can be categorized according to the following classes: data sources, content format, data storage, data stage and data processing [40]. That means:

- Data sources: Web and social network, machines, sensors, transactions and Internet of Things (IoT), etc.
- Content format: structured, semi-structured and unstructured.
- Data storage: Oriented document, column-oriented, graph based on a key-value, etc.
- Stage of data: cleaning, standardization and transformation.
- Data Processing: Lot (patch) and Real Time (streaming).

Table 1 indicates that big data has a strong potential for value creation in various industry sectors. Multiple domains benefit from the capabilities and opportunities offered by Big Data.

### 3.2 History and Evolution of Big Data

Although the emergence of big data has appeared only in recent years, the process of collecting and storing data goes back to the 1950s with the use

**Table 1** Overview of big data opportunities and 5Vs accuracy [35]

| Organizations | Volume | Variety | Velocity | Veracity | Value |
|---|---|---|---|---|---|
| Social Media | High | High | High | Medium | High |
| Healthcare | High | High | Medium | Medium | High |
| IoT | High | High | High | High | High |
| Transportation | Medium | Medium | Medium | High | Medium |
| Utilities | Medium | Medium | Medium | Medium | Medium |
| Government | High | High | Medium | High | High |
| Education | High | High | Medium | High | High |
| Insurance | High | Medium | Medium | Medium | High |
| Manufacturing | High | High | High | Medium | High |
| Natural resources | High | High | High | Medium | High |
| Banking | High | Medium | High | Medium | High |

of the first commercial computers. From this period until the 1990s, data progressed slowly due to the high cost of computers and the lack of storage media. Also, data in this period were structured as they were intended to serve operational information systems. Nevertheless, some technologies showed a considerable capacity to satisfy the need for continuous processing and storage, especially in the 1980s. When parallel databases appeared, the systems architectures are based on clusters where each machine has its processor and storage disc [1]. In the early 1990s, the World Wide Web appeared and led to explosive data development that involved effective assistance in processing and analysing these growing masses of data [33]. This process has evolved into three main stages:

The first generation of big data 1.0 (1994–2004): this stage witnessed the advent of E-commerce; the big companies were the leading players in web content. The creation of compelling exploration techniques was a requirement to explore and analyse the online activities of users. To cope with the challenge of big data management, Google has created GFS and MapReduce programming models to manage and analyse data across the Internet [1]. Web content exploration was divided into three different types: usage, web structure and web content. Techniques such as information retrieval and natural language processing have been introduced to extract the desired information [33]. Web content extraction techniques were limited during the big data 1.0 era and required improvement to deal with the explosive volume of data.

The second generation, called big data 2.0 (2005–2014): it was engendered by Web 2.0 and social media, which allows users to interact with websites and share their content. The majority of big companies (EMC, Oracle, IBM, Microsoft, Google, Amazon, and Facebook, etc.) have launched their own big data projects. The sentiment analysis, cloud-based service and methods based on lexicon and machine learning were widely used to measure the structure of social networks, relationships and other properties in order to model the development and dynamism of networks such as Facebook, Twitter and LinkedIn [33].

The last generation or big data 3.0 (2015–. . . ): it includes data from the previous two generations. IoT applications are the essential contributors to the massive amount of data generated in the form of signals, images, audio and video [41]. These data have different characteristics from those of the general big data due to the different types of collected data, such as noise and high redundancy [1]. Most IoT applications send data collected by on-site sensors [42], and the analysis of this data is done in streaming unlike

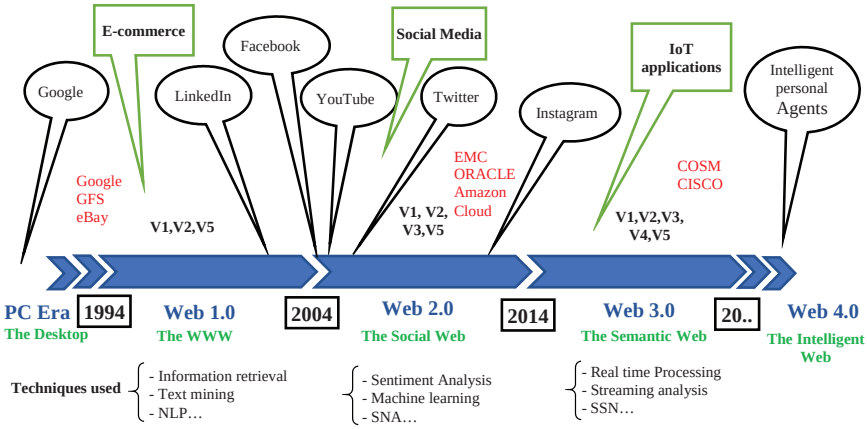


**Figure 4** Big Data development History, 5 Vs appearance, technologies and domains implication.

batch analysis of stored data used by social networks [43]. Streaming analysis requires real-time processing to discover patterns of interest being generated and collecting data as well as predicting future events that may occur or alert on time for critical applications [33]. Figure 4 represents Big Data timeline and their 5 Vs development, domains and technologies used [25].

### 3.3 Big Data Analysis

The use of advanced analytical techniques on big data determines their analysis quality. Currently, the available tools and techniques can ensure big data gathering and analysis. The analysis process requires a prerequisite stage, storage tools and capabilities for handling and managing large volumes of data [9]. According to their size, big data offer large statistical patterns and improve the experiment results. Finally, businesses and governments have clearly leveraged the benefits of investing in big data industries. Therefore, for an efficient analysis of big data, the use of various techniques becomes an absolute necessity. In this context, we quote the following domain analysis:

#### Text Analysis

Text analysis refers to the process of extracting valuable information [44] and knowledge from a massive amount of unstructured text. Text mining is located at the intersection of many disciplines, including information retrieval, machine learning, natural language processing (NLP), data mining

and statistics [1, 45]. Also, it can be used for topics modelling, questions answering [9] and search for information from e-mail, blogs [46], direct forums. This method of analysis involves statistical analysis and machine learning to extract meaningful data [8, 40], which provides the machine with a behavioural change capability based on empirical data [35].

### Audio Analysis

The purpose of the audio analysis is to extract the desired information from audio data [36], which have an unstructured format [36]. This task is carried out by using automatic speech recognition tools that are generally used at listening centres to analyse and understand the needs of customers in order to provide them with an adequate service and consequently increase the company's profitability [33].

### Video Analysis

The format of the videos is not the only challenge for the processing and analysis of this data type. The video size does not stop redoubling especially with the high quality (HD) and the high resolution afforded by modern cameras. This creates a significant challenge in dealing with these data and finding the practical tools that could support the analysis process in an efficient manner similar to other types of data. Video sharing sites are the main contributors responsible for web flooding. YouTube only powered by hundreds of video hours every minute [33]. Various techniques developed for real-time processing as well as pre-recorded videos generally used for automatic search and indexing of multimedia content, which is performed to facilitate search and retrieval of videos by the combination of audio analysis techniques and text that participate in the video indexing task. Among the areas of application of video analysis, we find marketing and operations management (crime control, border control, crowd control, etc.) [40].

### Social Networks Analysis

Web data analysis including Social Network Analysis (SNA) has raised as an active research era. Their main goal is to ensure an automatically retrieve, extract, and evaluate information from Web documents and services to discover useful knowledge [1, 29]. SNA has appeared with the birth of Web 2.0; it involves analysing both structured and unstructured data. Its goal is to see social relations in social network theory [34, 35].

Several categories of social media exist such as Social networks (Facebook, LinkedIn), Blogs (BlogSpot, WordPress), Microblogs (Twitter, Tumblr), Media sharing (Instagram, YouTube), Wiki (Wikipedia, Wikihow), etc. [40].

The analysis of social networks involves several disciplines, such as psychology, sociology, anthropology, computer science, mathematics and economics. Social network users generate data which are usually in the form of feelings, opinions, photos, videos, etc. [1] and interaction with other web entities (people, organisations and products) are the essential sources of data (structured and unstructured) [29]. The known techniques used in this field are community detection, social influence analysis, link prediction specification. Marketing is the leading application of social media analytics as it benefits from the widespread and growing social media by consumers around the world [33].

#### Predictive Analysis

It is mainly based on statistical methods. Its goal is to predict future actions and outcomes based on historical and current data. Predictive analytics seeks to discover patterns and find relationships between data [33]. Mobile means, linear regression and machine learning (neural networks) are the main techniques used in the predictive analysis. It is used in several sectors: e-commerce, weather forecast, epidemic spread, nuclear stations, election campaigns, etc. [5, 40].

### 3.4 Life Cycle of Data in Big Data

The process or the life cycle of data in big data usually contains the following steps [40]:

1. Data management: it includes acquisition and storing, extraction, cleaning and annotation, Integration and Aggregation and Representation.
2. Data Analysis: it consists in modelling, analysis and interpretation [47].

### 3.5 Big Data Technologies

The massive amount of exponentially generated and stored data and their characteristics, which widely differ from traditional databases, require new computer technologies capable of acquiring, storing, manipulating and analysing big data with an affordable cost and a very reasonable processing time. Enterprises are now more capable of handling the enormous volume of

data, which were previously processed by using expensive supercomputers with much cheaper tools [2, 9]. The use of new technologies is inevitable to manage and analyse this data in almost real time; this appears to be a crucial factor to benefit from the value of data generated by users.

The resulting solutions affected the data management market where several solutions, like NoSQL database [28], R and Hadoop, were developed. The Apache Hadoop Project (created by Doug Cutting in 2009) is a valuable tool [3, 10]. Hadoop is an open source (written in Java) that consists of two main modules: Distributed File System (HDFS) and MapReduce [1]. It offers the ability to process massive amounts of data efficiently in a reasonable time with low cost, regardless of their structure [30].

### Hadoop Ecosystem

The primary objective of Hadoop is to remedy the deficit recorded by the traditional tools of processing and analysing massive data. Hadoop offers many privileges to big data thanks to its fundamental components and its ecosystem that has a multitude of annexed software. The parallelisation of data processing between compute nodes characterises the solution of Hadoop distribution as it speeds up execution tasks and decreases latency through less expensive servers. It can also offer the possibility of processing all types of data (structured, semi-structured, unstructured) [3].

The HDFS architecture is based on a master-slave type represented by NameNode and DataNode. The former is responsible for managing the namespace, the file hierarchy and the metadata of each file while the latter contains the stored data [48].

MapReduce, introduced by Google in 2004 [48], relies on a master-slave architecture [9]. It provides efficiency and speed by parallel processing of data batches across cluster nodes. The treatment process is divided into two functions, namely Map and Reduce. The Map phase is used to analyse the problem and split it into sub-problems that are sent to other cluster nodes by using a Map function (key, value) [28]. In the Reduce phase, the lowest nodes return their results to the parent node. It calculates a partial result with the Reduce function and returns it to its parent node until the end of the process. The originating node establishes the final result by the function: Reduce (key, list (value)) $\rightarrow$ value [2, 48]. MapReduce greatly reduces data traffic across the network by moving computing processes to data on HDFS. For example, a problem that takes a few days of treatment can be solved in hours or minutes. Nevertheless, MapReduce has its disadvantages such as the

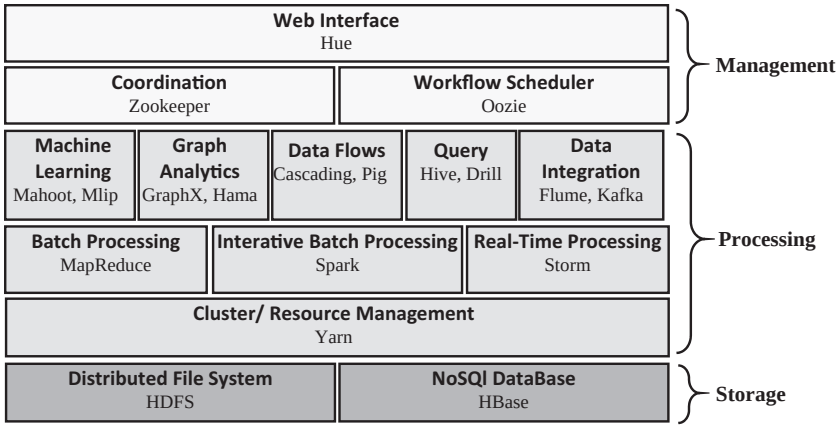


**Figure 5** The Hadoop Ecosystem [48].

difficulty of processing streaming data and many algorithms do not translate easily into its models.

YARN Hadoop is a developed example that allows batch and streaming data processing, it offers scalability, improved parallelism and advanced resource management compared to MapReduce. Around HDFS and MapReduce there are dozens of projects (Apache Pig, Apache Hive, Apache HBase, Apache Phoenix, Apache Spark, Apache Zookeeper, Cloudera Impala, Apache Flume, Apache Sqoop, Apache oozie, Apache Storm, etc.) [2, 10]. Each module corresponds to a specific feature and is intended for its own community of developers to accomplish particular tasks. In Figure 5 projects are classified according to their capabilities (storage, processing and management) [48]. Several companies such as Cloudera, Hortonworks and MapR offer Hadoop distributions that include a number of these projects.

## Hadoop Capabilities

- Storage capacity and massive data processing of different types, which do not stop expanding exponentially usually coming from social media and IoT [43].
- Computational power thanks to Hadoop's distributed computing architecture that offers fast processing of large data.
- Fault Tolerance, with data replication and multitudes of nodes committed to processing the system switches automatically in case of failure or unavailability of nodes [6].

- Flexibility, the ability to store without limitation structured or unstructured data then applying different types of processing.
- Low cost of storage and processing guaranteed by the open-source framework.
- System scalability, by adding storage nodes and processing easily without affecting the system.

## 4 Using Ontologies and Semantic Web with Big Data

Nowadays, the Web becomes a rich and complex source of information with billions of web pages. Unfortunately, most of them are neither readable nor exploitable by the individuals since they have a multitude of formats which are not readable by the machine and cannot deal with conventional software. To remedy this shortcoming and benefit the maximum of Web content, researchers have spent too much time in order to develop ontologies and the Semantic Web which aim to describe the structure of the information (syntactical aspect) and the meaning of data (semantic aspect) by the integration of a knowledge layer in the systems. Consequently, it offers the possibility of processing and exploiting the manipulated information.

Ontologies appeared in the 1990s to allow the representation of knowledge especially in the field of knowledge engineering and artificial intelligence to help the machine to exploit this knowledge [49]. Semantic Web technology aims to strengthen the representational aspect of data in the Web that makes searching, exchanging and exploiting the content easily [50]. This process is feasible by tagging or marking up, and it is possible to annotate the content of the Web by metadata that can be processed by software agents [51]. The appearance of the semantic web permits to exploit the content, which is considered as a qualitative development and a promoter of the standard Web. According to Tim Berners Lee "The Semantic Web provides a model that allows data to be shared and reused across multiple applications, enterprises, and user groups". It also provides automation, integration and reuse of data between various applications [51]. Unlike the classic Web, which is mainly based on the structure of the document and the links between the pages where extraction and exploitation of information by the machine is impossible [52]. Their advantages facilitated the machine's understanding and processing of information usually written in a natural language. This improvement mainly relieves the users of tasks that were previously of their prerogatives and adds flexibility in the relationship between computers and people, who became able to work and cooperate effectively.

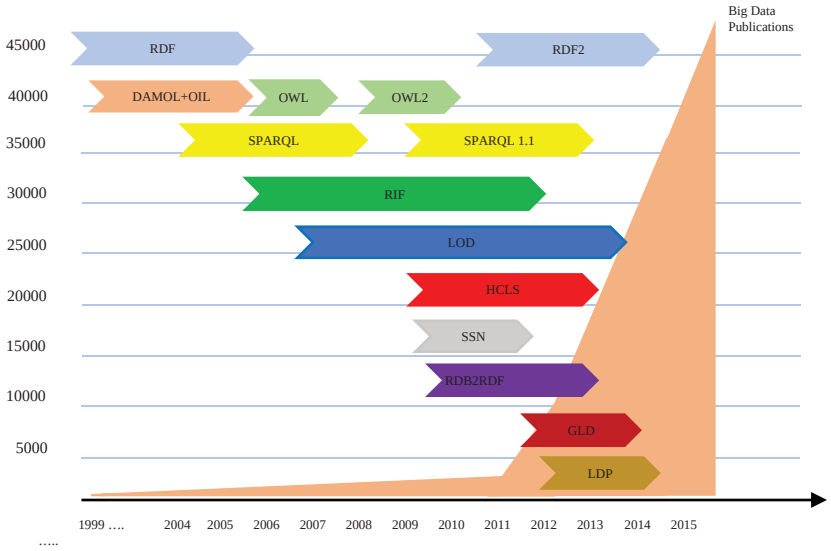


**Figure 6** Publications Number in The Big Data Domain and the Emergence of Semantic Web Technologies.

It becomes evident that after the appearance of the notion of Linked data by Tim Berners Lee the Semantic Web can build a network of structured information. This network allows the availability online and the reusability by many applications and domains including big data, the cloud, smart web services, open data, and so on. Other researchers predict that the Web of tomorrow will have a form of global and collective intelligence more than today's Web, it is the Web 3.0.

*<<The Semantic Web will enable machines to comprehend semantic documents and data, not human speech and writings>>* Tim Berners Lee [52]. Figure 6 clearly shows the fascinating growth of Big Data publications over the decades and the emergence of semantic technologies which increase respectively [49, 53].

## 4.1 Reasons for Using Ontologies

To ensure semantic interoperability in a heterogeneous information system the meaning of information shared across systems must be understood. Semantic conflicts appear when two different contexts use the same interpretation of information. Ontologies provide a unifying framework and provide primitives that facilitate communication between individuals, between

individuals and systems, and between systems [49]. They can be used to eliminate the ambiguity of the terms. Also, it can offer potentials for inference. Various fields use Ontologies to [54]:

- Share information and knowledge and facilitate the common understanding of the information structure between systems as well as between people and the machine [49].
- Standardise vocabulary across a domain and have a common understanding.
- Improve information retrieval processes [55].
- Allow the reuse of knowledge on a domain (define once but use in several applications of the same domain).
- Explicit what is considered implicit in a domain (define classes, relationships and instances).
- Distinguish knowledge in a field of operational knowledge.
- Analyse knowledge on a domain (the sense of association between objects).
- Execute and process queries expressed in a natural language.

### 4.2 The Process of Ontology Construction

Ontologies development process is also intended to use in information systems, it must follow similar procedures to those used in software engineering. In fact, the used ontology is intended to support automatic manipulation by software agents that must know the different steps to achieve the appropriate tasks needed to accomplish their objectives, either to call a procedure written in its code or to make a correspondence between the description tasks and procedures. This description is performed by using a formal language that is characterised by an ontological representation, which starts from a definition of domain representation primitives, then defines the meaning and semantics of the concepts which are used later for the representation of knowledge [56]. For an ontology to be sustainable and to reply to the objectives envisaged, it is apparent to follow and respect ontology construction rules. Starting with the definition of the application domain, ensure the reuse of existing ontologies to take advantage of what is already available and exploit the maximum of standard languages. In the literature, several methodologies are proposed for ontology construction. However, no consensus is adapted to agree on the best practice. Nevertheless, all the followed processes share certain primordial steps, namely semantic specification, conceptualisation and formalisation [56].

Other authors add an ontologisation phase and an operationalisation phase to the formalisation phase. The ontology construction process involves back and forth between conceptualisation and formalisation steps. So, the creation process is not linear since it has an iterative and incremental aspect [57]. Among the best-known methods, we cite On-To-Knowledge and Methontology, which include the main stages of the ontology lifecycle. In order to obtain a more coherent ontology, it is necessary to respect some general recommendations and principles. These recommendations should be used to guide the modeller during ontology construction such as [56]: clarity and objectivity, completeness and perfection, scalability [16], Minimal ontological commitment, Minimal encoding deformation and Ontological distinction [58, 59].

### 4.3 How Can Big Data Help to Overcome Drawbacks of Ontologies?

Large ontologies generally suffer from certain deficits, such as the ratio between the sizes of the instances and the working memory, besides the management of these instances in memory which appears to be difficult. These disadvantages are related to the size of the storage memory in which the execution is done. The updating process of the ontology requires reloading data that have already been recorded and then entirely rewritten at the end of the session (in traditional systems). This task lacks flexibility and performance (regarding execution time) especially in the case where the size of the ontology (number of instances) becomes enormous. It is necessary to find sufficient storage means for these massive number of instances effectively in order to enhance the performance and reliability requirements needed for many applications which are becoming an obligation. Therefore, using query languages to support these structures becomes a challenge for the SW community. To improve the flexibility and performance of ontology-based systems, the tools offered by big data (and NoSQL database) prove to be very promising in resolving these constraints. Indeed, these tools make it possible to store a large volume of ontology instances regardless of their size (HDFS) and run it quickly thanks to their processing potentials (MapReduce). As a consequence, using query languages, such as SPARQL, capable of supporting these structures becomes a serious challenge to the SW community.

In big data, we can store and manage a large quantity of information, regardless of their volume, without affecting the ease of data management. For the SW community, linking ontology to big data allows to get more flexibility and benefit from the functionality of big data ecosystem [28].

The use of big data in several ontological systems allows the possibility to use a large number of instances and store them in big data environment. Consequently, it leads to reliable and efficient data management and provides advantages for enterprises to use ontology in a large number of data and improve their meaning to extract new knowledge and facilitate information sharing between individuals and systems.

## 5 Semantic Integration in Big Data

This part will address the main purpose of this article. Hence, it is necessary to see first the meaning of integration notion, its role, advantages provided, different approaches used and also a diversity of works with a brief discussion.

### 5.1 From Data Integration Systems to Semantic Integration in Data

Since the advent of databases in the 1960s, researchers have worked extensively to combine different heterogeneous data sources to provide a standard query interface for querying and sharing data. This combination is performed according to the location of the data: a virtual integration of data by the mediator architecture, and a materialised integration of data where data is stored in a data warehouse. However, the current web and the massive proliferation of big data generate new challenges for sources heterogeneity. This latter appears in two aspects: heterogeneity of patterns and heterogeneity of data. To overcome this heterogeneity of schemas and data, the work focuses on the information present in schemas and data by syntactical comparison techniques such as similarity measures, heuristics, etc. [60]. Until this point, the problem of heterogeneity is not yet solved and an urgent need to explicit the semantic of the integrated data becomes inevitable. In other words, the semantic integration of the data can reduce the impact of sources heterogeneity and make the data more coherent and exploitable [50]. That is to say, the queries can be expressed regarding a single ontology and also queried via a single query interface [61].

### 5.2 Why Do We Need for Semantic in Big Data?

Today, big data hosts massive data generated from billions of people and systems stored in a myriad of sources [36]. However, they still suffer from

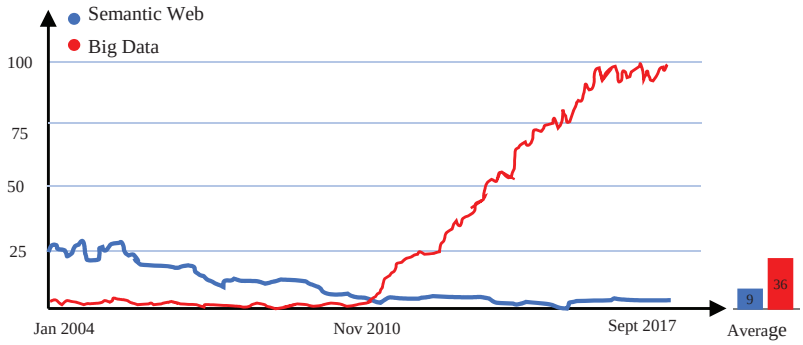


**Figure 7** Shows the ratio between the level of interest of Big Data and the Semantic Web [63].

a lack of the semantic aspect. The impact provided by big data is overturning the domains of industry [62], science and society leading people to think seriously to adopt new more efficient practices for commercial plans and government policies, especially in vital sectors such as energy, health, transportation, and so on. However, data are often stored in an incomplete, unstructured and heterogeneous format, which make much of them inaccessible by users. This tedious situation compels the research community to find technologies and tools to facilitate retrieving, exploiting and sharing these data to make them more understandable and readier for decision-makers [20]. Ontologies integration in big data proves to be the appropriate solution to overcome this challenge and enrich big data with semantics. Also, it can ensure semantic interoperability between systems using big data. For this reason, the existing tools based on the ETL process are no longer sufficient to consider the integration of many distributed data sources. The exploitation of ontologies and the Semantic Web capabilities can realise the promise of big data [15].

Over the last two decades, the Semantic Web had about 4X more advantages in Google searches than Big Data. Currently, Big Data searches are 25X higher than Semantic Web. The cross-over point appeared around April 2011: previsions expect that Big Data interest would exceed by 35X the Semantic Web in near years (as shown in Figure 7) which express the excessive need of semantic in Big Data domains [63].

## Semantic Integration: Is It a Challenge?

The data storage and processing in big data fields are no longer a worry for researchers, but the main issue that concerns them is how to endow

these different data with semantics for analysing and extracting the relevant information. Therefore, an approach aims to annotate and organise the data and their metadata in a way that permit the concepts to have a similar meaning non-contradictory [16]. For the search engines and analytics tools, this process offers an opportunity to reliably and efficiently find and extract information. Furthermore, ontologies provide a means that facilitate the information understanding exchanged between interoperable systems (semantically) through a standardised perspective of concepts representation and their relationships.

### What Does Ontology Integration Mean?

Ontologies are often considered as a formal framework used to provide a set of data with semantics. They not only allow to describe and represent the data but also make them understandable and shareable between different systems and provide capacities for communication and interoperability between them. Therein, semantics focus on the organisation and action of information that is seen as an intermediary between heterogeneous data sources, which can produce conflict not only with the structure but also with the context or the data value. For this reason, one must first understand the meaning of integration in the field of ontology. We can understand that data integration concerns the unification of data sharing some common semantics that are usually stored in unrelated sources. This definition involves the combination of data to provide a uniform view for users [11]. Moreover, this word refers to three meanings [64]:

- *Integration:* integration of ontologies during the construction of a new ontology by reusing other existing ontologies answering to the appropriate requirements by the envisaged ontology. Sometimes one builds a complete ontology by assembling other ontologies which must satisfy some criteria such as specification, adaptation, and so on.
- *Fusion:* used to build a new ontology by merging different ontologies into a unifying one that exploits their capabilities by merging ideas, concepts, axioms... In other words, mixing the knowledge of different ontologies in a single ontology aims at the same subject in order to unify concepts, terminologies, definitions, etc. [61].
- *Use/application:* the basic idea is to use one or more ontologies shared between different applications in an application intended to specify or implement a knowledge-based system (KBS), usually sharing the

same types of knowledge used in different applications, but each one is dedicated for specific needs [61].

### 5.3 Ontology Integration Process

Ontologies are introduced into materialised and virtual integration systems to solve semantic and syntactic conflicts. Three approaches can be used to manage this conflict [23]: (1) manual, with the help of a human expert (Tsimmis system), (2) semi-automatic based on linguistic ontologies (Momis system, WordNet) and (3) automatic, by the incorporation of conceptual ontologies [47] (Buster project, Picsel or SHOE). The use of ontologies is extended to design tasks (conceptual modelling, multidimensional, ETL processes, etc.) that offered effective and automatic management of encountered conflicts [65]. The choice of candidate ontologies for the integration process is done either by finding available ontologies or choosing among available ontologies those that are eligible [66]. The ontology integration process includes different steps:

1. Identify the possibility of integration, the modules needed to build the future ontology, identify ontological assumptions and commitments and identify the knowledge to be represented in each module.
2. Identify the candidate ontologies either by looking for available ontologies or by selecting the available ontologies from those that are eligible.
3. Obtain the candidate ontologies in an appropriate form that includes their representations and all available documentation.
4. Study and analyse the candidate ontologies through two activities: the technical evaluation by domain experts using specialised criteria, and the ontology evaluation, by ontologists with specialised criteria oriented towards integration.
5. Choose source ontologies whose development features are related to how the ontology was built [67].
6. Apply the integration operations by one of the followings: reuse, adaptation, specialisation and generalisation.
7. Analyse and evaluate the resulting ontology after knowledge integration.

Several methods proposed for the construction of ontologies each of which has its own approach. Nevertheless, the majority share common main steps to achieve the goal of data integration derived from various heterogeneous sources into a schema or model that provides sufficient semantics to perform intelligent queries and design more efficient applications.

This objective remains an open debate in the field of big data and the Semantic Web. In this context, the construction process often consists of the following steps [66]:

- Specification: identify the purpose and determine the future use of the ontology.
- Conceptualisation: structure the domain knowledge in a conceptual model.
- Formalisation: transform the conceptual model into a formal model.
- Integration: reuse existing ontologies, when possible, to accelerate the development process.
- Implementation: construct and validate an operating model usable by a computer.
- Maintenance: update the ontology when needed.

### How Can We Improve Integration?

An efficient and reliable integration of data coming from different heterogeneous sources into meaningful data models that enable smarter queries and facilitate application development is needed. However, this remains a significant challenge in the big data field [37]. To achieve this goal, a human expert responsible for domain identification and manual data analysis is mandatory. Unfortunately, this process consumes too much time [22]. Therefore, the need for algorithms able to generate automatic semantic data models is becoming paramount. Likewise, the following suggestions can improve integration:

- Identify existing ontologies through a process of data integration.
- Reuse an existing ontology with relevant classes, properties and relationships or design a new ontology.
- Align existing ontologies with rules designed for databases.
- Produce alignment rules between the concerned semantic data model and existing ontologies such as FOAF, DBPedia and Wordnet.

## 5.4 Ontologies Role in Big Data

Ontologies play a crucial role in the Semantic Web [68] as they are further applied to facilitate the processing and understanding big data and to reduce syntactic and semantic heterogeneities [65]. Semantic integration can automate communication between different computer systems; this process addresses the issues that arise from big data aspects such as variety because each software using big data must ensure that no semantic conflict arises.

Thus, the variety can be mitigated by annotating the data and using shared metadata. Furthermore, the standardisation of terms plays an important role that facilitates their reuse. The use of ontologies for data analysis shows high efficiency in data management. Currently, most big data projects process data on an ad hoc rather than a systematic basis [16]. In the literature, there are ontologies to describe data, such as PROV-O ontology. The process models typically used for the development of standard ontologies such as the OODA loop and the JDL / DFIG fusion models [16]. Big data analysis techniques provide semantic that can significantly facilitate the dissemination of results and they can correlate and relate to extending semantic schemas.

### Mapping as a Solution for Ontology Integration

Ontology reuse requires the integration of several ontologies, which seems as difficult as the development of a new ontology. The key idea is the creation of integrable modules that merge the semantics of reused components. The mapping role consists of linking ontologies to the actual content of information sources [21] which is very important in ontology reuse when there are several ontologies potentially used. According to Noy, "ontology mapping is a process that specifies a semantic convergence between different ontologies in order to extract them" [60]. Correspondence between certain entities integration aims to reduce variety issue in big data, where ontologies can help with data and metadata annotation. In several big data applications, the terminologies used may have a different interpretation. For this reason, ontologies have the capabilities of mitigating this problem significantly by providing a standard model independent from the used terminologies and the particular data represented by the mapping process [16]. This process plays an important role in reducing the heterogeneity between the different ontologies by aligning them through a semantic correspondence between the entities of these ontologies to ensure semantic interoperability [69]. Various applications, such as semantic Web, agent-to-agent communication, web services composition, use this principle [65].

### Big Data 5 vs Implications from Semantic and Linked Data Point of View

Semantic and linked data can play an important role in big data applications. Since they focused on the organization of data, their relationships. This process requires the construction of a conceptual representation used in

the application domain. For Big Data, the implication of semantics can be envisaged both from a management and technical perspectives.

Ontology is intended to describe terms and concepts structure, behaviour or functional systems, as well as facilitating the interactions and communications among Big Data systems. Ontology is suitable to abstract and decomposes complex concepts, it can also help to understand, manage and share knowledge. Techniques offered by ontologies help to mitigate the issues related to handling data integration, which requires careful modelling to ensure adequate representation of the real-world knowledge. These techniques allow an accepted level of abstraction, representation, extraction, processing, analysis and visualization of results. The issues of modelling and managing these systems are organized around the 5Vs of big data. Semantic integration, in general, contributes to the 5Vs of big data as follows [15, 17]:

- Volume: ontology can organise, structure, identify and describe important data and metadata. It can change the level of abstraction from data low-level to the highest level to make information more meaningful and suitable for decision-making. This task is realised by using a semantic perception, which aims to integrate semantics and perform perceptual inference in order to extract relevant and exploitable information from the massive amount of data [44].
- Variety: ontology can model the variety, hierarchies, data relations, integrate the data, and overcome the engendered challenges of big data warehousing. Using semantic metadata and semantic models ensure data annotation to describe and integrate data. This task can facilitate interoperability of systems and surmount the semantic and syntactic data heterogeneity [6], especially if the annotation of data is automatically generated. Or by using a domain ontology to define each source then automate the modelling of each individual sources [47].
- Velocity: data must be filtered, semantic can help to extract relevant data. Continuous semantics discussed in new works that use dynamic models for new concepts, objects, and their relationships for capture data [44]. Otherwise, researchers are looking for the possibility of using a semantic of content to create specific models that are able to know the new encountered concepts and entities. Nevertheless, online algorithms are highly recommended for real-time data analysis and filtering [24].
- Veracity: semantic use can check for data quality, completeness, and consistency. The exploration of trust semantic models and methods

that guarantee reliability by checking multimodal data using semantic constraints [28]. The exploitable information comes from multiple sources (sometimes conflicting and unreliable data) and this requires the abstraction and integration of heterogeneous data. Semantic integration can solve this weakness.

- Value: ontologies ensure an efficient management of Big Data projects; the analysis process allows to extract the value and evaluate the obtained results in a clear way. The implication of the semantic aspect: perfectly evolves the value of big data, which has enriched the semantic models and becomes complete and more expressive [16]. Also, knowledge bases are further strengthened and efficiently participate in decision-making intelligence and greatly assist in the process of decision-making.

## 5.5 Approaches to Building Ontologies

Different approaches are developed for setting up an ontology to explicit the semantics of information source. Each architecture deeply affects the representational formalism of ontology. So, the choice of an alternative is adapted to the needs of the developed application. In the field of data integration, there exist three possible variants, namely the mono-ontology approach, the multi-ontology approach and the hybrid approach [21].

### Mono-Ontology Approach (single-ontology approach)

This approach involves the use of a single distributed ontology between different knowledge-based systems. The use of a shared vocabulary is intended for the specification of semantics [54]. Global ontology links all the sources of information in such a way that each source has an independent model established for linking these objects to the global model, as shown in Figure 8 [64]. The SIMS and Picsel approach are examples of this type. One disadvantage of this integration structure is the need to redefine the ontology in case

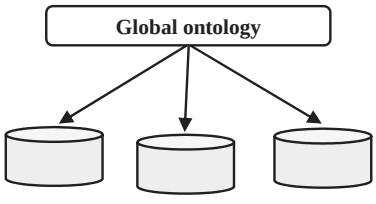


**Figure 8** Single-Ontology Approach [64].

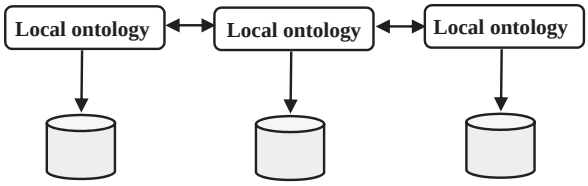


**Figure 9** Multi-Ontologies Approach [64].

of addition of a new source. The absence of a schematic autonomy of the sources is another drawback as well [65].

### Multi-Ontology Approach

Sometimes the modification of the sources can affect the global ontology in an interoperable system. The multi-ontologies approach is used to endow each source of information by a local ontology and provides liberty of definition for this source without considering the other ontologies (as shown in Figure 9). The use of this approach is generally in the case where it is impossible to find a consensual ontology generated by the glaring semantic difference between the systems. However, the absence of a standard vocabulary is shown as a major constraint limiting communication between sources. To overcome this limitation, an inter-ontology mapping is set up to make a correspondence between the different terms of the ontologies [68]. Unfortunately, this mapping is difficult to concretise due to the semantic heterogeneity that can be encountered and the complexity of mapping [67]. For N sources, the complexity is N (N-1) / 2 [44]. The OBSERVER approach used in this context [21].

### Hybrid Approach

This approach (illustrated in Figure 10) comes from the combination of the previous two approaches, which use local and shared ontologies. Each semantic source is defined by a local ontology, a common vocabulary is built to facilitate the mapping and make the different ontologies sources comparable [54]. The alignment between local ontologies and shared ontology can be done a priori or posterior [65]. Domain terms and primitives compose this vocabulary in a way that allows these primitives to construct complicated terms by combining them with operators, making the terms comparable and shareable between ontologies without any conflict [21]. This shared vocabulary gives a representation of an ontology. The COIN project represents the

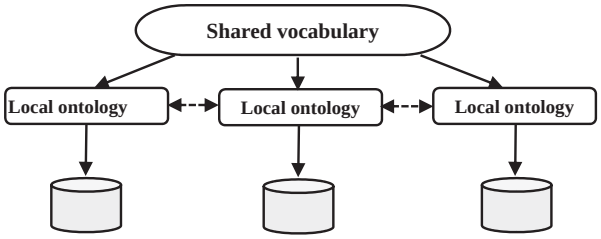


**Figure 10** Hybrid Ontologies Approach [64].

context of information by a vector of attribute values, METACOTA annotates each source by labels.

Ontologies sources are seen as a refinement of general ontology in BUSTER. The brought contribution of this approach is the possibility of adding new sources without affecting or modifying the common vocabulary, thus ensuring the evolution of the ontology [64].

Table 2 shows a comparison between the three integration approaches.

### 5.6 Opportunities and Challenges of Semantic Integration in Big Data

Various Big Data applications benefit from the semantic integration which offers promising opportunities, particularly in the standardization process of frameworks, it can also ensure the interoperability through the different systems in order to extract the relevant knowledge from the huge raw data.

We can cite the following opportunities for different fields such as:

- Share and reuse knowledge among systems and users.
- unify the vocabulary meaning used in several domains such as Social media, IoT, Healthcare, E-commerce, etc.
- Improve information retrieval and offer relevant information [55].
- Explicit the implicit information to ease the representation of linked data which offer quick and efficient analysis of large datasets in numerous domains: politics, healthcare, crisis and disasters, social movements, environment, decision making, etc.
- Execute and process queries expressed in a natural language [54].
- Reduce data traffic through the network and avoid bottleneck by moving the ontology to data sources.

**Table 2** A comparison between ontology integration approaches

| Ontology Integration Approach | Ontology Used | Vocabulary Used | Ontology Model | Implementation Feasibility | Advantages | Disadvantages | Examples |
|---|---|---|---|---|---|---|---|
| **Mono ontology approach** | Only one for all sources (global) | Shared | One model Connecting all sources | Difficult Implementation | – Interoperability between different sources<br>– Possibility of mapping between sources | – Absence of sources autonomy<br>– Extensibility requires the redefinition of ontology | SIMS<br>Picsel |
| **Multi ontology approach** | Local ontology for each source | Suitable for each source and (standardized with other sources) | Independent for each source + inter-mapping scheme | Easy implementation (Except the normalization process) | – Autonomy of definition for each source<br>– Extensibility of source | – Mapping difficult between sources<br>– Lack of common vocabulary | Observer |
| **Hybrid approach** | Local ontologies + global | Appropriate + shared global | Hybrid model | Implemented relatively easy | – Possibility of mapping<br>– Extensibility of sources<br>– Communication between sources | – The difficulty of reusing local ontologies | COIN<br>BUSTER<br>MECOTA |

- Basic knowledge can use a perceptual inference to extract knowledge and automate queries to offer smarter data serviceable for decision-makers [70].
- Ontologies used to require the security models for Big Data services when we harness the relationships of concept meaning. It can deploy semantic security risk management tools available with dynamic web service strategy contexts. Using ontologies decrease security risks (Malware Intrusion Ontology) when it used to specify communication relations, administration domains, and associated threats and vulnerabilities [71].

Semantic integration remains suffers from many challenges caused by the difficulty encountered in the integration process. The massive data sources increasing, their diversity, organisations and individuals need to create new challenges:

- The difficulty to manage large ontologies instances and optimise performance with Big Data [25].
- Highly-connected entities make difficult to distinguish data from noise and to combine data from external sources in a single global model [72].
- Difficulty to instantiate the abstract framework and to run queries over the different ontologies when reaching data from various sources [70].
- The absence of automated support needs significant time and quantity of input from domain experts for key tasks during large ontologies design [73].
- Challenge to determinate the complementary information (uncertain and incomplete) on the same entity that is dispersed across several data sources. The same entity is represented by different identifiers (URIs) which affects scalability performance on large ontologies in Big Data [70].
- To support the detailed semantic annotation of data sets when specific ontologies (application ontologies) used, it is not easy to conceive a general ontology (upper ontologies) and to support the detailed semantic annotation of data sets (e.g. medical ontologies [15], industrial ontologies [62]).
- Adopting high-speed data streams with appropriate ontologies capable of capturing and exploiting hidden knowledge remain an issue which can limit the performance of reasoners and inferences. Powerful methods and tools needed to integrate data streams [74].

## 5.7 Comparison of Works-Based Ontology Integration in Big Data

In this part, we quote some works from different fields of application such as healthcare, industry, business data, management, social networks, IoT, security and defense (Table 3). These works share the idea of integrating ontologies into big data. However, each author uses big data tools, and ontological techniques specific to the field explored according to the solicited needs and objectives. The following table shows a diversity of works including some criteria presented to give a general idea describing the covered subject and the big data tools used with the chosen ontology techniques. Also, a description of the approach and the steps followed to develop each proposal are presented. It is obvious to note the advantages provided by these works as well as their disadvantages and the encountered constraints.

## 5.8 Discussion

The works presented above cover various areas of actuality where the quantity and variety of data is an important aspect characterizing the handling and the analysing of processed data. This challenge is solved by using the means and capacities offered by big data tools (Hadoop, HDFS, MapReduce, HBase, Hive, MongoDB, etc.). The need to address the heterogeneity of different types of data generated from a multitude of resources requires the use of the ontology paradigm, these standardised and harmonised data provide intelligent data, easy to share and reuse. Moreover, it brings out the hidden semantics in the vast databases that are not directly available from distributed sources. At this stage, individuals and systems can easily communicate these semantised data, which can ensure high interoperability and improve the functionality of applications with new services. Furthermore, the addition of inference modules provides intelligent systems capable of extracting new knowledge and having a certain level of autonomy. The majority of illustrated works introduce big data tools for data processing and storage and use a domain ontology to model the knowledge of these fields.

In [75], the authors proposed a platform for the integration of heterogeneous mobile data in the health area. In this architecture, they used the cloud environment and MapReduce with WH ontology in addition to the KaaS application to integrate data from multiple wearable devices by describing a layer of semantic knowledge and using SMW with other extensions to annotate the features of the wearable devices and visualise the patient profile. Williams et al. [19] deal with the industry sector. They developed a big data

**Table 3** Comparison of works-based ontology integration in big data

| Article Ref Domain | Description of Work | Big Data Tools | Ontologies Used | Other Tools | Approaches' Used | Advantages | Drawbacks |
|---|---|---|---|---|---|---|---|
| [75] **Healthcare** | Development of an architecture enriching the NIST big data model with generic semantics to intelligently understand the data collected through wearable devices that will assist physicians to monitor the patient's health evolution and update the patients about their health. (Case diabetics). | – NIST Big Data<br>– Cloud environment<br>– Data processing algorithms<br>– MapReduce | – WH Ontology<br>– RDF<br>– SPARQL | – Semantic MediaWiki (SMW)<br>– KaaS<br>– Apache Jena Fuseki | Application of KaaS to integrate data from multiple wearable devices by describing a layer of semantic knowledge and using SMW with other extensions to annotate wearable devices features and visualize the patient's profile. | – Provide a scalable solution for storing the large volume of health care data generated from multiple sources.<br>– Support sharing and integration of data for better preventive decision-making.<br>– Extract valuable information from heterogeneous data. | – Challenge of security and confidentiality when transmitting data remains a challenge.<br>– Difficulty for the scalability of wearable KaaS for other scenarios. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [19] **Industry** | Development of a big data infrastructure through the use of semantics to create a domain model that provides access to unified data and to derive more value from data and perform analysis to improve engineering design of Gas turbines. | – Hive<br>– HDFS | – SPARQL<br>– RDF triple store<br>– URI<br>– Semantic triple store Virtuoso | – SPARQL graph<br>– SQL-like<br>– Python-based analytics | The storage layer contains the semantic model and the data instantiated to perform the test and the mapping of variables, the user interface based-web allows to create queries executed on the semantic basis and the time series database. The user can perform analyses on time series data. | – System offers significant savings in productivity and costs.<br>– Architecture scalable horizontally for storing time series data.<br>– System offers speed of processing and analysis and permits to derive more value from data through years to improve technical designs. | – Architecture limited for one type of data.<br>– System does not support storage on other NoSQL, HBase or Apache Cassandra platforms that allow scalability.<br>– System does not explore evolutionary semantics. |

(*Continued*)

**Table 3** Continued

| Article Ref Domain | Description of Work | Big Data Tools | Ontologies Used | Other Tools | Approaches' Used | Advantages | Drawbacks |
|---|---|---|---|---|---|---|---|
| [11] **Data management** | Illustration an approach to build a modular ontology integrating big data by exploiting a NOSQL database (MongoDB) in order to compose a global ontology from local ontologies by wrapping the data sources to the MongoDB database and generating local ontologies (scenario proposed for two companies). | – MongoDB<br>– NOSQL database | – OWL<br>– OWL-DL | – Talend<br>– MOOM M2Onto<br>– DBRef (MongoDB API)<br>– DBList | It starts with schema-level integration by normalizing data sources in the used schemas (MongoDB) to represent the data and then mapping the MongoDB data to an OWL ontology module in order to merge them into a global ontology module. | – Data integration provides an integrated view that facilitates access and reuses information.<br>– Approach is applicable in several areas.<br>– Possibility to integrate new sources that are emerging in the global ontology. (Horizontal evolution). | – Difficulty to update the global ontology.<br>– No vertical evolution for data sources.<br>– Slowness to handle a large number of sources (lack of big data tools ex Hadoop). |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [78] **Social networks** | Work based on ontology and NoSQL databases combine sentiment analysis (textual and visual) allowed to extract the emotions expressed in the data exploited from social networks to estimate public reaction towards a specific subject (French elections 2017). Which offer the possibility to classify the candidates (opinion poll) and to know the regions of high representativeness for each candidate. | – MongoDB<br>– NoSQL database engine<br>– Spring Data MongoDB | – SenticNet<br>– SentiBank<br>– JSON | – Spring Data MongoDB<br>– Twitter Streaming API<br>– Graph Explorer API<br>– Spring boot | – Extracting data via Facebook and Twitter APIs based on keywords, storing this data in MongoDB.<br>– Implementation of semantic paradigms and domain ontologies.<br>– Analysis of data by reasoning on predefined ontologies in order to produce the sentiment results. | – Combination of semantic web and sentiment analysis tools provide accurate results on public emotions.<br>– System can be used in other areas to assess the emotions and sentiments that can be used to guide this opinion.<br>– System offers a predicting tool for social behaviour. | – System is limited to the analysis of specific data (text, image)<br>– System is unable to perform real-time processing.<br>– System cannot determine the different communities that compose social networks. |

(*Continued*)

**Table 3** Continued

| Article Ref Domain | Description of Work | Big Data Tools | Ontologies Used | Other Tools | Approaches' Used | Advantages | Drawbacks |
|---|---|---|---|---|---|---|---|
| [41] **Internet of Things** | Proposal for an enhanced framework integrating semantic web technologies, big data and IoT. With new semantic features and big data analysis through a learning module. This system provides efficient support for all types of sensors to store and apply data reasoning to get the best results. | – NoSQL databases<br>– Apache Spark | – SSN ontology<br>– RDF, RDFS<br>– JSON-LD | – RDFS reasoning rules<br>– Jena rule syntax<br>– Machine learning methods | Acquire different data types from the sensors, apply the ETL process on the data to semantise them by using the RDF and SSN tools.<br>– Apply the reasoning rules on the RDF data to serve the system objectives.<br>– Learning phase using machine learning to improve results.<br>– Execute the actions. | – Framework provides the best features to sensor networks through the use of semantics and big data tools.<br>– System offers a scalable, intelligent and interoperable framework supporting different application areas.<br>– Learning and reasoning use can improve system results. | – Difficulty to choose the right tools and methods to implement a given system.<br>– Adaptation of different technologies in a domain requires a powerful strategy.<br>– Difficulty of integrating different types of sensors with data heterogeneity.<br>– Challenge of data availability and security. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [18] **Security (intelligence)** | Proposal of scalable intelligence data analysis and integration platform (MIDIS) through the use of semantic and big data web technologies to improve the integration and analysis of heterogeneous data. It provides timely and relevant information for security and intelligence services through intuitive search and discovery mechanisms. | – HDFS<br>– Hbase | – Semantic annotation<br>– Mapping between the terms of ontology<br>– RDF<br>– SPARQL | – GATE Platform<br>– SOA Technologies<br>– Platform cloudera | – Data acquisition from heterogeneous sources and integrated them into unified storage segments (HDFS, Hbase).<br>– Semantic enrichment based on ontology (mapping and annotation).<br>– Interrogation of data and analysis (search, filtering, notification, alert).<br>– Interactions with external reasoning modules. | Objectives:<br>– Provide timely and relevant information for analysts through intuitive search and discovery mechanisms.<br>– Provide a framework facilitating the integration of heterogeneous data and the preparation of various data to the analysis for a better knowledge of the situations. | – Semantic enrichment offers only a horizontal evolution of data integration.<br>– Lack of learning module can decrease system efficiency.<br>– Platform does not exploit the data of social networks which are critical and important information for this domain. |

infrastructure handled by HDFS and Hive; they used semantic technologies (SPARQL, RDF triple store, Virtuoso semantic triple store) to create a domain model that offers access to unified data and allows more data values to be derived and performs the analysis to improve the engineering design of gas turbines. In relation of the industrial domain [76] propose an approach aims to design generic semantic models for diagnostic applications aims to allow efficient adaptation of analytical procedures across multiple abstractions domains in remote monitoring of Siemens gas and steam turbines. Ontology-based data access (OBDA) techniques used for semantic interpretation, SSN ontology and SPARQL employed to query the instance data. These tools permit to tackle the main difficulty for analysis of different machines performance which has various sensors and a huge amount of produced data. Saettler et al. [77] propose an ontology-driven semantic framework to address interoperability issue between different data sources and dynamic service composition in Oil and Gas process plant construction. By combining web services with an ontology, an approach based on a Service Oriented Architecture provides a good description of data domain and business processes. This approach can be embedded in other systems from a diversity of domains by using of the domain ontology.

Abbes and Gargouri [11] proposes an approach destined to build a modular ontology integrating big data by exploiting a NoSQL database (MongoDB). To accomplish the objective of composing a global ontology from the data sources into the base of MongoDB, then generating local ontologies (scenario proposed for two companies), to achieve this approach, the author requires the use of OWL DL and other tools. El Hamdouni et al. [78] deals with another area (social networks). This work is based on ontologies and NoSQL databases (MongoDB), the combination of sentiment analysis (textual and visual) allowed to extract the emotions expressed in the data exploited from social networks. In order to estimate the reaction of the public towards a specific subject (French elections 2017) by the implementation of semantic paradigms and domain ontologies. This offers the possibility to classify the candidates (opinion poll) and to know the regions of high representativeness for each candidate. The authors [41] developed an augmented framework integrating semantic web technologies, big data and IoT. With new semantic functionality and big data analysis through a learning module, this system provides adequate support to all types of sensors to store and apply reasoning on the data to obtain the best results. The use of a NoSQL database, Apache Spark, SSN ontology, RDF, RDFS, RDFS

reasoning rules, etc. clearly has improved the system performance. In [18], the author designs a framework for the analysis and integration of evolving intelligence data (MIDIS) through the use of semantic web technologies RDF (annotation, mapping) and big data (HDFS, HBase) to improve the integration and analysis of heterogeneous data. Therefore, it provides timely and relevant information to security and intelligence services through intuitive search and discovery mechanisms.

In addition, real-time processing and analysis is another important aspect of Big Data which differ from regular applications [79]. In recent years, it takes a special care since it can minimise the hazards of human lives and resources, saves human lives and improves life quality and efficient resources management. The main feature of Real-time applications based on instantaneous input and immediate analysis to get a decision within a very short timeline [80]. Transportation, Clinical care, crowd control, Natural Disasters and Defense are the most field used in real-time Big Data analytics. Leenen et al. [81] highlights the argues that both semantic technologies and big data analytics combine together to provide a reliable countermeasure against cyber threats. Systems used in this domain can gather immense amounts of information then process, analyse, visualise and interpret results to be exploited later for prediction and stopover cyber-attacks. Various technologies used, such as stream reasoning, real-time monitoring, Intelligence-Led approaches, etc. Big data technologies reinforced by semantic technologies can improve cybersecurity, it can provide support for the understanding, processing and analysing of the gigantic amounts of raw information in the cyber environment.

Recent Big Data architecture called Lambda Architecture appeared to solve latency, throughput, and fault-tolerance issues by using batch and stream processing to provide comprehensive and precise views of batch data. In real-time stream processing Lambda approach provides views of online data by using three layers (batch layer exploits Pig and hive, the streaming layer employs Spark streaming and Spark SQL and serving layer). The rise of lambda architecture is associated with Big Data evolution, real-time analytics and the need to address the latencies of map-reduce [82].

## 6 Conclusion and Future works

In recent years, the massive evolution of data in all areas creates a real challenge to discover, exploit and share these enormous varieties of data.

This explosion of data is due to the development of the companies' number and activities. The popularisation of the Internet and the multi-varied offered features, facilitate the production of data, especially the emergence and the remarkable progress of social media (Facebook, Twitter, Instagram, YouTube, etc.), which have offered to the broad public the opportunity and the capability of continuously generating the data. This phenomenon makes each person as a source of a daily production of data of all types (video, photo, audio, text...). This reality provokes a paradoxical situation: how to allow the evolution of big data and to exploit and benefit from these data at the same time? The semanticisation of data appears as a promising solution to overcome this challenge. The integration of semantics into big data becomes a primary necessity in most contemporary applications as it offers the possibility of interaction and sharing knowledge as well. The use of ontology gives a great ability to emerge and share heterogeneous information.

The contribution of our work is to give a general description of semantics in big data context while showing the position of each domain with respect to the other and how semantics can help to overcome the drawbacks associated to the development of big data. Therefore, it provides the reader, who may not be very familiar with this domain of research, with a general overview of semantic integration in big data.

This paper gives, in the first place, a definition of big data with a description of their aspects, evolution, technologies and tools as well as their fields of application. Then, a general overview is given on the semantic web and ontologies: different types of ontologies, their components, construction processes, languages and editors. Finally, the article discusses the concept of semantic integration in big data while clarifying the role and the importance of this process, the used ontology approaches. This part is concluded by a comparison and discussion of some works that used the integration techniques in actuality domains.

As a future work, we aim to propose an approach of semantic integration in big data that ensures the semantic interoperability between different heterogeneous resources. The object is to extract the hidden knowledge and facilitate their sharing, in order to provide a support framework for operators and make knowledge interpretable (intelligent data) by web agents to assist in decision-making in a specific domain.

## References

[1] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mob. networks Appl.*, 2014, 9 (2), 171–209.

[2] A. Oussous, F. Benjelloun, A. Ait, and S. Belfkih, "Big Data Technologies: A Survey," *J. King Saud Univ. – Comput. Inf. Sci.*, 2017.

[3] J. Kim, "A Survey of Big Data Technologies and How Semantic Computing Can Help," *International J. Semant. Comput.*, 2014, 8 (1), 99–117.

[4] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big Data Analytics: A Survey," *J. Big Data*, 2015, 1–32.

[5] B. Eine, M. Jurisch, and W. Quint, "Ontology-Based Big Data Management," *Systems*, 2017, 5 (3), 45.

[6] H. M. Safhi, B. Frikh, B. Hirchoua, B. Ouhbi, and I. Khalil, "Data Intelligence in the Context of Big Data: A Survey," *J. Mob. Multimedia.*, 2017, 13 (1–2), 1–27.

[7] B. Andrew, McAfee. Erik, "Big Data: The Management Revolution," *Harv. Bus. Rev*, 2012, 90 (10), 60–68.

[8] A. Labrinidis and H. V. Jagadish, "Challenges and Opportunities with Big Data," *Proc. VLDB Endow*, 2012, 5 (12), 2032–2033.

[9] C. K. Emani, N. Cullot, and C. Nicolle, "Understandable Big Data: A Survey," *Comput. Sci. Rev*, 15, 70–81.

[10] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. IEEE,* 2010, 1–10.

[11] H. Abbes and F. Gargouri, "MongoDB-Based Modular Ontology Building for Big Data Integration," *J. Data Semant*, 2018, 7 (1), 1–27.

[12] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, "What Are Ontologies, and Why Do We Need Them?," *IEEE Intell. Syst. their Appl*, 1999, 14 (1), 20–26.

[13] A. Ben Salem, F. Boufares, and S. Correia, "Semantic Recognition of a Data Structure in," *J. Comput. Commun*, 2014, 2 (9), 93–102.

[14] P. Hitzler, K. Janowicz, "Linked Data, Big Data, and the 4th Paradigm," *Semant. Web*, 2013, 4 (3), 233–235.

[15] D. Obrst, N. Rychtyckyj, and M. Kim, "Integration of Big Data Using Semantic Web Technologies," in *Semantic Computing (ICSC)*, 2016, 382–385.

[16] L. Obrst, M. Grüninger, K. Baclawski, M. Bennett, D. Brickley, G. Berg-Cross, and C. Lange, "Semantic Web and Big Data Meets Applied Ontology," in *Ontology Summit 2014*, 2014.

[17] K. Thirunarayan and A. Sheth, "Semantics-Empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications," in *Semantics for Big Data: Papers from the AAAI Symposium. AAAI Technical Report FS-13-04*, 2013, 68–75.

[18] Anne-Claire Boury-Brisset, "Managing Semantic Big Data for Intelligence,", In STIDS, 2012, 41–47.

[19] J. W. Williams, P. Cuddihy, J. Mchugh, K. S. Aggour, A. Menon, S. M. Gustafson, T. Healy, and C. Control, "Semantics for Big Data Access & Integration: Improving Industrial Equipment Design through Increased Data Usability," in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, 1103–1112.

[20] G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, and R. Rosati, "Using Ontologies for Semantic Data Integration," *Springer Int. Publ.*, 2018, 187–202.

[21] H. Wache, T. Voegele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner, "Ontology-Based Integration of Information: A Survey of Existing Approaches," in *IJCAI-01 Workshop: Ontologies and Information Sharing*, 2001, 108–117.

[22] S. K. Bansal and S. Kagemann, "Semantic Extract-Transform-Load framework for Big Data Integration," *Computer (Long. Beach. Calif)*, 2015, 48 (3), 42–50.

[23] A. L. Guido and R. Paiano, "Semantic Integration of Information Systems," *Int. J. Comput. Networks Commun*, 2010, 2 (1), 48–64.

[24] V. K. Kiran and R. Vijayakumar, "Ontology-Based Data Integration of NoSQL Datastores," in *Industrial and Information Systems (ICIIS)*, 2014, 1–6.

[25] I. Lee, "Big Data: Dimensions, Evolution, Impacts, and Challenges," *Bus. Horiz*, 2017, 60 (3), 293–303.

[26] E. Zikopoulos, Paul. Chris, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data," in *McGraw-Hill Osborne Media*, 2011.

[27] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," *Rep. – McKinsey Glob. Inst*, 2011.

[28] V. C. Storey and I. Song, "Data & Knowledge Engineering Big Data Technologies and Management: What Conceptual Modelling Can Do," *Data Knowl. Eng*, 2017, 108 (2), 50–67.

[29] Q. Quboa and N. Mehandjiev, "Creating Intelligent Business Systems by Utilising Big Data and Semantics," in *Business Informatics (CBI), 2017 IEEE 19th Conference on. IEEE*, 2017, 39–46.

[30] L. R. C. Rodríguez-enríquez, J. Luis, S. J. Cervantes, J. Luis, and G. Alor-hernández, "A General Perspective of Big Data: Applications, Tools," *J. Supercompu.*, 2016, 72 (8), 3073–3113.

[31] H. J. Hadi, A. H. Shnain, S. Hadishaheed, and A. H. Ahmad, "Big Data and Five V's Characteristics," *Int. J. Adv. Electron. Comput. Sci*, 2015, 2 (1), 16–23.

[32] M. Panahiazar, V. Taslimitehrani, and A. Jadhav, "Empowering Personalized Medicine with Big Data and Semantic Web Technology: Promises, Challenges, and Use Cases," in *IEEE International Conference on Big Data 2014*, 2014, 790–795.

[33] S. R. Jeong, I. Ghani, S. Korea, and J. Bahru, "Semantic Computing for Big Data: Approaches, Tools, and Emerging Directions," *KSII Trans. INTERNET Inf. Syst*, 2014, 8 (6), 2022–2042.

[34] V. M. Rao, V. V. Kumari, and N. Silpa, "An Extensive Study on Leading Research Paths on Big Data Techniques and Technologies.," *Int. J. Comput. Eng. Technol*, 2015, 6 (12), 20–34.

[35] I. Yaqoob, I. Abaker, T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N. Badrul, and A. V. Vasilakos, "Big Data: From Beginning to Future," *Int. J. Inf. Manage*, 2016, 36 (6), 1231–1247.

[36] M. K. Saggi and S. Jain, "A Survey Towards an Integration of Big Data Analytics to Big Insights for Value-creation," *Inf. Process. Manag*, 2018, 54 (5), 758–790.

[37] C. Bizer, P. Boncz, M. L. Brodie, and O. Erling, "The Meaningful Use of Big Data: Four Perspectives – Four Challenges," *Acm Sigmod Rec*, 2012, 40 (4), 56–60.

[38] C. A. Knoblock, P. Szekely, and M. Rey, "Semantics for Big Data Integration and Analysis," in *Semantics for Big Data: Papers from the AAAI Symposium. AAAI Technical Report FS-13-04*, 2013, 28–31.

[39] T. Aruna, K. Saranya, and B. Chetna, "A Survey on Ontology Evaluation Tools," in *Process Automation, Control and Computing (PACC), 2011 International Conference on. IEEE*, 2011, 1–5.

[40] H. Özköse, P. L. Q. Uõ, and C. Gencer, "Yesterday, Today and Tomorrow of Big Data," in *Procedia-Social and Behavioral Sciences*, 2015, 195, 1042–1050.

[41] O. B. Sezer, E. Dogdu, M. Ozbayoglu, and A. Onal, "An Extended IoT Framework with Semantics, Big Data, and Analytics," in *IEEE International Conference on Big Data (Big Data)*, 2016, 1849–1856.

[42] P. Wira, I. Szilagyi, and P. Wira, "Ontologies and Semantic Web for the Internet of Things – a Survey," in *42nd IEEE Industrial Electronics Conference (IECON 2016), Florence, Italy*, 2016, 10, 6949–6954.

[43] O. B. Sezer, E. Dogdu, and A. M. Ozbayoglu, "Context-Aware Computing, Learning and Big Data on the Internet of Things: A Survey," *IEEE Internet Things J*, 2018, 5 (1), 1–27.

[44] A. Sheth, "Transforming Big Data into Smart Data: Deriving Value via Harnessing Volume, Variety, and Velocity Using Semantic Techniques and Technologies," in *Data Engineering (ICDE), 2014 IEEE 30th International Conference on. IEEE, 2014*, 2014, 2–2.

[45] G. Bello-orgaz, J. J. Jung, and D. Camacho, "Social Big Data: Recent Achievements and New Challenges," *Inf. Fusion*, 2016, 28, 45–59.

[46] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A Survey of Open Source Tools for Machine Learning with Big Data in the Hadoop Ecosystem," *J. Big Data*, 2015, 2 (1), 24.

[47] C. A. Knoblock and P. Szekely, "Exploiting Semantics for Big Data Integration," *AI Mag*, 2015, 36 (1), 25–38.

[48] C. Hsinchun, Roger H. L. Chiang, and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly,bus. Intell. Res. Bus*, 2012, 36 (4), 1165–1188.

[49] H. Wu and A. Yamaguchi, "Semantic Web Technologies for the Big Data in Life Sciences," *Biosci. Trends*, 2014, 8 (4), 192–201.

[50] S. Ahmad, C. Bukhari, K. Malik, "Semantic Web in the Age of Big Data: A Perspective," *OSF Prepr*, July 2018.

[51] S. Bourekkache, O. Kazar, L. Kahloul, F. Gargouri, and B. Aïcha-Nabila, "Un Environnement Sémantique à Base d'Agents pour la Formation à Distance (E-Learning).," in *In 10ième édition de la conférence sur Avancés des Systèmes Décisionnels-ASD 2016,* 2016.

[52] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Sci. Am,* 2001, 284 (5), 34–43.

[53] Ravil I. Muhamedyev, Maksat N. Kalimoldaev and Raisa K. Uskenbayeva, "Semantic Network of ICT Domains and Applications. Institute of Problems of Information," in *In Proceedings of the 2014 Conference*

*on Electronic Governance and Open Society: Challenges in Eurasia. ACM.*, 2014, 11, 178–186.

[54] F. Z. Laallam, M. L. Kherfi, and S. M. Benslimane, "A Survey on the Complementarity Between Database and Ontologies: Principles and Research Areas," *Int. J. Comput. Appl. Techno*l, 2014, 49 (2), 166–187.

[55] A. P. Junior and L. D. C. Botega, "Ontological Semantic Agent in the Context of Big Data: a Tool Applied to Information Retrieval in Scientific,", In New Advances in Information Systems and Technologies springer , 2016, 307–308.

[56] A. Vandecasteele, "Modélisation ontologique des connaissances experts pour l ' analyse de comportements à risque - Application à la surveillance maritime – Doctoral dissertation, Ecole Nationale Supérieure des Mines de Paris," 2012.

[57] N. Guarino, "Semantic Matching: Formal Ontological Distinctions for Information Organization , Extraction, and Integration," in *International Summer School on Information Extraction. Springer, Berlin, Heidelberg*, 1997, 139–170.

[58] T. R. Gruber, "Toward Principles for the Design Toward Principles for the Design of Ontologies Used for Knowledge Sharing," *Int. J. Human-Computer*, 1995, 43 (5–6), 907–928.

[59] A. Gómez-pérez, O. Corcho, and U. P. De Madrid, "Ontology Languages for the Semantic Web," *IEEE Intell. Syst.*, 2002, 17 (1), 54–60.

[60] N. F. Noy, "Semantic Integration: A Survey Of Ontology-Based Approaches," *ACM Sigmod Rec*, 2004, 33 (4), 65–70.

[61] H. S. Pinto, P. Martins, B. Monte, and A. R. Pais, "Some Issues on Ontology Integration," *IJCAI Scand. AI Soc. CEUR Work. Proceedings,* 1999.

[62] D. A. Koutsomitropoulos and A. K. Kalou, "A Standards-Based Ontology and Support for Big Data Analytics in the," *ICT Express*, 2017, 3 (2), 57–61.

[63] B. Mike,"SemWeb getting crushed by Big data in search popularity," *http://www.mkbergman.com/1803/pulse-big-data-smokes-semweb/.* accessed November 05, 2018, 2014.

[64] S. A. Ghafour, "Méthodes et Outils pour l'Intégration des Ontologies," in *Laboratoire d'InfoRmatique en Images et Systèmes d'information LIRIS, Lyon*, 2004.

[65] D. Zouhir, "Donner une Autre vie à Vos besoins fonctionnels: une approche dirigée par l ' entreposage et l ' analyse en ligne,"

*Doctort. dissertation. ISAE-ENSMA Ec. Natl. Supérieure Mécanique d'Aérotechique-Poitiers*, 2017.

[66] H. S. Pinto and P. Martins, "A Methodology for Ontology Integration," in *Proceedings of the 1st international conference on Knowledge capture, ACM*, 2001, 131–138.

[67] I. Horrocks, M. Giese, E. Kharlamov, and A. Waaler, "Using Semantic Technology to Tame the Data Variety Challenge," *IEEE Internet Comput*, 2016, 20 (6), 62–66.

[68] L. Ding, P. Kolari, Z. Ding, S. Avancha, T. Finin, and A. Joshi, "Using Ontologies in the Semantic Web: A Survey," *Ontol. Springer, Boston, MA*, 2007, 79–113.

[69] J. Hui, L. Li, and Z. Zhang, "Integration of Big Data: A Survey," in *International Conference of Pioneering Computer Scientists, Engineers and Educators. Springer, Singapore*, 2018, 1, 101–121.

[70] D. Calvanese, "Ontologies for Data Integration," in *IJCAI Workshop on Formal Ontologies for Artificial Intelligence (FOfAI). Italy*, 2015, 1–67.

[71] F. T. Imam, "Application of Ontologies in Cloud Computing: The State-Of-The-Art," *arXiv Prepr. arXiv1610.02333*, 2016.

[72] M. Y. Mehta, "Big Data Mining and Semantic Technologies: Challenges and Opportunities," *Int. J. Recent Innov. Trends Comput. Commun*, 2015. 3 (7), 4907–4913.

[73] H. Liyanage, P. Krause, and S. de Lusignan, "Using Ontologies to Improve Semantic Interoperability in Health Data," *Innov Heal. Inf*, 2015, 22 (2), 309–305.

[74] M. Bermúdez-Edo, E. Della Valle, and T. Palpanas, "Semantic Challenges for the Variety and Velocity Dimensions of Big Data," *Int. J. Semant. Web Inf. Syst*, 2016, 12 (4), 2016.

[75] E. Mezghani, E. Exposito, K. Drira, and M. Da Silveira, "A Semantic Big Data Platform for Integrating Heterogeneous Wearable Data in Healthcare," *J. Med. Syst*, 2015, 39 (12), 185.

[76] G. Mehdi, S. Brandt, M. Roshchin, and T. Runkler, "Semantic Framework for Industrial Analytics and Diagnostics," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2012, 4016–4017.

[77] A. M. Saettler, K. R. Llanes, P. Ivson, D. L. M. Nascimento, and E. T. L. Corseuil, "An Ontology-Driven Framework for Data Integration and Dynamic Service Composition: Case Study in the Oil & Gas Industry," in *16th International Conference WWW/Internet · ICWI 2017*, 2017, May 2018.

[78] M. El Hamdouni, H. Hanafi, A. Bouktib, and M. Bahra, "Sentiment Analysis in Social Media with a Semantic Web-Based Approach: Application to the French Presidential Elections 2017," in *Proceedings of the Mediterranean Symposium on Smart City Applications. Springer, Cham,* 2018, 470–482.

[79] A. I. Jony, "Applications of Real-Time Big Data Analytics," *Int. J. Comput. Appl*, 2016, 144 (5), 1–5.

[80] N. Mohamed and J. Al-jaroodi, "Real-Time Big Data Analytics: Applications and Challenges," in *High-Performance Computing & Simulation (HPCS), 2014 International Conference on IEEE.*, 2014, 305–310.

[81] L. Leenen, C. Peninsula, C. Town, and S. Africa, "Semantic Technologies and Big Data Analytics for Cyber Defence," *Int. J. Cyber Warf. Terror*, 2016, 6 (3), 53–64.

[82] M. Kiran, P. Murphy, I. Monga, J. Dugan, and S. S. Baveja, "Lambda Architecture for Cost-Effective Batch and Speed Big Data Processing". In Big Data (Big Data), 2015 IEEE International Conference on, 2785–2792.

## Biographies

**Zaoui Sayah** is a Ph.D. Student and research associate in the Artificial Intelligence and Information Technologies Laboratory (LINATI). Department of Computer Science and Information Technologies, Ouargla University, Algeria. He has a scientific BAC in 2001. He received his DEUA in computer science from Biskra University, Algeria in 2004. He earned his license and Master degrees in distributed system and artificial intelligence in 2014 and 2016 respectively from El oued University, Algeria. His current research interests include Big Data, Ontology, MAS, IoT, energy saving.

**Okba Kazar** obtained his magister degree in 1997 from the Constantine University (Algeria) by working on artificial intelligence field. He obtained his PhD degree from the same university in 2005. He is member of editorial board of some Journals. He published more than 307 papers in international journals and communication in international conference. He participate as a session chair in international conferences, and he also published a book "Manual d'Intelligence artificielle", Bigdata security" and five chapters book. His main research field is artificial intelligence, and he is interested in the multiagents systems and their applications, PHM in medical and industrial fields, ERP, advanced information systems, Web services, semantic Web, bigdata, Internet of things, and cloud computing. Actually, Okba KAZAR is a full professor at computer science department of Biskra University and director of smart computer science laboratory (LINFI).

**Ahmed Ghenabzia** is a Ph.D. student and research associate in the Artificial Intelligence and Information Technologies Laboratory (LINATI) at Kasdi Merbah University-Ouargla, Algeria. he received his IT engineering from ESI in Algiers, Algeria in 2013. He earned his licence and Master degrees in distributed system and artificial intelligence in 2016 respectively from El Oued University, Algeria. His current research interests include Data science, Big Data, multi-agent systems, IoT, Ontology and renewable energy.