


Towards a Taxonomy for Big Data Technological Ecosystem

Vitor Afonso Pinto¹ ^a and Fernando Silva Parreiras² ^b

¹Technology Department, Operational Technology for Mine, Plant and Expedition, Vale Mozambique, Tete, Mozambique

²Laboratory for Advanced Information Systems, FUMEC University, Rua do Cobre, Belo Horizonte, Brazil

Keywords: Big Data, Taxonomy, Mapping Study, Cross-case Analysis.

Abstract: Data is constantly created, and at an ever-increasing rate. Intending to be more and more data-driven, companies are struggling to adopt Big Data technologies. Nevertheless, choosing an appropriate technology to deal with specific business requirements becomes a complex task, specially because it involves different kinds of specialists. Additionally, the term Big Data is vague and ill defined. This lack of concepts and standards creates a fuzzy environment where companies do not know what exactly they need to do and on the other hand consultants do not know how to help them to achieve their goals. In this study the following research question was addressed: Which essential components characterize Big Data ecosystem? To answer this question, Big Data terms and concepts were first identified. Next, all terms and concepts were related and grouped creating a hierarchical taxonomy. Thus, this artifact was validated through a classification of tools available in the market. This work contributes to clarification of terminologies related to Big Data, facilitating its dissemination and usage across research fields. The results of this study can contribute to reduce time and costs for Big Data adoption in different industries as it helps to establish a common ground for the parts involved.

1 INTRODUCTION


Emerging technologies have made all devices, equipment, and systems to be smart, communicable, and integrated. Because of that, data is constantly created, and at an ever-increasing rate. Mobile phones, social media, imaging technologies and other examples create new data which must be stored somewhere for future usage (Dietrich et al., 2015). The total amount of data in the world increased from 2.8 zettabytes in 2012 to 8 zettabytes by 2015 (Duckett, 2016) and is expected to reach 44 zettabytes by 2020. Organizations are now carrying out studies that were impossible to conduct in the past due to data availability (Liu et al., 2016).


If on the one hand there is data available, on the other hand a solid definition of Big Data is still required. The term Big Data is vague and ill defined. It is not a precise term and does not carry a particular meaning other than the notion of its size (Demchenko et al., 2014). Big Data has been variously defined in the literature however these definitions lack ontological clarity. This lack of concepts along with an increasing list of new technologies creates a fuzzy en-

vironment for organizations that want to process data in their best interest.

Data-driven technologies continue to evolve at a rapid pace, with an ever vibrant ecosystem of startups, products and projects. The "2019 Data & AI Landscape", compiled by Matt Turck, includes 1335 Big Data companies (Turck, 2018). From the organizational point of view, decision makers need to navigate the myriad choices in compute and storage infrastructures as well as data analytics techniques, and security and privacy frameworks. Thus, choosing appropriate technologies to deal with specific business requirements may become a complex task. This study aims to explain existing roles for technologies inside a Big Data technological ecosystem.

This research intends to answer the following research question: "Which essential components characterize Big Data ecosystem?". By addressing this research question, this study make three contributions: first, main components related to Big Data are identified in the literature. Second, a taxonomy for classifying Big Data tools is formalized. Finally, the taxonomy is evaluated and used to classify real-world Big Data tools. By using the proposed taxonomy, organizations should be able to understand the role of each Big Data tool and how they fit in a Big Data technological ecosystem. This study is structured as

^a  <https://orcid.org/0000-0002-2731-0952>

^b  <https://orcid.org/0000-0002-9832-1501>

follows: Section 2 presents methods used in this research. In section 3, research results are presented. These results are discussed in section 4. This study is concluded and future work is suggested in section 5.

2 METHODS

This study follows the Design Science Research (DSR) approach. Design science research is a research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the body of scientific evidence (Hevner and Chatterjee, 2010). In this scenario, this research intends to formalize an artifact capable to answer the following research question: *"Which essential components characterize Big Data ecosystem?"*. In order to answer this research question, this study is divided in three steps. Firstly, Big Data constructs are raised through a systematic mapping study. Then, these constructs are ordered through a taxonomy creation process. Finally, the taxonomy artifact is validated through a cross-case study using existing tools available in the market. Figure 1 presents the methods applied on each research step and highlights how overall results are related to outcomes of each step.

2.1 Systematic Mapping Study

The outcome of a mapping study is an inventory of papers on the topic area, mapped to a classification. Hence, a mapping study provides an overview of the scope of the area, and allows to discover research gaps and trends. The main research question defined to be addressed by this research step was: *"What prior knowledge is available about Big Data taxonomy?"*. As research strategy, string "Big Data Taxonomy" was used to search on Mendeley catalog and references were manually added. Studies returned by automatic search were included, except those not written in English, not related to research topic or duplicated. For data extraction, we developed an extraction form using software START, intending to identify main constructs related to Big Data (Fabbri et al., 2016). Figure 2 illustrates transformation of raw results in the final list of primary studies.

2.2 Taxonomy Creation

In this research step, we analyzed the dataset created in section 2.1 to collect terms, group similar concepts and add relationships. The main research question defined to be addressed by this research step was: *"How*

can we represent a Big Data taxonomy?". In this step, guidelines proposed by (Redmond, 2013) were applied to build a Big Data taxonomy. Thus, scope was defined and then terms and concepts were collected. Next, concepts were grouped and related to each other.

2.3 Cross-case Analysis

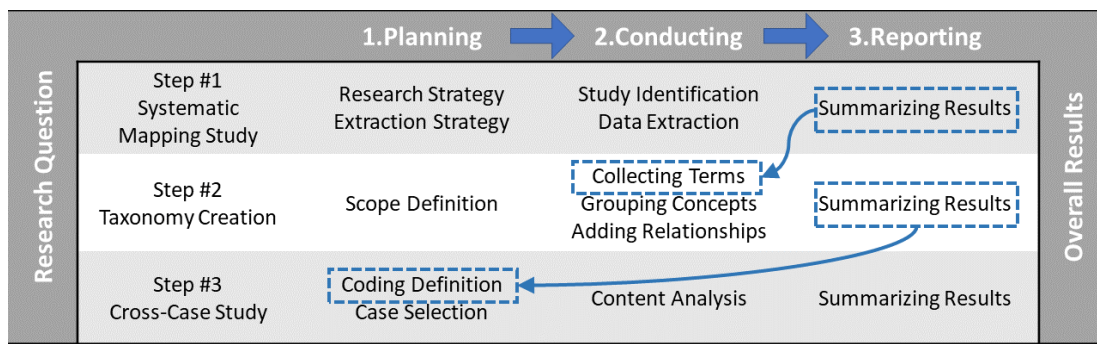
In this step of the research, procedures followed a qualitative approach through an exploratory descriptive cross-case study research. The main research question of this step was defined as: *"Is the proposed taxonomy sufficient for describing components of existing tools?"*. For data collection, websites of Big Data products were analysed and scrutinised according to categories proposed in section 2.2. We assessed fifty websites of randomly selected tools from landscape compiled by Matt Turck (Turck, 2018). Table 1 presents a list of tools selected for taxonomy validation. Considering the volume of collected data, content analysis techniques were used to facilitate the understanding and this process was supported by WebQDA software (WebQDA, 2019).

3 RESULTS

This section presents answers for each research question raised on each step of this study. Firstly Big Data constructs raised through the systematic mapping study conducted in section 2.1 are detailed. Then, the taxonomy artifact built in section 2.2 is showed. Finally, results of cross-case study conducted in section 2.3 are presented.

3.1 What Prior Knowledge is Available about Big Data Taxonomy?

The Systematic Mapping Study, presented in section 2.1 unveiled ten major roles technologies may perform inside a Big Data technological ecosystem: **data creation, data acquisition, data transmission, data ingestion, data storage, data preprocessing, abstraction middleware, data analytics, data applications** and **computing infrastructure**. Figure 3 shows how these components were covered by selected studies. Next subsections present qualitative details of each role considered by this study along with its categories.



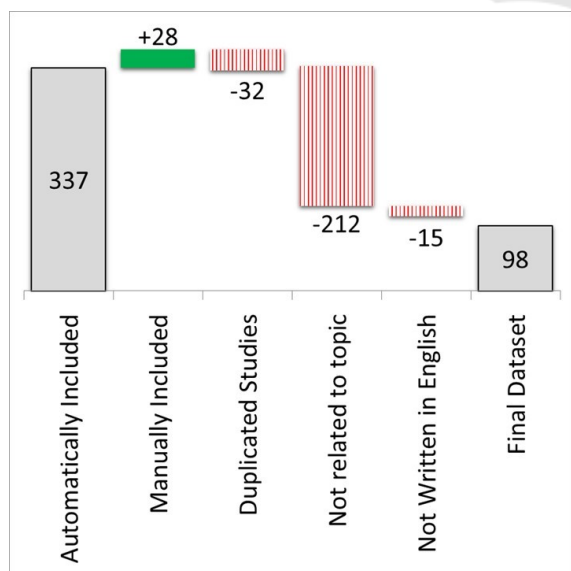
Source: Authors

Figure 1: Applied methods on this study.

3.1.1 Technologies for Data Creation

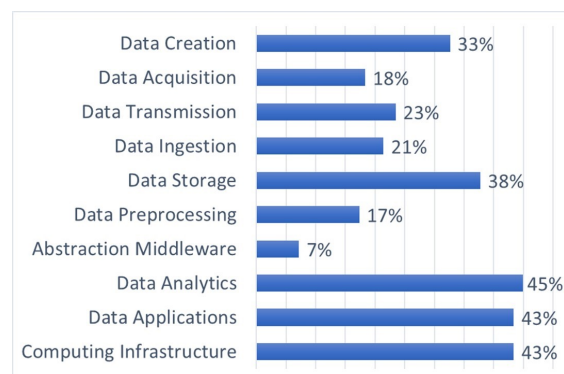
In order to discuss Big Data, it is relevant to understand how data is created. Technologies that perform this role are responsible for creating data. Considering just the internet, data is constantly being generated by posts in forums, blog messages, searching queries, among others. Online social networks are immensely prevalent and have now become a ubiquitous and important part of the modern, developed society (Howden et al., 2014). With the rapid proliferation of various social network services, it has become very common for people to express their thoughts or opinions on various issues using brief comments (Lee, 2016). At the same time, with the trend going on in ubiquitous computing, everything is going to be connected to the Internet and its data will be used for various progressive purposes. Most real-world data are not in

a form that can be directly recorded by a computer. These quantities typically include temperature, pressure, distance, velocity, mass, and energy output (such as optical, acoustic, and electrical energy) (Austerlitz, 2003). A physical quantity must first be converted to an electrical quantity (voltage, current, or resistance) using a sensor or transducer. Thus, transducers and sensors are used to convert a physical phenomenon into an electrical signal (voltage or current) that will be then converted into a digital signal used for the next stage such as a computer, digital system, or memory board (Emilio, 2013). In parallel, value also relies on corporate data, that is, all data maintained by any of the companies including, but not limited to, data related to its finances, taxes, employees, customers, suppliers and the business. Corporate data may reveal answers to most problems organizations face. Additionally, data generated by automated interactions between systems can also generate value as lots of events are registered in log files and each event can be related to a specific situation or to meta-data, for instance. Big Data typically consists of data from a variety of related and unrelated sources that can be quite complex. Table 2 presents a summary for each category of Data Creation, found on literature.



Source: Authors

Figure 2: Selection of Primary Studies.



Source: Authors

Figure 3: Coverage of Big Data Constructs by Publications.

Table 1: Selected Tools for Taxonomy Validation.

ID	Assessed Tool	Reference	ID	Assessed Tool	Reference
1	1010Data	(1010Data, 2019)	26	Estimote	(Estimote, 2019)
2	23andMe	(23andMe, 2019)	27	Garmin	(Garmin, 2019)
3	3Scan	(3Scan, 2019)	28	Cloud Dataflow	(Google, 2019)
4	6Sense	(6Sense, 2019)	29	Helium	(Helium, 2019)
5	Actifio	(Actifio, 2019)	30	Human API	(HumanAPI, 2019)
6	ActionIQ	(ActionIQ, 2019)	31	Illumio	(Illumio, 2019)
7	Active.AI	(Active.Ai, 2019)	32	MariaDB	(MariaDB.org, 2019)
8	Acxiom	(Acxiom, 2019)	33	MEMSQL	(MemSQL, 2019)
9	Aerospike	(Aerospike, 2019)	34	Microstrategy	(MicroStrategy, 2019)
10	Affirm	(Affirm, 2019)	35	MongoDB	(MongoDB, 2019)
11	AiCure	(AiCure, 2019)	36	Neo4j	(Neo4J, 2019)
12	Airobotics	(Airobotics, 2019)	37	Objectivity	(Objectivity, 2019)
13	Airtable	(Airtable, 2019)	38	OpenTSDB	(OpenTSDB, 2019)
14	Apache Drill	(Apache, 2019a)	39	PTC	(PTC, 2019)
15	Apache Flink	(Apache, 2019d)	40	Recorded	(RecordedFuture, 2019)
16	Apache Hive	(Apache, 2019b)	41	Riak	(Riak, 2019)
17	Apache Mesos	(Apache, 2019c)	42	Samsara	(Samsara, 2019)
18	Augury	(Augury, 2019)	43	SecurityScorecard	(SecurityScorecard, 2019)
19	Berkeley	(Oracle, 2019)	44	SentinelOne	(SentinelOne, 2019)
20	Birst	(Birst, 2019)	45	Sentry	(Sentry, 2019)
21	Cignifi	(Cignifi, 2019)	46	Sift	(Sift, 2019)
22	Cloudera	(Cloudera, 2019)	47	Signifyd	(Signifyd, 2019)
23	CyberX	(CyberX, 2019)	48	SlamData	(SlamData, 2019)
24	Darktrace	(Darktrace, 2019)	49	SparkCognition	(SparkCognition, 2019)
25	Elastic	(Elastic, 2019)	50	Uptake	(Uptake, 2019)

3.1.2 Technologies for Data Acquisition

Technologies that perform this role acquire Big Data from multiple sources. Data may have different structures, depending on how they are created, for instance. Structured data is both highly-organized and easy to digest. Traditional structured data, such as the transaction data in financial systems and other business applications, conforms to a rigid format to ensure consistency in processing and analyzing it. Unstructured data cannot simply be recorded in a data table, thus typically it is not a good fit for a mainstream relational database and requires more specialized skills and tools to work with. Semi-structured data is a data type that contains semantic tags, but does not conform to the structure associated with typical relational databases (Sawant and Shah, 2013). Big Data is the amount of structured, semi-structured and unstructured data coming from multiple sources such as online access, mobile devices, social media, scientific devices, and other inputs in addition to existing, traditional data sources (Bari et al., 2014). Acquiring data from separate data sources consists in accessing several applications. Some applications may be custom developed in-house while others are bought from

third-party vendors. The applications probably run on multiple computers, which may represent multiple platforms, and may be geographically dispersed. Some of the applications may be run outside of the enterprise by business partners or customers. Some applications may need to be integrated even though they were not designed for integration and cannot be changed. These issues and others like them are what make data acquisition difficult. Table 3 presents a brief description for each category of Data Acquisition.

3.1.3 Technologies for Data Transmission

The movement of data between origin and destination is possible because of technologies that perform this role. Information from new sources needs to be transmitted from their origin to a place where it can be processed or consumed. Bandwidth should be a critical element of Big Data strategies, because it is not possible to support the heavy traffic demands of streamed or file-based Big Data payloads without the pipelines needed to carry them. Bandwidth means the maximum amount of data transmitted through a communication channel at one time. Data transmis-

Table 2: Categories of Data Creation Component.

Category	Description
Social Interactions	Data generated by individuals or group of individuals. It also includes social network interactions, scientific researches, among others
Sensor Readings	Data generated by sensors, also known as “things”. Includes data that come from industry, agriculture, traffic, transportation, medical care, public departments, among others
Corporate Data	Content generated by enterprises operations. Includes production data, inventory data, sales data, financial data, etc
Systems Interactions	Data automatically generated by systems interactions. Includes metadata, log files, online trading data, etc
Data as a Service	Data collected, assessed and sold for usage as input for data-consumers applications

Source: Authors

Table 3: Categories of Data Acquisition Component.

Category	Description
Web Search	Log analysis, page tagging, linked data, among others
Messaging	Transferring of data packets in a frequently, immediately, reliably and asynchronously way, using customizing formats
Remote Procedure Invocation	Interfaces that allow interaction between any application to running applications
Database Integration	Two or more applications share common tables. One may write to it while the other while the other simply reads from these tables
File Transfer	Two or more applications share a common file system. One may write to it while the other may poll the file system to read it

Source: Authors

sion is the movement of data (bits) between at least two digital devices. Wired transmission can be implemented through twisted-pair wire (copper wire used for telephone and data communication), coaxial cable (consists of copper wire surrounded by insulation and braided wire), Fiber-optic cable (consists of thin strands of glass or plastic that carry data through pulses of light), among others. Wireless transmission

Table 4: Categories of Data Transmission Component.

Category	Description
Wired	Depends on physical cabling to enable data transmission. May include: Ethernet, Serial, USB, FireWire, among others
Wireless	Does not depend on physical cabling to enable data transmission. May include: Bluetooth, WiFi 802.11, GSM, ZigBee, RFID, LoRa, NB-IoT, Satellite, among others.

Source: Authors

can be implemented through infrared (Wireless transmission medium that carries data through the air using light beams), radio and Bluetooth (enables music, photos, and voice to travel through the air as radio frequency or radio waves), microwaves (transmit data via electromagnetic radio waves with short frequencies), satellites (microwave relay stations in space that transmit data through microwave signals), among others. Table 4 presents a brief description for each category of Data Transmission.

3.1.4 Technologies for Data Ingestion

Technologies that perform this role determine strategies for handling acquired data. Different dynamics of data may require different security approaches or different computing platforms to provide meaningful insights. The type of technology required to deal with data at rest or data in motion may be different. Data at rest is placed in storage rather than used in real time and requires batch processing. Data in motion (or streaming data) moves across a network or in-memory for processing in real time. Streaming data means high speed both in data arrival rate and in data processing. There are examples of streaming data ranging from data coming from equipment sensors to medical devices to temperature sensors to stock market financial data and video streams. This aspect also determines latency strategies and data workflow. **Latency** is the time it takes for data packets to be stored or retrieved. It needs to be adjusted to answer business requirements. An adequate data latency allows organizations to make business decisions in a timely manner. **Workflows** are used to allocate and schedule execution of Big Data applications in an optimized manner (Rani and Babu, 2015). These engines provide an effective tool to define and manage large sets of processing tasks (Palazzo et al., 2015). Workflow management systems enable the creation and the execution of adaptive analytics (Kantere and Filatov, 2015). Increased volume of streaming data as well as the demand for more complex real-time analytics

Table 5: Categories of Data Ingestion Component.

Category	Description
Data in motion	Data in motion is collected and processed in real-time as the data-creating event happens
Data at rest	Data at rest is collected in batches as sets of records and processed as a unit after the data-creating event has occurred

Source: Authors

require for execution of processing pipelines among heterogeneous event processing engines as a workflow (Ishizuka et al., 2016). Table 5 presents a brief description for each category of Data Ingestion.

3.1.5 Technologies for Data Storage

Technologies that perform this role organize elements of data. A storage model is the core of any big-data related systems. It affects the scalability, data-structures, programming and computational models for the systems built on top of any Big Data-related systems. Different data systems implement different storage models. Each one has advantages and disadvantages. NoSQL data models: Key-Value, Column families and Document-based models has looser consistency constraints as a trade-off for high availability and/or partition-tolerance in comparison with that of relational data models. In addition, NoSQL data models have more dynamic and flexible schemas based on their data models while relational databases use predefined and row-based schemas. Lastly, NoSQL databases apply the BASE models while relational databases guarantee ACID transactions. There are two major categories to represent stored data: relational and NoSQL, acronym for "not only SQL". Table 6 presents a brief description for each category of Data Storage.

3.1.6 Technologies for Data Preprocessing

Technologies that perform this role deal with quality of data used to create models. Data preprocessing is one of the most time-consuming steps in a typical data mining project (Luis, 2017). Real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results (Han et al., 2012). Much of the raw data contained in databases is unprocessed, incomplete, and noisy. For example, the databases may contain: a) fields that are obsolete or redundant; b) missing values; c) outliers; d) data in

Table 6: Categories of Data Storage Component.

Category	Subcategory	Description
Relational	Conventional	Data is represented in terms of tuples and grouped into relations with high focus on relational operations and transactions
	In Memory	Also known as NewSQL. Data is represented in terms of tuples and grouped into relations with lower focus on relational operations and transactions
No SQL	Key Value	Key-Value pairs, in which, keys are unique IDs for each data and also work as indexes during accessing the data
	Document	The unit of data is called a document which can contain an arbitrary set of fields, values and even nested objects and arrays
	Column	Data are considered as tables with rows and column families in which both rows and columns can be split over multiple nodes
	Graph	Uses graph structures with nodes, edges, and properties to represent and store data

Source: Authors

a form not suitable for the data mining models; e) values not consistent with policy or common sense (Larose and Larose, 2015). Dimensionality reduction is relevant to decrease the computational cost of models, increase the performance of models, reduce irrelevant and redundant dimensions (Garcia et al., 2015). Table 7 presents a brief description for each category of Data Preprocessing.

3.1.7 Technologies for Data Analytics

Technologies that perform this role enable the creation, improvement and deployment of models and applications. A model is a representation of a state, process, or system that we want to understand and reason about. Models can be equations linking quan-

Table 7: Categories of Data Preprocessing Component.

Category	Description
Transformation	Technologies for enriching data in order to make analysis more effective, focusing on cleaning up the data or creating new variables that may bring useful information for the analysis steps
Reduction	Technologies to reduce volume of data by using: principal component analysis, factor analysis, multidimensional analysis, sampling rows, variable selection, feature and instance selection, among others

Source: Authors

tities that we can observe or measure. They can also be a set of rules (Forte, 2015). When approaching a data mining problem, a data mining analyst may already have some a priori hypotheses that he or she would like to test regarding the relationships between the variables. However, analysts do not always have a priori notions of the expected relationships among the variables (Larose and Larose, 2015). If the client is a human, it is common to use a variety of models, tuned in different ways, to examine different aspects of data. If the client is a machine though, it will be probably needed to zero in on a single, canonical model that will be used in production (Cady, 2017). Model evaluation is the process of assessing a property or properties of a model in terms of its structure and data inputs so as to determine whether or not the results can be used in decision making. It encompasses: (1) verification, validation, and quality control of the usability of the model and its readiness for use, and (2) investigations into the assumptions and limitations of the model, its appropriate uses, and why it produces the results it does. (Gass and Harris, 2001). The model is assessed in three stages: business evaluation, statistical validation and application on the full population including the corresponding target variables. (Ahlemeyer-Stubbe and Coleman, 2014). Once there is confidence on the quality of data mining procedures, they need to be communicated. This frequently involves: 1) some sort of reporting to other people within some organization and/or 2) trying to deploy the outcome of data mining workflow. (Luis, 2017) Building a model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a re-

Table 8: Categories of Data Analytics Component.

Category	Description
Data Modeling	Technologies for data modeling. Core components of a model are: a set of equations with parameters that need to be tuned; some data that are representative of a system or process; a concept that describes the model's goodness of fit; a method to update the parameters to improve the model's goodness of fit
App Development	Technologies for development, validation and deployment of applications that encapsulate models. Validation include: holdout, random subsampling, cross validation, bootstrap estimates, lift and gain charts, model stability, sensitivity analysis, threshold analytics and confusion matrix, ROC curves, model complexity, among others. Deployment provides the option to deploy the analytical results in to every day decision making process

Source: Authors

peatable data mining process. (Sayad, 2010). Table 8 presents a brief description for each category of Data Analytics.

3.1.8 Technologies for Abstraction Middleware

Technologies that perform this role provide an abstraction layer interposed between the IT infrastructure and the applications. A Middleware platform aims to hide the technological details to enable the application developers to focus on the development of the applications (Chaqfeh and Mohamed, 2012). When billions of sensors are connected to the Internet, it is not feasible for people to process all the data collected by those sensors. Context-awareness computing techniques, such as middleware, are proposed to better understand sensor data and help decide what data needs to be processed (Xu et al., 2014). Table 9 presents a brief description for each category of Abstraction Middleware.

3.1.9 Technologies for Data Applications

Technologies that perform this role extract worthy insights from low-value data. According to (Gartner, 2015), Big Data applications can be grouped in four dimensions: Descriptive, Diagnostic, Predictive and

Table 9: Categories of Abstraction Middleware Component.

Category	Description
Interoperation	Technologies that allow the usage and sharing of information across diverse domains of applications using diverse communication interfaces
Context Detection	Technologies for characterizing the situation of an entity where an entity can be person, place, or object relevant to the interaction between a user and an application, including the user and applications themselves
Security	Technologies to ensure confidentiality, authenticity, and nonrepudiation across diverse domains of applications
Portability	Technologies to enable organizations migrating their applications and services to different platforms.
Device Discovery	Technologies for enabling any device in the IoT network to detect all its neighbouring devices and make its presence known to each neighbour in the network.

Source: Authors

Prescriptive. **Descriptive** dimension means the examination of data or content, usually manually performed, to answer the question “What happened?” (or “What is happening?”). **Diagnostic** dimension is a form of advanced analytics which examines data or content to answer the question “Why did it happen?”. **Predictive** dimension is a form of advanced analytics which examines data or content to answer the question “What is going to happen?” or more precisely, “What is likely to happen?”. **Prescriptive** dimension examines data or content to answer the question “What should be done?” or “What can be done to make a specific result happen?”. Table 10 presents a brief description for each category of Applications.

3.1.10 Technologies for Computing Processing

Technologies that perform this role deal with data processing and supports all technologies inside Big Data technological ecosystem. Computing paradigms on Big Data currently differ at the first level of abstraction on whether the processing will be done in batch mode, or in real-time/near real-time on streaming data (data that is constantly coming in and needs to be processed right away). If an application demands “immediate” response to each event as it occurs, some

Table 10: Categories of Applications Component.

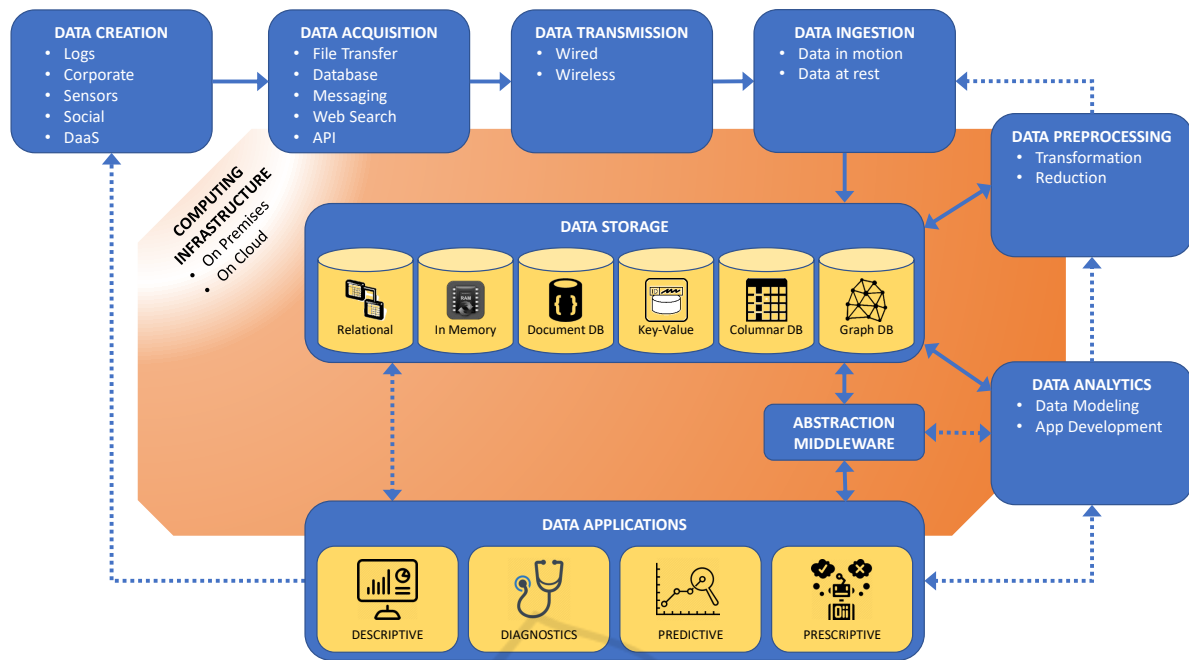
Category	Description
Descriptive	Characterized by traditional business intelligence (BI) and visualizations such as pie charts, bar charts, line graphs, tables, or generated narratives
Diagnostic	Characterized by techniques such as drill-down, data discovery, data mining and correlations
Predictive	characterized by techniques such as regression analysis, forecasting, multivariate statistics, pattern matching, predictive modeling, and forecasting
Prescriptive	Characterized by techniques such as graph analysis, simulation, complex event processing, neural networks, recommendation engines, heuristics, and machine learning

Source: Authors

form of stream processing is needed, which essentially processes the data as it comes in (Murthy et al., 2014). For this reason, emerging technologies are largely engaged in processing Big Data using different computational environments. The parallel and cloud computing platforms are considered a better solution for Big Data. The concept of parallel computing is based on dividing a large problem into smaller ones and each of them is carried out by one single processor individually. In addition, these processes are performed concurrently in a distributed and parallel manner (Tsai et al., 2016). In data parallelism, each node executes the same task on different pieces of distributed data. Task parallelism focuses on distributing execution processes across different parallel computing nodes. Graph parallelism consists of systems that encode computation as vertex programs which run in parallel and interact along edges in the graph. Computing processing in Big Data can be implemented on premises or on cloud. Table 11 presents a brief description for each category of Computing Processing.

3.2 How Can We Represent a Big Data Taxonomy?

Based on the literature, we consider Big Data as a group of technological components that interact to each other. **Data acquisition** component intends to collect data from several data sources. **Data Transmission** component deals with transferring data from an origin to a destination. **Data Ingestion** handles data sending them to **Data Storage**. **Data Prepro-**



Source: Authors

Figure 4: The Big Data Technological Ecosystem.

Table 11: Categories of Computing Processing Component.

Category	Description
On Premises	Each company manage each own Big Data computing processing using internal infrastructure
On Cloud	Companies pay for using computing processing infrastructure from vendors. In public clouds, vendors manage their proprietary data-centers delivering services built on top of them. In private clouds, vendors provide services deployed over a company intranet or in a private data center. In hybrid clouds there is a composition of two or more (private or public) clouds that remain different entities but are linked together

Source: Authors

cessing component interacts with stored data to address missing and noisy data. Depending on the process, an interaction between Data Preprocessing and Data ingestion may occur in order to adjust data processing. **Data Analytics** component cares about data exploration, development and deployment of applications. It can request adjustments in Data Preprocessing. **Data Applications** component extracts value from stored data. It can create more data, creating

an interaction with **Data Creation** aspect. It can request adjustments in deployed applications, creating interaction with Data Analytics component. Data Applications component can also interact with **Abstraction Middleware** which may interact with Data Storage aspect. **Computing Infrastructure** aspect interacts with all other aspects as every component needs computing power to perform its tasks.

After identifying main constructs related to Big Data, a hierarchical taxonomy for Big Data components was created as presented in figure 4. This taxonomy considers the roles performed by technologies inside the Big Data technological ecosystem, that is, all Big Data components and their interactions in a high level.

3.3 Is the Proposed Taxonomy Sufficient for Describing Existing Tools?

Considering the proposed taxonomy as reference we assessed fifty randomly selected Big Data tools to verify if this taxonomy is sufficient for describing real-world existing tools, as described in Table 1. In a general way, all Big Data tools were fully explained and their features were categorized into one or more categories. Most part of tools presented features that allow the development, deployment or the usage of applications. There was a single category not used to explain any of the selected tools: Wired Data Trans-

Table 12: Results of Big Data Taxonomy Validation.

Component	Category	Ratio
Data Creation	Sensors	26%
	DaaS	16%
	Logs	10%
	Social	10%
	Corporate	6%
Data Acquisition	Web Search	24%
	API	20%
	Messaging	16%
	File Transfer	10%
	Database Integration	10%
Data Transmission	Wireless	6%
	Wired	0%
Data Ingestion	Data in Motion	22%
	Data at Rest	18%
Data Storage	In Memory	10%
	Key Value Pair	8%
	Columnar DB	8%
	Relational	6%
	Graph DB	6%
	Document DB	4%
Data Preprocessing	Transformation	20%
	Reduction	8%
Abstraction Middleware	Security	22%
	Context Detection	10%
	Interoperation	8%
	Portability	6%
	Device Discovery	2%
Data Analytics	Data Modeling	28%
	App Development	24%
Data Applications	Predictive	30%
	Prescriptive	24%
	Descriptive	22%
	Diagnostic	20%
Computing Infrastructure	On Cloud	18%
	On Premises	8%

Source: Authors

mission. Table 12 summarizes results of this qualitative assessment. Column "Ratio" shows percentage of tools classified on each category.

4 DISCUSSION

Big Data can be defined as an integrated ecosystem of technologies performing formal roles with the purpose to create technical conditions for the delivery of value-added applications based on data. As an integrated ecosystem, each technology has direct or indirect effect both in other technologies and also in final applications. The roles of technologies in Big Data Technological Ecosystem can be divided in three ma-

major groups: 1) data gathering; 2) application development; 3) computing infrastructure. The first group enables the creation, acquisition, transmission, ingestion and storage of data. The second group enables both preprocessing and analysis of data as well as the development and deployment of value-added applications. Computing infrastructure supports the whole ecosystem.

In this regard, each technology should be minutely chosen in order to extract the most of Big Data initiatives and to avoid loss of effectiveness and unnecessary investments. Value-added applications depend on certain technical conditions to be developed and deployed. On the one hand, Big Data is focused on the technological perspective and explains how technologies should be organized in order to enable value creation in Data Science initiatives, for instance. On the other hand, it seems that organizations need to create processes and structures to stimulate the development and deployment of value-added applications on top of Big Data technologies.

It is important to highlight that, according to this definition, there is not an ideal Big Data technological ecosystem. This may explain the difficulties to define the term Big Data so far. Although Big Data comprises technologies performing formal roles, the technologies chosen to perform each role may vary among organizations. This creates infinite possibilities as each organization is free to define its own Big Data technological ecosystem based on the most appropriate technologies for their case. In this regard, taxonomy proposed in this study may be used to help organizations to choose technologies that best suit their own interest as it can be used as a reference for comparison of Big Data technologies. In practical terms, organizations may use this study to compare features of different Big Data technologies, keeping in mind the importance of defining at least one tool for each role of Big Data technological ecosystem.

5 CONCLUSION

Components of Big Data technological ecosystem can be classified in ten categories: data creation, data acquisition, data transmission, data ingestion, data storage, data preprocessing, data modeling, abstraction middleware, data applications and computing infrastructure. The taxonomy proposed in this study explained terms and concepts related to these aspects based on literature. More than that, taxonomy was sufficient to classify fifty Big Data tools, randomly selected for this study. This work contributes to the clarification of concepts and terminologies related to Big

Data and facilitates dissemination and usage of Big Data across research fields. Additionally, this work helps to establish a common ground for all parts involved in the whole Big Data technological ecosystem. In this regard, knowing the Big Data taxonomy proposed in this study may direct attention for each required aspect. The results of this study can contribute to reduce the lack of vocabulary related to Big Data and help companies to leverage Big Data initiatives. Taxonomy proposed in this study may be used to help organizations to choose technologies that best suit their own interest as it can be used as a reference for comparison of Big Data technologies.

This study has limits as it described Big Data ecosystem from a technological perspective only. In this context, no managerial, social or organizational aspect was considered. In this regard, word "ecosystem" was used to explain only the technological aspects of Big Data. Thus, management processes that oversight availability, usability, integrity and security of data were not discussed here. Additionally, only publications written in English had been considered. It is important to highlight that during the "collect terms and concepts" stage, it was necessary to interpret subjective information provided by publications as they did not present objective details regarding the topics analyzed. Future works could expand the proposed taxonomy, creating a Big Data ontology or thesaurus, extending this classification.

REFERENCES

- 1010Data (2019). Self-service platform for data management, analytics and application building.
- 23andMe (2019). Dna genetic testing & analysis.
- 3Scan (2019). Discover in 3d with 3scan whole tissue digitization and exploration.
- 6Sense (2019). Abm orchestration platform.
- Actifio (2019). Enterprise cloud data management.
- ActionIQ (2019). Enterprise customer data platform - cdp.
- Active.Ai (2019). Enterprise ai platform for financial services.
- Axiom (2019). Identity resolution & people-based marketing.
- Aerospike (2019). Aerospike.
- Affirm (2019). Affirm.
- Ahlemeyer-Stubbe, A. and Coleman, S. (2014). *A practical guide to data mining for business and industry*. John Wiley & Sons.
- AiCure (2019). Intelligent observation. better care.
- Airobotics (2019). Automated industrial drones.
- Airtable (2019). Organize anything you can imagine.
- Apache (2019a). Apache drill - schema-free sql for hadoop, nosql and cloud storage.
- Apache (2019b). Apache hive tm.
- Apache (2019c). Apache mesos.
- Apache (2019d). Stateful computations over data streams.
- Augury (2019). Machines talk, we listen.
- Austerlitz, H. (2003). *Data acquisition techniques using PCs*. Academic Press, San Diego, CA.
- Bari, A., Chaouchi, M., and Jung, T. (2014). *Predictive Analytics for Dummies*. Wiley, [s.l.].
- Birst (2019). Business intelligence & analytics, bi software.
- Cady, F. (2017). *The data science handbook*. John Wiley & Sons, Inc., Hoboken, NJ.
- Chaqfeh, M. A. and Mohamed, N. (2012). Challenges in middleware solutions for the internet of things. In *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, pages 21–26. IEEE.
- Cignifi (2019). Cignifi transforms mobile behavior into financial opportunity.
- Cloudera (2019). The enterprise data cloud company.
- CyberX (2019). Ics, scada & ot security for the industrial network.
- Darktrace (2019). Darktrace.
- Demchenko, Y., de Laat, C., and Membrey, P. (2014). Defining architecture components of the big data ecosystem. In *2014 International Conference on Collaboration Technologies and Systems (CTS)*, pages 104–112.
- Dietrich, D., Heller, B., and Yang, B. (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting*. Wiley, [s.l.].
- Duckett, G. (2016). *Data Science: Questions and Answers*. CreateSpace Independent Publishing Platform, [s.l.].
- Elastic (2019). Open source log management.
- Emilio, M. D. P. (2013). *Data Acquisition Systems From Fundamentals to Applied Design*. Springer, New York, NY.
- Estimote (2019). Indoor location with bluetooth beacons and mesh.
- Fabbri, S., Silva, C., Hernandez, E., Octaviano, F., Di Thommazo, A., and Belgamo, A. (2016). Improvements in the start tool to better support the systematic review process. In *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*, pages 1–5.
- Forte, R. M. (2015). *Mastering predictive analytics with R: master the craft of predictive modeling by developing strategy, intuition, and a solid foundation in essential concepts*. Packt Publishing, Birmingham, UK.
- Garcia, S., Herrera, F., and Luengo, J. (2015). *Data preprocessing in data mining*. Springer International Publishing, Cham.
- Garmin (2019). Garmin international.
- Gartner (2015). Descriptive analytics.
- Gass, S. I. and Harris, C. M. (2001). *Encyclopedia of operations research and management science*. Kluwer Academic, Boston.
- Google (2019). Cloud dataflow - stream & batch data processing — cloud dataflow — google cloud.

- Han, J., Kamber, M., and Pei, J. (2012). *Data mining: concepts and techniques*. Morgan Kaufmann, Amsterdam.
- Helium (2019). An internet for machines. built and owned by you.
- Hevner, A. and Chatterjee, S. (2010). *Design research in information systems: theory and practice*, volume 22. Springer Science & Business Media.
- Howden, C., Liu, L., Li, Z., Li, J., and Antonopoulos, N. (2014). Virtual vignettes: the acquisition, analysis, and presentation of social network data. *Science China Information Sciences*, 57(3):1–20.
- HumanAPI (2019). Get health data from everywhere.
- Illumio (2019). Illumio.
- Ishizuka, Y., Chen, W., and Paik, I. (2016). Workflow transformation for real-time big data processing. In *Big Data (BigData Congress), 2016 IEEE International Congress on*, pages 315–318. IEEE.
- Kantere, V. and Filatov, M. (2015). A workflow model for adaptive analytics on big data. In *Big Data (Big-Data Congress), 2015 IEEE International Congress on*, pages 673–676. IEEE.
- Larose, D. T. and Larose, C. D. (2015). *Data mining and predictive analytics*. Wiley, Hoboken, NJ.
- Lee, C. (2016). Guest editorial: Automated big data analysis for social multimedia network environments. *Multimedia Tools and Applications*, 75(20):12663–12667.
- Liu, J., Li, J., Li, W., and Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:134 – 142. Theme issue 'State-of-the-art in photogrammetry, remote sensing and spatial information science'.
- Luis, T. (2017). *Data mining with R: learning with case studies*. CRC Press, Taylor and Francis Group, Boca Raton.
- MariaDB.org (2019). Mariadb.org.
- MemSQL (2019). Memsql is the no-limits database powering modern applications and analytical systems.
- MicroStrategy (2019). Business analytics & mobility solutions.
- MongoDB (2019). The most popular database for modern apps.
- Murthy, P., Bharadwaj, A., Subrahmanyam, P. A., Roy, A., and Rajan, S. (2014). Big data taxonomy.
- Neo4J (2019). Neo4j graph platform – the leader in graph databases.
- Objectivity (2019). Infinitograph - distributed graph database.
- OpenTSDB (2019). A distributed, scalable monitoring system.
- Oracle (2019). Oracle berkeley db.
- Palazzo, C., Mariello, A., Fiore, S., D'Anca, A., Elia, D., Williams, D. N., and Aloisio, G. (2015). A workflow-enabled big data analytics software stack for escience. In *High Performance Computing & Simulation (HPCS), 2015 International Conference on*, pages 545–552. IEEE.
- PTC (2019). Digital transformation solutions to unlock the value of iiot.
- Rani, B. K. and Babu, A. V. (2015). Scheduling of big data application workflows in cloud and inter-cloud environments. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2862–2864. IEEE.
- RecordedFuture (2019). Threat intelligence powered by machine learning.
- Redmond, A. (2013). *SLA 2013 Annual Conference*. SLA, San Diego, CA.
- Riak (2019). Enterprise nosql database.
- Samsara (2019). Internet-connected sensors.
- Sawant, N. and Shah, H. (2013). *Big Data Application Architecture Q & A: a Problem-Solution Approach*. Apress, Berkeley, CA.
- Sayad, S. (2010). Model deployment.
- SecurityScorecard (2019). Cyber security scores & risk analysis.
- SentinelOne (2019). Endpoint security software.
- Sentry (2019). Error tracking software - javascript, python, php, ruby, more.
- Sift (2019). Digital trust & safety: Go beyond fraud prevention with sift.
- Signifyd (2019). Guaranteed fraud protection and charge-back recovery.
- SlamData (2019). The mra powered etl solution provider.
- SparkCognition (2019). Sparkcognition.
- Tsai, C.-F., Lin, W.-C., and Ke, S.-W. (2016). Big data mining with parallel computing: A comparison of distributed and mapreduce methodologies. *Journal of Systems and Software*, 122:83–92.
- Turck, M. (2018). Great power, great responsibility: The 2018 big data & ai landscape.
- Uptake (2019). Industrial ai and iot for global industry.
- WebQDA (2019). webqda – qualitative data analysis software.
- Xu, L. D., He, W., and Li, S. (2014). Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, 10(4):2233–2243.