Chapter 17:
DIPAR: A Framework for Implementing Big Data Science in Organizations

Luis Eduardo Bautista Villalpando[1,2], Alain April[2], Alain Abran[2]

[1]Department of Electronic Systems, Autonomous University of Aguascalientes, Aguascalientes, Mexico

[2]Department of Software Engineering and Information Technology, ETS – University of Quebec, Montreal, Canada

Abstract: Cloud computing is a technology aimed at processing and storing very large amounts of data which is also known as Big Data. One of the areas that have contributed to the analysis of Big Data is Data Science. This new study area is called Big Data Science (BDS). One of the challenges on implementing BDS into organizations is the current lack of information which helps to understand BDS. Thus, this chapter presents a framework to implement Big Data Science in organizations which describe the requirements and processes necessaries for such implementation.

Keywords: Cloud Computing, Big Data Science, System Requirements, Security

17.1 Introduction

Cloud Computing (CC) is a technology aimed at processing and storing very large amounts of data. According to the ISO subcommittee 38 – the study group on Cloud Computing, CC is a paradigm for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable cloud resources accessed through services, that can be rapidly provisioned and released with minimal management effort or service provider interaction [1].

One of the most important challenges in CC is how to process large amounts of data (also known as Big Data - BD) in an efficient and reliable form. In December 2012, the International Data Corporation (IDC) released a report titled "The Digital Universe in 2020" which mentions that at the end of 2012, the total data generated was 2.8 Zettabytes (ZB) (2.8 trillions of Gigabytes) [2]. Furthermore, IDC predicts that the total data for 2020 will be 40 ZB - this is roughly equivalent to 5.2 terabytes (TB) of data generated by every human being alive at that year. In

addition, the report mentions that only 0.5 % of data have been analyzed until today and a quarter of all currently available data may contain valuable information if this is analyzed. Thus, *Big Data (BD) processing* will be a topic of great relevance in next year's.

One of the main areas which contribute to the analysis of BD is *Data Science (DS)*. Although the term "Data Science" has emerged recently, this has a long history because is based on techniques and theories from fields such as mathematics, statistics, data engineering, etc. [3], that integrated into the BD paradigm has resulted in a new study area called *Big Data Science (BDS)*.

BDS has recently become a very important topic within organizations because the value it can generate to costumers and themselves. However, one of the main challenges in BDS is the current lack of information which helps in understanding, structuring and defining how to integrate this study area into organizations and how to develop the processes to its implementation.

Unlike DS, BDS adds new challenges during its implementation such as; integration of large amounts of data from different sources, data transformation, storage aspects, security factors, the analysis of large data sets by means of high-performance processing technologies, and the representation of analysis results (visualization) - only to mention a few.

Thus, this chapter presents the DIPAR framework which proposes a mean to implement BDS in organizations defining the requirements and involved elements. The framework consists of five stages; *Define, Ingest, Pre-process, Analyze and Report (DIPAR)* and describes how to implement it as well as its components.

The rest of this chapter is organized as follows. The section 17.2 presents an overview of the Big Data Science, its definition, its history and its relationship with other study areas like Data Mining and Data Analysis. Section 17.3 presents the ISO 15939 Systems and Software Engineering – Measurement Process whose purpose is to collect, analyze, and report data relating to products to be developed. The section 17.4 is the core of this book chapter and presents the DIPAR framework. This framework proposes a mean to implement BDS in organizations as well as its stages. Section 17.5 describes the relationship between the DIPAR framework and the ISO 15939 Measurement Process. In addition, this section presents how it is integrated the stages of the DIPAR framework into the measurement processes defined in the ISO 15939 standard. The section 17.6 presents a case study which uses the DIPAR framework to develop a BD product for the performance analysis of cloud computing applications. Finally, section 17.7 presents a summary and conclusions of this chapter.

17.2 Big Data Science

The term *"Big Data Science" (BDS)* has been in common usage for over three years but it is in part an evolution of the term *"Data Analysis" (DA)*. In 1962, Jhon Tukey [4] wrote that while mathematical statistics evolved, it is possible being applied to "very extensive data" which is the central interest in DA. Moreover,

Tukey mentions that DA includes, among other things; procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, and so on.

In recent years, *Data Mining (DM)* has been the area of knowledge that has been responsible for data analysis in organizations. Authors like Han [5], describes DM as an interdisciplinary subject which includes an iterative sequence of steps for the *knowledge discovery*. These steps are; data cleaning, data integration, data selection, data transformation, pattern evaluation and presentation of results. Han mentions that such steps can be summarized into the ETL (extraction/transformation/loading) process. Extraction is the stage in which data is collected from outside sources, the transformation stage applies methods and functions to data in order to generate valuable information and then be loaded into the end target to generate outputs reports.

Although the ETL process has been applied in organizations for some years, this approach cannot be completely used in the same manner in BD. This because the traditional data warehouse tools and processes related to it are not designed to work on very large amount of data. Some authors like Lin [6] mentions that "big data mining" is about much more than what most academics would consider data mining. Furthermore, Lin sentences that a significant amount of tooling and infrastructure is required to operationalize vague strategic directives into concrete, solvable problems with clearly-defined metrics of success. Other authors like Thusoo [7] notes that the BD processing infrastructure has to be flexible enough to support optimal algorithms and techniques for the very different query workloads. Moreover, Thusoo mentions that what makes this task more challenging is the fact that the data under consideration continues to grow rapidly, just to mention one example, in 2010 Facebook generated more than 60TB of new data every day.

17.3 Big Data Science as a measurement process

One of the most important challenges for organizations is to turn available data into final products which generate value and create a competitive advantage for the enterprises and institutions. For this, it is necessary to develop measurement processes which allow us to analyze the information related to the original data in order to define the types of products that can be developed. According to the ISO 15939 [8] *Systems and Software Engineering – Measurement Process*, the purpose of a measurement process, is to collect, analyze, and report data relating to the products developed and processes implemented within the organizational unit, to support effective management of the process, and to objectively demonstrate the quality of the products.

The ISO 15939 standard defines four sequential activities to develop such measurement which are: establish and sustain measurement commitment, plan the measurement process, perform the measurement process, and evaluate the measurement. These activities are performed in an iterative cycle that allows for con-

tinuous feedback and improvement of the measurement process, as shown in Figure 1.

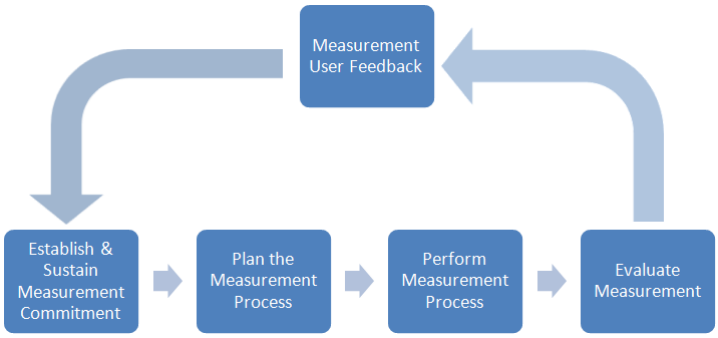


Fig. 17.1. Sequence of activities in a measurement process
(Adapted from the ISO 15939 measurement process model [8])

Next, it is described each activity performed during the measurement process:

- Establish & Sustain Measurement Commitment. This activity consists of two tasks, (1) accept the requirements for measurement and (2) assign resources. Accept the requirements for measurement involves defining the scope of measurement such as a single project, a functional area, the whole enterprise, etc., as well as the commitment of management and staff to measurement; this means that the organizational unit should demonstrate its commitment through policies, allocation of responsibilities, budget, training, etc. In addition, the assign resources task involves the allocation of responsibilities to individuals as well as to provide resources to plan the measurement process.
- Plan the Measurement Process. This activity consists of a series of activities such as identify information needs, select measures, define data collection, define criteria for evaluating the information of products and process. Also it includes the activities to review, approve and provide resources for measurement tasks.
- Perform the Measurement Process. This activity performs the tasks defined into the planning of measurement process across of the following sub-activities: integrate procedures, collect data, analyze data and development of information products and finally communicate results.
- Evaluate Measurement. This activity evaluates the information products against the specified evaluation criteria providing conclusions on strengths and weaknesses of the information products and the measurement process. Also, this activity must identify potential improvements to the information products. For instance, changing the format of an indicator, changing from linear measure to an area measure, minutes to hours, or a line of code size measure, etc.

The next section presents the DIPAR framework, its stages and its process of implementation. Also it is presented the relationships that exist between each stage and how it is integrated into the measurement processes defined in the ISO 15939 standard.

17.4 DIPAR framework

The DIPAR framework integrates the four activities described in the ISO 15939 standard and its main objective is to design BD products that have high impact in the performance of organizations. The Fig. 17.2 presents the stages to perform during the DIPAR framework implementation as well as the order in which should be executed.

Fig. 17.2 Stages to develop during the implantation of the DIPAR framework

The next sub-sections, describes each stage of the DIPAR framework and its involved elements.

17.4.1 Re-define the New Product to Develop

The first step in the DIPAR framework is to define whether is necessary a new BD product or not. If it is not necessary, all the analytical work developed to create the product will be a waste of time and resources. Sometimes cannot be clearly

possible to establish the type of product to be developed, this because there is no knowledge of the type of data that can be collected and analyzed. Patil [9] mentions that a good idea to overcome this issue is to take some clever shortcuts to get products off the ground. Patil sentences that these shortcuts will survive into finished products because they represent some fundamentally good ideas that might not have seen otherwise; resulting in more complex analytic techniques which will be the baseline of better products to be developed in the future. Moreover, these basic ideas are normally aligned with strategic objectives of the organization, e.g., *"it is necessary to improve the user experience in the online store, in order to increase sales"*. Thus, this could be to origin of the development of a new product such as a recommender system for sales.

17.4.2 Ingest the Big Data System

In order to clearly define the boundaries of the new product, it is necessary to collect large amounts of data to analyze. One of the main challenges during the ingestion process of a BD system is to define the ingestion sources because most of the time, data comes from different source services. Therefore, it is very difficult to know what type of data will be ingested in the BD system. For instance, an organization can gather behavioural data from users from very different sources such as web pages logs that users visit, links that they click, social media, location systems included these like mobile devices, etc. In addition, many of these sources of data (services) are loosely-coordinated systems giving as result the creation of a large number of isolated data stores. This distributed scheme makes it difficult to know the type of data that is being collected as well as its state. In addition, services provided by different systems change over time and functionalities evolve, sometimes replaced by new systems or merging into a new. This results in inconsistencies in the data to be analyzed that it is necessary to keep in mind.

17.4.3 Big Data Pre-processing

One of the main problems after the process of ingestion of a BD system is the *"sanity of data"*. For this reason it is necessary to verify the data quality in order to be able to perform the *Big Data Analysis (BDA)*. As example, Lin [6] sentences that during the data collection process in Twitter, it was never encountered a large real-world data set that was directly usable without a procedure of data cleaning. Moreover, some of the main data quality issues to consider during the process of data cleaning in a BDS are; corrupted records, erroneous content, missing values or inconsistency in formatted data (just to name a few). An example of the work involved in BD pre-processing is *Inconsistency;* formatted data is one of the main issues in the process of assurance of data quality because the very different forms that data can take. For example, in one data service the property *product ID* could be defined as *product_id* and in other service as *productID*. Furthermore, the data

type defined in the same property in the first could be defined as a numeric value while in the latter as an alphanumeric value.

Thus, one of the main challenges during the pre-processing stage is how to structure data in standard formats in order to analyze them in a more efficient form. This is often easier said than done, this because during the process of structuring and merging data into common formats there is a risk of losing valuable information. In fact, this is a current topic of investigation for many researchers.

Other issues before to start the Big Data analysis, is to determine what data fields are more relevant in order to construct analysis models [10]. One solution is to use *"sampling"* to get an overview of type of data collected in order to understand the relationships among features spread across multiples sources. However, it is important to mention that training models on only a small fraction of data does not always give an accurate indication of the model´s effectiveness at scale [6].

17.4.4 Big Data Analysis

Once the data has been pre-processed, it is possible to analyze such information in order to obtain relevant results. For this, it is necessary to develop models which can be used on the creation of new products. One of the main problems during the design of such models is to understand which of the available data is most relevant to an analysis task. During a study of the process of implementation of big data analysis in organizations [10], Kandel found that almost 60 % of data scientists have difficult to understand the relationship among features spread across multiple databases. Moreover, he found that the main challenge in this stage is feature selection which allows developing more accurate models. Kandel also mentions that most data scientists has problem with the size of their data sets because most of the existing analytic packages, tools or algorithms do not scale with such BD sets.

One form to address this problem is using new BD technologies which allow to process and analyze large data sets in reasonable time periods. New technologies like Hive [11] allows to perform this type of tasks in a very fast form. Hive is a data warehousing framework which was created at Facebook for reporting ad hoc queries and analysis of their repositories. Other products like Mahout [12], helps to build scalable machine learning libraries which can be used on large data sets. Mahout supports four use cases were machine learning techniques are used: recommendation mining process, clustering, classification and market basket analysis.

Once it has been able to develop complex models and algorithms for data analysis, we have the possibility to create products that create value to the organization. However, in order to establish the direction to be taken during the product development process, it is necessary to understand the results of the previous analyzes. For example, once that Amazon analyzed its large data set, they found that with the historical record of web pages visited by users, they were in possibility of

create a recommender system like "*People who viewed the product X, also viewed the product Y*".

Thus, it is necessary to have a mechanism to present and consult the analysis results in order to understand them, and the same time communicates clearly to stakeholders involved in the product design such results. Next section describes aspects that are necessary to consider while reporting the analysis results.

17.4.5 Report of Results

Once BD is ingested, pre-processed and analyzed, users need to be able to access, evaluate and consult the results. These results have to be presented in a way that they are understood and "humanized". Often, they are presented in statistical charts and graphs that contain too much information which is not descriptive for the end user. Although a number of BD analysts still deliver their results only by means of statics reports, some final users complain that they are too inflexible and do not allow interactive verification in real time [10]. The Networked European Software and Services Initiative (NESSI) in its technical paper titled *"Big Data, A New World of Opportunities"* [13], mentions that reports generated from the analytics can be thought of documents. These documents frequently contain varying forms of media in addition to textual representation. The NESSI determines that when trying to present complex information, the interface of such information needs to be "humane", it means, responsive to human needs and closely linked to the knowledge of the users. For this the NESSI proposes the use of Visual Analytics (VA) which combine the strengths of human and electronic data processing. The main idea behind visual analytics is to develop knowledge, methods, technologies and practices that exploit the human capacities and electronic data processing. Furthermore, the NESSI lists key features of visual analytics as follow:

- Emphasis on data analysis, problem solving, and/or decision making
- Leveraging computational processing by applying automated techniques for data processing, knowledge discovery algorithms, etc.
- Active involvement of a human in the analytical process through interactive visual interfaces
- Support for the provenance of analytical results, and
- Support for the communication of analytical results to relevant recipients

Furthermore, authors like Yau [14] mentions that data visualization is like a story, in which the main character is the user and he can take two paths. A story of charts and graphs might read a lot like a textbook; however a story with context, relationships, interactions, patterns, and explanations reads like a novel. However, it is important to mention that the first is not better than the latter, in fact, what is pretended is to present something in between the textbook and a novel to visualize BD. What it is wanted is to present the facts but also it is wanted to provide context.

On the other hand, authors like Agrin [15] have focused on point what are the real challenges that BD visualization developers face and what should be avoided when implement BD visualization. Agrin mentions that simplicity must be the goal in data visualization; he sentences that at risk of sounding regressive there are good reasons to work with charts that have been in continuous use since the 18th century. In addition, he notes that Bar charts are one of the best tools available for facilitating visual comparisons leveraging our innate ability to precisely compare side-by-side lengths. Agrin lists a number of tools and strategies which can be useful in design of data visualization which is presented below:

- Don't dismiss traditional visualisation choices if they represent the best option for your data
- Start with bar and line charts, and look further only when the data requires it
- Have a good rationale for choosing other options
- Compared to bar charts, bubble charts support more data points with a wider range of values; pies and doughnuts clearly indicate part-whole relationships; tree maps support hierarchical categories
- Bar charts have the added bonus of being one of the easiest visualisations to make - you can hand-code an effective bar chart in HTML using nothing but CSS and minimal JavaScript, or make one in Excel with a single function

Thus, it is necessary to consider the type of results to be presented in order to determine what scheme of visual representation will be used. For example, if what is desired is to present the degree of relationships between persons, representation by graph charts can be the best option to use. On the other hand if what is required is to present the influence degree of certain factors into the performance of Cloud Computing (CC) systems, maybe the best option is to use bar charts.

The next section presents, the relationship between the elements of DIPAR framework and the standard ISO 15939 Systems and Software Engineering - Measurement Process.

17.5 DIPAR Framework and ISO 15939 Measurement Process

One of the main characteristics of the DIPAR framework is that it was designed taking into account the ISO 15939 measurement process activities. Each stage presented in the DIPAR framework, match with the activities described in the ISO 15939 standard. In addition, both – stages and activities - follow the sequence defined in the ISO 15939 standard. The Fig. 17.3 shows the relationship that exists between the DIPAR framework stages and the activities defined in the ISO 15939 measurement process standard.

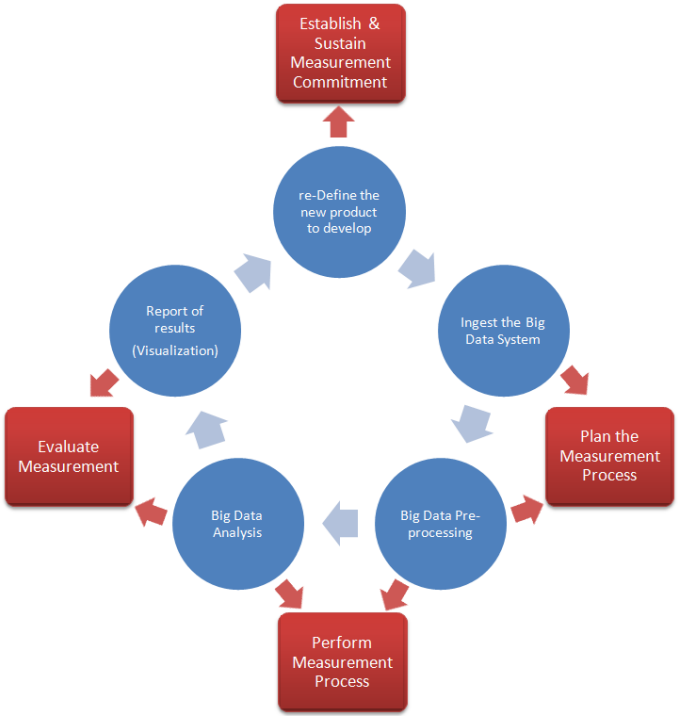


Fig. 17.3 Relationship between the DIPAR framework and the ISO 15939 standard

Once that the DIPAR framework has been matched to the ISO 15939 standard, it is necessary to present in a detailed form what stages described in the DIPAR framework are part of the ISO 15939 activities. The Table 17.1 presents the relationship between the DIPAR framework and the ISO 15939 measurement process.

Table 17.1 Relationship between DIPAR framework and ISO 15939 measurement process

ISO 15939 Activitie	DIPAR Stage	Activities to Develop in the DIPAR Stages
1. Establish & Sustain Measurement Commitment	1. Define the new BD product to develop	• Define the new BD product requirements • Align the product with strategic objectives of the organization • Define the scope of the product • Define a development plan • Assign resources to the develop of the product
2. Plan the Measurement Process	2. Ingest the Big Data system 3. Pre-processing	• Define the sources of data collection • Sketch the type of data to collect

	Big Data system	• Define interfaces to merge the collected data from different sources • Verify the data quality • Perform data cleaning
3. Perform the Measurement Process	4. Pre-processing of Big Data 5. Big Data Analysis	• Get an overview of the relationship between the collected data *(e.g. sampling)* • Develop models and algorithms to the data analysis • Implant the models by means of Big Data processing technologies • Prepare results in order to be reported to users
4. Evaluate Measurement	6. Report of results (visualization)	1. Select the type of format to use to present results (graphs charts, bar charts, etc.) 2. Design flexible reports in order to update them in real-time 3. Design friendly user interfaces to present results 4. Support of results by means of a human analytical process
5. Measurement user feed back	7. Re-define the product to develop	• Use results to create or re-define more complex products such as recommender systems, market basket products, etc. • Re-structure new data to develop new products

The next section presents a case study which uses the DIPAR framework to develop a BD product for the performance analysis of cloud computing applications.

17.6 Case Study

17.6.1 Introduction

One of the most important challenges in delivering Cloud Services is to ensure that they are fault tolerant, as failures and anomalies can degrade these services and impact their quality, and even their availability. According to Coulouris [16], a failure occurs in a distributed system (DS), like a CC system (CCS), when a process or a communication channel departs from what is considered to be its normal or desired behavior. An anomaly is different, in that it slows down a part of a CCS without making it fail completely, impacting the performance of tasks within nodes, and, consequently, of the system itself.

Developing products for CCS, and more specifically for Cloud Computing Applications (CCA) which propose a means to identify and quantify "normal application behavior," can serve as a baseline for detecting and predicting possible anom-

alies in the software (i.e. jobs in a Cloud environment) that may impact Cloud application performance.

Cloud Services (CS) use different technologies to offer storing, processing, and developing through different frameworks for managing CCA. Hadoop is one of the most used technologies within CS because it offers open source tools and utilities for Cloud Computing environments. Hadoop includes a set of libraries and subsystems which permit the storage of large amounts of information, enabling the creation of very large data tables or summarize data with tools of data warehouse infrastructure. Although there are several kinds of application development frameworks for CC, such as GridGain, Hazelcast, and DAC, Hadoop has been widely adopted because of its open source implementation of the MapReduce programming model which is based on Google's MapReduce framework [17].

According to Dean [17], programs written in MapReduce are automatically parallelized and executed on a large cluster of commodity machines. In addition, Lin [18] mentions that the approach to tackling large-data problems today is to divide and conquer, in which the basic idea is to partition a large problem into smaller sub problems. Thus, those sub problems can be tackled in parallel by different workers for example, threads in a processor core, cores in a multi-core processor, multiple processors in a machine or many machines in a cluster. In this form, intermediate results from each individual worker are then combined to yield the final output.

Cloud Computing systems in which MapReduce applications are executed, are exposed to common-cause failures (CCF) which are a direct result of a common cause (CC) or a shared root cause, such as extreme environmental conditions, or operational or maintenance errors [19]. Some examples of CCF in CC systems are; memory failures, storage failures and processes failures. Thus, it is necessary to develop a product which is capable of identify and quantify "*normal application behavior"* by means of collection of base measures specific to CCA performance, such as application processing times, memory used by applications, number of errors in network transmission, etc.

The next section presents the implantation process of the DIPAR framework for the creation of a product which identifies and quantifies the Normal Application Behavior (NAB) of Cloud Computing Aplications (CCA).

17.6.2 Define the Product to Develop

The first stage in the implantation process of DIPAR framework is the definition of the product to be developed. Table 17.2 shows the BD product definition stage as well as the items involved in it.

Table 17.2 Product definition stage and involved items

Product Name: *Application for Performance Analysis of CCA*	**DIPAR Stage:** *Product Definition*

Item	Values
1. Product Requirements	• The product must improve the performance of CCA • The product must include a Performance Measurement Process (PMP) • The PMP must be able to measure Hadoop performance characteristics
2. Align the product with strategic objectives of the organization	• The product must improve the performance of the organization by increasing the quality of provision services in BD processing
3. Scope of the product	• The product must provide performance analysis for users, developers and maintainers • The product must be able to measure MapReduce and Hadoop system performance characteristics • The product must not include analysis of elastic or virtualized cloud systems
4. Define a development plan	• The product development will be performed through the following steps: o Installation a Hadoop test cluster o Collection of system and application performance measures o Develop of a performance analysis model o Report of analysis model results
5. Allocation of resources	• Hadoop test cluster • BD scientist • MapReduce developer • BD visualization developer

17.6.3 Ingest the Big Data System

The second stage in the implementation process of DIPAR framework is ingesting the Big Data System. In this stage, it is defined the type of data to collect as well as their sources. Table 17.3 presents the elements involved in the BD system ingestion stage.

Table 17.3 BD system ingestion stage and involved items

Product Name: *Application for Performance Analysis of CCA*	**DIPAR Stage:** *BD system ingestion*
Item	**Values**
1. Sketch the type of data to collect	• Two data types must be collected; a) Measures of the Hadoop clusters and b) Measures of execution of MapReduce applications
2. Sources of data collection	• Hadoop system logs • MapReduce logs • Measures obtained from system monitoring tools *(e.g. Ganglia, Nagios, etc.)* • MapReduce execution statistics

3. Interfaces to merge in data collection	• The data collected from the sources will be merged and stored into a BD repository as HBase [20]

17.6.4 Big Data Pre-processing

As it was said, one of the main problems after the stage of ingestion in the BDS is *"sanity of data"*. For this reason it is necessary to verify the data quality in order to be able to perform the *Big Data Analysis (BDA)*. Table 17.4 presents the elements involved in the BD Pre-processing stage and steps to execute.

Table 17.4 BD system ingestion stage and involved items

Product Name: *Application for Performance Analysis of CCA*	**DIPAR Stage:** *BD Pre-processing*
Item	**Values**
1. Data Quality	• The data collected from different sources such as logs, monitoring tools and application statistics, were parsed and examined using the cleaning process provided by the Hadoop Chukwa [21] libraries. • Chukwa is a large-scale log collection and analysis system supported by the Software Apache Foundation.
2. Data cleaning	• The data cleaning process was performed using the Chukwa raw log collection and aggregation workflow. • In Chukwa, a pair of MapReduce jobs runs every few minutes, taking all the available logs files as input to perform the data cleaning process [22]. • The first job simply archives all the collected data, without processing or interpreting it. • The second job parses out structured data from some of the logs then clean and loads this structured data into a data store (HBase).

17.6.5 Big Data Analysis

Once the data has been pre-processed, it is possible to analyze such information in order to obtain relevant results. In this case study, a performance measurement framework for cloud computing applications [23] is used in order to determine the form in which system performance characteristics should be measured. Table 17.5 presents the elements involved in the BD Analysis stage and steps to execute.

Table 17.5 BD analysis stage and involved items

Product Name: *Application for*	**DIPAR Stage:** *BD Analysis*

Performance Analysis of CCA

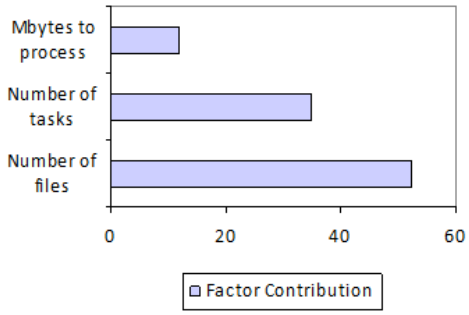
Item	Values
1. Overview of the relationship between collected data (*sampling)*	• The *Performance Measurement Framework for Cloud Computing* [23] defines the elements necessary to measure a *cloud system behavior* using software quality concepts. • The framework determines that the performance efficiency and reliability concepts are closely related in performance measurement. • In addition, the framework determines five function categories to collect performance measures which are: failures function, faults function, task applications function, times function, transmission function.
2. Develop models and algorithms to analyze data	• In order to analyze and determine the type of relationship that exist in the measures collected from Hadoop, a methodology for performance analysis of CCA is used [24]. • This methodology uses the Taguchi method for the design of experiments for identifying the relationships between the various parameters (base measures) that affect the quality of CCA performance. • One of the goals of this framework is to determine what type of relationship exists between the various base measures. For example, *what is the extent of the relationship between CPU processing time and amount of information to process?*
3. Implant the models by means of Big Data processing technologies	• Once the analysis method is determined, it was implanted by means of Pig and Hive technologies in order to implant it the BD repository. • Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems. • Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

17.6.6 Report of Results (visualization)

Once BD was analyzed, it was necessary to evaluate and consult the results. These results have to be presented in a way that they are understood in statistical charts and graphs that contain information that is descriptive for the end user. Ta-

ble 17.6 presents the elements involved in the Report of results stage and the involved elements.

Table 17.6 Report of results stage and involved items

Product Name: *Application for Performance Analysis of CCA*	**DIPAR Stage:** *BD Report of Results*
Item	**Values**
1. Type of format to use to present results	• To present the relationships between the different performance factors, it was determined to use bar charts, line charts and scatter charts. • e.g. Chart of the percentage of factor contribution 
2. Design of user interfaces to present results	• A web-based data visualization scheme was selected to present results. JavaScript and D3.js libraries were selected to design the data visualization web site.
3. Support of results by means of a human analytical process	• The charts were supported with textual explanations, and classroom presentations.

Once finalized the implantation process of the DIPAR framework, it was found that the original product can be re-defined to create a new product. This new product has two main functions; the first is a recommender system, and the second a fault prediction system. The results of the analysis of performance can be used to implant different algorithms of machine learning in both systems.

The recommender system would propose different Hadoop configurations to improve the performance of the CC applications, and the failure prediction system would propose different cases or scenarios in which the CC system may fail or simply degrades its performance.

17.7 Summary

Cloud Computing (CC) is a technology aimed at processing and storing very large amounts of data. One of the most important challenges in CC is how to pro-

cess large amounts of data which also known as Big Data (BD). In December 2012, the International Data Corporation (IDC) released a report titled "The Digital Universe in 2020" which mentions that at the end of 2012, the total data generated was 2.8 Zettabytes (ZB). Thus, BDS has recently become a very important topic in organizations because the value it can generate to costumers and themselves. However, one of the main challenges in BDS is the current lack of information which helps in understanding, structuring and defining how to integrate BDS into organizations. BDS adds new challenges during its implementation such as; integration of large amounts of data from different sources, data transformation, storage aspects, security factors, etc. This chapter presents the DIPAR framework which consists of five stages; Define, Ingest, Pre-process, Analyze and Report. This framework proposes a mean to implement BDS in organizations defining its requirements and its elements. The DIPAR framework is based on the ISO 15939 Systems and Software Engineering – Measurement Process whose purpose is to collect, analyze, and report data relating to products to be developed. In addition, this chapter presents the relationship between the DIPAR framework and the ISO 15939 measurement process standard. Finally, this chapter presents a case study to implant the DIPAR framework for the creation of a new BD product. This BD product identifies and quantifies the Normal Application Behavior (NAB) of Cloud Computing Aplications (CCA). Once finalized the implantation of the DIPAR framework, it was found that the original product can be redefined to create a new product. This new product has two main functions; the first is a recommender system, and the second a fault prediction system. The DIPAR framework can be implanted in different areas of BD and we hope that it contributes to the development of new BD technologies.

References

1. ISO/IEC, *ISO/IEC JTC 1 SC38:Study Group Report on Cloud Computing*, 2011, International Organization for Standardization: Geneva, Switzerland.
2. Gantz, J. and D. Reinsel, *THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, 2012, IDC: Framingham, MA, USA. p. 16.
3. Press, G. *A Very Short History Of Data Science*. 2013 [cited 2013 May, 2013]; Available from: http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/.
4. Tukey, J.W., *The Future of Data Analysis*. The Annals of Mathematical Statistics, 1962. **33**(1): p. 1-67.
5. Han, J., M. Kamber, and J. Pei, *Data Mining, Concepts and Techniques*, ed. Elsevier2012, Waltham, MA, USA: Morgan Kaufmann. 633.

6. Lin, J. and D. Ryaboy, *Scaling Big Data Mining Infrastructure: The Twitter Experience*, in *Conference on Knowledge Discovery and Data Mining 2012*, B. Goethals, Editor 2012, Association for Computing Machinery: Beijing, China. p. 6-19.
7. Thusoo, A., et al. *Data warehousing and analytics infrastructure at facebook*. in *ACM SIGMOD International Conference on Management of data 2010*. 2010. Indianapolis, Indiana, USA: Association for Computing Machinery.
8. ISO/IEC, *ISO/IEC 15939:2007 Systems and software engineering — Measurement process*, 2008, International Organization for Standardization: Geneva, Switzerland.
9. Patil, D., *Data jujitsu: The Art of Turning Data Into Product*, 2012: Sebastopol, CA, USA.
10. Kandel, S., et al. *Enterprise Data Analysis and Visualization: An Interview Study*. in *IEEE Visual Analytics Science & Technology (VAST)*. 2012. Seattle, WA, USA: IEEE Xplore
11. Thusoo, A., et al. *Hive-a petabyte scale data warehouse using Hadoop*. in *26th International Conference on Data Engineering*. 2010. Long Beach, California, USA: IEEE Xplore.
12. A.S.F. *What is Apache Mahout?* 2012; Available from: https://cwiki.apache.org/confluence/display/MAHOUT/Overview.
13. N.E.S.S.I., *Big Data, A New World of Opportunities*, 2012, Networked European Software and Services Initiative: Madrid, Spain.
14. Yau, N., *Seeing Your Life in Data*, in *Beautiful Data, The Stories Behind Elegant Data Solutions*, T.S.a.J. Hammerbacher, Editor 2009, O'Reilly Media, Inc.: Sebastopol, CA. p. 1-16.
15. Agrin, N. and N. Rabinowitz. *Seven Dirty Secrets of Data Visualisation*. 2013 February, 18; Available from: http://www.netmagazine.com/features/seven-dirty-secrets-data-visualisation#null.
16. Coulouris, G., et al., *Distributed Systems Concepts and Design*. 5th ed. Pearson Education2011, Edinburgh: Addison Wesley.
17. Dean, J. and S. Ghemawat, *MapReduce: simplified data processing on large clusters*. Communications of the ACM, 2008. **51**(1): p. 107-113.
18. Lin, J. and C. Dyer, *Data-Intensive Text Processing with MapReduce*2010, University of Maryland, College Park: Manuscript of a book in the Morgan & Claypool Synthesis Lectures on Human Language Technologies.
19. Xing, L. and A. Shrestha. *Distributed Computer Systems Reliability Considering Imperfect Coverage and Common-Cause Failures*. in *11th International Conference on Parallel and Distributed Systems*. 2005. Fuduoka, Japan: IEEE Computer Society.
20. A.F.S. *Apache HBase, the Hadoop database, a distributed, scalable, big data store*. 2013 June 6th]; Available from: http://hbase.apache.org/.

21. Rabkin, A. and R. Katz, *Chukwa: a system for reliable large-scale log collection*, in *Proceedings of the 24th international conference on Large installation system administration*2010, USENIX Association: San Jose, CA. p. 1-15.
22. Boulon, J., et al. *Chukwa, a large-scale monitoring system*. in *In Cloud Computing and its Applications (CCA '08)*. 2008. Chicago, IL.
23. Bautista, L., A. Abran, and A. April, *Design of a Performance Measurement Framework for Cloud Computing*. Journal of Software Engineering and Applications, 2012. **5**(2): p. 69-75.
24. Bautista, L., A. Abran, and A. Abran, *A Methodology for Identifying the Relationships Between Performance Factors for Cloud Computing Applications*, in *Software Engineering Frameworks for the Cloud Computing Paradigm*, M. Zaigham and S. Saqib, Editors. 2013, Springer: London, England. p. 111-117.