# Advancing Big Data Warehouses Management, Monitoring and Performance

Nuno Silva[1][0000-0002-9160-5923]

[1] ALGORITMI Research Centre, University of Minho, Guimarães, Portugal
nuno.silva@dsi.uminho.pt

**Abstract.** In recent years, Big Data technologies have proven their capabilities and usefulness. Different organizations are testing and using several Big Data technologies to support their business. The development of new methodologies, frameworks and data models by the academic community is helping practitioners to use these technologies and apply them in their business contexts. Still, some areas are starting to emerge, as happens in the area of Big Data Warehouse monitoring, which is extremely relevant for managing and maintaining such complex data repository. The main motivation for this work is to find a set of metrics for monitoring Big Data Warehouses in order to allow their evolution and improve performance.

**Keywords:** Big Data Warehouse, Big Data, Monitoring, Performance, Metrics.

## 1 Introduction

Nowadays organizations are competing at a global scale, due to that, they are trying to find business advantages in different areas and for that investing in information technologies such as Data Warehouses (DWs) is of upmost importance.

DWs allow organizations to get insights about their business, but nowadays data has additional characteristics (such as velocity, volume and variety) that are constantly defying the capabilities of traditional DWs.

This need for technologies capable of dealing with new data characteristics pushes forward a new paradigm called Big Data. Big Data is frequently defined by several characteristics, often framed the Vs of Big Data. There are different opinions regarding how many Vs are necessary to characterize Big Data[1][2][3].

Big Data technologies have the capability to deal with a huge amount of data arriving from different sources, with various schemes (or even schema-less sources). Moreover, they are capable of real-time analytics, allowing a constant analysis of the actual state of the organization.

This is significantly difficult to ensure using traditional DWs, due to the fact that they are more suitable for structured and historical data analyses and to their difficulties in scaling horizontally [4]. Big Data Warehouses (BDWs) can scale horizontally more easily and can perform real-time data analysis [5].

Intensive data-systems, such as BDWs, need to be managed and monitored to maintain or increase their performance. As organizations typically measure their performance using Key Performance Indicators (KPI) to ensure their efficiency in the business context, these systems also need specific KPIs to monitor and analyse their performance. Despite the fact that nowadays one can find in the literature the necessary architectures to build a BDW [6]–[8], as well as comparisons between technologies that help the practitioners in their choices [9], [10], managing and monitoring a BDW is a topic which currently has less attention.

Monitoring the performance of BDWs is essential to maintain or even improve their performance through time, even with the increase of more data and business processes. Moreover, if we are able to monitor BDWs (through the use of metrics such as CPU, RAM, and network utilization, user-friendliness, among others) we can manage their resources and adjust definitions, in order to improve their performance. For example, if we know that in some period of time we are overloading the hardware resources, we can try to reschedule some processes that are running in the same period of time. Also, a complete BDW monitoring process can identify if there are users or processes that need more resources or even retrieve insights to adjust the BDW's data model.

The main focus of this research process is to standardize and make available a set of metrics that should be considered for the adequate BDWs managing and monitoring.

To accomplish this task, this work discusses the relevance of a system capable of not only verifying the bottlenecks of BDWs but also prescribing changes to the BDW, being this essential to maintain a BDW under a sustainable and performant growth.

## 2 Related Work

DWs are suitable for structured data and for data that does not change frequently, but DWs have difficulties dealing with large amounts of semi-structured and unstructured data. These data properties are common in the era of Big Data and are making DWs inadequate [11]. To solve this problem, BDWs based on Hadoop components began to be developed by the community using flexible data models, and scalable technologies that can vastly improve the data processing tasks [12].

One of the boosters of BDWs is Hive [13]. Using the principles of DWs but with scalability in mind, Hive has the capability of querying huge volumes of data stored in a distributed system using a SQL-Like language [14], making Hive a system that is familiar to the practitioners and can support flexible data models [13].

Furthermore, the data model has a relevant role in a Big Data Warehousing system's efficacy and efficiency and can be considered the central piece of the BDW [15]. Even in the context of semi-structured or unstructured data, relevant characteristics or features can be extracted from the datasets and need to be stored with some structure (even though with more flexible schemas) to be capable of answering the business questions [15].

BDW's data models should be flexible and scalable. Costa et al. [15] propose an approach for modelling BDW with four types of objects: i) analytical objects, ii) descriptive attributes, iii) factual attributes, and iv) predictive attributes. The use of these objects allows the creation of more flexible and efficient data models [15]. There are other approaches for modelling data for BDWs, such as, for example, total denormalization of data into NoSQL databases [16] or data models based on graphs [17].

Big Data raises new challenges in the development of DWs (originating the concept of BDWs), but new strategies and technologies continue to be developed to mitigate those challenges. One of those challenges is how to monitor and manage a BDW ensuring that its performance maintains the expected level for interactive analyses and decision support tasks, mainly when new business processes and/or data are added.

To verify if the system is running as required, to optimize the system, or to change the data model due to new data sources or processes, it is necessary to have tools that show how the system is currently performing. It is difficult to optimize a system without knowing in advance how it is behaving [18]. It is possible to use intuition or experience to verify and optimize the system but, normally, intuition is not an adequate counsellor [19].

To properly analyze the behavior of the system it is necessary an adequate and objective set of metrics that can measure the performance of the system, monitoring it and helping to identify how or in which components it can be optimized, in case that is needed [19].

Due to the youth of the research related to BDWs, there is a lack of methodologies, frameworks, or metrics able to monitor the BDW's performance. However, being a BDW a complex system that is composed of multiple known and studied components, different metrics can be identified and used to monitor these components. If needed, new metrics can be proposed. For example, to measure the performance of a Cassandra NoSQL database, Bagade et al. [18] use the following metrics: CPU usage, memory usage, thread pool statistics, read-write counts and latencies for each keyspace and column family. In-memory databases can be monitored using the time taken to complete operations, and how efficiently they use memory during operations [20].

For measuring the understandability of the DW conceptual models, Serrano et al. [19] use the number of dimension tables, maximum depth of the hierarchy relationships, number of hierarchy relationships, among others. Khan et al. [21] use several metrics (e.g., execution time, CPU usage and session connection time) to compare their framework for efficient data retrieval in virtual DWs environments with other solutions. Execution time is also used to compare different technologies being queried [22] or to compare different configurations [10]. In order to suggest techniques to improve query performance in a DW, AlHammad and Taha [23] observe the accessed number of rows and disk I/O.

Is possible to measure the performance of a system by measuring its utility. In [24], questionnaires are used to evaluate the usefulness and user-friendliness of the system. To measure the implementation success of information systems, Delone–Mclean present several dimensions that need to be analyzed, such as Information Quality, System Quality, Service Quality, Net Benefits, among others [25].

The literature review shows that there are different ways of measuring the performance of a complex system and some of them can be useful to measure the performance of BDWs. The same can help changing BDWs' configurations, characteristics, data models, among others, when we are looking for better performance and robustness. However, specific literature focusing on BDW monitoring was not identified.

## 3 Expected Contributions

Managing and monitoring a BDW is complex and difficult. A BDW is not stationary, as it is evolving through time. Research for ensuring that the BDW evolution is smooth, not complex, and not impactful for performance is still missing in the academic community.

The focus of this doctoral thesis is to promote and evolve this research topic of BDW management and monitoring, to help researchers and practitioners in the maintenance and evolution (being able to perform modifications) of BDWs. For that, this doctoral thesis has the following research goal:

*"Design and implement a system capable of monitoring and managing the evolution of a BDW."*

This thesis will be developed in a business context that will be used as a demonstration case, more specifically in the BDW of an organization with substantial representativity in the realm of multimedia car parts manufacturing. In this context, the following objectives have been defined:

- Identify a set of dimensions and metrics to monitor, manage and evaluate the performance of a BDW; this first objective is at an early stage of development and is later described in section 5;
- Propose a generic KPIs tree, based on the result of the previous objective, to integrate the monitoring dimensions and their corresponding metrics capable of being implemented in any organization with targets defined within that specific context;
- Propose a prescriptive system, based in the KPI tree, capable of recommending changes in the BDW data model, to improve its performance.

The accomplishment of these three objectives will lead to a set of metrics to be used in the BDW monitoring, in order to identify which changes can be done in the BDW to improve its performance.

The following section presents the used methodology and its activities, in addition to the tasks that will be performed in this doctoral thesis and the artefacts that will be developed.

## 4 Research Methodology

To perform scientific research, it is necessary to follow a methodology that helps us performing structured research that can be replicated, evaluated, and validated by others. One of the methodologies that can be used to perform this research is the De-

sign Science Research Methodology for Information Systems (DSRM-IS). This methodology was proposed by Peffers, Tuunanen, Rothenberger, & Chatterjee [26], and allows performing research in this field by using structured methods and guidelines.

**Fig. 1** represents the Design Science Research process that should be followed in Information Systems research [26]. **Fig. 1** shows six main activities, but the process does not need to be straight forward from activity 1 to activity 6. Moreover, there are four main entry points depending on the starting context of the research process. It is also possible to have various iterations [26].

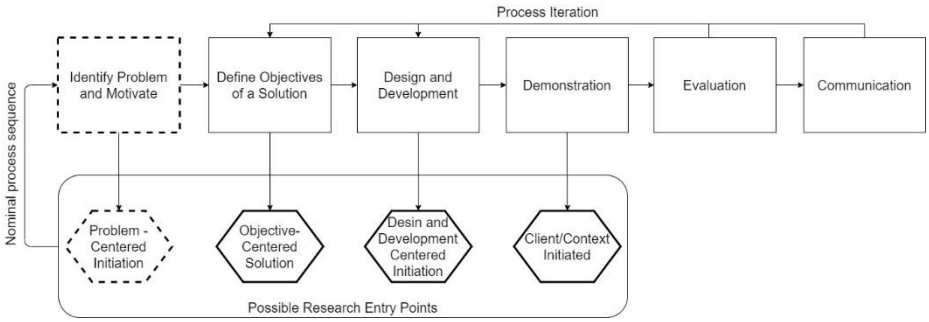


**Fig. 1.** DSRM-IS process model. Adapted from [26]

The first activity is "Problem identification and motivation". In this activity, it was observed the absence of literature related to monitoring and managing of BDWs. That can lead to a decrease in performance and an increase in costs for these data infrastructures.

The second activity is "Define the objectives for a solution". Due to the problem identified in the previous activity, this doctoral thesis has three objectives mentioned in section 3, in order to develop artefacts that can guide the practitioners in the monitoring and managing of BDWs.

The third activity is "Design and development". In this activity, it is expected the creation of 3 artefacts: i) a framework that provides a group of dimensions and metrics for monitoring, managing the BDW's performance; ii) KPIs tree based in the previous identified metrics and dimensions; and, iii) prescriptive system for recommending changes in the BDW data model. Each one of these artefacts are directly connected to each objective mentioned in section 3.

In the fourth activity, "Demonstration", the researcher is invited to conduct case studies, simulations or other demonstrations to show the artefact's capability to solve the problem [26]. In this doctoral thesis, two demonstration cases will be presented, one with synthetic data and one with real data from a car parts manufacturing company.

The fifth activity is "Evaluation". The artefact evaluation can be qualitative or quantitative relying on the nature of the artefacts that will be developed. For example, the dimensions and metrics will be evaluated qualitatively, while the prescription system can be evaluated using a quantitative approach comparing the performance

before and after the implementation of the suggested alterations. Moreover, the third objective will allow the verification of the usefulness and efficacy of the different metrics used to monitor a BDW.

The sixth and last activity is "Communication". It is expected that this doctoral thesis lead to scientific papers with the accomplished results, to be published and validated by the scientific community.

This is the last step in the DSRM-IS, but it is an iterative process that can be restarted to optimize and improve the solution.

Furthermore, Esearch et al. [27] present seven guidelines to follow when researchers use Design Science Research (DSR). The main goal of these guidelines is to help the scientific community understanding what is necessary for an adequate DSR process. The guidelines mention that the research must: produce a viable artefact; be relevant and solve business problems; be evaluated; have contributions in its body of knowledge; apply rigorous methods; be designed as a search process; and, be communicated to the proper audiences.

Following this methodology and guidelines, one expects to produce artefacts that are able to solve the identified problems, fulfilling the objectives proposed in section 3.

## 5 Proposed Approach and Current Results

The work starts with the definition of different dimensions to monitor and managing the performance of a BDW. Each dimension will have a set of metrics to highlight the current status of the BDW. This organization allows for: 1) verifying the health of the BDW; 2) understanding in which dimension the BDW is failing; 3) tackling a well-defined problem to improve the overall performance of the BDW. Moreover, this organization into different dimensions allows the development of a future visualization system with dashboards capable of showing the current status of the BDW.

The dimensions were already defined and the resulting structure is presented in **Fig. 2**. In this figure, one defines the basic elements of the 2ME (Monitoring, Managing and Enhancing) framework. From top to bottom, the framework 2ME has different stages, that are monitored by dimensions, wherein each one has different metrics or group of metrics.

Each stage is the representation of a different process present in BDWs that has different characteristics and requirements to be monitored. Dimensions are related to "what" we want to monitor in each process, such as hardware, software, among others. Metrics are the components at the lowest level of detail for monitoring these dimensions.

**Fig. 3** presents the current status of the 2ME Framework developed using the components described before (**Fig. 2**). This figure shows the three BDW's stages to be monitored. The first one is the data collection stage, where data is retrieved from the data sources and moved to the staging repositories that support the BDW. The second one is the processing and enrichment of raw data, a task usually performed in staging areas supporting the BDW. The third stage is the use of the BDW, where the end-user

can perform distinct analyses over the data. Each one of these stages can be monitored through different dimensions to provide an overall overview of their performance and, if needed, the improvement capabilities. During the three stages, we can have up to four different dimensions: time, network, hardware, and software. As each stage is independent, we can have different dimensions for each one. **Fig. 3** highlights that the first stage (Data Gathering) has only three dimensions, while the other two have four dimensions. The second stage is Data Processing and Enrichment, and the third one is Data Analytics.

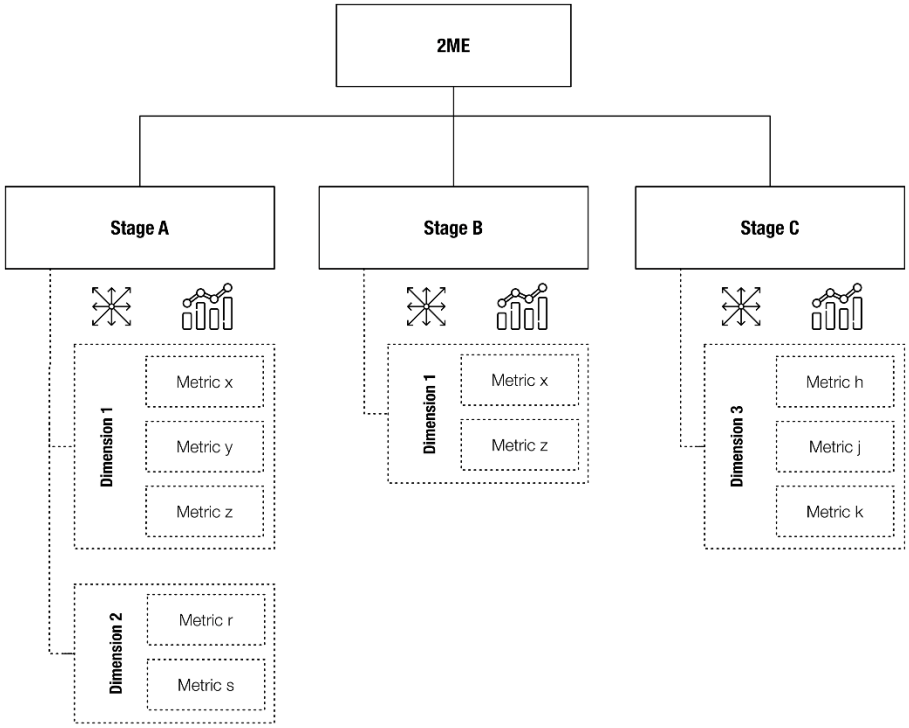


**Fig. 2.** 2ME Framework Structure

Similar to dimensions in stages, different metrics can be included in the dimensions. Therefore, the same dimension can have different metrics in different stages. For example, the dimension “Software” in the Data Analytics stage is related with metrics oriented towards the data model analysis.

At this early stage of the 2ME framework development, the metrics presented in **Fig. 3** are related to common ways of monitoring different systems, such as the capability of a network for transferring data between two points or server hardware usage. This is possible as a BDW, as a complex system, is built different already studied systems.


**Fig. 3.** 2ME Framework (stages, dimensions, and metrics definition)

## 6 Conclusions and Future Work

The management and performance evolution of BDWs should be a concern for all organizations that make use of them. With time, entropy increases, which will lead to a disorganized BDW with lack of performance. To maintain a suitable, or even to increase, the performance of BDWs, we need to develop an artefact capable of indicating in a simple form if there are some processes of the BDW that can be improved. This improvement can lead to an increase in performance and manageability.

Therefore, the research goal of this thesis is the design and implementation of a system capable of managing and monitoring the evolution of a BDW. To accomplish that, it is necessary to identify the useful dimensions to monitor and managing the BDW performance. Moreover, the development of a KPIs tree is necessary to understand the relationship between the metrics and to provide a more comprehensible form of visualization to the practitioners. Considering this as a starting point, it is possible to further propose a prescriptive system to evolve the BDW data model in a semi-autonomous way, considering how it is currently used in the organization. These artefacts will be evaluated using two demonstration cases – one based on synthetic data and the other supported by real data.

This paper presents the current development stage of this doctoral thesis, demonstrating the initial stage of the 2ME framework and describing its main components (stages, dimensions and metrics).

As future work, the first step will continue the development of the 2ME framework, improving or redefining the metrics of each dimension, proposing the KPIs tree (and the goals to achieve) and validating the framework through the two demonstrations cases.

**Acknowledgements.** "This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020" and by the Doctoral scholarship PD/BDE/142900/2018. It is supervised by Professors Maribel Yasmina Santos and Carlos Costa. This paper uses icons made by Freepik, geotatah, itim2101, Pixel perfect from www.flaticon.com.

## References

1. D. Laney, "3D data management: Controlling data volume, velocity and variety," 2001.
2. T. Furche, G. Gottlob, L. Libkin, G. Orsi, and N. W. Paton, "Data Wrangling for Big Data: Challenges and Opportunities," *Int. Conf. Extending Database Technol.*, pp. 473–478, 2016.
3. Ishwarappa and J. Anuradha, "A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology," *Procedia Comput. Sci.*, vol. 48, no. Iccc, pp. 319–324, 2015.
4. A. B. M. Moniruzzaman and S. A. Hossain, "NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison," *New Sci.*, vol. 216, no. 2895, pp. 43–45, Jun. 2013.

5. F. Di Tria, E. Lefons, and F. Tangorra, “Evaluation of Data Warehouse Design Methodologies in the Context of Big Data,” in *Concurrency and Computation: Practice and Experience*, vol. 28, no. 15, 2017, pp. 3–18.
6. M. Y. Santos *et al.*, “A Big Data Analytics Architecture for Industry 4.0,” Springer, Cham, 2017, pp. 175–184.
7. N. Marz and J. Warren, *Principles and best practices of scalable real-time data systems*. Manning Publications Co., 2015.
8. N. Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 6, Reference Architecture (Technical Report No. NIST SP 1500-6),” Gaithersburg, MD, Oct. 2015.
9. E. Costa, C. Costa, and M. Y. Santos, “Evaluating partitioning and bucketing strategies for Hive-based Big Data Warehousing systems,” *J. Big Data*, vol. 6, no. 1, p. 34, Dec. 2019.
10. J. Correia, M. Y. Santos, C. Costa, and C. Andrade, “Fast Online Analytical Processing for Big Data Warehousing,” in *2018 International Conference on Intelligent Systems (IS)*, 2018, pp. 435–442.
11. N. Jukić, A. Sharma, S. Nestorov, and B. Jukić, “Augmenting Data Warehouses with Big Data,” *Inf. Syst. Manag.*, vol. 32, no. 3, pp. 200–209, Jul. 2015.
12. C. Costa and M. Y. Santos, “Big Data: State-of-the-art concepts, techniques, technologies, modeling approaches and research challenges,” *IAENG Int. J. Comput. Sci.*, vol. 44, pp. 285–301, 2017.
13. J. Camacho-Rodríguez *et al.*, “Apache Hive: From MapReduce to Enterprise-grade Big Data Warehousing,” *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pp. 1773–1786, Mar. 2019.
14. Hive, “Hive,” 2019. [Online]. Available: https://hive.apache.org/index.html. [Accessed: 22-Jul-2019].
15. C. Costa, C. Andrade, and M. Y. Santos, “Big Data Warehouses for Smart Industries,” in *Encyclopedia of Big Data Technologies*, S. Sakr and A. Zomaya, Eds. Cham: Springer International Publishing, 2018, pp. 1–11.
16. K. Dehdouh, F. Bentayeb, O. Boussaid, and N. Kabachi, “Using the column oriented NoSQL model for implementing big data warehouses,” *Int. Conf. Parallel Distrib. Process. Tech. Appl.*, pp. 469–475, 2015.
17. S.-M.-R. Beheshti, B. Benatallah, and H. R. Motahari-Nezhad, “Scalable graph-based OLAP analytics over process execution data,” *Distrib. Parallel Databases*, vol. 34, no. 3, pp. 379–423, Sep. 2016.
18. P. Bagade, A. Chandra, and A. B. Dhende, “Designing performance monitoring tool for NoSQL Cassandra distributed database,” in *International Conference on Education and e-Learning Innovations*, 2012, pp. 1–5.
19. M. Serrano, J. Trujillo, C. Calero, and M. Piattini, “Metrics for data warehouse conceptual models understandability,” *Inf. Softw. Technol.*, vol. 49, no. 8, pp. 851–870, Aug. 2007.
20. A. T. Kabakus and R. Kara, “A performance evaluation of in-memory databases,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 29, no. 4, pp. 520–525, Oct. 2017.
21. F. A. Khan, A. Ahmad, M. Imran, M. Alharbi, Mujeeb-ur-rehman, and B. Jan, “Efficient data access and performance improvement model for virtual data warehouse,” *Sustain. Cities Soc.*, vol. 35, pp. 232–240, Nov. 2017.
22. M. S. Wiewiórka, D. P. Wysakowicz, M. J. Okoniewski, and T. Gambin, “Benchmarking distributed data warehouse solutions for storing genomic variant information,” *Database*, vol. 2017, pp. 1–16, Jan. 2017.

23. N. AlHammad and Y. Taha, “Performance evaluation study of data retrieval in data warehouse environment,” in *Proceedings of the 2nd International Conference on Communication and Information Processing - ICCIP ’16*, 2016, pp. 48–54.
24. P. Mondino and J. L. Gonzalez-Andujar, “Evaluation of a decision support system for crop protection in apple orchards,” *Comput. Ind.*, vol. 107, pp. 99–103, May 2019.
25. W. H. DeLone and E. R. McLean, “The DeLone and McLean Model of Information Systems Success: A Ten-Year Update,” *J. Manag. Inf. Syst.*, vol. 19, no. 4, pp. 9–30, Apr. 2003.
26. K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, Dec. 2007.
27. S. Y. R. Esearch, B. A. R. Hevner, S. T. March, and J. Park, “Design Science in Information Systems Research,” *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004.