XV itAIS Conference, Pavia, Italy, Oct. 2018

Relating Big Data Business and Technical Performance Indicators

Barbara Pernici, Chiara Francalanci, Angela Geronazzo, Lucia Polidori, Stefano Ray, Leonardo Riva[1], Arne Jørgen Berre[2] and Todor Ivanov[3]
[1] Politecnico di Milano, Italy
[2] SINTEF, Norway
[3] University of Frankfurt, Germany

Abstract. The use of big data in organizations involves numerous decisions on the business and technical side. While the assessment of technical choices has been studied introducing technical benchmarking approaches, the study of the value of big data and of the impact of business key performance indicators (KPI) on technical choices is still an open problem. The paper discusses a general analysis framework for analyzing big data projects wrt both technical and business performance indicators, and presents the initial results emerging from a first empirical analysis conducted within European companies and research centers within the European DataBench project and the activities of the benchmarking working group of the Big Data Value Association (BDVA). An analysis method is presented, discussing the impact of confidence and support measurements and two directions of analysis are studied: the impact of business KPIs on technical parameters and the study of most important indicators both on the business and on the technical side, for specific industry sectors, with the goal of identifying the most relevant design and assessment criteria.

Keywords: Big Data, Benchmarking, Key Performance Indicators

1 Introduction

The use of big data in organizations implies numerous decisions on the business and technical side. While the assessment of technological choices has been studied introducing technical benchmarking approaches, the study of the value of big data and of the impact of business key performance indicators (KPI) on technical choices is still an open problem. This is mentioned as an IS research challenge: "What design theories do we need to guide big data architectures based on organizational and industry-level contexts?" in [1], discussing research challenges for Big Data.

Giving an answer to this question is one of the goals of the H2020 DataBench research project, funded by the European Commission, started in January 2018. The aim of the project is to provide objective, evidence-based methods to measure the correlation between Big Data Technology (BDT) benchmarks and business benchmarks for an organization and to demonstrate return on investment, developing tools to support this analysis. The identification of adequate benchmarks can support Value management practices in an organization, as described in [10], and in particular

in structural practices such as the Value management office and in process practices, namely benefits management and risk management.

The paper discusses a general analysis framework for analyzing big data projects, discussing both business performance indicators and IT technical indicators emerging from the analysis of ongoing European research projects on Big Data, and presenting the first results emerging from an initial empirical analysis conducted within European companies and research centers within the European DataBench project and the activities of the benchmarking working group of the Big Data Value Association (BDVa)[1]. The issue of relating IT performance and measuring value [10] and managing value [9] of information systems has widely been debated in the literature, in which studies deriving indicators based on case studies are proposed.

In this paper, an analysis method is presented, discussing the impact of confidence and support measurements and two directions of analysis are studied: the impact of business KPIs on technical parameters and the study of most important indicators both on the business and on the technical side, for specific industry sectors, with the goal of identifying the most relevant design criteria. With reference to the classification presented in [9], as our approach is oriented to consider indicators to evaluate Big Data systems benchmarks, we consider indicators with a Business Operations focus, including external service delivery and IT operations indicators.

While IT technical indicators have been analyzed in the literature [9], and can be derived from reference models, such as the ones introduced by BDVa in [2] and NIST in the NIST Big Data Reference Architecture) [8], performance indicators from the business perspective still need further investigation in this area.

As discussed in [5], starting from empirical evidence, industries in the IT sector and highly competitive industries are able to extract value from Big Data, while in other industry groups there is a need to find a measurable impact of this technology.

The goal of the paper is to define business and technical indicators and to study how to find relationships among indicators. The main aim is to profile industry sectors wrt Big Data Analytics (BDA) and to find the significant indicators for assessing its value to organizations.

The developed methodology is based on desk analysis and a questionnaire to collect data from the European research space, in particular from participants in Projects on Big Data within the Private-Public-Partnerships (PPP) in 2014-15[2]. The questionnaire has been developed within the DBVa Benchmarking working group with the goal of collecting information about both business and technical aspects.

The paper is structured as follows. Section 2 introduces a first new framework developed within DataBench to classify business performance indicators. Then, in Section 3, the technical indicators derived from the analysis of existing reference architectures are illustrated. Section 4. describes the methodological approach followed to analyze the results of the questionnaire to collect data about ongoing projects and Section 5 presents and discusses the first conclusions that can be derived from the analysis.

[1] http://bdva.eu/

[2] https://ec.europa.eu/digital-single-market/en/big-data-value-public-private-partnership

2 Business performance indicators

The literature on the relationship between IT (information technology) and business benefits is vast. A largely accepted assumption of this literature is that if a company makes a major investment in IT, the benefits of the investment should be measurable with a business performance indicator [10]. IT is attributed an important organizational role and IT's impact is considered pervasive [12], tangible [14] and measurable with both financial and non financial business performance indicators, often referred to as business KPIs (key performance indicators, cf. [11]). The next section provides a classification of business KPIs, grounded on previous literature.

2.1 Categories of indicators

Business KPIs have been classified in several different ways in previous literature. A fundamental distinction is made between financial, or economic, and non financial KPIs [10]. There is general agreement that a correct evaluation of benefits from a major IT investment should be based on multiple KPIs. For example, authors in [11] have introduced the concept of *balanced scorecard* as a basis for the design of management control dashboards in the design of executive information systems. Similarly, [13] considers the combined use of financial and non-financial KPIs as more effective in the assessment of strategic decisions.

In DataBench, we focus on use cases of big data & analytics projects and aim at the assessment of benefits at a use-case level. An example use case could be the application of machine learning techniques in loyalty marketing and a corresponding benefit could be the reduction of customer churn. In turn, the measurable business KPIs that can be associated with a reduction of churn could be customer satisfaction and revenue growth. In DataBench, we are conducting a desk analysis to collect and classify big data & analytics use cases. So far, we have classified 75 use cases in 9 different industries. The next section discusses how these KPIs represent a fundamental dimension of the more general framework that we have used to classify use cases and to contextualize the measure of business KPIs.

2.2 Modeling business indicators

Figure 1 shows a table where different dimensions represent characteristics of use cases that have to be assessed in order to support the high-level design of the technology architecture and the selection of corresponding technical benchmarks (see Section 3). These characteristics have emerged from the analysis of a total of 75 big-data projects based on our preliminary desk analysis. For example, the *industry* has emerged as an important factor driving high-level technical choices and the corresponding selection of technical benchmarks. We have observed that in the retail industry, the adoption of non-relational technologies is not seen as a business enabler, as retail data are mostly structured and data schema changes are not frequent. Consequently, technical benchmarks designed for non-relational technologies are less (or not) needed in retail, compared to other industries, such as financial services,

where handling documents and applying varying tag sets with semantic technologies can result in frequent data schema changes.

Current work in the DataBench project is focusing on completing the classification of big data project characteristics based on the desk analysis and experimenting them in field studies. As shown in Figure 2, business indicators are grouped in characteristics. Business indicators represent a classification dimension that has a relationship with the choice of technical benchmarks that is mediated by other big-data project characteristics. A project classified with a multi-dimensional model is likely to use specific technical benchmarks. In turn, the correct design of the technology architecture aided by the technical benchmarks represents an enabler of specific business KPIs.

Industry	Big Data Maturity	KPI	Scope of Big Data & Analytics	Data User	DB & Analytics Application	Size of Business	Data size	Datasource
Finance	Currently using	Cost reduction	Decision optimization task	Data Enterpreneurs	Sales	5000 or more	Gigabytes	Distributed
Manufacturing	Piloting or implementing	Time efficiency	Data driven business processes	Vendors in the ICT industry	Customer service & support	2500 to 4999	Terabytes	Centralized
Retail & Wholesale	Considering or evaluating for future use	Product/service quality	Data oriented digital transformation	User companies	IT & data operation	1000 to 2499	Petabytes	
Telecom/ Media	Not using and no plan to do so	Revenue growth			Governance risk & compliance	250 to 999	Exabytes	
Transport/ Accomodation		Customer satisfaction			Product management	50 to 249		
Utility/Oil&Gas/ Energy		Business model innovation			Marketing	10 to 49		
Professional services		Lauch of new products and/or services			Maintencance & logistics	less than 10		
Governamental/ Education					Product innovation			
Healthcare					HR & Legal			
					R&D			
					Finance			

Fig. 1. Big data business indicators

3 Technical indicators

Figure 2 shows the mediated relationship between technical benchmarks and business KPIs discussed in the previous section. Different technical benchmarks evaluate different technical features and provide different output metrics, accordingly. The goal of DataBench is to understand the decision variables that should be considered to choose the right technical benchmark, which, in turn, can help delivering business benefits. Section 3.1 reports a classification of technical benchmarks and related output metrics. Section 3.2 shows a preliminary technical decision framework.

Fig. 2. The mediated relationship between business KPIs and technical benchmarks.

3.1 Categories of benchmarks and output metrics

Figure 3 shows a matrix positioning the Big Data benchmarks being developed in [15] according to different criteria defined in the BDVA Reference Model [2]. On the left (in blue) are listed the different industry application domains, data types and technology areas. On the bottom (in green) in a release time order all the main Big Data benchmarks are listed. Different technical benchmarks show a clearly different focus in terms of features that are benchmarked. There is no complete benchmarking suite and companies have to make a decision on which benchmarking tool is best suited to their application purposes. However, there is no clear correlation between the characteristics of the technical benchmark and the architectural choices that the company should make, which, in turn, depend on the characteristics of the big-data project.

3.2 Modeling technical decision variables

Figure 4 represents a first attempt developed in DataBench to classify technical indicators to select key decision variables in the choice of the technical benchmark. In addition to characteristics related to the output metric and the system, the nature of the task to be accomplished seems to represent a key decision variable. For example, in some cases companies have very complex predictive analytics to execute and need to make sure that the algorithm that they choose is efficient at using available computing capacity. In other cases, they are concerned with more traditional benchmarks evaluating the response time of a DBMS at retrieving information from large SQL tables with a different schema design. As for business indicators, the table provides a classification of indicators in characteristics (from Metrics to Platform features).

4 Relating indicators

The objective of this section is to define systematic analyses that have to be performed in order to gather evidence about the importance of single indicators in Big Data systems and their relationships. The analysis also aims to profile the gathered information by focusing on some specific aspects, such as for instance industry sectors, or specific technical or business characteristics.

Fig. 3. Classification of technical benchmarks (source: [15])

Metrics	Data Types	Benchmark Data Usage	Storage Type	Processing Type	Analytics Type	Architecture Patterns	Platform Features
Execution time/ Latency	Business Intelligence (Tables, Schema…)	Synthetic data	Distributed File System	Batch	Descriptive	Data Preparation	Fault-tolerance
Throughput	Graphs, Linked Data	Real data	Databases/ RDBMS	Stream	Diagnostic	Data Pipeline	Privacy
Cost	Time Series, IoT	Hybrid (mix of real and synthetic) data	NoSQL	Interactive/(near) Real-time	Predictive	Data Lake	Security
Energy consumption	Geospatial, Temporal		NewSQL/ In-Memory	Iterative/In-memory	Prescriptive	Data Warehouse	Governance
Accuracy	Text (incl. Natural Language text)		Time Series			Lambda Architecture	Data Quality
Precision	Media (Images, Audio and Video)					Kappa Architecture	Veracity
Availability						Unified Batch and Stream architecture	Variability
Durability							Data Management
CPU and Memory Utilization							Data Visualization

Fig. 4. Characteristics of technical benchmarks.

In this section, the analysis process is delineated, while in next section the first outcomes of the DataBench project, obtained combining desk analysis and the results of an online questionnaire, are illustrated.

In the following, N will indicate the number of collected responses. Multiple responses for an indicator are possible. In this section, indicators are the possible values for each category (e.g., small, medium, large for size category are considered as three indicators).

POS(Ii) indicates a positive answer to one value of an indicator, POS(I1,...,In) indicates the number of positive answers to a question in the questionnaire, where POS(I1,I2) indicates positive answers to both indicators I1 and I2.

4.1 Identifying common goals in Big Data Projects

A first goal is to identify the most popular indicators in Big Data projects. For each category, the most popular answers will be identified. There are two elements to be considered: the percentage of answers supporting the indicator within a decision variable and a threshold to establish when the percentage is significant to support the indicator. To this purpose two formulas are used:

$$confidence: POS(Ii) / \sum_{1,n} POS(Ii)$$

to indicate the significance of the indicator within a decision variable with n possible values and

support: POS(Ii)/N

to indicate the support for a given indicator, i.e., the percentage of positive answers supporting the indicator on the collected data.

4.2 Analyzing dependencies among indicators

The goal of this analysis is to capture significant dependencies among pairs of indicators. The analysis is therefore based on POS(Ii, Ij) values.

Depending whether the interest is in analyzing the impact of indicator Ii on Ij (or vice versa), the relative importance for the indicator of interest is assessed with the following formula:

cross-significance: $POS(Ii,Ij) / \sum_{1,n} POS(Ii)$ (if we focus on Ii, otherwise the sum is over Ij), where Ii and Ij are indicators belonging to different types of characteristics.

This analysis is useful to find significant relationships between indicators belonging to different categories, e.g., to assess if a given business indicator influences technical choices, or if given technical choices are more common in given business situations. An example is shown in Figure 5, derived from the field analysis illustrated in Section 5 (in this first questionnaire also margin growth was considered, which is been considered as a more detailed indicator linked to cost reduction and revenue growth and therefore not shown in Figure 1.

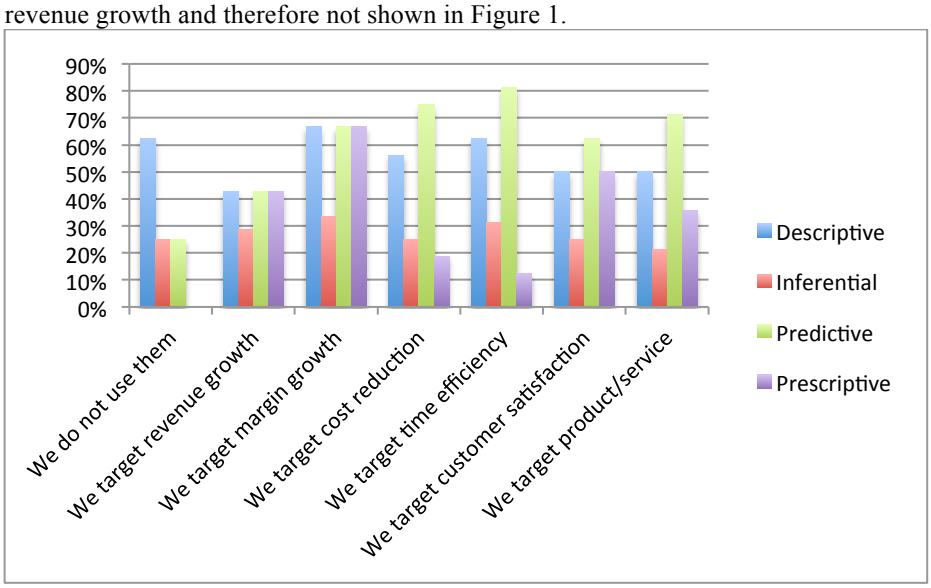

Fig. 5. Example of cross-significance analysis business KPI (x-axis) vs Analysis type (y-axis)

4.3 Profiling on a pivot indicator

The analysis techniques presented above can be used to focus on one characteristic and analyze its implications on other characteristics.

For instance, starting from the 'Industry' characteristic, each industry type can be profiled, selecting all most significant indicators in other characteristics for such industry type.

The analysis is based on the use of one indicator Ii as the pivot indicator, and identifying the cross-significance for the other indicators. The most significant for each category are selected as representative indicators in the profile, based on a threshold. The threshold can be set considering the significance and precision of the indicator in the data set.

5 Results from a first field analysis

In the following, we perform our analysis considering some of the above-mentioned indicators, analyzing the results of a questionnaire on business, technical, and benchmarking aspects developed within the BDVa Benchmarking group and for which answers were collected in the period March-May 2018. Respondents where mainly participants in European PPP Big Data projects, for a total of 36 responders, representing 37 different projects. The questionnaire is synthetically reported in the appendix.

In the questionnaire, we analyzed the most important indicators using the profiling technique illustrated in Section 4.3, and the indicator category [D5] "What are your Big Data application domains", which can assume the following values:
Energy, Financial Services, Manufacturing, Construction, Food/Agriculture, Retail, Wholesale/Professional services, Transport Services, Public Administration, Healthcare, Education, Telecom/IT/Media, Utilities.

We present in Figure 6. the profile obtained for the Manufacturing domain, selecting the indicators that have high confidence in the domain, i.e., for which most of the respondents in the sectors indicated an interest.
From the analysis illustrated in Section 4.1, we also derive that some of these indicators are generally significant across industry sectors, e.g., the answer indicating compliance wrt business requirements and specifications for D10 is common to most sectors.

6 Concluding remarks

In this paper, we have discussed our preliminary results in the definition of a framework to tie the use of technical benchmarks to business indicators. The assumption underlying this study is that technical choices play a strategic role in big data projects and the use of technical benchmarks is of pivotal importance to help architectural choices. The link between technical benchmarks and business indicators can be used in both directions, to help the selection of the right technical benchmarks and to maximize business KPIs with a correct interpretation of the results of technical benchmarking. In the DataBench project, the framework will help the selection of technical benchmarks with a toolbox that will embed key decision variables and will either integrate or link to the most appropriate technical benchmarks. On the other

MANUFACTURING [7 respondents]

- [D10] What are your big data benchmarking goals/plans?
  - **Check whether an implementation fulfils given business requirements and specifications.** [86%]
- [D11] Which aspects of Big Data are you benchmarking or planning to benchmark? (ref. BDV Reference Model)
  - **Data Analytics** [100%]
  - **Data Management** [71%]
  - **Data Processing** [71%]
- [D12] What kind of data are you using/planning to use?
  - **Real Data** [86%]
  - **Hybrid** [86%]
- [D15] What type of Data Storage (Storage/Querying/Discovery) are you benchmarking/considering?
  - **Graph Stores** [67%]
  - **NoSQL** [67%]
- [D16] What is the most important type of Data Processing in your platform?
  - **Interactive/(near) Real-time processing** [71%]
- [D17] What types of data problems are you tackling?
  - **Descriptive** [71%]
  - **Predictive** [86%]
- [D18] What types of machine learning approaches do you typically use?
  - **Supervised** [86%]
- [D19] Which modelling techniques do you typically use?
  - **Deep Learning** [67%]
- [D20] What types of data are stored and processed in your system/platform? (Ref. BDV Reference Model types)
  - **Time Series including IoT Data** [86%]
- [D21] What are the technical key performance metrics that you (want to) measure in your system/platform/service?
  - **End-to-end execution time (Runtime)** [100%]
  - **Throughput** [67%]
- [D22] Which of the following qualitative features are important for your application/platform?
  - **Fault-Tolerance** [71%]

Fig. 6. Profiling key performance indicators in the Manufacturing domain

hand, the results of technical benchmarking activities will be tied back to business KPIs and benefits with in depth case studies.

Results from the questionnaire support the assumption that there is a relationship between technical benchmarks and business KPIs. They also indicate that this relationship is mediated by other variables, such as the industry where a company operates and the specific characteristics of the big data project that is performed. In this research we started from a desk analysis and an initial questionnaire to explore the field, in our future research we will systematically analyze these indicators validating them with an extensive survey and selected case studies in order to design our decision framework for Big Data benchmarking.

Acknowledgements. This work has been partially funded by the European Commission H2020 project DataBench - Evidence Based Big Data Benchmarking to Improve Business Performance, under project No. 780966. This work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work. The authors thank all the participants in the project for discussions and common work.

References

1. Abbasi, A., Sarker, S., Chiang, R.H.L.: Big Data Research in Information Systems: Toward an Inclusive Research Agenda. J. AIS **17:2** (2016) 3
2. Big Data VAlue Association: European big data value - Strategic research and innovation agenda, vers. 4.0. (2017) http://www.bdva.eu/node/874 last accessed 6/2/2018
3. Fox, G.C, Jha, S., Qiu, J., Ekanayake, S., Luckow, A.: Towards a comprehensive set of big data benchmarks. In Proc. High Performance Computing Workshop 2014, in L. Grandinetti, G. Joubert, M. Kunze, V. Pascucci (eds.), Big Data and High Performance Computing, IOS Press (2015) 47 - 66
4. Grover, V., Roger H.L. Chiang, R.H.L, Liang, T.-P., Zhang, D.: Creating Strategic Business Value from Big Data Analytics: A Research Framework. Journal of Management Information Systems **35:2** (2018) 388-423
5. Maes, K., De Haes, S., & Van Grembergen, W.: Developing a Value Management Capability: A Literature Study and Exploratory Case Study, Information Systems Management, **32:2** (2015) 82-104
6. Mitra, S., Sambamurthy, V., & Westerman, G.: Measuring IT performance and communicating value. MIS Quarterly Executive, **10:1** (2011).
7. Müller, O., Fay, M., vom Brocke, J.: The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. Accepted for publication in Journal of Management Information Systems (2018).
8. NIST Big Data Public Working Group (NBD-PWG), NIST Big Data Interoperability Framework: Vol. 6, Reference Architecture, Volume 6, Reference Architecture, NIST Special Publication 1500-6, http://dx.doi.org/10.6028/NIST.SP.1500-6 (2015)
9. Pääkkönen, P., Pakkala, D.: Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems. Big Data Research **2:4** (2015) 166-186
10. Ping-Ju Wu, S., Straub, D.W., and Liang, T.-P.: How Information Technology Governance Mechanisms and Strategic Alignment Influence Organizational Performance: Insights from a Matched Survey of Business and IT Managers. MIS Quarterly **39** (2015) 497-518
11. Hoque, Z.: 20 years of studies on the balanced scorecard: Trends, accomplishments, gaps and opportunities for future research, The British Accounting Review **46:1** (2014) 33-59
12. Gupta, P. and Moitra, D.: Evolving a pervasive IT infrastructure: a technology integration approach. Pers. Ubiquit. Comput. **8:1** (2004) 31-41

13. Shen, Y-C., Chen, P.S., Wang, C.H.: A study of enterprise resource planning (ERP) system performance measurement using the quantitative balanced scorecard approach, Computers in Industry **75** (2016) 127-139
14. Lange, M., Mendling, J., Recker, J.: An empirical analysis of the factors and measures of Enterprise Architecture Management success, European Journal of Information Systems **25:5** (2017) 411-431
15. DataBench Team: D3.1 DataBench Architecture (2018)

Appendix - Benchmarking Big Data Benchmarks questionnaire structure

The structure of the Big Data Benchmarking WG questionnaire is as follows:

General questions

- What is your current role/position?
- Are you participating in EU research projects? If yes, which ones?
- Are you affiliated with an organization? If yes, which one?
- Which societal challenges do you target?
- What are your Big Data application domains?
- Do you use business indicators to measure the performance of your big data & analytics initiatives?
- Are your big data & analytics in real-time and integrated with business processes?
- In which role do you perform benchmarking?
- Are you currently evaluating software using benchmarking technologies?
- What are your big data benchmarking goals/plans?
- Which aspects of Big Data are you benchmarking or planning to benchmark? (ref. BDV Reference Model)
- What kind of data are you using/planning to use?
- Which dataset sizes do you target in your application(s)?

Additional technical questions

- What type of Data Storage (Storage/Querying/Discovery) are you benchmarking/ considering?
- What is the most important type of Data Processing in your platform?
- What types of data problems are you tackling?
- What types of machine learning approaches do you typically use?
- Which modelling techniques do you typically use?
- What types of data are stored and processed in your system/platform? (Ref. BDV Reference Model types)
- What are the technical key performance metrics that you (want to) measure in your system/platform/service?
- Which of the following qualitative features are important for your application/platform?
- What are the key technologies that you are using in your big data infrastructure? For example, Big Data platforms such as Cloudera, HortonWorks, MapR or others offering Hadoop distributions, Spark, Flink, Storm or similar for batch and stream processing, Hive, Spark SQL, Presto or similar for SQL capabilities on top of Hadoop.