A Survey on Big Data Analytics Solutions Deployment

Camilo Castellanos[1], Boris Pérez[1,2], Carlos A. Varela[3], María del Pilar Villamil[1], and Dario Correal[1]

[1] Systems Engineering and Computing Department
Universidad de los Andes, Bogotá, Colombia
{cc.castellanos87, br.perez41, mavillam, dcorreal}@uniandes.edu.co
[2] Systems Engineering and Computing Department
Universidad Francisco de Paula Santander, Cúcuta, Colombia
borisperezg@ufps.edu.co
[3] Computer Science Department
Rensselaer Polytechnic Institute, Troy, NY, USA
cvarela@cs.rpi.edu

Abstract. There are widespread and increasing interest in big data analytics (BDA) solutions to enable data collection, transformation, and predictive analyses. The development and operation of BDA application involve business innovation, advanced analytics and cutting-edge technologies which add new complexities to the traditional software development. Although there is a growing interest in BDA adoption, successful deployments are still scarce (a.k.a., the "Deployment Gap" phenomenon). This paper reports an empirical study on BDA deployment practices, techniques and tools in the industry from both the software architecture and data science perspectives to understand research challenges that emerge in this context. Our results suggest new research directions to be tackled by the software architecture community. In particular, competing architectural drivers, interoperability, and deployment procedures in the BDA field are still immature or have not been adopted in practice.

1 Introduction

With recent big data proliferation, enterprises can use analytics to extract valuable insights from large-scale data sources, something not possible a few years ago. Traditional big data analytics (BDA) methodologies [1,2] involve three knowledge domains: business, analytics, and technology. In the business domain, business users have to define the business goals to drive the analytics project. In the analytics domain, these business goals are translated by data scientists into specific analytics tasks such as data cleaning, model building, and evaluation. This model development is performed within the data lab. Finally, in the technology domain, the IT (Information Technology) team take the analytics model as an input for software implementation and deployment in the production environment respecting Quality Attributes (QA). This migration of the analytics model from data lab to production environment is called a *BDA deployment*.

Despite the growing interest of companies in BDA adoption, actual deployments are still scarce. Chen et al. in [3] coined this phenomenon as the "Deployment Gap". Later, Chen et al. in [4] summarized a set of technical, organizational, and technology challenges that must be handled when developing BDA projects. Previous works have tackled BDA adoption and challenges in analytics practices, and they will be reviewed in Section 2, but little research has been carried out to identify practices, behavior, and procedures from the perspective of software engineering and architecture.

The aforementioned aspects motivate the development of a survey whose objective is to identify the practices, techniques, and tools used in the design, development, and deployment of BDA projects from a software architecture perspective. We conducted a survey among practitioners following a methodology proposed by Kitchenham et al. in [5] defining objectives, designing, developing, and evaluating the survey, then obtaining data, and finally, analyzing the results. We collected answers from 76 practitioners engaged with cross-industry BDA projects in Colombia. The objectives of this survey are framed in the BDA development and deployment context, and they are stated as follows: i) To determine used practices and methods. ii) To determine used techniques and tools. iii) To identify perceived challenges. iv) To identify considered quality attributes.

The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 describes our research methodology. Sections 4 presents the survey results. Section 5 discusses the findings. Section 6 presents the threats to validity. Finally, Section 7 draws conclusions and describes future work.

2 Related work

Chen et al. [3] identified 11 factors which affect BDA adoption, and these factors include organizational, innovation, and technology. They presented the status and strategies to deploy BDA solutions based on 25 European enterprise case studies, but specific behaviors, practices, and tools used in the current deployment of such solutions were not reviewed.

Previous industry surveys (e.g. [6] [7]) have focused on understanding analytics practices using questionnaires directed to a wide number of data scientists. They reported trends about algorithms, tools, data scientist roles, and analytics deployments. These works confirmed low rates (half of the respondents) of analytics projects being deployed, and delayed time of deployment—25% of deployments take months or even years. On the other hand, the survey results presented in [8] were focused specifically on the deployment of BDA solutions. That survey inquired about procedures for packaging, retraining and monitoring BDA solutions, finding that 50% of their respondents stated the level of difficulty of analytics model deployment was more than six (from 1 to 10). Real-time scoring showed a higher level of difficulty, and projects with issues on data quality and pipeline development presented also delayed deployment. Those surveys offer important statistics about deployment and operation of analytics solutions,

but they are not framed in the BDA life cycle, and they do not consider either software engineering or architecture, highly implicated in those processes.

Lavalle et al. presented in [9] challenges and opportunities in business analytics, and highlight the need for analytics capabilities to achieve competitive advantages and make informed decisions. In addition, they compared analytics adoption level, practices, and challenges to organization performance to offer some recommendations to improve analytics adoption across the organization. Although their research analyses general organizational and technology facets, detailed practices and techniques related to deployment, software engineering and architecture are not considered.

3 Methodology

According to Easterbrook et al. [10], the research method depends on the research questions. Based on the above, we decided to use a survey research method to identify the practices in industry and academy about how they develop and deploy BDA solutions. This survey follows the methodology proposed by Kitchenham and Pfleeger [5] for survey designing in empirical software engineering.

3.1 Research questions

We formulate the research questions (RQs) of this survey based on the objectives presented in Section 1.

RQ1: *What are the practices, methods, techniques, and tools used in BDA development and deployment?* By answering this question, we intend to characterize practices, techniques and tools used in BDA design, development, deployment, and operation.

RQ2: *What are the main challenges faced in BDA development and deployment?* By answering this question, we aim at identifying the challenges practitioners have to face in this context.

RQ3: *What are the main quality attributes considered in BDA modeling, evaluation, and deployment stages?* By answering this question, we aim at characterizing QAs which drive BDA's software architecture.

3.2 Sample and population

In our survey, the target population entails practitioners who have participated in BDA projects, playing a range of roles such as project manager, business expert, requirements engineer, data scientist/analyst, data engineer software designer/developer, software/IT/solution architect and IT administrator. We employed *Convenience sampling* (a non-probabilistic sampling method [5]) for selecting the population because of our access to participants involved in BDA projects. Participants were available through the master programs in Information Engineering and IT Architecture offered by Universidad de Los Andes, and

the Colombian Center of Excellence and Appropriation in Big Data Analytics (CAOBA). These participants were involved in industry BDA projects and they were available to collaborate in this research. The master students were signed up for IT Architecture and Data Science Applied courses.

Inclusion and exclusion criteria enable us to choose valid answers regarding experience in BDA practice and consistency. This survey considered the following Inclusion criteria: (i) The respondent has industry experience in BDA projects, and (ii) The respondent has academic experience in BDA projects. The exclusion criteria were (i) There are inconsistent (i.e. contradictory) answers and, (ii) respondents that answered less than 50% of the questions.

3.3 Survey design

This survey can be classified as descriptive research because: 1) This survey was preplanned and structured, and 2) the information collected can be statistically inferred over a population. This type of research uses closed-ended questions allowing us to get a better understanding of opinion or attitude by a group of people on a specific topic.

This survey is a self-administered questionnaire, where a research participant is given a set of questions to answer via paper-based questionnaire. Our survey includes an opening paragraph to introduces the purpose, concepts, and considerations needed to answer the instrument. The questionnaire was reviewed externally by two other researchers and they checked the content, meaning, and understandability. Additionally, 9 practitioners on BDA projects answered a pilot to refine the instrument and estimate the time needed to complete the survey.

Our questionnaire consisted of 5 parts and 24 questions as presented in Figure 1 written in Spanish, the participant's native language. Eighteen questions corresponded to closed-ended questions with single choice, and seven questions included multiple-choice grids to specify the respondent's level of agreement or disagreement on a Likert scale. All questions were mandatory. The 5 parts of the survey were: (a) demographic questions, (b) questions about practices, behavior and challenges in BDA context, (c) questions about techniques and tools used in BDA projects, (d) questions about BDA deployment, and (e) questions about how practitioners dealt with quality attributes. Figure 1 also details how each questionnaire's part is related to the Research Questions (RQ).

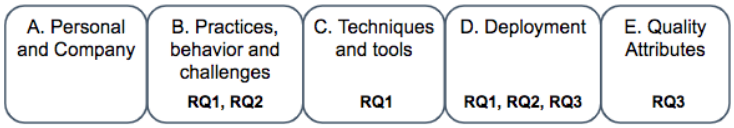

Fig. 1. Questionnaire sections and Research Questions

Demographic questions asked for job, role, level of education and experience of the subjects. These questions also asked for company information like industry

sector, size, experience, and maturity. This first section helped us to understand the participants' background. Remaining parts were used to collect data about the general perception of deployment of BDA projects.

Data analysis were done through the following steps: (i) collection of responses into a single spreadsheet, (ii) analysis of the spreadsheet using descriptive statistics for quantitative answers for each given response, and (iii) identification of key findings from results of the statistical analyses. In order to enable the fully replication of this research, a package with the questionnaire and raw answers is publicly available[4].

4 Survey results

This Section reports the survey results based on collected data, and the following four subsections address the questionnaire's sections detailed in Fig. 1.

In total, 115 answers were collected of which 39 (33.9%) were excluded by criteria detailed in Section 3.2. The remaining 76 (66.1%) valid answers were further analyzed. Hereinafter the 76 subjects who respond valid answers are denominated "respondents".

4.1 Personal and company data

This subsection describes the background information of the respondents. This background can influence the perspective and perception of BDA development and deployment process. This information includes respondent's profession, the role played in BDA projects, educational background and specific experience in this kind of projects.

Regarding respondent's profession, the vast majority of them (84.2%) are IT professionals, followed them by mathematicians/statistics (5.2%), engineers Non-IT and business administrators (3.9%).

The respondent's role played in BDA allows us to know how is represented the stakeholders introduced in Section 1, IT managers corresponds to 26.3%, software architects: 19.7%, developers: 15.7%, data scientists: 14.4%, and IT operators: 6.5%.

We also asked respondents the level of education. Most of them (40.7%) hold an M.Sc degree, 35.5% have a B.Sc. degree, 22.3% a specialization degree and one respondent holds a Ph.D. degree.

The question related to work experience in BDA projects shows that most of the respondents are in junior level hence 67.1% have got involved between 1 and 2 projects, 22.3% have participated between 3 and 5 projects, and 10.5% in more than 5 projects. Regarding the years of experience, half of the respondents have worked between 1 and 3 years, 32.8% less than 1 year, 10.5% between 3 and 6 years. Finally, 6.5 percent of the participants have 6 years of experience or more.

[4] https://storage.cloud.google.com/ccastellanos/BDA-Survey-package.zip

We asked the company's sector to the respondents to understand the business environment in which BDA projects are developed, and education (23.6%) is the most common sector, technology is the second-most popular sector with 22.3%. Both Financial and Government sectors are in the third place with 13.1% of participation, while Communication (9.2%) and Transport (5.2%) sectors complete the list of the top six.

Questions 8 and 9 inquire about the company size and experience by measuring the number of employees and projects undertaken within the company. Most respondents (63.1%) work in large companies (more than 250 employees), 18.4 in small (between 11 and 50), 13.1% in medium (between 51 and 250 employees) and only the 5.2% in micro-enterprises (less than 11 employees). With regard to the number of BDA projects, 47.3% of all participants work in companies with 1 to 3 projects, 15.7% in companies with more than 9 projects, and 14.4% in companies between 4 and 6 projects. Finally, 4 respondents answer that their companies have not developed such projects (5.2%), and 2.6% (2 out of of 76) between 7 and 9 projects.

To know the appropriation level of BDA in the Companies, we asked the current status of BDA projects. As a result, pilot projects were reported in progress by 32.8% of respondents, 23.6% have at least an active program in production, 17.1% in exploration, 9.2% have no a plan and 5.2% have a defined plan to be implemented.

4.2 Practices, Behavior, and Challenges

Fig. 2 depicts the perception of collaboration and teamwork among the stakeholders involved in the BDA environment. This perception is measured ranging from 1 to 5 (1- Difficult and disjointed and 5- Very fluid and articulated). Analytics and IT collaboration and teamwork have the best scoring with a rank greater than 3 for 56.5% of the respondents. Business/IT, and Business/Analytics interactions report the worst rating with only 26.3% and 22.3% of positive evaluations (i.e. greater than 3) respectively.

Fig. 2. Collaboration and Teamwork.

We also inquired about the difficulty to carry out each BDA phase to identify the most challenging activities in the BDA life cycle regarding traditional

methodologies [1,2]. This difficulty score ranges from 1 to 10, and the results are presented in Fig. 3 as boxplot graphs, including mean ($\overline{x}$) and standard deviation (σ). Six out nine activities observe the highest medians (8 points of difficulty): *1) Define project's business goals*, *3) Align analytics tasks to business goals*, *4) Collect data*, *5) Prepare data*, *8) Deploy BDA solution* and *9) Operation*. Among these six activities, those that present the highest means are: *1) Define project's business goals* ($\overline{x}$=7.7, σ=2.1), *3) Align analytics tasks to business goals* ($\overline{x}$=7.2, σ=2.4), and *8) Deploy BDA solution* ($\overline{x}$=7.6, σ=1.9). The boxplots of these three challenging activities show that *8) Deploy BDA solution* activity has the smallest Interquartile Range (between 7 and 9) while the other two activities exhibit more dispersed values. It implies that deployment activity presents jointly the highest mean and the least disperse difficulty score.

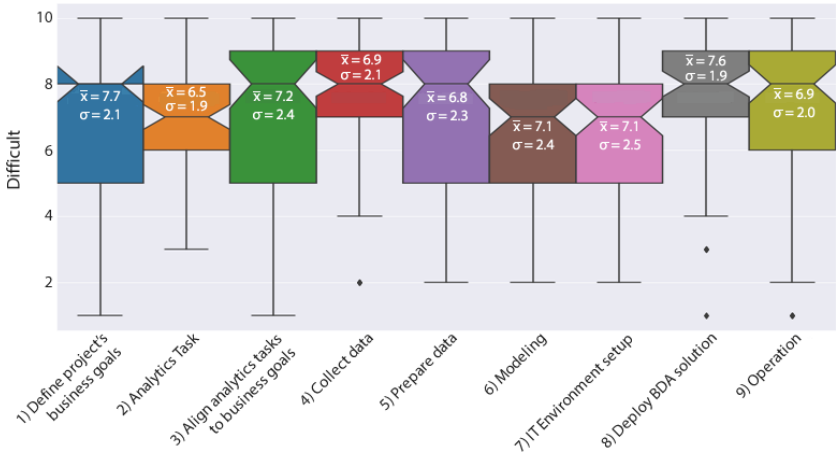

Fig. 3. Level of difficulty to perform BDA activities

4.3 Techniques and Tools

We asked respondents to categorize the usage of an arrangement of techniques to know how data scientists deal with and work with a myriad of options. Fig. 4b describes the frequency of use of analytics techniques/algorithms to build analytics models in a scale from 1 (rarely used) to 5 (frequently used). The five most popular techniques are, in descending order: aggregations (sum, count, means, etc), regression, clustering, anomalies (detection) and Principal Component Analysis (PCA). Aggregations are not actually ML algorithms, but they are the most used when data analysis is required. The most novelty techniques such as Deep Learning and Support Vector Machines (SVM) present a low level of usage in the respondents' context.

In addition to the techniques, we also asked about technology tools usage in BDA development through the same scale from 1 to 5 and Fig. 4b summarizes the results obtained. It is worth noting that this question comprised from spreadsheets to distributed processing engines including self-service Business Intelligence (BI) tools. This can be explained by the data scientist's need to explore, model, visualize and process data. Excel and Standard Query Language (SQL) to access relational databases predominate in the respondent's toolbox with 78.9% of high use frequency. The following eight-most used technologies are in descending order: Tableau, R, Power BI, Click view, Spark, SAS, IBM SPSS and Oracle Data mining. Except in the case of R, big data and ML open source frameworks such as Apache Spark, Scikit Learn, and Mahout are not widely utilized. And some IT big players such as Microsoft (Power BI), SAS, IBM (SPSS) and Oracle rank in the top ten of the technology preferences.

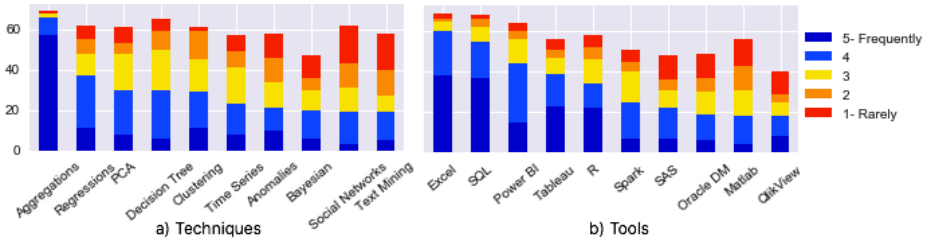

Fig. 4. Usage Frequency of a) Techniques and b) Technology Tools

4.4 Deployment

In Fig. 5a the frequency of BDA deployments on a production environment is shown. As can be noted, few times a year (34.2%), several times a year (18.4%) and "None yet" (18.4%) are the predominant answers, thus confirming the low frequency in our study's context.

During maintenance and operation stages is necessary to retrain/adjust models and software to have up-to-date services. Fig. 5b depicts the procedures used to do such retraining. 22.3% of respondents retrain the model in data lab environments and they upgrade the production model using a manual procedure. Other respondents group reports that they do not retrain models, but they have to rewrite the code (18.4%), 14.4% retrain the model and export the new parameters to production, and only the 6.5% use a DevOps approach.

The respondents were consulted about the procedure or methodology to package/migrate the analytics models and data transformations from the data lab to production and Fig. 6a shows these results. Noteworthy, 31.5% of the respondents did not know or answer which deployment procedure is used. The 28,9% of respondents reported they do not have a procedure because they have a single

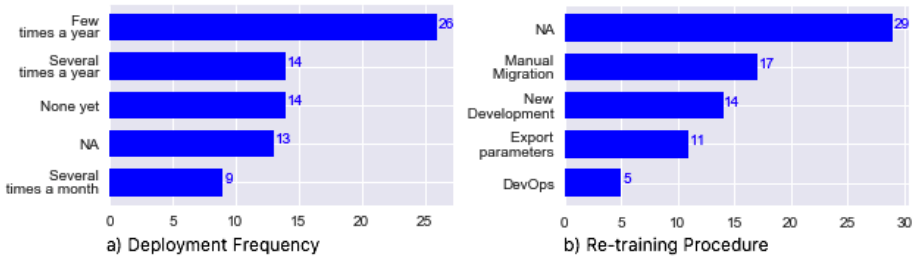

Fig. 5. Frequency of a) BDA deployment in productive environment and b) Re-training Procedure.

environment of BDA, use an ad-hoc procedure (25%), or have to rewrite whole source code (9.2%). Only 1 respondent (1.3%) reported the use of interoperable models such as PMML or PFA.

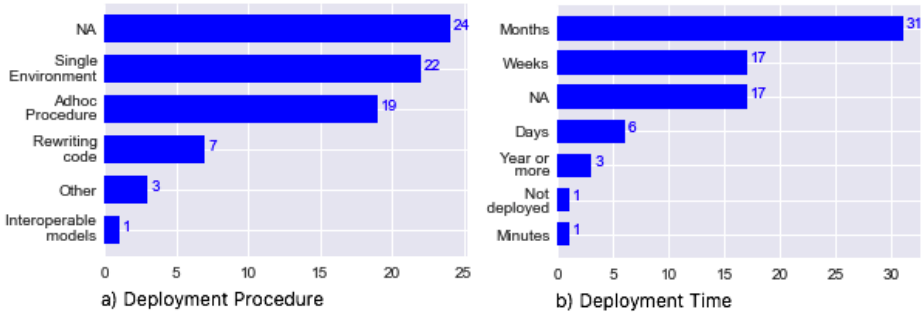

Fig. 6. Frequency of a) Deployment Procedure and b) Deployment Time.

To gain first-hand knowledge about the lag time in the deployment of BDA solutions, we also asked the time elapsed between model development and its deployment in production. Fig. 6b details the time scales invested in this deployment. The most common time scale is *months* (40.7%), followed by *weeks* (22.3%), and in a lower proportion, *days* (7.8%).

To understand the relationship between deployment procedure and frequency, we compare such questions results in Fig. 7. Ad hoc procedure is the most common both in monthly (44.4%, 4 out of 9) and yearly deployments (42.3%, 11 out of 26). Although maintaining a single environment is highly used (35.7%, 5 out of 14) in projects with several deployments a year, also it is the most common procedure (50%, 7 out of 14) among projects which have no deployments yet. Specifications for sharing and interoperating predictive models are not used or scarcely used, displaying a lack of knowledge about these de facto standards.

Figures 8 and 9 compare the appropriation level of the company with the deployment time and deployment procedure. Companies with active BDA pro-

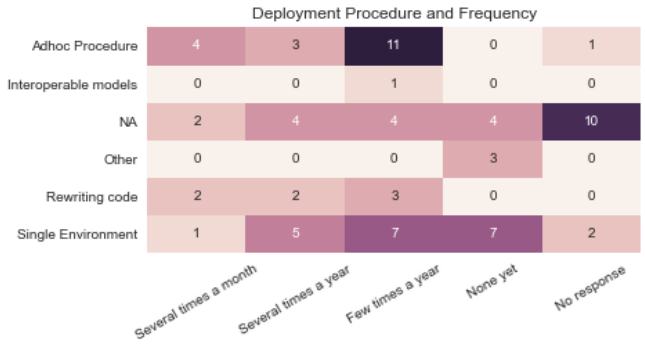

Fig. 7. Deployment Procedure/methodology and Frequency.

grams take weeks 46.6% (7 out of 15) and months 24.6% (4 out of 11). While organizations with a BDA plan to be implemented take months (4 out of 4), pilot project exhibits monthly deployment (53.8%), and companies in the exploration phase take months to deploy their applications. Considering deployment procedures, it is noticeable that companies with active programs use mainly (50%) ad hoc procedures. Something similar occurs with companies with project pilots, where 28% (7 of 25) use ad hoc procedures, no-answer 28% (7 of 25), and rewriting code 20% (5 out of 25). Finally, most of the projects in the exploration phase (53%) or without a BDA plan (71.4%) use a single environment approach (i.e. data lab and production are the same environment).

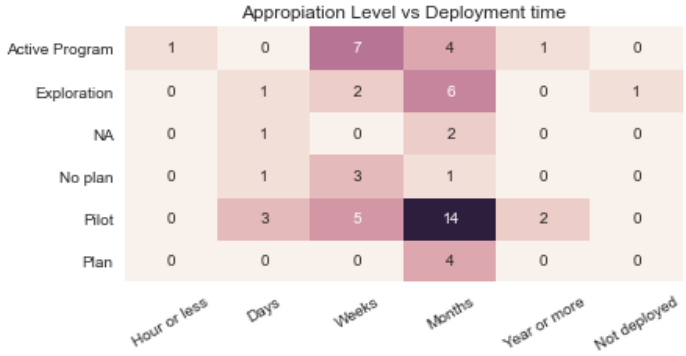

Fig. 8. Appropriation Level and Deployment Time.

4.5 Quality Attributes

The quality attributes drive the architecture of software solutions. In BDA context, it is also true. Hence it is valuable in this research to know how stakeholders

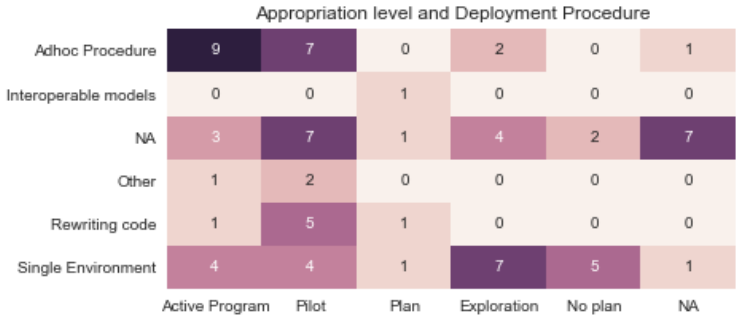

Fig. 9. Appropriation Level and Deployment Procedure.

deal with the trade-offs among quality attributes. For this reason, we formulated a set of questions oriented to answer RQ1.

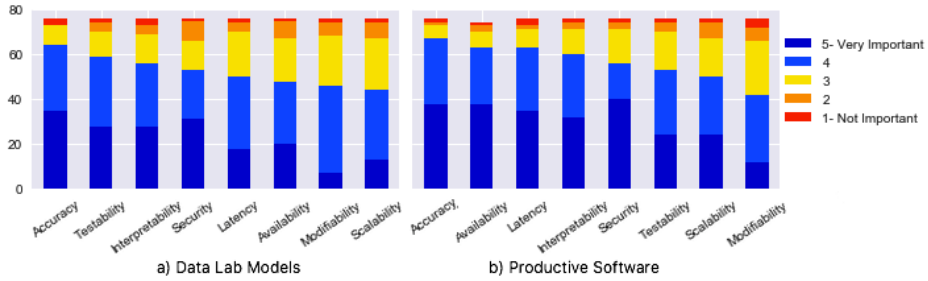

Fig. 10. Quality Attributes Relevance in the a) Data Lab and b) Productive Software Solution.

Fig. 10a details the weights of relevance (from 1 to 5) for each QA when analytics techniques and models are selected, built and evaluated in the data lab environment. The most weighted QA is accuracy with 84.2% of positive ratings (i.e. greater than 3), followed by testability (77.6%), interpretability (73.6%), and, security (69.7%) and response time (65.7%) complete the top 5. Availability and scalability observe the lower ratings (63.1%, 60.5%, and 57.8% respectively) of relevance inside the data lab.

On the other hand, the same question about QA's relevance was made, but in the production environment to compare the quality's priorities. Fig. 10b shows that accuracy continues in the first place with 88.1% of respondent's positive ratings (i.e. greater than 3). The second and third places are occupied by performance QAs: availability (82.8%) and response time (82.8%). Interpretability fall to fourth place with 78.9% of positive ratings and security ends the top 5 list with 73.6%. Despite the fact scalability and modifiability maintain the last

two places (65.7% and 55.2% respectively), it is worth to note that scalability increases the rating of *Very important* from 17.1% to 31.5%.

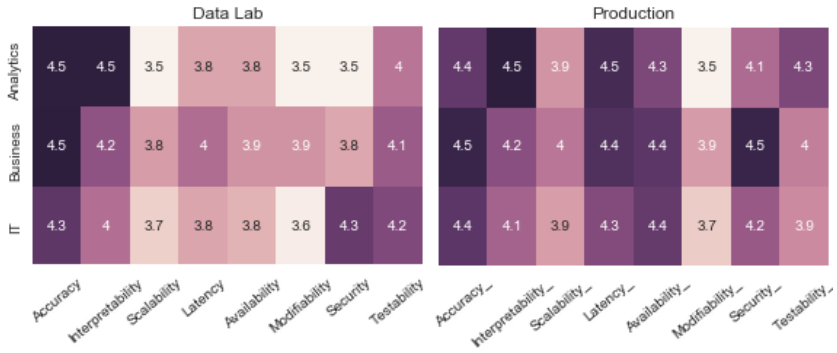

Fig. 11. Quality attributes relevance regarding stakeholder domains.

Fig. 11 reports QA relevance averages (from 1-Not Important to 5-Very Important) in the data lab and production regarding the stakeholder domains. In the data lab, accuracy observes the highest relevance for all stakeholders with slight differences in magnitude. On a second level, analytics (data scientists) and business stakeholders rank interpretability and testability, while IT stakeholders prioritize security and testability, respectively. In the production environment, the picture changes significantly. Data scientists give more relevance to interpretability and latency, while business users prioritize accuracy and security. IT users rate accuracy and availability with the highest scores. Comparing the relevance scores between data lab and production, the differences in latency, availability, scalability, and security for all stakeholders are remarkable, evidencing a clear change of QA consideration between environments.

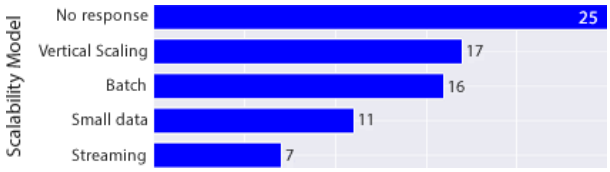

Fig. 12. Scaling approaches.

Finally, we included a question to know how is the scalability capacity to support the BDA context and Fig. 12 summarizes the respondent's answers. The most noticeable result is that most of the respondents do not know/do not respond (32.8%, 25 out of 76), which could reflect the lack of knowledge or in-

terest about the technical capabilities to support big data processing. Vertical scaling based on robust appliances is the most used approach with 22.3%. Distributed batch processing using big data frameworks such as Hadoop or Spark is used by 21.1% of respondents, 14.4% declared do not have scaling capabilities because they only work with small data. Distributed streaming processing is only required by 9.2% of the respondents.

5 Discussion

The BDA adoption and appropriation among companies is incipient as shown by results in which 47% have only developed between 1 and 3 projects, and only 23.6% have an active BDA program. This situation is slightly better compared to a report presented by the Colombian IT Ministry [11] that calculates the adoption of big data technologies of 16.8% in big enterprises. Compared to a previous worldwide report in 2016 [7], our survey reports better levels of appropriation in terms of the proportion of active programs in organizations (23.6% versus 17%), pilot programs (32.8% versus 17%) and "no–BDA plans" (9.2% versus 23%). In contrast, we find lower indicators regarding organizations in phases of exploration (17.1% versus 32%) and plans to be implemented (5.2% versus 11%). These results could suggest a growing interest in companies for BDA adoption and their respective progress over time.

This survey found that classic analytics techniques such as aggregations, regression, and clustering are the most used by companies. These results are similar to previous studies [6,7], the only exception is that in our survey, the decision tree is not ranked in the top three of the most used algorithms. The most basic tools like Excel and SQL scripts are in the first places, followed by Tableau and R. These preferences are different from specific data science studies where R, SPSS, SAS, and Tableau occupied the top positions. This can suggest unfamiliarity or lack of skills in data science-oriented tools in the Colombian context. This survey also reports a lack of standard procedures to deploy and operate BDA solutions which frequently implies manual code rewriting and configuration, confirming findings presented previously in [8]. It is noticeable the lack of knowledge and use of de-facto standards (1.3%) for sharing analytics models across technologies (such as PMML or PFA) compared to previous studies (19%) such as [8], what can promote the cumbersome and delayed process of putting analytics services in operation. These findings allow us to argue that DevOps practices in these specific domains are still unknown, immature, or under-used, and some recent works such as [12,13] have addressed this concern.

Activities involved in BDA development, such as business objectives and analytics goals definition, data collection, and deployment, are considered "hard" on average. Specifically, deployment seems to be a challenging stage, probably due to different factors such as software development driven by competing QAs in different environments, tools heterogeneity, and the lack of mature deployment procedures, even in organizations with active BDA programs. These factors have also been identified in previous works [8,7]. Teamwork and collaboration between

data scientists and IT stakeholders are better ranked compared to business/IT and business/data scientist interaction.

In terms of deployment challenges, our results confirm issues in different facets: scarcity of deployments into production leading to low operationalization of BDA solutions and long delays for deployment which range from weeks to months (63%). This scenario can be caused partly by technical reasons such as inadequate tools, and inadequate procedures to deploy and retrain BDA solutions in production environments. These findings coincide with conclusions reported in [7] and [8] where they reported low rates of deployment, lack of procedures to deploy BDA solutions, and long deployment times. Even companies in a more mature BDA stages (i.e. with active programs) reported deployment times from weeks to months.

Relevant QAs during the data analytics modeling are not the same as those during the software development phase. The reason for this is that both artifacts (models and software) pursue different objectives, while the analytics model's quality is measured by the accuracy, interpretability, and testability, BDA software must achieve expected performance metrics such as availability, response time, and scalability. This can lead to competing drivers when the software architect makes decisions (i.e. patterns, tactics, technologies) which may differ for the same analytics solution in different environments. This situation could also lead to heterogeneity of technology tools reported along the BDA life cycle.

6 Threats to validity

In our study, the research methodology was validated to avoid biases as much as possible. In the following, construct validity, internal validity, external validity and reliability are presented together with their mitigation strategies as reported by Runeson and Martin schema [14].

Construct validity. It reflects the relation between operational measures studied and researcher's main idea, according to the research questions [14]. The phrasing used in sentences for closed-ended questions could be the most recurrent threat in questionnaire-based surveys. In order to mitigate this thread, we first piloted the survey internally several times and then piloted the survey externally with practitioners involved in BDA projects through an online survey what allow as to refine the used language.

Another risk is related to participants did not finding any suitable response in the set of available ones. For this, our strategy was included an "Other" answer for each question. In our results, we had a relatively low number of respondents using this alternative answer.

Internal validity. It reflects the presence of causal relations affecting the investigated factor [14]. For this, we performed analysis of the data using basic descriptive statistics and performed cross-analysis of the responses of each participant. We also provided definitions that are used consistently in the survey allowing the respondents to fully understand the questions asked.

External validity. It reflects the possibility of generalize the findings, and to discover if the findings are of interest to other people outside the investigated case [14]. For our study, a potential threat refers to the demographic distribution of response samples. We applied Convenience sampling to helped us in selecting study participants. However, we are aware that this sampling technique could have had a negative impact on the size of the set of respondents. To mitigate this potential threat, we ensure that the set of respondents were an heterogenous sample in terms of demographic information, such as professional experience, educational background, number of projects, etc (Section 4.1).

Reliability. It reflects the independence between the extracted data and the obtained results [14]. To mitigate this threat, we employed observer triangulation, having all authors participating in the data extraction and analysis processes. Due to the non-statistical nature of convenience sampling used in this study, we cannot give strong inferences, and we also avoid performing any statistical correlation analysis because we are aware our sample size is small and too centered in practitioners who have participated in BDA projects. Despite of this fact, our results can open new discussions and research lines.

7 Conclusions

We have presented an empirical study of how practitioners deal with the development and deployment of BDA solutions. We first developed and evaluated a pilot to design a paper-based survey. The data extracted from the questionnaires' answers provide clues for understanding activities, behavior, practices, and challenges faced by practitioners.

Our results open new research directions within the software architecture and software engineering community related to BDA procedures, methodologies, and design. The definition of the project's business goals, alignment between business goals and analytics task, and solution deployment were reported as the most challenging activities in BDA life cycle. We found communication and interoperability concerns across knowledge domains within BDA life cycle. Our results also found competing QAs (e.g. testability and interpretability vs performance) when developing analytics models compared to BDA software. Heterogeneity of technology tools and immature or little-known deployment procedures could lead to delayed and sporadic deployments which hinder BDA appropriation.

Regarding the practice of software architecture, our results offer insights about how to plan and design BDA solutions regarding the related challenges and procedures, and the deployment barriers to be tackled in advance. In addition, the most common methodologies, techniques, and tools in the industry could be a starting point to define a BDA adoption road map.

As future work, we can extend this survey by applying it on a wider and varied population in a regional or worldwide scale. We are researching on methodologies and frameworks in the BDA context which consider separation of concerns among the knowledge domains to reduce the deployment gap by integrating and interoperating business, analytics, software, and IT specifications.

8 Acknowledgment

This research is supported by Fulbright Colombia and the Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA), supported by the Ministry of Information Technologies and Telecommunications of the Republic of Colombia (MinTIC) through the Colombian Administrative Department of Science, Technology, and Innovation (COLCIENCIAS) within contract No. FP44842-anexo46-2015.

References

1. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 Step-by-step data mining guide. Technical report, The CRISP-DM consortium (August 2000)
2. IBM: Foundational methodology for data science. `http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IMW14824USEN` (2015) Accessed: 2017-07-11.
3. Chen, H.M., Kazman, R., Matthes, F.: Demystifying Big Data Adoption: Beyond IT Fashion and Relative Advantage. In: Twentieth DIGIT Workshop, Texas, US (2015) 1–14
4. Chen, H.M., Schütz, R., Kazman, R., Matthes, F.: How Lufthansa Capitalized on Big Data for Business Model Renovation. MIS Quarterly Executive **1615**(14) (2017) 299–320
5. Kitchenham, B., Pfleeger, S.: Personal Opinion Surveys. In: Guide to Advanced Empirical Software Engineering. Springer, London (2008) 63–92
6. Rexer, K.: 2013 Data Miner Survey. Technical report, Rexer Analytics (2013)
7. Rexer, K., Gearan, P., Allen, H.: 2015 Data Science Survey. Technical report, Rexer Analytics (2016)
8. Dataiku: Building Production-Ready Predictive Analytics. `http://asiandatascience.com/wp-content/uploads/2017/12/Production-Survey-Report.pdf` (2017) Accessed: 2017-07-11.
9. LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big bata, analytics and the path from insights to value. MIT Sloan Management Review **52**(2) (2011) 21
10. Easterbrook, S., Singer, J., Storey, M.A., Damian, D. In: Selecting Empirical Methods for Software Engineering Research. Springer London, London (2008) 285–311
11. Katz, R.L.: El Observatorio de la Economía Digital de Colombia. Technical report, Ministerio de Tecnologías de la Información y las Comunicaciones (2017)
12. Castellanos, C., Correal, D., Rodriguez, J.D.: Executing Architectural Models for Big Data Analytics. In Cuesta, C.E., Garlan, D., Pérez, J., eds.: Software Architecture, Springer International Publishing (2018) 364–371
13. Lechevalier, D., Ak, R., Lee, Y.T., Hudak, S., Foufou, S.: A Neural Network Meta-Model and its Application for Manufacturing. In: 2015 IEEE International Conference on Big Data. (2015)
14. Runeson, P., Höst, M.: Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering **14**(2) (Dec 2008) 131