UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Alex Bruno Paranhos da Silva
Daniel Amador dos Santos

PROTOCOLO DO ESTUDO DE CASO
DATA ANALYTICS NA FAPESB

SALVADOR – BA
Setembro de 2016

1. Background

According to Menzies & Zimmermann (2012), software analytics refers to the use of data analysis and exploration to help decision makers extract important informations and insights. The use of software analytics is frequently in the area of software development, helping engineerings and project managers make decisions about their software projects. Zhang et al. (2013) depicts software analytics as a trinity of research topics: development process, system, users.

Menzies & Zimmermann (2012) list some good principles to engineer software analytics systems in a fashionable manner. They actually use Menzies et al. (2011) list in a paper in which these authors produce a manifesto for what they call Inductive Software Engineering. These good principles are not intended to be a formal maturity model, but rather good practices that the authors judged to be valuable based on their experiences during many years in contact with both academia and industry performing research in data mining.

So these are the principles: “users before algorithms”, “plan for scale”, “early feedback”, “be open-minded”, “do smart learning”, “live with the data you have” and “broad skill set, big toolkit”. Along with the principles themselves, the authors of the manifesto describe more tips and guidance to identify each one and apply it effectively. Such information is available in Menzies et al. (2011).

Since Menzies & Zimmermann encourage the usage of these principles to software analytics also, we are able to infer that, if a software development team is aware of them when developing for analytics, these principles are then a good indicator that this team might be on a right track in order to implement a full-fledged software analytics system when it wishes so.

The objects of this study are the practices of inductive software engineering and practices to deal with the implementation of non-functional requirements, both in a small software development team. The main goal of this case study is to **understand the software engineering practices used to create data analytics solutions.** To achieve this objective we decide to answer the following research questions:

RQ1 Are Inductive Software Engineering Principles followed by the development team?

RQ2 What are the non-functional requirements used in the development of data analytics solutions?

RQ2.1 How these non-functional requirements are specified, designed, implemented and tested?

2. Design

The goal of this study is to understand the software engineering practices used during data analytics system's development. We also want to investigate if from the developers' point of view they consider that those practices and principles bring advantages to data analytics software production process.

Since there is only one development team in FAPESB, and since we do not know significant hierarchical subdivisions beforehand, the study will be conducted as a holistic single-case one. As soon as we identify concurrent teams, we might consider them as different units of analysis, which would make the case as a holistic embedded case study.

For RQ2 and RQ2.1, we also would like to collect some feedback from the developers. The interview questions could induct them to opinate how they think some principle affects (if followed) or would affect (if not followed) the target organization systems development process. For the latter case, they might think that some propositions are not worthy to be applied, and/or would not bring any real advantage comparing to the existing developing procedures. We would like them to elaborate about those concerns. Moreover, we would like to analyze the artifacts produced during the development.

3. Selection

For the case study to be performed, FAPESB (Fundação de Amparo à Pesquisa do Estado da Bahia) was selected. FAPESB is a government organization that aids scientific research in the state of Bahia, Brazil. This organization has a software development team. The FAPESB team was chosen to be subject of the case study because:

- The evident rising need for research data analytics also in small development teams.
- The current president of the organization, Dr. Eduardo Almeida has authority in the organization so he can give the researchers access to both personnel and physical locations.
- The team is small enough so the researchers can take up all the interviewing and analyzing process feasibly in the schedule proposed. Yet, this team still develops real world data-oriented systems, which can bring valuable insights.

4. Procedures and roles

In order to achieve our objective, the research team will conduct a number of interviews with the development team. Each member will answer questions regarding the principles of the Inductive Software Engineering Manifesto. However, this manifesto or even terminologies like software analytics, data mining or anything related will not be explicit for the subjects.

For the interview session, a semi-structured session will be used. Besides the issues themselves we are intending to find out, using this kind of structure may give us surprisingly useful insights. Also, we consider important that the interviewees, to some extent, feel that they are in a conversation, and less in a highly rigid methodical interview. The questions will be placed following the funnel general principle: from less complex/specific to more complex/specific. The complete interview questions outline is yet to be developed.

Prior to the interview, a pilot will be conducted with a single developer member (since the team is small, it is not wise to waste more subjects with the pilot). This so called pre-session will be useful to assess if the interview procedures and the questions fit well for the researchers' needs. If they do not, the procedures identified as inadequate or not working properly might be reviewed and redesigned, as well as the questions.

The interview's core questions are the ones indicated in Figure 1. In order to make the interviewees comfortable and to gather information about their context, they will be asked about how is the workplace and their views about FAPESB. After this phase, the researchers start asking the central questions, that relate directly to the principles.

The research team is composed by the authors of this protocol, namely: Alex Bruno Paranhos da Silva and Daniel Amador dos Santos. The researchers will conduct the interviews with the developers alternatively along the session. While one perform the interview, the other one assist the session with possible forgotten issues or may raise additional questions if the interviewee behave friendly and open.

5. Data collection

The data to be collected will be the actual recorded answers from the interviewees, in audio format. These data will be stored in a repository using a Concurrent Versioning System (CVS).

The audio files containing the sessions will be heard by the researchers separately. Both will take their conclusions. Later, the conclusions will be compared by the researchers, and in a case of disagreement of the results, the researchers will then work to get in a consensus. This might make emerge different point of views of the declarations of the interviewees and might expose both researchers to a different point of view about the answers.

We intend to verify some documentation in order to confirm or rebuke the statements made by the developers. However, at this point we still do not know how much FAPESB team documents its work.

6. Analysis

After listening to the audio sessions, the researchers will identify from the answers if the Inductive Software Engineering principles are being applied. Each principle may be classified as applied, not applied and partially, according to each developer's perception. After that, we can come to a conclusion if the team's feeling is that the principles are used, not used and why so. A further reading of the situation might point if the team consider those principles valuable or not.

Each question can be tracked to a Inductive Software Engineering principle. The correspondence of the central questions can be seen in Figure 1. Since FAPESB does not make use of data analytics, i.e. the usage of data relations to make insightful information for developers, we consider that the principles "live with the data you have" and "broad skill set, big toolkit" do not apply.

Later, possible explanations of why some principle is applied or not will be considered. We can assess if many developers indicate the same cause for the usage or not of the principles. Similar answers might indicate a likely pattern. It is also expected the likelihood that the explanations may contradict the actual answer sometimes. All of this will be debated in the final report. From the answers of the question that asks how something is implemented or not we will build explanations (when the answers agree) or we will state the differences on the developer views on a certain principle (when the answers are divergent).

Interview Question	Relates to Principle
IQ1 - Before implementing a specific algorithm or data processing method, do you think first if this is valuable for your client? How?	Users before Algorithms
IQ2 - Do you consider your current systems can deal scalably with data? In which way?	Plan for scale
IQ3 - When developing a new feature, how early do you check your client for feedback? Do you consider it makes a big difference in the development process?	Early feedback

IQ4 - For developing systems that deals with data, do you adopt new approaches for new types or data or do you have a pre-established way to deal with them? Do you think it could be different?	Be open-minded
IQ5 - Do you acquire new knowledge from the data you get? If not, do you imagine something valuable you could take from that?	Do smart learning

Figure 1: Interview Questions

If documentation is available, it will be used as additional data to make information triangulation. There is also a possibility that the documentation does not bring anything relevant to build a solid conclusion. In this case, the documents will not be used either. Nevertheless, we will report why the documentation was relevant or not.

7. Plan validity

About the construct validity, the communicability of the questions will be verified in terms of the questions are really asking what the researchers intend to. That will be assessed in the pilot. If the answers in this session relates to something substantially different from what is asked, the questions might be not clear enough. If they are not, the set of questions will be then refined for the actual interview session. Also, if the documentation brings strong relevant evidence hence will be used.

It is essential that the subjects do not communicate the answers to each other. This could make later interviewees follow the responses of the first ones. That phenomenon would put the internal validity at risk. So we should ask the participants explicitly to keep the answers in secret until the end of the interview session. Even so, there is no way to obligate the subjects to not communicate their opinions each other. Therefore, we will apply this measure to mitigate this risk, but it still exists as threat to internal validity.

Another issue concerning internal validity is if the subjects feel comfortable enough to share testimonials about the daily routine inside the organization. Before the actual interview

session, the interviewers must state clearly that they not intend to supervise their job, and the information about each employee will not be used to harm them in any ways.

With this case study the researchers can draw conclusions about if an organization that manages to do some work with data analyzing with a small development might have seen the need to adopt some practices that agrees with the Inductive Software Engineering Manifesto. If so, it would be interesting observe that totally isolate contexts have led to similar ideas or principles about solving data mining challenges. The researchers are assuming that FAPESB team has never heard anything about the Inductive Software Engineering before due the novelty of this research topic. A future work reporting the implementation of software analytics could be interesting as it would be possible to investigate if the pre-existing using of the principles have facilitated this process somehow. A further contribution would be compare this case against other ones.

8. Study Limitations

According to Singer et al. (2007) the study often reveals a couple of limitations. Therefore, if the they can be spotted early enough, these difficulties can be defeated or mitigated in the case study design. That is why the pilot session is so necessary.

One limitation is the development team size. Working with a more numerous group could ground a more consistent theory with the opinions of the developers concerning the principles presented. Therefore, a further similar case study can bring more insight to the issues raised by this research.

Another factor that could make the case more appealing would be if the organization already had a full operational data analytics operation. If that happen, this study could give more substantial contribution to existing research in this topic.

9. Reporting

In Singer et. al (2007) is discussed about the importance of reporting the findings in the study, as well as others documents, for example, the protocol and the outcomes collected from the study, observing ethical principles.

At the end of the study a report will be produced in the IEEE Software format containing all the findings and the lessons learned through the research. Moreover, all the versions of this protocol, the interviews transcripts, the results of the data analysis, as well as others relevant artifacts, will be available.

The audience of this report will be the academic community of software analytics and Inductive Software Engineering. The findings in this study can help researchers to analyse methods and techniques applied at the development of the systems and relate with the principles of the Inductive Software Engineering described in Menzies et al. (2011), to support replication, as well as to judge the credibility and the value of the study.

Another target of this report are the stakeholders from FAPESB, which may get advantage if they intend to implement a full-fledged Software Analytic system. This report could make thems verify how far they are from implementing the Inductive Software Engineering principles. By applying them, FAPESB could have a set of guidance tips from experienced researchers in the area of data mining, so a report assessing their current status towards these practices could be quite handy to them.

10. Schedule

1. Write the first version of the protocol (until September 08, 2016).
2. Define the final interview script (until September 12, 2016).
3. Execute the interviews and collect data (2 days, depending on the interviewee's availability).
4. Transcribe interviews (7 days)*.
5. Analyze the data (5 days)*.

6. Write report (until October 21, 2016)*.

* Activities with an asterisk may be performed in parallel.

11. References

Forrest Shull, Janice Singer, and Dag I.K. Sjøberg. 2007. Guide to Advanced Empirical Software Engineering. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Menzies, T. and Zimmermann, T., 2013. Software analytics: so what?. *IEEE Software*, *30*(4), pp.31-37.

Menzies, T., Bird, C., Zimmermann, T., Schulte, W. and Kocaganeli, E., 2011, November. The inductive software engineering manifesto: principles for industrial data mining. In *Proceedings of the International Workshop on Machine Learning Technologies in Software Engineering* (pp. 19-26). ACM.

Zhang, D., Han, S., Dang, Y., Lou, J.G., Zhang, H. and Xie, T., 2013. Software analytics in practice. *IEEE software*, *30*(5), pp.30-37.