Submitted By: Areeba Nadeem

Title:

From Ad-Hoc Data Analytics to DataOps

Author:

Aiswarya Munappy, David Issa Mattos, Jan Bosch, Helena Holmström Olsson, Anas Dakkak

Conference name:

International Conference on Software and Systems Process

Introduction And Motivation:

Data is the key asset for the organizations as it helps in better decision making, analyses performance and solving problems, to analyses the consumer behavior and market and so on. Moreover, data is the backbone for many hot and trending technologies like machine learning and deep learning [1].Increased importance of data lead to the acquisition and storage of data in higher volumes which in turn gave rise to fields like Big Data, data mining and data warehousing. Due to the operations initiated by the engineers or by the change in data sources, continuous data changes happen and there is the requirement for data versioning and sharing techniques. Thus, data management becomes vital for all organizations to collect, store, organize, protect, verify, and process essential data. Data being the fuel for digital economy need for data products like machine learning datasets, dashboards and visualizations is tremendously increasing. Organizations invest in data science and data analytics to solve problems with the collected data. Organizations realize that data is the key factor of success and as a result they invest an enormous amount of money in the development of data products. Data products are built through a sequence of steps called data life cycle wherein for each step there will be both hardware and software requirements. Due to this reason, it is very essential to find the right balance of investment on requirements in different stages of data life cycle. Data management, data life cycle management, data pipeline robustness, fast delivery of high quality insights are some of the major data problems that prevents companies to achieve their full potential. Dev Ops is a methodology adapted in Software Engineering to aid agile software development. Agile methodology focuses on empowering individuals, rapid production of working software, close collaboration with customers and quick response to the change in customer requirements. Agile development is directly facilitated by CI/CD practices because it aids in software changes reaching production more frequently and rapidly. Consequently, customers get more opportunities to experience and provide feedback on changes [6]. Industries apply agile methodology, Data Ops and CI/CD methodologies in the software development. Data being an artifact like code, data analytics can also be benefitted by the application of best practices of these methodologies in data analytics. Data Ops is a process oriented methodology which is derived

from Data Ops, continuous integration/continuous delivery and agile methodology for the quick delivery of high quality results to the customers. The contribution of this paper is three-fold. First, it analyses the various definitions of Data Ops from the literature as well as from the interviewers and then derive a definition for Data Ops including the main components identified. Second, we present the phases the teams at Ericsson evolved through for better delivery of insights. Third, we create a stairway of evolution process. The rest of this paper is organized into six sections. Section II is a description of the background and related work. In section III, the research methodology adopted for conducting the study is introduced. Section IV focuses on findings of the case study, framing the definition for Data Ops and the evolution stages. Section V details threats to validity and finally section VI summarizes our conclusions and completes this paper.

Research Methodology:

The goals of this study was to formulate a definition for Data Ops and to identify the phases of Data Ops evolution.

A. Setting the RQs The RQs defined in the study are as given below:-

• RQ1. How do practitioners define "Data Ops"? The goal of this research question is to achieve an aligned understanding on the concept of Data Ops, aiding communication among researchers and practitioners when presenting results related to Data Ops.

• RQ2. What are the different maturity stages Ericsson has gone through while trying to evolve from ad-hoc data analysis to Data Ops? This RQ seeks to identify data strategies that are used at Ericsson to do Data Analytics in industrial systems and also the drivers for adoption at each stage to move to the successive stages. overlap between the two data analytic approaches and the practical difficulties Ericsson encounter while trying to completely adopt Data Ops as their analytic approach To set the basic understanding on Data Ops concepts and the essential components, we adopted the Multi-Vocal Literature Review approach following the instructions given by [10]. Then we conducted an interpretive single-case study, following the guidelines by [11], The overall research design and major steps in the process of the study are described below.

B. Multi-Vocal Literature Review An MLR is a form of a Systematic Literature Review (SLR), which includes the Grey literature in addition to the published literature (e.g., journal and conference papers) [13]. Grey literature in SE can be defined as any material about Software Engineering that is not formally peer-reviewed nor formally published. The multi-vocal literature review approach was selected because it allowed us to gain more understanding on Data Ops practices, we analyzed if there is a great potential for benefiting from grey literature in Data Ops study and we identified that clearly this approach is the best suited one for studying Data Ops. Because, the formal literature on the other hand Data Ops is highly limited and on the other hand, there are quite a number of blogs, video media and technical reports. Moreover, MLRs are useful since they can provide summaries of both the state-of-the art and practice in a given area. We searched the academic literature using the Google Scholar, IEEE Explore, ACM digital library and the grey literature using the regular Google search engine.

C. Need for MLR:

To learn more about the concept of Data Ops, we did an initial search for the formal academic literature in different databases such as Google Scholar, IEEE Explore, ACM digital library, Web of Science, Scopus and Science Direct. However, we could not find a considerable number of peer reviewed papers on the topic. Consequently, we decided to conduct a Multi vocal Literature Review, based on all available literature on a topic. According to Ogawa et.al a broader view about a particular topic can be obtained by using this wide spectrum of literature as they include the voices and opinions of academics, practitioners, independent researchers, development firms, and others who have experience on the topic [14]. Grouse et al. states that the practitioners produce literature based on their experience, but most of them are not published as academic literature. Also, voice of the practitioners better reflect the important current state-of-the-art practice in SE. Therefore, it is important to include Grey literature too in the systematic review

D. Process of MLR: The Multi-vocal literature review procedure adopted for the study is demonstrated in Fig. 1. The systematic review employs string-based database search to select relevant studies from the literature. All retrieved literature was exported to MS Excel for further processing. The exported references were screened based on inclusion exclusion criteria. The inclusion and exclusion criteria considered in our study are as shown below.

• Inclusion Criteria:

(1) Papers and Google links describing the steps of Data Ops approach, essential components of Data Ops, benefits and challenges.

(2) Papers describing the Bigdata pipelines, Big data processing pipelines

• Exclusion Criteria: (1) Duplicates and non-English

E. Exploratory Case study: The study was conducted in collaboration with Ericsson. Ericsson is a Swedish multinational network and telecommunications company. The company provides services, software and infrastructure in information and communications technology. The objective of the study is to explore the essential stages of Data Analytic approach which Ericsson follows in their real-world settings and also to investigate its similarity to the popular Data Ops approach. Each case in the study refers to a team at Ericsson working with the data they collect from different sources. For the study, a sample pool of Data Fig. 1. Multi-vocal literature review procedure applied in the study Scientists, Data Analysts and Data Engineers were selected by one of the authors according to their expertise in the area of Data Analytics. Request to participate in the study was sent and after the interviews, interviewees suggested some of their colleagues and all together 10 interviews were conducted for this study. Table 1 illustrates the role of our interviewees and the use cases. Data Scientist team R7 Data Scientist H Building data pipelines for CI and CD data R8 Program Manager

F. Data Collection:

Empirical data was collected through semi-structured interviews were used to acquire qualitative data. Based on the objective of research to explore data analytic approach employed at Ericsson, an interview guide with 45 questions categorized into six sections was formulated. The first and second sections concentrated on the background of the interviewee. The third and fourth sections focused on the data collection and processing in various use-cases and the last section inquired in detail about data testing and monitoring practices and the impediments faced during every phase of the data pipeline. The interview guide was prepared by the firs Exploratory Case study t author and was reviewed by all the other authors. Based on the comments and recommendations some additional questions were added, a few similar questions were merged together and some irrelevant questions were removed finally forming an interview protocol with 30 questions spread across six different categories. All interviews were conducted via video conferencing except for three which were done face-to-face and each interview lasted 50 to 100 minutes. All the interviews were recorded with the permission of respondents and were transcribed later for analysis. One of the authors of this paper is an Ericsson employee who works quite a lot with the data teams. First two authors of this paper are consultants at Ericsson and attend weekly meetings with Data Scientists and Data Analysts. Data collected through the meetings and discussions are also incorporated. The contact points at Ericsson were also a great help while validating the collected data.

G. Data Analysis: After the interviews, audio recordings of interview were sent for transcription and a summary of each interview was prepared by the first author highlighting the important focus points of the interview. The investigated points from the summary were cross-checked several times with the audio recordings and interview transcripts obtained after transcription. A theoretical thematic data analysis approach was selected for coding [15]. The first author coded each relevant segment of the interview transcript in Vivo. For the first iteration, the objective was to identify the use-cases discussed by each interviewee and phases of data analytics used by their team. After identifying the phases, a second iteration was performed to investigate the impediments encountered to completely set up Data Ops practices at Ericsson. Thematic coding was performed by setting high level themes as (i) Data Collection, (ii) Data Analytics (iii) Data Ops (iv) Automation (v) Data Testing, (vi) Data monitoring, (vii) Agile environment. After careful analysis of collected data, the first two authors agreed on the presentation of results in the paper. From the analysis, results were tabulated and sent to the other authors for their reflections and then the final summary of the cases and results were sent to the interviewees for validating the inferred results.

Conclusion:

The collection of high quality data gives companies a significant competitive advantage in their decision making process. It helps in understanding customer behavior and enables the use and deployment of new technologies based on machine learning.