Software Engineering for Big Data Projects: Domains, Methodologies and Gaps

Vijay Dipti Kumar  
David R. Cheriton School of Computer Science  
University of Waterloo  
Waterloo, Canada  
vdkumar@uwaterloo.ca

Paulo Alencar  
David R. Cheriton School of Computer Science  
University of Waterloo  
Waterloo, Canada  
palencar@cs.uwaterloo.ca

Abstract—Context: Big data has become the new buzzword in the information and communication technology industry. Researchers and major corporations are looking into big data applications to extract the maximum value from the data available to them. However, developing and maintaining stable and scalable big data applications is still a distant milestone. Objective: To look at existing research on how software engineering concepts, namely the phases of the software development project life cycle (SDPLC), can help build better big data application projects. Method: A literature survey was performed. A manual search covered papers returned by search engines resulting in approximately 2,000 papers being searched and 170 papers selected for review. Results: The search results helped in identifying data rich application projects that have the potential to utilize big data successfully. The review helped in exploring SDPLC phases in the context of big data applications and performing a gap analysis of the phases that have yet to see detailed research efforts but deserve attention.

Keywords—Big data, software engineering, application projects, domains, methodologies, gaps, literature review.

I. INTRODUCTION

There have been several success stories of big data being used by technology giants to dominate their competitors in areas such as social media, search engines, e-commerce and video streaming services. Few of the popular players and leaders in these fields leveraging big data are Facebook, LinkedIn, Twitter, Google, Amazon, and Netflix. The success of these big data use cases have piqued the interest of and spurred numerous companies to take an interest in developing big data applications, to extract maximum value from all the data that is becoming available.

According to a Gartner survey1, 64% out of 720 respondents of the survey had invested or planned to invest in big data applications in 2013. However, less than 8% had actually deployed at the time of the survey. As shown in a case study of optimizing the manufacturing process of digital displays [1], it is possible to enhance process efficiency using big data. Different sectors such as healthcare, trading, agriculture, tourism and politics, in which big data is being used to stakeholders’ advantage, are detailed in [2]. Enriching customer experience using big data has been exemplified by the “People you may know” option offered by Facebook and LinkedIn, movie recommendations by Netflix, and even the “Customers Who Bought This Item Also Bought” service provided by Amazon.

A. Big Data and Software Engineering

Big data has been characterised by 4Vs: Volume, Velocity, Variety and Veracity. Volume implies the data explosion that has come to mark the last decade in the world of computing. Velocity is the constraint that demands real time processing of the data available, the failure of which would result in its loss or obsolescence. Variety means data format could be structured, semi-structured or unstructured or have multiple sources and applications must be capable of handling these sources. Finally, veracity implies that the available data, historical or even real-time streaming data, might have to be cleaned prior to processing to ensure its usefulness to the application.

The whole process of developing software is fraught with errors which arise due to changes in requirements, the environment or due to communication issues between the stakeholders. This is true for software development for ordinary applications even today despite the fact we have been developing software for more than 20 years. When factors such as the 4Vs are taken into account, the complexities involved in developing software for big data application projects only increase.

Despite the complexities involved in following the SDPLC phases in building a big data application, these phases and concepts must be leveraged to build robust and scalable big data applications projects. It would be to the advantage of the stakeholders and developers involved in building a big data application that the best practices and methodologies laid down by the software engineering research community be applied to build systems that are fault tolerant and capable of handling even more data than envisioned at the time of its creation.

B. Motivation

There have been numerous reviews on big data itself. One discusses state of the art in architecture and large scale data analysis platforms [168], one is a comprehensive big data survey [169], another describes the related technologies and the acquisition and applications of big data [170] and one even lists the different definitions of big data [171]. The guest editor’s introduction to one of the issues of the IEEE Software magazine [172] discussed the software engineering challenges in building data-intensive, or big data software

---

1Big Data Gartner survey—http://www.gartner.com/newsroom/id/2593815
TABLE I. APPLICATION DOMAIN CATEGORIES & CLASSIFICATION

<table>
<thead>
<tr>
<th>Application Domain</th>
<th>Papers</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Technology</td>
<td>[1], [3–99]</td>
<td>98</td>
</tr>
<tr>
<td>Healthcare</td>
<td>[100–112]</td>
<td>13</td>
</tr>
<tr>
<td>Geospatial Data Processing/Geographic Information Systems</td>
<td>[4], [101], [113–122]</td>
<td>12</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>[123–133]</td>
<td>11</td>
</tr>
<tr>
<td>Transport</td>
<td>[76], [123], [128], [131], [134–139]</td>
<td>10</td>
</tr>
<tr>
<td>Retail/Tourism/Commerce</td>
<td>[36], [76], [101], [140–144]</td>
<td>8</td>
</tr>
<tr>
<td>Social Networks</td>
<td>[7], [115], [116], [145–148]</td>
<td>7</td>
</tr>
<tr>
<td>Environmental Monitoring/Conservation</td>
<td>[4], [85], [120], [149–151]</td>
<td>6</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>[152–154]</td>
<td>3</td>
</tr>
<tr>
<td>Meteorology</td>
<td>[118], [155], [156]</td>
<td>3</td>
</tr>
<tr>
<td>Cyber Physical Systems</td>
<td>[157–159]</td>
<td>3</td>
</tr>
<tr>
<td>Law &amp; Order/Criminal Investigation/Forensic Analysis</td>
<td>[23], [160], [161]</td>
<td>3</td>
</tr>
<tr>
<td>Agriculture</td>
<td>[162], [163]</td>
<td>2</td>
</tr>
<tr>
<td>Banking and Financial Industry</td>
<td>[164], [165]</td>
<td>2</td>
</tr>
<tr>
<td>Military</td>
<td>[161], [166]</td>
<td>2</td>
</tr>
<tr>
<td>Aviation Industry</td>
<td>[159]</td>
<td>1</td>
</tr>
<tr>
<td>Astronomy</td>
<td>[167]</td>
<td>1</td>
</tr>
<tr>
<td>National Security</td>
<td>[161]</td>
<td>1</td>
</tr>
</tbody>
</table>

systems. However, no comprehensive study reviewing existing software engineering research methodologies for enabling development of big data application project was found.

C. Goals and Contributions

The main goal of this study was to look into the existing research on applying software engineering to big data application projects. The results from this literature survey help identify the domains that have been studied in detail for big data application development. It also helped recognize the SDPLC phases that were most commonly being utilized in developing big data applications. Additional results include pinpointing the domains that have the potential to deploy big data application projects with advantageous outcomes and have seen early research efforts in the field but remain underexplored and definitely deserve more attention from researchers.

II. RESEARCH METHOD

For conducting the literature review, popular academic search engines, namely Scopus, Web of Science and IEEE Xplore Digital Library were targeted. The search was performed using the Command option under Advanced Search section of these search engines. A combination of keywords related to SDPLC phases were selected from the software engineering standard textbook *Software Engineering: A Practitioner’s Approach* [175] to understand which subfields were popular among researchers.

In the context of software engineering, the search terms used were architecture, evolution, process, quality, reuse, specification, requirement(s) engineering, design, “domain modeling”, testing, verification, validation, maintenance, quality, analysis, framework, process, and patterns.

The idea behind formulating the queries was to search for papers that combined topics about software engineering and big data. We have adopted an extensive search and selection process to identify a set of studies that is as complete as possible. Our search process involved automatic and manual searching. An example of the pattern of queries used in our study is as follows:

“big data” AND (engineering OR requirement OR specification OR design OR architecture OR analysis OR testing OR verification OR validation OR maintenance OR framework OR quality OR design OR evolution OR patterns OR process OR reuse OR “domain modeling”)”

The abstract, introduction and conclusion of each search result was examined in detail to ensure the papers were eligible to be included in the study. Reproducibility was not given a priority because a manual search was the main method applied.

Admittedly, the quality of our results depends on the search queries used and the efficiency of the search engines. Numerous publications from relevant conferences related to big data did not show up in the search results because of the SDPLC specific search terms used in the queries. For example, the use of big data in the field of advertising [176] was not covered because there is no specific SDPLC phase referenced or used in this study despite the fact that the paper dealt with an interesting application domain, namely advertising.
<table>
<thead>
<tr>
<th>Software Engineering Subfield</th>
<th>Papers</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements</td>
<td>[17],[18],[21],[24],[26],[29],[55],[70],[74],[81],[98],[136],[142],[150],[153],[159]</td>
<td>16</td>
</tr>
<tr>
<td>Design</td>
<td>[1],[5],[13],[20],[22],[26],[30],[38],[42],[44],[47–49],[55],[57],[60],[65–67],[72],[75],[77],[84],[85],[96],[110],[123],[131],[144],[157],[159]</td>
<td>31</td>
</tr>
<tr>
<td>Architecture</td>
<td>[1],[3–6],[9],[12],[17],[22],[23],[27],[29–33],[35],[40],[43],[50],[52],[54],[56],[61],[64],[68],[73],[78],[79],[82],[85],[89],[92],[93],[95],[97],[99],[101],[102],[104–106],[110–112],[117],[119],[120],[127–129],[133],[137],[138],[141],[146–149],[151],[156],[160],[161],[162],[163],[165],[166],[174]</td>
<td>68</td>
</tr>
<tr>
<td>Testing</td>
<td>[28], [39], [43], [86], [51], [55], [79], [81], [109], [173]</td>
<td>10</td>
</tr>
<tr>
<td>Validation/Verification</td>
<td>[46], [109]</td>
<td>2</td>
</tr>
<tr>
<td>Maintenance</td>
<td>[55], [166]</td>
<td>2</td>
</tr>
<tr>
<td>Quality Assurance</td>
<td>[14], [28], [87], [98], [109], [147]</td>
<td>6</td>
</tr>
<tr>
<td>Domain Specific Languages/Ontology</td>
<td>[4], [19], [41], [53], [72], [94], [108], [121], [127], [141], [146], [149], [160]</td>
<td>13</td>
</tr>
</tbody>
</table>

**A. Research Questions**

The main research questions that were addressed through the study were:

RQ1. Which application domains have received attention for the development of big data application projects and which domains require more attention?

RQ2. Which SDPLC phases were used to enable big data applications and which fields need more research efforts?

**B. Classification Criteria**

The following were the main criteria taken into account when analyzing and categorizing each paper. Under each criteria, multiple categories were identified. The first criteria categorized the papers on the basis of the application domain addressed by the paper and the second criteria identified the SDPLC phase studied.

C1. Which application domain does the paper belong to?

C2. Which SDPLC phase is utilized in the paper?

**III. LIMITATIONS**

This literature review was conducted through a manual search of the results of targeted queries in popular search engines like Scopus, Web of Science and IEEE Xplore Digital Library. Since the main method for reviewing and selecting/discarding the papers was through a manual process, reproducibility of this study was not taken into consideration.

Admittedly, there is room for improvement and opportunity to include more relevant papers. More papers may have been added to the these search engines after the queries for this study were run. Any relevant papers so missed would potentially be due to a matter of timing rather than oversight.

**IV. RESULTS**

The main purpose of this literature review in addition to looking into the existing research of applying software engineering to big data, was to perform a gap analysis of the research till date. The gap analysis helped identify the application domains and the SDPLC phases in the context of big data that have not yet received much attention from researchers but have huge potential.

The application domains identified through this study are illustrated in Table I and the classification of the SDPLC phases for big data addressed in each paper is listed in Table II.

**A. Study Analysis**

RQ1. Which application domains have received attention for the development of big data application projects and which domains require more attention?

Table I lists all the application domains found by analyzing the papers in this study. The papers that proposed new methods or customized versions of existing technology but did not mention a specific domain, such as healthcare, military, or infrastructure, were classified into the “Information Technology” category. Out of the 170 papers selected in this review, the number of papers dealing with this category was 98, as can be seen in the column “Count” of Table I. Nearly 57% of the analyzed papers focused on the Information Technology domain, which indicates that most of the papers are focusing on topics that would directly affect the world of computing. The primary observation was that researchers focused on improving existing technologies to better suit current requirements.

Application domains such as healthcare and the banking and financial industry are commonly believed to be data rich. The healthcare industry is a source of huge amounts of data sourced from the electronic medical records of patients. Data from hospitals, clinics, medical governing bodies and even insurance providers can be mined to study disease affliction rates, patterns and susceptibility trends. Analyzing medical
data can also help in developing innovative treatment methods, customize more effective and economical treatment plans, or even help healthcare professionals in dispensing medication to groups of patients suffering from similar afflictions that have identical medical history of symptoms and reactions.

The banking and financial industry has access to the monetary blueprint of the world. The tremendous amount of transactional data that commercial banks handle on a daily basis can be used to better understand the spending patterns of its customers. Credit card usage, mobile banking application usage patterns, mortgage and credit history of customers can give banks greater insight into the needs of its customers and help tailor their products accordingly, thereby enhancing customer experience and satisfaction.

Other data rich domains that have seen early research are infrastructure, transport and manufacturing. Metropolitan authorities responsible for traffic management and building infrastructure facilities can use current as well as historical data to build smart cities and green buildings that use minimum amounts of energy and water for heating and maintenance.

The aviation industry and the military also have huge potential for using big data for better performance and maintenance of equipment, aircrafts and vehicles. Interestingly, Geographic Information Systems (GIS) have attracted the attention of big data researchers. This domain can enhance and benefit more from the data available from location services that come built into several mobile and Internet applications used today.

Global warming is causing serious damage to our environment and wildlife and meteorologists can use big data from the global weather sensors to make more accurate weather predictions and provide timely natural disaster alerts.

RQ2. Which SDPLC phases were used to enable big data applications and which fields need more research efforts?

Table II provides a breakdown of the SDPLC phases that were addressed by each paper. Numerous papers dealt with methodologies involving software architectures and frameworks and a good number dealt with design. This does not come as a surprise since building big data applications with the latest cutting edge technological developments would primarily deal with the design and architecture of a system. Thus, we see a lot of research done that specifically deals with designs and software architectures or frameworks, which act as templates for building similar software systems.

SDPLC phases such as maintenance, validation, verification and quality assurance of big data applications have not seen much research. Only two papers were identified that explicitly reported research on the topic of verification [46] [109] and were classified under “Verification/Validation” in Table II. Verification and validation processes ensure that the software system has been designed and will function according to the requirements and design foundations laid down before development.

Similarly, only two papers were found that directly dealt with “Maintenance” [55] [166]. Maintenance of the software systems is what keeps the software application running smoothly even after years of deployment. These areas are extremely important to ensure that once a big data application is built, it continues to meet its goals and satisfy its stakeholders’ expectations without compromising performance and remains scalable in the future when even more data could be leveraged.

The requirements subfield fares comparatively better, which could be expected since, although the requirements gathering for big data applications due to the 4V properties of the data involved is a complex process, application data and processing requirements need to be analyzed, to some extent, before the applications are built. Conducting requirements research can provide pointers to new adopters of big data technology, and help to establish benchmarks and standard practices.

B. Classification Procedure

The papers were classified on the basis of the classification criteria C1 and C2 mentioned in an earlier section. Each paper was analyzed and categorized on the basis of application domain and SDPLC phase. For example, if a paper discussed the implementation of big data in the healthcare industry then its application domain was classified as “Healthcare”. If a paper proposed a domain specific language or described an ontology technique, it was categorized under “Domain Specific Language/Ontology”.

Papers that did not mention any application domain explicitly were classified under “Information Technology”. Several papers provided a framework approach for their system as well as an architectural implementation. However, if the author(s) of these papers identified the contribution of the paper as a framework-oriented method then the paper was designated the category “Framework” and not “Architecture”.

V. DISCUSSION

As illustrated by the results of this review in Tables I and II, there are noticeable differences in the amount of research attention received by the different application domains and SDPLC phases. The Information Technology application domain received much more research attention when compared to other important and promising data rich domains such as Healthcare and the Banking and Financial Industry.

There is a lot of potential for making technological advances using big data in domains identified by this review such as Aviation, Infrastructure, Transport and Environmental Monitoring/Conservation. These domains and many others have not yet witnessed much work from researchers but we need to prioritize these domains. More focused research utilizing big data applications has the ability to transform the software management and development of these domains.

Research on software requirements methodologies for big data applications would help reduce the chances of system errors, project failures and unsatisfied stakeholder expectations. The technologies and environment around big data continues to evolve constantly and big data applications designed today have to deal with these unknown but inevitable changes, which makes the need for requirements research even more urgent.

Similarly, more research needs to be conducted into enhancing the existing methods and practices and developing novel methodologies for maintenance, testing, validation/verification and quality assurance of big data applications.
The high stakes and risks involved, such as system unpredictability and project failure in big data applications, can be mitigated by conducting thorough system testing, validation and verification and laying the foundations for good software quality assurance practices.

VI. CONCLUSION & FUTURE WORK

This is mostly likely to be the first comprehensive study into existing software engineering research in context of big data applications. The purpose of this literature review was to understand how software engineering research till date was enabling big data application projects and to focus on research that highlights use of the SDPLC phases in building robust, scalable big data applications.

A gap analysis was performed to identify the more popular application domains among researchers in this field. This analysis also aimed at revealing the main SDPLC phases that have seen significant research efforts and the domains and phases that need more research attention in the future.

This paper aims to provide perspective to future researchers looking into big data applications from a software engineering point of view. It can help potential researchers identify promising but underexplored application domains and focus on using specific software engineering methodologies to develop better big data applications. More research in these areas should motivate and help big data application developers and project managers to contribute more time, effort and resources in the different SDPLC phases of big data application development.

Future work encompasses widening the net to look for more papers. There may be more research published after the time of the initial search for this study was performed. By widening the search, more promising application domains can be identified that could benefit from utilizing big data applications.

With respect to the search method, the next step is to supplement the existing manual search with an automated search to make this review reproducible by other researchers. This would also help in widening the range covered by the search to get more relevant results and avoiding false positives.

REFERENCES

[21] H. Eridaputra, B. Hendradiyana, and W. D. Sunindyo, “Modeling the requirements for big data application...


[52] C. Wang, X. Li, and X. Zhou, “SODA: software defined fpga based accelerators for big data,” in *Proceedings of


