# Software Engineering for Big Data Projects:

# Domains, Methodologies and Gaps

- Vijay Dipti Kumar & Paulo Alencar

Presenter: Ivens Portugal

# Overview

- Motivation
- Research Questions
- Approach
- Sample Query
- Goal
- Results
- Observations
- Gap Analysis

# Motivation

64% organizations have invested/plan to invest in Big Data in 2013

- 30% already invested in big data
- 19% plan to invest within the next year
- 15% plan to invest within two years

Less than 8% of Gartner's 720 respondents, however, have actually deployed big data technology.

http://www.gartner.com/newsroom/id/2593815

# Motivation (contd.)

- Developing software itself is difficult and fraught with problems, even more complex for Big Data.
- Problems faced in developing big data applications would only be manifold due to the nature of the data like:
  - volume, velocity, variety
  - veracity, validity, volatility
  - value
- There has been no literature survey till date (that we could find) about software engineering for Big Data projects.

# Goal

- A literature survey to look into existing research on applying software engineering methodologies in building better Big Data applications.

- Gap analysis of which application domains and software development life cycle (SDPLC) phases need to be focused on.

# Research Questions

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 phases need more research efforts?

# Approach

An extensive literature survey was performed.

Academic search engines used:

- Scopus,
- IEEE Xplore Digital Library,
- Web of Science

Approximately, 2,000 papers were searched and 170 papers were selected.

# Sample Query

"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")

# Results – RQ1

| Application Domain | Count |
|---|---|
| Information Technology | 98 |
| Healthcare | 13 |
| Geospatial Data Processing/Geographic Information Systems | 12 |
| Infrastructure | 11 |
| Transport | 10 |
| Retail/Tourism/Commerce | 8 |
| Social Networks | 7 |
| Environmental Monitoring/Conservation | 6 |
| Manufacturing | 3 |
| Meteorology | 3 |
| Cyber Physical Systems | 3 |
| Law & Order/Criminal Investigation/Forensic Analysis | 3 |
| Agriculture | 2 |
| Banking and Financial Industry | 2 |
| Military | 2 |
| Aviation Industry | 1 |
| Astronomy | 1 |
| National Security | 1 |

# Results – RQ2

| SDPLC Phases | Count |
|---|---|
| Requirements | 16 |
| Design | 31 |
| Framework | 51 |
| Architecture | 68 |
| Testing | 10 |
| Validation/Verification | 2 |
| Maintenance | 2 |
| Quality Assurance | 6 |
| Domain Specific Languages/Ontology | 13 |

# Observations

- Majority of the papers(57%) are directly related to application design and optimization of existing technologies.

- Majority of the papers(59%) discussed or proposed system architecture and frameworks.

# Gap Analysis

Scarcity of research in data rich domains like

- Banking and Finance,
- Transport,
- Aviation,
- Meteorology

# Gap Analysis(contd.)

Dearth in research on topics like:

- validation or verification,
- maintenance,
- quality assurance,
- testing

# Conclusions

- First comprehensive study in context of Big Data.
- To provide perspective to future researchers.
- Widening the range to cover more papers after the time of this review.

# Thank You!

# References


1. A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications.
2. Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems.
3. A modular software architecture for processing of big geospatial data in the cloud.
4. On the Management and Analysis of Our LifeSteps.
5. TerraFly GeoCloud: An Online Spatial Data Analysis and Visualization System.
6. Norming to Performing: Failure Analysis and Deployment Automation of Big Data Software Developed by Highly Iterative Models.
7. Guiding the Introduction of Big Data in Organizations: A Methodology with Business- and Data-Driven Ideation and Enterprise Architecture Management-Based Implementation
8. CAP: Community Activity Prediction Based on Big Data Analysis
9. Knowle: A semantic link network based system for organizing large scale online news events
10. Investigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health conditions
11. Occupancy schedules learning process through a data mining framework
12. Design and Development of a Medical Big Data Processing System Based on Hadoop
13. The Solid architecture for real-time management of big semantic data
14. A Generalized Scalable Software Architecture for Analyzing Temporally Structured Big Data in the Cloud
15. Modeling coordinated multiple views of heterogeneous data cubes for urban visual analytics
16. The Evolvement of Big Data Systems: From the Perspective of an Information Security Application
17. Managing a Big Data project: The case of Ramco Cements Limited
18. A knowledge-based platform for Big Data analytics based on publish/subscribe services and stream processing
19. Cloud Based Big Data Analytics Framework for Face Recognition in Social Networks using Machine Learning
20. A framework for processing large scale geospatial and remote sensing data in Map Reduce environment
21. The Use of Distributed Processing and Cloud Computing in Agricultural Decision-Making Support Systems
22. Mapping the data shadows of Hurricane Sandy: Uncovering the socio-spatial dimensions of ‘big data’
23. KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications
24. Advancing big data for humanitarian needs
25. Development of an intelligent environmental knowledge system for sustainable agricultural decision support
26. A data-driven framework for archiving and exploring social media data
27. Applying data models to big data architectures

# References


28. Modeling and Supporting ETL Processes via a Pattern-Oriented, Task-Reusable Framework
29. Model oriented system design on big-data
30. iCARE: A framework for big data-based banking customer analytics
31. Trajectory Patterns Mining Towards Lifecare Provisioning
32. Intelligent operational dashboards for smarter commerce using big data
33. Context-based Ontology-driven Recommendation Strategies for Tourism in Ubiquitous Computing
34. WaaS: Wisdom as a Service
35. Improving rail network velocity: A machine learning approach to predictive maintenance
36. SemantEco: A semantically powered modular architecture for integrating distributed environmental and ecological data
37. Exploiting semantic technologies in smart environments and grids: Emerging roles and case studies
38. Extending ER models to capture database transformations to build data sets for data mining
39. DICE: Quality-Driven Development of Data-Intensive Cloud Applications
40. Service innovation and smart analytics for Industry 4.0 and big data environment
41. Intelligent services for Big Data science
42. Early Experience with Model-driven Development of MapReduce based Big Data Application
43. An Open Framework for Dynamic Big-Data-Driven Application Systems (DBDDAS) Development
44. Moving code – Sharing geoprocessing logic on the Web
45. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud
46. Smart Traffic Cloud: An Infrastructure for Traffic Applications
47. A Model-Driven Prototype Evaluation to Elicit Requirements for a Sensemaking Support Tool
48. Towards Model-Driven Engineering for Big Data Analytics – An Exploratory Analysis of Domain-Specific Languages for Machine Learning
49. ELTA: New Approach in Designing Business Intelligence Solutions in Era of Big Data
50. Railway assets: A potential domain for big data analytics
51. A scalable framework for spatiotemporal analysis of location-based social media data
52. Perspectives of Emerging Museum Professionals on the Role of Big Data in Museums
53. A Framework to Model Big Data Driven Complex Cyber Physical Control Systems
54. Toward the digital water age: Survey and case studies of Australian water utility smart-metering programs
55. Assembling Cloud-Based Geographic Information Systems: A Pragmatic Approach Using Off-the-Shelf Components

# References


56. Web based visualization of large climate data sets
57. Modeling The Requirements for Big Data Application Using Goal Oriented Approach
58. Breeze graph grammar: a graph grammar approach for modeling the software architecture of big data-oriented software systems
59. Architecture Dedicated to Data Integration
60. A domain model of Web recommender systems based on usage mining and collaborative filtering
61. CloudExp: A comprehensive cloud computing experimental framework
62. Engineering Privacy for Big Data Apps with the Unified Modeling Language
63. Big Data System Development: An Embedded Case Study with a Global Outsourcing Firm
64. Testing Big Data (Assuring the Quality of Large Databases)
65. An Architecture to Support the Collection of Big Data in the Internet of Things
66. Embrace the Challenges: Software Engineering in a Big Data World
67. Distribution, Data, Deployment: Software Architecture Convergence in Big Data Systems
68. Towards Service-oriented Enterprise Architectures for Big Data Applications in the Cloud
69. Towards an Architecture for Managing Big Semantic Data in Real-Time
70. Sustainability Data and Analytics in Cloud-Based M2M Systems
71. Fog Computing: A Platform for Internet of Things and Analytics
72. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications
73. Building Open Environments to Meet Big Data Challenges in Earth Sciences
74. Software Development Support for Shared Sensing Infrastructures: A Generative and Dynamic Approach
75. Towards Mega-Modeling: A Walk through Data Analysis Experiences
76. An Integrated System for Regional Environmental Monitoring and Management Based on Internet of Things
77. Leveraging Big Data for the Development of Transport Sustainability Indicators
78. Using the Web to Monitor a Customized Unified Financial Portfolio
79. Big-data for building energy performance: Lessons from assembling a very large national database of building energy use
80. A Proposed Case for the Cloud Software Engineering in Security
81. A Scalable Big Data Test Framework

# References


112. A Model Architecture for Big Data applications using Relational Databases
113. Cloud Computing for Extracting Price Knowledge from Big Data
114. A Big Data Modeling Methodology for Apache Cassandra
115. A Domain-Driven, Generative Data Model for BigPetStore
116. Big Data Search for Environmental Telemetry
117. Advanced Control Distributed Processing Architecture (ACDPA) using SDN and Hadoop for Identifying the Flow Characteristics and Setting the Quality of Service(QoS) in the Network
118. UniMiner: Towards a Unified Framework for Data Mining
119. 5Ws Model for Big Data Analysis and Visualization
120. Towards a Big Data Exploration Framework for Astronomical Archives
121. A Holistic Architecture for the Internet of Things, Sensing Services and Big Data
122. Enabling proactive data management in virtualized Hadoop clusters based on predicted data activity patterns
123. Ontology-Based Workflow Generation for Intelligent Big Data Analytics
124. A reference architecture for big data systems in the national security domain
125. Exploring a framework for identity and attribute linking across heterogeneous data systems
126. Model-Driven Observability for Big Data Storage
127. Streaming software analytics
128. Towards a Domain Specific Language for Geospatial Data Visualization Maps with Big Data Sets
129. Towards a model-driven design tool for big data architectures
130. Understanding Quality Requirements in the Context of Big Data Systems
131. Analysis of Requirements for Big Data Adoption to Maximize IT Business Value
132. Strategic Prototyping for Developing Big Data Systems
133. A Deep Intelligence Framework for Online Video Processing
134. A Big Data Approach for Proactive Healthcare Monitoring of Chronic Patients
135. Continuous Architecting of Stream-Based Systems
136. Adaptable Enterprise Architectures for Software Evolution of Smartlife Ecosystems
137. Electronic Health Record Error Prevention Approach using Ontology in Big Data
138. Facilitating Twitter Data Analytics- Platform, Language and Functionality
139. Sanitizing And Minimizing Databases For Software Application Test Outsourcing

# References


137. A Cooperative Sensing and Mining System for Transportation Activity Survey
138. A Framework for Ensuring the Quality of a Big Data Service
139. Research and Application of One-Key Publishing Technologies for Meteorological Service Products
140. Nonparametric Discovery of Contexts and Preferences in Smart Home Environments
141. State of Art in Testing For Big Data
145. A Cloud Based Architecture for Distributed Real Time Processing of Continuous Queries
146. Building Pipelines for Heterogeneous Execution Environments for Big Data Processing
147. Big Data Transformation Testing based on Data Reverse Engineering
148. Linked Enterprise Data Model and its use in Real Time Analytics and Context-Driven Data Discovery
149. Research of Performance Test Technology for Big Data Applications
150. An Empirical Study on Quality Issues of Production Big Data Platform
151. A Framework for Composition and Enforcement of Privacy-Aware and Context-Driven Authorization Mechanism for Multimedia Big Data
152. Toward Semantically Enabled Development of Service-Oriented Architectures for Integration of Socio-Medical Data
153. A Smart Polyglot Solution for Big Data in Healthcare
154. EPIC- OSM - A Software Framework for OpenStreetMap Data Analytics
155. Improving Power Grid Monitoring Data Quality - An Efficient Machine Learning Framework for Missing Data Prediction
156. The EMBERS Architecture for Streaming Predictive Analytics
157. Evaluating the Quality of Social Media Data in Big Data Architecture
158. A Pervasive Framework for Real-Time Activity Patterns of Mobile Users
159. An Automatic Discovery Framework of Cross-source Data Inconsistency for Web Big Data
160. A Reference Architecture for Social Media Intelligence Applications in the Cloud
161. Towards a Data Processing Architecture for the Weather Radar of the INTA Anguil
162. BDCaM - Big Data for Context-aware Monitoring - A Personalized Knowledge Discovery Framework for Assisted Healthcare
163. An Architecture to Process Massive Vehicular Traffic Data
164. SMASH - A Cloud-based Architecture for Big Data Processing and Visualization of Traffic Data
165. Multi-Disciplinary Ontological Geo-Analytical Incident Modeling
166. A big data processing framework for uncertainties in transportation data
167. Building a Big Data Platform for Smart Cities - Experience and Lessons from Santander
168. Agile Big Data Analytics for Web-based Systems - An Architecture-centric Approach
169. A Cloud Service Architecture for Analyzing Big Monitoring Data
170. A Preliminary Survey on Domain-Specific Languages for Machine Learning in Big Data