Software Engineering for Big Data
- A Literature Survey

- Vijay Dipti Kumar
Overview

• Motivation

• Research Questions

• Approach

• Goal

• Progress

• Gap Analysis

• Observation
Motivation

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

• 30% have already invested in big data technology

• 19% plan to invest within the next year, and

• another 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.

• 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.
Research Questions

What is the state of the art in Software Engineering specific to Big Data applications?

RQ1. What software engineering research areas have been used to enable big data applications?

RQ2. What are the gaps in software engineering research on big data applications done till date?
Approach

An extensive literature survey:

- 83 papers listed on the course website
- 40 papers from my own search
Goal

- A literature survey using which research till date is classified on the basis of the following:
  - Software engineering subfields
  - Application domains
  - Types of data used
  - Big data technologies used

- Gap analysis of which areas need to be focused on
## Progress

<table>
<thead>
<tr>
<th>Application Domain</th>
<th># of Research Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Technology</td>
<td>72</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>9</td>
</tr>
<tr>
<td>Geospatial Data Processing / Geographic Information Systems/ Astronomy</td>
<td>9</td>
</tr>
<tr>
<td>Retail, Tourism and Commerce</td>
<td>7</td>
</tr>
<tr>
<td>Healthcare</td>
<td>7</td>
</tr>
<tr>
<td>Transport</td>
<td>5</td>
</tr>
<tr>
<td>Environmental monitoring/Protection/Conservation</td>
<td>5</td>
</tr>
<tr>
<td>Social Networks</td>
<td>4</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>3</td>
</tr>
<tr>
<td>Cyber Physical Systems</td>
<td>3</td>
</tr>
<tr>
<td>Agriculture</td>
<td>2</td>
</tr>
<tr>
<td>Banking and Financial sector</td>
<td>2</td>
</tr>
<tr>
<td>Military/Aviation Industry</td>
<td>2</td>
</tr>
<tr>
<td>Law &amp; Order/ Criminal Investigation/ Forensic Analysis</td>
<td>1</td>
</tr>
</tbody>
</table>
# Progress (contd.)

<table>
<thead>
<tr>
<th>Software Engineering Methodology</th>
<th># of Research Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements</td>
<td>13</td>
</tr>
<tr>
<td>Design</td>
<td>23</td>
</tr>
<tr>
<td>Frameworks</td>
<td>38</td>
</tr>
<tr>
<td>Architecture</td>
<td>45</td>
</tr>
<tr>
<td>Maintenance</td>
<td>2</td>
</tr>
<tr>
<td>Testing</td>
<td>6</td>
</tr>
<tr>
<td>Domain Specific Languages/Ontology</td>
<td>8</td>
</tr>
<tr>
<td>Quality Assurance</td>
<td>2</td>
</tr>
<tr>
<td>Verification/Validation</td>
<td>1</td>
</tr>
</tbody>
</table>
Gap Analysis

- Majority of the papers are directly related to application design and optimization of existing technologies

- Majority of the papers were about system architecture and frameworks
Observation

• Not enough papers focusing on requirements gathering, testing, quality assurance, validation or verification and maintenance of big data applications.

• Not enough papers focusing on data rich domains like health care, banking and financial sectors or transportation sector.
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.
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
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
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
References

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
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
References

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
82. AllJoyn Lambda: an Architecture for the Management of Smart Environments in IoT
83. An Adaptable Framework to Deploy Complex Applications onto Multi-cloud Platforms
84. Big data Analysis Solutions using MapReduce Framework
85. Bugarium: 3D Interaction for Supporting Large-Scale Bug Repositories Analysis
86. Design Principles for Effective Knowledge Discovery from Big Data
87. DIVE: A Graph-Based Visual- Analytics Framework for Big Data
88. DOT: A Matrix Model for Analyzing, Optimizing and Deploying Software for Big Data Analytics in Distributed Systems
89. Formal Verification Problems in a Big Data World: Towards a Mighty Synergy
90. Getting an Intuition for Big Data
91. Implementing Conceptual Search Capability in a Cloud-Based Feed Aggregator
92. Metamorphic Runtime Checking of Applications without Test Oracles
93. Multilayer Big Data Architecture for Remote Sensing in Eolic Parks
94. Store, Schedule and Switch – A New Data Delivery Model in the Big Data Era
<table>
<thead>
<tr>
<th>Reference Number</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>95.</td>
<td>Research on Warship Communication Operation and Maintenance Management based on Big Data</td>
</tr>
<tr>
<td>96.</td>
<td>A Distributed Randomization Framework for Privacy Preservation in Big Data</td>
</tr>
<tr>
<td>97.</td>
<td>Research of Performance Test Technology for Big Data Applications</td>
</tr>
<tr>
<td>98.</td>
<td>SODA: Software Defined FPGA based Accelerators for Big Data</td>
</tr>
<tr>
<td>99.</td>
<td>The Application of Semantic-based Classification on Big Data</td>
</tr>
<tr>
<td>100.</td>
<td>A Layer Based Architecture for Provenance in Big Data</td>
</tr>
<tr>
<td>101.</td>
<td>Big Picture of Big Data Software Engineering</td>
</tr>
<tr>
<td>102.</td>
<td>Designing Big Data Driven Cyber Physical Systems Based on AADL</td>
</tr>
<tr>
<td>103.</td>
<td>Low Power and Scalable Many-Core Architecture for Big-Data Stream Computing</td>
</tr>
<tr>
<td>104.</td>
<td>I/O Performance Modeling for Big Data Applications over Cloud Infrastructures</td>
</tr>
<tr>
<td>105.</td>
<td>Breaking the Boundary for Whole-System Performance Optimization of Big Data</td>
</tr>
<tr>
<td>106.</td>
<td>Building a Big Data Analytics Service Framework for Mobile Advertising and Marketing</td>
</tr>
<tr>
<td>107.</td>
<td>Enrichment Patterns for Big Data</td>
</tr>
<tr>
<td>108.</td>
<td>Architecture Knowledge for Evaluating Scalable Databases</td>
</tr>
<tr>
<td>109.</td>
<td>Big Data-as-a-Service: Definition and architecture</td>
</tr>
<tr>
<td>110.</td>
<td>GERBIL: MPI+YARN</td>
</tr>
<tr>
<td>111.</td>
<td>A Cognitive Oriented Framework for IoT Big-data Management Prospective</td>
</tr>
<tr>
<td>112.</td>
<td>A Model Architecture for Big Data applications using Relational Databases</td>
</tr>
<tr>
<td>113.</td>
<td>Cloud Computing for Extracting Price Knowledge from Big Data</td>
</tr>
<tr>
<td>114.</td>
<td>A Big Data Modeling Methodology for Apache Cassandra</td>
</tr>
<tr>
<td>115.</td>
<td>A Domain-Driven, Generative Data Model for BigPetStore</td>
</tr>
<tr>
<td>116.</td>
<td>Big Data Search for Environmental Telemetry</td>
</tr>
<tr>
<td>117.</td>
<td>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</td>
</tr>
<tr>
<td>118.</td>
<td>UniMiner: Towards a Unified Framework for Data Mining</td>
</tr>
<tr>
<td>119.</td>
<td>5Ws Model for Big Data Analysis and Visualization</td>
</tr>
<tr>
<td>120.</td>
<td>Towards a Big Data Exploration Framework for Astronomical Archives</td>
</tr>
<tr>
<td>121.</td>
<td>A Holistic Architecture for the Internet of Things, Sensing Services and Big Data</td>
</tr>
<tr>
<td>122.</td>
<td>Enabling proactive data management in virtualized Hadoop clusters based on predicted data activity patterns</td>
</tr>
<tr>
<td>123.</td>
<td>Ontology-Based Workflow Generation for Intelligent Big Data Analytics</td>
</tr>
</tbody>
</table>
Thank You!