Big Picture of Big Data Software Engineering
With example research challenges

Nazim H. Madhavji, UWO, Canada
Andriy Miranskyy, Ryerson U., Canada
Kostas Kontogiannis, NTUA, Greece

madhavji@gmail.com
avm@ryerson.ca
kkontog@softlab.ece.ntua.gr
Introduction

• The Big Data technology and services market:
  – compound annual growth rate of 27%
    • $9.8 billion in 2012 to projected $32.4 billion in 2017

  – anticipated to generate data exponentially.

Big Data is on the rise ....
Observation -- 80:20 Bias

• Technical and business communities are focusing:
  – **More** on “data analytics” and Big Data infrastructures (76%)
  – **Less** on new “application” areas (22%) involving Big Data
    (the so-called “**Big Data Software Engineering (BDSE)**”).
Observation -- 80:20 Bias

• No inherent reason for this bias.
• Perhaps, more a function of:
  – availability of data, infrastructure, mining, and analysis tools
  – Relatively quick payback
If we know our History ....

• A predominant model of software development in the ‘60s was “Code & Fix”.
  – There was the so-called, “software crisis”.
  – We have learnt over the years:
    • Waterfall, iterative, evolutionary, CBSE, SoS, agile, etc., models of development.
Question

• Are we repeating this mentality all over again?
  – The current paradigm is that of “Mine and Analyse”.
  – Confusion abounds in industry.
  – Are we missing a Big Picture?
Big Data Engineer

• We are looking for someone who has A LOT of experience (10k hours) in coding and building software. Experience with Big Data stacks like Elasticsearch, HBase, Hadoop and Cassandra (or large databases in general).

• [Link to Indeed job posting](http://www.indeed.com/viewjob?jk=6ca00b23587bc40c&from=api&q=title%3A+%22big+data%22&atk=19cplk2050ni3ag&sclick=1&sjdu=bHIslJc3OwpxOxYFLbYJAGGGh4l1GGQPT4rYranXcPF8ktqmPQB5SbCesna_YdpNYpg8Rvg9Sy6uCl4M2jH2jH3MJzfB7c2xnydF8CDweHp7JlFvCOgv4MBPS67pvzRlz3RyntWiRqJnLMjSMGbsuhmQhA7rEqHOJvA4jnRnR6V1ZVuvjpFoqGuyGqEaaOyedbzo0p5zcdJEeb0NleR9uiLF7iNn0vCkkuXsX5rXbKBA-ur7uS888ICqqnz14s2c8ru4nevKZREByoIDCWDODYlo-tLF4Xmbi8afr3kojgq8DtYdE9jceLhUBqXoVqvllX9J6i7GBIYZzZpDaTGCKg8WwzM6e5ZsrXGi09xo&pub=4009563241607923)
Big Data Engineer

• Research, design, and implementation of new technologies which will provide a solid ground for Data Processing and Machine Learning tools.

  • http://www.indeed.com/viewjob?jk=c1eeb42674926be3&from=api&q=title%3A+%22big+data%22&atk=19cpkk2050nki3ag&sclk=1&sjdu=bHIslJc3OwpxOxYFLbYJAGGGhl1GGQPT4rYranXcPF_baxBgYZGNbBqJ5rol-SE_CBCJG4E6DeEPvps-KYZ1-NKQCOam3hEBL-kPhGgqDVvlCNrhv4-ZZPb3eJY1alR6DxaZYW71PQZS1jJf5zPs7WircXR0Z-EUfLyGZpWOL20pEz8c40C8i6w3z1_TOWIPRdfa0wob-2aBDIf3ABy9d12YYZcHI9qM_bV6CkLuZ0m0yTm3bKcWrWxLHwYzeb_KYVKUGy2RZX7qrzbfO8LLzL1_JJsH2kAEc4NZgztkCCtdNgEPtjHRj-z9oKg5y_SH6C7lSpNjIxdaTSymFk3zNfA&pub=4009563241607923

• Advert 1/2 and 2/2 are both “Big Data Engineer” yet have completely different tasks.
Big Data Architect

• Responsibilities -- Collaborate with senior management at the largest financial institutions in the world, and the Clutch senior stakeholders advising them, to translate business challenges into technical requirements

• http://www.indeed.com/viewjob?jk=d186ffc9abee2ed0&from=api&q=title%3A+%22big+data%22&atk=19cpkk2050nki3ag&sclk=1&sjdu=bHIslJc3OwpxOxYFLbYJAGGGhI1GGQPT4rYranXcPF9nU-axYmrmEG0GRgHtiyduBuYsH8OgnPcc3TEk53cUJUvl99Pzwdt7Fjj18XVa9LDqtoc37jflOt9_fZU2i-
Big Data Architect

• Position Purpose -- Engineer, deploy, and support Hadoop clusters in production and lower environments. Provide hands-on administration of running Hadoop jobs and optimizing daily tasks.

  • http://www.indeed.com/viewjob?jk=ddcd956b9e8ae114&from=api&q=title%3A+%22big+data%22&atk=19cpmp0ee0nr17te&sclk=1&sjdu=bHIsIjc3OwpxOxYFLbYJAGGGhl1GGQPT4rYranXcPF9nU-axYmrmEG0GRgHtiydj_6r7WPyR-HwKeEmfutq6ECoA-XVrXAQ7ijoynCOxhBWBxtw6gbG6mt6WJd7LSR5jDOnafN0Qs4-8Fmx6lef8GChkZVVSQ50u9NchkJTvKA-EX3VfjtvUu2T-wLZp4t74-3QSh_CyoFarQJzuBSr_vBZnkAjTnU4FcWg21WX0QHTx0aHowDb4btV_qpBfEtJDGZLTUVMp9jstwRePDwAz_1ulXv77kvOMpHMJAgkLy10VNUbRFgdQvGnY4iPLVbFHeOd7_Q&pub=4009563241607923

• Advert 1/2 and 2/2 are both “Big Data Architect” yet have completely different tasks.
We envisage a Big Picture of Big Data Software Engineering

It is anticipated to help in at least two ways:

(a) in organisational structuring
   — agent roles, processes, relationships, etc., among the interacting elements

(b) in precipitating research in targeted areas of the Big Data environment.
The evolution of understanding and structure of theories, models, procedures, and laws of application and system domains.

Requirements analysis leads to computational procedures and algorithms.

Corporate management and marketeers are supported by user support.

Project and process managers are affected by exogenous change.

Big Data systems & services, and analytics (e.g., Financial sector) are used in business and client scenarios.

22 May, 2015

BIGDSE 2015, Florence, Italy

ACTION ON BIG DATA
(Practice)

STRATEGIC THINKING
(Practice)

STRATEGY IMPLEMENTATION
(Practice)

RESEARCH
(Theory)
Corporate Decision-making

Business and Client Scenarios

Big Data SE practice

CS research on Big Data

SE research on Big Data Environments

Streamed/Historical data, Technologies, Infrastructures, etc.

Used in

Big Data systems & services, and Analytics (e.g., Financial sector)

Deployed applications and services

(Theory)

Usage
Illustrative Research Challenges

• Requirements Engineering (RE)
• Architectures
• Testing and Maintenance
Requirements Engineering (RE)

- **RE process involves**, amongst others:
  - modelling the application domain,
  - eliciting scenarios and requirements from stakeholders and from other sources,
  - developing functional and behavioural models,
  - analysis, prioritisation, and validation of requirements.

- Real-world characteristics of the elements (e.g., dimensions, weight, cost, transfer rates, etc.) need to be represented in requirements notation along with operations on them.
Requirements Engineering (RE)

- Thus, when eliciting scenarios of Big Data system responses, obviously, we need to capture properties (such as \textit{volume}, \textit{velocity}, \textit{variety}, \textit{varacity}, etc.) in RE models.

- Currently, RE for Big Data applications is an emerging area.
  - Need to separate requirements for:
    - infrastructures,
    - analytic tools, and
    - end-user or business applications.
Requirements Engineering (RE)

• Example business scenario:
  – customer arrives in a department store,
  – possible previous experiences in the store,
  – real-time processing,
  – personalised and time-sensitive offers to this customer,
  – at specific points of displays where the customer would likely be shopping.

• How should a requirements analyst exploit video footage, images, historical data, and info.?
  – Needs to address all the “Vs” of Big Data.
Example RE challenges

1. Envisaging innovative application scenarios in the context of Big Data.
   - Focus is on the application domain, stakeholder-value, and creative ways to utilise the “Vs” of Big Data including combinations thereof.
Example RE challenges

2. Creating underlying technologies, libraries and frameworks that provide primitive operations on Big Data characteristics.
   — Analogous to creating mathematical theories.
   — Example:
     — Fast comparisons of videos
     — Encryption of data of various types (e.g., videos, images, voice, etc.) for secure transmission of information.
     — Etc.
Example RE challenges

3. **Envisaging system solutions (at RE-time)** (taking into account primitive operations):
   – In order to **specify, analyse, validate, prioritise** requirements for Big Data applications.
Illustrative Research Challenges

• Requirements Engineering (RE)
• Architectures
• Testing and Maintenance
Reference Architectures for Big Data Applications

• Limited work on reference architectures and patterns for Big Data analytics applications. There is also a need to investigate:

  – How these reference architectures can be mapped onto existing technologies, frameworks, and tools to yield relevant and optimal application deployment.

  – What is, and how to assess, the impact of different architectural design decisions on functional and non-functional requirements in Big Data analytics applications

• Note this “middle-out” development process; not always top-down.
Dynamic Resource Deployment and Management

• Big data analytics applications encompass the use of multiple data sources and processes implementing complex algorithms.

• **Use of virtualisation technology, and data distribution.** These technologies pose new issues related to dynamic resource allocation and resource management. Some interesting points to investigate include:

  – Novel techniques and frameworks that allow for moving computation logic closer to where different data sources reside, rather than moving massive data to centralised data centers

  – Novel techniques to assess and verify whether key policies (e.g. privacy, security, accuracy) **hold in a given deployment and ensure compliance** within an organisation’s governance framework
Illustrative Research Challenges

- Requirements Engineering (RE)
- Architectures
- **Testing and Maintenance**
  - How to scale down the resources to produce realistic BDS test system?
  - How to scale/synthesize data representatively?
  - How to capture workload on a production BDS and replay it on a test BDS?
Summary

• Current Big Data paradigm is biased in favour of “data analytics”.

• We propose a “Big Picture” of Big Data that extends the paradigm into “application engineering”.

• We give illustrative research challenges in the areas of:
  – Requirements engineering
  – System architectures
  – Creating representative test system for BDS