DICE: Developing Data-Intensive Cloud Applications with Iterative Quality Enhancements

Giuliano Casale

*Imperial College London*

*Project Coordinator*
DICE Project

- Horizon 2020 Research & Innovation Action
  - Quality-Aware Development for Big Data applications
  - Feb 2015 - Jan 2018, 4M Euros budget
  - 9 partners (Academia & SMEs), 7 EU countries

©DICE 3/27/2015
Motivation

- **Software market rapidly shifting to Big Data**
  - 32% compound annual growth rate in EU through 2016
  - 35% Big data projects are successful [CapGemini 2015]

- **European call for software quality assurance (QA)**
  - ISTAG: call to define environments “for understanding the consequences of different implementation alternatives (e.g. quality, robustness, performance, maintenance, evolvability,...)”

- **QA evolving too slowly compared to the trends in software development (Big data, Cloud, DevOps ...)**
  - Still crucial for competitiveness!
Quality-Aware MDE Today

Platform-Indep. Model

Architecture Model

Platform-Specific Model

Domain Models

QA Models

Analytical Models

Cost-Quality Models

Platform Description

Code generation

C++
Java
C#
Challenge 1: QA for Big Data

- 5Vs:
  - Volume,
  - Velocity,
  - Variety,
  - Veracity,
  - Value

- Problem: today no QA toolchain can reason on the quality of complex Big Data applications

- Heterogeneous Big Data Technologies
  - NoSQL, Spark, Hadoop/MapReduce, Storm, CEP, ...

- Cloud infrastructure adds complexity
  - Cloud storage, auto-scaling, private/public/hybrid, ...
Challenge 2: Embracing DevOps

- QA must become lean as well
  - Continuous quality checks and model versioning
- Modelling of the operations
  - Dev needs awareness of infrastructure and costs
- Continuous feedback
  - Forward and backward model synchronisation
  - Tracking of self-adaptation events (e.g. auto-scaling)
- Big data coming from continuous monitoring
  - QA has its own Big data, use machine learning?
An Holistic Approach: DICE

DICE IDE

Platform Description

Deployment & Continuous Integration

Big Data

Platform-Indep. Model

Architecture Model

Platform-Specific Model

Domain Models

Data Awareness

Continuous Validation

QA Models

Continuous Monitoring
Benefits

- Tackling skill shortage and steep learning curves
  - Data-aware methods, models, and OSS tools
- Shorter time to market for Big Data applications
  - Cost reduction, without sacrificing product quality
- Decrease development and testing costs
  - Select optimal architectures that can meet SLAs
- Reduce number and severity of quality incidents
  - Iterative refinement of application design
DICE QA: Quality Dimensions

- Reliability
  - Availability
  - Fault-tolerance

- Efficiency
  - Performance
  - Time behaviour
  - Costs

- Safety & Privacy
  - Risk of harm
  - Privacy & data protection
DICE Platform Independent Model (DPIM)
DICE Profile: PIM Level

- Functional approach to data to be expanded
  - Data dependencies
    - graph relationships between data, archives and streams
- QA focuses on quantitative aspects of data
  - Static characteristics of data
    - volumes, value, storage location, replication pattern, consistency policies, data access costs, known schedules of data transfers, data access control / privacy, ...
  - Dynamic characteristics of data
    - cache hit/miss probabilities, read/write/update rates, burstiness, ...
DICE Platform and Technology Specific Model (DTSM)
DICE Profile: PSM Level

- Need for technology-specific abstractions
  - Hadoop: Number of mappers and reducers, ...
  - In-memory DBs: Peak memory and variable threading
  - Streaming: merge/split/operators, networking, ...
  - Storage: Supported operations, cost/byte, ...
  - NoSQL: Consistency policies, ...

- Generation of deployment plan
  - Proposed Chef + TOSCA extension

- Interest is both on private and public clouds
  - Private clouds more relevant for batch processing
  - Public clouds more relevant for streaming
## Demonstrators

<table>
<thead>
<tr>
<th>Case study</th>
<th>Domain</th>
<th>Features &amp; Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributed data-intensive media system (ATC)</td>
<td>• News &amp; Media</td>
<td>• Large-scale software</td>
</tr>
<tr>
<td></td>
<td>• Social media</td>
<td>• Data velocities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data volumes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data granularity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Multiple data sources and channels</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Privacy</td>
</tr>
<tr>
<td>Big Data for e-Government (Netfective)</td>
<td>• E-Gov application</td>
<td>• Data volumes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Legacy data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data consolidation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data stores</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Privacy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Forecasting and data analysis</td>
</tr>
<tr>
<td>Geo-fencing (Prodevelop)</td>
<td>• Maritime sector</td>
<td>• Vessels movements</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Safety requirements</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Streaming &amp; CEP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Geographical information</td>
</tr>
</tbody>
</table>
Thanks!

www.dice-h2020.eu