Bug bites Elephant?

Test-driven Quality Assurance in Big Data Application Development

Dr. Dominik Benz, inovex GmbH
2013/06/03, Berlin Buzzwords
Who speaks... ... the Elephant language?

Class A extends Mapper...

ROI, $$, ...

Write/execute tests, specify acceptance criteria, ...

apt-get install...

TDD!
The road... ... to Big Data QA

our Big Data QA problem

the FitNesse approach

test data definition / selection

job & workflow control

result inspection
Web Intelligence @ 1&1

~ 1 billion log events / day, ~ 1 TB (thrift) logfiles

BI reporting, web analytics, …

chains of MR jobs, running on 20 nodes / 8 cores / 96 GB RAM (CDH)
An exemplary workflow

- QA problem
- Log Files (thrift)
- MR job 1
- Intermediate result (avro)
- MR job 2 ...
- DWH (RDBMS)
- create (sample) input data
- inspect (binary) formats
- control workflows
<table>
<thead>
<tr>
<th>method</th>
<th>tests what?</th>
<th>issues for our usecase</th>
</tr>
</thead>
<tbody>
<tr>
<td>JUnit</td>
<td>isolated functions</td>
<td>no integration, Java syntax</td>
</tr>
<tr>
<td>MRUnit</td>
<td>1 mapper + 1 reducer</td>
<td>„little“ integration, Java syntax</td>
</tr>
<tr>
<td>iTes</td>
<td>hadoop jobs/workflows</td>
<td>Java / Groovy syntax</td>
</tr>
<tr>
<td>Scripts/CLI</td>
<td>(manual) scripting/inspect.</td>
<td>„script chaos“, syntax</td>
</tr>
</tbody>
</table>

→ **FitNesse** as suitable addition / solution!
The road... to Big Data QA

Big Data QA is different!

the FitNesse approach

test data definition / selection

job & workflow control

result inspection
FitNesse

In a nutshell

“fully integrated standalone wiki and acceptance testing framework”

• „executable“ Wiki-Pages (returning test results)
• (almost) natural language test specification
• connection to SUT via (Java-)“Fixtures“
FitNesse Architecture Overview

Browser

| script | check | num results | 3 |

FitNesse Server

Fixtures

public int numResults

{ ... }

→ „calling java methods from wiki“, compare return values
→ Integrates with REST, Jenkins...
FitNesse
An Exemplary Test

FrontPage
HadoopLogFileProcessingTest

Assertions: 3 right, 0 wrong, 0 ignored, 0 exceptions (0.054 seconds)

<table>
<thead>
<tr>
<th>script</th>
<th>Hadoop</th>
</tr>
</thead>
<tbody>
<tr>
<td>upload</td>
<td>/home/inovex/viewLog.csv to hdfs /testdata</td>
</tr>
<tr>
<td>hadoop job from jar:</td>
<td>ViewLogCounter.jar /testdata/.*.csv output</td>
</tr>
<tr>
<td>show</td>
<td>job output</td>
</tr>
<tr>
<td>check</td>
<td>number of output files 3</td>
</tr>
</tbody>
</table>

Starting MR job [...]
<table>
<thead>
<tr>
<th>path</th>
<th>/home/inovex/lib/*.jar</th>
</tr>
</thead>
<tbody>
<tr>
<td>script</td>
<td>Hadoop</td>
</tr>
<tr>
<td>upload</td>
<td>viewLog.csv</td>
</tr>
<tr>
<td>hadoop job from jar</td>
<td>viewLog.jar</td>
</tr>
<tr>
<td>show</td>
<td>job output</td>
</tr>
<tr>
<td>check</td>
<td>number of output files</td>
</tr>
</tbody>
</table>
public class Hadoop {

    public boolean uploadToHdfs(String localFile, String remoteFile) {...}

    public boolean hadoopJobFromJar(String jar, String input, String output) {...}

    public String jobOutput() {...}

    public String numberOfOutputFiles() {...}
}

The road... ... to Big Data QA

Big Data QA is different!

Fitnesse Wiki test execution!

test data definition / selection

job & workflow control

result inspection
## Table: Log File

/home/inovex/custom_logs/input/viewLog.csv

<table>
<thead>
<tr>
<th>date</th>
<th>user</th>
<th>product</th>
<th>browser</th>
<th>os</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-03-12</td>
<td>john</td>
<td>1</td>
<td>ff</td>
<td>win</td>
</tr>
<tr>
<td>2013-03-12</td>
<td>john</td>
<td>2</td>
<td>ff</td>
<td>win</td>
</tr>
<tr>
<td>2013-03-13</td>
<td>john</td>
<td>2</td>
<td>ff</td>
<td>win</td>
</tr>
<tr>
<td>2013-03-14</td>
<td>lisa</td>
<td>1</td>
<td>ie</td>
<td>win</td>
</tr>
<tr>
<td>2013-03-14</td>
<td>peter</td>
<td>1</td>
<td>ff</td>
<td>lin</td>
</tr>
<tr>
<td>2013-03-15</td>
<td>lisa</td>
<td>2</td>
<td>ie</td>
<td>lin</td>
</tr>
<tr>
<td>2013-03-15</td>
<td>peter</td>
<td>2</td>
<td>ff</td>
<td>mac</td>
</tr>
<tr>
<td>2013-03-16</td>
<td>lisa</td>
<td>1</td>
<td>ff</td>
<td>mac</td>
</tr>
</tbody>
</table>
- Big Data: **Efficient** data transfer among **heterogeneous sources**
- Define Interface via **IDL**, Compiler for many languages

<table>
<thead>
<tr>
<th>Table: Thrift Log File</th>
</tr>
</thead>
<tbody>
<tr>
<td>/home/inovex/viewLog.thrift</td>
</tr>
<tr>
<td>date</td>
</tr>
<tr>
<td>2013-03-12</td>
</tr>
<tr>
<td>2013-03-12</td>
</tr>
<tr>
<td>2013-03-13</td>
</tr>
<tr>
<td>2013-03-14</td>
</tr>
</tbody>
</table>
Dev/Test Hadoop Cluster: **Identical Hardware** like Prod, but fewer nodes

(random/biased) **sampling** e.g. on daily basis

**Feedback loop:**
- identify „**special cases“** from real data
- include them in (manual) data definition
- Gradually **increase test coverage** / artefact quality
The road... ... to Big Data QA

Big Data QA is different!

FitNesse Wiki test execution!

Define CSV / thrift / real-world test data!

result inspection

job & workflow control
- Execute arbitrary (shell) commands
- Mainly a **wrapper** around `apache.commons.exec.CommandLine`

<table>
<thead>
<tr>
<th>script</th>
<th>Shell</th>
</tr>
</thead>
<tbody>
<tr>
<td>show</td>
<td>exec</td>
</tr>
<tr>
<td></td>
<td>hadoop fs -put /home/inovex/custom_logs/input/viewLog.csv /user/inovex/input/</td>
</tr>
<tr>
<td>show</td>
<td>exec</td>
</tr>
<tr>
<td></td>
<td>hadoop jar /home/inovex/showcase/ViewLogCounter.jar /user/inovex/input/* .csv /user/inovex/output</td>
</tr>
</tbody>
</table>
Hide complexity from test authors

„define“ appropriate test language via (Java) method names

re-use other fixtures (Shell, ...) internally
FitNesse allows to group tests into **suites**

- Can be used to simulate MR **processing chains**
  - **SetupSuite** / **TearDownSuite** for creating / destroying test conditions

- Tests can still be executed **individually**
The road... to Big Data QA

- Big Data QA is different!
- FitNesse Wiki test execution!
- Define CSV / thrift / real-world data!
- Use suites & fixtures for jobs/workflows!
- Result inspection
Validate **RDBMS contents** (via JDBC)

- E.g. for checking the **final** result

- Or use **Hive** + Hive-Server to query raw data
Execute arbitrary **pig commands** from Wiki page

- Inspect e.g. **binary intermediate results** (avro, ...)

<table>
<thead>
<tr>
<th>script</th>
<th>Pig Console</th>
<th>MAPREDUCE</th>
<th>fixtures.jar</th>
<th>avro.output.codec=snappy</th>
</tr>
</thead>
<tbody>
<tr>
<td>load avro file</td>
<td>/data/viewlog.avro</td>
<td>using alias</td>
<td>viewlog</td>
<td></td>
</tr>
<tr>
<td>execute</td>
<td>flights = foreach viewlog generate key.campaign as flight:int</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>execute</td>
<td>filtered = filter flights by flight = 5283</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>check</td>
<td>number of records from alias</td>
<td>35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
public class PigConsole extends PigServer {

    public void loadAvroFileUsingAlias(String filename, String alias) {
        this.registerQuery(
            alias + "= LOAD" + filename + "USING" + AVRO_STORAGE_LOADER + ";";");
    }
}

Results

Server Infrastructure

- **Fitnesse Master**
  - **TestEnvironments**
    - ProjA
    - ProjB
  - **TestConfigurations**
    - ProjA
      - dev
      - qs
      - live
    - ProjB
      - dev
      - qs
      - live

- **Import / edit tests remotely**
- **Import / edit config remotely**

ProjA Slave

Dev

QS

Live
Thank you!

Big Data QA is different!

FitNesse Wiki test execution!

Define CSV / thrift / real-world data!

Inspect results via Pig/Hive

Use suites & fixtures for jobs/workflows!
Want more? Inovex trains you!

- **Android Developer Training** (3 days, Karlsruhe/München)
- **Certified Scrum Developer Training** (5 days, Köln)
- **Hadoop Developer Training** (3 days, Karlsruhe/Köln)
- **Liferay Portal-Developer Training** (4 days, Karlsruhe)
- **Liferay Portal-Admin Training** (3 days, Karlsruhe)
- **Pentaho Data Integration Training** (4 days, München/Köln)

Information and registration at

[www.inovex.de/offene-trainings](http://www.inovex.de/offene-trainings)