inovex

Bug bites Elephant?

Test-driven Quality Assurance

in Big Data Application Development

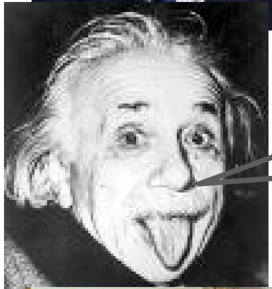
Dr. Dominik Benz, Inovex GmbH

2013/06/03, Berlin Buzzwords

Who speaks... ... the Elephant language?

Class A extends Mapper…

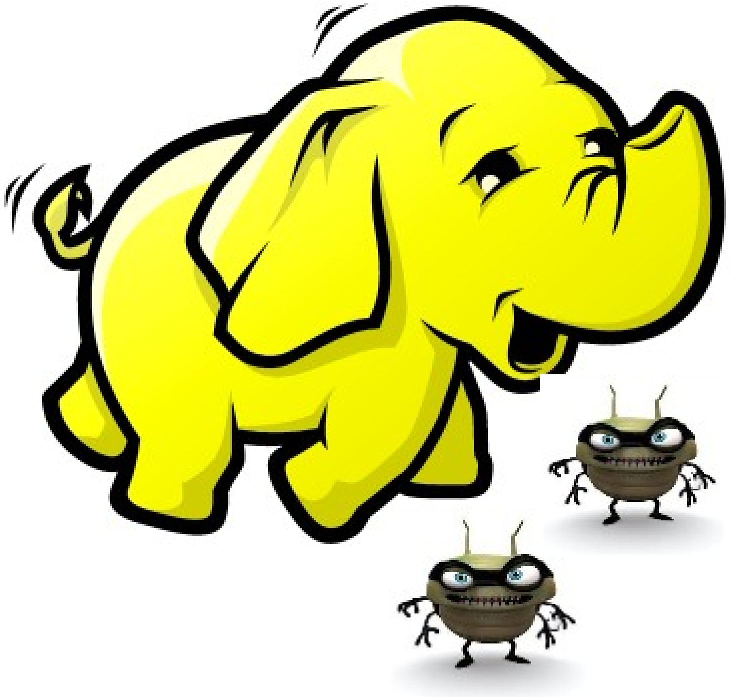
ROI, $$, …



$$D_{chord}(x_1, x_2) = \sqrt{\sum_1^p \left(\frac{x_{1j}}{\sqrt{\sum_1^p x_{1j}^2}} - \frac{x_{2j}}{\sqrt{\sum_1^p x_{2j}^2}}\right)^2}$$

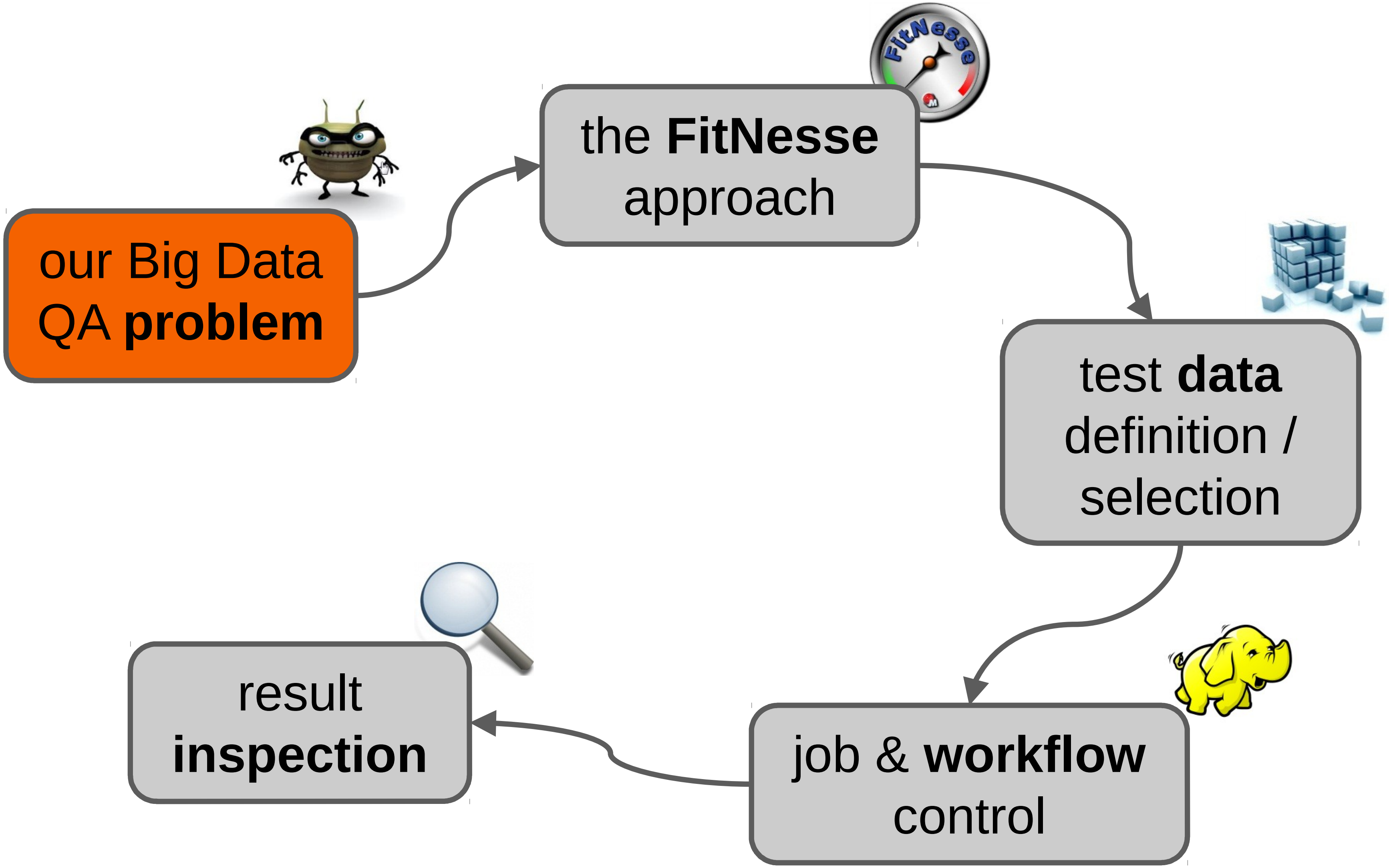
apt-get install…

TDD!



Write/execute tests, specify acceptance

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



our Big Data QA **problem**

the **FitNesse** approach

test **data** definition / selection

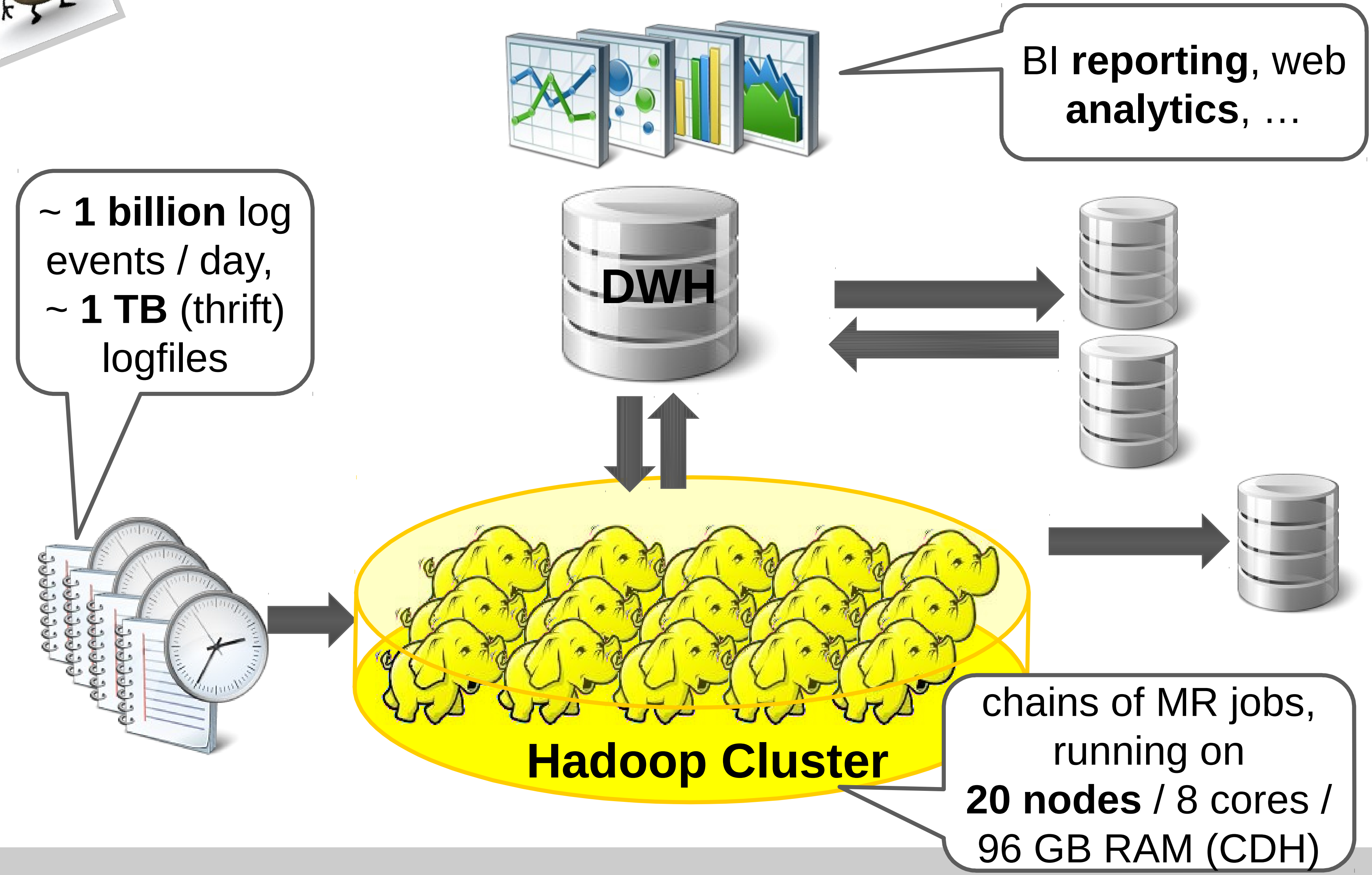
job & **workflow** control

result **inspection**

QA
em

Web Intelligence @ 1&1

inovex



BI **reporting**, web **analytics**, …

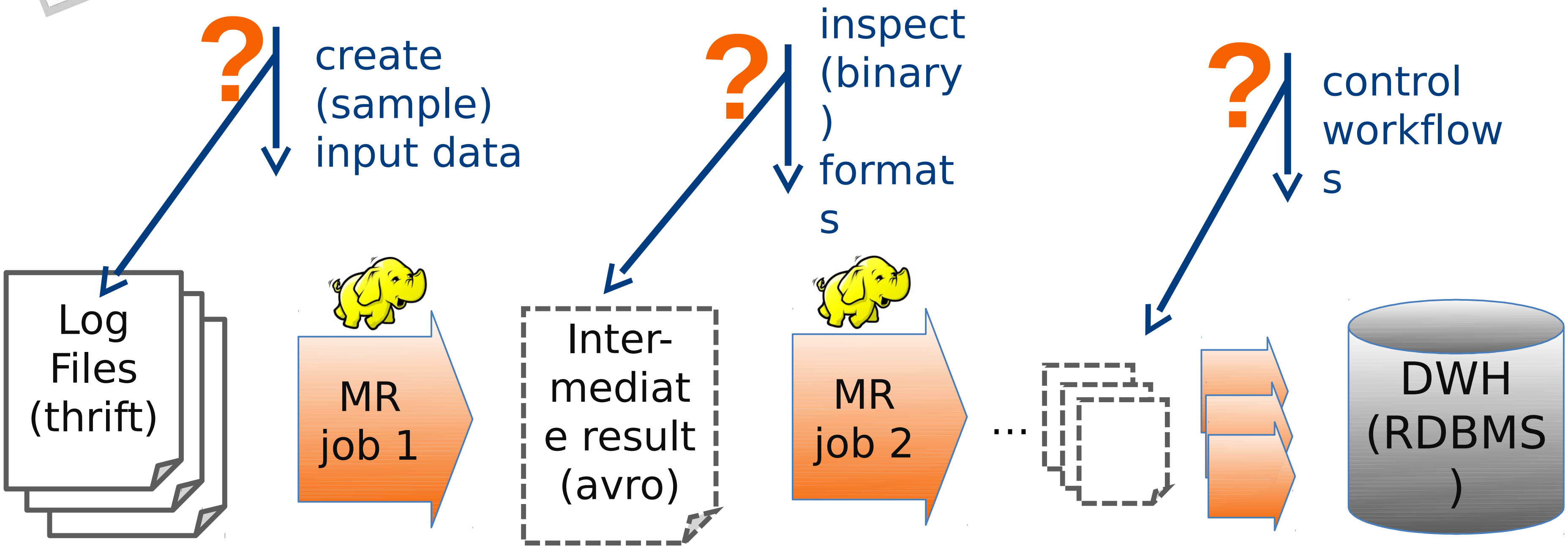
~ **1 billion** log events / day, ~ **1 TB** (thrift) logfiles

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

QA

em

An exemplary workflow



?

create (sample) input data

?

inspect (binary) formats

?

control workflows

Log Files (thrift)

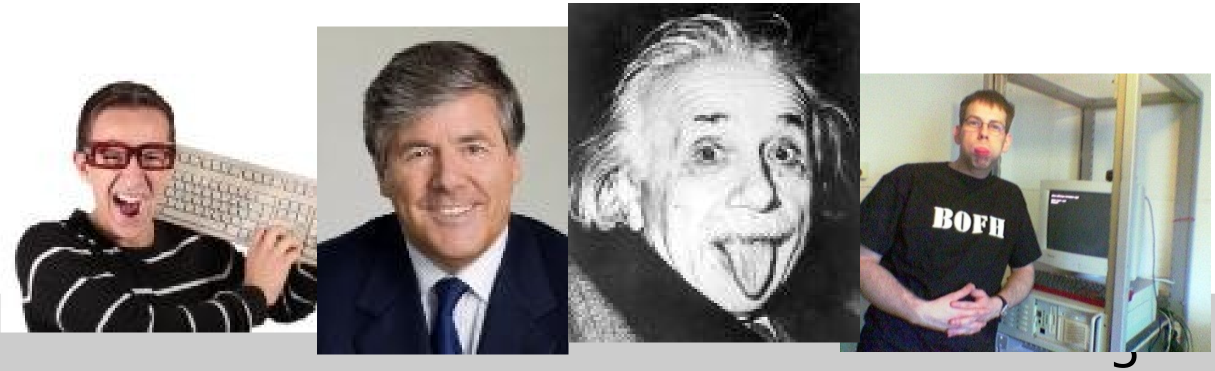
MR job 1

Intermediate result (avro)

MR job 2

…

DWH (RDBMS)

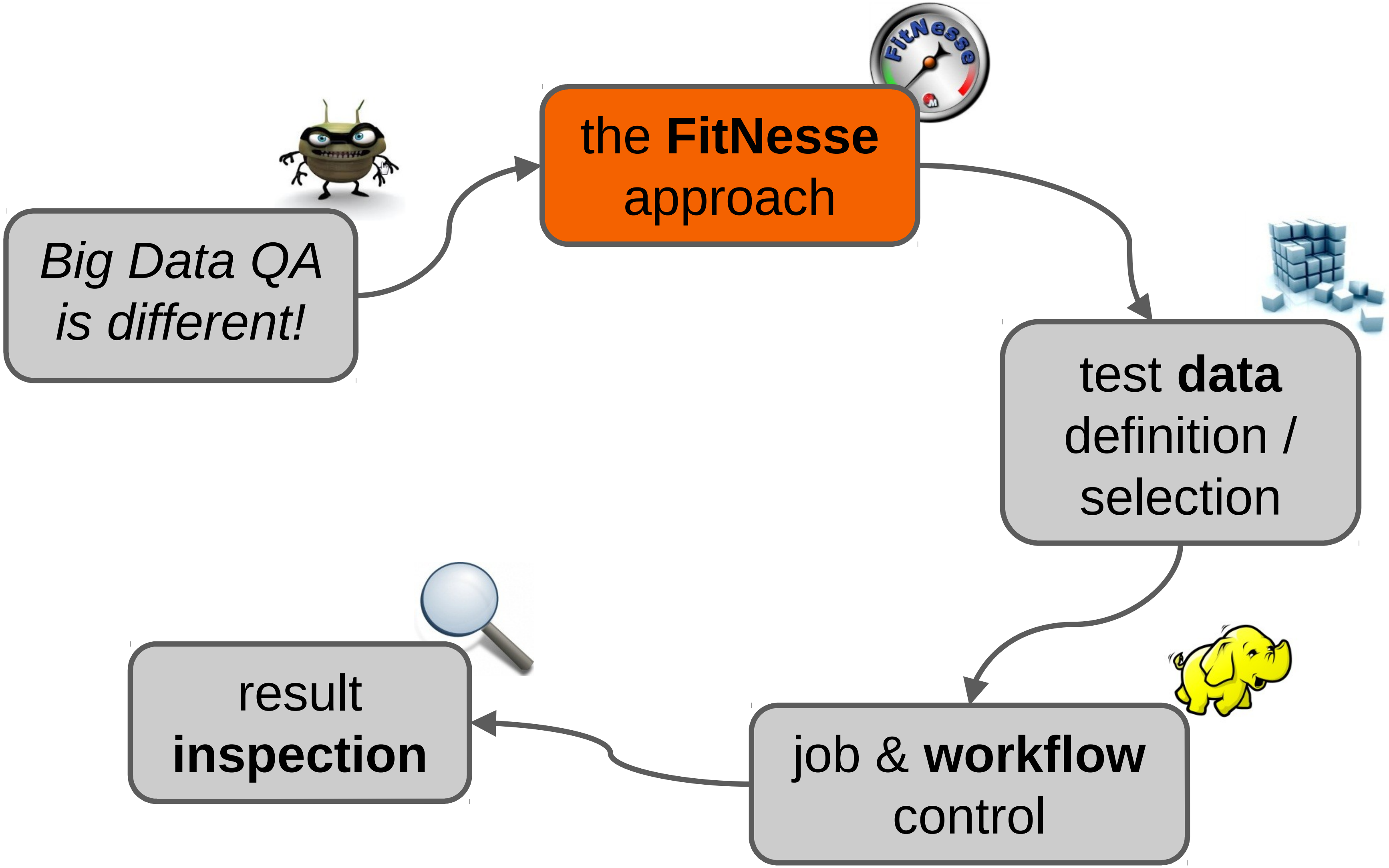


Existing Approaches

metho d	tests what?	issues for our usecase
JUnit	isolated functions	no integration, Java syntax
MRUnit	1 mapper + 1 reducer	„little“ integration, Java syntax
iTest	hadoop jobs/workflows	Java / Groovy syntax
Scripts/ CLI	(manual)	„script chaos“,

⇒ **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**

In a nutshell

„fully integrated standalone wiki and acceptance testing framework”

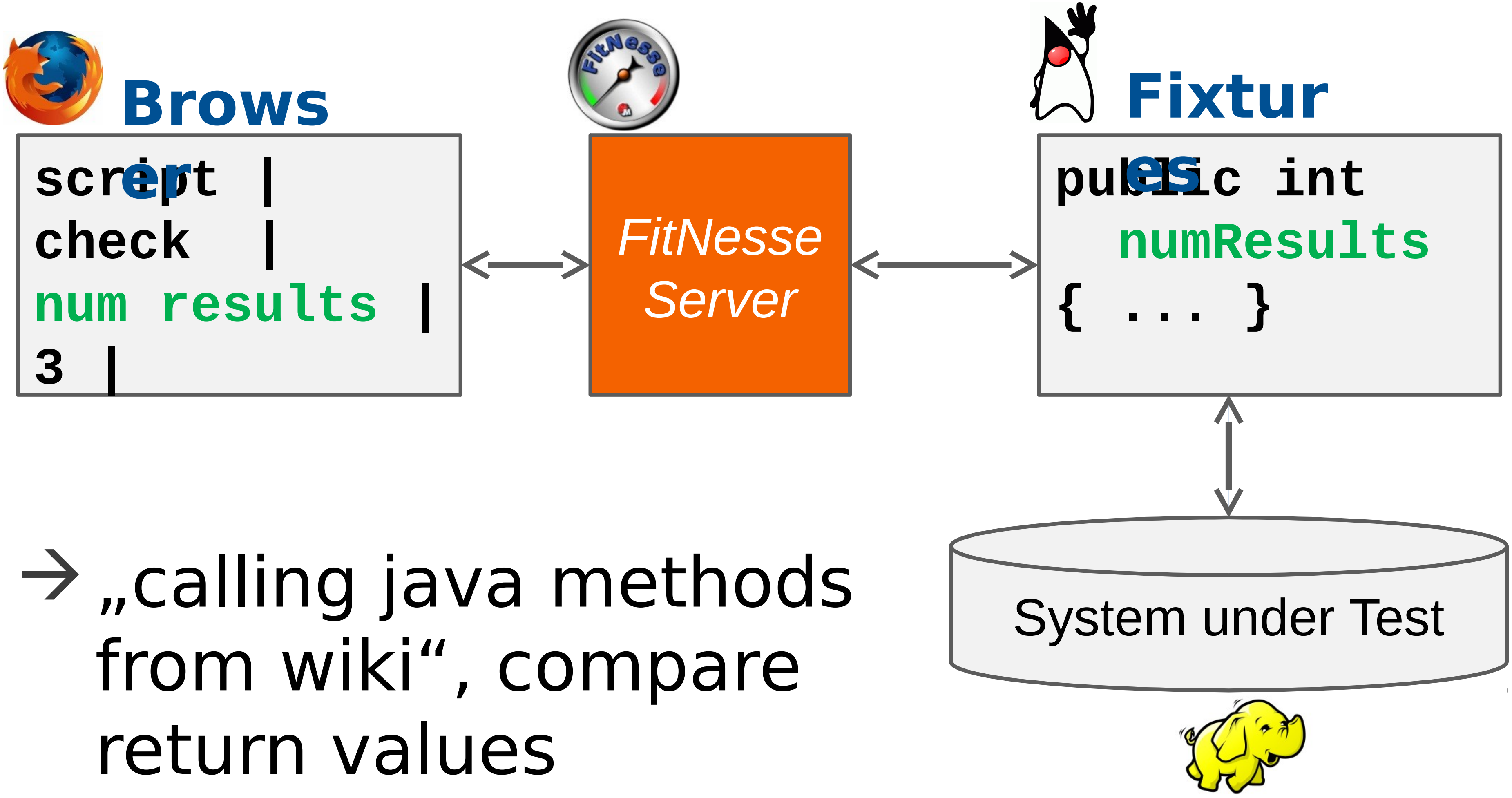
„executable“ **Wiki-**Pages (returning test results)

(almost) **natural language** test specification

connection to SUT via (Java-)“**Fixtures**“

FitNesse Architecture Overview

inovex



Browser

```
script |
check   |
num results |
3 |
```

FitNesse Server

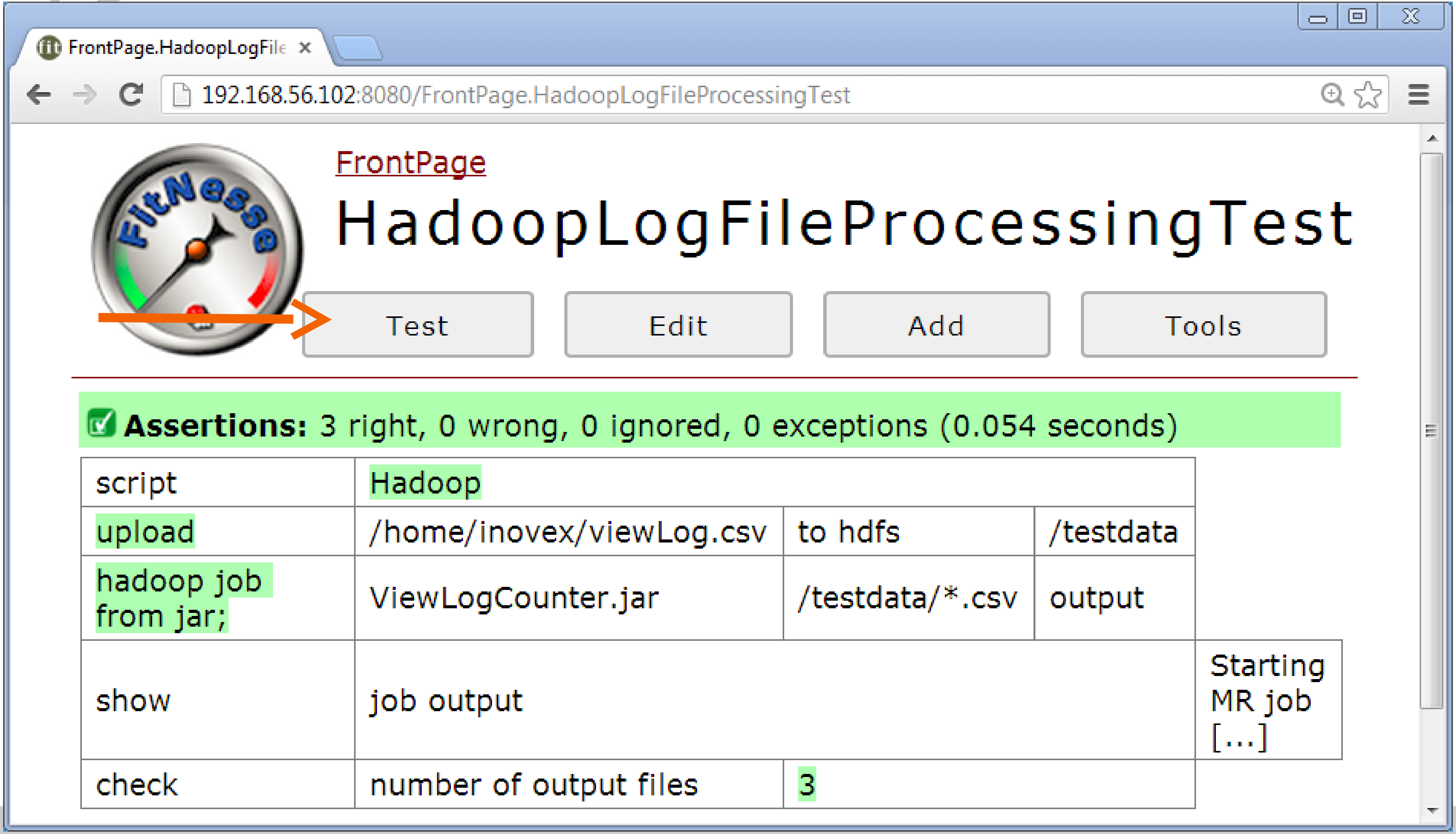
Fixtures

```
public int
   numResults
{ ... }
```

System under Test

- → „calling java methods from wiki“, compare return values
- → Integrates with REST, Jenkins

An Exemplary Test



192.168.56.102:8080/FrontPage.HadoopLogFileProcessingTest

FrontPage

HadoopLogFileProcessingTest

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

script	Hadoop			
upload	/home/inovex/viewLog.csv	to hdfs	/testdata	
hadoop job from jar;	ViewLogCounter.jar	/testdata/*.csv	output	
show	job output			Starting MR job [...]
check	number of output files	3		

Exemplary Test Source

```
!path /home/inovex/lib/*.jar


| script | Hadoop |
| upload | viewLog.csv | to hdfs | /testdata/ |
| hadoop job from jar  | viewLog.jar | [...] |
| show   | job output |
| check  | number of output files | 3 |
```

Hadoop Fixture Java Code

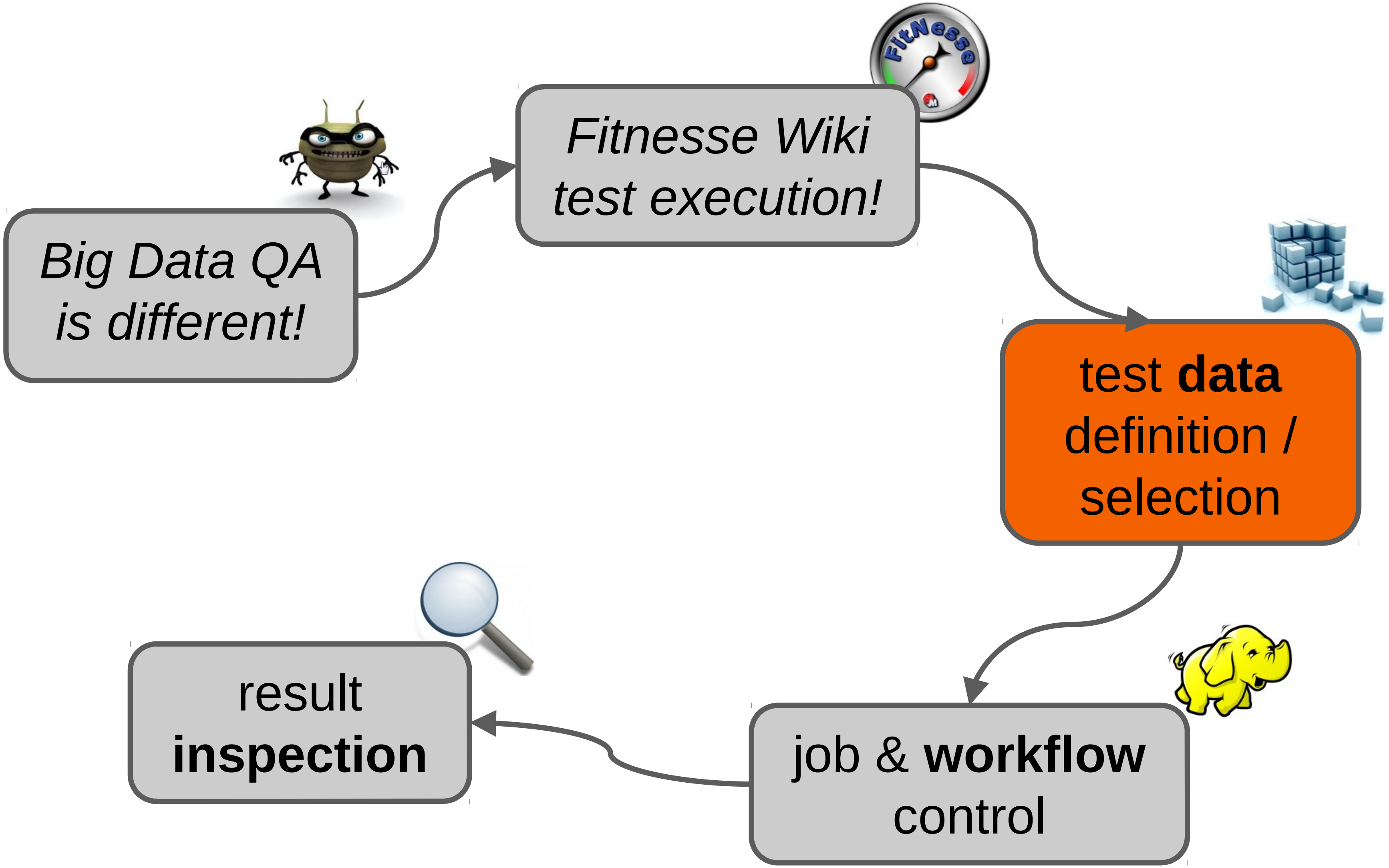
```
public class Hadoop {
  public boolean uploadToHdfs(String localFile,
                       String remoteFile) {...}

  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**

Test D

CSV

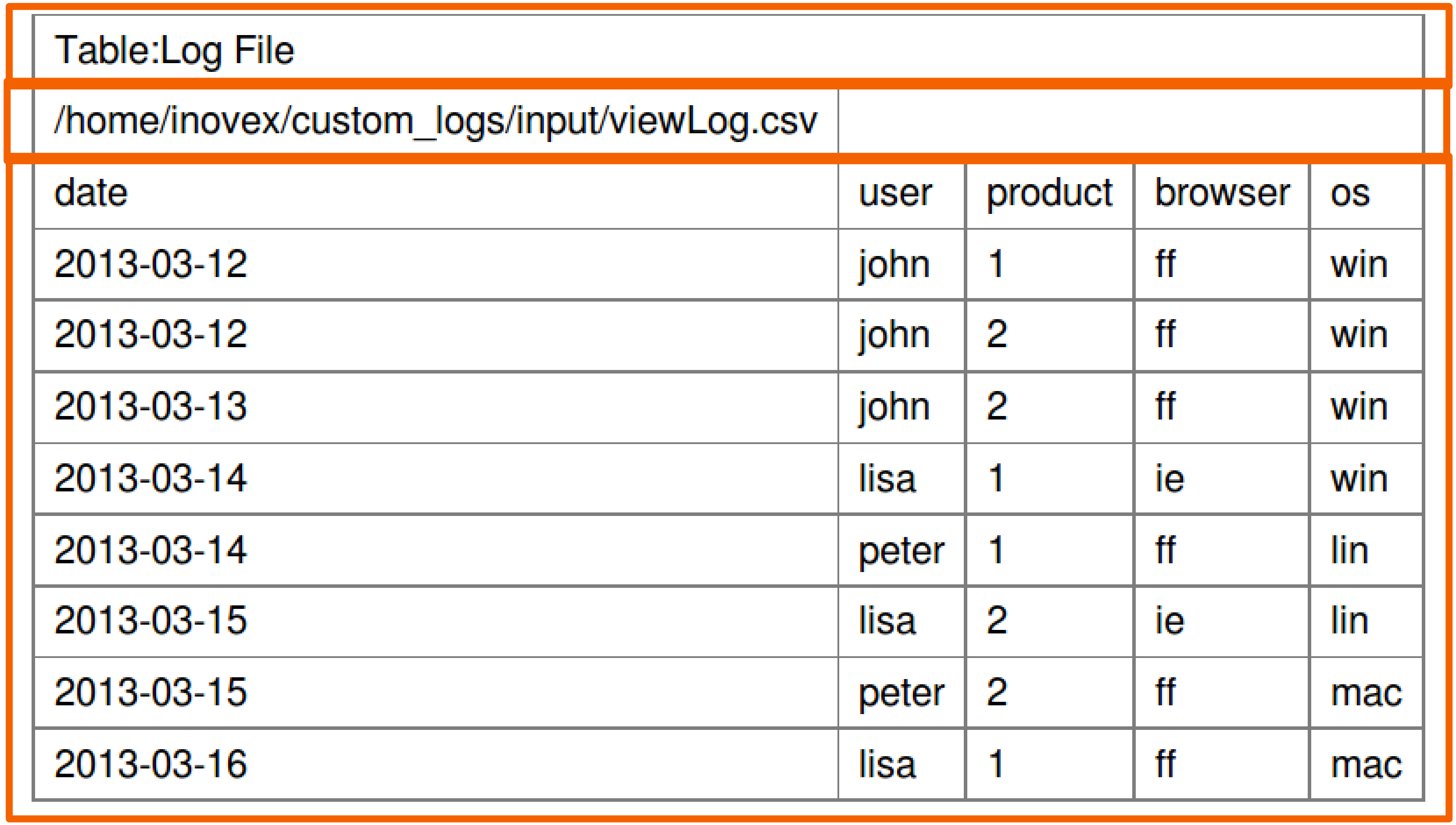


Table:Log File				
/home/inovex/custom_logs/input/viewLog.csv				
date	user	product	browser	os
2013-03-12	john	1	ff	win
2013-03-12	john	2	ff	win
2013-03-13	john	2	ff	win
2013-03-14	lisa	1	ie	win
2013-03-14	peter	1	ff	lin
2013-03-15	lisa	2	ie	lin
2013-03-15	peter	2	ff	mac
2013-03-16	lisa	1	ff	mac

Thrift

- Big Data: **Efficient** data transfer among **heterogeneous sources**
- Def
lan

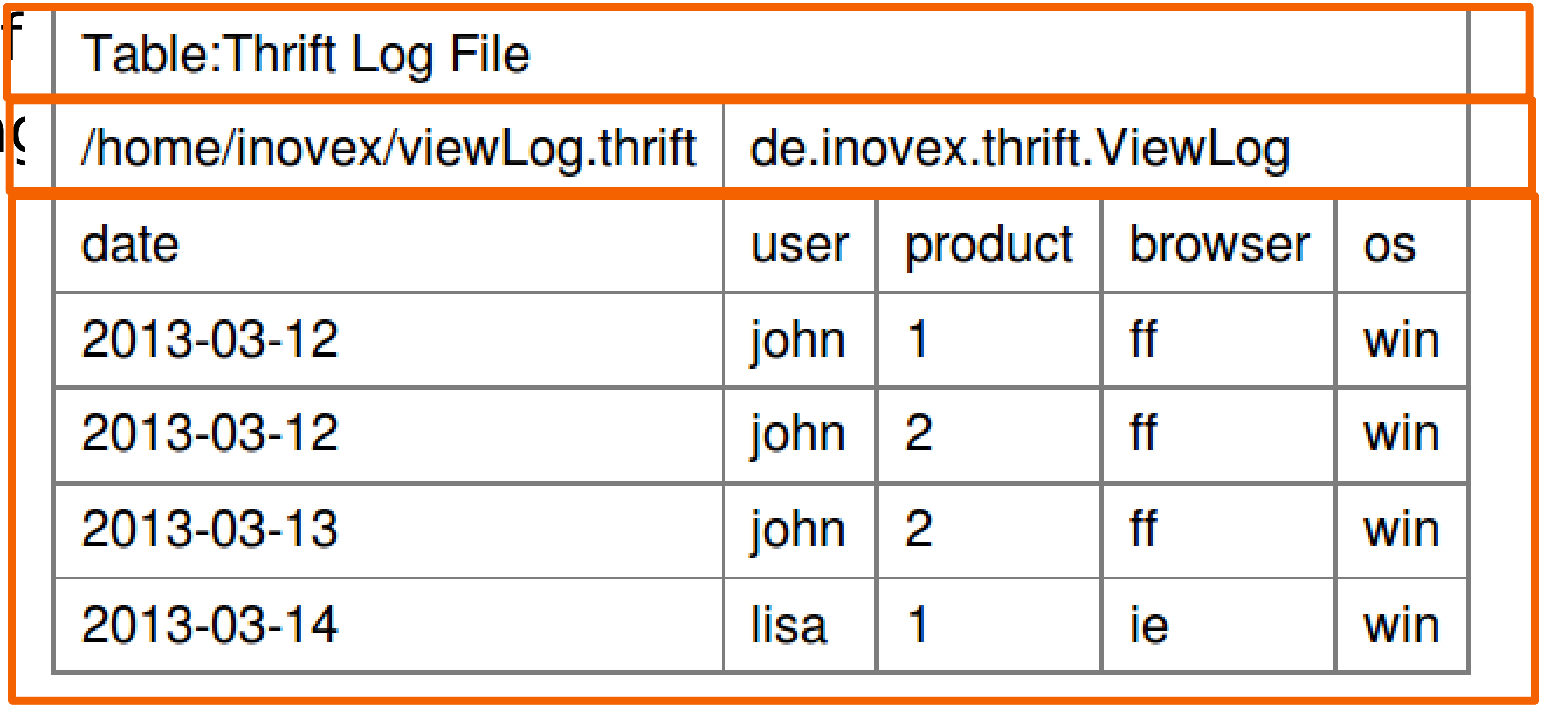


Table:Thrift Log File

/home/inovex/viewLog.thrift	de.inovex.thrift.ViewLog			
date	user	product	browser	os
2013-03-12	john	1	ff	win
2013-03-12	john	2	ff	win
2013-03-13	john	2	ff	win
2013-03-14	lisa	1	ie	win

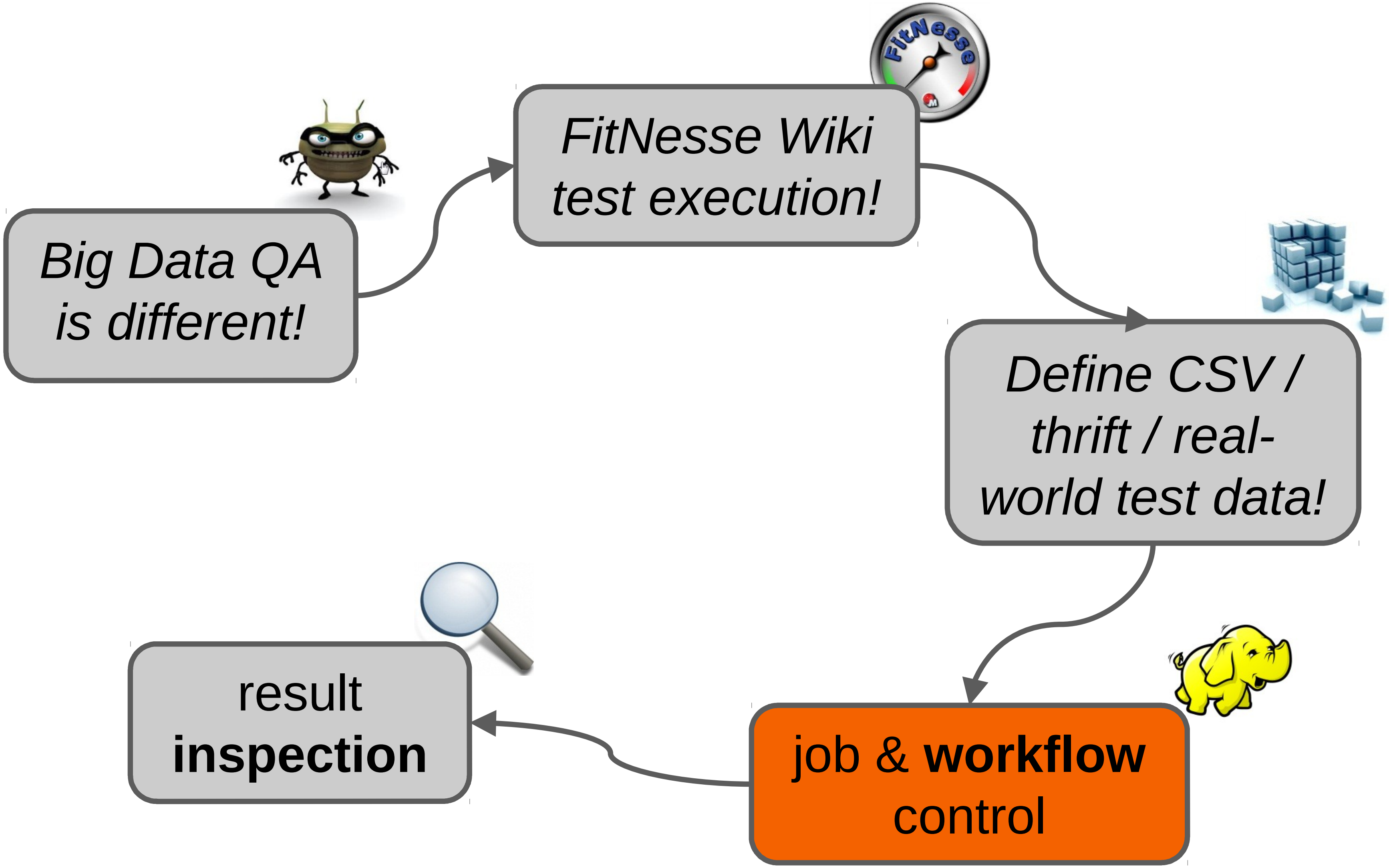
Real World Data

- 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!

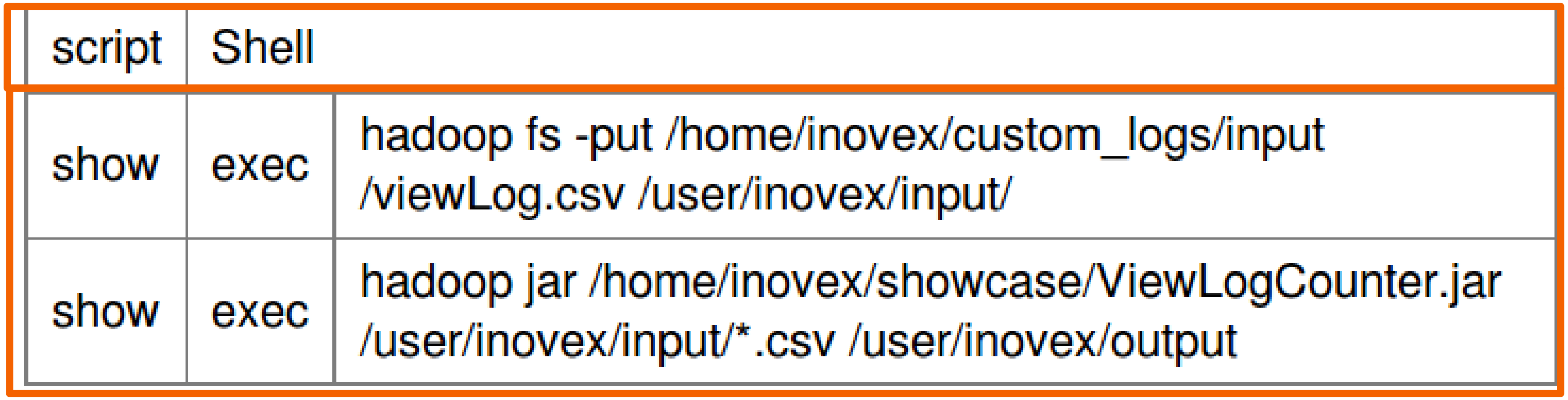
FitNesse Wiki test execution!

Define CSV / thrift / real-world test data!

job & **workflow** control

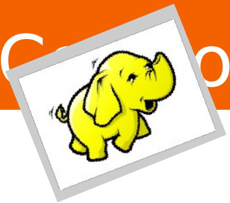
result **inspection**

Swiss Army Knife: Shell

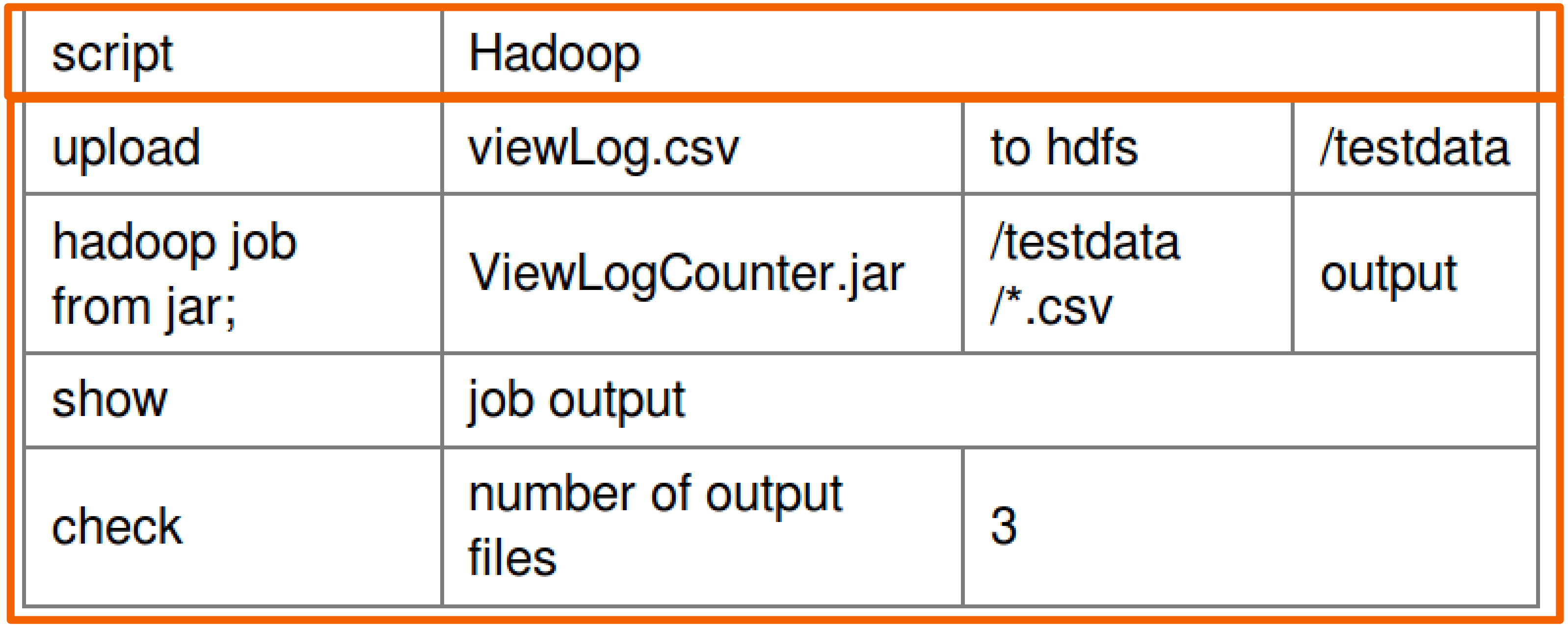


script	Shell	
show	exec	hadoop fs -put /home/inovex/custom_logs/input /viewLog.csv /user/inovex/input/
show	exec	hadoop jar /home/inovex/showcase/ViewLogCounter.jar /user/inovex/input/*.csv /user/inovex/output

- Execute arbitrary (shell) commands
- Mainly a **wrapper** around **apache.commons.exec.CommandLine**

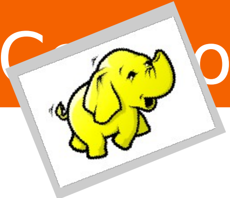


Hadoop Fixture



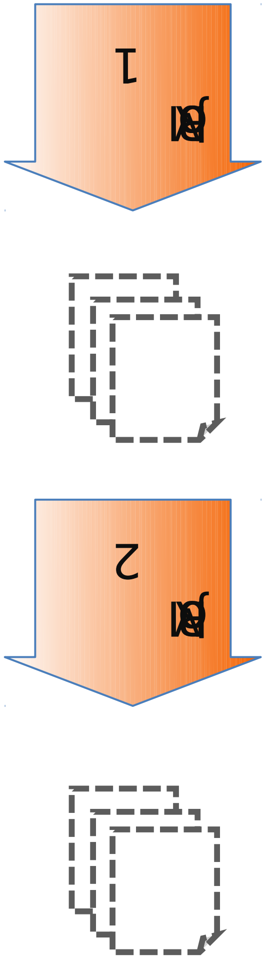
script	Hadoop		
upload	viewLog.csv	to hdfs	/testdata
hadoop job from jar;	ViewLogCounter.jar	/testdata /*.csv	output
show	job output		
check	number of output files	3	

- **Hide complexity** from test authors
- „define“ appropriate **test language** via (Java) method names
- **re-use** other fixtures (Shell, ...) internally

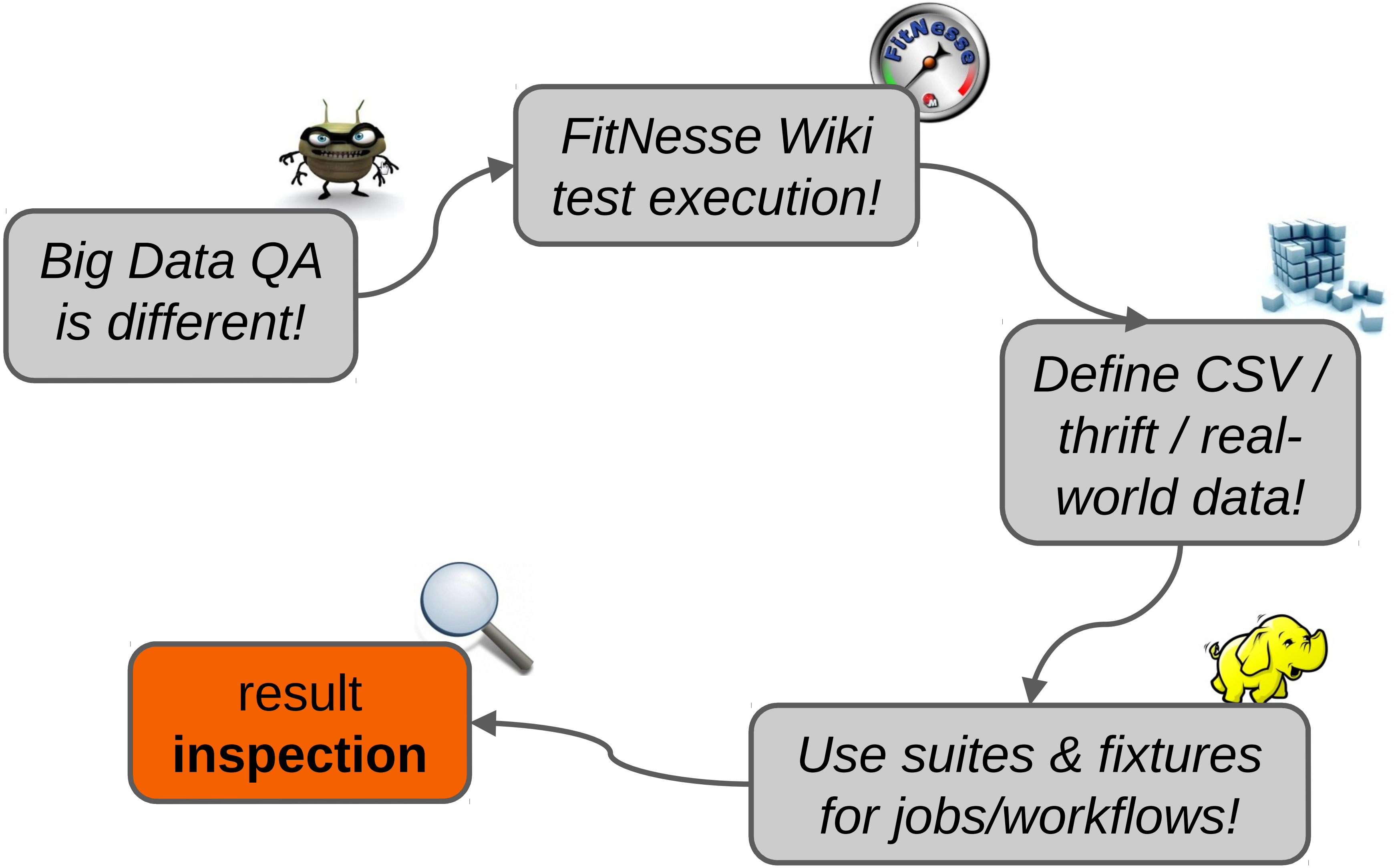


Workflows & Suites

- 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!

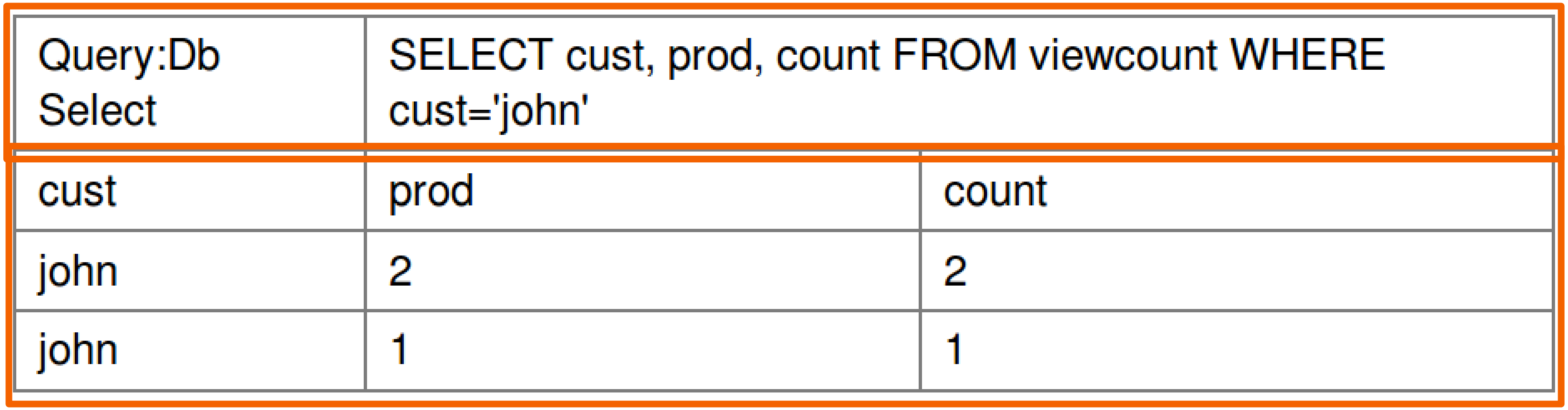
Define CSV / thrift / real-world data!

Use suites & fixtures for jobs/workflows!

result **inspection**

Results

Data Warehouse / Hive

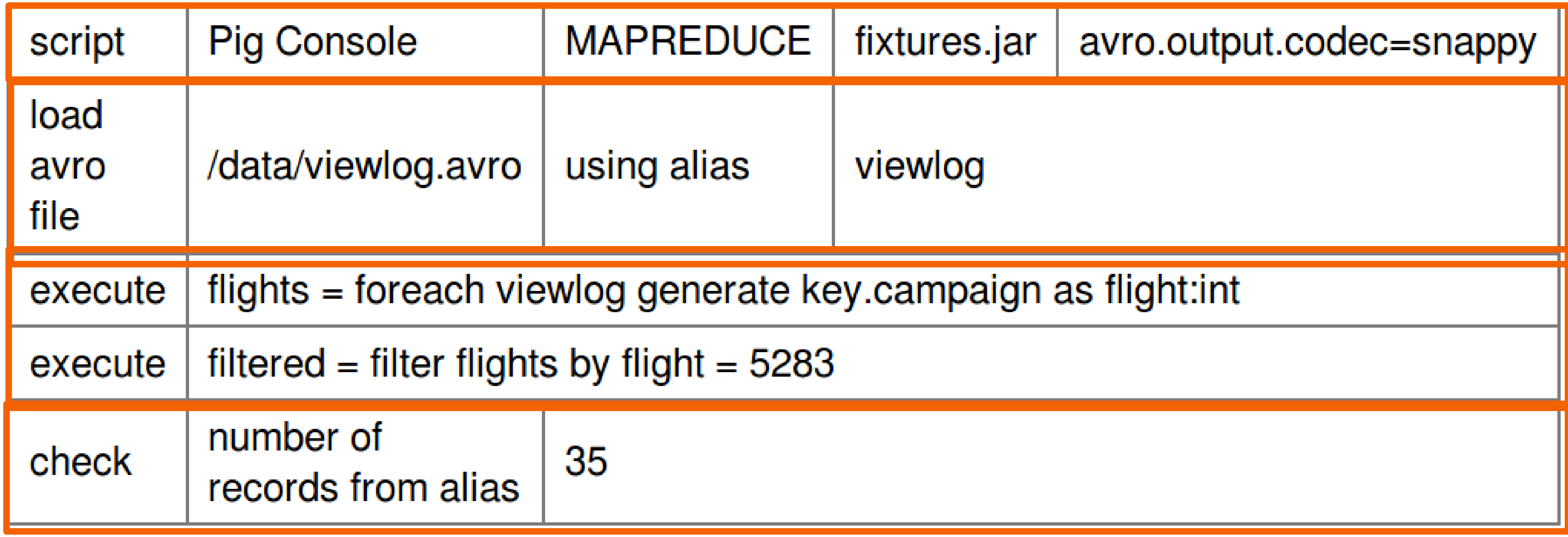


Query:Db Select	SELECT cust, prod, count FROM viewcount WHERE cust='john'	
cust	prod	count
john	2	2
john	1	1

- Validate **RDBMS contents** (via JDBC)
- E.g. for checking the **final** result
- Or use **Hive** + Hive-Server to query raw data

Results

Pig



script	Pig Console	MAPREDUCE	fixtures.jar	avro.output.codec=snappy
load avro file	/data/viewlog.avro	using alias	viewlog	
execute	flights = foreach viewlog generate key.campaign as flight:int			
execute	filtered = filter flights by flight = 5283			
check	number of records from alias	35		

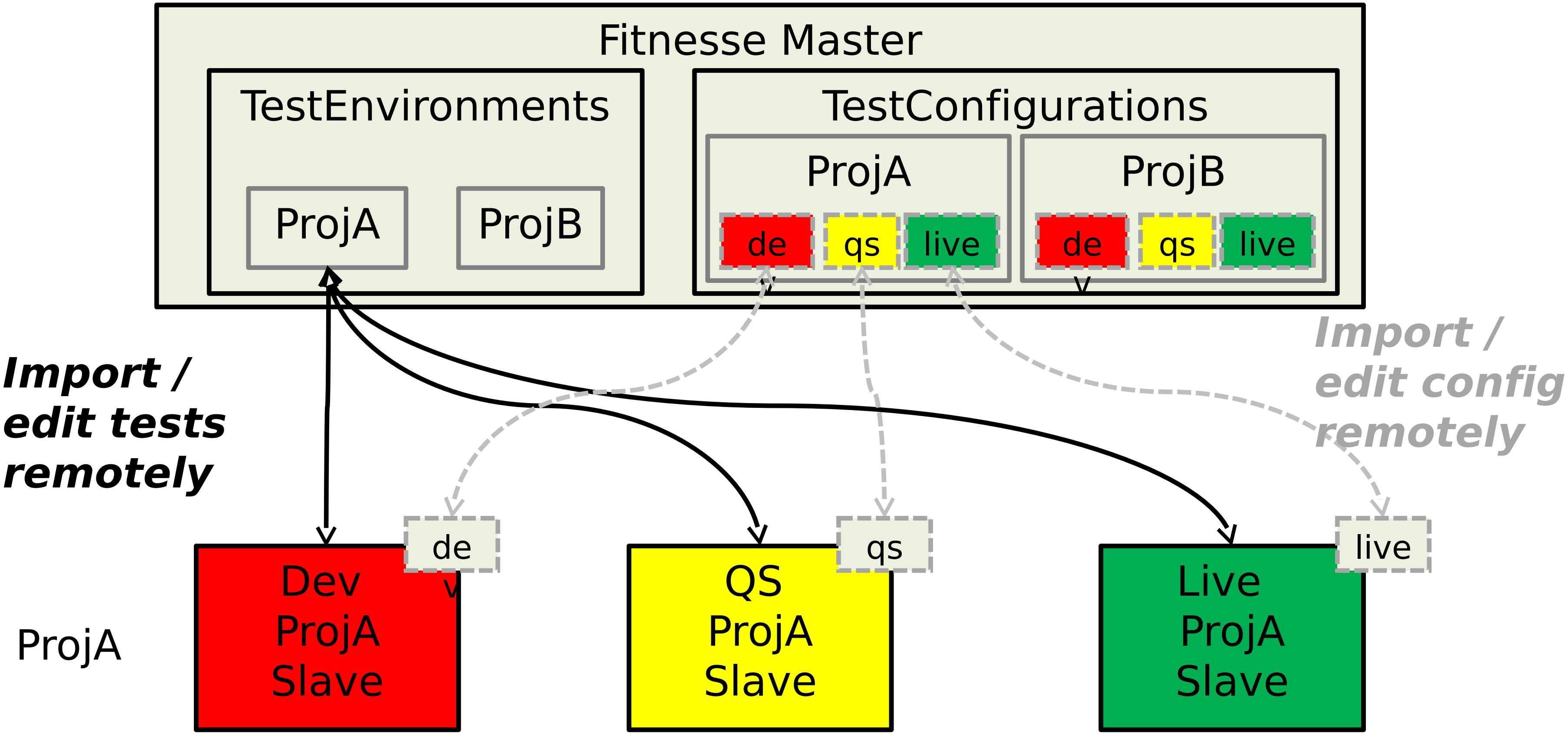
- Execute arbitrary **pig commands** from Wiki page
- Inspect e.g. **binary intermediate results** (avro, …)

Pig Fixture extends PigServer

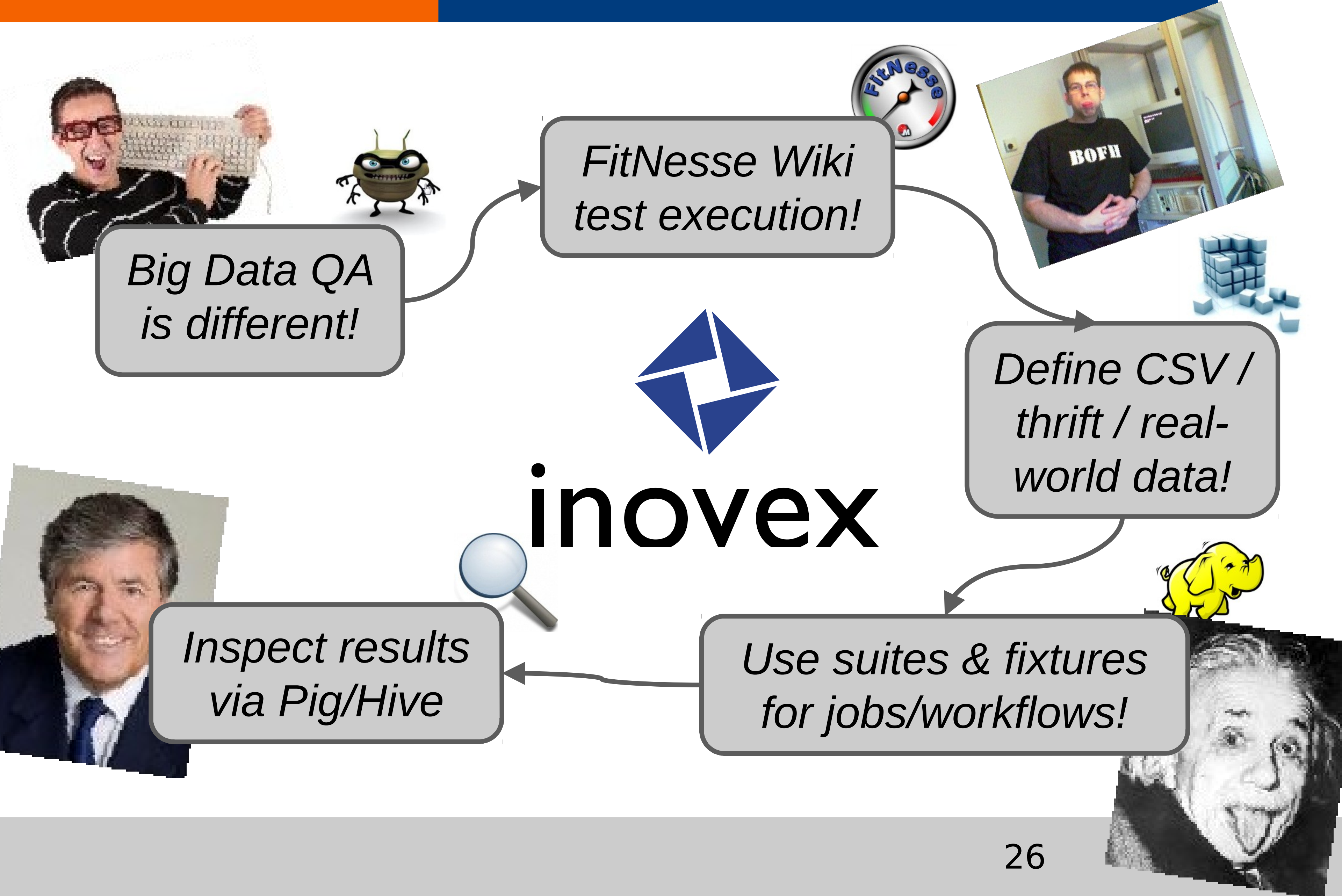
```
public class PigConsole extends PigServer {

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

Server Infrastructure



Thank you! dominik.benz@inovex.de



Big Data QA is different!

FitNesse Wiki test execution!

Define CSV / thrift / real-world data!

Use suites & fixtures for jobs/workflows!

Inspect results via Pig/Hive

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

Inovex @bbuzz

inovex

Stefan

Kathrin

Bernhard

Jörg

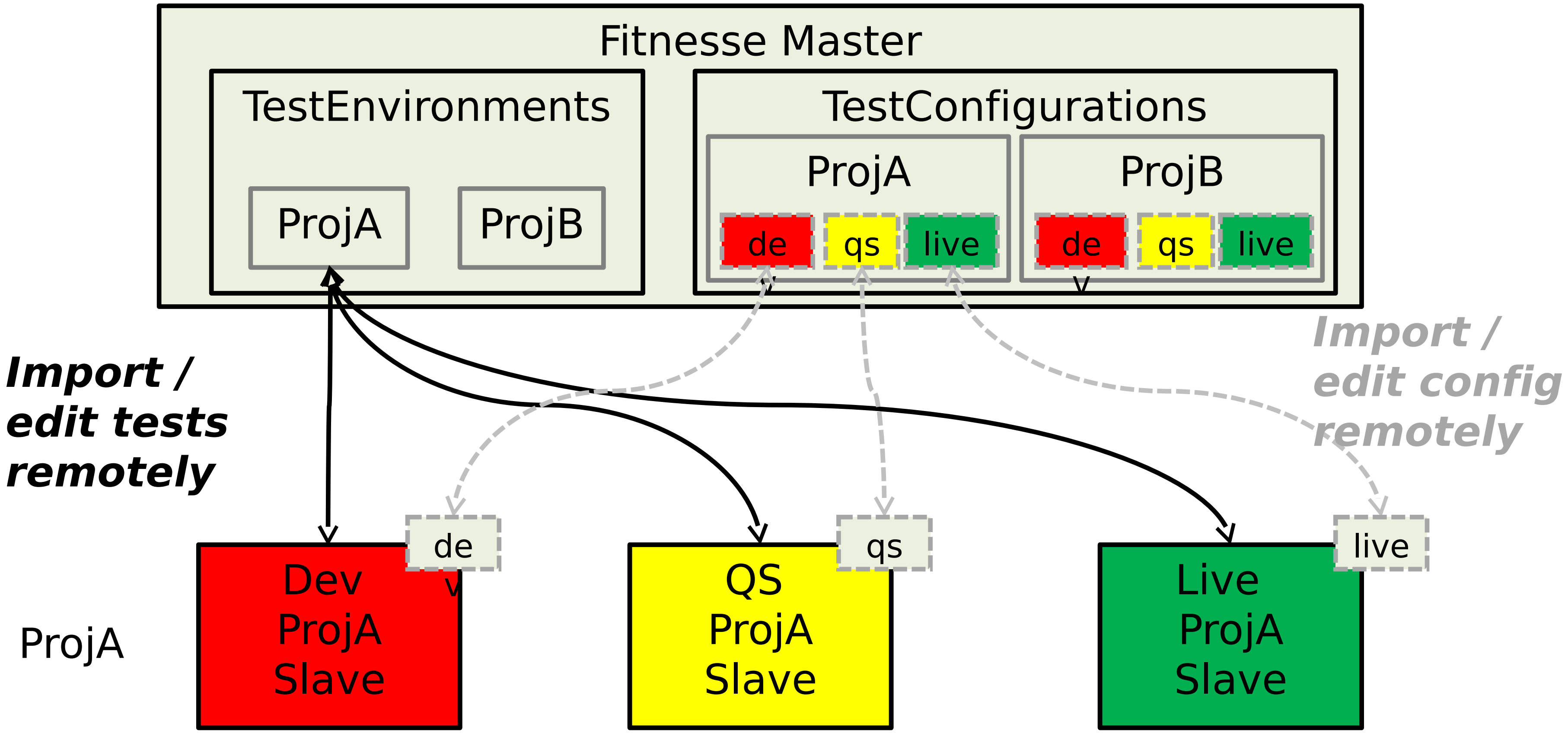
Andrew

Christian

Christian

BACKUP

FitNesse Server Infrastructure



Demo

- Download & install FitNesse server
- Create csv log file
- Run hadoop job which counts viewed items
- Inspect Results with Hive

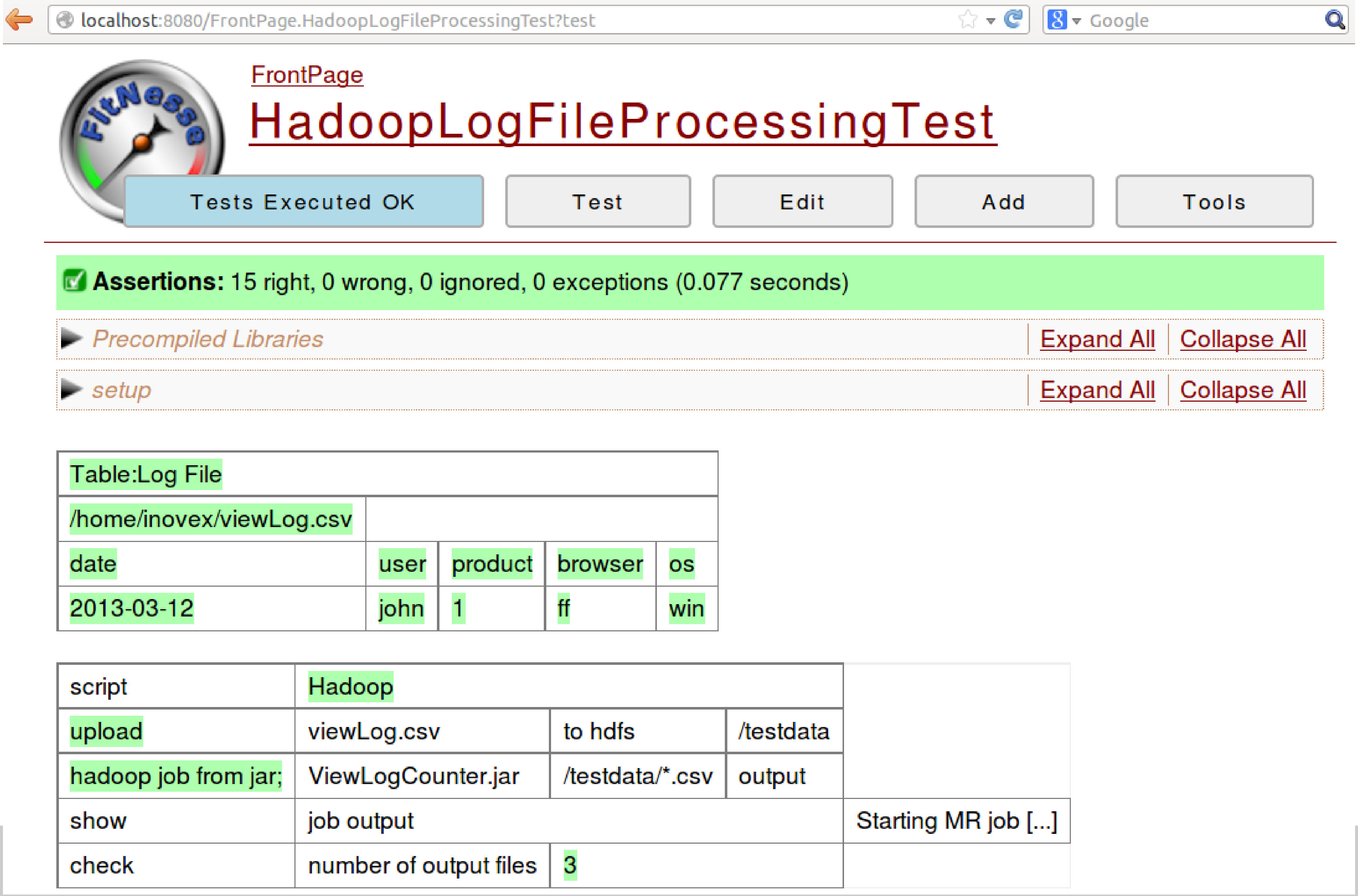
Exemplary Test Source

```
!path /home/inovex/lib/*.jar


| Table:Log File             |
| /home/inovex/viewLog.csv | |
| date       | user  | product | browser | os  |
| 2013-03-12 | john  | 1       | ff      | win |


| script | Hadoop |
| upload | viewLog.csv | to hdfs | /testdata/ |
| hadoop job from jar  | viewLog.jar | [...] |
| show    | job output |
| check   | number of output files | 3 |
```

An Exemplary Test



FrontPage

HadoopLogFileProcessingTest

☑ **Assertions:** 15 right, 0 wrong, 0 ignored, 0 exceptions (0.077 seconds)

▶ *Precompiled Libraries*

▶ *setup*

Table:Log File				
/home/inovex/viewLog.csv				
date	user	product	browser	os
2013-03-12	john	1	ff	win

script	Hadoop			
upload	viewLog.csv	to hdfs	/testdata	
hadoop job from jar;	ViewLogCounter.jar	/testdata/*.csv	output	
show	job output			Starting MR job [...]
check	number of output files	3		