



Agile Methoden als Erfolgsfaktor für BI und Big Data Projekte

Best Practices aus unseren Projekten

Dr. Stefan Igel
Karlsruhe, 20.02.2014

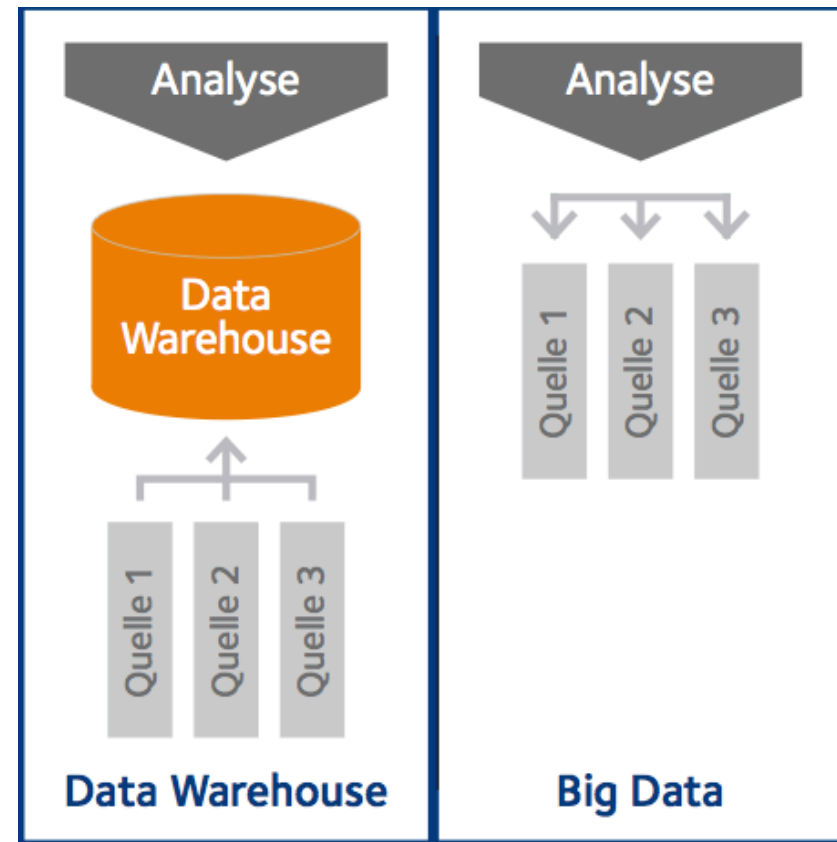
Agenda

1. Agile Ziele
2. Agile Teams
3. Agil BI-Projekte managen
4. Agile BI-Lösungen entwickeln
5. Agil testen
6. Agil dokumentieren

Ziele Agiler Business Intelligence

Analysen in BI und Big Data

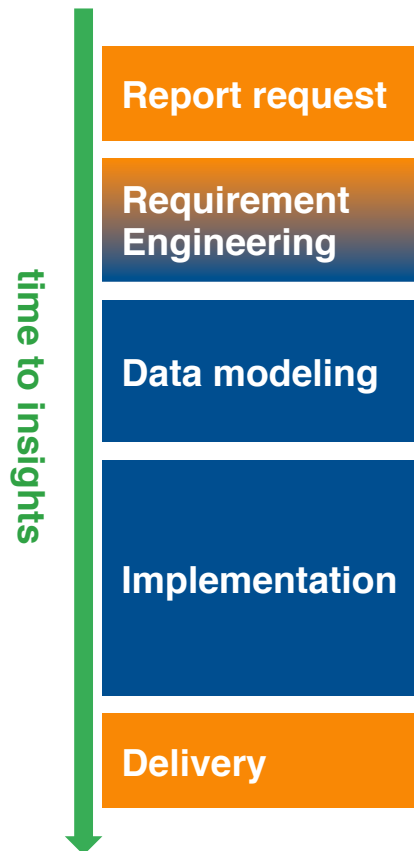
- ▶ Regelbewirtschaftung vs. Ad-hoc Analysen
- ▶ Strukturierte vs. semi- und unstrukturierte Daten
- ▶ Aggregate vs. Rohdaten



Ziele Agiler Business Intelligence

Optimierung der „Time to Insights“

Klassisches Vorgehen



Agiles Vorgehen



KEFs für BI-Lösungen

- ▶ Adaptiv und flexibel
- ▶ Schnell, ausfallsicher, skalierbar
- ▶ Wartungsfreundlich
- ▶ Investitions- und innovationssicher

Das Agile Team

Data Scientist - The „sexiest job of the 21. century“ *



- Statistik
- Mathematik
- Machine Learning
- Data Mining
- IT Engineering
- Daten Architektur
- Software Engineering
- Visualisierung
- Business Analyst

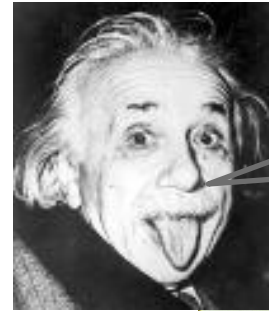


*<http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/>

Das Agile Team

Cross-functional

- Statistik
- Mathematik
- Machine Learning
- Data Mining
- IT Engineering
- Daten Architektur
- Software Engineering
- Visualisierung
- Business Analyst



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



`apt-get
install...`



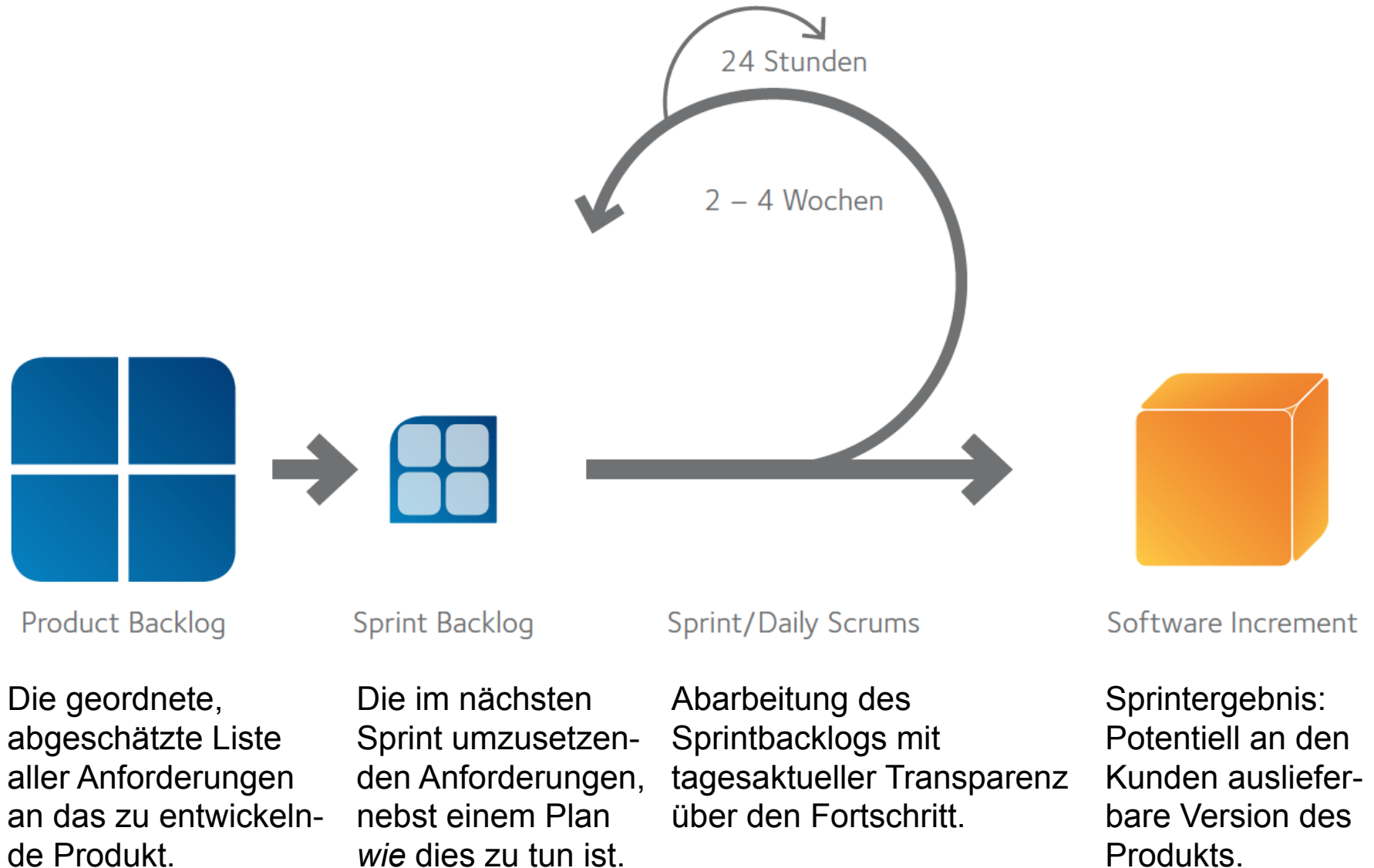
Class A
extends
Mapper...



ROI, \$\$,
...

Agiles Projektmanagement

Scrum im Überblick



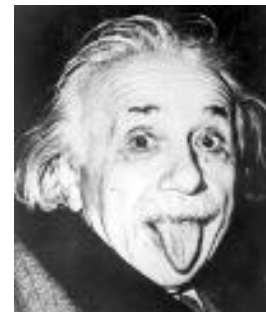
Agiles Projektmanagement

Rollen in Scrum

Product Owner



Team

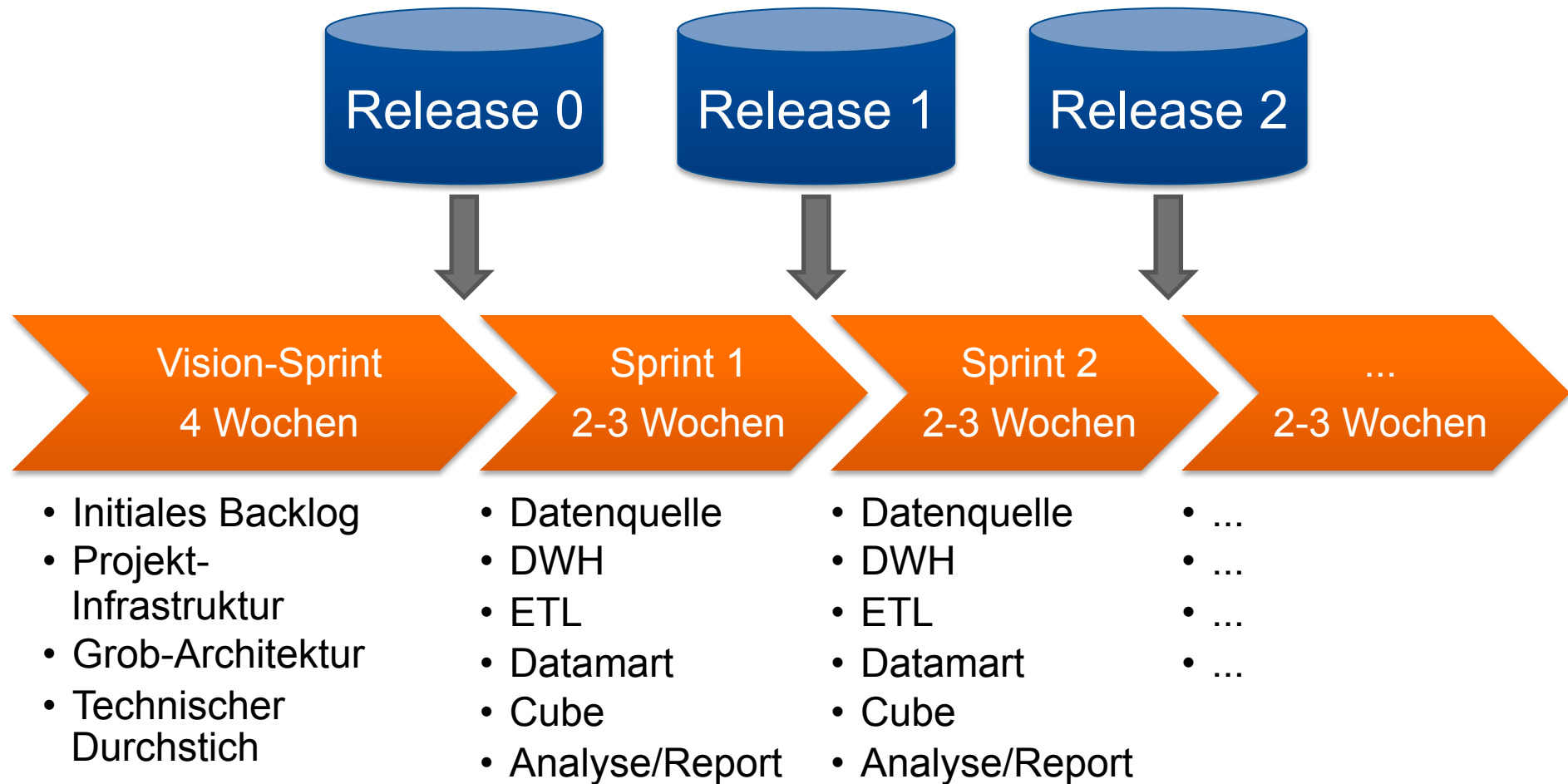


Scrum Master

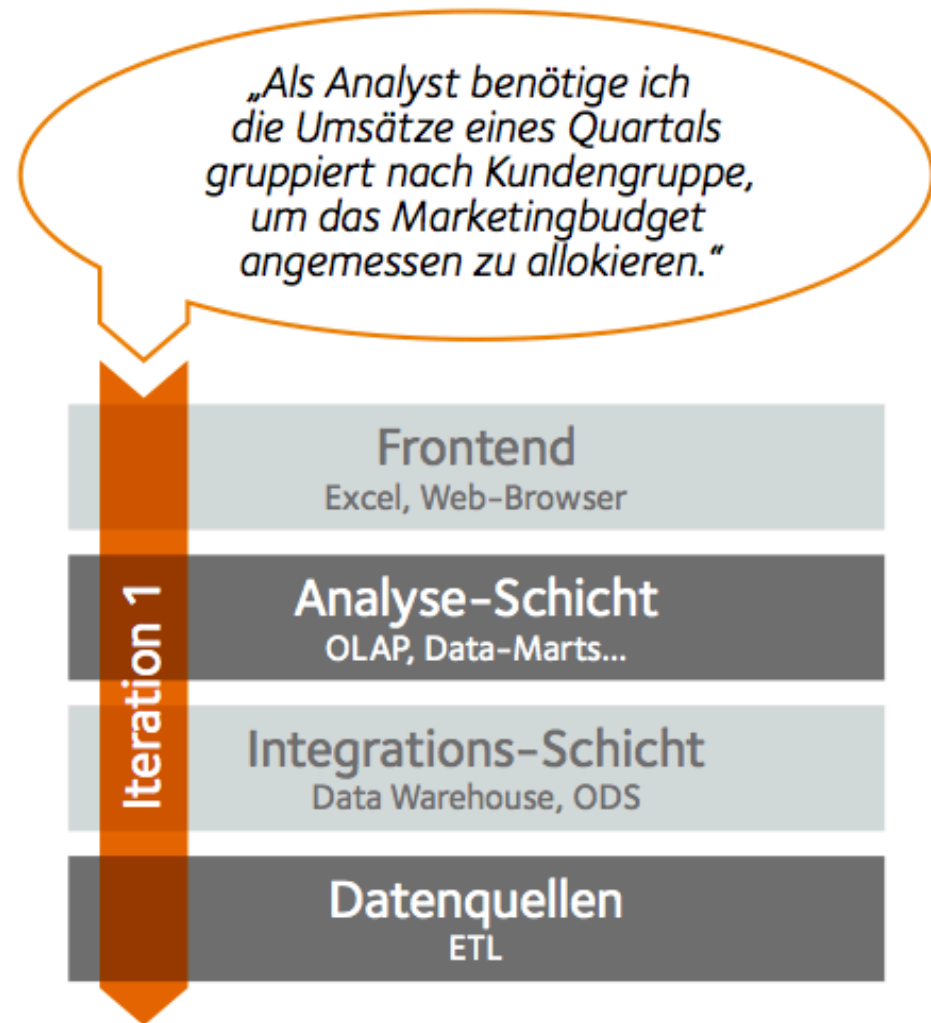


Agiles Projektmanagement

Projektablauf konkret



- Auf Anwender-Nutzen fokussieren
- „Vertikal“ statt „horizontal“ erzählen
- Akzeptanz-Kriterien formulieren
- Definition of Done beachten



Agiles Projektmanagement

PM-Tools: Redmine



Übersicht - BigData-Showcase - inovex Redmine

https://redmine.inovex.de/projects/big-data-showcase

Hauptseite Meine Seite Projekte Scrum Statistiken Hilfe

Angemeldet als sigel Mein Konto Abmelden

BigData-Showcase

Suche: BigData-Showcase

Übersicht Aktivität Roadmap Backlogs Aufgabenliste Releases Tickets Neues Ticket Gantt-Diagramm Kalender News Dokumente Wiki Datelen Konfiguration

Übersicht

BigData-Showcase

Tickets

- Bug: 0 offen / 0
- Task: 8 offen / 8
- Story: 27 offen / 31
- Support: 0 offen / 0
- Performance: 0 offen / 0
- Refactoring: 0 offen / 0

Alle Tickets anzeigen | Kalender | Gantt-Diagramm

BigData-Showcase - inovex Redmine

BigData-Showcase » Master Backlog Darstellungsoptionen (2) Disable Auto-refresh Refresh Multiline

sprint 3: Social Media Content 2014-02-14 2014-02-21 35

13799	Als Data Analyst möchte ich die Twitter-Tweets des inovex-Accou	New	3.0
13800	Als Data Analyst möchte ich alle öffentlich zugänglichen Twitter-T	New	3.0
13801	Als Data Analyst möchte ich die Facebook-Events des inovex Acc	New	2.0
13802	Als Data Analyst möchte ich alle		
13803	Als Marketing-Manager möchte		
13804	Als Marketing-Manager möchte		
13805	Als Marketing-Manager möchte		
13806	Als Marketing-Manager möchte		
13808	Als Marketing-Manager möchte		
13809	Als Marketing-Manager möchte		
13810	Als Marketing Manager möchte		
14041	Als MS-affiner Data Analyst mö		

Produkt Backlog Schließe abgeschlossene Sprints 0

5750	Definition der eingesetzten Software	New
5774	Definition des Formats der eingelesenen Daten (analog Staging-t	New
5775	Definition zusätzlicher Angaben (Dimensionen)	New

BigData-Showcase - inovex Redmine

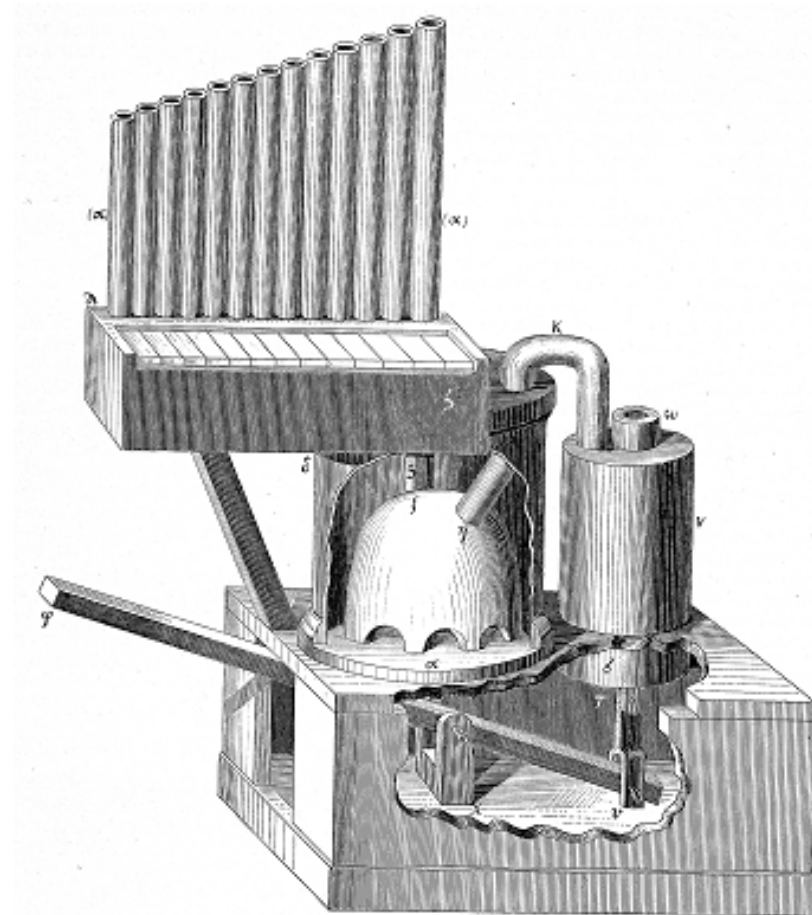
https://redmine.inovex.de/rb/taskboards/377

Story	New	Accepted	In Progress	Resolved
Sprint Blockaden				
Hans-Peter Zorn 13463 (0 Stunden) Konzept für Wiederverwendbarkeit von Connectors/Datacollectors				
Hans-Peter Zorn 13465 (8 Stunden) Evaluierung von SQL-On-Hadoop Lösungen		13509 TestSuite für SQL-On-Hadoop-Evaluierung Hans-Peter Zo 4.0	13459 Presto integrieren/Evalu	13454 Spark/Shark in Showcase integrieren Hans-Peter Zo 2.0
Hans-Peter Zorn 13504 (0 Stunden) TextAnalysis-Plugin-PoC für Datameer				

Agile BI-Entwicklung

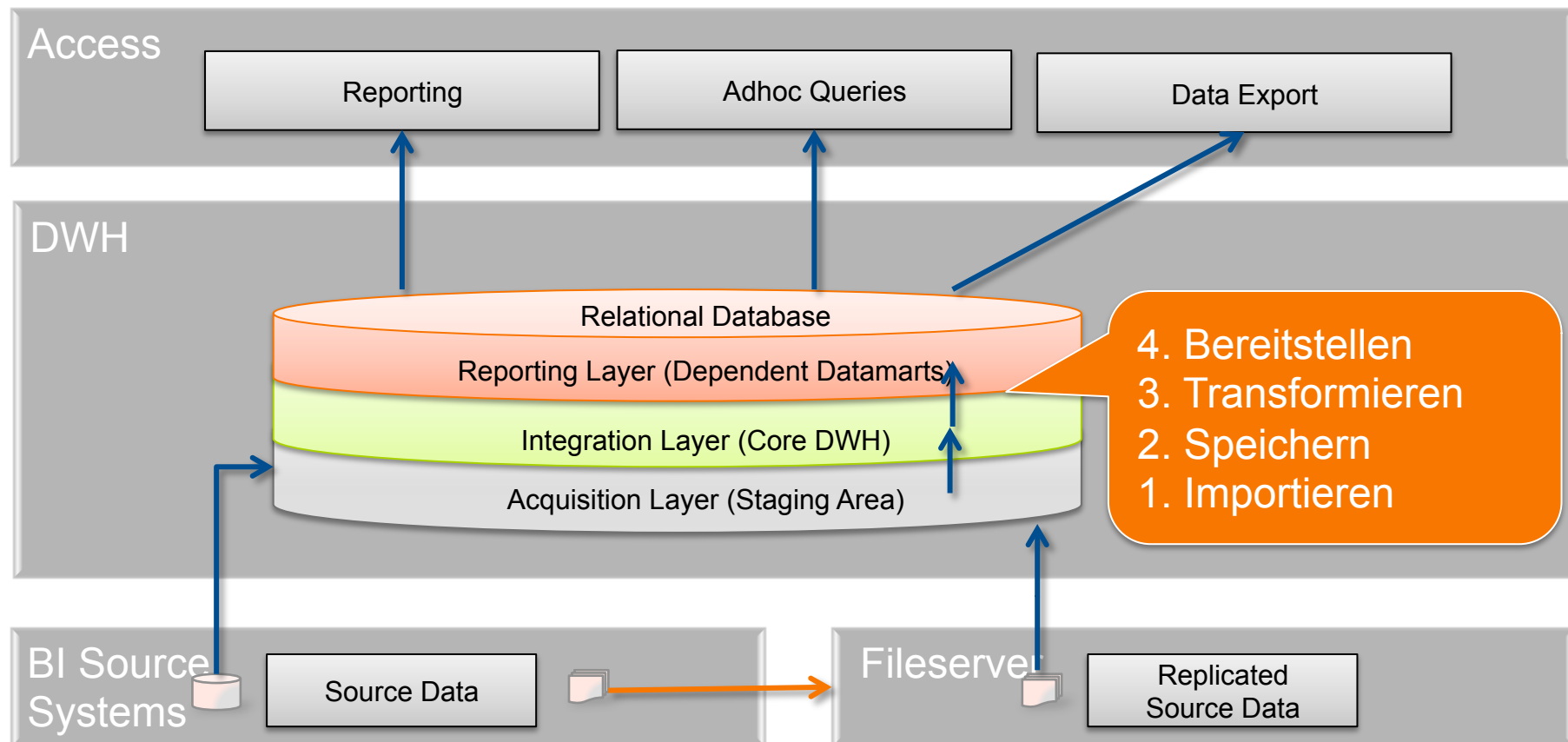
Änderungsfreundlichkeit durch Automatisierung

- Automatisierung der Integration
- Automatisierung der Tests
- Automatisierung der Deployments
- Automatisierung der Dokumentation



Agile BI-Entwicklung

DWH-Architektur



- Etablierte Design-Pattern
- Standardisiert

Agile BI-Entwicklung

Big Data Architektur: Make your Choice!

Applications
and
Analytics



Systems Management



Batch
Processing &
Storage



redis



cassandra

Server



Neo4j



mongoDB



Transport &
Speed

STORM



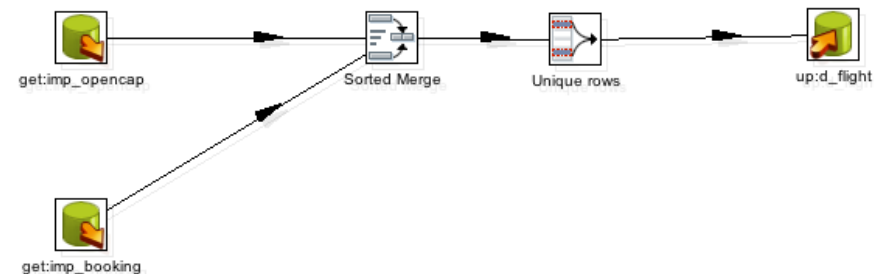
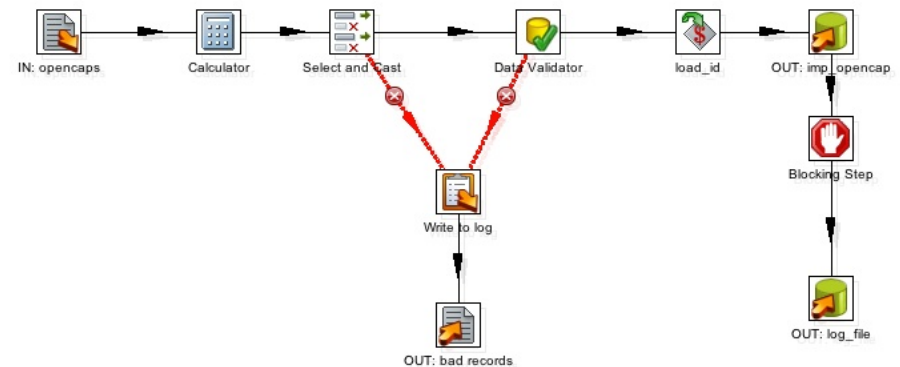
- Extrem hoher Innovationsgrad
- System-Design individuell

Agile BI-Entwicklung

Änderungsfreundliches Software Design Pattern



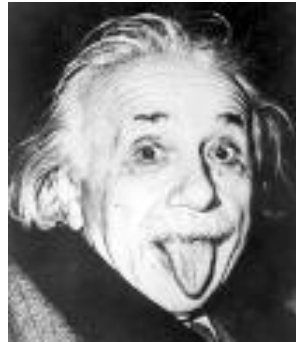
- ▶ ETL als kleine, modulare Einheiten (PDI: Jobs, Transformationen) entwickeln
- ▶ Datenbank-Änderungen pro Release als ALTER TABLE implementieren
- ▶ Fein-granular versionieren, Quellcode-Repository verwenden (z.B. git)
- ▶ Continuous Integration: Täglich einchecken, nächtlich integrieren
- ▶ Hinreichende Testabdeckung sicherstellen



Beispiel: PDI Transformationen

Agiles Testen von BI und Big Data Lösungen

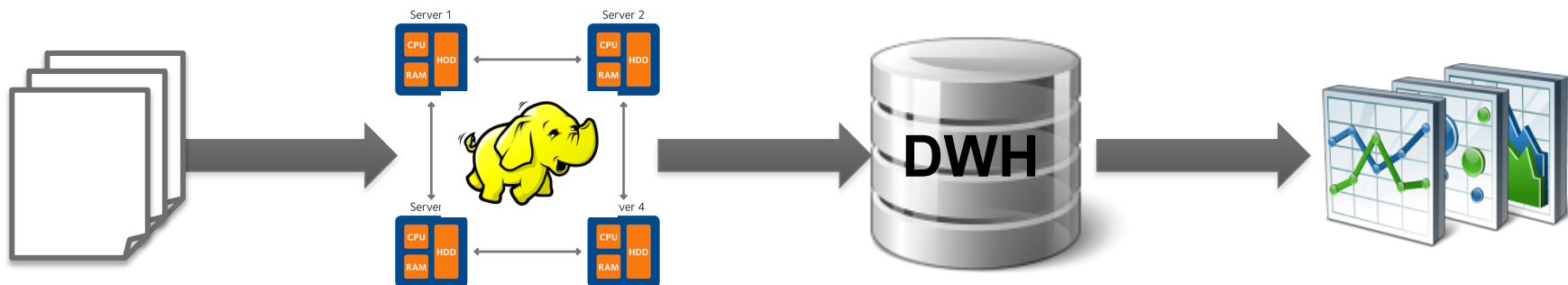
Viele Rollen – unterschiedliche Schwerpunkte



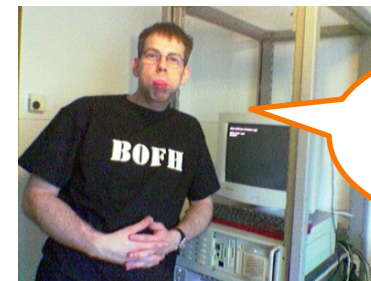
Daten-
Validierung



Fachliche
Tests



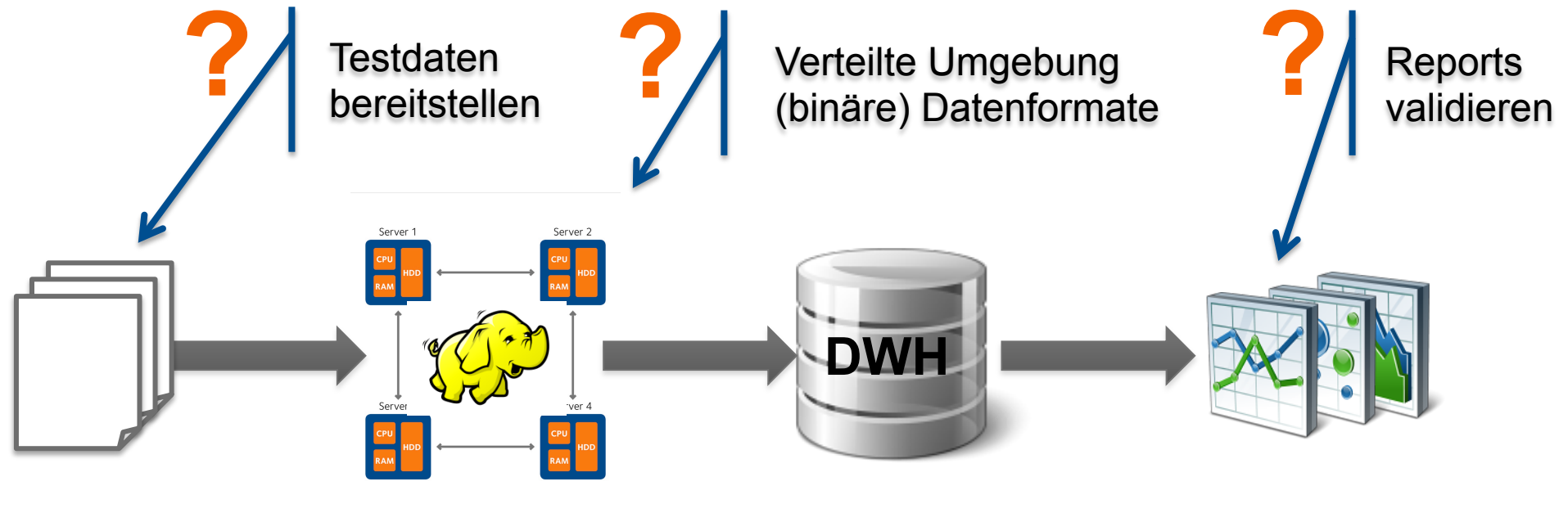
ETL-/
MR-Jobs



Workflows
Deploy-
ments

Agiles Testen von BI und Big Data Lösungen

Herausforderungen von „Test Driven BI“



Modultests



E2E-Tests



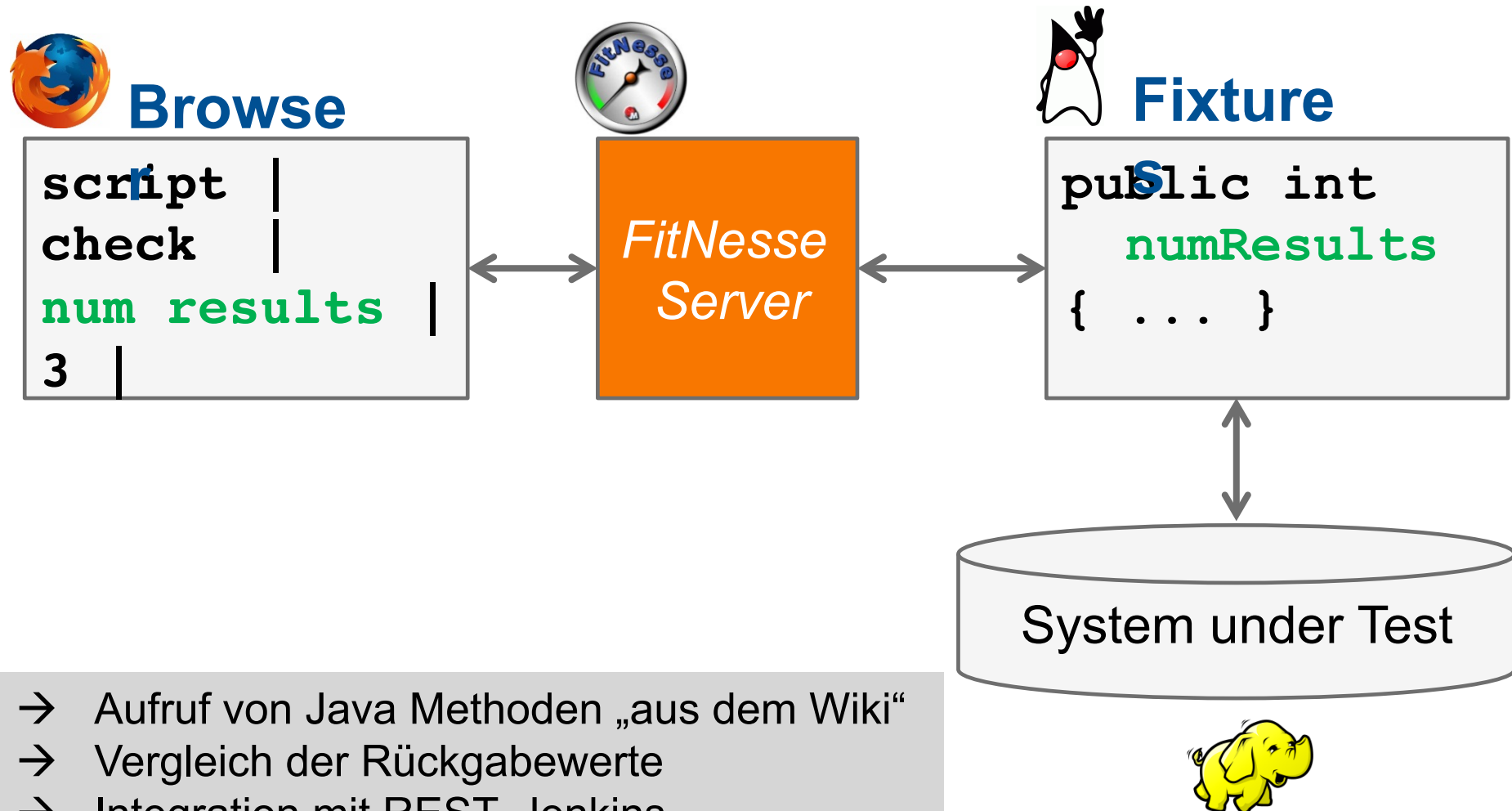
- ▶ „ausführbare“ **Wiki**-Seiten
(Test Ergebnisse als Rückgabewerte)
- ▶ (fast) **natürlichsprachliche**
Test Spezifikation
- ▶ Verbindung zum SUT via
(Java-)“**Fixtures**“



*„fully integrated
standalone wiki and
acceptance testing
framework“*

Agiles Testen von BI und Big Data Lösungen

FitNesse Architektur



- Aufruf von Java Methoden „aus dem Wiki“
- Vergleich der Rückgabewerte
- Integration mit REST, Jenkins...

‣ Synthetische Daten

- Definition in FitNesse-Wiki
- Quellen (Files, DBs, ...)
- Formate (CSV, Thrift, Avro, ...)

‣ Reale Daten

- Zufällige oder verzerrte Stichproben, z. B. eines Tages

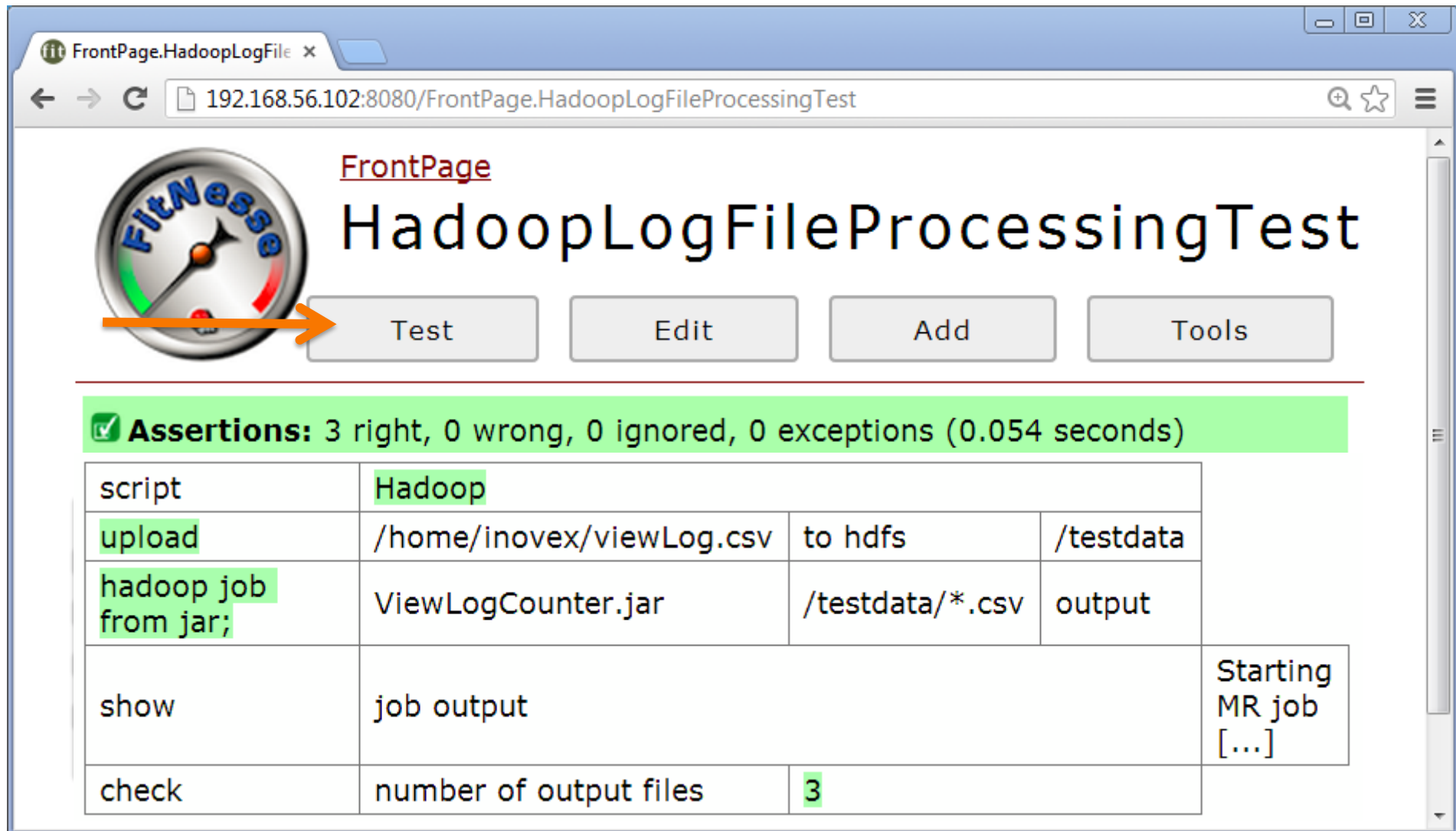
‣ Feedback Schleife:

- Spezial- bzw. Fehlerfälle aus den realen Daten identifizieren
- In synthetische Testdatenmenge aufnehmen
- Inkrementelle Erhöhung der Testabdeckung

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

Agiles Testen von BI und Big Data Lösungen

Ein exemplarischer FitNesse Test



FrontPage

HadoopLogFileProcessingTest

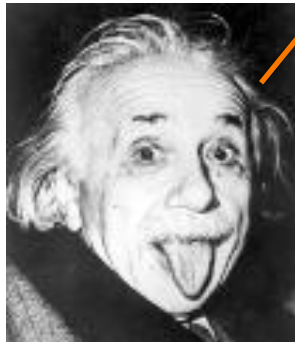
Test Edit Add Tools

✓ **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	

Agiles Testen von BI und Big Data Lösungen

FitNesse: Erweiterbar durch Fixtures



Pig-Fixture
DBSlim



Kettle-Fixture



Hadoop-Fixture
Kettle-Fixture



Shell-Fixture
Git pull

Agiles Testen von BI und Big Data Lösungen

Continuous Integration mit FitNesse



- ▶ Tests können zu Test-Suites zusammengefasst werden
- ▶ Nächtliche Ausführung einplanen
- ▶ Integrationssystem aktualisieren
 - ▶ Git pull
 - ▶ DDLs (*.sql) mit DB-Deploy ausführen
 - ▶ Pentaho Files kopieren
 - ▶ ETL (*.kjb, *.ktr) Files
 - ▶ Cubes (*.mondrian)
 - ▶ Reports (*.prpt, *.analyzer)
- ▶ Ggf. Testdaten aktualisieren
- ▶ Tests ausführen

FrontPage

10.10.0.54:8085

Reader

FrontPage

Test

Edit

Add

Tools

Test Structure

Expand All

Collapse All

Summary of the nightly test results

	View as Overview	Purge > 30 days	Purge > 7 days	Purge all	Cancel																
Page	Pass	Fail	Latest	Last 20 Results																	
FrontPage.CallsLoad.DailyCallsLoad	49	3	18 Feb 14, 20:10	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.AllBookingsCube	38	9	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.AllCassPotentialCube	39	8	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.IntrepCallStatistics	30	7	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.IntrepInvoiceReport	33	17	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.IntrepMultiUser	20	5	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.IntrepOperationalPerformance	25	27	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.IntrepQualityReport	39	8	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.IntrepSalesConceptReport	28	28	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.IntrepTaskReport	36	16	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.IntrepVisitReport	35	10	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckCubes.WfsCampaignActivitiesCube	42	5	18 Feb 14, 20:20	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckReports.BookingsTrendOverview	41	6	18 Feb 14, 20:15	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckReports.CallStatisticsPerDay	30	12	18 Feb 14, 20:19	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckReports.CassPotentialReport	41	6	18 Feb 14, 20:15	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckReports.ChannelTrendOverview	40	7	18 Feb 14, 20:15	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckReports.ProductTrendOverview	40	7	18 Feb 14, 20:15	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckReports.ShortTermMarketingOutbound	39	8	18 Feb 14, 20:19	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.CheckReports.UpSelling	37	11	18 Feb 14, 20:19	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.DwhInconsistencies.BreakAllReferences	38	0	18 Feb 14, 20:21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
FrontPage.DwhInconsistencies.CheckDwhConsistency	45	32	18 Feb 14, 20:21	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.DwhInconsistencies.FixDwhConsistency	36	2	18 Feb 14, 20:21	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.InitialLoad.DatabaseChecks	54	5	18 Feb 14, 20:07	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.InitialLoad.RunInitialLoad	60	1	18 Feb 14, 20:01	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.LcagLoad.DailyLcagLoad	47	6	18 Feb 14, 20:15	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
FrontPage.LcagLoad.DatabaseChecksAfterDaily	40	12	18 Feb 14, 20:19	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	

- ▶ BI-Projekte haben typischerweise viele Schnittstellen
- ▶ Mittlerweile oft parallele Entwicklung von Quellsystem und BI-System
- ▶ Basis für Entwicklung und Synthetische Testdaten
- ▶ Inhalt
 - ▶ Gesamtkontext
 - ▶ Fachliche Spezifikation
 - ▶ IT-Technische Spezifikation
 - ▶ Betriebskonzept (OSLA)
- ▶ inkrementelle Fortschreibung

3	INTERFACE DESCRIPTION	4
3.1	FUNCTIONAL INTERFACE DESCRIPTION	4
3.1.1	Content	4
3.1.2	Interface criticality	1
3.1.3	IT Security	1
3.1.4	Source System(s)	1
3.1.5	Expected Data Volume	1
3.1.6	Delivery cycle	1
3.1.7	Data Types	6
3.1.8	Record definition	6
3.2	TECHNICAL INTERFACE DEFINITION	9
3.2.1	Interface Name	9
3.2.2	Data Transfer	9
3.2.3	DataFormat	9
3.2.4	Process description	10
4	OPERATIONAL SERVICE LEVEL AGREEMENT	11
4.1	SERVICE TIMES	11
4.1.1	Application Management	11
4.1.2	Interface Availability	11
4.1.3	Exceptions	11
4.1.4	Planned Downtimes	11
4.2	NECESSARY SUPPORT FROM INTERFACE PARTNERS	12
4.2.1	Applications And Infrastructure	12
4.2.2	TTR and TTF	12
4.3	COMMUNICATION CHANNELS	12
4.3.1	Incident Management	12
4.3.2	Escalation and Communication Matrix	12
4.3.3	Change Management	13
4.4	RESPONSIBILITIES	13
4.4.1	Provider Responsibilities	13
4.4.2	Consumer Responsibilities	13
4.5	IT SECURITY	13
4.5.1	Provider Responsibilities	13
4.5.2	Consumer Responsibilities	13
5	ACCEPTANCE OF INTERFACE CONVENTION	13
6	OPEN ISSUES	13

- ▶ Dokumentation so nah wie möglich am Artefakt
 - ▶ DB-Kommentare
 - ▶ Kommentare in Kettle
- ▶ Dokumentation generieren
 - ▶ Kettle-Step für ETL-Dokumentation
- ▶ Kettle-Jobs/Transformationen
 - ▶ Systemtabellen auslesen
 - ▶ Pentaho XML-Files (*.kjb, *.ktr, *.mondrian, *.prpt, *.analyzer) auswerten
 - ▶ Schnittstellen-Dok auswerten

Object: cdb.xml

Type/Owner: CUBE/-

Object	Type	References
cdb.xml	CUBE	D_AIRCRAFT
cdb.xml	CUBE	D_CASS_WEIGHTBREAK
cdb.xml	CUBE	MV_CASS_UNSERVICED
MV_CASS_UNSERVICED	MATERIALIZED VIEW	D_CASS_MARKET
MV_CASS_UNSERVICED	MATERIALIZED VIEW	MV_CASS_UNSERVICED
MV_CASS_UNSERVICED	MATERIALIZED VIEW	F_CASS

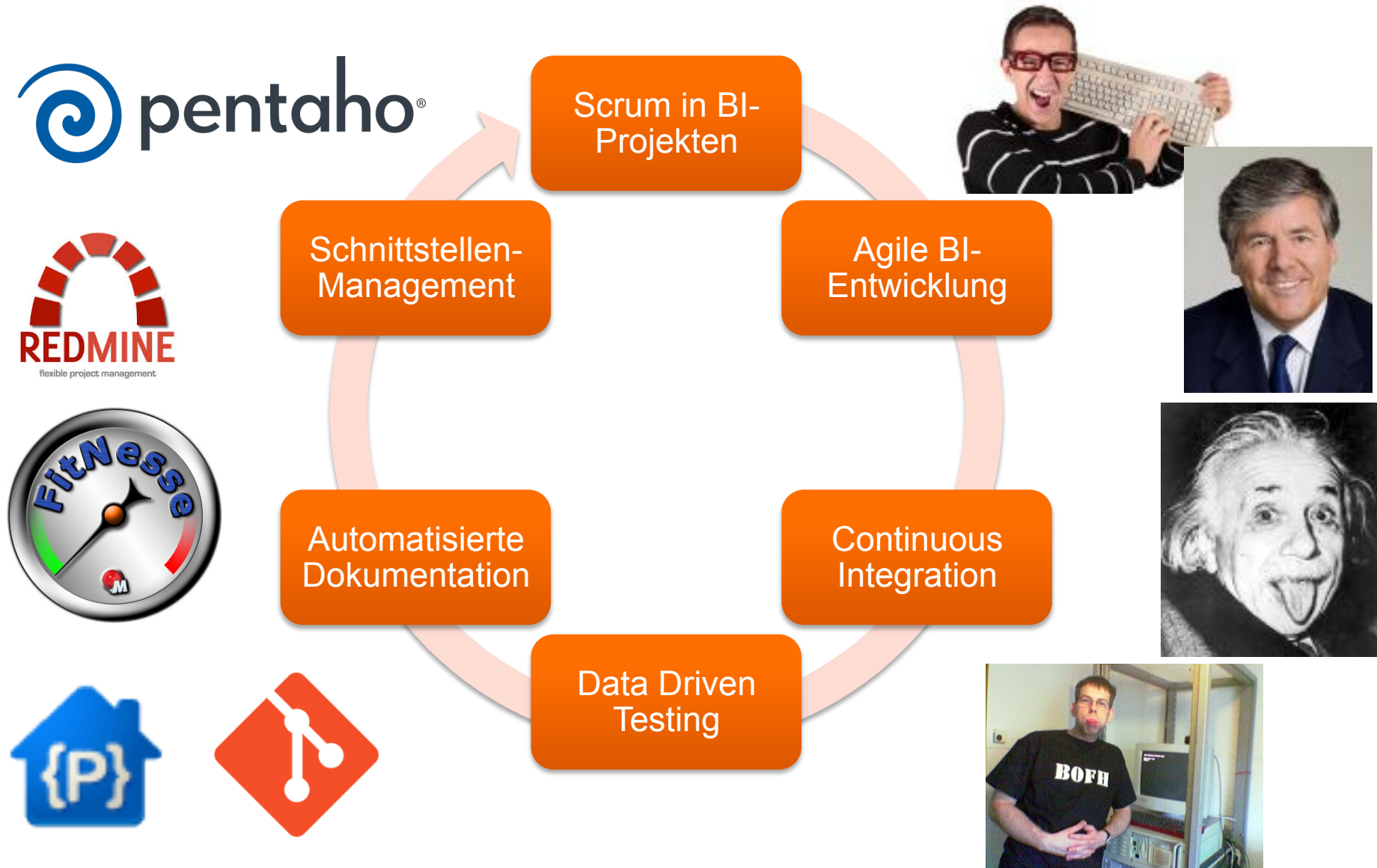
Object: D_CASS_MARKET

Type/Owner:

Referenced by (1st level)	Type	Owner
MV_ACCOUNT_ACTIVITY	MATERIALIZED VIEW	CDB_REPORT
MV_CASS_REWRITE	MATERIALIZED VIEW	CDB_REPORT
MV_CASS_UNSERVICED	MATERIALIZED VIEW	CDB_REPORT
MV_STM_CASS	MATERIALIZED VIEW	CDB_REPORT
V_CASS	VIEW	CDB_REPORT
V_DWH_CHECKS	VIEW	CDB_REPORT
cdb.xml	CUBE	-

Agilität in BI Projekten

Resume



Durchstarten bei inovex

Wir suchen dich

- für das Schreiben deiner Abschlussarbeit (Bachelor oder Master)
- für eine Werkstudententätigkeit
- für verschiedene Einstiegspositionen nach dem Studium

ab sofort in den Bereichen

- Consulting
- Entwicklung (Web/Mobile)
- Business Intelligence
- IT Engineering & Operations.

Wir freuen wir uns über deine Bewerbung
per Mail an jobs@inovex.de

Deine Ansprechpartnerin ist
Jasmina Weilbeer, Tel. 0721 619 021-50

Mehr über uns: www.inovex.de

inovex auf
Platz 3
der beliebtesten Arbeitgeber
der IT-Branche in Deutschland

kununu 
die jobs, die wir lieben

Team inovex.
Wir nutzen
Technologien,
um unsere Kunden
glücklich zu machen.
Und uns selbst.



Vielen Dank für Ihre Aufmerksamkeit



Kontakt

Dr. Stefan Igel
Head of Big Data Solutions

inovex GmbH
Office Karlsruhe
Ludwig-Erhard-Allee 6
D-76131 Karlsruhe

+49 173 3181030
Stefan.igel@inovex.de

