# Big Data en Testen

## samen in een veranderend speelveld

Testnet 10 april 2014

Paul Rakké

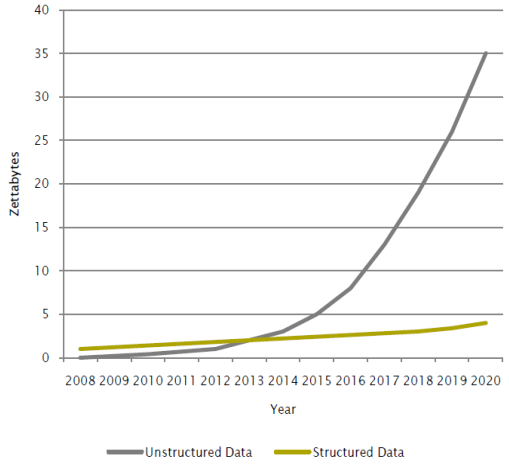
# Kernvraag

Is het testen van Big Data omgevingen, applicaties en de data anders dan het testen van meer traditionele c.q. BI/DWH-omgevingen met bij behorende applicaties en data

# Wat is Big Data

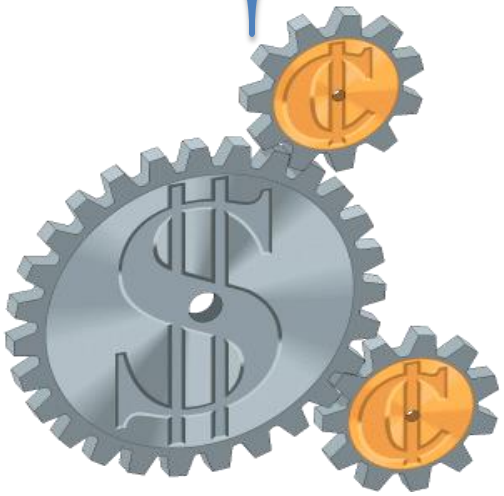
- “Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze (**McKinsey**)
- Big data is the collection of techniques and technologies that produce actionable insight from source data at extremes of scale using commodity resources and massive parallel processing (**Forrester**)
- “Big data” is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making (**Gartner**)

# Big Data: verschil in karakteristieken




volume variety velocity

**value**



# Big Data - wat statements vooraf

- Big Data is meer dan Analytics en Hadoop
- Big Data gaat ook nog steeds over gestructureerde data met:
  - traditionele column-based DWH/BI (maar extra **Volume** !)
  - sensor data en locatie data etc.
  - migratie van gestructureerd naar ongestructureerd (**Variety**)
- maar met sterk veranderde (verwerkings)karakteristieken
  - ETL tools/eisen veranderen
  - veranderde Analytics (known versus unknown)
  - andere eisen aan **Velocity**
  - platforms (b.v. DWH/BI versus Hadoop)
- en verder extra aandacht nodig voor:
  - skills
  - requirements
  - use cases
  - privacy
  - **testen**

# Big Data Testing - veranderingen

- Plaats van het testen in proces:
  - onderdeel van nieuw ‘ETL’-proces
  - requirements en use cases testen en gebruik business rules
  - hypothese testing (A/B testing)

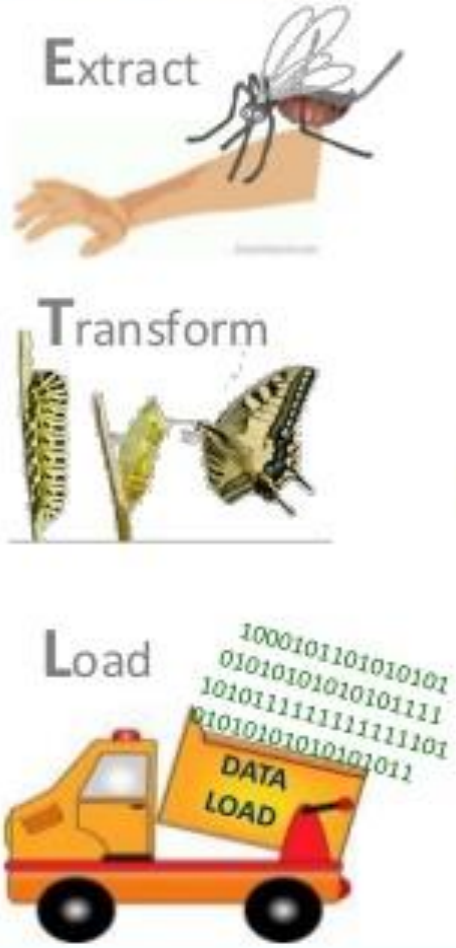
- Andere rol/invulling bij:
  - testen use cases
  - testen requirements
    - functioneel
    - non-functioneel
    - security
    - known versus unknown
  - data kwaliteit (‘traditionele’ data profiling niet echt bruikbaar meer)
  - test data (management)

# Data Warehouse – the ETL process

QuerySurge

## Source Data


## ETL Process




## Target DWH

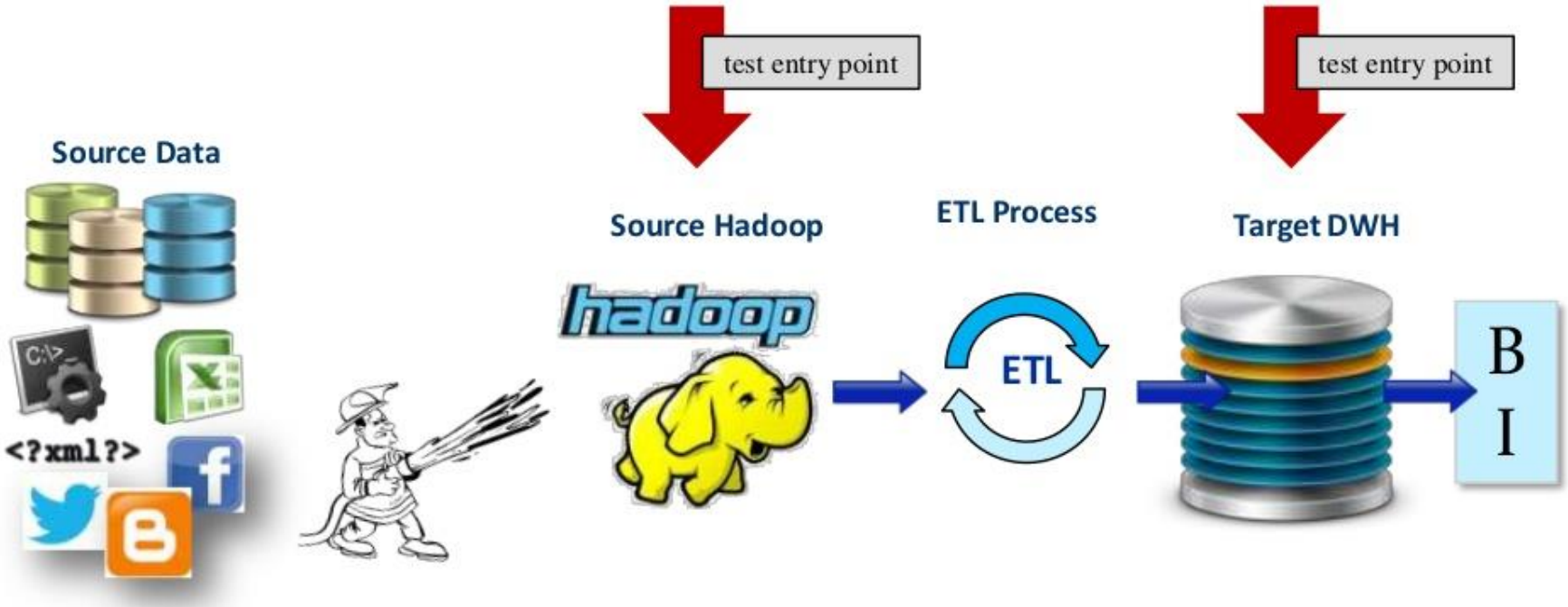
built by RTTS

# Testing Big Data: Entry Points

QuerySurge

**Recommended functional test strategy:** Test every **entry point** in the system (feeds, databases, internal messaging, front-end transactions).

The goal: provide ***rapid localization*** of data issues between points



# Hadoop in verwerkingsproces (voorbeeld)


*Figure 1: Big Data Testing Focus Areas* *Source: Infosys Research*

# Polarion


**Webinar:**
**Big Data Testing - Die Grenzen von Excel Spreadsheets**

**Wo und Wann**

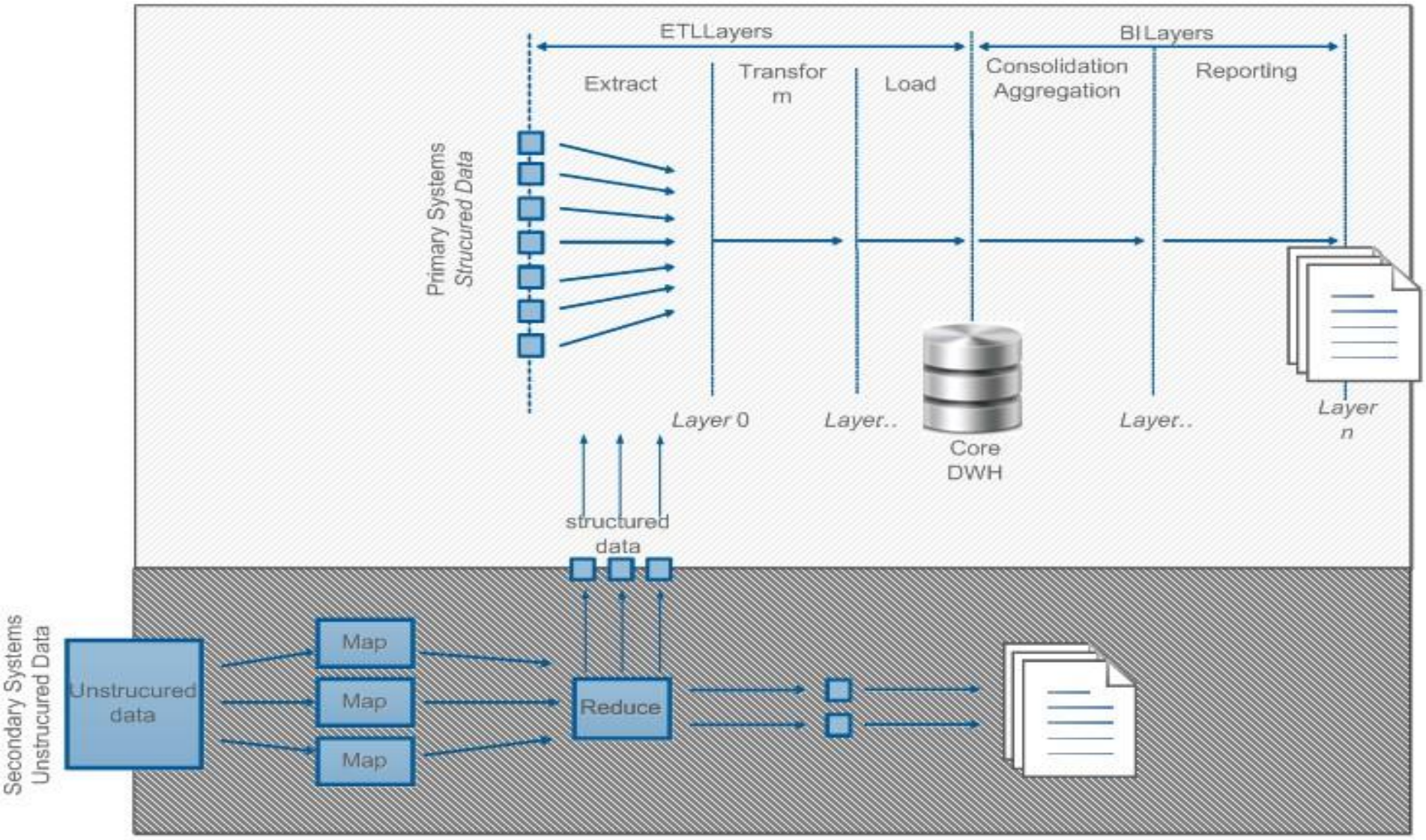
Dienstag, 4. Februar 2014, 11:00 Uhr - 11:45 Uhr.

Webinar auf Ihrem Computer.

**Ihr Sprecher: Stefan Schuck** ist Mitarbeiter des Polarion Software Professional Service Teams. Als Consultant unterstützt er dort Tag für Tag bestehende und potentielle Kunden bei der Lösung Ihrer Probleme in unterschiedlichsten Branchen.

TRICENTIS

# Unstructured Big Data Use



# Scope Big Data & Testing

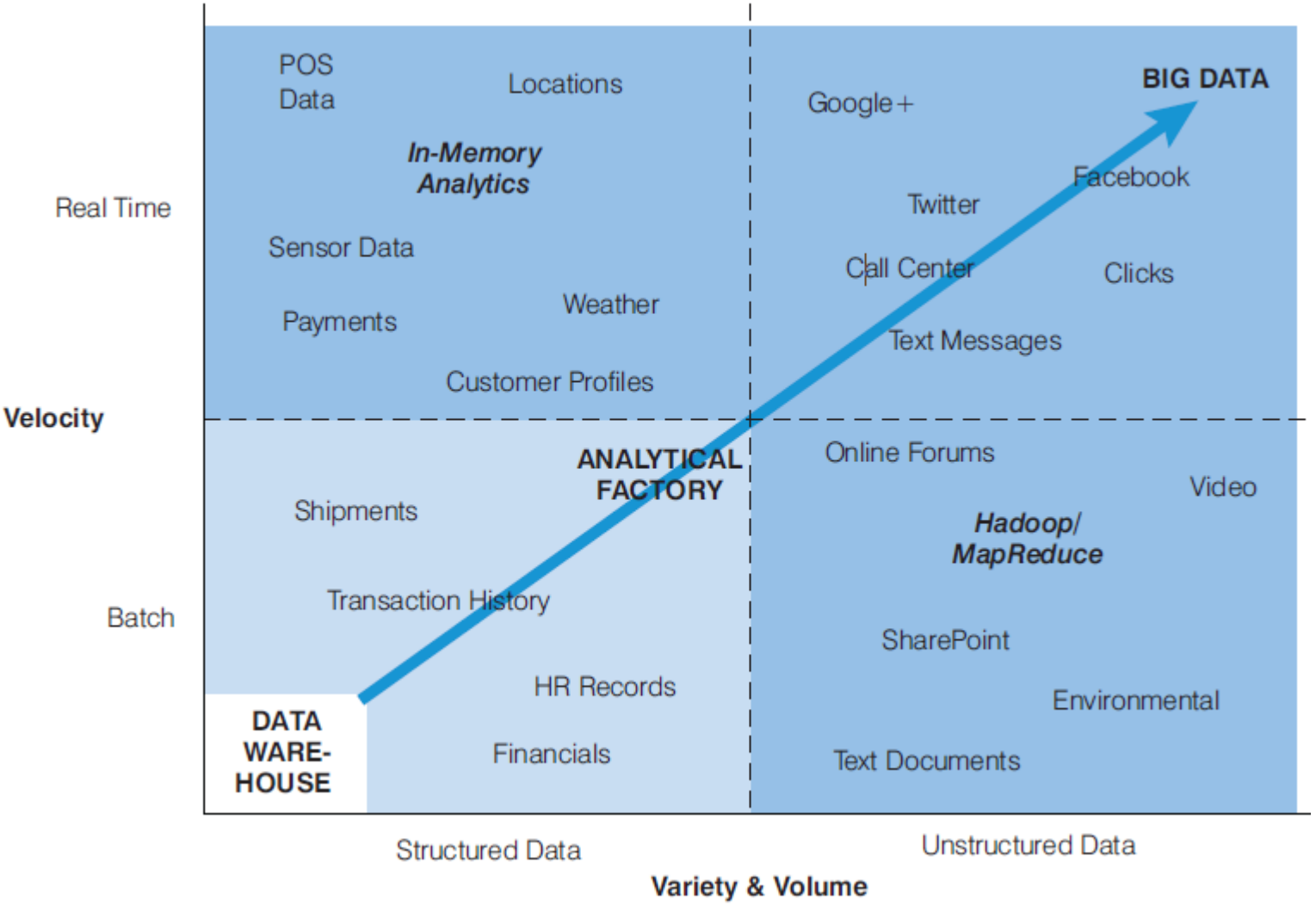
- Test data
  - samenstellen testsets
  - anonimisering (bredere data privacy)
  - subsetting
  - data generatie
- Test Data Management
- Uit te voeren tests
  - standaard lijst test typen
  - extra hypothese testen (A/B testing)
- Domeinen
  - infrastructuur/platform
  - applicaties
  - data
  - business
- Big Data test strategie

INFORMATION INTEGRITY SOLUTIONS

# What is PbD?


- The design and implementation of systems and processes to **respect individual privacy** while **meeting business objectives**, finding greatest expression in an organisation's:
  - Information technology
  - Physical design and networked infrastructure
  - Accountable business practices

# Andere Positionering (bron Booz & Company)



# Verdere positionering

| Nr | **Big Data Approach** | **Traditional BI/DWH Approach** |
|---|---|---|
| 1 | Opportunity oriented | Requirements based |
| 2 | Bottom-up experimentation | Top-down design |
| 3 | Immediate use | Integration and reuse |
| 4 | Tool proliferation | Technology consolidation |
| 5 | "World of Hadoop and NoSQL" | World of data warehouse and BI |
| 6 | Hackathons | Competence centers |
| 7 | **Better business** | **Better decisions** |
| 8 | Marketing/operations | Enterprise focus |
| 9 | Exploratory and informal | Rigorous and formal |
| 10 | **Bringing analysis to data** | **Bringing data to analysis** |
| 11 | **Data model 'on the fly'** | **Fixed data model in early phase** |
| 12 | Key adaptations to governance model | Traditional governance model |

*Source of 1-9: Gartner 2013*

# Issues bij Big Data Testing

- Testen van ongestructureerde data
- Transformatie van ongestructueerd naar gestructureerd (en omgekeerd)
- Aanpassing ETL-processen
- Hadoop en testing
- Velocity van traditionele DWH's versus Big Data omgeving
- De rol van NoSQL en SQL
- Rol van Privacy (profiling)

Merkbaar bij:

- uit te voeren testen door primaire verschuiving naar data
- uitvoering van type testen
- verschuiving van gevraagde skills
- beschikbaarheid van de gevraagde skills

# Onderscheiden rollen

- Big Data developer
- Big Data architect
- Big Data analist
- Big Data administrator
- Big Data project manager
- Big Data designers
- Data scientist

- Waar is de specifieke rol van de Tester ?

# Antwoord op kernvraag

Is het testen van Big Data omgevingen, applicaties en de data anders dan het testen van meer traditionele c.q. BI/DWH-omgevingen met bij behorende applicaties en data en het gebruik van beschikbare tooling

**JA**

Vragen ?

# Dank U

Paul Rakké
06-51282217

10/04/2014