





Testen in de wereld van **Big Data**



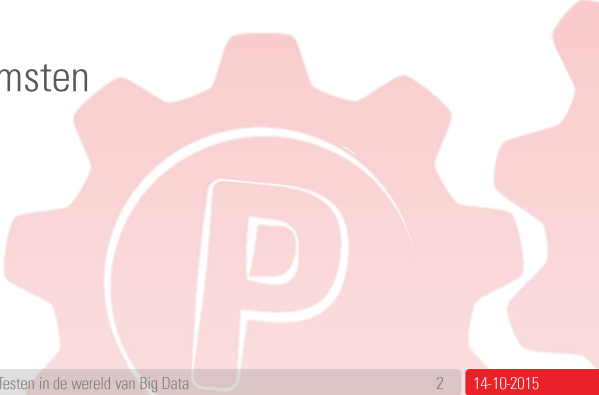
TestNet Najaarsevenement
14-10-2015 – Eibert Dijkgraaf



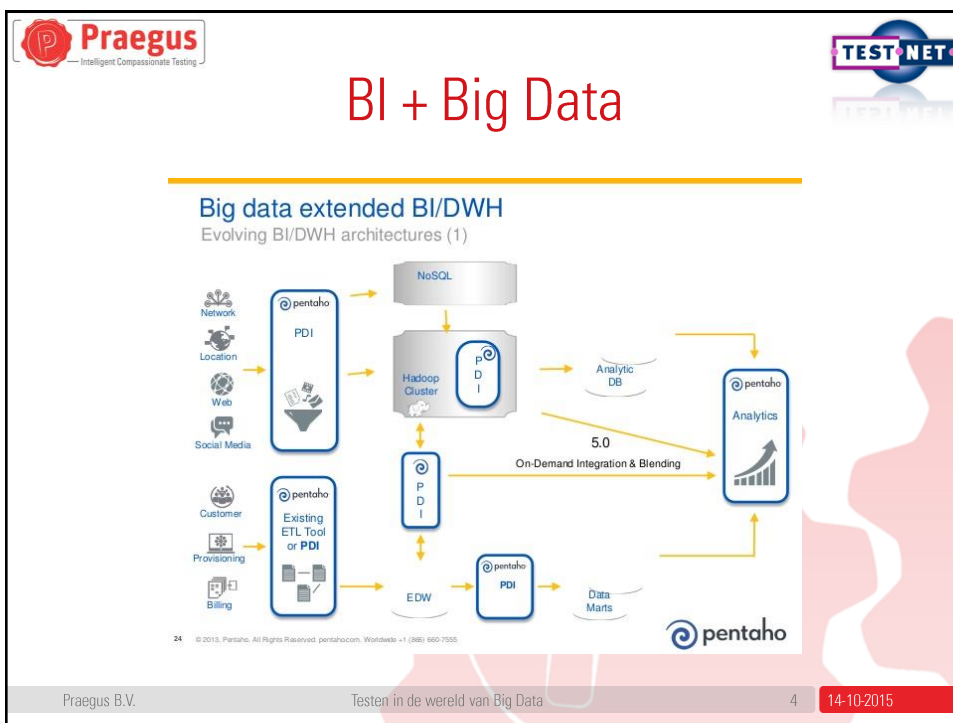
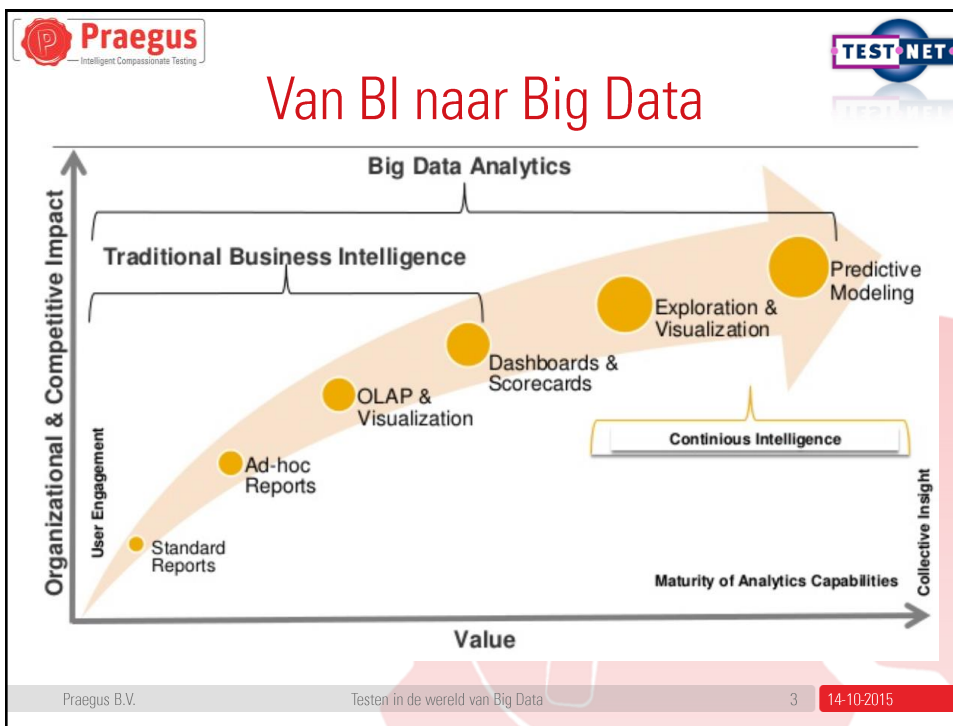
INLEIDING



Komend uit de wereld van:

- Structuur
- Voorspelbaar
- Verwachte uitkomsten
- Subsets
- Anonimiseren
- BI is groot(s)...



Praegus B.V. Testen in de wereld van Big Data 2 14-10-2015



Data groei

September 2015

- (IEEE, Hu et al, 2014)

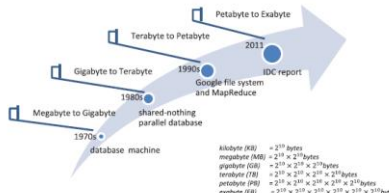


FIGURE 2. A brief history of big data with major milestones. It can be roughly split into four stages according to the data size growth of order, including Megabyte to Gigabyte, Gigabyte to Terabyte, Terabyte to Petabyte, and Petabyte to Exabyte.

Record aan data door Amsterdams internetknooppunt

© GISTEREN, 09:41 ECONOMIE, TECH

Internetknooppunt Amsterdam Internet Exchange (AMS-IX) heeft gisteravond meer dan 4 terabit aan data per seconde verwerkt. Dat is een mijlpaal voor de Amsterdamse organisatie.



Bij AMS-IX zijn meer dan zevenhonderd partijen aangesloten die internetverkeer uitwisselen. De hoeveelheid data die wordt uitgewisseld stijgt ieder jaar opnieuw. Afgelopen jaar alleen al met 30 procent.

Praegus B.V.


Testen in de wereld van Big Data

5


14-10-2015


IDC - IoT




4
BILLION
Connected People




\$4
TRILLION
Revenue Opportunity




25+
MILLION
Apps



25+
BILLION
Embedded and Intelligent Systems



50
TRILLION
GBs of Data





Source: Mario Morales, IDC.

Praegus B.V.






Testen in de wereld van Big Data

6

14-10-2015






Big Data is here

- Forbes, 2015
 - » 89% of business leaders believe Big Data will revolutionize business operations in the same way the Internet did.
 - » 83% have pursued Big Data projects in order to seize a competitive edge.
- InformationWeek's 2015 Analytics & BI Survey
 - » Finding correlations across multiple disparate data sources (clickstream, geospatial, ...) (48%), predicting customer behavior (46%) and predicting product or services sales (40%) are the three factors driving interest in Big Data analytics.
- InformationWeek, 2015
 - » Just two years ago, big data was at this peak of the hype cycle. It was replaced last year by the Internet of Things, a ranking that IoT still holds in this report. Indeed, big data is nowhere to be found on the current Gartner Hype Cycle.

Praegus B.V.
Testen in de wereld van Big Data
7
14-10-2015

Wat is het?

- De 4 V's
 - » Volume
 - » Velocity
 - » Variety
 - » Veracity
- Gestructureerd vs ongestructureerd

We brachten de data naar....

We brengen het nu naar de data...

Laten we alles opslaan, mogelijk kunnen we het nog gebruiken...

Praegus B.V.
Testen in de wereld van Big Data
8
14-10-2015



Praegus
Intelligent Compassionate Testing



Wat kun je ermee? 1

- **Webshop: order – klikgedrag**

eCommerce Example: Web Sales

- Fully Structured
- The Sale Transaction typically carries all fundamental dimensions:
 - Time
 - Customer
 - Referring URL / Search Phrase
 - Product
- Purchase and/or Shipment (Geo or URL) Locations
- Promotion / Campaign
- Etc.
- And "How Many" Measures
- Unit and Price Quantities / Amounts
- Discount Amounts
- Etc

eCommerce Example: Clickstream

- Semi-Structured
- Recording of every page request made by a user
- Includes some structural elements – such as when the request was made and who the user is
- Requires significant prep work in order to fit into a traditional row-based relational database
- Apples and Oranges: Pre-Sessionized Page Visits, Detailed Product Views, Catalogue Requests, Shopping Cart Adds / Deletes / Abandons, etc.
- Needs to be converted into separate-but-reliable dimensional facts - with many shared (conformed) dimensions

```

Raw Clickstream Data
-----
35 52 164 242 274 328 368 448 538 561 636 667 730 775 821
824
39 130 124 000 401 581 704 814 825 834
36 238 674 7 737 739 864 900
39 422 443 924 825 867 895 92 784 968
15 229 262 258 266 262 261 704 738 748 853 963 966 974
36 104 143 320 380 400 739
7 388 218 362 538 664 667 762 859 848 833 947 970 979
327 380
71 160 204 272 279 280 300 333 494 539 530 587 616 674 671
720 85 8 14 932
163 163 27 274 276 277 37 4 74 433 436 512 529 426 653 1
878 939
161 175 177 424 482 571 59 7423 746 756 863 910 960
126 130 137 608 609 839
380 461 638 641 852
27 78 104 17 733 779 781 841 902 921 938
102 147 239 263 411 468 523 726 827 826 276 303 84 2 929
71 208 217 246 279 290 454 479 52 14 14 768 853 888 884 944
42 70 176 204 227 234 268 480 510 703 728 833 874 885
25 32 278 730
161 423 634 868 880
71 73 178 27 4 3 10 327 388 419 449 468 484 706 722 786 810
844 848 91
130 274 4 10 288 367
188 507 230 261 483 525 526 727 774 788 789 834 903 975
89 116 184 201 333 356 603 720 846
10 171 287 298 483 538 541 623 87 1 101 825 848 964
143 162 17 271 487 623 634 645 676 736 780 863 884 936
17 242 41 176 263 637 896
52 146 161 263 375 385 678 721 731 730 759 888
98 209 276 629
84 84 84 84 84
                    
```

- **Marketing etc**

Facebook kijkt via 'Like'-knop mee met je webgedrag

🕒 GISTEREN, 11:27 IN TECH

De Facebook-knoppen die op websites te vinden zijn, worden vanaf volgende maand ook gebruikt om gerichtere advertenties te leveren.



Praegus B.V.

Testen in de wereld van Big Data

9

14-10-2015



Praegus
Intelligent Compassionate Testing



Wat kun je ermee? 2

- **Volle maan**

Week by week Google Search volume for "full moon"

Event ● Actual week of full moon ● Searches for full moon





Praegus B.V.

Testen in de wereld van Big Data

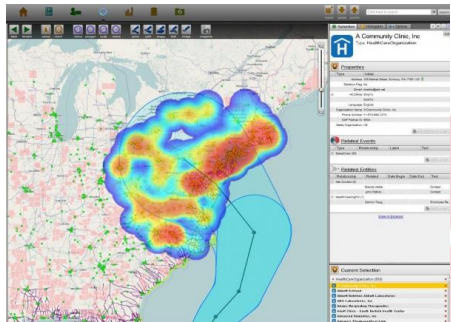
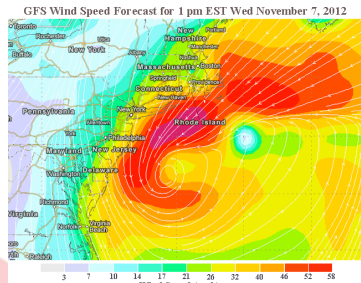
10


14-10-2015



Wat kun je ermee? 3

- Hurricane Sandy
 (<https://hbr.org/2013/04/the-hidden-biases-in-big-data/>)



Praegus B.V.
Testen in de wereld van Big Data
11
14-10-2015

Big Data Value Chain

Big Data Value Chain

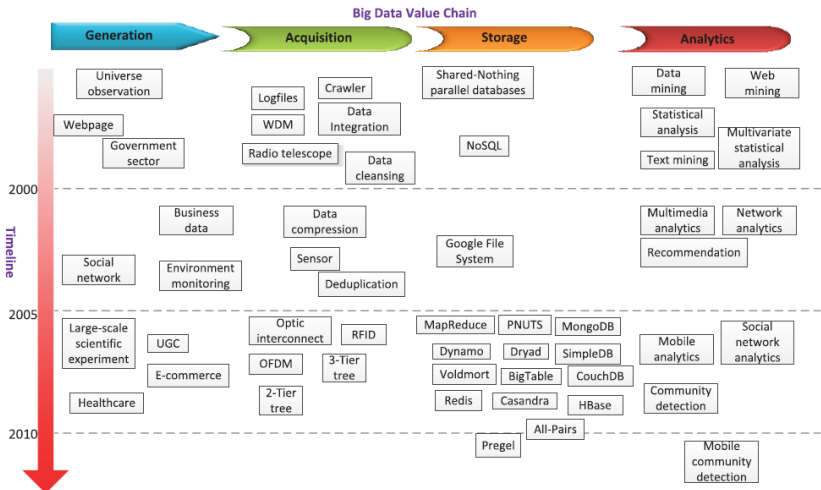
Generation

Acquisition

Storage



Analytics

Timeline



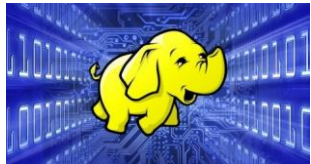
The diagram shows a vertical timeline from 2000 to 2010. A red arrow points downwards. Technologies are placed in boxes corresponding to the four stages of the value chain. For example, in the 2000-2005 period, 'Webpage' and 'Government sector' are in the Generation stage, while 'Statistical analysis' and 'Text mining' are in the Analytics stage. By 2010, technologies like 'Pregel' and 'All-Pairs' are in the Storage stage, and 'Mobile community detection' is in the Analytics stage.

FIGURE 3. Big data technology map. It pivots on two axes, i.e., data value chain and timeline. The data value chain divides the data lifecycle into four stages, including data generation, data acquisition, data storage, and data analytics. In each stage, we highlight exemplary technologies over the past 10 years.

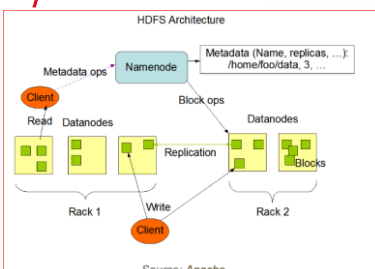



Het Hadoop Ecosysteem

- Apache™ Hadoop®: open source software raamwerk
- HDFS: Hadoop Distributed File System



HDFS Architecture



Source: Apache

- Gedistribueerd / replicatie
- Schaalbaar
- Fouttolerant / betrouwbaar
- Gestructureerde en ongestructureerde data

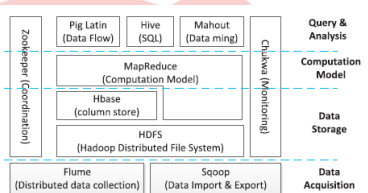




FIGURE 14. A hierarchical architecture of Hadoop core software library, covering the main function of big data value chain, including data import, data storage and data processing.

Praegus B.V.

Testen in de wereld van Big Data


13

14-10-2015

Hadoop Distributions

- HortonWorks, 2011



GOVERNANCE INTEGRATION

Data Lifecycle & Governance

Falcon
Atlas

Data Workflow

Sqoop
Flume
Kafka
NFS
WebHDFS

DATA ACCESS

Batch MapReduce	Script Pig	SQL Hive	NoSQL HBase Accumulo Phoenix	Stream Storm	Search Solr	In-Mem Spark	Others... ISV Engines
Tez	Tez	Tez	Shade	Shade	HDP	HDP	HDP

YARN: Data Operating System

HDFS Hadoop Distributed File System

DATA MANAGEMENT

SECURITY

Administration
Authentication
Authorization Auditing
Data Protection

Ranger
Knox
Atlas
HDFS Encryption

OPERATIONS



Provisioning, Managing, & Monitoring

Ambari
Cloudbreak
ZooKeeper

Scheduling

Oozie

- Cloudera, 2008
- MapR, 2011






Praegus B.V.

Testen in de wereld van Big Data

14

14-10-2015

Een overzicht (momentopname)

- **Hadoop Common:** The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.

Other Hadoop-related projects at Apache include:

- **Ambari™:** A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually along with features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™:** A data serialization system.
- **Cassandra™:** A scalable multi-master database with no single points of failure.
- **Chukwa™:** A data collection system for managing large distributed systems.
- **HBase™:** A scalable, distributed database that supports structured data storage for large tables.
- **Hive™:** A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™:** A Scalable machine learning and data mining library.
- **Pig™:** A high-level data-flow language and execution framework for parallel computation.
- **Spark™:** A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Tez™:** A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive™, Pig™ and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop™ MapReduce as the underlying execution engine.
- **ZooKeeper™:** A high-performance coordination service for distributed applications.

Zoek op
"Hadoop Ecosystem Table"

Praegus B.V.
Testen in de wereld van Big Data
15
14-10-2015




Een voorbeeld 1


ElasticSearch: full text search and analytics engine


The screenshot below is from a **social media monitoring application** which uses *ElasticSearch* not only to search and mine the data, but also to provide data aggregation for the interactive dashboard.



Bron:
www.elastic.co

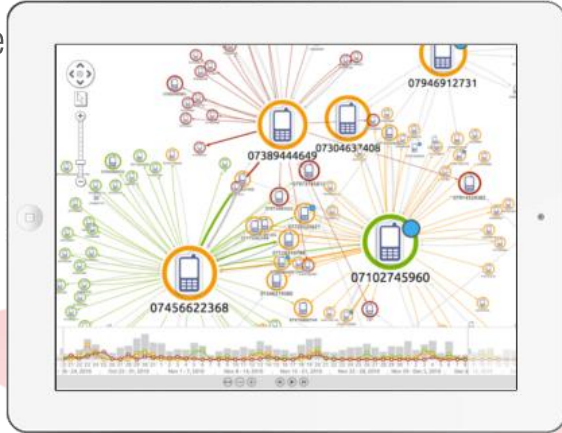
Praegus B.V.
Testen in de wereld van Big Data
16
14-10-2015





Een voorbeeld 2

Keylines:
Netwerk visualisatie



Praegus B.V.
Testen in de wereld van Big Data
17
14-10-2015





pffffffff

\ Sport

Nooit eerder won Oranje op woensdag in het jaar van de draak na een priemgetal aan doelpunten

Door Richard van der Toren en Jochem van den Berg • Tuesday 12 June 2012

Nederland speelt morgen voor de eerste keer in de geschiedenis op een woensdag in de tweede poulewedstrijd tegen een land waarvan alle elf spelers in het basisteam een andere voorletter hebben, tijdens een EK waarvan 1 van de gastlanden twee dagen eerder van een Scandinavisch land wist te winnen.

. . .



Fans zien het rooskleuriger in. Ben Elker (32), data-analist te Bilthoven, licht toe: "Het Nederlands elftal maakte tijdens de kwalificatie 37 doelpunten, een priemgetal. Nu spelen we tegen Duitsland, die in diezelfde kwalificatie zeven doelpunten tegen kreeg: Ook een priemgetal. En wanneer gebeurde dat de vorige keer? Precies, in '88."

Bron:
www.despeld.nl



Relatief veel tov eerdere EK's

Praegus B.V.
Testen in de wereld van Big Data
18
14-10-2015

En dan ... Testen ?!

- De 4 V's
 - » Volume
 - » Velocity
 - » Variety
 - » Veracity

We brachten de data naar....

We brengen het nu naar de data...



- Gestructureerd vs ongestructureerd

Praegus B.V.

Testen in de wereld van Big Data

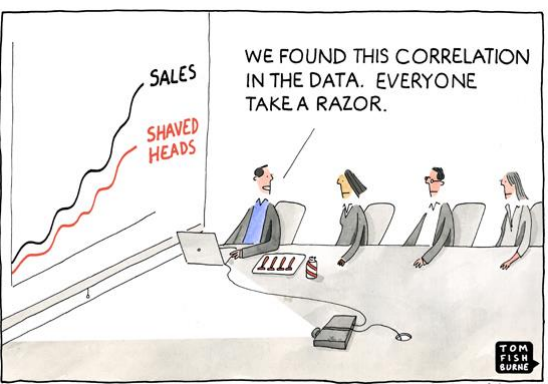
19

14-10-2015

Enkele steekwoorden

- Technology
- Agile
- Testdata management
- Data Governance
- Metadata
- Design Principles
- Data Science
- Predictive modelling
- Causale verbanden
- Machine Learning
- Ethics ➔
- Internet-of-Things ➔
- ... ➔



© marketoonist.com

Praegus B.V.

Testen in de wereld van Big Data

20


14-10-2015

Praegus — Intelligent Compassionate Testing

TEST NET

Kwaliteit bij Big Data

- Kwaliteit? Testen?
 - » Functionaliteit
 - » Performance
 - » Scalability
 - » Security
 - » Auditability
 - » Actualiteit
 - » Ethiek; privacy
 - » Betrouwbaarheid



Praegus B.V. Testen in de wereld van Big Data 21 14-10-2015

Praegus — Intelligent Compassionate Testing

TEST NET

Behoefte aan houvast

Verschillende onderdelen:



- de infrastructuur
- het laden van de data
- Data Analytics
- Data Virtualization
-

- Generation (source)
 - Acquisition
 - Storage
 - Analytics

Houd rekening met de 4 V's:

- Volume
- Velocity
- Variety
- Veracity

Praegus B.V. Testen in de wereld van Big Data 22 14-10-2015





Behoeftte aan houvast

Verschillende onderdelen:

- **de infrastructuur**
- het laden van de data
- Data Analytics
- Data Virtualization
-

- Basisfunctionaliteit van de infrastructuur.
- Scalability
- Performance
- Security
- Gebruik partities
- Inrichten OTAP

Praegus B.V. Testen in de wereld van Big Data 23 14-10-2015





Behoeftte aan houvast

Verschillende onderdelen:

- de infrastructuur
- **het laden van de data**
- Data Analytics
- Data Virtualization
-

- Validatie
 - » Volledigheid
 - » Juistheid
- Performance
- Betrouwbaarheid

Praegus B.V. Testen in de wereld van Big Data 24 14-10-2015

Behoefte aan houvast



Verschillende onderdelen:

- de infrastructuur
- het laden van de data
- **Data Analytics**
- Data Virtualization
-

- Functionaliteit?
- Performance?
- Betrouwbaarheid

gebruikers is laag

Praegus B.V. Testen in de wereld van Big Data 25 14-10-2015

Behoefte aan houvast



Verschillende onderdelen:

- de infrastructuur
- het laden van de data
- Data Analytics
- **Data Virtualization**
-

- Functionaliteit
standaard?
- Performance
- Security
- Bruikbaarheid
- Gebruiksvriendelijkheid
- Betrouwbaarheid

gebruikers is hoog

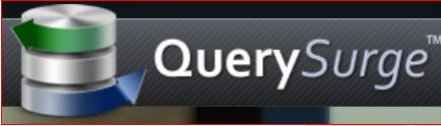

Praegus B.V. Testen in de wereld van Big Data 26 14-10-2015



Testtooling

Tools gespecialiseerd in:

- het omgaan met Hadoop ecosysteem,
- gericht op datakwaliteit,
- en klaar voor Volume/Variety/Velocity

Praegus B.V.
Testen in de wereld van Big Data
27
14-10-2015

Nog wat cijfers

- IT Executive (9-2015; bron KPMG)
 - » 60% worstelt met kwaliteit en betrouwbaarheid data
 - » > 40% heeft te weinig ervaring met Analytics
- Forrester (8-2015): 67% van de NL en Be. bedrijven heeft geen goed proces om de datakwaliteit te borgen.
- Gartner (2015):

DID YOU KNOW?

Through 2015, 85% of Fortune 500 organizations will be unable to exploit big data for competitive advantage.
- Big Data Alliantie: In 2018 heeft NL 10.000 data scientists nodig, maar kan er maar 2000 leveren.

Praegus B.V.
Testen in de wereld van Big Data
28
14-10-2015

Praegus — Intelligent Compassionate Testing

TEST NET

CONCLUSIES – Big Data...

- ...kent veel mogelijkheden
- ...staat nog in de kinderschoenen
- ...is onvermijdelijk
- ...kent weinig ervaring
- ...is niet ongrijpbaar
- ...vraagt ook om inzicht in kwaliteit / betrouwbaarheid
- ...definieer een passende aanpak
- ...en blijf kritisch in het geweld van de 4 V's

Praegus B.V. Testen in de wereld van Big Data 29 14-10-2015

Praegus — Intelligent Compassionate Testing

TEST NET

Overwegingen – BIG DATA

- Snel veranderende wereld.
- Weinig structuur – wel grip.
- Focus!

Gaan wij het verschil maken?

Een tester is geen toolboer of pseudo-data-scientist

TIME WELL SPENT™ by Tom Fishburne

TELL ME AGAIN HOW THE BRAVE KNIGHT CONQUERED BIG DATA AND RESCUED REAL-TIME BUSINESS INSIGHTS.

KRONOS © 2013 Workforce Innovation That Works™ KRONOS.COM/TIMEWELLSPENT

Praegus B.V. Testen in de wereld van Big Data 30 14-10-2015



Eibert Dijkgraaf
Testadviseur / Testmanager

T +31(0)6 43 57 30 23
E eibert.dijkgraaf@praegus.nl

Transistorstraat 31
1322 CK Almere
+31 (0) 36 751 9748

Prins Willem-Alexanderlaan 705
7311 ST Apeldoorn
info@praegus.nl

Praegus B.V.
KVK 59482745