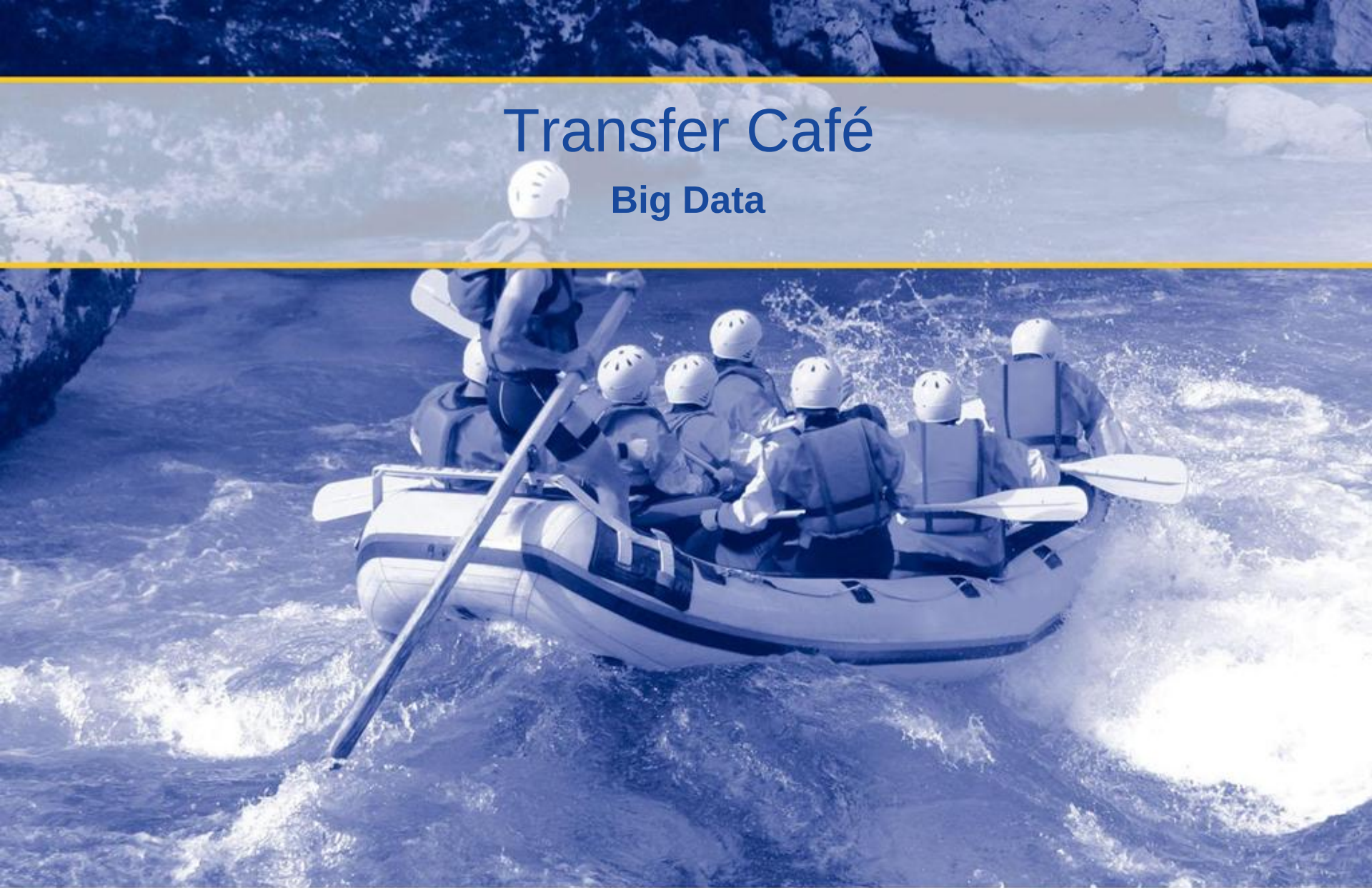


Transfer Café

Big Data



Spreker(s) : René La Haye
Datum : 14 april 2015
E-mail : rhaye@transfer-solutions.com

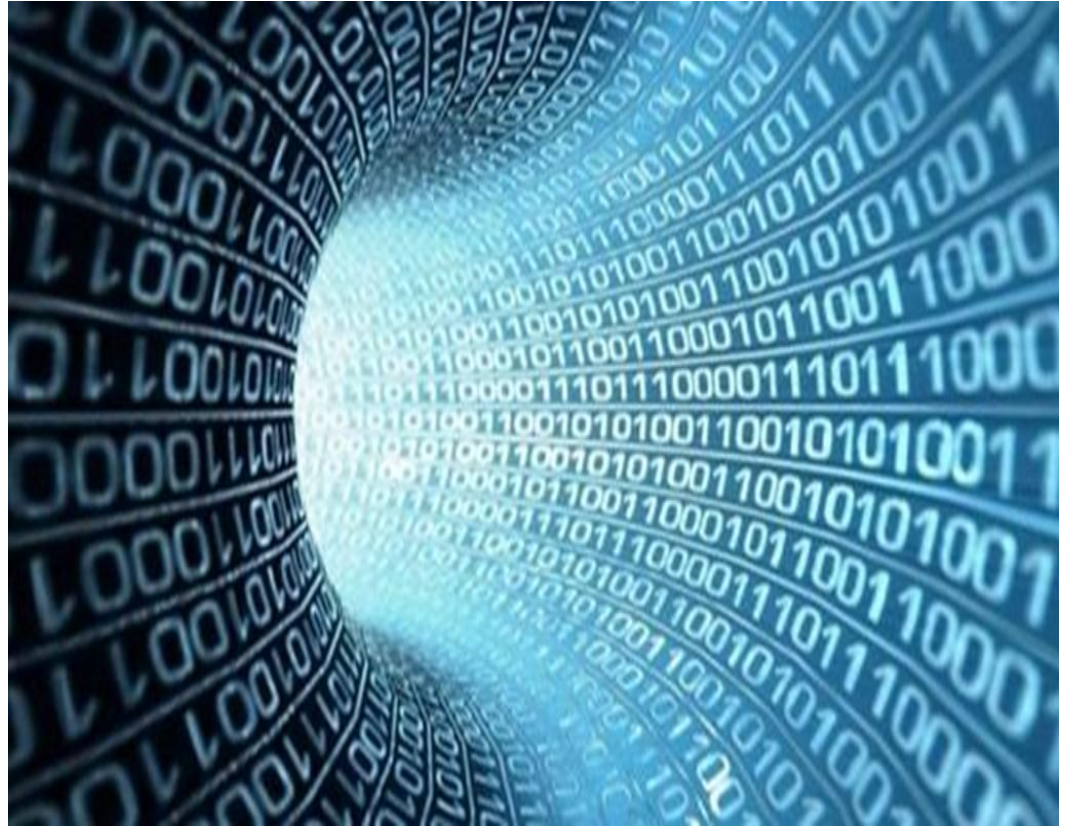
TRANSFER

WWW.TRANSFER-SOLUTIONS.COM

Big Data

Agenda

- Definitie(s)
- Business Cases
- Instrumenten
- Methoden
- Oracle Big Data Appliance



Big Data Timeline

- 1991: www
- 1995: Java (middle-tier, recording traffic) / GPS
- 1998: NoSQL / Google
- 1999: Term Internet of Things
- 2001: Wikipedia
- 2002: Bluetooth v1.1
- 2003: LinkedIn
- 2004: Facebook
- 2005: Hadoop
- 2007: iPhone
- 2008: Devices op het Internet > Wereldbevolking
- 2011: IPv4 adressen zijn gebruikt
- 2012: 2.8 Zetabytes
- 2013: Democratization van Data

Big Data

Definitie - 1

Data volume

Petabytes
Exabytes
Zettabytes
Yottabytes

90% of total data in the world was created in the last 2 years

Walmart's hourly transaction data **67x** the books in the Library of Congress

Would take **1,000 years** to watch all the videos on YouTube

Data variety



New forms of content coming online like **GPS data, augmented reality, speech to text, QR codes etc**

Plus infinite varieties of user content, eg. **posts, pics, videos, links etc**

Data velocity



Visa has cut processing time to spot credit card fraud from **1 month to 13 min**

Amazon's 'cloud' holds >260B objects & handles **200K requests / sec**

Data veracity

1 in 3 Business Leaders don't trust the information they use to make a decision

Poor data quality costs the US economy around \$3.1 Trillion a Year

Data value



\$200M EBIT opportunity at an international energy co around 'Smart Home' & cross sell based on new data

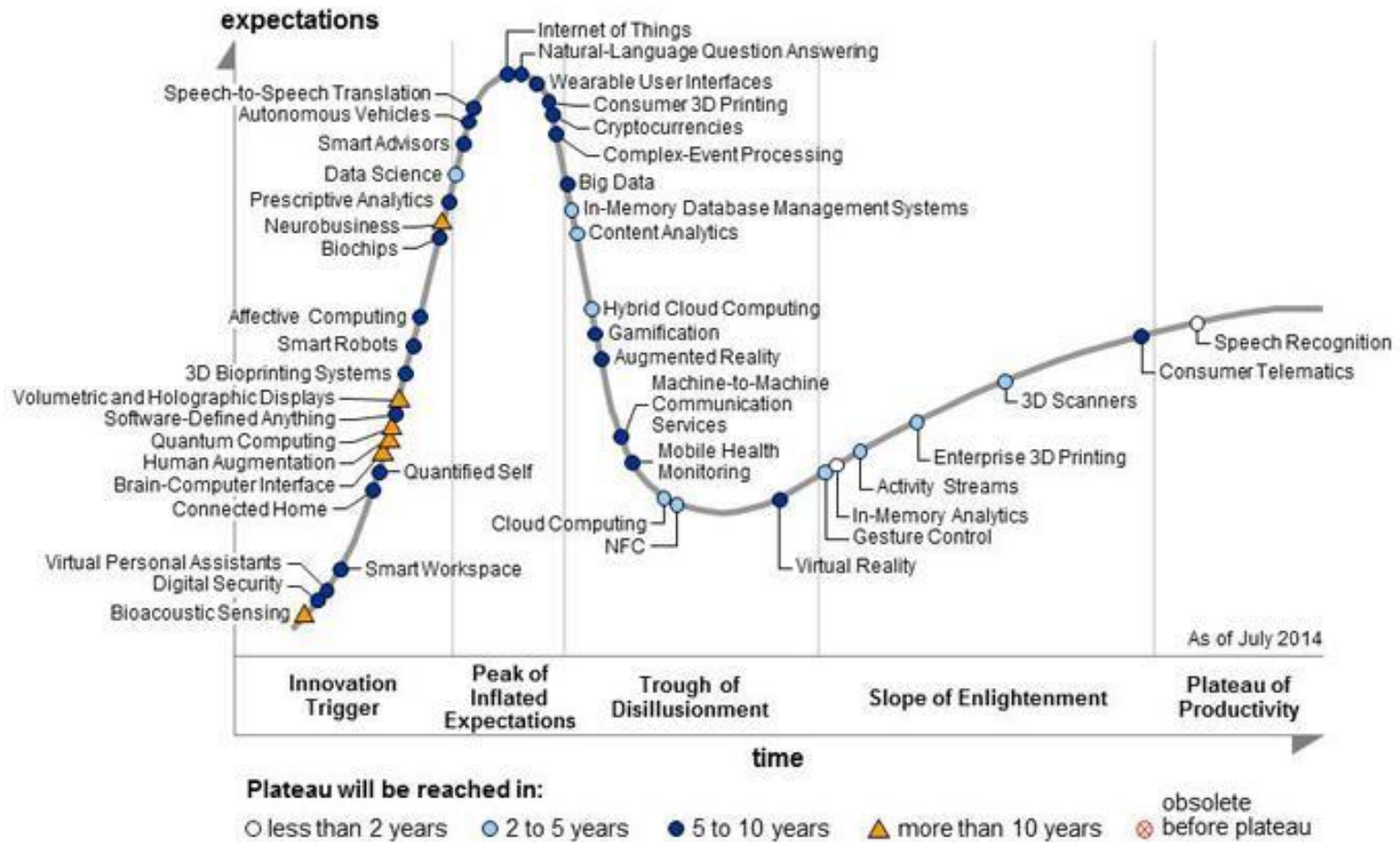
Sales up **7-12%**, page views **25%** when Shopzilla dropped speed of loading website pages from seven seconds to two

Cost of sequencing DNA has fallen by a factor of 1,000x as speed has increased by 1,000,000x

Big Data is when data itself becomes part of the problem.

TRANSFER

Big Data Hype?



Gartner Hype Cycle Aug-2014

It is a very hot topic today...

...as measured by mentions in News Headlines

Google

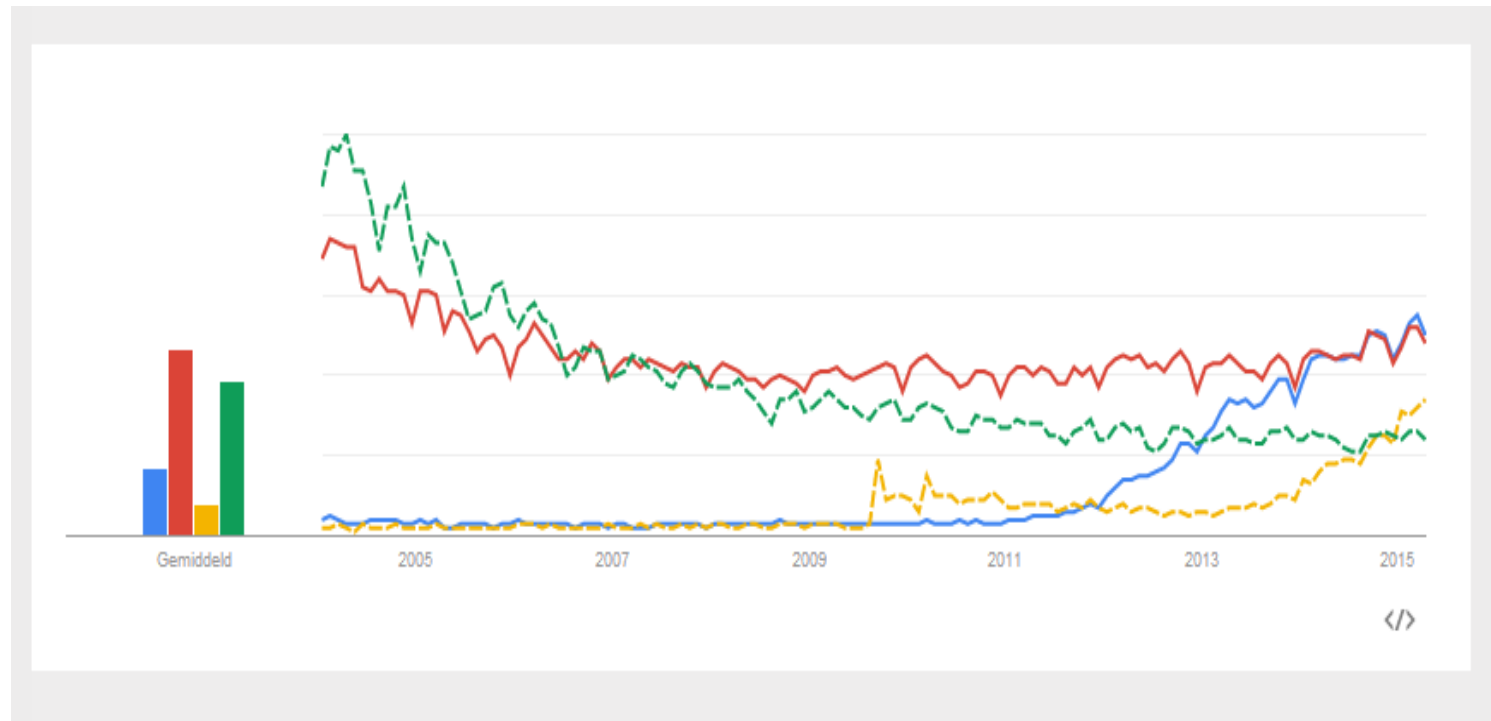
Trends

Big data

eCommerce

Data Mining

Internet of Things



Big Data

Definitie - 2

An all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications.

The ability of society to harness information in novel ways to produce useful insights or goods and services of significant value” and “...things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value.

Big Data

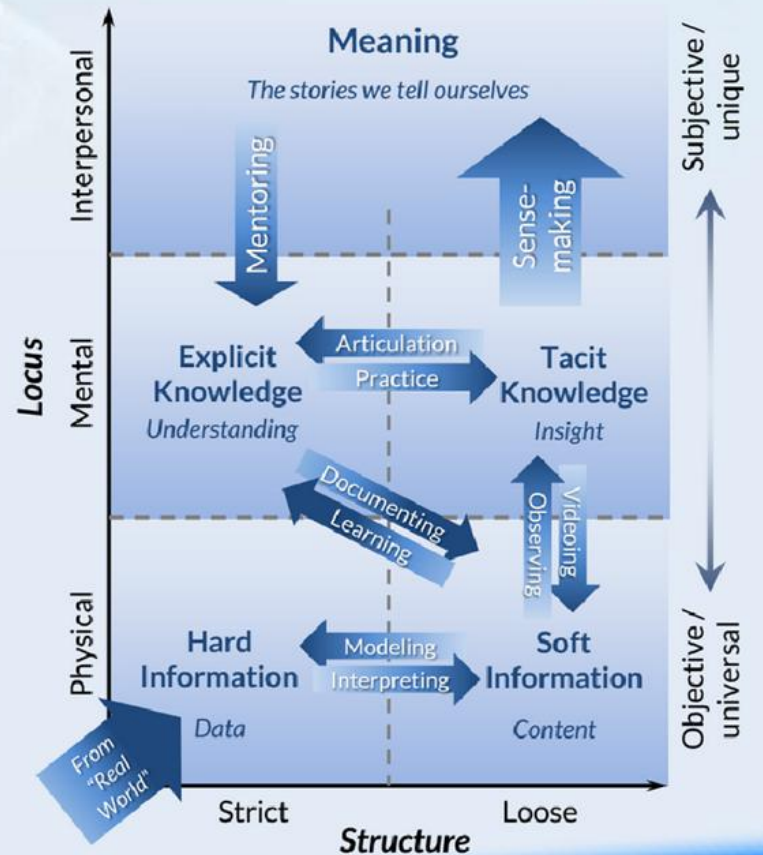
Om over na te denken

M³: THE MODERN MEANING MODEL

- Ackoff's DIKW pyramid is no longer viable



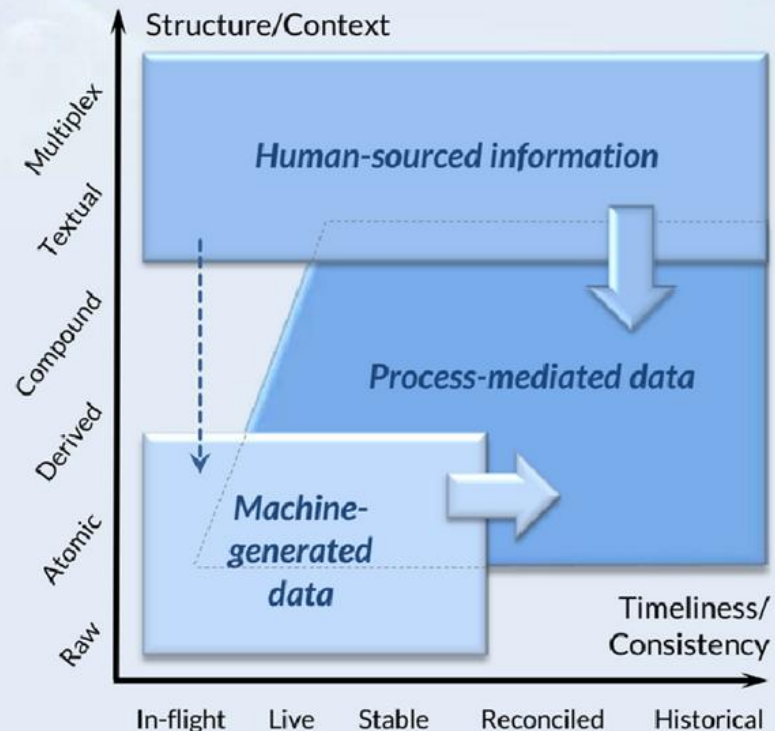
- Information precedes data
 - Data is simply information optimized for computers
 - The Web has fully devalued "facts"
 - People process information



Big Data de bronnen

THE TRI-DOMAIN INFORMATION MODEL

- **Process-mediated data**
 - “Traditional” operational & informational data
 - Via data entry and cleansing processes
- **Machine-generated data**
 - Output of machines and sensors
 - The Internet of Things
- **Human-sourced information**
 - Subjectively interpreted record of personal experiences
 - From Tweets to Videos



[In the context of these domains, “data” signifies well-structured and/or modeled and “information” is more loosely structured and human-centric.]

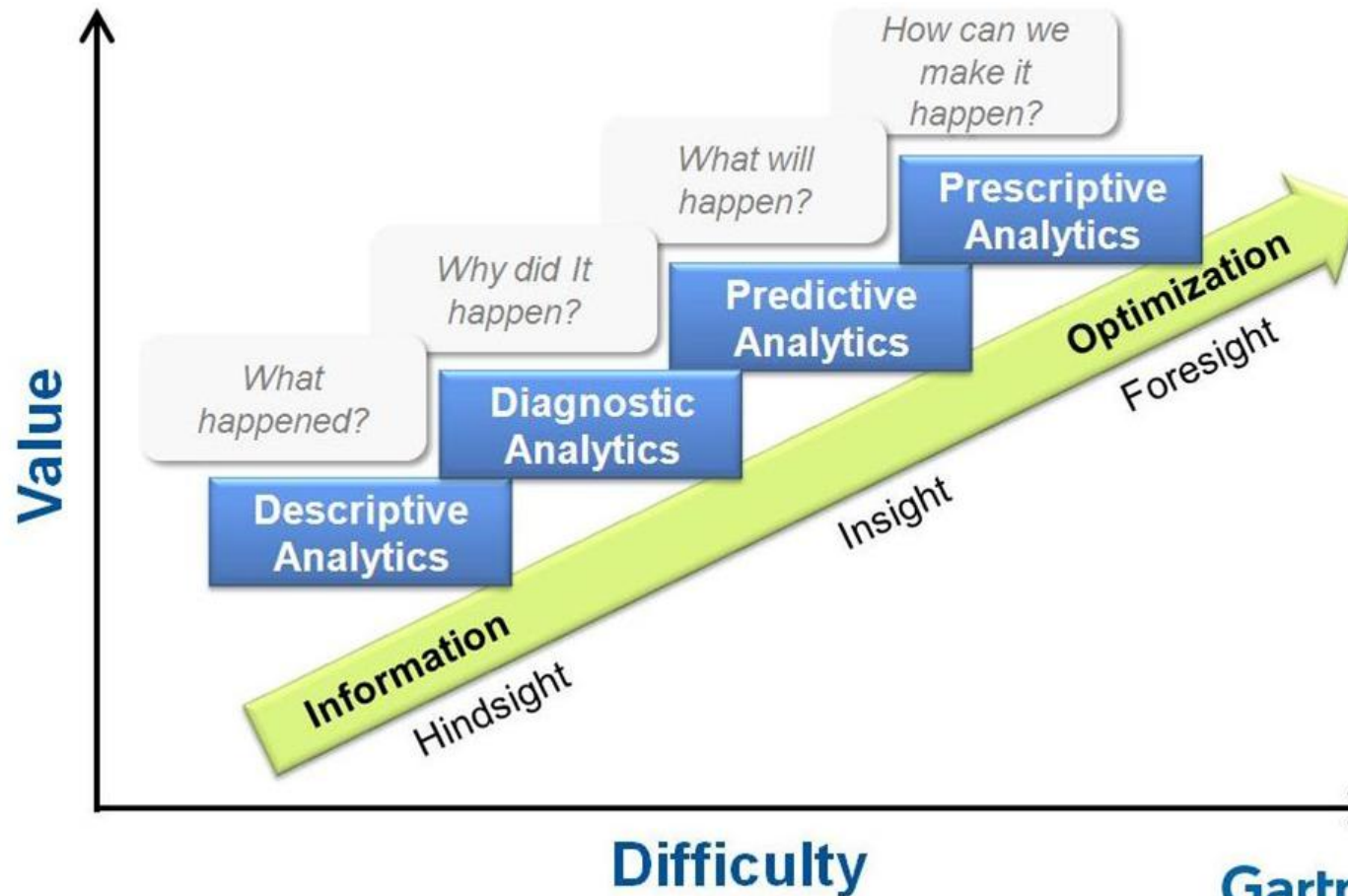
Big Data & Data Mining

- Ontdekken van relaties/patronen/modellen en daarmee kennis
- Machine Learning
 - op basis van data (data en output) wordt een model gevonden, meer data levert een beter model en daarmee is predictive analytics mogelijk (bv. handschrift herkenning, foto herkenning)



- Representatieve sample set of... de hele set
- Big Data: het vinden van uitzonderingen

Business Intelligence

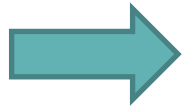


Big Data

Business Cases

Klantsegmentaties

- Eigen klantinformatie aangevuld met
- Process: Klantgedrag op websites (klikken en aankopen)
- Human Sources information: Klanteigenschappen op sociale media



Gerichte productaanbiedingen

Dynamic Pricing

- Inspelen op Capaciteit en Vraag en Aanbod
- Maar ook op klantgegevens ?



Winstoptimalisatie

Big Data

Business Cases

Fraude detectie en risico analyse

- Klantinformatie aangevuld
- Afwijkend gedrag wordt opgemerkt (bv. Locatie)
- Fraudegevoelige omstandigheden worden ontdekt -> voeding voor de risico analyse
- Gaming

Voorspellingen

- Verkoop van producten
- Voorkomt out-of-stock situaties

Service

- Tekst analyse op Commentaar op sites
- Reageren op Social Media (Twitter, Facebook)

Big Data

Business Cases

Gezondheidszorg

- Voorspellingen van epidemieën
- Effectiviteit van behandelingen en medicijnen
 - Verzamelen van patientengedrag
 - Gebruik van DNA
- Verkorten van wachttijden
- Specialisaties van ziekenhuizen

A **20% decrease in patient mortality** can be achieved by analyzing patient data

A couple of weeks **before there is an increase in flu patients** coming to hospital emergency rooms in a region, **there is a spike in Google™ search requests for terms like “flu symptoms” and “flu treatments”**

Big Data

Business Cases

Internet of Things

- Locatie informatie en Navigatie optimalisatie
- Service op machines
 - Vb.: Windmolens, Auto's
- Klimaatbeheersing

Combinatie Klant en Machine data

- Slimme thermostaat (gecombineerd met navigatie)

Uitstapje naar

- Foto herkenningsoftware

Big Data

Business Cases

Hypothese:

*Weinig data met een goed model levert slechtere resultaten
dan veel data met een slecht model .*

Voorbeeld: Google Translate

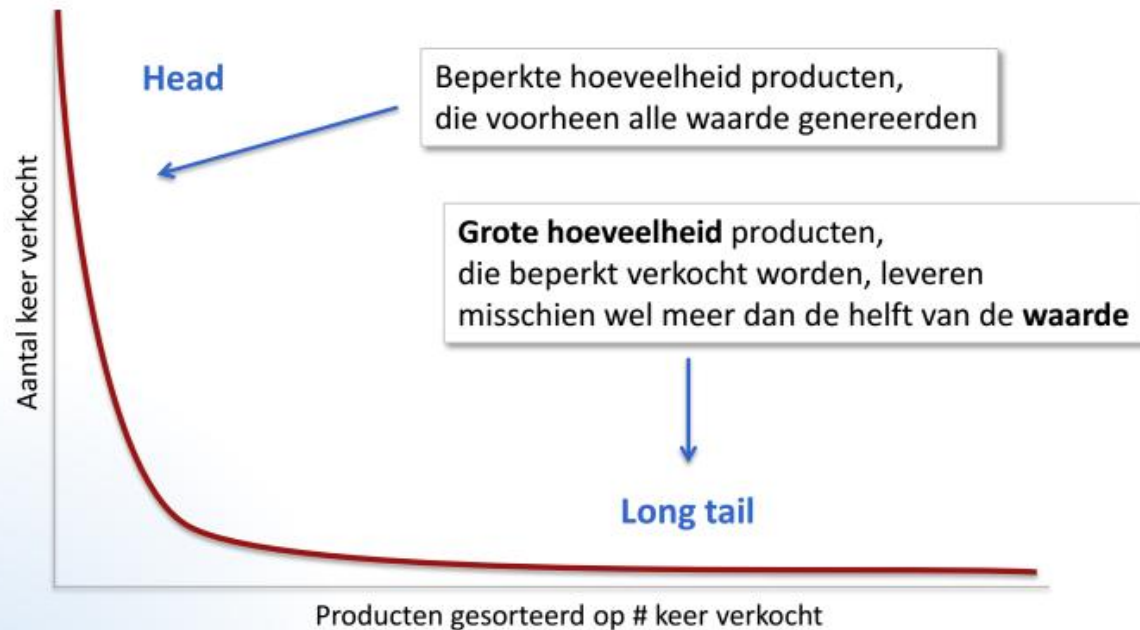
Ingewikkeld: mapping van grammatica's en woordenboeken

Benadering: Statistische analyses op een enorme bron van vertalingen

Big Data

Business Cases

The Long Tail



Bron: 'The Long Tail', Chris Andersen

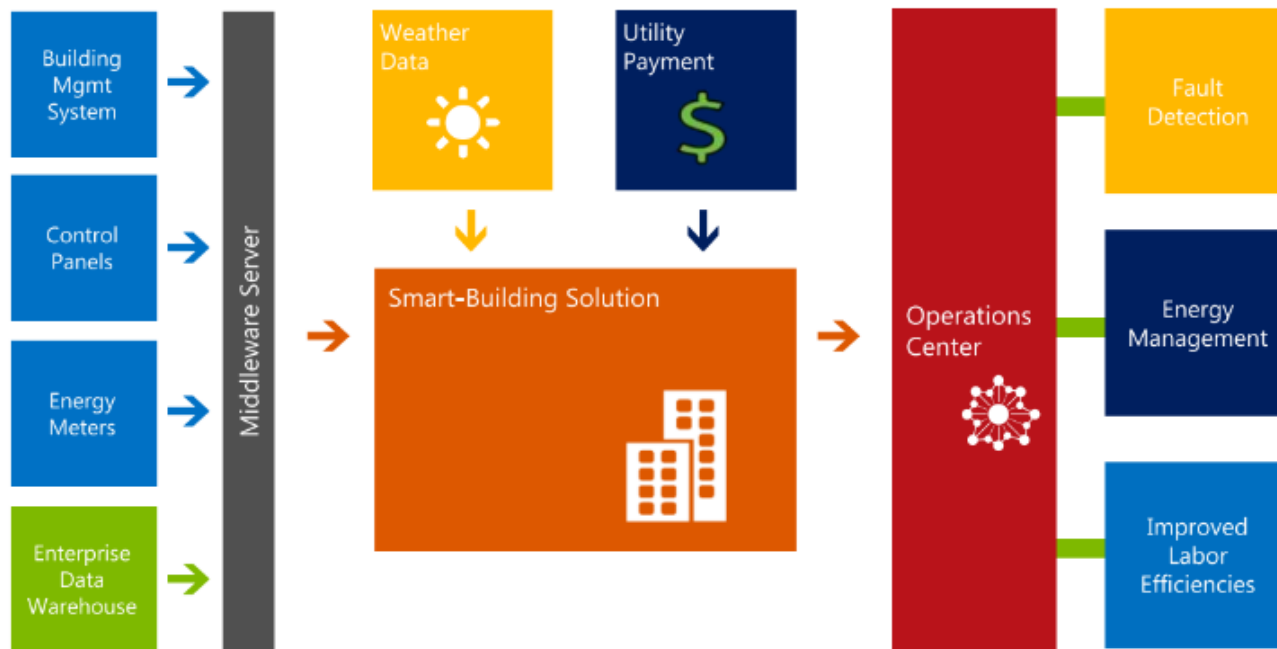
bol.com

TRANSFER

Big Data

Klimaatbeheersing (Microsoft)

Architecture



Big Data

Dynamic Pricing

- Ethisch verantwoord?
- Gebaseerd op correcte data
- Gebruik van de juiste algoritmes
- Wijzigen van klantgedrag
- Gebruik van cookies



Bij reizen veel toegepast

- Capaciteit is beperkt (stoelen, kamers)
- Mensen willen soms betalen voor comfort (tijdstip, stoelkeuzes)
- Doel: hotels en vluchten zoveel mogelijk vol met uiteindelijk de hoogste omzet

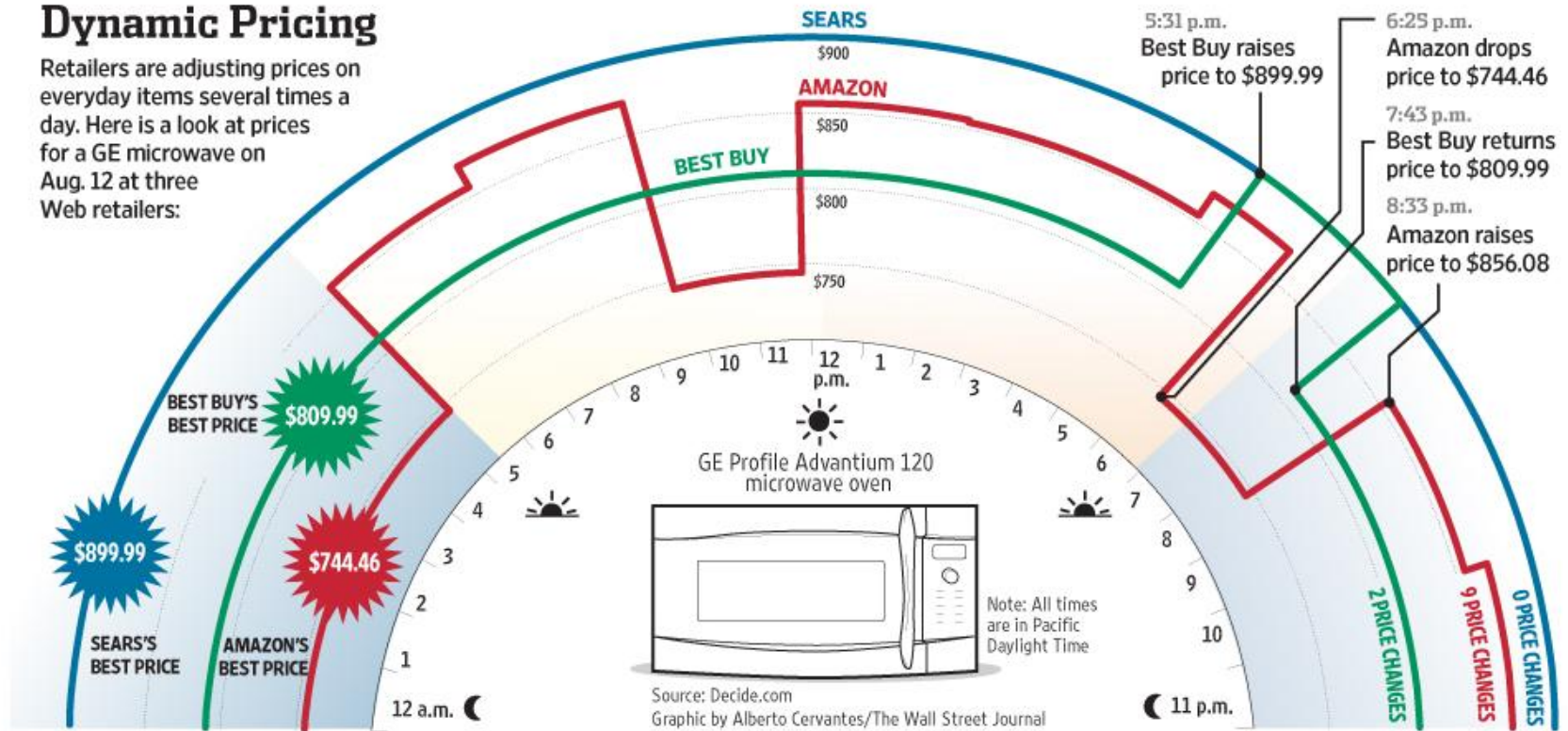
In their white paper, Dynamic Pricing in the Airline Industry (California Institute of Technology) cite one estimate that American Airlines "changes half a million prices per day." Even if that number is off by a factor of 10, that's 50,000 prices changes a day.

Big Data

Dynamic Pricing

Dynamic Pricing

Retailers are adjusting prices on everyday items several times a day. Here is a look at prices for a GE microwave on Aug. 12 at three Web retailers:

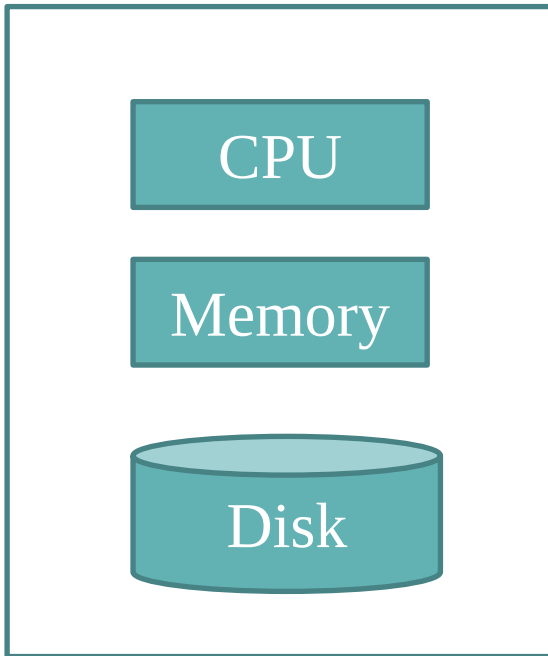


Data opslag en verwerking



Data verwerking

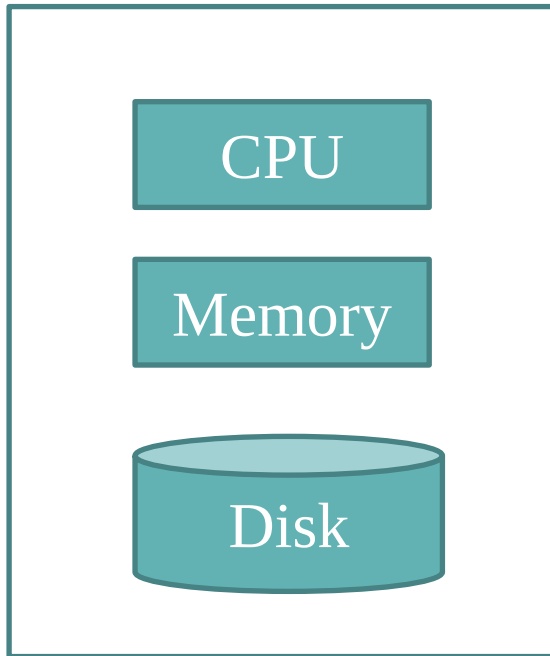
Single Node Architecture



- Machine Learning, Statistics
- “Klassieke” Data Mining

Data verwerking

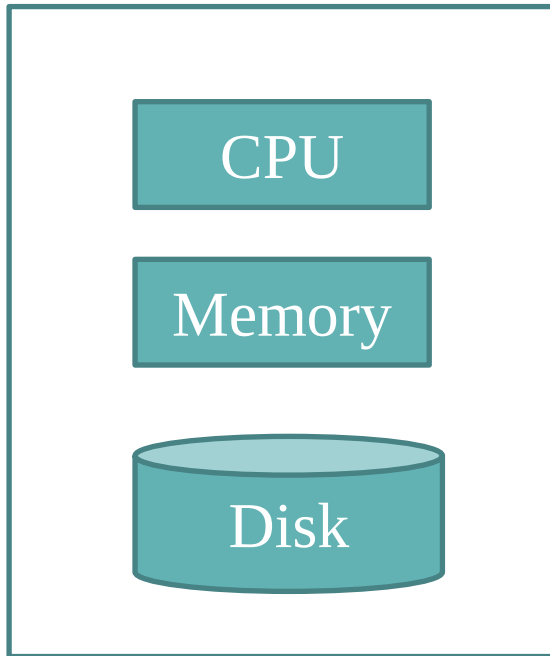
Single Node Architecture



- Google Case
 - 10 miljard (M) webpagina's
 - Gem. webpagina: 20 KB
 - $10\text{ M} * 20\text{ KB} = 200\text{ TB}$
 - Disk read bandwidth = 50 MB/sec
 - Benodigde tijd = 4 milj sec = 46+ dagen
 - Zonder verdere analyse
- SSD: 450 MB/sec

Data verwerking

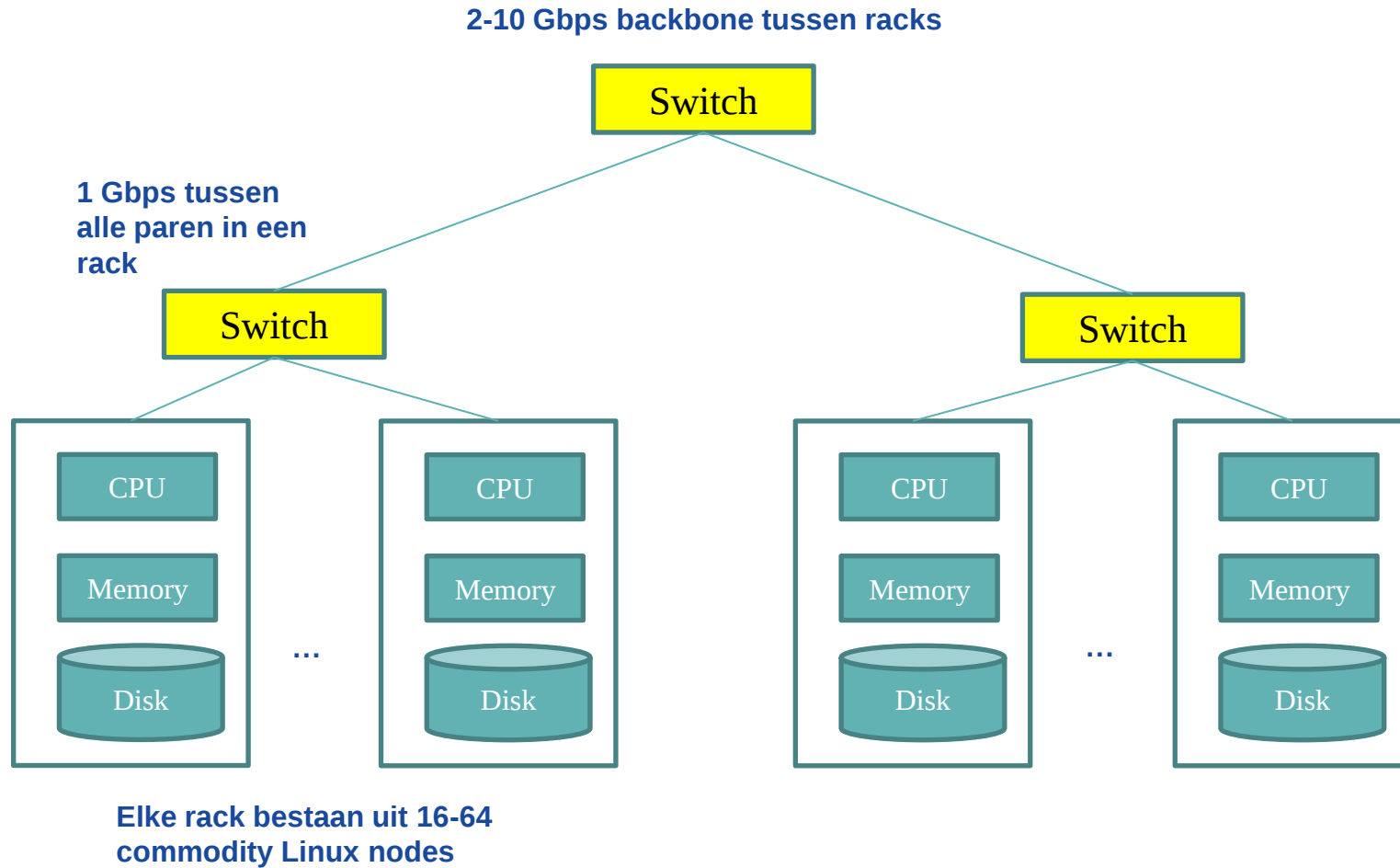
Single Node Architecture



- Geheugen wordt veel goedkoper: 12 tot 30 GB/sec
- High-bandwidth storage is nog duur: 6 GB/sec
- Netwerk: 4GB/sec
- Vergelijk Oracle Exadata:
 - 230 TB flash memory
 - Infiniband: 40 GB/sec

Data verwerking

Cluster Architecture



Data verwerking

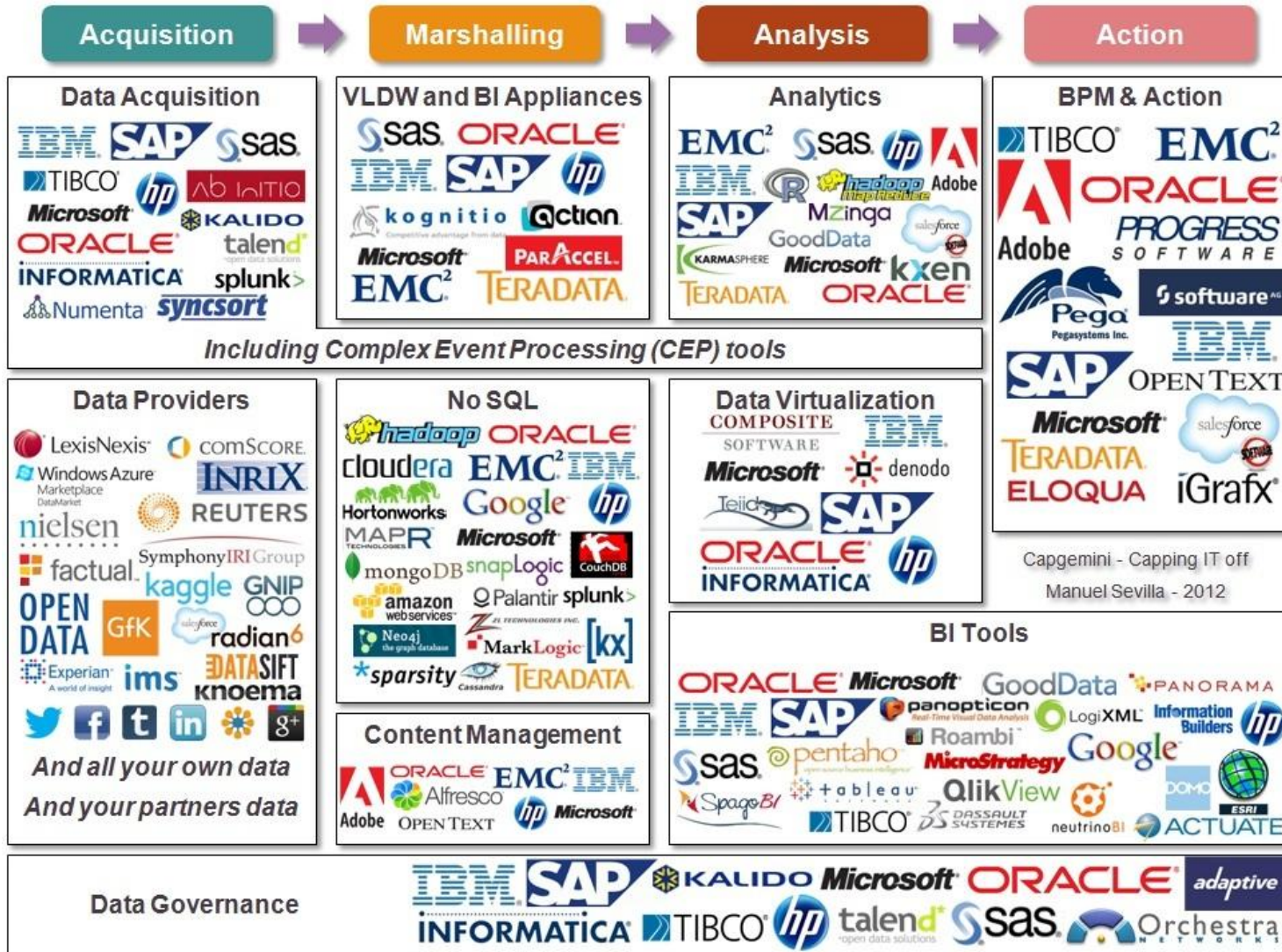
Cluster Architecture

- Node failures
 - 1 server kan wel 3 jaar blijven draaien
 - 1000 servers = 1 failure per dag
 - 1M servers = 1000 failures per dag
- Uitdaging:
 - Hoe kan je data persistent opslaan?
 - Hoe ga je om met langdurige rekenprocessen?
 - Netwerk is een trage factor

Big Data Technologie oplossingen

	In memory DB	MPP DB	Big Data Appliance	Hadoop	NoSQL DB
Consistent					
Available					
Fault tolerant					
Real-time transactions					
Analytics					
Extreme Big Data					
Unstructured Data					

Big Data Landscape



Big Data Landscape

Big Data Landscape (Version 2.0)



© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures

Big Data

Verticale vs Horizontale schaalbaarheid



Schaalbaar
rekenen:
ieder begin is
klein

Big Data

Schaalbaarheid



Big Data

Schaalbaarheid



Big Data Schaalbaarheid



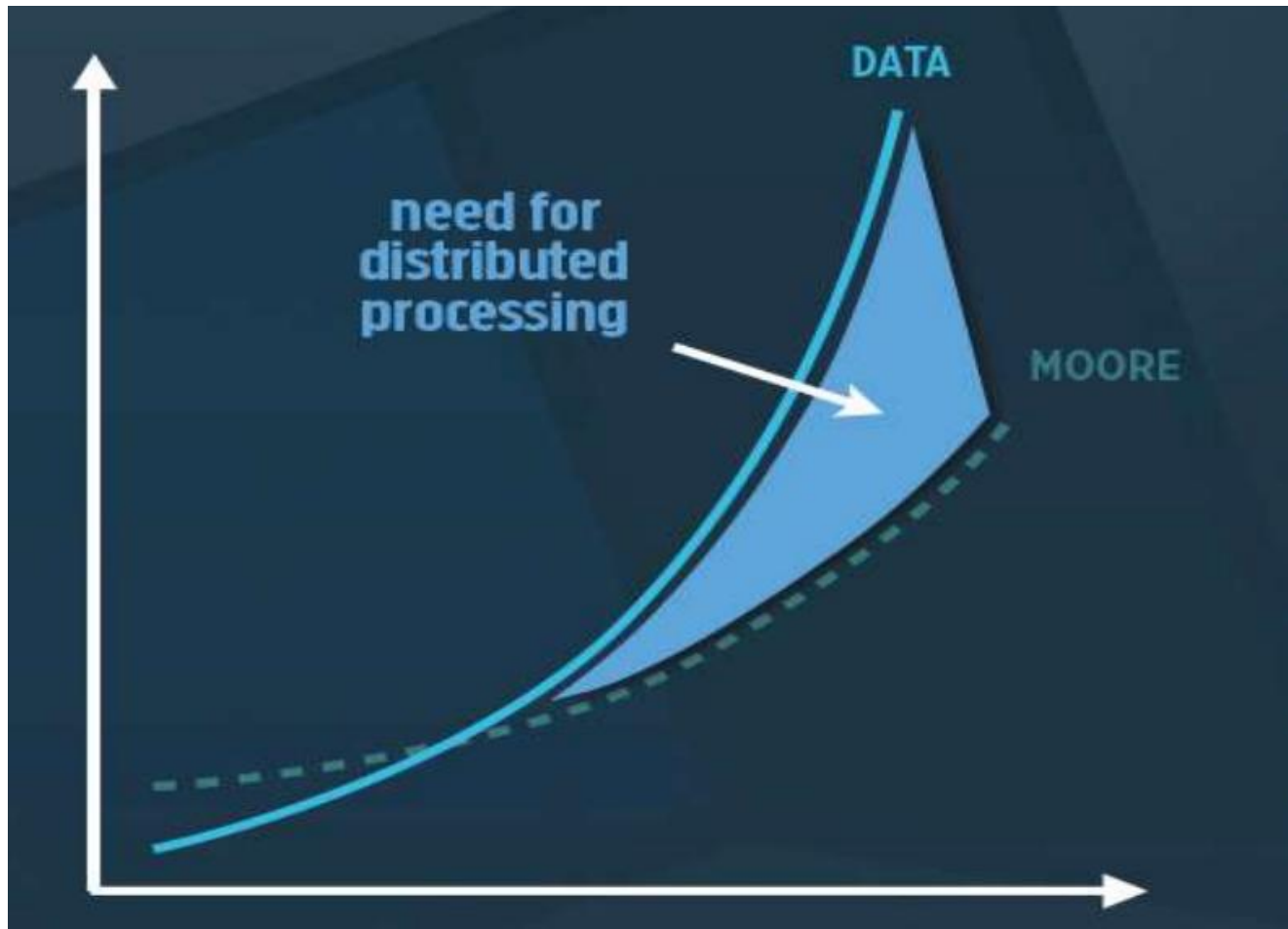
Big Data

Schaalbaarheid



Big Data

Schaalbaarheid



Big Data

Schaalbaarheid



Lineair horizontale schaalbaarheid: klein
beginnen, groot eindigen

Big Data

NoSQL databases

- NoSQL: term bedacht voor non-relational databases (alternatief voor databases waar relational databases niet geschikt voor zijn)
 - Verminderen van complexiteit zoals ACID
 - Hoge verwerkingssnelheid
 - Horizontale schaalbaarheid
 - Geen dure object/relational mapping
 - Geschikt voor eenvoudige data structuren
 - Performance boven betrouwbaarheid
- Drivers achter NoSQL: Google, Amazon en OpenSource (vanaf 2000)
- Bijzondere toepassingen:
 - Stream processing
 - Text processing
 - Scientific-oriented (Arrays vs Tables)
- Tegenwoordig beter: Not Only SQL

Big Data

NoSQL databases

Aantal: 150 (nosql-databases.org)

Types:

- Key-Value



- Document Store



- Entity-Attribute-Value Store

▪ Relational Data

UserID	UserName	Email	Password	FirstName	LastName
1	Chuck.Norris@icb.bg	Chuck.Norris@icb.bg	[enc]	Chuck	Norris
2	Agent.007@icb.bg	Agent.007@icb.bg	[enc]	Agent	007

▪ EAV Tables

▪ Entity

EntityId	EntityName
1	Chuck.Norris@icb.bg
2	Agent.007@icb.bg

▪ Entity Attribute

AttributeId	Name
1	Email
2	Password

▪ Value

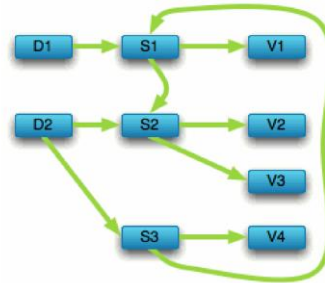
EntityId	AttributeId	Value
1	1	Chuck.Norris@icb.bg
1	2	[enc]

Big Data

NoSQL databases

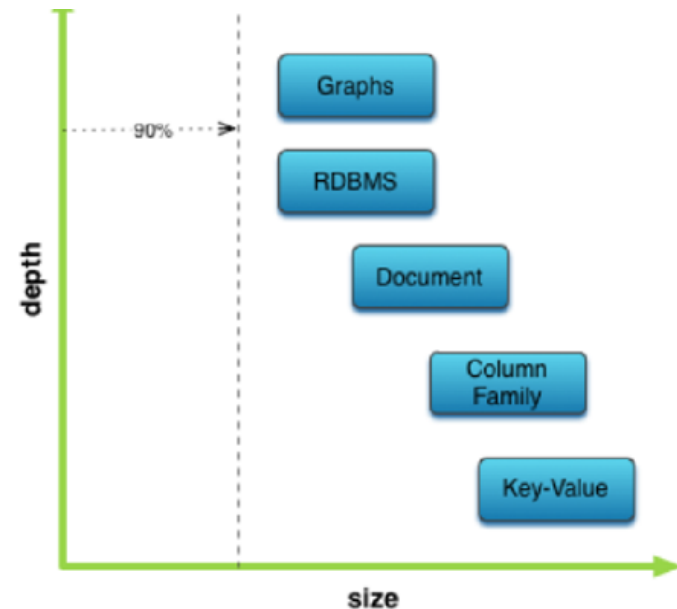
Types (vervolg)

- Graph

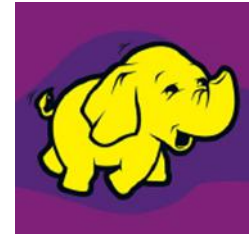


Not **only** SQL, but modern choices which excel at answering different kinds of questions.

- Average income? Ask an RDBMS
- Shopping cart? Use a Key-Value Store
- How did you get here? Ask a graph



Big Data Instrumenten



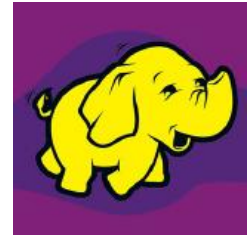
Hadoop

- Open Source project
- Gedistribueerde verwerking van grote hoeveelheden data (batch)
- Gebaseerd op Google MapReduce
- Data wordt opgeslagen op HDFS (HBase)

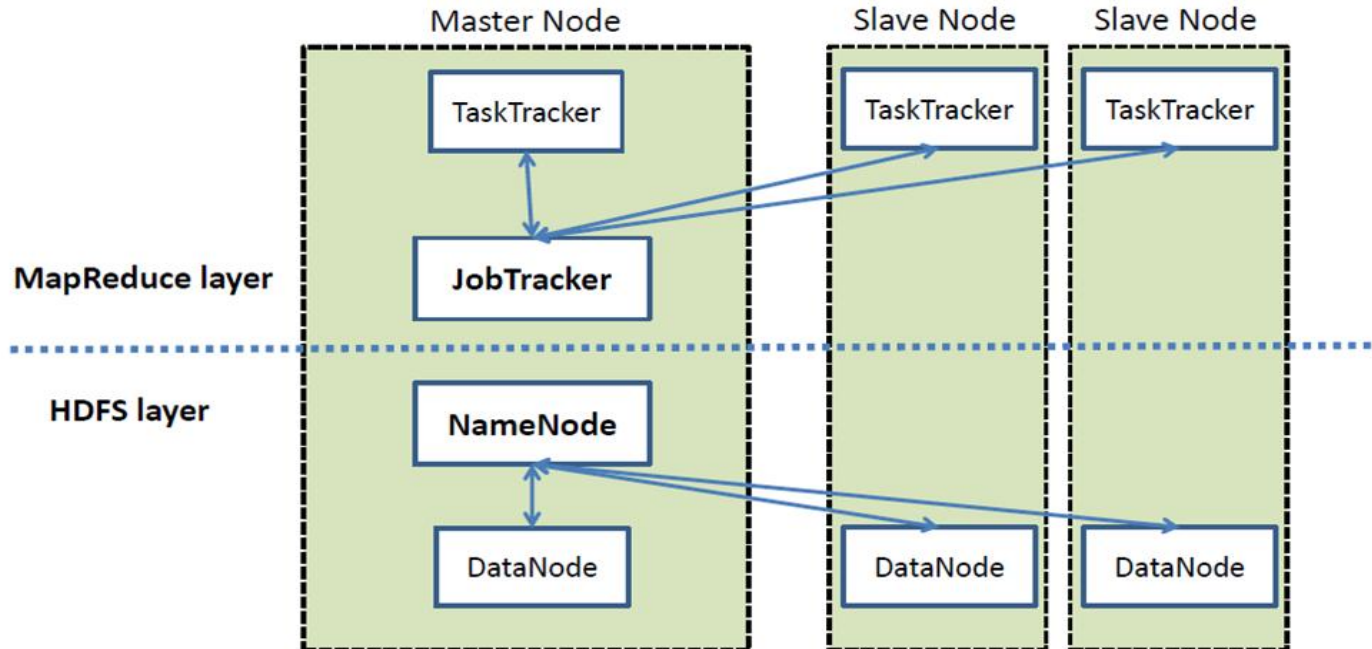
Uitgangspunten:

- Verwerking en opslag van grote hoeveelheden data (PetaBytes)
- Performance, storage, process scale linair
- Berekeningen verplaatsen naar data
- Simple core, modulair en uitbreidbaar
- Failure is normaal en verwacht
- Beheerbaar en zich zelf corrigerend
- Ontworpen om op normale (commodity) hardware te draaien – kosten effectief
- Write once, read many times

Big Data Instrumenten

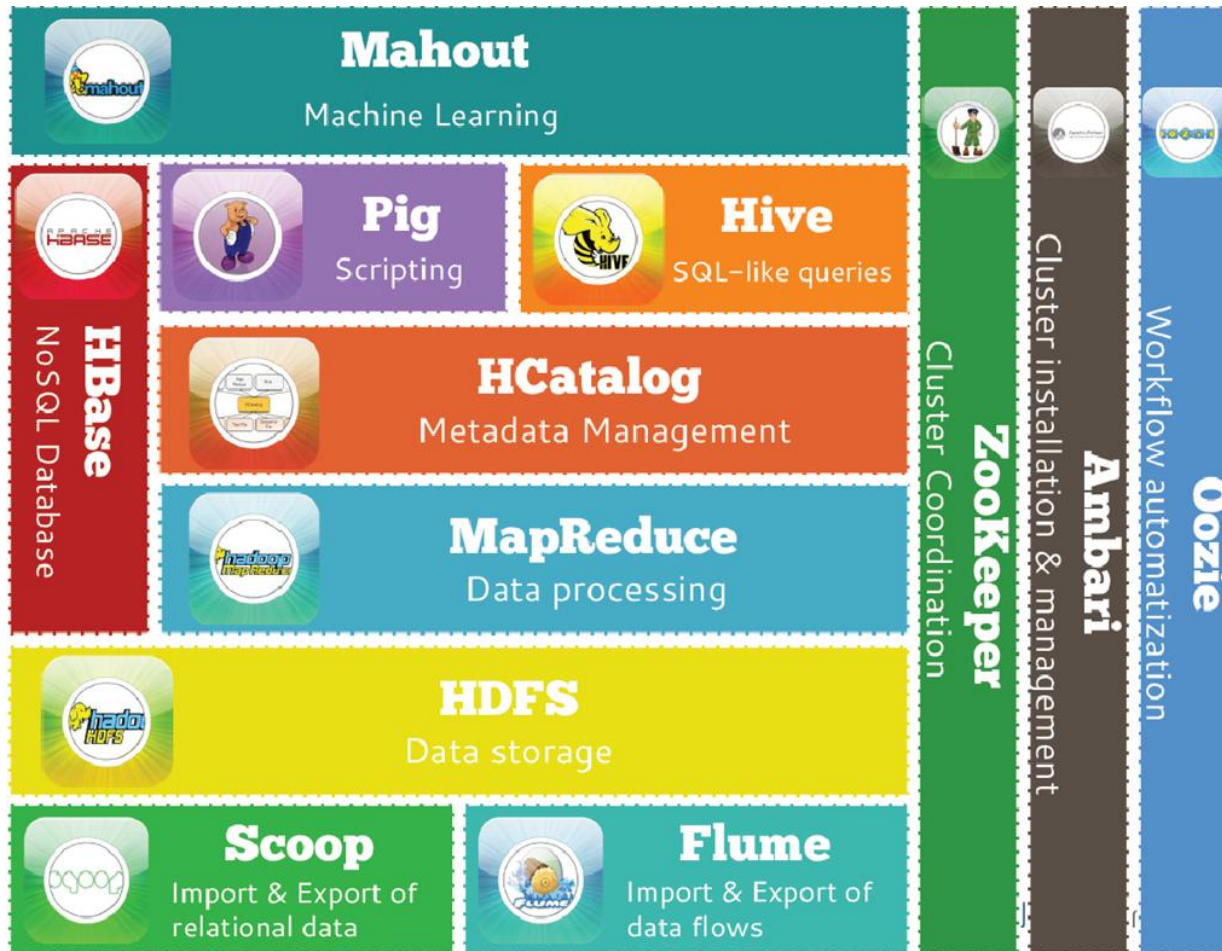


High Level Architecture of Hadoop



Big Data

Hadoop Core



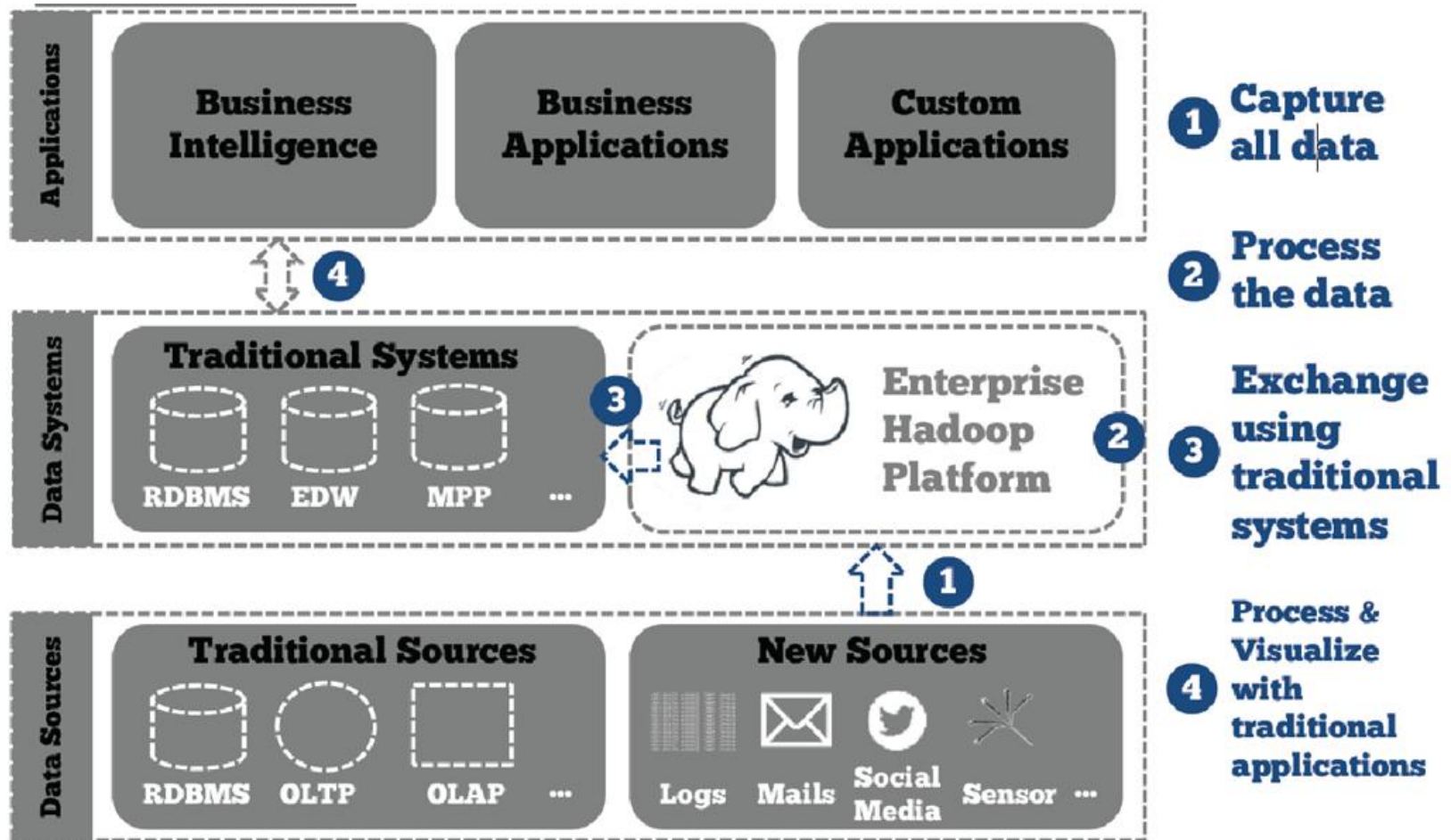
Big Data

Hadoop Store



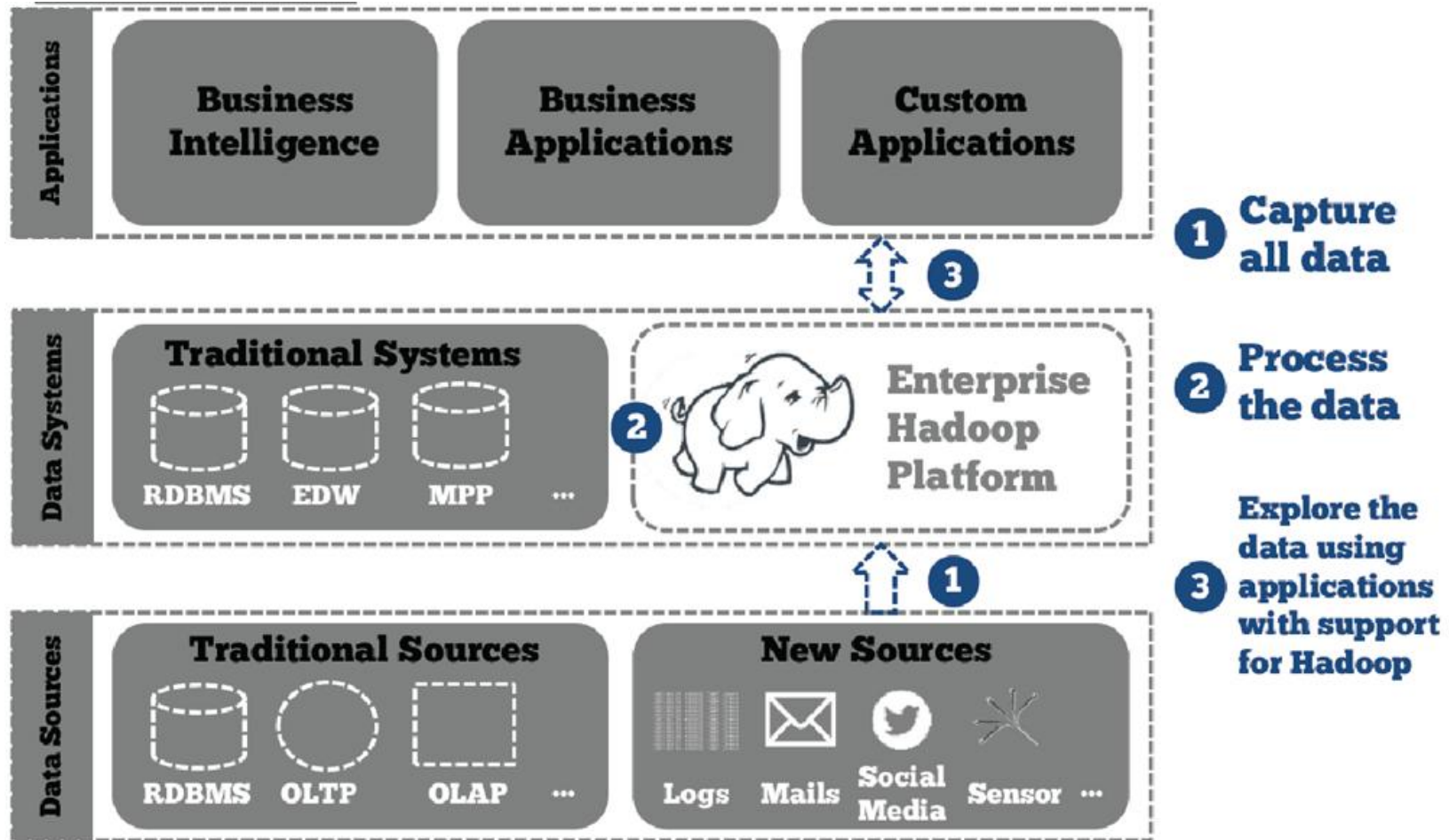
Big Data

Hadoop Refine Data



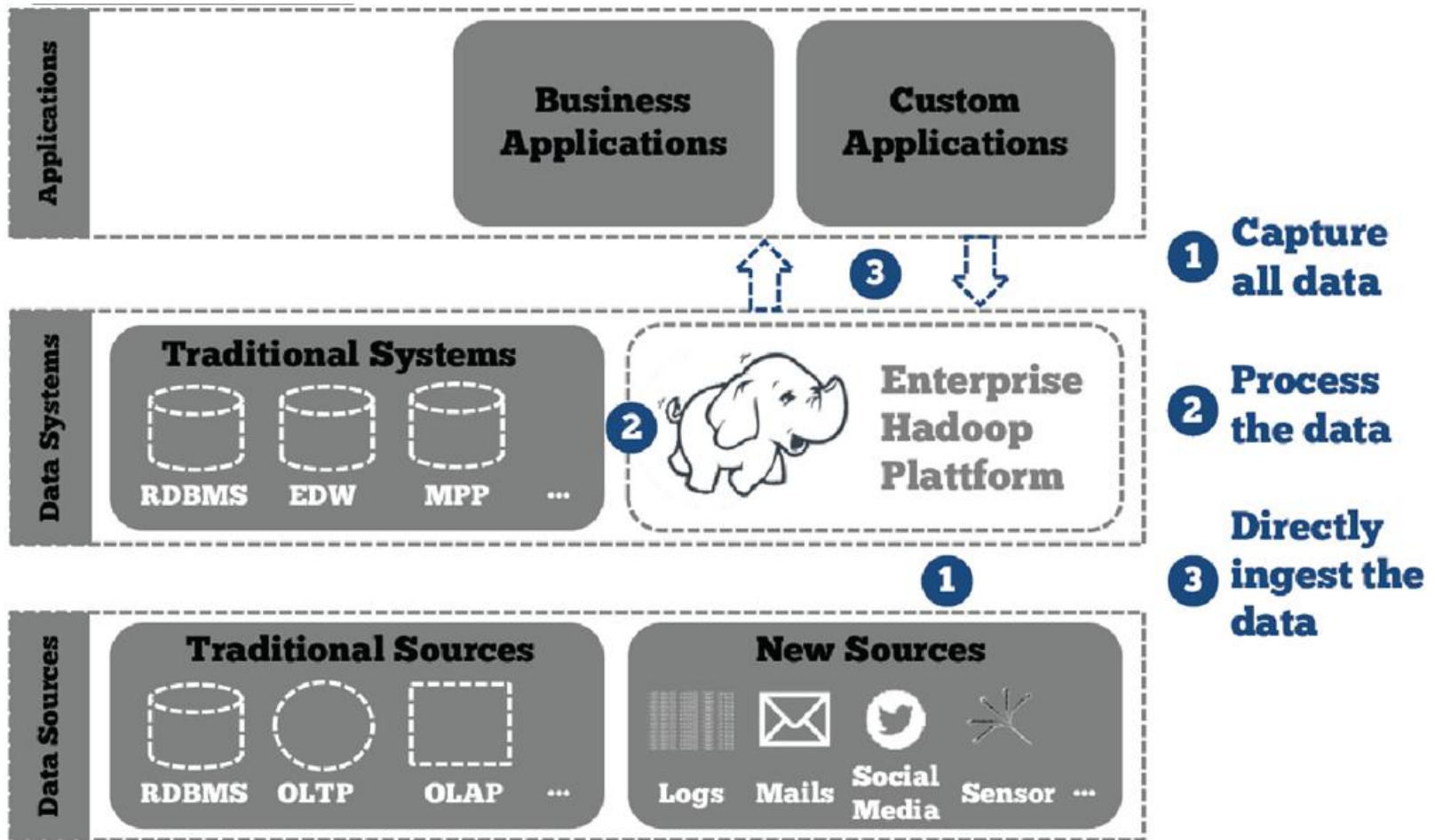
Big Data

Hadoop Explore Data



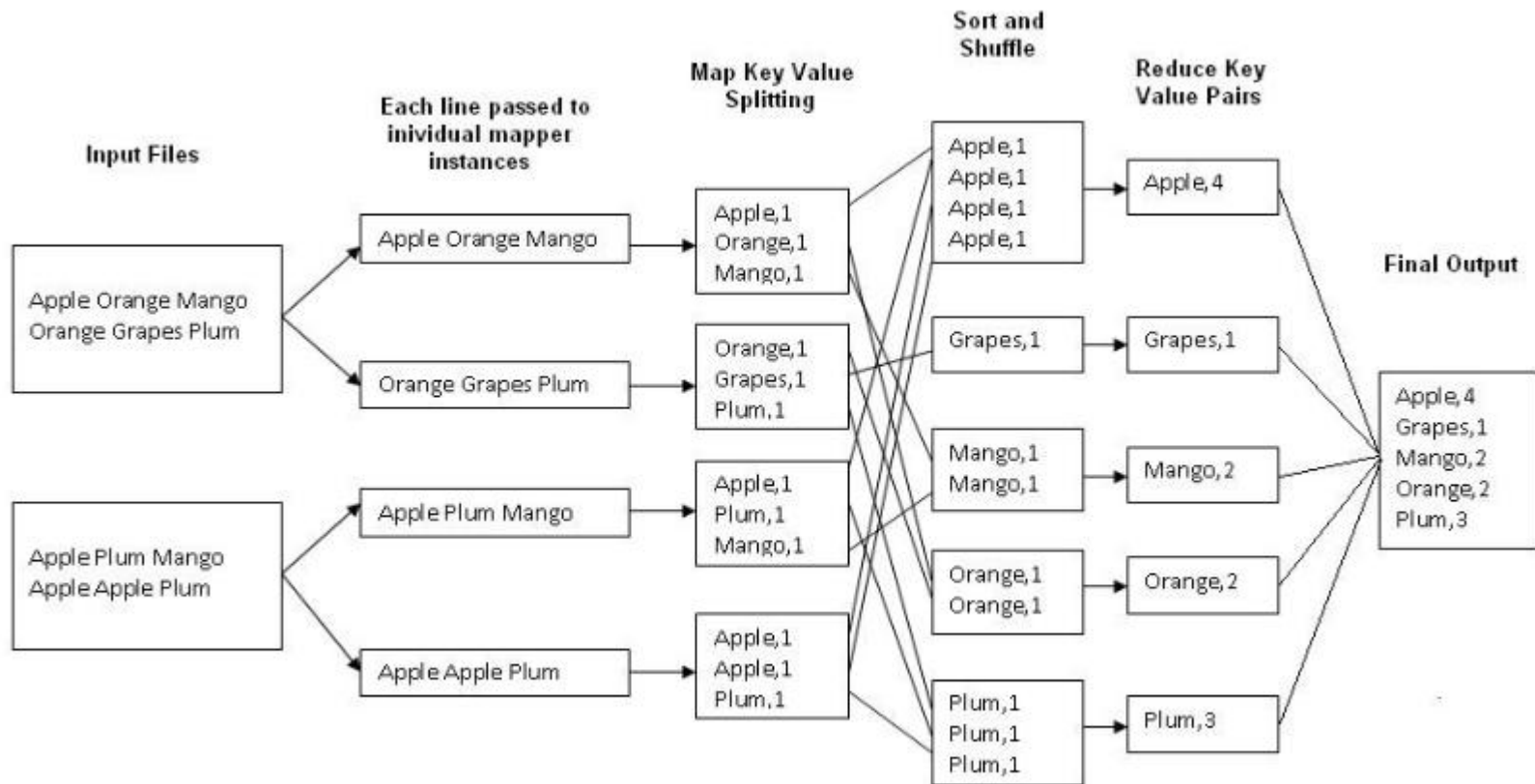
Big Data

Hadoop Enrich Data



Big Data Methoden

Map Reduce



Big Data

Methoden

Tekst analyse

- Sentiment analyse
- Samenvattingen
- Herkennen van onderwerpen
- Eigenschappen van onderwerpen/personen
- Context van het onderwerp

Big Data

Oracle Solutions

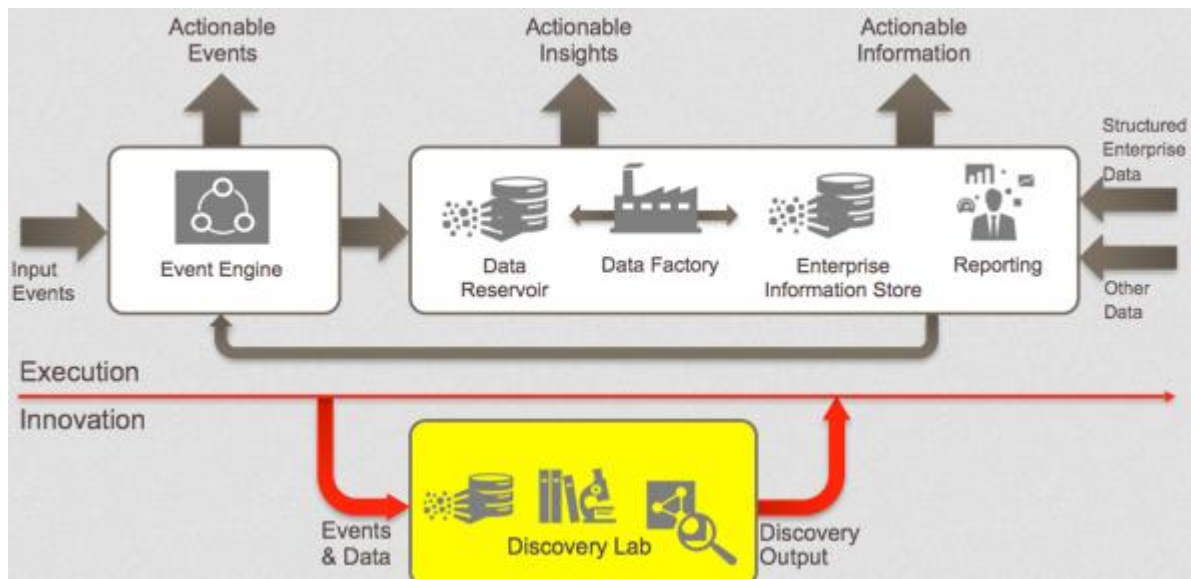
- Oracle Big Data Appliance
 - Hardware (vgl. Exadata)
 - Cloudera CDH
 - Cloudera Manager
 - Cloudera RTQ (Impala)
 - Oracle NoSQL Databases
 - Oracle R Distribution
 - Oracle Big Data SQL



Big Data

Oracle Solutions

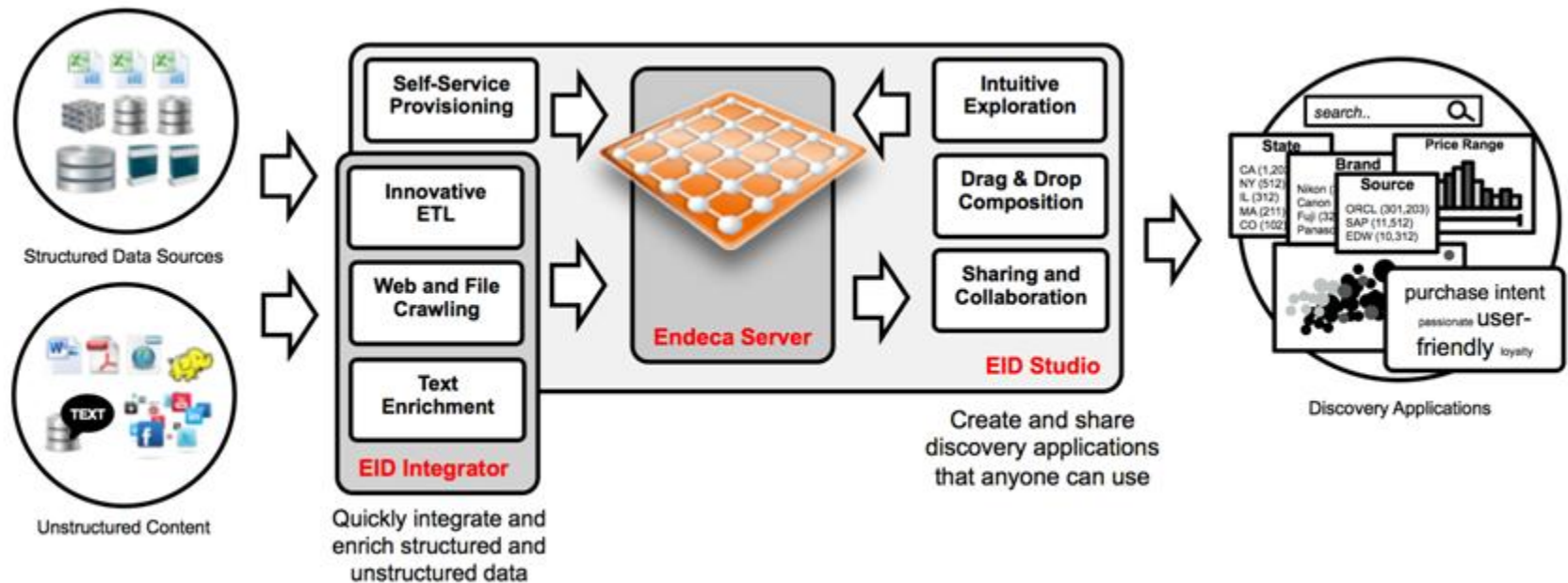
- Oracle Big Data Discovery
 - Visualisaties op Hadoop
 - Data transformatie en verrijking (Apache Spark)



Big Data

Oracle Solutions

- Oracle Endeca
 - BI op ongestructureerde en tekst data
 - 'click-and-refine' user interface





CONSULTING | MANAGED SERVICES | EDUCATION

WWW.TRANSFER-SOLUTIONS.COM

