

# Streamlining Analytics

[Jan.Meskens@smals.be](mailto:Jan.Meskens@smals.be)

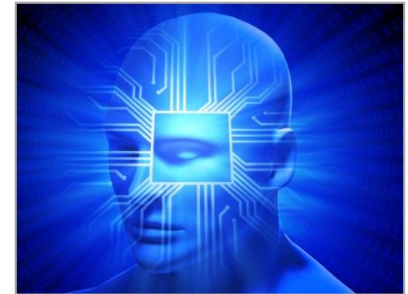
[Dries.VanDromme@smals.be](mailto:Dries.VanDromme@smals.be)

Onderzoek - [@SmalsResearch](https://twitter.com/SmalsResearch)

<http://blogresearch.smalsrech.be>

# Streamlining Analytics

**Predictive analytics**  
**De data supply chain**



**Barrières bij de introductie van analytics**



**Hardware appliances voor analytics**

**Data quality**

**Analytics project management**



# Streamlining Analytics

**Predictive analytics**  
De data supply chain



**Barrières** bij de introductie van analytics



**Hardware appliances** voor analytics

**Data quality**

**Analytics project management**





**Data**



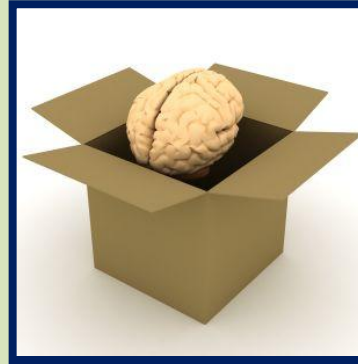
**Voorspellingen  
over de  
toekomst**



# Fase 1: Training

Predictief Model

Algoritmes



Historische data

Nieuwe data

# Fase 2: Toepassing



Voorspellingen  
over de  
toekomst



# PA voorbeeld: marketing analytics

Instant Order Update for Jan Meskens. You purchased this item on March 1, 2013. [View this order.](#)

**Voorspel** wat kopers nodig hebben

**Toon advertenties** voor deze producten



Click to **LOOK INSIDE!**

**PREDICTIVE**  
or  
Sign in to turn on 1-Click ordering

[Add to Wish List](#)

**Sell Us Your Item**  
For a **\$5.60** Gift Card

[Trade in](#)

[Learn more](#)

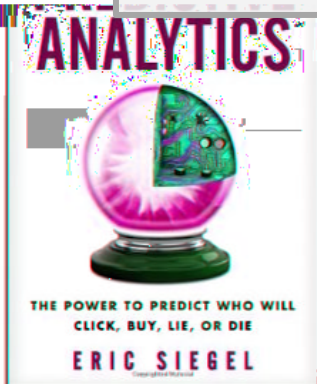
**More Buying Choices**  
**64 used & new from \$10.25**

Have one to sell? [Sell on Amazon](#)

**Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die**  
[Hardcover]  
Eric Siegel (Author), Thomas H. Davenport (Foreword)

★★★★★ (5 customer reviews)

List Price: ~~\$28.00~~



Click to open expanded view

[See all 2 customer images](#)

[Share your own customer images](#)

Price: **\$15.07** & **FREE Shipping** on orders over \$25. [Details](#)  
You Save: **\$12.93 (46%)**

**In Stock.**  
Ships from and sold by Amazon.com. Gift-wrap available.

**44 new** from \$10.25 **20 used** from \$11.95

Formats	Amazon Price	New from	Used from
Kindle Edition	\$21.18	--	--
Hardcover	\$15.07	\$10.25	\$11.95

money&markets

**Shop the Money & Markets Store**

Are you a finance, investing, economics or accounting professional? Find books, read blog and discover new authors and thought-leaders in [Money & Markets](#), a new home for finance industry professionals on Amazon.com. [> Shop now](#)

posts, e

## Customers Who Bought This Item Also Bought

Page 1 of 17

**asure Anything: Value of ...**  
W. Hubbard  
★ (58)  
Hardcover  
\$70.40

**Handbook of Statistical Analysis and Data ...**  
> Robert Nisbet  
★★★★★ (29)  
Hardcover  
\$70.40

**Big Data: A Revolution That Will Transform ...**  
Viktor Mayer-Schonberger  
★★★★★ (65)  
Hardcover  
\$15.84

**Predictive Analytics: Microsoft Excel**  
> Conrad Carlberg  
★★★★★ (11)  
Paperback  
\$23.55

**Big Data Analytics: Disruptive ...**  
Dr. Arvind Sathi  
★★★★★ (5)  
Paperback  
\$11.24

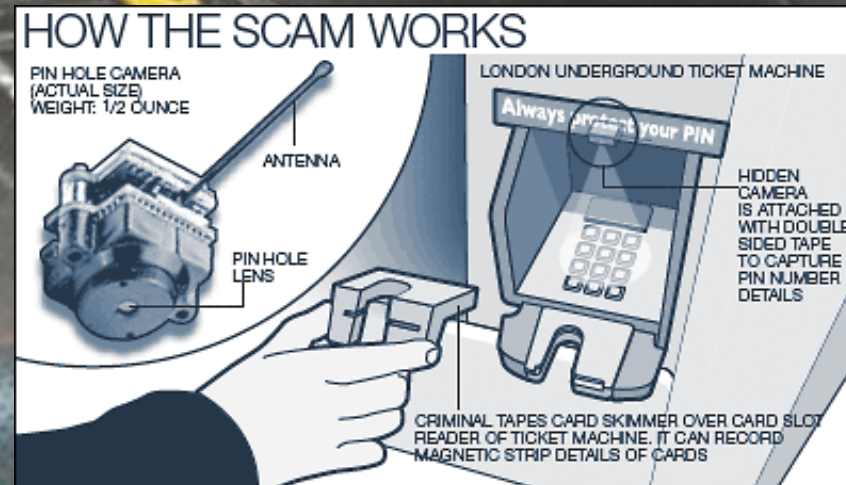
**Big Data, Big Analytics: Emerging Business ...**  
> Michael Minelli  
★★★★★ (8)  
Hardcover  
\$34.38

**How to Me Finding the ...**  
> Douglas ...  
★★★★★  
Hardcover  
\$31.94

# PA voorbeeld: **fraudebestrijding**

**Voorspel** of een transactie frauduleus is

**Blokkeer** automatisch kaarten, rekeningen,...



# PA voorbeeld: **churn prediction**

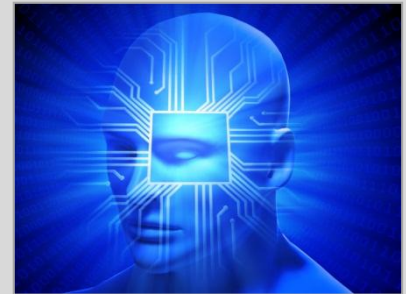
**Voorspel** of een klant gaat overlopen naar de concurrentie

**Aangepaste acties** om potentiële overlopers te behouden



# Streamlining Analytics

**Predictive analytics**  
**De data supply chain**



**Barrières bij de introductie van analytics**



**Hardware appliances voor analytics**  
**Data quality**  
**Analytics project management**



# Analytics vs. Productie-applicaties

Applicatie

Data

Analytics

Address

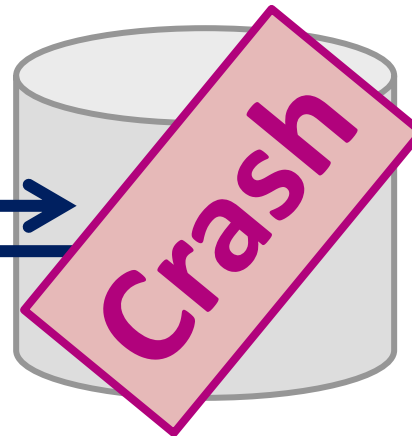
Street Address

Address Line 2

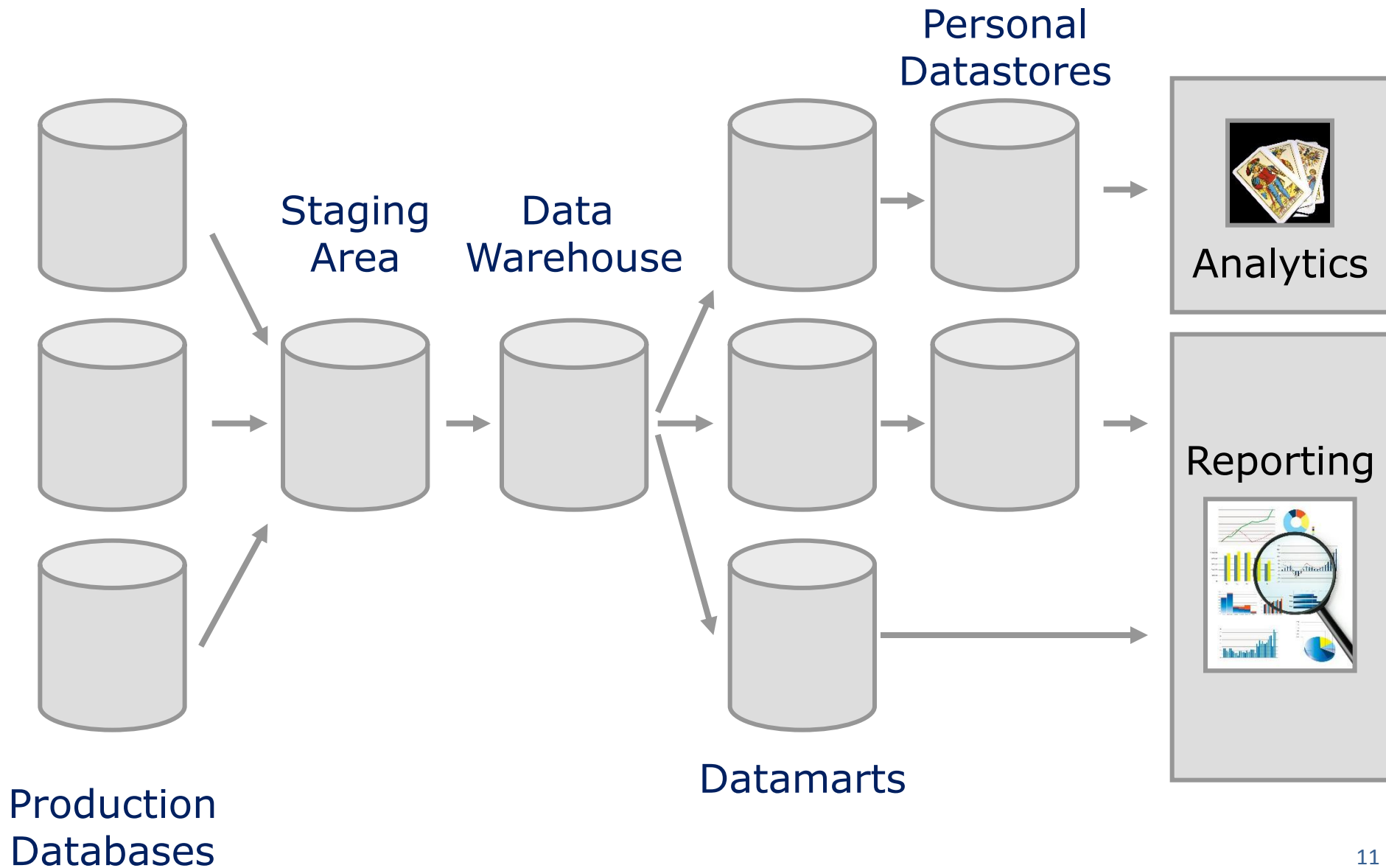
City Zip Code

Submit

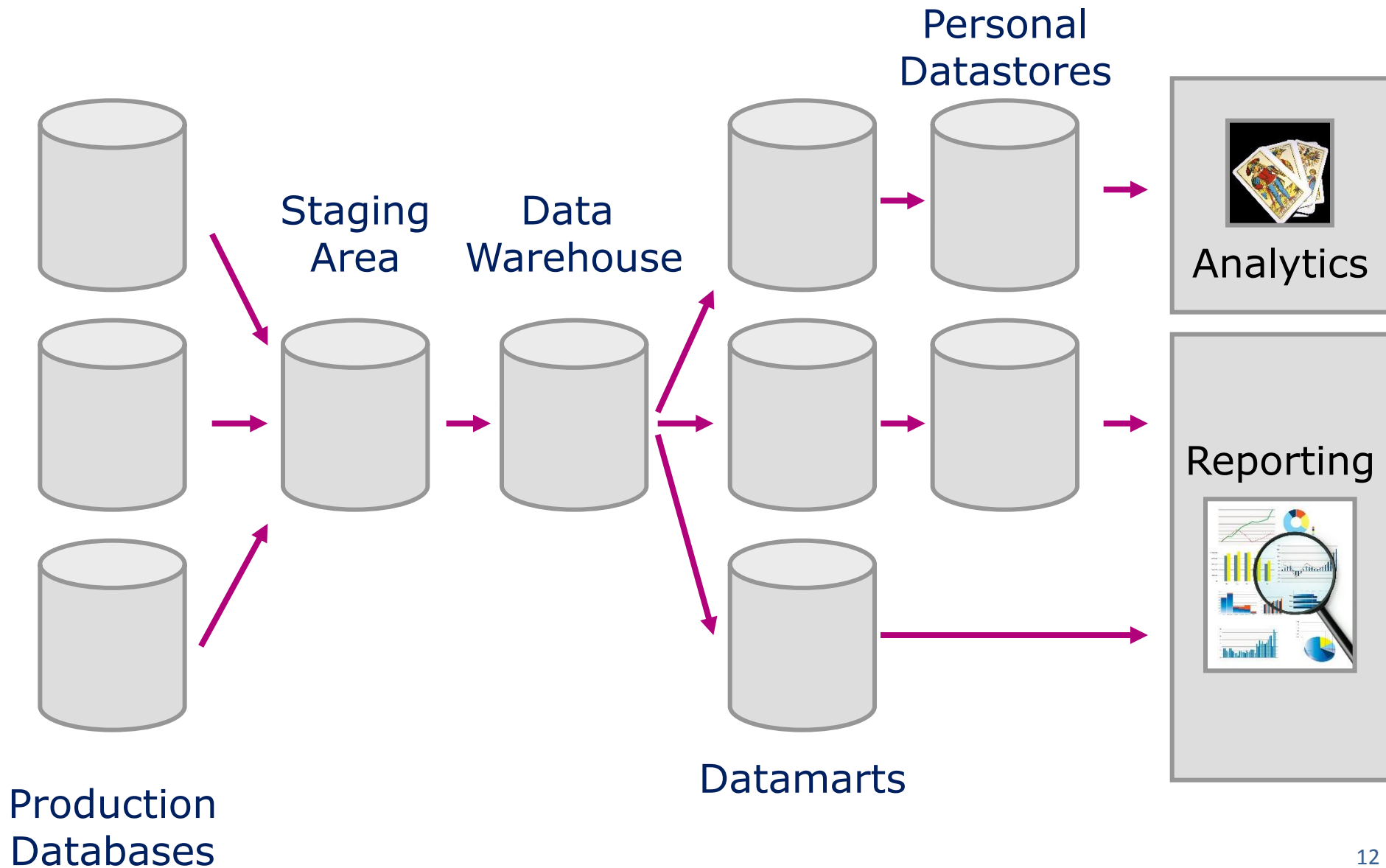
**Crash**



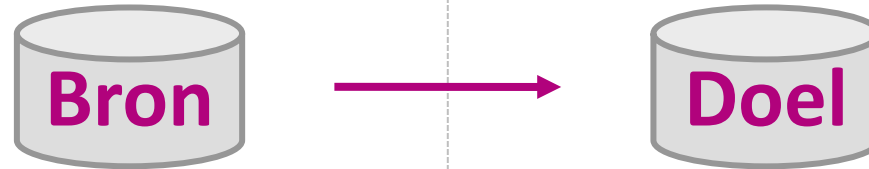
# De « data supply chain »



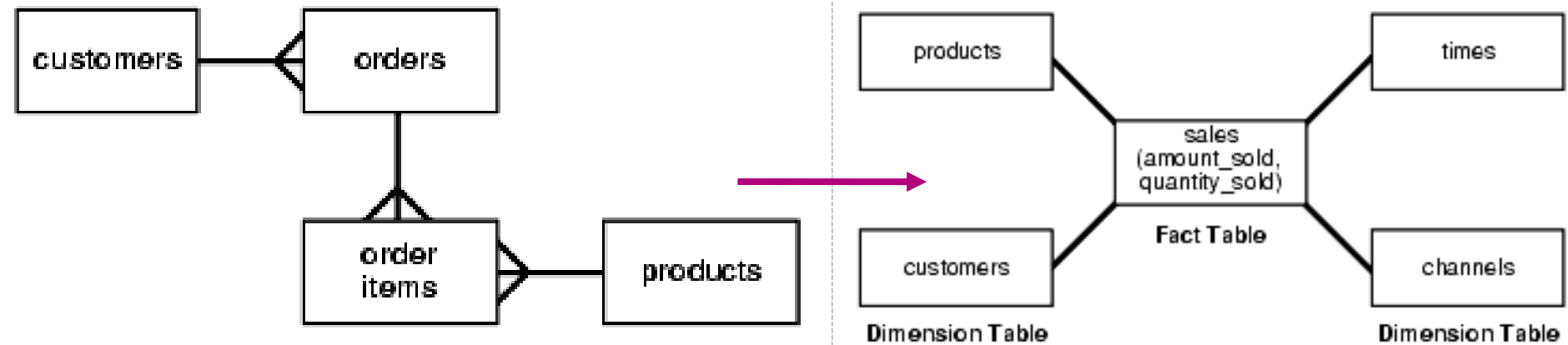
# De « data supply chain »



# ETL: Extract Transform Load



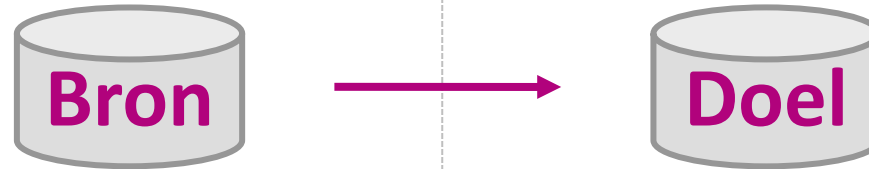
Omvormen data-schema



Derde normaalvorm Ster-schema



# ETL: Extract Transform Load



**Data-quality operaties**

**Adres correcties**

Fonsnylaan/Avenue Fonsny 20 → Fonsnylaan/~~Avenue Fonsny~~ 20

**Harmonisatie afkortingen**

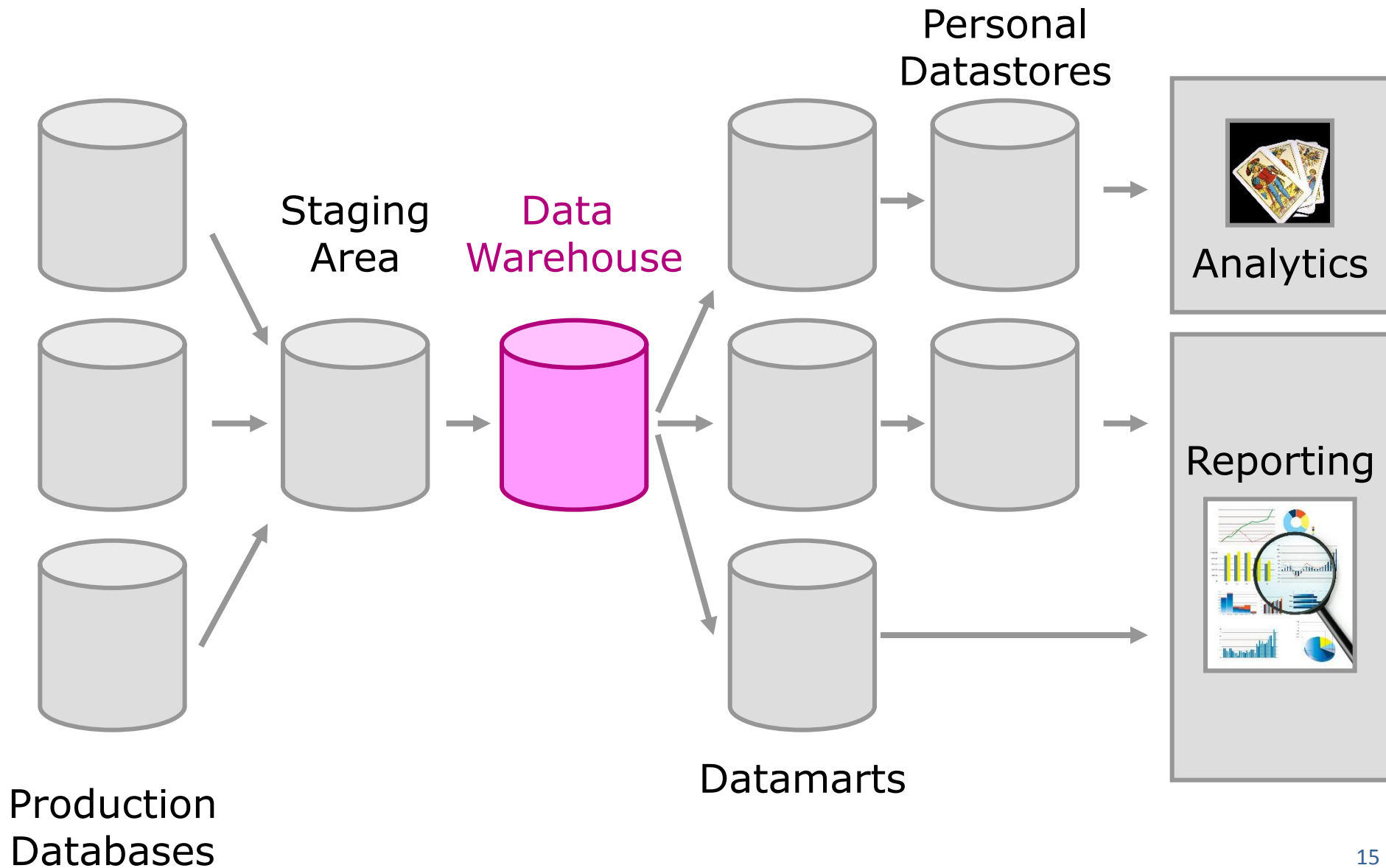
SPRL → 2200

N.v.. → NV

**Missing values?**

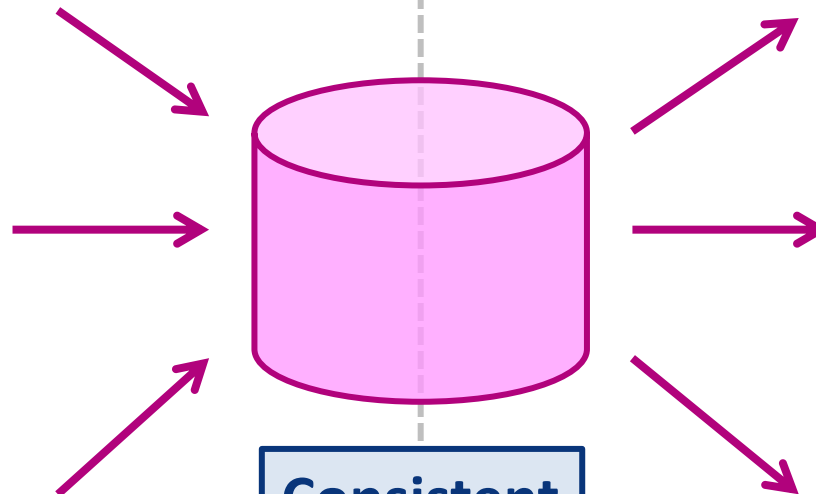
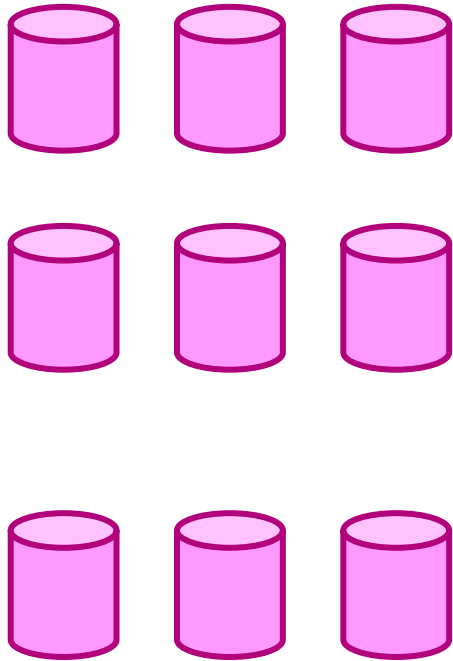


# De « data supply chain »



# Het data warehouse (DWH)

Veel operationele  
data sources



Veel toepassingen



Consistent

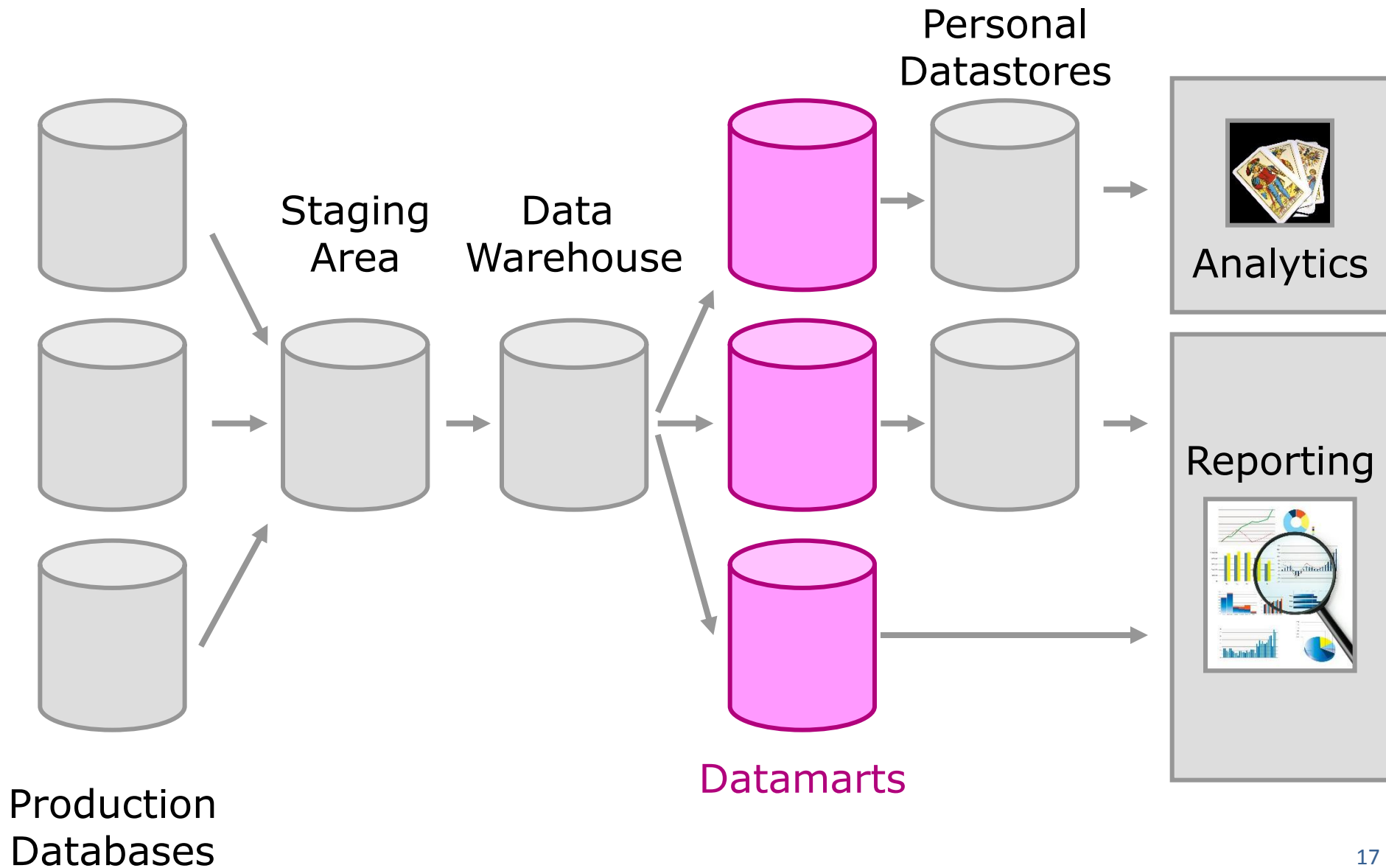
Uniform data-schema

Historieken - versies

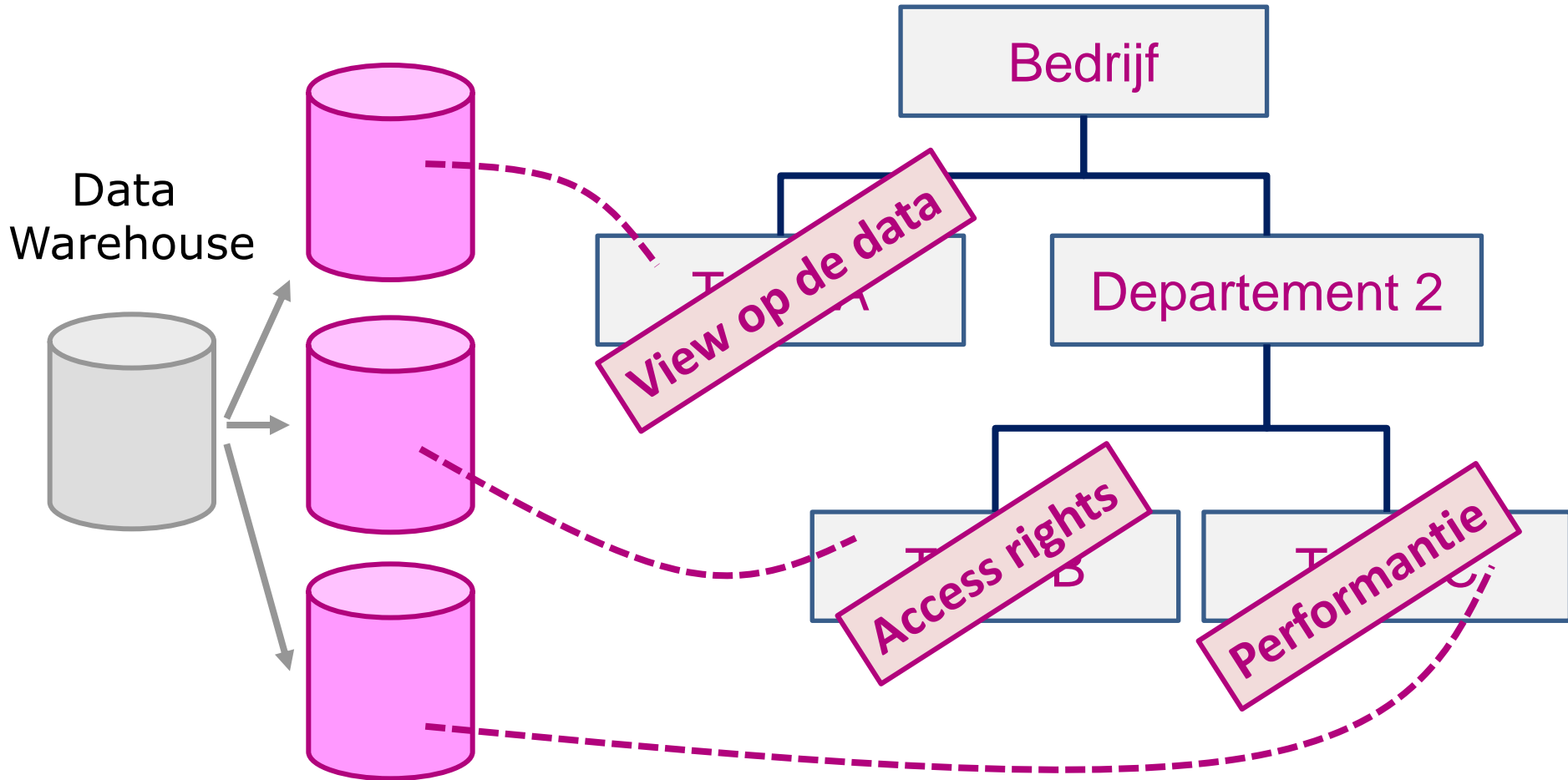
Geoptimaliseerd voor output



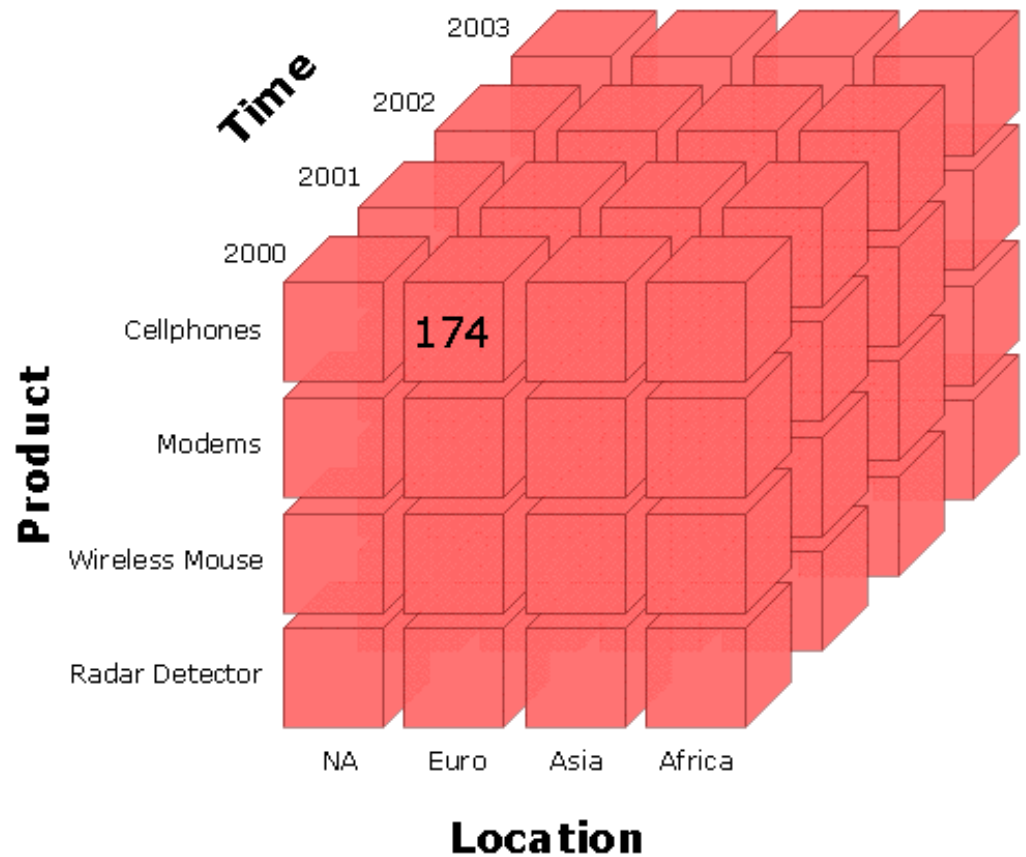
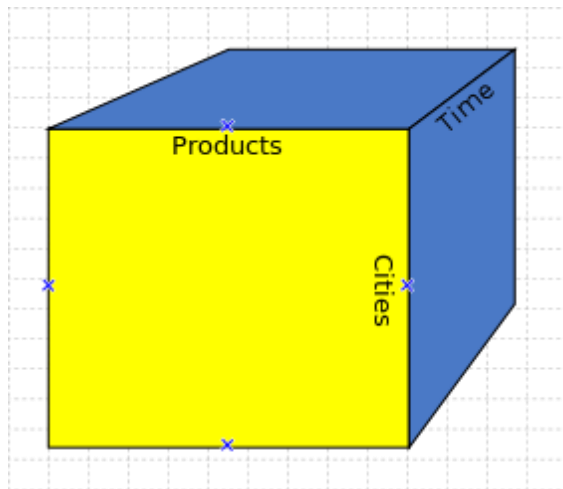
# De « data supply chain »



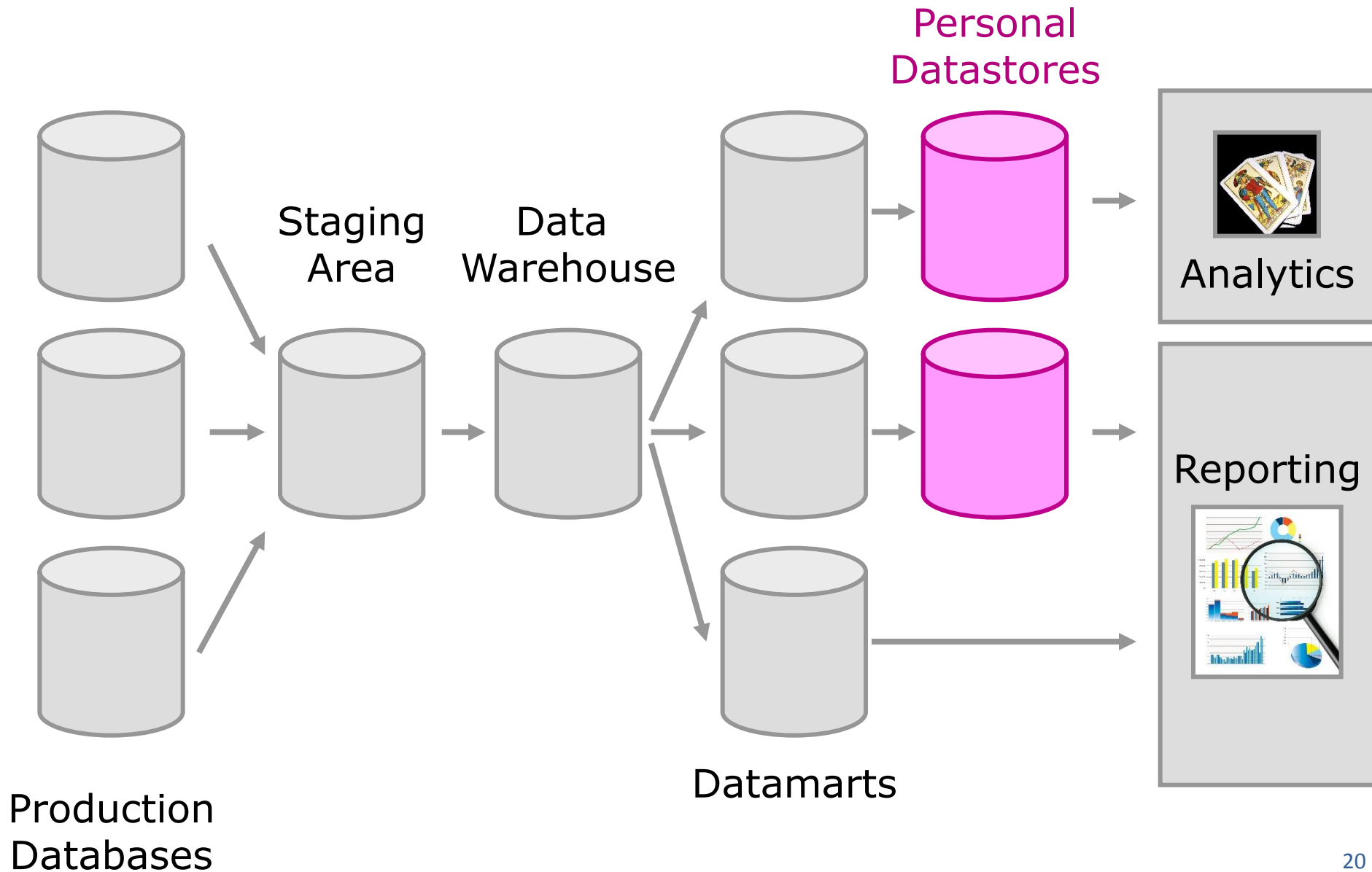
# Data marts



# Vaak gebruikt **data-model:** **OnLine Analytical Processing (OLAP)**



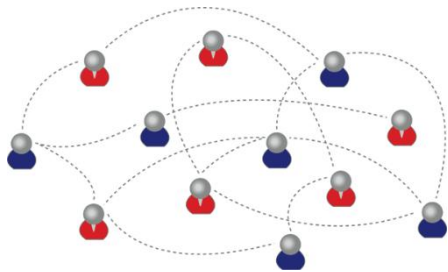
# De « data supply chain »



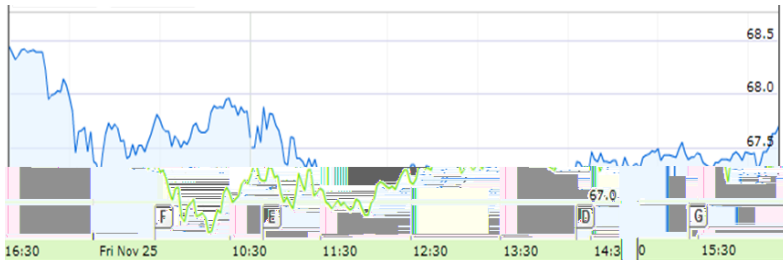
## Data-warehouse / data-mart



**Images,  
Video**



**Network data**



**Continuous data**

**Time series**

## Analytics personal datastore

	A	B	C	D	E
1	<b>X1</b>	<b>X2</b>	<b>X3</b>	<b>X4</b>	<b>Y?</b>
2	1	2	1	1	<b>TRUE</b>
3	1	1	1	1	<b>TRUE</b>
4	0	3	1	1	<b>FALSE</b>
5	1	2	1	1	<b>TRUE</b>

**Set independent variables  
( $X_1, \dots, X_n$ ) en 1 dependent  
variable (Y)**

Data-warehouse /  
data-mart

Analytics  
personal datastore



Images,  
Video



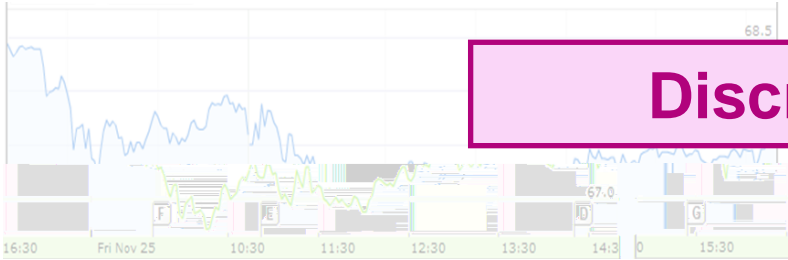
Network data

	A	B	C X3	D X4	E Y?
		2	1	1	TRUE
3	1	1	1	1	TRUE
4	0	3	1	1	FALSE
5	1	2	1	1	TRUE

Feature extraction

Graaf algoritmes

Set independent variables  
( $X_1, \dots, X_n$ ) en 1 dependent  
variable ( $Y$ )



Discretisatie

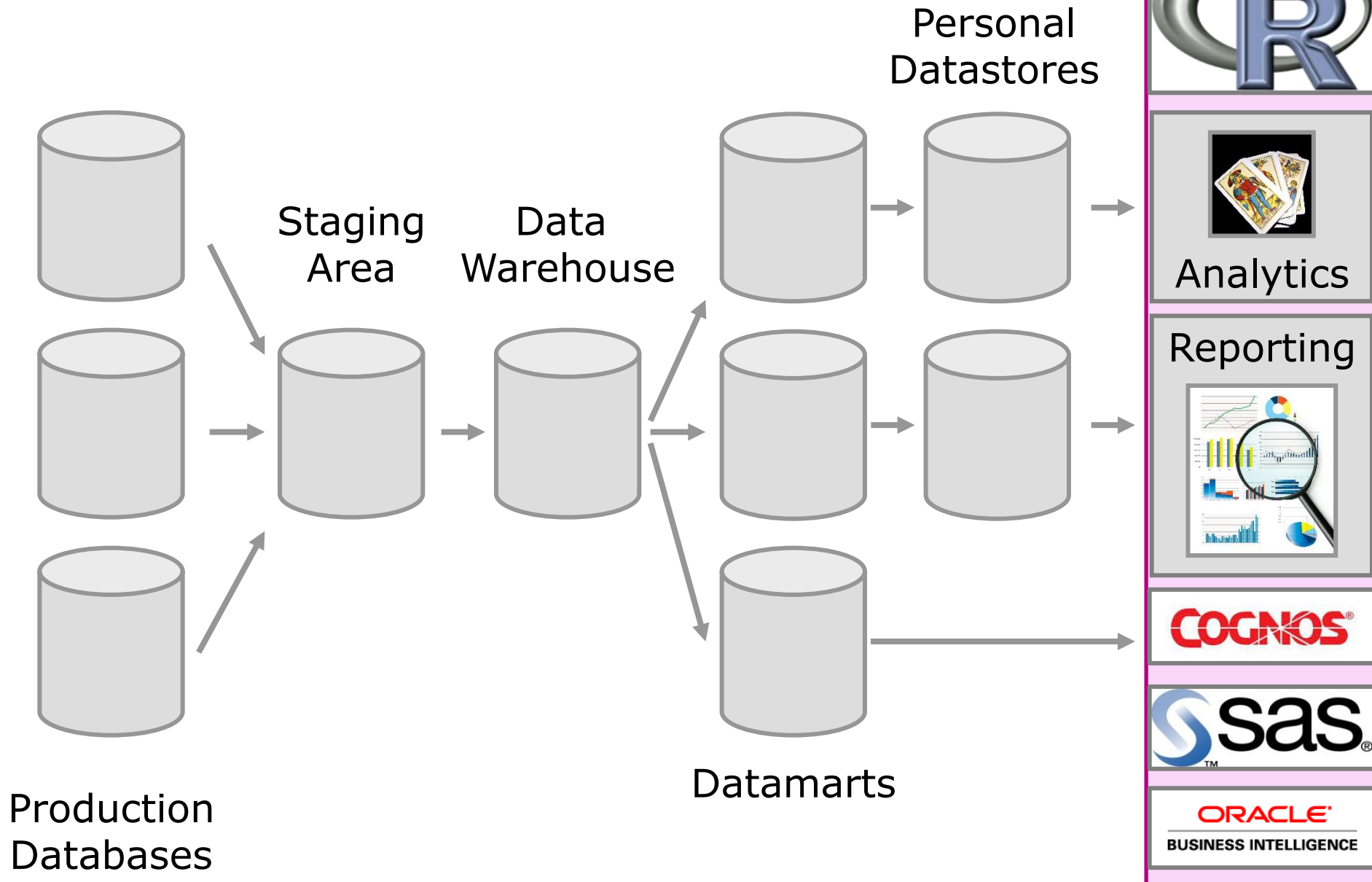
Continuous data

Time series

Herschaling

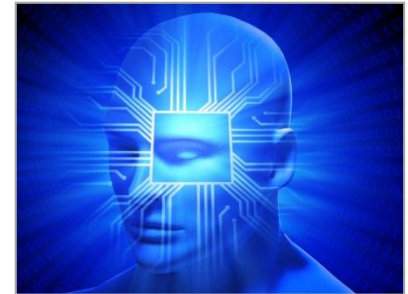


# De « data supply chain »



# Streamlining Analytics

**Predictive analytics**  
**De data supply chain**



**Barrières bij de introductie van analytics**



**Hardware appliances voor analytics**  
**Data quality**  
**Analytics project management**



# Lancering van het predictive analytics project!

**Uitstel, uitstel, uitstel, ... , afstel?**

« Atypische/onbekende techniek »

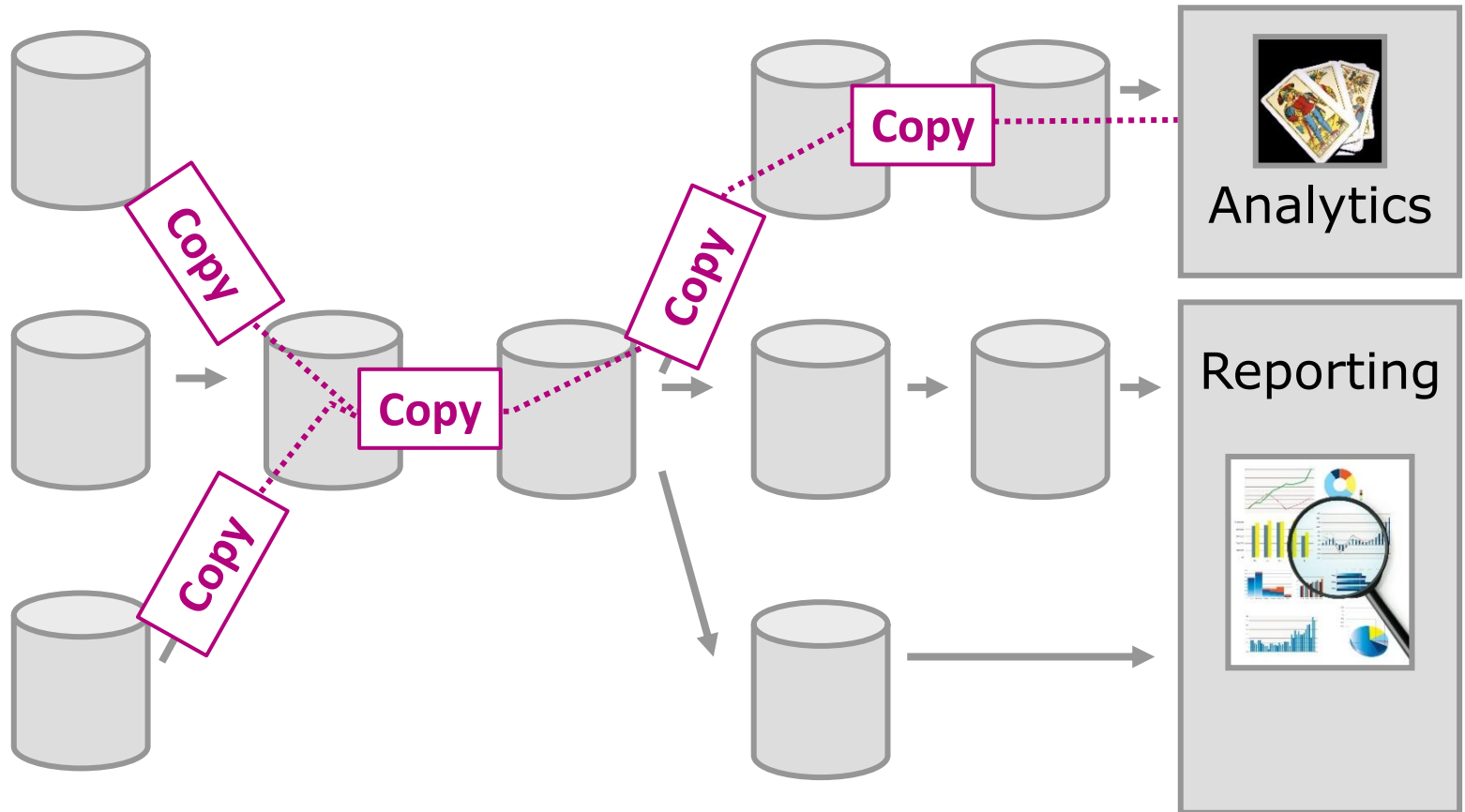
« Modellen verouderd »

« Servers niet krachtig genoeg »

« Waar is de data? »

« Betrokkenheid van veel rollen/teams »

# Barrière 1: Complexe en trage architectuur

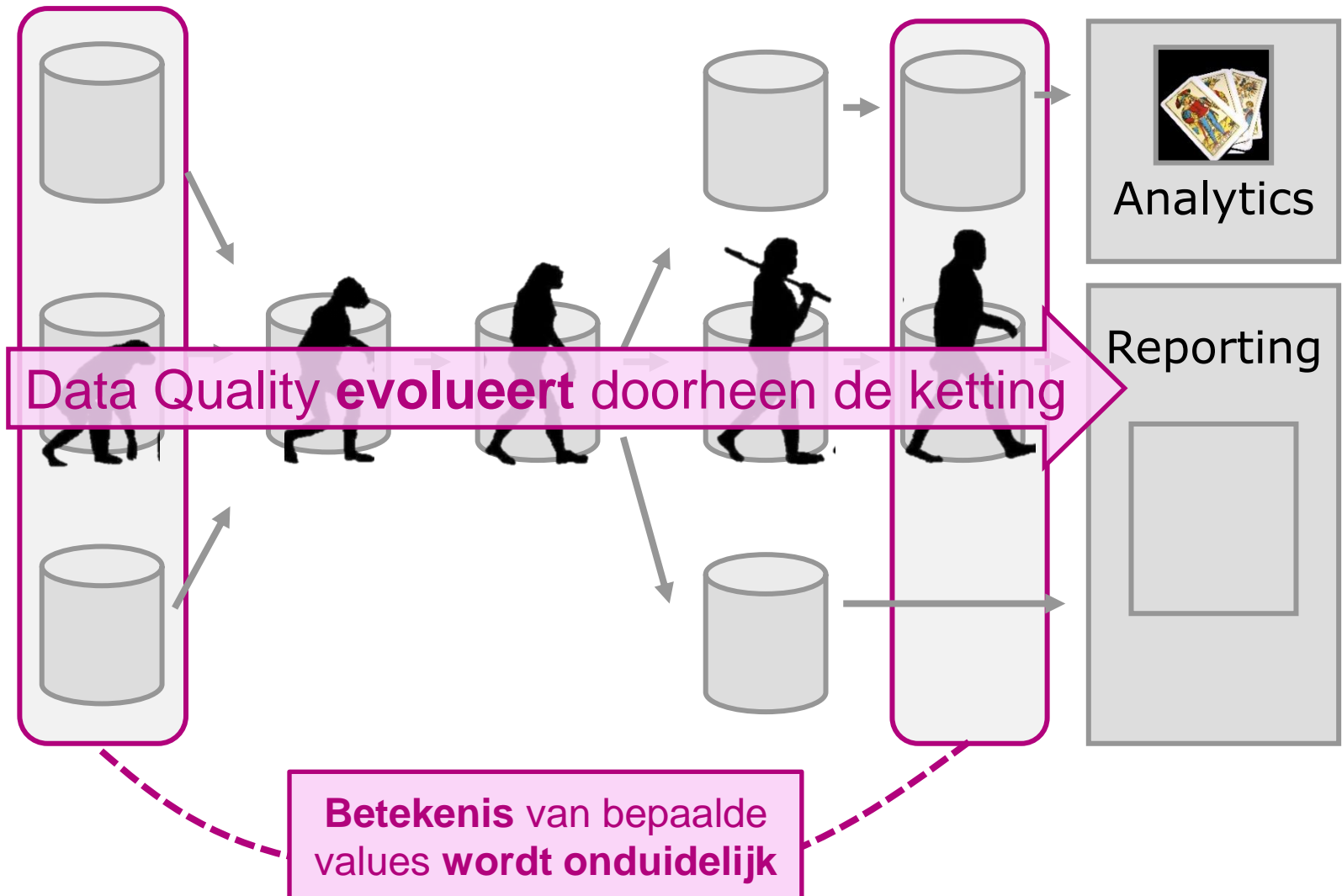


# Barrière 1:

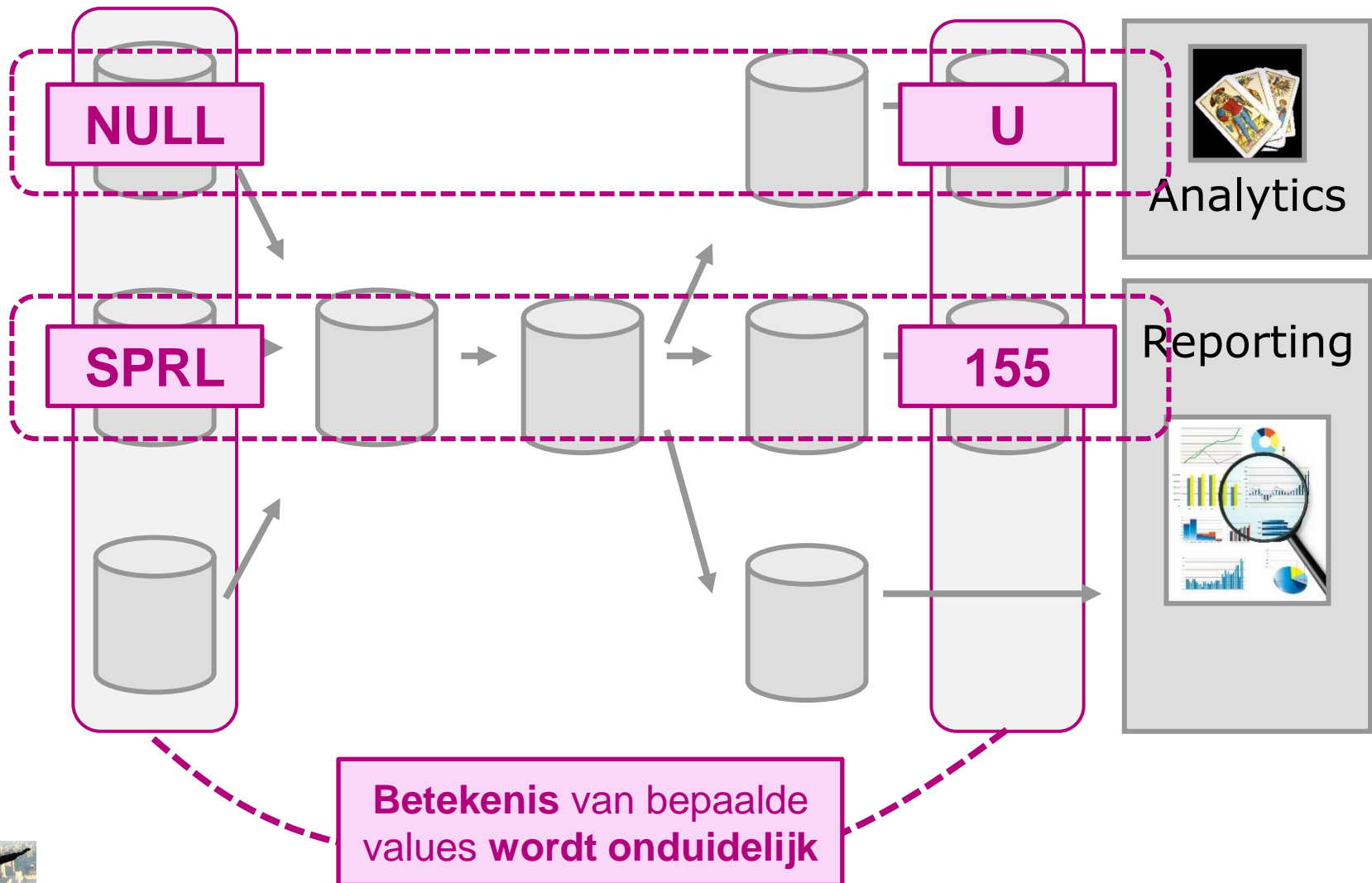
## Complexe en **trage** architectuur



# Barrière 2: Dataquality management

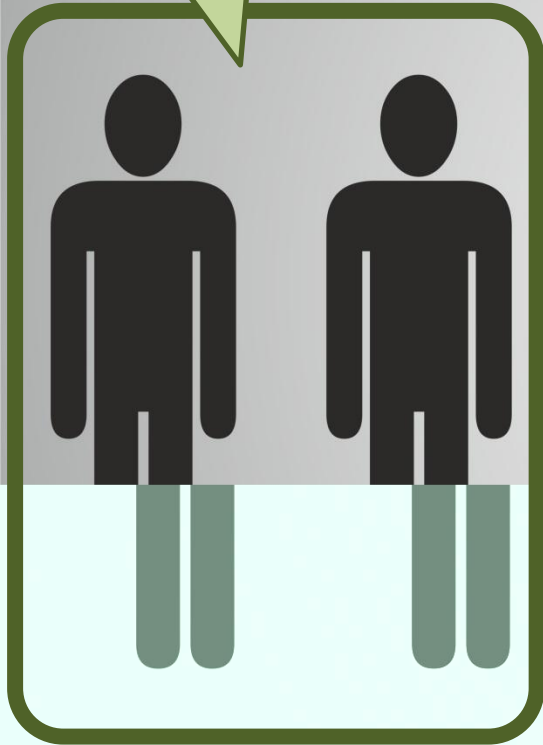
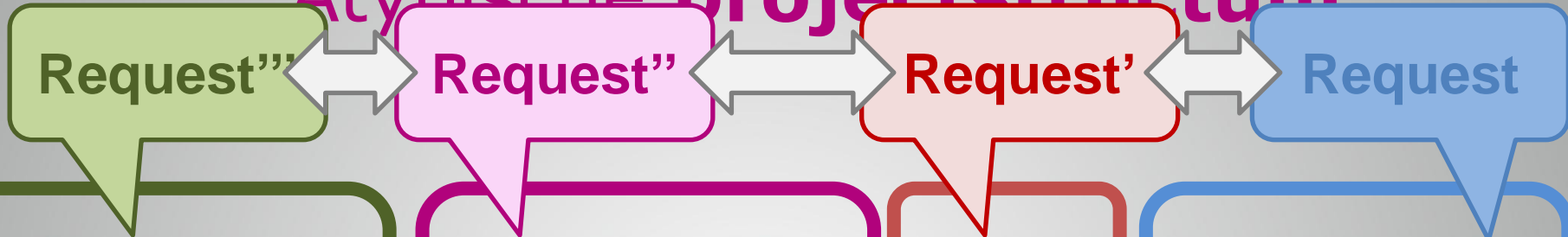


# Barrière 2: Dataquality management

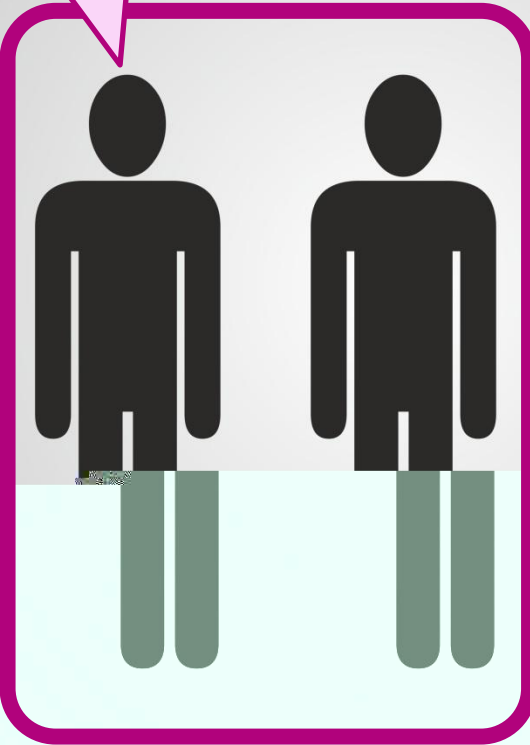


# Barrière 3:

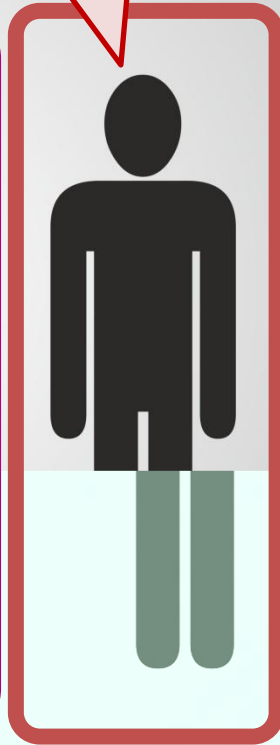
Atypische projectstructuur



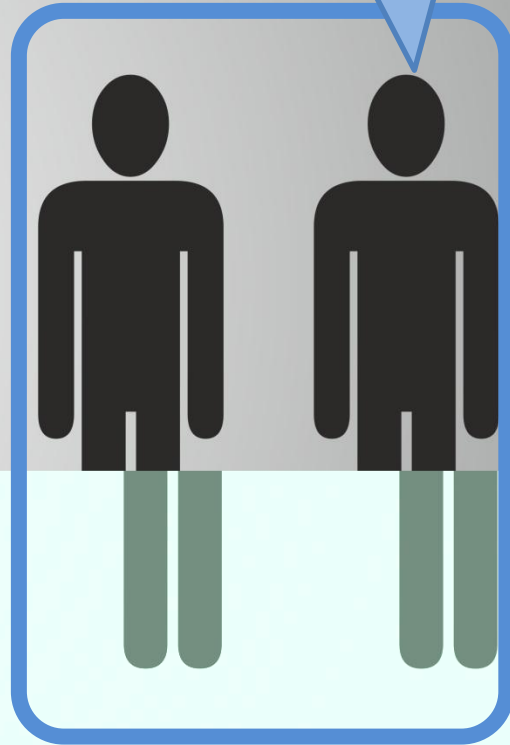
Operationele systemen



Data warehouse

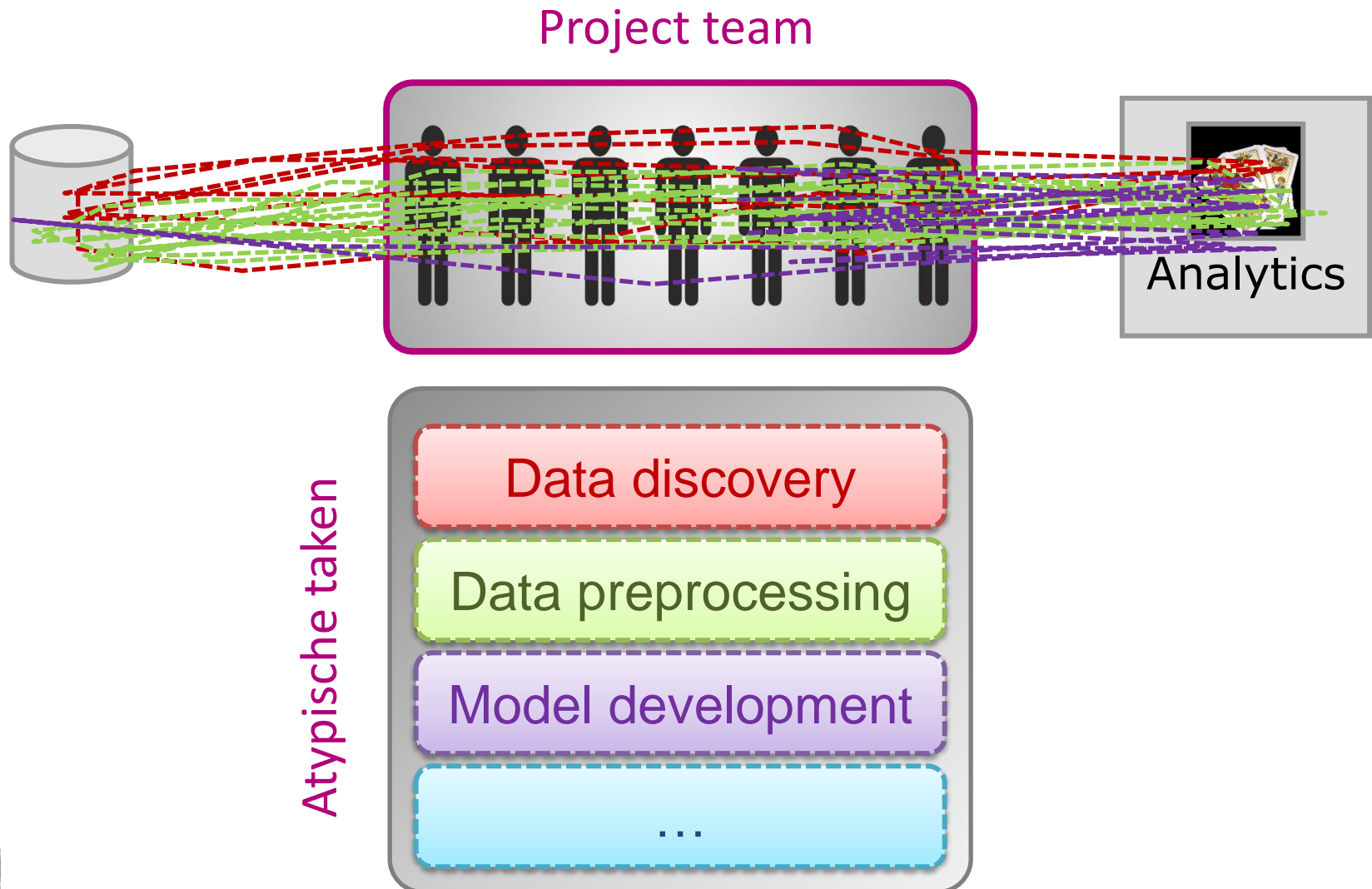


Data-mart

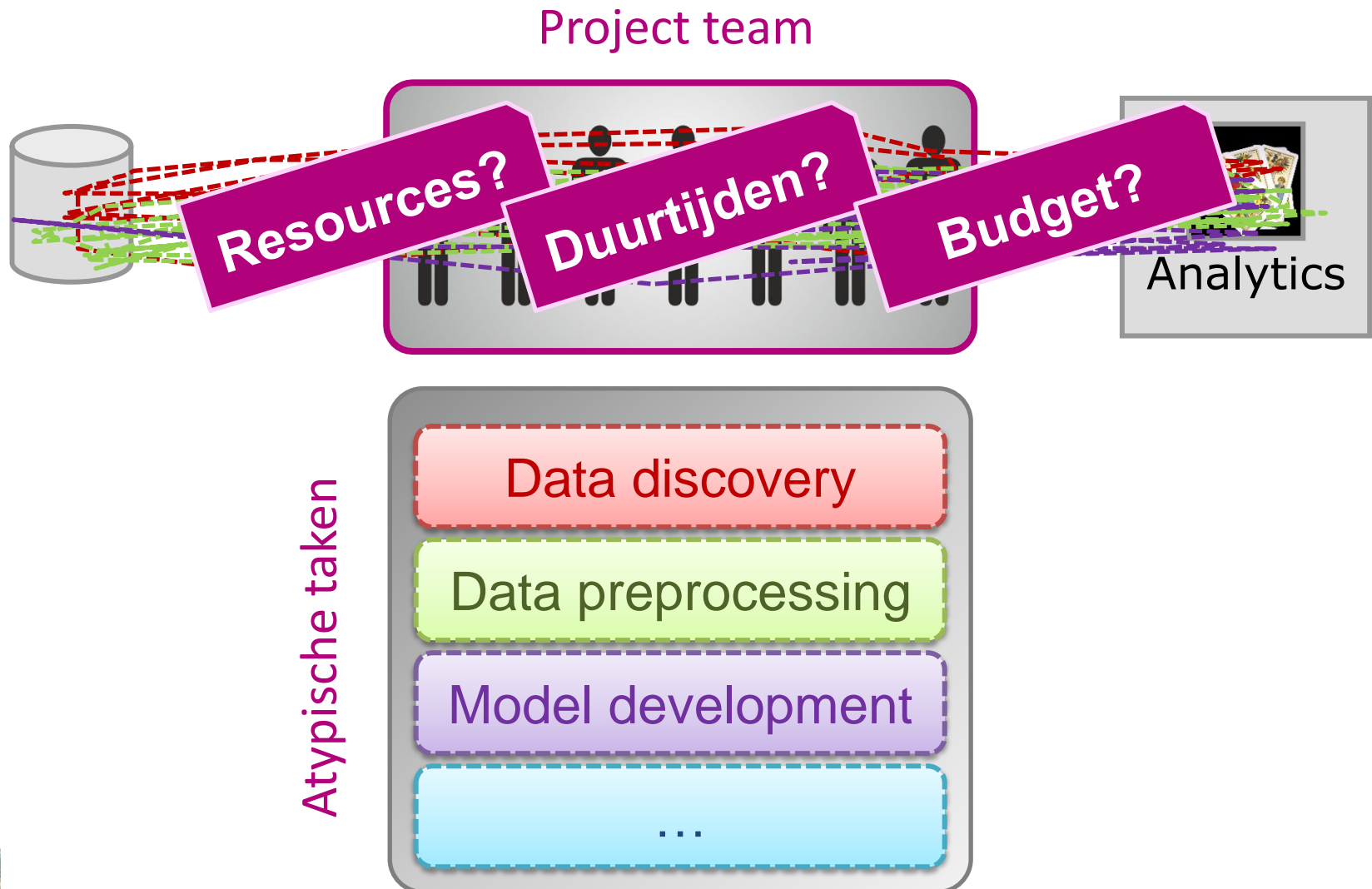


Analytics

# Barrière 3: Atypische projectstructuur



# Barrière 3: Atypische projectstructuur



# Streamlining Analytics

**Predictive analytics**  
**De data supply chain**



**Barrières bij de introductie van analytics**



**Hardware appliances voor analytics**

**Data quality**

**Analytics project management**



# Hardware appliances?



# Hardware appliances?

An appliance has **one function** (toasting, making coffee, ...), is **easy to use** and **its internals are not relevant**



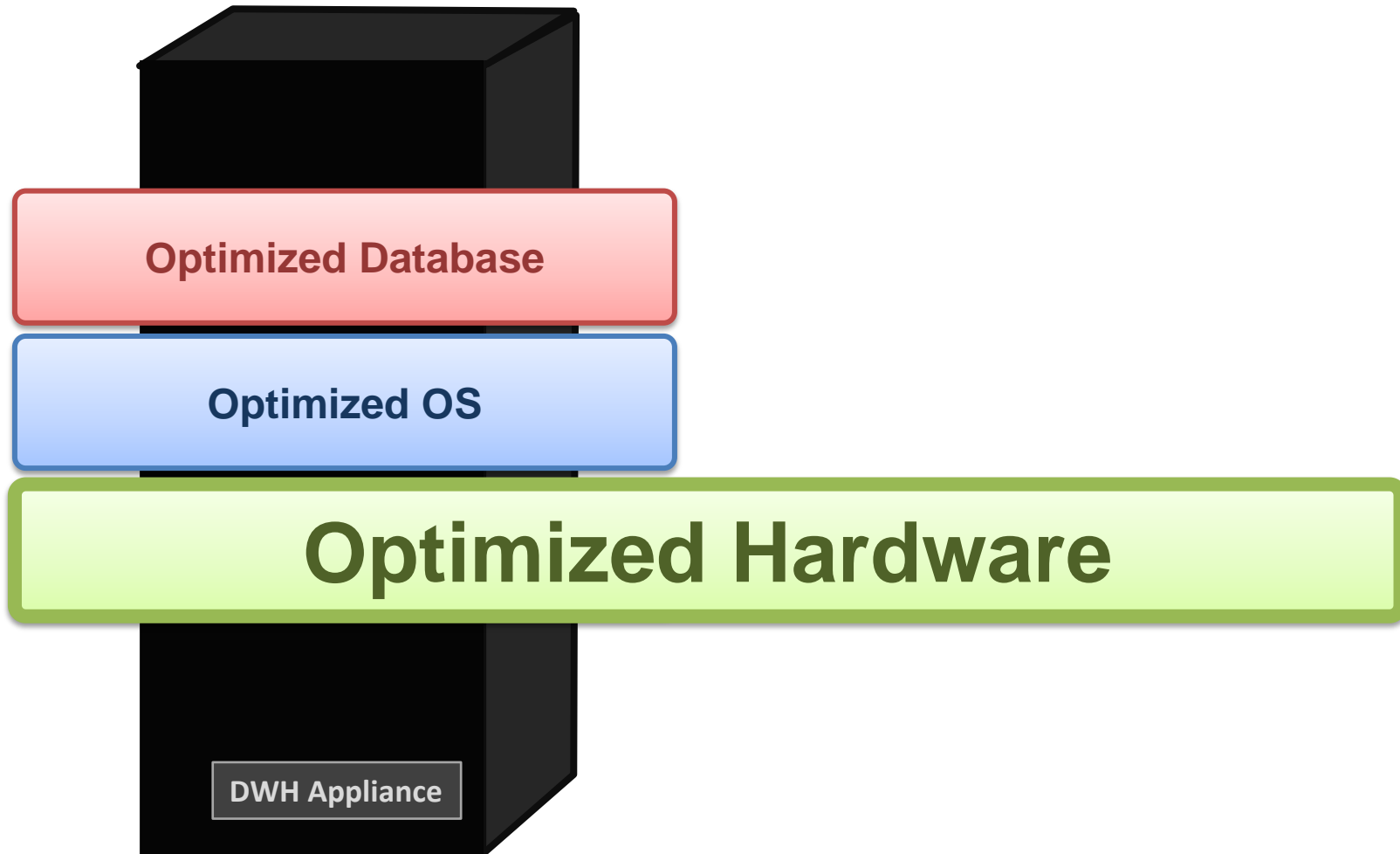
A **hardware appliance for analytics** has one function (executing complex queries), is **easy to install, manage, tune, ...** and its internals are not relevant



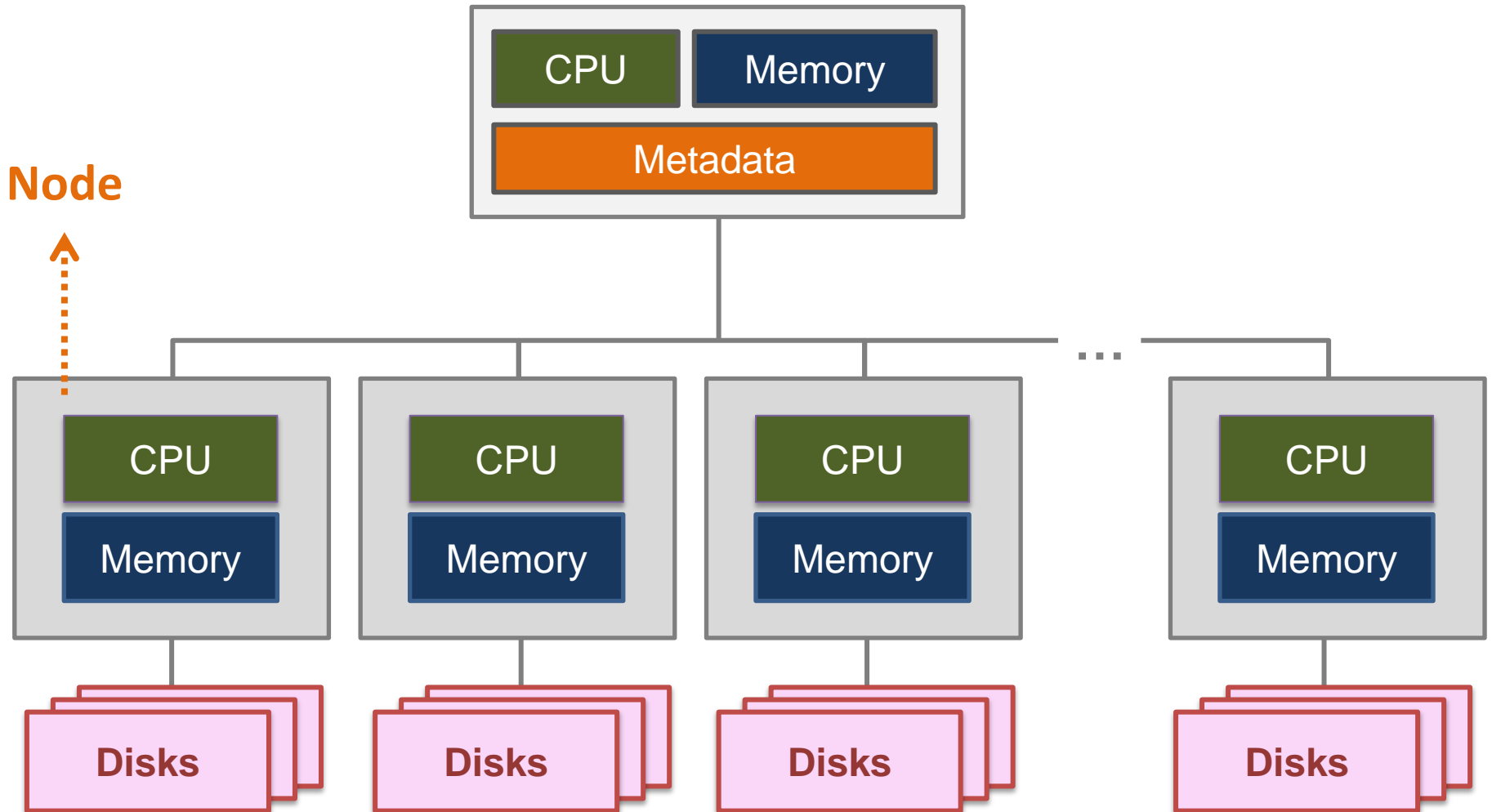
# « Supercomputers voor analytics »



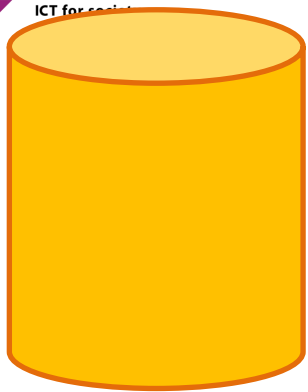
# Hardware appliance voor analytics (Data Warehouse appliance)



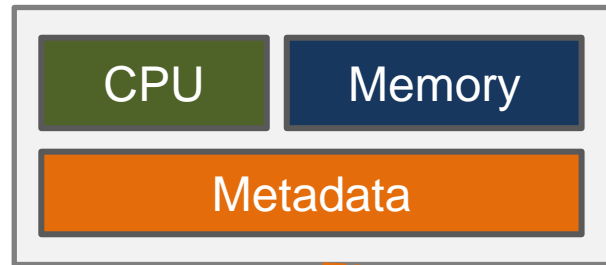
# Massively Parallel Processing (MPP)



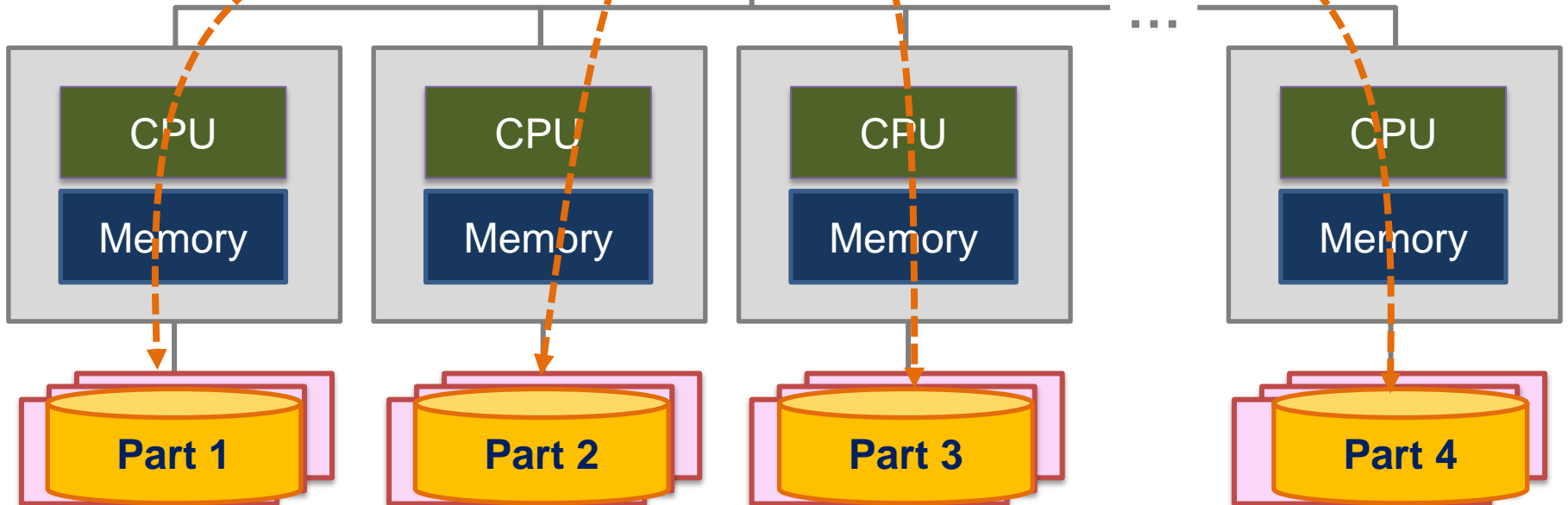
# Data loading



1. Profile data



2. Partition & distribute data

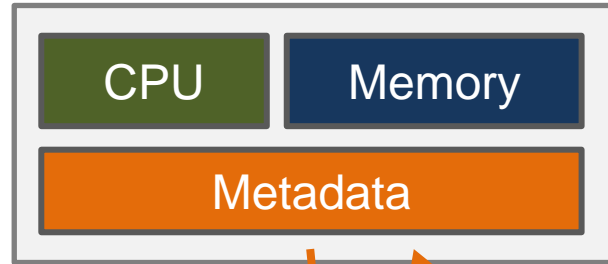


# Data querying

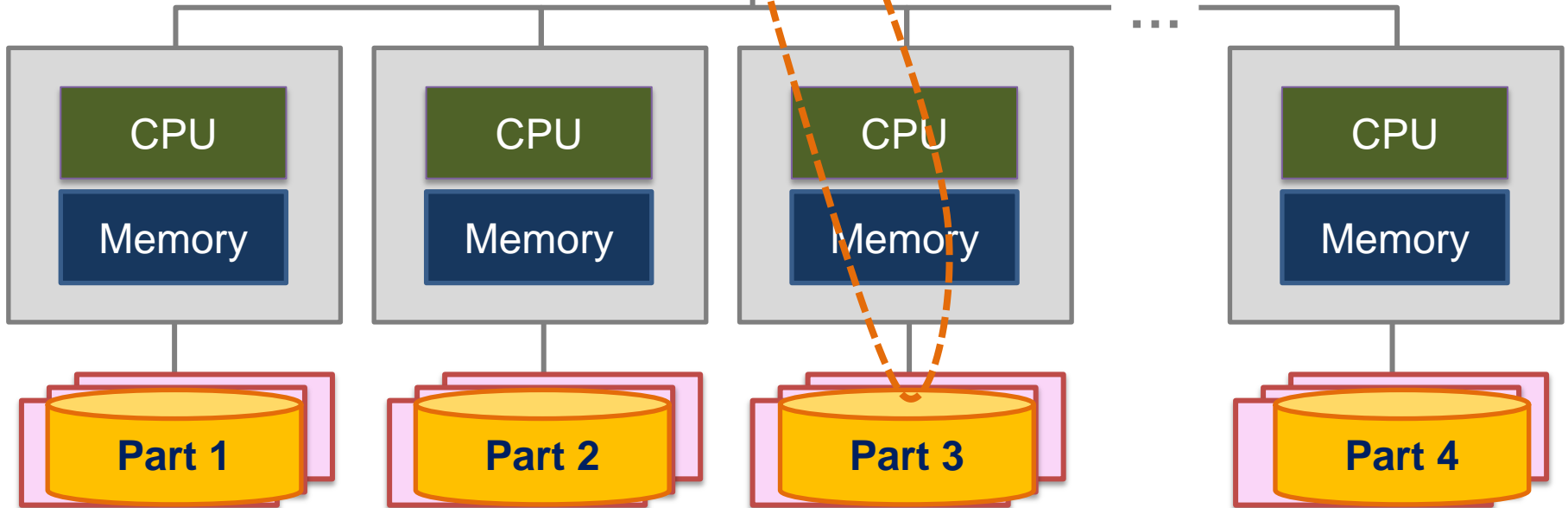
```
SELECT * FROM X  
WHERE id=1000
```

3. Combineer resultaten

1. Bouw queryplan



2. Voer query uit



# Data querying

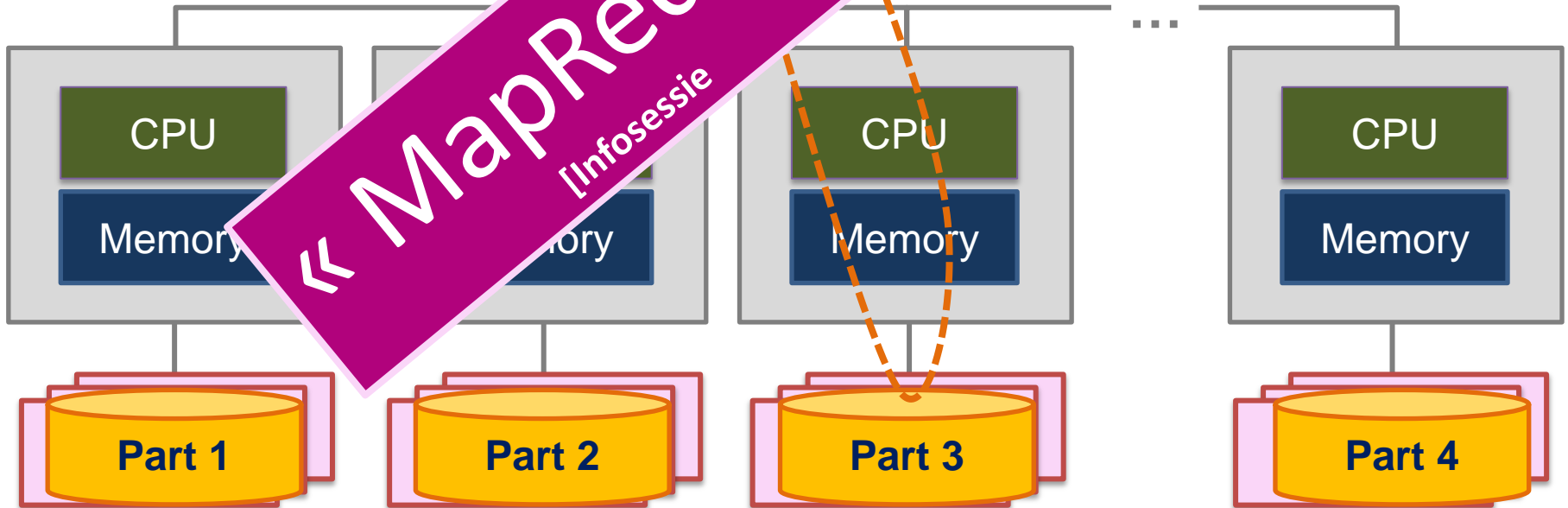
```
SELECT * FROM X  
WHERE id=1000
```

1. Bouw queryplan



3. Computatie resultaten

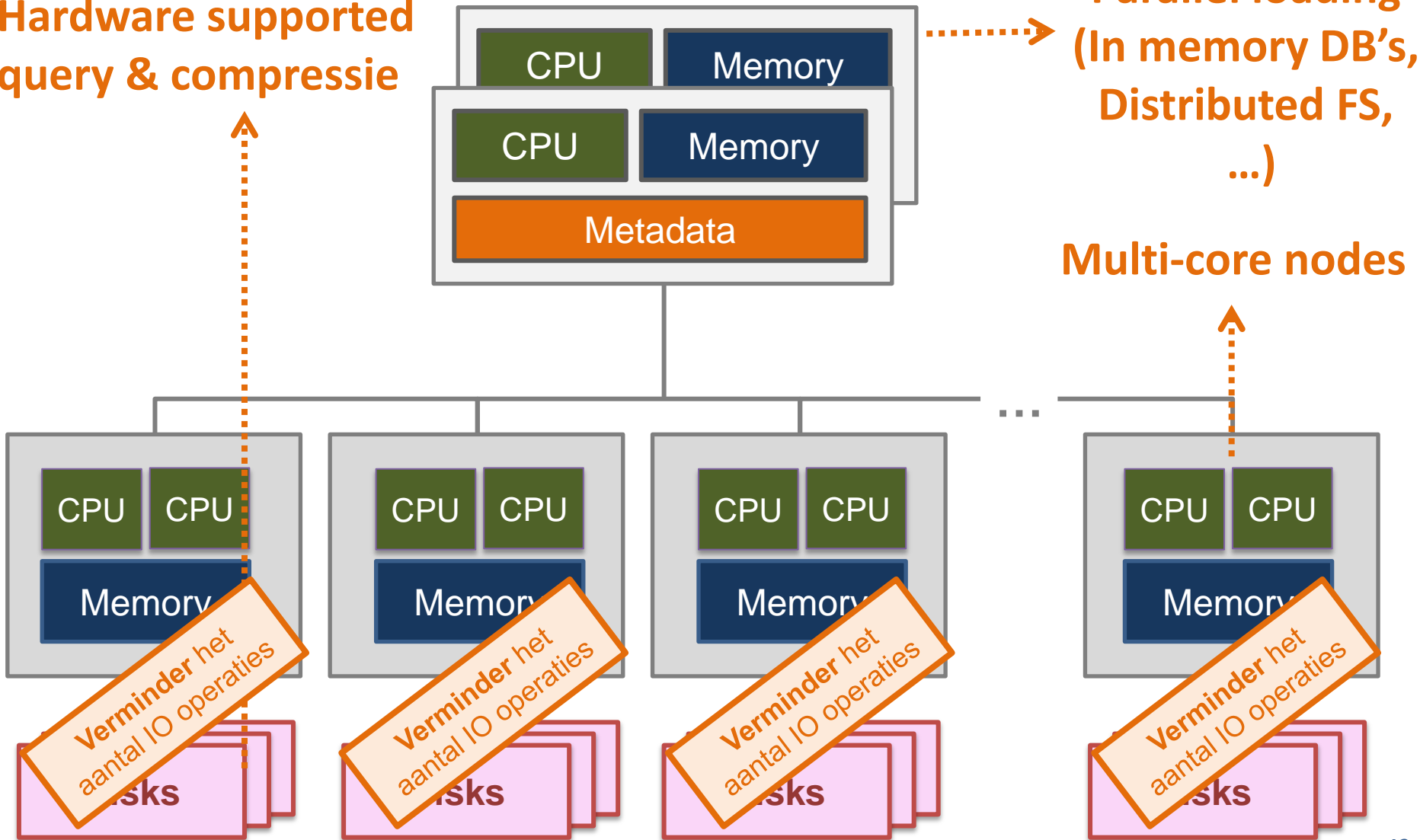
2. Voer query uit



« MapReduce in a box »  
[Infosessie]

# Extra optimalisaties

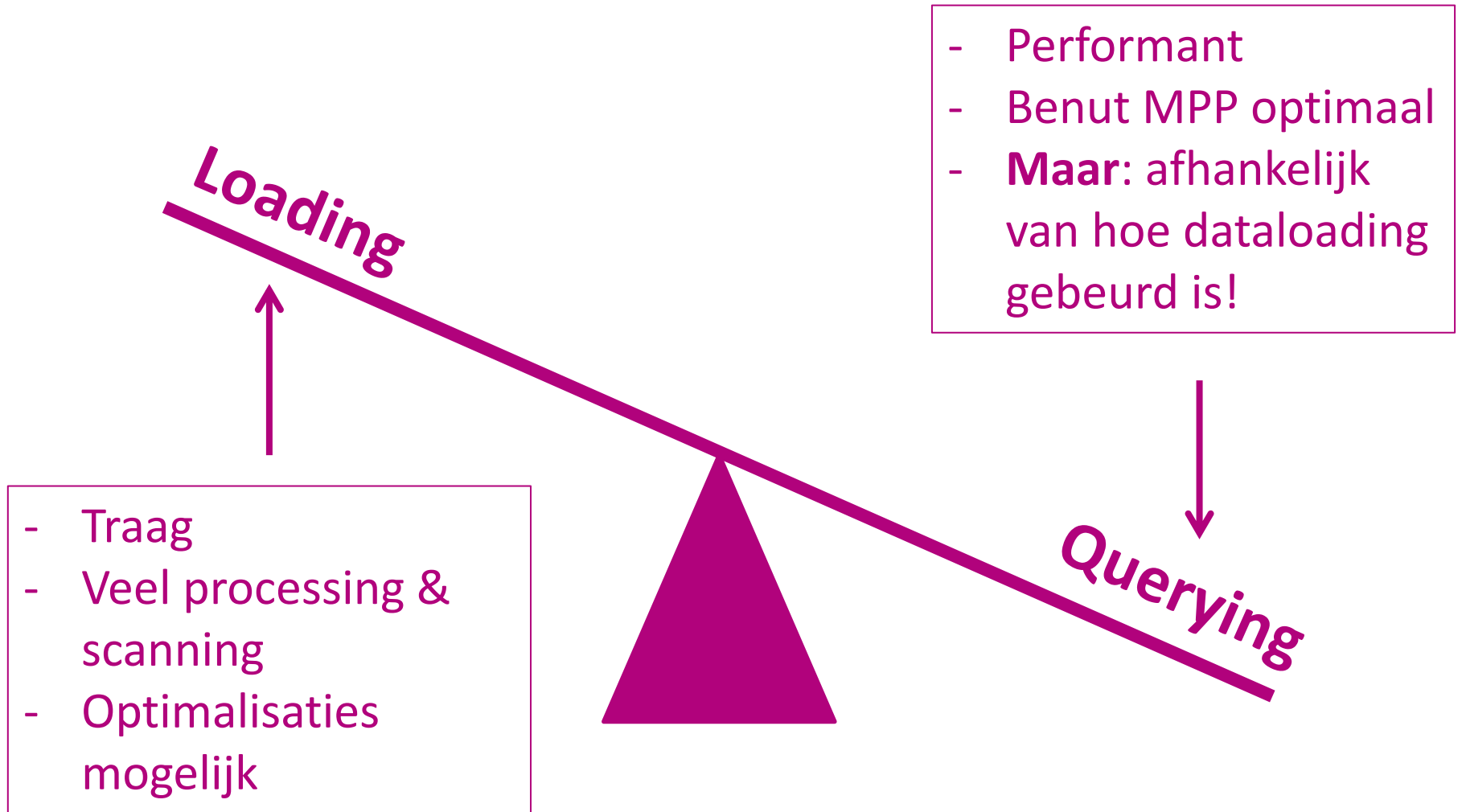
Hardware supported query & compressie



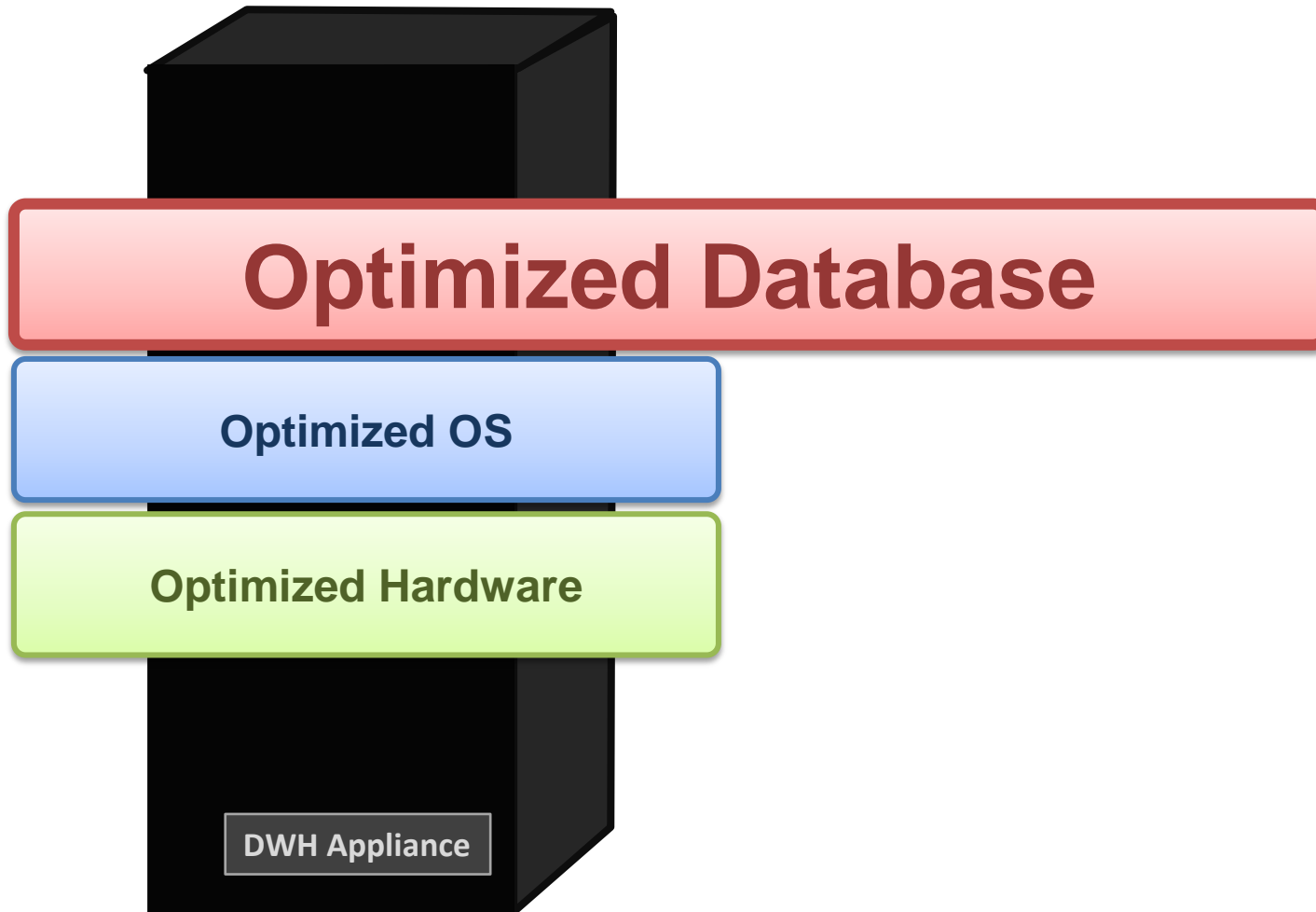
Parallel loading  
(In memory DB's,  
Distributed FS,  
...)

Multi-core nodes

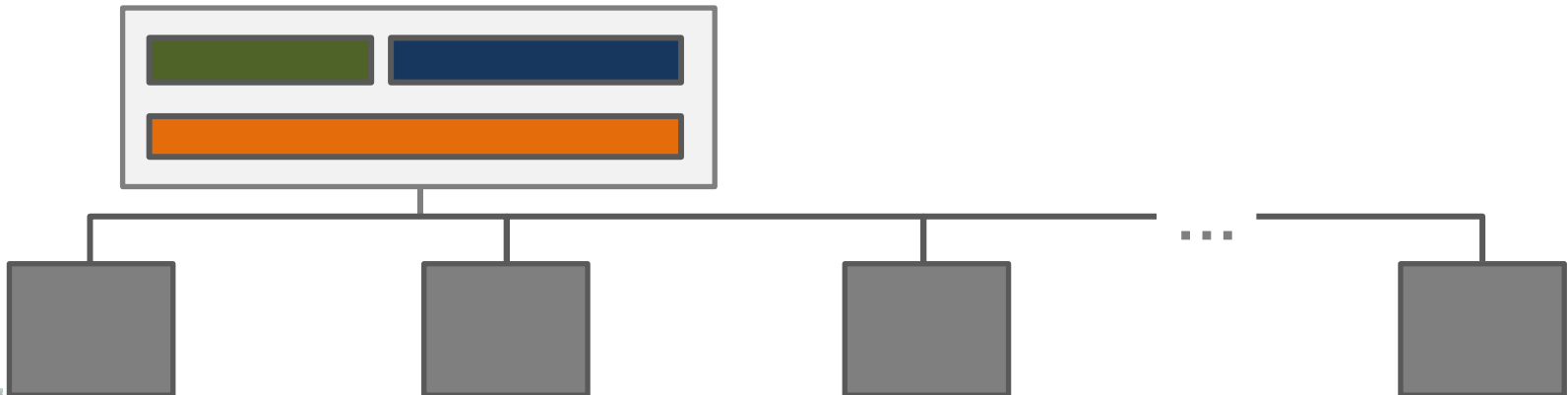
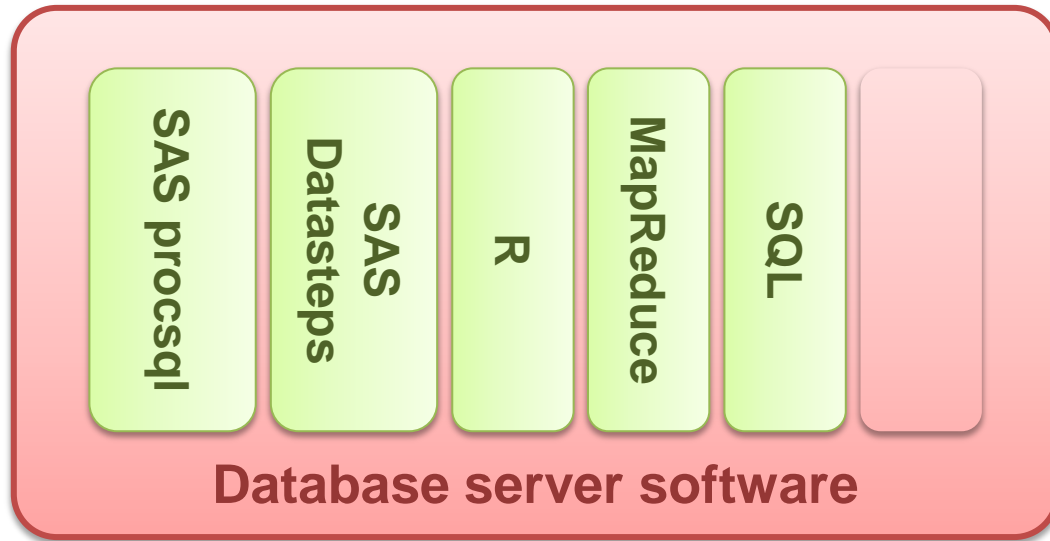
# Data querying vs. data loading



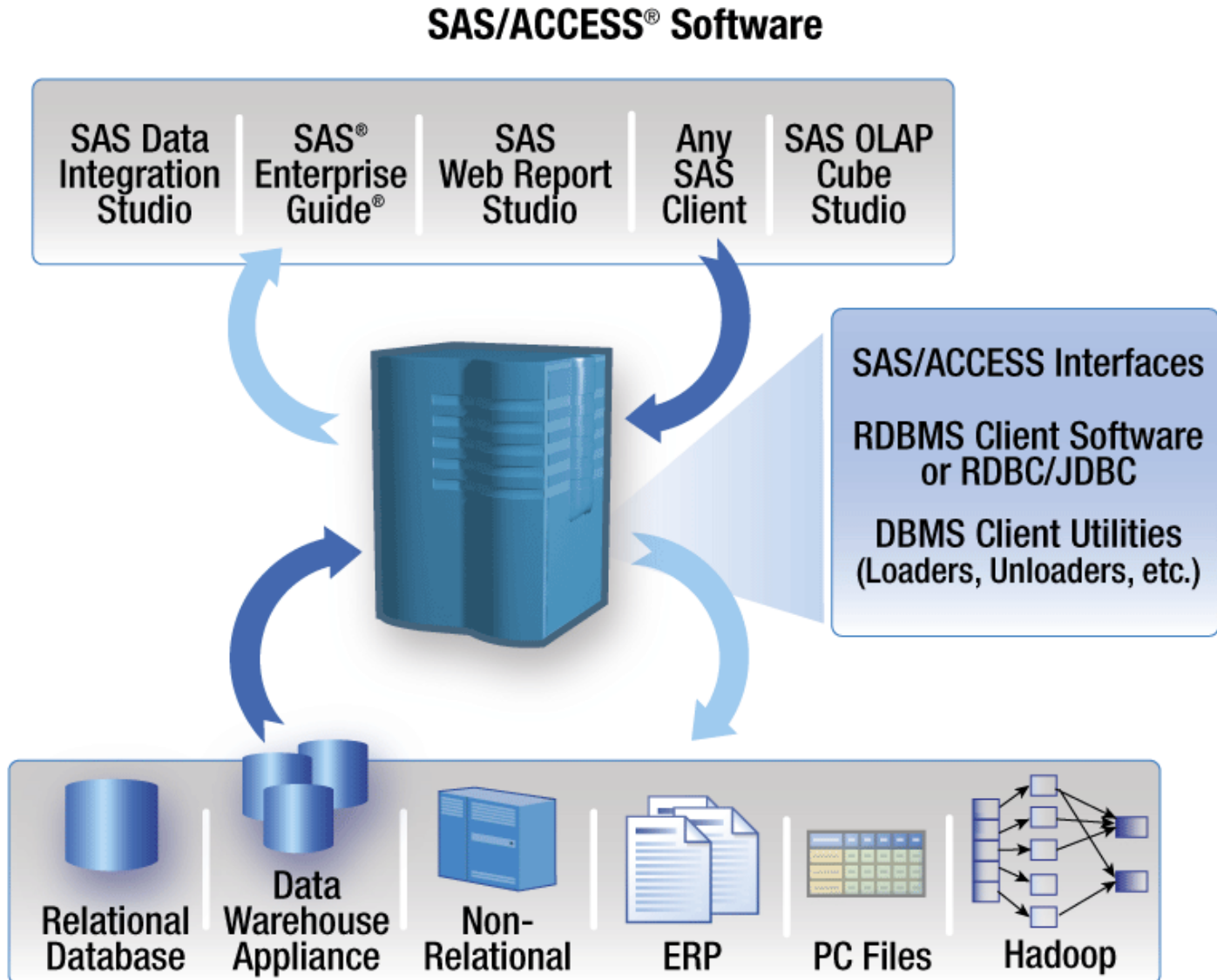
# Hardware appliance voor analytics (Data Warehouse appliance)



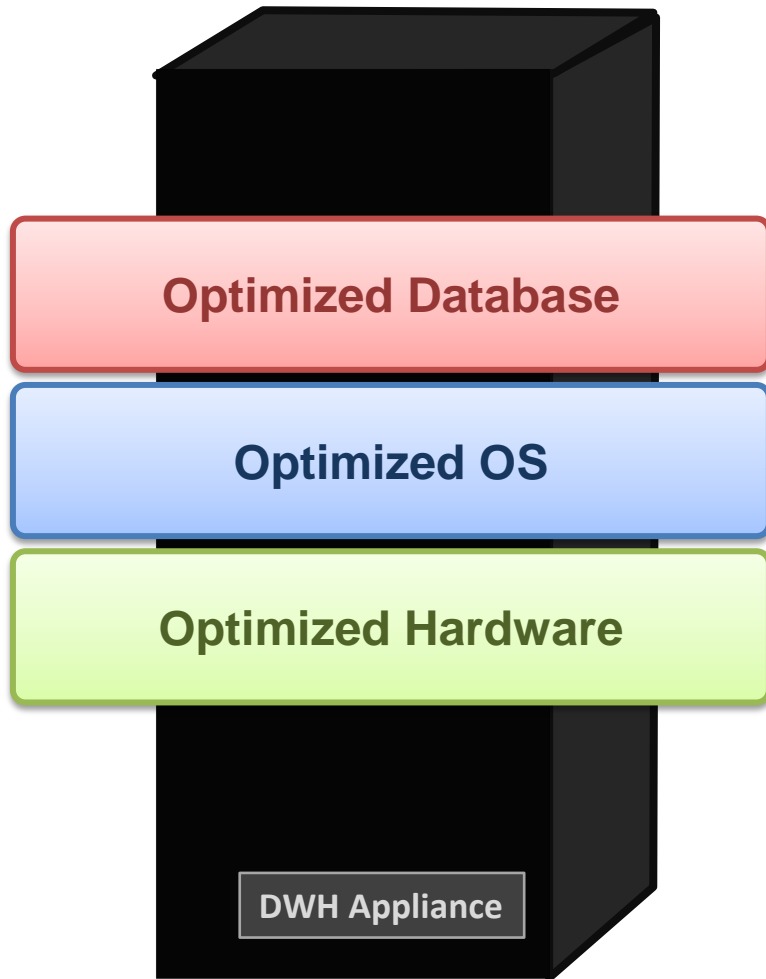
# In-database analytics



# SAS connectoren



# DWH Appliances samengevat



- Heel performante MPP architectuur
- Queries worden snel uitgevoerd
- Data loading gaat vrij traag
- Connectoren voor 1 of meerdere (analytics) softwares



# Marktoverzicht

EMC<sup>2</sup>



GREENPLUM<sup>®</sup>

TERADATA. ASTER



NETEZZA<sup>®</sup>

an IBM<sup>®</sup> Company

ORACLE<sup>®</sup>

EXADATA

READY

SAP  
HANA



VERTICA

An HP Company



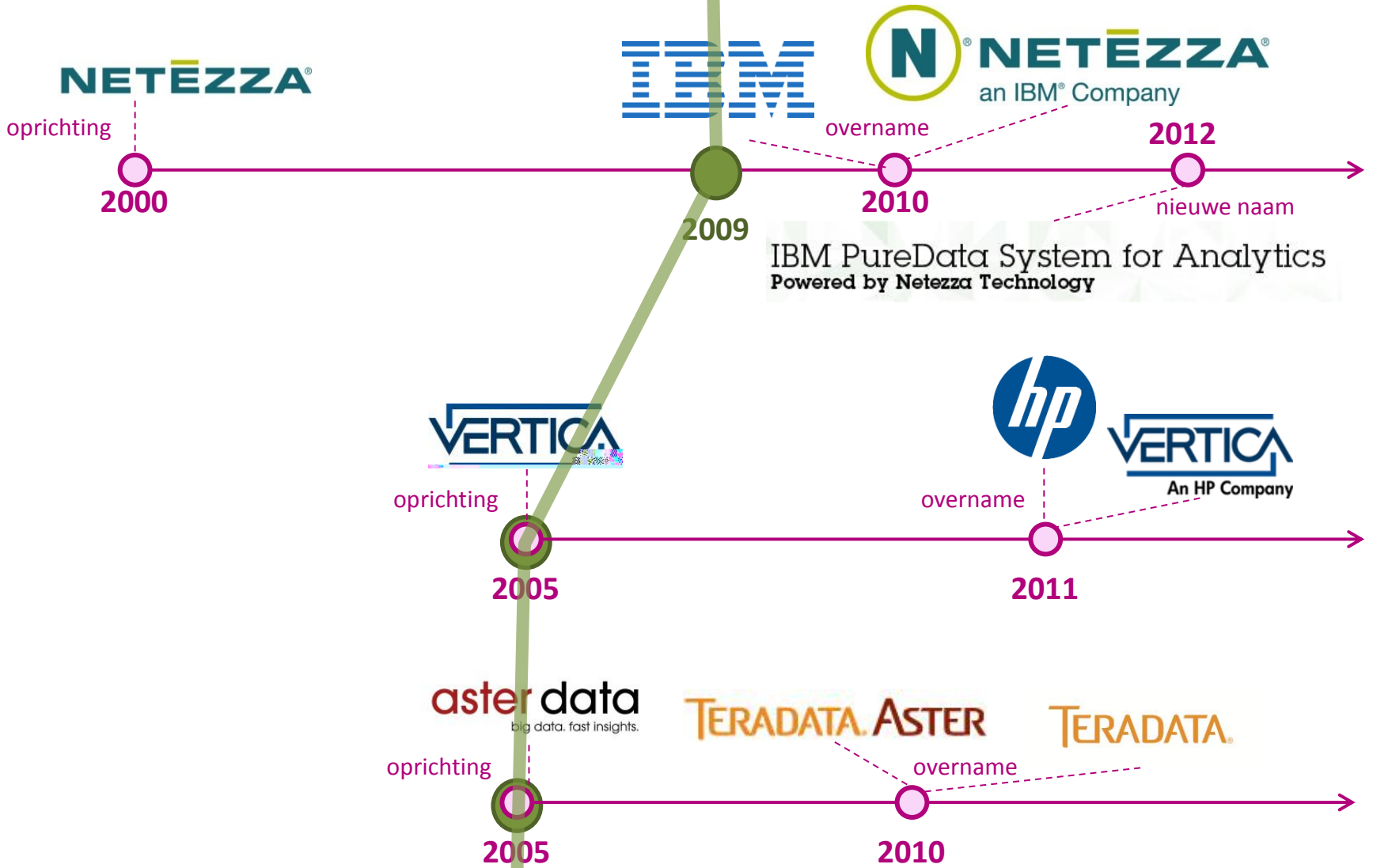
Microsoft<sup>®</sup>

SQL Server<sup>®</sup> 2012

Parallel Data Warehouse

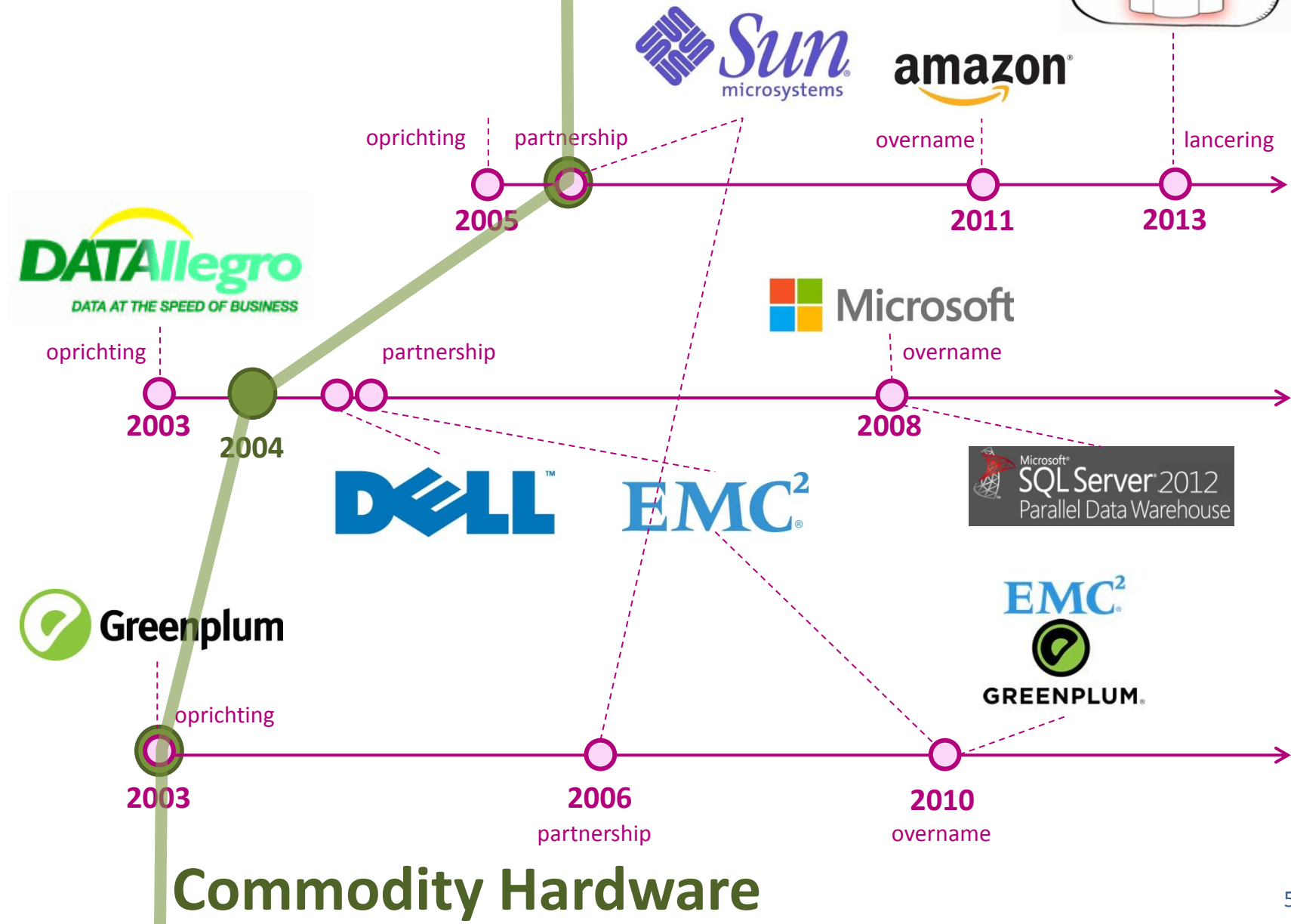


# Markt in beweging



Proprietary Hardware | Commodity Hardware

# Markt in beweging



## Commodity Hardware

**Generic**

**Meer info over dit kwadrant:  
contacteer Smals Onderzoek**

**Software-stack**

**Accelerator**

**Architectuur**

**MPP**

**Specific**



# TERADATA ASTER nCluster

## Analytic Applications & Front-End Tools

Custom Apps



SAS, Others



BI Tools



## Aster Data nCluster

Pattern

Time Series

Graph

Text

Custom Analytics

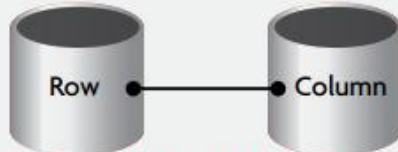
Partner Analytics

...

SQL

SQL-MapReduce

Data Analytic Services: Workload Management, Data Flow, Scaling



Massively Parallel Data Stores

Java

Perl

C/C++



MapReduce

Python

R

Embedded Analytic Processing



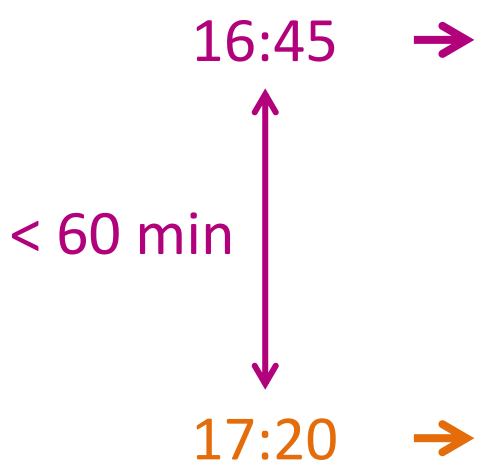
# Get all the flights to London for which another flight exists to London that leaves within an hour on the same day

Alle vluchten | Mijn vluchten | Laatste bijgewerkt op: 13:52

Tijd | Vluchtnr. | Bestemming

Vluchten op: 17/06/2013

Tijd	Vlucht	Bestemming	Status	Track
16:40	LH5579	Hamburg	Voorzien	16:40 ...
16:45	SQ2825	Londen Heathrow	Voorzien	16:45 ...
17:00	JAF1711	Palma Mallorca	Voorzien	17:00 ...
17:05	SN2705	Bazel Mulhouse	Voorzien	17:05 ...
17:05	LX4537	Bazel Mulhouse	Voorzien	17:05 ...
17:05	SN2721	Geneve	Voorzien	17:05 ...
17:05	LX4559	Geneve	Voorzien	17:05 ...
17:10	LH2291	München	Voorzien	17:10 ...
17:10	A31429	München	Voorzien	17:10 ...
17:10	SN7059	München	Voorzien	17:10 ...
17:15	SN2307	Stockholm Bromma	Voorzien	17:15 ...
17:15	EY7217	Stockholm Bromma	Voorzien	17:15 ...
17:15	TF2222	Stockholm Bromma	Voorzien	17:15 ...
17:20	EZ81526	Geneve	Voorzien	17:20 ...
17:20	BA397	Londen Heathrow	Voorzien	17:20 ...
17:25	H847	Cork International	Voorzien	17:25 ...
17:30	SK1594	Copenhagen Kastrup	Voorzien	17:30 ...
17:30	LH1045	Frankfurt	Voorzien	17:30 ...
17:30	SN7007	Frankfurt	Voorzien	17:30 ...
17:30	HQ1992	Kos	Voorzien	17:30 ...
17:30	DE198	Kos	Voorzien	17:30 ...
17:30	JAF1603	Enfidha	Voorzien	17:30 ...
17:30	SN3661	Strasbourg	Voorzien	17:30 ...
17:30	TP7476	Strasbourg	Voorzien	17:30 ...
17:30	UA9905	Strasbourg	Voorzien	17:30 ...
17:35	JAF1651	Pristina	Voorzien	17:35 ...
17:40	AT833	Casablanca	Voorzien	17:40 ...



# Get all the flights to London for which another flight exists to London that leaves within an hour on the same day

## SQL

```
SELECT *
FROM DEPARTURES AS D1

WHERE
DESTINATION = 'London'
AND
DEP_TIME + 60 MINUTES >=
(
SELECT MIN(DEP_TIME)
FROM DEPARTURES AS D2
WHERE DESTINATION = 'London'
AND D2.DEP_TIME > D1.DEP_TIME
AND D2.DEP_DAY = D1.DEP_DAY
)
ORDER BY DEP_TIME
```

## SQL-MapReduce

```
SELECT *
FROM GET_NEXT_FLIGHT_1HR
(ON DEPARTURES PARTITION BY
DESTINATION)
WHERE
DESTINATION = 'London'
```

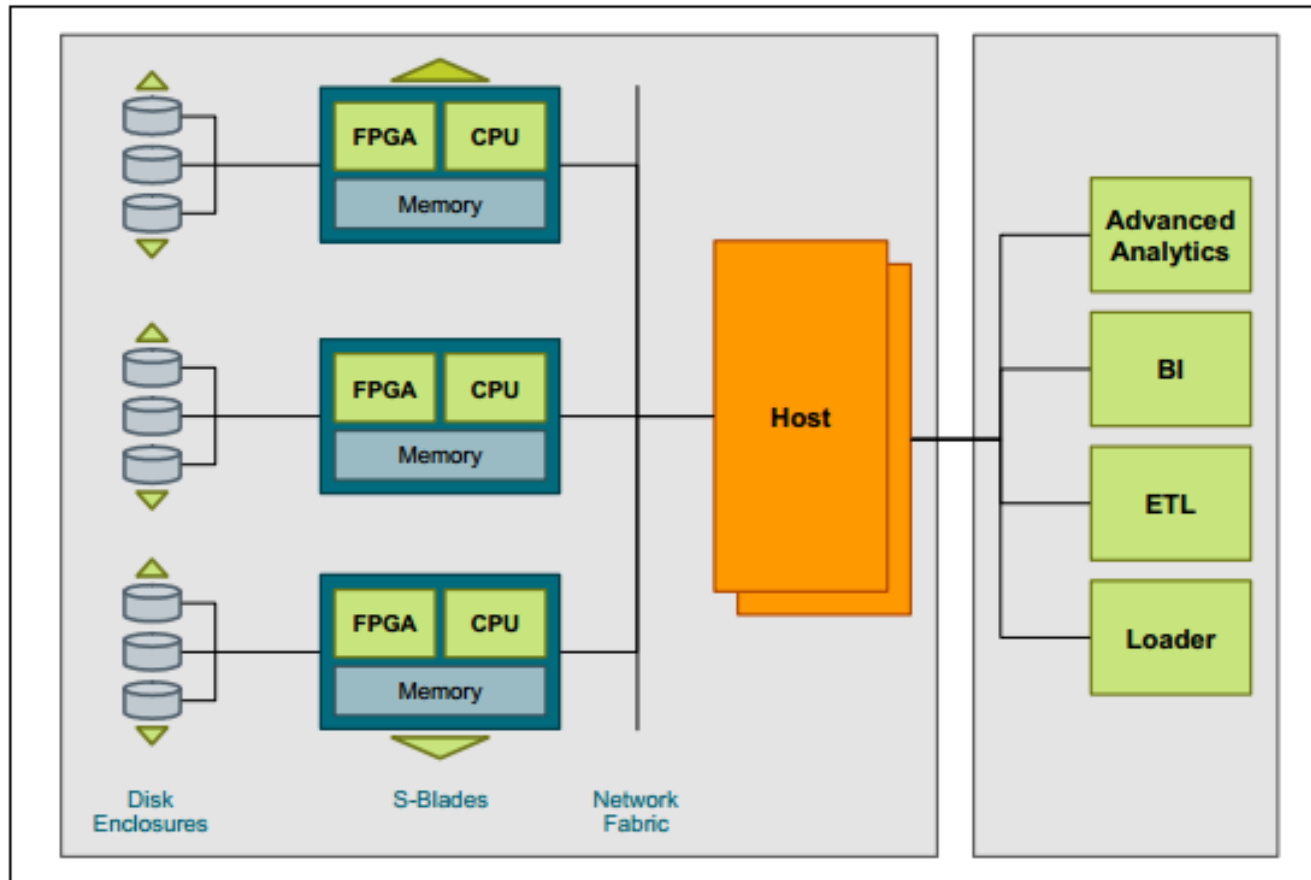
↓

MapReduce table function  
(kunnen zelf gedefinieerd worden)

```
ORDER BY DEP_TIME
```

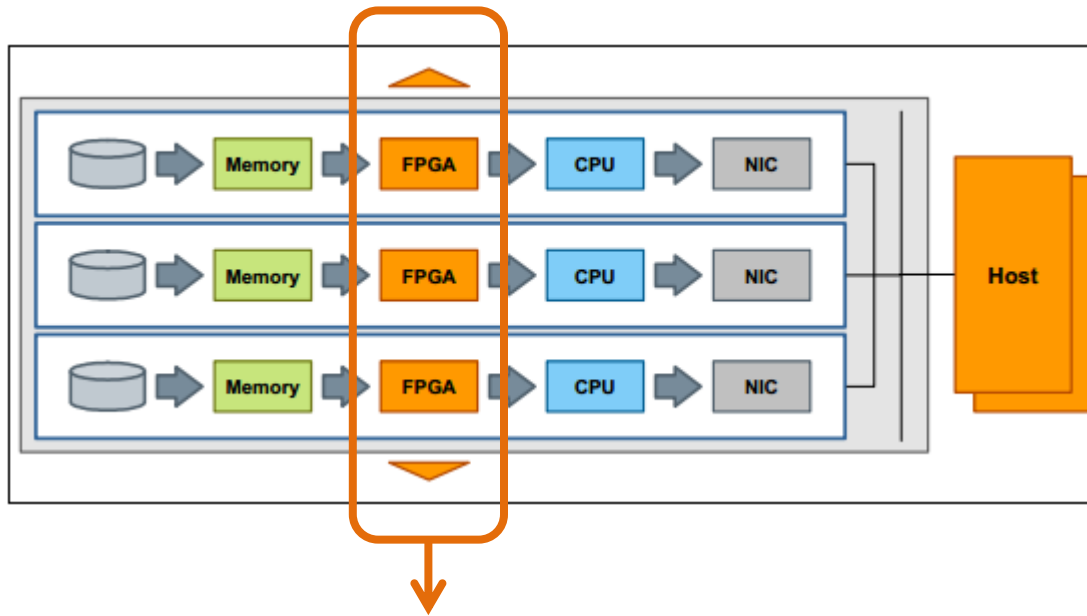


## IBM<sup>®</sup> PureData<sup>™</sup> for Analytics, powered by Netezza technology



IBM® PureData™ for Analytics, powered by Netezza technology

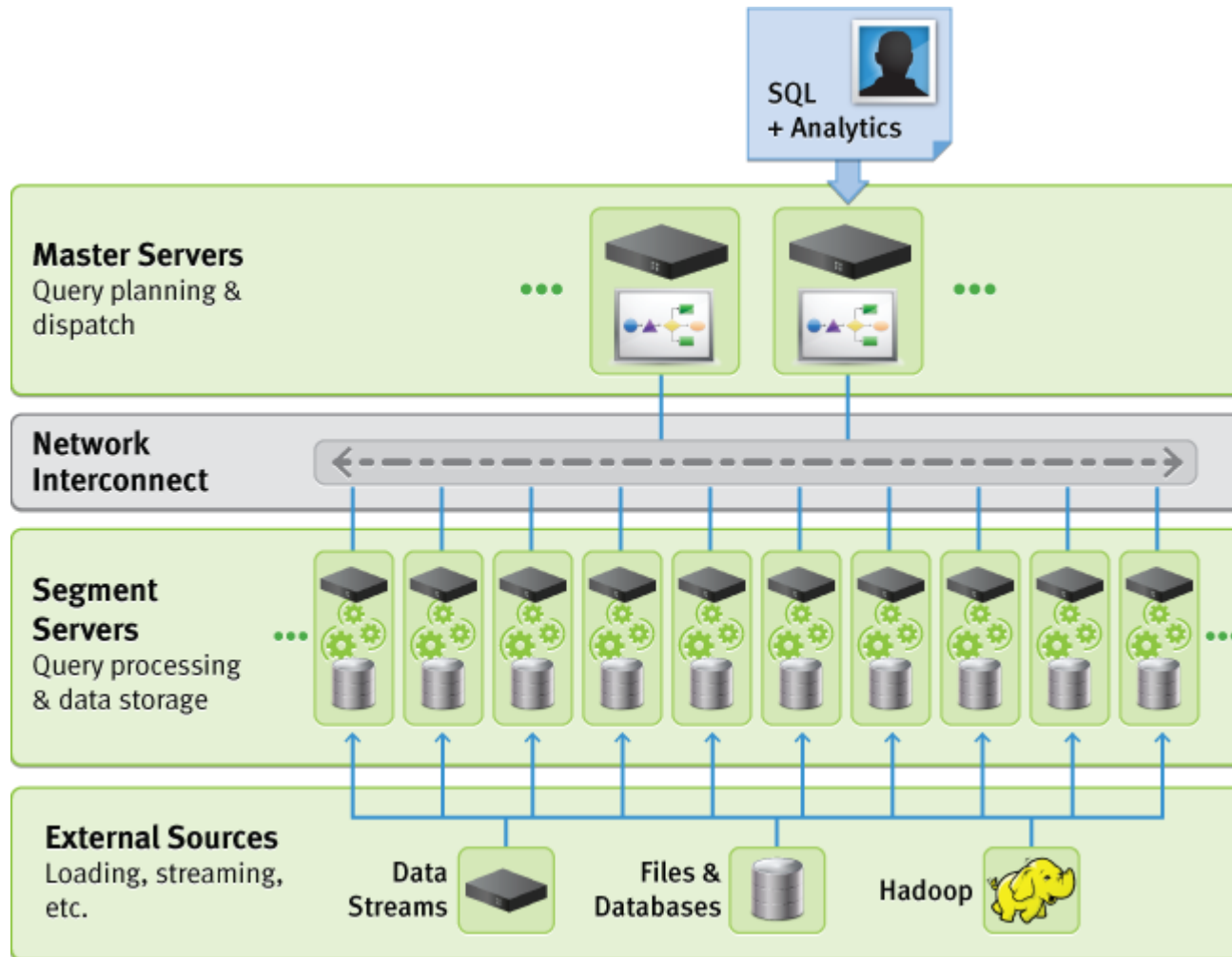
In de worker nodes:



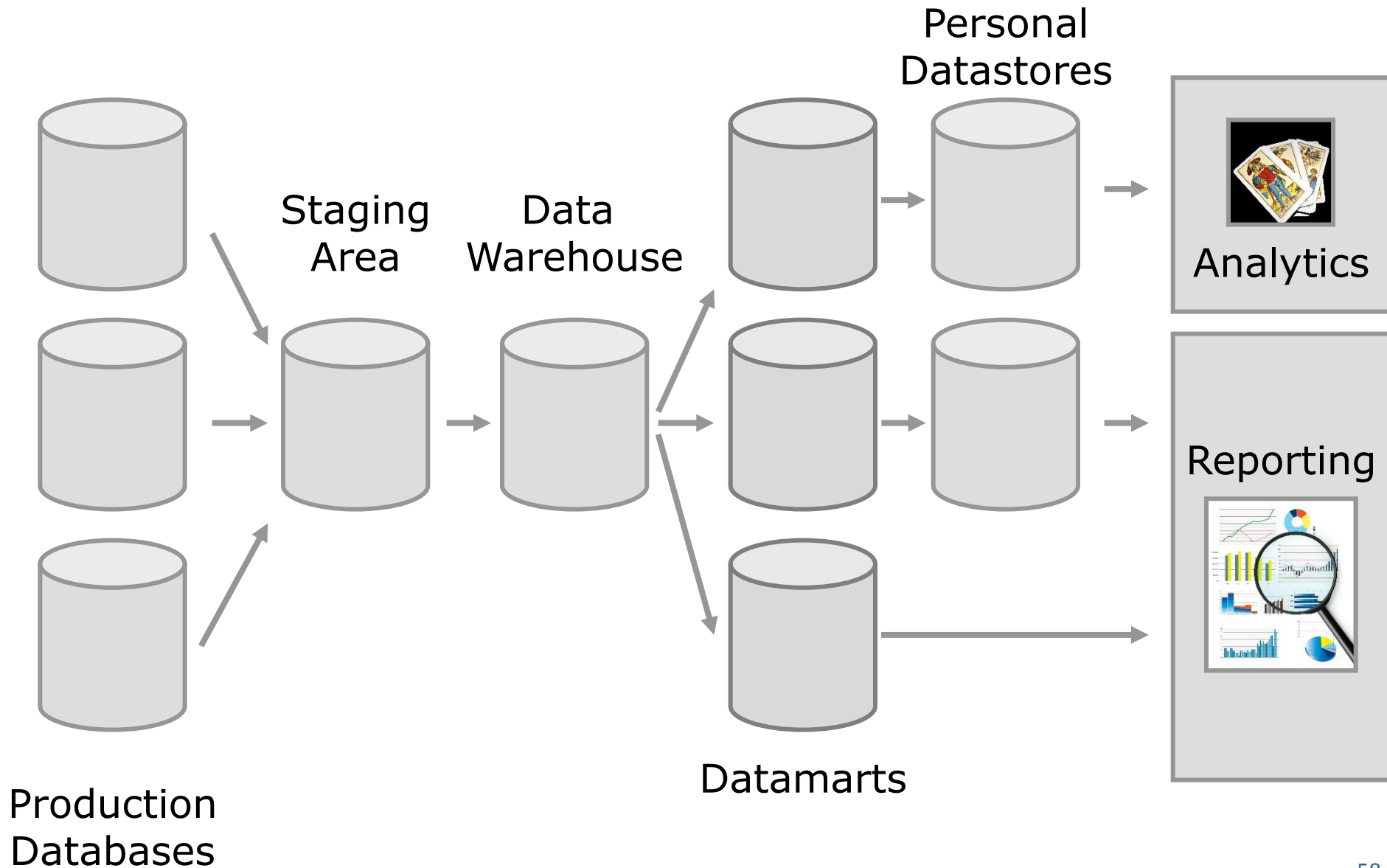
Zeer snelle decompressie en windowing



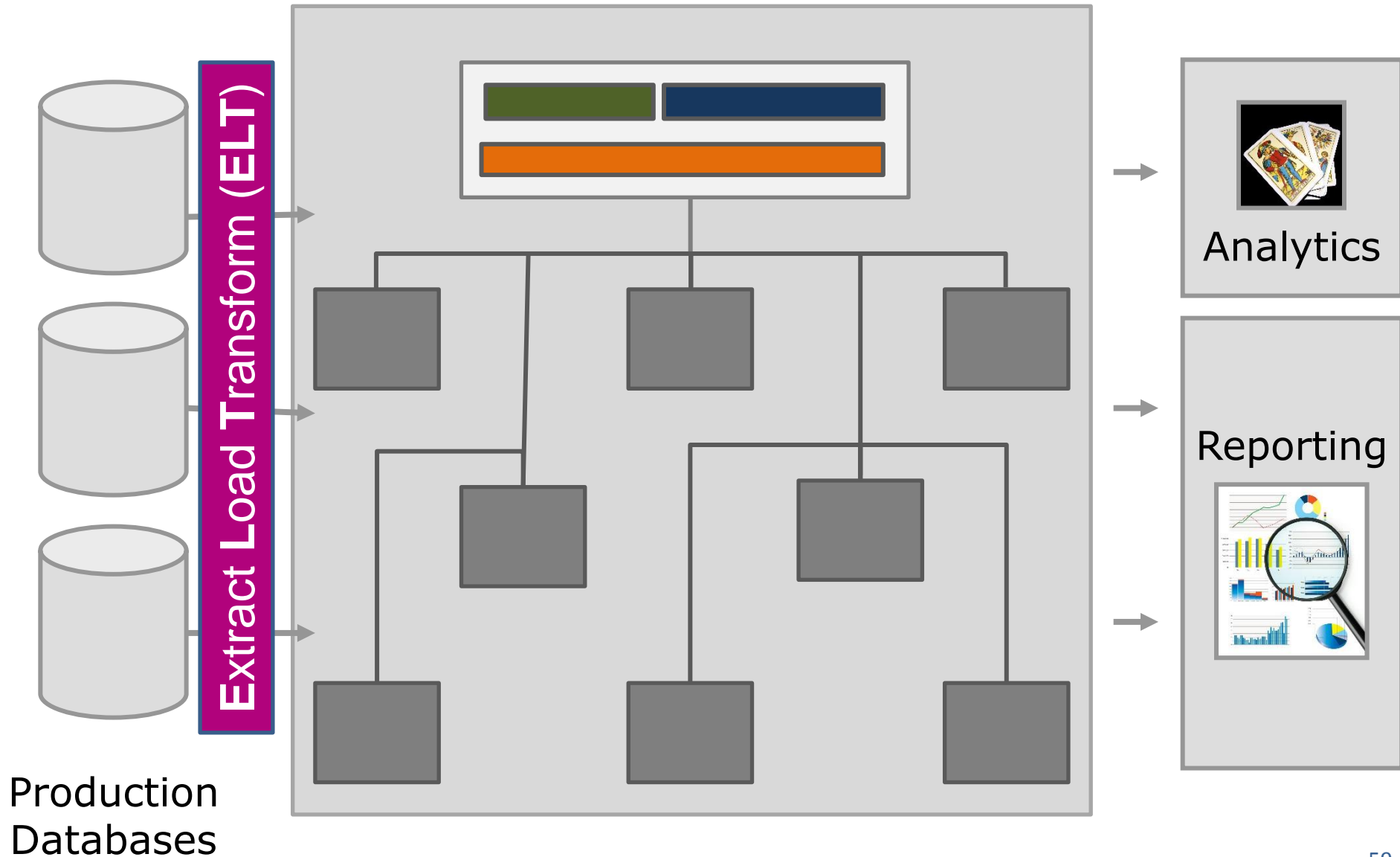
# EMC<sup>2</sup> GREENPLUM.



# DWH Appliances in de **data supply chain**



# DWH Appliances in de data supply chain



### Barrière 1:

# Complexe en trage architectuur

« Data-supply chain » wordt **eenvoudiger** en **performanter**  
met DWH appliances

---

### Barrière 2:

# Dataquality management

**Grote** verbetering mogelijk dankzij vereenvoudigde architectuur  
mits Information/Master Data Management (Documentatie)

---

### Barrière 3:

# Atypische projectstructuur

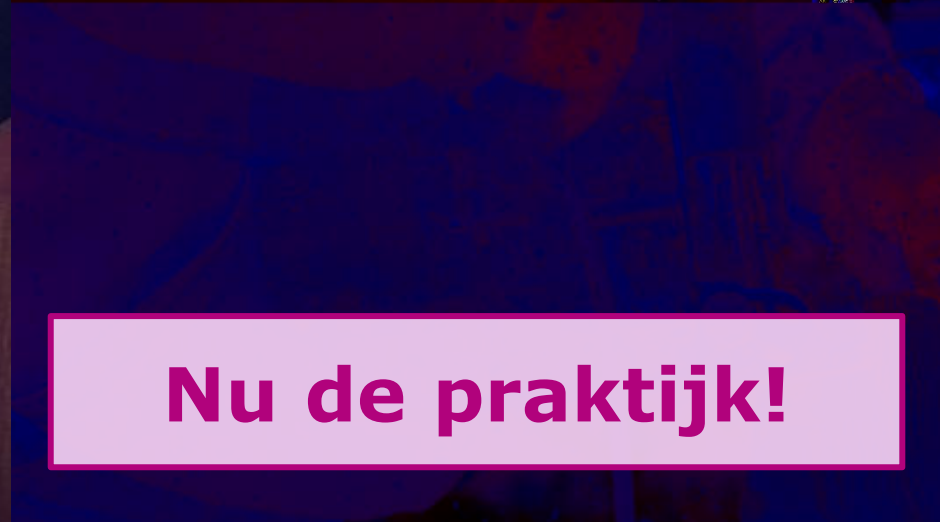
**Een stap in de goeie richting:**  
Analytics-experten hebben makkelijker toegang tot **alle** data  
(met respect voor security/privacy) <sup>60</sup>



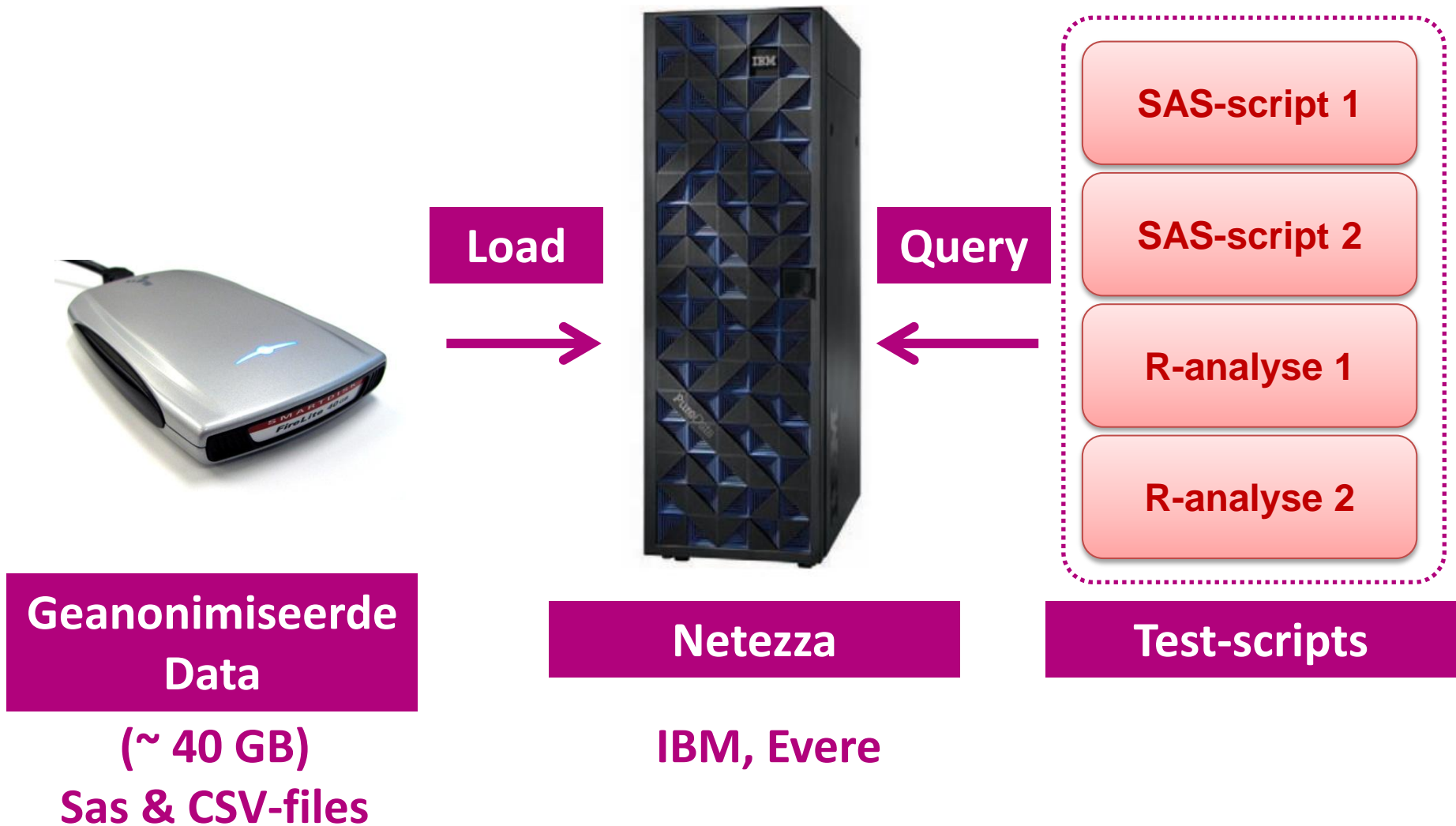
**Dit was de theorie...**



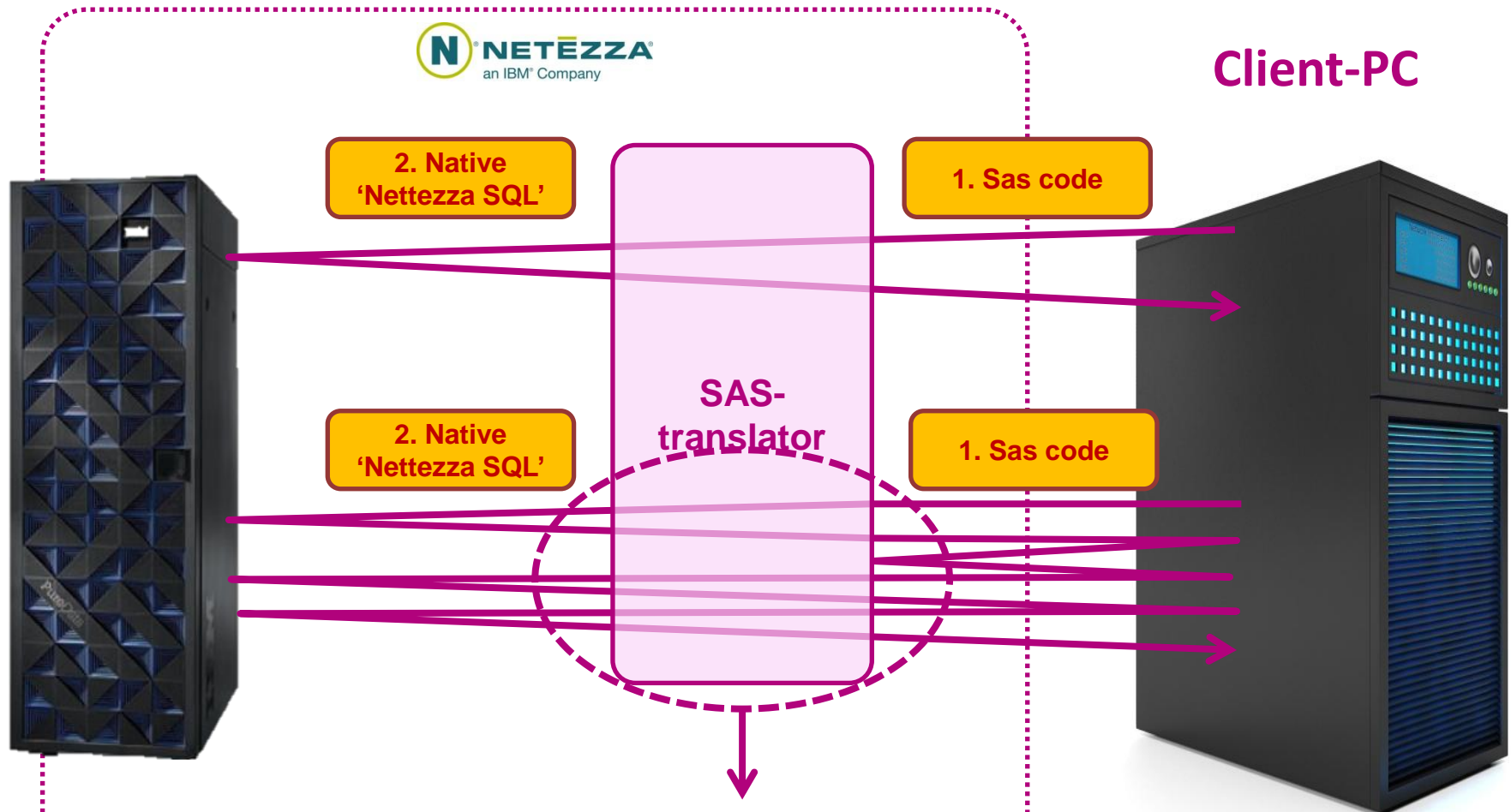
**Nu de praktijk!**



# POC met IBM Netezza



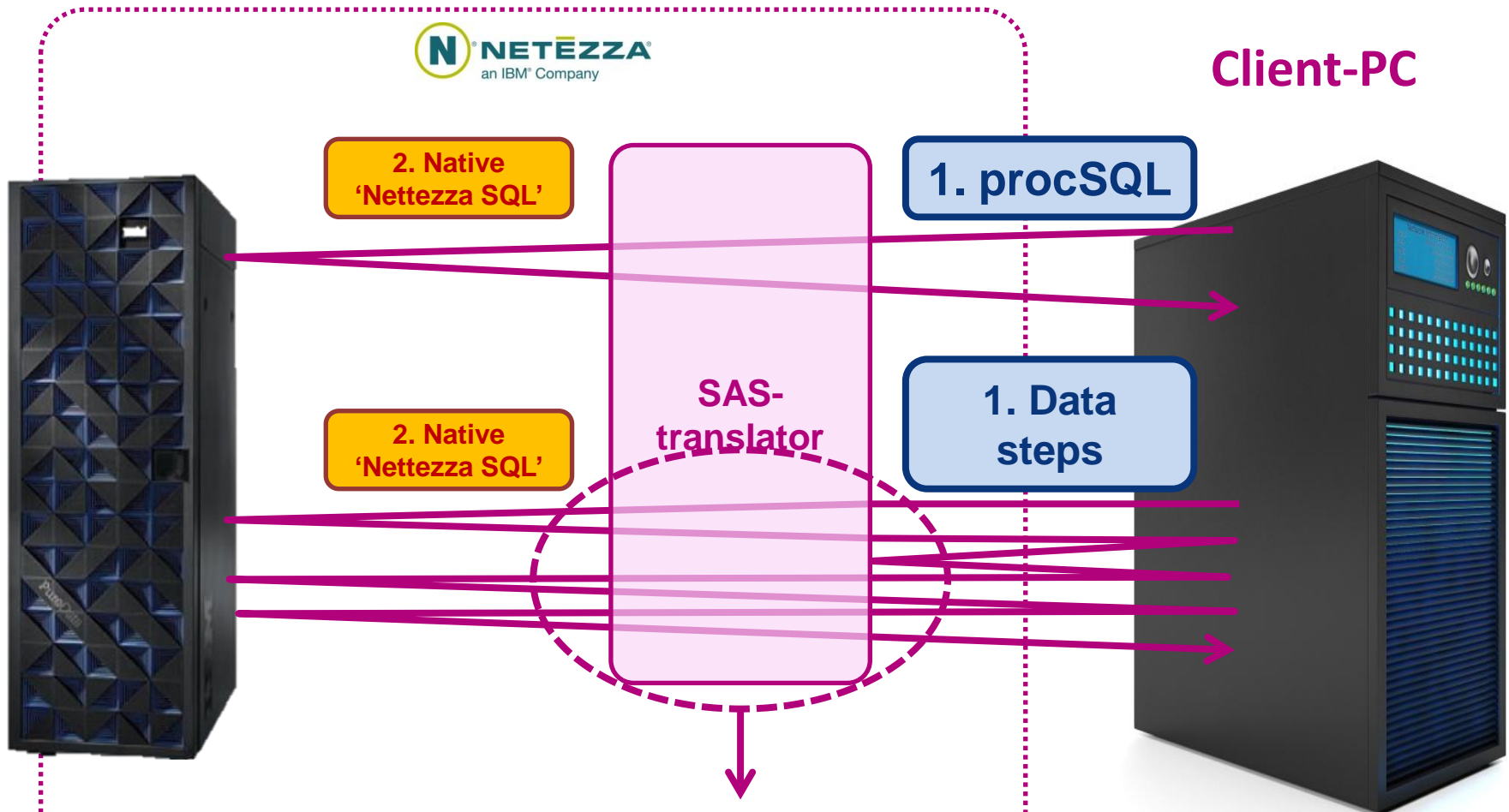
# SAS code aanpassingen



Onaangepaste code kan **niet volledig** vertaald worden :  
Berekeningen worden gespreid over Client-PC & Netezza  
(= veel performantieverlies)



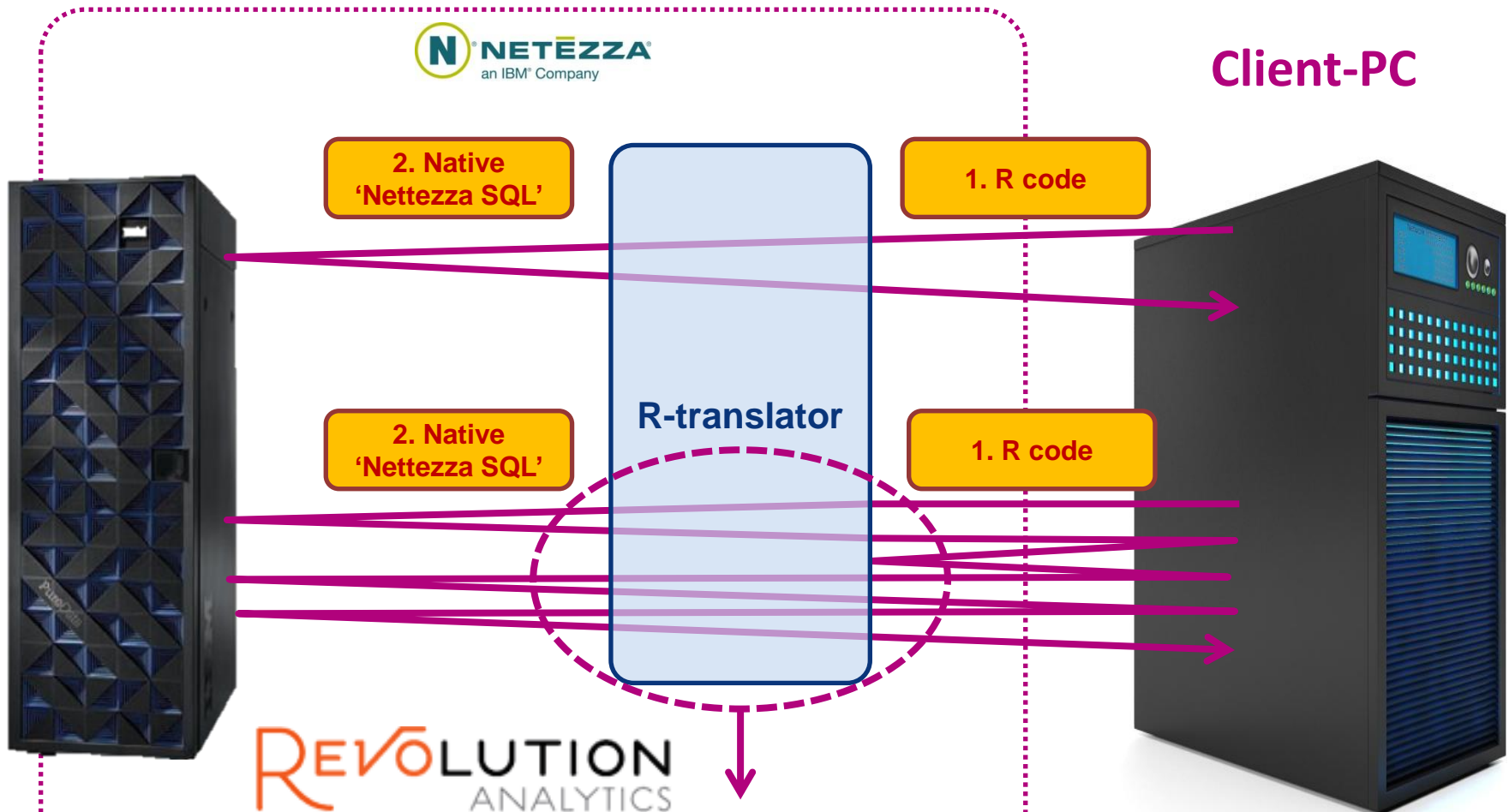
# SAS code aanpassingen



Onaangepaste code kan **niet volledig** vertaald worden :  
Berekeningen worden gespreid over Client-PC & Netezza  
(= veel performantieverlies)



# R code aanpassingen



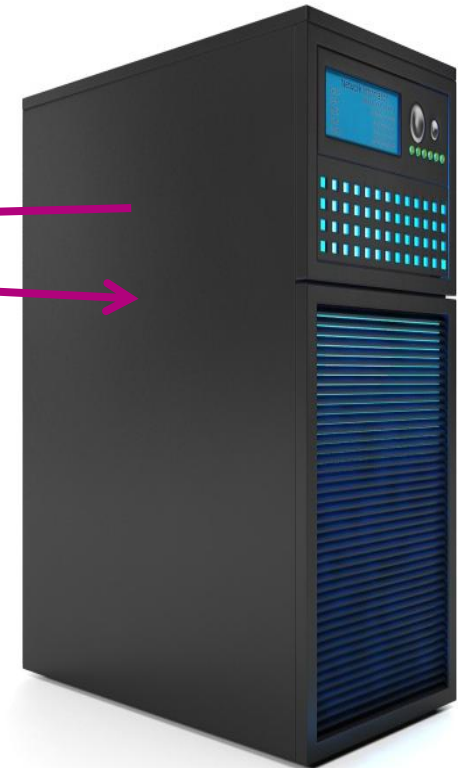
Onaangepaste code kan **niet volledig** vertaald worden :  
Berekeningen worden gespreid over Client-PC & Netezza  
(= veel performantieverlies)



# SQL is **native** ondersteund



Client-PC



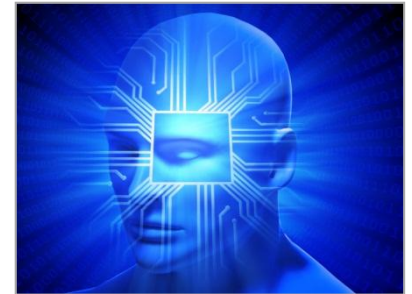
SQL

SQL hoeft niet vertaald te worden  
= **enorm performant**



# Streamlining Analytics

**Predictive analytics**  
**De data supply chain**



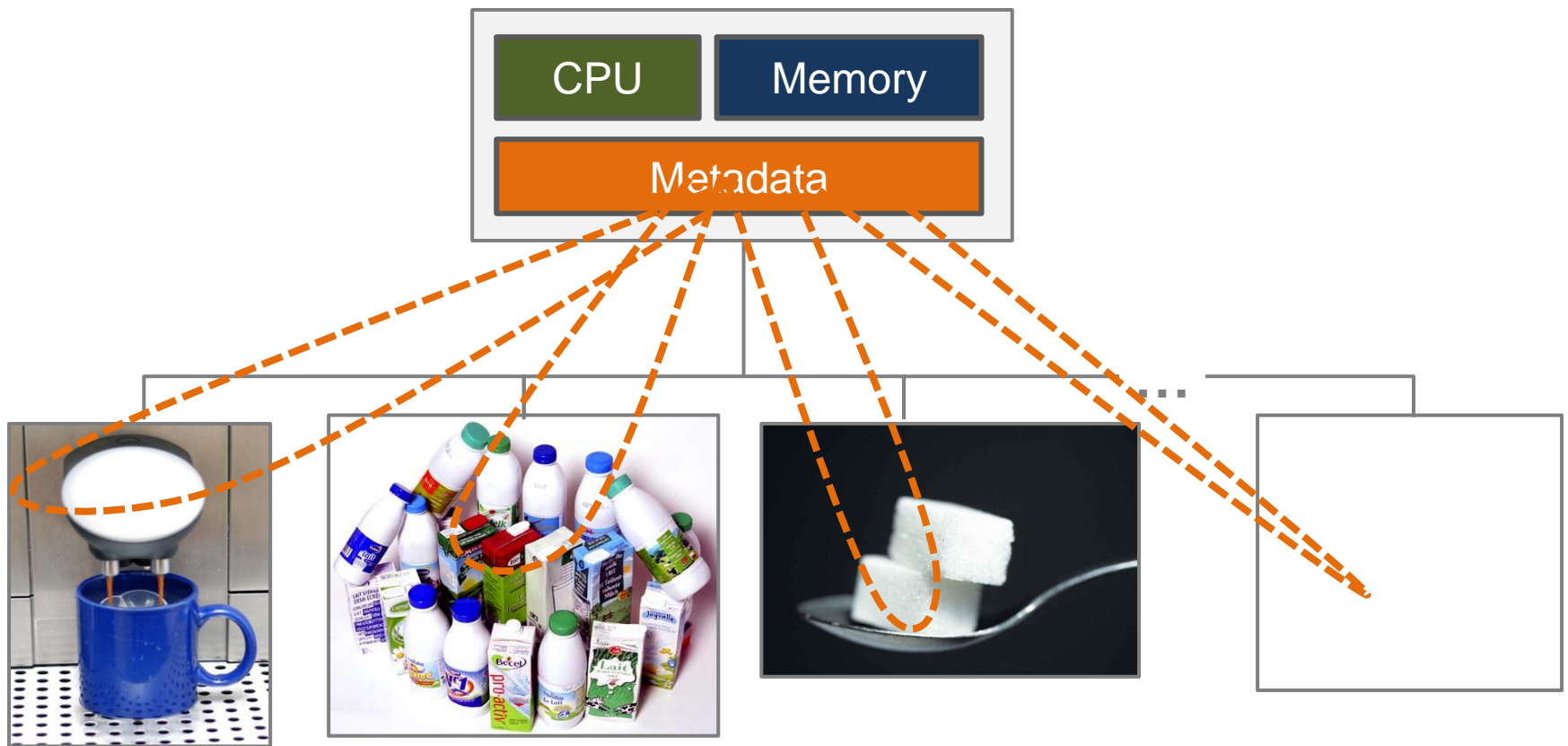
**Barrières bij de introductie van analytics**



**Hardware appliances voor analytics**  
**Data quality**  
**Analytics project management**

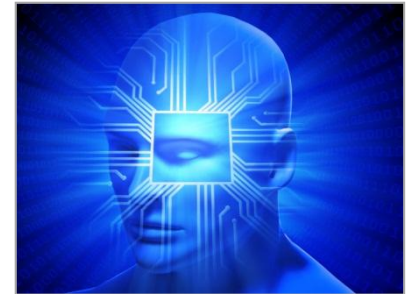


```
SELECT coffee, milk, sugar, cookies  
FROM regular_break  
INNER JOIN infoession_break ON rld = ild
```



# Streamlining Analytics

**Predictive analytics**  
**De data supply chain**



**Barrières bij de introductie van analytics**



**Hardware appliances voor analytics**

**Data quality**

**Analytics project management**



# Analytics & Data Quality

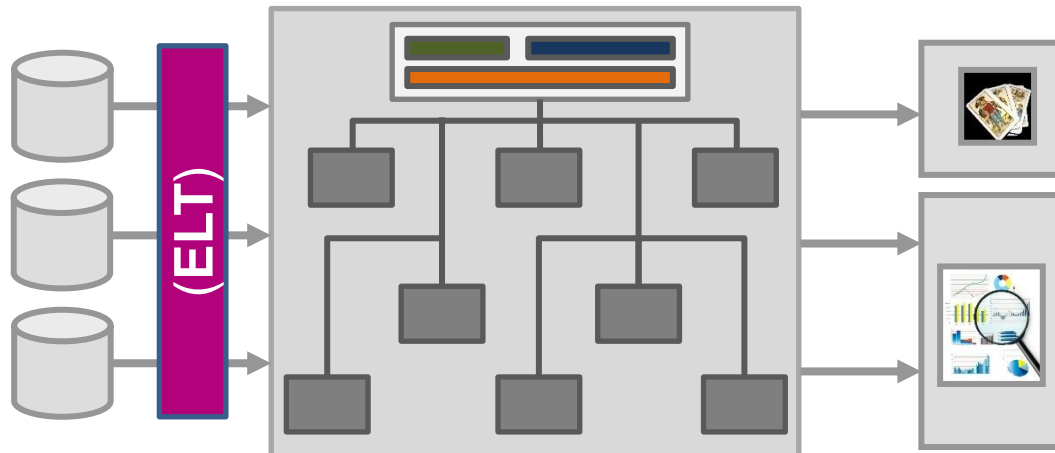
- De ideale wereld (voorgaande studies)
  - best practices
- De realiteit (praktijkvoorbeelden uit projecten)
  - typische problemen van data quality bij analytics-projecten
  - hoe data quality tools de aanpak ondersteunen



# Analytics & Data Quality: de ideale wereld (1)

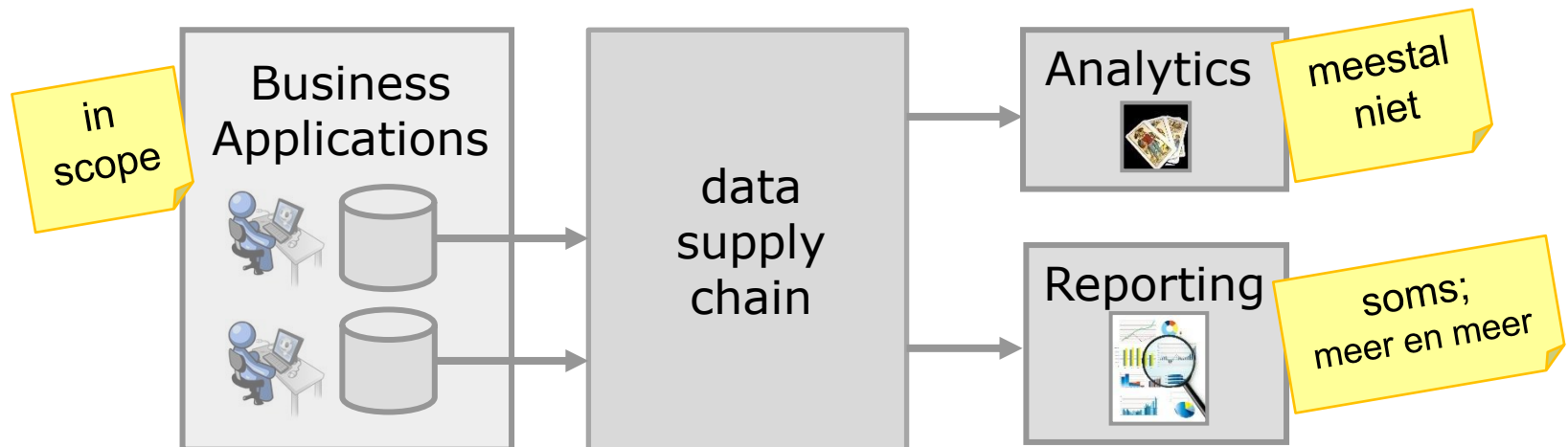
- Vereenvoudigde data supply chain
    - minder ETL, minder vertaalslagen (cfr. barrière 2)
    - opportuniteit voor data governance:
      - Data quality management
      - Master data management
      - Information management
- op één plaats enten

→  
perfecte  
documentatie



# Analytics & Data Quality: de ideale wereld (2)

- Best practice: « fitness for use »
  - 100% data quality is onbereikbaar
  - kosten vs. baten
  - *good enough for its (all) intended use*
    - **in scope voor intended use?**



# Analytics & Data Quality: de ideale wereld (3)

- Best practice: DQ aanpakken aan de

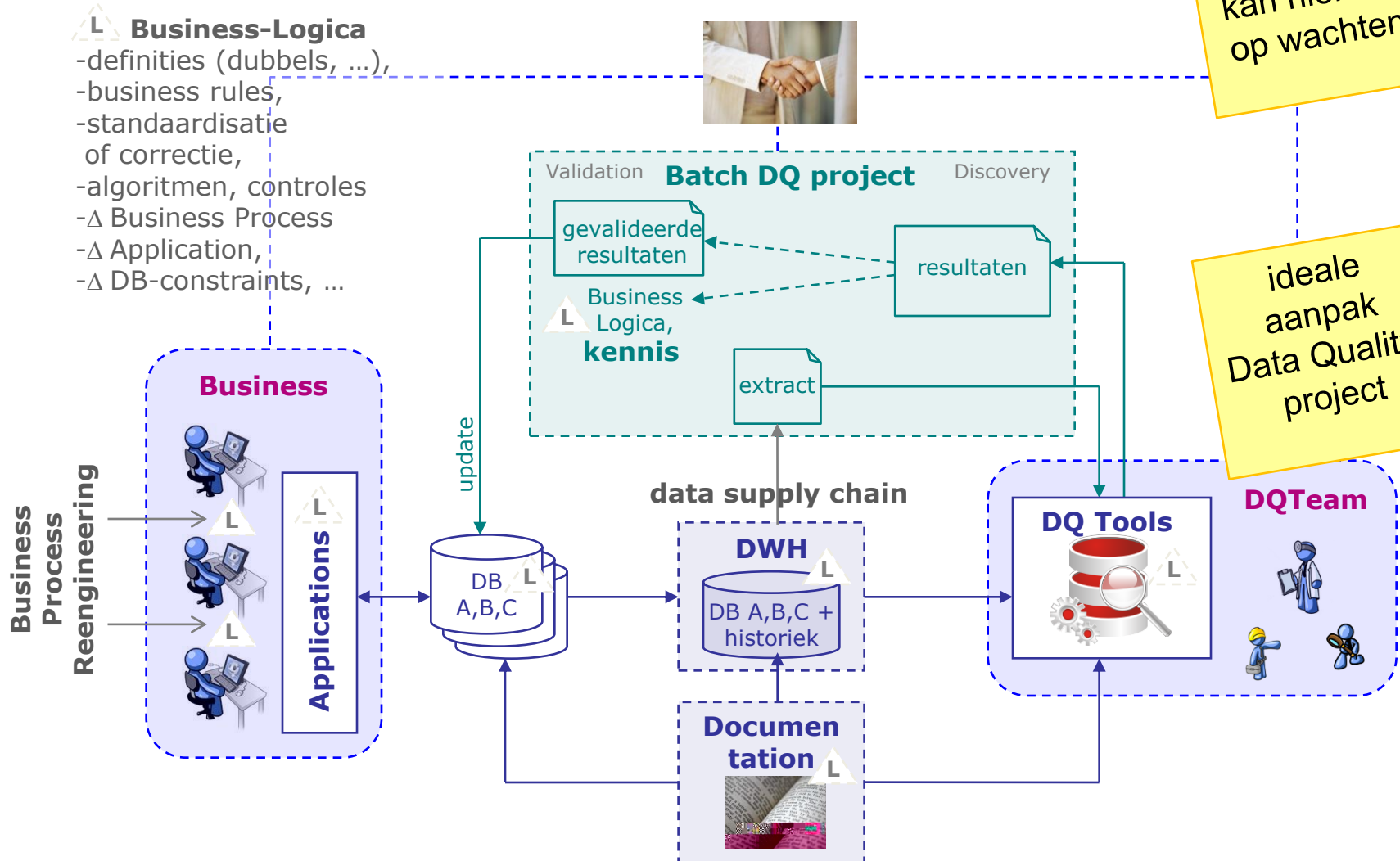
## Business-Logica

- definities (dubbels, ...),
- business rules,
- standaardisatie of correctie,
- algoritmen, controles
- Δ Business Process
- Δ Application,
- Δ DB-constraints, ...



Analytics-project: kan hier niet op wachten!

ideale aanpak Data Quality-project



# Analytics & Data Quality: de realiteit

- De realiteit
  - gebrek aan data quality / best practices
  - hoe data quality tools de aanpak ondersteunen
    - a.h.v. praktijkvoorbeelden uit voorgaande projecten
    - typische problemen van data quality bij analytics-projecten
    - **waarom typisch voor analytics-projecten?**
      - net bij interessante groepen (risico's, nieuwe opportuniteiten, ...) zijn applicatie en databank (nog) onaangepast en is de data quality slechter
      - uitzonderingen, outliers zijn belangrijk (bv. fraude)
      - data quality wordt soms misbruikt om aan controles te ontsnappen

# 1. Gebrek aan (up to date) documentatie

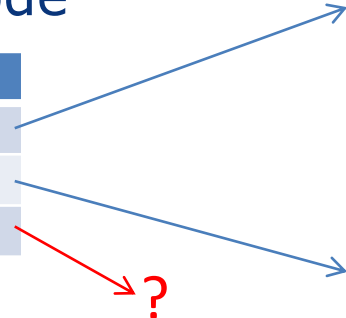
- Werkelijke gebruik is geëvolueerd
  - documentatie echter niet aangepast
    - bv. niet-gedocumenteerde code

REC_ID	REC_poorlyDocumentedVar
456	01
457	61
458	51

**DESCRIPTION :**  
Statut indiquant s'il faut traiter l'e réimmatriculation ou un cas d'en

---

**DOMAINE DE DEFINITION :**  
01 : nouvel e[redacted] avec numéro est absent du répertoire provisoire  
02 : nouvel e[redacted] dont le numé  
11 : réimmatriculation d'un e[redacted]  
l'e[redacted] existe dans le répertoire  
12 : réimmatriculation d'un e[redacted]  
l'e[redacted] est en archive (off-lme)  
13 : réimmatriculation d'un e[redacted] complémentaire; l'e[redacted] existe  
61 : enquête avec statut 81 résolue  
62 : enquête avec statut 82 résolue  
63 : enquête avec statut 83 résolue



- Na een vertaalslag in de data supply chain
  - (E)T(L) → bv. codes gewijzigd

REC_ID	OLTP1_CODACT
456	
457	1
458	

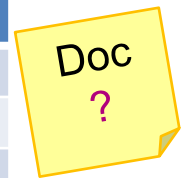
**OLTP1\_CODACT**

**DOMAINE DE DEFINITION :**  
Valeurs possibles [NULL , 1].  
NULL : actif.  
1 : supprimé

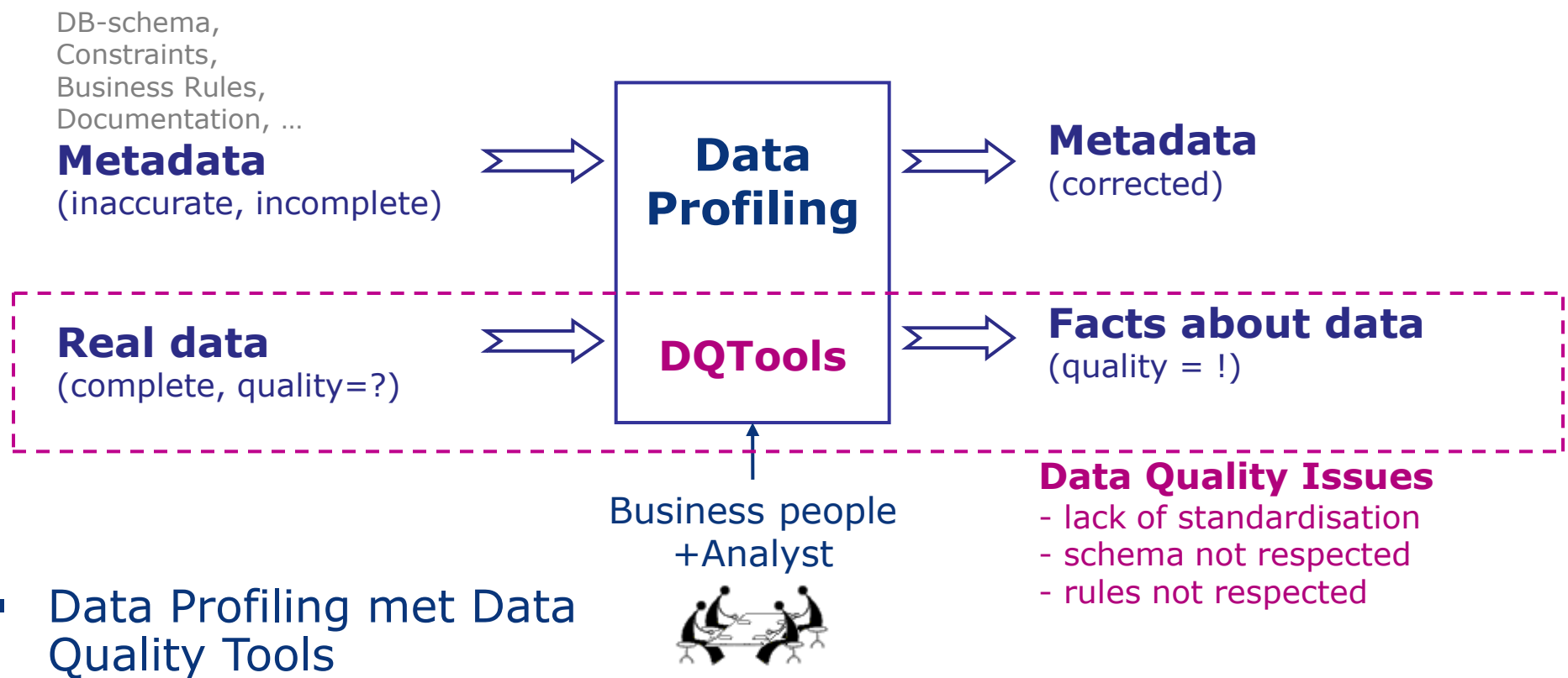
---

**LONGUEUR :** 1  
**TYPE :** Alpha

REC_ID	DWH_CODACT
456	A
457	S
458	A



# Data Quality Tools ondersteuning: Data Profiling (formele audit)



- Data Profiling met Data Quality Tools
  - **Veld per veld**: Column Property Analysis
  - **Structuur**: referentiële constraints
  - **Consistency**: business rules



Jack E. Olson,  
"Data Quality – The Accuracy Dimension" 76

# Data Profiling – Column Property Analyse, drill-down (1)

Entities | Analysis | Findings |

unt: 6) | Statistic

- Source\_secontaire(60) Rows=246496
  - Metadata Keys=0 Deps=0
  - Attributes Count=16
    - Reeksnummer(1) Distribution=100.000
    - Identificatienummer(2) Distribution=99.763
    - Atypemp(3) Distribution=0.014
    - Anatjur(4) Distribution=0.041
    - Ara(5) Distribution=0.002
    - Arem(6) Distribution=0.001
    - Adataffil(7) Distribution=3.876
    - Adatsup(8) Distribution=0
    - Adenomemp(9) Distribution=98.672
    - Aadresemp(10) Distribution=82.704
    - Aboxemp(11) Distribution=0.148
    - Apostemp(12)** Distribution=...
    - Acomemp(13) Distribution=2.401

**Apostemp(12) Properties:**

- Type: real
- Schema Name: Apostemp
- Compliance%: 100.000%
- Inferred type: Integer (Unkn)
- Unique Values: 1155 (0.469%)
- Patterns: 2
- Min: 0
- Max: 9992
- Min Len: 1
- Max Len: 4
- Masks: 2**
- Integers: 1155 (100.00%)
- Dependencies(Discovered): 0 (2)
- Attribute Created Date: 2008/09/08 1...

- (1)-(16) geanalyseerd tijdens load
  - gegeneerde metadata  
Min, Max, Lengtes,  
Null Values, Unique Values  
Patronen,  
Distributies,  
Business Rules, ...
  - detectie Keys & Dependencies
- **Apostemp(12):** postcode werkgever
  - vb. patronen, a.h.v. **Masks**
- Performantie: < 5 min op 1 miljoen records

**Background Tasks**  
Timestamp = Mon Jun 17 15:47:09 CEST 2013

Activity Name	Progress	State	Duration
Discover Keys and Dependencies	Processed rules	Completed	0 Days, 00:00:02
Analyze Statistics	Analyzed 25 Attributes	Completed	0 Days, 00:01:33
/GREPperTest	Rows Loaded: 848589	Completed	0 Days, 00:02:35

# Data Profiling – Column Property Analyse, drill-down (2)

Entities | Analysis | Findings |

unt: 6)      Statistic

- Source\_secontaire(60) Rows=246496
  - Metadata Keys=0 Deps=0
  - Attributes Count=16
    - Reeksnummer(1) Distribution=100.000
    - Identificatienummer(2) Distribution=99.763
    - Atypemp(3) Distribution=0.014
    - Anatjur(4) Distribution=0.041
    - Ara(5) Distribution=0.002
    - Arem(6) Distribution=0.001
    - Adataffil(7) Distribution=3.876
    - Adatsup(8) Distribution=0
    - Adenomemp(9) Distribution=98.672
    - Aadresemp(10) Distribution=82.704
    - Aboxemp(11) Distribution=0.148
    - Apostemp(12) Distribution=0**
      - Type real
      - Schema Name Apostemp
      - Compliance% 100.000%
      - Inferred type Integer (Unkn)
      - Unique Values 1155 (0.469%)
      - Patterns 2
      - Min 0
      - Max 9992
      - Min Len 1
      - Max Len 4
      - Masks 2**
      - Integers 1155 (100.00%)
      - Dependencies(Discovered) 0 (2)
      - Attribute Created Date 2008/09/08 1...
    - Acomemp(13) Distribution=2.401

### Unique Masks

Attribute = Source\_secontaire(60).Apostemp

Mask	Mask Pattern	Value Count	Frequency	Dist %
NNNN	N4	1154	239983	97.358
<b>N</b>	<b>N1</b>	<b>1</b>	<b>6513</b>	<b>2.642</b>

Quality Project Unique Masks

### Unique Values

Attribute = Source\_secontaire(60).Apostemp

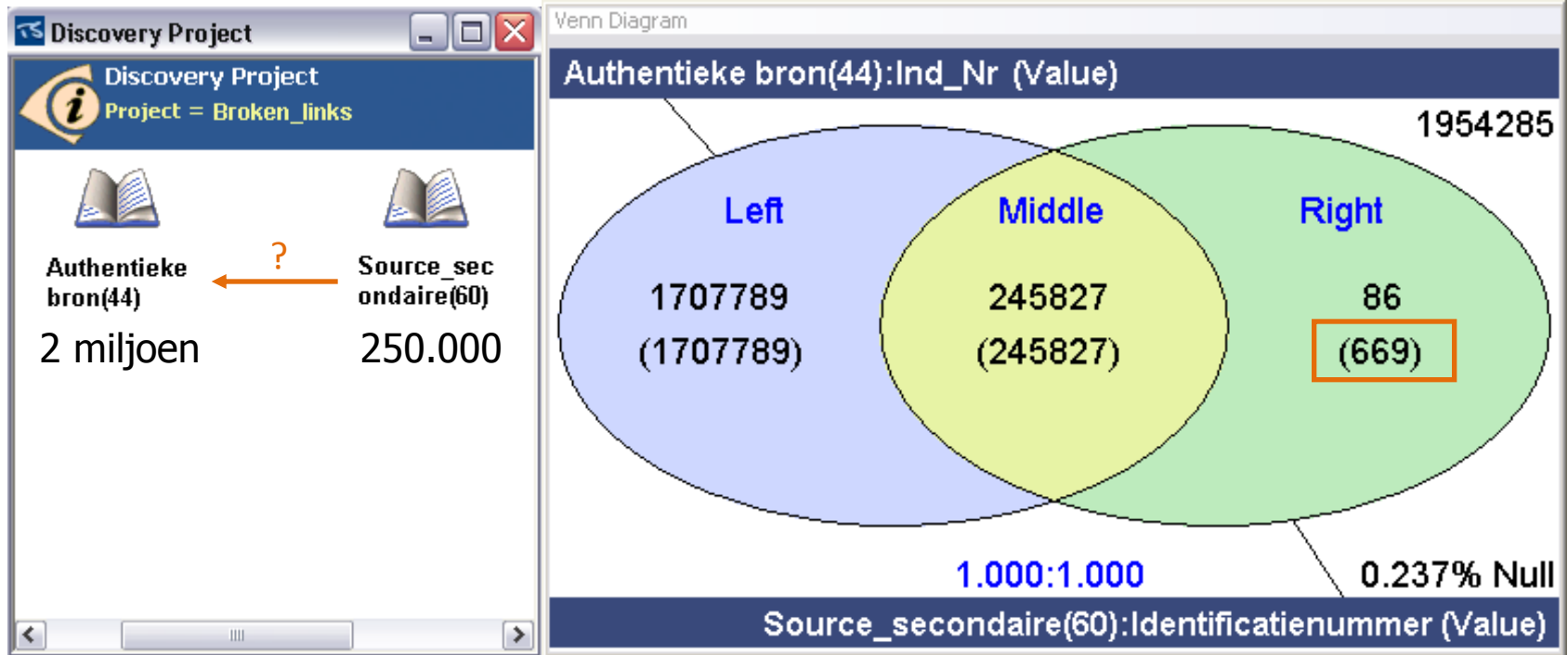
Value	Frequency	Dist %	Length	Soundex	Metaphor
<b>0</b>	<b>6513</b>	<b>2.642</b>	<b>1</b>		

Quality Project Unique Values

### Data rows filtered by selected values in 'Apostemp'

Row	Aadresemp	Apostemp	Acomemp
51	[REDACTED]	0	JUNGLINSTER LU
77	[REDACTED]	0	JA TIEL NL
90	[REDACTED]	0	2215RW VOORHOUT NL
137	[REDACTED]	0	4131 LX VIANEN NL
147	[REDACTED]	0	THORN NL
156	[REDACTED]	0	LEIDSCHENDAM NL
206	[REDACTED]	0	92150 SURESNES FR
220	[REDACTED]	0	7665SG ALBERGEN NL
296	[REDACTED]	0	MADRID ES
302	[REDACTED]	0	VENDIN LE VIEL FR
335	[REDACTED]	0	BREDA NL
481	[REDACTED]	0	SAINT FULGENT FR

# Data Profiling – Referentiële integriteit, foreign keys (1)



- bron 1: Authentieke bron(44)  
sleutel: Ind\_nr
- bron 2: Source\_sec ondaire(60)  
Identificatienummer verwijst naar Ind\_nr's

- **Confrontatie bron 1 – bron 2**  
86 waarden niet in authentieke bron  
in 669 records (er zijn dus dubbels)



# Data Profiling – Functional Dependencies

**Dependencies (Discovered)**  
Entity Adres Communicatie(350) Detectie, tijdens load quality | sample 10.000 | conflicten

Lh Attrs	Rh Attr	Status	Verified	Job	Quality %	Confirming LR Values	Conflicting LH Values	Conflicting Rows	Verifi
C Taalcode,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
C Taalregime,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
D Begindatum,C Postcode	C Nis Gemeentecode	Discovered	No	60	98.760	9876	62	129	
D Begindatum,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.990	9999	1	2	
D Ts Lwizj	D Ts Creatie	Discovered	No	60	98.450	9845	109	232	
Gemeentenaam,Landnaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
Straatnaam 21	Straatnaam Voll	Discovered	No	60	99.350	9935	29	65	
Straatnaam Voll	Straatnaam 21	Discovered	No	60	99.140	9914	68	136	

**Dependencies (Verified)**  
Entity Adres Communicatie(350) Verificatie, on demand exhaustief, 2.9 miljoen

Lh Attrs	Rh Attr	Status	Verified	Job	Quality %	Confirming LR Values	Conflicting LH Values	Conflicting Rows	Verified Date	Verified By
Landnaam	C Landcode	Permanent	Yes	-	99.960	2935363	5	10	2010/04/06 16:17:19	smals

drill-down naar overzicht van conflicten indien Quality < 100%

**Dependency Conflicts**  
Adres Communicatie(350) Dependency Landnaam -> {C Landcode}

Frequency	Landnaam	C Landcode
2854		150
233		
1292	Allemagne	173
945	Allemagne	134
3	Angola	341
1	Angola	381
20	Bahamas	425
5	Bahamas	484
18	Tchécoslovaquie	130
5	Tchécoslovaquie	171

**Drill down to Matching Rows**

Resolve Conflicts...  
List Corrections...

---

Filter...  
Bookmark  
Convert view to Entity...

---

Back  
Forward

---

Change View ▶

---

Export ▶  
Export to Server ▶

---

Copy

# Data Profiling – Business Rules (1)

Sub Contractor5(490) Rows=104930

Metadata Keys=0 Deps=0

- Attributes 12
- Rows Loaded 104930
- Business Rules(Passed) 2(0)**
- Keys(Discovered) 0 (3)
- Dependencies(Discover... 0 (5)
- Date Created 2010/07/...
- Entity State Fully Loa
- Attributes Count=12
  - Denomination(1) Distributi
  - Id\_1(2) Distributi
  - Id\_2(3) Distributi
  - Id\_3(4) Distributi
  - Rue(5) Distributi
  - Rue Num(6) Distributi
  - Rue Box(7) Distributi
  - Commune(8) Distributi
  - Postcode(9) Distributi
  - Country Code(10) Distributi
  - Creation Date(11) Distributi
  - Max Work End Date(12) Distributi

### Entity Business Rule

Enabled  Name Présence\_Identifiant

Description NOT (Id\_1 = "" AND Id\_2 = "" AND Id\_3 = "")

Au minimum 1 des 3 types d'identifiant doit être rempli

### Entity Business Rule

Enabled  Name Enregistrement à temps

Description [Creation Date] < [Max Work End Date]

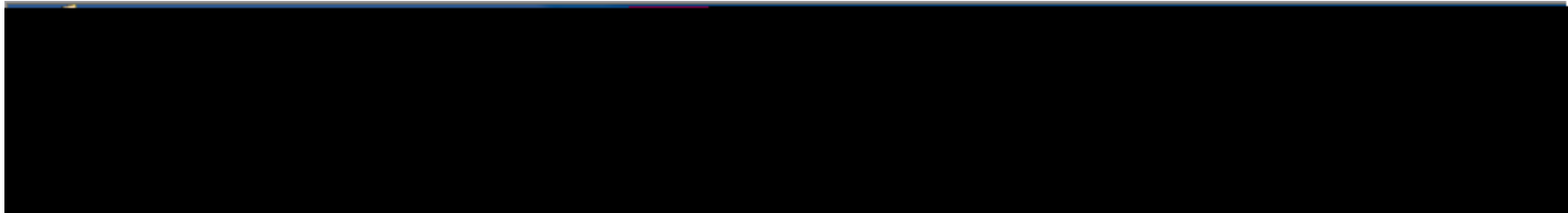
Date de déclaration < date de fin des derniers travaux

Set the threshold to be 100 % of

The test passed on 67.785% of

Choose expression elements from the lists below

Attributes	Bce
Functions	Commune
Literals	Country Code
Operators	Creation Date
	Denomination
	Max Work End Date



# Data Profiling – Business Rules (2)

**Failing Rows (Présence\_Identifiant)**  
Entity = Sub Contractor5(490)

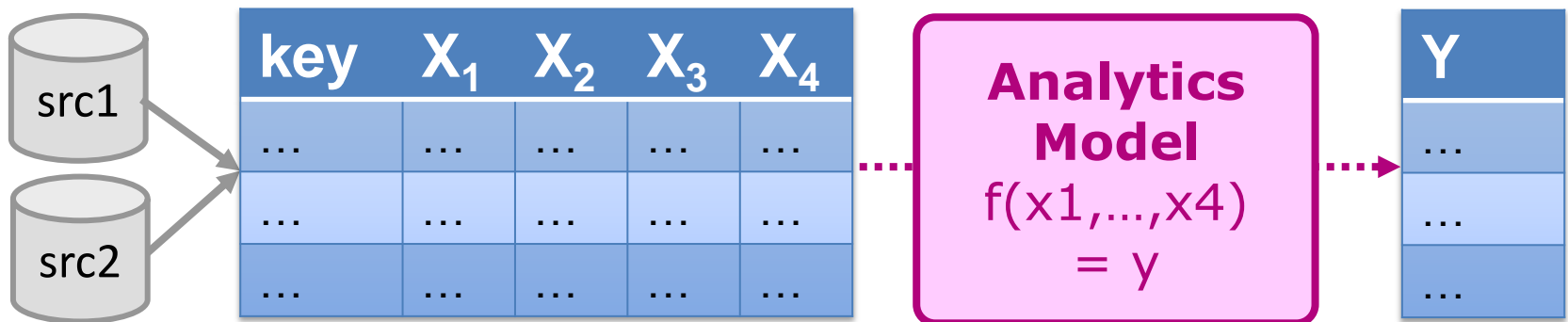
Denomination	Id_1	Id_2	Id_3	Rue	Rue Num	Rue Box	Commune	Postcode	Country...	Creation Date	Max Work End Date
ENTREPRIS...				RU...	9		ANDENNE	5300	150	2005-08-18 ...	2006-06-22 00:00:0...
CONCEPT ...				AV...	15		FOREST	1190	150	2004-02-19 ...	2003-12-01 00:00:0...
ADRIAN SPRL				ZO...	11	n/a	MONT D...	7750	150	2009-05-20 ...	2009-03-31 00:00:0...
UNION PR...				KL...	6A		OUDEN...	4730 AE	129	2007-09-26 ...	2007-12-31 00:00:0...
MONTI B...				G...	37		HASSELT	3511	150	2007-03-20 ...	2011-07-31 00:00:0...
JON...				RU...	80		CHARLE...	6031	150	2005-12-02 ...	2006-09-28 00:00:0...
...				BE	10		ZOULIE	3440	150	2006-02-09 ...	2006-07-15 00:00:0...

**Failing Rows (Enregistrement à temps)**  
Entity = Sub Contractor5(490)

Denomination	Id_1	Id_2	Id_3	Rue	Rue Num	Rue Box	Commune	Postcode	Country...	Creation Date	Max Work End Date
WUUST...	76...			BA...	1		WUUST...	2990	150	2007-10-09 ...	2007-08-31 00:00:0...
LIBRAM...	46...	124...		RU...	n/a	n/a	LIBRAM...	6800	150	2009-05-20 ...	2003-11-28 00:00:0...
ETTERB...	41...	127...		R ...	63	n/a	ETTERB...	1040	150	2009-05-20 ...	2008-05-31 00:00:0...
HOESELT	73...		78297...	TO...	7	n/a	HOESELT	3730	150	2007-06-07 ...	2007-06-04 00:00:0...
HOESELT	73...		78297...	DA...	2	n/a	HOESELT	3730	150	2007-06-07 ...	2007-06-04 00:00:0...
NIEL	46...	171...		P...	74	n/a	NIEL	2845	150	2009-05-20 ...	2006-04-14 00:00:0...
ST KATE...	74...		78242...	DU...	n/a	n/a	ST KATE...	2860	150	2005-06-16 ...	2005-01-31 00:00:0...
BEER ...	44	160		G...	27	n/a	BEER ...	3990	150	2009-05-20 ...	2003-05-15 00:00:0...

## 2. Data-integratie vanuit heterogene bronnen

- Data Preparation & transformation
  - naar een vorm geschikt voor Analytics



src1_key	src2_key
406798006	0406.798.006
206731645	0206.731.645
...	...

src1_creationdate	src2_datstart
20060401	01APR06
19501001	01OCT50
...	...

# Ondersteuning data integratie met dqtools:

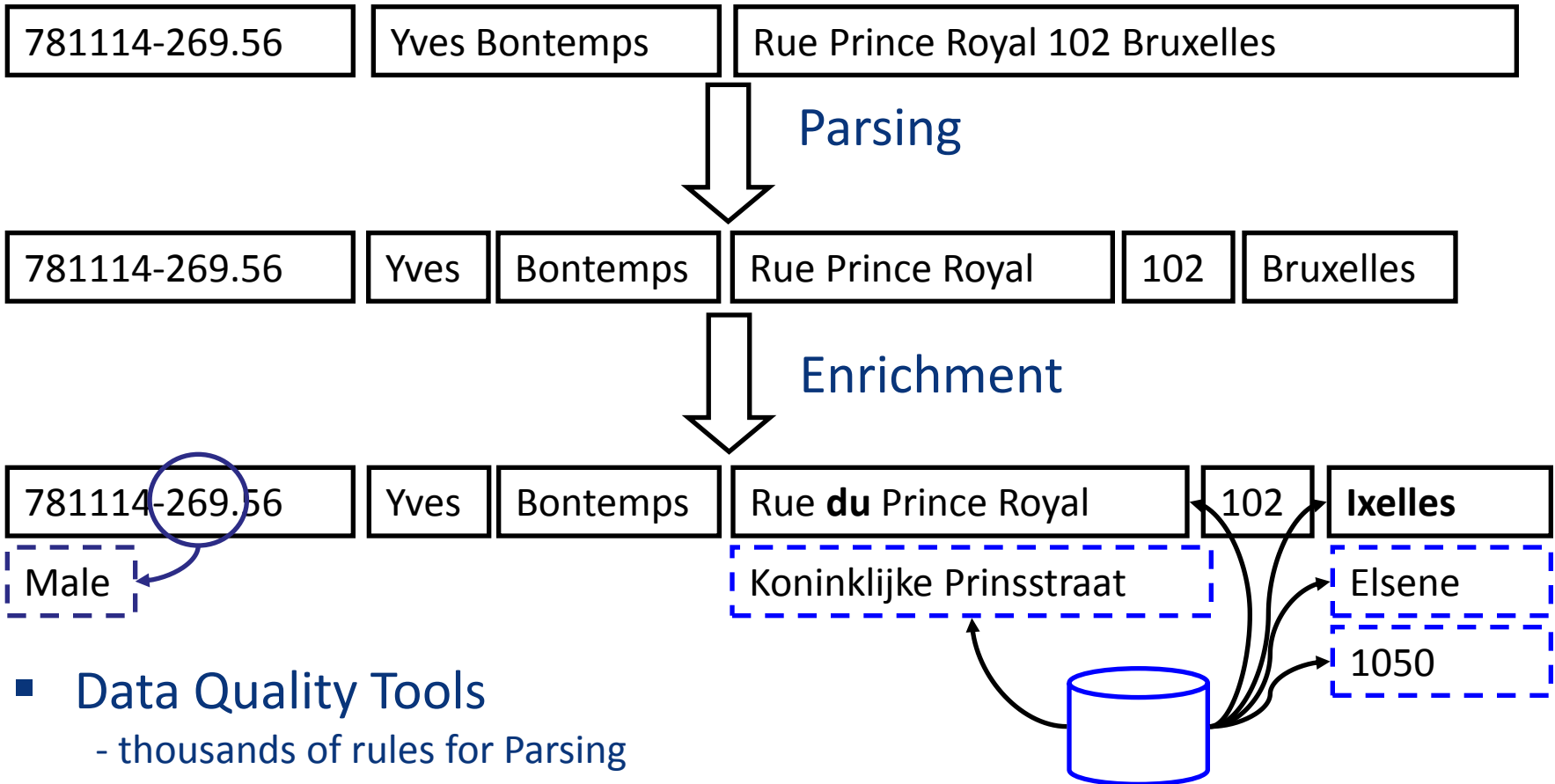
## Data Standardisation – simple domains

### Example: country codes

Data Rows (Dynamic)											
Entity = pl tranfmr p26(513)											
origineel						toegevoegd m.b.v. data quality tool					
Tsq Denom	Adres	Boite	Tsq Postcd	Tsq Commune	Orig Landcd	Iso3166 Cd	Iso3166 2l	Iso3166 3l	Iso3166 NI	Ins Cd	Ins NI
<del>GREENS ...</del>					150	056	BE	BEL	België	150	België
<del>JANSEN ...</del>	MAH...			BLACKROC...	116	372	IE	IRL	Ierland	116	Ierland /Eire/
<del>SCHEN ...</del>	OLD...			WATERSFO...	IRL	372	IE	IRL	Ierland		
<del>BONJOU ...</del>	268		067 82	MODRA NAD...	141	703	SK	SVK	Slowakije	141	Slovaakse Republiek
<del>BONJOU ...</del>	HIA...		067 82	MODRA NAD...	SK	703	SK	SVK	Slowakije		
<del>BERND ...</del>	Vizi...		1031	BUDAPEST	115	348	HU	HUN	Hongarije	115	Hongarije ( Rep. )
<del>BERND ...</del>	VIZI...		1031	BUDAPEST	H	348	HU	HUN	Hongarije		
<del>VOOREM ...</del>	SICI...		1045 AX	AMSTERDAM	129	528	NL	NLD	Nederland	129	Nederland
<del>VOOREM ...</del>	SICI...		1045	AMSTERDAM	NL	528	NL	NLD	Nederland		
<del>DEWERT ...</del>	MAD...		1131	BUDAPEST	H	348	HU	HUN	Hongarije		
<del>DEWERT ...</del>	POS...		1440	AP PURMER...	NL	528	NL	NLD	Nederland		
<del>DEWERT ...</del>	POS...		1440	AP PURMER...	NL	528	NL	NLD	Nederland		
<del>SANRA ...</del>	czer...	4	20 349	LUBLIN	122	616	PL	POL	Polen	122	Polen ( Rep. )
<del>SANRA ...</del>	CZE...		20 349	LUBLIN	PL	616	PL	POL	Polen		
<del>TRABEL ...</del>	Via ...		24129	BERGAMO	128	380	IT	ITA	Italië	128	Italië
<del>TRABEL ...</del>	VIA ...		24129	BERGAMO	I	380	IT	ITA	Italië		
<del>VERBA ...</del>	Stee...		2407 BD	ALPHEN AA...	129	528	NL	NLD	Nederland	129	Nederland
<del>VERBA ...</del>	STE...		2407	BD ALPHEN ...	NL	528	NL	NLD	Nederland		
<del>VERBA ...</del>	STE...		2407	BD ALPHEN ...	NL	528	NL	NLD	Nederland		

# Ondersteuning data integration met dqtools: Data Standardisation – complex domains

Example: names and adresses



- **Data Quality Tools**

- thousands of rules for Parsing
- knowledge bases for Enrichment, often Regional

# Data standardisation with Data Quality Tools

## What view would you rather build analytics upon? ...

P/N	DESCRIPTION
1774-5674	TUBE, CENTRIFUGE POLY S 15ML (CS/500)CONICAL-BOTTOM
1774-5675	TUBE, CENTRIFUGE PPL 15ML (CS/500)CONICAL-BOTTOM
1774-4532	TUBE, CENTRIFUGE PPL 50ML (CS/500)CONICAL-BTTMPCK 25/RACK
1774-4538	TUBE, CENTRIFUGE POLY S 50ML (CS/500)CONICAL-BTMPK 25/RACK
645-4556	PIPET, CLEAR SEROLOGICAL 2ML (CASE/500)
195-7934	NUT, LOCK RH,11"
3324-7955	VIAL, WHEATON 33* CLEAR 4ML (CS/144)



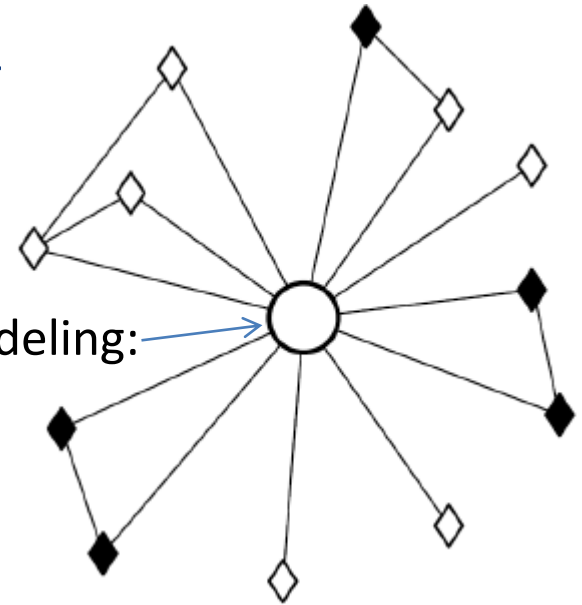
P/N	ITEM NAME	MATERIAL	SIZE	UOM	DESCRIPTOR	PACKAGE	PACK METHOD
1774-5674	CENTRIFUGE TUBE	POLYSTERENE	15	ML	CONICAL	CASE/500	BOTTOM PACKED
1774-5675	CENTRIFUGE TUBE	POLYPROPYLENE	15	ML	CONICAL	CASE/500	BOTTOM PACKED
1774-4532	CENTRIFUGE TUBE	POLYPROPYLENE	50	ML	CONICAL	CASE/500	BOTTOM PACKED 25/RACK
1774-4538	CENTRIFUGE TUBE	POLYSTERENE	50	ML	CONICAL	CASE/500	BOTTOM PACKED 25/RACK
ML	CLEAR	CASE/500			0645-4556	SEROLOGICAL PIPET	2
IN	RIGHT HAND				0195-7934	LOCK NUT	11
ML	CLEAR	CASE/144			3324-7955	WHEATON VIAL	4

# 3. Netwerk-analytics met fuzzy matching

- Eenvoudig geval: als de relaties in een (social) netwerk rechtstreeks uit de RDBMS-structuur kunnen gehaald worden
  - klanten die met elkaar getelefoneerd hebben
  - werkgever-werknemersrelatie
  - producten die samen gekocht zijn
- Complex geval: «fuzzy matching»
  - gevestigd op «gelijkaardig adres»
  - Hoe?
    - voorbereid met data quality tools

voorwerp van modeling: →

- risico,
- opportuniteit,
- ...



# Fuzzy matching met data quality tools (1)

Source	match type	Denomination	Adres	Boite	Postcd	Commune	Cdpays
L	100	PROJEKT SERWIS [REDACTED]	BOHATEROW [REDACTED] 63	43	05-100	NOWY DWOR [REDACTED]	122
L	100	PROJEKT SERWIS [REDACTED]	BOHATEROW [REDACTED] 63	43	05-100	NOWY DWOR [REDACTED]	122
R	115	PROJEKT SERWIS ([REDACTED] [REDACTED])	UL. BOHATEROW [REDACTED] 63	42	05-100	NOWY DROW [REDACTED]	PL
R	115	PROJEKT SERWIS [REDACTED] [REDACTED] A NOWY DWOR	UL BOHATEROW [REDACTED] 63	43	05-100	NOWY DWOR [REDACTED]	PL
R	135	PROJEKT SERWIS [REDACTED] [REDACTED]	BOHATEROW [REDACTED] 63/43		05-100	NOWY DWOR [REDACTED]	PL
R	106	PROJEKT SERWIS [REDACTED] [REDACTED]	BOHATEROW 63LOK	43	05-100	NOKY DWOR [REDACTED]	PL
R	106	PROJEKT SERWIS [REDACTED] [REDACTED]	BOHATEROW [REDACTED] 63LOK	43	05-100	NOKY DWOR [REDACTED]	PL
R	128	PROJEKT SERWIS [REDACTED] [REDACTED]	BOHATEROW [REDACTED] 63LOK	43	05-100	NOKY DWOR [REDACTED]	PL
R	128	PROJEKT SERWIS [REDACTED] [REDACTED]	BOHATEROW [REDACTED] 63LOK	43	05-100	NOKY DWOR [REDACTED]	PL
R	128	PROJEKT SERWIS [REDACTED] [REDACTED]	BOHATEROW [REDACTED] 63LOK	43	05-100	NOKY DWOR [REDACTED]	PL
R	138	SOCIETE PROJEKT SERWIS	BOHATEROW [REDACTED] 63/43	N/A	05-100	NOWY DWOR [REDACTED]	PL

## 2 databanken, L en R

- ✓ er bestaat geen 'vreemde sleutel'-relatie tussen L en R
- ✓ DQTool detecteert dubbels en organiseert in clusters
- ✓ DQTool legt link tussen beide databanken
- ✓ mogelijke fraude ?

# Fuzzy matching met data quality tools (2)

C Postcode	Tq Gout Postal Code	Straatnaam Voll	Tq Gout Street Name	Huisnummer	Pr House N...	Gemeentenaam	Tq Gout Postal City
1020	1020	<u>RUE E VANDER AA</u>	<u>RUE ERNEST VANDER AA</u>	1	1	Brussel	BRUSSEL
1020	1020	<u>rue Vander Aa</u>	RUE ERNEST VANDER AA	3	3	Bruxelles	BRUXELLES
1050	1050	<u>91 R VAN AA</u>	<u>RUE VAN AA</u>	—	<u>91</u>	Elsene	ELSENE
1050	1050	<u>27 R.VAN AA</u>	RUE VAN AA	—	<u>27</u>	Elsene	ELSENE
1050	1050	RUE VAN AA	RUE VAN AA	94	94	Ixelles	IXELLES
1050	1050	RUE VAN AA	RUE VAN AA	94	94	Elsene	ELSENE
<u>1020</u>	<u>1050</u>	rue Van Aa	RUE VAN AA	2	2	<u>Bruxelles</u>	<u>IXELLES</u>
1050	1050	<u>2 R VAN AA</u>	RUE VAN AA	—	<u>2</u>	Ixelles	BRUXELLES
1000	1000	R JOSEPH II <u>40</u>	RUE JOSEPH II	—	<u>40</u>	Bruxelles	BRUXELLES
1000	1000	rue Joseph II <u>71 (...)</u>	RUE JOSEPH II	—	<u>71</u>	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II	RUE JOSEPH II	71	71	Brussel	BRUSSEL
<u>1040</u>	1000	Rue Joseph II <u>5-7</u>	RUE JOSEPH II	—	<u>5-7</u>	Bruxelles	BRUXELLES
<u>1040</u>	1000	Rue Joseph II <u>67A</u>	RUE JOSEPH II	—	<u>67A</u>	Bruxelles	BRUXELLES
<u>1030</u>	1000	rue JOSEPH II, <u>114 -</u>	RUE JOSEPH II	<u>116</u>	<u>114 - 116</u>	Schaarbeek	BRUXELLES

## Resultaat (fuzzy matching + adresvalidatie, adrescleansing)

- ✓ gecorrigeerde postcode
- ✓ gestandaardiseerde straatnaam
- ✓ correct ingedeelde adreselementen (parsing)
- ✓ gecorrigeerde gemeentenaam
- ✓ dubbels gedetecteerd en georganiseerd in clusters

# Fuzzy matching met data quality tools (3)

- Performantie is cruciaal
  - in principe kwadratisch in  $f(\text{aantal records})^*$
  - voorbeeld: « gelijkaardige adressen »
    - parsing
    - adresvalidatie
    - matching
    - post-processing (inspectiedistricten)
  - 250.000 actieve: 10 min
  - 850.000 +historiek: 1u



Ref	Activity Name	State	Duration
93	TSQ Process 'WG_zelfde_adres' execution	Completed	0 Days, 01:01:23
92	TSQ Process 'WG_zelfde_adres' execution	Completed	0 Days, 00:10:15

\* Cfr. Performance, Blocking → Smals Research publicatie « Data Quality II: Tools », D. Van Dromme, 2007

# Analytics & Data Quality: conclusie

- Belangrijke taken van de Analytics-expert, hebben te lijden aan een gebrek aan dq
  1. Kennis opbouwen bij gebrek aan (up to date) documentatie en standaardisatie
  2. Data-integratie vanuit heterogene bronnen
  3. Netwerk-analytics met fuzzy links
- Data Quality Tools-functionaliteiten kunnen de Analytics-expert helpen
  - Profiling – Standardisation – Fuzzy Matching

# Streamlining Analytics

**Predictive analytics**  
**De data supply chain**



**Barrières bij de introductie van analytics**



**Hardware appliances voor analytics**  
**Data quality**

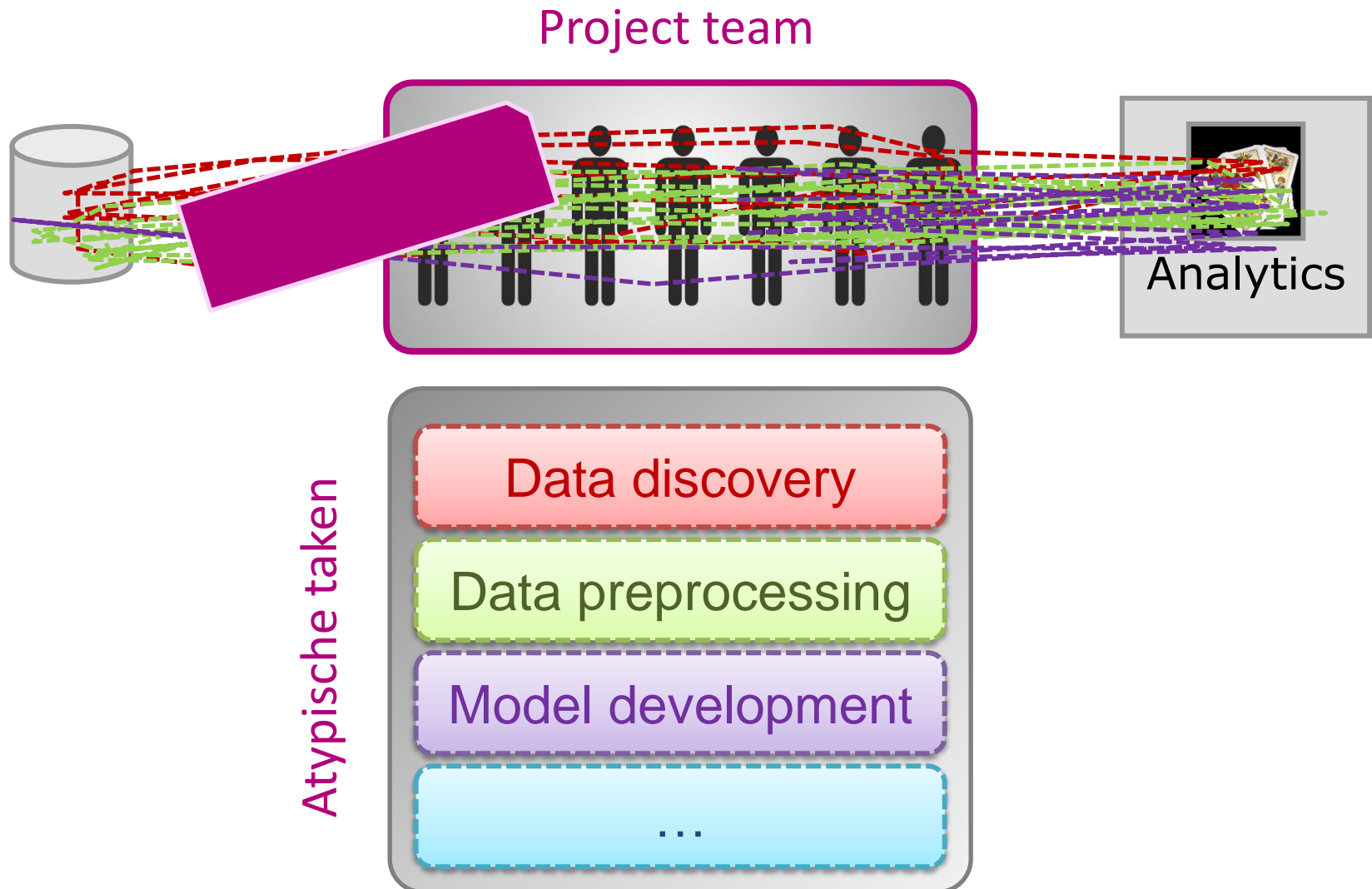
**Analytics project management**



# Analytics **project management**

- Probleemstelling (barrière 3)
- Data Mining Methodologie
- Resource & project planning
- Critical data miner tasks

# Probleemstelling - barrière 3: atypische projectstructuur



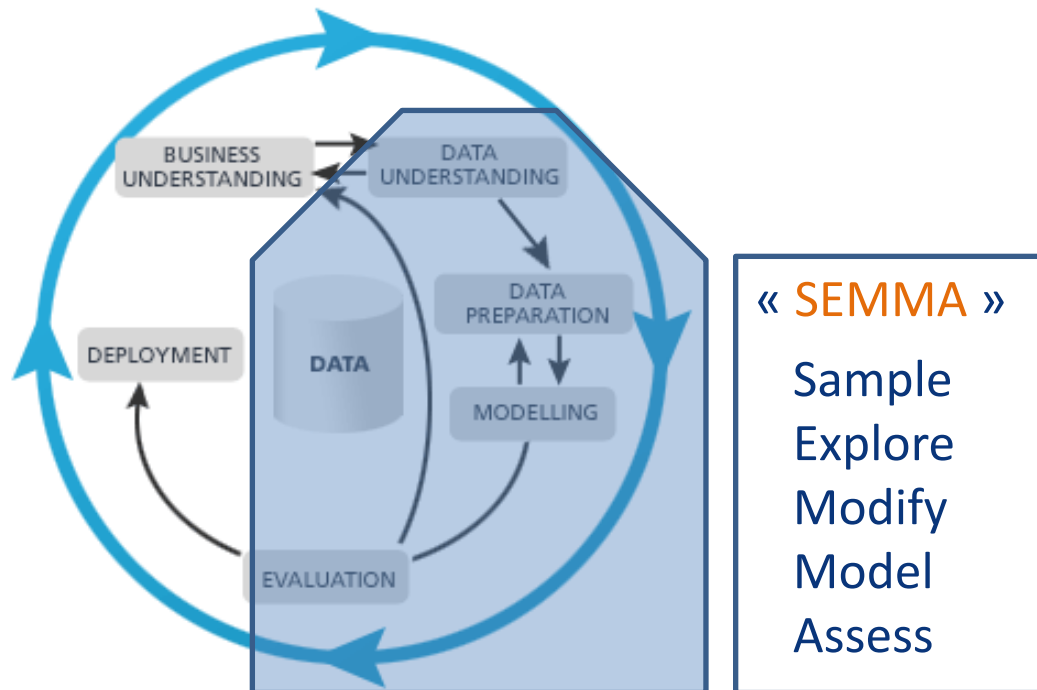
# Analytics-project-methodologie

- Gelijkaardige, iteratieve methodologieën
  - **KDD** original approach: 5 key processes (Fayyad, 1996)
  - **CRISP-DM** 1.0 (1999+), 2.0 (2006+)
  - **SEMMA** by SAS

KDD	SEMMA	CRISP-DM
Pre KDD	- -	Business understanding
Selection	Sample	Data understanding
Pre-processing	Explore	
Transformation	Modify	Data preparation
Data Mining	Model	Modeling
Interpretation	Assess	Evaluation
Post KDD	- -	Deployment

# Analytics-project-methodologie

- « CRISP-DM » Cross-Industry Standard Process for Data Mining



Karakteristieken: **iteratief** proces, **multidisciplinair**

→ samenstelling projectteam

→ projectfasen

# DM Methodologieën in tools

## SAS Enterprise Miner (SEMMA)

The screenshot shows the SAS Enterprise Miner (SEMMA) interface. The main window is titled 'Enterprise Miner - SAS Global Forum'. The menu bar includes File, Edit, View, Actions, Options, Window, and Help. The toolbar contains various icons for file operations and analysis. The left sidebar shows a tree view with categories like SAS Global Forum, Data Sources, Diagrams, Model Packages, and Users. The main workspace is divided into a 'Sample' tab and a workflow diagram. The 'Sample' tab has a large 'S' icon and a list of actions: Input Data, Sample, File Import, Data Partition, Merge, Append, Time Series, and Filter. The workflow diagram shows a sequence of steps: Input Data, Sample, File Import, Data Partition, Merge, Append, Time Series, and Filter, leading to a 'Neural Network' model. The bottom left shows a 'General' properties panel with a table of properties and values.

Property	Value
<b>General</b>	
Node ID	CNTRL
Imported D	...
Exported D	...
<b>Status</b>	
Create Tim	
Run Id	
Last Error	
Last Status	
Last Run Ti	
Run Duratic	
Grid Host	

**General**  
General Properties

- **Data-input**
  - via csv, sas7bdat, ..., SQL, SQL-pushback (via DWH/appliance?)
- **Omgaan met beperkingen**
  - wat als de interessantste patronen (bv. fraude) in de minderheid zijn?
  - sampling-strategieën, ...
- **Training – Validation – Test sets**
- **Niet: Welke concepten spelen een rol?**  
→ welke databronnen → input via welke weg, in welke vorm?

# DM Methodologieën in tools

## SAS Enterprise Miner (SEMMA)

The screenshot shows the SAS Enterprise Miner (SEMMA) interface. The main window is titled 'Enterprise Miner - SAS Global Forum'. The menu bar includes File, Edit, View, Actions, Options, Window, and Help. The toolbar contains various icons for file operations and analysis. The left sidebar shows a tree view with 'SAS Global Forum', 'Data Sources', 'Diagrams', 'Model Packages', and 'Users'. The main workspace is titled 'intel' and displays the 'Explore' tab. A list of data mining techniques is shown, including Association, Cluster, Graph Explore, Market Basket, MultiPlot, Path Analysis, SOM/Kohonen, StatExplore, Variable Clustering, and Variable selection. A background diagram illustrates the SEMMA process flow: Sample -> Explore -> Modify -> Model -> Assess -> Utility -> Credit Scoring. The diagram also shows various modeling techniques like Neural Networks, Decision Trees, Regression, and Association.

Property	Value
<b>General</b>	
Node ID	CNTRL
Imported D	...
Exported D	...
<b>Status</b>	
Create Tim	
Run Id	
Last Error	
Last Status	
Last Run Ti	
Run Duratic	
Grid Host	

- interactief, overleg
- visualisatie
- bestuderen van distributies
- preparatie van transformaties
- unsupervised learning techniques
  - associatie, clustering
- inzicht in de data verhogen

# DM Methodologieën in tools

## SAS Enterprise Miner (SEMMA)

The screenshot shows the SAS Enterprise Miner interface. The top menu bar includes File, Edit, View, Actions, Options, Window, and Help. Below the menu is a toolbar with various icons. The left sidebar shows a tree view with folders for SAS Global Forum, Data Sources, Diagrams, Model Packages, and Users. The main workspace is titled 'intel' and displays the 'Modify' tab. On the left of the workspace is a 'Modify' panel with a list of actions: Transform variables, Replacement, Impute, Interactive Binning, Principal Components, Drop, and Rules Builder. Below this is a table with 'Property' and 'Value' columns, showing details for 'General' and 'Status' properties. To the right of the workspace is a workflow diagram with nodes for 'Variable Selection', 'Tree (2) - Gini', 'Online Regression', 'Tree - Entropy', and 'Regression', all leading to an 'AutoML' node.

Property	Value
<b>General</b>	
Node ID	CNTRL
Imported D	...
Exported D	...
<b>Status</b>	
Create Tim	
Run Id	
Last Error	
Last Status	
Last Run Ti	
Run Duratio	
Grid Host	

- Transformatie
  - naar geschikte vorm voor Data Mining
- Extraheren van « gedrag », « events » en « netwerkvariabelen » uit de ruwe variabelen

# DM Methodologieën in tools

## SAS Enterprise Miner (SEMMA)

The screenshot shows the SAS Enterprise Miner (SEMMA) interface. The main window is titled 'Enterprise Miner - SAS Global Forum'. The menu bar includes File, Edit, View, Actions, Options, Window, and Help. The toolbar contains various icons for file operations and model management. The left sidebar shows a tree view with 'SAS Global Forum', 'Data Sources', 'Diagrams', 'Model Packages', and 'Users'. The main workspace is titled 'intel' and displays the 'Model' tab. A large 'M'odel icon is visible. Below it, a grid of algorithm buttons is shown, including Decision Tree, Gradient Boosting, AutoNeural, Neural Network, DMNeural, Regression, Dmine Regression, MBR, Partial Least Squares, LARS, Model Import, Rule Induction (2), Ensemble, and woStage. A workflow diagram is overlaid on the right side of the workspace, showing a sequence of steps: 'Neural Network', 'Model Computer', 'Tree (2) - Grid', 'Stepwise Regression', 'Tree - Strategy', and 'Regression'. The diagram also includes 'Model Import' and 'AutoNeural' at the end of the flow.

■ Welk algoritme?

- interpreteerbaarheid
- karakteristieken van de inputvariabelen en gezochte patronen
- meervoudige classificatie
- ...

Property	Value
<b>General</b>	
Node ID	CNTRL
Imported D	...
Exported D	...
<b>Status</b>	
Create Tim	
Run Id	
Last Error	
Last Status	
Last Run Ti	
Run Duratic	
Grid Host	

**General**  
General Properties

# DM Methodologieën in tools

## SAS Enterprise Miner (SEMMA)

The screenshot shows the SAS Enterprise Miner interface. The top menu includes File, Edit, View, Actions, Options, Window, and Help. Below the menu is a toolbar with various icons. The main workspace is titled 'intel' and displays the 'Assess' tab. On the left, there is a 'Property Value' table and a 'General' section. The 'Assess' tab contains several buttons: Model Comparison, Score, Segment Profile, Decisions, and Cutoff. In the background, a workflow diagram is visible, showing nodes for Neural Network, Tree (2) - Gini, Regression, and AutoML, connected by arrows.

Property	Value
<b>General</b>	
Node ID	CNTRL
Imported D	...
Exported D	...
<b>Status</b>	
Create Tim	
Run Id	
Last Error	
Last Status	
Last Run Ti	
Run Duratic	
Grid Host	

**Assess**

- Model Comparison
- Score
- Segment Profile
- Decisions
- Cutoff

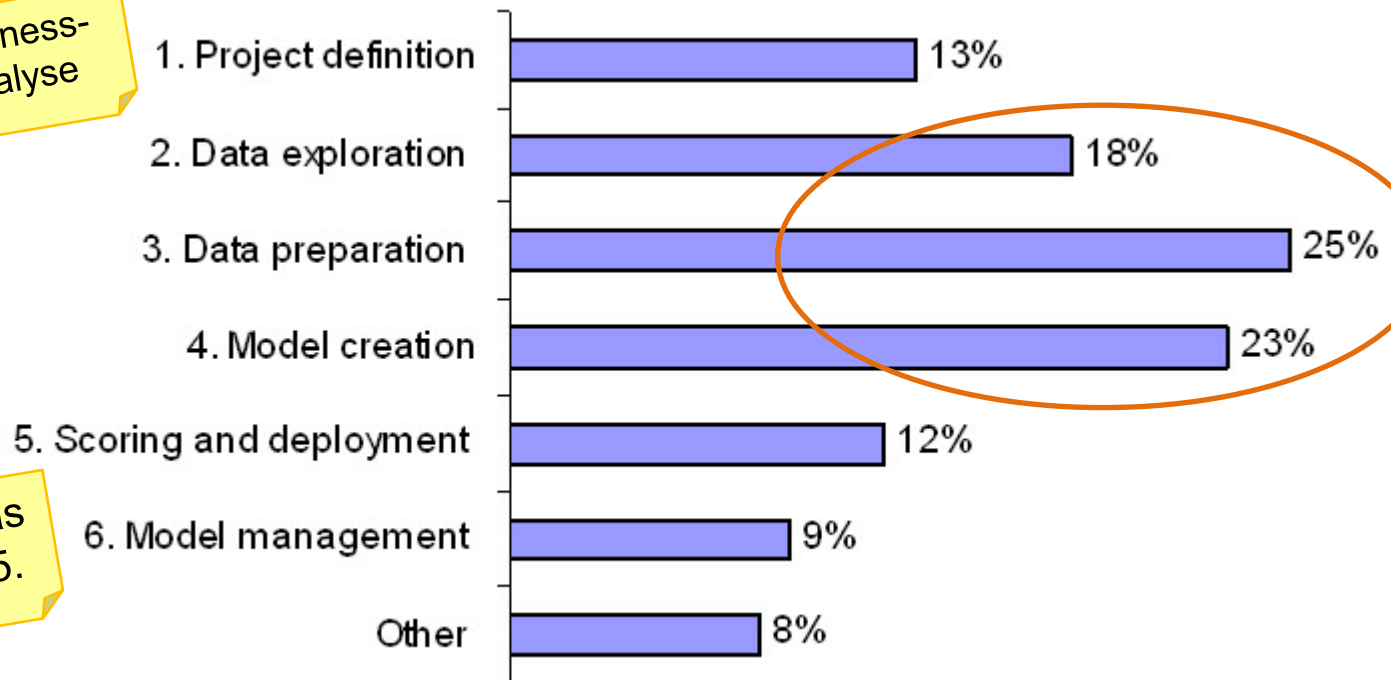
- « model scoring »
  - hoe goed *fit* het model de data?
  - training - validation - test
- Niet: feedback vanuit de acties ten gevolge van het toepassen van het model → « Monitor & maintain »

Analoog bij andere tools, zoals IBM-SPSS (CRISP-DM)

# Analytics project (resource) planning *naar een inschatting van benodigde effort ...*

## ■ Predictive modeler tasks

Business-  
analyse



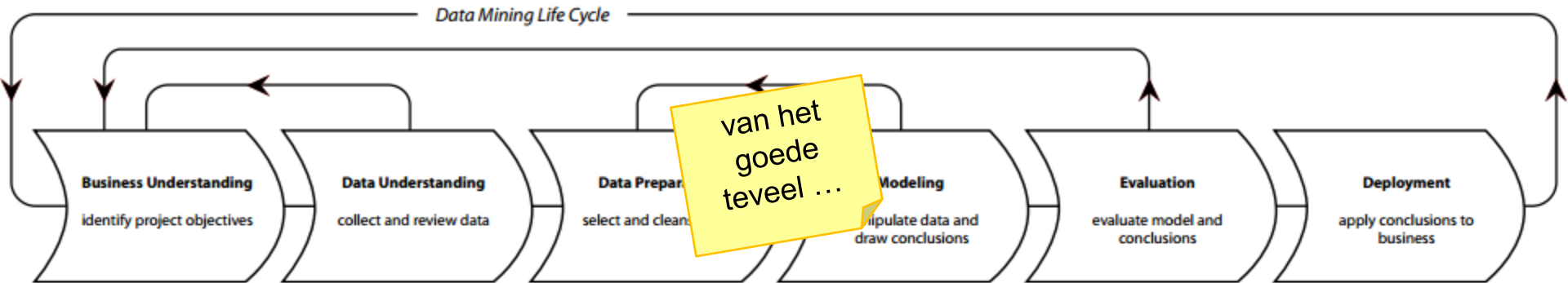
klant ziet pas  
wat vanaf 5.

Ontbreken:

- betrokkenheid andere stakeholders (Business, IT)
- solution architecture
- deeltaken

# CRISP-DM: onder de motorkap ...

## Phases



Determine Business Objectives	Collect Initial Data	Data Set	Select Modeling Technique	Evaluate Results	Plan Deployment
Initial Data Collection Report Modeling Assumptions (Log and Report Process)	Initial Data Collection Report Modeling Assumptions (Log and Report Process)	Data Set Description Mining Results (Log and Report Process)	Approved Models Review Process Review of Process (Log and Report Process)	Final Report Final Presentation (Log and Report Process)	Deployment Plan (Log and Report Process)
Select Data Rationalization (Log and Report Process)	Generate Test Design Test Design (Log and Report Process)	Determine Next Steps List of Possible Actions Decision (Log and Report Process)	Produce Final Report Experience Documentation (Log and Report Process)	Assess Situation Inventory of Resources, Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits (Log and Report Process)	Describe Data Data Description Report (Log and Report Process)
Clean Data Data Cleaning Report (Log and Report Process)	Build Model Parameter Settings Models Model Description (Log and Report Process)				Explore Data Data Exploration Report (Log and Report Process)
Construct Data Derived Attributes Generated Records (Log and Report Process)	Assess Model Model Assessment Revised Parameter (Log and Report Process)				Verify Data Quality Data Quality Report (Log and Report Process)
Integrate Data Merged Data (Log and Report Process)				Determine Data Mining Goals Data Mining Goals Data Mining Success Criteria (Log and Report Process)	
Format Data Reformatted Data (Log and Report Process)				Produce Project Plan Project Plan Initial Assessment of Tools and Techniques (Log and Report Process)	

## a visual guide to CRISP-DM methodology

SOURCE CRISP-DM 1.0  
<http://www.crisp-dm.org/download.htm>  
DESIGN Nicole Leaper  
<http://www.nicoleleaper.com>

**Generic Tasks**  
*Specialized Tasks*  
(Process Instances)

# werkbare CRISP-DM based project breakdown

Effort	Proces/fase	Taken	Stakeholder		
			Business	Analyst	IT
5%-10%	<b>Problem understanding</b>	Determine & refine objective	B	A	
		Define success criteria	B		
		Determine data mining goals	B	A	
10%-15%	<b>Data Understanding</b>	Collect initial data	B	A	
		Explore data & enhance insight	B	A	
		Verify data quality		A	I
30%-60%	<b>Data Preparation</b>	Select data, extract 'events' & 'behaviour'		A	
		Cleanse data		A	I
		Format data		A	I
20%-30%	<b>Modeling</b>	Select modeling technique(s)		A	
		Build models		A	
		Revise sampling strategy & cost functions		A	
20%-30%	<b>Evaluation of Results</b>	Compare & select model		A	
		Validate model	B	A	
		Explain model	B	A	
5%-10%	<b>Deployment</b>	Deploy model		A	I
		Score deployment (feedback)			I
		Monitor & maintain	B	A	

met enkele toevoegingen op basis van onze ervaring

# Analytics Project Management - caveat

Effort	Proces/fase	Taken	Stakeholder		
					IT
5%-10%	<b>Problem understanding</b>	Determine & refine objective			
		Define success criteria			
		Determine data mining goals			
10%-15%	<b>Data Understanding</b>	Collect initial data	B	A	
		Explore data & enhance insight	B	A	
		Verify data quality		A	I
30%-60%	<b>Data Preparation</b>	Select data, extract 'events' & 'behaviour'		A	
		Cleanse data		A	I
		Format data		A	I
20%-30%	<b>Modeling</b>	Select modeling technique(s)		A	
		Build models		A	
		Revise sampling strategy & cost functions		A	
20%-30%	<b>Evaluation of Results</b>	Compare & select model		A	
		Validate model	B	A	
		Explain model	B	A	
5%-10%	<b>Deployment</b>	Deploy model		A	I
		Score deployment (feedback)			I
		Monitor & maintain	B	A	

al te vaak te vaag gedefinieerd

Een project is immers niet « Data Mining » of « Fraudebestrijding »

→ dat is een dienstverlening

→ een project richt zich op een **welbepaalde modeling-taak** (risico, fraudetype, ...)

# Analytics Project Management - caveat

Effort	Proces/fase	Taken	Stakeholder		
			Business	Analyst	IT
5%-10%	<b>Problem understanding</b>	Determine & refine objective	B	A	
		Define success criteria	B		
		Determine data mining goals			
10%-15%	<b>Data Understanding</b>	Collect initial data			
		Explore data & <b>enhance insight</b>			
		Verify data quality			I
30%-60%	<b>Data Preparation</b>	Select data, <b>extract</b> 'events' & 'behaviour'		A	
		Cleanse data		A	I
		Format data		A	I
20%-30%	<b>Modeling</b>	Select modeling technique(s)		A	
		Build models		A	
		<b>Revise sampling strategy &amp; cost functions</b>		A	
20%-30%	<b>Evaluation of Results</b>	Compare & select model		A	
		Validate model	B	A	
		Explain model	B	A	
5%-10%	<b>Deployment</b>	Deploy model		A	I
		Score deployment ( <b>feedback</b> )			I
		Monitor & maintain	B	A	

Data quality & Visualisation tools, clustering & association



**Pitfall:** blijven exploreren en verfijnen

Remedie: « agile », « timeboxing », milestones ← bv. 1ste model na 3 maand  
 iteratief karakter maakt oplevering van telkens accuratere modellen mogelijk

# Analytics Project Management - caveat

Effort	Proces/fase	Taken	Stakeholder		
			Business	Analyst	IT
5%-10%	<b>Problem understanding</b>	Determine & refine objective	B	A	
		Define success criteria	B		
		Determine data mining goals	B	A	
10%-15%	<b>Data Understanding</b>	Collect initial data	B	A	
		Explore data & enhance insight	B	A	
		Verify data quality		A	I
30%-60%	<b>Data Preparation</b>	Select data, <b>extract 'events' &amp; 'behaviour'</b>			
		Cleanse data			I
		Format data			I
20%-30%	<b>Modeling</b>	<b>Select modeling technique(s)</b>			
		Build models			
		<b>Revise sampling strategy &amp; cost functions</b>		A	
20%-30%	<b>Evaluation of Results</b>	Compare & select model		A	
		Validate model	B	A	
		Explain model	B	A	
5%-10%	<b>Deployment</b>	Deploy model		A	I
		Score deployment (feedback)			I
		Monitor & maintain	B	A	

individuele Skills van de Analytics-expert zijn bepalend

Extract 'events', 'behaviour', 'network' vars uit ruwe variabelen

→ bv. Aantal adreswijzigingen in bepaald tijdsvenster vóór event X

Een woordje extra uitleg over het selecteren van de juiste modeling-techniek

# Selecteren van modeling-techniek: *bestaat een « beste » techniek?*

- Een ideale techniek
  - moet kunnen omgaan met:
    - mixed datatypes (continu, discreet, ordinaal, categorisch)
    - missing values
    - irrelevante input
    - gecorreleerde input
    - grote datasets
  - en is bovendien:
    - robuust t.o.v. outliers in de input
    - ongevoelig voor monotone transformaties van invoervariabelen
    - vlot interpreteerbaar
    - accuraat & stabiel
- Zo'n techniek zou dan zonder veel pre-processing van invoerdata en zonder veel tuning van parameters toepasbaar zijn

# Selecteren van modeling-techniek

## vuistregels Smals Research Predictive Analytics Competence Center

Criterion								

# Analytics Project Management - caveat

Effort	Proces/fase	Taken	Stakeholder		
			Business	Analyst	IT
5%-10%	<b>Problem understanding</b>	Determine & refine objective	B	A	
		Define success criteria	B		
		Determine data mining goals	B	A	
10%-15%	<b>Data Understanding</b>	Collect initial data	B	A	
		Explore data & enhance insight	B	A	
		Verify data quality		A	I
30%-60%	<b>Data Preparation</b>	Select data, extract 'events' & 'behaviour'		A	
		Cleanse data		A	I
		Format data		A	I
20%-30%	<b>Modeling</b>	Select modeling technique(s)		A	
		Build models		A	
		Revise sampling strategy & cost functions		A	
20%-30%	<b>Evaluation of Results</b>	Compare & select model		A	
		Validate model	B	A	
		Explain model	B	A	
5%-10%	<b>Deployment</b>	Deploy model			I
		Score deployment ( <b>feedback</b> )			I
		Monitor & maintain			

Niet te onderschatten!

Succes van het project is slechts meetbaar, traceerbaar ALS ook de feedback (de resultaten van acties ten gevolge van het toegepaste model) deftig geregistreerd en opnieuw geëxploiteerd geraakt

# Analytics Project Management - caveat

Succes van het project is slechts meetbaar, traceerbaar ALS ook de feedback (de resultaten van acties ten gevolge van het toegepaste model) deftig geregistreerd en opnieuw geëxploiteerd geraakt

- **Dit vergt:**
  - **nieuwe business-processen**
    - afhandelprocessen
    - actie t.g.v. modelpredicties
  - **registratie van feedback**
    - gestructureerd → exploiteerbaar
    - gestuurd → de juiste informatie met de juiste granulariteit
  - **bijsturen van modellen** in  $f(\text{feedback})$
  
- Dit betekent meestal dat de **solution architecture** voor dit volledige proces **een stuk ruimer** is dan typisch voorzien
  - positief is dat vele elementen hiervan herbruikbaar zijn voor een volgende predictief model

# Analytics Project Management - caveat

*"You can't model what you don't have (examples for)"*

- De meeste methodologieën en succesverhalen gaan ervan uit dat supervised learning-technieken toepasbaar zijn

- **Dit veronderstelt:**

- voldoende voorbeelden zijn beschikbaar
  - voor de trainingsfase van een eerste (predictief) model
  - van het welbepaald type probleem waarvoor men een predictief model wenst te deployen

evenals tegen-voorbeelden

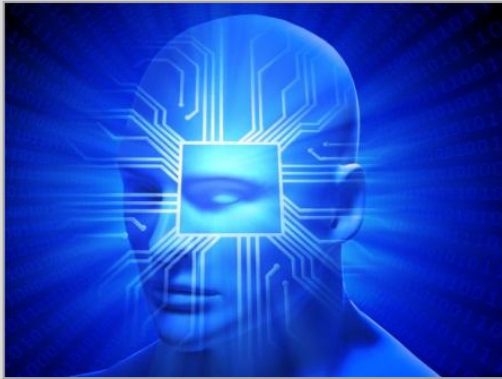
- **Zoniet**

- dient men te vertrekken van hypothesen,
- zullen andere technieken ingezet worden, en
- zullen de eerste resultaten slechts dienen om echte feedback / de gezochte feedback te verzamelen

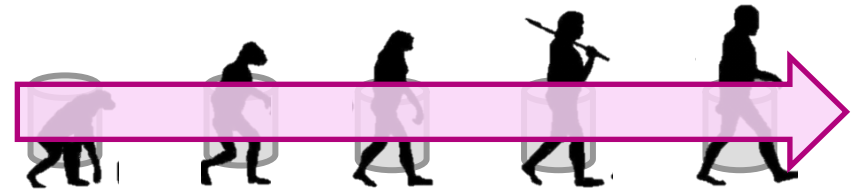
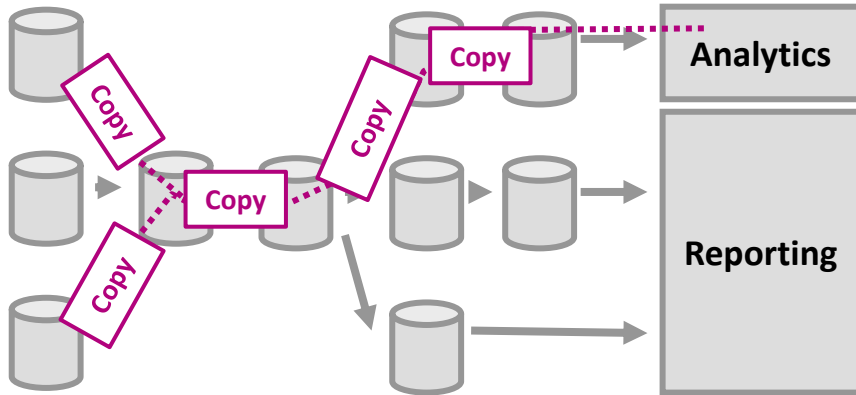
dit komt vaker voor dan men denkt!

- Ook dit zijn **nieuwe business-processen**
  - waarvoor de organisatie doorgaans niet klaar is

# Conclusie

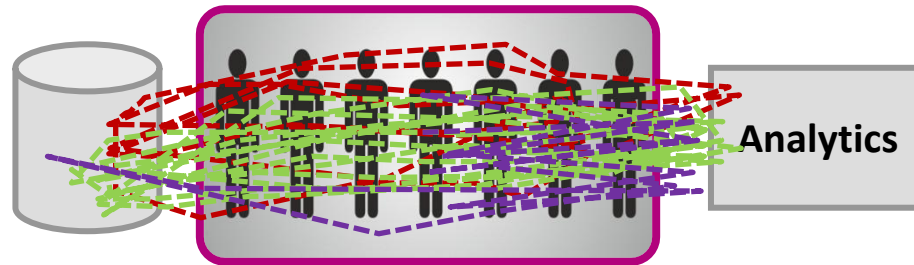


# Barrières bij de introductie van analytics



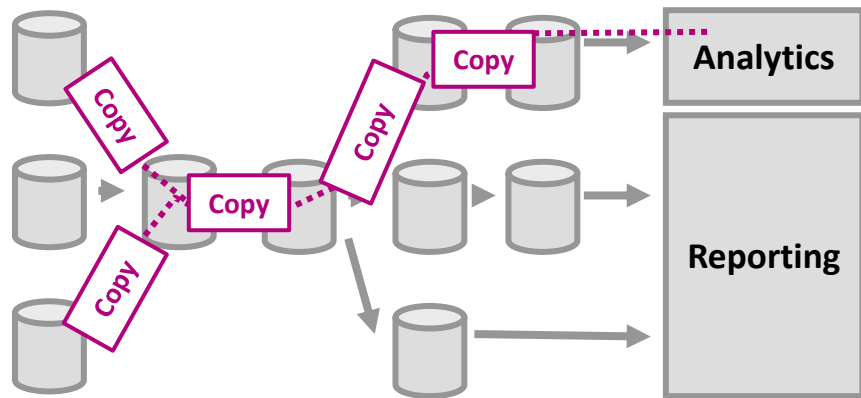
1. Complexe & trage architectuur

2. Data quality doorheen de supply chain

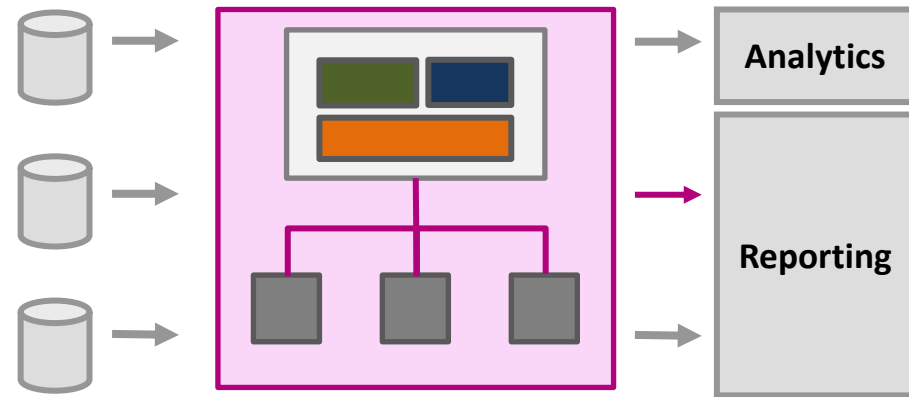


3. Atypische projectstructuur

# Streamlining analytics

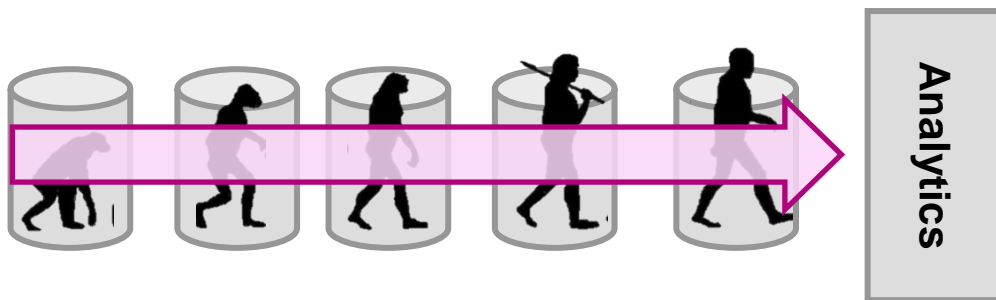


Complexe & trage architectuur



Hardware appliance

# Streamlining analytics

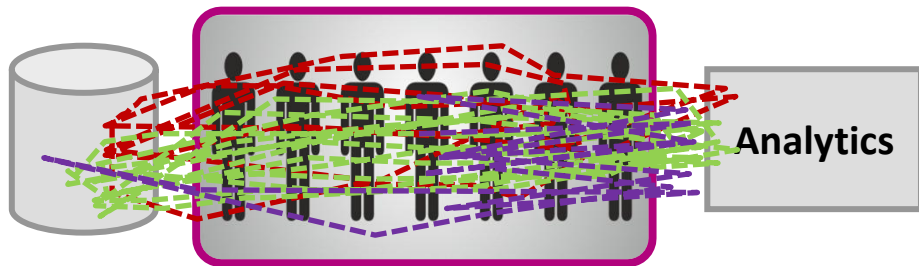


**Data quality doorheen de supply chain**



**Data quality aan de bron**

# Streamlining analytics

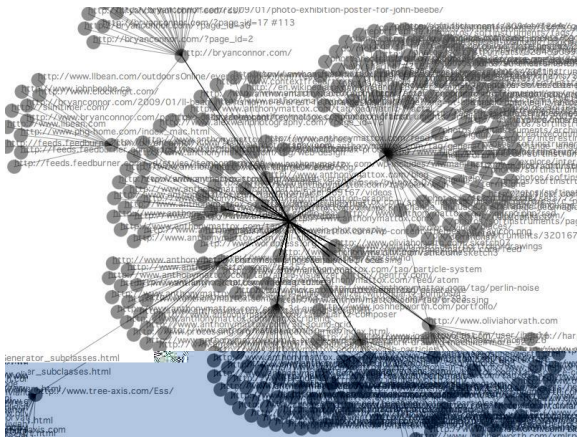


Effort	Proces/fase
5%-10%	<b>Problem understanding</b>
10%-15%	<b>Data Understanding</b>
30%-60%	<b>Data Preparation</b>
20%-30%	<b>Modeling</b>
20%-30%	<b>Evaluation of Results</b>
5%-10%	<b>Deployment</b>

**Atypische projectstructuur**

**CRISP-DM based project  
breakdown**

# Barrières hangen samen met de analytics/BI behoeftes



Interactieve visualisatie van Big Data



Operational BI



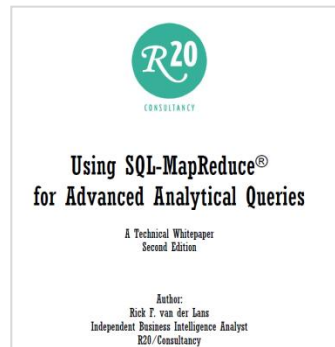
Mobile BI

# Bibliografie



Jack E. Olson,  
"Data Quality – The Accuracy Dimension"

<http://www.b-eye-network.com/view/13506>







## The Advantages of Data Warehouse Appliances Revisited

by Rick van der Lans

ORIGINALLY PUBLISHED MAY 4, 2010

Much has already been written about data warehouse appliances and their advantages. In most articles, query performance improvement is described as the primary advantage. In this article, we take a slightly different viewpoint. Data warehouse appliances can definitely improve performance, but query performance improvement is not really an advantage – it's a property. However, this performance improvement property is the basis for certain advantages, and this article focuses on those advantages.

-  PRINTER-FRIENDLY
-  EMAIL TO A FRIEND
-  EMAIL TO MYSELF
-  LISTEN NOW
-  DOWNLOAD MP3
-  COMMENTS

<http://www.teradata.be/white-paper/Using-SQL-MapReduce-for-Advanced-Analytical-Queries/>



# Bibliografie



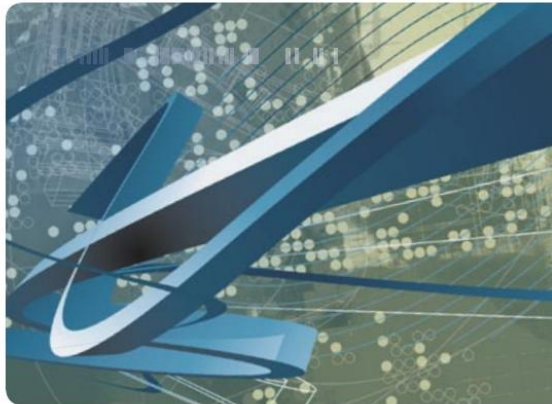
<http://www.kdd.org>

## PREDICTIVE ANALYTICS

Extending the Value of Your  
Data Warehousing Investment

By Wayne W. Eckerson

**Wayne Eckerson,**  
**"Predictive Analytics: Extending the Value of Your Data  
Warehousing Investment,"**  
**2007**  
**[166 respondents]**



# Documentatie.smals.be

- *Dries Van Dromme, 09/2007, Data Quality: Tools - Evaluer et améliorer la qualité des données*
- *Dries Van Dromme, 03/2011, Gestion intégrée des anomalies - Evaluer et améliorer la qualité des données*
- *Isabelle Boydens, 05/2006, Data Quality – Best Practices*
- *Jan Meskens, 12/2011, Predictive Analytics*
- *Grégory Ogonowski, NoSQL – Hype ou Innovation?*

# Vragen?

[Jan.Meskens@smals.be](mailto:Jan.Meskens@smals.be)  
[Dries.VanDromme@smals.be](mailto:Dries.VanDromme@smals.be)

Onderzoek - [@SmalsResearch](https://twitter.com/SmalsResearch)  
<http://blogresearch.smalsrech.be>