Streamlining Analytics

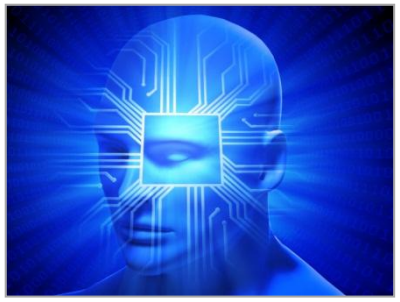
Jan.Meskens@smals.be
Dries.VanDromme@smals.be

Onderzoek - @SmalsResearch
http://blogresearch.smalsrech.be

Streamlining Analytics

Predictive analytics

De **data supply chain**



Barrières bij de introductie van analytics

Hardware appliances voor analytics

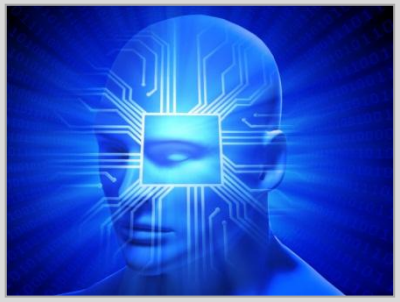
Data quality

Analytics **project management**

Streamlining Analytics

Predictive analytics

De **data supply chain**



Barrières bij de introductie van analytics

Hardware appliances voor analytics

Data quality

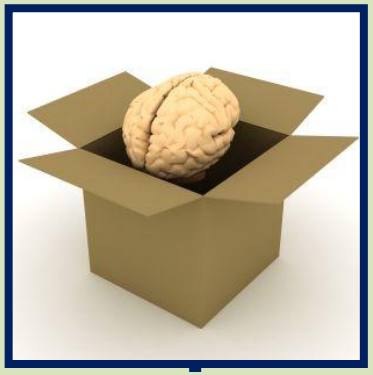
Analytics **project management**

Data

Voorspellingen over de toekomst

Fase 1: Training

Predictief
Model



Algoritmes

Historische data

Nieuwe data

Voorspellingen
over de
toekomst

Fase 2: Toepassing

PA voorbeeld: **marketing analytics**

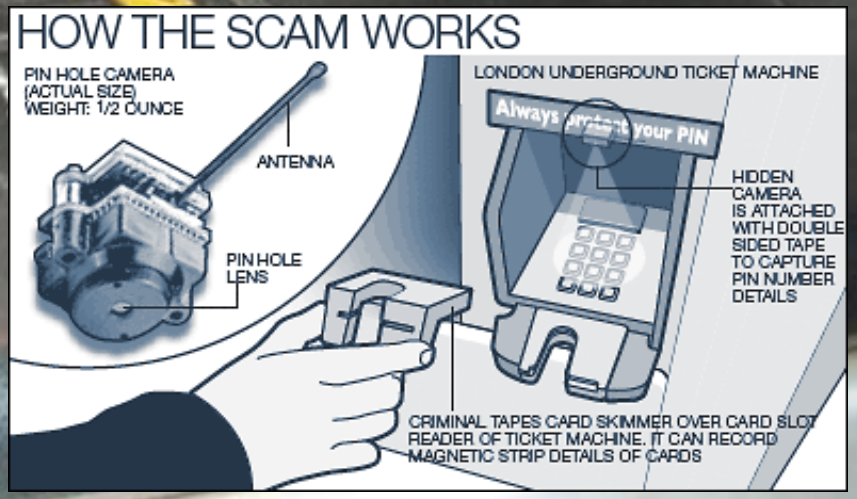
Voorspel wat kopers nodig hebben

Toon advertenties voor deze producten

PA voorbeeld: **fraudebestrijding**

Voorspel of een transactie frauduleus is

Blokkeer automatisch kaarten, rekeningen,…



PA voorbeeld: **churn prediction**

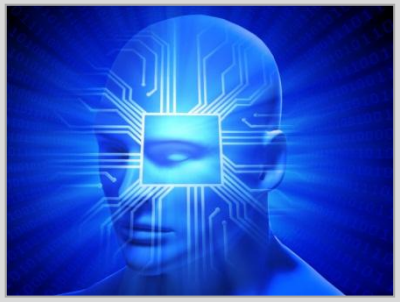
Voorspel of een klant gaat overlopen naar de concurrentie

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

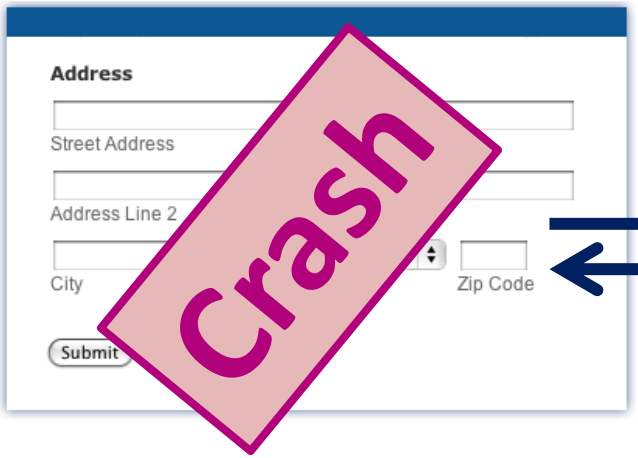
Analytics **project management**

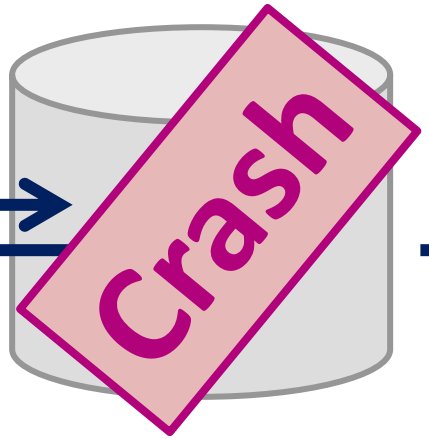
Analytics vs. **Productie**-applicaties

Applicatie

Data

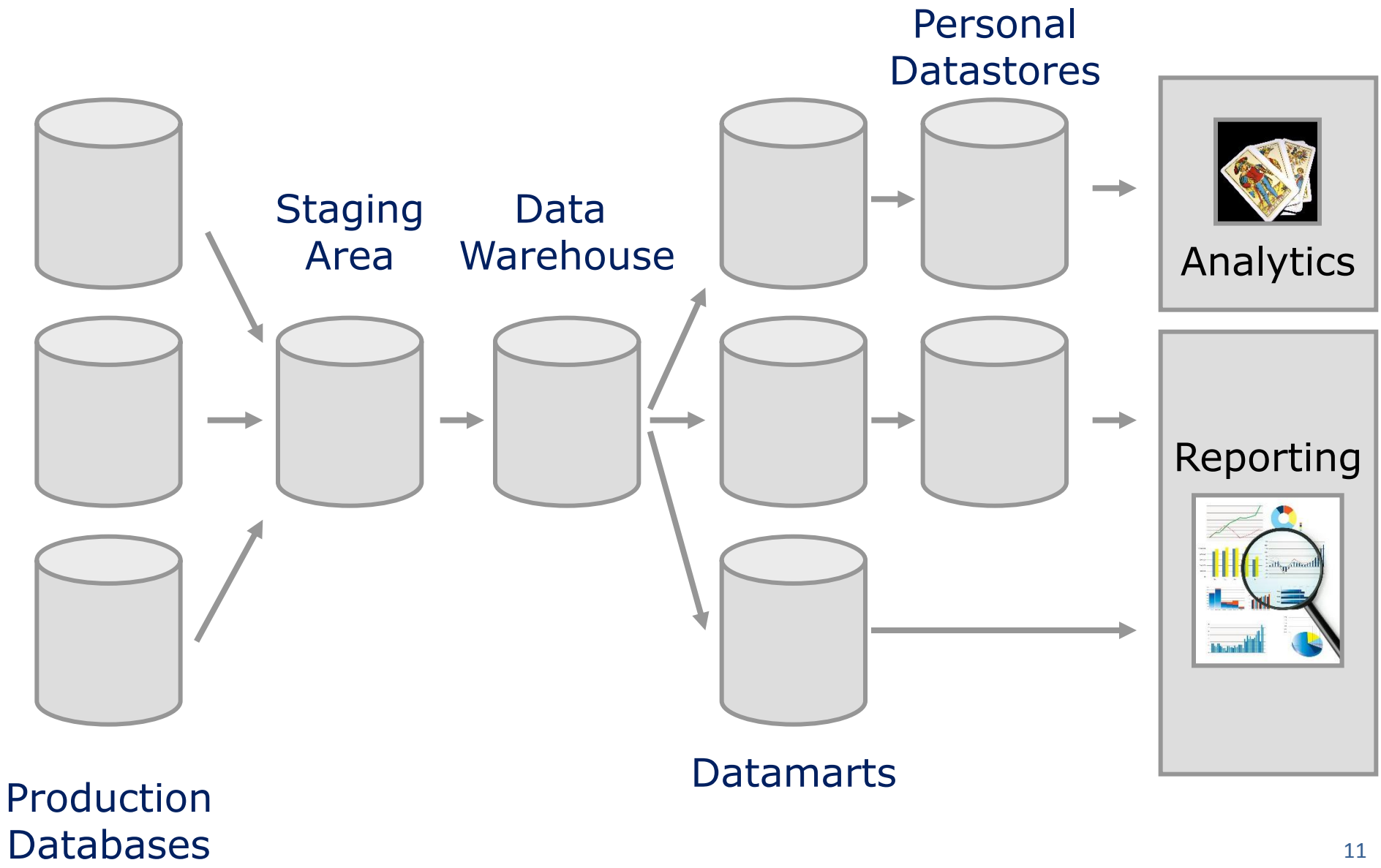
Analytics



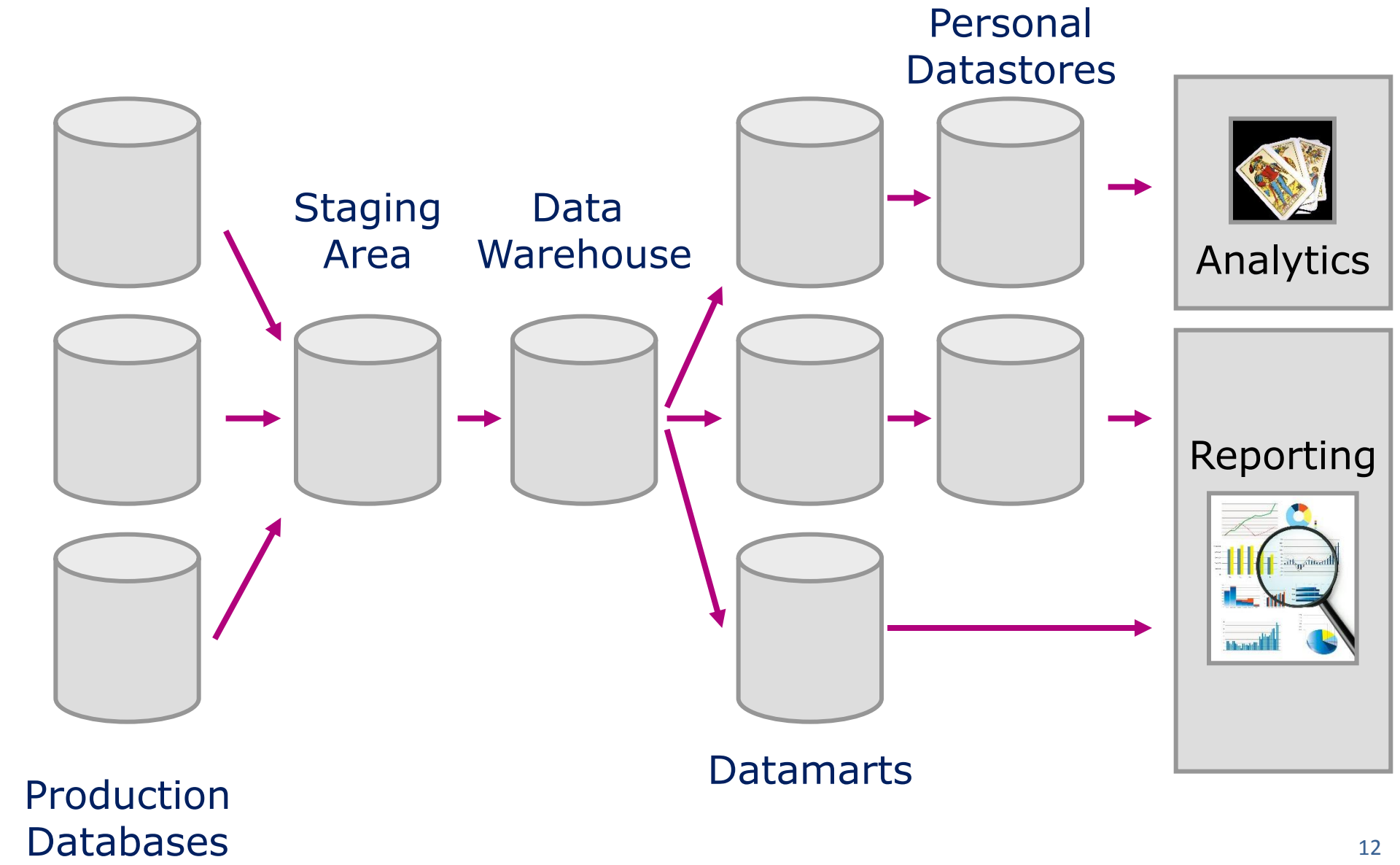


10

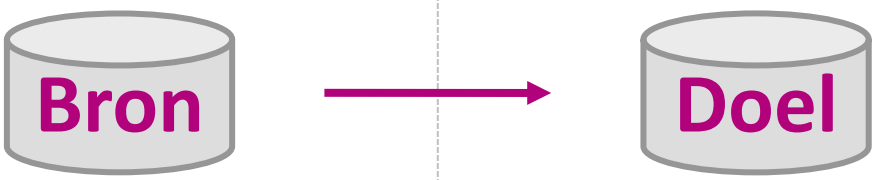
De « **data supply chain** »



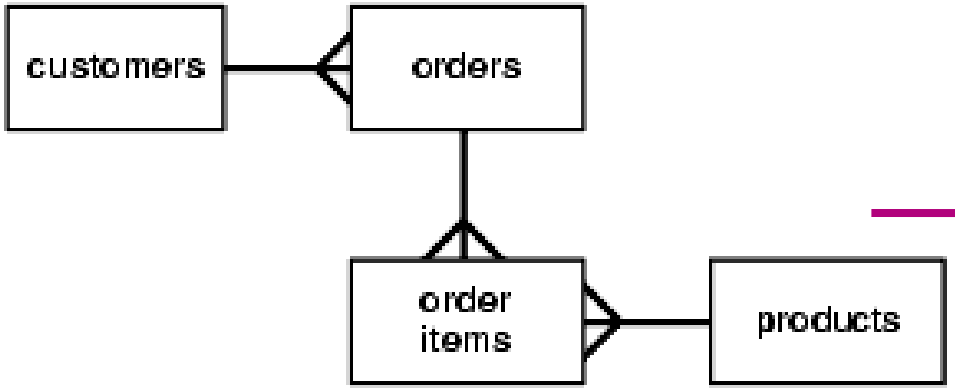
De « **data supply chain** »

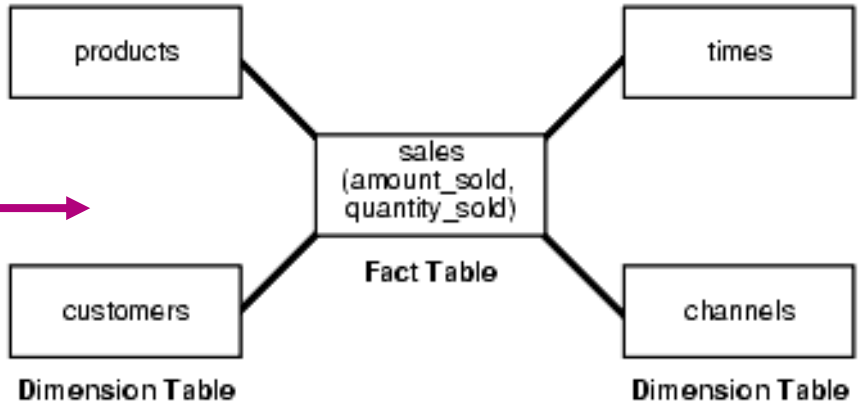


ETL: Extract Transform Load



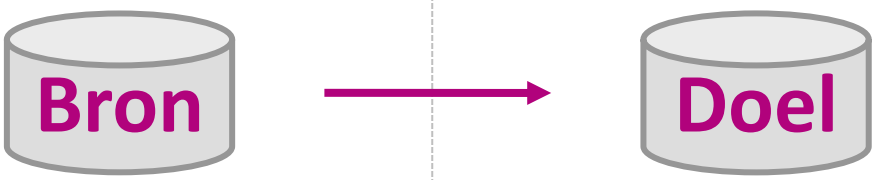
Omvormen **data-schema**





Derde normaalvorm **Ster-schema**

ETL: Extract Transform Load



Data-quality operaties

Adres correcties

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

Harmonisatie afkortingen

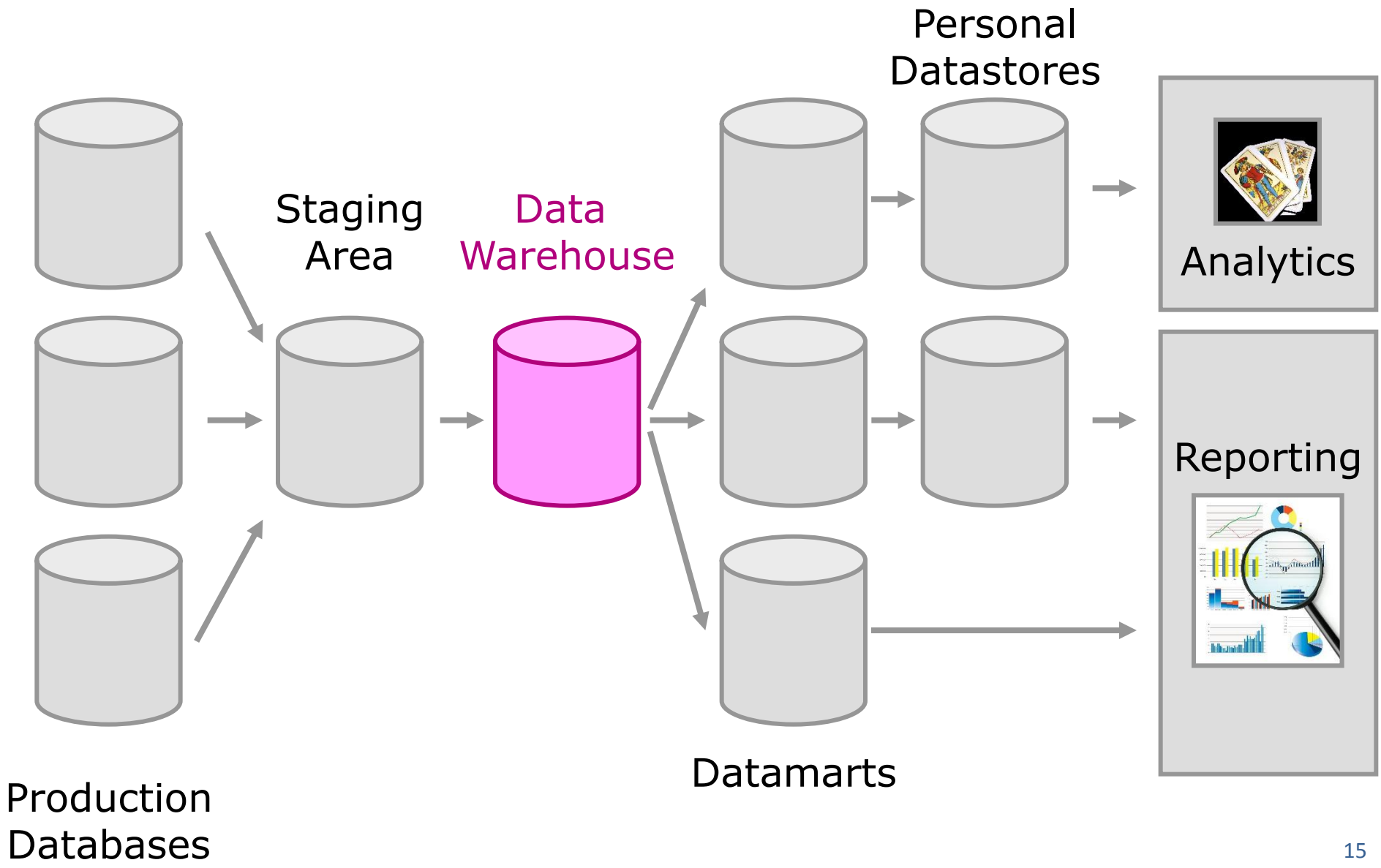
SPRL → 2200

N.v.. → NV

Missing values?

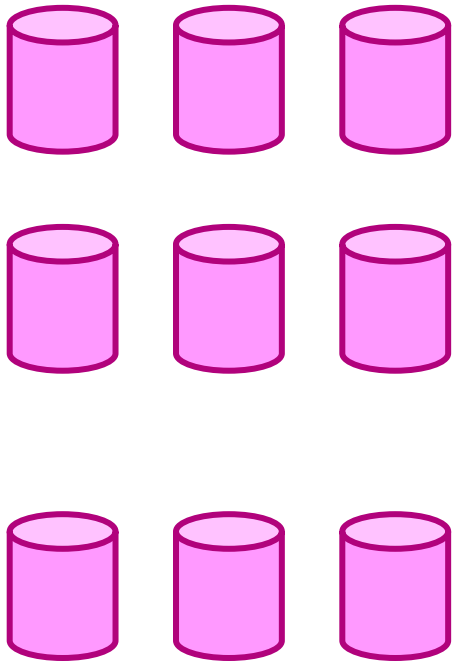
...

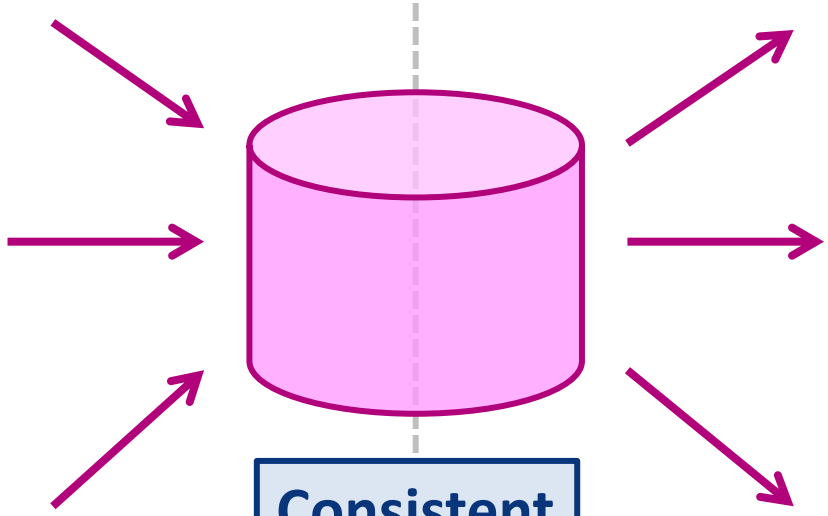
De « **data supply chain** »



Het **data warehouse (DWH)**

Veel operationele **data sources**





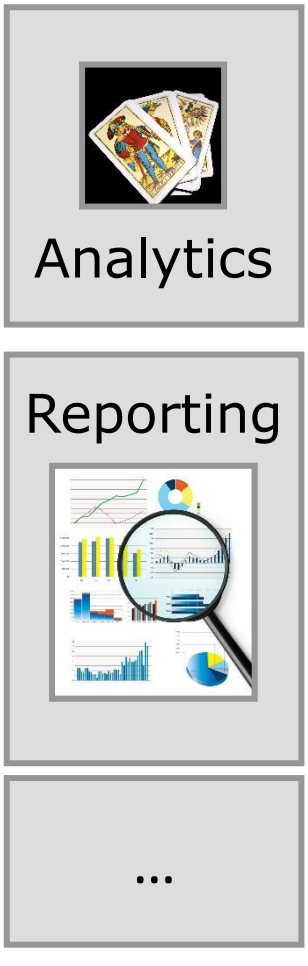
Consistent

Uniform data-schema

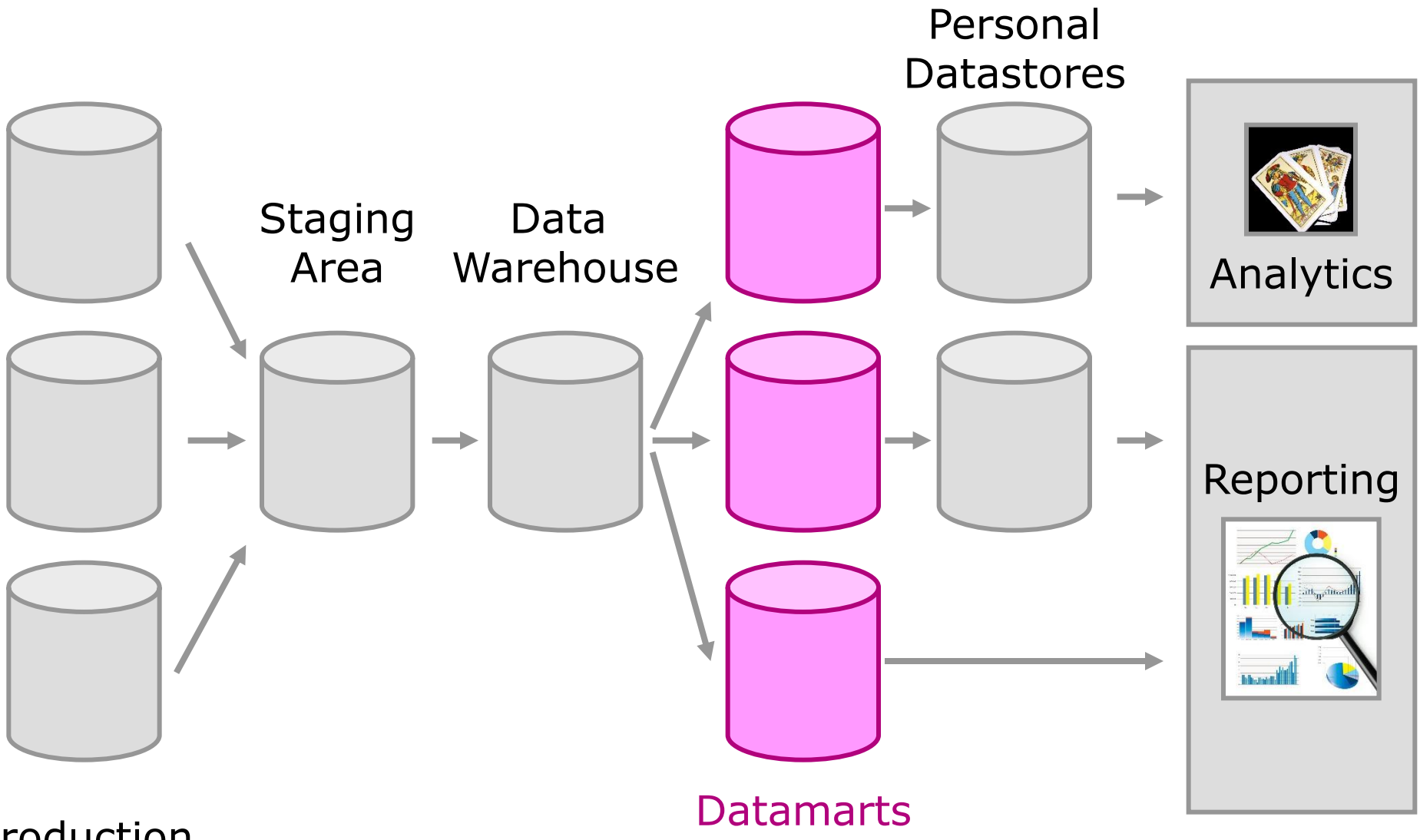
Historieken - versies

Geoptimaliseerd voor output

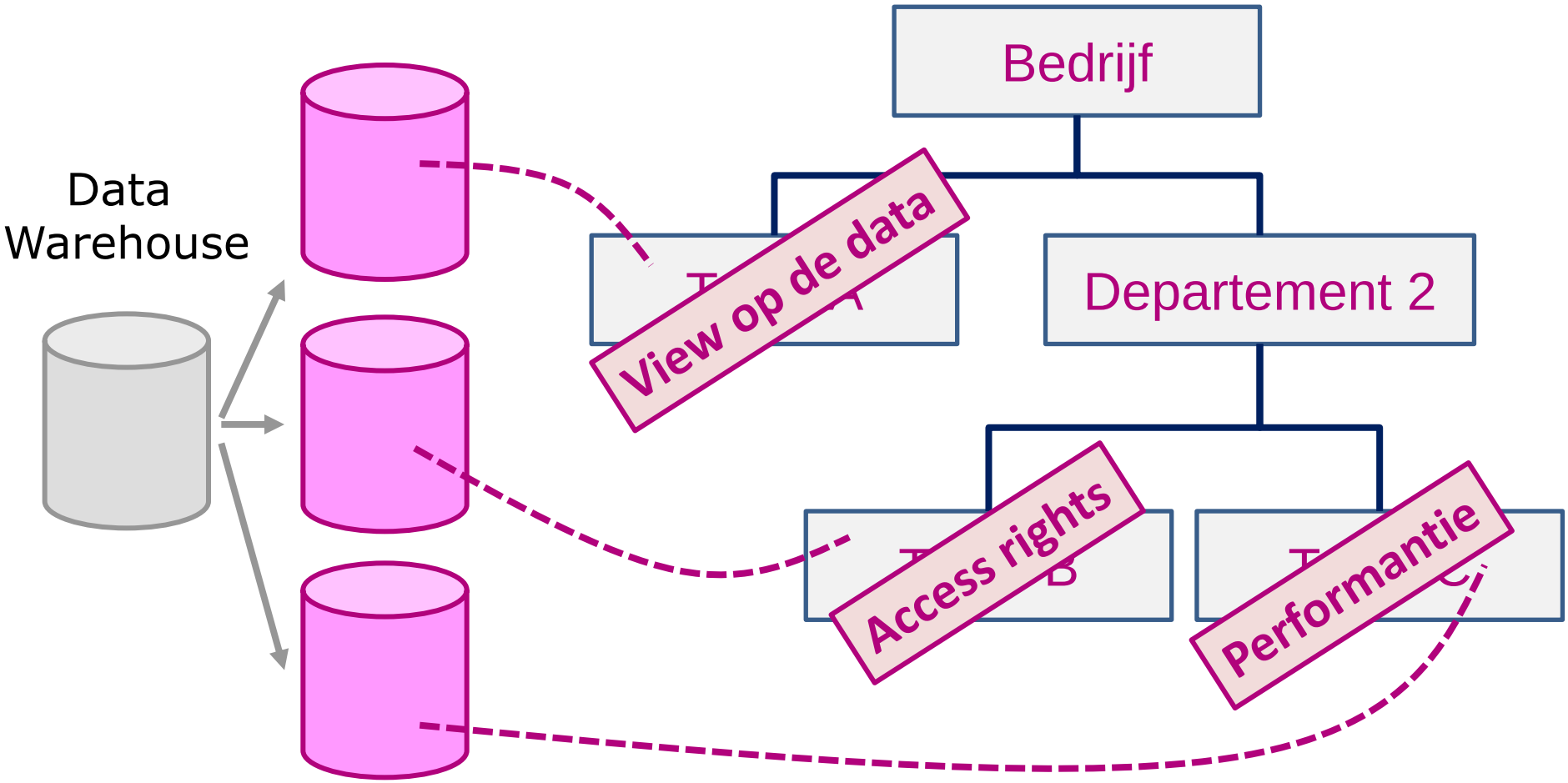
Veel **toepassingen**



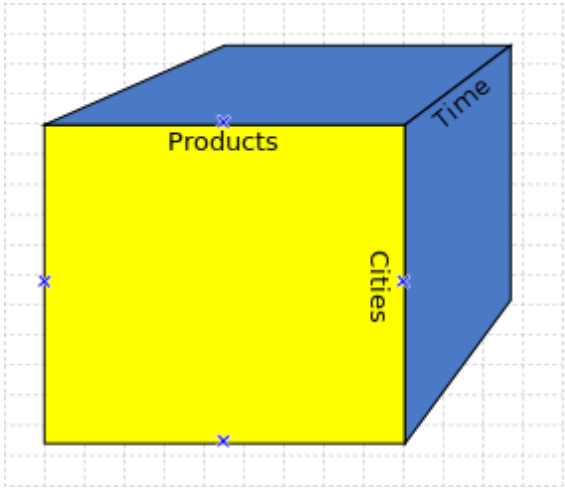
De « **data supply chain** »

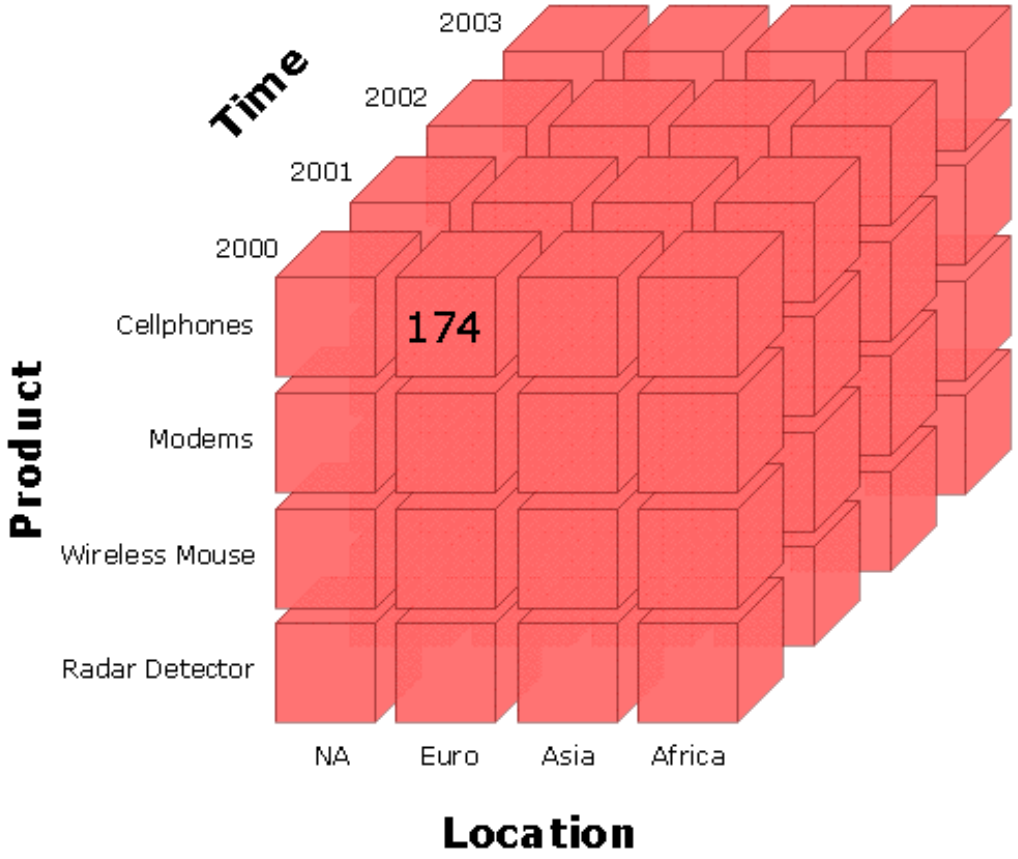


Data marts

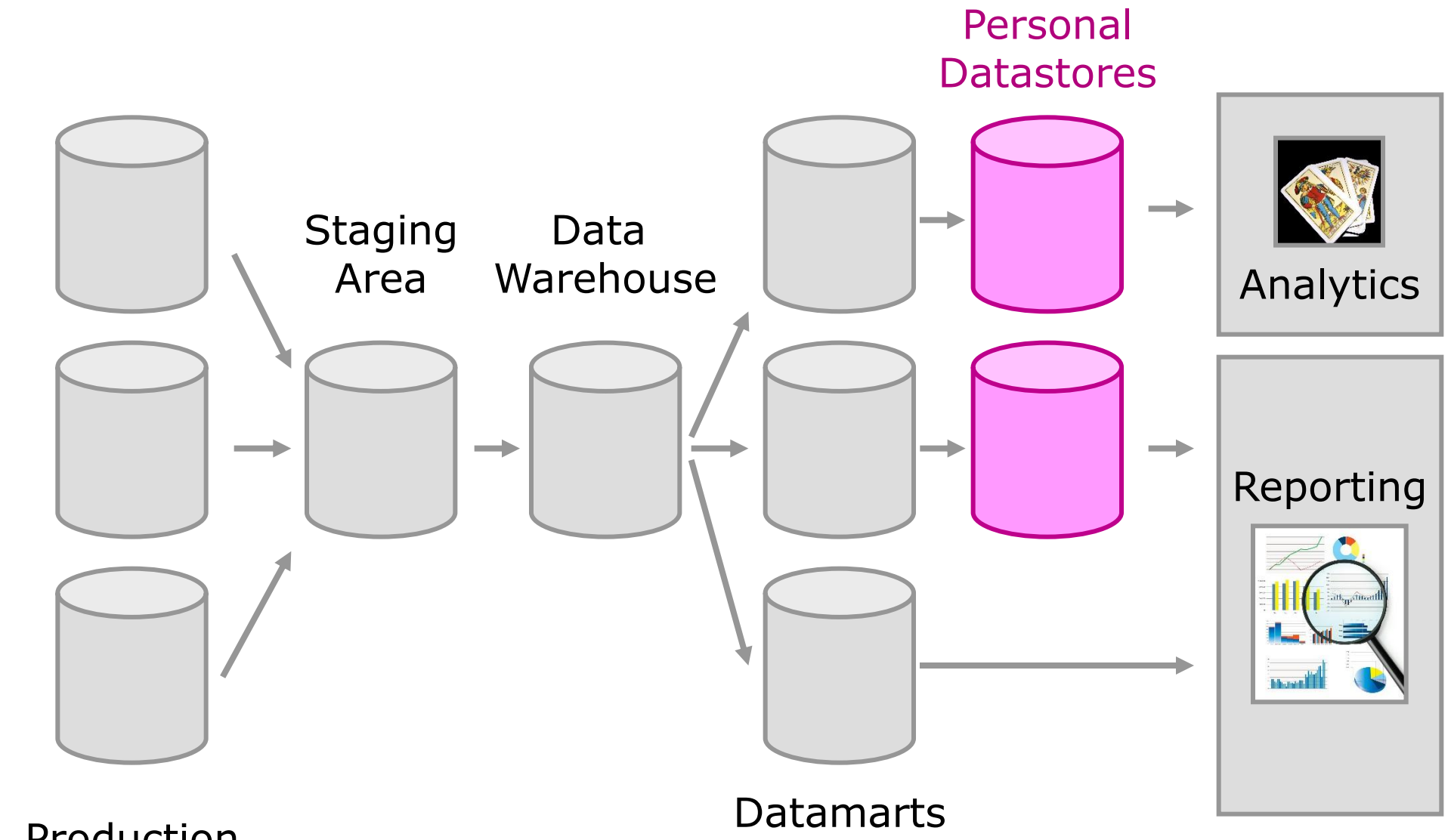

Data Warehouse
Bedrijf
Departement 2
View op de data
Access rights
Performantie

Vaak gebruikt **data-model:** **O**n**L**ine **A**nalytical **P**rocessing (OLAP)





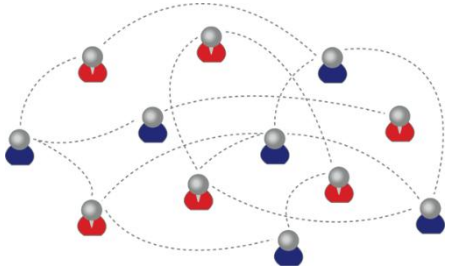
De « **data supply chain** »



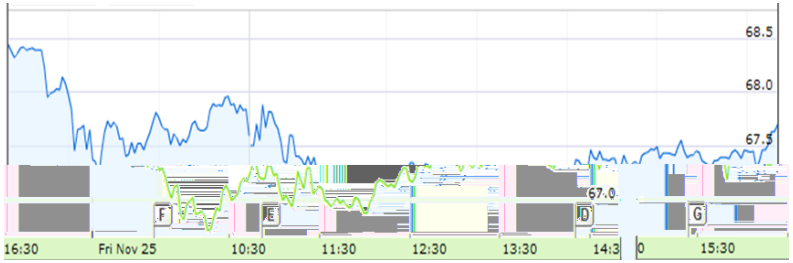
Smals
ICT for society

Data-warehouse / data-mart

Images, Video



Netwerk data



Continuous data

Time series

Analytics
personal datastore

	A	B	C	D	E
1	**X1**	**X2**	**X3**	**X4**	**Y?**
2	1	2	1	1	**TRUE**
3	1	1	1	1	**TRUE**
4	0	3	1	1	**FALSE**
5	1	2	1	1	**TRUE**

Set independent variables (X1,…,Xn) en 1 dependent variable (Y)

Smals
ICT for society

Data-warehouse / data-mart

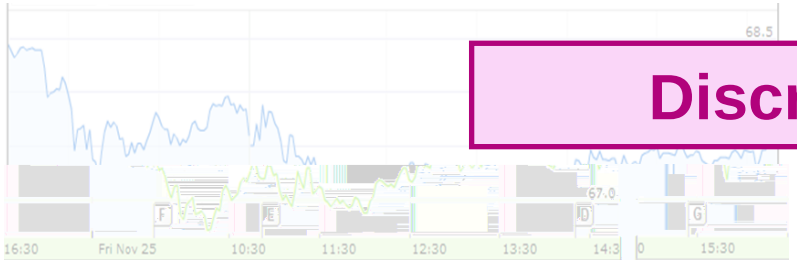
Analytics
personal datastore

Images, Video

Netwerk data

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

Set independent variables (X1,...,Xn) en 1 dependent variable (Y)



Continuous data

Time series

Feature extraction

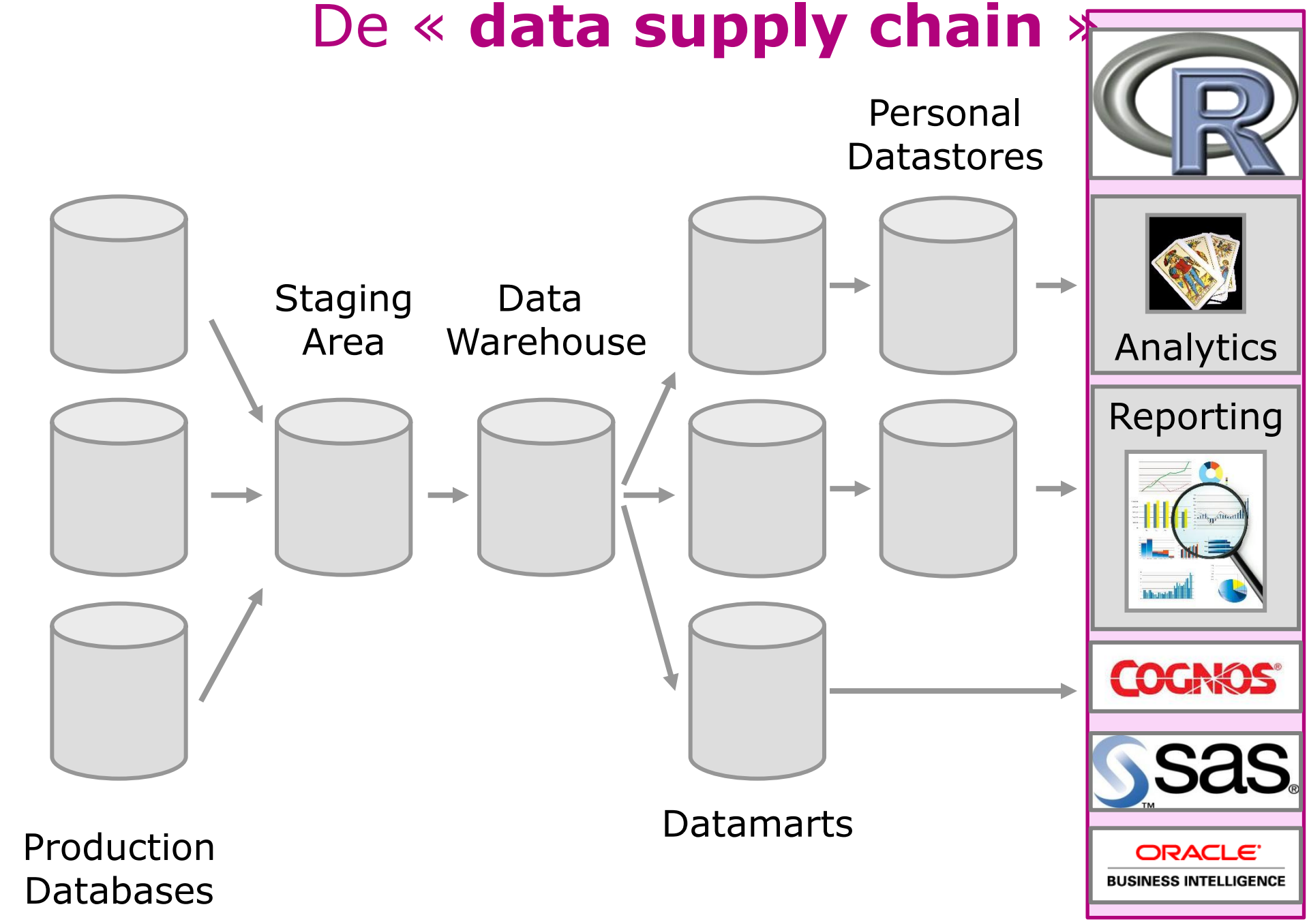
Graaf algoritmes

Discretisatie

Herschaling

Smals
ICT for society

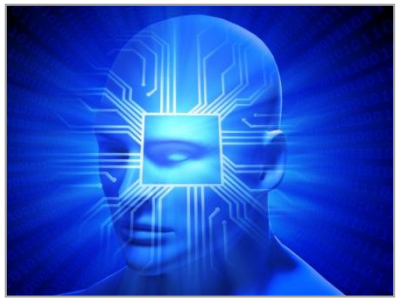
De « **data supply chain** »



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 »

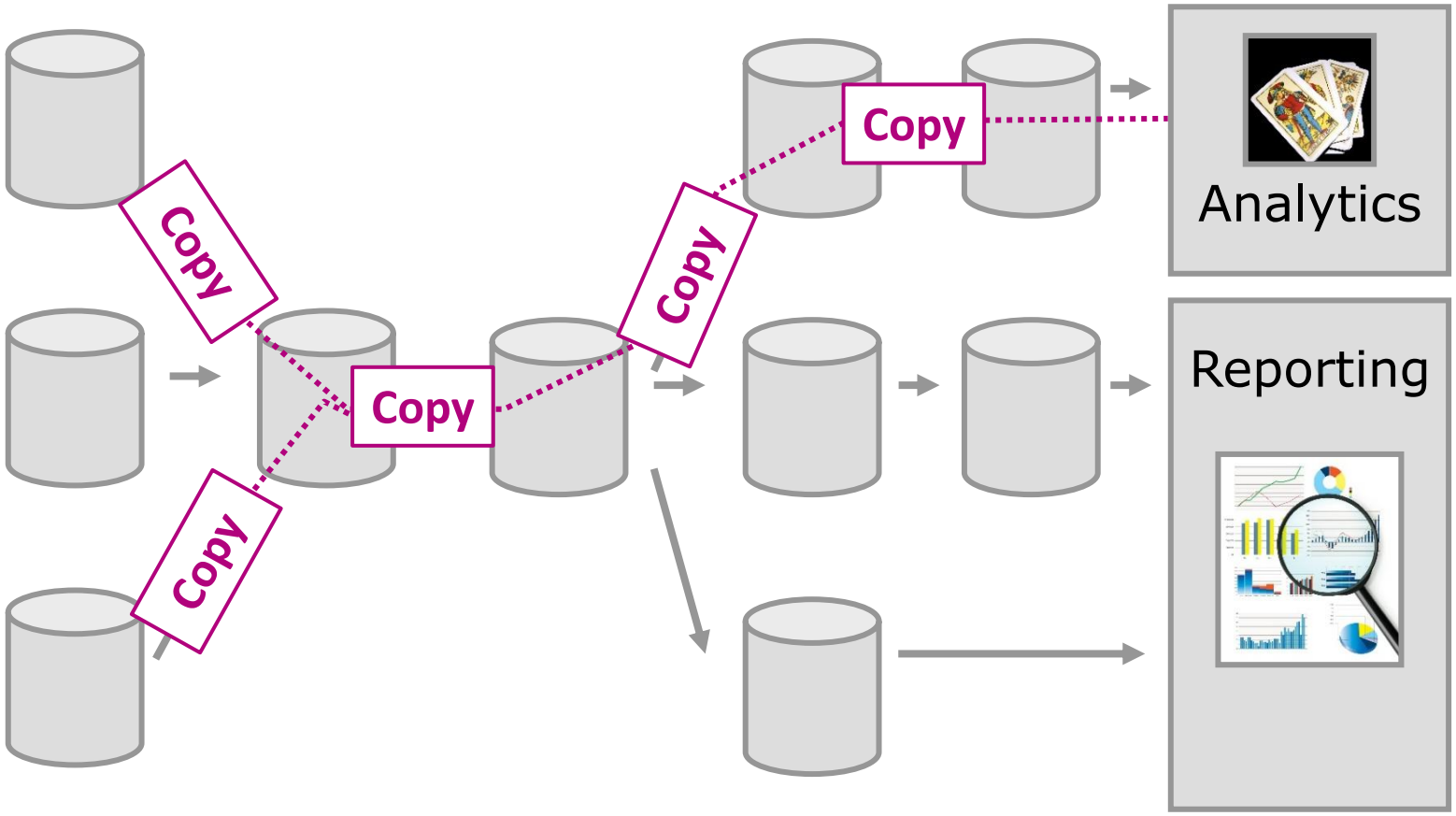
« Servers niet krachtig genoeg »

« Waar is de data? »

« Betrokkenheid van veel rollen/teams »

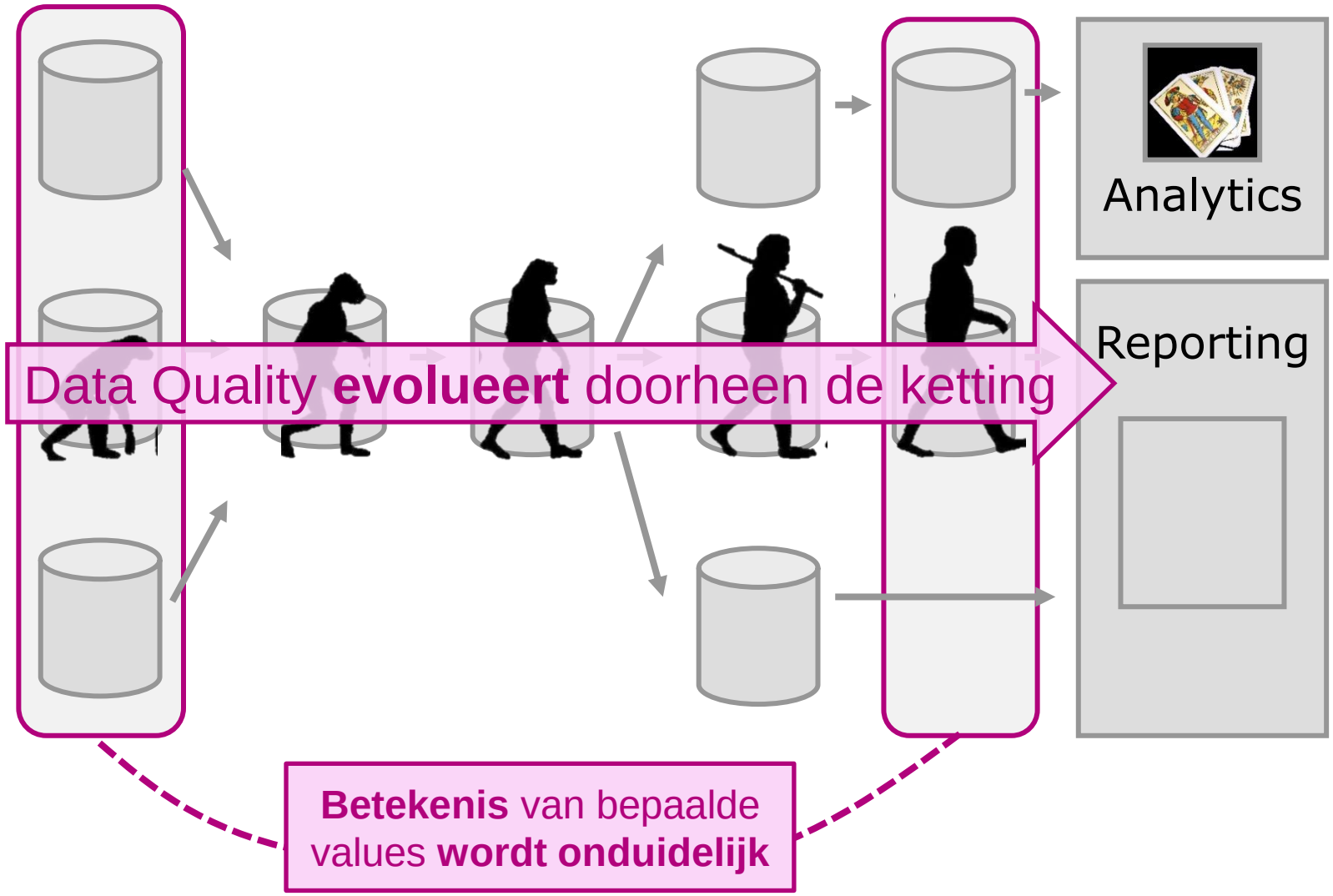
[Eric Siegel, Predictive Analytics, 2013]

Barrière 1:
Complexe en trage architectuur

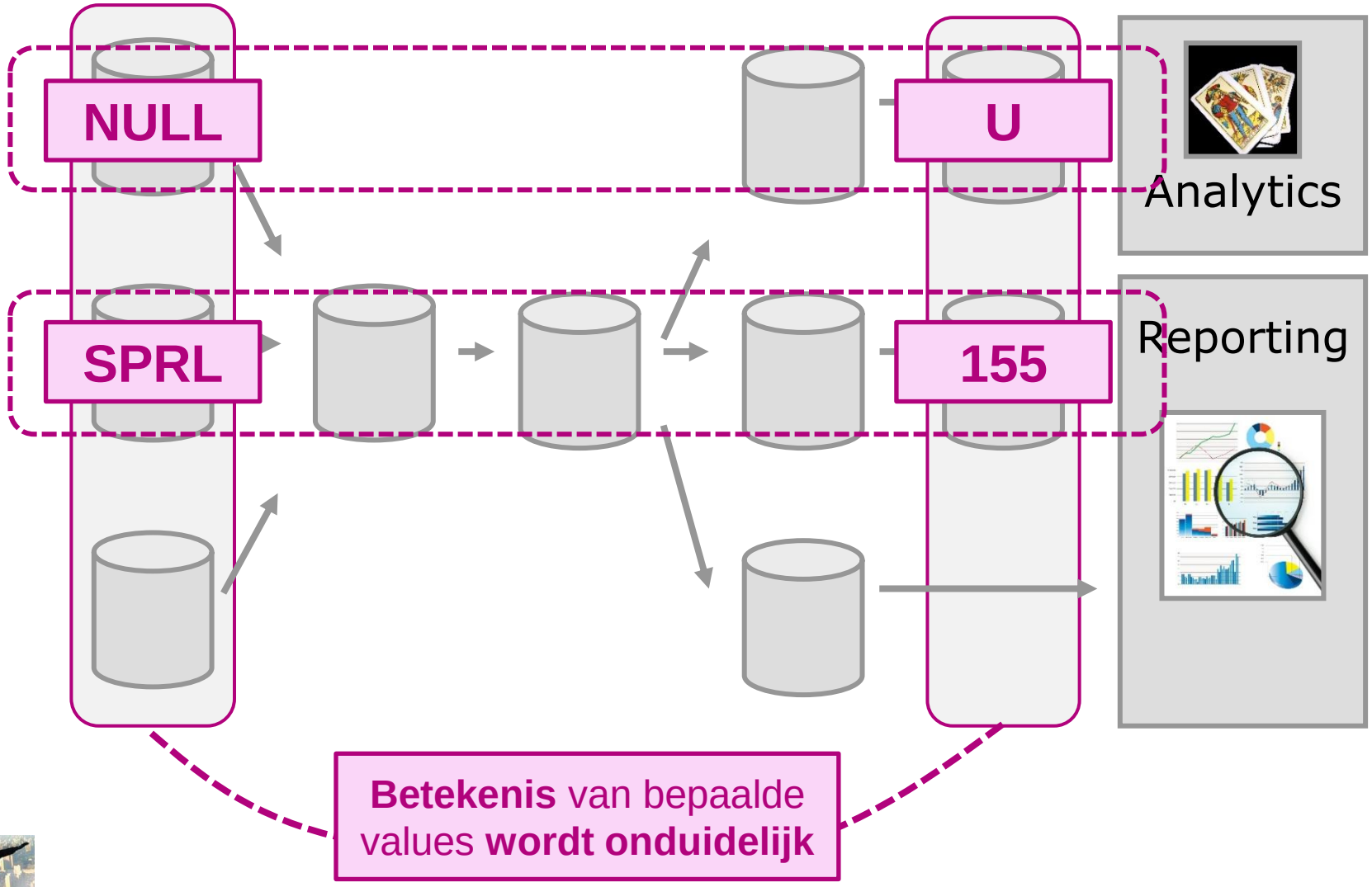


Barrière 1:
Complexe en **trage** architectuur

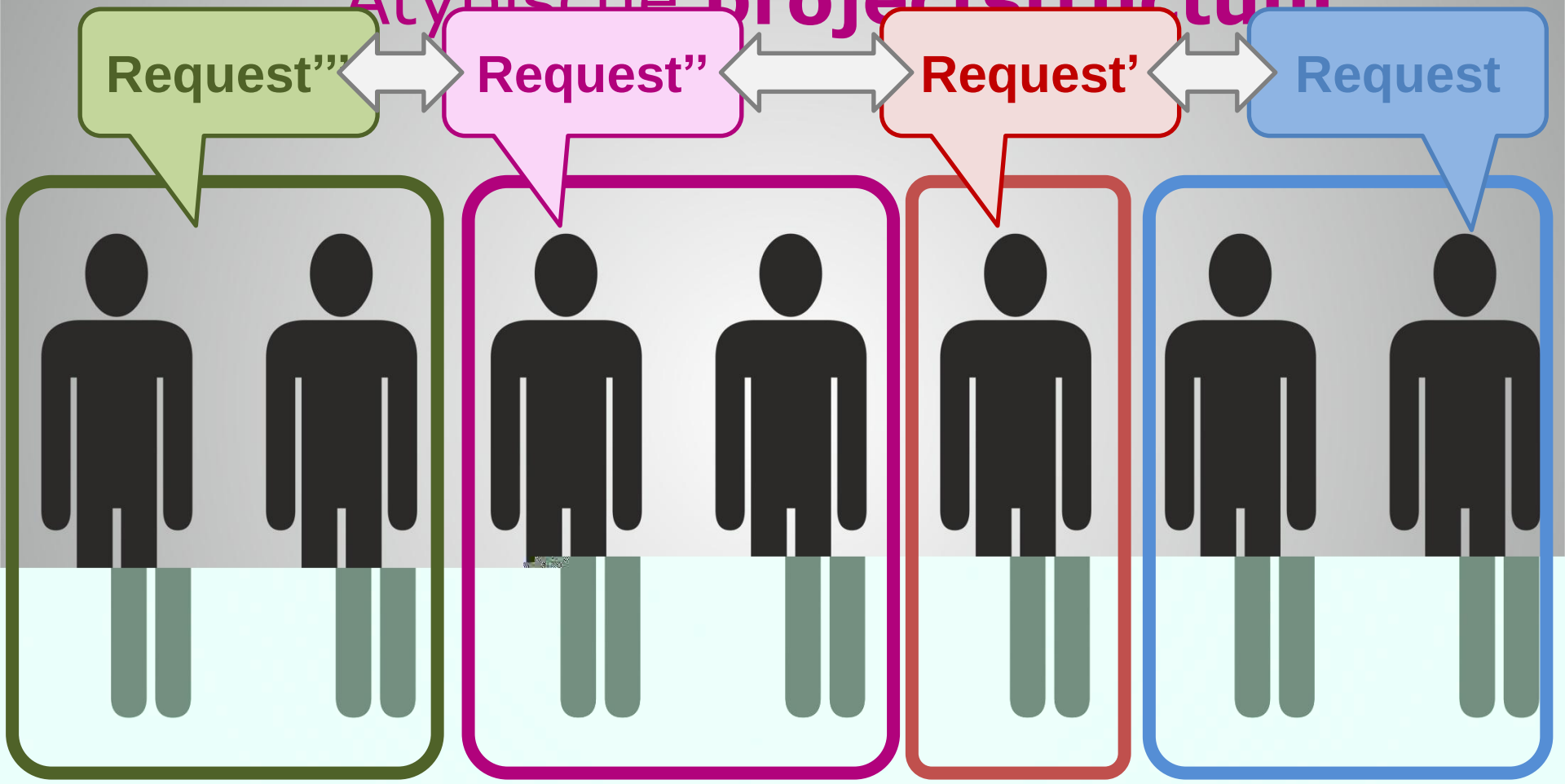
Barrière 2:
Dataquality management



Barrière 2: **Dataquality** management



Barrière 3:
Atypische projectstructuur



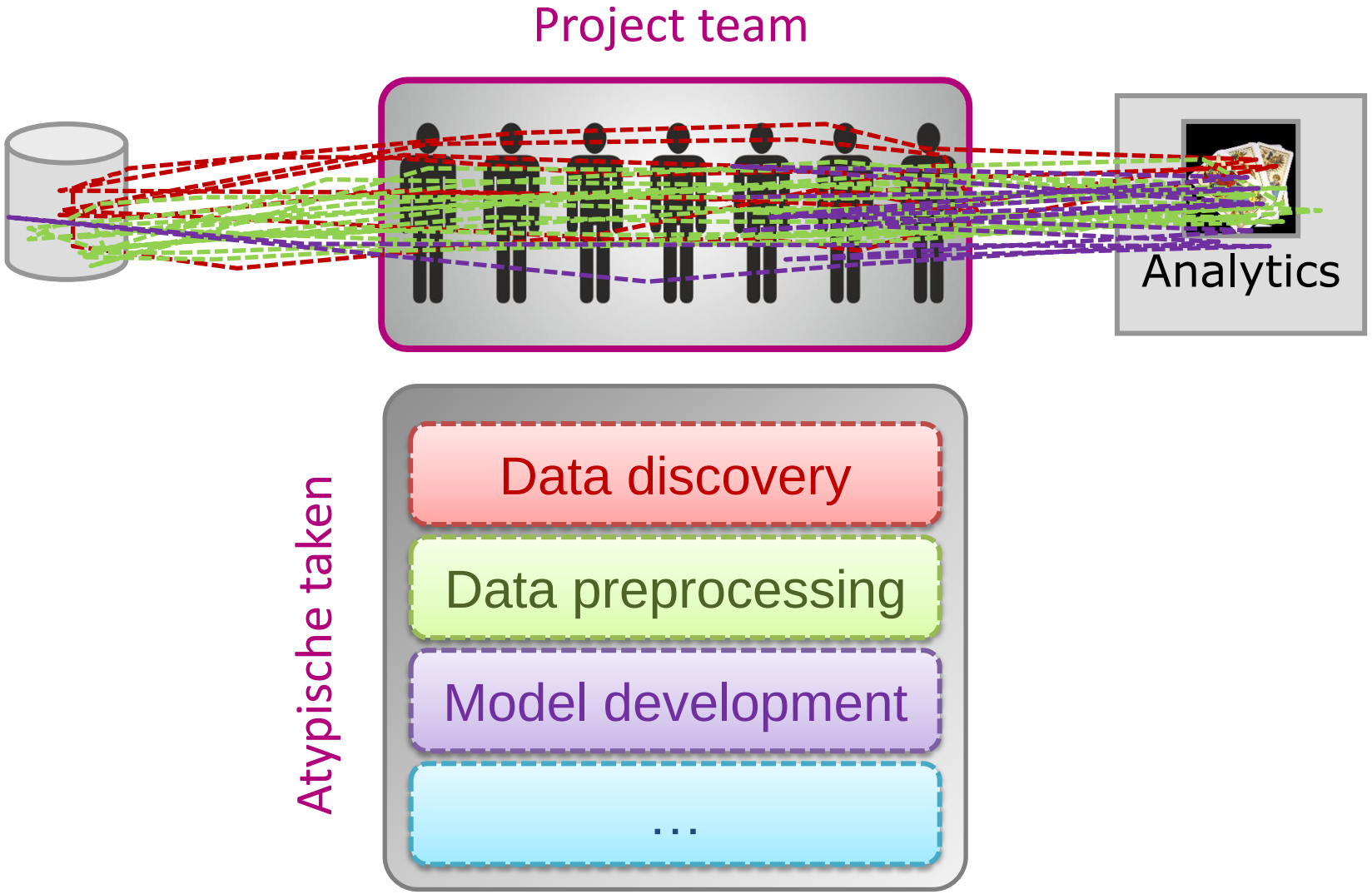
Operationele systemen

Data warehouse

Data-mart

Analytics

Barrière 3:
Atypische **projectstructuur**



Barrière 3:
Atypische **projectstructuur**

Atypische taken

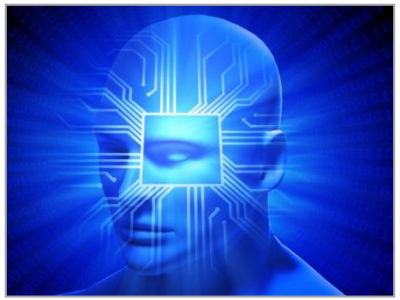
- Data discovery
- Data preprocessing
- Model development
- …

Smals
ICT for society

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?**

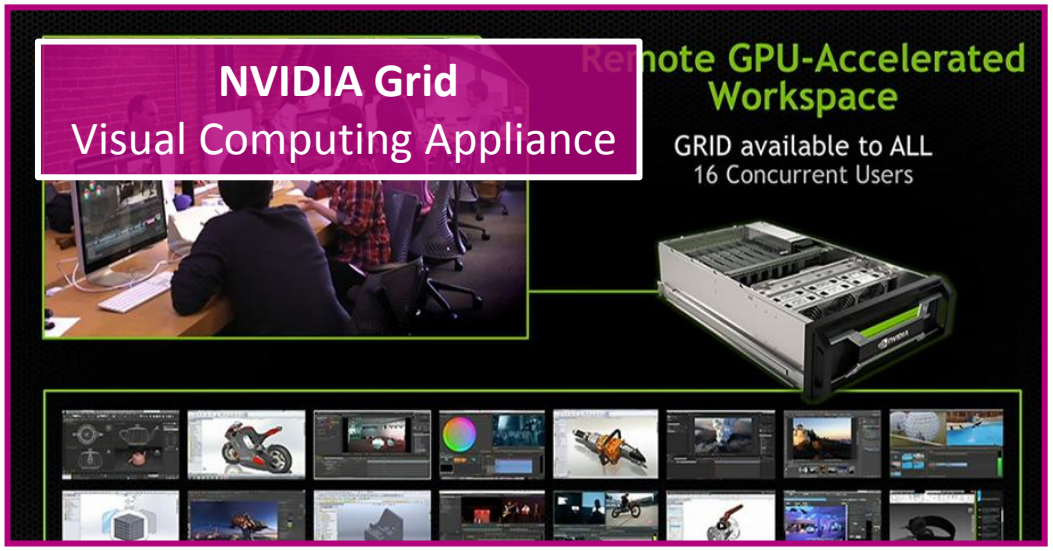
An appliance has **one function** (toasting, making coffee, …), is **easy to use** and **its internals are not relevant**

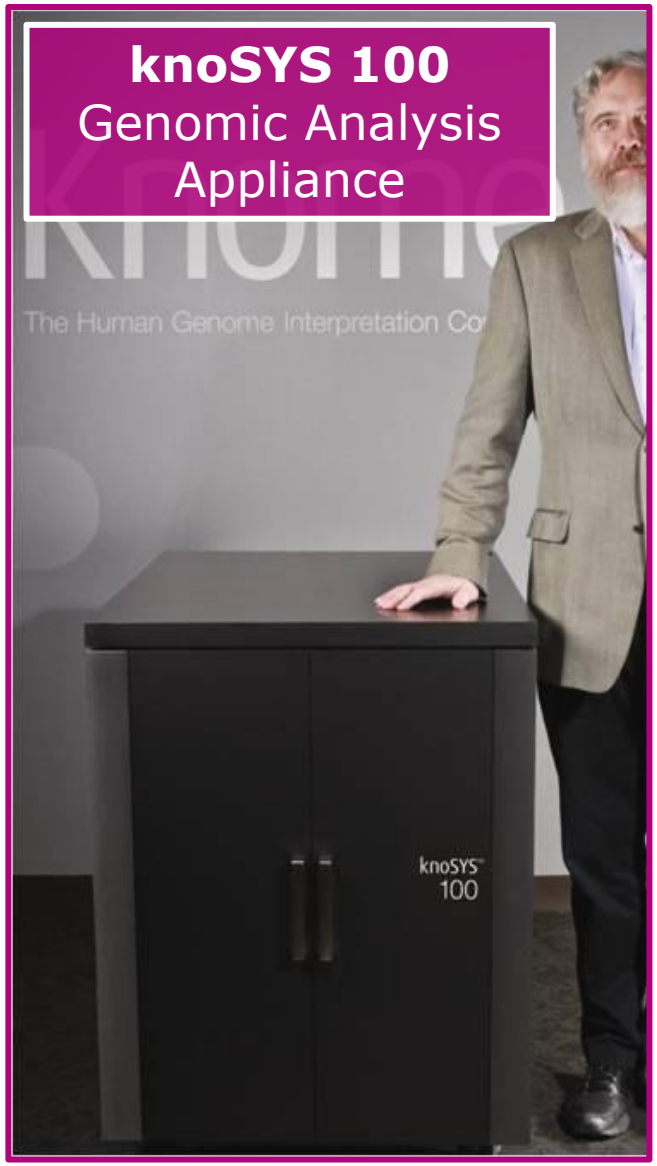
A **hardware appliance for analytics** has one function (executing complex queries), is easy to install, manage, tune, … and its internals are not relevant

[Rick van der Lans, http://www.b-eye-network.com/view/13506]

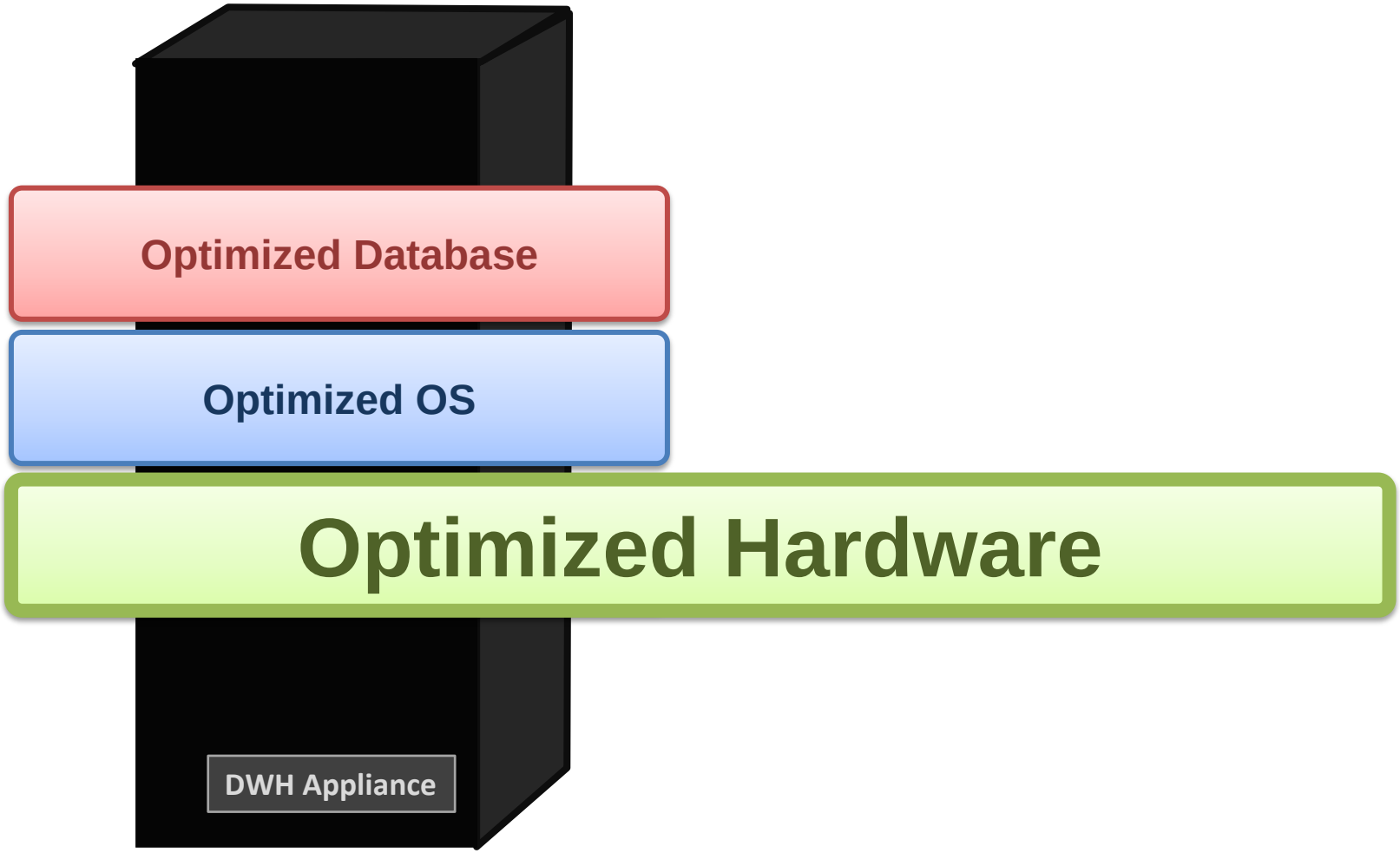
Smals
ICT for society

« Supercomputers voor analytics »

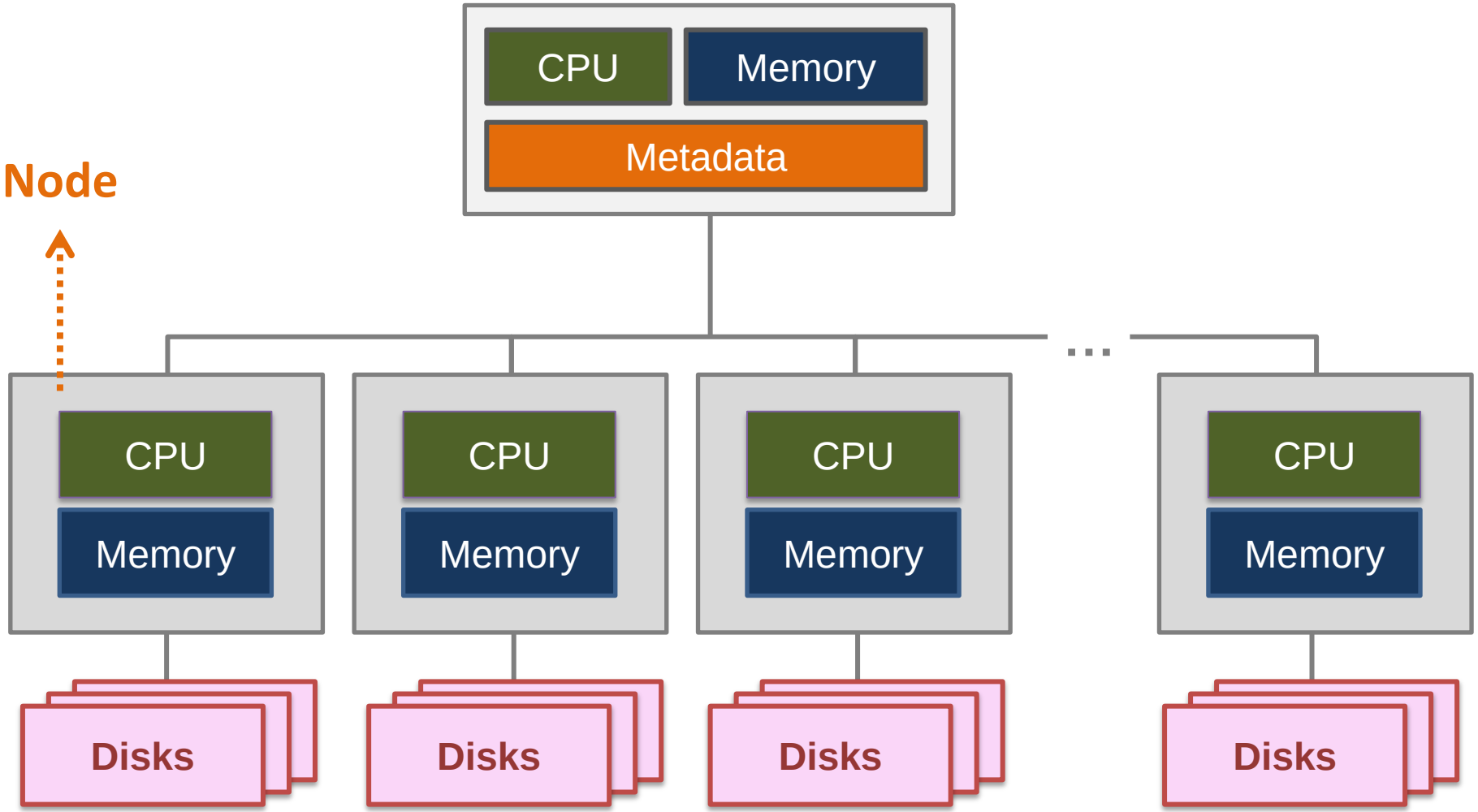




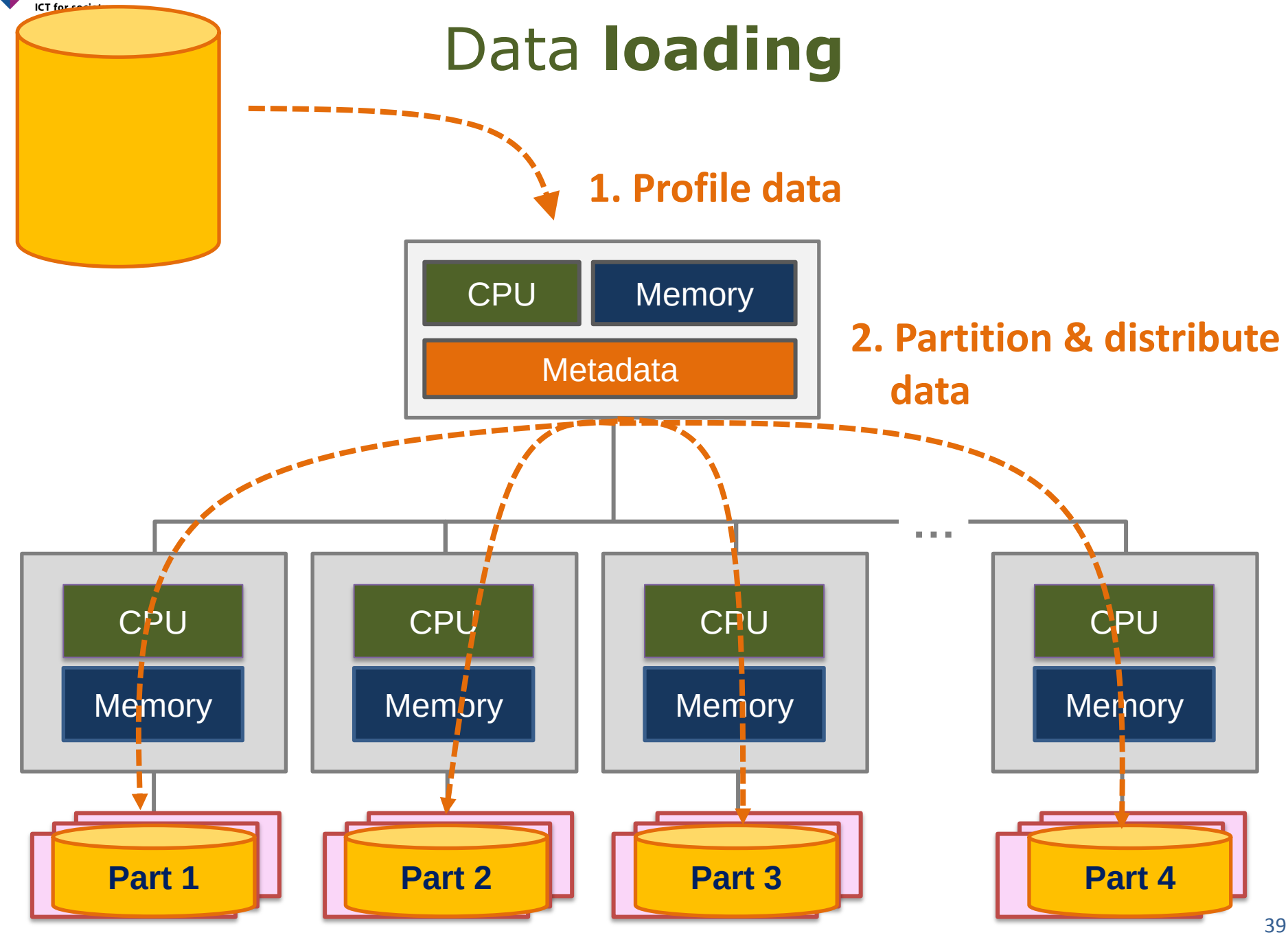
Hardware **appliance voor analytics**
(Data Warehouse appliance)



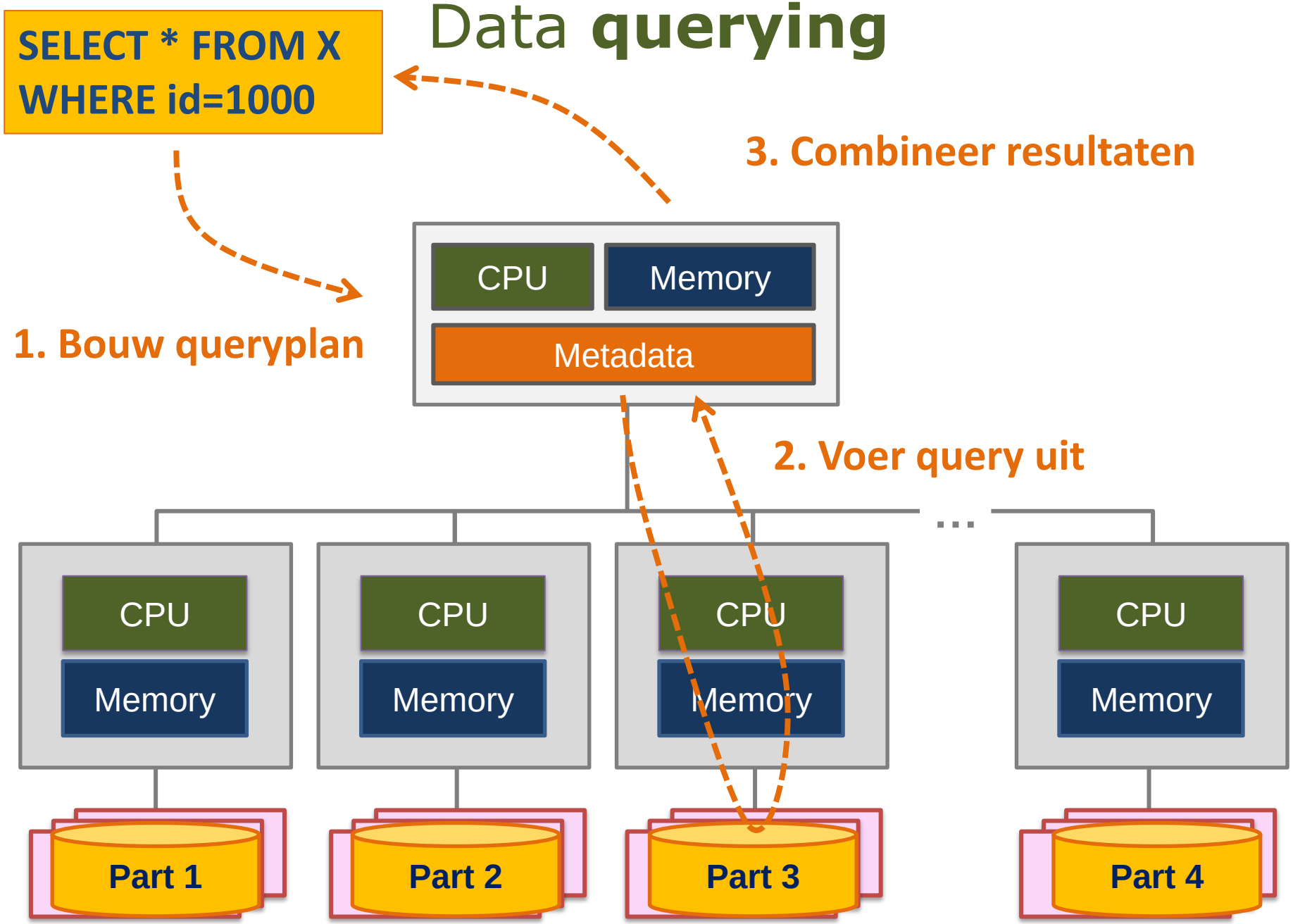
Massively Parallel Processing (MPP)



Data **loading**



Data **querying**



Data **querying**

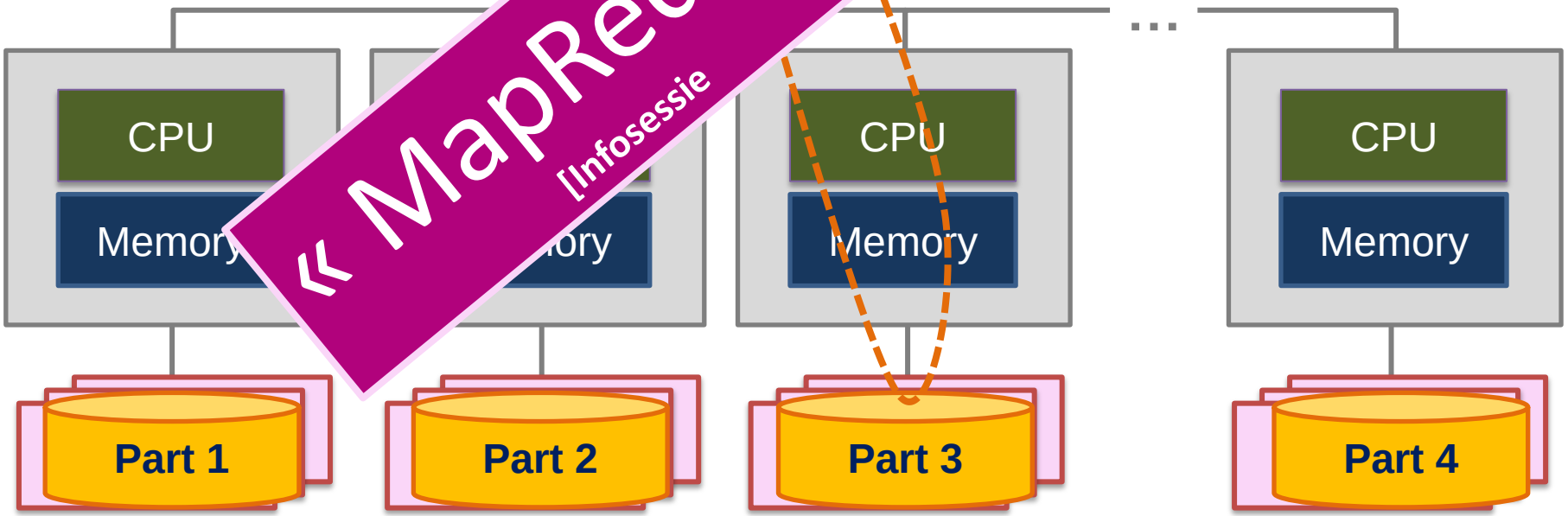
SELECT * FROM X
WHERE id=1000

3. Cor… …ltaten

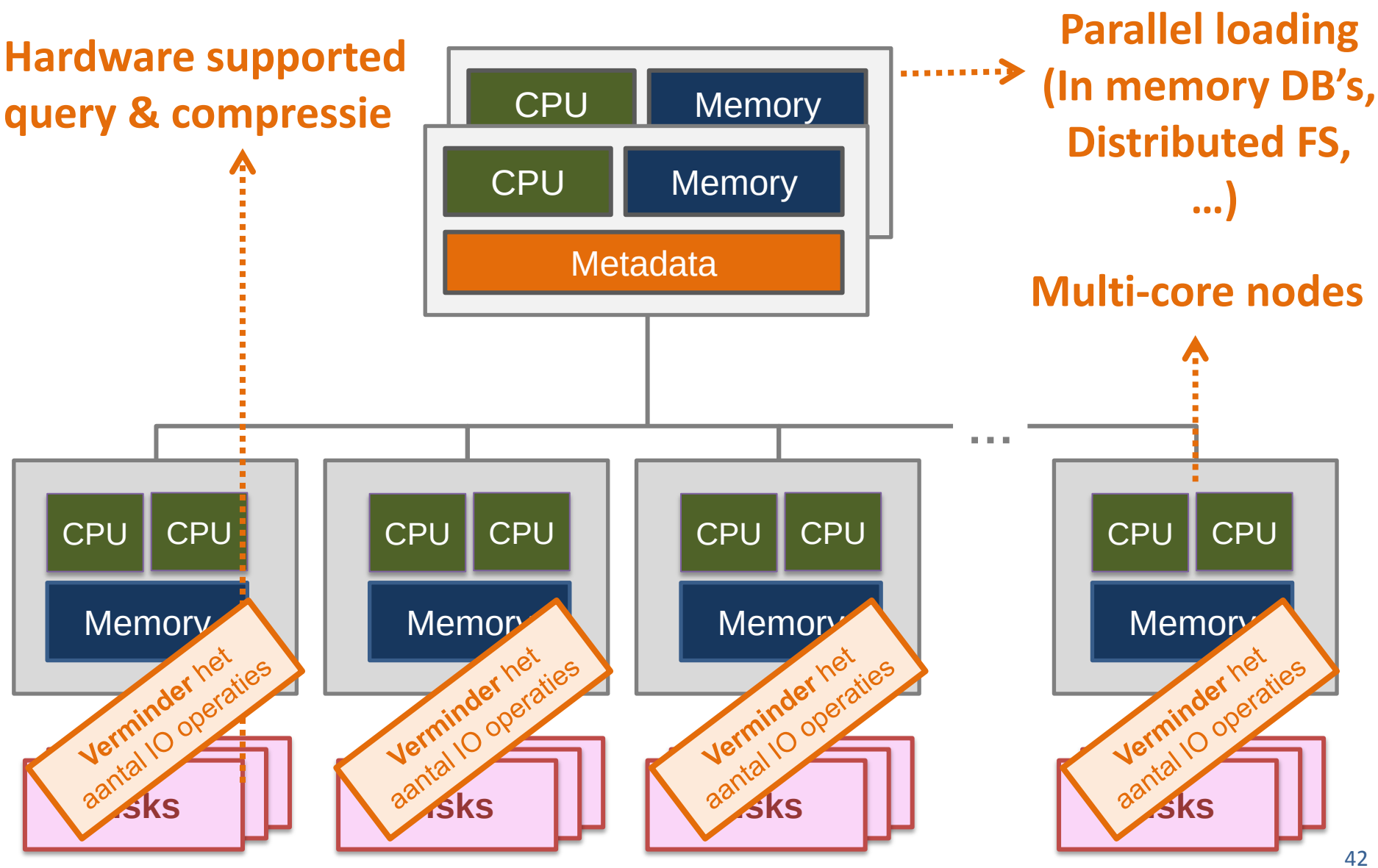
1. Bouw queryplan

« MapReduce in a box »
[Infosessie

2. Voer query uit



Extra **optimalisaties**

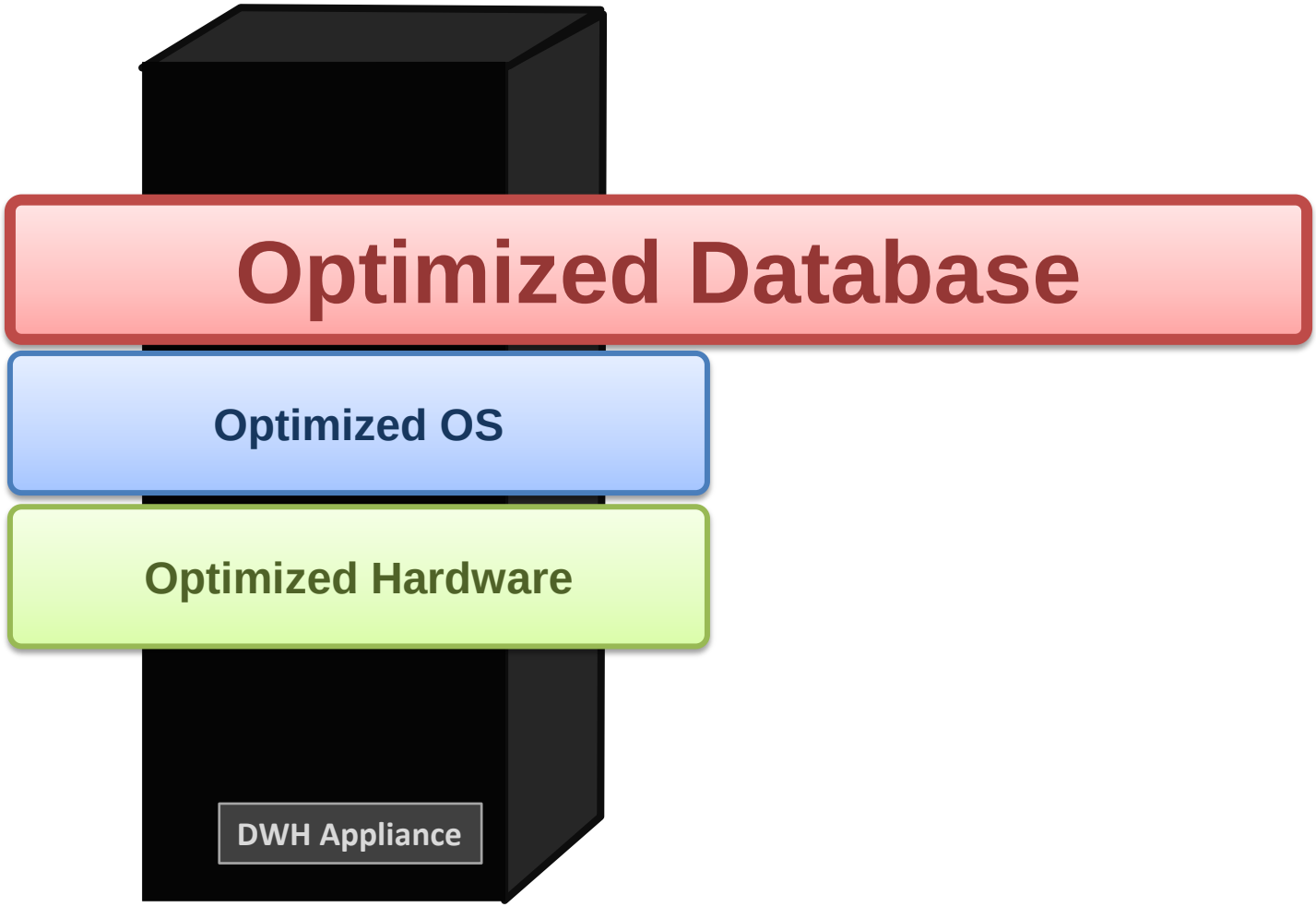


Data **querying** vs. data **loading**

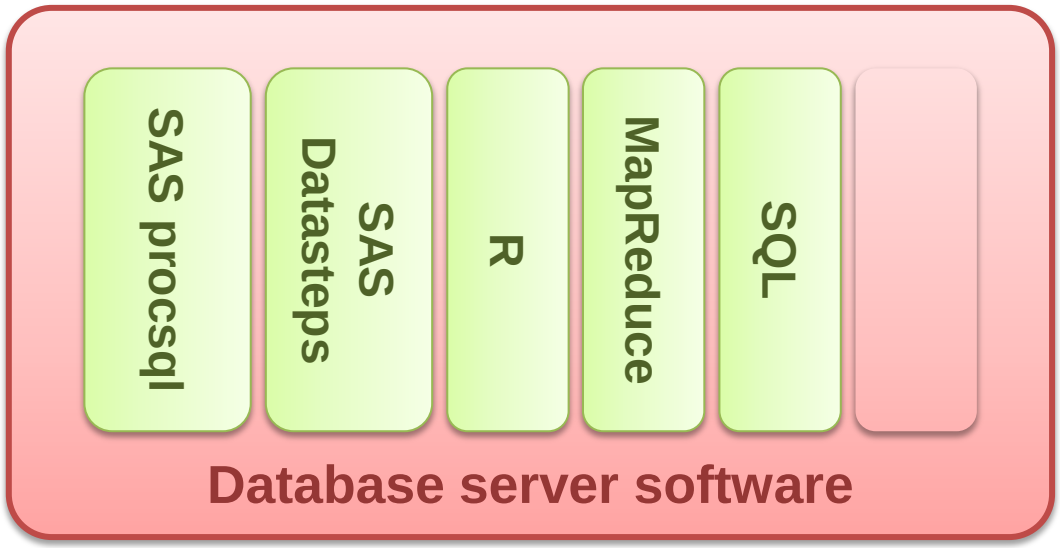
- Traag
- Veel processing & scanning
- Optimalisaties mogelijk

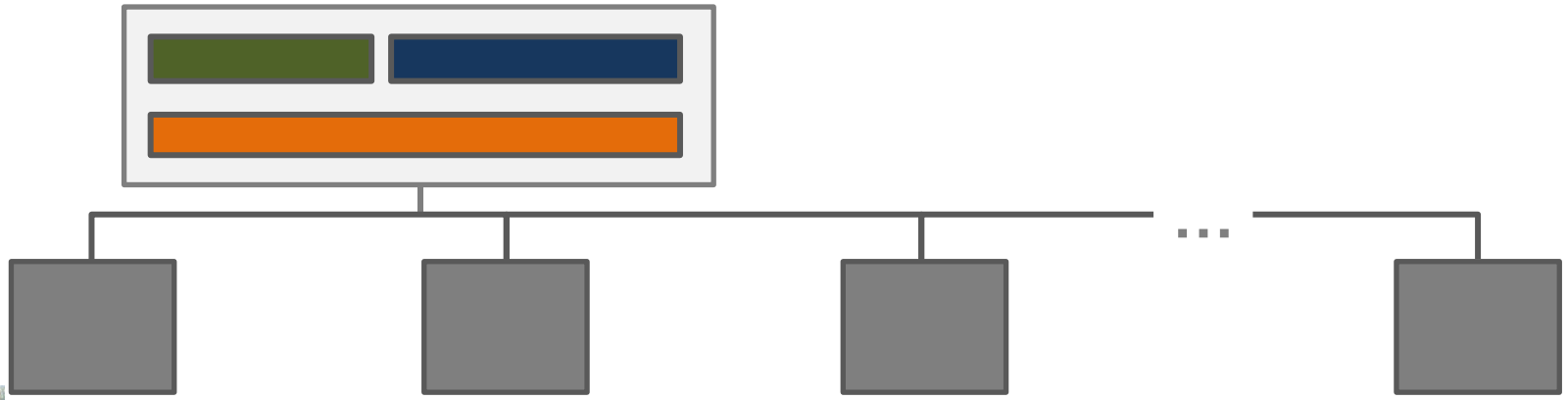
- Performant
- Benut MPP optimaal
- **Maar**: afhankelijk van hoe dataloading gebeurd is!

Hardware **appliance voor analytics**
(Data Warehouse appliance)

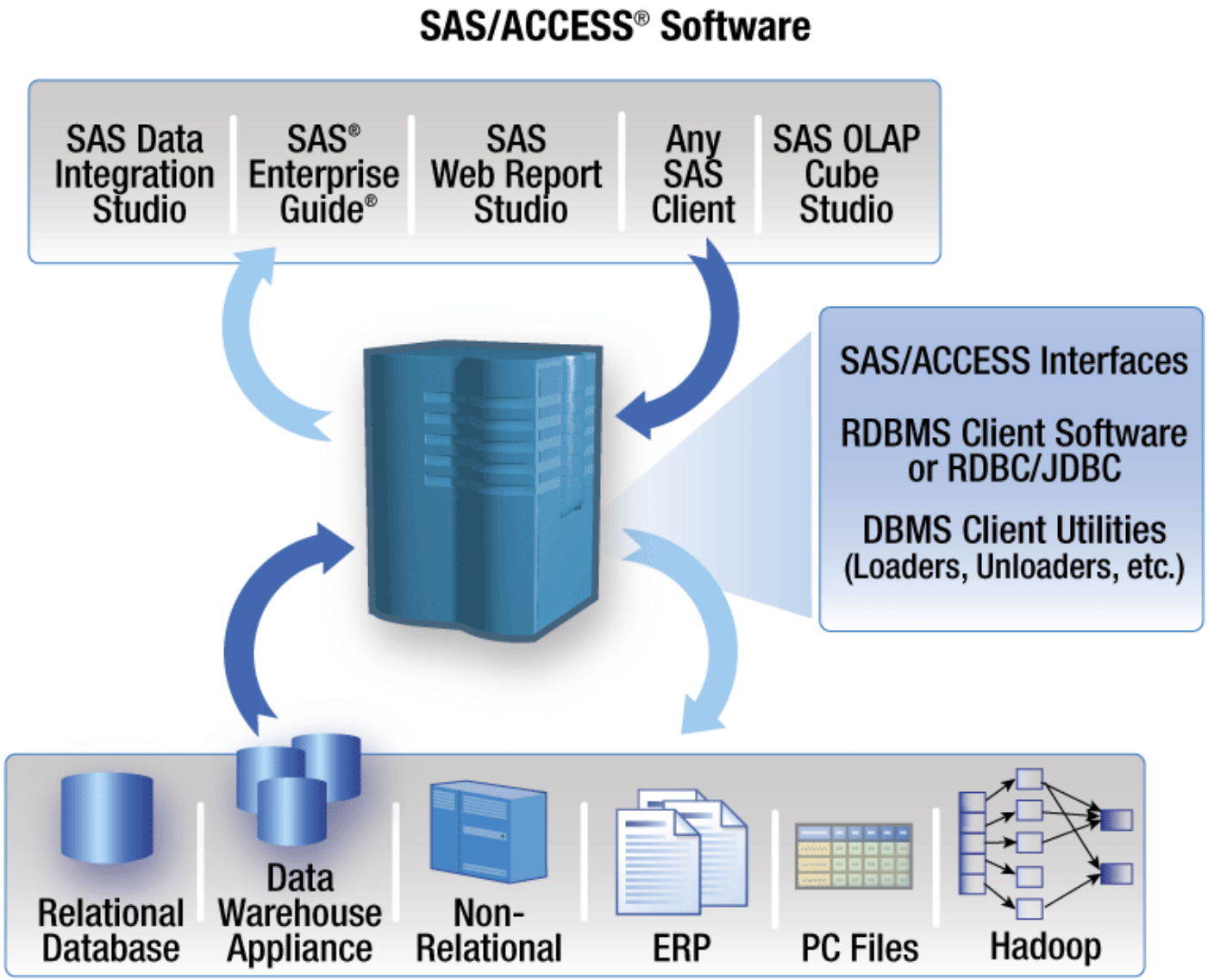


In-database analytics

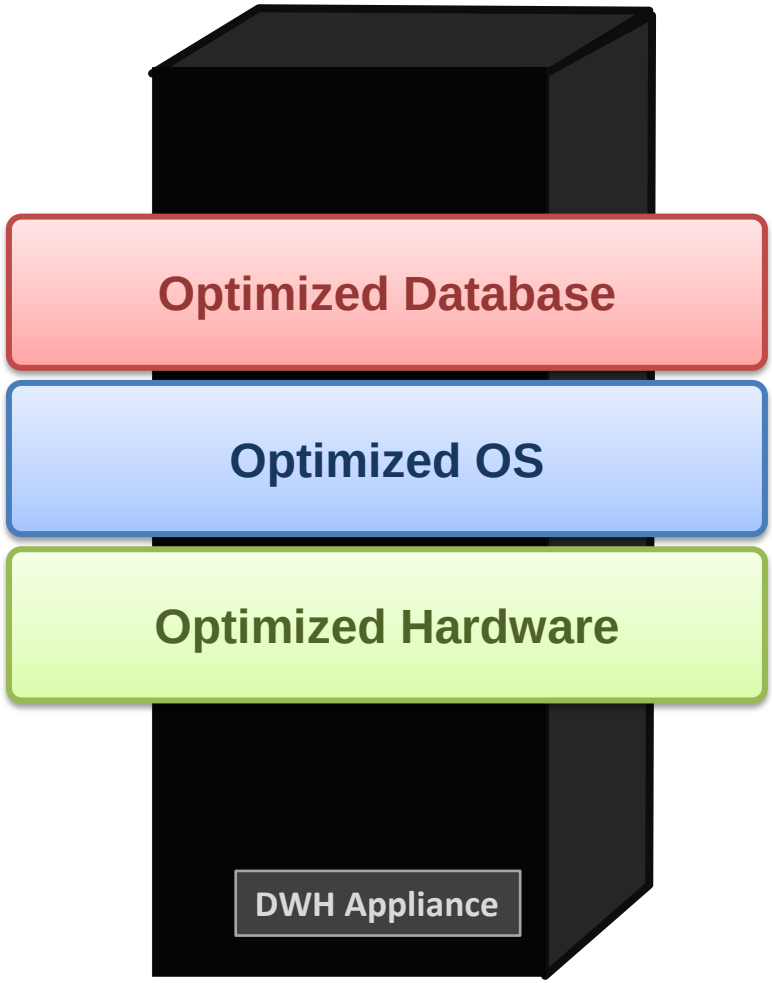




SAS connectoren



DWH Appliances samengevat



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

47

Marktoverzicht

EMC²

GREENPLUM

TERADATA. ASTER

NETEZZA
an IBM® Company

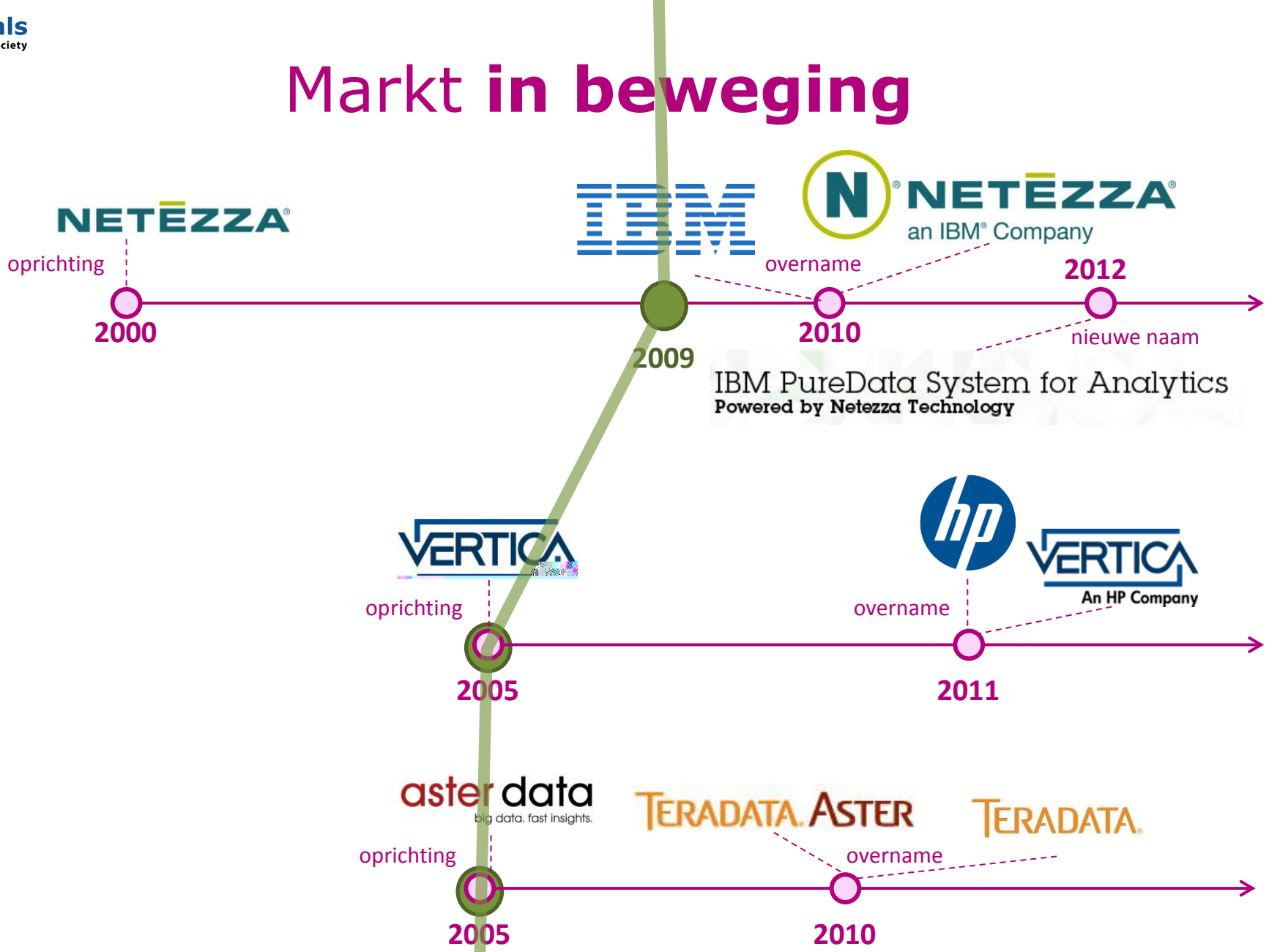
ORACLE
EXADATA
READY

SAP HANA

VERTICA
An HP Company

Microsoft®
SQL Server® 2012
Parallel Data Warehouse

Markt **in beweging**



Proprietary Hardware | **Commodity Hardware**

Markt **in beweging**

Sun microsystems

amazon

oprichting — partnership — overname — lancering

2005 — 2011 — 2013

DATAllegro
DATA AT THE SPEED OF BUSINESS

Microsoft

oprichting — partnership — overname

2003 — 2004 — 2008

DELL

EMC²

Microsoft SQL Server 2012 Parallel Data Warehouse

Greenplum

EMC² GREENPLUM

oprichting — 2006 partnership — 2010 overname

2003

Commodity Hardware

Smals
ICT for society

Generic

Meer info over dit kwadrant: contacteer Smals Onderzoek

Software-stack

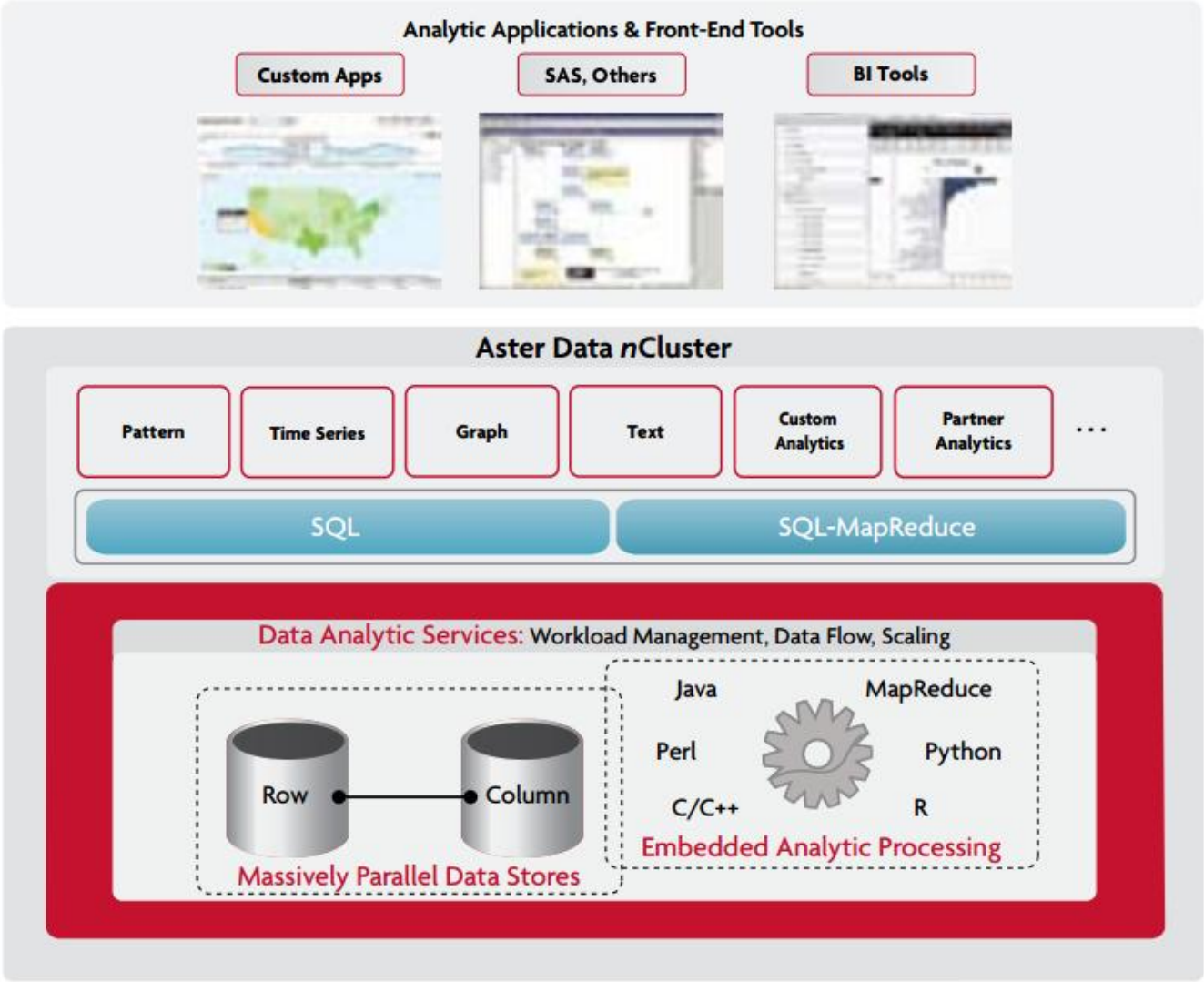
Accelerator

Architectuur

MPP

Specific

TERADATA. ASTER *n*Cluster

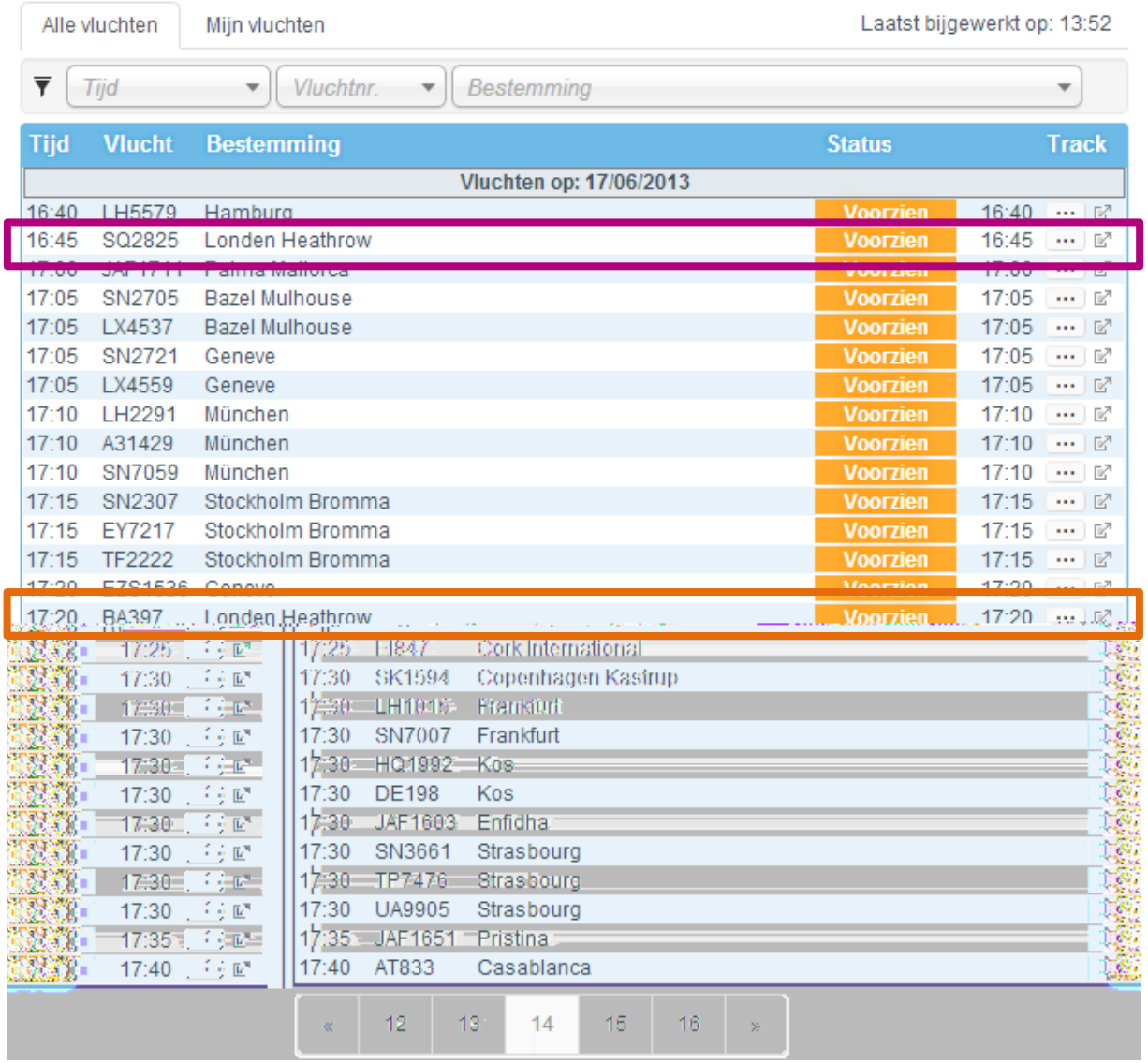


Get all the flights to London for which another flight exists to London that leaves within an hour on the same day

16:45

< 60 min

17:20



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

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)

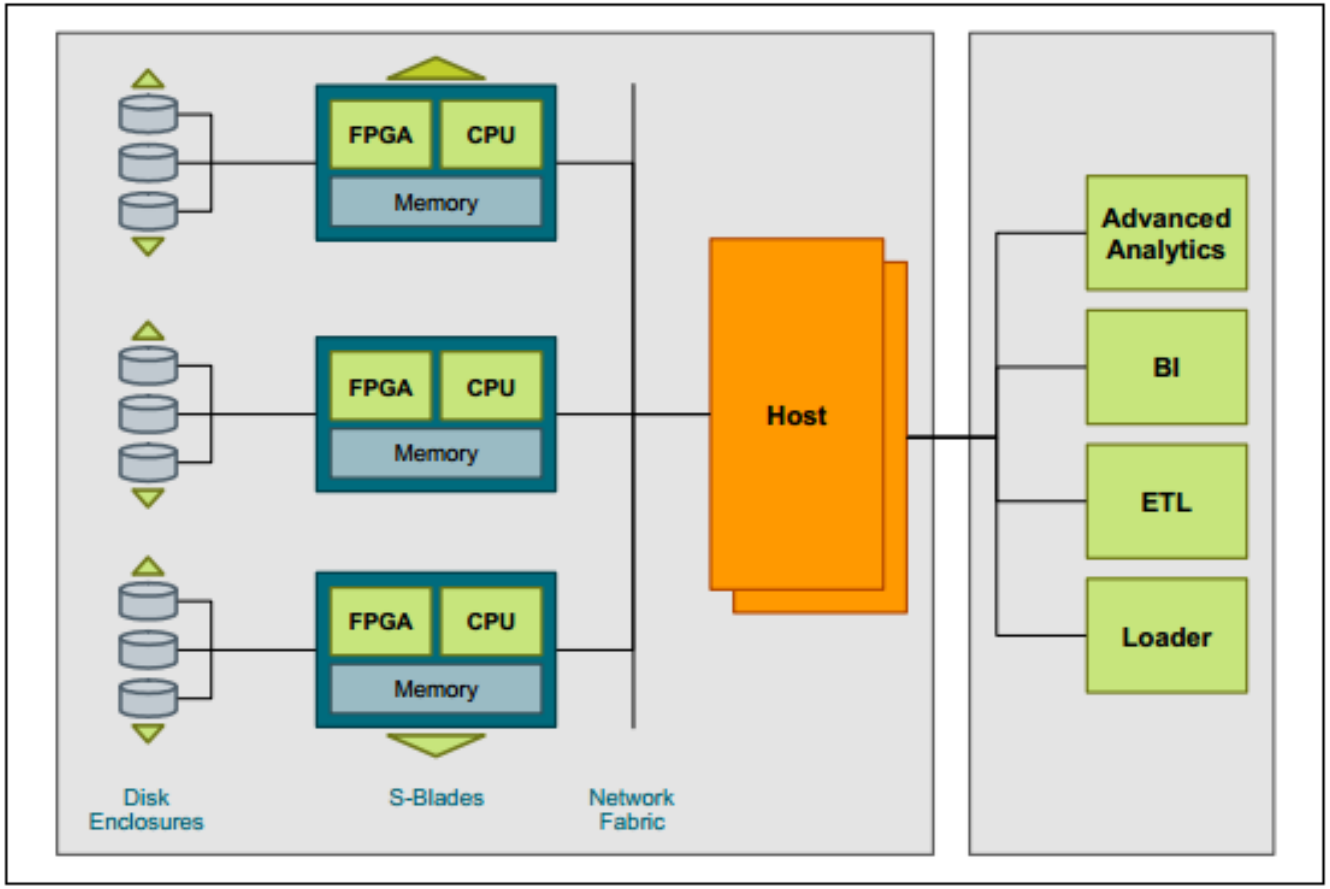
```
ORDER BY DEP_TIME
```

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

Smals
ICT for society

NETEZZA®
an IBM® Company

IBM® PureData™ for Analytics, powered by Netezza technology

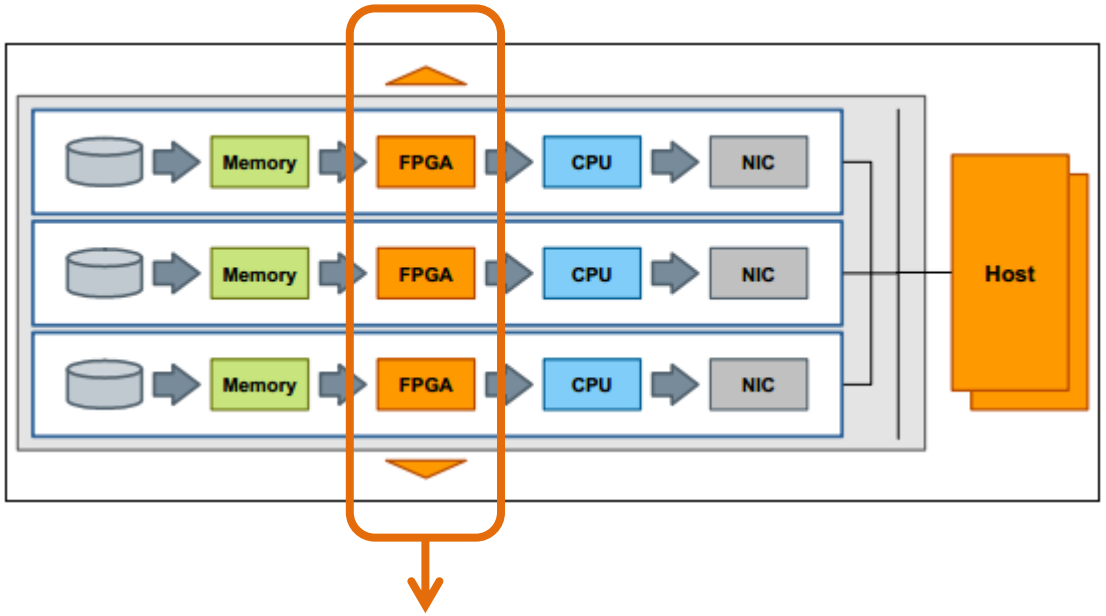


Smals
ICT for society

NETEZZA
an IBM Company

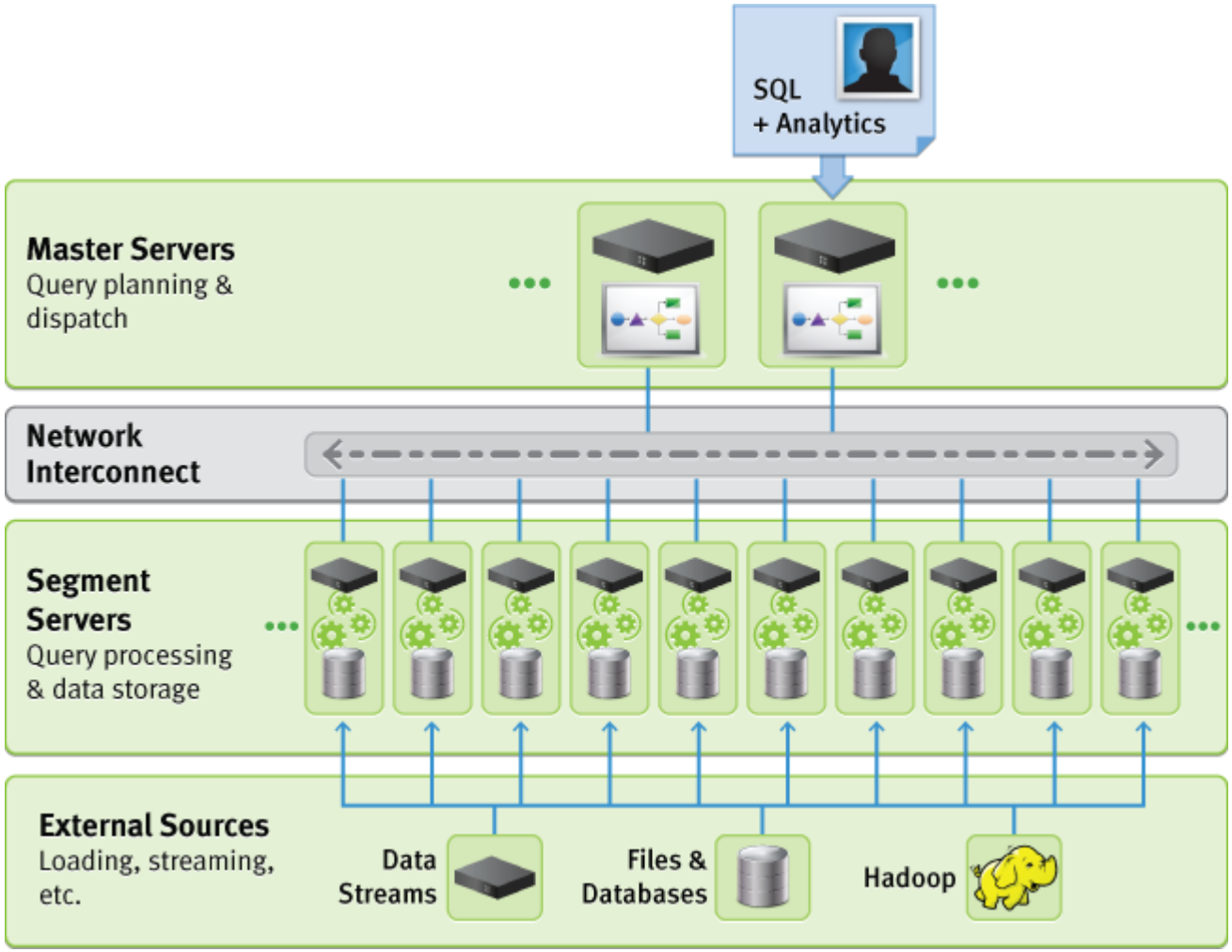
IBM® PureData™ for Analytics, powered by Netezza technology

In de worker nodes:

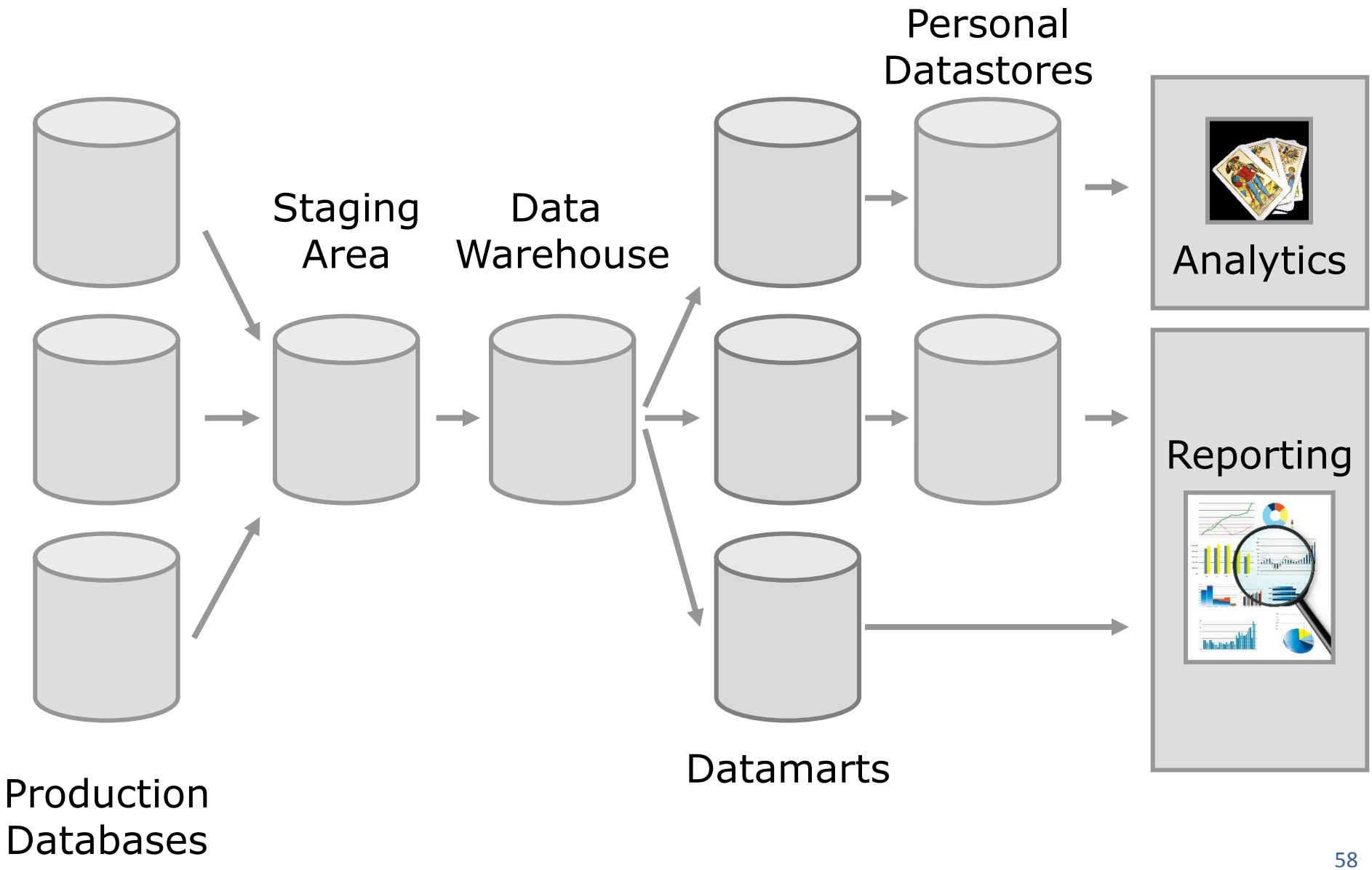


Zeer snelle decompressie en windowing

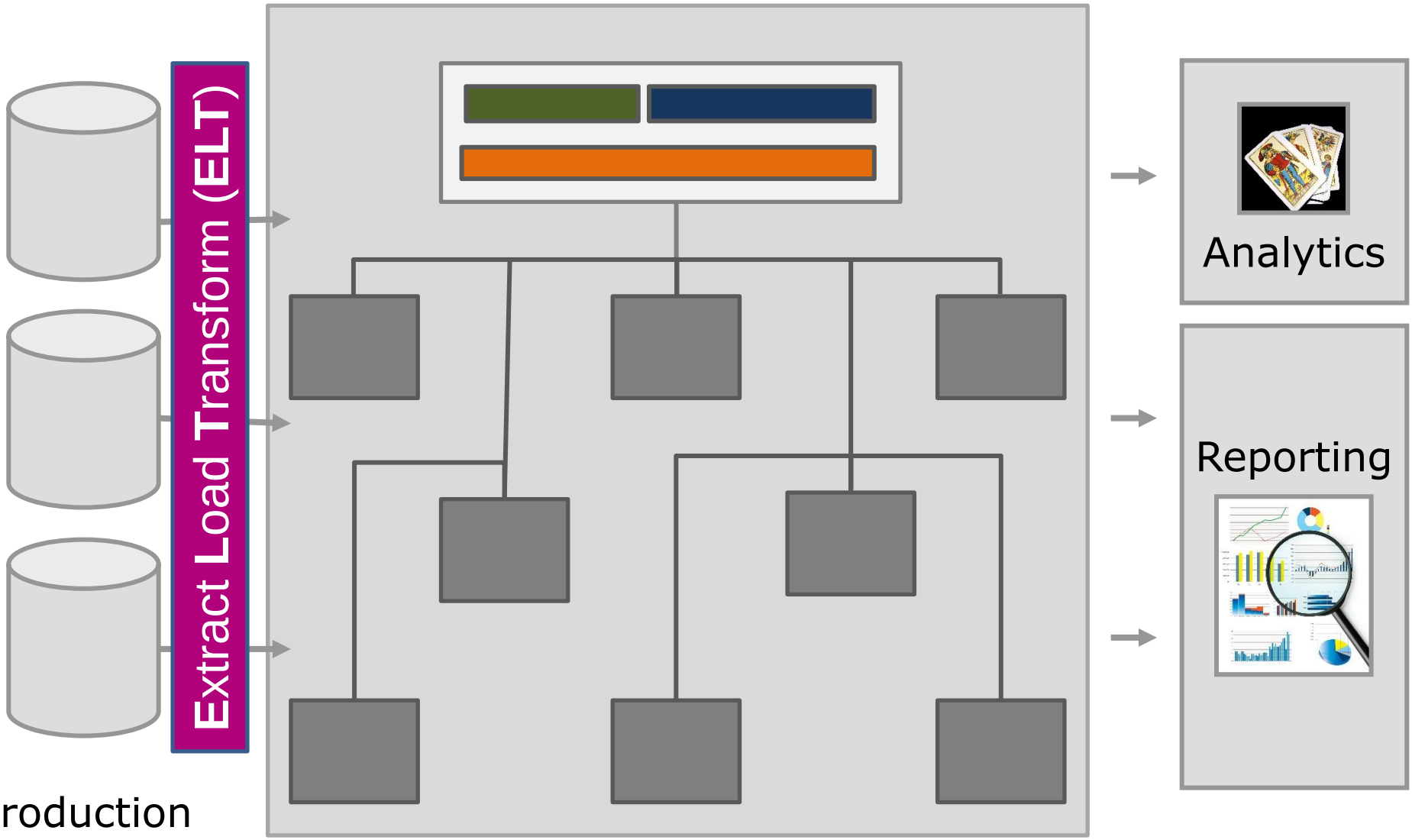
EMC²
GREENPLUM



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)

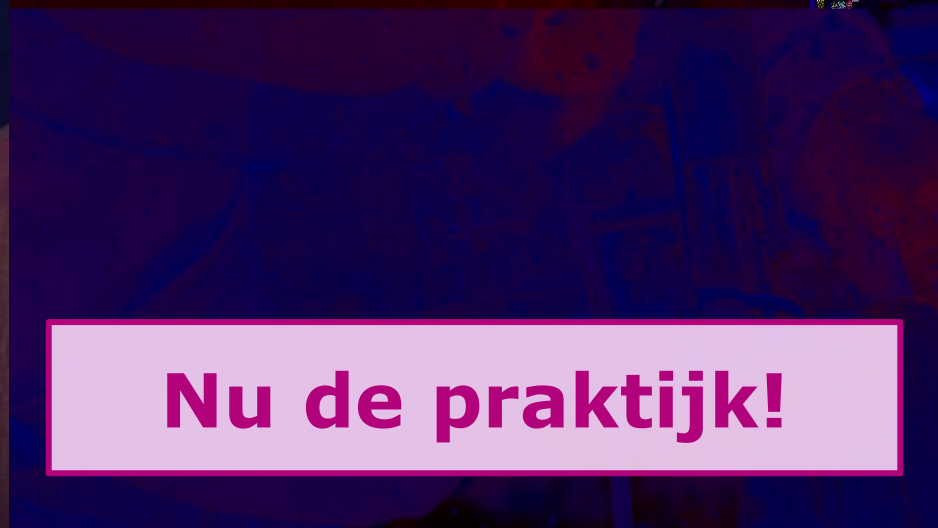
Barrière 3:

Atypische projectstructuur

Een stap in de goeie richting:

Analytics-experten hebben makkelijker toegang tot **alle** data (met respect voor security/privacy)

Dit was de theorie…



Nu de praktijk!

Meer info over deze POC: contacteer Smals Onderzoek

POC met IBM Netezza

Load

Query

- SAS-script 1
- SAS-script 2
- R-analyse 1
- R-analyse 2

Geanonimiseerde Data

(~ 40 GB)

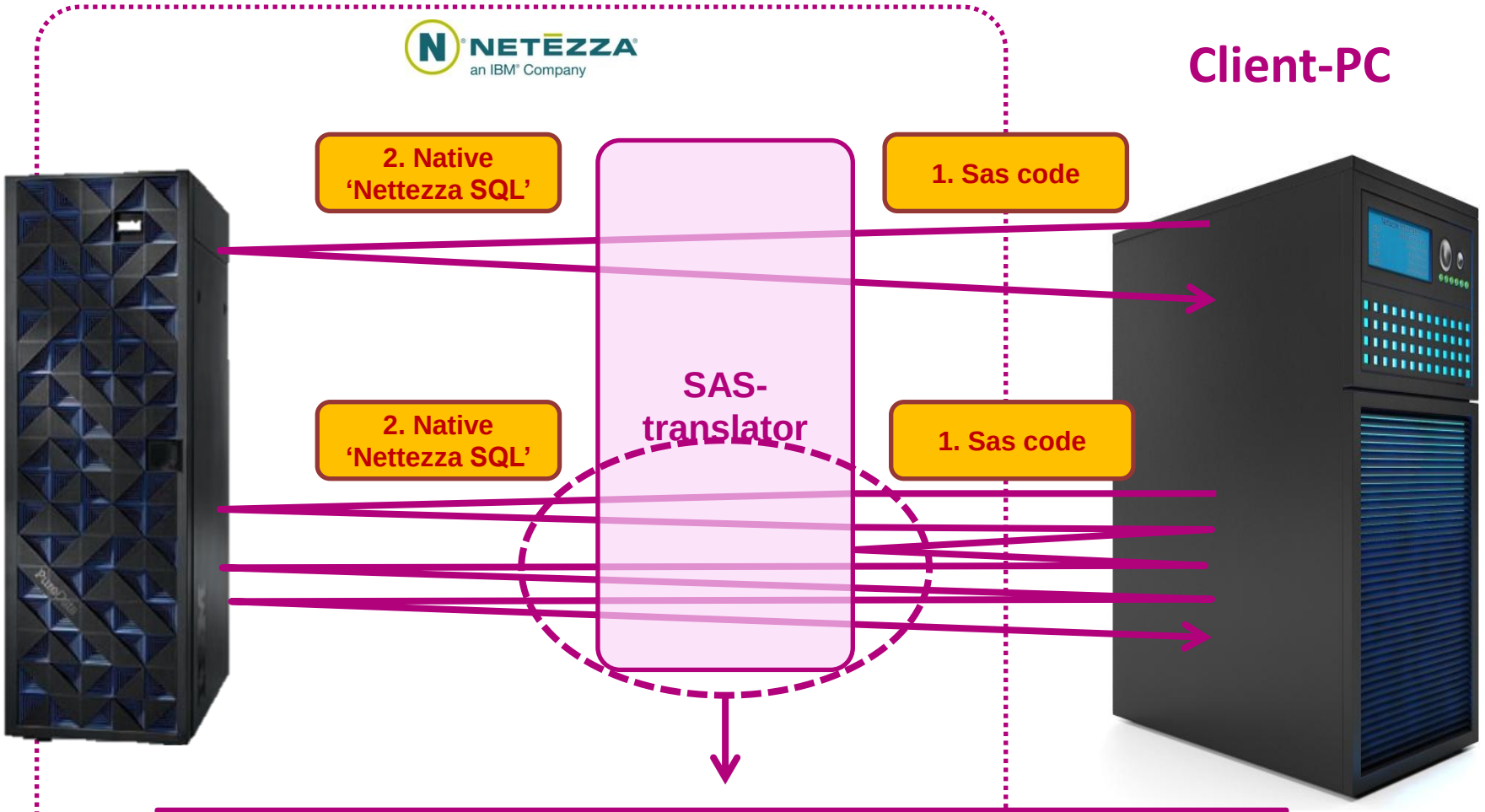
Sas & CSV-files

Netezza

IBM, Evere

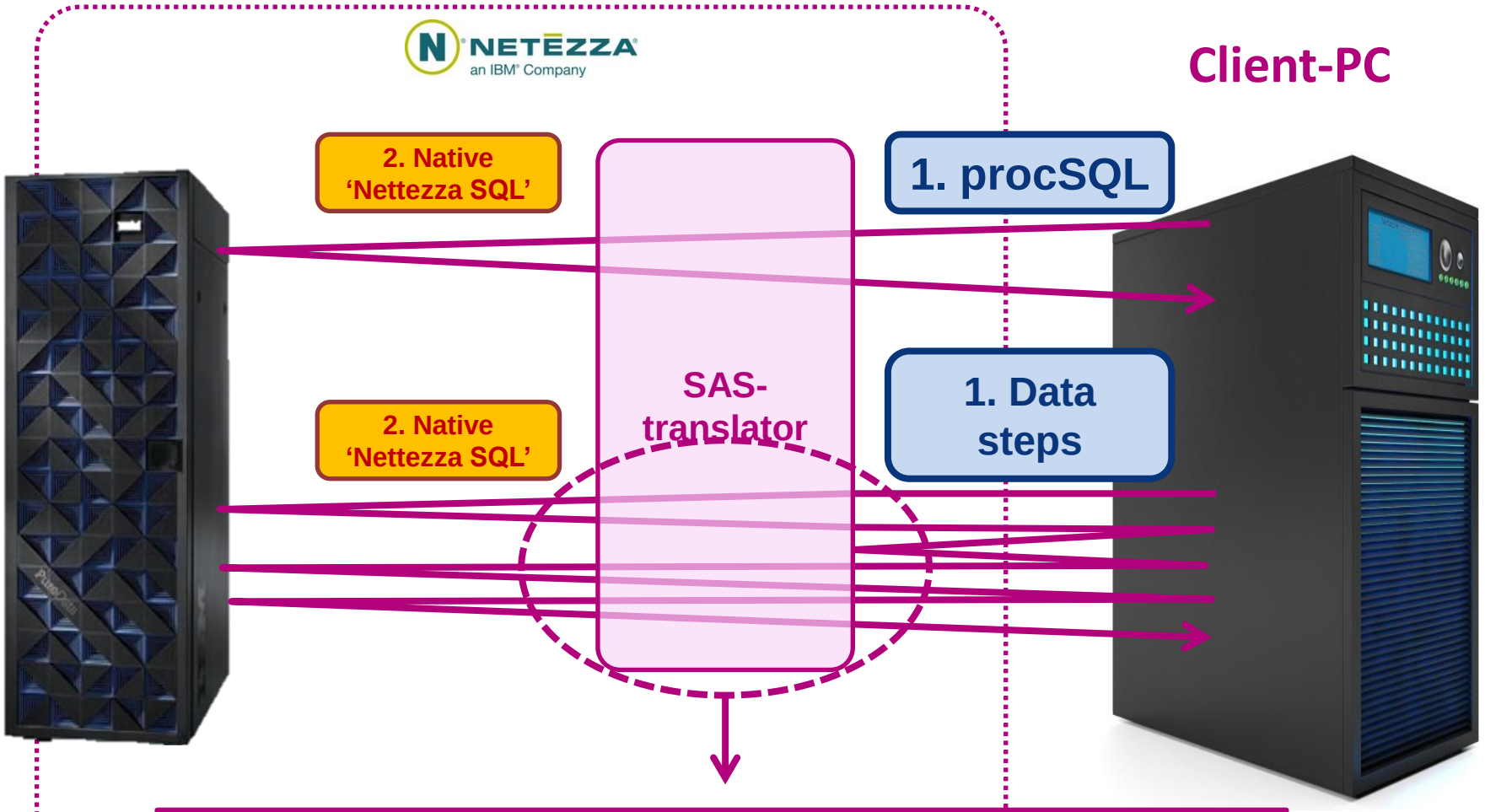
Test-scripts

SAS **code aanpassingen**



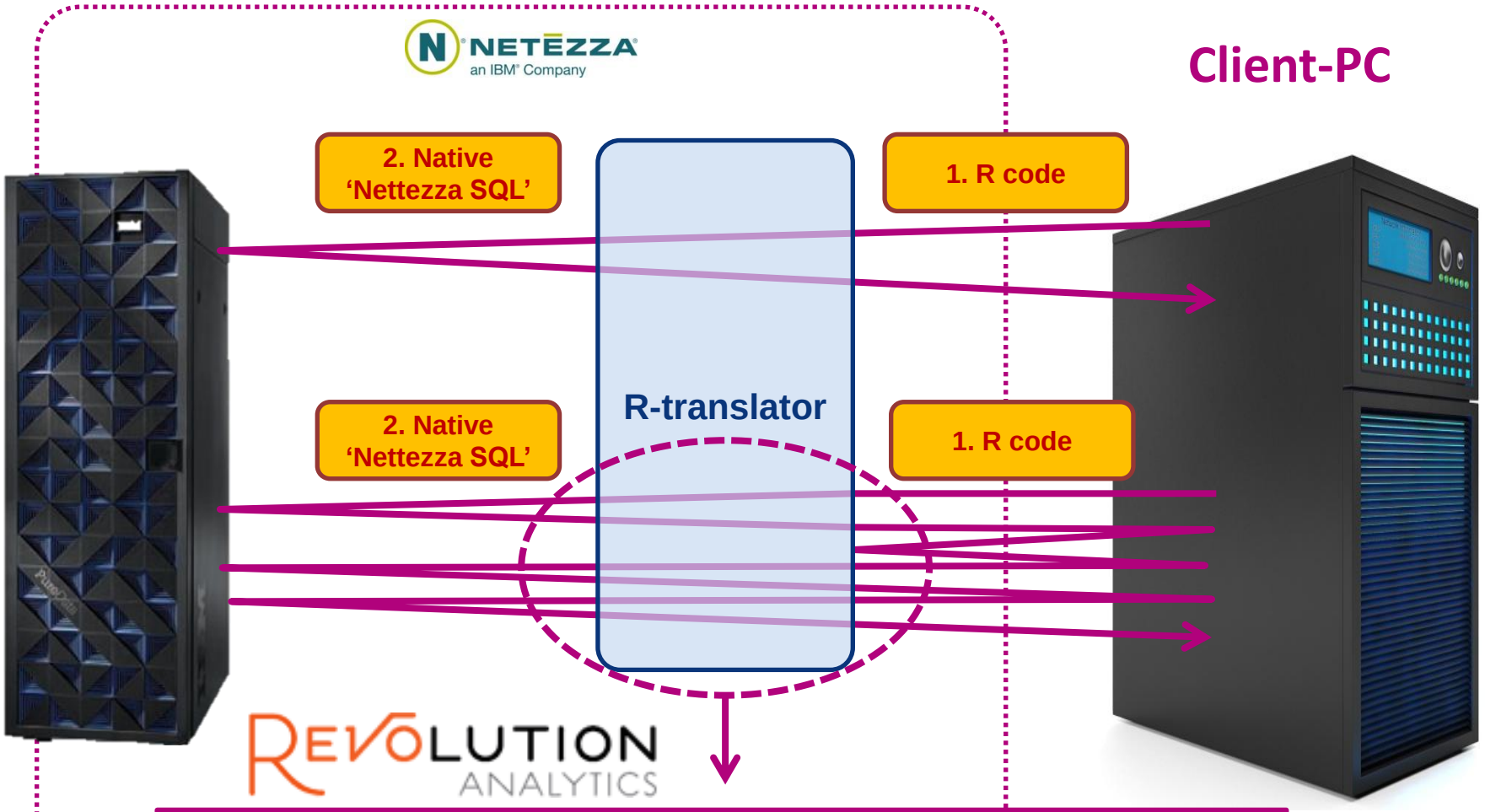
Onaangepaste code kan **niet volledig vertaald** worden :
Berekeningen worden gespreid over Client-PC & Netezza
(= veel performantieverlies)

SAS **code aanpassingen**



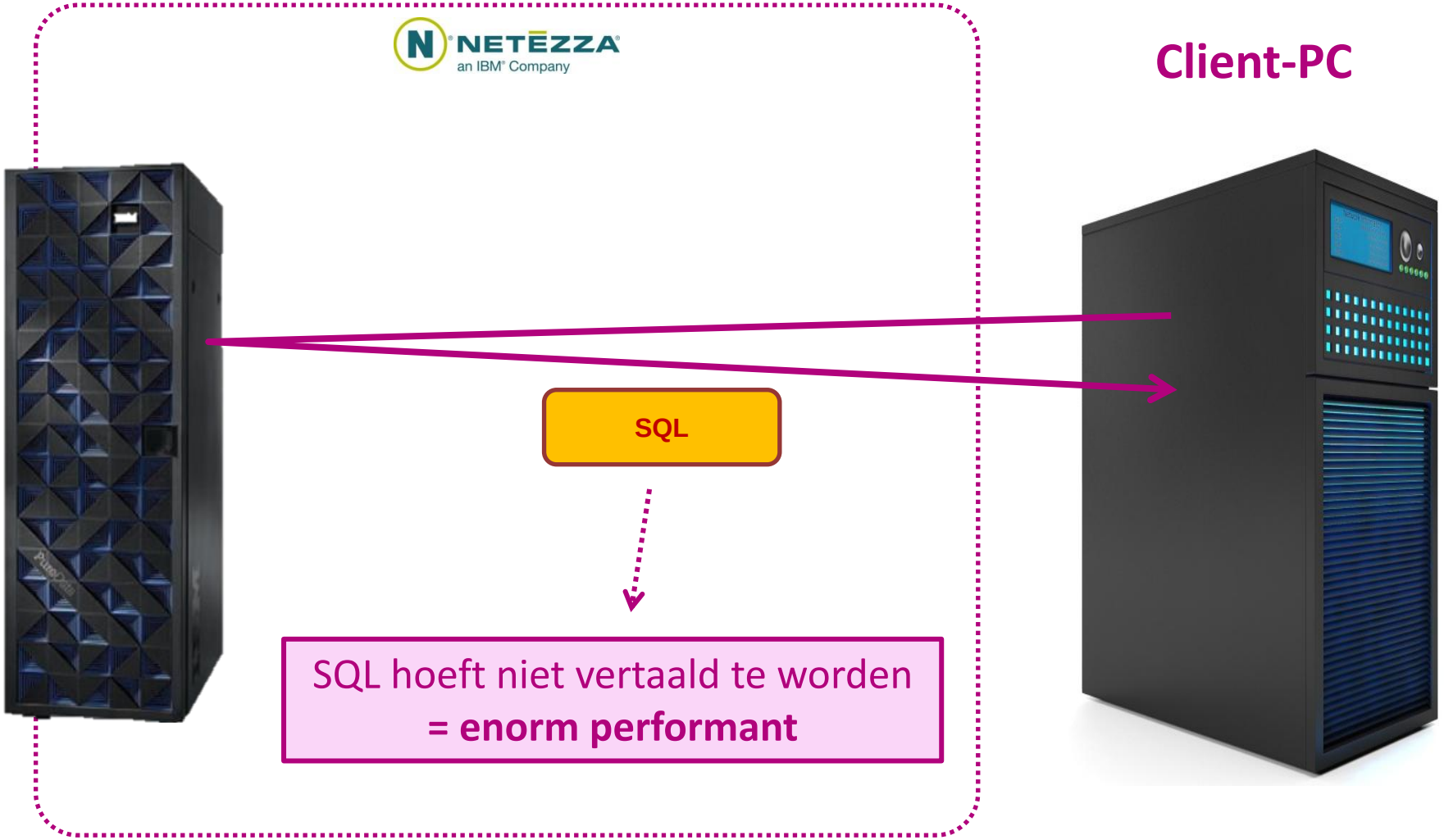
Onaangepaste code kan **niet volledig vertaald** worden :
Berekeningen worden gespreid over Client-PC & Netezza
(= veel performantieverlies)

R code aanpassingen



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

SQL is **native** ondersteund

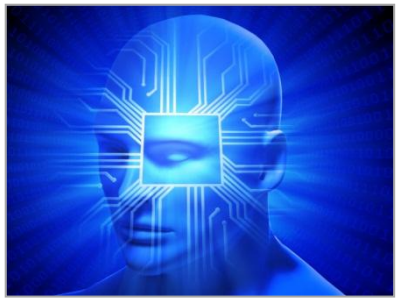


Smals
ICT for society

Streamlining Analytics

Predictive analytics

De **data supply chain**



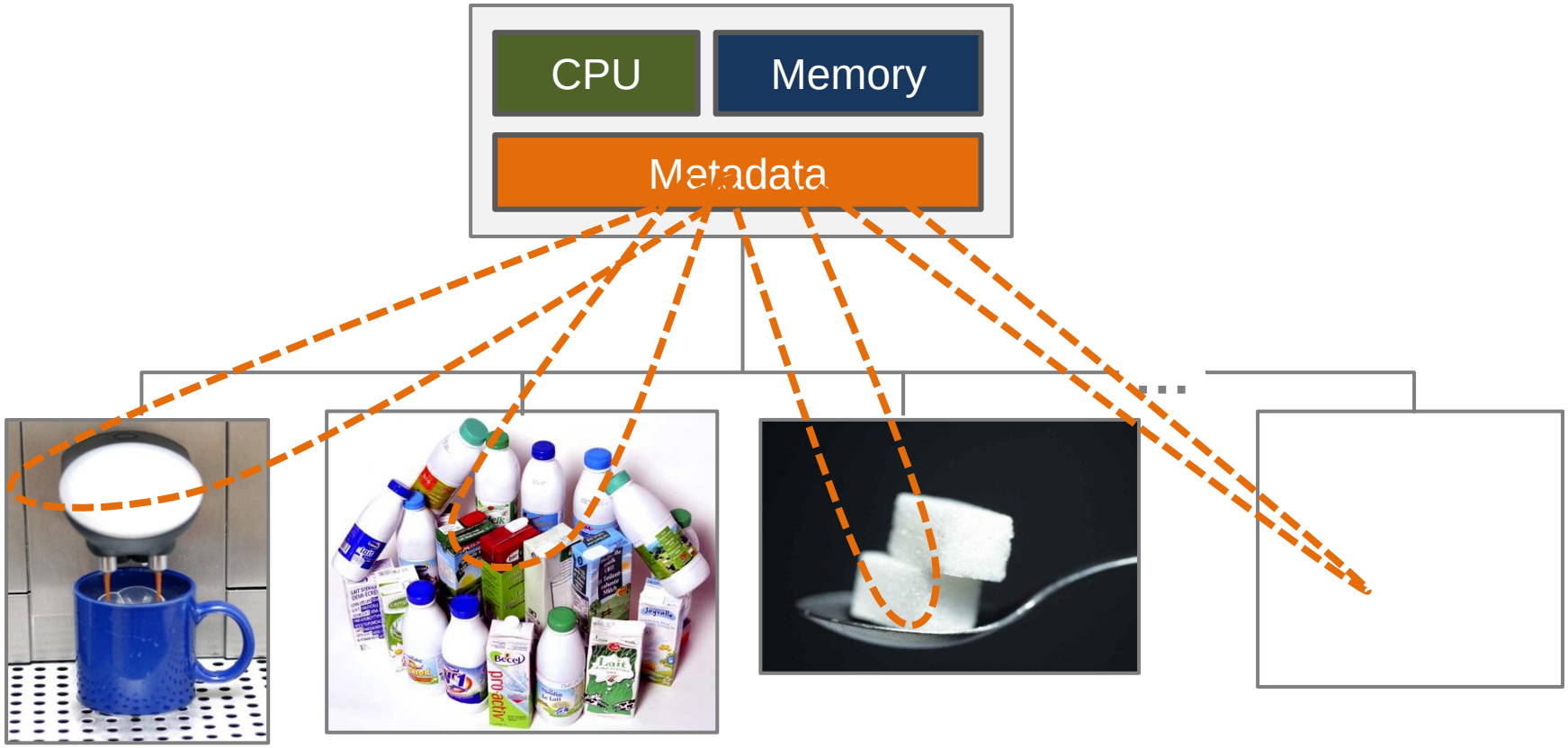
Barrières bij de introductie van analytics

Hardware appliances voor analytics

Data quality

Analytics **project management**

```
SELECT coffee, milk, sugar, cookies
FROM regular_break
INNER JOIN infosession_break ON rId = iId
```

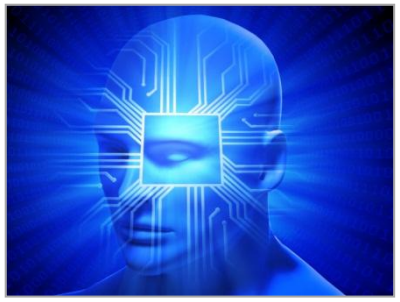


Smals
ICT for society

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

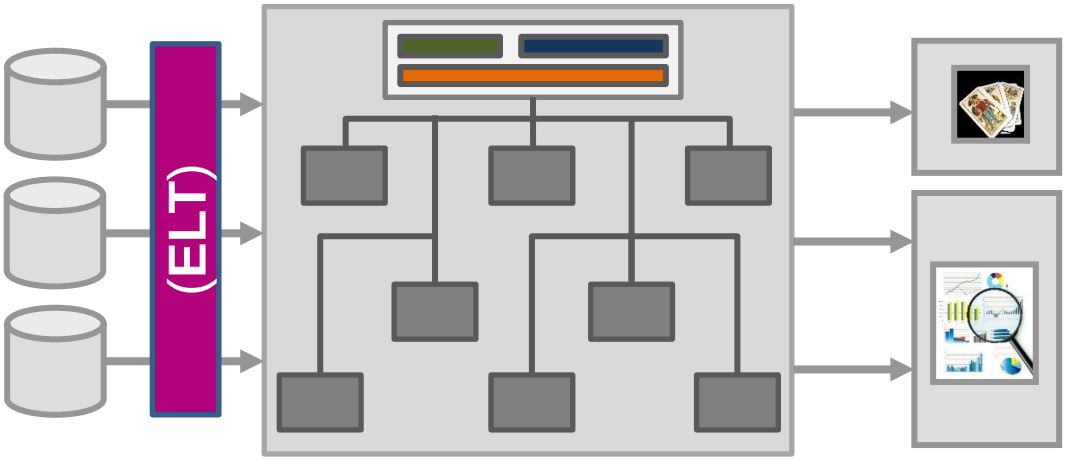
- De ideale wereld (voorgaande studies)
  - best practices
- De realiteit (praktijkvoorbeelden uit projecten)
  - typische problemen van data quality bij analytics-projecten
  - hoe data quality tools de aanpak ondersteunen

Analytics & Data Quality: de ideale wereld (1)

- Vereenvoudigde data supply chain
  - minder E**T**L, minder vertaalslagen (cfr. barrière 2)
  - opportuniteit voor data governance:
    - Data quality management
    - Master data management
    - Information management

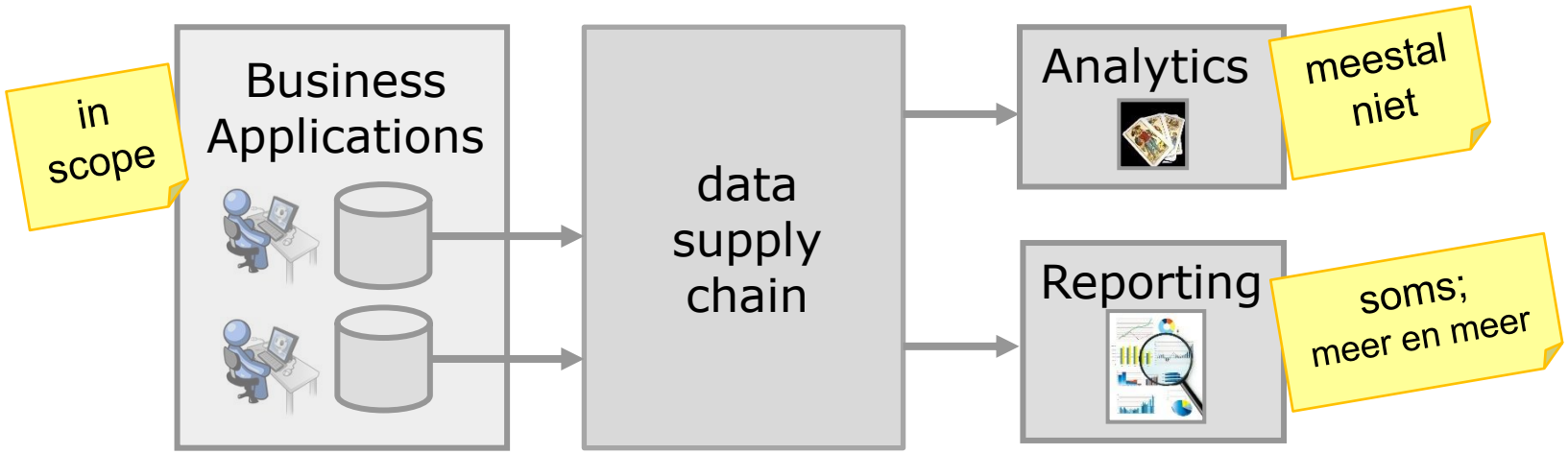
  → op één plaats enten

→ perfecte documentatie



Analytics & Data Quality: de ideale wereld (2)

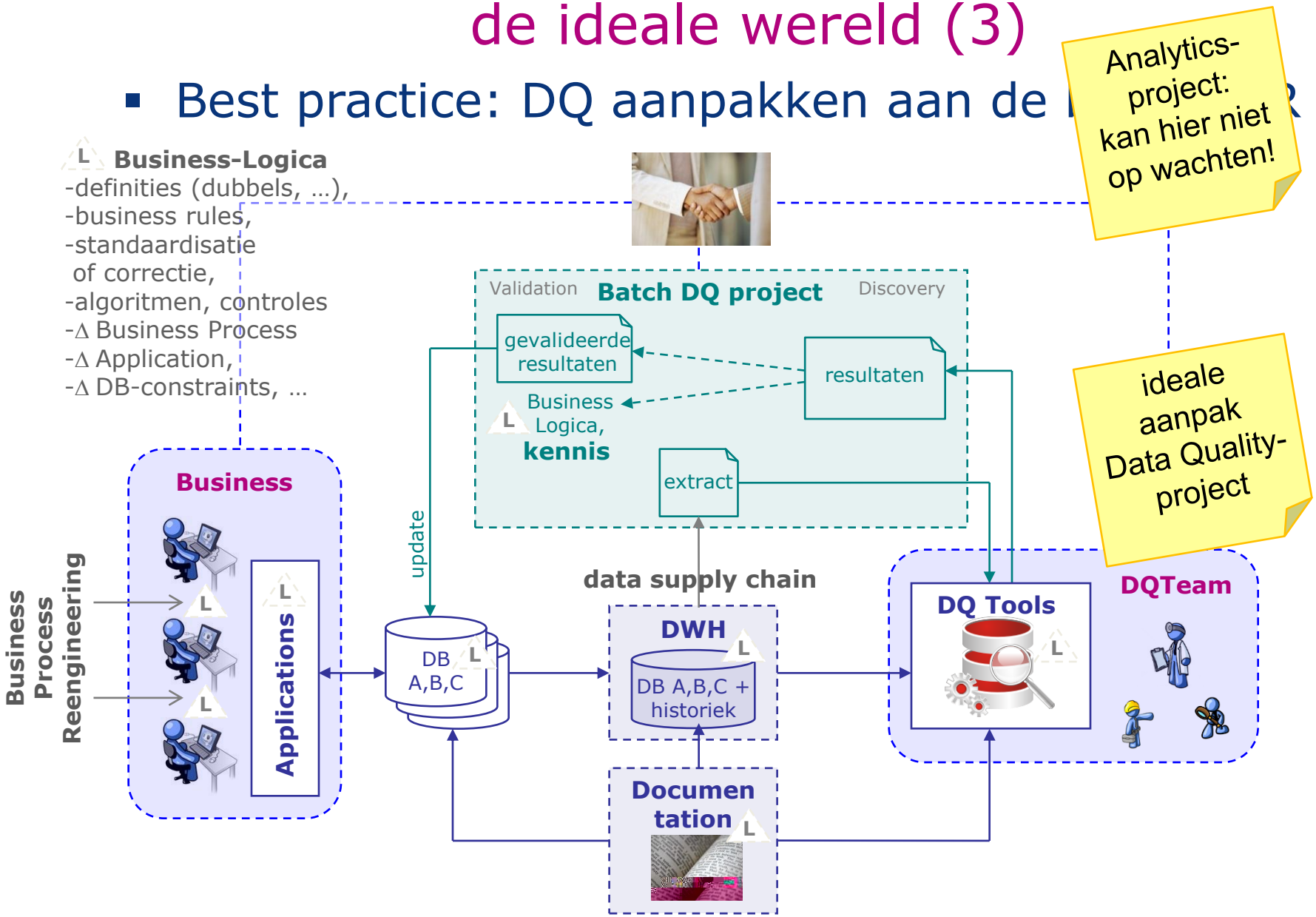
- Best practice: « fitness for use »
  - 100% data quality is onbereikbaar
  - kosten vs. baten
  - *good enough for its (all) intended use*
    - **in scope voor *intended use*?**



Smals Research publicatie « Data Quality: Best Practices », I.Boydens, 2006

Analytics & Data Quality:
de ideale wereld (3)

- Best practice: DQ aanpakken aan de



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-line)
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

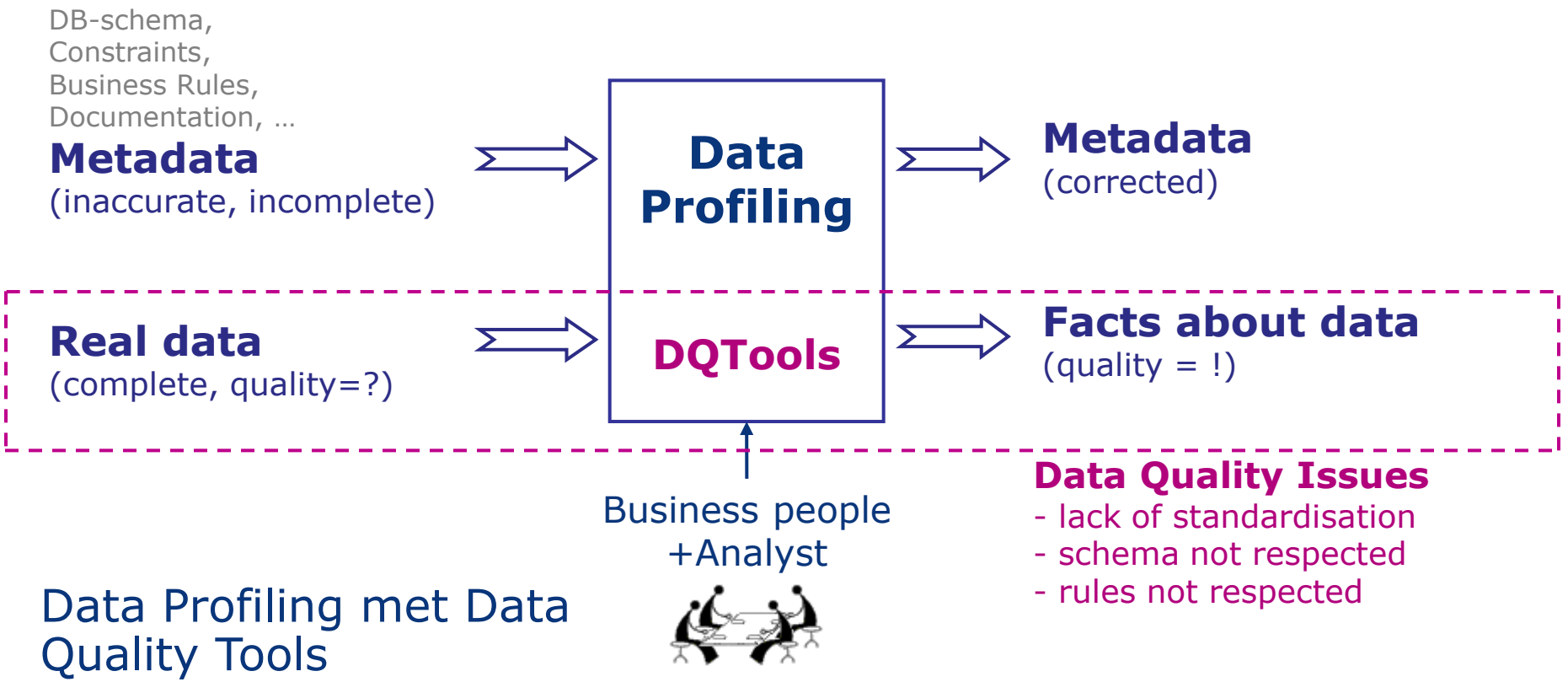
DOMAINE DE DEFINITION :
Valeurs possibles [NULL , 1].
NULL : actif.
1 : supprimé

LONGUEUR : 1
TYPE : Alpha

REC_ID	DWH_CODACT
456	A
457	S
458	A

Doc
?

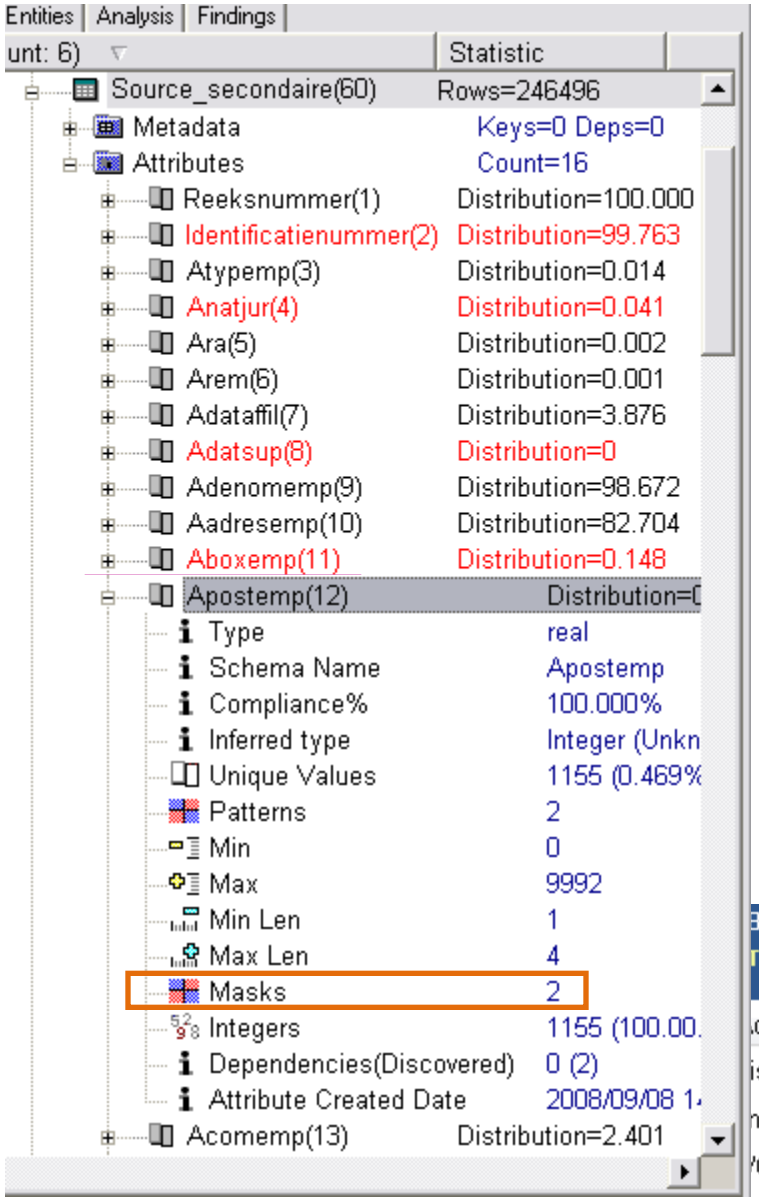
Data Quality Tools ondersteuning: Data Profiling (formele audit)



- 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"

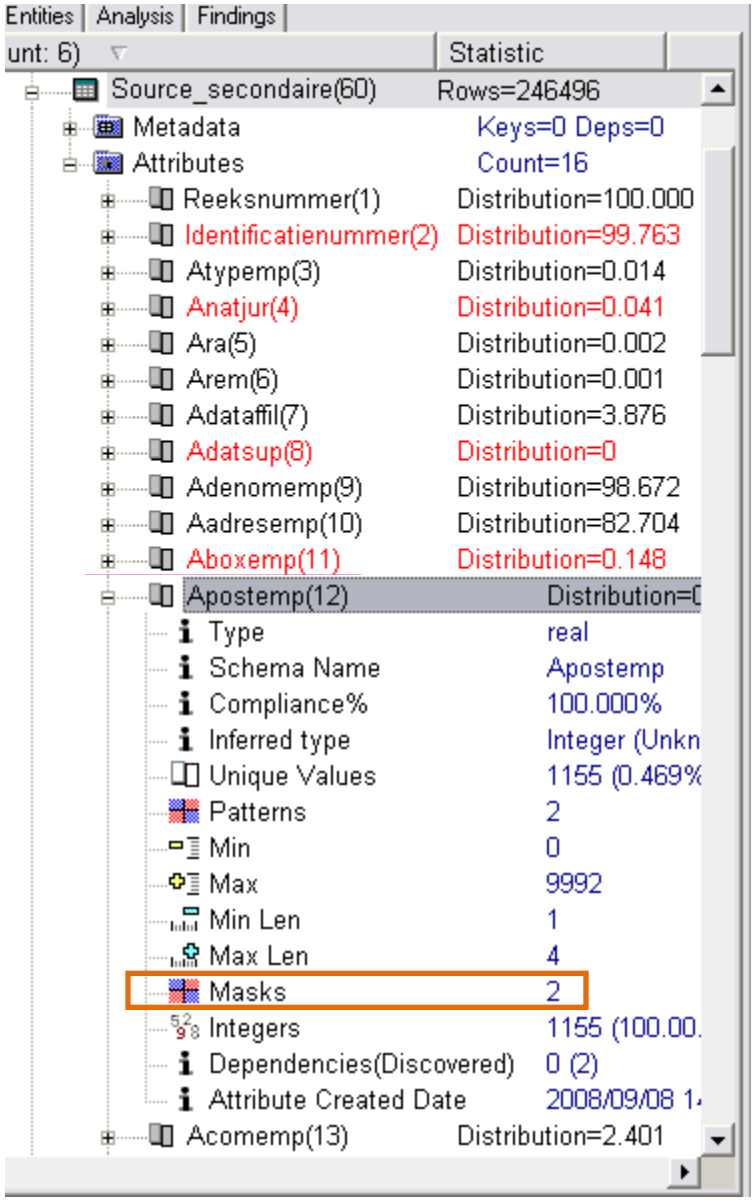
Data Profiling – Column Property Analyse, drill-down (1)

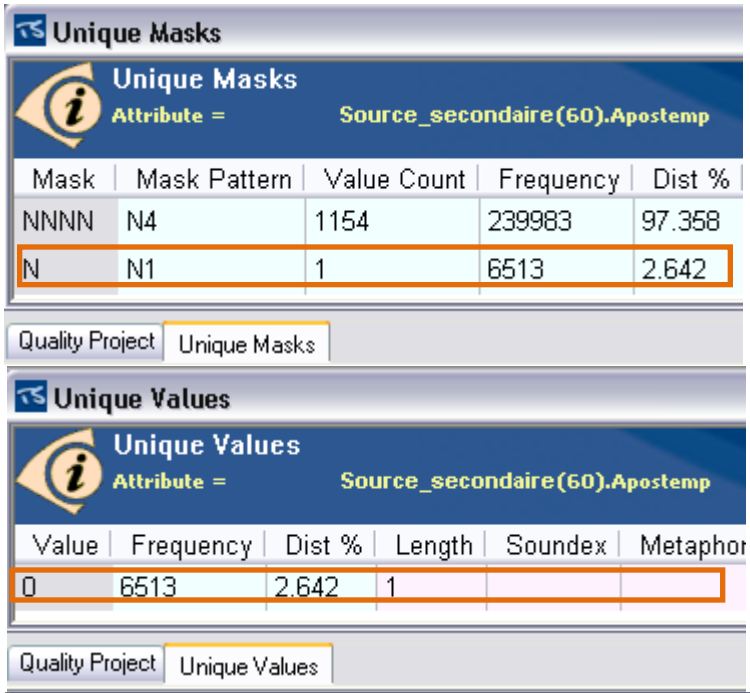


- (1)-(16) geanalyseerd tijdens load
  - gegenereerde metadata
    - Min, Max, Lengtes,
    - Null Values, Unique Values
    - Patronen,
    - Distributies,
    - Business Rules, …
  - detectie Keys & Dependencies
- Apostemp(12): postcode werkgever
  - vb. patronen, a.h.v. Masks

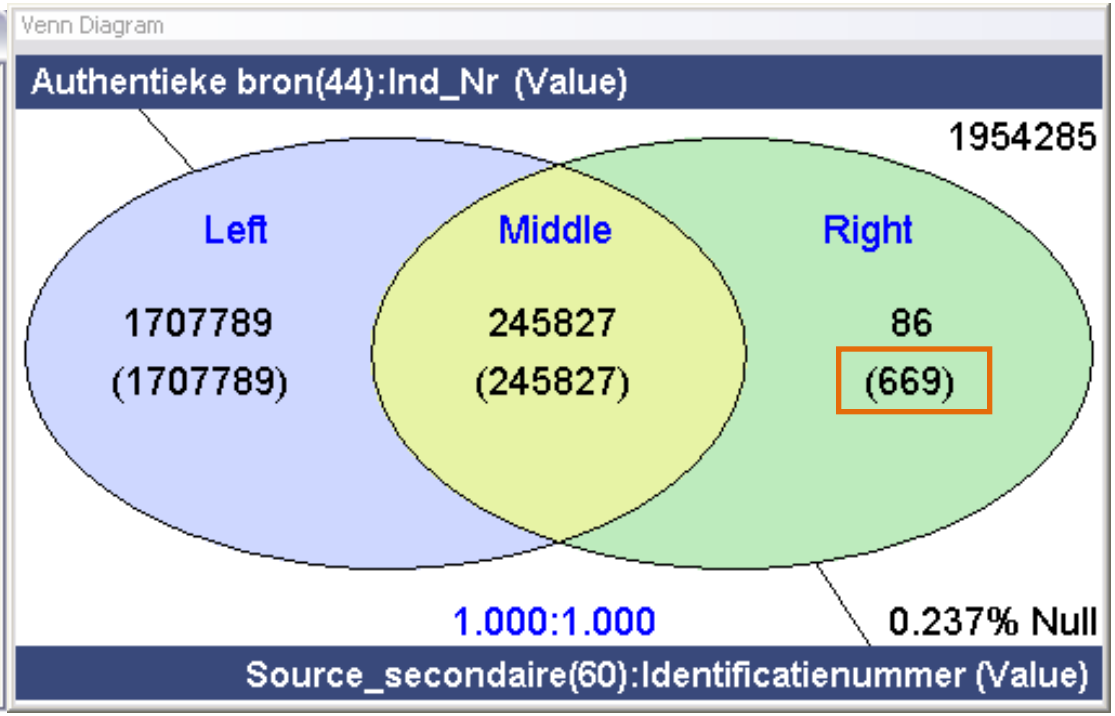
- Performantie: < 5 min op 1 miljoen records

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





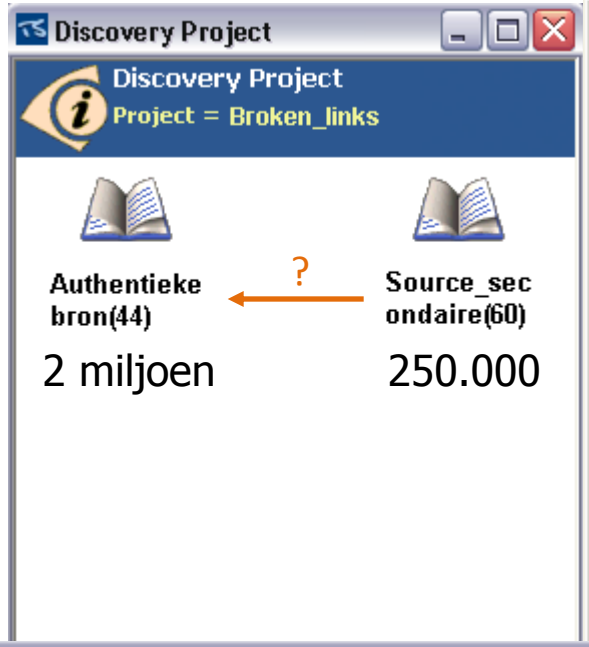
Data Profiling – Referentiële integriteit, foreign keys (1)

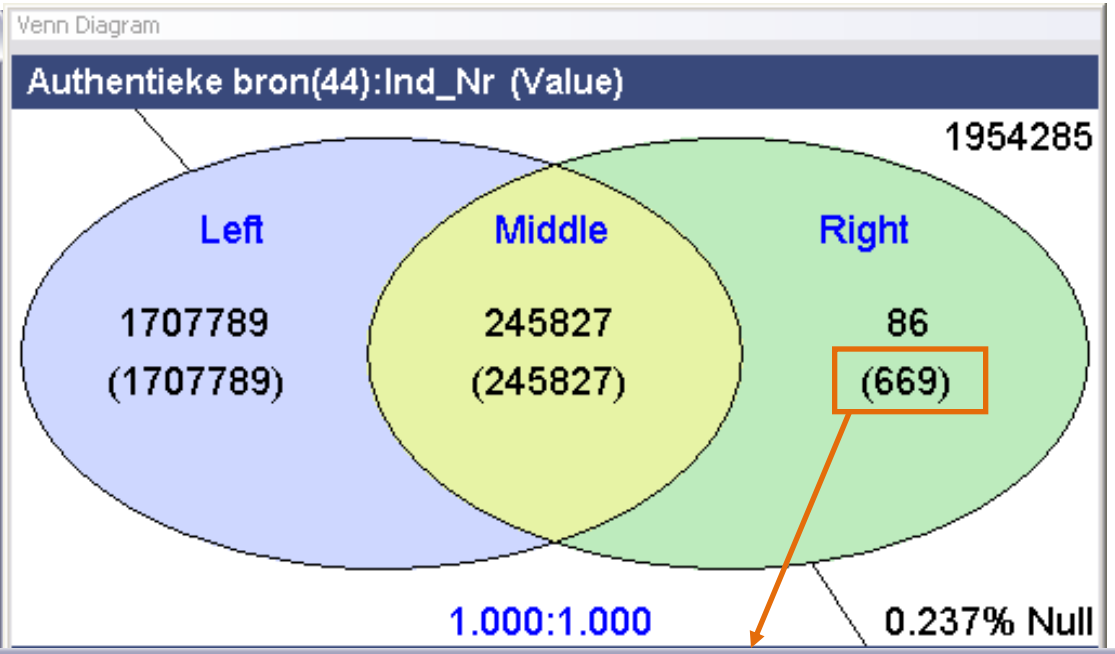


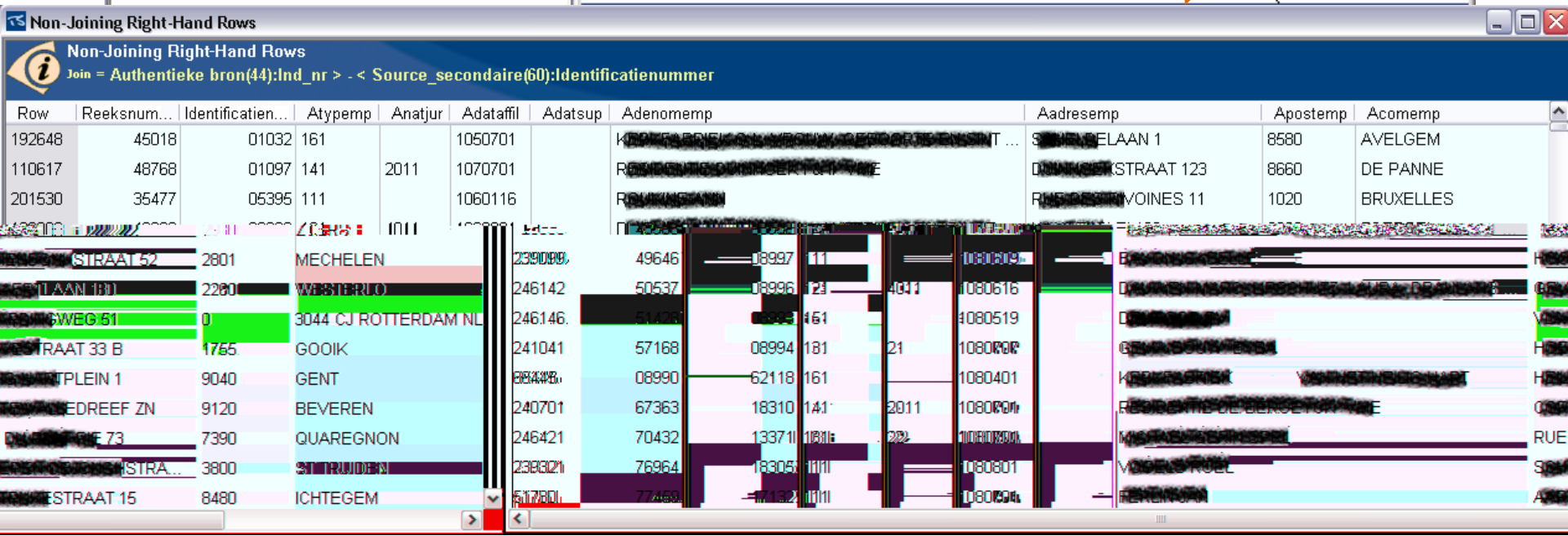
- bron 1: Authentieke bron(44)
  - sleutel: Ind_nr
- bron 2: Source_secondaire(60)
  - Identificatienummer verwijst naar Ind_nr's

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

Data Profiling – Referentiële integriteit, foreign keys (2)







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 Lwijz	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

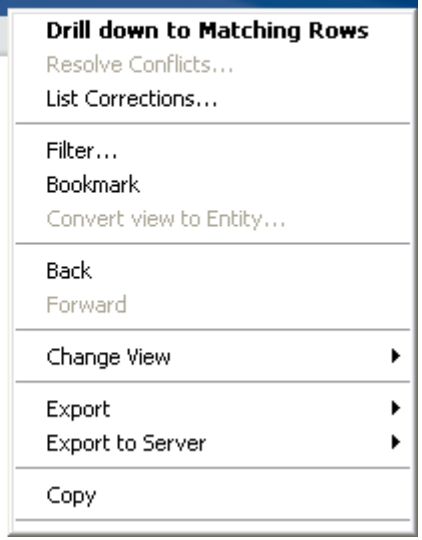
exhaustief, 2.9 miljoen

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

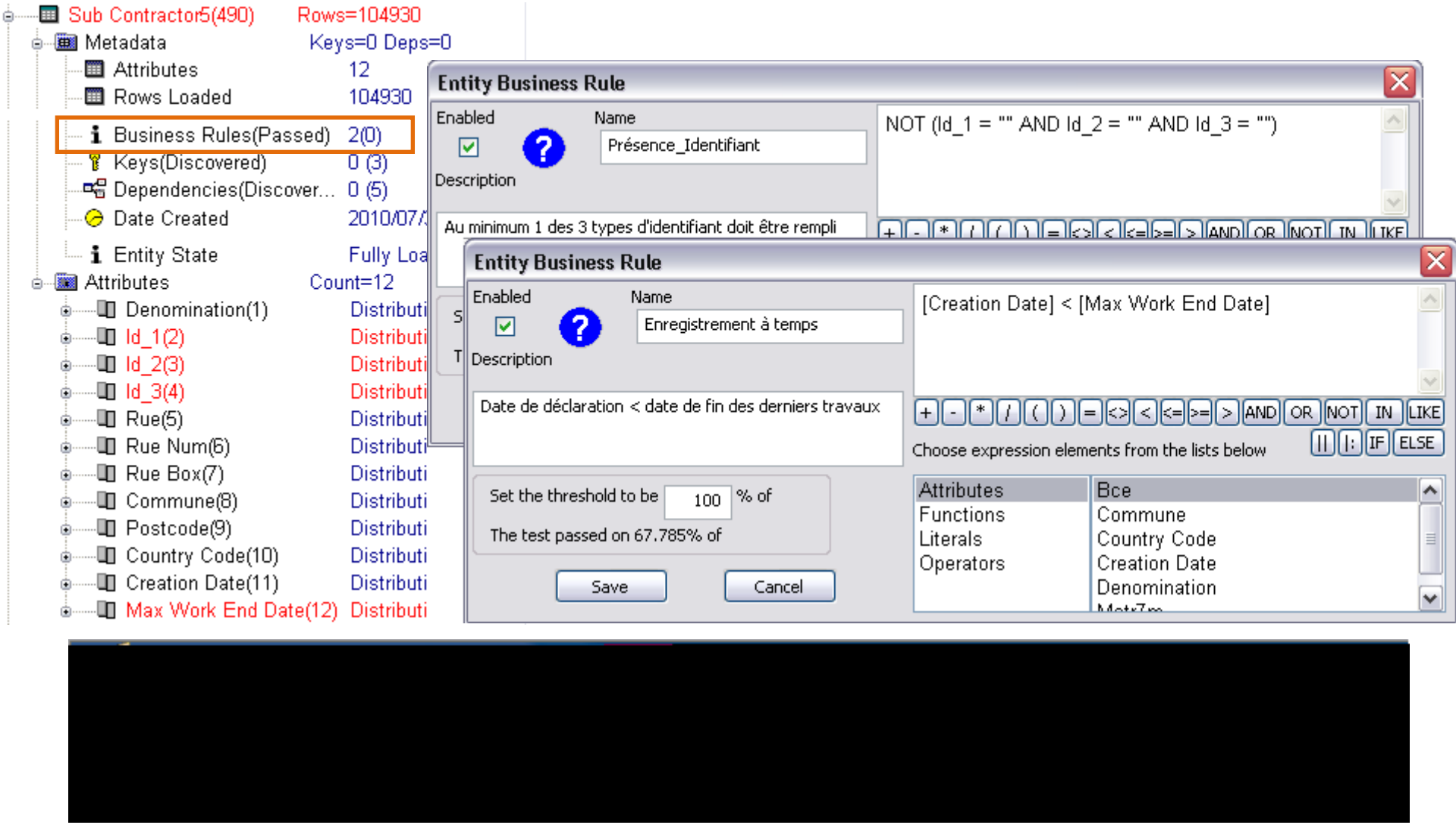
drill-down naar
overzicht van conflicten indien
Quality < 100%

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



Data Profiling – Business Rules (1)



Data Profiling – Business Rules (2)

Failing Rows (Présence_Identifiant)
Entity = Sub Contractor5(490)

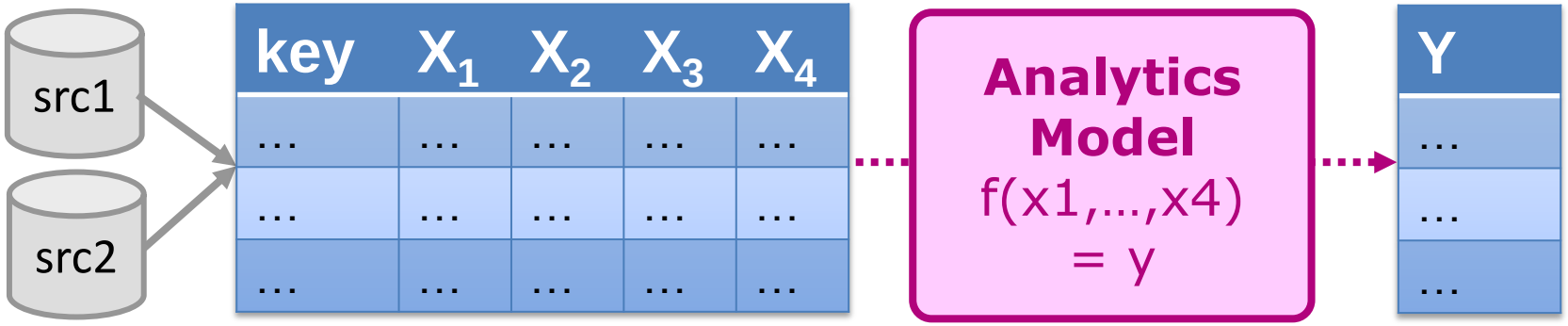
Denomination	Id_1	Id_2	Id_3	Rue	Rue Num	Rue Box	Commune	Postcode	Country...	Creation Date	Max Work End Date
[illegible] SARL				R ...	37/A		59300VA...	0	111	2009-05-20 ...	2011-10-21 00:00:0...
ENTREPRIS...				RU...	9		ANDENNE	5300	150	2005-08-18 ...	2006-06-22 00:00:0...
CONCEPT ...				AV...	15		FOREST	1190	150	2004-02-19 ...	2003-12-01 00:00:0...
[illegible] SPRL				ZO...	11	n/a	MONT D...	7750	150	2009-05-20 ...	2009-03-31 00:00:0...
[illegible] PR...				KL...	6A		OUDEN...	4730 AE	129	2007-09-26 ...	2007-12-31 00:00:0...
[illegible] B...				G...	37		HASSELT	3511	150	2007-03-20 ...	2011-07-31 00:00:0...
[illegible] JON...				RU...	80		CHARLE...	6031	150	2005-12-02 ...	2006-09-28 00:00:0...
[illegible]				BE	10		ZOUTLE...	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
[illegible] ...	76...			BA...	1		WUUST...	2990	150	2007-10-09 ...	2007-08-31 00:00:0...
[illegible] ...	46...	124...		RU...	n/a	n/a	LIBRAM...	6800	150	2009-05-20 ...	2003-11-28 00:00:0...
[illegible] ...	41...	127...		R ...	63	n/a	ETTERB...	1040	150	2009-05-20 ...	2008-05-31 00:00:0...
[illegible] ...	73...		78297...	TO...	7	n/a	HOESELT	3730	150	2007-06-07 ...	2007-06-04 00:00:0...
[illegible] ...	73...		78297...	DA...	2	n/a	HOESELT	3730	150	2007-06-07 ...	2007-06-04 00:00:0...
[illegible] ...	46...	171...		P...	74	n/a	NIEL	2845	150	2009-05-20 ...	2006-04-14 00:00:0...
[illegible] ...	74...		78242...	DU...	n/a	n/a	ST KATE...	2860	150	2005-06-16 ...	2005-01-31 00:00:0...
[illegible]	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
…	…

Smals
ICT for society

Ondersteuning data integratie met dqtools: Data Standardisation – simple domains

Example: country codes

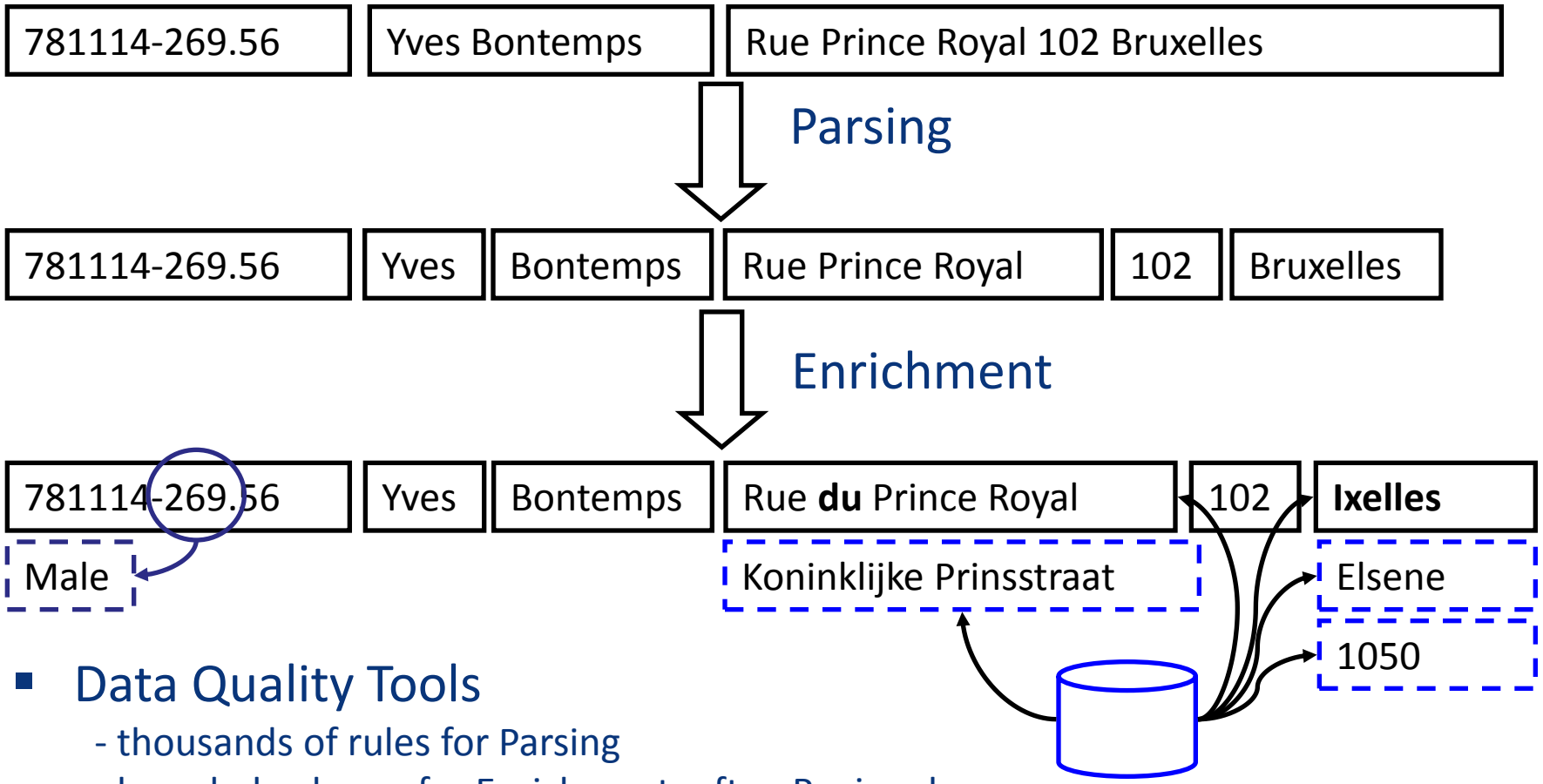
Data Rows (Dynamic)
Entity = pl tranfrmr p26(513)

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 Nl	Ins Cd	Ins Nl
[illegible] ...					150	056	BE	BEL	België	150	België
[illegible] ...	MAH...			BLACKROC...	116	372	IE	IRL	Ierland	116	Ierland /Eire/
[illegible]...	OLD...			WATERSFO...	IRL	372	IE	IRL	Ierland		
[illegible]...	268		067 82	MODRA NAD...	141	703	SK	SVK	Slowakije	141	Slovaakse Republiek
[illegible]...	HIA...		067 82	MODRA NAD...	SK	703	SK	SVK	Slowakije		
[illegible]...	Vizi...		1031	BUDAPEST	115	348	HU	HUN	Hongarije	115	Hongarije (Rep.)
[illegible]...	VIZI...		1031	BUDAPEST	H	348	HU	HUN	Hongarije		
[illegible]...	SICI...		1045 AX	AMSTERDAM	129	528	NL	NLD	Nederland	129	Nederland
[illegible]...	SICI...		1045	AMSTERDAM	NL	528	NL	NLD	Nederland		
[illegible]	MAD...		1131	BUDAPEST	H	348	HU	HUN	Hongarije		
[illegible]...	POS...		1440	AP PURMER...	NL	528	NL	NLD	Nederland		
[illegible].	POS...		1440	AP PURMER...	NL	528	NL	NLD	Nederland		
[illegible]...	czer...	4	20 349	LUBLIN	122	616	PL	POL	Polen	122	Polen (Rep.)
[illegible]...	CZE...		20 349	LUBLIN	PL	616	PL	POL	Polen		
[illegible] ...	Via ...		24129	BERGAMO	128	380	IT	ITA	Italië	128	Italië
[illegible]...	VIA ...		24129	BERGAMO	I	380	IT	ITA	Italië		
[illegible]...	Stee...		2407 BD	ALPHEN AA...	129	528	NL	NLD	Nederland	129	Nederland
[illegible]	STE...		2407	BD ALPHEN ...	NL	528	NL	NLD	Nederland		
[illegible]	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

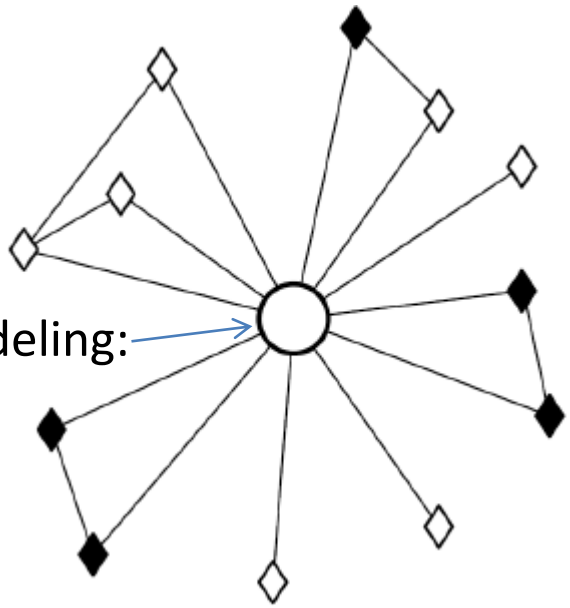
What view would you rather build analytics upon? ...

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	POLYPROPILENE	15	ML	CONICAL	CASE/500	BOTTOM PACKED
1774-4532	CENTRIFUGE TUBE	POLYPROPILENE	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



Fuzzy matching met data quality tools (1)

Source	match type	Denomination	Adres	Boite	Postcd	Commune	Cdpays
L	100	PROJEKTSERWIS	BOHATEROW 63	43	05-100	NOWY DWOR	122
L	100	PROJEKTSERWIS	BOHATEROW 63	43	05-100	NOWY DWOR	122
R	115	PROJEKT SERWIS (A)	UL. BOHATEROW 63	42	05-100	NOWY DROW	PL
R	115	PROJEKT SERWIS A NOWY DWOR	UL BOHATEROW 63	43	05-100	NOWY DWOR	PL
R	135	PROJEKT SERWIS	BOHATEROW 63/43		05-100	NOWY DWOR	PL
R	106	PROJEKT SERWIS	BOHATREROW 63LOK	43	05-100	NOKY DWOR	PL
R	106	PROJEKT SERWIS	BOHATREROW A 63LOK	43	05-100	NOKY DWOR	PL
R	128	PROJEKT SERWIS	BOHATREROW A 63LOK	43	05-100	NOKY DWOR	PL
R	128	PROJEKT SERWIS	BOHATREROW A 63LOK	43	05-100	NOKY DWOR	PL
R	128	PROJEKT SERWIS	BOHATREROW A 63LOK	43	05-100	NOKY DWOR	PL
R	138	SOCIETE PROJEKTSERWIS	BOHATERROW 63/43	N/A	05-100	NOWY DWOR	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	RUE E VANDER AA	RUE ERNEST VANDER AA	1	1	Brussel	BRUSSEL
1020	1020	rue Vander Aa	RUE ERNEST VANDER AA	3	3	Bruxelles	BRUXELLES
1050	1050	91 R VAN AA	RUE VAN AA		91	Elsene	ELSENE
1050	1050	27 R.VAN AA	RUE VAN AA		27	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
1020	1050	rue Van Aa	RUE VAN AA	2	2	Bruxelles	IXELLES
1050	1050	2 R VAN AA	RUE VAN AA		2	Ixelles	BRUXELLES
1000	1000	R JOSEPH II 40	RUE JOSEPH II		40	Bruxelles	BRUXELLES
1000	1000	rue Joseph II 71 (...	RUE JOSEPH II		71	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II	RUE JOSEPH II	71	71	Brussel	BRUSSEL
1040	1000	Rue Joseph II 5-7	RUE JOSEPH II		5-7	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II 67A	RUE JOSEPH II		67A	Bruxelles	BRUXELLES
1030	1000	rue JOSEPH II, 114 -	RUE JOSEPH II	116	114 - 116	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(aantal records)*
  - voorbeeld: « gelijkaardige adressen »
    - parsing
    - adresvalidatie
    - matching
    - post-processing (inspectiedistricten)
  - 250.000 actieve: 10 min
  - 850.000 +historiek: 1u

Background Tasks
Timestamp = Tue Jun 18 02:06:15 CEST 2013

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

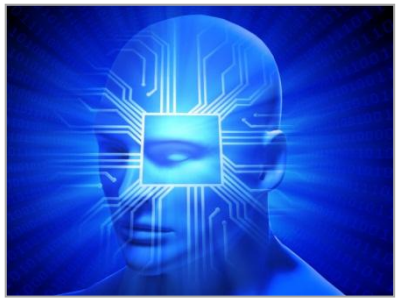
- Belangrijke taken van de Analytics-expert, hebben te lijden aan een gebrek aan dq
  1. Kennis opbouwen bij gebrek aan (up to date) documentatie en standaardisatie
  2. Data-integratie vanuit heterogene bronnen
  3. Netwerk-analytics met fuzzy links
- Data Quality Tools-functionaliteiten kunnen de Analytics-expert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

Smals
ICT for society

Streamlining Analytics

Predictive analytics

De **data supply chain**



Barrières bij de introductie van analytics

Hardware appliances voor analytics

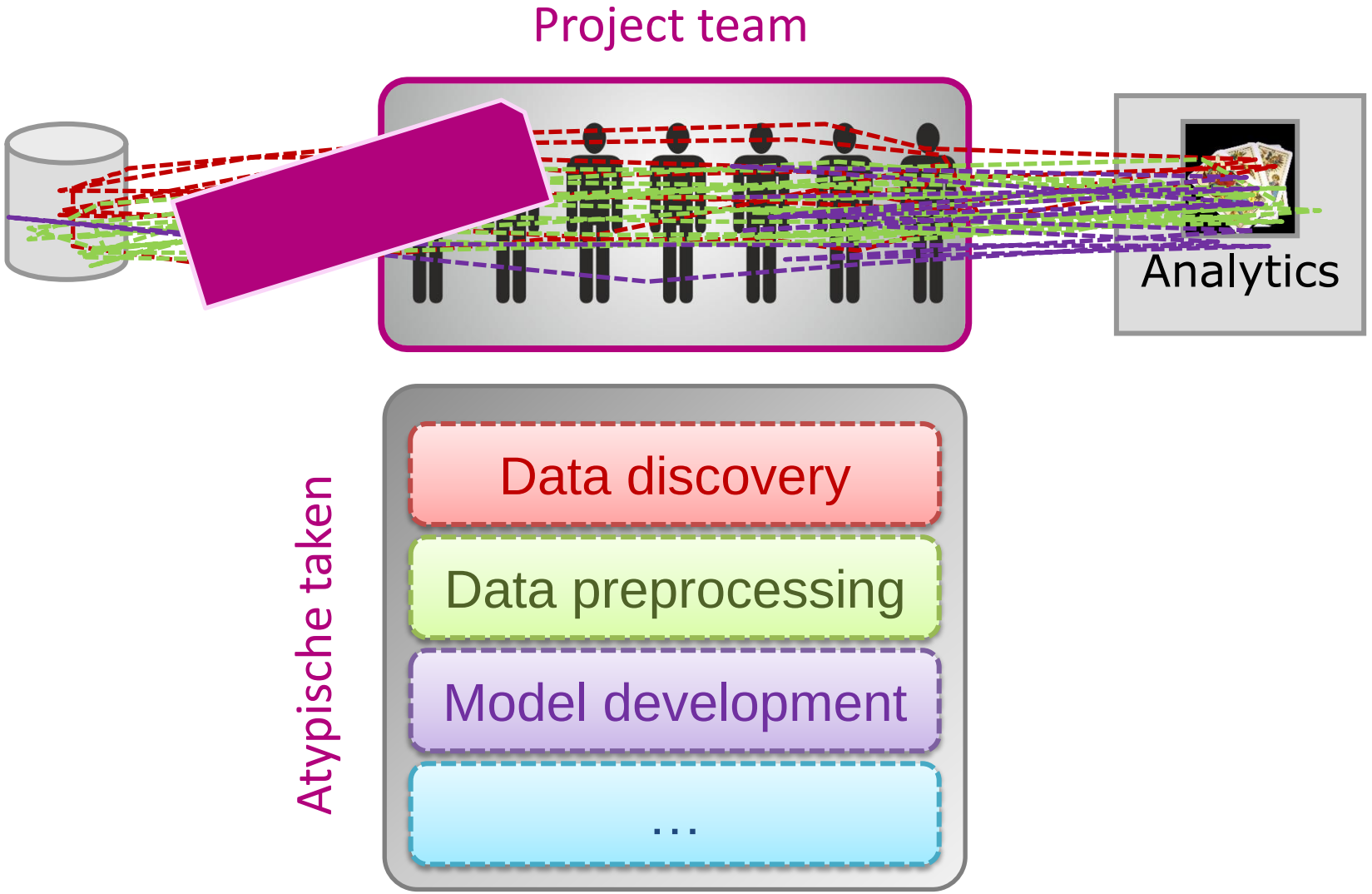
Data quality

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**

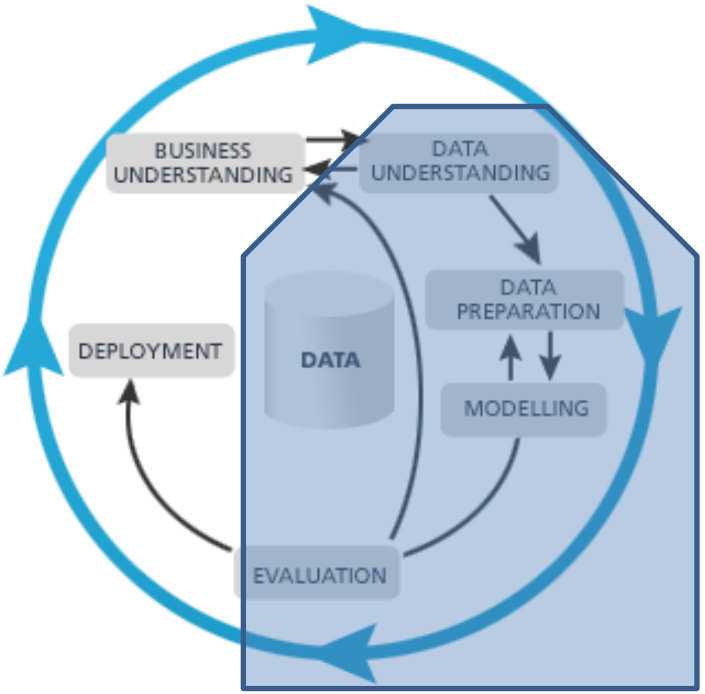
- Gelijkaardige, iteratieve methodologieën
  - KDD original approach: 5 key processes (Fayyad, 1996)
  - CRISP-DM 1.0 (1999+), 2.0 (2006+)
  - SEMMA by SAS

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

http://www.kdd.org/

Analytics-project-**methodologie**

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



« SEMMA »

Sample
Explore
Modify
Model
Assess

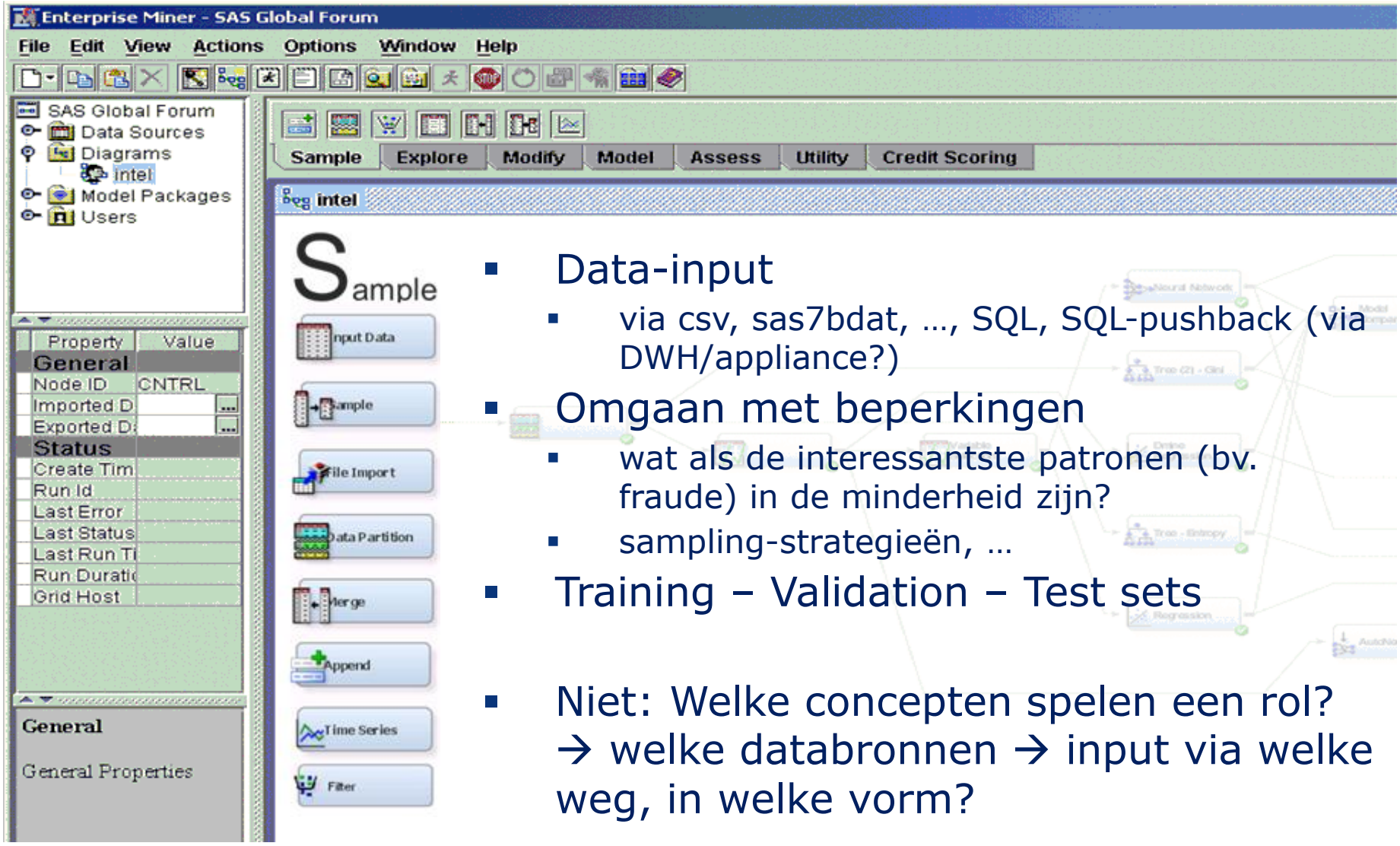
Karakteristieken: iteratief proces, multidisciplinair

→ samenstelling projectteam

→ projectfasen

DM Methodologieën in tools

SAS Enterprise Miner (SEMMA)

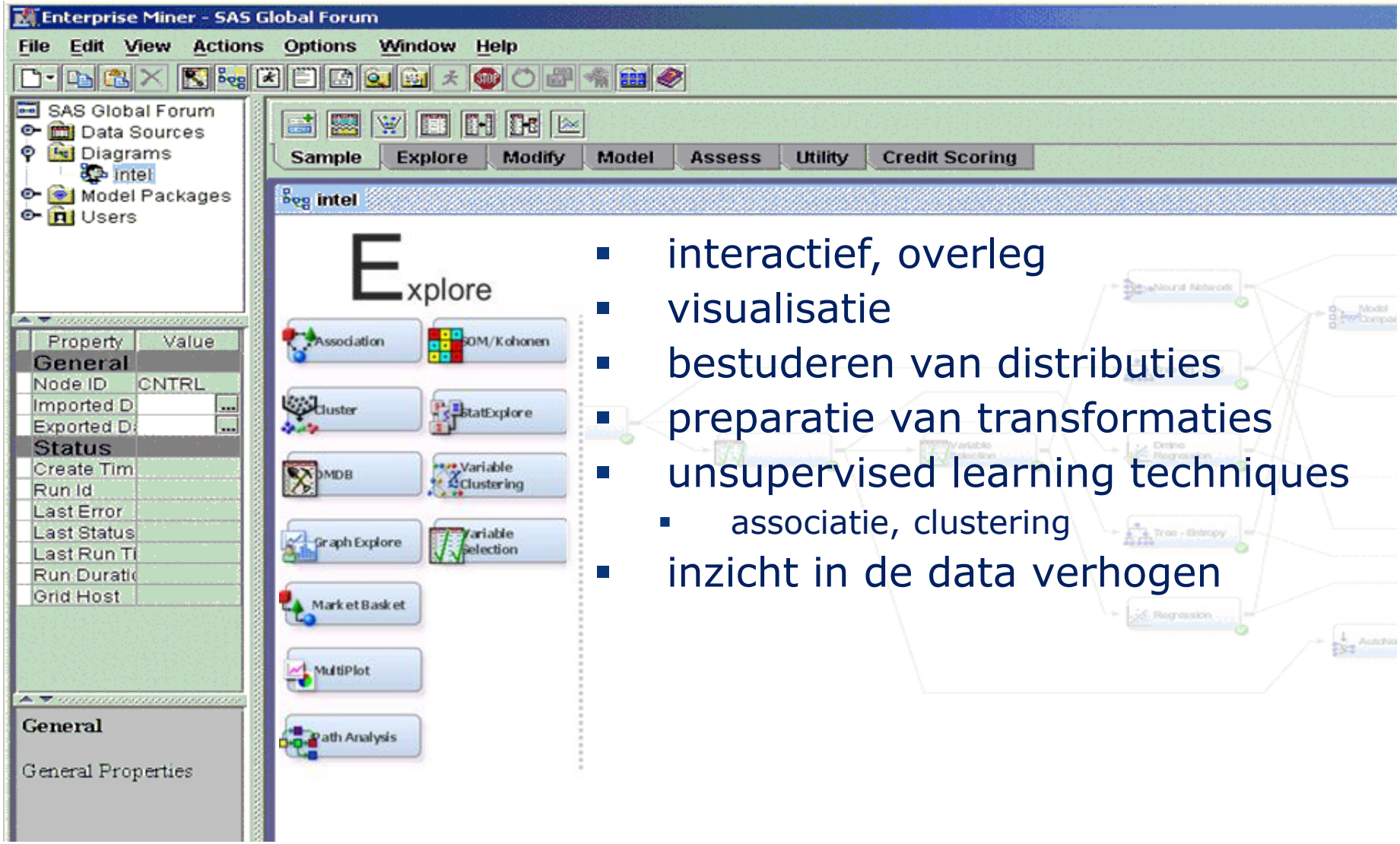


- Data-input
  - via csv, sas7bdat, …, SQL, SQL-pushback (via DWH/appliance?)
- Omgaan met beperkingen
  - wat als de interessantste patronen (bv. fraude) in de minderheid zijn?
  - sampling-strategieën, …
- Training – Validation – Test sets

- Niet: Welke concepten spelen een rol? → welke databronnen → input via welke weg, in welke vorm?

DM Methodologieën in tools

SAS Enterprise Miner (SEMMA)



- 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)

- 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)

- Welk algoritme?
  - interpreteerbaarheid
  - karakteristieken van de inputvariabelen en gezochte patronen
  - meervoudige classificatie
  - …

DM Methodologieën in tools

SAS Enterprise Miner (SEMMA)

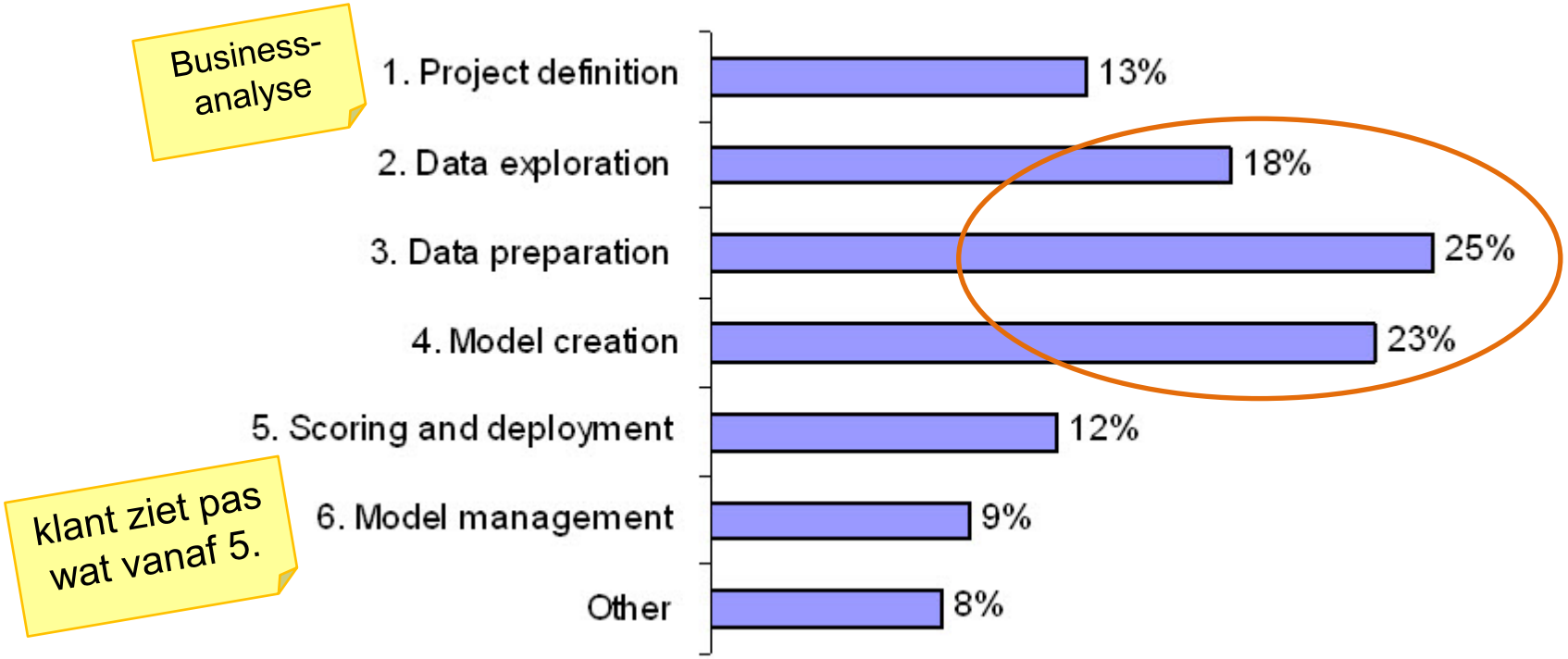
- « model scoring »
  - hoe goed *fit* het model de data?
  - training – validation – test
- Niet: feedback vanuit de acties ten gevolge van het toepassen van het model → « Monitor & maintain »

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

Analytics project (resource) planning

naar een inschatting van benodigde effort …

- Predictive modeler tasks



Ontbreken:

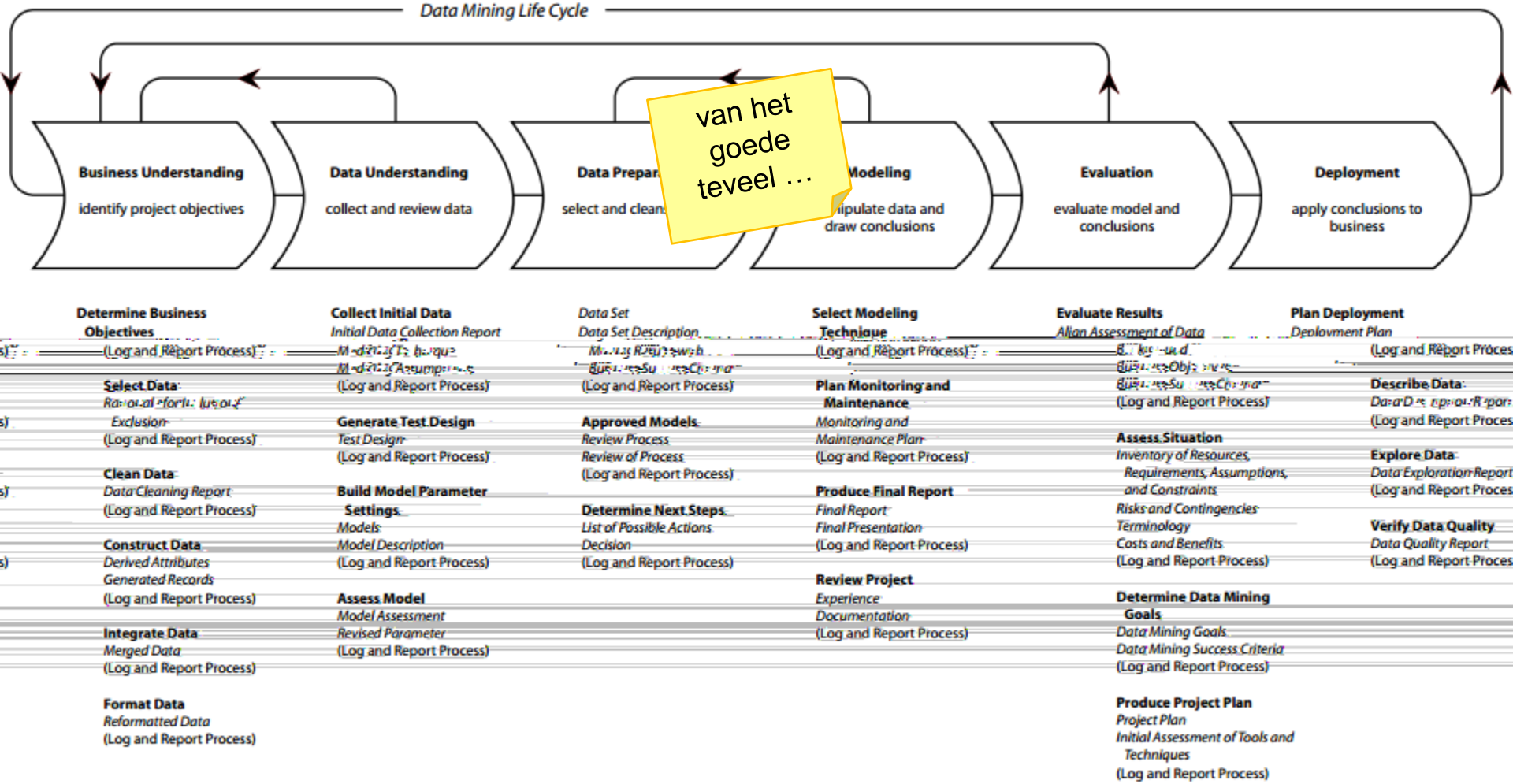
- betrokkenheid andere stakeholders (Business, IT)
- solution architecture
- deeltaken

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

Smals
ICT for society

CRISP-DM: onder de motorkap …

Phases



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)

Smals
ICT for society

werkbare
CRISP-DM based project breakdown

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

Criterium								

Analytics Project Management - caveat

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

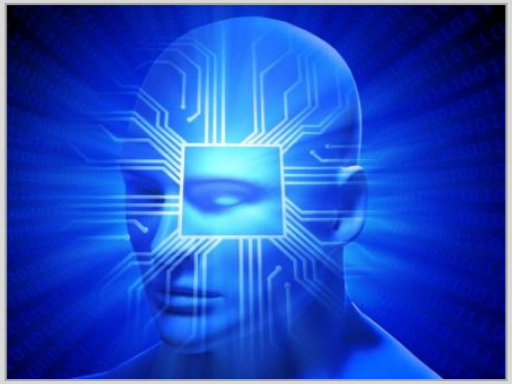
evenals tegen-voorbeelden

- **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

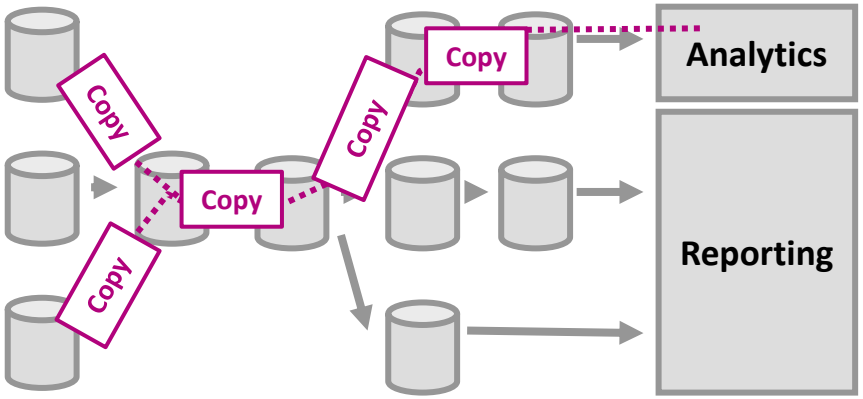
dit komt vaker voor dan men denkt!

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

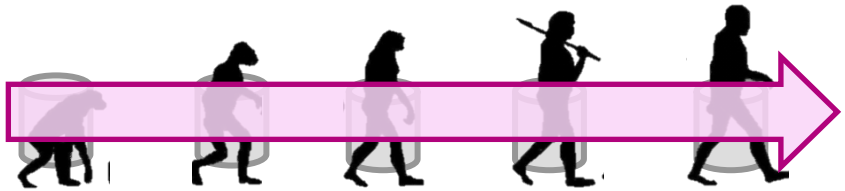
Conclusie



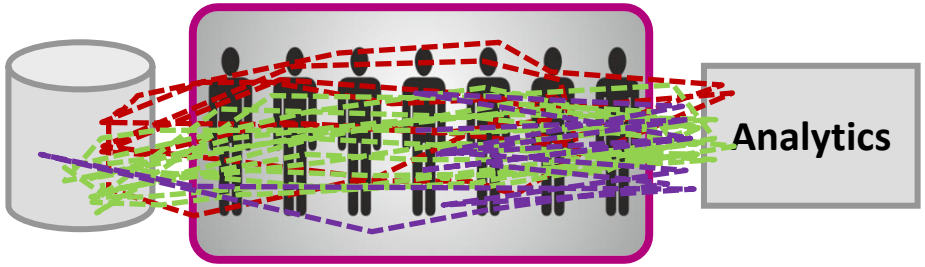
Barrières bij de introductie van analytics



1. Complexe & trage architectuur

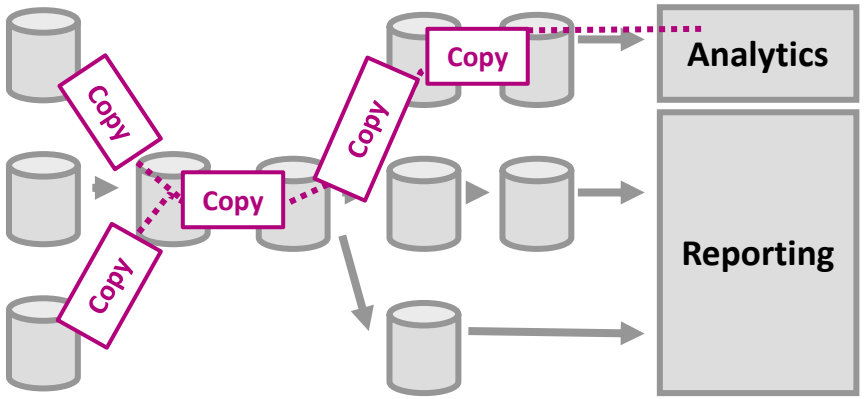


2. Data quality doorheen de supply chain

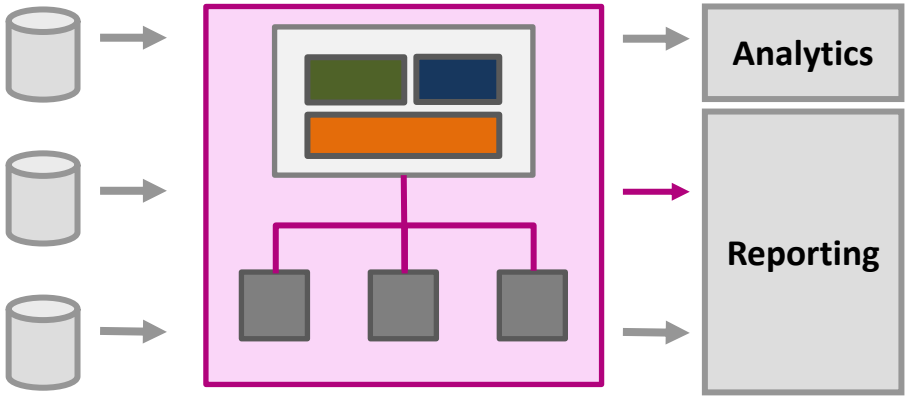


3. Atypische projectstructuur

Streamlining analytics

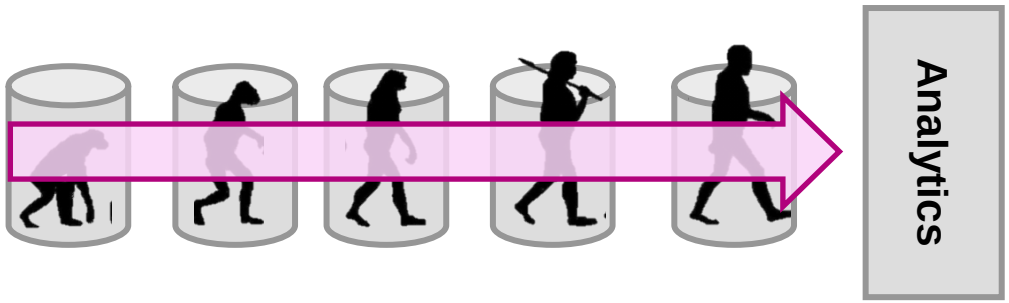


Complexe & trage architectuur



Hardware appliance

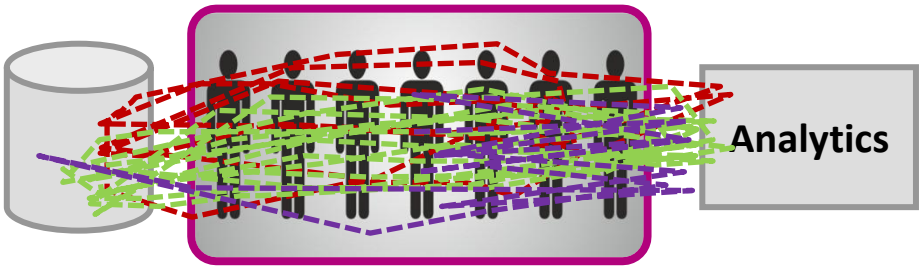
Streamlining analytics



Data quality doorheen de supply chain

Data quality aan de bron

Streamlining analytics

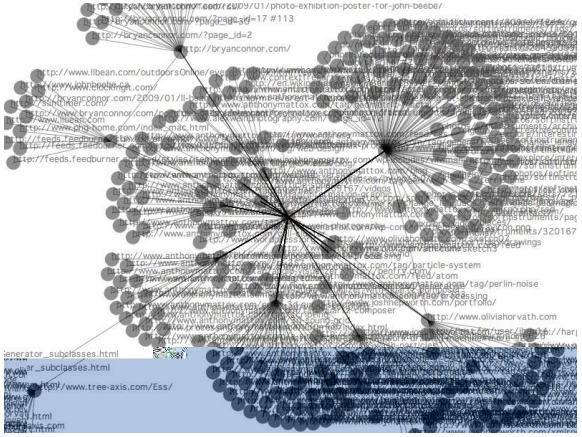


Atypische projectstructuur

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

CRISP-DM based project breakdown

Barrières hangen samen met de **analytics/BI behoeftes**



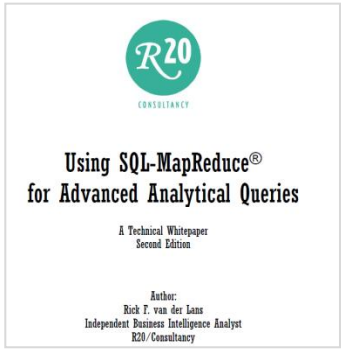
Interactieve visualisatie van Big Data

Operational BI

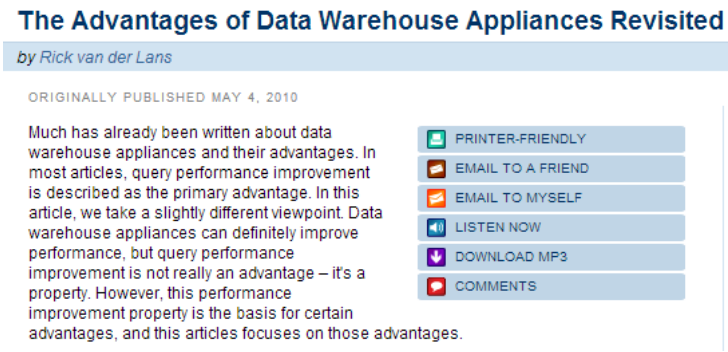
Mobile BI

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

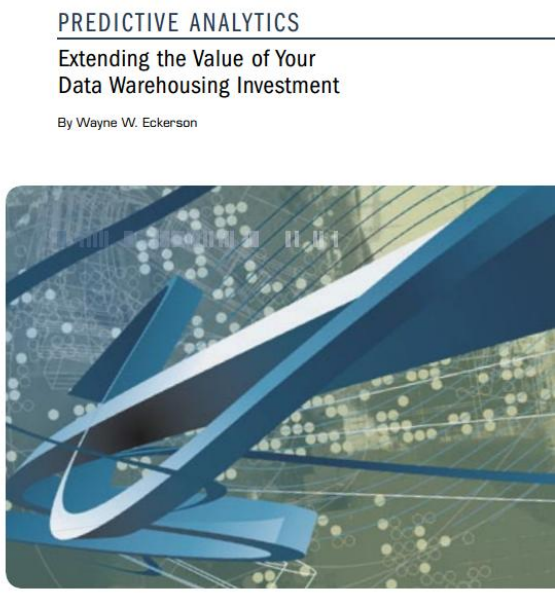
Much has already been written about data warehouse appliances and their advantages. In most articles, query performance improvement is described as the primary advantage. In this article, we take a slightly different viewpoint. Data warehouse appliances can definitely improve performance, but query performance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



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
Dries.VanDromme@smals.be

Onderzoek - @SmalsResearch
http://blogresearch.smalsrech.be