

Älykkäämpien päätösten perusta - integrointi, laatu & varastointi

Pekka Leppänen

Myyntijohtaja

IBM Software Group, Information Management





Älykkäämpi päätöksenteko

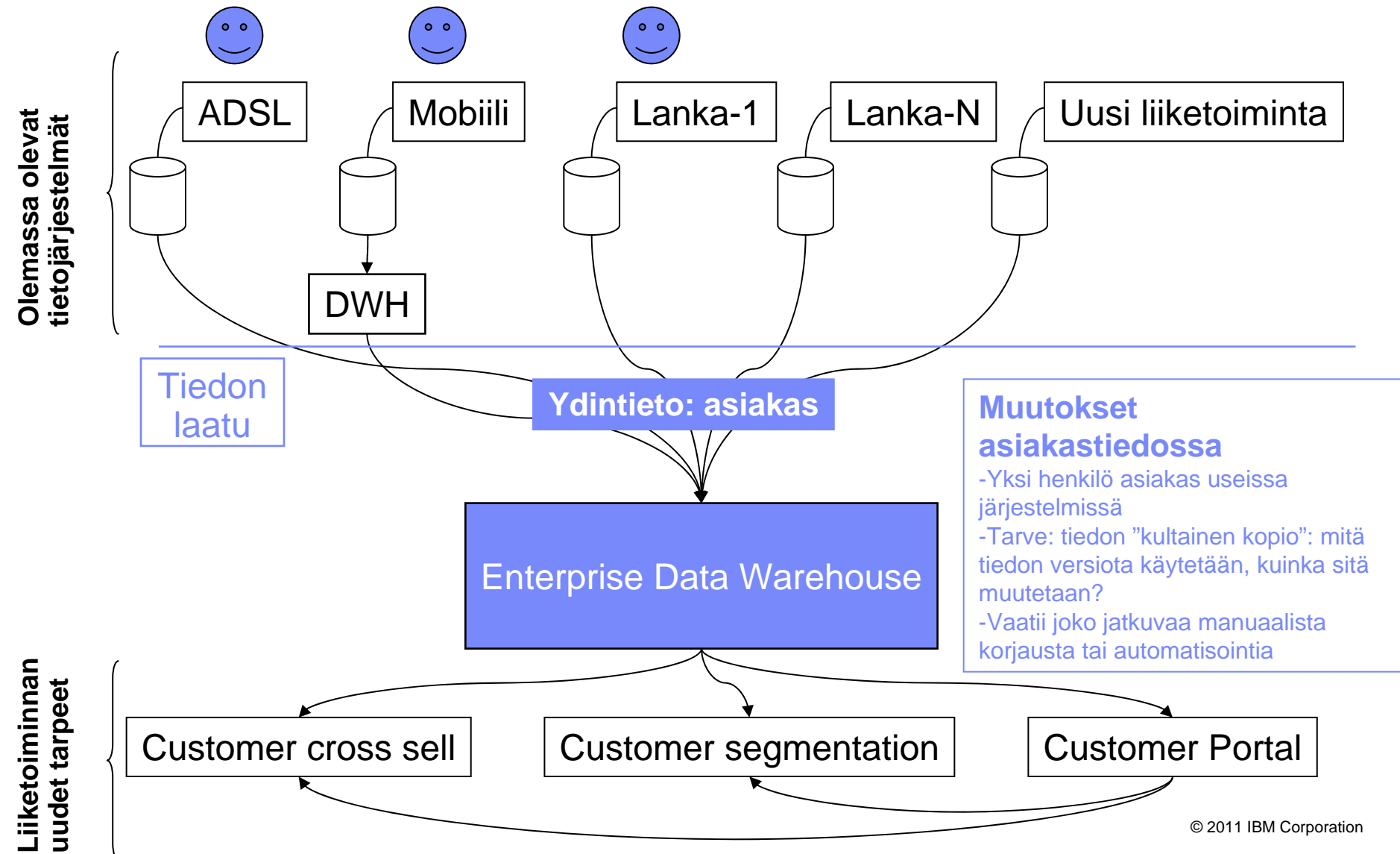
kartoita > mallinna > integroi > tietovarastoi > analysoi > ennakoi

- Esimerkki: uudet liiketoimintatarpeet olemassa olevassa ympäristössä
- Analyttiset tietovarastot
- Tiedon laatu
- Ydintiedon hallinta
- Yhteenveto: tiedon hallinta alusta loppuun



(c) Pekka Leppanen 2010

Esimerkki: tietoliikenneoperaattori





Älykkäämpi päätöksenteko

kartoita > mallinna > integroi > tietovarastoi > analysoi > ennakoi

Esimerkki: uudet liiketoimintatarpeet olemassa olevassa ympäristössä

Analyttiset tietovarastot

Tiedon laatu

Ydintiedon hallinta

Yhteenveto: tiedon hallinta alusta loppuun





Tapahtumankäsittely

Asiakas



tapahtuma



OLTP-tietokanta

Tietokanta



Item: 'Shoes'
Cost: '\$34'
Cust: 'James'



Item	Cost	Cust
Shoes	\$34	James

Tapahtumankäsittelyn työkuorma

Suuri voluumi, suuri nopeus, yksinkertainen tapahtuma



Analyttinen työkuorma

Liiketoiminta analyttikko

Tietovarasto

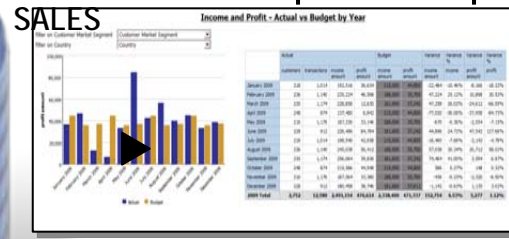
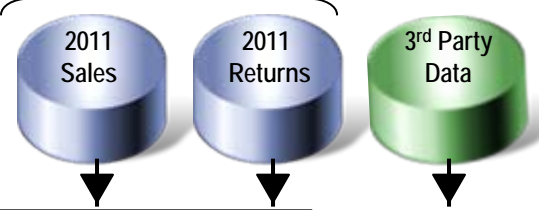


Minimutkainen kysely

Sales & Profit for Shoes & Belts Year >= 2005



Tapahtumatiedot



2005

Raportit ja analyysit

Analyttinen työkuorma

Monimutkainen, moniulotteinen, suuret historialliset tietomäärät,

Perinteisen ja analyttisen työkuorman profilointia: tyypillisiä vaatimuksia

	Tapahtuman- käsittely	Analytiikka
Kysely	Yksinkertainen	Monimutkainen
Käytettävyys	Erittäin korkea	Korkea
Aikajakso	Reaaliaika	Historia <i>(ja nykyhetki)</i>
Koko	Suuri	Erittäin suuri
Sisältö	Raaka	Puhdistettu
Toteutus	OLTP	Data Warehouse
Käyttö	Operatiivinen	Päätöksenteko

The TwinFin™ Appliance – Revolutionizing Analytics



- Purpose-built analytics engine
- Integrated database, server & storage
- Standard interfaces
- Low total cost of ownership
- Speed: 10-100x faster than traditional systems
- Simplicity: Minimal administration and tuning
- Scalability: Peta-scale user data capacity
- Smart: High-performance advanced analytics

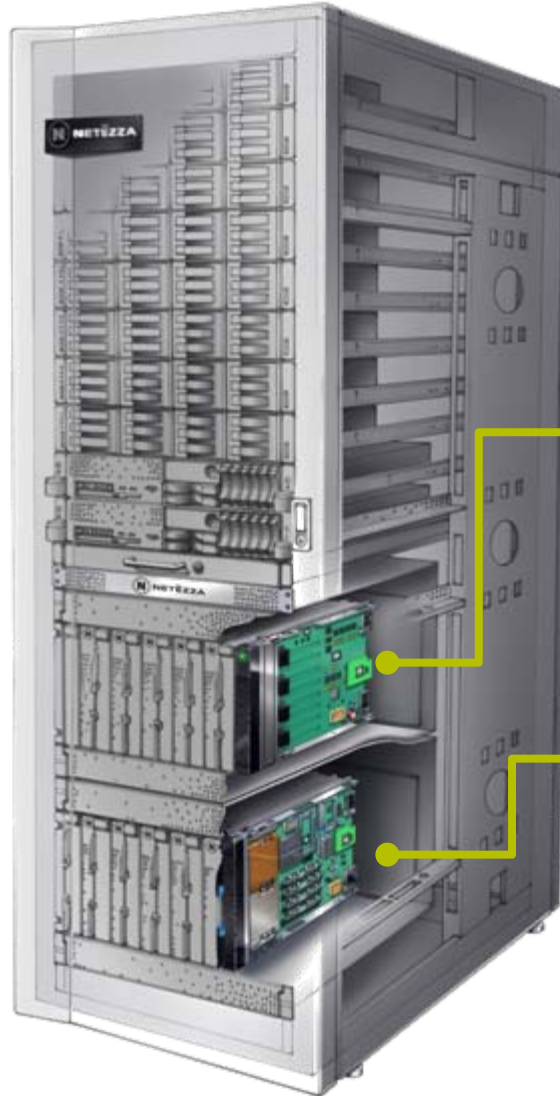
The IBM Netezza TwinFin™ Architecture

Optimized Hardware+Software

Purpose-built for high performance analytics; requires no tuning

True MPP

All processors fully utilized for maximum speed and efficiency



Streaming Data

Hardware-based query acceleration for blistering fast results

Deep Analytics

Complex analytics executed in-database for deeper insights



42



Älykkäämpi päätöksenteko

kartoita > mallinna > integroi > tietovarastoi > analysoi > ennakoi

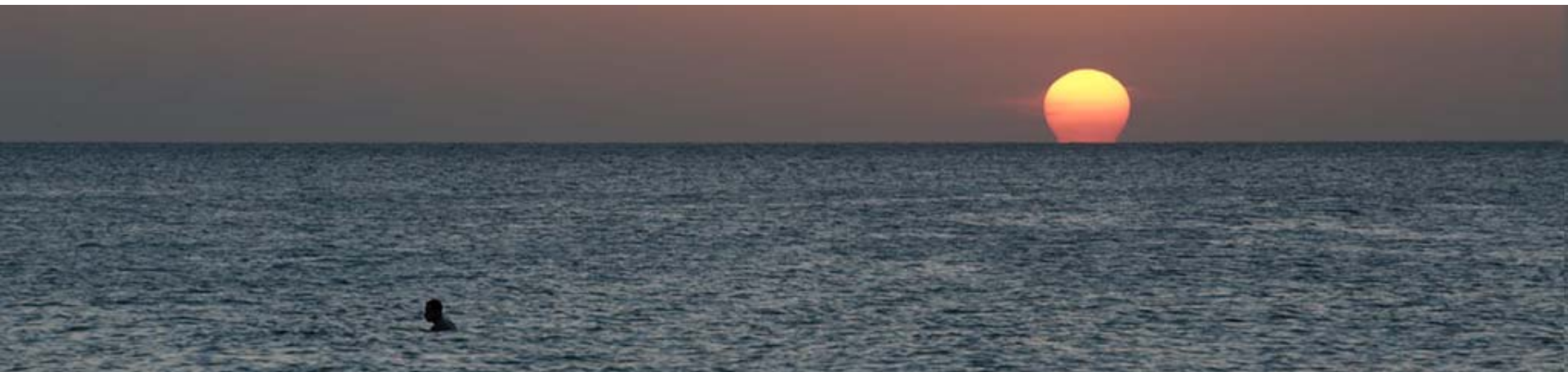
Esimerkki: uudet liiketoimintatarpeet olemassa olevassa ympäristössä

Analyttiset tietovarastot

Tiedon laatu

Ydintiedon hallinta

Yhteenveto: tiedon hallinta alusta loppuun





Common Data Problems

Lack of information standards

- Different formats & structures across different systems

Kate A. Roberts	416 Columbus Ave #2, Boston, Mass 02116
Catherine Roberts	Four sixteen Columbus APT2, Boston, MA 02116
Mrs. K. Roberts	416 Columbus Suite #2, Suffolk County 02116

Data surprises in individual fields

- Data misplaced in the database

Name	Tax ID	Telephone
J Smith DBA Lime Cons.	228-02-1975	6173380300
Williams & Co. C/O Bill	025-37-1888	415-392-2000
1st Natl Provident	34-2671434	3380321
HP 15 State St.	508-466-1200	Orlando

Information buried in free-form fields

```
WING ASSY DRILL 4 HOLE USE 5J868A HEXBOLT 1/4 INCH
WING ASSEMBY, USE 5J868-A HEX BOLT .25" - DRILL FOUR HOLES
USE 4 5J868A BOLTS (HEX .25) - DRILL HOLES FOR EA ON WING ASSEM
RUDER, TAP 6 WHOLES, SECURE W/KL2301 RIVETS (10 CM)
```

Data myopia

- Lack of consistent identifiers inhibit a single view

19-84-103	RS232 Cable 6' M-F Cands
CS-89641	6 ft. Cable Male-F, RS232 #87951
C&SUCH6	Male/Female 25 PIN 6 Foot Cable

The redundancy nightmare

- Duplicate records with a lack of standards

90328574	IBM	187 N.Pk. Str. Salem NH 01456
90328575	I.B.M. Inc.	187 N.Pk. St. Salem NH 01456
90238495	Int. Bus. Machines	187 No. Park St Salem NH 04156
90233479	International Bus. M.	187 Park Ave Salem NH 04156
90233489	Inter-Nation Consults	15 Main Street Andover MA 02341
90345672	I.B. Manufacturing	Park Blvd. Bostno MA 04106



So, What Constitutes Data Quality?

Data is standardized

Data is fit for purpose (conforms to rules)

Each record is unique

View of information is complete

Records are certified against authoritative sources

Lineage is understood

Data quality is measured over time





IBM InfoSphere Information Server

A Platform Enabling Enterprise Data Quality

IBM InfoSphere Information Server

Unified Deployment

Understand



Discover, model, and govern information structure and content

Cleanse



Standardize, merge, and correct information

Transform



Combine and restructure information for new uses

Deliver



Synchronize, virtualize and move information for in-line delivery

Unified Metadata Management

Parallel Processing

Rich Connectivity to Applications, Data, and Content



InfoSphere Foundation Tools



*Manage Business
Terms*



*Discover Data
Relationships*



*Design Enterprise
Models*



*Assess, Monitor,
Manage Data Quality*



*Capture Design
Specifications*



Monitor Data Flows

- Terminologia: liiketoiminnasta tekniikkaan
- Liiketoimintatiedon löytäminen: raaka-datasta liiketoimintaobjekteihin
- Tiedon mallintaminen ja visualisointi
- Tiedon profilointi ja laadun analysointi
- Tiedon vaaimusten hallinta
- Tiedon vaikutusala: mistä se tulee ja mihin sitä käyteään



Data Cleansing: InfoSphere QualityStage

Provides specialized data quality processing

- Ensures clean, standardized, de-duplicated information
- Enables a single version of the truth

Provides visual tools for designing quality rules and matching logic

- Seamlessly integrated with DataStage (one engine, one metamodel, one UI)
- Precisely calibrates matching rules

Allows quality logic to be deployed seamlessly within ETL, or as shared services



Subject Matter Experts



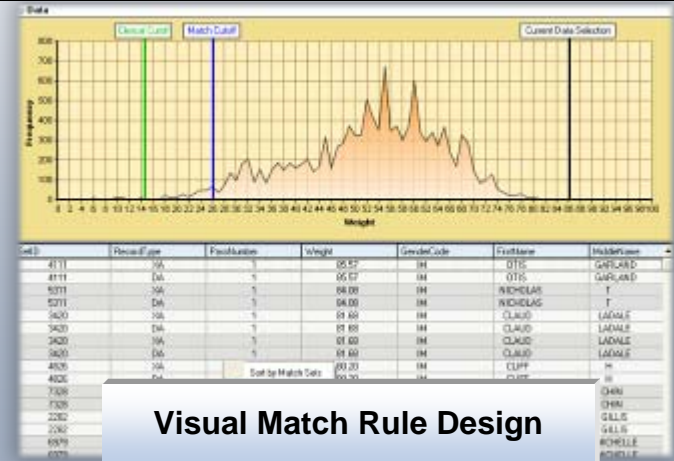
Data Analysts

Cleanse



InfoSphere QualityStage™

Standardize and correct source data fields, and match records together across sources to create a single view



Visual Match Rule Design



Measuring Data Quality Over Time

Complete analysis of structure and content

View differences between current state and the baseline

Analysis can be run on a scheduled basis, or embedded in batch processes

The screenshot shows the IBM Information Server interface. The left pane displays a tree view of data sources, with 'STATE_ABBREVIATION' selected. The main area shows the 'Differences' tab, which compares the current state against a baseline. Two tables are visible:

Value & Format Profile		
Name	Checkpoint	Baseline
Cardinality	42	41
# Distinct Values	1027	1026
# Distinct Formats	2	2
Standard Deviation Value Frequency	0	0
Standard Deviation Format Frequency	0	0
# Null	3	3
% Null	7.142857	7.317073

Completeness & Validity Measures			
Name	Checkpoint	Baseline	
# Incomplete	3	3	
% Incomplete	7.142857	7.317073	
# Invalid	0	0	
% Invalid	0	0	
# Format Violations	0	0	
% Format Violations	0	0	



Älykkäämpi päätöksenteko

kartoita > mallinna > integroi > tietovarastoi > analysoi > ennakoi

Esimerkki: uudet liiketoimintatarpeet olemassa olevassa ympäristössä

Analyttiset tietovarastot

Tiedon laatu

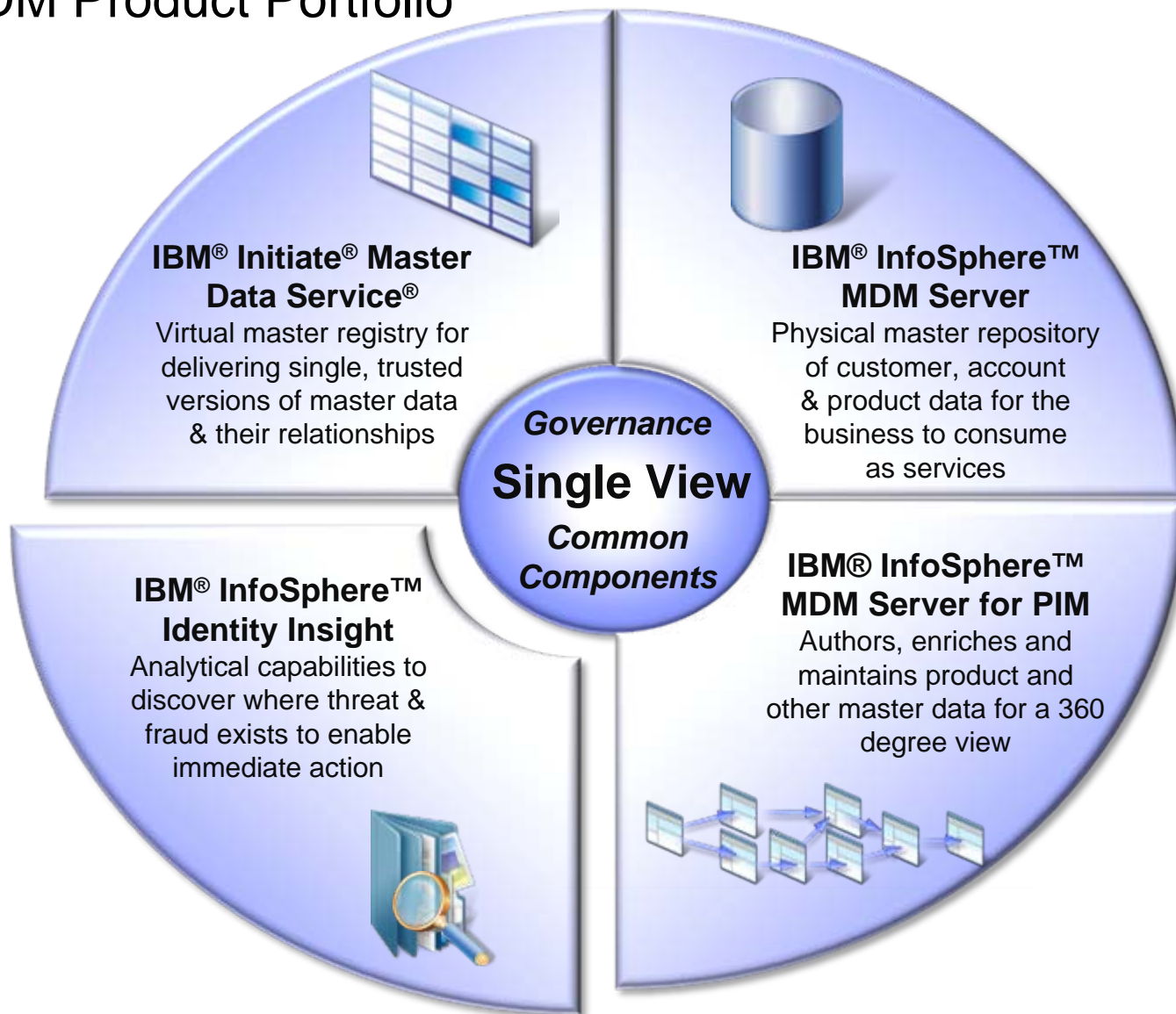
Ydintiedon hallinta

Yhteenveto: tiedon hallinta alusta loppuun





IBM® MDM Product Portfolio





Älykkäämpi päätöksenteko

kartoita > mallinna > integroi > tietovarastoi > analysoi > ennakoi

Esimerkki: uudet liiketoimintatarpeet olemassa olevassa ympäristössä

Analyttiset tietovarastot

Tiedon laatu

Ydintiedon hallinta

Yhteenveto: tiedon hallinta alusta loppuun



Information Management End to End

