Stockholms universitet

Presentation:
DW Architechtures

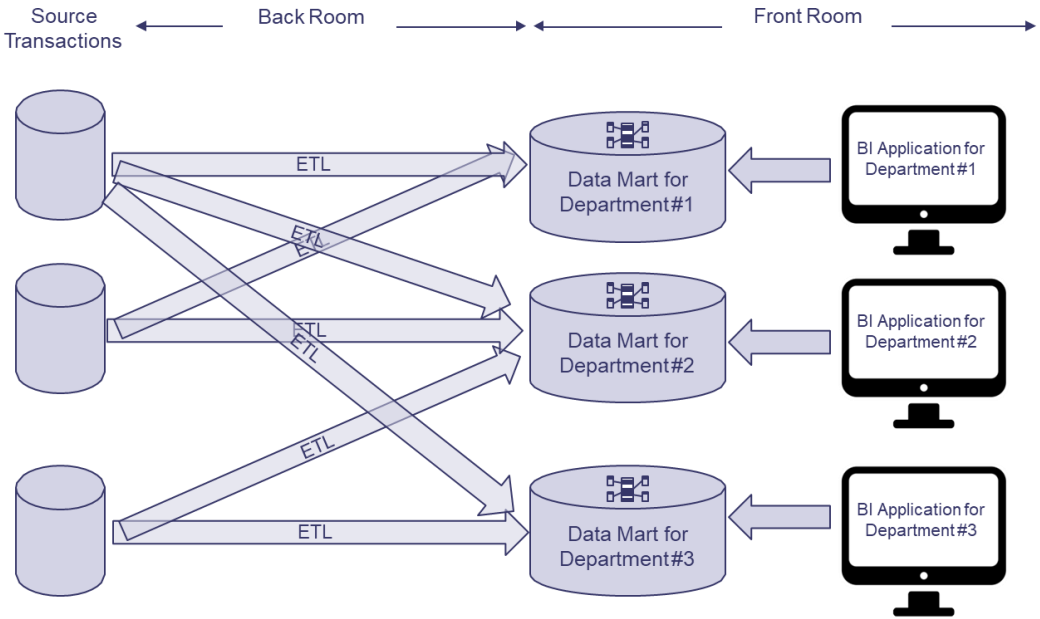
Erik Perjons

DSV, Stockholm University

Stockholms universitet

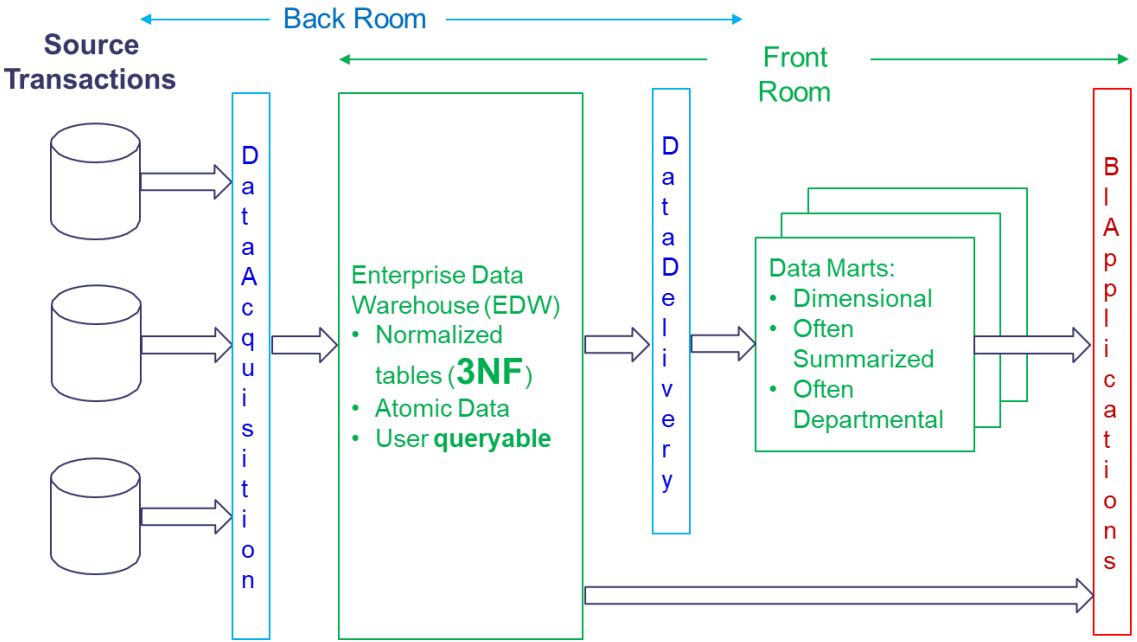
Four types of Data Warehouse architectures

Independent Data Mart Architecture



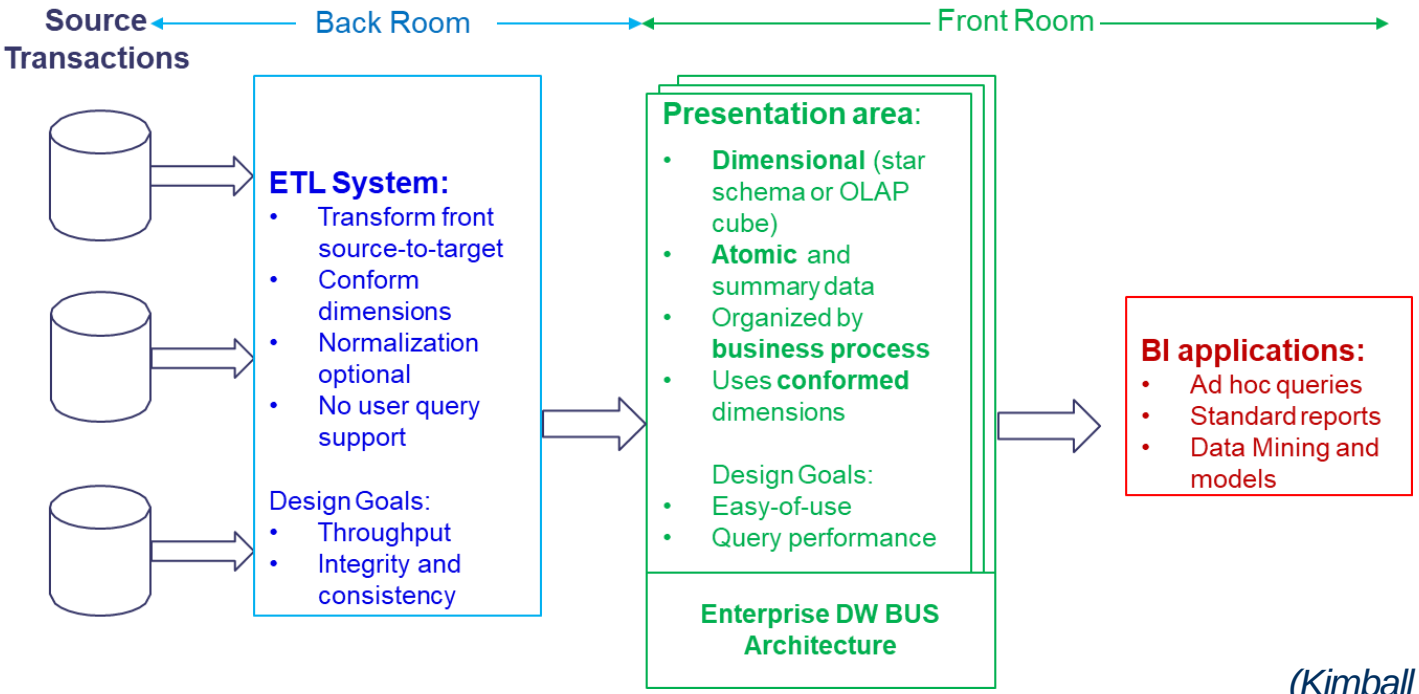
(Kimball & Ross, 2013)

Hub-and-Spoke Corporate Information Factury Archiecture (Inmon)



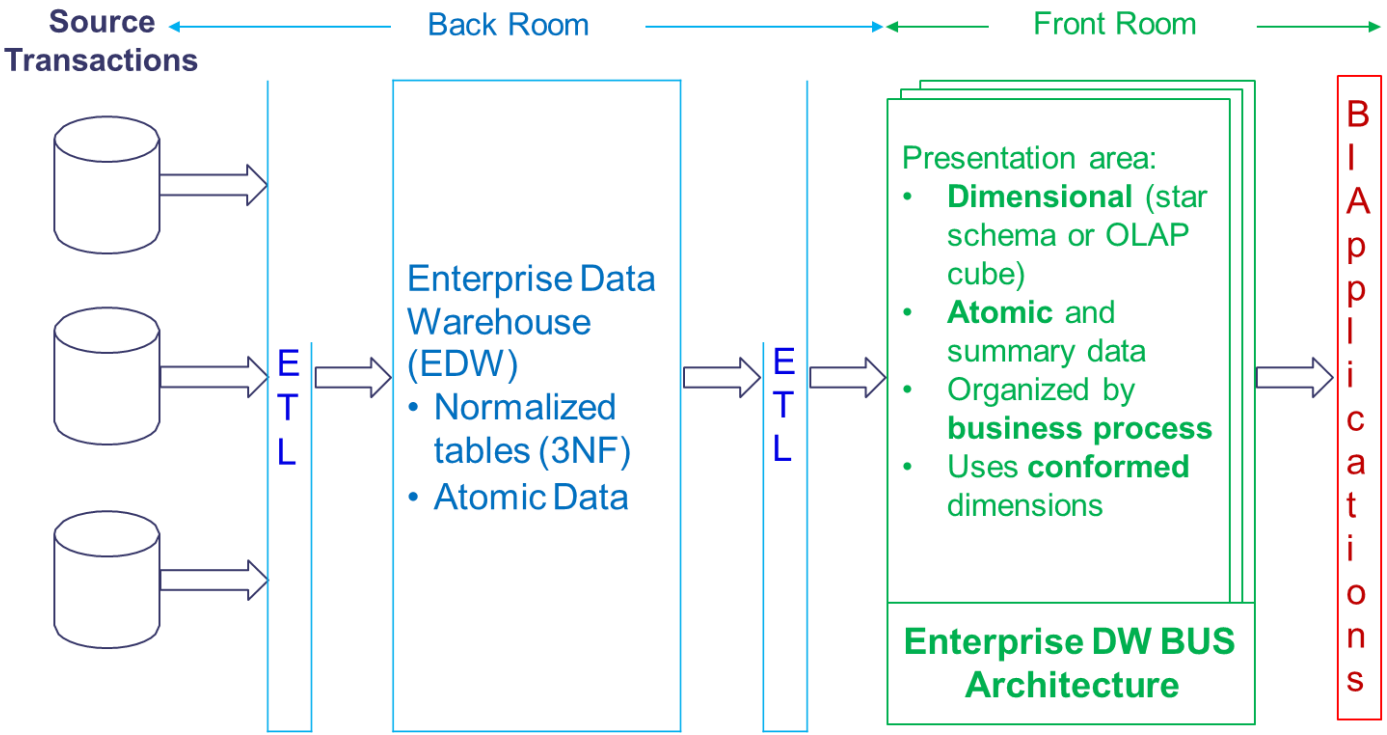
(Kimball & Ross, 2013)

Kimball/Ross DW/BI Architecture



(Kimball & Ross, 2013)

Hybrid Architecture

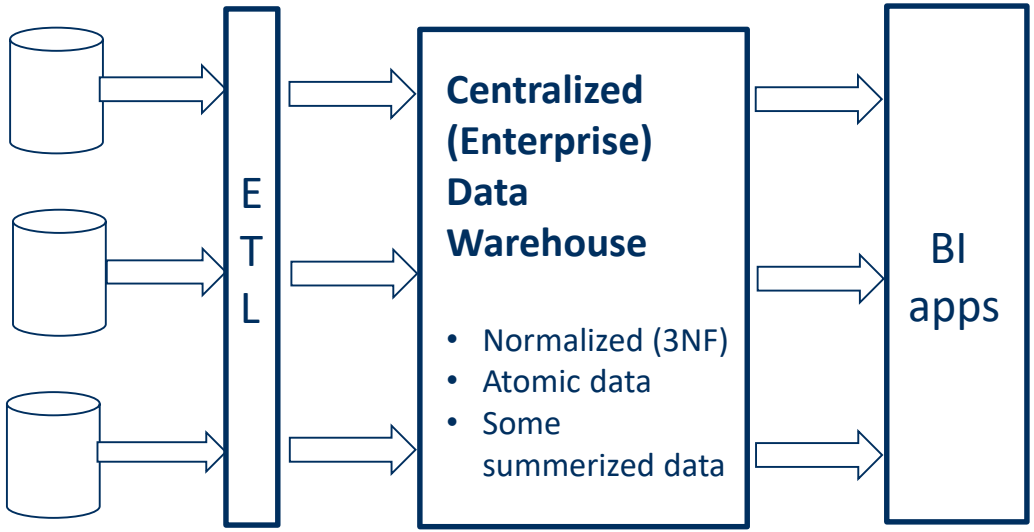


(Kimball & Ross, 2013)

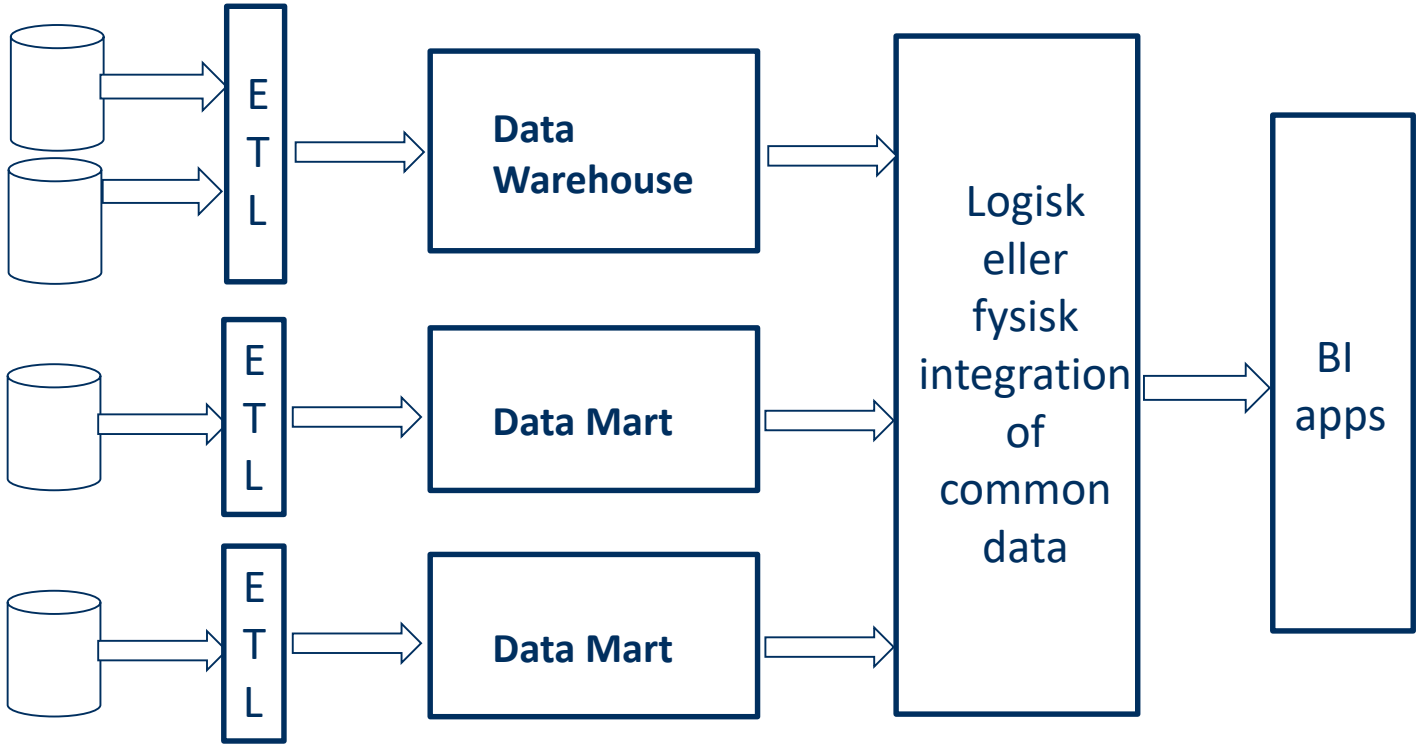
Stockholms universitet

Two additional types of Data Warehouse architectures

Centralized Architecture



Federated Architecture



Stockholms universitet

Which factor impact the selection of data warehouse architechture?

Which factors affect architecture selection?

Which of these factors affect?

- High interdependence between departments, units, employees?
- High urgency for a DW solution?
- Tasks in the organization are relatively less routine?
- DW as a short-term point solution or strategic infrastructure project?
- Low resource availability?
- IT staff with low perceived ability?
- Upper management sponsorship?

(Ariyachandra, T., & Watson, H. (2010). Key organizational factors in data warehouse architecture selection. *Decision support systems*, *49*(2), 200-212.)

Stockholms universitet

Data Lake and Data Warehouse

What is a Data Lake?

"A data lake is a **storage repository** that holds a **vast amount of raw data in its native format**, including structured, semi-structured, and unstructured data."

(https://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html)

What is a Data Lake”

A data lake is “a **methodology** enabled by a **massive data repository** based on **low cost technologies that improves the capture, refinement, archival, and exploration** of raw data within an enterprise.”

"**Yesterday's unified storage** is **today's enterprise data lake**"

(Huang Fang, Managing Data Lakes in Big Data Era: What’s a data lake and why has it became popular in data management ecosystem, The 5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, June 8-12, 2015, Shenyang, China.)

What is a Data Lake?

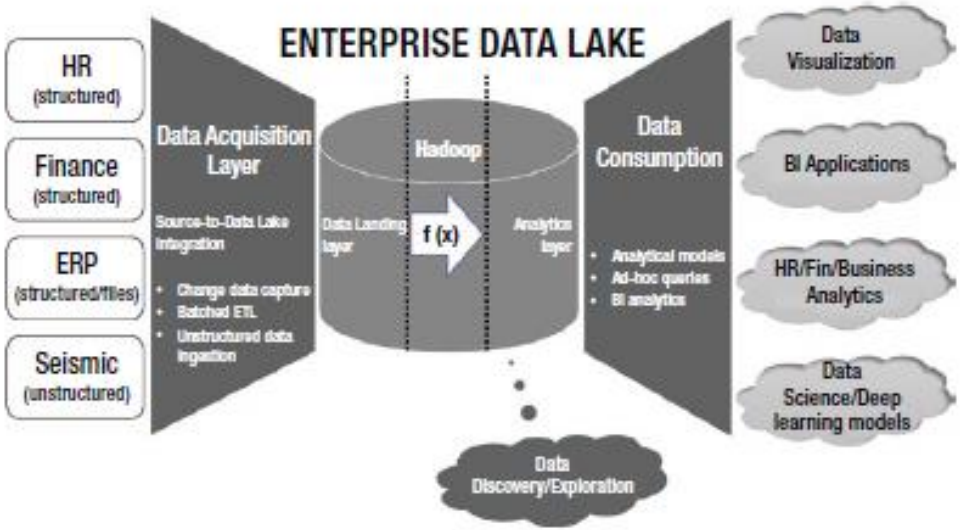
"The basic idea of Data Lake is simple, **all data emitted by the organization will be stored in a single data structure** called Data Lake."

"Data will be stored in the lake in their **original format**."

"Once data are placed in the lake, it's **available for analysis by everyone in the organization.**"

(Khine, P. P., & Wang, Z. S. (2018). Data lake: a new ideology in big data era. In *ITM Web of Conferences* (Vol. 17, p. 03025). EDP Sciences.)

An example of a Data Lake Architecture



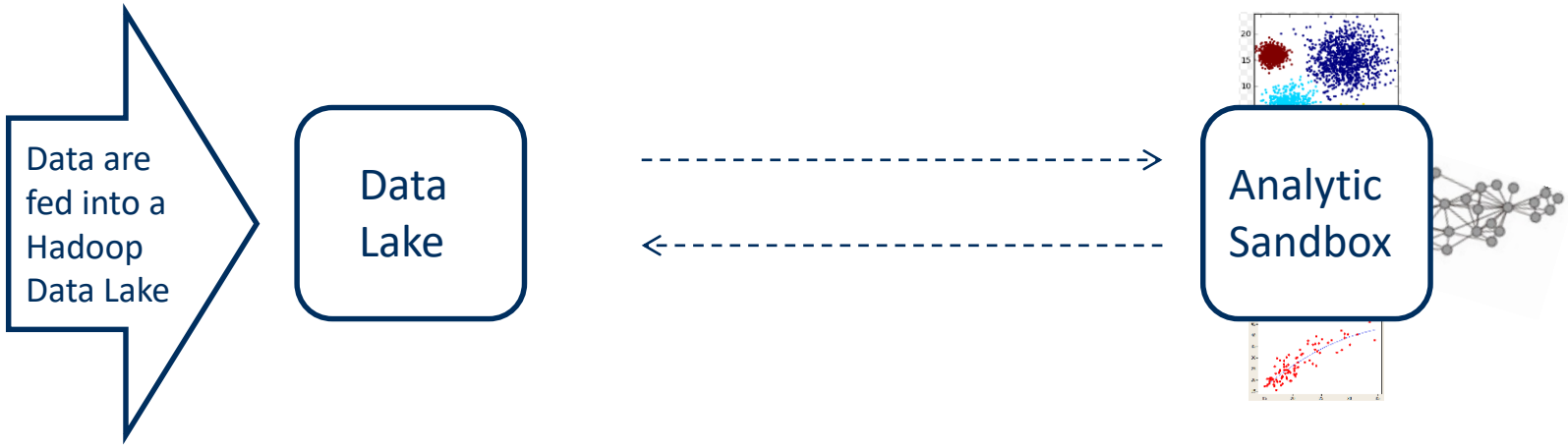
Gupta, S., & Giri, V. (2018). Introduction to Enterprise Data Lakes. In *Practical Enterprise Data Lake Insights* (pp. 1-31). Apress, Berkeley, CA.

Data Lake vs Data Warehouse

Stockholms universitet

Comparison	Data Warehouse	Data Lake
Data	Structured, processed data	Structured, semistructured and unstructured data, raw data, unprocessed data
Processing	Schema-on-write	Schema-on-read
Storage	Expensive, reliable	Low cost storage
Agility	Less agile, fixed configuration	High agility, flexible configuration
Security	Matured	Maturing
Users	Business professional	Data scientists

Data Lake



(Bill Schmarzo, Big Data MBA, Wiley, 2016)

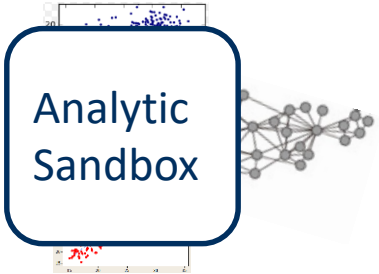
The analytics dilemma

How does an organization supports both the data warehouse/BI and an analytics environment?

DW/BI environment

- Static and pre-planned process – a production process
- Pre-planned use of data (via ETL)

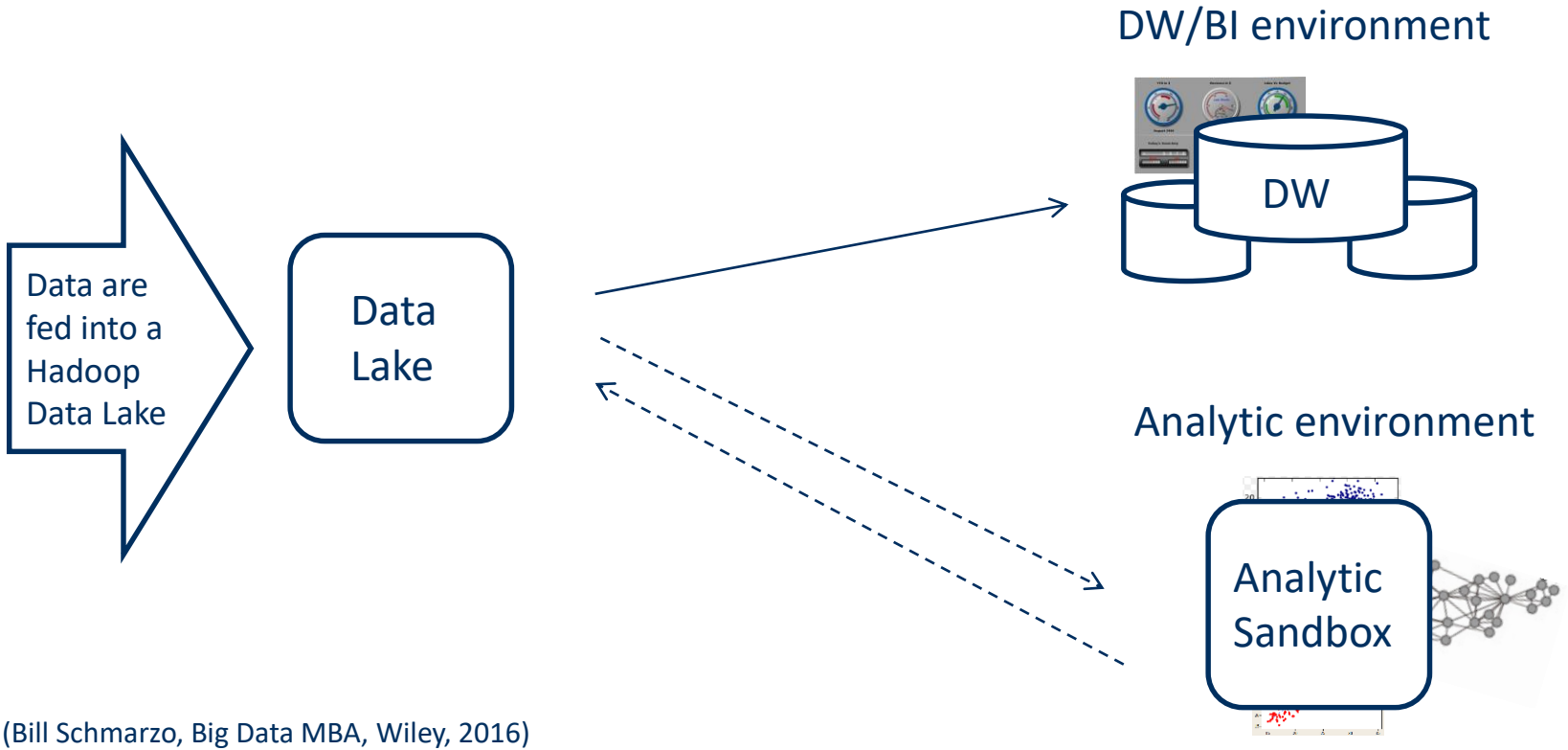
Analytic environment



- Exploratory, experimental process
- On the fly use of large amount of data
- Loosely governed

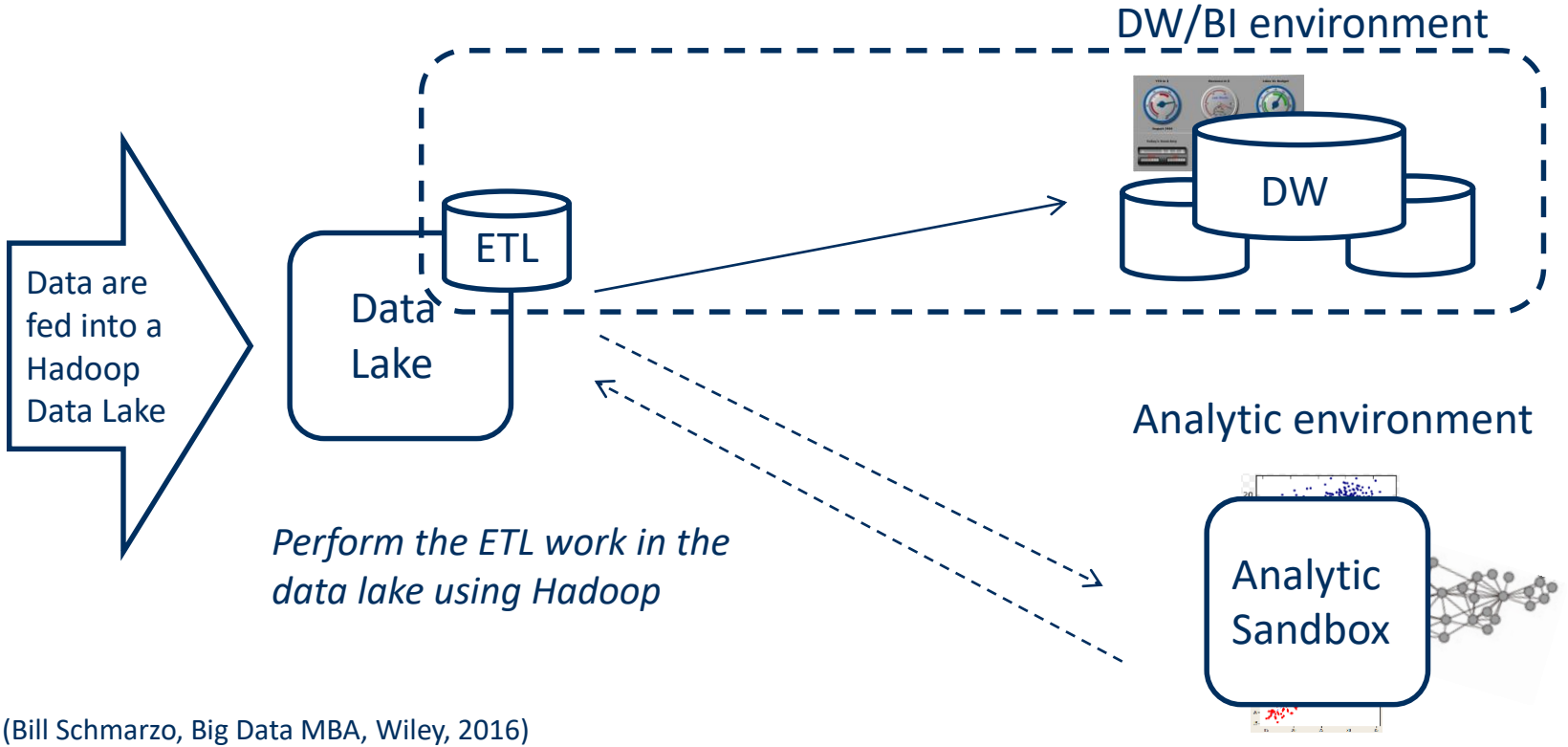
(Bill Schmarzo, Big Data MBA, Wiley, 2016)

Data Lake can support both environments



(Bill Schmarzo, Big Data MBA, Wiley, 2016)

Off-load ETL processes from DW



(Bill Schmarzo, Big Data MBA, Wiley, 2016)

Actions to exploit the value of Data Lake

- Action 1: Create a Hadoop-Based Data Lake
- Action 2: Introduce the Analytics Sandbox
- Action 3: Off-load ETL Processes from Data Warehouses

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

Lessons learned

- **There shall be one Data Lake, not several** – facititating sharing of the corporate data assets across the organization
- **Data governance is a life cycle, not a project**
- **Data Lake sits before your data warehouse, not after it**

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

Benefits of using a Data Lake

- **Eliminate data silos** – consolidating the data in one repository result in increased data use and sharing
- **Reduce cost for IT infrastructure** – according to Schmarzo, it is 20 to 50 times cheaper to store, manage and analyze a hugh amount of data in a big data/analytics environment, compared to store, manage and analyze data in a traditional data warehouse environments

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

Benefits of using a Data Lake

- **Provide a scalable, flexible and shared storage platform** - that support both BI, analytic and next generation environments
- **Store all data** - Store data even if the organization has not decided if it is going to use the data and how to use the data

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

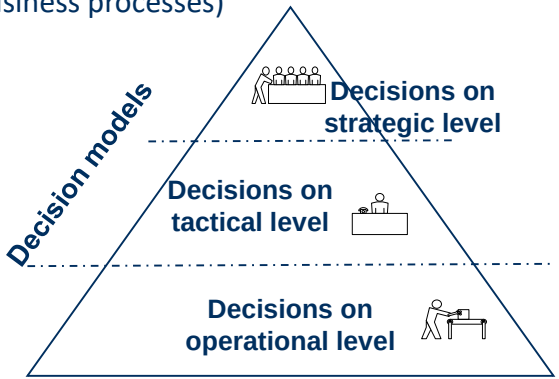
Stockholms universitet

Business Intelligence vs Big Data/Data science

Business intelligence – an overview

Goals
(vision, enterprise goals, objectives)

Means
(mission, strategies, tactics, business processes)



BI systems/tools

- Dashboard systems
- Analytical decision support tools
- Expert systems
- Data warehouse systems
- OLAP tools
- Data mining tools
- Reporting tools

System integration tools

- ETL tools

BI related methods

- Performance management
- Six sigma
- Balanced scorecard
- Lean
- Activity based costing
- Data and process mining
- Decision and risk analysis

Visualization

- Visualization techniques

Other areas related

- Business processes
- Business rules
- Models for decision
- Data governance
- Master data management
- Meta data management
- Org structures
- IT architecture
- Data repository
- Cloud computing
- Security
- Business case/ROI

Systems supporting decision making and data sources

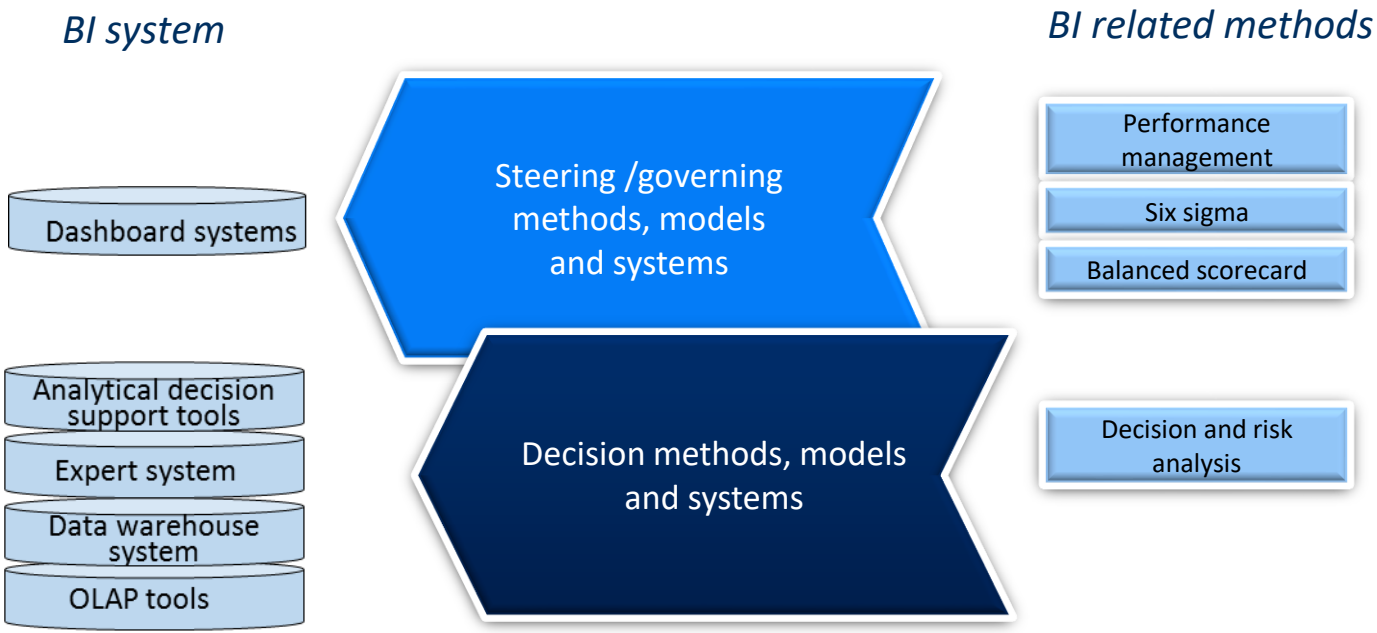
Operational systems: System that support the daily business, such as business systems (ERPs), BPM system, CRM system, etc

Business Intelligence – a definition

Business intelligence (BI) is an umbrella term that is commonly used to describe the technologies, applications, and processes for gathering, storing, accessing, and analyzing data to help users make better decisions.

Wixom and Watson, 2010

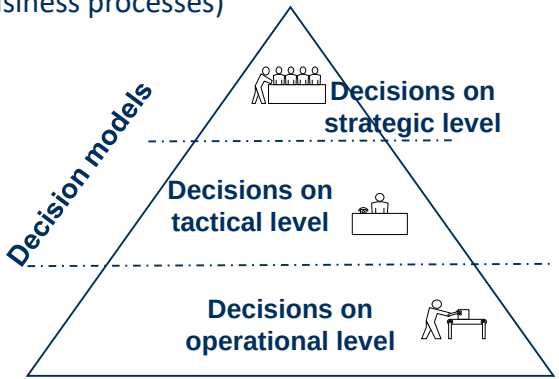
Two different approaches within BI



Big data/Data science – an overview

Goals
(vision, enterprise goals, objectives)

Means
(mission, strategies, tactics, business processes)



Big data related systems/techniques

- Hadoop
- MapReduce YARN/SPARK
- SAS/R

Additional data sources

- Social media
- Internet of Things
- Web
- Video/Camera
- Machines/Devices
- Maps
- Satellites

Big Data related methods for data analysis

- Data mining
- Machine learning
- Language technology
- Process mining
- Web mining

Processes for data analysis

- CRISP-DM
- Knowledge Discovery in Databases (KDD)

Visualization

- Visualization techniques

Other areas related

- Business processes
- Business rules
- Models for decision
- Data governance
- Master data management
- Meta data management
- Org structures
- IT architecture
- Data repository
- Cloud computing
- Security
- Business case/ROI

Systems supporting decision making and data sources

Operational systems: System that support the daily business, such as business systems (ERPs), BPM system, CRM system, etc

Data Science – a definition

Data science is about finding new variables and metrics that are better predictors of performance

Lewis (2004) Moneyball: The Art of Winning an Unfair Game

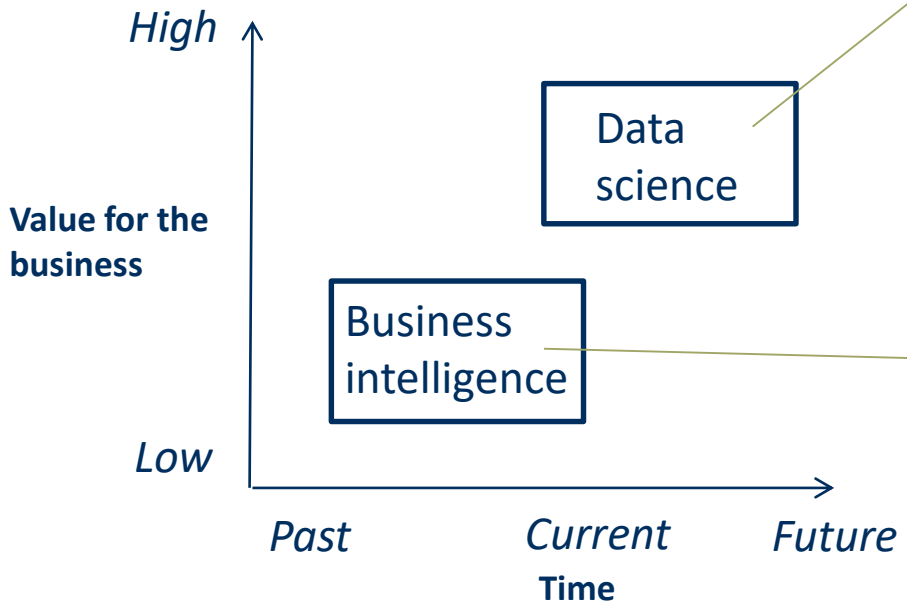
Big Data – a definition

Stockholms universitet

Big data is a key enabler of a new discipline called data science that seeks to leverage new sources of structured and unstructured data, coupled with predictive and prescriptive analytics, to uncover new variables and metrics that are better predictors of performance

Bill Schmarzo, 2016, based on Lewis (2004) Moneyball: The Art of Winning an Unfair Game

Business intelligence vs. Data science



- Predictive analytics
- Prescriptive analytics
- On the fly use of large amount of data
- Exploratory, experimental process
- Focus on pattern, correlations
- Data model – schema on query/read
- Data quality – good enough, propabilities

- Desciptive analytics
- Pre-planned use of data (via ETL)
- Static and pre-planned process
- Focus on trends, use of KPIs
- Data model – schema on load
- Data quality – high quality, single source of truth

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

Business intelligence vs. Data science

The differences between BI and Data science

- The questions are different
- The analyst characteristics are different
- The analytic approaches are different
- The data models are different
- The views on business are different

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The questions are different

Business Intelligence

- ***Focus on descriptive analytics:*** ”What happend?” type of questions: How many units of products X did we sell in Jan 2017

Data Science

- ***Focus on predictive analytics:*** ”What is likely to happend? type of questions: How many units of products X will we sell in Jan 2018?
- ***Focus on prescriptive analytics:*** ”What should we do?” type of question: How many components A, B, C should I order to support the sales of product X?

To answer the predictive and prescriptive questions, the data scientist build analytic models in order to quantify cause and effect relationships

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The analysts' characteristics are different

The attitude and work approach among BI analysts and data scientists differs:

AREA	BI ANALYST	DATA SCIENTIST
Focus	Trends, KPIs	Pattern, Correlations, Models
Process	Static	Exploratory, experimentation, visual, agile
Data sources	Pre-planned, added slowly	On the fly, as needed
Transform data	Carefully planned	On demand, enrichment
Data quality	Single version of truth	"Good enough", probabilities
Data model	Schema on load	Schema on query
Analysis	Descriptive	Predictive, prescriptive

(Bill Schmarzo, Big Data MBA, Wiley, 2016, based on EMC)

BI vs. Data science: The analytic approaches are different 1(2)

Stockholms universitet

BI analytic approach

- Step 1: Pre-build a data model (Schema on load)
- Step 2: Make use of (visualisation) tools that automatically generated SQL commands from drag and drop using attributes/dimensions/facts
- Step 3: Make use of the generated SQL commands to generate reports automatically

Data science analytic approach

- Step 1: Define hypothesis (test/prediction)
- Step 2: Gather data (Data Lake)
- Step 3: Build data model (Schema on query)
- Step 4: Build analytic models (SAS, R)
- Step 5: Evaluate model goodness of fit

What would happend if you want to add new data into the data warehouse/BI environment? What is the benefit with schema on the load?

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The analytic approaches are different 2(2)

Schema on load

– a schema must be built prior to loading data into the data warehouse

Schema on query/read

– a schema is defined as needed based on data being used, and the data scientist will go through different versions of the schema until finding a schema that support the analytical model

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The data models are different

Business Intelligence

- Schema on load
- Often star join schemas – multiples tables, many (comparable slow) joins

Data Science

- Schema on query/read
- Often flattened tables – few (flattened) tables with a lot of data, few joins

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

BI vs. Data science: The views on business are different

Business Intelligence

- Aggregated data on business entities, such as customers, products

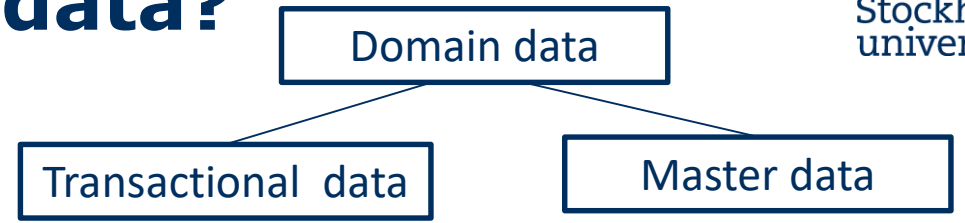
Data Science

- Build **analytic profiles** on each business entity. Example of business entities are customers, partners/suppliers, devices, machines
- For exampel, analytic profiles for customers could be used for managing customer rentension/attrite rate

(Bill Schmarzo, Big Data MBA, Wiley, 2016)

Master Data

What is master data?



- **Domain data** – is data about the business and the area of the business, and it can be devided in:
- **Transactional data** – is data about transactions, such as order request, ATM machine transcations, mobile calls. Transaction data represent different kinds of business events in the organization. Transactional data must be related to master data to receive a meaning
- **Master data** – is data about central business entities that are be used in several processes and systems, such as customers, suppliers, products, assets, locations

What is master data?

- Master data has the following characteristics:
  - **Master data are not transactional** – but master data are linked to transactions to give the transaction meaning
  - **Master data have meaning independent of transactional data** –master data are linked to transactional data but they have meaning also without the transactional data
  - **Master data have known provenance/place of origin** – that is, you know where the data come from

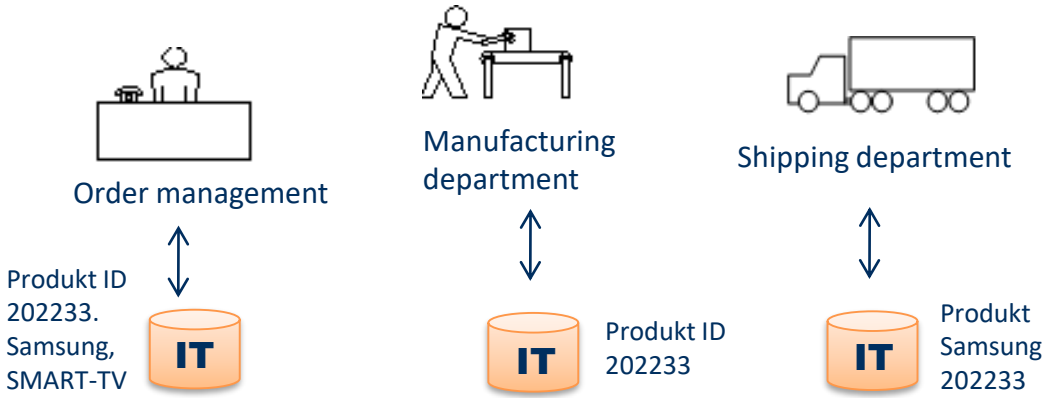
[Graham A. (2015) Mastering You Data, Koios Associated Ltd]

Stockholms universitet

Master Data Management System

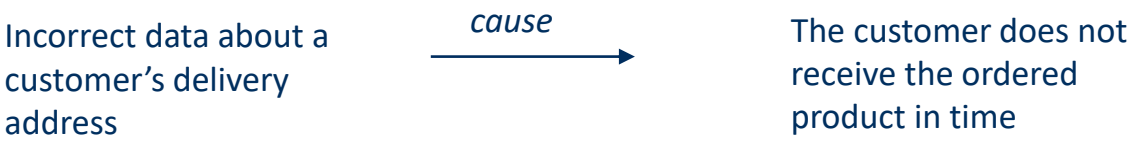
Why a master data management system (MDM-system)?

- **Problems addressed by a MDM-system:** When you have several IT systems, there may exist incorrect and/or inconsistent data about customers, suppliers, and products in the systems. For example, a name of a customer or a product can be incorrectly spelled in some systems, a customer can have two different delivery addresses in two different systems



Why a master data management system (MDM-system)?

- **Business consequenses of incorrect data in different IT system**
  - Customers may not receive in time what has been ordered
  - If the same product has different spellings or there are incorrect info about the customer – the trust of the company can get damaged
  - Reports to governmental organizations may be harder to perform



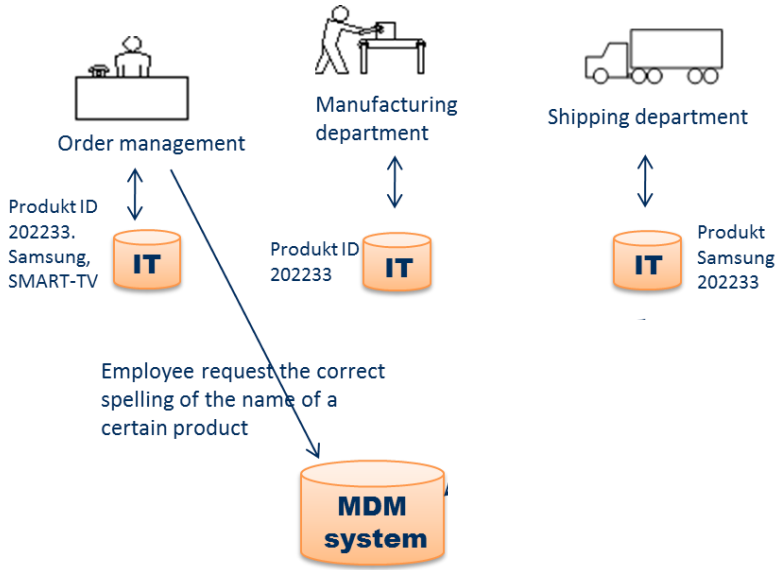
What is a MDM system?

- **A MDM-system –** is a system that contain correct info about customers, suppliers and products, etc
- In the MDM system, you can find the correct spelling, abbrevations, descriptions of customers, suppliers and products
- In the MDM system, you can also find allowed categorizations, and hierarchies of categorizations (for example the hierarchy: food - fruit – berry - strawberry eller the hierarchy: food – perishable - strawberry)

”One version of the truth”

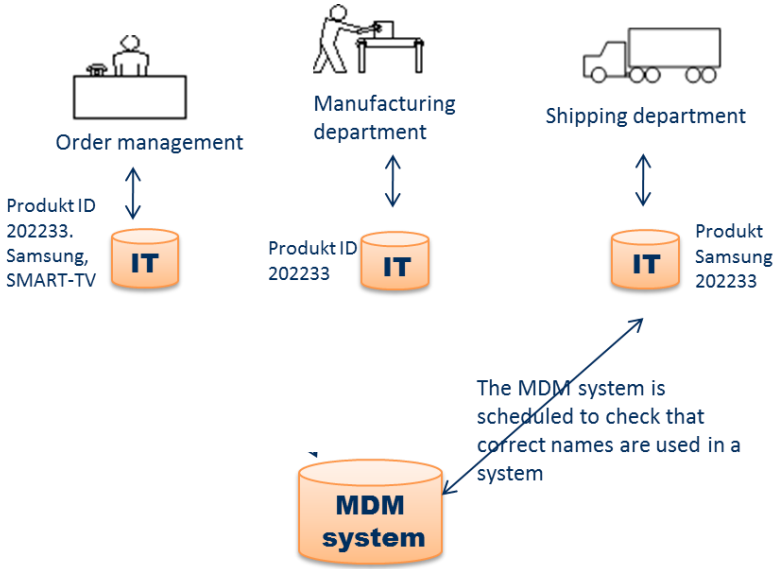
Hur to use a MDM system?

- Employees can request correct spelling and attributes of data element

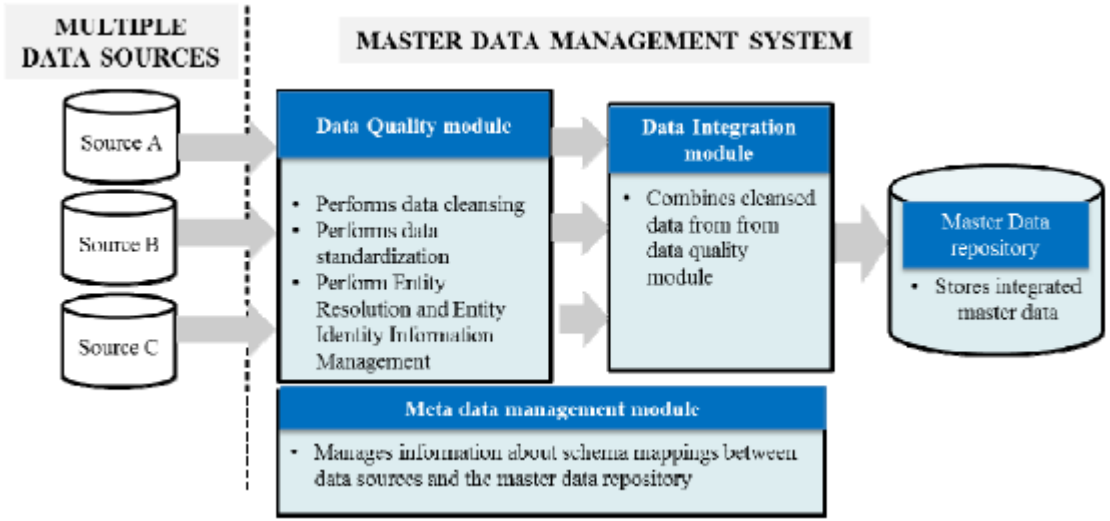


Hur to use a MDM system?

- The MDM system could be used for automatically correct info in the operational systems

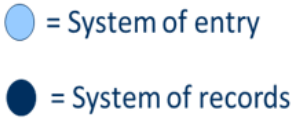


Architecture of a MDM system



Haneem, F., Ali, R., Kama, N., & Basri, S. (2017, July). Resolving data duplication, inaccuracy and inconsistency issues using Master Data Management. In *Research and Innovation in Information Systems (ICRIIS), 2017 International Conference on* (pp. 1-6). IEEE.

Define where your single source of truth for data elements are?



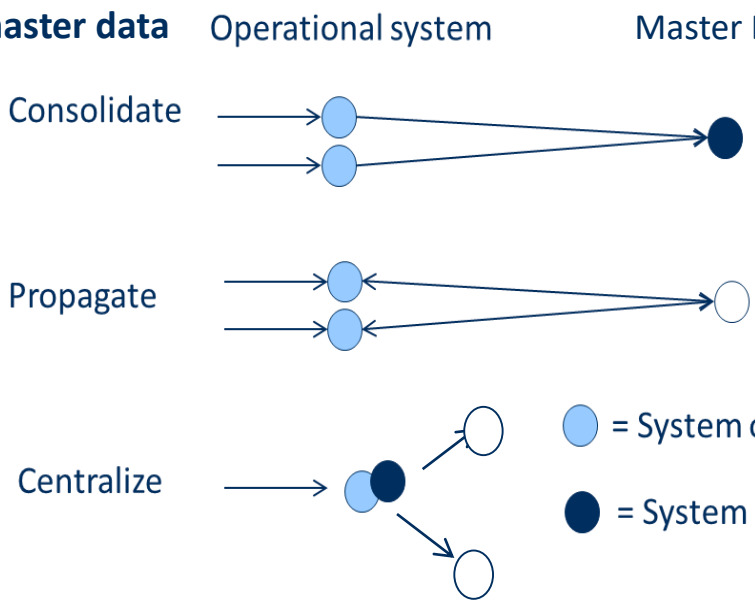
= place where the master is created and maybe also maintained

= master data = golden copy = single place where master data i garanteed to be accurate and up to date

[White C. (2007) Using master data i business intelligence, BI research]

Define where your single source of truth for data elements are?

Approaches for managing master data



Syncronize master data changes between systems so master data in all systems are kept consistens

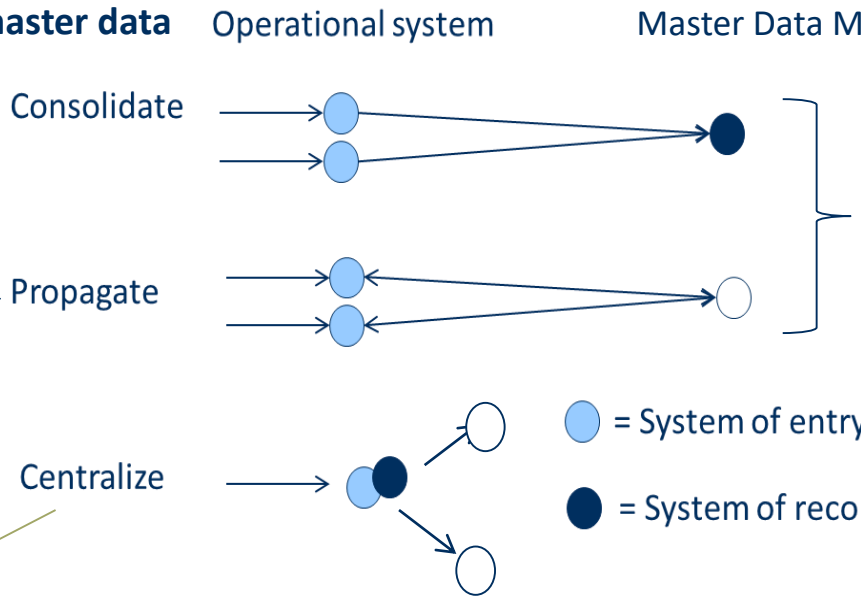
= System of entry = place where the master is created and maybe also maintained

= System of records = master data = golden copy = single place where master data i garanteed to be accurate and up to date

[White C. (2007) Using master data i business intelligence, BI research]

Define where your single source of truth for data elements are?

Approaches for managing master data



Syncronize master data changes between systems so master data in all systems are kept consistens

Another approach could also be to use both approaches = consolidate and propagate

= System of entry = place where the master is created and maybe also maintained

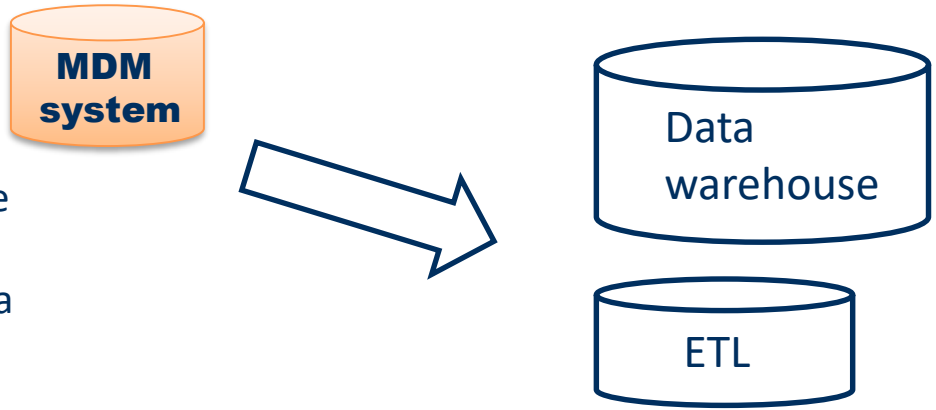
= System of records = master data = golden copy = single place where master data i garanteed to be accurate and up to date

Often the long term goal

[White C. (2007) Using master data i business intelligence, BI research]

Hur to use a MDM system?

How could the MDM system support a Data Warehouse?



Hur to use a MDM system?

How could the MDM system support a Data Lake?