Data analytics

(basic concepts and traditional DWH architecture)

Jana Kostičová

Outline

1. Data analytics concepts
2. Traditional data warehouses (BI application)

3. Modern data analytics
4. New technologies
5. New architectures
6. Summary

Data analytics concepts

Data analytics definition

“Data analytics is any activity that involves applying an analytical process to data to derive insight from the data.” – Ridge (2015)

“Data analytics is the science of integrating heterogeneous data from diverse sources, drawing inferences, and making predictions to enable innovation, gain competitive business advantage, and help strategic decision-making.” – Gudivada (2017)

“Data analytics converts raw data into actionable insights. It includes a range of tools, technologies, and processes used to find trends and solve problems using data.” – AWS Cloud Computing Concepts Hub

data analytics vs data analysis

Data analytics - applications

Traditional – analytics primarily provides insights to support human analysis and decision-making

- Business intelligence (BI) - DWH architecture
- Science (bioinformatics, astronomy, climate research, …)
- Healthcare
- Social sciences
- …

Modern – analytics increasingly becomes embedded in products, services, and automated processes

- Recommendation systems
- Dynamic pricing
- Real-time monitoring and prediction
- Fraud detection (real-time / predictive analytics)
- Personalized services
- …

Both traditional and modern applications of data analytics are important today.

Four types of data analytics

1. **Descriptive analytics:**

   = to summarize and describe data. It is often used to identify trends and patterns in data.

   RETAILER: What have been our best-selling products in last 6 months? What has been the demographics of our customers in last year? (based on sales data)

2. **Diagnostic analytics**

   = to understand the cause of a problem. It is often used to identify root causes of issues.

   CAR MANUFACTURER: What has been the major cause of defects in SUV product line within last 3 years? (based on various data: test results, claims, product data, component data, supplier data, …)

Four types of data analytics

3. **Predictive analytics:**

   = to predict future outcomes. It is often used to forecast sales or customer churn.

   BANK: What is the likelihood of customers defaulting on loans? (based on existing customer/loan data)

4. **Prescriptive analytics:**

   = to recommend a course of action. It is often used to optimize processes or make decisions

   TRANSPORTATION COMPANY: How can we optimize routes and reduce delivery times? (based on existing route data)

These four types of data analytics are often used in combination to provide a comprehensive understanding of data.

General requirements of data analytics

- Integrate data from multiple heterogeneous data sources
- Store potentially large volumes of data and provide scalability
- Provide historical data
- Organize data for analytical queries
- Support efficient execution of complex analytical queries (aggregations, filtering, joins, …)

(all are relevant to BI; for other applications only a subset might be relevant)



With technology advances and new use cases, the new requirements appear:

- Integrate broader range of relevant data sources — including semi-structured and unstructured data, sensors, web data, etc.
- Handle larger data volumes
- Support larger-scale and more sophisticated analytics
- Satisfy stricter latency requirements — including real-time and streaming scenarios

References

- Ridge (2015): Guerrilla Analytics: A Practical Approach to Working with Data

  Chapter 1 available at sciencedirect.com

- Gudivada, V. N. (2017). Data Analytics: Fundamentals. In M. Chowdhury, A. Apon, & K. Dey (Eds.), Data Analytics for Intelligent Transportation Systems, pp. 31–67. Elsevier.

  Available at academia.edu

Traditional data warehouses (Business Intelligence)

Business Intelligence (BI)

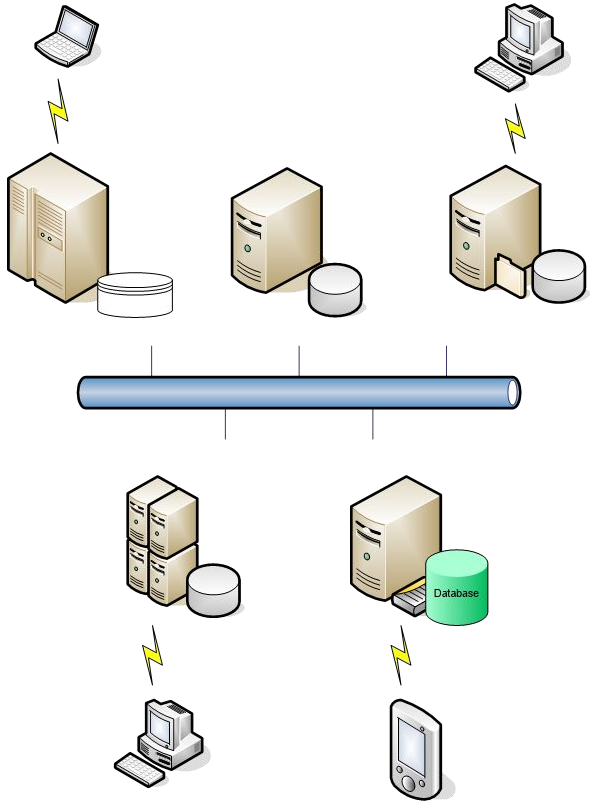
Business Intelligence - providing insights to support business decision-making

We will look at BI application in more detail, because it provides the classical setting in which traditional data warehousing concepts emerged.

Corporate IT architecture (1)

Billing

- Invoicing, payment processing, revenue collection



HR - Human Resource System

- Employees, positions, salaries, vacations, trainings, career progress, attendance, utilization, benefits,...

Content Management System (CMS)

- Intranet, information for employees
- Internet, web portal, information for customers

Customer Relationship Management (CRM)

- Interactions with customers / potential customers - sales, marketing, customer service

… and more

Corporate IT architecture (2)

Operational systems (production systems, transaction systems)

- Store data needed to support the organization's day-to-day operations
- Support operational business processes
- OLTP (Online Transaction Processing)
  - Focused on transaction processing
  - Optimized for frequent, relatively small reads and writes / inserts, updates, and deletes
  - Typically require low latency and high concurrency
- May include legacy systems as well as modern applications
- Different operational systems store data from different business/application perspectives
  - Application-oriented data

Analytical queries vs operational systems

- Did the effectivity of salespersons (the number of products sold per month) increased after a specific training?

Analytical queries vs operational systems

- Did the effectivity of salespersons (the number of products sold per month) increased after a specific training?



HR system, CRM system

Analytical queries vs operational systems

- Did the effectivity of salespersons (the number of products sold per month) increased after a specific training?

  ⇨ HR system, CRM system

- What is the relationship between the length of employment of a given salesperson and the number of products sold per month?

Analytical queries vs operational systems

- Did the effectivity of salespersons (the number of products sold per month) increased after a specific training?

  ⇨ HR system, CRM system

- What is the relationship between the length of employment of a given salesperson and the number of products sold per month?

  ⇨ HR system, CRM system

Analytical queries vs operational systems

- Did the effectivity of salespersons (the number of products sold per month) increased after a specific training?

  ⇨ HR system, CRM system

- What is the relationship between the length of employment of a given salesperson and the number of products sold per month?

  ⇨ HR system, CRM system

- Which customer segment has the largest amount of unpaid invoices?

Analytical queries vs operational systems

- Did the effectiveness of salespersons (the number of products sold per month) increased after a specific training?

  ⇨ HR system, CRM system

- What is the relationship between the length of employment of a given salesperson and the number of products sold per month?

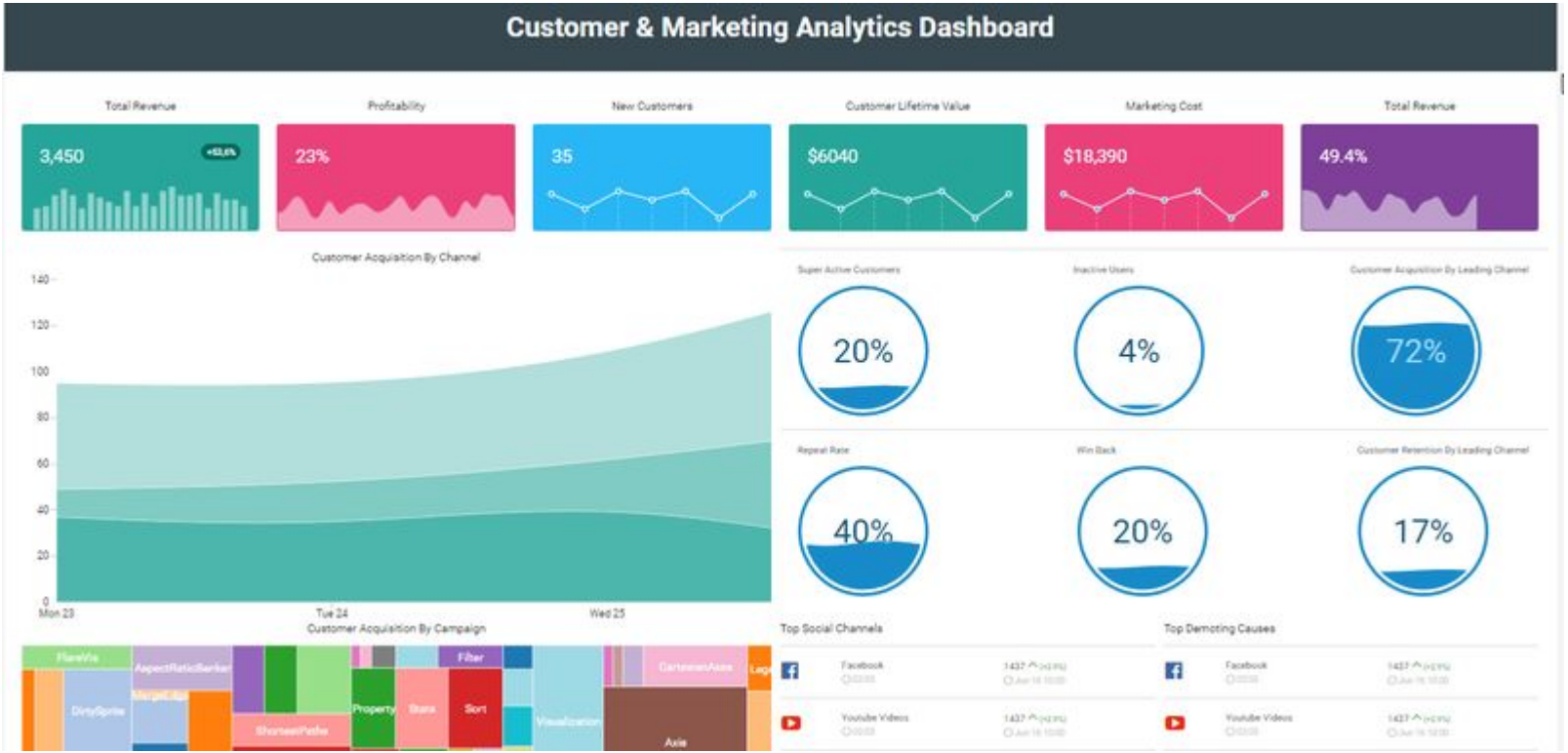
  ⇨ HR system, CRM system

- Which customer segment has the largest amount of unpaid invoices?

  ⇨ Billing system, CRM system

Analytical queries generally need data from different operational systems.

Analytics Dashboards



Examples

- Power BI demo: LINK
  - Market Share report, comparison to competitors' portfolios
  - Fictional company VanArsdel
- Google Analytics demo accounts: LINK
  - Traditional web site analytics

Enterprise data

There is hidden wealth in enterprise data

- How to gain it?
- How to use it?



We need an unified view of the subjects (such as Customer, Product, Employee)

Enterprises can make better decisions on how to achieve its ultimate goals (mostly increasing the profit)

- Decision support systems (DSS)

Operational databases and data analytics

Data in operational systems is not stored in accordance with requirements of BI

Data integration
- Data is distributed across multiple operational systems
- Systems may use different formats, identifiers, definitions, and levels of detail
- BI often requires a consistent, integrated view of the organization

Data volume
- Not optimized for storing and analyzing large amounts of historical data

Complexity
- Optimized primarily for transactional queries rather than complex analytical queries

Performance impact
- Running analytical queries can impact performance for transactional workloads

Data design
- Application-oriented data, it can be challenging to perform analytical queries (including joins across multiple tables)

Before data warehouses

Ad-hoc methods used for data analytics:

Manual methods:

- E.g., using spreadsheets, manual data exports and integration
- Long lasting data preparation and less time for analysis itself
- Error-prone (human factor)

Custom built-in applications:

- Cobol / Fortran / C, …
- Inflexible, difficult to handle new or changed requirements (both changes driven by business requirements or by changes in data sources)

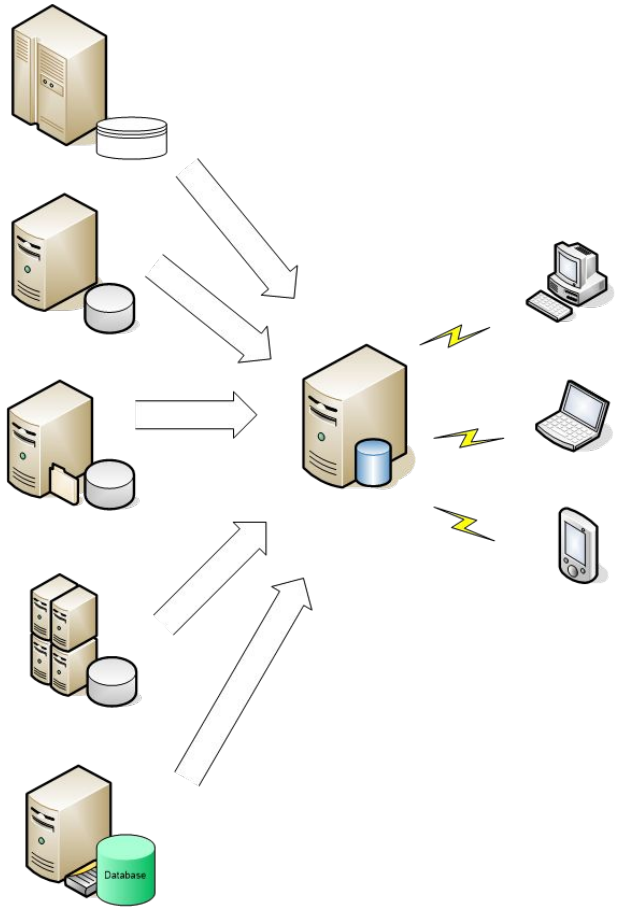
Issues arises especially with larger data volumes, more complex analyses and the need to integrate many systems.



** Designed by Wannapik*

Data warehouse

- Systematic, modular approach
- Data preparation separated from the data analysis
- Data preparation is (mostly) automatic
  - Short-term process
  - The error rate is minimized
- Data analysts get pre-prepared data and they can focus solely on analysis



Data warehousing era (1990s-2000s)

Data warehouse (**DWH**, DW, EDW) =

“a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process” – Inmon (1992)

- Bill Inmon: Building the Datawarehouse, 1992 (4th edition - 2005 [1])
- Ralph Kimball: The Data Warehouse Toolkit, 1996 (4th Edition - 2013 [2])

“Fathers” of data warehousing

Subject-oriented

- Data is organized around major entities of interests of an organization (e.g., Customer, Product, Supplier)

Integrated

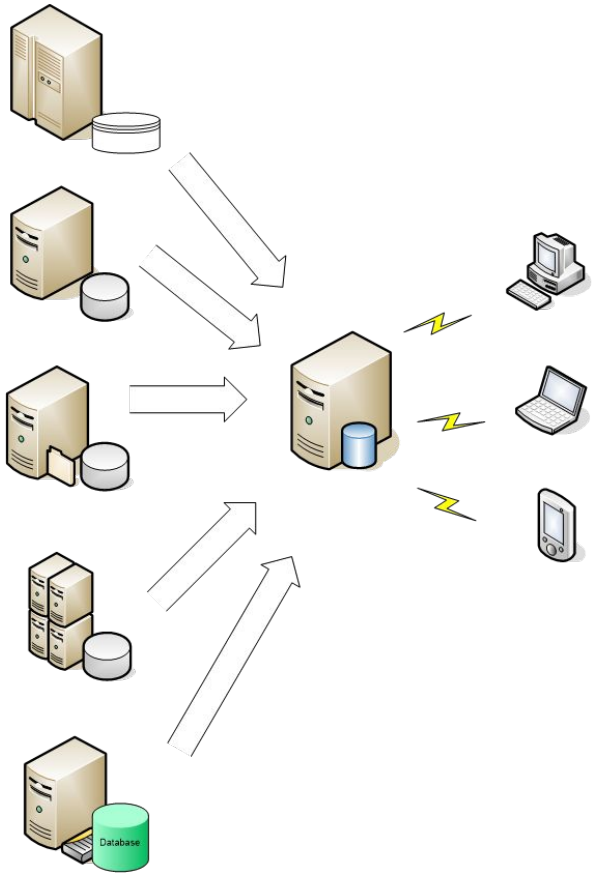
- Integrated data from various sources: operational databases, external databases, even meta-data

Time-variant

- Multiple years of data (even > 10 years), data is identified by timestamp.

Non-volatile

- Data is stable. Data is added, but usually not updated or deleted.



Characteristics of traditional DWHs

Processing type

- Batch processing – data were typically loaded periodically (e.g., daily, weekly, or monthly)

Technologies:

- Initially - **relational databases** running on individual servers
- Later also **relational MPP (Massive Parallel Processing) DWHs** running across multiple servers, e.g., Teradata

Environment:

- Primarily on-premise large-scale systems, built mostly by large corporations with significant IT budgets (SK - banks, telco operators)

Analytics types:

- Focused mostly on descriptive / diagnostic analytics
- Data mining techniques also enabled some predictive analytics

Are traditional data warehouses obsolete?

No - they still remain a fundamental component of many organizations' data architectures.

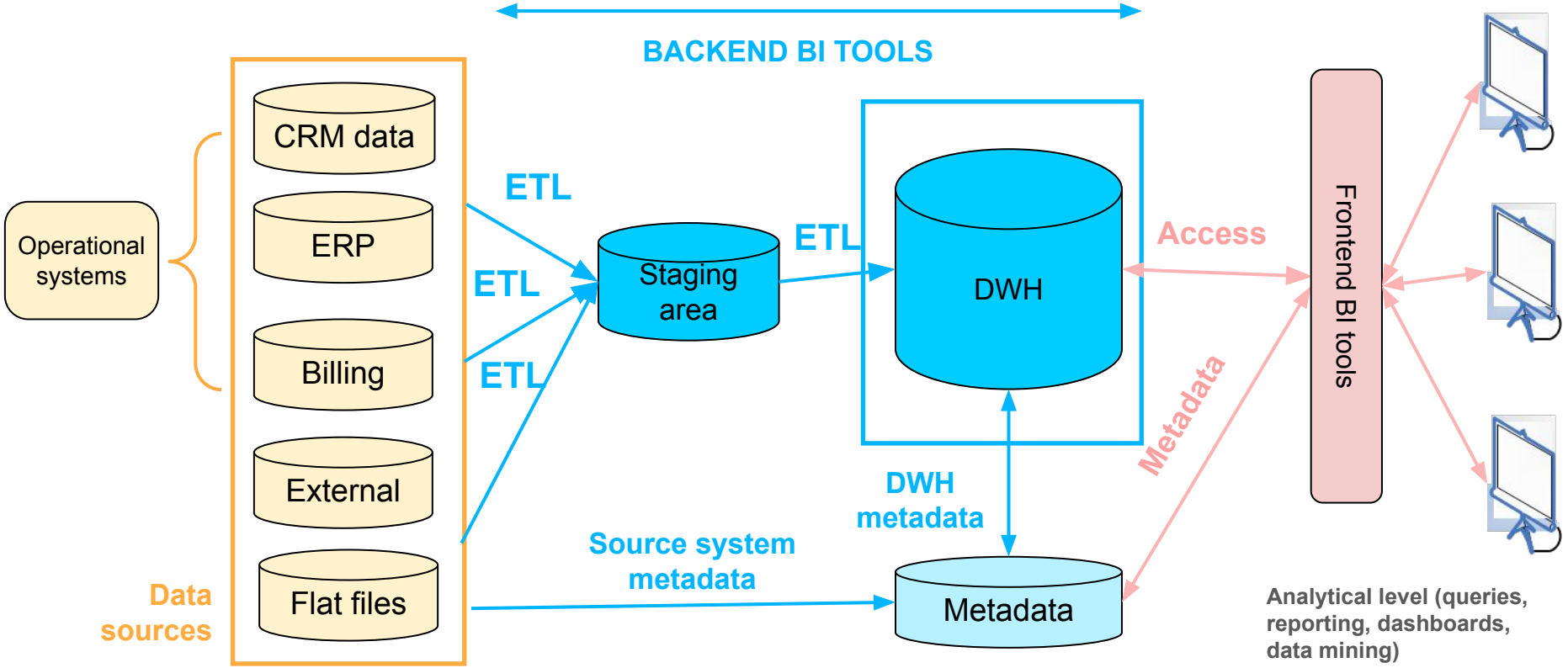
But, they are evolving

- **to integrate** with new architectures, and
- **to coexist (and cooperate)** with new technologies.

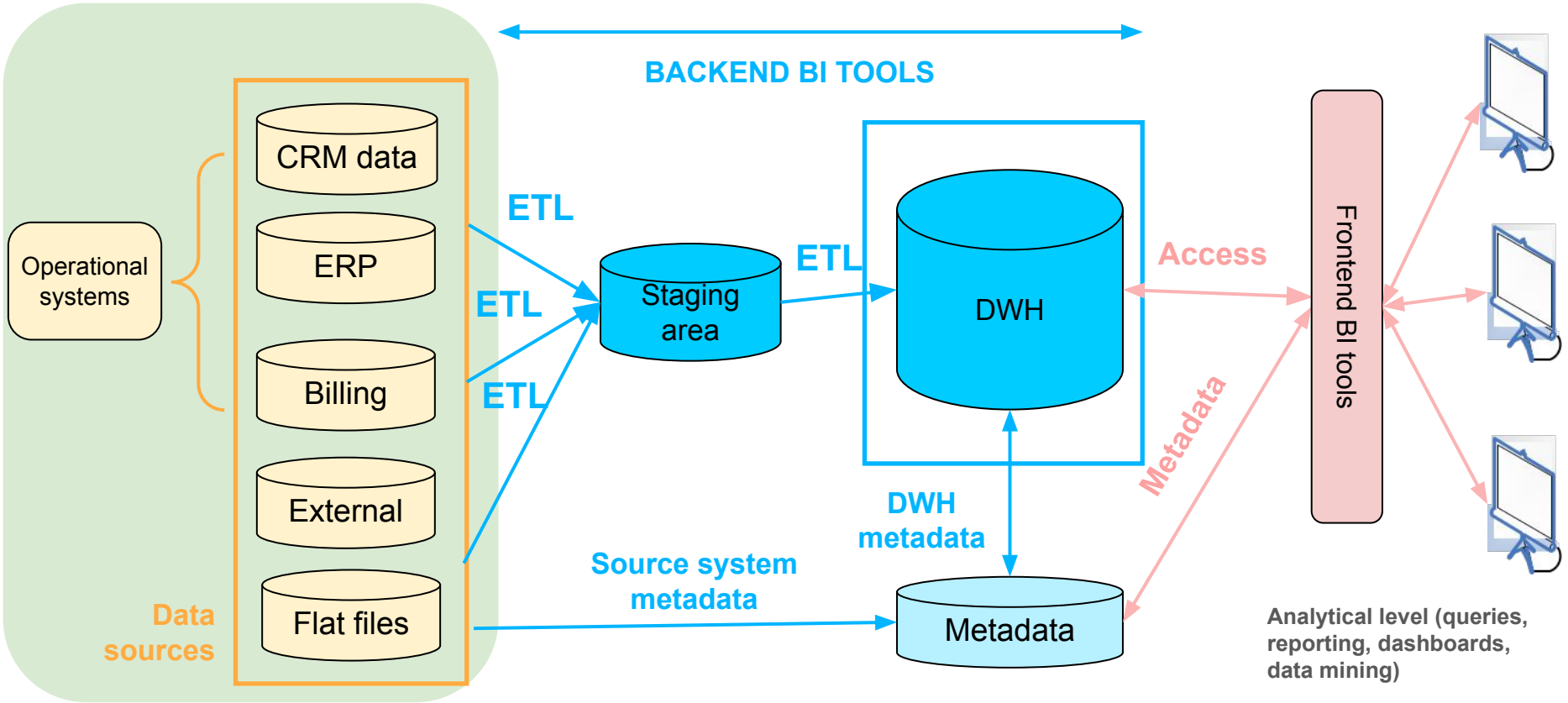
In addition, they provide an important conceptual foundation for understanding core data management and analytics issues.



Traditional BI architecture



Traditional BI architecture - data sources



Data sources

- Contain enterprise data
- Various form: databases, files (mostly XML, CSV, …)
- Operational systems, external data

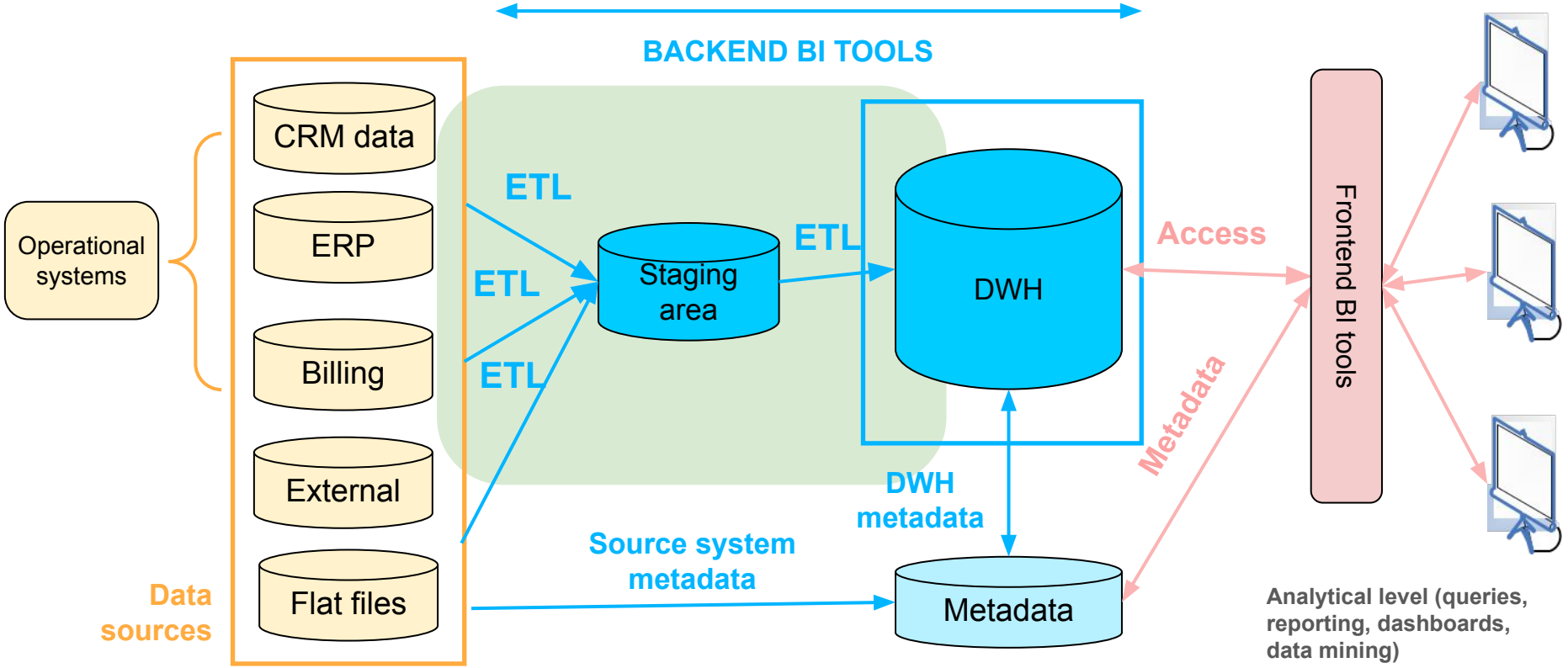
Problems with operational systems

- They have to be up and running “all the time”
  - We can't burden them with data extraction at any time (usually ok during the night)
- Data is not stored in accordance with BI requirements

External data

- External DBs or external files
- Examples of external files: addresses according to the regions (LINK), postal codes (LINK)

Traditional BI architecture - ETL



ETL (Extract - Transform - Load)

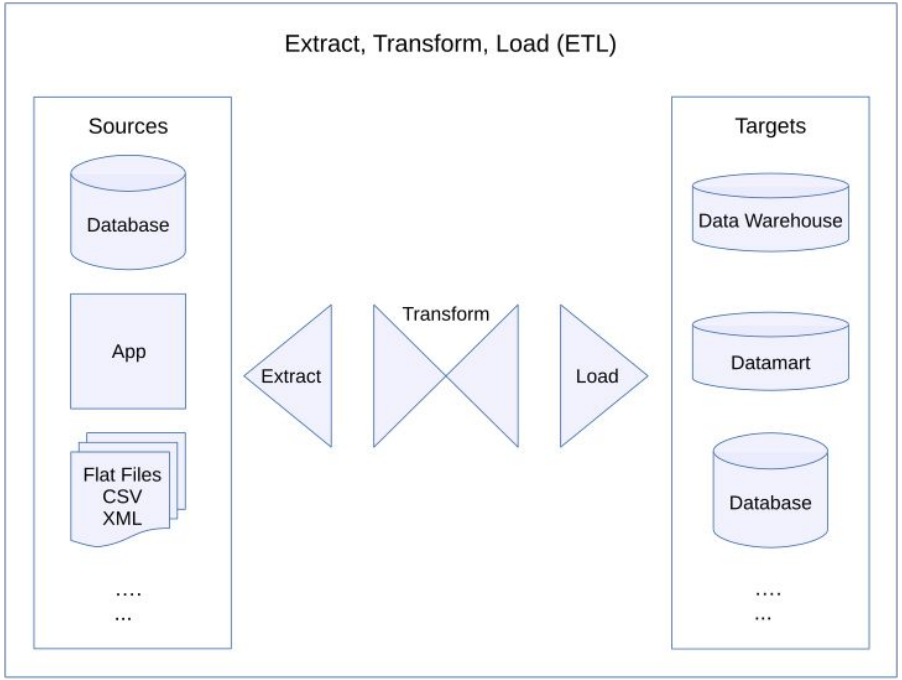
= a process transporting (and transforming) data from the data sources to DWH

Extract: Data is extracted from source systems in its raw format.

Transform: Data is transformed to match the data warehouse schema and requirements.

Load to DWH: The transformed data is loaded into the final data warehouse tables.

Daily / weekly / monthly / … loads



Extract

- Form of data storage (DB, CSV files, …)
- Form of access (DB native, ODBC, app, …)
- Protocols (HTTP, FTP, SFTP, ..)
- Full load vs incremental load (CDC - change data capture)
- Automatic loads vs manual loads

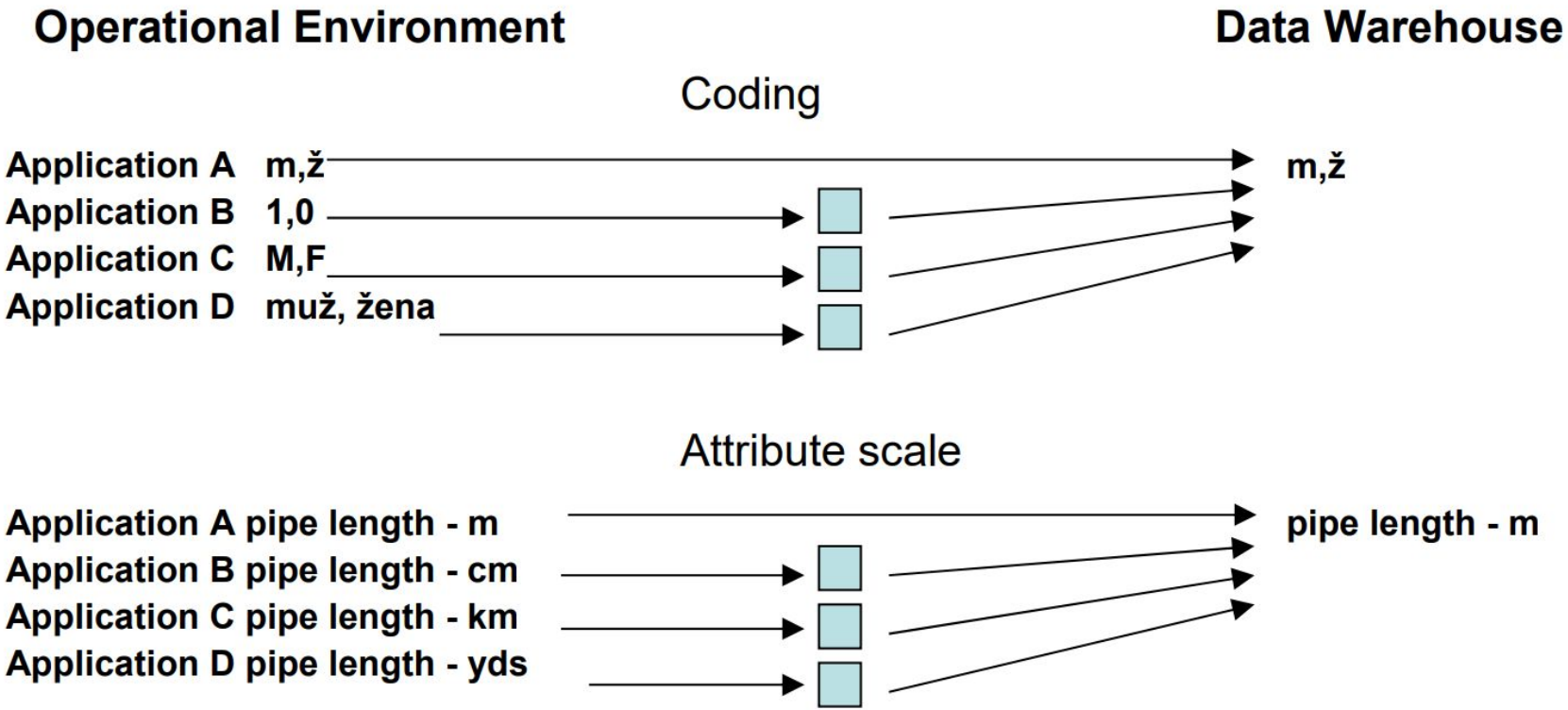
Transform

- Data validation - checking for accuracy, completeness, consistency
- Data cleansing - handling missing values and inconsistencies, deduplication, etc.
- Data standardization - consistent format for dates, currencies, etc.
- Data masking - protecting sensitive data by replacing or obfuscating certain values
- Data enrichment - add calculated fields, derived attributes
- Data aggregations
- Data transformations - filtering, sorting, joining

Load

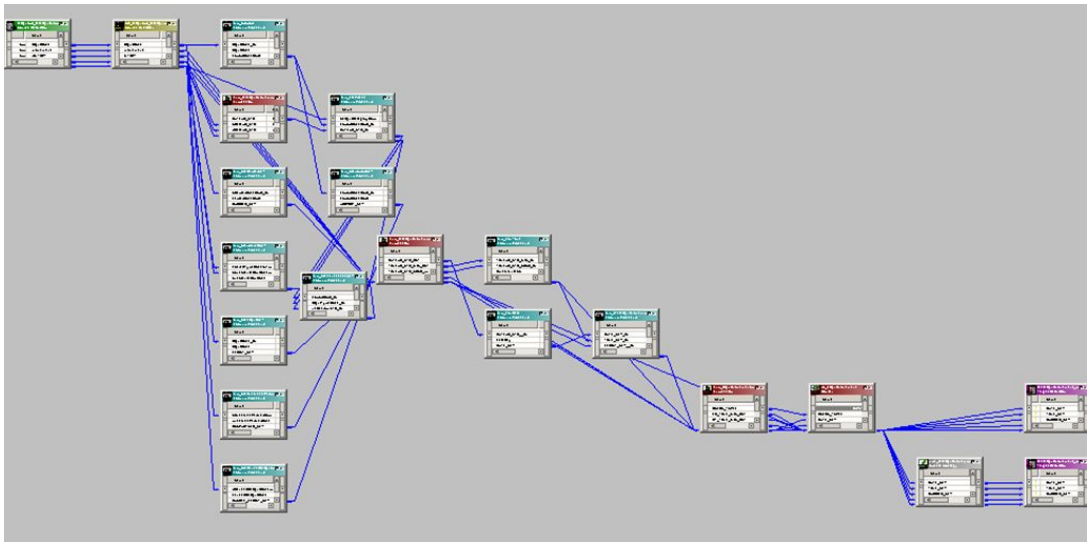
- Target identification
- Loading mechanism: bulk for large datasets, incremental, CDC,...

Example - data standardization



ETL - implementation

- Hand-coded scripts vs ETL tool
- Scheduler: daily, weekly, monthly, event-driven, manual
- Exception handling
- Task recovery and restart
- Metadata
- Security (sensitive data)



Example: Sorting data in Informatica PowerCenter: LINK

Staging area (stage, STG)

Stage is not necessarily needed in smaller DWHs

= temporary data storage, “snapshot” of source systems

Main goal

- to quickly load large volumes of raw data from source systems

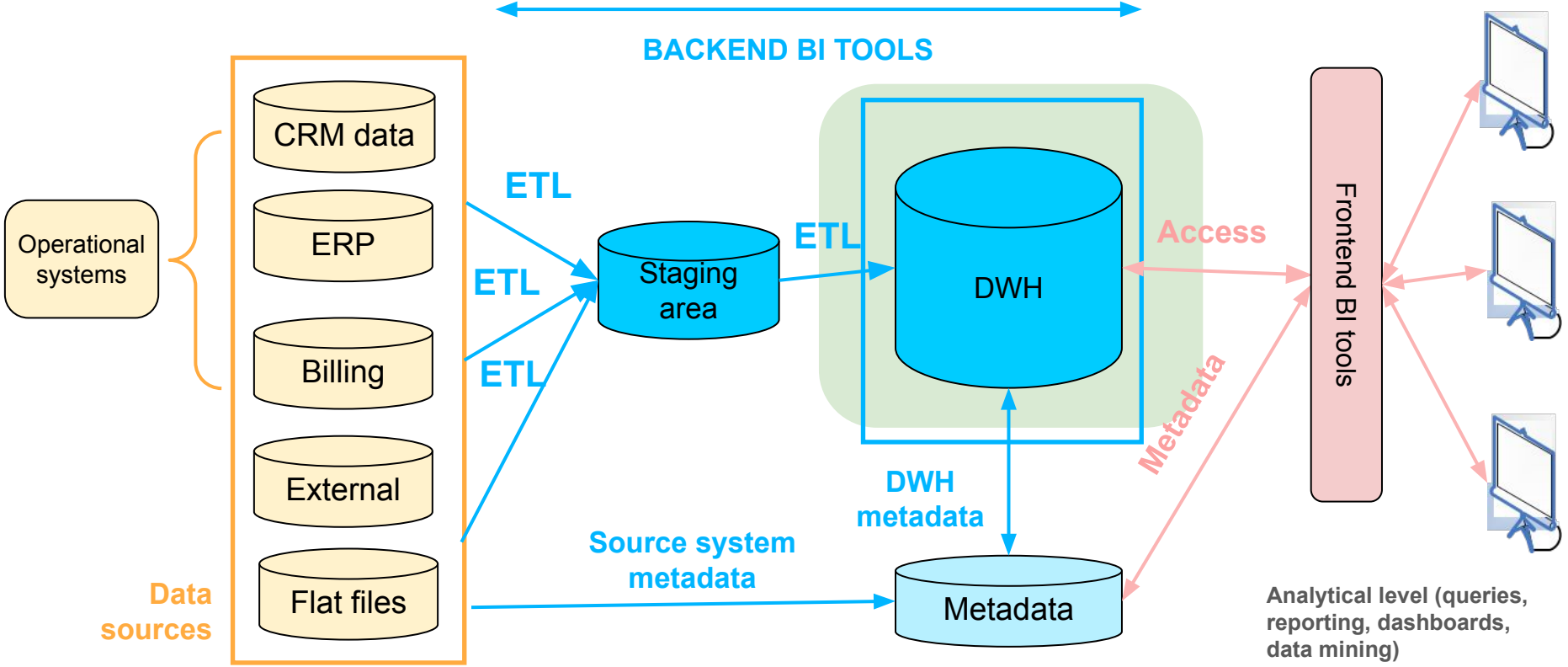
Benefits

- Handles inaccessibility of source systems
- Source systems are not burdened with complex operations
- Easy data reloads

Properties of STG tables

- They should have the same structure as the tables in the source systems
- Typically no referential integrity
- Some lightweight transformations can be performed before loading data to stage such as
  - Data extraction and filtering, basic data cleansing, data masking

Traditional BI architecture - DWH

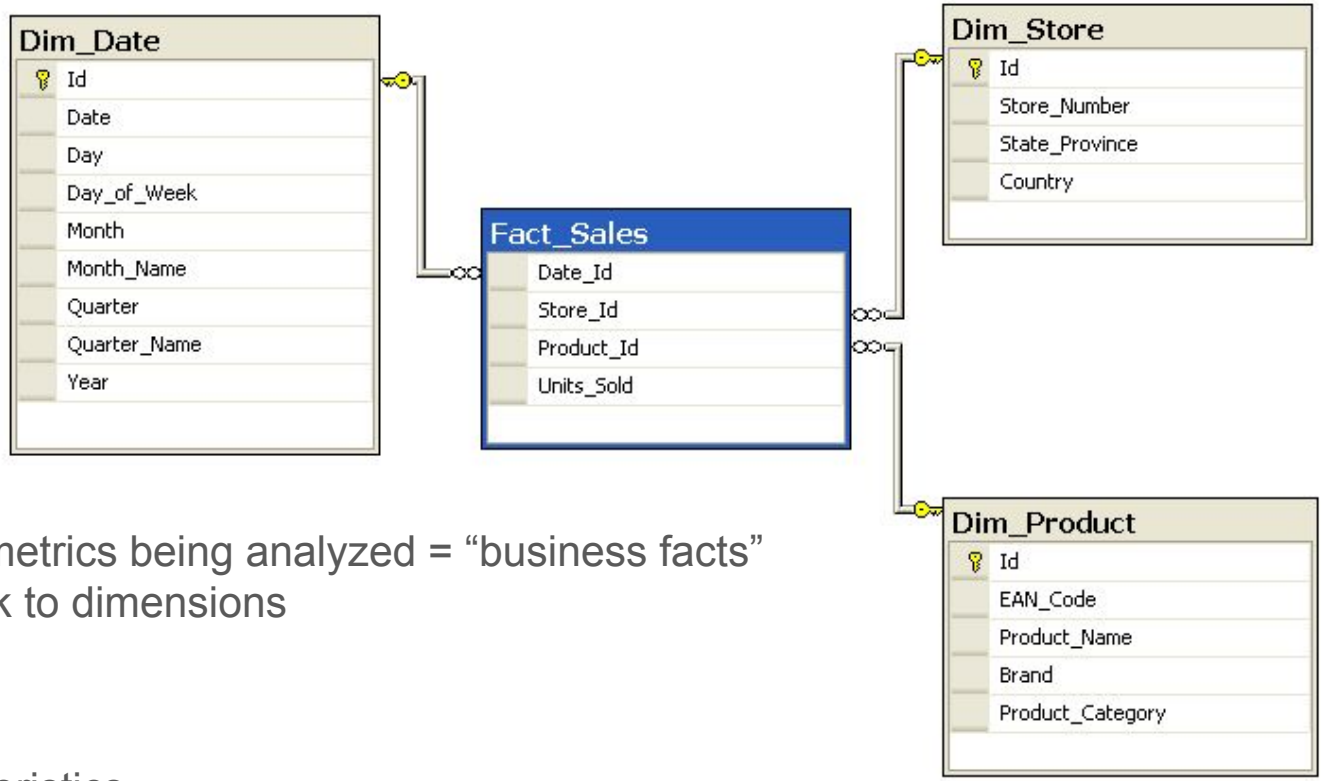


DWH

- Pre-prepared data for data analytics
  - Mostly dimensional model is used
- Table types (dimensional model)
  - Dimensions
  - Facts
  - Aggregation tables (pre-computed less detailed data for faster queries)
  - .. other
- Common data organization
  - Star schema, snowflake schema

Star schema

A fact table at the center connected to multiple dimensions.



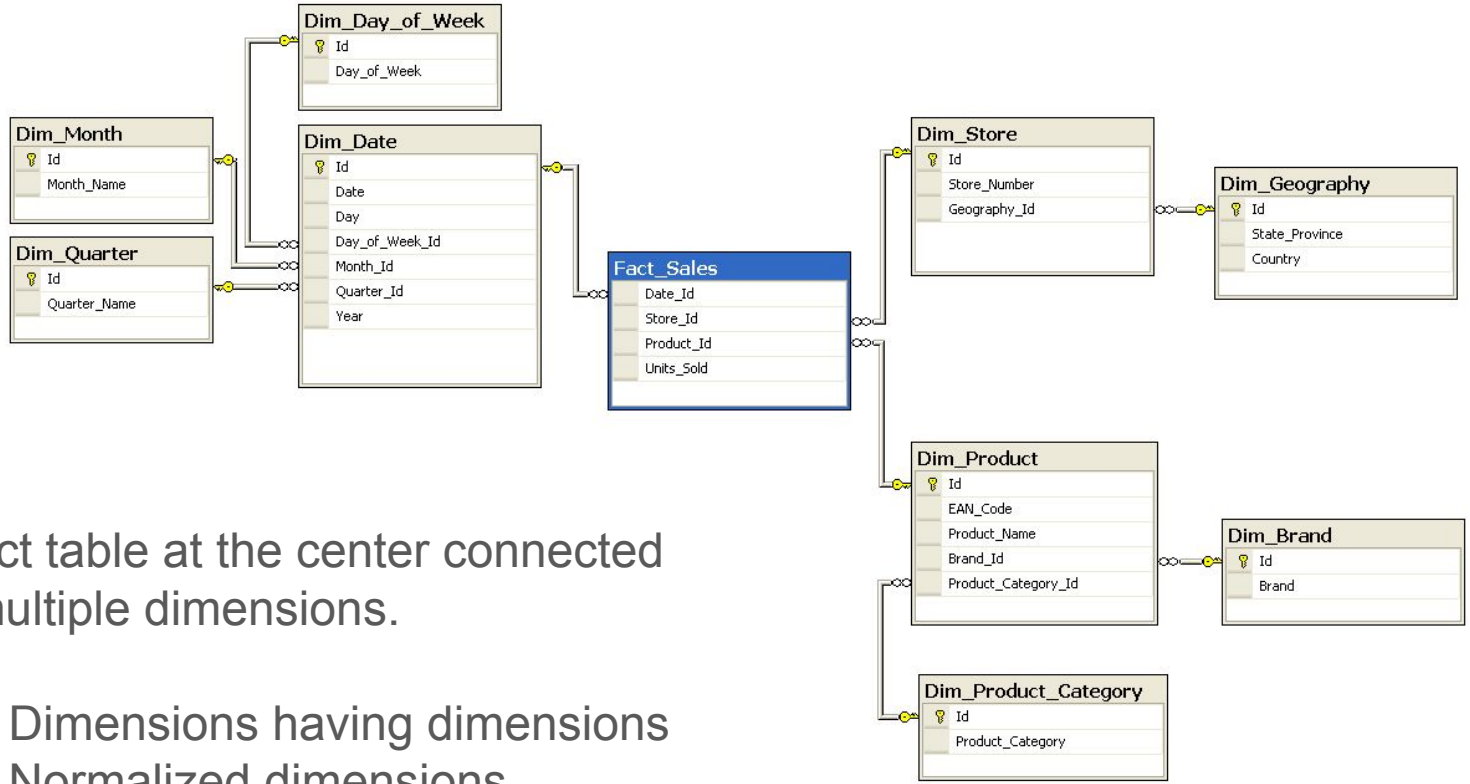
Fact table:

- Measurements or metrics being analyzed = “business facts”
- Numeric values + fk to dimensions
- Can grow fast

Dimension table:

- Attributes / characteristics
- Smaller number of records
- Can have many attributes
- Denormalized dimensions

Snowflake schema



A fact table at the center connected to multiple dimensions.

- Dimensions having dimensions
- Normalized dimensions

Example

Google analytics dimensions and facts (metrics): LINK

Data mart

- A subset of DWH data
- Focuses on a specific subject area or business department (e.g. finance, marketing, sales, management, ..)
- Designed to provide targeted data and analysis for a particular group of users, such as a sales department or a marketing team.

Two different principles

- Inmon’s vs Kimball’s approach

Inmon’s approach

Top-down approach:

- First create a centralized DWH and then create data marts as needed

CIF – Corporate Information Factory

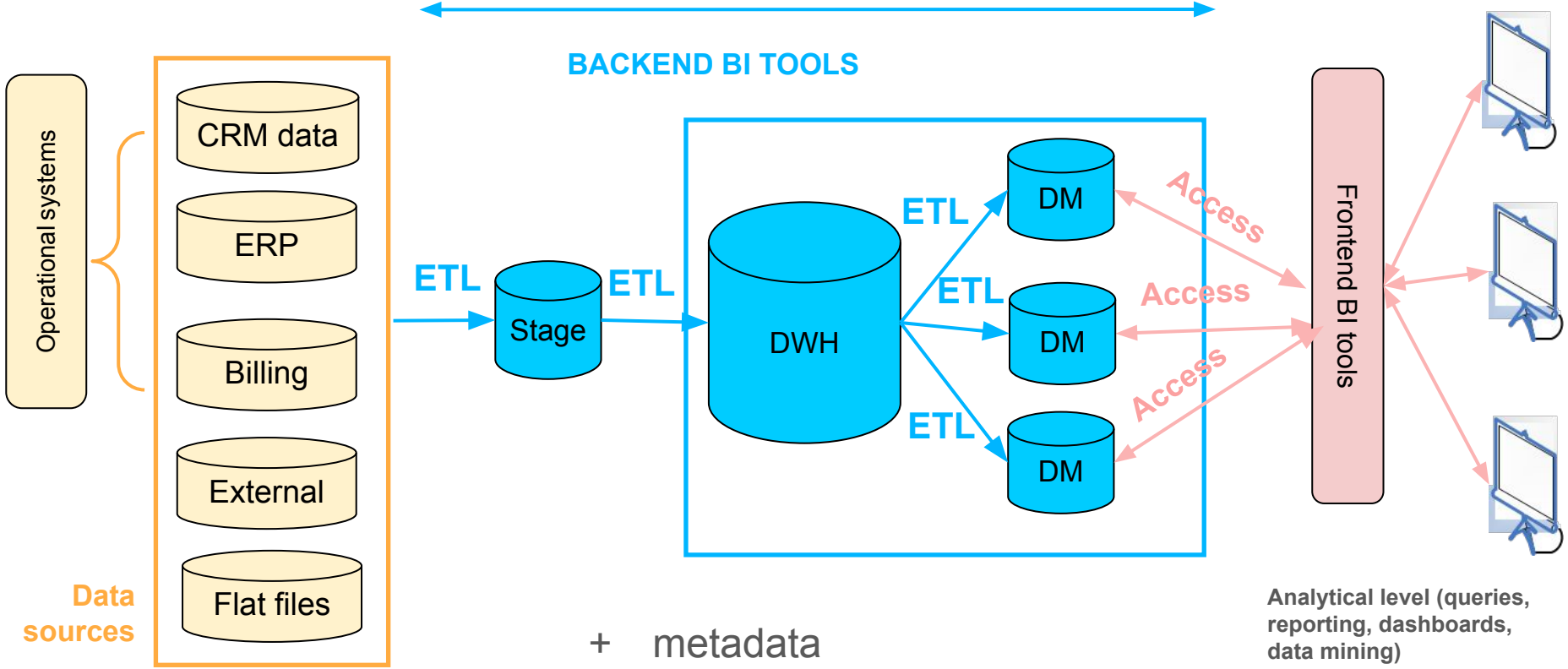
3-layer architecture

1. Staging area
2. Central data warehouse (normalized)
3. Data marts – independent, fed by the data from central DWH, each department has its own, customized data mart

(+) DWH = a single source of truth, consistent and accurate

(-) Time-consuming to build and maintain

Inmon's approach



Kimball’s approach

Bottom-up approach:

- First create data marts for specific business needs and then integrate them into large DWH

DWHBus – Data Warehouse Bus

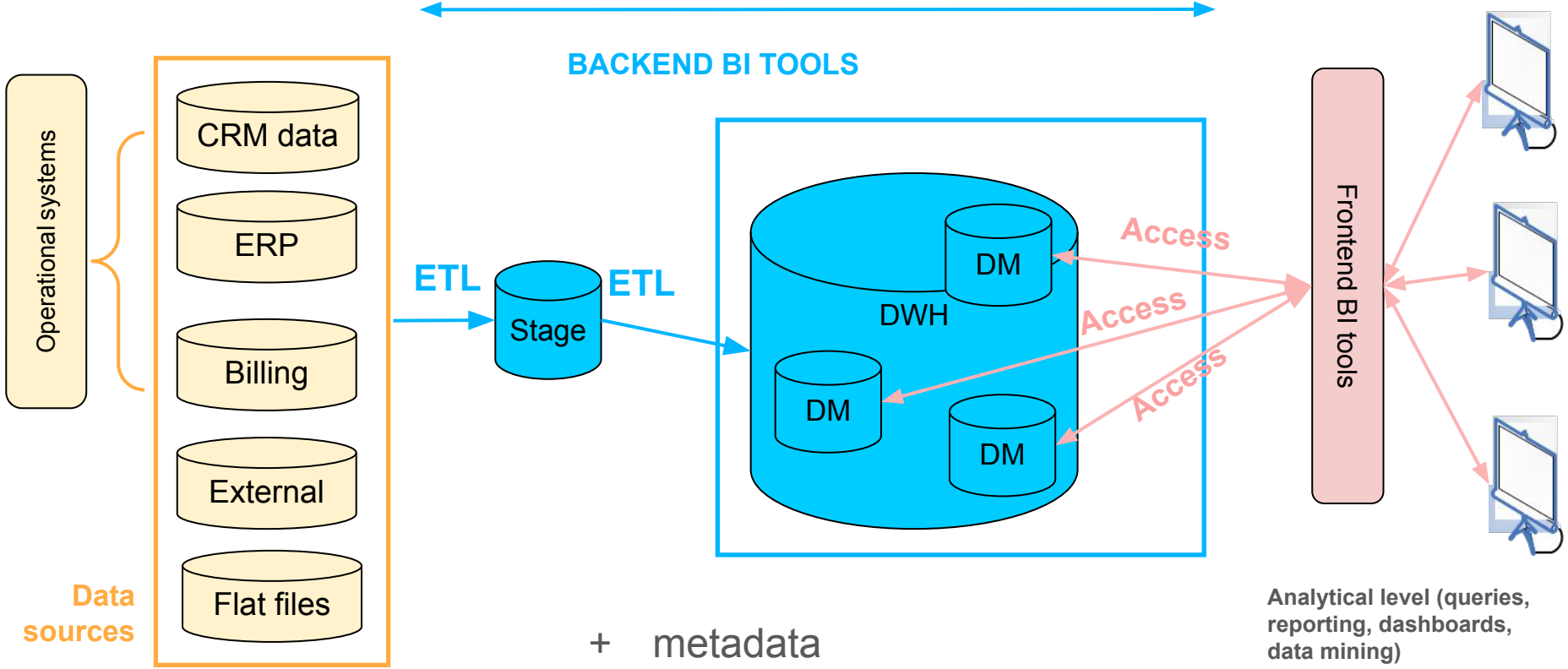
2-layer architecture

1. Staging area
2. DWH = collection of integrated data marts, conformed dimensions
    - Consolidated, shared dimensions
    - Data marts are often denormalized to improve query performance - star schema
    - Enteprise Data Warehouse Bus Matrix - plan of which business processes (marts) share which conformed dimensions.

(+) More agile approach

(-) DWH may become inconsistent as more DMs are added, data redundancy due to denormalization

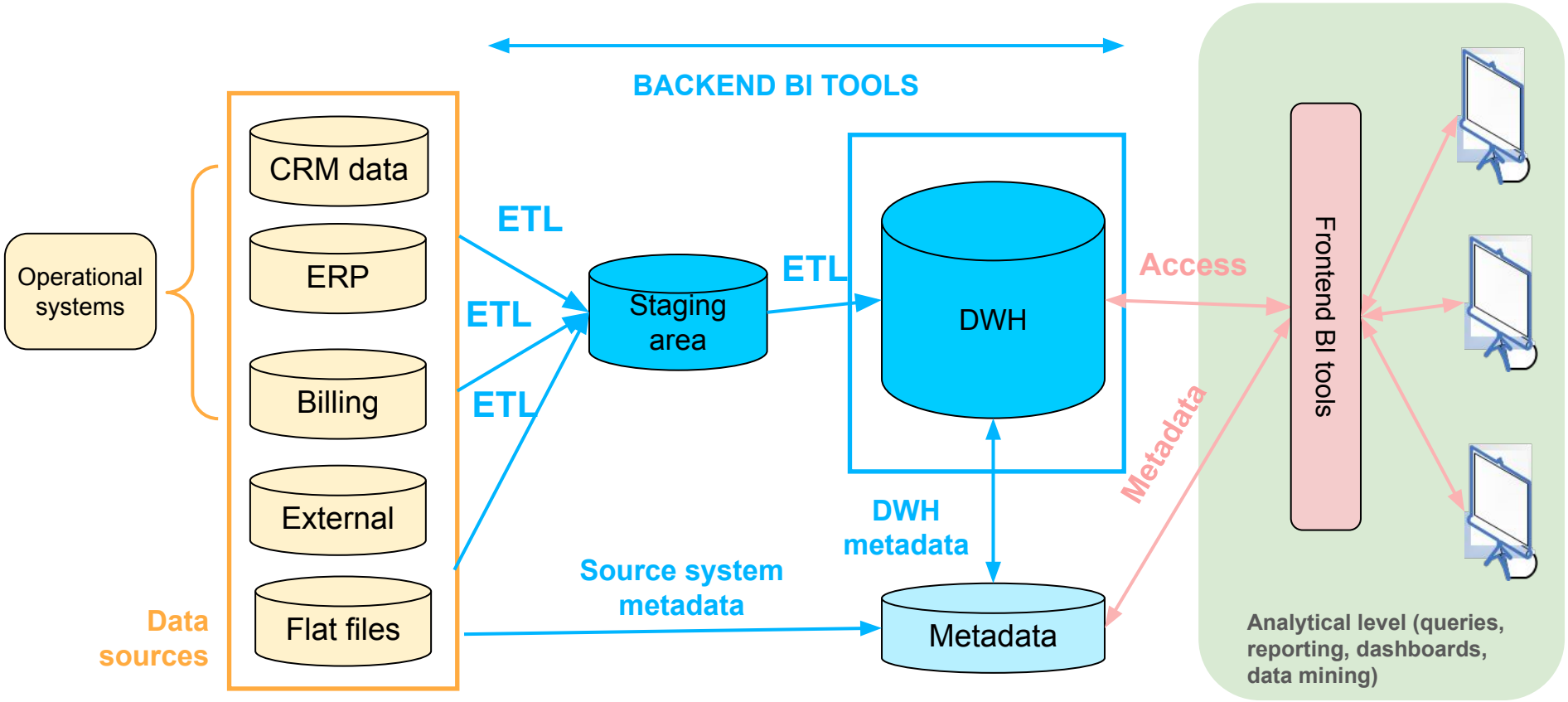
Kimball's approach



OLTP vs OLAP

OLTP - Online Transactional Processing	**OLAP - Online Analytical Processing**
Current and recent data	Historical data
Detailed data	Both detailed and summarized data
Application oriented	Analytics oriented (Subject oriented)
Insert, update, delete	Read (including complex queries)
Optimized for fast transactions	Optimized for complex queries (dimensional model)

Traditional BI architecture - front-end BI tools



Frontend BI tools

- Make information available to the users
- Tables + data visualizations: dashboards, charts, graphs
- Predefined reports vs ad-hoc (custom) reports
  - Slicing and dicing, drill-down, drill-up

  Microsoft Power BI, SAP Business Objects, Tableau, Qlik Sense
- Advanced users can access DWH / Data marts directly
  - They must know the data model and query language
- Data mining tools
  - Discover patterns, relationships, and trends in data, build predictive models
  - Statistical methods, decision trees, clustering, association rules, neural networks, …
  - Later: **data science** - broader term
    - Data mining + data management, data preparation, visualization, machine learning, and domain-specific knowledge

- Google Analytics demo accounts: LINK
- Power BI demo: LINK



OLAP cube

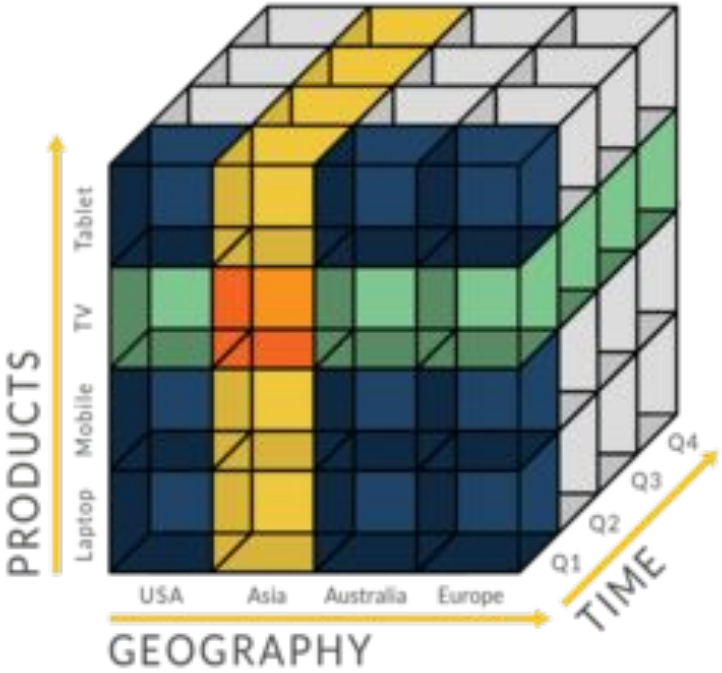
OLAP ≠ OLAP cube !

= a precomputed data structure for quick data analysis through multiple dimensions

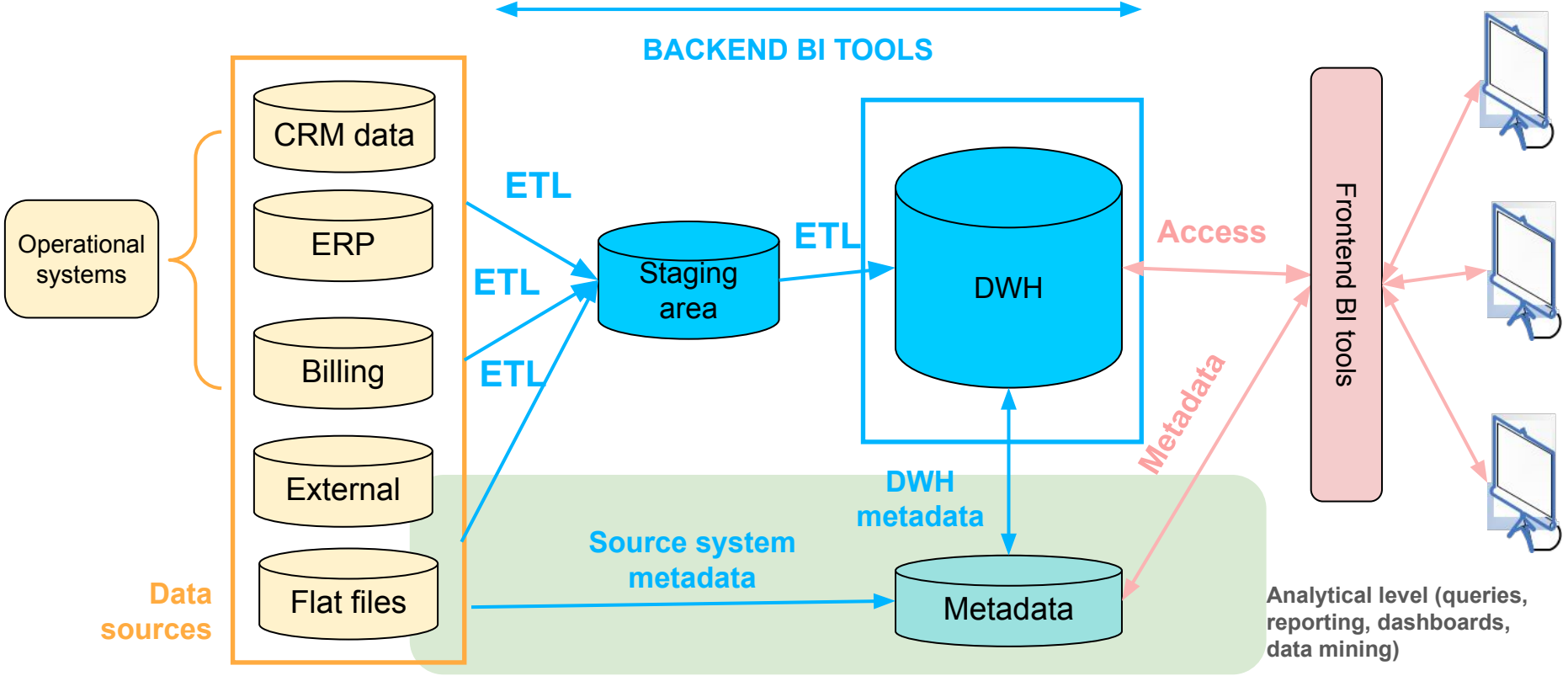
- In traditional DWHs - explicitly created for fast access (materialized views, multidimensional storage)
- In modern solutions - querying directly DWH

Example: How many TV service sales did we have in Asia in Q1?

- “Asia” – **yellow cut**
- “TV“ – **green cut**
- “Q1” – **blue cut**



Traditional BI architecture - metadata



Metadata

= data describing data

Technical

- Table: creation date, author, column names, database name
- Column: data type, maximum length, table
- Report: type, creation date, data sources
- Data lineage: source table/column(s), transformations, target table/column

Business

- Business definitions, business rules
  - Table name, table description, data preparation rules

Process Execution

- Data “freshness”
- ETL (ELT) execution time
- Length and result (success / fail) of data processing

Managed for each data storage / ETL separately or in centralized metadata management tool

- Automatic metadata extraction (if possible), metadata integration, visualizations

Example: Data lineage with QlikSense Nodegraph LINK

Example

- Integrating loan banking products
- CRM, LOS (Loan Origination System - approval processes), Core banking system (Contracts, Payments / Transactions)
- Integrating products into a conformed dimension DIM_PRODUCT

Discussion

- Why don't we want source-system-specific product dimensions in the DWH?
- What type of dimensional model does this solution represent?
- The solution hardcodes two product hierarchies into the DIM_PRODUCT table schema. What would be a more flexible approach?
- The solution hardcodes the logic for deriving the product name. What would be a more flexible approach?
- Is "interest rate" in LOS an attribute of DIM_PRODUCT, or a measure in a fact table? What determines the answer?

Integration of dimensional data

Dimensional data may be scattered across many source systems

- Unclear mapping
- Different coverage
- Different granularity
- Different hierarchies

For analytical purposes, we want a conformed dimension

- Business input is often needed: mapping tables / master data
  - Master Data: “Data about the business entities that provide context for business transactions” *[DAMA-DMBOK: Data Management Body of Knowledge]*

Nowadays

- Integration of traditional source systems is often already in place, but requires maintenance and, in some cases, migration to newer technologies
- Integration of new data sources and systems is an ongoing requirement

Slowly changing dimensions

SCD type 1:

- Overwrite old value
- Historical analyses may be inconsistent

SCD type 2:

- Create a new version of given dimension record
- Validity attributes: VALID_FROM, VALID_TO
- Consistent historical analyses

SCD type 3:

- Store old and new value in columns
- PRODUCT_NAME, PREV_PRODUCT_NAME
- Limited history -> only the values explicitly stored are available
- Useful when only a limited amount of history is needed

References

[1] Inmon, W.H.: Building the Data Warehouse, 4th edition, 2005

[2] Kimball, R.: The Data Warehouse Toolkit, 4th Edition, 2013

[3] Song, IY. (2009). Data Warehouse. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA, 2009. https://doi.org/10.1007/978-0-387-39940-9_882

[4] DAMA-DMBOK: Data Management Body of Knowledge. Data Management Association. 2017. ISBN 978-1634622349.

Credit:

- Wannapik
- Slide 41: Example of a database star schema. A central fact table links records in dimension tables using unique identifiers to key related records.by SqlPac (talk), CC BY-SA 3.0
- Slide 42: Example of a Snowflake schema, by SqlPac (talk), CC BY-SA 3.0
- Slide 52: OLAP cube by Margaret Rouse, olap.com, Guru99, CC BY-SA 4.0
- Vectorportal.com, CC BY