Data Warehousing

Architecture
Outline of the Course

- Introduction
- **DWH Architecture**
  - DWH-Design and multi-dimensional data models
- Extract, Transform, Load (ETL)
- Metadata
- Data Quality
- Analytic Applications and Business Intelligence
- Implementation and Performance
Lecture Outline

1. What is “Architecture”?
2. DWH-Architectures
3. DWH-Reference Architectures
4. DWH-Platforms
What is “Architecture”

- Description of the important elements and their relationships in a system
- Architectural styles
- Architecture standards
Enterprise Architecture

- No (apparent) planning
- Anything can be built anywhere

- Overall plan
- Infrastructure by design

Source:
General Architecture Goals

- (Design) blueprints
- Standardization
- Target state of a strategy
- Reuse:
  - Of components
  - Of designs, experiences, entire architectures
- Understanding
- Communication
Lecture Outline

1. What is “Architecture”?  
2. **DWH-Architectures**  
3. DWH-Reference Architectures  
4. DWH-Platforms
Scope of a DWH

- Different architecture styles are adequate for different scopes
- Departmental DWHs implement some or all of the requirements of a single department
- Enterprise DWHs integrate all the company’s data and make them accessible to analytic applications
Strategic Role of a DWH

- **Strategic DWH:**
  - Regarded as mission-critical tool for the analysis and optimization of business processes
  - Built and used with a long-term view

- **Tactical DWH:**
  - Built for a concrete, single, and restricted purpose
  - Built as fast as possible to make analysis available as soon as possible

 depending on the strategic role, different architecture approaches are reasonable and practicable
DWH Architecture Requirements: Functional

- **Integration**
  - Sourcing of data from relevant data sources
  - Integration and homogenization
  - Historization of data
  - Current data
  - Data quality assurance
  - "single source of truth" respectively "single version of the truth"

- **Analysis**
  - Support for adequate analysis technologies
  - Multi-dimensionality
  - Consistent reporting
DWH Architecture Requirements > Non-functional

- Extensibility
  - New sources
  - New analysis applications

- Performance
  - Analysis performance
  - Processing time

- Scalability

- Time-to-Market

- No impact on operational applications

- Availability

- Cost effectiveness
DWH-Architecture: ad-hoc

DWH without explicit Architecture
- Flexible for some time, then turns into barely extensible system
  - Most of the efforts goes into maintenance of the current state (KTLO)
- isolated
- expensive
- Difficult to maintain
- No single version of the truth
- Inconsistent reporting
- Probably incomplete

⚠️ Only adequate for tactical and departmental DWHs
DWH-Architectures (Ariyachandra & Watson 2005)

- Independent data marts
- Data Mart Bus
- Hub and Spoke
- Central Data Warehouse
- Federated Data Warehouse
Data Integration

- **virtual vs. physical**
  - physical: data are copied from sources into the DWH
  - Virtual: data are kept only in sources and are integrated at runtime with other data

- **End-of-day vs. Real-time**
  - EOD: data are extracted regularly (e.g., at end of the day)
  - (Near) real-time: data are loaded continuously
Extraction, Transformation, Loading

- ETL-Processes and tools
- **E**xtraction of data (out of sources)
- **T**ransformation (→ Integration)
- **L**oading of the transformed data into the DWH
- Here ETL is shown with an own staging area for transformations
Single-Layer DWH-Architecture

- Integration and ...
- Analysis in the same database (but separated from OLTP)
- Centralized data warehouse
Single-Layer DWH-Architecture

- Impact for enterprise-wide DWHs:
  - Integration of all data sources
  - Support for all analysis needs

- A single system must satisfy all (even diverging) requirements

- Integration and analysis requirements may conflict

- Analysis requirements (from different users) may also conflict
Independent Data Marts

- Multiple data marts (small data warehouses)
- No common, shared integration -> independent data marts
- No central, physical data warehouse
Data Marts

- DWHs are typically designed independently from analysis needs
- Typical DWH structures can be hard (inefficient) to use for analysis
- Build database that is tailored for analytic use
  - Dimensional view
  - Aggregation
  - Selection
- Separation of concerns between DWH (integration) and data mart (analysis)
- “dependent” Data Marts
- Hub and Spoke
DWH-Architecture with Multiple Data Marts

- Can support a broad spectrum of users and analysis needs
- Customized data mart design
- Can even support different technologies for data marts (relational, non-relational)
DWH-Architecture with Multiple Data Marts

😊 Higher redundancy
😢 Higher processing overhead
😊 Data mart tailored for analysis
  - complexity
  - Performance
😊 Loading of data marts can be parallelized
Structure of the DWH / Integration Layer

- DWH-DB is logically (!) structured into multiple logical databases
- Logical databases are defined according to domains (Business Areas, Subjects)
  ➜ Sample domains in a bank: payments, credits, customers, ...
DWH-Architecture with Virtual Data Marts
DWH-Architecture with Virtual Data Marts

- Data are not materialized (i.e., stored) in data marts
- Data marts are defined as database views over base tables in the DWH
- ETL logic coded in view definitions
  - higher complexity
  - challenging requirements against the DBMS
  - complex view definitions
- lower storage consumption
- shorter processing time
- better time-to-market
Federated DWH-Architecture
Federated DWH-Architecture

- Access to OLTP (!) and DWH databases via federation layer
  - \(\square\) short time-to-market
  - \(\square\) current data (real-time view of data)
  - \(\times\) additional dependencies
  - \(\times\) impact on OLTP processing
  - \(\times\) data are typically not historized in OLTP systems
Lecture Outline

1. What is “Architecture”?
2. DWH-Architectures
3. DWH-Reference Architectures
4. DWH-Platforms
The Notion of "Reference Architecture"

- "A reference model is a division of functionality together with data flow between the pieces. A reference model is a standard decomposition of a known problem into parts that cooperatively solve the problem. Arising from experience, reference models are a characteristic of mature domains." [Bass et al., Software Architecture in Practice]

- "A reference architecture is a reference model mapped onto software elements (that cooperatively implement the functionality defined in the reference model) and the data flows between them." [Bass et al., Software Architecture in Practice]
Credit Suisse DWH Reference Architecture V4

Legend:
- relational database
- multidimensional database
- file
- logic; extract, transform, load
- logic (no ETL)
- data flow

Layer: Data Integration | Data Enrichment | Analysis
---|---|---
DWH Data Universe | Data Marts | Analysis Services
Integration, Historization | Selection, Aggregation, Calculation | Presentation
Subject Matter Areas | Reporting, OLAP, Data Mining | Front End
Staging Area | Reusable Measures & Dimension Areas | Web/App Servers
Metadata Management | GUI
Landing Zone

Spring Term 2018 Slide 30
Credit Suisse Reference Architecture: Principles

- **P1**: Separation of integration and analysis
- **P2**: Ownership of DWH-applications
- **P3**: Managed data quality
- **P4**: Corrections and adjustments
- **P5**: Encapsulation of data access across layers

**general principles**

- **P6**: Logical integration of base data into centrally-managed schemas

**data integration principles**

- **P7**: Reusable complex logic in enrichment applications
- **P8**: Independence of enrichment applications

**data enrichment principles**

- **P9**: Independence of data marts
- **P10**: Private data marts

**analysis principles**
Lecture Outline

1. What is “Architecture”?  
2. DWH-Architectures  
3. DWH-Reference Architectures  
4. **DWH-Platforms**
Application Platforms - Definition

**Application Platform (AP):** Set of integrated *technical components* and *processes* for the *development* and *operation* of similar applications

- Frees developers and operations from many infrastructure and operation concerns
- Focuses on efficient processes and lean operation
- Developing an AP is substantially more expensive than building the infrastructure for a single application. An adequate number of applications sharing the platform is necessary to amortize the effort
- An AP is designed and supported as a whole (allows optimizations not possible on a per component basis, e.g. availability, BCP, security, audit)
- Standard products, processes, and guidelines
  - HW, OS, middleware, network for test levels and production
  - System Management (deployment, change, monitoring, administration, ...)
  - Security (authentication, authorization, encryption, firewalls, ...)
  - Development (tools, frameworks, construction guidelines, ...)
  - Overall platform support, release management & roadmap, security concept, operating manual, usage criteria
  - Cross-AP communication via integration infrastructure
DWH-Platforms

- Standard components, especially
  - DBMS
  - ETL-Tool
  - BI-Suite

- Integration of these components
  - Which each other
  - Into security infrastructure
  - Into Systems Management (e.g. monitoring)

- Architecture- and development standards and guidelines
  - Reference architecture
  - Detail concepts (e.g., metadata, data quality)
  - Modeling guidelines
Conclusion

- Different architecture styles exist
- There is no universally best solution
  - Best architecture depends on scope and role of the DWH as well as on the maturity of the organization
- Architecture and concept are critical especially for enterprise DWHs
- Reference architectures are useful as blueprints and for the enforcement of standards
- Big, enterprise-wide DWHs tend to have two or even three layers
- (Application) Platforms are an approach to industrialize and standardize, which is particularly important in the DWH area (why?)