Presentation:
Summary

Erik Perjons

DSV, Stockholm University

The goals of and requirements on the data warehouse

Stockholms universitet

Kimball/Roll's bedrock requirements for the DW/BI system can be interpreted as the goals of a data warehouse (since requirements on a system can seen as a type of a system)

- The DW/BI system must make information **easily accessable (simpe and fast)**
- The DW/BI must **present information consistently**
- The DW/BI system must **present information in a timely way**
- The DW/BI system must **adapt to change**
- The DW/BI system must serve as the **authoritative and trustwothy foundation for improved decision making**
- The **business community must accept the DW/BI system** to deem it successfully

The goals of and requirements on the data warehouse

Also Inmons' data warehouse definition can be interpreted as containing requirements on - or goal of - a data warehouse

- A data warehouse is a **subject oriented, integrated, non-volatile, and time-variant collection** of data in support of management's decisions".
- This definition can be interpreted as following:
  - A data warehouse - aims to be - or needs to be:
    - subject oriented
    - integrated
    - non-volatile
    - time-variant

The goals of and requirements on the data warehouse

In the beginning of Lecture 1, I also presented some requirements on - or goals of – a Data warehouse, based on some problem stated by Kimball/Ross (page 3)

- We need **Accessibility** to the data for users not very familiar with IT and data structures
- We need **Data integration** on the basis of a standard enterprise model
- We need **Query flexibility** to maximise advantages obtained on the run
- We need **Information conciseness** allowing target oriented and effective analysis
- We need **Correctness and Completeness** of integrated data
- We need a solution that captures **Historical data**
- We need to optimize **Read performance**

The benefits of data warehouse

- Time saving for users
- More and higher quality of information
- Better decisions
- Improvement of business processes
- Support for accomplish strategic business goals

(Watson, H. J., Goodhue, D. L., & Wixom, B. H. (2002). The benefits of data warehousing: why some organizations realize exceptional payoffs. *Information & Management*, *39*(6), 491-502.)

DW characteristics

What is characteristics regarding data warehouse, according to Chaudhuri&Dayal (1997):

- SQL extensions (operators like Cube, Crossjoin)
- Index structures (bit map indexes, join indexes)
- Materialized views (pre-calculated aggregations)

(Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM Sigmod record*, *26*(1), 65-74.)

Stockholms universitet

More about DW characteristics

- The data is organised differently, i.e. “multidimensional”
  - star-joins schemas
  - snowflake schemas
- The data is viewed differently
  - multidimensional view
- The data is stored differently
  - vector (array) storage for MOLAP solutions
- The data is indexed differently
  - bitmap indexes
  - join indexes

Different DW Archtectures

- Kimball/Ross' DW/BI Architecture
- Inmon's Hub-and-Spoke Corporate Information Factury Archiecture?
- Kimball's/Ross' suggested Hybrid Architecture
- Centralized Architecture
- Federated Architecture
- Independent Data Mart Architecture

(See Lecture 5, papers as part of the course literature, and Kimball & Ross, chapter 1)

Other archtectural and strategic issues

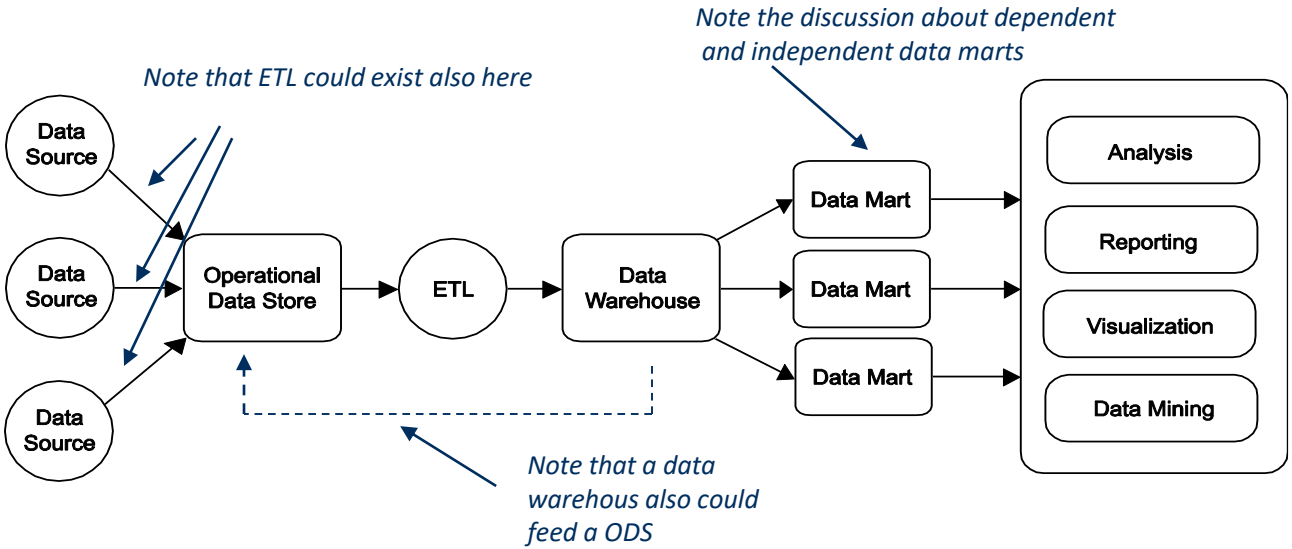
- Data marts – independent and dependent
- Operational data store (note, not the same as staging tables used in the ETL process)
- Master data and master data management system

Other archtectural and strategic issues

- BI vs. Big Data
- Data lake
- The future for DW in business intelligence
- BI strategy and DW
- Data brokers
- Metadata repository

Operational Data Store (ODS)

- Operational data store is a system sitting between the operational systems and the data warehouse providing operational reporting, especially when neither other operational (OLTP) systems or the data warehouse can provide near real time – and somewhat integrated – operational data



Four steps to design a star schema

1. • Select a Business Process
2. • Declare the Grain
3. • Identify the Dimensions
4. • Identify the Facts

Dimensional modelling patterns

Stockholms universitet

- Transaction Fact Table, Periodic Snapshot Fact Table, Accumulating Snapshot Fact Table
- Factless fact table
  - Event tracking tables
  - Coverage tables
- Heterogenous products: Core and custom fact tables

Dimensional modelling patterns

- Slowly changing dimensions Type 1, 2 and 3
- Minidimension
- Degenerated dimensions
- Junk dimensions
- Time dimension

Other dimensional modelling concepts

- Conformed dimension, conformed fact
- Denormalized dimensions
- The use of surrogate key, concatenated key, natural key, smart keys etc.
- Additive, semi-additive and non-additive facts
- Sparse fact table

Other dimensional modelling concepts

- Textual values in fact vs dimensional columns/attributes
- Hierarchies of attributes
- Preaggregated fact tables
- The concept of cube
- Data integrity
- Declare the grain
- Fiscal year

Other dimensional modelling concepts

- Enterprise Data Warehouse Buss Matrix
- Outriggers
- Different kind of aggregations
- Snowflake schema

Why dimensional modelling?

Stockholms universitet

- the logical model is easy understand
- a predictable standard framework for end user applications
- the logical design can be done nearly independent of expected query pattern
- handle changes easy - adding new dimensional attributes, and adding new dimensional tables

Why dimensional modelling?

Stockholms universitet

- high performance “browsing” across the attributes and query performance, eliminating joins and make use bit vector indexes
- strategy to handling aggregates, e.g. summery records that are logical redundant with base table to enhance query performance
- the database engine can make strong assumption how to optimise
- strategies for handling slowly changing dimensions, heterogeneous products, event-handling (“factless fact tables”)

ETL architetures

- Different architecture
  - Staggered staging
  - Persistent staging
  - Pipeline staging
  - Chuncked Accumulated staging
  - Ackumulated staging
- Staging tables

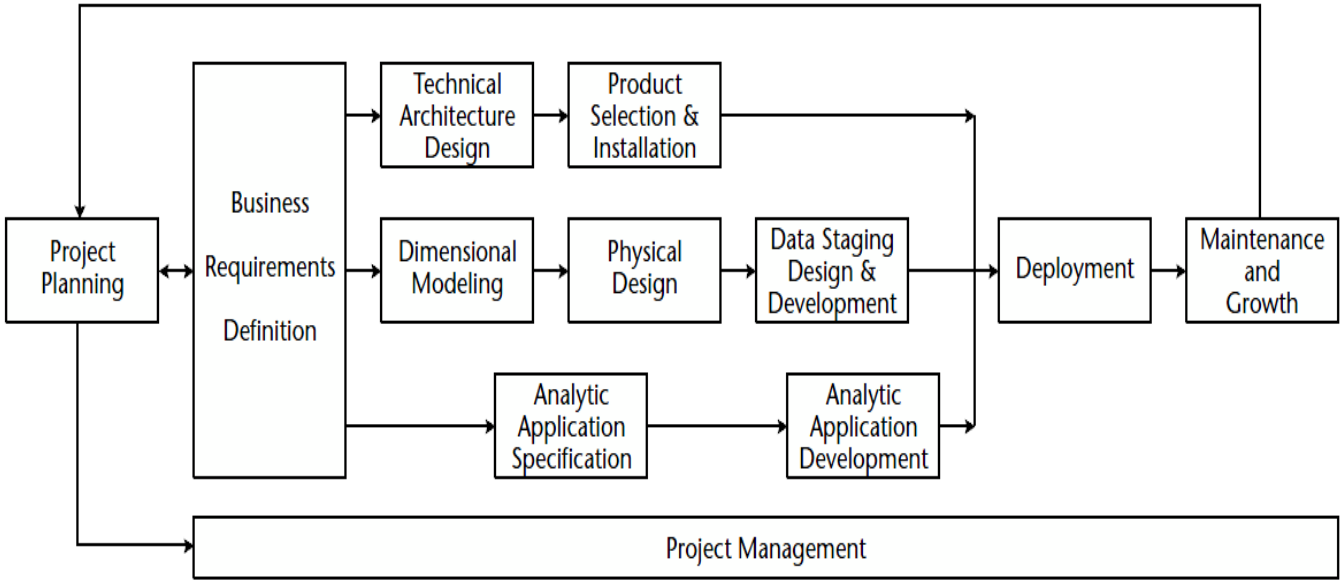
ETL issues

- Source system issues
- Extract issues
- Data cleansing issues
- Dimensional building issues
- Fact building issues
- One time historic load processing vs. incremental load processing

Kimball/Ross' ETL process

- Subprocess 1: Develop a plan (and its substeps 1-4)
- Subprocess 2: Develop on-time historic load processing (and its substeps 5-6)
- Subprocess 3: Develop Incremental ETL Processing (and its substeps 7-10)

DW/BI Lifecycle



OLAP

- OLAP – the concept of OLAP
- MOLAP, ROLAP, HOLAP solutions
- Loading the cube in MOLAP