Spatial Data Warehouse : Why and What

국민대학교 정보관리학부

정 승 렬

Data Warehouse – Why (I)

기존 시스템 환경

- 신속한 분석, 예측이 어려움
- 통합 보고서 작성이 어려움
- 비정형화된 보고서 작성이 오래 걸림

Batch 방식의 **Report**

- **Long Time**
- **Redo for Similar Job**
- **No More Past Data (We delete data sets after batch jobs)**

5년간 년도별 매출실적
-> 5년간의 테잎을 프로세싱

카드회사의 고객카드 사용대금 명세서
-> 고객의 카드 사용내역 분석불가

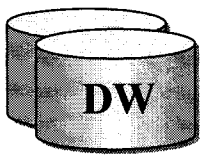
Prof. Seung Ryul Jeong

Data Warehouse – Why (II)

통합 데이터 베이스

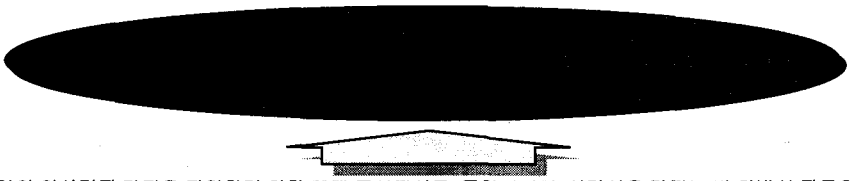
- **Log file**을 이용하여 **Operational Data**를 추출 (**Deferred Log** 추출방식)
- 운영계 시스템의 거의 모든 데이터를 보유
- 사용자를 위해 프로그래밍된 화면을 통해 데이터에 접근

- 수백개 이상의 화면 디자인 -> 개발기간이 길어짐
- 개발비용이 높음
- 비정형화된 보고서는 다시 프로그래밍 해야 함
- 분석, 예측기능이 없음



Prof. Seung Ryul Jeong

Data Warehouse – What (I)

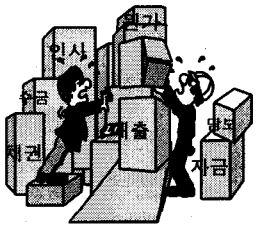


□ 기업의 의사결정 과정을 지원하기 위한 주제 중심적이고, 통합적이며, 시간성을 가지는 비 휘발성 자료의 집합
- Bill Inmon -

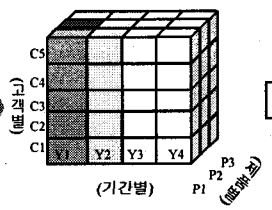
□ 의사결정 지원용으로 특별히 디자인된 주제 중심적인 정보 저장고 - Meta Group -

□ 회사의 각 부문에 흩어져 있는 데이타를 의미있는 정보로 바꾸어 모아놓은 창고 - Informix -

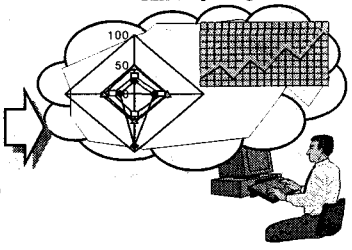
분산/단절된 정보



정리된 Data창고



End User Computing실현



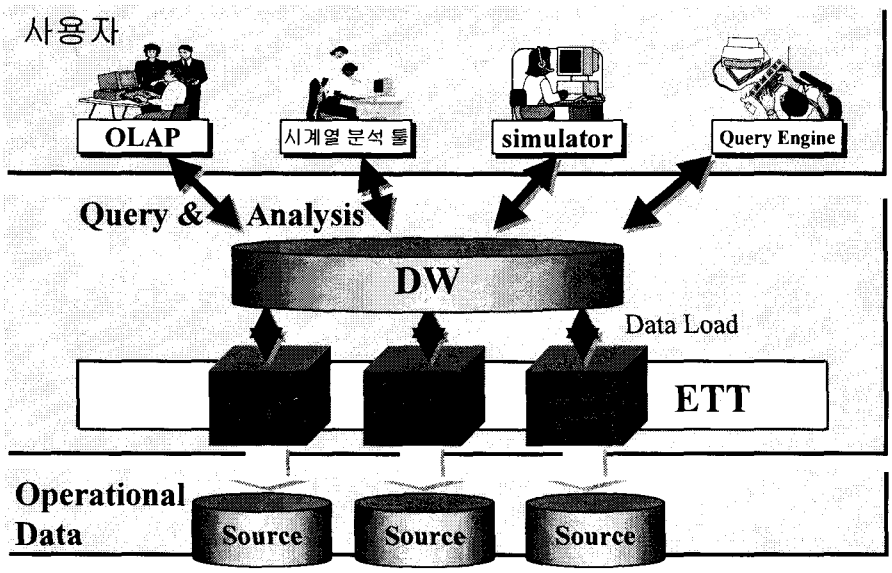
Prof. Seung Ryul Jeong

Data Warehouse – What (II)

- **Stored Collection of Diverse Data**
  - Solution to data integration problem
  - Single repository of information
- **Subject-Oriented**
  - Organized by subject, not by application
  - Used for analysis, data mining
- **Optimized Differently from Transaction DB**
- **User Interface Aimed at Executive**
- **Large Volume (GB, TB) of Data**
- **Non-Volatile**
  - Contents are stable for long periods of time
  - Enables long analysis transactions
  - Often updates are append-only
- **Time Variant Kept**
  - History: Set of snapshots
  - Time attributes important

Prof. Seung Ryul Jeong

Data Warehouse – Architecture



Prof. Seung Ryul Jeong

DW and Related Issues

❑ OLAP (Online Analytical Processing)

- ➢ 저장고에 있는 데이터를 사용자가 접근하여 분석하는 도구
- ➢ **MOLAP vs. ROLAP vs. HOLAP**
  - **MOLAP:** 다차원 데이터베이스 구조, 고급분석, 시간과 비용이 없는 경우, 부서용 활용 목적
  - **ROLAP:** 관계형 데이터베이스 구조, 다차원 모델링 **(Star Schema, Snowflake Schema)** 대용량의 전사적 **DW**, 원시데이터의 출력이 필요한 경우
- ➢ **Multi-Dimensional Modeling**
  - **Fact Table : Numerical measurements, usually additive** 예) 매출액, 매출 수량
  - **Dimension : Textual Descriptions such as product specification. Each dimension joins with Fact table.** 예) 기간 (년/분기/월), 조직 (본부, 지사, 영업소)

❑ Warehouse Loading

- ➢ **Extraction, Transformation (format, merge, code inconsistencies, etc.), Cleansing**

❑ ODS (Operational Data Store)

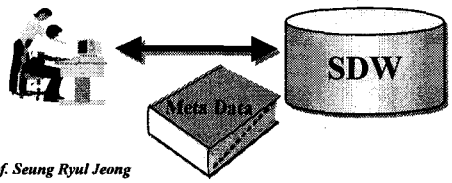
- ➢ 운영계에서 추출, 정제한 데이터를 보관.
- ➢ **E/R Modeling,** 사용자의 직접적인 접근 불허용

❑ Top Down vs. Bottom Up

Prof. Seung Ryul Jeong

Spatial Data Warehouse - Why

•Many corporate & user applications
•Various kinds of technology and formats
•Most applications built vertically with little cross application data standards or integration
•Users do not know what information exists
•Users can not easily integrated data
•Cannot afford to rebuild those applications.
•Many are meeting specific operational needs.

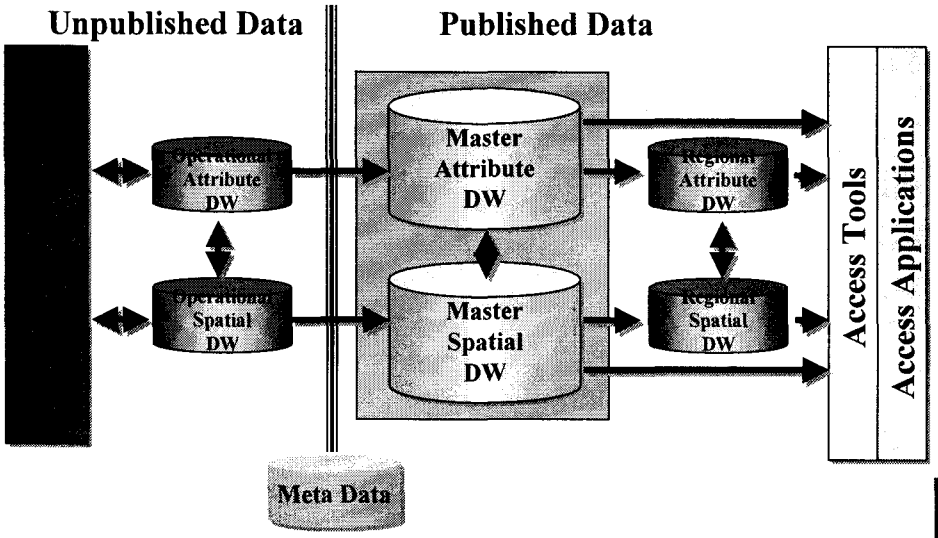


The Solution

•One central repository
•One common technology format
•Data is integrated
•Common data standards followed
•Database tuned for ad hoc query
•Fast, easy, and consistent access to all data sets
•Data is "cleansed" / anomalies are known

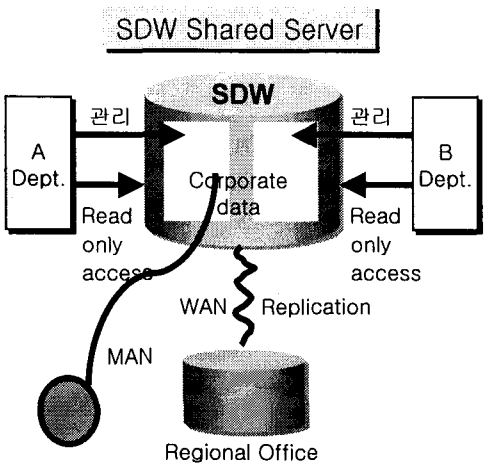
Prof. Seung Ryul Jeong

Spatial Data Warehouse - What



Prof. Seung Ryul Jeong

Multi-Agency Data sharing Through SDW



- ❑Publishing Concept
- ❑Significant improvements in GIS data currency
- ❑Require Standards on GIS Projections, ground units, naming conventions, meta-data, etc → greater efficiencies for cooperation
- ❑Data sharing not impact existing processes of data capture & quality assurance → use the existing technology environment
- ❑Data integrity & synchronization of data sets → CSF
- ❑Data replication service > current GIS data translation & data Mgt. Processes
- ❑Data Security for data replication must be standardized

Prof. Seung Ryul Jeong

Example: Weather Pattern Analysis

- **Input:**
  - a map with about 3,000 weather probes scattered in B.C.
  - daily data for temperature, precipitation, wind velocity, etc.
  - attributes are organized in hierarchies
- **Output:**
  - a map that reveals patterns: merged (similar) regions!
- **Goals:**
  - interactive analysis (drill-down, slice, dice, pivot, roll-up)
  - fast response time
  - minimizing storage space used
- **Challenge:**
  - a merged region may contain hundreds of "primitive" regions (polygons).

Prof. Seung Ryul Jeong

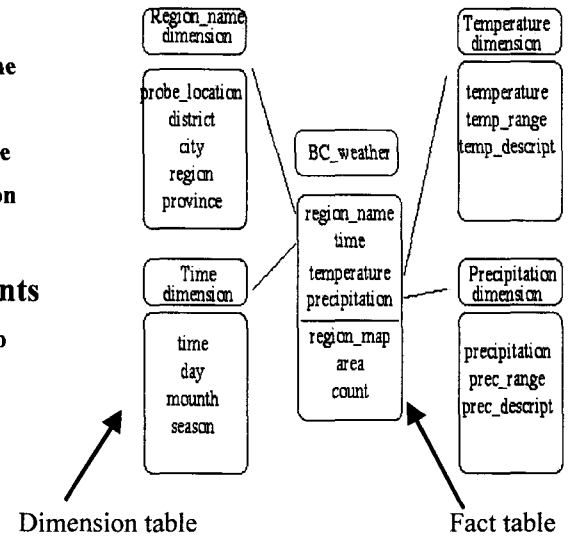
A Model of SDW

- **Dimension modeling:**
  - nonspatial
    - (e.g. temperature: 25-30 degrees generalizes to *hot*)
  - spatial-to-nonspatial
    - (e.g. region "B.C." generalizes to description "*western provinces*")
  - spatial-to-spatial
    - (e.g. region "Burnaby" generalizes to region "Lower Mainland")
- **Measure formation:**
  - numerical
    - distributive (e.g. count, sum)
    - algebraic (e.g. average)
    - holistic (e.g. median, rank)
  - spatial
    - collection of spatial pointers (e.g. pointers to all regions with 25-30 degrees in July)

Prof. Seung Ryul Jeong

Star Model of a SDW

- **Dimensions**
  - **region_name**
  - **time**
  - **temperature**
  - **Precipitation**
- **Measurements**
  - **region_map**
  - **area**
  - **count**



Prof. Seung Ryul Jeong

SDW 구축 사례 (I)

- **EPA (Environmental Protection Agency), U.S.A.**
  - **Envirofacts Data Warehouse** 를 구축하여 환경과 **human population** 간의 관계분석을 수행
  - 공간정보 + **regulatory data** + **Census Bureau**의 인구통계학적 데이터
  - **EDW**에 구축된 규제데이터는 6개의 **EPA National Systems** 으로부터 전송되어 통합
    - CERCLIS (The Comprehensive Environmental Response, Compensation, and Liability Information System)
    - PCS (The Permit Compliance System) / RCRIS (The Resource Conservation and Recovery Information system)
    - TRIS (The Toxic Release Inventory System) / GICS (The Grants Information and Control System)
    - AFS (The Envirofacts Aerometric Information Retrieval System / AIRS Facility Subsystem)
  - 현재 구축중인 시스템으로 **SDWIS (The Safe Drinking Water Information System)**가 있음.
  - **EDW**의 주요 구성요소
    - Facility Index System: 다종의 regulatory DB에서 나타나는 시설물들을 상호 연계시켜 통합 분석이 가능하도록 함
    - Locational Reference Tables: 규제 시설물(regulated facilityes)에 대한 위치 정보를 저장함.
    - EPA Spatial Data Library System: 약 50GB의 지리공간 정보를 저장하고 있는 공간 데이터 저장소임.
  - **EPA**는 Envirofacts Data Warehouse를 이용하여 특정 지역 또는 규제 시설물 등에 대한 각종 환경 오염 등을 측정하여 원인지역에 대한 분석, 영향평가 등을 수행하고 있음.
  - 또한 규제 data와 각종 공간 데이터(도로, 강, 국립공원, 학교 등) 및 인구통계학적 데이터(인구밀도, 1인당소득, 빈곤도 등)와의 결합을 통해 다양한 분석용 지도를 생성하고 있기도 함.

Prof. Seung Ryul Jeong

SDW 구축 사례 (II)

❑ MELP (Ministry of Environment, Lands, and Parks), BC, Canada

- ➢ **To make decisions on land use, environmental regulation, and resource planning.**
- ➢ **SDW provides instant access to current versions of spatial and attribute data from many sources to users throughout the Ministry.**
- ➢ **Warehouse Contents – spatial data & Attributes data**
  - Topography, Transportation, Toponymy, Administrative Boundaries, Forestry, Cadastre, Water and Drainage, Wildlife, etc.
  - Financial records, Water licenses, Fish Observations and activities, Lake surveys and chemistry, Hunt Kills, etc.
- ➢ **Each information source has its own refresh frequency**
- ➢ **Changes to spatial data are automatically copied to regional sites each night**
- ➢ **SDW is designed for 3 classes of users**
  - MELP employees directly access to data via LAN. Victoria staff access the warehouse over the MAN while regional staff access local copies replicated by overnight copies.
  - External agencies and companies download files from the public FTP site. Some predefined queries are also available to fetch attributes from the data warehouse.
  - Public access via WEB is in the planning stages.
- ➢ **SDW is effective for delivery of large volumes of current data to a variety of users. It simplifies data management, and encourages cooperation between the agencies which use it.**

Prof. Seung Ryul Jeong