Data-Warehouse-, Data-Mining- und OLAP-Technologien

Chapter 1: Introduction

Bernhard Mitschang
Universität Stuttgart

Winter Term 2014/2015
Overview

- Motivation
  - Retail Scenario
  - Heterogeneous Data Sources
  - The Basic Idea
- Operational vs. Analytical Systems
- Fields of Application
- Information Services
- Other Approaches
- Overview Chapters 2-7
Retail Scenario

- Retail company runs many department stores in Europe (Berlin, London, Paris, Rome, ...).
- The goods come from global suppliers.
- Business people (marketing, inventory management, ...) need to know:
  - What are the gross sales for an article in each country?
  - What are the gross sales for an article per month?
  - What are the gross sales for an article per quarter?
  - What is the sales-per-square-meter of an article?
  - Did our last sales promotion push the gross sales of an article?
  - Which articles are typically sold together?
  - Which customer groups are most likely interested in a special sales promotion?
  - ...
Heterogeneous Data Sources

- The retail company runs several database systems that capture all the operational data:
  - one point-of-sale (POS) database per department store
  - one supplier database per country
  - one central article database
  - …
Heterogeneous Data Sources

- Broad range of data sources:
  - Database systems (relational, object-relational, hierarchical, XML, …)
  - Enterprise Resource Planning Systems
  - External information sources (other companies, surveys, …)
  - Files of standard applications (Excel, …)
  - Other documents (Word, WWW, …)

- Data Sources may differ in:
  - schema
  - coding schemes
  - time frame
  - treatment of history
  - aggregation level
  - database management system (data model, vendor, version, …)
Data Warehouse: The Basic Idea

- separation of operational and analytical data
- several analytical applications supported
Information Pyramid

Transaction Processing (OLTP)
(heterogeneous information systems,
isolated information islands
constantly increasing data sets)

Data Warehouse

OLAP, Data Mining

strategic
planning
operational
effectiveness
performance

All this data must be worth something

Too much data, not enough information

We are drowning in information, but starving for knowledge
# Operational vs. Analytical Systems

<table>
<thead>
<tr>
<th></th>
<th>operational</th>
<th>analytical</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>focus</strong></td>
<td>read, write, update, delete</td>
<td>read, periodical update</td>
</tr>
<tr>
<td><strong>transaction</strong></td>
<td>short read/write transaction</td>
<td>long read transaction</td>
</tr>
<tr>
<td><strong>queries</strong></td>
<td>simple</td>
<td>complex</td>
</tr>
<tr>
<td><strong>records per</strong></td>
<td>tens</td>
<td>millions</td>
</tr>
<tr>
<td><strong>transaction</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>data model</strong></td>
<td>flat relation</td>
<td>multidimensional</td>
</tr>
<tr>
<td><strong>data sources</strong></td>
<td>mainly one</td>
<td>many</td>
</tr>
<tr>
<td><strong>properties</strong></td>
<td>current, up-to-date, detailed, isolated</td>
<td>historical, summarized, consolidated</td>
</tr>
<tr>
<td><strong>volume</strong></td>
<td>MB - GB</td>
<td>GB - TB</td>
</tr>
<tr>
<td><strong>data access</strong></td>
<td>single rows</td>
<td>ranges</td>
</tr>
<tr>
<td><strong>user type</strong></td>
<td>clerk, IT professional</td>
<td>knowledge worker, manager</td>
</tr>
<tr>
<td><strong>number of</strong></td>
<td>$&gt; 10^3$</td>
<td>$&gt; 100$</td>
</tr>
<tr>
<td><strong>response time</strong></td>
<td>ms - s</td>
<td>s - min</td>
</tr>
<tr>
<td><strong>metric</strong></td>
<td>transaction throughput</td>
<td>query throughput, response time</td>
</tr>
</tbody>
</table>
Operational vs. Analytical Systems

- Performance and safety considerations
- Logical interoperability problems
- Temporal and granularity mismatch

Butler
Chester Road
+03 6543 2323
xyz@netmail
Fields of Application

- **Business Management**
  - Marketing Management
  - Enterprise Information Portals
  - One-to-one-Marketing, Electronic Customer Care, Customer Relationship Management (CRM)
  - Electronic Procurement, Supply Chain Management

- **Scientific Applications**
  - empirical studies, e.g. in earth observation, environmental studies
  - Statistical and Scientific Databases (SSDB)

- **Engineering Applications**
  - analyzing the reliability of products
  - analyzing substances and material
Introduction to Services

Definitions of ‘Service’

- A **repeatable business task** – e.g., check customer credit; open new account

- Discretely defined **set of** contiguous and autonomous business or technical **functionality**

- A **function** provided at a network address; available via **various transports, formats, and QoS**; “always on”

- A mechanism to enable **access** to capabilities, which is provided using a prescribed **interface** and is exercised consistent with **constraints and policies** as specified by the **service description**

Sources: IBM; Wikipedia; Frank Leymann; OASIS SOA Reference Model
Introduction to SOA

Definitions of ‘Service-oriented Architecture’

- An IT architectural style that supports integrating your business as linked services

- A method for systems development and integration where functionality is grouped around business processes and packaged as interoperable services

- An architectural style to realize service-oriented computing

- A paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains
Overview of Information Services

- **Service-oriented Architecture:**
  - Integrierter Ansatz zur Erneuerung der Geschäftsprozesse (Software AG)

- "You will waste your investment in SOA unless you have enterprise information that SOA can exploit." (Gartner, 2005)

- **Information is a Vital Component of SOA Strategy**

- **Information Service:**
  - **Equals:** Data Storage + Retrieval (and crunching)
  - **Definition:** Every capability needed to understand, cleanse, integrate and deliver information across heterogeneous systems
  - **Goal:** services that provide accurate, consistent, integrated information to business processes and people
  - **Starting point:** existing legacy, inconsistent & diverse data
  - **Approach:** Information as a Service
    - Exposing application logic as services
    - Data is only accessible through the corresponding application logic
    - If needed: Mediation (brokering) and orchestration (workflow) of application logic
Classification of Information Services

- Classification Criteria:
  - Level of abstraction:
    - Plain source data (SQL, ...) or meta data
    - Integrated data (data/information integration, ETL, ..., MDM)
    - Interpreted data (BI, ..., text analytics, content extraction)
    - Application data (CRM, ...)
  - Range
    - Intranet
    - Internet
  - Source
    - Single/multiple
    - Homogeneous/heterogeneous
  - Provisioning of the service
    - Proprietary service
    - Web/SOA service
  - …
Cloud Computing: One Big Picture

**Gartner:** “Cloud Computing is a style of computing where **massive scalable IT-enabled capabilities** are delivered **as a service** to external **customers** using Internet technologies.”

... **others** ...: Cloud Computing isn’t new. We have hosted apps for years. But now also custom-developed applications can be hosted in the cloud.

---

**A Service**

- **Business Properties**
  - No startup cost
  - Pay per use (pay as you go): costs grow linearly

- **Technical Properties**
  - Simple and easy to use programming model
  - Scalability and reliability as well as administration (for free) ➔ Cloud
  - Minimum TCO ➔ Autonomic/Cloud, Multi-tenancy
What is Missing?

1. Good and experienced Service Engineers
2. Good and experienced Cloud Engineers
3. Best Practices experiences

To 1.: International Master in Service Engineering (IMSE)
To 2.: International Master in Service Engineering (IMSE)
To 3.: International Master in Service Engineering (IMSE)
Other Approaches

<table>
<thead>
<tr>
<th>1960s</th>
<th>1970s</th>
<th>1980s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Support Systems</td>
<td></td>
<td>Management Decision Systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strategic Planning Systems</td>
</tr>
</tbody>
</table>

- **Goals**
  - Focused on providing managers with structured, periodic reports.
  - Integrated man/machine system for providing information to support operations, management, and decision-making functions in an organization.
  - Interactive information systems that used data and models to help managers analyze semi-structured problems.

- **Why they failed:**
  - lacking fast communication networks
  - lacking large, cheap, and fast data storage
  - lacking cheap, high-performance CPUs
  - lacking flexibility
  - lacking productivity
  - lacking trustworthiness
Defining 'Data Warehouse'

- A data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data in support of management's decisions.  
  
  (Source: [Inm05])

- A data warehouse is a copy of transaction data specifically structured for query and analysis.  

  (Source: [Kim96])

- Ein Data Warehouse ist ein physischer Datenbestand, der eine integrierte Sicht auf die zugrunde liegenden Datenquellen ermöglicht.  

  (Source: [Zeh03])
Other Important Terms

• **Data Warehousing:**
  - Process of integrating data from several source systems, storing it in the data warehouse database, and using this integrated data source.

• **Business Intelligence:**
  - General term that covers technology and tools to turn the volume of data a organization collects and stores into meaningful information in order to achieve better and timelier business decisions.

• **Decision Support:**
  - More general term referring to all kinds of analyses of existing data in order to make better decisions - like data mining, online analytic processing (OLAP), simulation, scenario analyses, ...
Issues in Data Warehousing

- Architecture
- Data Model
- Data Extraction
- Data Cleaning
- Data Transformation
- Database Support
- Data Warehouse Design
- Metadata
- Data Aggregation
- Query Optimization
- Data Warehouse Refreshment
- Project Management
- Data Warehouse Quality
- Data Mining
- Phases
- Online Analytic Processing
Goals of the Lecture

• Know what a data warehouse is and how it differs from other information systems
• Know the main components of a data warehouse
• Know the main processes in a data warehouse system
• Know how to design the schema of a data warehouse
• Know how olap and data mining is used to gain information from data warehouse data
• Know database technology that is adjusted to the needs of a data warehouse system
Overview

• Motivation
  ▪ Retail Scenario
  ▪ Heterogeneous Data Sources
  ▪ The Basic Idea
• Operational vs. Analytical Systems
• Fields of Application
• Other Approaches

• Overview Chapters 2-7
  ▪ Data Warehouse Architecture
  ▪ Data Warehouse Design
  ▪ Extraction, Transformation, Load
  ▪ OLAP
  ▪ Data Mining
  ▪ Database Support
Introduction

Anwendungssoftware

Architecture

End User Data Access

Data Warehouse

Load

Data Staging Area

Load

Extraction

Transformation

Monitor

Data Warehouse Manager

Metadata Manager

Metadata Repository

Data Warehouse System

Data flow

Control flow

(Source: [BG04])
Information Integration

End User Data Access

Data Warehouse

Integrated Schema

ETL

Data Sources

Query Execution

materialized II

data Sources

Global Schema

Schema Mappings

Wrapper

Data Sources

virtual II

Data Sources

Mediator

Wrapper

Data Sources

(Source: [LN07])
Metadata in Data Warehousing

- What data is available in the warehouse and where is the data located?
- **Data dictionary**: Definitions of the databases and relationship between data elements
- **Data flow**: Direction and frequency of data feed
- **Data transformation**: Transformations required when data is moved
- **Version control**: Changes to metadata are stored
- **Data usage statistics**: A profile of data in the warehouse
- **Alias information**: Alias names for a field
- **Security**: Who is allowed to access the data
Chapter 2: Data Warehouse Architecture

- Data Warehouse Architecture
  - Data Sources and Data Quality
  - Data Mart
  - Operational Data Store

- Information Integration
  - materialized vs. virtual
  - Federated Information Systems

- Metadata
  - Metadata Repository
  - Metadata in Data Warehousing
  - CWM Metamodel
Data Warehouse Design Process

- Requirement analysis and specification
  - Operational database schema
- Conceptual design
  - Semiformal business concept
  - Formal conceptual schema
- Logical design
  - Formal logical schema
- Physical design
  - Physical database schema

- time

- • interview
- • noun analysis
- • brainstorming
- • document analysis
- • ... (StarER)
- • dimensional fact model
- • ME/R
- • mUML
- • ... (Multidimensional)
- • relational
- • object-relational
- • ... (DB2, ORACLE, MS Server, Essbase, MS Analysis Services)
- ... (Source: [HLV00])
Data Warehouse Design

Multidimensional Model

Fact Data (sales)

„Cube“ Metaphor
Data Warehouse Design

Source systems

Customer_Berlin I
Name Address Age

Articles_Berlin I
ANR Name Price

Order_Berlin I
Cust ANR Count

POS_Berlin I
ZIP Town

Kunden_Berlin II
Name Adresse Alter

Artikel_Berlin II
Num Artikel

Bestellung_Berlin II
Kunde Anum Anzahl Farbe Preis Status

POS_Berlin II
ZIP Stadt Stadtteil

Customer_Rome
Name Address Age

Orders1_Rome
ONR Customer

Orders2_Rome
ONR Article Price

Orders_London
Cust Article Num Color Price Status

Customer_London
Name Address Age

Orders
CNR ANR SNR Count Status

Articles
ANR Name Price Color Size

Customers
CNR Name Age ZIP Address

Stores
SNR Name Address

data warehouse

schema?
Chapter 3: Data Warehouse Design

- Data Warehouse Design Process
- Conceptual Design
  - Multidimensional Model
  - Dimensional Fact Model, UML-based, starER...
- Logical Design
  - Star Schema
  - Snowflake Schema
  - …
- Dimensional Modeling
  - Extended Dimension Table Design
    - Slowly Changing Dimensions, Large Dimensions with Frequent Changes, …
  - Extended Fact Table Design
    - Modeling Events and Coverage, Factless Fact Table, …
- Physical Design
Extraction, Transformation, Load

source systems

Extraction Anwendersoftware

Anwendungssoftware

Articles

<table>
<thead>
<tr>
<th>ANR</th>
<th>Name</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Shirt</td>
<td>50€</td>
</tr>
<tr>
<td>102</td>
<td>Shirt</td>
<td>75€</td>
</tr>
<tr>
<td>113</td>
<td>Blouse</td>
<td>102€</td>
</tr>
<tr>
<td>114</td>
<td>Bluse</td>
<td>102€</td>
</tr>
<tr>
<td>204</td>
<td>Jeans</td>
<td>75€</td>
</tr>
</tbody>
</table>

Orders

<table>
<thead>
<tr>
<th>CNR</th>
<th>ANR</th>
<th>SNR</th>
<th>Count</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>7017</td>
<td>100102</td>
<td>1</td>
<td>5</td>
<td>open</td>
</tr>
<tr>
<td>7098</td>
<td>100104</td>
<td>1</td>
<td>3</td>
<td>ok</td>
</tr>
<tr>
<td>7564</td>
<td>210704</td>
<td>1</td>
<td>12</td>
<td>ok</td>
</tr>
<tr>
<td>7017</td>
<td>100105</td>
<td>2</td>
<td>1</td>
<td>open</td>
</tr>
<tr>
<td>7017</td>
<td>210204</td>
<td>2</td>
<td>1</td>
<td>open</td>
</tr>
<tr>
<td>7017</td>
<td>100101</td>
<td>2</td>
<td>1</td>
<td>open</td>
</tr>
<tr>
<td>7537</td>
<td>120113</td>
<td>2</td>
<td>1</td>
<td>open</td>
</tr>
<tr>
<td>7456</td>
<td>120513</td>
<td>2</td>
<td>1</td>
<td>open</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Articles

<table>
<thead>
<tr>
<th>ANR</th>
<th>Name</th>
<th>Price</th>
<th>Color</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>100101</td>
<td>Shirt</td>
<td>50€</td>
<td>yellow</td>
<td>L</td>
</tr>
<tr>
<td>100102</td>
<td>Shirt</td>
<td>75€</td>
<td>blue</td>
<td>XL</td>
</tr>
<tr>
<td>100103</td>
<td>Shirt</td>
<td>80€</td>
<td>white</td>
<td>M</td>
</tr>
<tr>
<td>100104</td>
<td>Shirt</td>
<td>75€</td>
<td>black</td>
<td>M</td>
</tr>
<tr>
<td>100105</td>
<td>Shirt</td>
<td>65€</td>
<td>black</td>
<td>S</td>
</tr>
<tr>
<td>120113</td>
<td>Blouse</td>
<td>102€</td>
<td>pink</td>
<td>XS</td>
</tr>
<tr>
<td>120513</td>
<td>Blouse</td>
<td>150€</td>
<td>yellow</td>
<td>S</td>
</tr>
<tr>
<td>210204</td>
<td>Jeans</td>
<td>75€</td>
<td>blue</td>
<td>L</td>
</tr>
<tr>
<td>210704</td>
<td>Jeans</td>
<td>85€</td>
<td>blue</td>
<td>XL</td>
</tr>
<tr>
<td>340101</td>
<td>Socks</td>
<td>5€</td>
<td>black</td>
<td>6</td>
</tr>
</tbody>
</table>

Transformation

<table>
<thead>
<tr>
<th>Price (€)</th>
<th>Color</th>
<th>Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>75€</td>
<td>blue</td>
<td>Shirt</td>
</tr>
<tr>
<td>75€</td>
<td>black</td>
<td>Shirt</td>
</tr>
<tr>
<td>85€</td>
<td>blue</td>
<td>Jeans</td>
</tr>
</tbody>
</table>

Load

<table>
<thead>
<tr>
<th>Name</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>113</td>
<td>102€</td>
</tr>
<tr>
<td>704</td>
<td>-</td>
</tr>
<tr>
<td>113</td>
<td>Blouse</td>
</tr>
<tr>
<td>113</td>
<td>102€</td>
</tr>
<tr>
<td>113</td>
<td>Blouse</td>
</tr>
<tr>
<td>113</td>
<td>102€</td>
</tr>
<tr>
<td>704</td>
<td>Jeans</td>
</tr>
<tr>
<td>704</td>
<td>-</td>
</tr>
<tr>
<td>704</td>
<td>Jeans</td>
</tr>
<tr>
<td>704</td>
<td>75€</td>
</tr>
<tr>
<td>704</td>
<td>75€</td>
</tr>
<tr>
<td>704</td>
<td>75€</td>
</tr>
<tr>
<td>704</td>
<td>75€</td>
</tr>
<tr>
<td>704</td>
<td>75€</td>
</tr>
<tr>
<td>704</td>
<td>75€</td>
</tr>
</tbody>
</table>

same article? yes

missing article? yes

same customer? yes

missing price? yes

same coding (€/£)? yes

Ich empfehle Ihnen, den Artikel mit Artikelnummer 113 zu wählen, da er identisch mit Artikel 704 ist. Der Artikel ist ein Bluse und kostet 102€.


Zusammenfassend kann gesagt werden, dass die Integration von Extraktion, Transformation und Laden von Daten eine Grundvoraussetzung für die effiziente Datenverarbeitung ist. Diese Vorgehensweise ermöglicht es Unternehmen, die Daten in ihren Systemen zu optimieren und in der Lage zu sein, schnell auf Veränderungen in der Wirtschaft zu reagieren und ihre Geschäftsaufgaben effektiv durchzuführen.
Chapter 4: Extraction, Transformation, Load

• Monitoring
• Extraction
  ▪ Export, Import, Filter, Load
  ▪ Direct Integration
• Load
  ▪ Bulk Load
  ▪ Replication
  ▪ Materialized Views
• Transformation
  ▪ Schema Integration
  ▪ Data Integration
  ▪ Data Cleansing
• Tools
Introduction

**OLAP**

- **Online Analytic Processing**
- Technologies and tools that support (ad-hoc) analysis of multi-dimensionally aggregated data
- Individual analysis is supported, i.e., the user is not restricted to available standard reports/analysis
- Graphical user interface is available for analysis specification
- Knowledge of a query language or programming language is not required
- Result information is given graphically and made available for incorporation into other applications
- Users: Analysts, Manager, “knowledge worker”
- Typical Analysis scenarios:
  - Multi-dimensional views, e.g. turnover per product group and month
  - Comparisons, e.g. turnover in Q4 compared to that of Q3
  - Ranking, e.g. top 10 product in a certain group ranked by turnover
OLAP

- **Examples:**
  - **Slicing:** Analysis of a certain product group
  - **Aggregation:** Sum of turnover for all products of a given product group
  - **Drilling:** Based on a previous analysis of a certain business area, further analysis on a level of product group and even on the level of a single product are performed
OLAP

ToP-CHain - THE retail company

predefined reports

define individual analysis
Chapter 5: Online Analytic Processing

- Introduction
- OLAP Operations
- OLAP Characteristics
  - OLAP Product Evaluation Rules
  - FASMI Test
- Multidimensional Database Systems
  - Multidimensional Arrays
  - Sparse Cubes
  - Multidimensional Query Language
- Architecture
  - MOLAP, ROLAP, HOLAP
Data Mining

- **Data mining** is the process of discovering hidden, previously unknown and usable information from a large amount of data. The data is analyzed without any expectation on the result. Data mining delivers knowledge that can be used for a better understanding of the data.
  

- **Data Mining Techniques:**
  - **Association Rule Discovery**
    - \{beer, nappies\} $\rightarrow$ \{potato chips\}
    - support = 0.04
    - confidence = 0.81

  - **Clustering**

  - **Classification**

  - **Regression**
Data Mining
Chapter 6: Data Mining

• Introduction
  ▪ Terms & Definitions
  ▪ Disciplines & Applications

• Data Mining Techniques
  ▪ Classification
  ▪ Regression
  ▪ Association Rule Discovery
  ▪ Clustering

• Data Mining Systems
  ▪ Tool
  ▪ Trends
OLAP Support in SQL:1999

- WINDOWs allow to apply aggregate functions to the current row and its neighboring rows.

- GROUP BY clause is extended by the CUBE and ROLLUP keywords which allows multidimensional summaries.

<table>
<thead>
<tr>
<th>partno</th>
<th>date</th>
<th>sales</th>
<th>Sum(sales)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P050</td>
<td>2003-09-29</td>
<td>500</td>
<td>800</td>
</tr>
<tr>
<td>P050</td>
<td>2003-09-30</td>
<td>200</td>
<td>850</td>
</tr>
<tr>
<td>P050</td>
<td>2003-10-01</td>
<td>100</td>
<td>1550</td>
</tr>
<tr>
<td>P050</td>
<td>2003-10-02</td>
<td>550</td>
<td>950</td>
</tr>
<tr>
<td>P050</td>
<td>2003-10-03</td>
<td>400</td>
<td></td>
</tr>
<tr>
<td>P101</td>
<td>2003-09-30</td>
<td>600</td>
<td></td>
</tr>
<tr>
<td>P101</td>
<td>2003-10-01</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>
Introduction

CUBE

- Support Roll-up (dimension reduction)
- Compute:
  - Sales for all aggregation groups of attributes country, month and group.

query using grouping sets

```sql
SELECT country, month, group, SUM(sales)
FROM Sales
GROUP BY GROUPING SETS (
    (country, month, group),
    (country, month),
    (country, group),
    (month, group),
    (country), (month), (group), ()
GROUP BY country, month, group
GROUP BY country, group
GROUP BY country, month
GROUP BY month, group
GROUP BY country
GROUP BY month
GROUP BY group
```
CUBE

equivalent query using CUBE
SELECT country, month, group, SUM(sales)
FROM Sales
GROUP BY CUBE
  (country, month, group)

"-": NULL value
Evaluating Star Queries

Optimizer of the DBMS has to decide on join order, index usage, predicate push-down, ...

Alternative query plans for star queries:

SELECT D1.a, D2.d, …, Dn.x, f(F.fact)
FROM F, D1, D2, …, Dn
WHERE F.b = D1.b
AND F.e = D2.e

…
AND D1.c = …
AND D2.f = …

GROUP BY D1.a, D2.d, …, Dn.x
TPCH Benchmark Schema

PART (P_)
- SF*200.000
  - PARTKEY
  - NAME
  - MFG
  - BRAND
  - TYPE
  - SIZE
  - CONTAINER
  - RETAILPRICE
  - COMMENT

PARTSUPP (PS_)
- SF*800.000
  - PARTKEY
  - SUPPKEY
  - AVAILQTY
  - SUPPLYCOST
  - COMMENT

ORDERS (O_)
- SF*1.500.000
  - ORDERKEY
  - CUSTKEY
  - ORDERSTATUS
  - TOTALPRICE
  - ORDERDATE
  - ORDERPRIORITY
  - CLERK
  - SHIPPRIORITY
  - COMMENT

LINEITEM (L_)
- SF*6.000.000
  - ORDERKEY
  - PARTKEY
  - SUPPKEY
  - LINENUMBER
  - QUANTITY
  - EXTENDEDPRICE
  - DISCOUNT
  - TAX
  - RETURNFLAG
  - LINESIZE
  - SHIPDATE
  - COMMITDATE
  - RECEIPTDATE
  - SHIPINSTRUCT
  - SHIPMODE
  - COMMENT

CUSTOMER (C_)
- SF*150.000
  - CUSTKEY
  - NAME
  - ADDRESS
  - NATIONKEY
  - PHONE
  - ACCTBAL
  - MKTSEGMENT
  - COMMENT

SUPPLIER (S_)
- SF*10.000
  - SUPPKEY
  - NAME
  - ADDRESS
  - NATIONKEY
  - PHONE
  - ACCTBAL
  - COMMENT

NATION (N_)
- 25
  - NATIONKEY
  - NAME
  - REGIONKEY
  - COMMENT

REGION (R_)
- 5
  - REGIONKEY
  - NAME
  - COMMENT

Database Scaling:

<table>
<thead>
<tr>
<th>SF</th>
<th>database size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 GB</td>
</tr>
<tr>
<td>10</td>
<td>10 GB</td>
</tr>
<tr>
<td>30</td>
<td>30 GB</td>
</tr>
<tr>
<td>100</td>
<td>100 GB</td>
</tr>
<tr>
<td>300</td>
<td>300 GB</td>
</tr>
<tr>
<td>1000</td>
<td>1000 GB = 1 TB</td>
</tr>
<tr>
<td>3000</td>
<td>3000 GB</td>
</tr>
<tr>
<td>10000</td>
<td>10000 GB</td>
</tr>
<tr>
<td>30000</td>
<td>30000 GB</td>
</tr>
<tr>
<td>100000</td>
<td>100000 GB</td>
</tr>
</tbody>
</table>
Chapter 7: Database Support

• Support for OLAP and Data Mining in SQL
  ▪ OLAP:
    ROLLUP, CUBE, WINDOW, OLAP Functions
  ▪ Data Mining (SQL/MM)

• Database Support for OLAP
  ▪ Partitioning
  ▪ Materialized Views
  ▪ Indexes
  ▪ Optimization of OLAP Queries (Star Joins)

• Performance Evaluation: TPC Benchmarks
Books


Papers


