
OLAP Tuning

Outline

- OLAP 101
  - Data warehouse architecture
  - ROLAP, MOLAP and HOLAP
- Data Cube
  - Star Schema and operations
  - The CUBE operator
  - Tuning the cube
- Data Mining 101

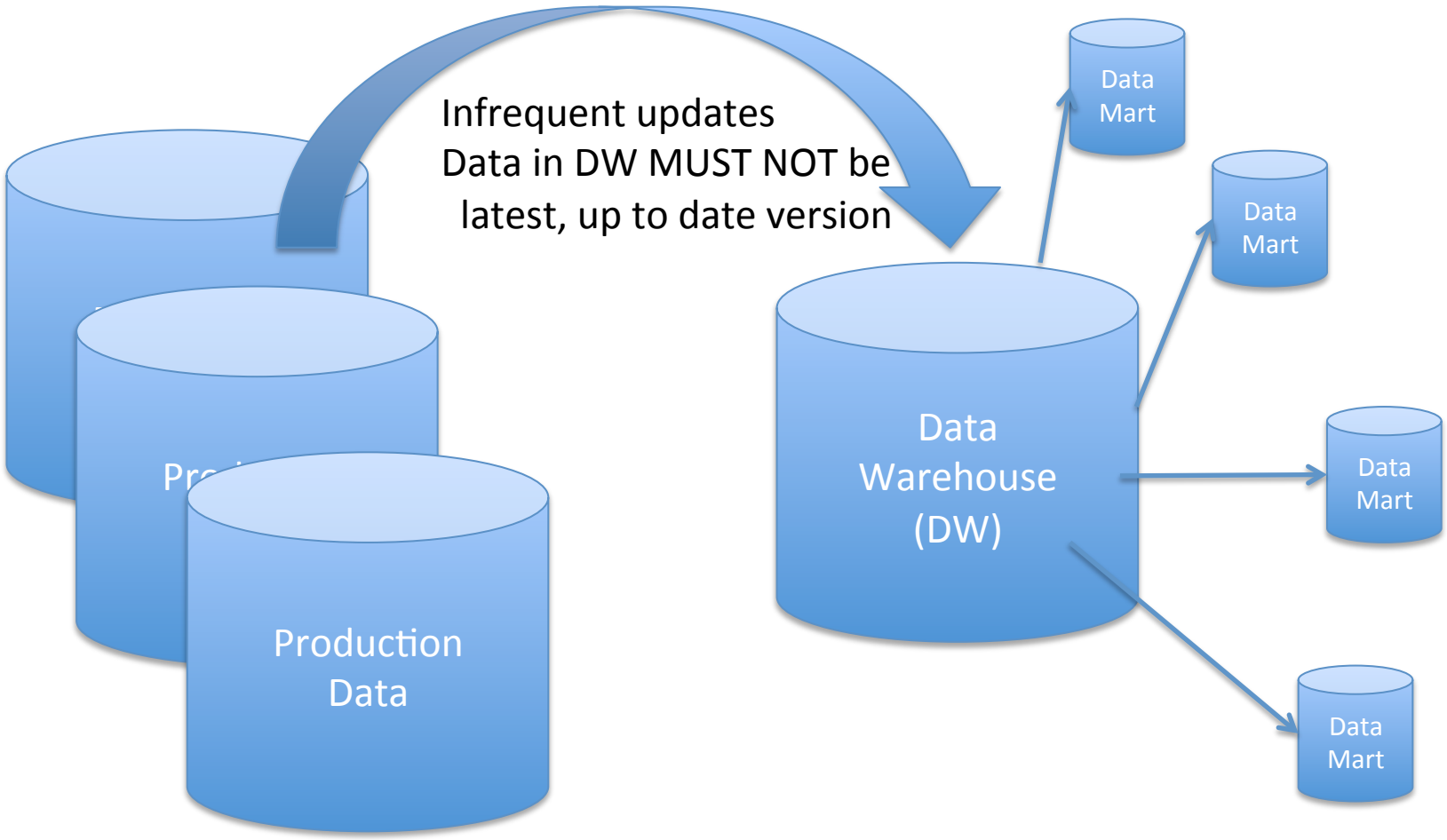
@ Dennis Shasha and Philippe Bonnet, 2013

OLAP

- Online Analytical Processing
  - OLAP enables a user to interactively and selectively extract and view data from different points-of-view.
  - Typical OLAP queries
    - Find sales for seniors in Copenhagen (selection)
    - Find sales per age group, per city (aggregation)
    - Find sales per age group, per country (aggregation)
    - Find total sales (aggregation)
    - Find sales for seniors, per country (selection, aggregation)

Selections & Aggregations on Multi-dimensional Data

Data Warehouse



Transactional Processing

Analytical Processing

ROLAP, MOLAP, HOLAP

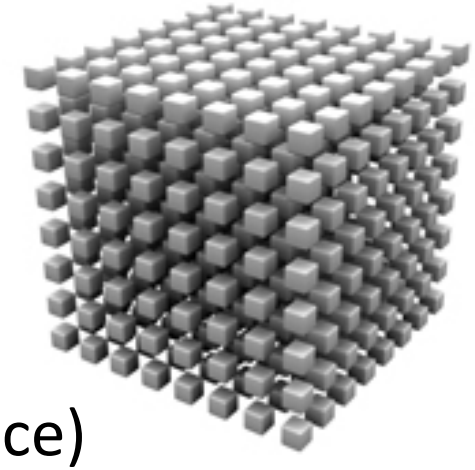
- MOLAP
  - DW is a proprietary system, tailored for multi-dimensional data manipulations
- Relational OLAP
  - Multi-dimensional data mapped onto tables, and manipulations mapped onto relational queries
- Hybrid OLAP
  - Relational systems extended with specific OLAP functionalities

Star Schema

- Fact table

  Sales(*Product_Id*, *Time_Id*, *City_id*, Amount)

  - Multi-dimensional data
  - Can be represented as a (hyper-)cube
    - 3 dimensions: Product, Time, City
    - The cube contains Amount values



- Dimensions

  Product(Product_id, Name, Category, Price)

  Space(City_id, City, Country, Region)

  Time(Time_id, Week, Month, Quarter)

  - Typically organized in a hierarchy

Drill Down and Roll Up

- Dimensions as aggregation hierarchy
- Drill down
  - Series of queries that moves down the aggregation hierarchy
  - E.g., per region, per country, per city
- Roll up
  - Series of queries that moves up the aggregation hierarchy
  - E.g., per week, per month, per year
- Same form of SQL query, different attributes
  - When **rolling up, query results can be re-used**
    - An aggregation can be used as a basis for an aggregation one or more levels up in the hierarchy

Pivoting

- Data as a cube which is pivoted so that a user can “see” its various faces
  - Pivoting on dimensions D1, D2, D3 means grouping by attributes from these dimensions
  - New pivot on D3, D2, D1
    - Interesting in case a visualization software is used to represent 3 dimensions as x, y, z in space
    - Interesting if there are N dimensions, and the pivot concerns a subset of these dimensions

Slicing and Dicing

- Slice
  - A value is given for a dimension attribute in the where clause
  - We take a “slice” of the cube
- Dice
  - Multiple values (or a range) are given for a dimension attribute in the where clause
  - We are dicing, i.e., reduce the size of, the original cube

Star Schema Operations

- Write the following sequence of queries in SQL on the sales star-schema
  - Original cube:
    - Sales amount per country, per week, per category
  - Roll-up on time
    - Sales amount per country, per month, per category
    - Sales amount per country, per year, per category
  - Drill-down on city
    - Sales amount per city, per year, per category
  - Pivot on product, time and space
    - Sales amount per category, per year, per city
  - Slice on year 2012
    - Sales amount per category, per city for 2012
  - Dice on the last three years
    - Sales amount per category, per city for 2010,2011,2012

Star Schema Operations

- What are the SQL queries you need to construct the following table

	Product 1	Product 2	Product 3	Total
City 1	520	230	100	850
City 2	10	15	10	35
City 3	1000	1200	1000	3200
Total	1530	1445	1110	4085

The CUBE Operator

SELECT *city_id*, *product_id*, SUM(amount) as sum_a
FROM SALES
GROUP BY CUBE (city_id, product_*id*)

- Defined by Jim Gray et al. in 1996
- Part of the SQL standard
- Supported in Oracle, DB2, SQL Server
- ROLAP implementation

City_id	Product_id	Sum_a
City1	Product 1	520
City1	product2	230
City1	product3	100
City1	ALL	850
City2	Product 1	10
City2	product2	15
City2	product3	10
City2	ALL	35
City3	Product 1	1000
City3	product2	1200
City3	product3	1000
City3	ALL	3200
ALL	Product 1	1530
ALL	product2	1445
ALL	product3	1110
ALL	ALL	4085

The ROLLUP Operator

SELECT *city_id*, *product_id*, SUM(amount) as sum_a

FROM SALES

GROUP BY city_id, product_*id*
with ROLLUP

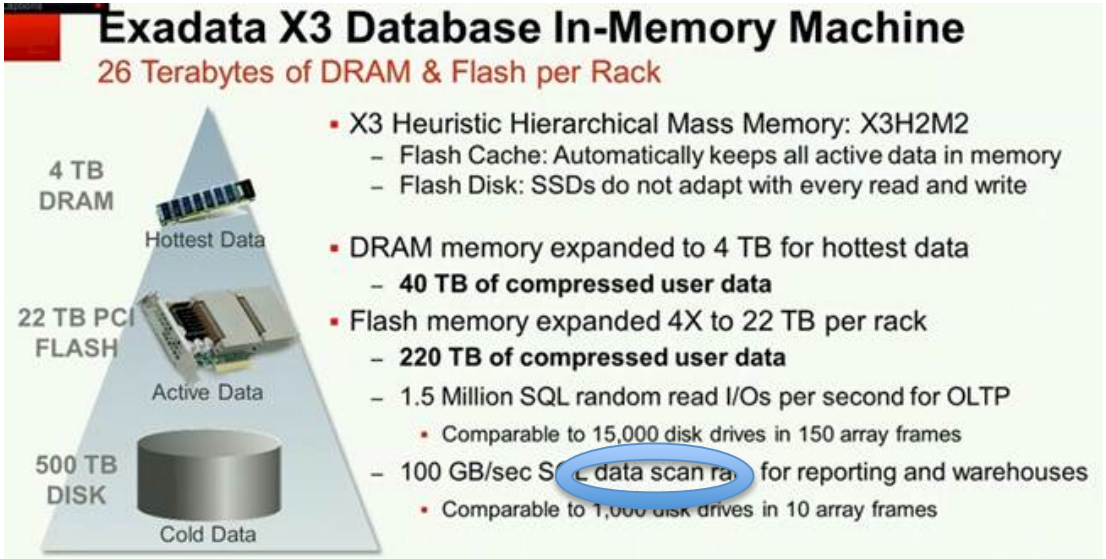
- Part of the SQL standard
- Supported in Oracle, DB2, SQL Server, MySQL

City_id	Product_id	Sum_a
City1	Product 1	520
City1	product2	230
City1	product3	100
City1	ALL	850
City2	Product 1	10
City2	product2	15
City2	product3	10
City2	ALL	35
City3	Product 1	1000
City3	product2	1200
City3	product3	1000
City3	ALL	3200
ALL	ALL	4085

Tuning the Cube

- Materialized Views
  - To materialize the original cube and the result of important cube manipulations (those that are re-used often)
- Indexes
  - Speeding up foreign-key/primary-key joins
    - Dimensions as index-organized tables (clustering index on primary key)
    - Non-clustered index on foreign key in fact table
  - Indexing low-cardinality attributes
    - Bitmap index (Oracle)
  - SQL Server columnstore indexes
- Compression
  - Speeding up scans, reducing DW footprint on 2nd storage
- Column-Oriented Representation
  - Great for slicing, dicing
  - Great for compression
  - Great for leveraging RAM, Processor cache
- Parallelism
  - Work on dimensions in parallel, Speeding up scans
  - Tuning degree of parallelism in ORACLE, DB2

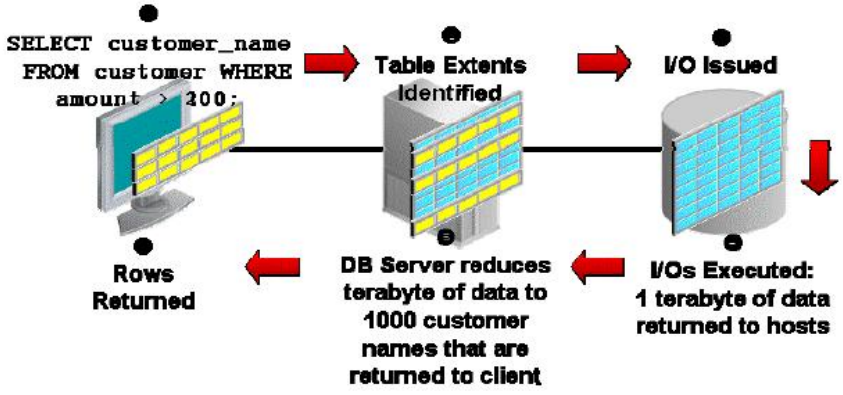
Data Warehouse Appliances

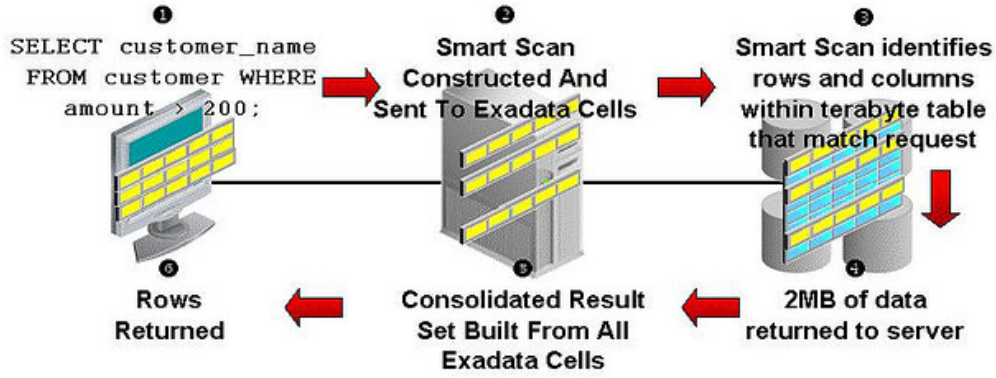


[Exadata Data Sheet]

Normal Table Scan vs. Exadata Smart Scan

[See B.Durrett's slides]





Column Stores

- Columnar representation
  - Compression & Scan efficiency
  - Tailored for RAM, processor cache utilization
- VectorWise
- SQL Server ColumnStore indexes

Data Mining 101

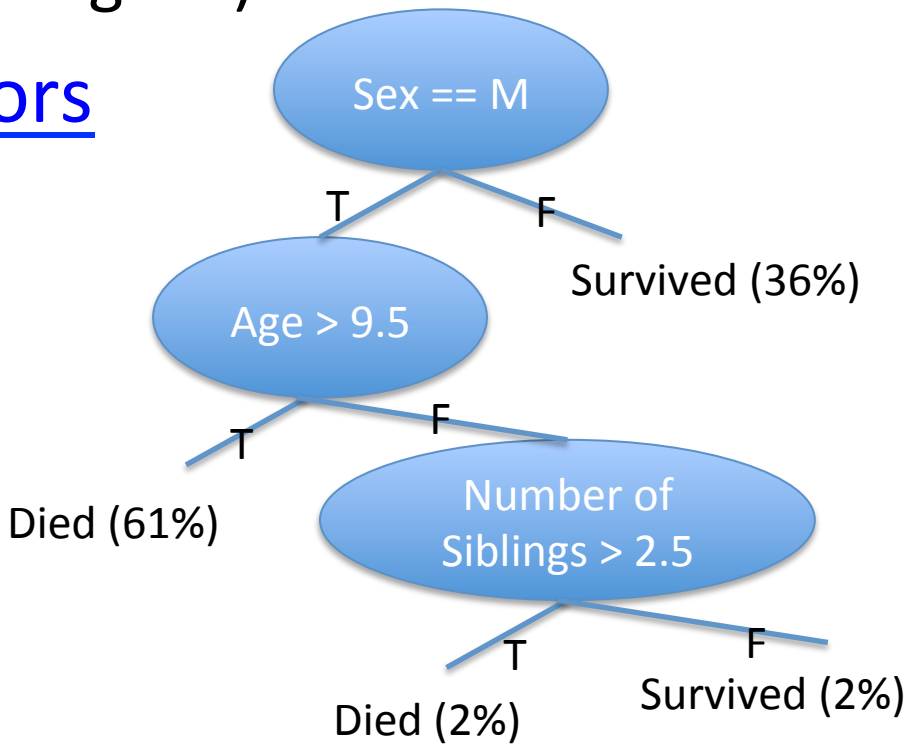
- Boundaries of Data Management, Statistics and Machine Learning
- Finding Patterns in Large Data Sets
  - Associations
    - Many buy Product1 and Product3 together
  - Classification
    - Given some predefined classes (e.g.., StaysInBusiness, GoesOutOfBusiness) train a classifier to distinguish in which class a store belongs based on its sales records
    - Might be used for prediction
  - Clustering
    - Like classfication but the classes are not given beforehand. They are discovered by the clustering algorithm.

Associations

- An association is a correlation between values in the same or different columns
  - Noted Predicate1 => Predicate2
  - Example: Purchases_Diaper => Purchases_Beer
- Confidence and Support
  - Confidence (rule): percentage of where Predicate2 is true when Predicate1 is true
  - Support (itemset): percentage of records where all attribute values needed by the rule are present
- Confidence and support must be over a given threshold so that an association holds

Classification

- Decision tree, Neural networks
  - Multiple variables analysis
  - Learning algorithm (training set)
- Example: Titanic survivors



Clustering

- K-means algorithm
  1. Each (of the given K) cluster is given a centroid
  2. Form clusters by assigning points to cluster with closest centroid (distance is defined)
  3. Recompute cluster centroid
  4. Repeat 2,3 until centroids do not move
- Other techniques
  - Hierarchical clustering (e.g., BIRCH), Support Vectors

Tuning for Mining

- Tuning Scans
  - Most algorithms require several passes over the data
  - Parallelism & compression
- Statistics features in systems
  - Predictive Analytics in SQL Server
  - Data mining features of DB2
  - Oracle DataMiner