Presentation:
OLAP and DW/BI Lifecycle, Part 1

Erik Perjons

DSV, Stockholm University

OLAP

OLAP – as an approach

- **OLAP (Online Analytical Processing) is an interactive approach for analyzing data from multidimensional perspectives/views/dimensions**
- OLAP is often contrasted to OLTP (Online Transaction Processing).
  - **OLAP** - aims to answer complex queries for business analysis, most often involving several transactions in order to provide aggregated information about the business. OLAP also aims to answer such complex queries quickly. Moreover, OLAP is optimized for answering such queries and not for operations such as update, delete and insert
  - **OLTP** - aims to process large volume of non-complex queries (including create, read, update, delete, insert) supporting operational business processes. OLAP queries often involve only one transaction at a time

OLAP – as an approach

- **OLAP use three basic operations** for analyzing data:
  - **roll-up/consolidation** (aggregation of base data computed in one or more perspectives/views/dimensions, usually by climbing up hierarchy or by dimension reduction)
  - **drill-down** (navigate from higher level summary to lower level summary or detailed data, or introducing new dimensions)
  - **slicing and dicing** (slicing/take out a specific set of data and dicing/view the slices from different perspectives/views/dimensions)

OLAP – and the cube

- **OLAP is closely related to the concept of a data cube**
- A **concept of the cube provide a way of structure the data** and express relationships between the data using **numeric facts/measures** and **dimensions/viewpoints**
- **The dimensions provide the context or metadata (or label) to the numeric facts/measures.**
- The **numeric facts/measures are placed in the intersection** (in the cell of a cube) **of user selected dimensional attributes and values**

OLAP – and the cube

Stockholms universitet

An example:

- Sales amount $ 30 (fact, measure) does not say so much.
- Dimensional attributes and values are added to give the $30 meaning, such as Month of Dec 2018 (dimensional attribute and value) and Store in the city of Falun, Sweden (dimensional attribute and value).
- That is $ 30 is the amout sales during Dec 2018 in the store in Falun.

OLAP – and the cube

- The **use of pre-aggregated measure** are an important part of OLAP cube
- This pre-aggregated measures is a **mechanism that provide rapid answers to complex analytic queries**
- An **aggregate function** is used to aggregate base data. This is called loading or building och processing the cube

OLAP – and the cube

- The aggregate function **aggregate data along dimensional attributes, organized in hierarchies** (such as date, month, and year in the date dimension, or product, and product category in the product dimension)
- **These loading or building the OLAP cube** with aggregated data is done in so called **MOLAP servers.**

OLAP - and the data model

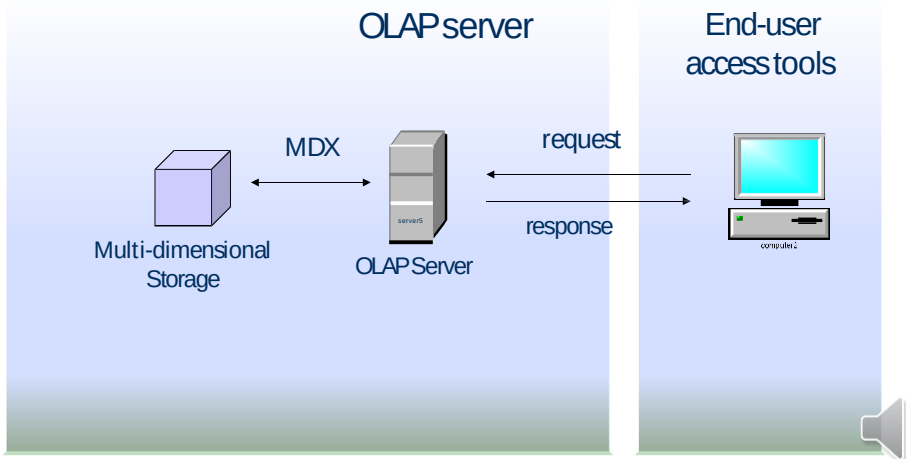
- **OLAP is supported by a specific data model** used by databases, called the **dimensional data model** (or the multidimensional data model)
- The dimensional data model makes it possible to answers complex and ad hoc queries rapidly
- The dimensional data model structures the data, as in the concept of cube described earlier, in facts (also called measures) and dimensions (context or metadata to the facts)

OLAP - as a system

- There are three basic forms of OLAP systems: ROLAP, MOLAP and HOLAP

Multidimensional OLAP (MOLAP)

- MOLAP (Multidimensional OLAP) – is a OLAP system that support the OLAP operations (that is, Roll up, Drill down and Slicing and Dicing) by loading data in a multidimensional array, called the MOLAP cube or data cube



Multidimensional OLAP (MOLAP)

Stockholms universitet

- MOLAP store data in an array that provide very fast answers to queries due to:
  - multidimensional indexing (natural indexes in MOLAP since the indexes can be part of the array)
  - pre-calculating aggregrations (carried out in an automated way during data load, based on the dimensional hierarchies). The result of such pre-calculating aggregrations is sometimes refered to as the data cube or OLAP cube. The data cube consist of all possible answers given a certain range of queries

Benefits of MOLAP

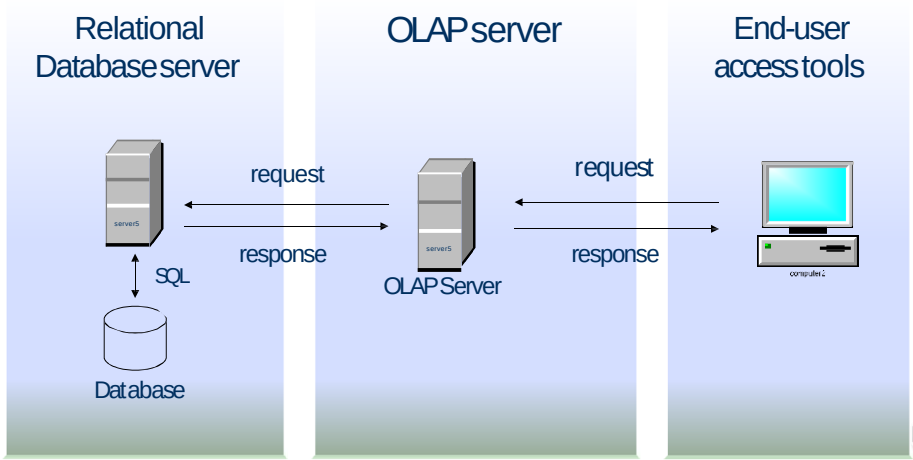
- Rapid answer to queries
- Less storage space due to smart compression techniques
- Possibility to automated the aggregation function, thereby perform pre-calulation of aggregations, that is loading the data cube or OLAP cube

Drawbacks of MOLAP

- The load of data into the cube is time consuming, especially if the data volumes are large (however, this can be manage by just updating the last changed data)

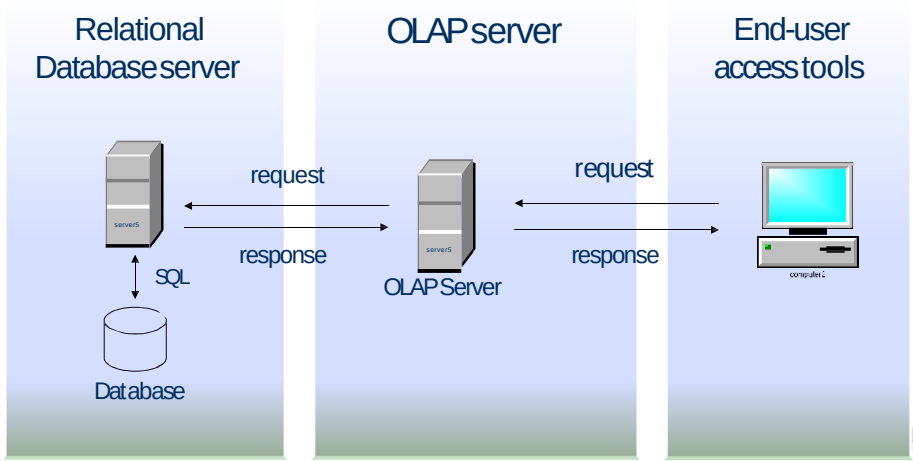
Relational OLAP (ROLAP)

- ROLAP (Relational OLAP) – is a OLAP system that support the OLAP operations (that is, Roll up, Drill down and Slicing and Dicing) by using a dimensional data model in a relational database



Relational OLAP (ROLAP)

- That is, ROLAP, manipulate the data stored in a relational database so that it looks like a MOLAP solution



Relational OLAP (ROLAP)

- ROLAP provide OLAP operations (Slicing and Dicining) by adding ”WHERE” clausuls in SQL statements
- ROLAP does not use pre-calculated data cubes

Relational OLAP (ROLAP)

- ROLAP can provide fast aggregated result (Roll up) by either:
  - carry out aggregation on the fly by using the dimensional data model, and calculate the facts given the dimensional attributes and attribute values selected
  - make use of pre-aggregated data by using the dimensional data model again, but this time adding new fact tables with pre-aggregrated data. Also some new dimensional tables need to be added, for exampel, a dimensional table where the grain is based on month and not date

Benefits of ROLAP

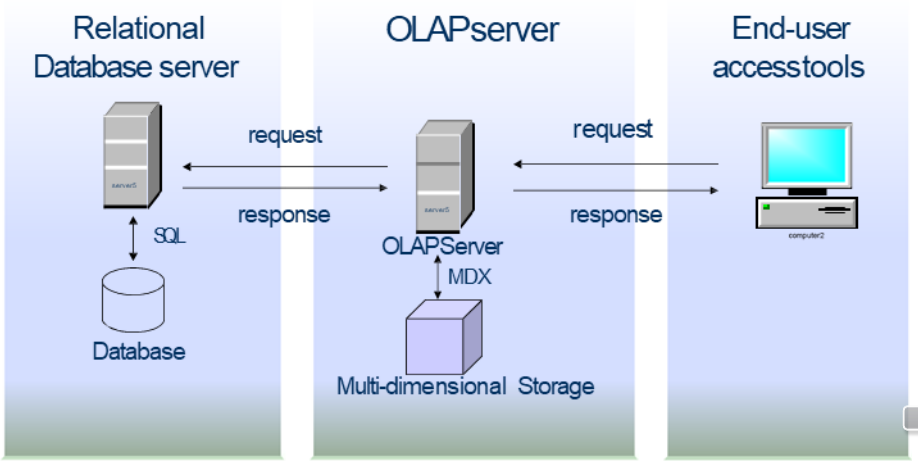
- ROLAP solutions scale better than MOLAP if the data volumes are large
- No cube to load, as in MOLAP. Such data load in MOLAP is time consuming, especially if the data volumes are large
- SQL tools can be used for manipulating the data since data is stored in a relational database
- Datamodelling become simpler in ROLAP if the data do not fit in the dimensional data model

Drawbacks of ROLAP

- Lower query performance than MOLAP
- Preaggregation must be manage via the ETL process (you do not load a cube), which require more time to develop the ETL system
- It is not practical to create a large number of pre-aggregregated fact and dimensional tables to store pre-calulated aggregation

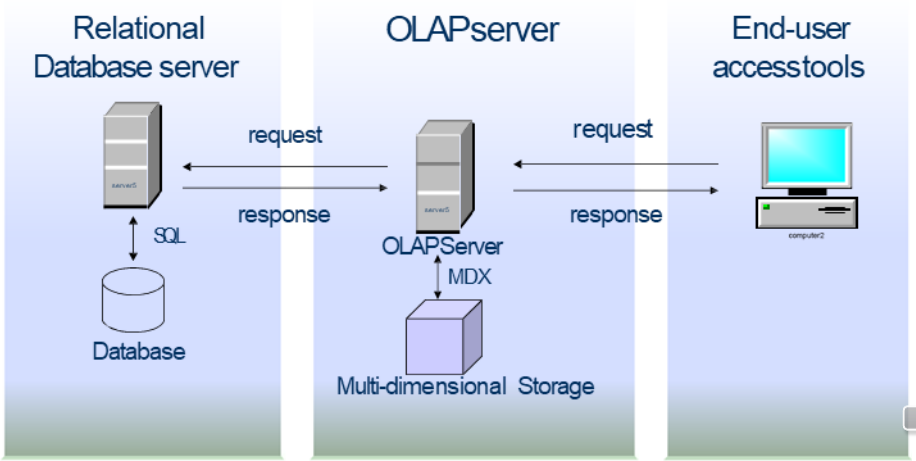
Hybrid OLAP (HOLAP)

- HOLAP (Hybrid OLAP) is a OLAP system that combine both MOLAP and ROLAP functionalities, that is, use both relational database sources for data (ROLAP) and pre-calculated cubes of data (MOLAP)



Hybrid OLAP (HOLAP)

- The designers/developers can select which data to be stored in the MOLAP solution and which data to be stored in the ROLAP solution



Hybrid OLAP (HOLAP)

Stockholms universitet

Two strategies for combining MOLAP and ROLAP solutions:

- Vertical partitioning:
  - HOLAP store aggregations in MOLAP – for fast querying
  - HOLAP store detailed data in ROLAP – for lessen the time to load data in a cube by not loading detailed data in the cube, and for better scalability by using ROLAP

- Horizontal partitioning
  - HOLAP store most recent data (or some other slice of data) in MOLAP – for fast querying
  - HOLAP store the rest of the older data i ROLAP – for lessen the time to load data in a cube by not loading detailed data in the cube, and for better scalability by using ROLAP

Aggregates

What is aggregates?

- Aggregates are pre-calculated summary data

Why aggregates?

- Aggregates are introduced to answer complex queries more rapidly

Different types of aggregations

Stockholms universitet

MOLAP aggregates

- pre-calculated aggregrates in the cube (pre-calulated during data load of the cube)

ROLAP aggregates

- Pre-calculated aggregrates stored in new aggregated fact tables in star schemas (pre-calculated during the ETL process)

ROLAP – Aggregation on the fly

- Aggregations on the fly (summing up additive and semiadditive facts using the star schema structure)

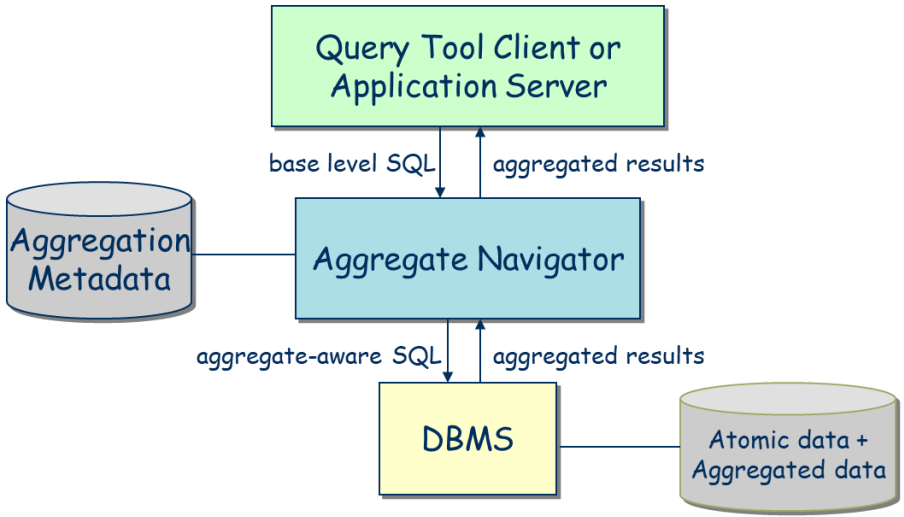
Aggregated fact tables (ROLAP)

Stockholms universitet

- Pre-calculated aggregates are stored in new aggregated fact tables in star schemas (pre-calculated during the ETL process)
- These aggregate are stored together with atomic base fact tables in star schemas
- That is, in a dimensional data model there will be both star schemas with aggregated fact tables and star schemas with atomic fact tables
- Also new, so called, rolled up dimensional tables, need to be added to fit the aggrated fact tables. Rolled up dimensions are shrunken versions of the dimensions associated with the granular base facts.

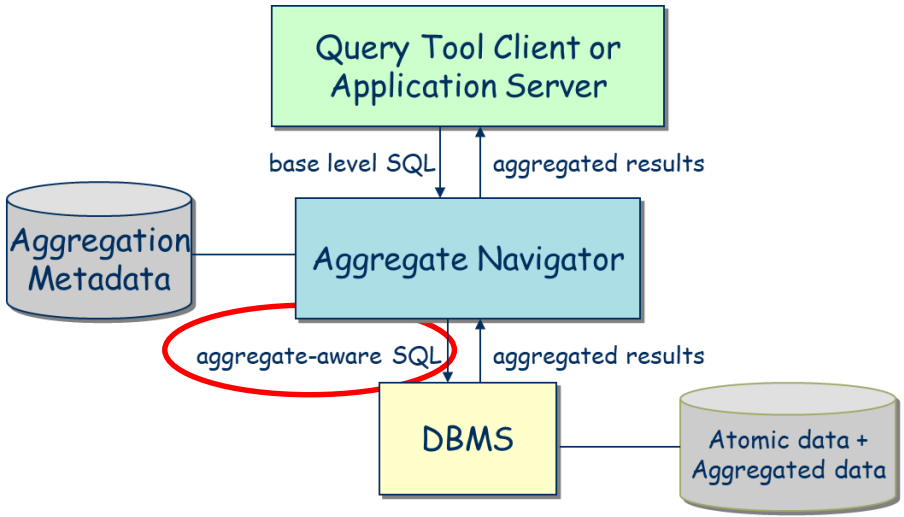
Aggregation navigator

- An **aggregation navigator** is a software that are using **aggregation metadata** to identify if there exist **aggregated tables** that could handle a query or if **atomic base tables** are need for handling the query



Aggregation navigator

- If the **aggregation navigator** identify – using aggregation metadata - that a SQL query could be send to aggregrated tables, the aggregation navigator will automatically change the SQL code to fit the aggregrated tables

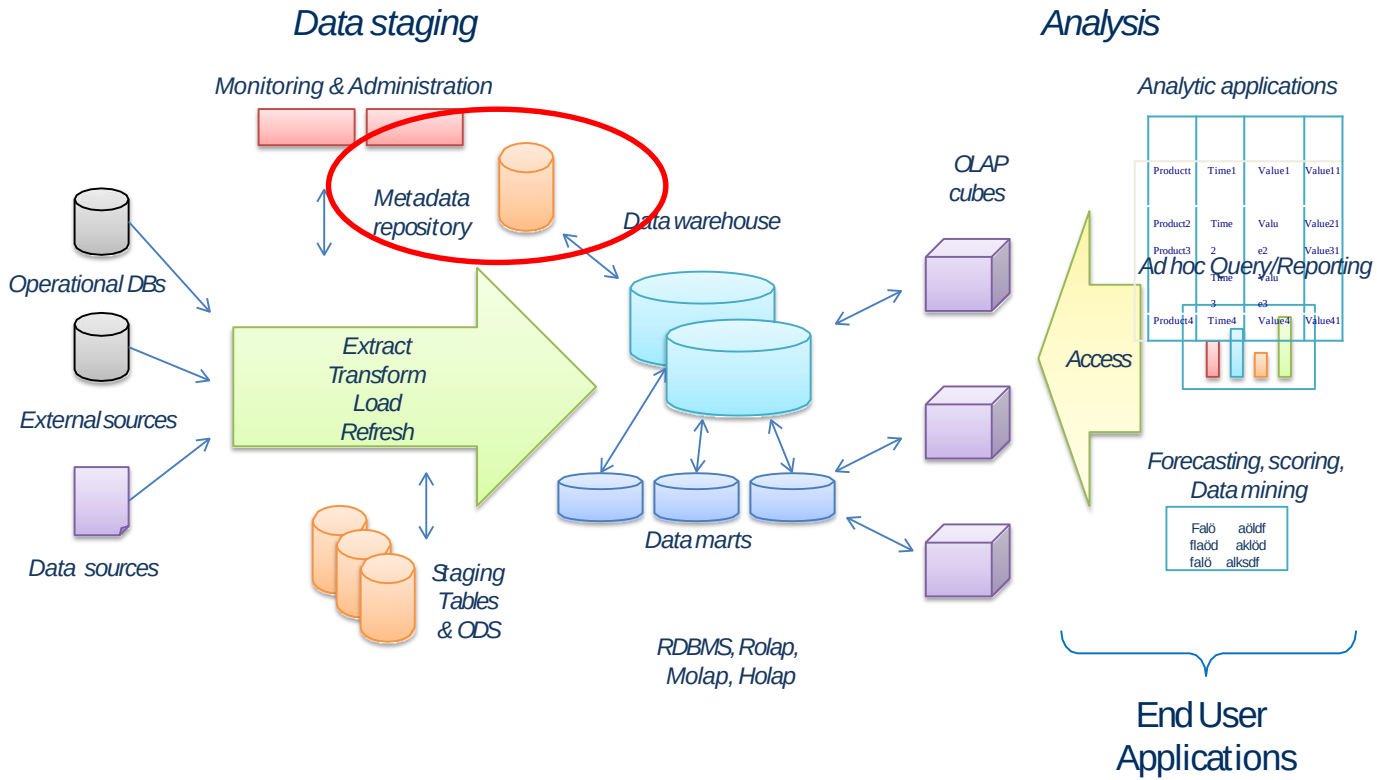


HOLAP solutions

- A similar solution as the aggregation navigator could be used for deciding if a query should be send to a MOLAP or a ROLAP system in a HOLAP solution

Metadata repository

Metadata repository



Metadata repository

- A metadata repository is integrated complete source of metadata
- The metadata repository is at the heart of the data warehouse architecture
- The metadata repository supports the information needs of system developers, data administrators, system administrators, users, applications of the data warehouse
- The metadata repository must always be up to date

Different types of metadata 1(2)

- **Administrative metadata** - information necessary for setting up and managing the data warehourse:
    - source databases
    - fact and dimension tables
    - aggregates
    - hierarchies
    - predefined queries
    - physical organisation
    - rules, script, staging tables for ETL
    - back-end and front end tools

Different types of metadata 2(2)

- **Business metadata**:
    - business terms and definitions
    - ownership of data
    - quality rules
    - etc
- **Operational metadata** - information collected during the operations of the DW:
    - usage statistics
    - error reports

Index

What is a database index?

- A database index is a data structure that speed up the data retrieval operations in database tables
- A database index is a type of copy of the data in database table columns and this copy can be searched efficiently
- The index also contain a link or address to the real data in table column

Why using indexes?

- Without indexes, the DBMS may be forced to conduct a full table scan, i.e., reading every row in the table to locate the right data.
- Instead, index scans are much more efficient than searches of data base tables

Indexes used in DW and OLAP

- The increased focus on DW and OLAP and complex queries has resulted in the development and use of “bit map indexes” and “join indexes”

Bitmap Index

Stockholms universitet

Base Table

Cust	Region	Rating
C1	N	H
C2	S	M
C3	W	L
C4	W	H
C5	S	L
C6	W	L
C7	W	H

Region Index

RowId	N	S	E	W
1	1	0	0	0
2	0	1	0	0
3	0	0	0	1
4	0	0	0	1
5	0	1	0	0
6	0	0	0	1
7	0	0	0	1

Rating Index

RowId	H	M	L
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	0	0	1
6	0	0	1
7	1	0	0

SELECT Cust
FROM Base Table
WHERE Region = W AND Rating = L

Bitmap Index

Stockholms universitet

Base Table

Cust	Region	Rating
C1	N	H
C2	S	M
C3	W	L
C4	W	H
C5	S	L
C6	W	L
C7	W	H

Region Index

RowId	N	S	E	W
1	1	0	0	0
2	0	1	0	0
3	0	0	0	1
4	0	0	0	1
5	0	1	0	0
6	0	0	0	1
7	0	0	0	1

Rating Index

RowId	H	M	L
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	0	0	1
6	0	0	1
7	1	0	0

Region = W AND Rating = L

Bitmap Index

Stockholms universitet

Base Table

Cust	Region	Rating
C1	N	H
C2	S	M
C3	W	L
C4	W	H
C5	S	L
C6	W	L
C7	W	H

Region Index

RowId	N	S	E	W
1	1	0	0	0
2	0	1	0	0
3	0	0	0	1
4	0	0	0	1
5	0	1	0	0
6	0	0	0	1
7	0	0	0	1

Rating Index

RowId	H	M	L
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	0	0	1
6	0	0	1
7	1	0	0

Region = W AND Rating = L

Bitmap Index

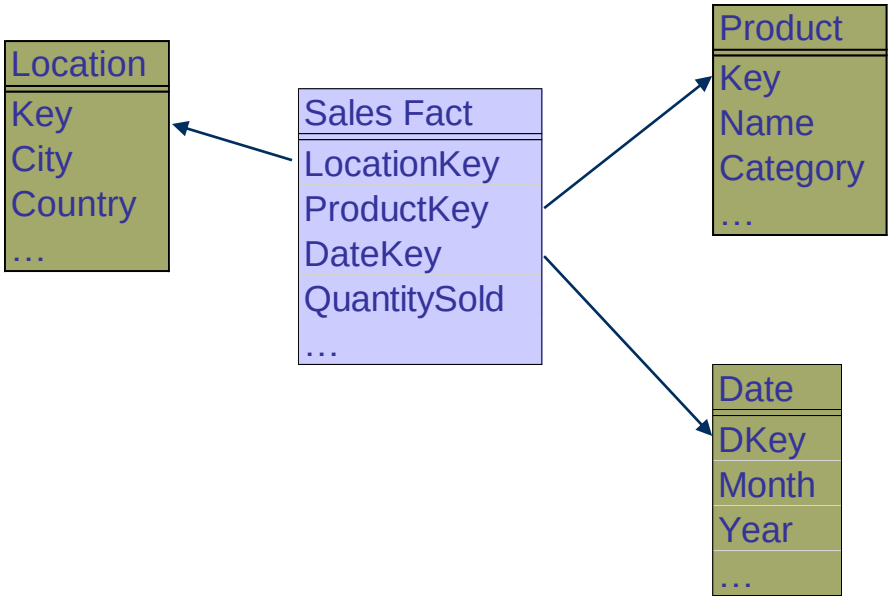
- An effective indexing technique for attributes with low-cardinality domains
- There is a distinct bit vector BV for each value V of the domain

Join Index

Stockholms universitet

- Providing good performance from queries that require multitable joins of large tables has been an ongoing challenge for DBMS architects
- Join index can be seen as a ”pre-computed” join. That is, the indexes consist of references to those rows in two or more tables that satified the join condition

Example


Location
Key
City
Country
…
Sales Fact
LocationKey
ProductKey
DateKey
QuantitySold
…
Product
Key
Name
Category
…
Date
DKey
Month
Year
…

Join Index - Ex

	Location Dimension		
	Key	City	…
rid1	1	Stockholm	…
rid2	2	London	…
rid3	3	Paris	…

	Product Dimension		
	Key	Name	…
rid22	1	# 5	…
rid23	2	Noah	…
rid24	3	Opium	…

	Sales Fact			
	LKey	PKey	DKey	Qnt
rid4	1	1	1	5
rid5	1	2	1	7
rid6	1	3	1	4
rid7	2	1	1	8
rid8	2	2	1	3
rid9	2	3	1	5
rid10	3	1	1	20
rid11	3	2	1	10
rid12	3	3	1	30
rid13	1	1	2	10
rid14	1	2	2	9
rid15	1	3	2	7
rid16	2	1	2	5
rid17	2	2	2	10
rid18	2	3	2	8
rid19	3	1	2	20
rid20	3	2	2	50
rid21	3	3	2	30

Stockholms universitet

Join Index - Ex

	Location Dimension		
	Key	City	…
rid1	1	Stockholm	…
rid2	2	London	…
rid3	3	Paris	…

	Product Dimension		
	Key	Name	…
rid22	1	# 5	…
rid23	2	Noah	…
rid24	3	Opium	…

	Sales Fact			
	LKey	PKey	DKey	Qnt
rid4	1	1	1	5
rid5	1	2	1	7
rid6	1	3	1	4
rid7	2	1	1	8
rid8	2	2	1	3
rid9	2	3	1	5
rid10	3	1	1	20
rid11	3	2	1	10
rid12	3	3	1	30
rid13	1	1	2	10
rid14	1	2	2	9
rid15	1	3	2	7
rid16	2	1	2	5
rid17	2	2	2	10
rid18	2	3	2	8
rid19	3	1	2	20
rid20	3	2	2	50
rid21	3	3	2	30

CityJI

CityK	Rid
1	rid4
1	rid5
1	rid6
1	rid13
1	rid14
1	rid15
2	rid7
2	rid8
2	rid9
2	rid16
2	rid17
2	rid18
…	

Join Index - Ex

	Location Dimension		
	Key	City	…
rid1	1	Stockholm	…
rid2	2	London	…
rid3	3	Paris	…

	Product Dimension		
	Key	Name	…
rid22	1	# 5	…
rid23	2	Noah	…
rid24	3	Opium	…

	Sales Fact			
	LKey	PKey	DKey	Qnt
rid4	1	1	1	5
rid5	1	2	1	7
rid6	1	3	1	4
rid7	2	1	1	8
rid8	2	2	1	3
rid9	2	3	1	5
rid10	3	1	1	20
rid11	3	2	1	10
rid12	3	3	1	30
rid13	1	1	2	10
rid14	1	2	2	9
rid15	1	3	2	7
rid16	2	1	2	5
rid17	2	2	2	10
rid18	2	3	2	8
rid19	3	1	2	20
rid20	3	2	2	50
rid21	3	3	2	30

City-Product JI

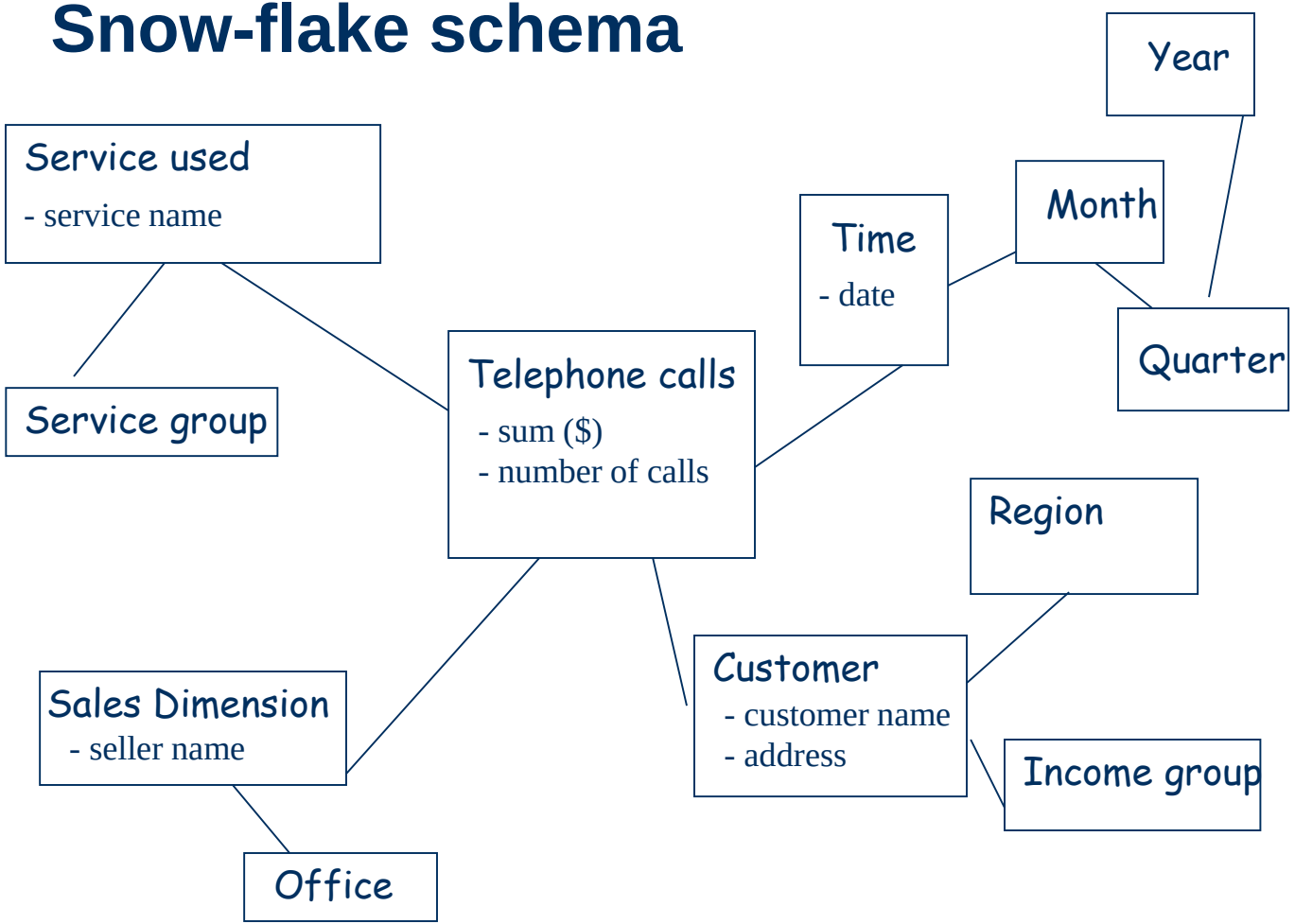
CityK	PrdK	Rid
1	1	rid4
1	1	rid13
1	2	rid5
1	2	rid14
1	3	rid6
1	3	rid15
…		

Snowflake schema

What is a snowflake schema?

- A snowflake schema is a star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake

Snow-flake schema



Kimball and snowflaking

Stockholms universitet

Kimball’s critics against snowflaking:

- a more complex structure
- numerous tables and joins usually translate into slower performance
- insignificant disk space savings
- slow down the users’ ability to brows within the dimension
- defeat the use of bit map indexes

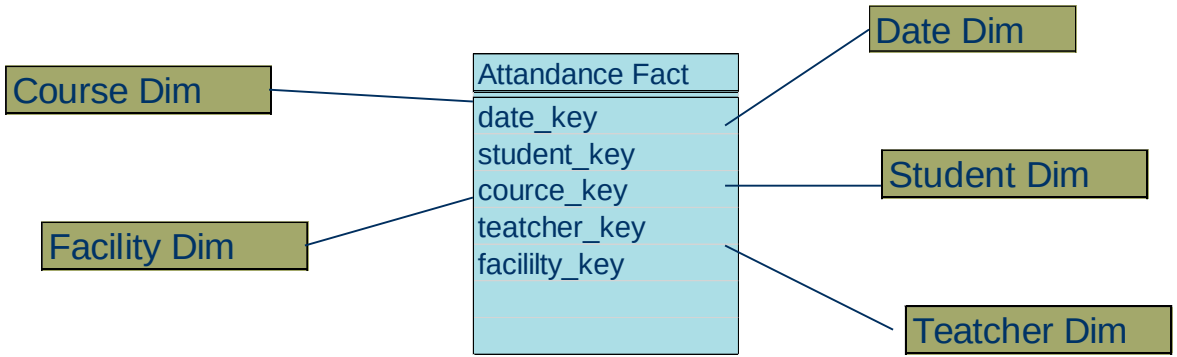
Coverage table

Factless fact tables

- Some fact tables quite simply have no measured facts
- These fact tables are useful to describe events and coverage, i.e. the tables contain information that something has (event tracking) or has not (coverage table) happened
- There are several types of factless fact tables, two of the most common are:
  - event tracking tables
  - coverage tables

Event tracking (factless fact) tables

- An event tracking table - records events, e.g. records every time a student attends a course (see figure), or people involved in accidents and vehicles involved in accidents



Coverage (factless fact) tables

- Covarage tables is a table that contain information of what has not happened

Coverage (factless fact) table

- An example: What products were on promotion but did not sell?
  - The sales fact table records only the SKUs actually sold.
  - Therefore, we need to create a factless fact table that cover all product that is part of the promotion