Multi-Dimensional Database Modeling and Querying: Methods, Experiences and Challenging Problems

Part II
Multi-dimensional Benchmark Design

Alfredo Cuzzocrea University of Trieste & ICAR
Rim Moussa University of Carthage & LaTICE

17th of November, 2016

The 35th Intl. Conference on Conceptual Modeling
@ Gifu, JAPAN

Tutorial Outline

- Introduction
- Part I: State-of-the-Art
- Part II: Experiences
  - TPC-H*d Experience
  - *AutoMDB*
  - *TPC-DS*d*
- Part III: Challenging Problems
- Conclusion

Problem

Given,

- A relational Warehouse schema
- A Workload -a set of SQL Statements,

  W = {Q1, Q2, …, Qn}

  where Qi is a parameterized query

  - **How to design the Multi-dimensional DB Schema?**
  - **How to define cubes?**
    - **Will there be a single cube or multiple cubes? Are there any rules for merging of cubes? Are there any rules for definition of virtual cubes?**
  - **Which Optimizations are suitable for performance tuning ?**
    - **Derived data calculus & refresh?**
    - **Data partitioning & parallel cube building?**

Idea

- Map each business question to an OLAP cube

  >> Obtain a multi-dimensional DB schema

- Recommend & Test Optimizations

  >> Derived Data

  >> Data partitioning

  >> Cube Merging

SQL Statement Template

```
SELECT  t1.col_a, t1.col_b, …, tn.col_a, tn.col_z,
  aggregate_function(column) as measure_1, …,
  aggregate_function(expression) as measure_m
FROM table_1  t1, table_2  t2, …, table_n tn
WHERE   ti.col_x  operator $query_parameter$
      AND   ti.col_y = tj.col_z
      AND  …
GROUP BY t1.col_a, t1.col_b, …, tn.col_a, tn.col_z
```

aggregate_function: min, max, sum, avg, count, count-distinct …
Operator: =, < , <=, >=, !=

OLAP Cube Design: Measures' Definition

- Measures feature aggregate functions,
  - e.g. `min, max, count, count-distinct, sum, average, …`
- Simple Measure
  - Defined over a single attribute,
  - e.g. `SUM(l_extendedprice),`
- Measure expressions
  - Involve more than one attribute,
  - e.g. `SUM(l_extendedprice*(1 - l_discount))`
- Computed Members
  - Involve already defined measures or measure expressions,
  - e.g. `M1=SUM(l_extendedprice), M2=COUNT(l_orderkey),`

    `CM = M1 / M2`

OLAP Cube Design
Fact Table Definition (1/9)

- All attributes involved in measures and measure expressions belong to the fact table,
- Example: Q10 of TPC-H benchmark

```
SELECT c_custkey,c_name,c_acctbal, n_name, c_address, c_phone,
       c_comment, SUM(l_extendedprice*(1-l_discount)) as rev
FROM customer, orders, lineitem, nation
WHERE c_custkey = o_custkey
  AND l_orderkey = o_orderkey
  AND o_orderdate >= date '[DATE]'
  AND o_orderdate < date '[DATE]' + '3' month
  AND l_returnflag = 'R'
  AND c_nationkey = n_nationkey
GROUP BY c_custkey,c_name,c_acctbal,c_phone,n_name,c_address,c_comment
ORDER BY revenue desc;
```

OLAP Cube Design
Fact Table Definition over multiple tables (2/9)

- Case measurable attributes belong to different tables then the fact table is a view defined as the join of relations, to which attributes belong!

  - Example 1: Q9 of TPC-H benchmark, where `l_extendedprice`, `l_discount` and `l_quantity` belong to `lineitem`, and `ps_supplycost` belongs to `partsupp`.

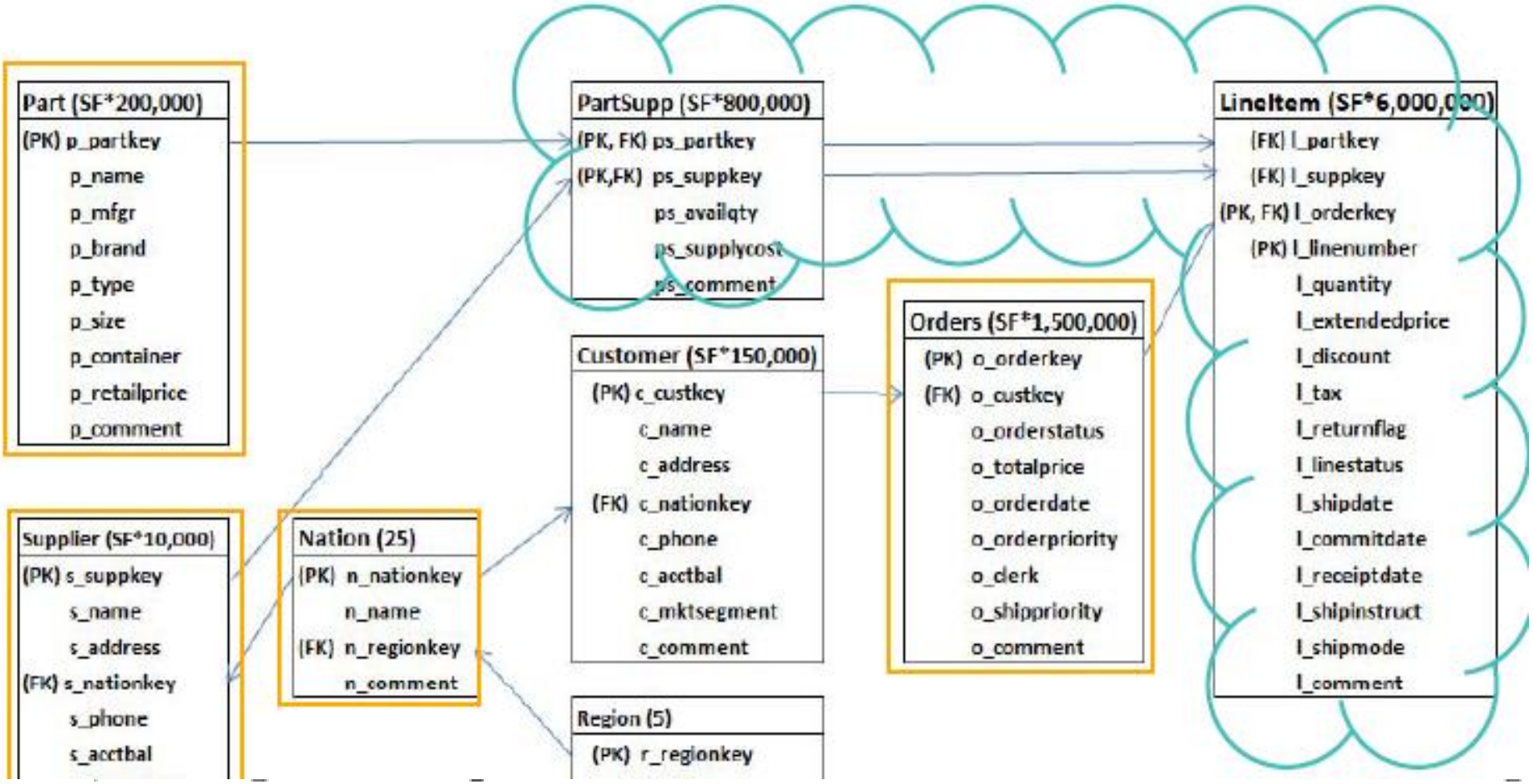
  - The fact table is the join of `lineitem` and `partsupp` tables. Only attributes needed for join with dimension tables (namely, `l_partkey`, `l_orderkey`, `l_suppkey`), and measurable attributes (namely `l_extendedprice`, `l_discount`, `l_quantity`, `ps_supplycost`) are selected.

OLAP Cube Design
Fact Table Definition over multiple tables (3/9)

Q9 SQL statement

```
SELECT nation, o_year, sum(amt) as sum_profit
FROM (SELECT  n_name as nation,
              extract(year from o_orderdate) as o_year,
              l_extprice * (1 - l_disc)-ps_suppcost * l_qty as amt
      FROM part, supplier, lineitem, partsupp, orders, nation
      WHERE s_suppkey = l_suppkey
        AND ps_suppkey = l_suppkey
        AND ps_partkey = l_partkey
        AND p_partkey = l_partkey
        AND o_orderkey = l_orderkey
        AND s_nationkey = n_nationkey
        AND p_name like '%[COLOR]%') as profit
GROUP BY nation, o_year
ORDER BY nation, o_year desc;
```

OLAP Cube Design
Fact Table Definition over multiple tables (4/9)



$$\prod_{\{l\_partkey, l\_ext.price, l\_disc, l\_qty, l\_orderkey, l\_suppkey, ps\_suppcost\}} LINEITEM \bowtie_{l\_suppkey=ps\_suppkey \; and l\_partkey=ps\_partkey} PARTSUPP$$

OLAP Cube Design
Fact Table Definition over multiple tables (5/9)

O14 SOL statement

```
SELECT 100.00 * sum(case when p_type like 'PROMO%'
                    then l_extendedprice*(1-l_discount)
                    else 0 end)
                / sum(l_extendedprice * (1 - l_discount)) as promo_rev
FROM lineitem, part
WHERE l_partkey = p_partkey
AND l_shipdate >= [DATE]
AND l_shipdate < [DATE] + interval '1' month;
```

- Example 2: Q14 of TPC-H benchmark, where `l_extendedprice`, `l_discount` and belong to `lineitem`, and `p_type` belongs to `part`.

- The fact table is the join of `lineitem` and `part` tables

OLAP Cube Design
Fact Table Definition and filters' processing (6/9)

- Filters Processing: the fact table is defined as a view of facts with filters
  - » Extract all filters involving the fact table from the WHERE clause, such as
    - (`attr_i operator attr_j`), where both `attr_i` and `attr_j` belong the fact table,
    - (`attr_k operator $value$`), such that `attr_k` belongs to the fact table,
    - `[not] exists (select … from … where attr_k …)`, such that `attr_k` belongs to the fact table,
    - `attr_k [not] in (list of values)`, such that `attr_k` belongs to the fact table,
- Example 1: Q10 of TPC-H benchmark
- Example 2: Q16 of TPC-H benchmark
- Example 3: Q21 of TPC-H benchmark

OLAP Cube Design
Fact Table Definition and filters' processing (7/9)

Q10 SQL statement

```
SELECT c_custkey,c_name,c_acctbal, n_name, c_address, c_phone,
       c_comment, SUM(l_extendedprice*(1-l_discount)) as rev
FROM customer, orders, lineitem, nation
WHERE c_custkey = o_custkey
  AND l_orderkey = o_orderkey
  AND o_orderdate >= date '[DATE]'
  AND o_orderdate < date '[DATE]' + '3' month
  AND l_returnflag = 'R'
  AND c_nationkey = n_nationkey
GROUP BY c_custkey,c_name,c_acctbal,c_phone,n_name,c_address,c_comment
ORDER BY revenue desc;
```

Q16 SQL statement

```
SELECT p_brand, p_type, p_size, count(distinct ps_suppkey) as supp_cnt
FROM partsupp, part
WHERE p_partkey = ps_partkey
AND p_brand <> ['Brand']
AND p_type not like ['Type%']
AND p_size in ([S1], [S2], [S3], [S4], [S5], [S6], [S7], [S8])
AND ps_suppkey not in (select s_suppkey
                       from supplier
                       where s_comment like '%Customer%Complaints%')
GROUP BY p_brand, p_type, p_size
ORDER BY supplier_cnt desc, p_brand, p_type, p_size;
```

OLAP Cube Design
Fact Table Definition and filters' processing (9/9)

Q21 SQL statement

```
SELECT n_name, s_name, count(*) as numwait
FROM supplier, lineitem l1, orders, nation
WHERE s_suppkey = l1.l_suppkey
AND o_orderkey = l1.l_orderkey
AND o_orderstatus = 'F'
AND l1.l_receiptdate > l1.l_commitdate
AND exists ( select *
             from lineitem l2
             where l2.l_orderkey = l1.l_orderkey
               and l2.l_suppkey <> l1.l_suppkey)
AND not exists ( select *
                 from lineitem l3
                 where l3.l_orderkey = l1.l_orderkey
                   and l3.l_suppkey <> l1.l_suppkey
                   and l3.l_receiptdate > l3.l_commitdate)
AND s_nationkey = n_nationkey
GROUP BY n_name, s_name
ORDER BY n_name, numwait desc, s_name;
```

OLAP Cube Design
Dimension Definition (1/7)

- **First**, consider all attributes in the SELECT, WHERE and GROUP BY clauses,
  - Discard measurable attributes, which figure out in measures, measure expressions, or computed members,
  - Discard attributes which figure out in the WHERE clause, and are used for joining tables or filtering the fact table with static values,
  - Compose time dimension along well known hierarchies,
    - Year, quarter, month
  - Compose geography dimension along well known hierarchies,
    - Region, nation, city

OLAP Cube Design
Dimension Definition (2/7)

- Example: Q10 of TPC-H benchmark
  - All highlighted attributes are considered for dimensions' mount
  - Time dimension *o_orderdate* requires *order_year* and *order_quarter* levels

```
SELECT c_custkey,c_name,c_acctbal, n_name, c_address, c_phone,
       c_comment, SUM(l_extendedprice*(1-l_discount)) as rev
FROM customer, orders, lineitem, nation
WHERE c_custkey = o_custkey
  AND l_orderkey = o_orderkey
  AND o_orderdate >= date '[DATE]'
  AND o_orderdate < date '[DATE]' + '3' month
  AND l_returnflag = 'R'
  AND c_nationkey = n_nationkey
GROUP BY c_custkey,c_name,c_acctbal,c_phone,n_name,c_address,c_comment
ORDER BY revenue desc;
```

OLAP Cube Design
Dimension Definition (3/7)

- **Second**, find out hierarchical relations, i.e., one-to-many relationships, and re-organize attributes along hierarchies to form dimensions' hierarchies,
  - Example: Q10 of TPC-H benchmark
    - each customer can be related to at most one nation, but a nation may be related to many customers,

      customer_dim:

      Customer nation → n_name

      Customer details→c_custkey, c_name, c_acctbal, c_address, c_phone, c_comment
    - **order_dim**

      order_year

      order_quarter

OLAP Cube Design
Dimension Definition (3/7)

- Third, distinguish levels from properties.
  - Properties are in functional dependency with levels,
  - Example: Q10 of TPC-H benchmark
    - For customer_dim, `c_custkey` is the level, and all of `c_name`, `c_acctbal`, `c_address`, `c_phone`, `c_comment` attributes are properties of `c_custkey` level.

OLAP Cube Design
Dimension Definition and Filters' processing (4/7)

- Filters Processing: not all tuples in the dimension table should be considered, so we have to extract filters defined over dimension tables from the WHERE clause not useful for multi-dimensional design,
  - Exple 1: Q12 of TPC-H benchmark
    - For each line shipping mode, year, Count the number of high priority orders (high line count) and the number of not high priority orders (low line count) over orders' facts and consider only lines such as `l_commit_date < l_receipt_date` and `l_ship_date < l_commit_date.` These are filters over dimension table.
  - Exple 2: Q19 of TPC-H benchmark
    - Calculate revenue for particular parts

OLAP Cube Design
Dimension Definition and Filters' processing (5/7)

Example 1: Q12 of TPC-H Benchmark

```
SELECT l_shipmode,
    SUM(case when o_orderpriority ='1-URGENT' or ='2-HIGH' then 1
                else 0 end) as high_line_count,
     SUM(case when o_orderpriority != '1-URGENT' and != '2-HIGH then 1
                else 0 end) as low_line_count
FROM orders, lineitem
WHERE o_orderkey = l_orderkey
  AND l_shipmode in ('[SHIPMODE1]', '[SHIPMODE2]')
  AND l_commitdate < l_receiptdate
  AND l_shipdate < l_commitdate
  AND l_receiptdate >= date '[DATE]'
  AND l_receiptdate < date '[DATE]' + '1' year
GROUP BY l_shipmode
ORDER BY l_shipmode;
```

OLAP Cube Design
Dimension Definition and Filters' processing (6/7)

Example 2: Q19 of TPC-H Benchmark.

```
select
        sum(l_extendedprice * (1 - l_discount) ) as revenue
from
        lineitem, ---> fact table
        part      ---> dimension table
where
        (
                p_partkey = l_partkey
                and p_brand = '[BRAND1]'
                and p_container in ( 'SM CASE', 'SM BOX', 'SM PACK', 'SM PKG')
                and l_quantity >= [QUANTITY1] and l_quantity <= [QUANTITY1] + 1
                and p_size between 1 and 5
                and l_shipmode in ('AIR', 'AIR REG')
                and l_shipinstruct = 'DELIVER IN PERSON'
        )
        or
```

OLAP Cube Design
Dimension Definition and Filters' processing (7/7)

```
(
        p_partkey = l_partkey
        and p_brand = '[BRAND2]'
        and p_container in ('MED BAG', 'MED BOX', 'MED PKG', 'MED PACK')
        and l_quantity >= [QUANTITY2] and l_quantity <= [QUANTITY2] + 10
        and p_size between 1 and 10
        and l_shipmode in ('AIR', 'AIR REG')
        and l_shipinstruct = 'DELIVER IN PERSON'
)
or
        (
        p_partkey = l_partkey
        and p_brand = '[BRAND3]'
        and p_container in ( 'LG CASE', 'LG BOX', 'LG PACK', 'LG PKG')
        and l_quantity >= [QUANTITY3] and l_quantity <= [QUANTITY3] + 10
        and p_size between 1 and 15
        and l_shipmode in ('AIR', 'AIR REG')
        and l_shipinstruct = 'DELIVER IN PERSON'
```

*TPC-H*d*

- Truly OLAP variant of TPC-H benchmark
- TPC-H SQL workload translated into MDX (MultiDimensional eXpressions)
- The workload is composed of 23 MDX statements for OLAP cubes and 23 MDX statements for OLAP business queries.
  - Each business question of TPC-H benchmark is mapped to an OLAP cube

TPC-H*d
Q8: From SQL statement to OLAP cube

```
SELECT o_year, sum(case when nation = '[NATION]' then volume
                   else 0 end) / sum(volume) as mkt_share
FROM (SELECT extract(year from o_orderdate) as o_year,
             l_extendedprice * (1-l_discount) as volume,
             n2.n_name as nation
      FROM part, supplier, lineitem, orders, customer, nation n1, nation n2, region
      WHERE p_partkey =
        AND l_orderkey =
        AND c_nationkey
        AND r_name = '[R
        AND o_orderdate
        AND p_type = '[T
GROUP BY o_year;
```

```
OLAP Cube 8
|
|-- Fact Table                                          ey
|     |-- LineItem Facts                                ey
|-- Measures
|     |-- M8.1: Σ(l_extendedprice×(1-l_discount))
|-- Dimensions
|     |-- D8.1: Part
|     |     |-- L0: type
|     |-- D8.2: Order Date
|     |     |-- L0: year
|     |-- D8.3: Customer Geography
|     |     |-- L0: region
|     |-- D8.4: Supplier Geography
|     |     |-- L0: nation
```

Market Share for each *supplier nation* within a *region of customers*, for each *year* and each *part type*

				Mesures	
Supplier Nation	Customer Region	Order Year	Part Type	Volume	Market Share
+All Supplier Nations	+All Customer Regions	-All Order Years	ECONOMY ANODIZED BRASS	13 591 860,657	1
			ECONOMY ANODIZED COPPER	17 577 009,763	1
			ECONOMY ANODIZED NICKEL	6 437 355,496	1
			ECONOMY ANODIZED STEEL	11 684 602,546	1
FRANCE	-All Customer Regions	-All Order Years	-All Part Types	43 574 220,136	0,021
			ECONOMY ANODIZED BRASS	196 397,15	0,014
			ECONOMY ANODIZED COPPER	250 634,275	0,014
			ECONOMY ANODIZED NICKEL		
			ECONOMY ANODIZED STEEL		

```
WITH MEMBER Measures.[Market Share] AS 'Measures.[Volume] /
   (Measures.[Volume], [Supplier Nation].[All Supplier Nations].Lead(0))'
SELECT {[Measures].[Volume], [Measures].[Market Share]} ON COLUMNS,
   Crossjoin(Crossjoin(Crossjoin([Supplier Nation].members,
         [Customer Region].members),
        {[Order Year].members}), [Part Type].members)  ON ROWS
FROM [Cube8]
```

TPC-H*d OLAP Query Q8

Market Share for each *RUSSIAN Suppliers* within AMERICA region,
Over the years 1995 and 1996 and for *part type ECO. ANODIZED STEEL*

			Mesures	
Supplier Nation	Customer Region	Order Year	Volume	Market Share
RUSSIA	AMERICA	1995	28 637,136	0,076
		1996	94 173,948	0,152

Slicer: [Part Type=ECONOMY ANODIZED STEEL]

```
WITH MEMBER [Measures].[Market Share] AS '([Measures].[Volume] /
  ([Measures].[Volume], [Supplier Nation].[All Supplier Nations].Lead(0)))'
SELECT {[Measures].[Volume], [Measures].[Market Share]} ON COLUMNS,
  Crossjoin({[Supplier Nation].[RUSSIA]}, Crossjoin({[Customer Region].[AMERICA]},
  {[Order Year].[1995], [Order Year].[1996]})) ON ROWS
FROM [Cube8]
WHERE {[Part Type].[ECONOMY ANODIZED STEEL]}
```