Benchmarks para DW

Processamento Analítico de Dados

Profa. Dra. Cristina Dutra de Aguiar Ciferri

Motivação

- Literatura: diferentes técnicas
  - melhoria do desempenho no processamento de consultas analíticas
- Como medir o ganho de desempenho?
  - realização de testes
    - geração de dados sintéticos (ou artificiais)
    - baseados em um *benchmark* padrão
    - uso de volumes de dados distintos e significativos

Benchmark

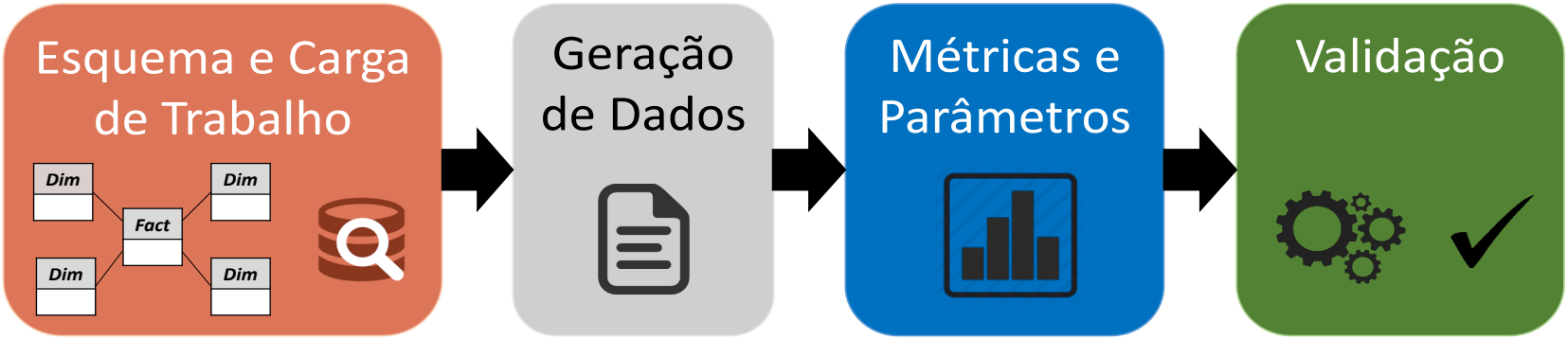
- Técnica experimental
- Definição genérica
  - consiste da execução de um conjunto conhecido de consultas e de operações de inserção, de remoção e de modificação de dados (i.e., a sua carga de trabalho) em um conjunto de dados também conhecido e em geral gerado artificialmente

CIFERRI, R. R. Análise da Influência do Fator Distribuição Espacial dos Dados no Desempenho de Métodos de Acesso Multidimensionais. Tese (Doutorado em Ciência da Computação). Centro de Informática, Universidade Federal de Pernambuco, Recife, PE, Brasil, 2002.

Benchmark

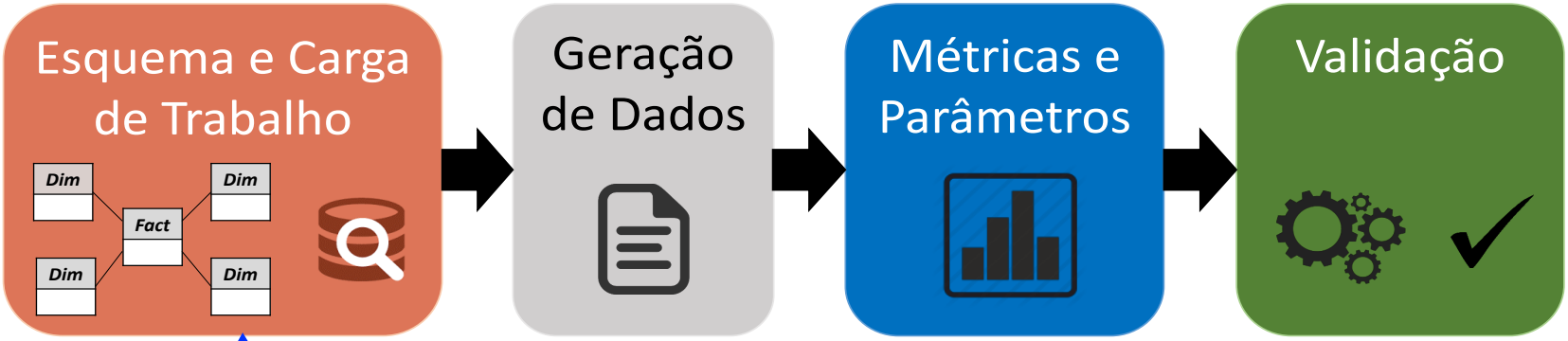
- Usado para avaliar o desempenho de aplicações de banco de dados
- Funcionalidades oferecidas
  - esquema de dados
  - diferentes tipos de consulta
  - diferentes operações de inserção, remoção e atualização
- Aspectos considerados
  - seletividade
  - volume de dados

Etapas de um *Benchmark*



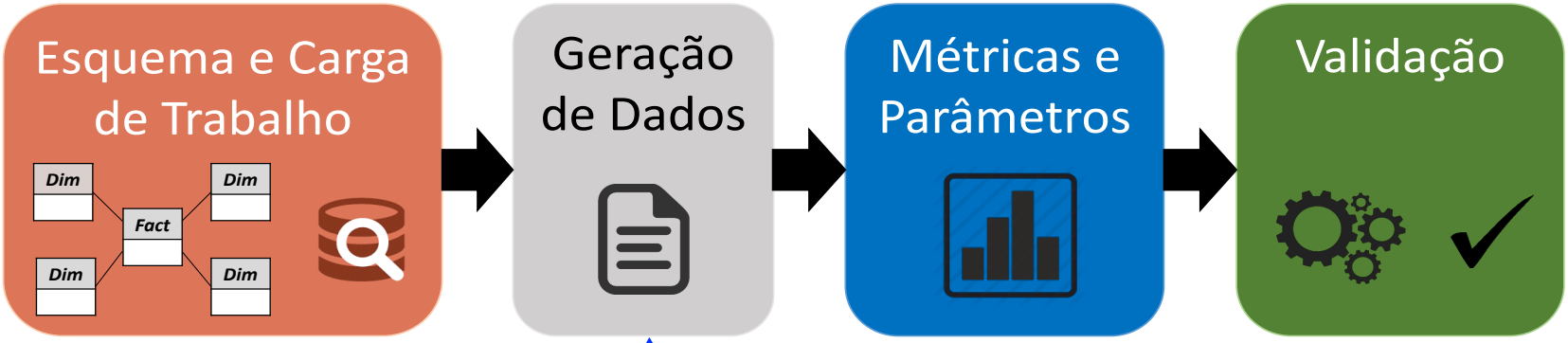
SCABORA, L. C. Avaliação do *Star Schema Benchmark* aplicado a bancos de dados NoSQL distribuídos e orientados a colunas. Dissertação (Mestrado em Ciência da Computação). Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil, 2016.

Etapas de um *Benchmark*



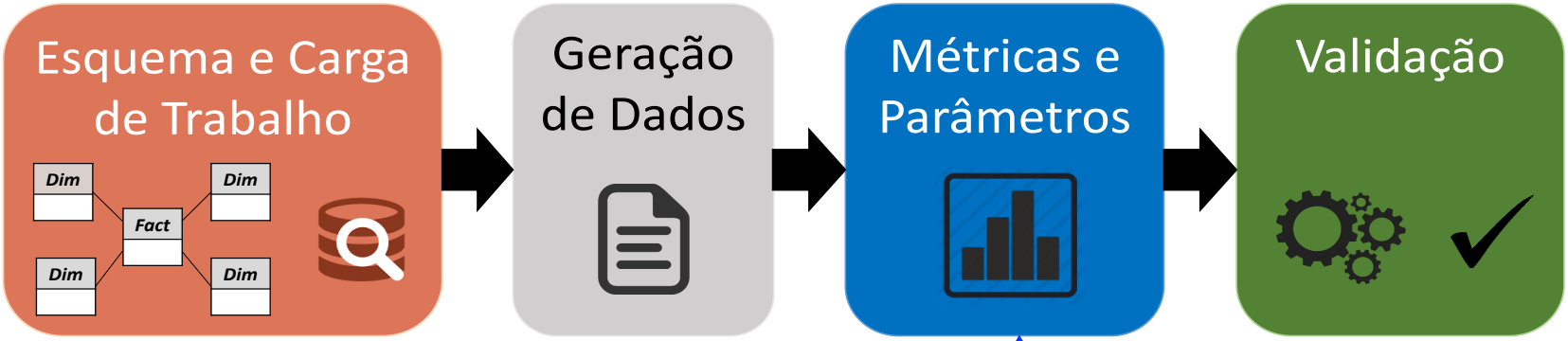
- esquema
  - esquema estrela ou suas variações
- carga de trabalho
  - consultas OLAP

Etapas de um *Benchmark*



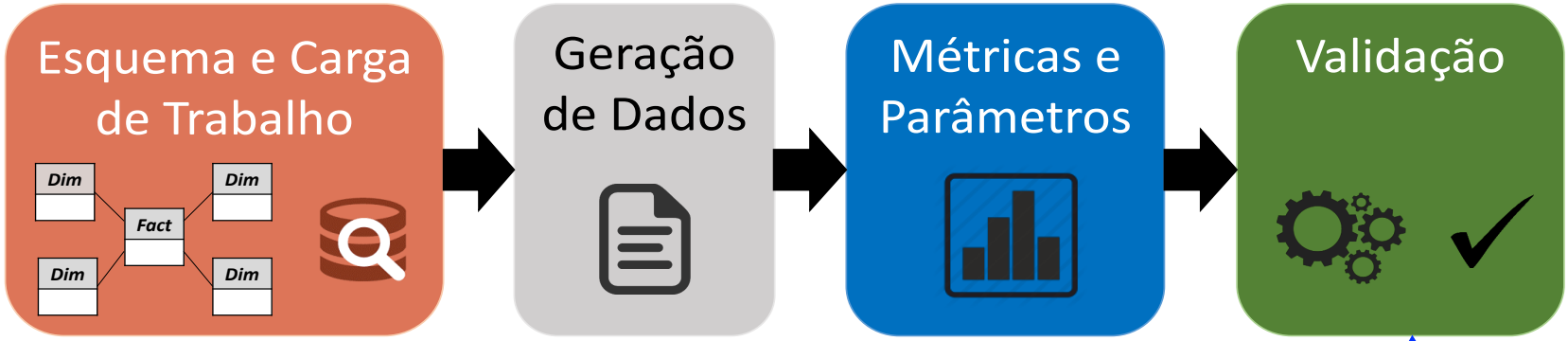
- conjunto de dados relativos à aplicação que se deseja avaliar
  - dados reais *versus* dados sintéticos
- dados sintéticos
  - geração artificial de dados

Etapas de um *Benchmark*



- métricas: informações qualitativas e quantitativas
  - tempo de resposta, *etc*
- parâmetros: valores configuráveis
  - seletividade, volume de dados, *etc*

Etapas de um *Benchmark*



aplicação do esquema e da carga de trabalho juntamente à geração de tabelas e à inserção de dados usando os dados para explorar os parâmetros definidos a fim de coletar as métricas definidas, necessárias à comparação entre os diferentes sistemas analisados

Benchmark TPC-H

http://www.tpc.org/tpch/

- Define uma aplicação de DW
  - relacionada a dados históricos referentes a pedidos e vendas de uma empresa de varejo, durante um certo período de tempo
- Esquema
  - constelação de fatos
    - tabelas de fatos: LineItem e PartSupp
    - tabelas de dimensão: Orders, Part, Supplier, Customer, Nation e Region

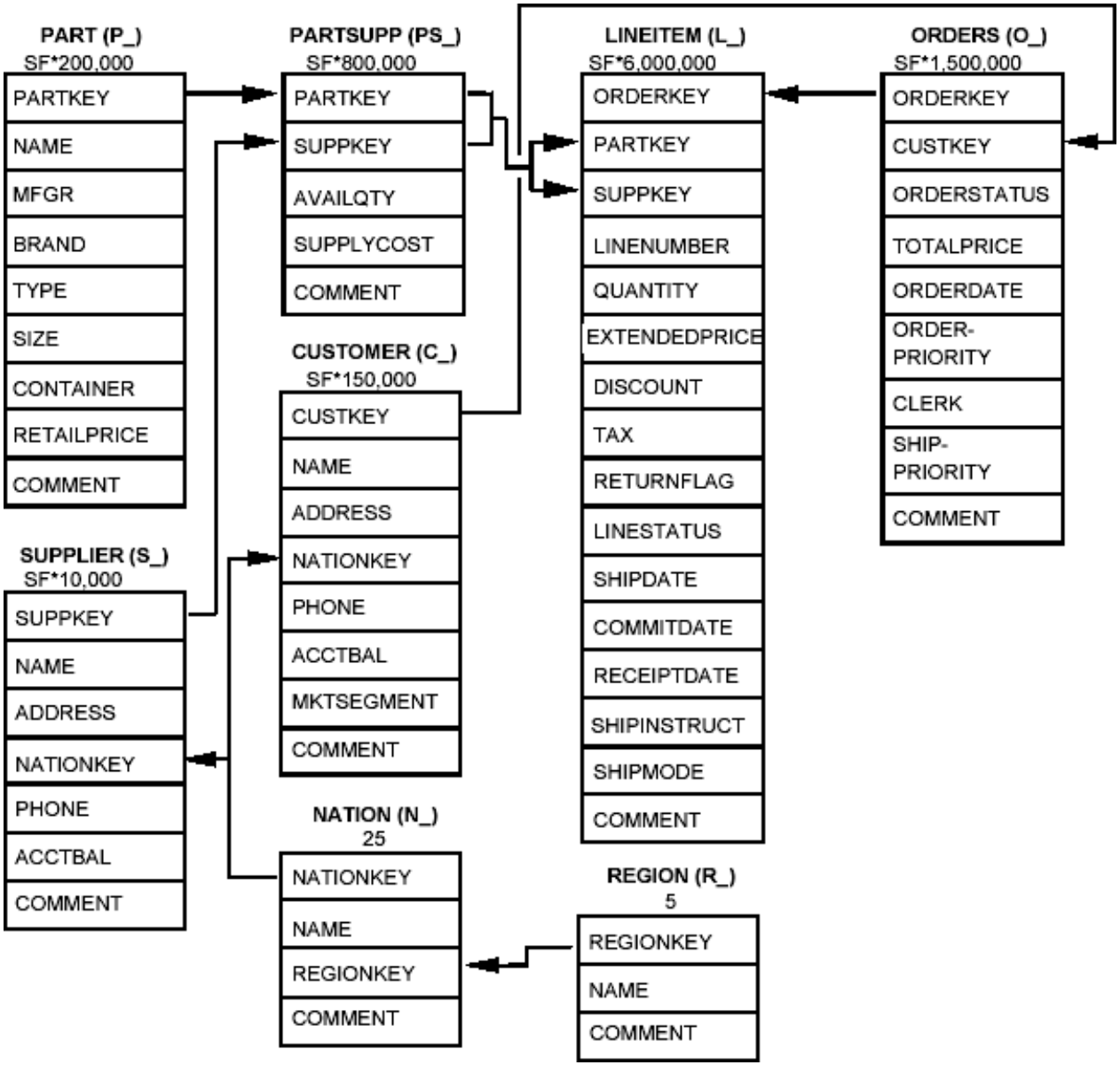
Benchmark TPC-H

- Carga de trabalho
  - 22 consultas
  - aspectos analisados
    - determinação de preços e promoções
    - gerenciamento de oferta e procura
    - gerenciamento de lucros e ganhos
    - estudo da satisfação do consumidor
    - estudo da participação no mercado
    - gerenciamento de remessas

Benchmark TPC-H

- Gerador de dados: dbgen
  - gera dados sintéticos
  - cardinalidade das tabelas, ou seja, a quantidade de dados gerados para cada tabela
    - pode ser estática *ou*
    - pode ser baseada no fator de escala (SF)
- Métrica
  - desempenho, calculada por meio do número de consultas executadas por hora

Esquema

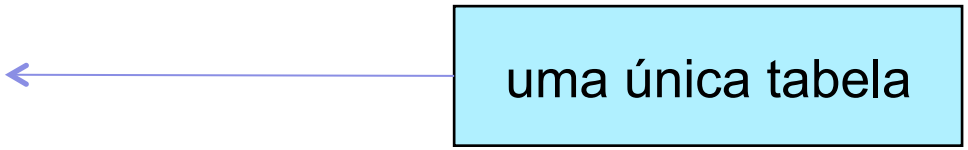


O'NEIL, P.; O'NEIL, E.; CHEN, X.; REVILAK, S. The star schema benchmark and augmented fact table indexing. In: Performance Evaluation and Benchmarking, 2009. p. 237–252..

Exemplo: Consulta Q1

- Lista diversas informações de interesse relacionadas ao negócio.

```
select
        l_returnflag,
        l_linestatus,
        sum(l_quantity) as sum_qty,
        sum(l_extendedprice) as sum_base_price,
        sum(l_extendedprice*(1-l_discount)) as sum_disc_price,
        sum(l_extendedprice*(1-l_discount)*(1+l_tax)) as sum_charge,
        avg(l_quantity) as avg_qty,
        avg(l_extendedprice) as avg_price,
        avg(l_discount) as avg_disc,
        count(*) as count_order
from
        lineitem
where
        l_shipdate <= date '1998-12-01' - interval '[DELTA]' day (3)
group by
        l_returnflag,
        l_linestatus
order by
        l_returnflag,
        l_linestatus;
```



Exemplo: Consulta Q3

- Recupera os 10 pedidos ainda não enviados que possuem os valores mais elevados.

```
select
        l_orderkey,
        sum(l_extendedprice*(1-l_discount)) as revenue,
        o_orderdate,
        o_shippriority
from
        customer,
        orders,
        lineitem
where
        c_mktsegment = '[SEGMENT]'
        and c_custkey = o_custkey
        and l_orderkey = o_orderkey
        and o_orderdate < date '[DATE]'
        and l_shipdate > date '[DATE]'
group by
        l_orderkey,
        o_orderdate,
        o_shippriority
order by
        revenue desc,
        o_orderdate;
```

junção de tabelas
- esquema estrela -

Exemplo: Consulta Q9

- Obtém as receitas dos fornecedores agrupadas por nação e ano, realizando a junção das tabelas de fatos LineItem e PartSupp por intermédio das tabelas de dimensão compartilhadas Customer e Supplier.

```
select
        nation,
        o_year,
        sum(amount) as sum_profit
from (
        select
                n_name as nation,
                extract(year from o_orderdate) as o_year,
                l_extendedprice * (1 - l_discount) - ps_supplycost * l_quantity as amount
        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;
```

junção de tabelas
- floco de neve –
junção de 2 tabelas de fatos

Benchmark SSB

http://www.cs.umb.edu/~poneil/StarSchemaB.pdf

- Baseado no *benchmark* TPC-H
- Esquema
  - modifica o esquema do TPC-H para um esquema estrela
- Alterações
  - criação da tabela LineOrder, que junta as tabelas Lineitem e Order, e possui todos os atributos dessas duas tabelas originais

Benchmark SSB

- Alterações
  - exclusão da tabela PartSupp
    - atualizações feitas nas tabelas LineOrder e PartSupp não ocorrem ao mesmo tempo
  - exclusão de alguns atributos das tabelas LineItem e Order
    - atributos do tipo texto não estruturados, que não podem ser agrupados ou sumarizados, e que não contextualizam os fatos analisados
  - adição da tabela de dimensão Date
    - para armazenar histórico de vendas da companhia

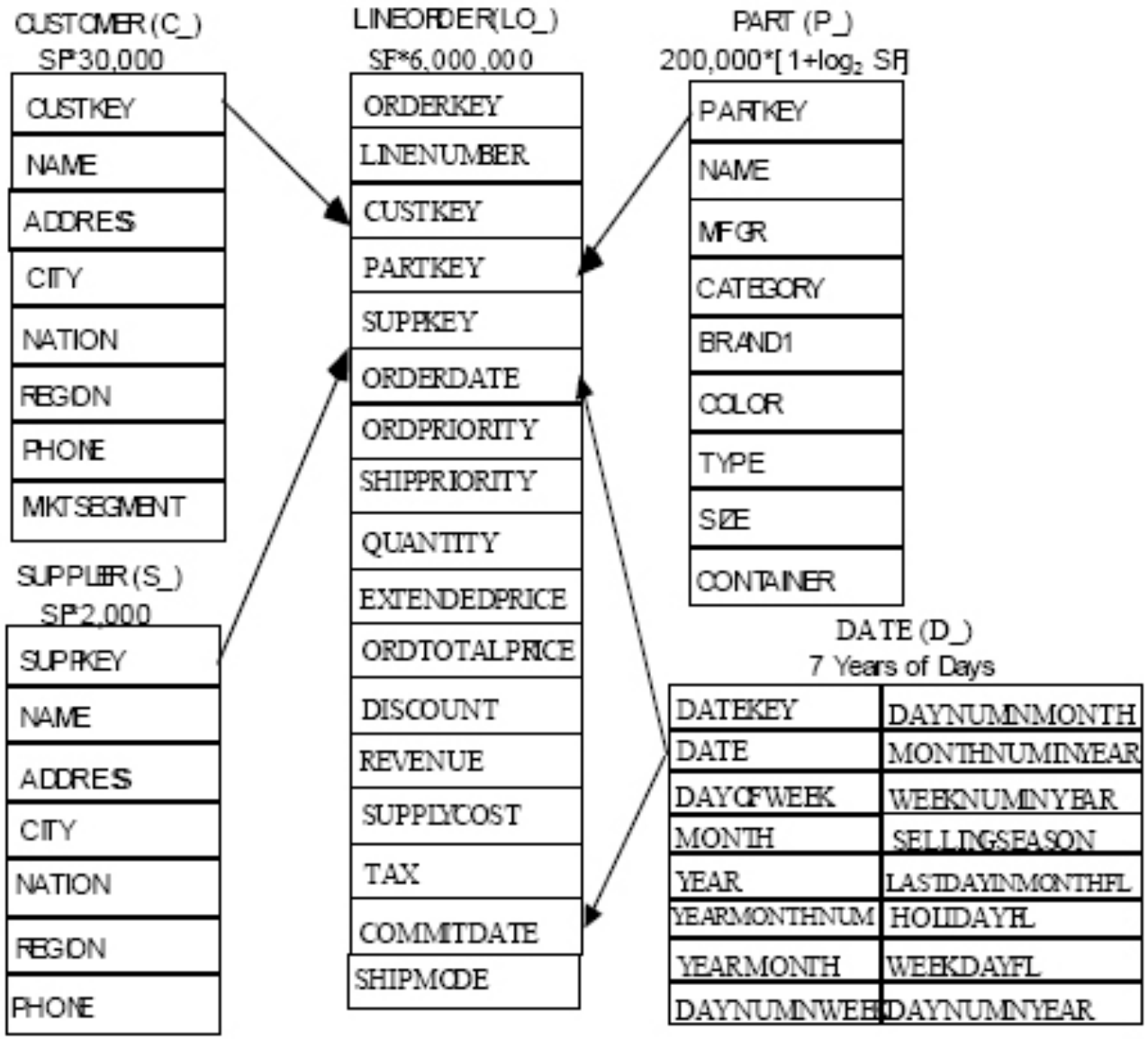
Benchmark SSB

- Carga de trabalho
  - 13 consultas
  - consultas
    - adaptadas do TPC-H, considerando as adaptações no esquema
    - novas, criadas para validar o novo esquema
  - divisão em classes
    - realizam cobertura funcional (diferentes tipos de consultas comuns ao esquema estrela) e testam variações tanto de seletividade quanto de nível de agregação

Benchmark SSB

- Gerador de dados: dbgen
  - gera dados sintéticos
  - cardinalidade das tabelas, ou seja, a quantidade de dados gerados para cada tabela
    - pode ser estática *ou*
    - pode ser baseada no fator de escala (SF)
- Métrica
  - desempenho, calculada por meio do número de consultas executadas por hora

Esquema



O'NEIL, P.; O'NEIL, E.; CHEN, X.; REVILAK, S. The star schema benchmark and augmented fact table indexing. In: Performance Evaluation and Benchmarking, 2009. p. 237–252..

Exemplo de Consultas

- Classe Q1
  - quantifica a renda a partir da eliminação de certos descontos da empresa, dada uma porcentagem de produtos enviados em um determinado ano

Consulta Q1.1

Q1.1 YEAR = 1993, DISCOUNT = 2, QUANTITY = 25, so predicates are d_year = 1993, lo_quantity < 25, lo_discount between 1 and 3.

```
select sum(lo_extendedprice*lo_discount) as revenue
 from lineorder, date
 where lo_orderdate = d_datekey
  and d_year = 1993
  and lo_discount between1 and 3
  and lo_quantity < 25;
```

junção das mesmas tabelas, mas usando diferentes critérios de seleção

FF = (1/7)*0.5*(3/11) = 0.0194805. Number of lineorder rows selected, for SF = 1, is 0.0194805*6,000,000 ≈ 116,883.

Consulta Q1.2

Q1.2 d_yearmonthnum = 199401, lo_quantity between 26 and 35, lo_discount between 4 and 6.

```
select sum(lo_extendedprice*lo_discount) as revenue
 from lineorder, date
 where lo_orderdate = d_datekey
  and d_yearmonthnum = 199401
  and lo_discount between4 and 6
  and lo_quantity between 26 and 35;
```

junção das mesmas tabelas, mas usando diferentes critérios de seleção

FF = (1/84)*(3/11)*0.2 = 0.00064935. Number of lineorder rows selected, for SF = 1:
0.00064935*6,000,000 ≈ 3896.

Consulta Q1.3

Q1.3 d_weeknuminyear = 6 and d_year = 1994,
lo_quantity between 36 and 40, lo_discount between 5
and 7.

```
select sum(lo_extendedprice*lo_discount) as revenue
 from lineorder, date
 where lo_orderdate = d_datekey
  and d_weeknuminyear = 6
  and d_year = 1994
  and lo_discount between 5 and 7
  and lo_quantity between 26 and 35;
```

junção das mesmas tabelas, mas usando diferentes critérios de seleção

FF = (1/364)*(3/11)*0.1 = .000075. Number of lineorder rows selected, for SF = 1, is
.000075*6,000,000 ≈ 450.