VisualTPCH: Uma Ferramenta para a Geração de Dados Sintéticos para Data Warehouse

Gustavo Ramos Domingues
Cristina Dutra de Aguiar Ciferri
Ricardo Rodrigues Ciferri

Introdução

- 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 TPC-H

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

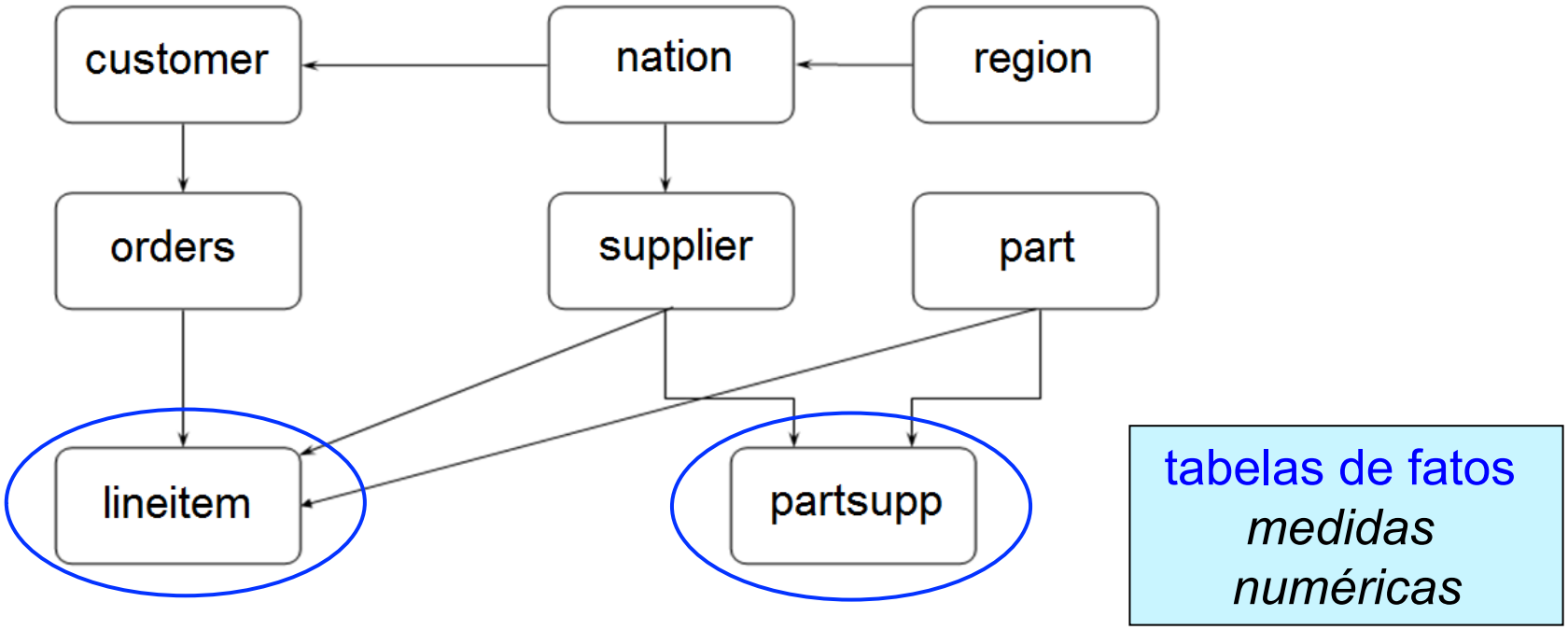
- Características
  - *benchmark* voltado à tomada de decisão
  - especifica como dados sintéticos para DW devem ser gerados
- Limitações
  - oferece um aplicativo de linha de comando
  - gera dados em arquivos texto de difícil manipulação
  - define um projeto de DW fixo e normalizado
  - não organiza os dados em níveis de agregação

Motivação

- Requisito
  - necessidade de suprir as limitações do TPC-H
  - oferecimento de uma ferramenta flexível para a geração de dados para testes
- A ferramenta proposta VisualTPCH
  - oferece uma interface gráfica
  - permite a manipulação do projeto do DW
  - facilita a geração de dados sintéticos
  - é baseada no *benchmark* TPC-H

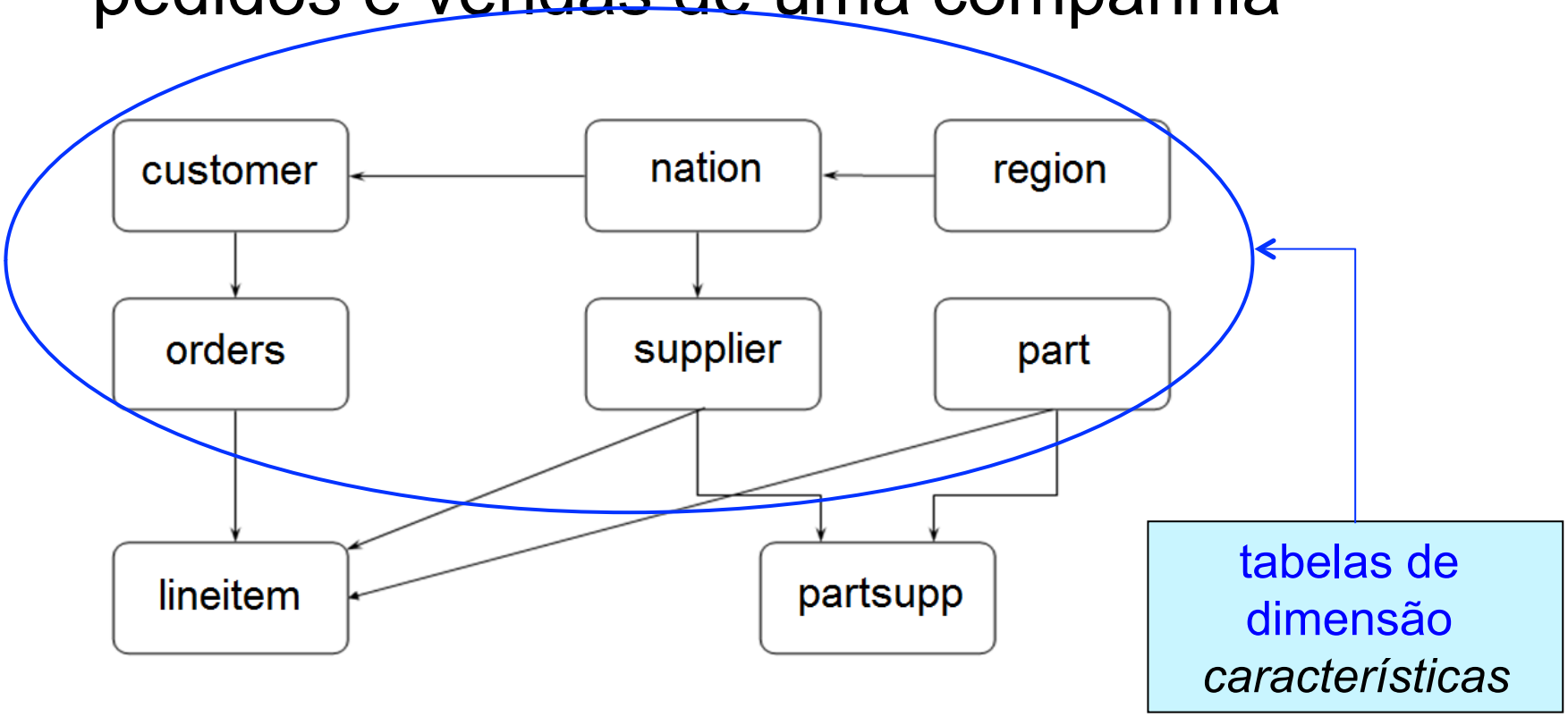
Esquema do TPC-H

- Armazena dados históricos relativos a pedidos e vendas de uma companhia



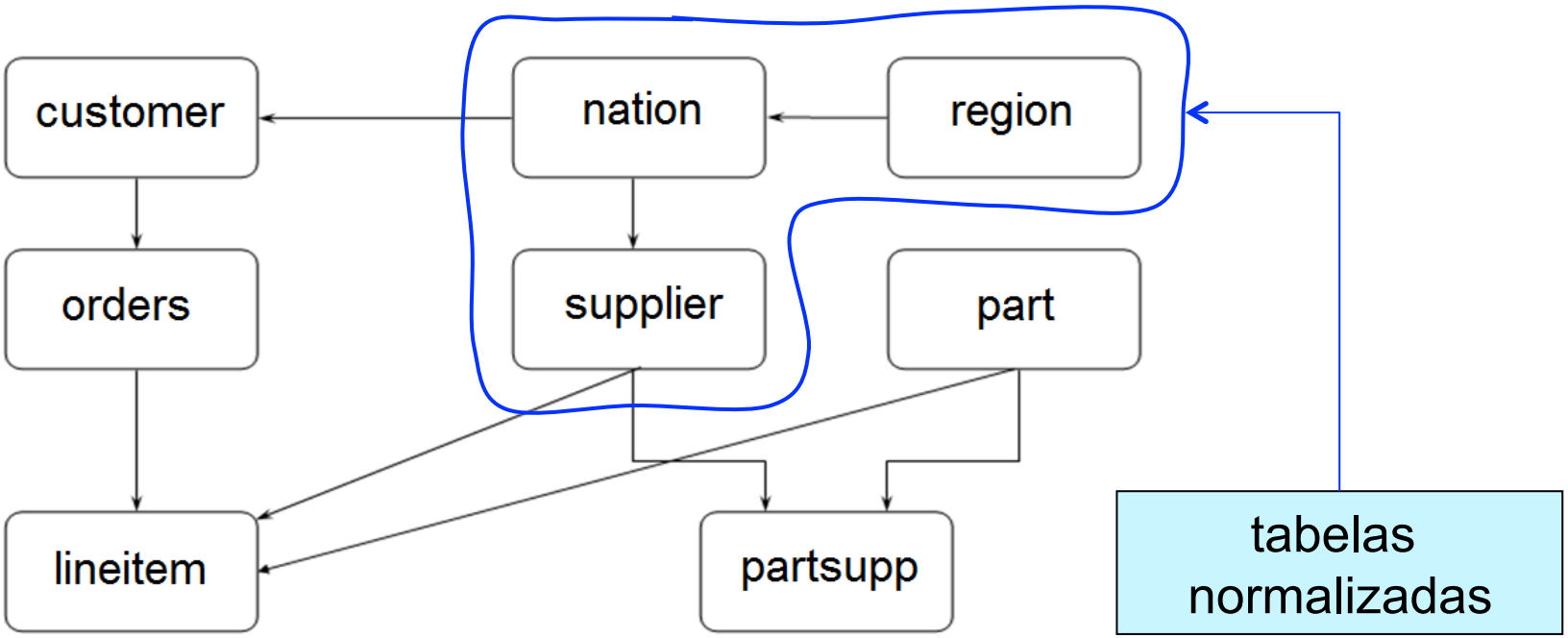
Esquema do TPC-H

- Armazena dados históricos relativos a pedidos e vendas de uma companhia



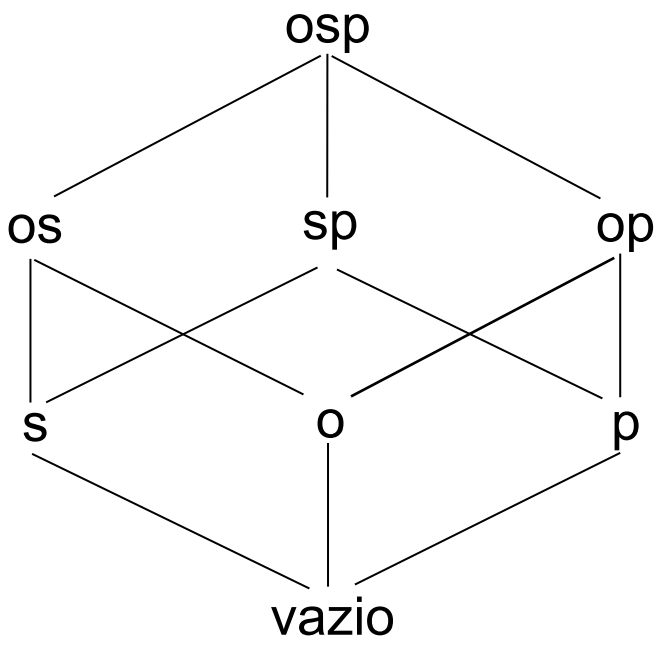
Esquema do TPC-H

- Armazena dados históricos relativos a pedidos e vendas de uma companhia



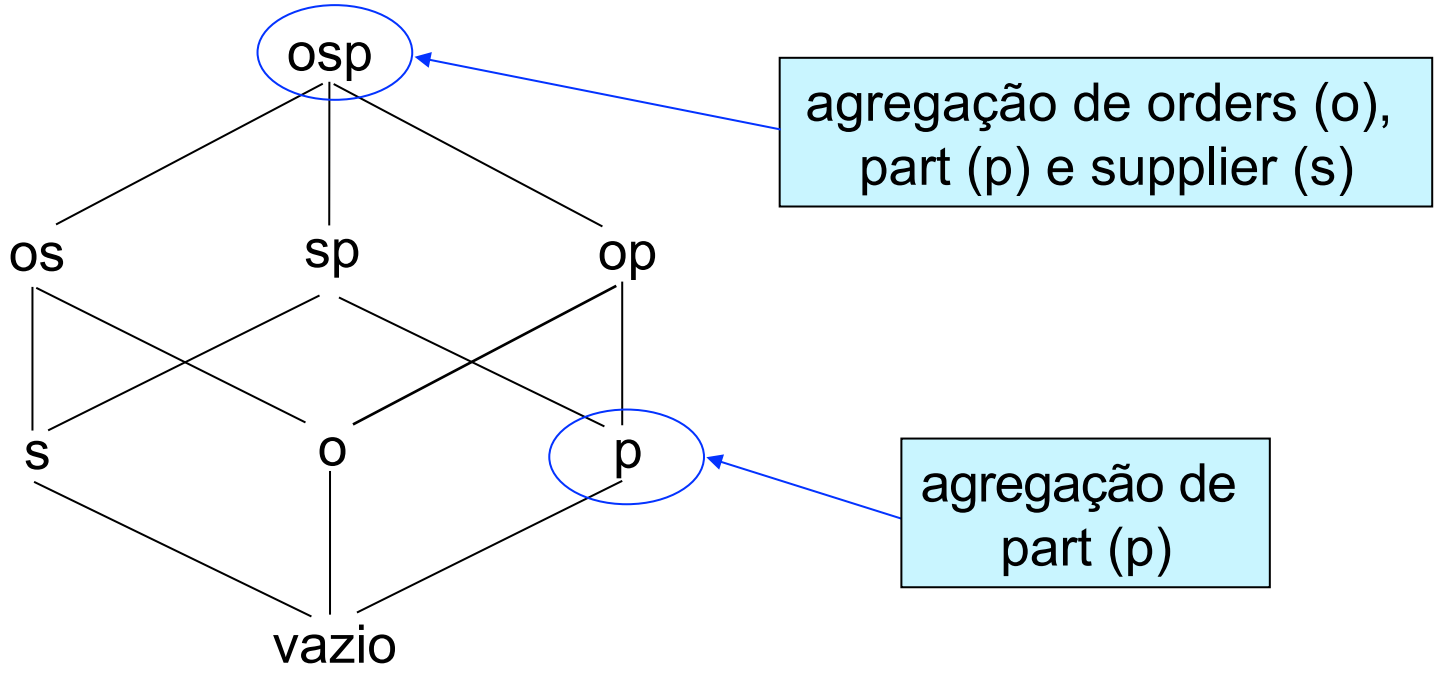
Grafo de Derivação: lineitem

- Vértices
  - agregação de medidas numéricas sobre as dimensões naquele vértice

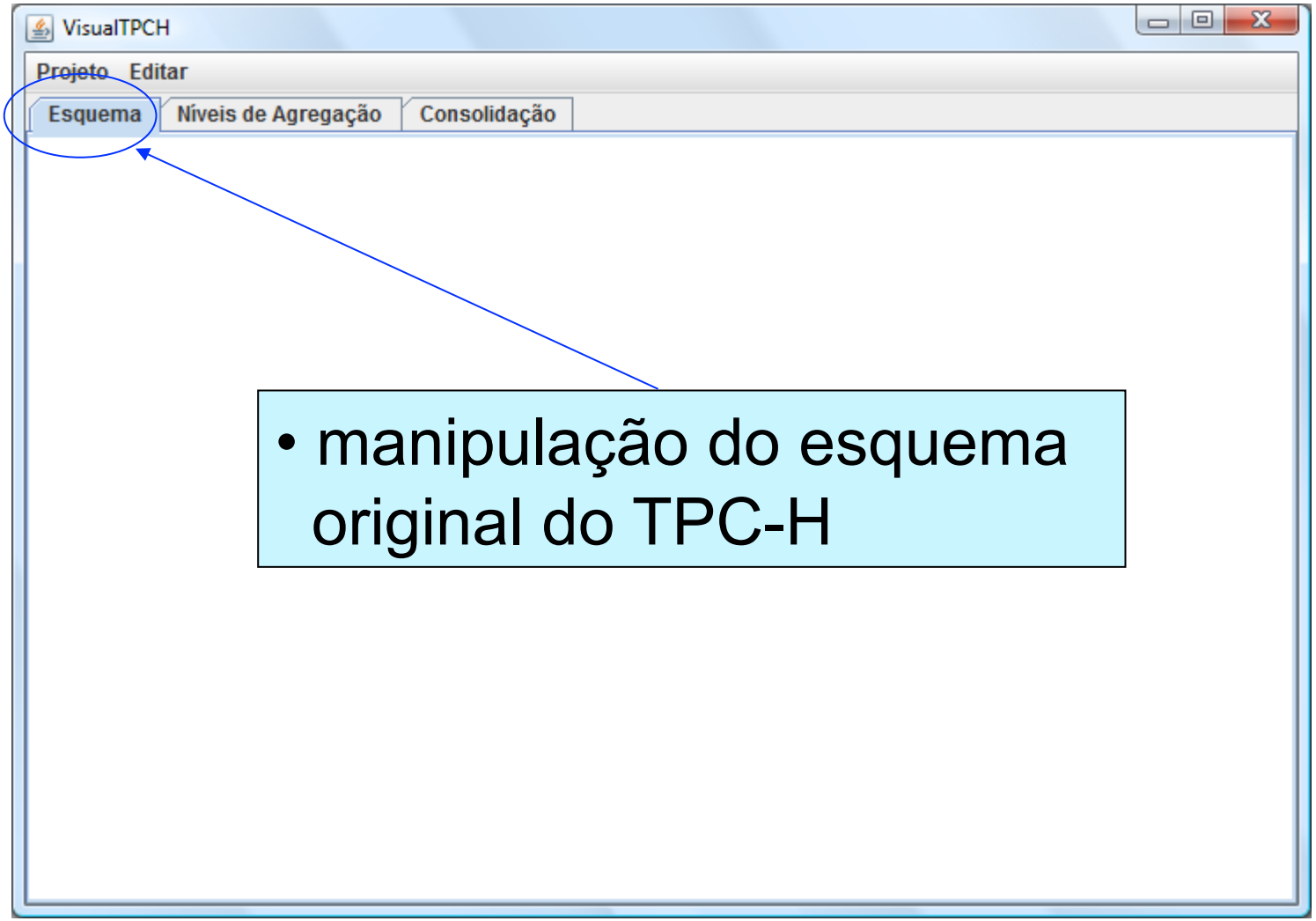


Grafo de Derivação: lineitem

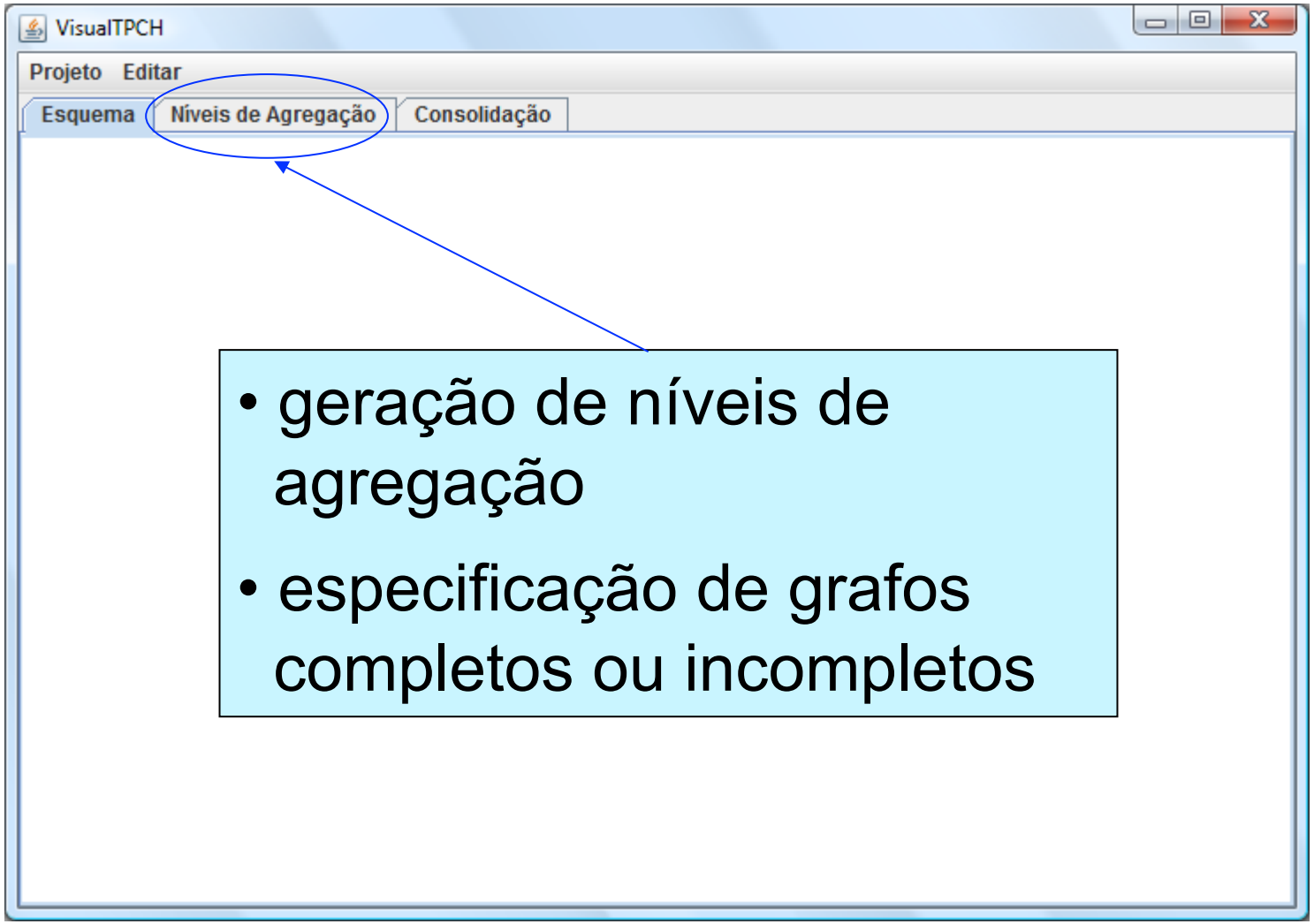
- Vértices
  - agregação de medidas numéricas sobre as dimensões naquele vértice



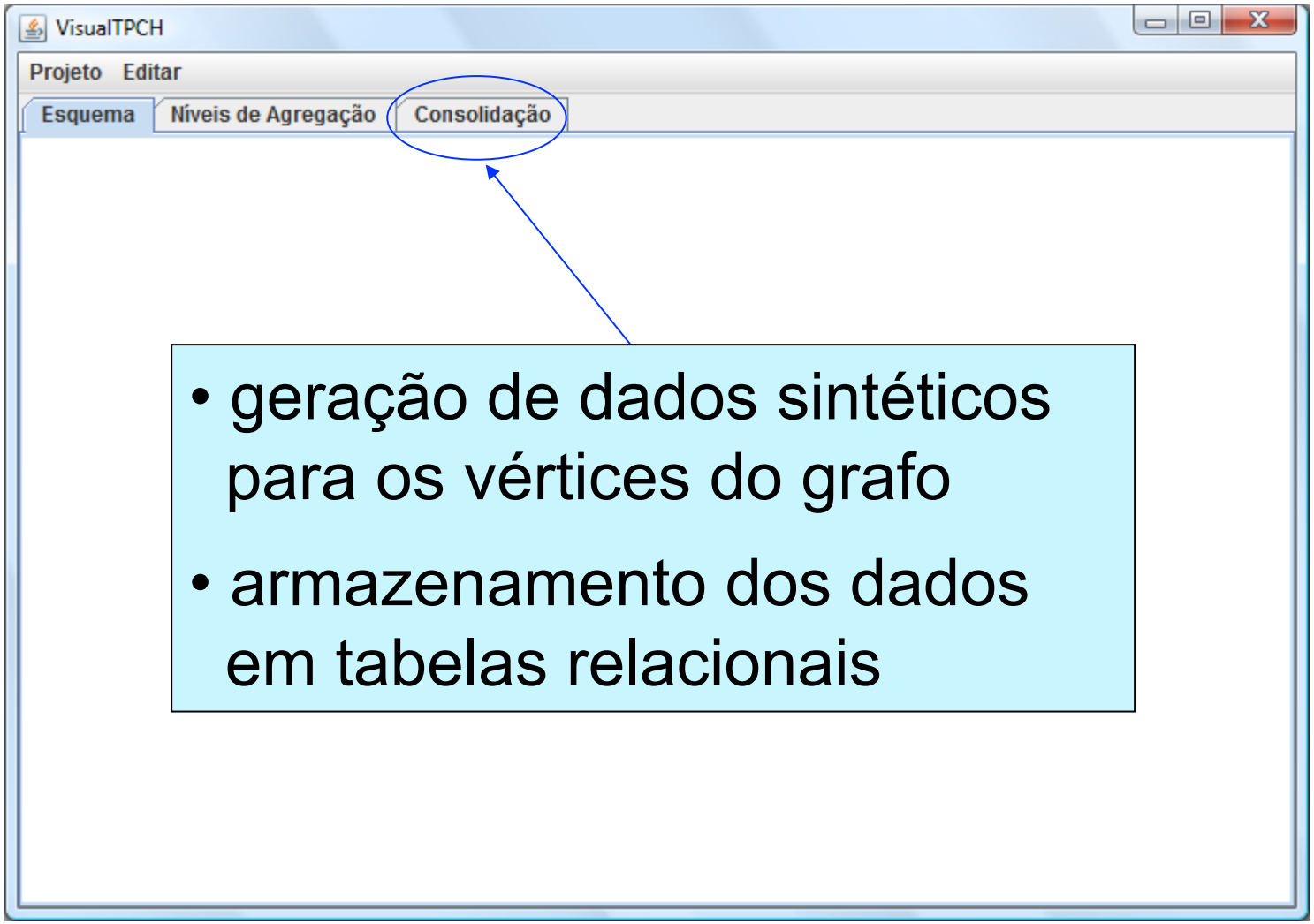
VisualTPCH


VisualTPCH
Projeto
Editar
Esquema
Níveis de Agregação
Consolidação
• manipulação do esquema original do TPC-H

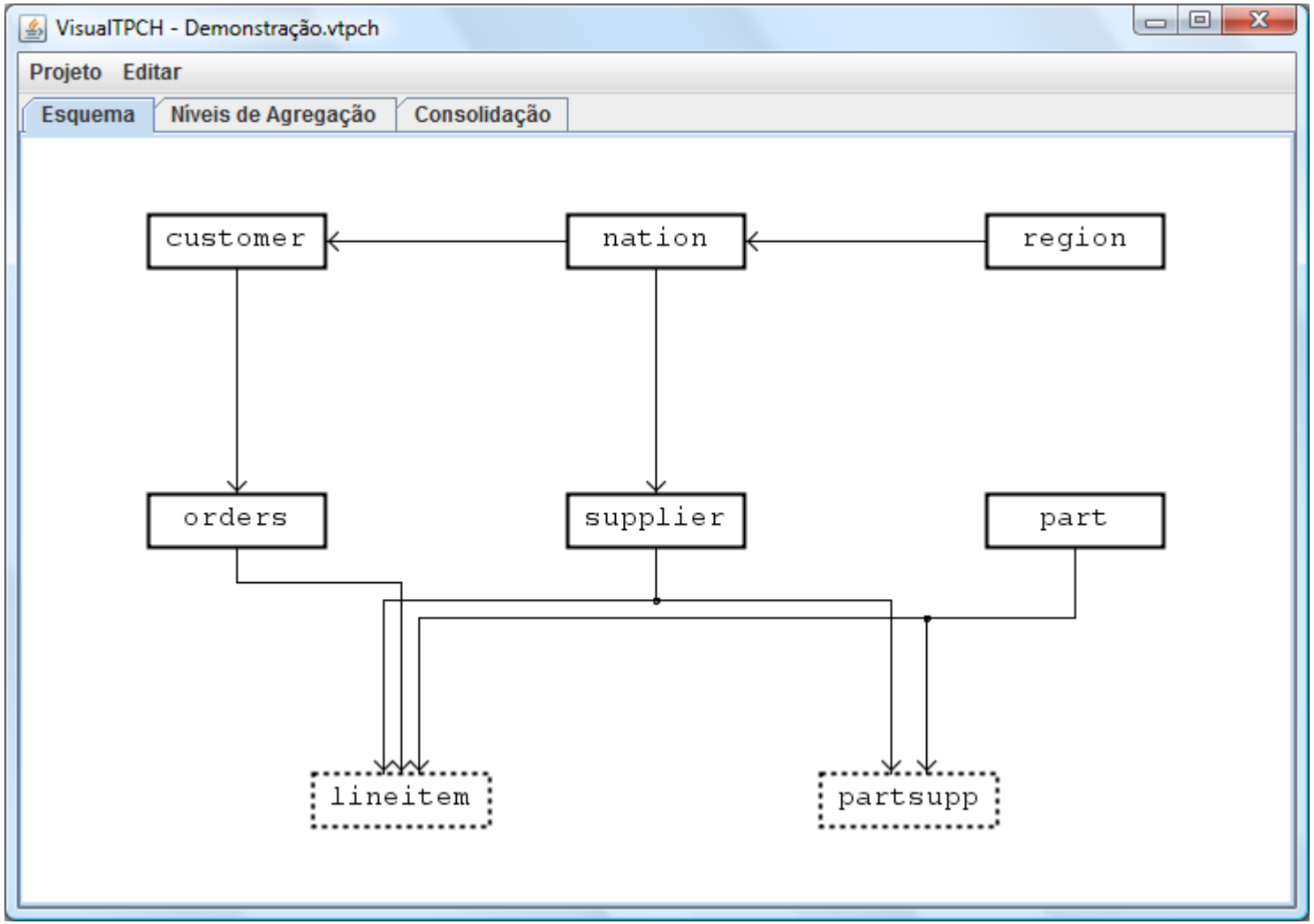
VisualTPCH


VisualTPCH
Projeto Editar
Esquema
Níveis de Agregação
Consolidação
• geração de níveis de agregação
• especificação de grafos completos ou incompletos

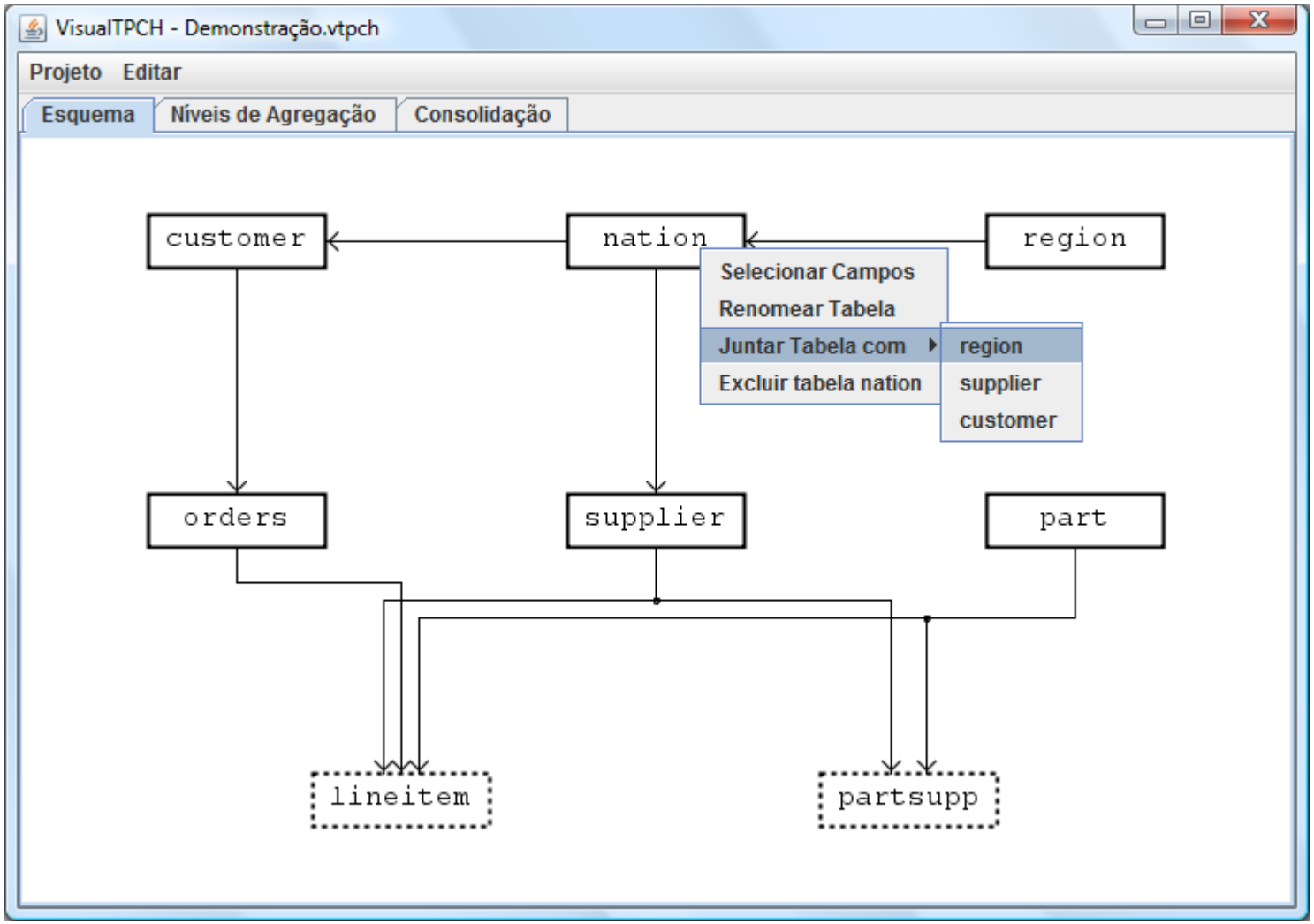
VisualTPCH


VisualTPCH
Projeto
Editar
Esquema
Níveis de Agregação
Consolidação
• geração de dados sintéticos para os vértices do grafo
• armazenamento dos dados em tabelas relacionais

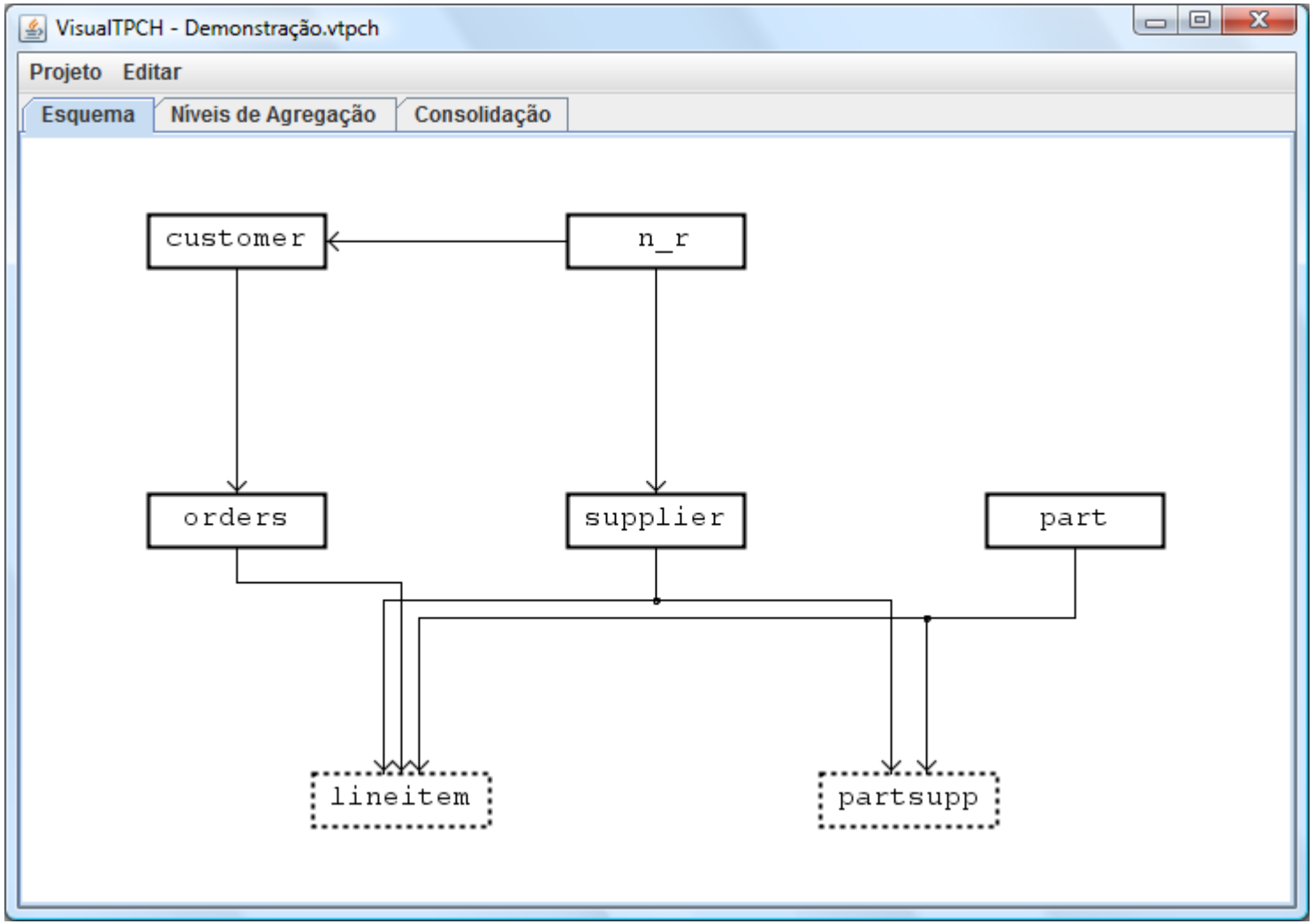
Esquema: Exibir TPC-H



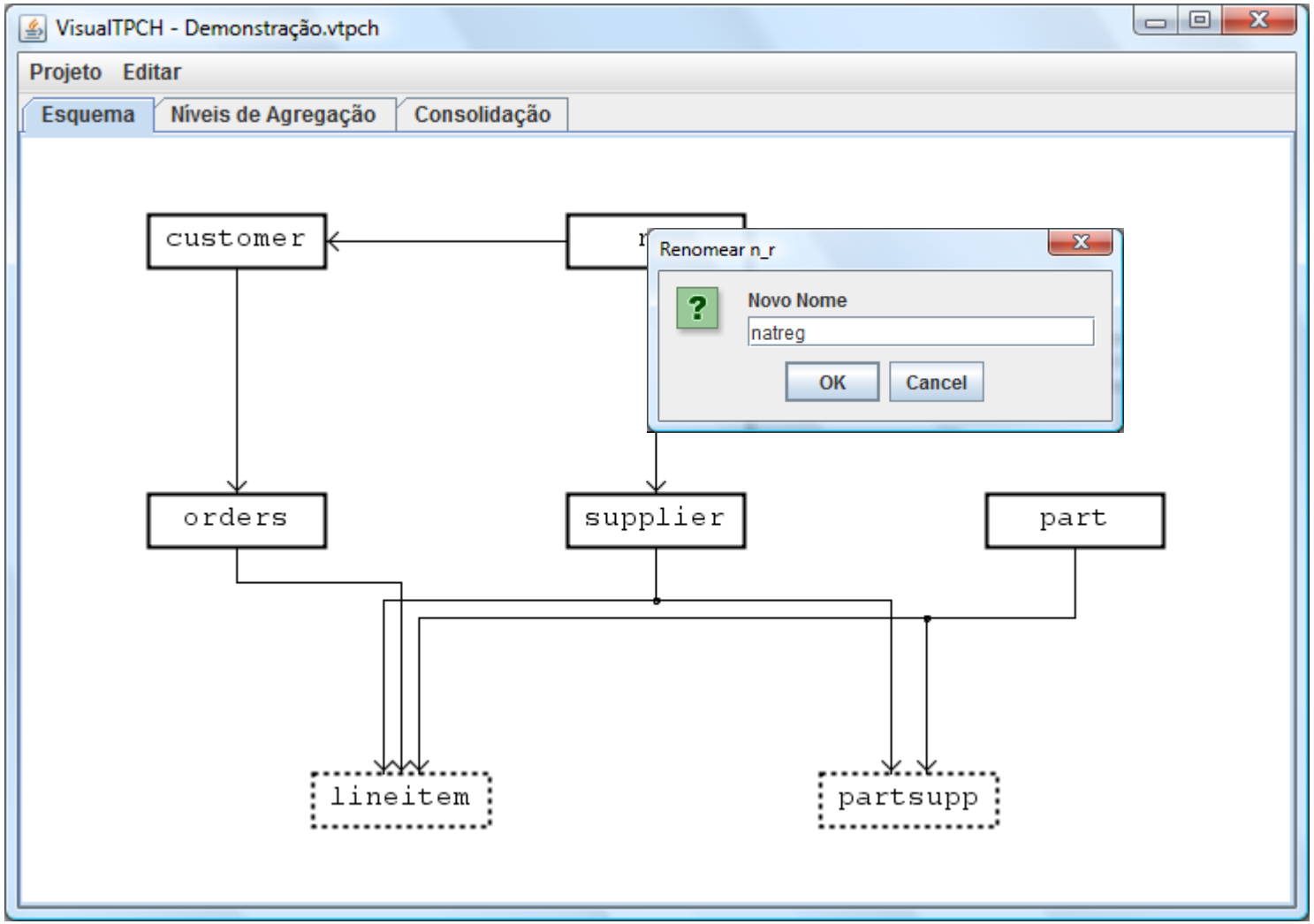
Esquema: Juntar Tabelas



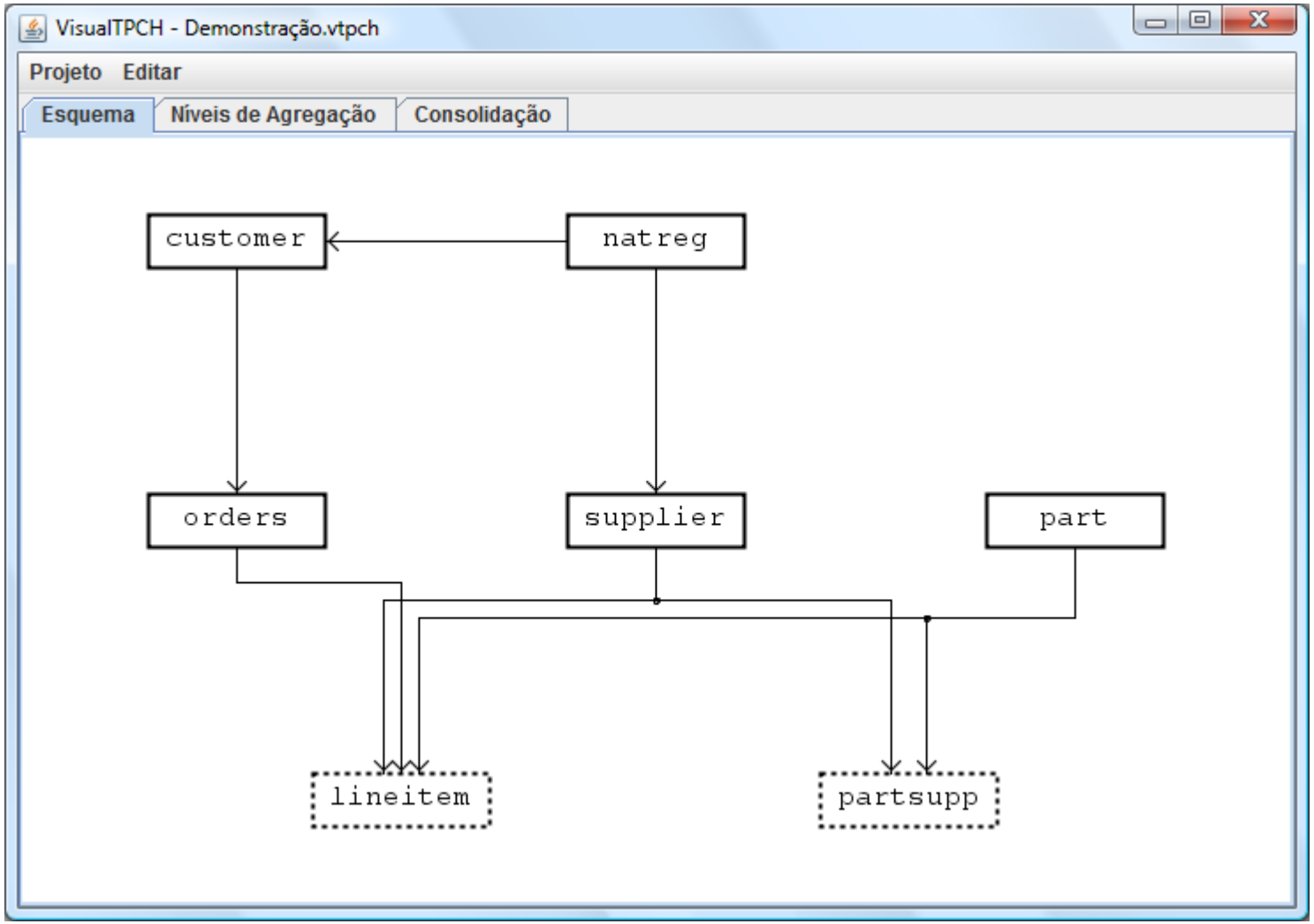
Esquema: Juntar Tabelas



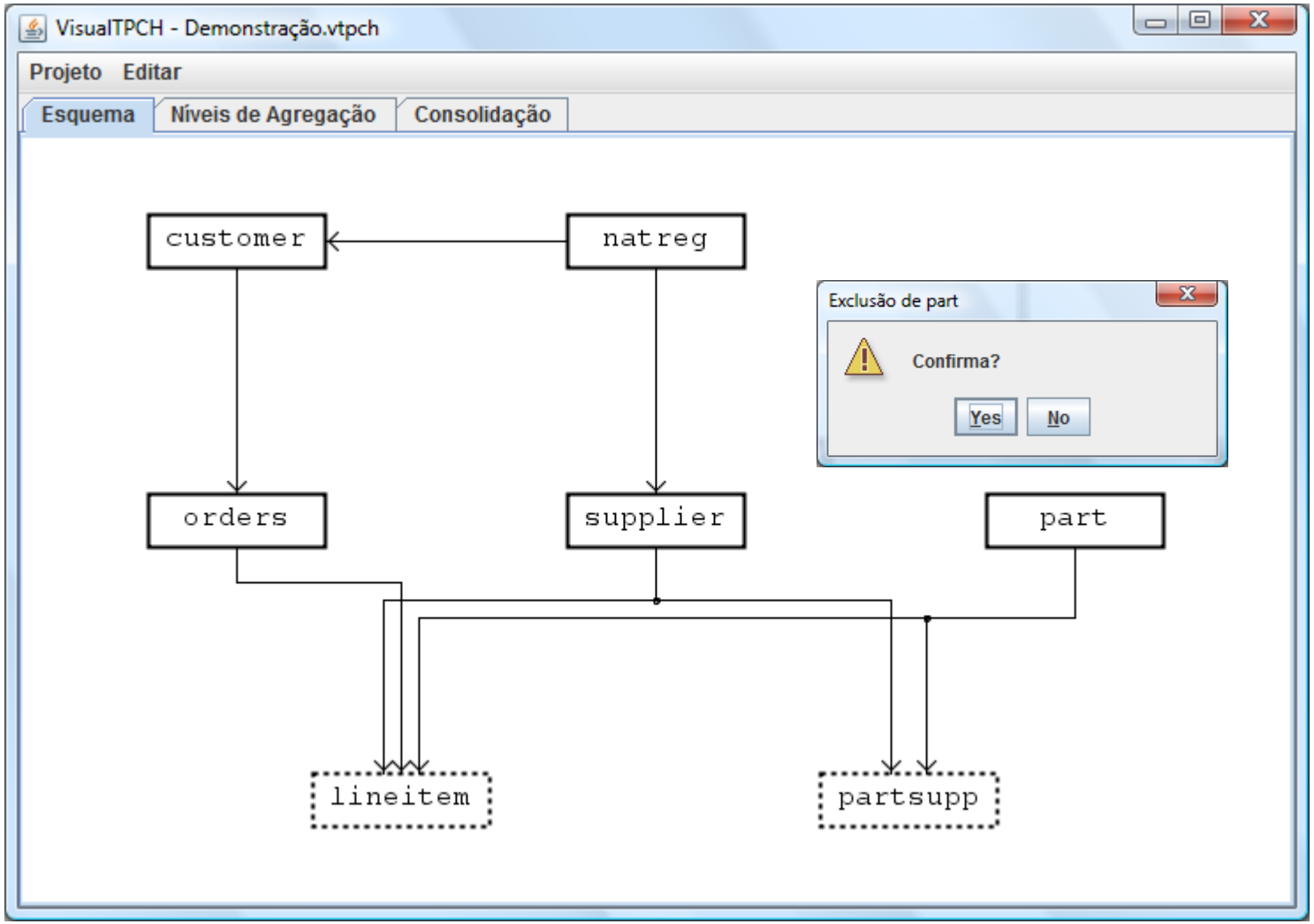
Esquema: Renomear Tabela



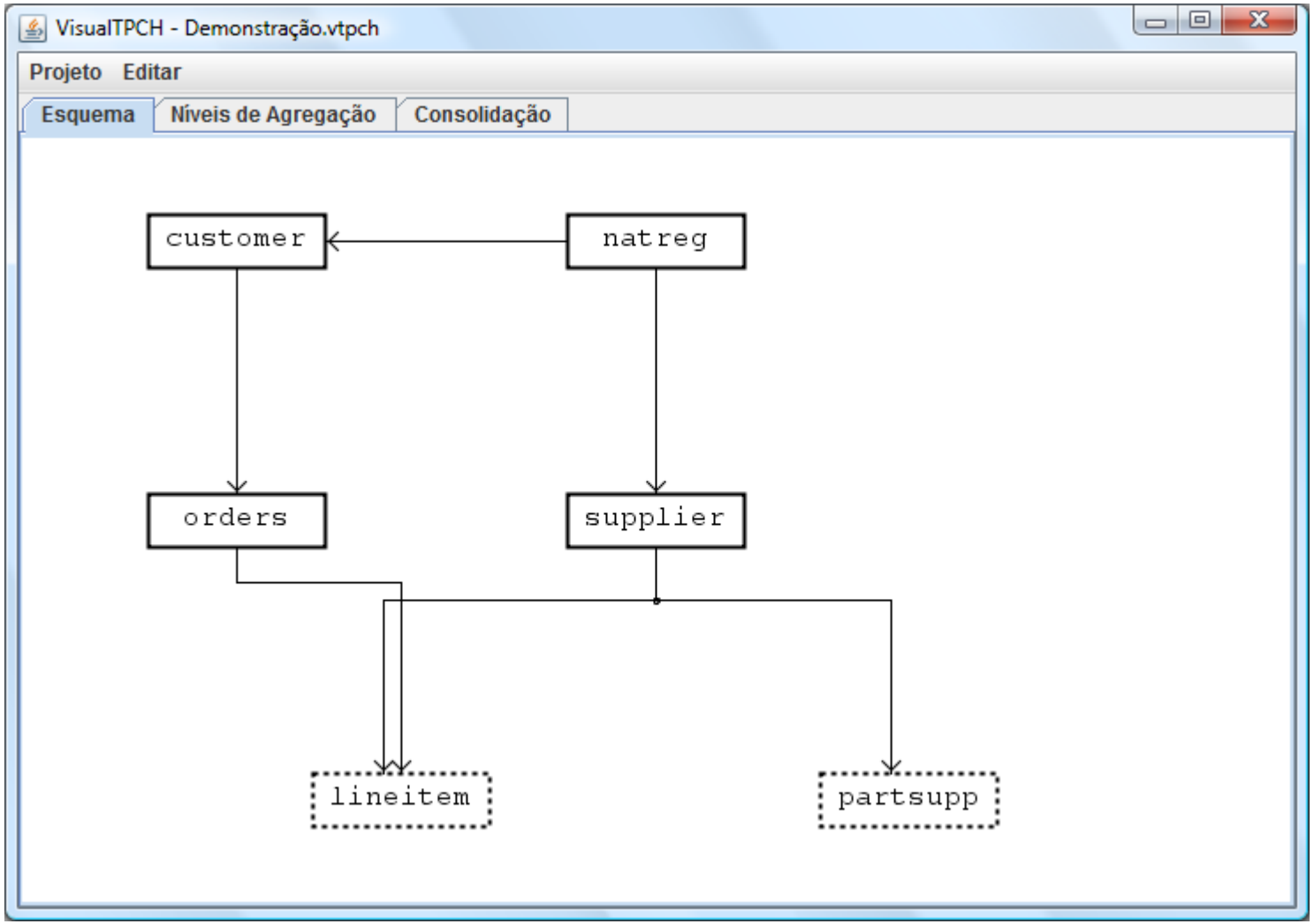
Esquema: Renomear Tabela



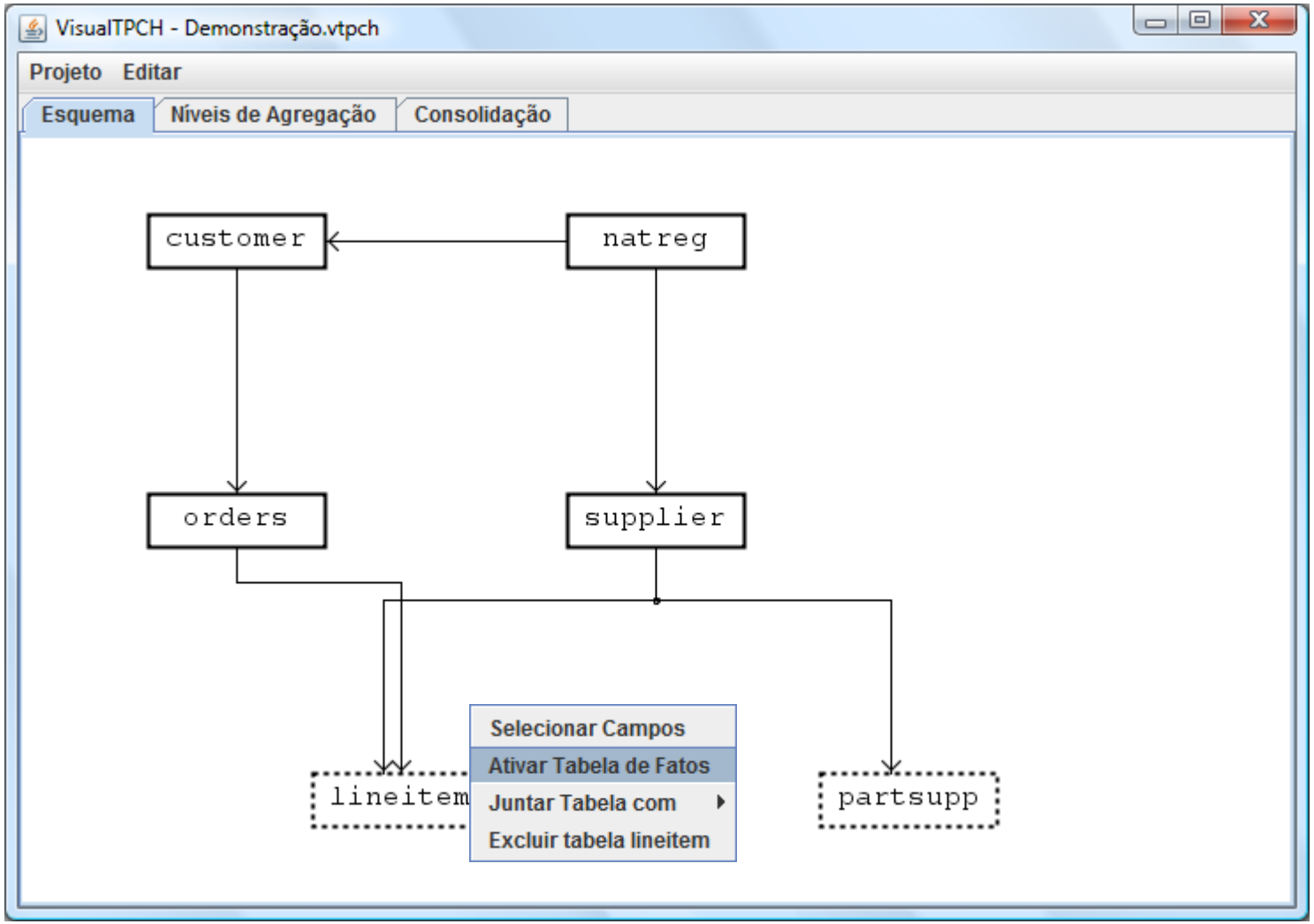
Esquema: Excluir Tabela



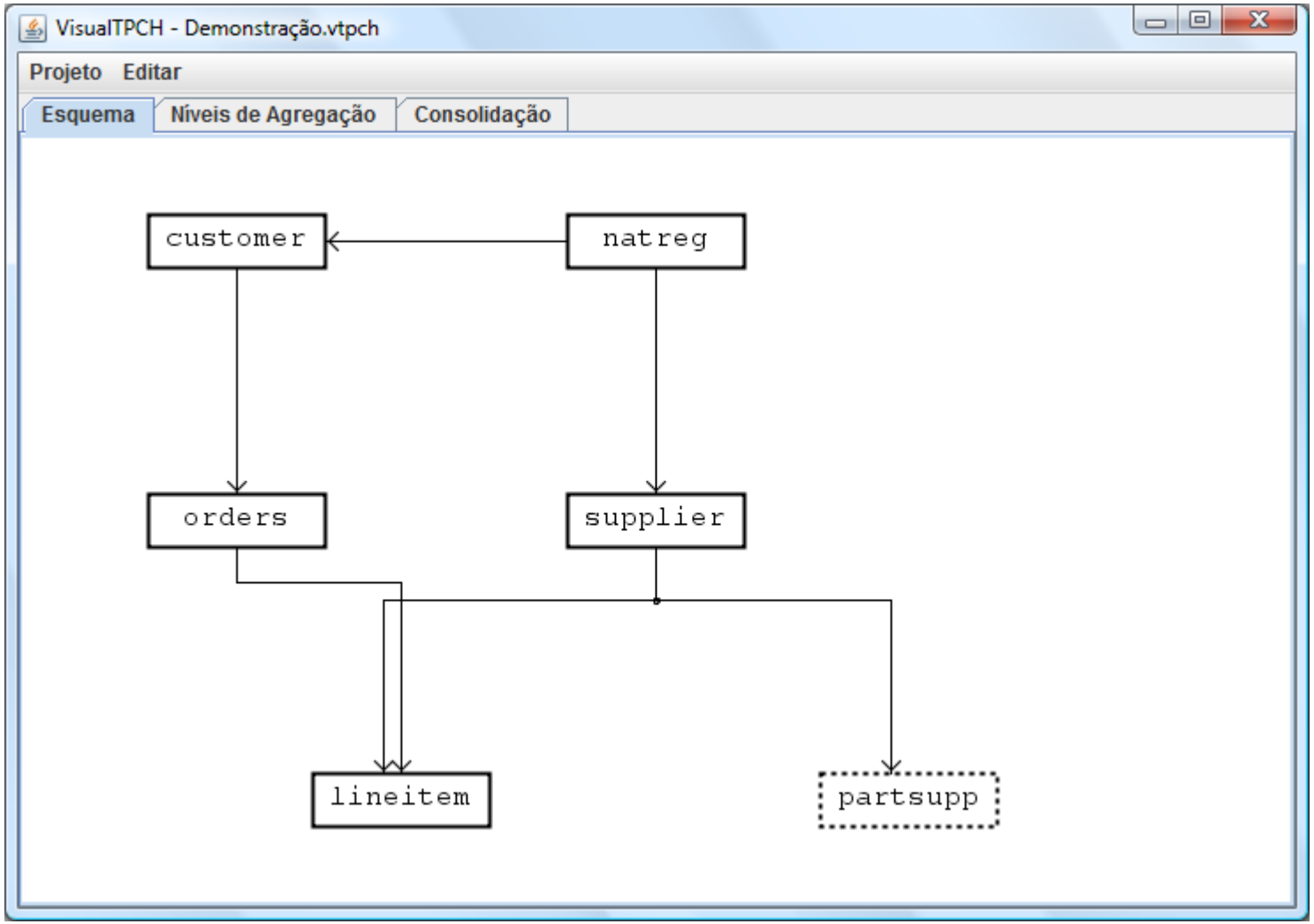
Esquema: Excluir Tabela



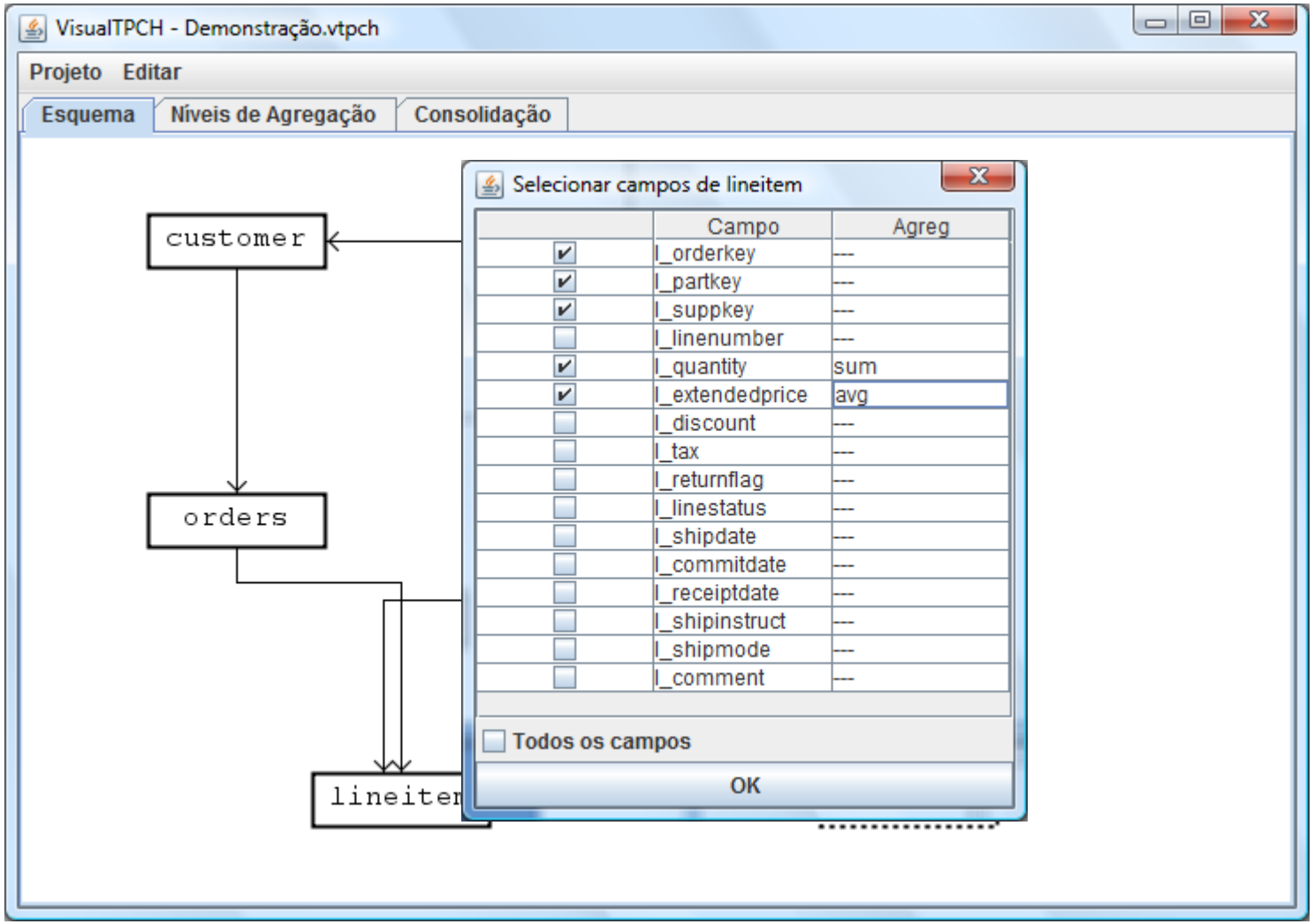
Esquema: Ativar Tabela de Fatos



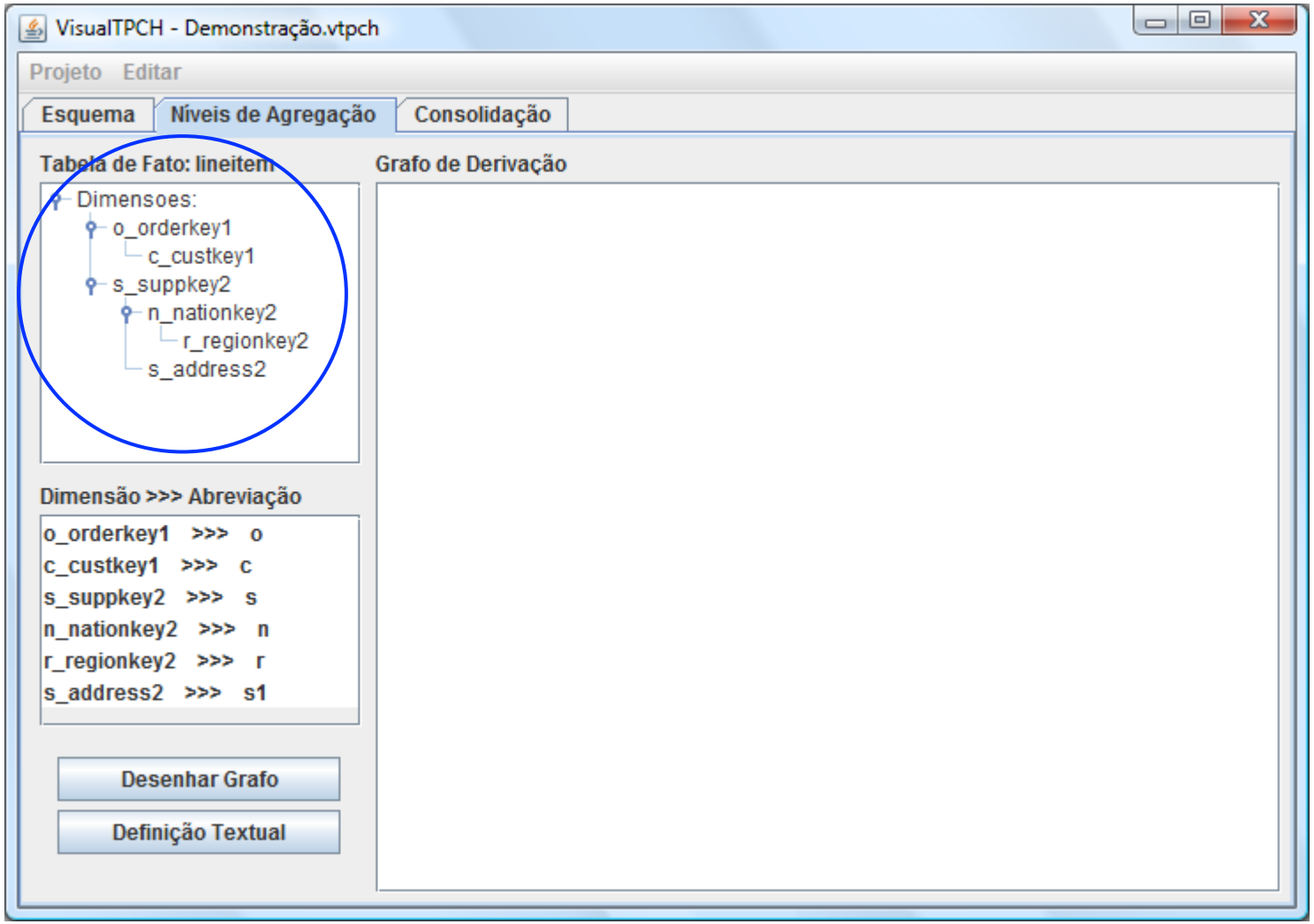
Esquema: Ativar Tabela de Fatos



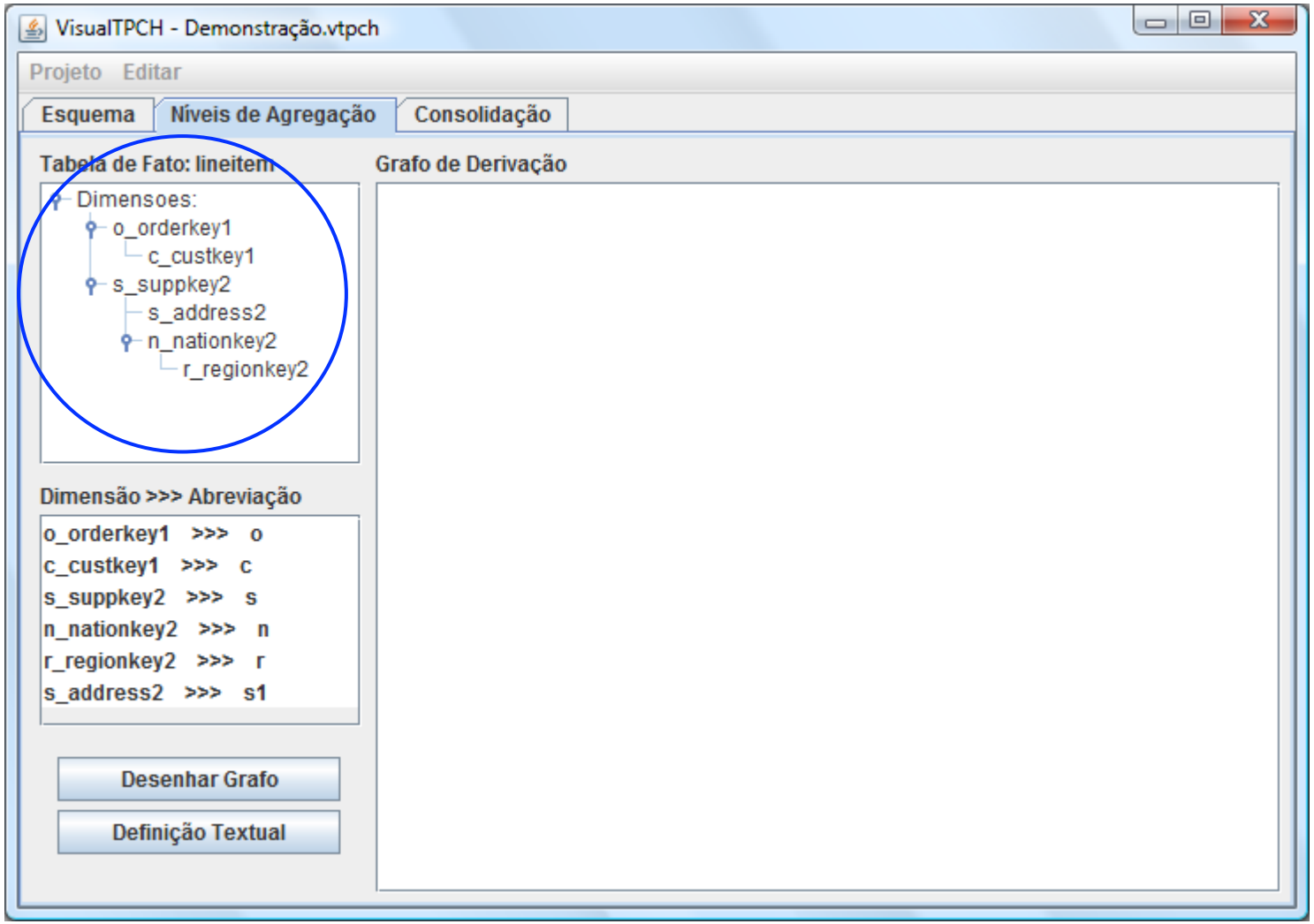
Esquema: Selecionar Campos



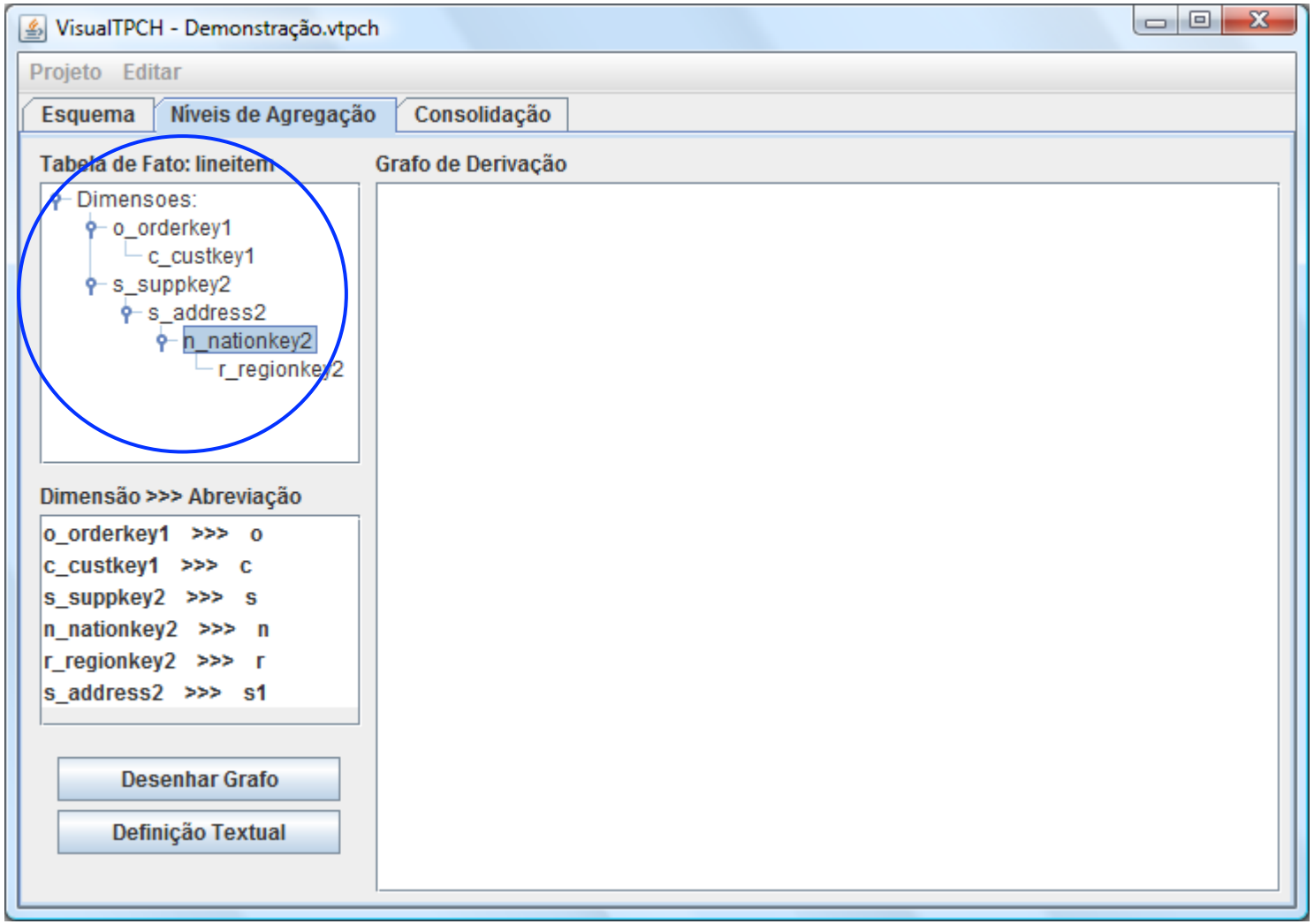
Níveis de Agregação: Dimensões



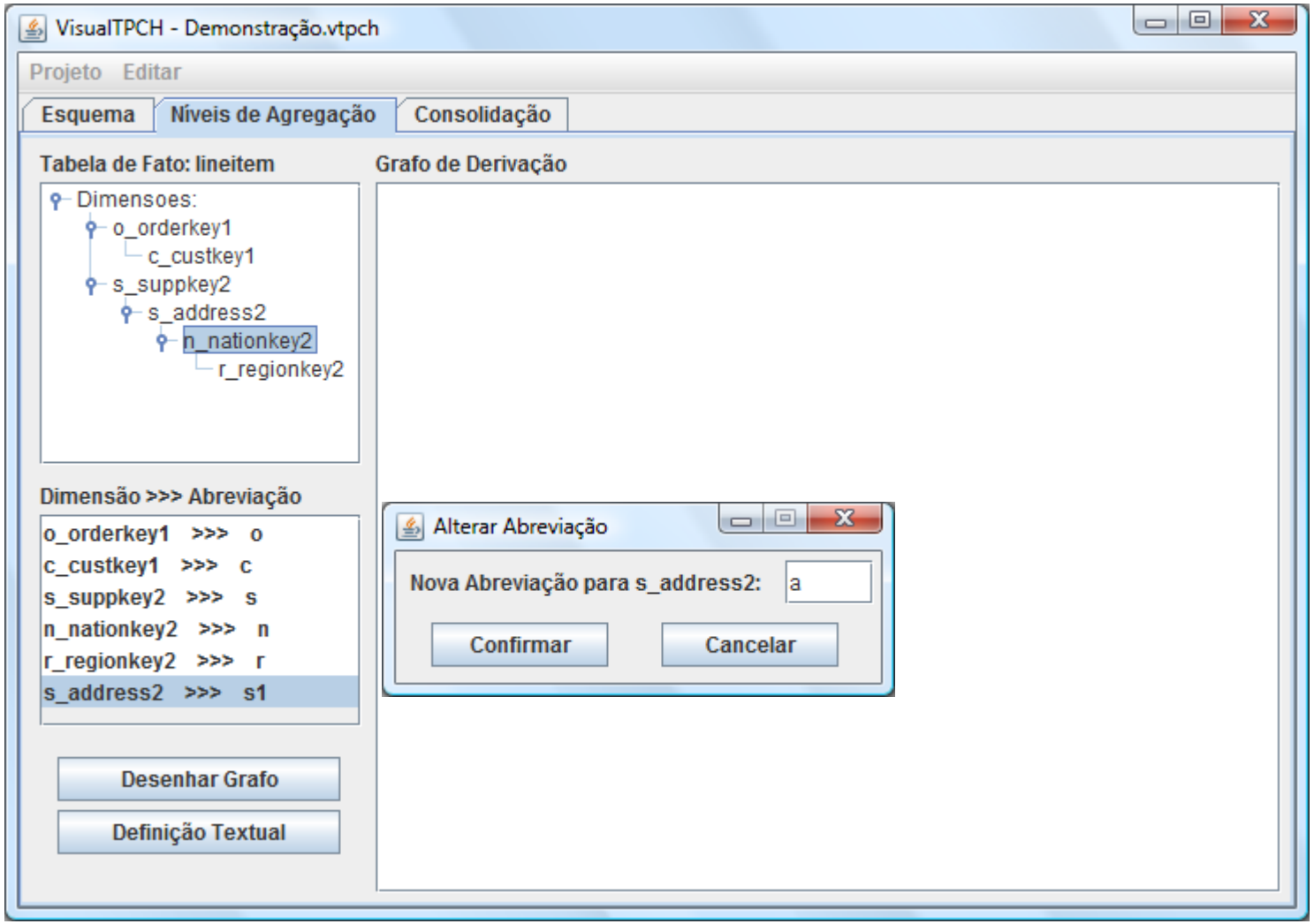
Níveis de Agregação: Dimensões



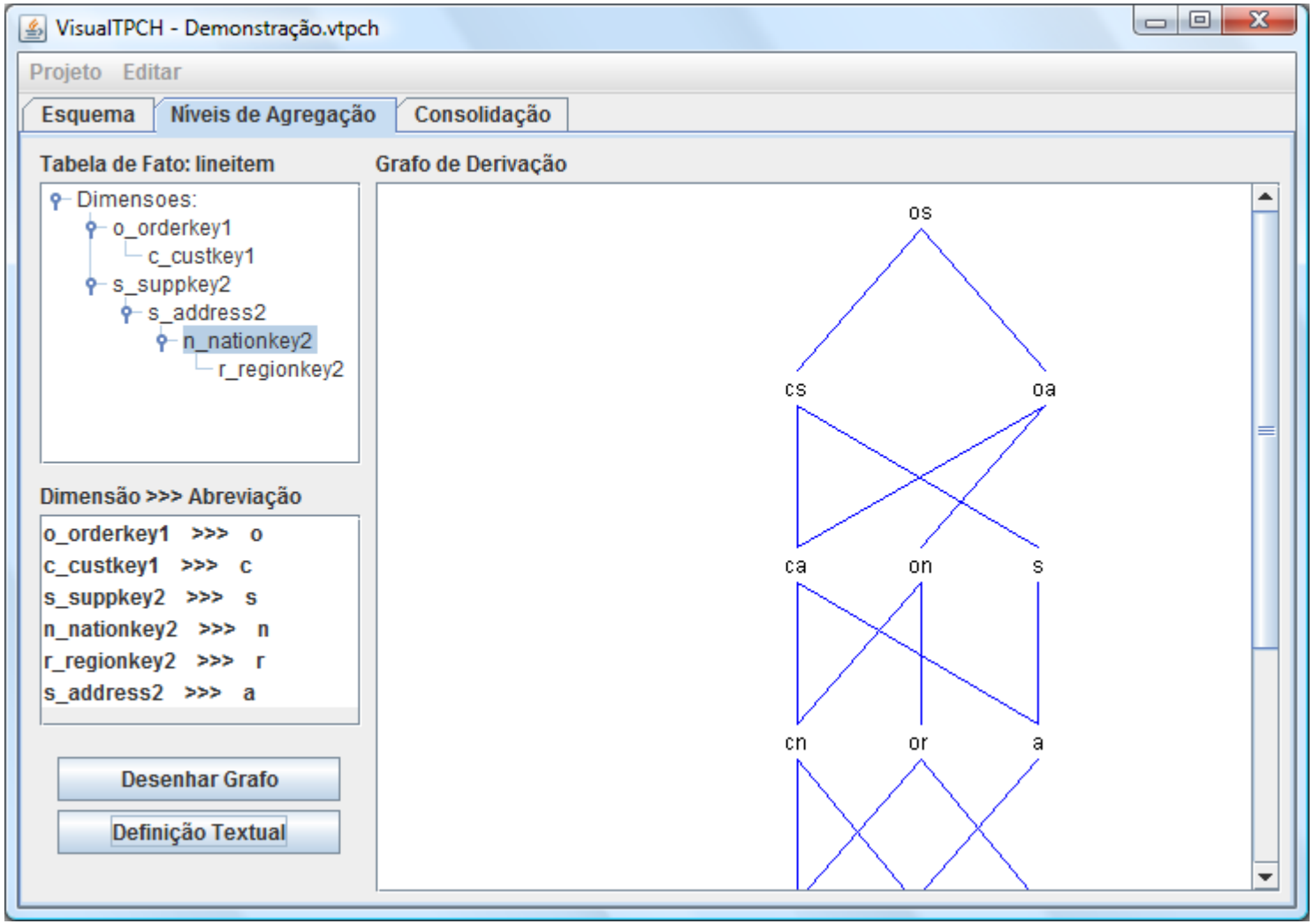
Níveis de Agregação: Dimensões



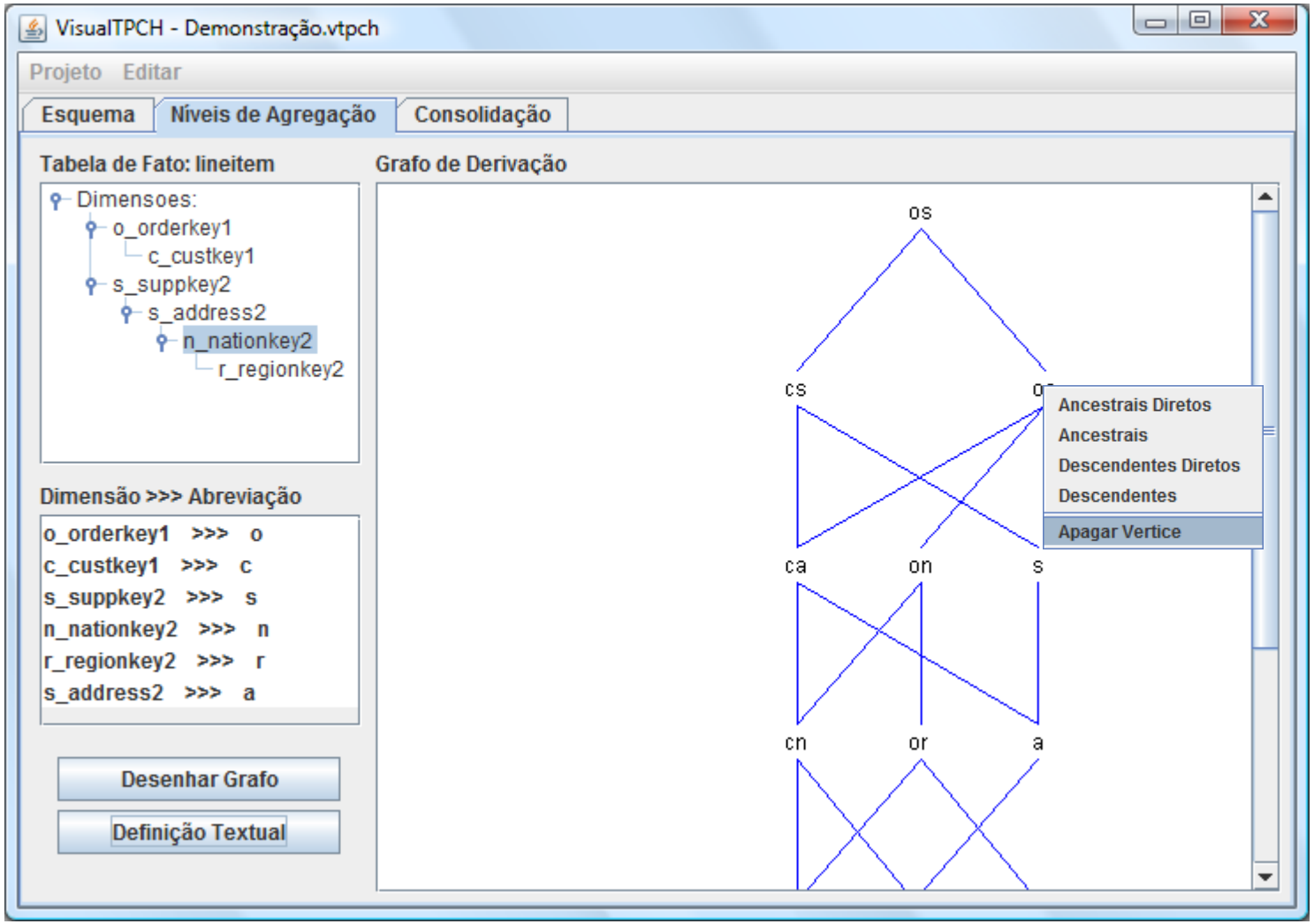
Níveis de Agregação: Abreviação



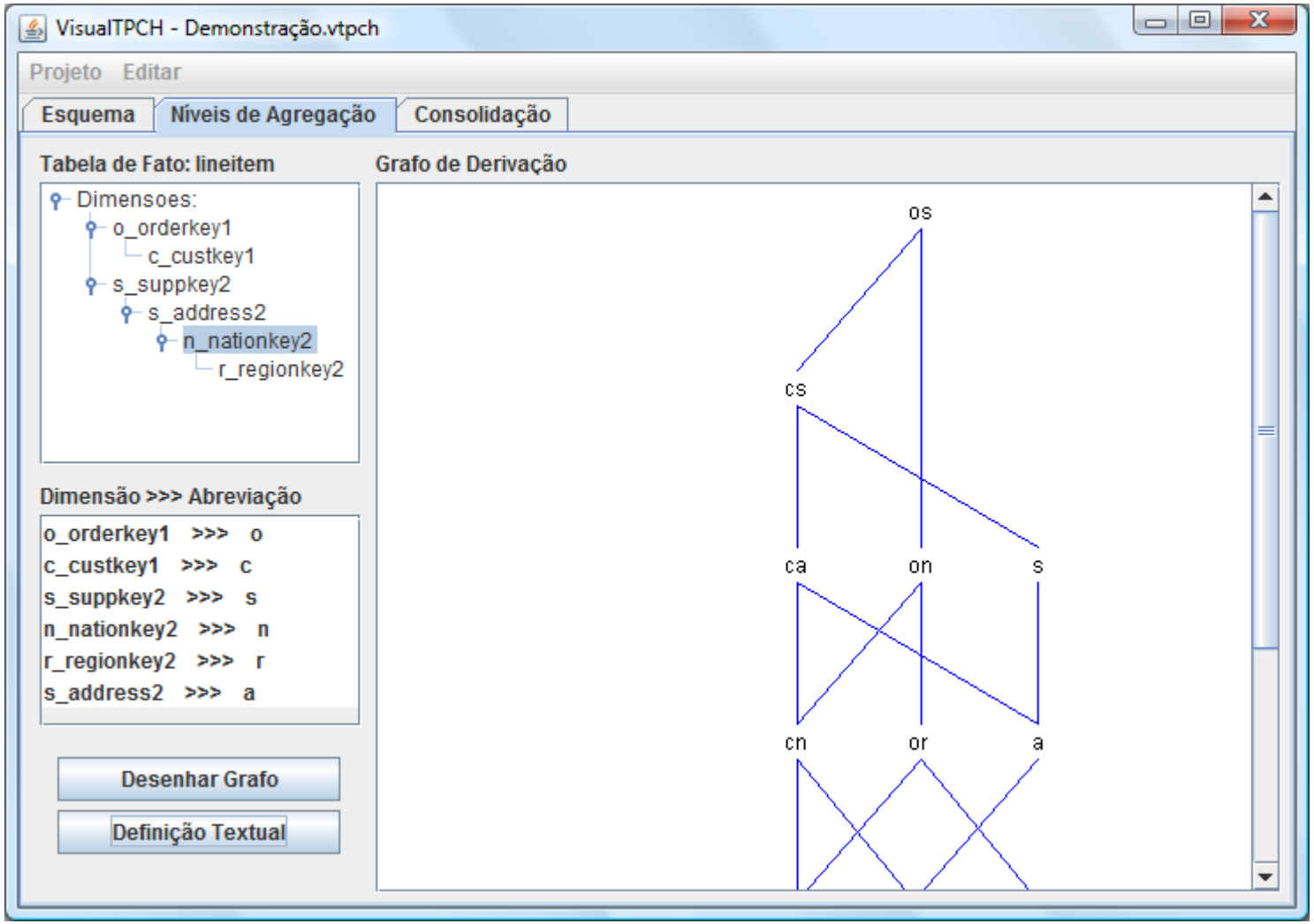
Níveis de Agregação: Grafo



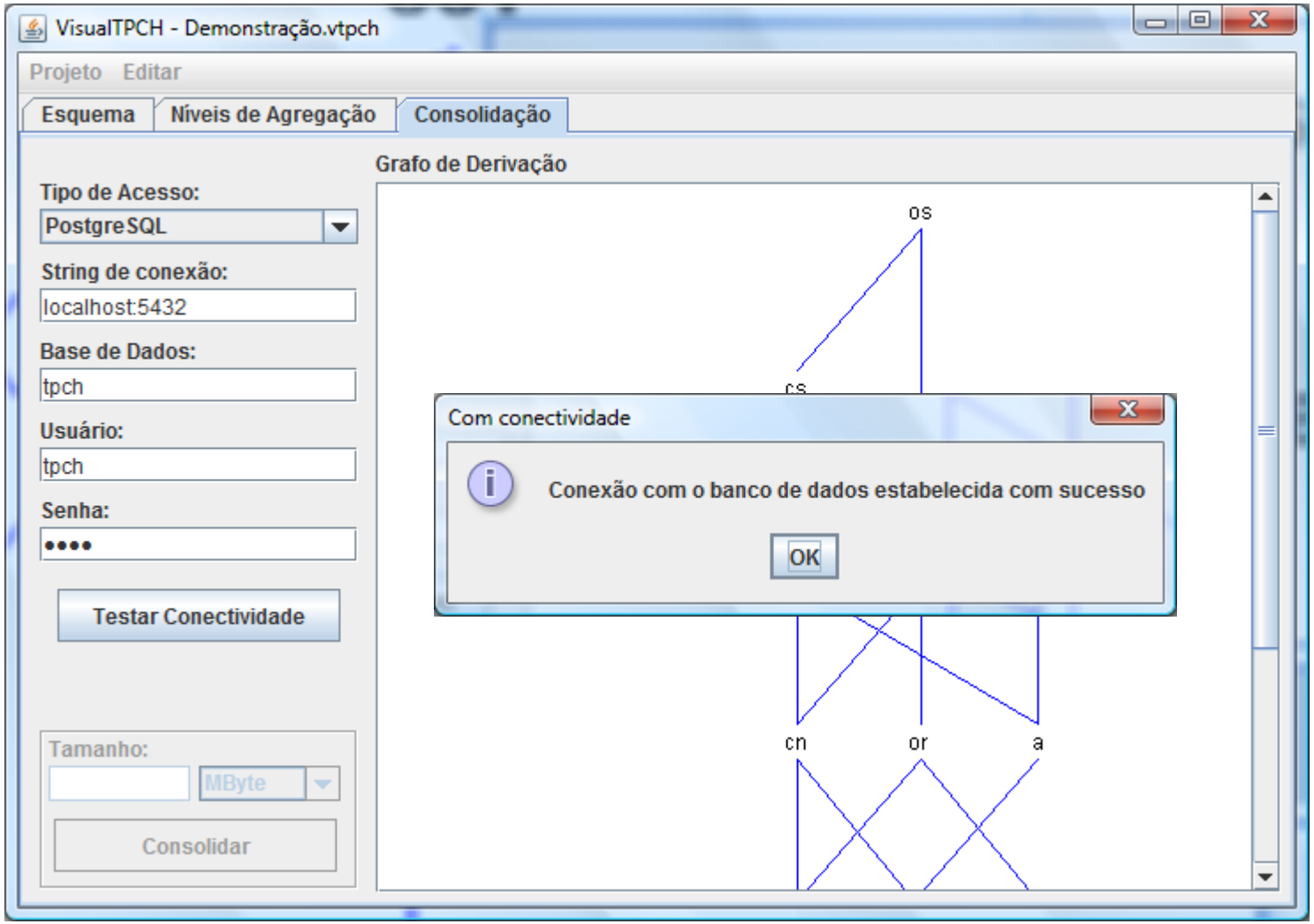
Níveis de Agregação: Grafo



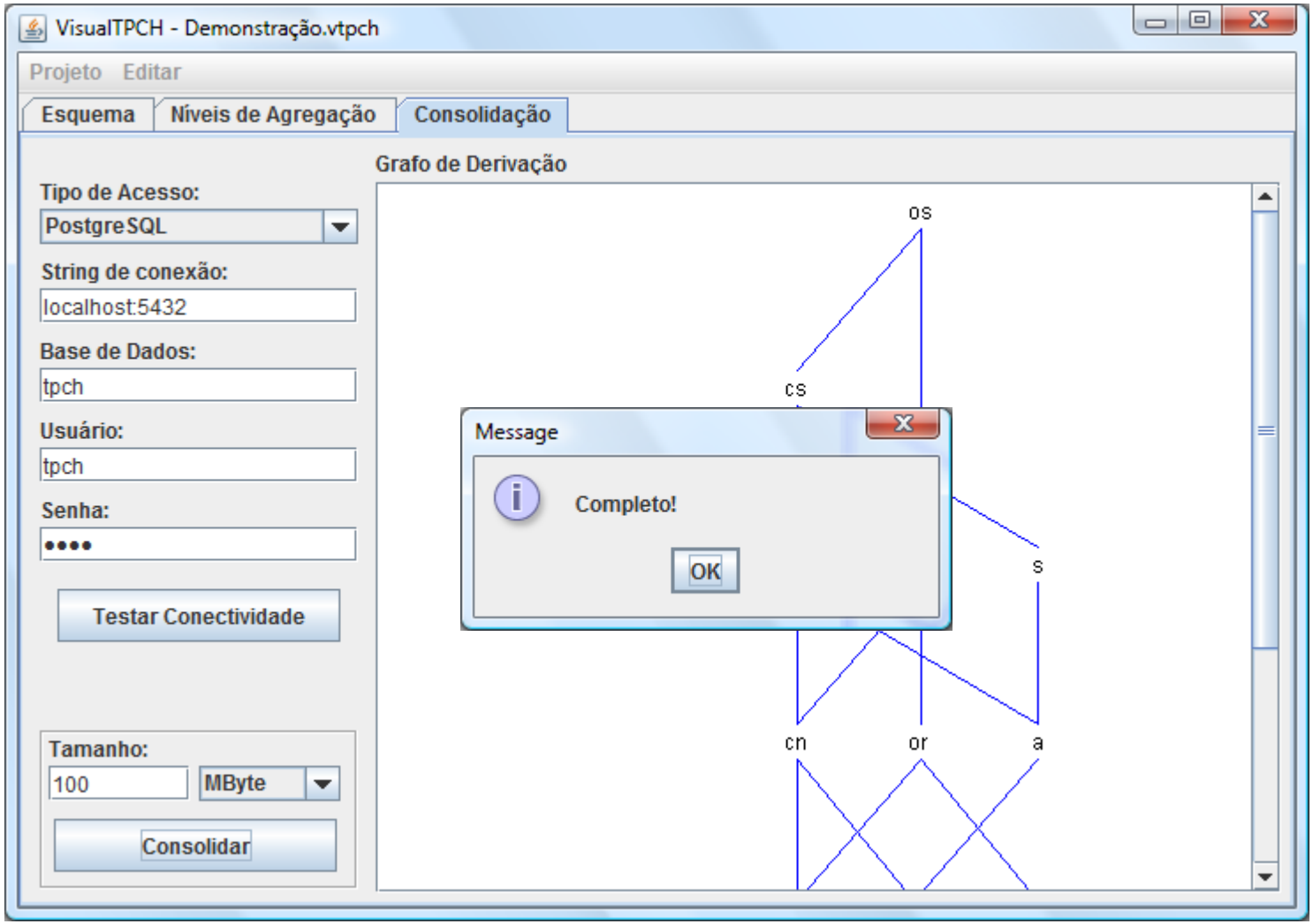
Níveis de Agregação: Grafo



Consolidação: Conectividade

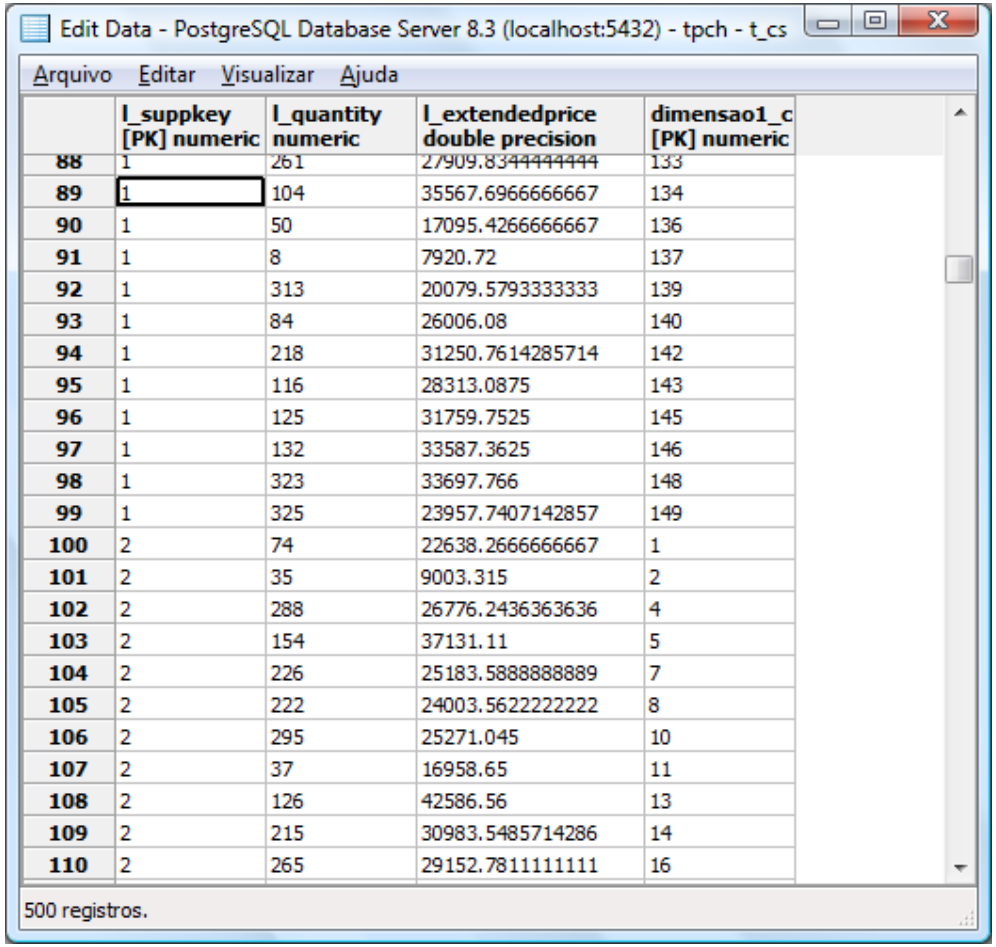


Consolidação: Geração de Dados



Tabelas e Dados

Tabela t_cs



	l_suppkey [PK] numeric	l_quantity numeric	l_extendedprice double precision	dimensao1_c [PK] numeric
88	1	261	27909.8344444444	133
89	1	104	35567.6966666667	134
90	1	50	17095.4266666667	136
91	1	8	7920.72	137
92	1	313	20079.5793333333	139
93	1	84	26006.08	140
94	1	218	31250.7614285714	142
95	1	116	28313.0875	143
96	1	125	31759.7525	145
97	1	132	33587.3625	146
98	1	323	33697.766	148
99	1	325	23957.7407142857	149
100	2	74	22638.2666666667	1
101	2	35	9003.315	2
102	2	288	26776.2436363636	4
103	2	154	37131.11	5
104	2	226	25183.5888888889	7
105	2	222	24003.5622222222	8
106	2	295	25271.045	10
107	2	37	16958.65	11
108	2	126	42586.56	13
109	2	215	30983.5485714286	14
110	2	265	29152.7811111111	16

Conclusão

- Portabilidade
  - linguagem JAVA
  - IDE Eclipse v.3.2

sistemas operacionais
Linux e Windows

- Características
  - alteração do esquema do TPC-H
  - geração de grafos e incompletos
  - automatização da geração e carga dos dados
  - suporte a diferentes SGBD

DB2, Oracle,
PostgreSQL