DATA WAREHOUSE – GOIÁS MARKET

O *Data Warehouse*, ou DW, é uma das partes mais importantes do projeto de *Business Intelligence*. Aqui, culmina-se todos os esforços de cargas e transformações de dados – ETL – necessários para armazenamento de dados concentrando-as no formato correto para análises, construção de *dashboards* e *reports* necessários para os analistas, gestores e tomadores de decisões no nível gerencial.

A partir do Data Warehouse, assim como os próximos projetos, irão tratar do ambiente OLAP, destinado à análises e automação de relatórios.

Um rápido paralelo aos conceitos de modelagem de um DW, pode-se citar os *Data Mart*, que são bancos de dados com a mesma função do DW, porém, com um assunto (ou Fato) específico. Como exemplo, *Data Mart* destinados ao Marketing, ou Logística, ou Financeiro, ou Comercial, etc.

Neste projeto, não serão modelados *Data Mart*. Apenas o *Data Warehouse* que irá armazenar todos os dados do projeto.

Portanto, após as etapas de ETL do sistema OLTP para a *STAGE AREA* e, em seguida, para o *Data Warehouse*, os dados já estão prontos para ingestão e análise.

Como será definido mais adiante, é importante ressaltar que o DW também possui a função de guardar o histórico de modificações das entradas do OLTP.

Uma vez que algumas informações podem ser modificadas, tal como endereço, local de residência, telefone, sexo,

OBJETIVO DO PROJETO – DATA WAREHOUSE

O projeto do *Data Warehouse* tem como objetivo o armazenamento dos dados oriundos do sistema transacional do negócio (banco OLTP), a fim de diminuir o tráfego de informações que causam lentidões, e disponibilizar esses dados de forma estruturada para serem analisados e interpretados pelos analistas e tomadores de decisões.

Para este projeto, o objetivo do DW visa responder as seguintes questões de negócio:

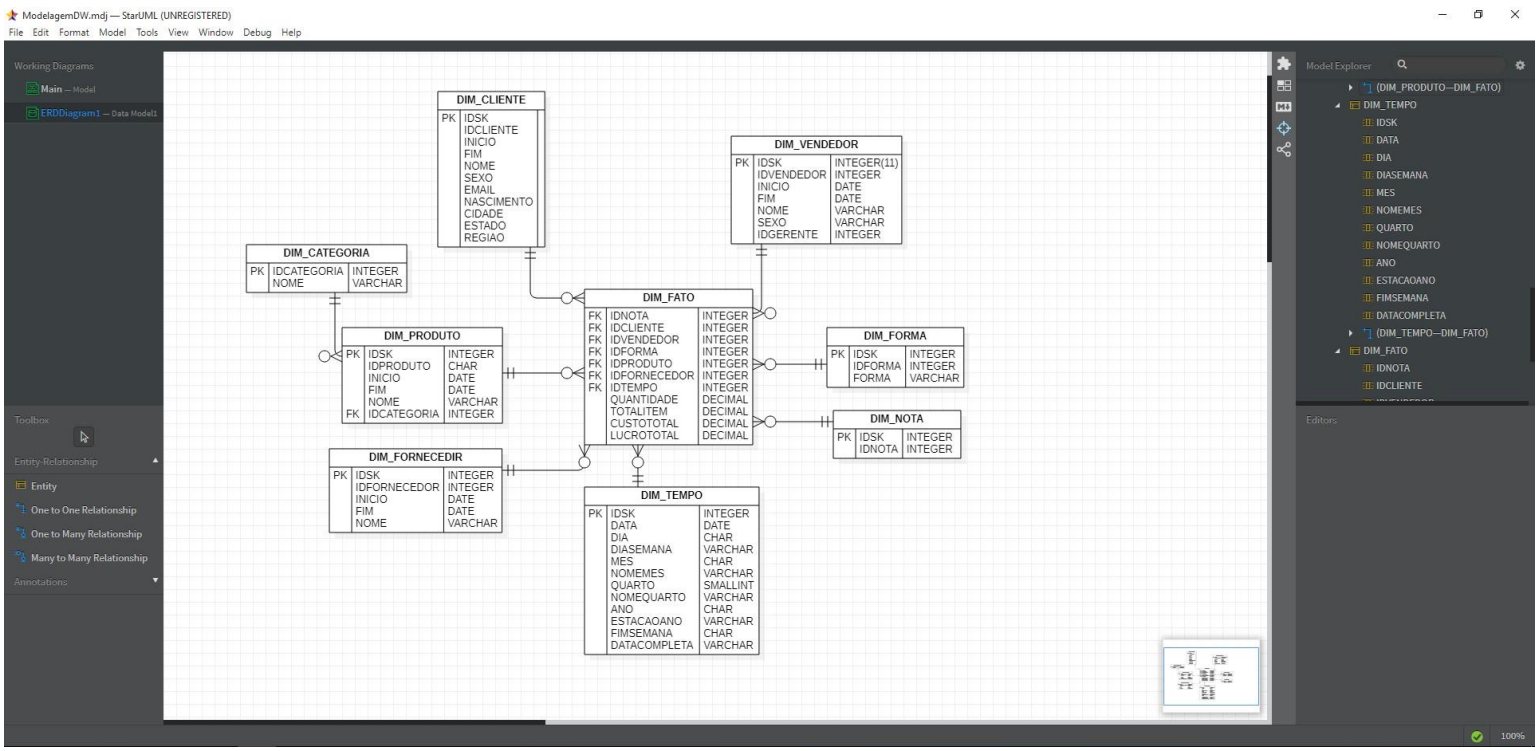
- Quem são os melhores clientes?
- Quem são os melhores vendedores?
- Qual categoria rende mais?
- Qual a minha relação com os fornecedores?
- Qual meu pior e melhor produto?
- Em qual região eu vendo mais?

MODELAGEM LÓGICA

A modelagem do *Data Warehouse* possui grande foco na integração das tabelas. A integridade relacional deve ser garantida a fim de que todos os dados estejam conectados corretamente através das *Primary Key* (PK, ou chave primária) e *Surrogate Key* (SK, ou chave substituta).

No modelo a seguir, pode-se destacar a tabela DIM_FATO, que é a tabela principal do DW. Essa tabela poderia ser facilmente escrita por um Query SQL, trazendo todos os dados na seleção. Entretanto, a tabela FATO consolida a função do *Data Warehouse*, facilitando a consulta de dados.

A tabela FATO, como demonstrado abaixo, recebe os relacionamentos de cardinalidade de 1 para muitos (1 x N). Isto significa que a tabela fato recebe os registros únicos de cada tabela relacionada, bem como o seu histórico de modificação, consolidando as informações de uma PK em um único lugar.



Os dados armazenados no *Data Warehouse* já estão prontos para consultas pelos analistas, sejam elas através de *queries* diretamente do banco de dados, por queries SQL, por exportação para um sistema de análise em ambiente OLAP (como o SQL Server Analysis Services) ou por uma conexão com alguma fermenta de *DataViz* (como o Microsoft Power BI, Tableu, Qlik Sense e etc).

Nesse ponto, é válido também a conexão com ferramentas de Data Science, como R e Python, que são capazes de acessar diretamente o DW através de bibliotecas específicas para conexão e interface com SQL.

Os pipelines em R e Python serão modelados futuramente neste mesmo projeto, para modelagem de dados, aplicação de modelos de *Machine Learning* e *Data Mining*.

Ao modelar o DW, há um novo atributo nas tabelas em relação ao banco OLTP e a *Stage Area*, são as colunas “INICIO” e “FIM, do tipo *DATE.* Estas colunas serão responsáveis por manter o registro de alteração de algum item.

O conceito que se aplica em tabelas com capacidade de armazenar histórico por alguma alteração, são definidas como dimensões mutáveis, ou SCD (*Slowly Changing Dimensions*)

Dessa forma, é possível guardar um registro de alteração de um item ao longo tempo sem perder a rastreabilidade, pois a chave primeira (PK) se manterá igual, e os outros atributos que serão alterados terão uma data de validade, representada na coluna “FIM”, que é a data em que foi feita a alteração e este item será substituído pela sua alteração.

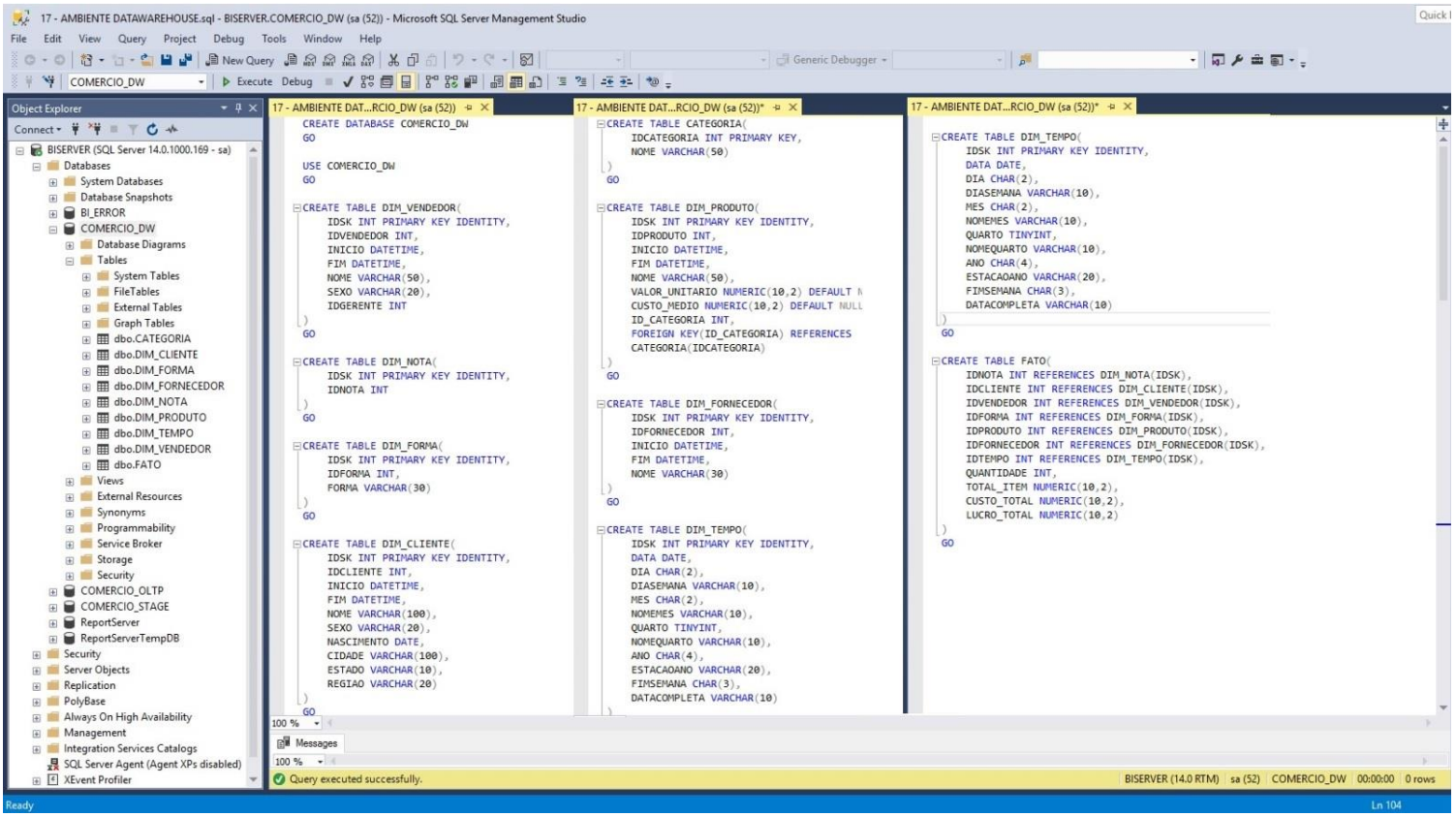
Feito isso, cria-se uma nova entrada no DW com a mesma *Primary Key* (PK), porém com uma *Surrogate Key* (SK) diferente, já com as devidas alterações.

As tabelas que podem sofrer esse tipo de alteração, são: DIM_PRODUTO, DIM_CLIENTE, DIM_FORNECEDOR E DIM_VENDEDOR. Todas essas tabelas podem sofrer alteração em alguma de suas entradas mediante a necessidade e serem salvas no *Data Warehouse* durante o processo de ETL. O *Integration Services* será responsável por esse processo e será exemplificado a seguir.

Com os conceitos já definidos e a modelagem lógica feita, partimos então para a modelagem física do *Data Warehouse*.

MODELAGEM FÍSICA

Começamos então a modelagem física no banco de dados, pelo SQL Server, com a criação do banco de dados do *Data Warehouse*. Neste projeto, o DW será chamado de COMERCIO_DW.



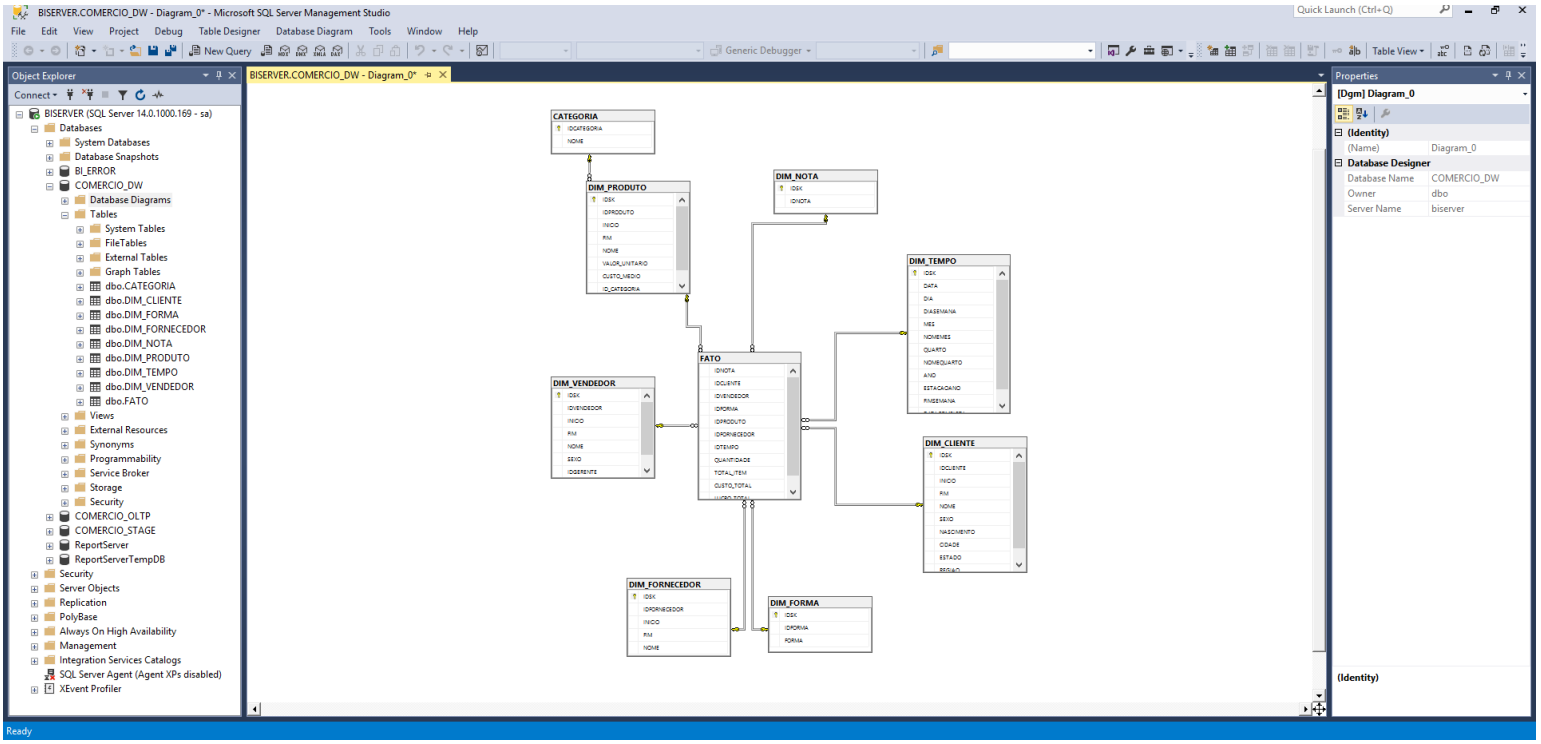
As tabelas foram criadas com o prefixo “DIM_”, indicando as dimensões do DW separadas por assunto.

As desse banco foram criadas em uma ordem específica já com os seus relacionamentos (ou *Constraints*) definidos, interligando as chaves primárias de cada uma. Apesar de que o procedimento de criação de *constraints* possa ser feito a parte, foram aplicados mediante a criação das tabelas para evitar qualquer fuga de relacionamentos.

A tabela DIM_FATO, ilustra bem a modelagem de *constraints* pois é a tabela mais importante do DW, que necessita estar totalmente referenciada com as dimensões que a compõe.

A função SQL utilizada para referenciar as tabelas durante a criação, segue na última *query*: REFERENCES

Neste momento, a banco já está criado, como segue o diagrama gerado pelo SQL Server, e as tabelas já estão prontas para receber os dados por ETL.

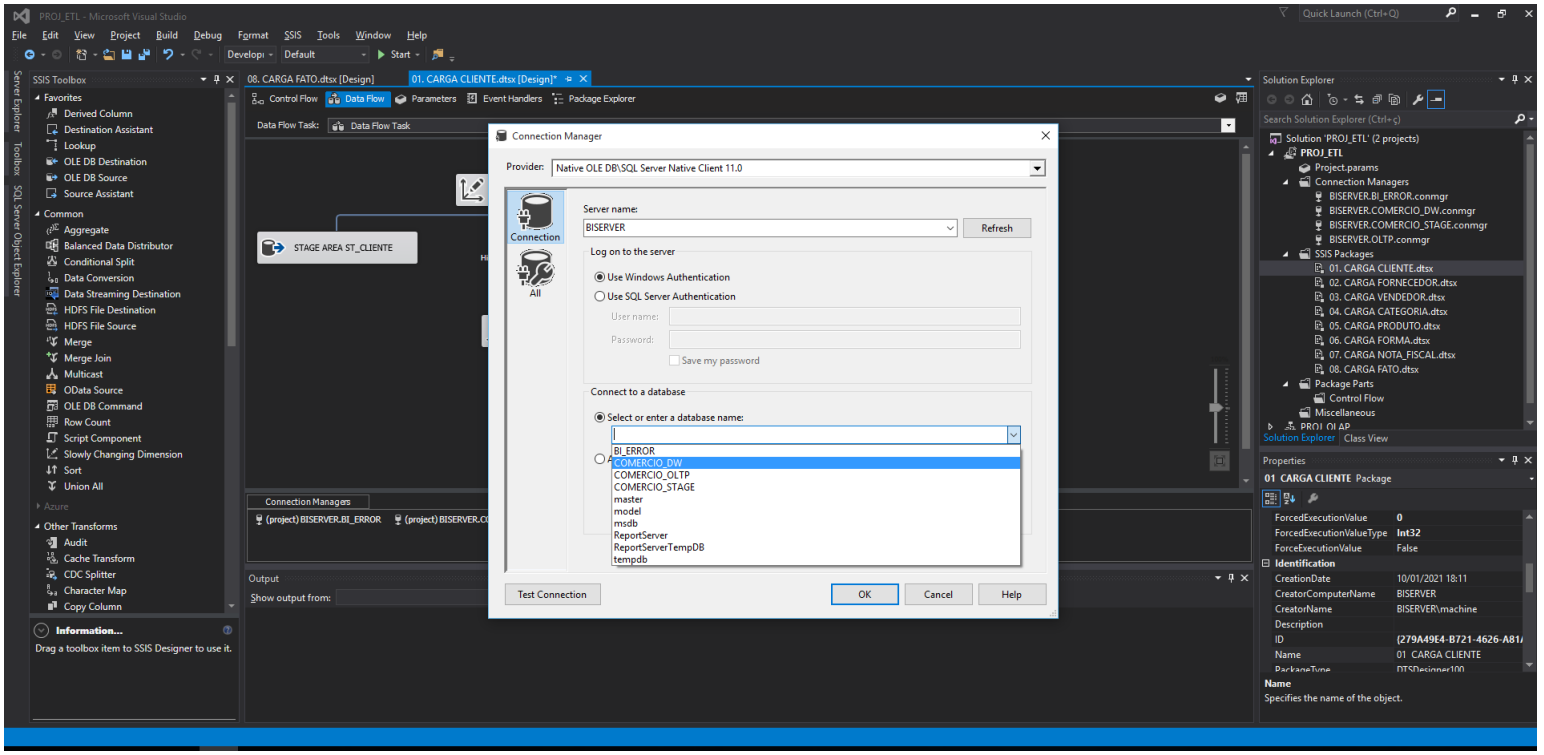


SINCRONIZAÇÃO DE ETL COM INTEGRATION SERVICES

O *Integration Service* (SSIS) é uma versátil ferramenta de ETL, construindo os pacotes de cargas e os gerenciando por projetos e por solução, como segue no menu à direita.

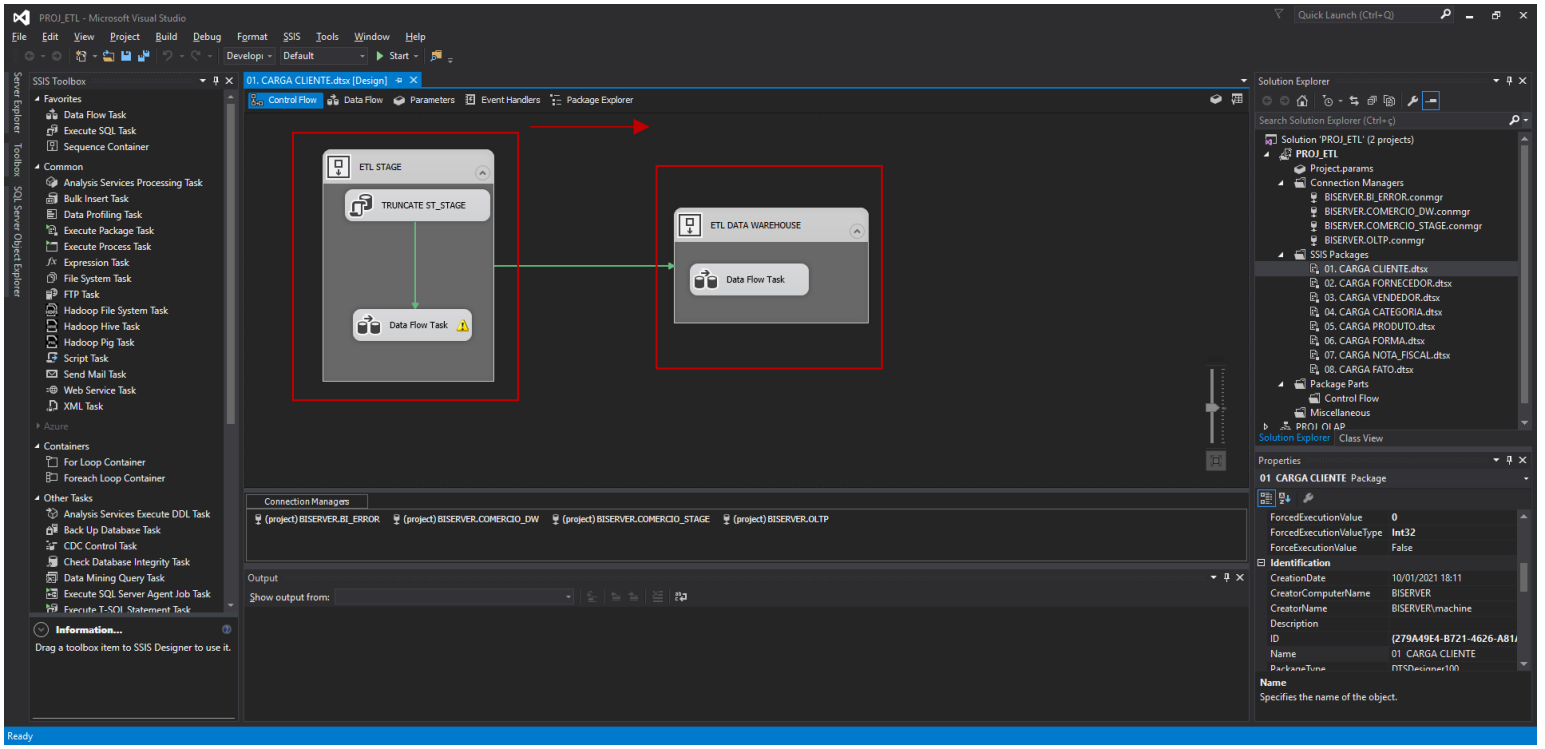
Nesse projeto, os pacotes terão o processo de ETL sincronizados com a carga da Stage Area. Isso significa que o pacote irá fazer separadamente e consequentemente as cargas dos bancos.

As conexões entre os bancos foram definidas previamente e foram compartilhadas como conexões de toda essa solução BI. Neste momento, é criada a conexão com o banco COMERCIO_DW, que é o *Data Warehouse*.



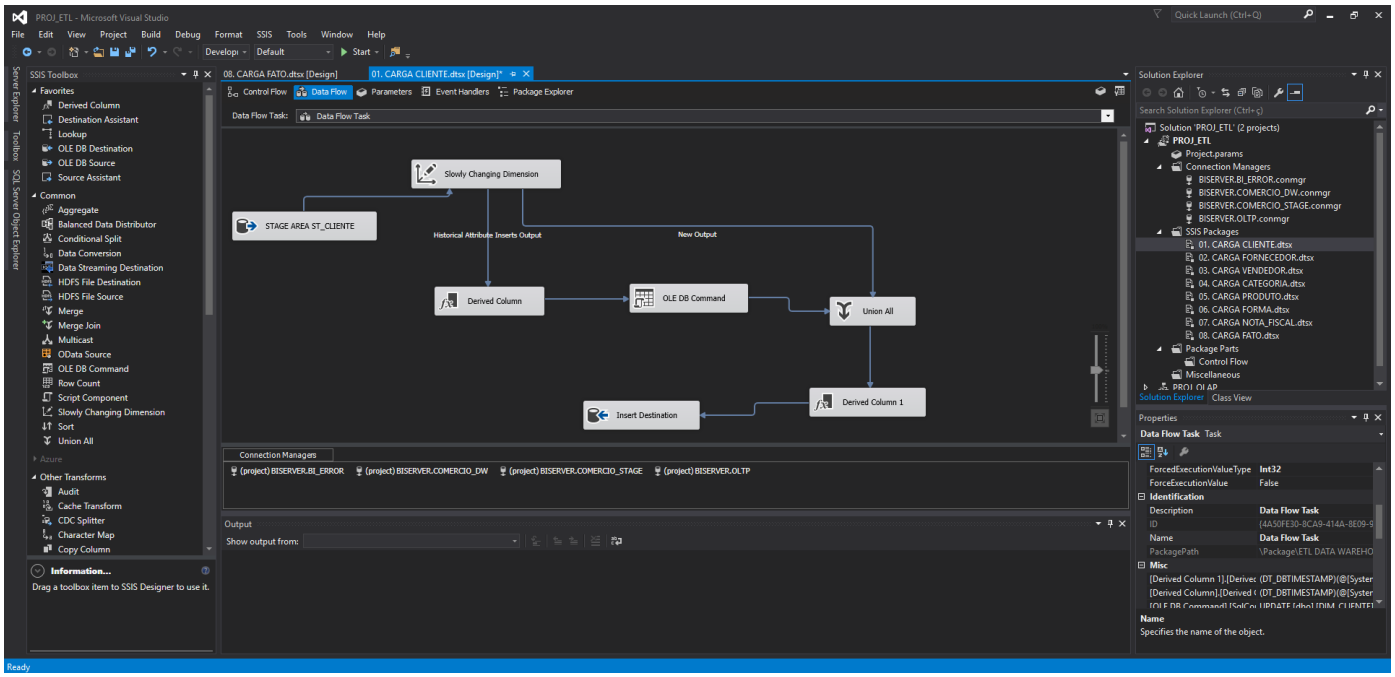
Como mostra a *screenshot* abaixo, os processos foram separados por uma ferramenta chamada ***Sequence Container***. Essa ferramenta é capaz de separar os processos de ETL da área de Stage, que é feita no primeiro momento, para então, depois de finalizado, fazer a carga do *Data Warehouse*.

Este processo faz todo sentido uma vez que a *Stage Area* serve apenas para carga e transformações necessárias dos dados. Então, ao finalizar a carga da *Stage* pelo pacote 'ETL_STAGE', inicia-se o container seguinte, '*ETL DATA WAREHOUSE*', com a carga do DW.

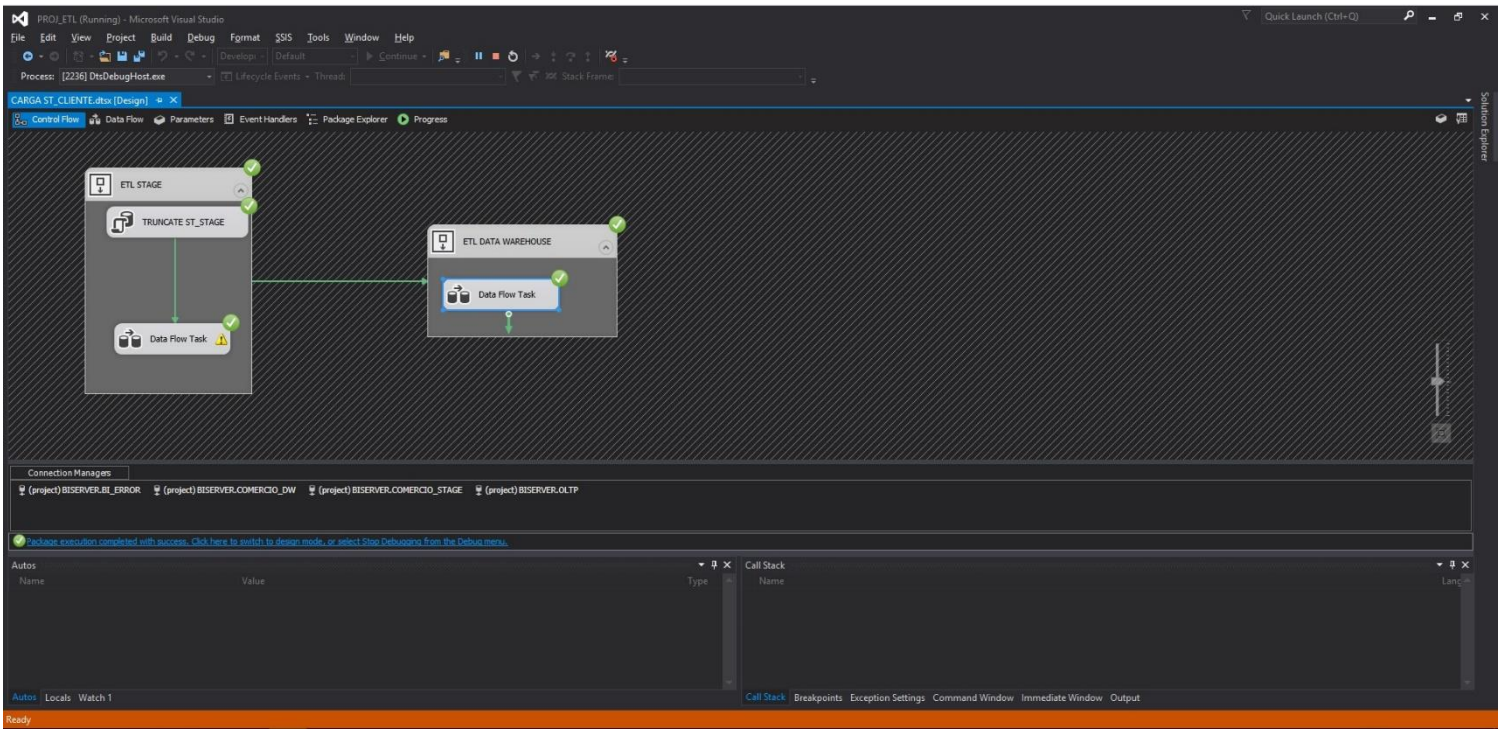


O processo interno de ETL do pacote do *Data Warehouse*, ou ***Data Flow***, possui um processo de carga um pouco diferente das demais, uma vez que esta carga se baseia em SCD (Slowly Changing Dimensions ou dimensões mutáveis) para garantir o armazenamento do histórico.

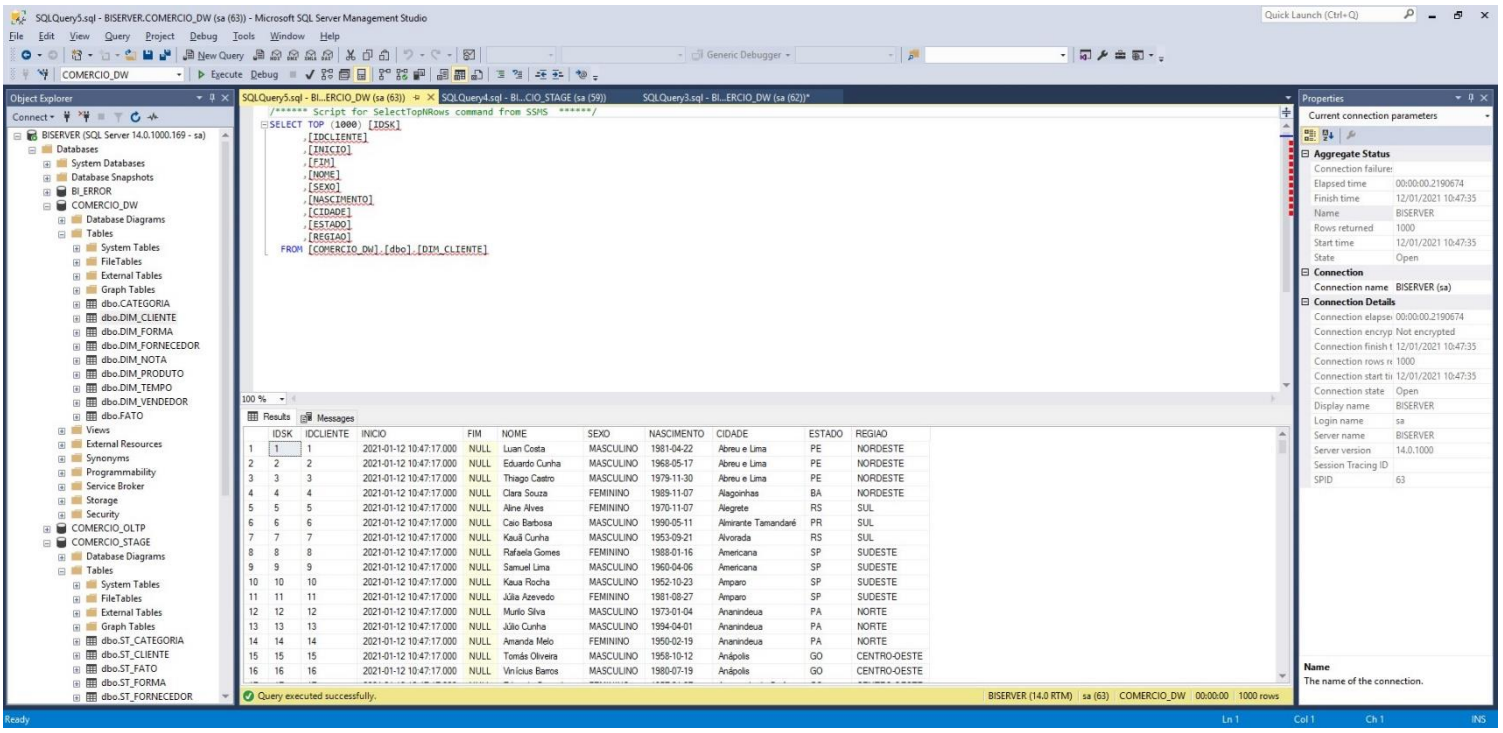
É adicionado uma ferramenta ao pacote chamada 'Slowly Changing Dimensions', encontrada no menu de transformações comuns do SSIS. O assistente da ferramenta define as opções para a tabela e cria automaticamente o *Data Flow*, já conectado ao source de dados ('STAGE AREA ST_CLIENTE').



Ao finalizar o processo de ETL, podemos ver a confirmação de êxito do processo, como na *screenshot* abaixo:



Ao finalizar o pacote do SSIS, podemos consultar no SQL Server a carga dos dados, realizando um SELECT simples na tabela DIM_CLIENTE



IDSK	IDCLIENTE	INICIO	FIM	NOME	SEXO	NASCIMENTO	CIDADE	ESTADO	REGIAO
1	1	2021-01-12 10:47:17.000	NULL	Luan Costa	MASCULINO	1981-04-22	Abreu e Lima	PE	NORDESTE
2	2	2021-01-12 10:47:17.000	NULL	Eduardo Cunha	MASCULINO	1968-05-17	Abreu e Lima	PE	NORDESTE
3	3	2021-01-12 10:47:17.000	NULL	Thiago Castro	MASCULINO	1979-11-30	Abreu e Lima	PE	NORDESTE
4	4	2021-01-12 10:47:17.000	NULL	Clara Souza	FEMININO	1989-11-07	Alagoinhas	BA	NORDESTE
5	5	2021-01-12 10:47:17.000	NULL	Aline Alves	FEMININO	1970-11-07	Alegrete	RS	SUL
6	6	2021-01-12 10:47:17.000	NULL	Caio Barbosa	MASCULINO	1990-05-11	Almirante Tamandaré	PR	SUL
7	7	2021-01-12 10:47:17.000	NULL	Kauã Cunha	MASCULINO	1953-09-21	Alvorada	RS	SUL
8	8	2021-01-12 10:47:17.000	NULL	Rafaela Gomes	FEMININO	1988-01-16	Americana	SP	SUDESTE
9	9	2021-01-12 10:47:17.000	NULL	Samuel Lima	MASCULINO	1960-04-06	Americana	SP	SUDESTE
10	10	2021-01-12 10:47:17.000	NULL	Kaua Rocha	MASCULINO	1952-10-23	Amparo	SP	SUDESTE
11	11	2021-01-12 10:47:17.000	NULL	Júlia Azevedo	FEMININO	1981-08-27	Amparo	SP	SUDESTE
12	12	2021-01-12 10:47:17.000	NULL	Murilo Silva	MASCULINO	1973-01-04	Ananindeua	PA	NORTE
13	13	2021-01-12 10:47:17.000	NULL	Júlio Cunha	MASCULINO	1994-04-01	Ananindeua	PA	NORTE
14	14	2021-01-12 10:47:17.000	NULL	Amanda Melo	FEMININO	1950-02-19	Ananindeua	PA	NORTE
15	15	2021-01-12 10:47:17.000	NULL	Tomás Oliveira	MASCULINO	1958-10-12	Anápolis	GO	CENTRO-OESTE
16	16	2021-01-12 10:47:17.000	NULL	Vinícius Barros	MASCULINO	1980-07-19	Anápolis	GO	CENTRO-OESTE

Nesse ponto, podemos ver o *Data Warehouse* já disponível para queries e conexão com as ferramentas de BI. Claro, após a carga de todas as outras tabelas.

O processo de ETL demonstrado aqui se repete igualmente para todas as outras dimensões. Portanto, não será demonstrado aqui todas as tabelas, uma vez que o processo é igual.

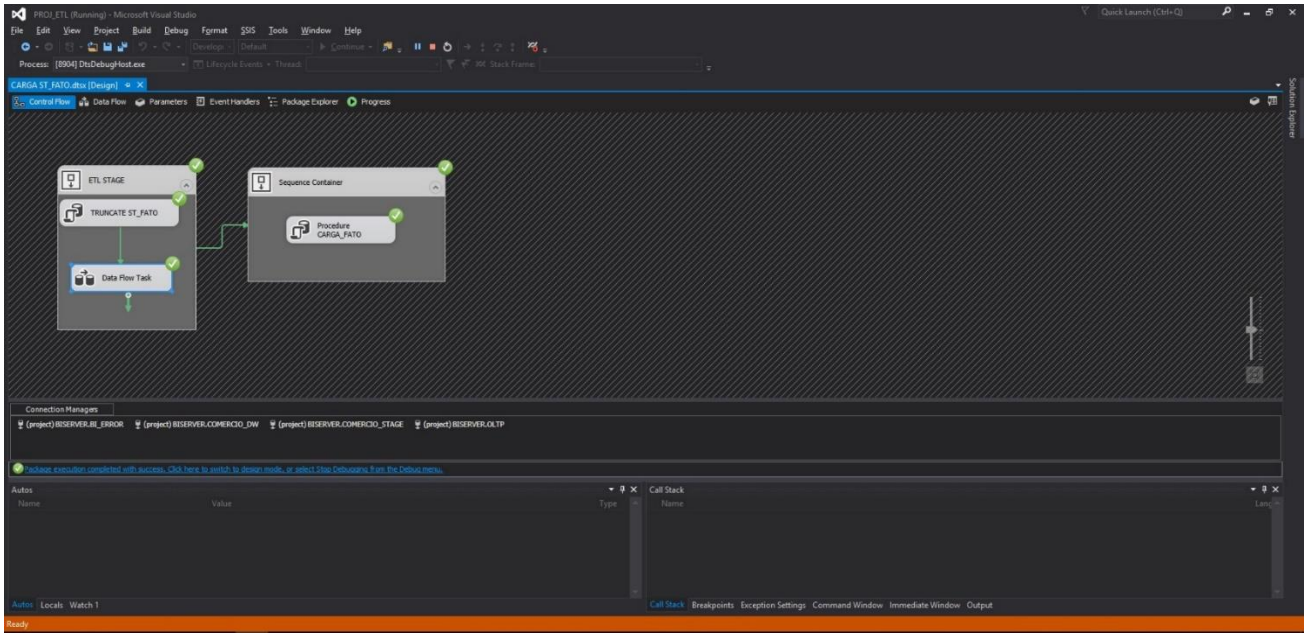
CARGA DA TABELA DIM_FATO

Por final, carregamos a tabela fato, que é a consolidação do Data Warehouse.

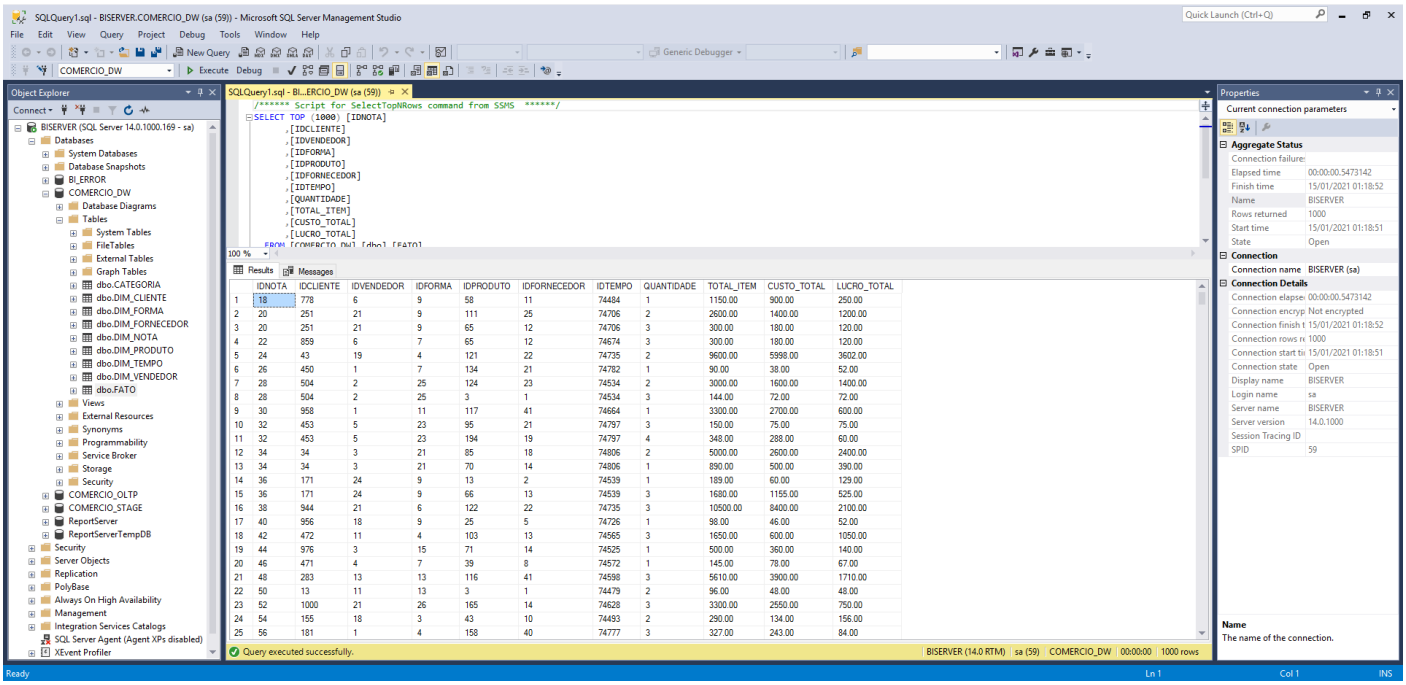
A tabela DIM_FATO é carregada utilizando um script em SQL que é capaz de filtrar os dados já existentes, comparando as entradas já existentes no DW com os novos resultados. Dessa forma, o ETL carrega apenas os novos dados inseridos no sistema OLTP.

Por motivos de privacidade, o script que foi fornecido pelo instrutor Felipe Mafra, não será descrito aqui nesse projeto. Entretanto, este pode ser obtido ao adquirir um de seus cursos na plataforma Udemy. O mesmo motivo é valido também para a DIM_TEMPO, que foi criada separadamente do SQL Server por um script de inteligência de data.

Os scripts, de forma genial, podem ser replicados em outros projetos de forma inteiramente funcional.



Após a conclusão do processo de carga da DIM_FATO, podemos verificar os dados através de uma query de consulta.



CONCLUSÃO DATA WAREHOUSE

O projeto de modelagem do *Data Warehouse* finalmente conclui-se com êxito em todos os processo de ETL e definição de relacionamentos.

O DW está pronto para ser consultados por queries em SQL e/ou receber conexões diretamente de ferramentas de *Data Visualization* para construção de dashboards.

Entretanto, nesse ponto, inicia-se os projetos do sistema OLAP, utilizando o *Analysis Services* (SSAS) e o *Reporting Services* (SSRS) para criar os cubos analíticos e automação dos relatórios.