Google Cloud

Workshop: Google Data Fusion

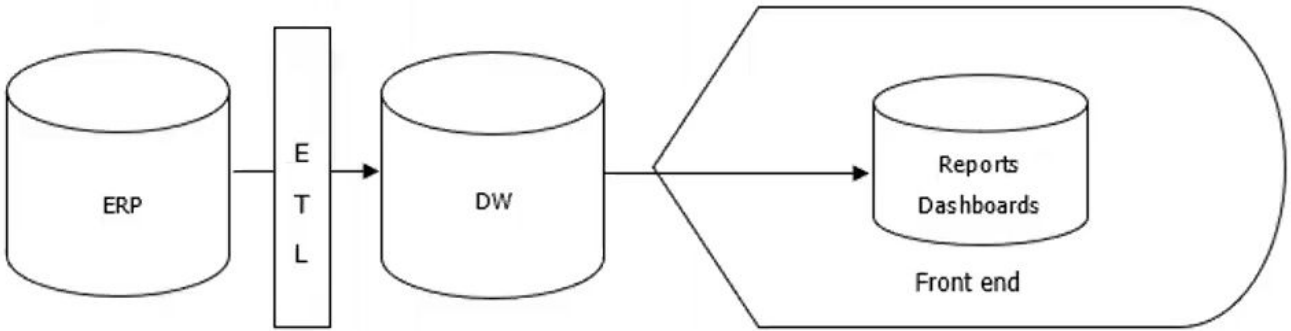
Google Cloud

Instrutor: Rafael Arruda

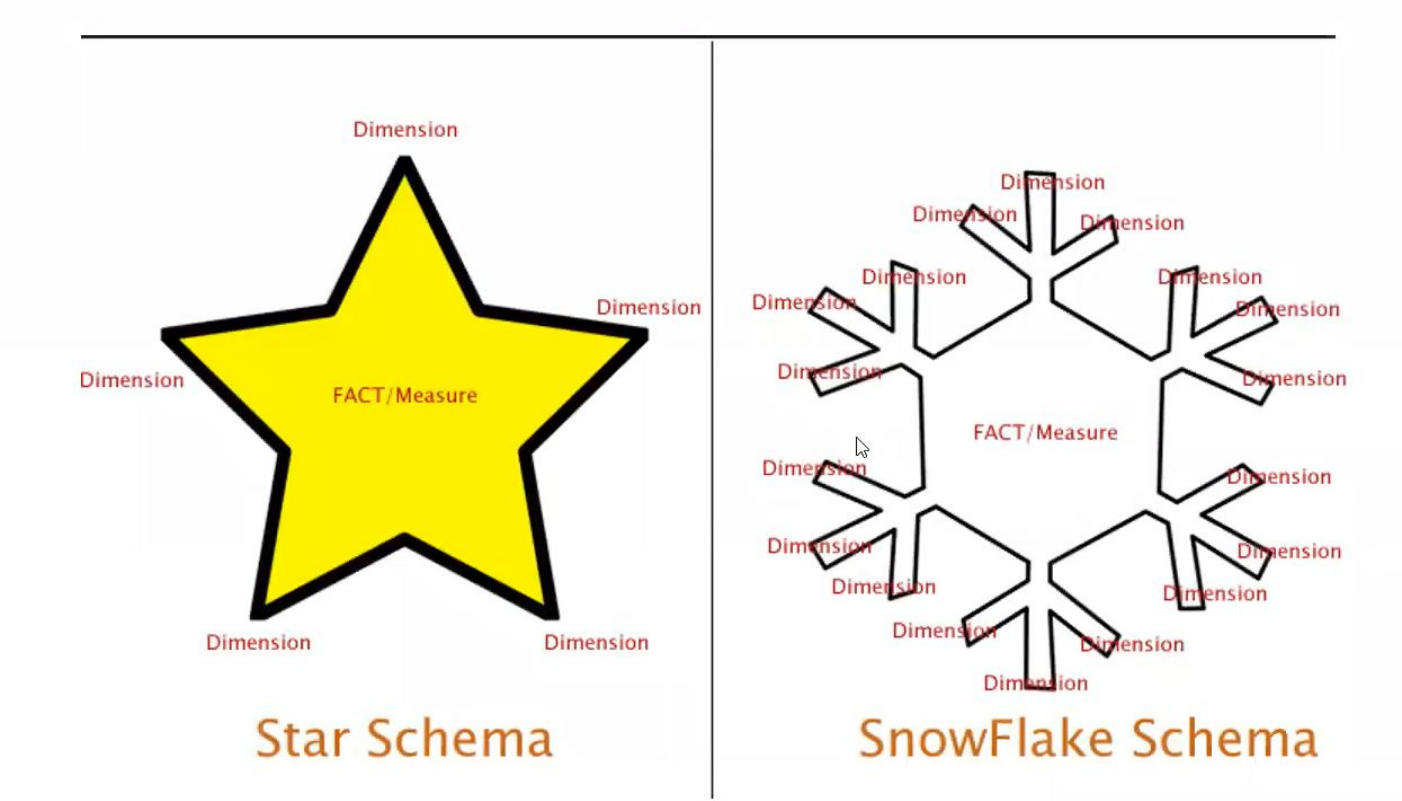
Atualização conteúdo: Ademário Nobre em 14/07/2023

Business Intelligence

- Processo de converter dados brutos em informocações relevantes para os gestores tomarem decisões.
- Arquitetura de um processo de B.I

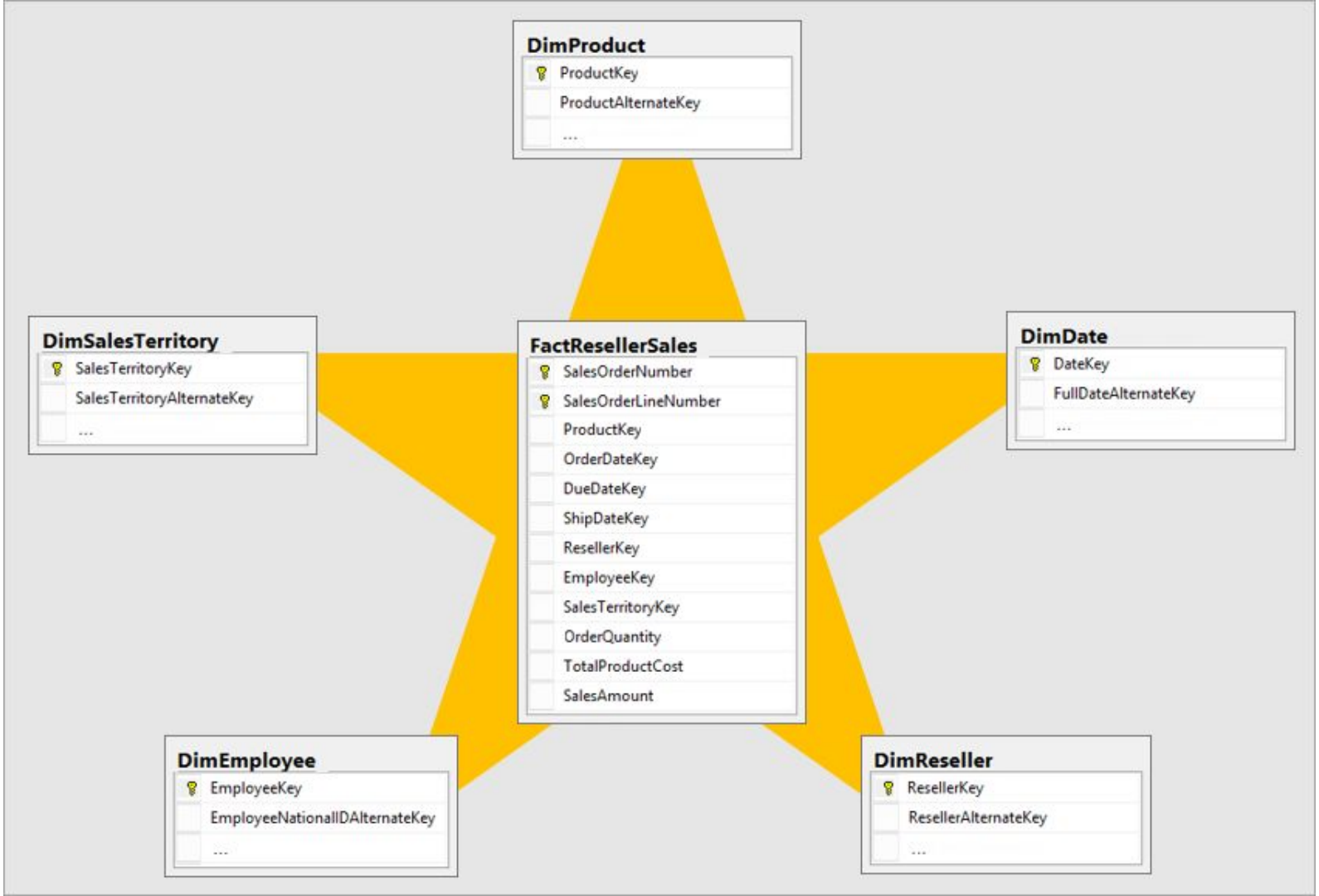


Tipos de topologia

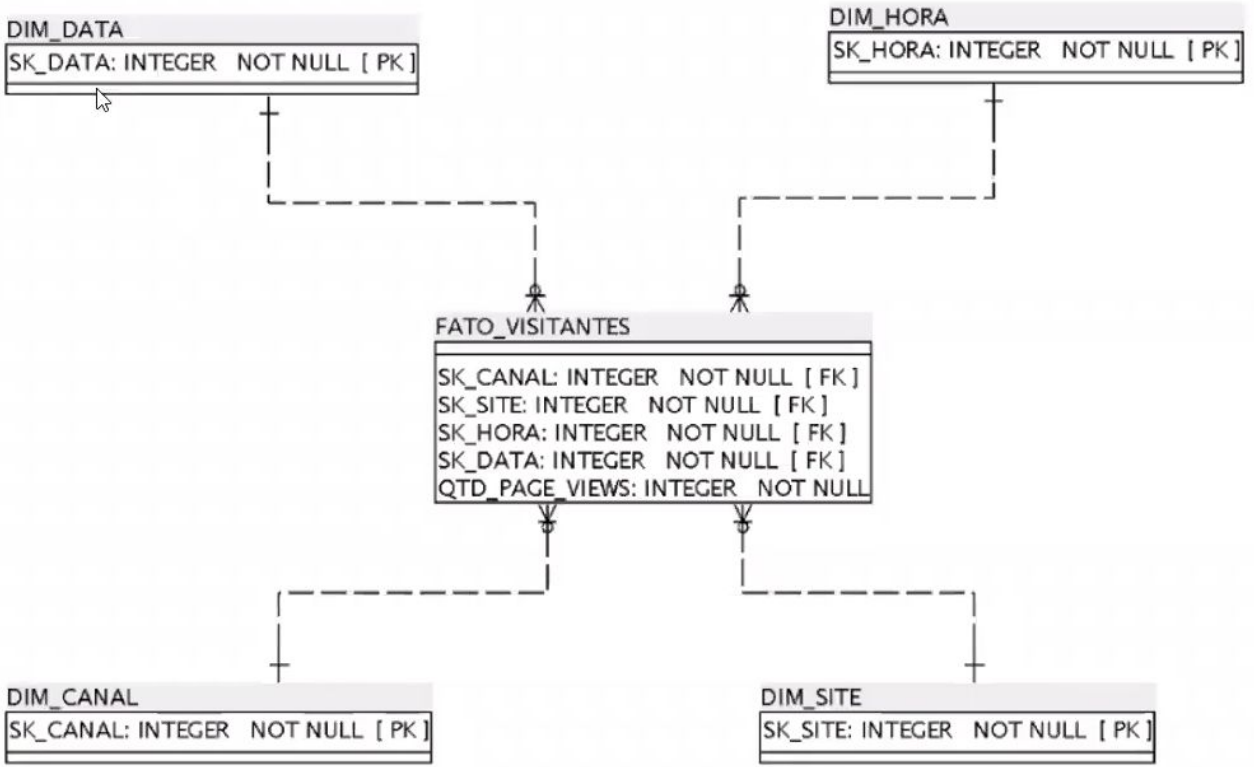


Tabelas

- Dimensões: Uma tabelas que armazena todos os registros descritivos que possuem referência com as informações de uma tabela fato. Contém uma ou mais colunas de chave, que atuam como um identificador exclusivo, e colunas descritiva
- Fatos: Tabelas de fatos armazenam observações ou eventos e podem ser ordens de vendas, saldos de ações, taxas de câmbio, temperaturas, etc. Principal tabela do data warehouse, pois conectam com as dimensões.

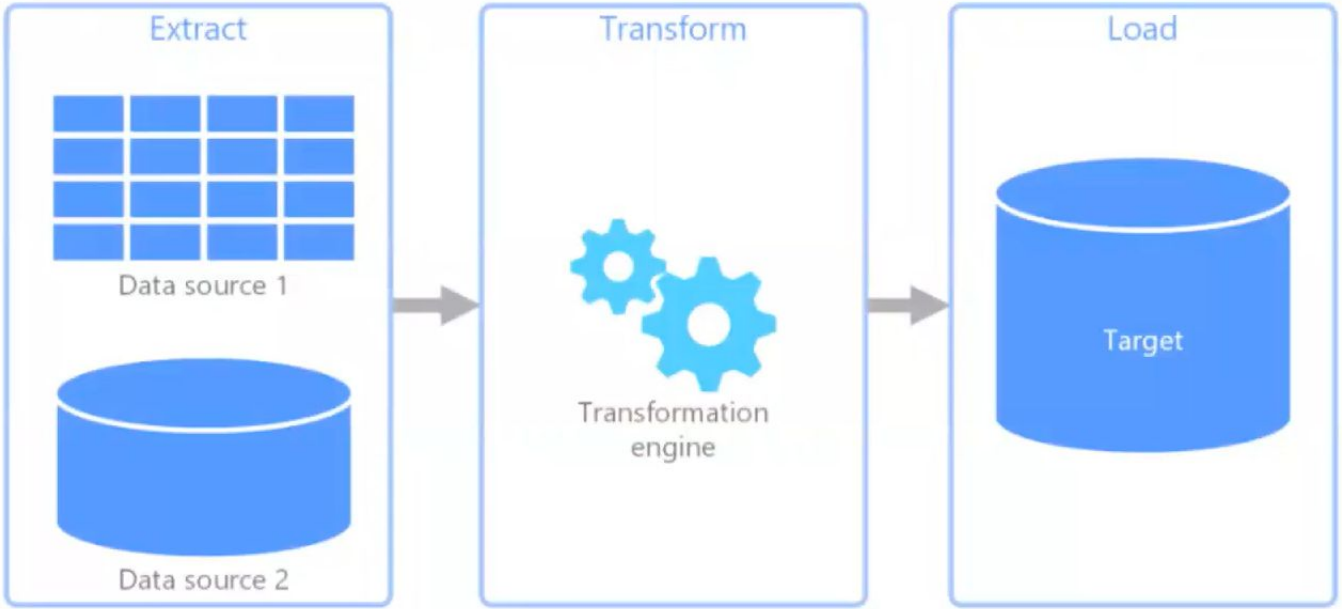

DimProduct
ProductKey
ProductAlternateKey
...
DimSalesTerritory
SalesTerritoryKey
SalesTerritoryAlternateKey
...
FactResellerSales
SalesOrderNumber
SalesOrderLineNumber
ProductKey
OrderDateKey
DueDateKey
ShipDateKey
ResellerKey
EmployeeKey
SalesTerritoryKey
OrderQuantity
TotalProductCost
SalesAmount
DimDate
DateKey
FullDateAlternateKey
...
DimEmployee
EmployeeKey
EmployeeNationalIDAlternateKey
...
DimReseller
ResellerKey
ResellerAlternateKey
...

Exemplo de modelo de DW

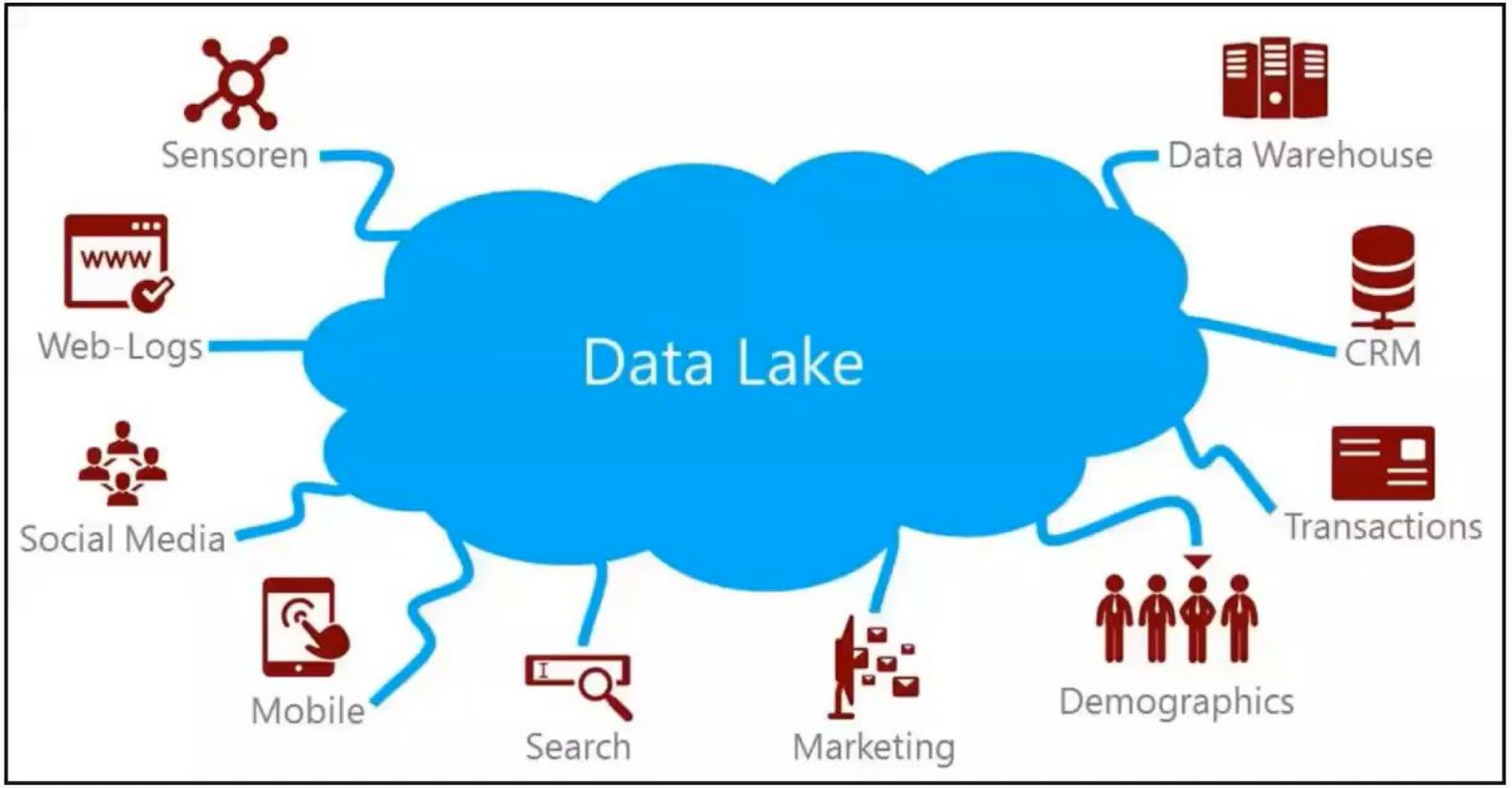


ETL

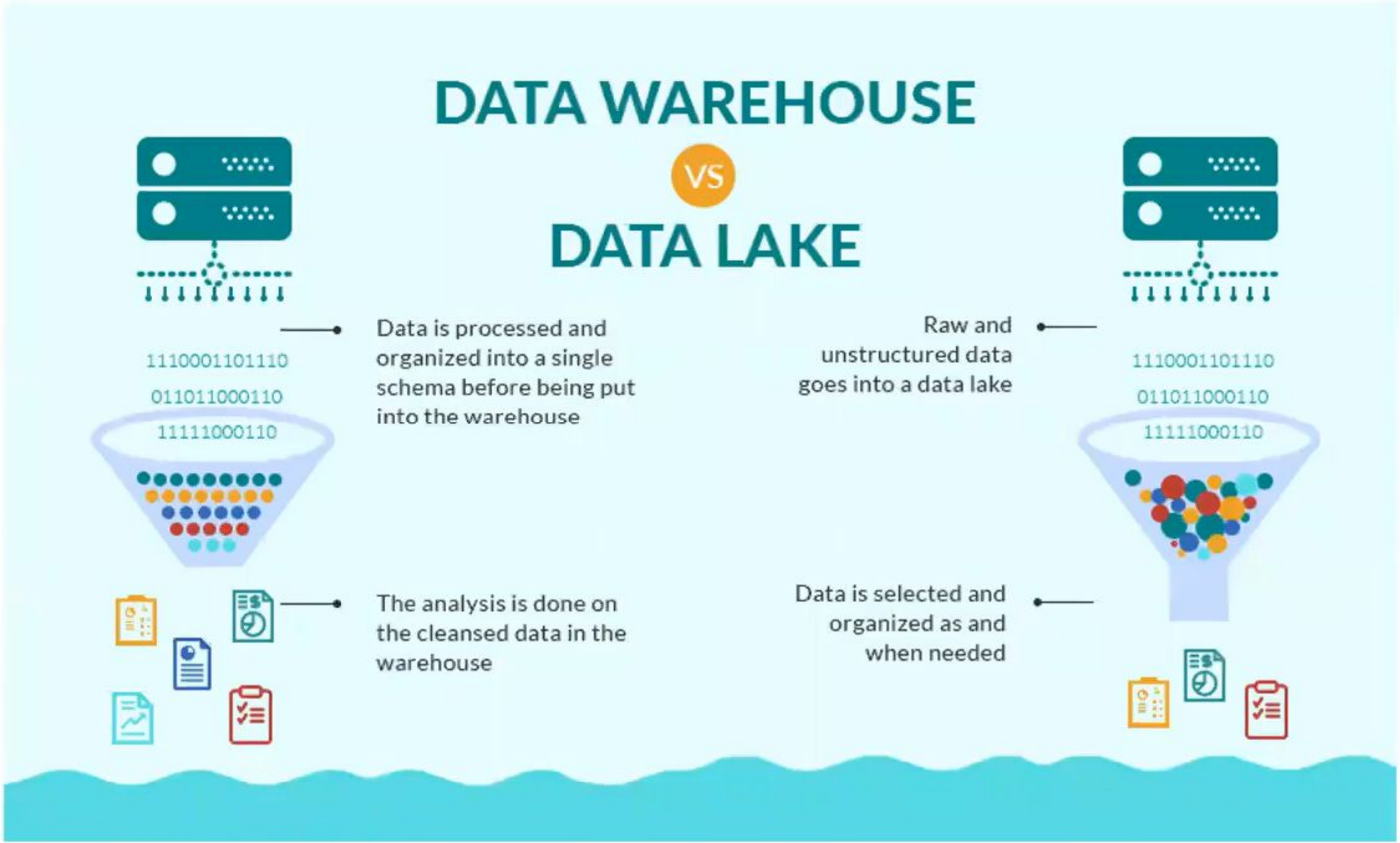
- Extração
- Transformação
- Load(carga)



DATA LAKE


Sensoren
Data Warehouse
www
Web-Logs
Data Lake
CRM
Social Media
Transactions
Mobile
Search
Marketing
Demographics

DATA WAREHOUSE X DATA LAKE



O que é Cloud?

O que é Cloud?

- A computação em nuvem é a entrega de recursos de TI sob demanda por meio da Internet com definição de preço de pagamento conforme o uso. Em vez de comprar, ter e manter datacenters e servidores físicos, você pode acessar serviços de tecnologia, como capacidade computacional, armazenamento e bancos de dados, conforme a necessidade, usando um provedor de nuvem como a GCP, Azure e AWS.

Benefícios da nuvem

- Agilidade
- Escalabilidade
- Redução de custos
- Disponibilidade e backup

Plataforma de dados no GCP

Cloud SQL

- Cloud SQL é uma base de dados relacional;
- Serviço de banco de dados relacional totalmente gerenciado para MySQL, PostgreSQL e SQL Server com coleções avançadas de extensões, além de sinalizações de configuração e ecossistemas de desenvolvedores.

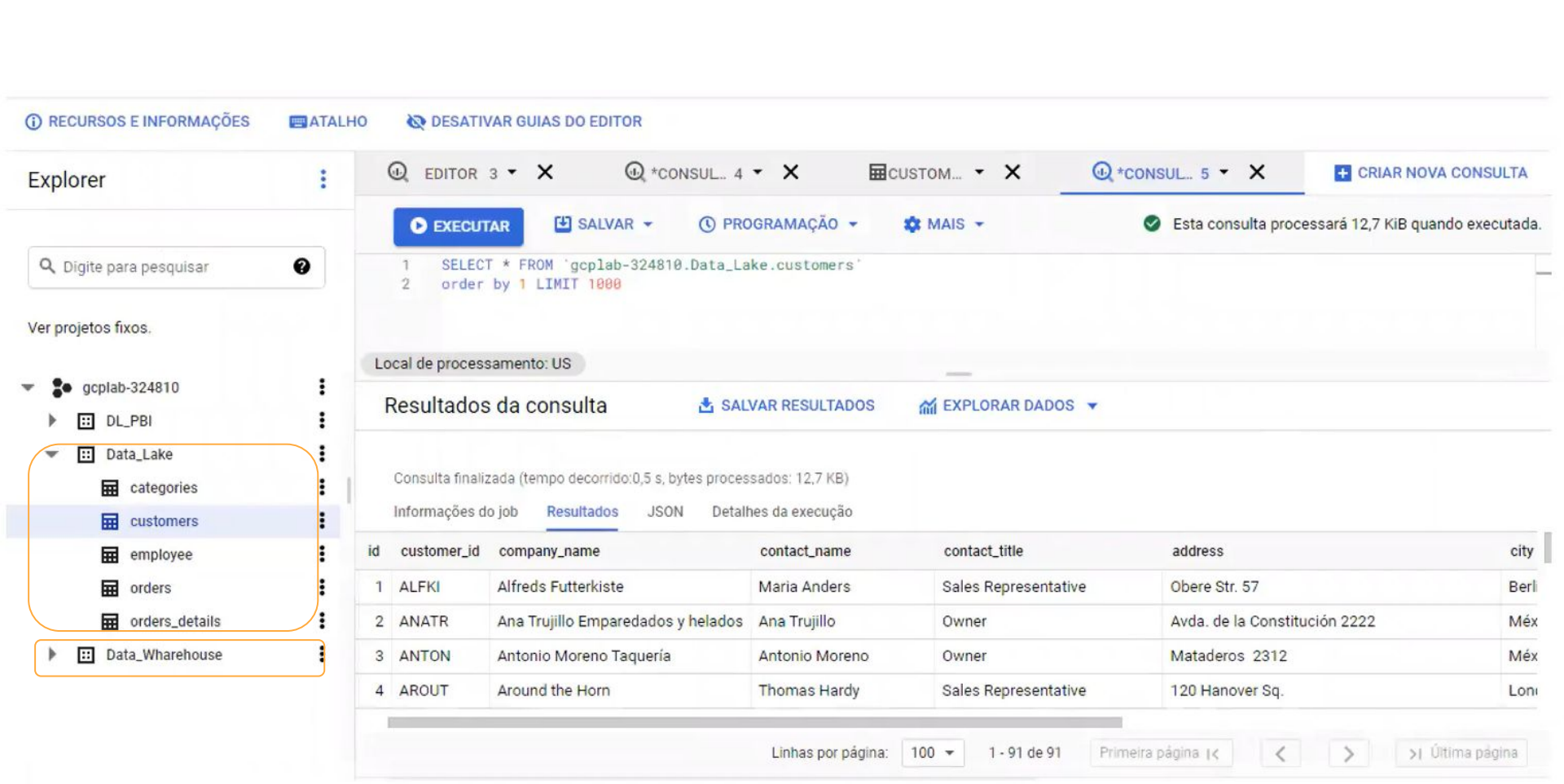
Google Cloud Storage

- Armazenamento do GCP, nosso Data Lake.
- O Cloud Storage é um serviço gerenciado para armazenar dados não estruturados. Armazene qualquer quantidade de dados e recupere-os quantas vezes quiser. Armazenamento de arquivos, também chamados de objetos ou blobs, com um foco em escalabilidade, disponibilidade, segurança e rendimento, com custo extremamente acessível, onde paga-se pelo armazenamento.
- O Cloud Storage é um serviço para o armazenamento de **objetos** no Google Cloud. Um objeto é um dado imutável composto de um arquivo em qualquer formato. Os objetos são armazenados em contêineres chamados de buckets. Todos os buckets estão associados a um projeto, e é possível agrupar os projetos em uma organização. Inclui o armazenamento em blocos, a transferência de dados e o armazenamento de arquivos.

Google Cloud Storage

Big Query

- Big Query é o melhor produto da Google para área de dados;
- É uma base de dados para consultar grandes volumes de dados;
- Pode ser utilizado para consultar dados diretamente em arquivos armazenados no Google.
- Também funciona como Data Warehouse.
- Podemos ler dados do Data Lake e do DW.

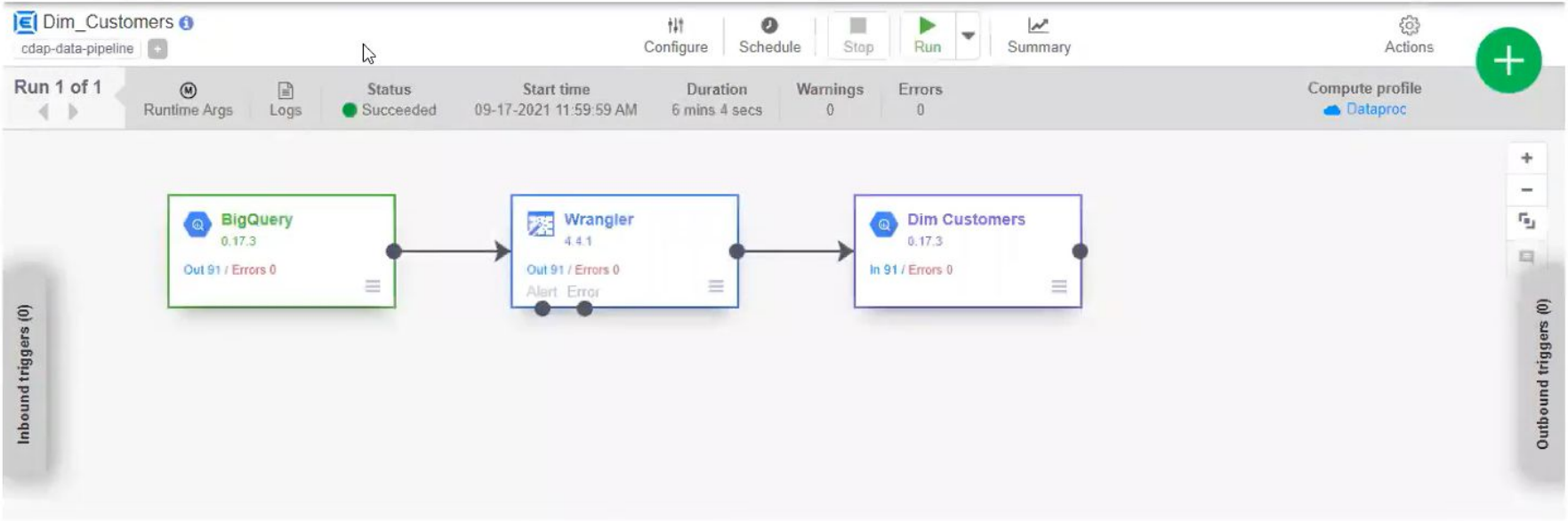

RECURSOS E INFORMAÇÕES
ATALHO
DESATIVAR GUIAS DO EDITOR
Explorer
EDITOR 3
*CONSUL... 4
CUSTOM...
*CONSUL... 5
CRIAR NOVA CONSULTA
EXECUTAR
SALVAR
PROGRAMAÇÃO
MAIS
Esta consulta processará 12,7 KiB quando executada.
Digite para pesquisar
Ver projetos fixos.
1 SELECT * FROM `gcplab-324810.Data_Lake.customers`
2 order by 1 LIMIT 1000
Local de processamento: US
gcplab-324810
DL_PBI
Data_Lake
categories
customers
employee
orders
orders_details
Data_Wharehouse
Resultados da consulta
SALVAR RESULTADOS
EXPLORAR DADOS
Consulta finalizada (tempo decorrido:0,5 s, bytes processados: 12,7 KB)
Informações do job
Resultados
JSON
Detalhes da execução
id customer_id company_name contact_name contact_title address city
1 ALFKI Alfreds Futterkiste Maria Anders Sales Representative Obere Str. 57 Berl
2 ANATR Ana Trujillo Emparedados y helados Ana Trujillo Owner Avda. de la Constitución 2222 Méx
3 ANTON Antonio Moreno Taquería Antonio Moreno Owner Mataderos 2312 Méx
4 AROUT Around the Horn Thomas Hardy Sales Representative 120 Hanover Sq. Lon
Linhas por página: 100
1 - 91 de 91
Primeira página
Última página

Data Fusion

- Solução de ETL da Google;
- Excelente solução para que querem ter um ETL serverless;
- Tem custo maior comparado ao Pentaho;
- Assemelha-se ao Data Factory do Azure.

Data Fusion

- Coleta, tratamento e gravação



https://arrudaconsulting.com.br/pentaho-gcp/