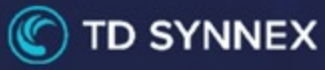
Diamante



Platina

Ouro

Prata

Apoio

Apresentação

DBA BRASIL
DATA & CLOUD

Gilson Martins

Oracle ACE Pro

- 17 anos de experiência na área de TI
- 15 anos focado em administração de banco de dados e tecnologias Oracle
- Consultor Oracle e Instrutor de GoldenGate há 11 anos
- Host no GoldenTalks e Golden Tips YouTube
- Alta disponibilidade com Grid Infrastructure, RAC, Data Guard, GoldenGate
- Migração e replicação de dados
- Certificações Database OCP 10g / 11g / 12c
- Certificações OGG 11g / 12c
- Atualmente → Engenharia de Dados e BigData

GOLDENGATEBR
Consultoria & Treinamentos

Apresentação

DBA BRASIL
DATA & CLOUD

Rafael Milanez

Enterprise Architect

Formado em Big Data com especialização em Engenharia de Dados pela FIAP

Introdução

Bancos Relacionais (Monolítos)

GOLDENGATEBR
Consultoria & Treinamentos

Bancos Relacionais (Monolítos)

GOLDENGATEBR
Consultoria & Treinamentos

Bancos Relacionais (Monolítos)

GOLDENGATEBR
Consultoria & Treinamentos

Data Warehouse

- ✓ Dados Históricos
- ✓ Dados Estruturados
- ✓ Banco de Dados Relacional

Data Lake

GOLDENGATEBR
Consultoria & Treinamentos

- ✓ Dados de várias fontes distintas e formatos diferentes
- ✓ Dados Não-Estruturados
- ✓ Banco de Dados Não-Relacional

Big Data

GOLDENGATEBR
Consultoria & Treinamentos

Os três Vs

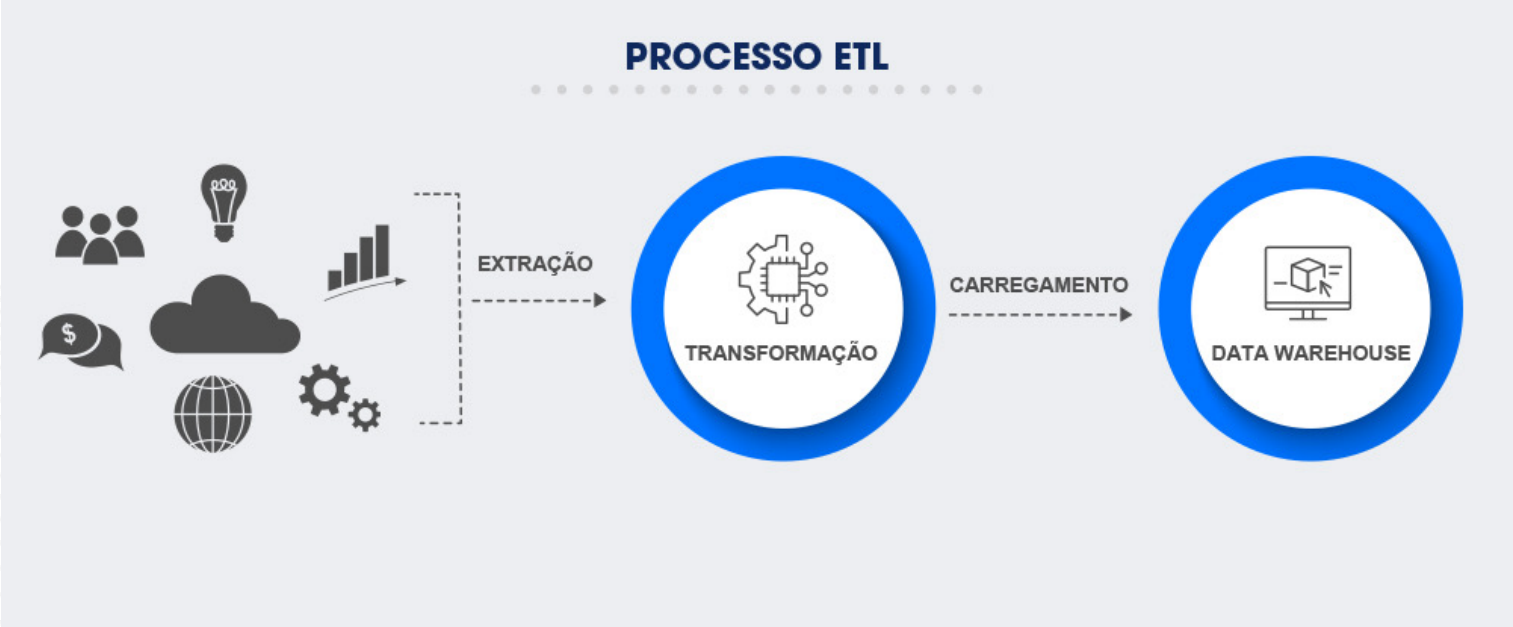
- Variedade
- Volume
- Velocidade

- Valor
- Veracidade
- Visualização

- ✓ Json
- ✓ Avro
- ✓ Parquet
- ✓ XML
- ✓ TXT
- ✓ Etc...

ETL vs REAL-TIME (Streaming Data)

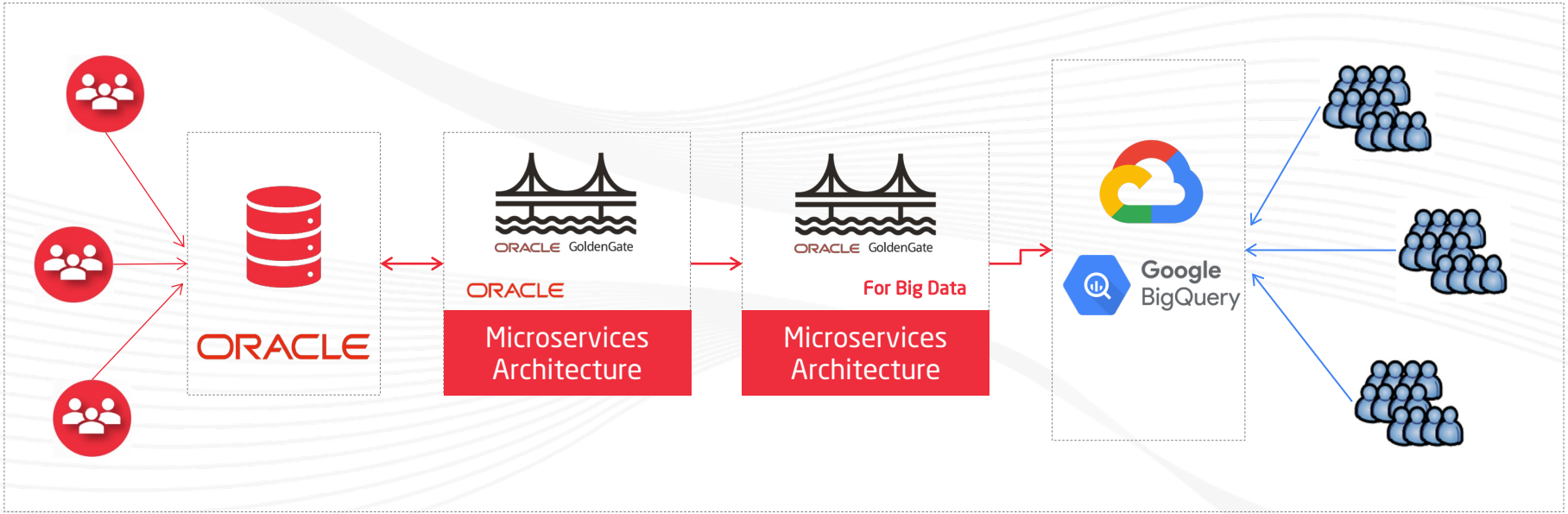
GOLDENGATEBR
Consultoria & Treinamentos



- ✓ Processamento Batch

Engenheiro de Dados - Streaming Data

GOLDENGATEBR
Consultoria & Treinamentos



- ✓ Criação e sustentação de Pipelines
- ✓ Melhores ferramentas e Tecnologias
- ✓ Ingestão de Dados Real-Time

O que é BigQuery?

O que é BigQuery?

GOLDENGATEBR
Consultoria & Treinamentos

1 Datawarehouse Corporativo

- O BigQuery é um serviço de Datawarehouse Corporativo, diponivel na Google Cloud.

2 Totalmente Gerenciado

- Não se preocupe em gerenciar ou dimensionamento infraestrutura, a Google faz isso para você.

3 Serverless

- Baseado em uma arquitetura que não demanda alocação de servidores.

4 Escala de acordo com sua demanda

- Permite dimensionar seus dados na escala de exabytes.

5 Utiliza Criptografia por padrão

- Todos os dados do BigQuery são compactados e criptografados automaticamente por padrão.

6 Suporta Linguagem SQL

- Suporta linguagem SQL compativel com ANSI:2011, ou seja, é o padrão SQL a partir de 2011 para frente.

7 Alta Disbonibilidade

- SLA de 99.99%

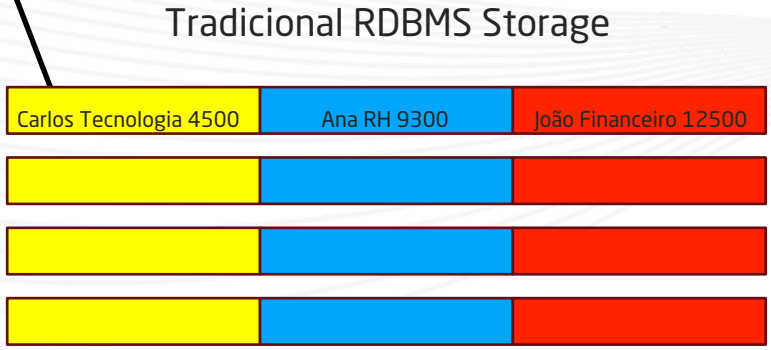
8 Recursos Integrados Disponíveis

- Recursos integrados, como Aprendizado de Máquina, Análise Geoespacial, compatível com streaming etc...

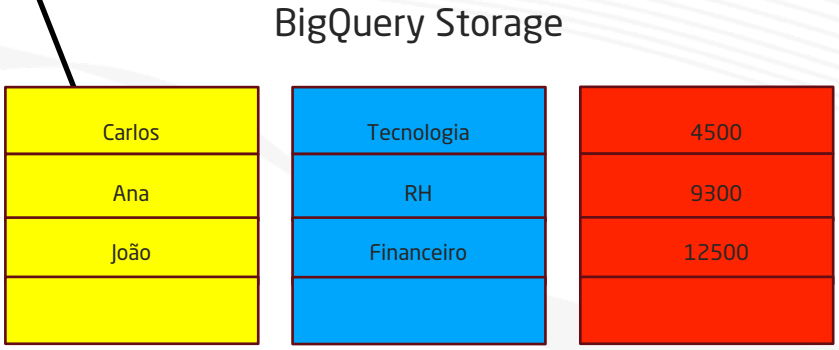
Por que o BigQuery é rápido?

GOLDENGATEBR
Consultoria & Treinamentos

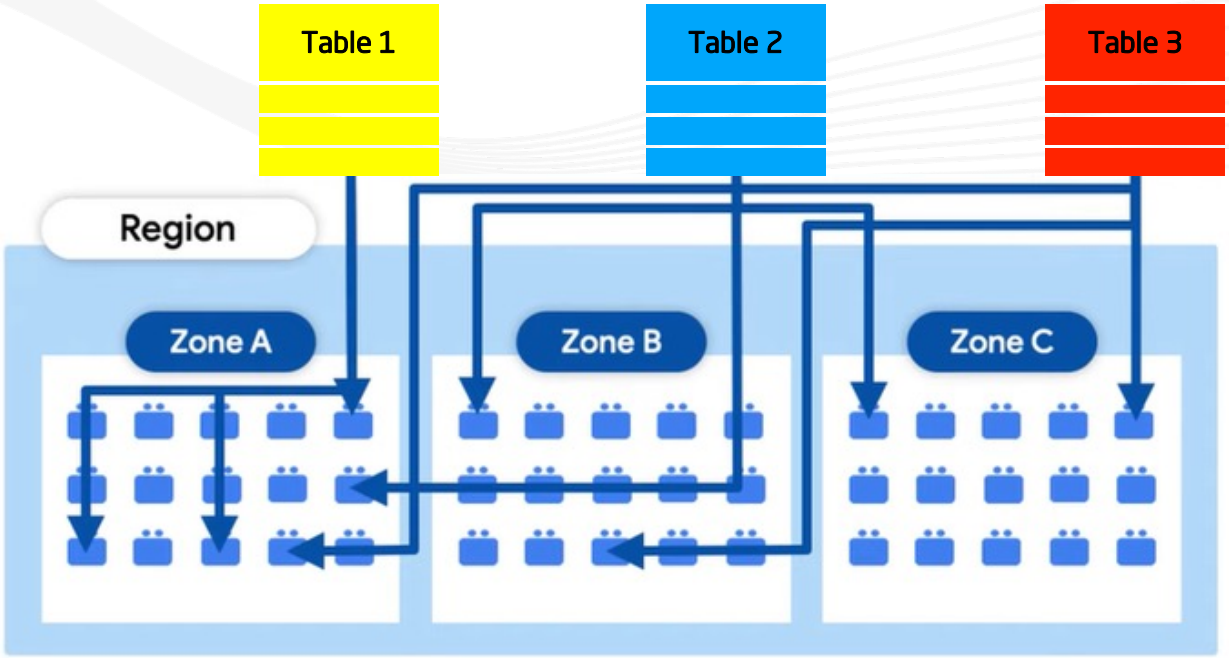
- Ótimo para sistemas transacionais e casos de uso OLTP.
- Compressão de dados: Uma linha pode ter diferentes tipos de dados, algoritmo de compactação menos eficiente.



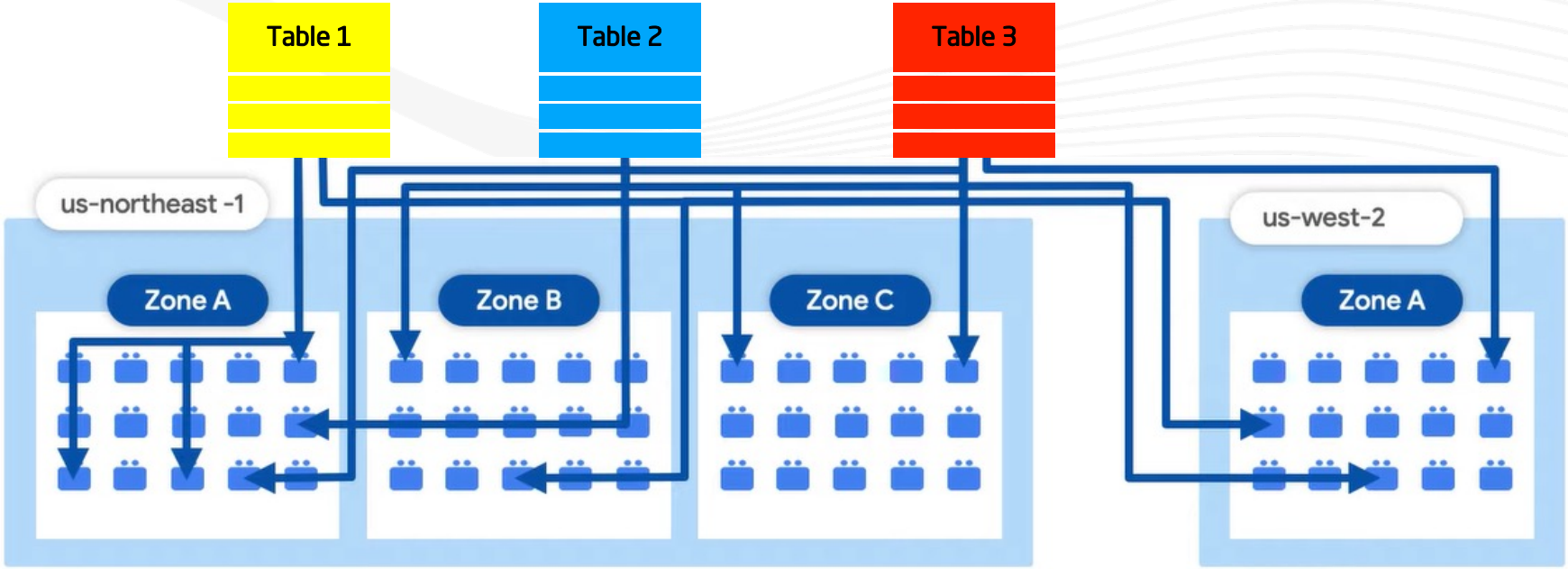
- As tabelas são armazenadas em formato colunar otimizado.
- Maior compressão, dados iguais são armazenados juntos.
- Excelente para operações de agregações e agrupamentos.



Armazenamento



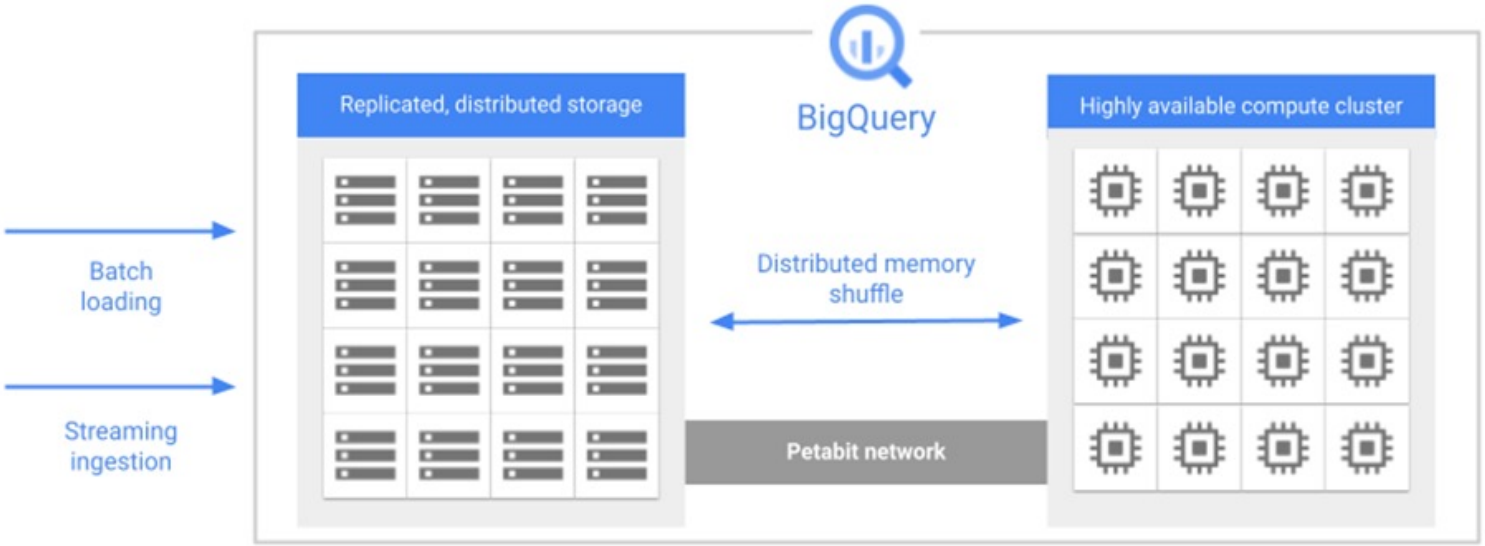
Armazenamento



Arquitetura

Arquitetura

GOLDENGATEBR
Consultoria & Treinamentos



Arquitetura

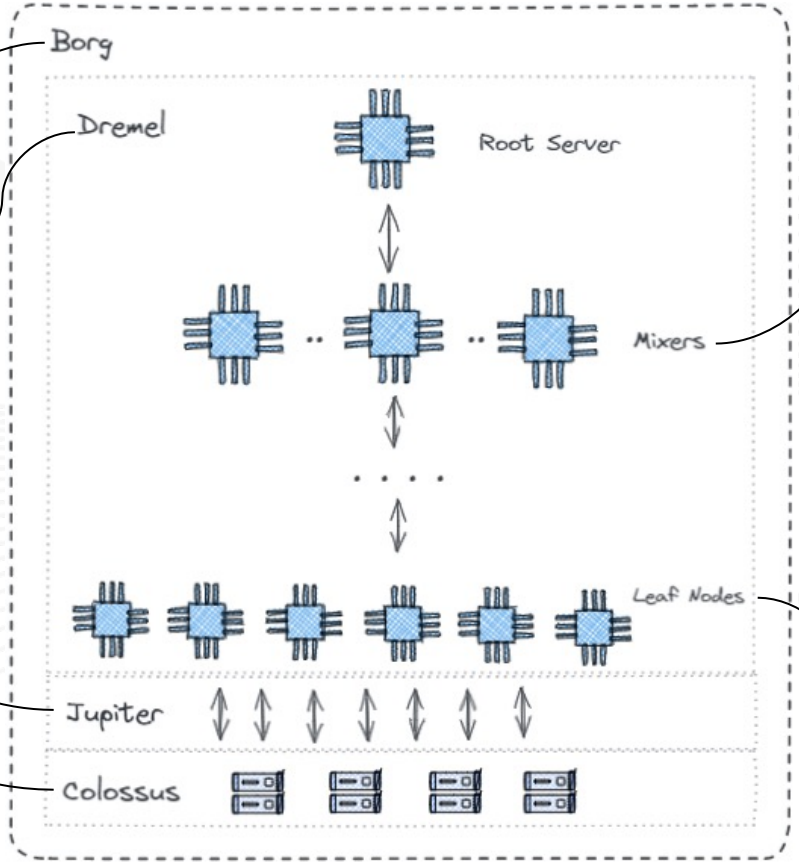
GOLDENGATEBR
Consultoria & Treinamentos

Plataforma que deu origem aos contenendores. Sistema de gerenciamento de cluster em larga escala do Google

Transforma a consulta SQL em uma árvore de execução

Rede Petabit

Sistema de arquivos distribuídos do Google **(columnar storage)**



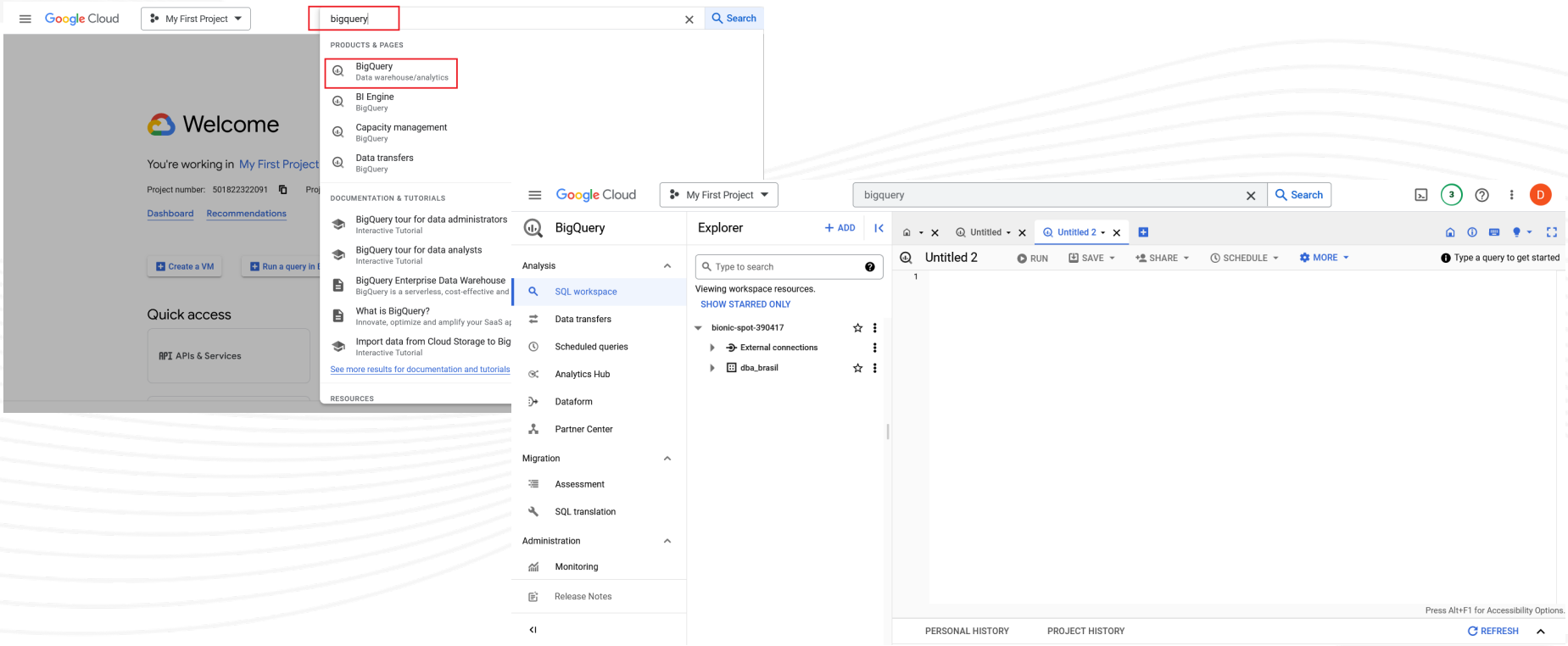
Após a leitura dos dados pelos nós folha, temos os galhos, também conhecidos como *mixers*, onde é realizada a agregação dos dados que são retornados para o nível acima e para o usuário.

São responsáveis pela leitura massiva e processamento dos dados que estão armazenados no *FileSystem* do Google

Interface com o BigQuery

UI da Web

GOLDENGATEBR
Consultoria & Treinamentos



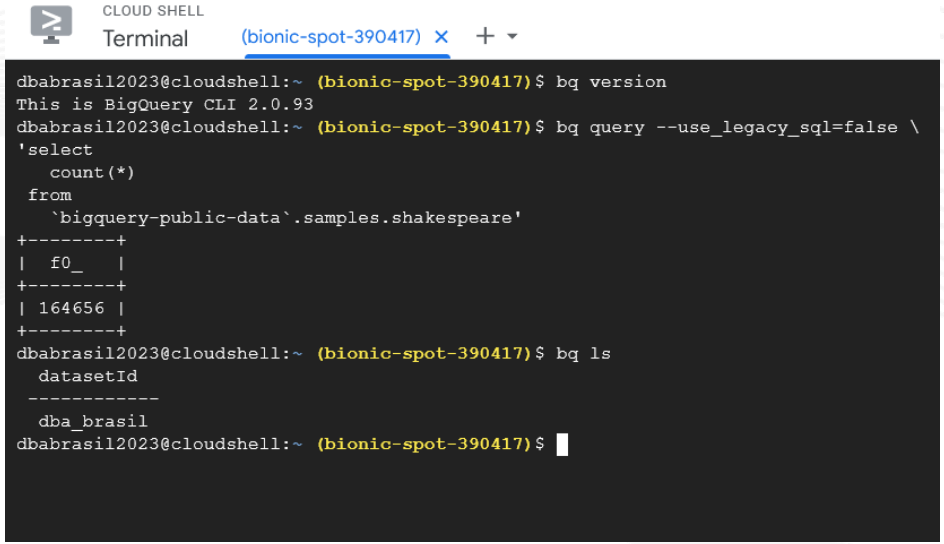
BQ Command Line

GOLDENGATEBR
Consultoria & Treinamentos

- Ferramenta de linha de comando baseada em Python do BigQuery.
- Cloud Shell ou Google SDK (gcloud) para interagir por meio de um terminal

Diversas operações:

- **bq ls** – listar objetos.
- **bq load** – carregar dados para tabelas.
- **bq mk** – criar datasets, tabelas, views etc...
- **bq rm** – apagar datasets, tabelas, views etc...
- **bq show** – exibir detalhes de uma tabela.
- **bq help** – listar opções.

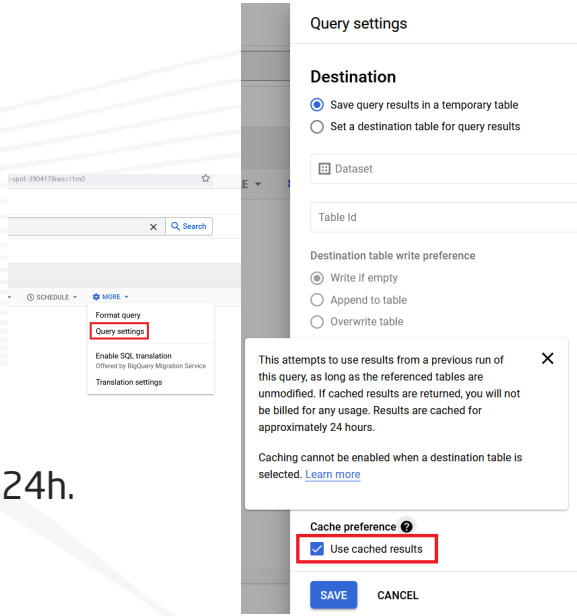


Recursos do BigQuery

SQL Cache

GOLDENGATEBR
Consultoria & Treinamentos

- Resultados das queries são armazenados no Cache.
- Execuções posteriores da mesma query acessam o Cache:
  - ✓ Melhor performance
  - ✓ Não é cobrado (sem custo)



Restrições:

- Query precisa ser uma réplica exata da query original.
- Cache é armazenado em tabelas temporárias por aproximadamente 24h.
- Consultas Federadas não utilizam recurso de Cache.
- Tabelas utilizadas em ingestão de dados por streaming.
- Consulta Curinga (Prefixo/Sufixo)
- Caso a consulta utilize funções não determinísticas: **`CURRENT_TIMESTAMP()`**, **`CURRENT_DATE`**, **`CURRENT_USER()`**, **`SESSION_USER`**() , retornam valores diferentes dependendo de quando uma consulta é executada, não considera Cache.

Consultas Federadas

GOLDENGATEBR
Consultoria & Treinamentos

Dados que estão armazenados fora do Bigquery:

- Cloud SQL
- Cloud Spanner
- BigTable
- Cloud Storage
- Google Drive
- BigQuery Omni

Formatos de arquivos compatíveis:

- JSON (apenas delimitado por nova linha)
- CSV
- Avro
- Parquet
- Iceberg
- ORC

- Use a função **EXTERNAL_QUERY** para consultar o banco de dados externo.
- A performance é **degradada** em relação a uma consulta nativa.
- Não utiliza recurso de **Cache.**
- **Regiões devem ser compatíveis,** uma região única do BigQuery só pode consultar um recurso na mesma região.
- Uma consulta executada na multi-region EUA do BigQuery pode consultar qualquer região única na região geográfica dos EUA, como **us-central1,**
- **us-east4** ou **us-west2.**

```
SELECT c.customer_id, c.name FROM
mydataset.customers AS c LEFT OUTER JOIN
EXTERNAL_QUERY('us.connection_id',
  '''SELECT customer_id, MIN(order_date) AS
first_order_date FROM orders
  GROUP BY .....................
```

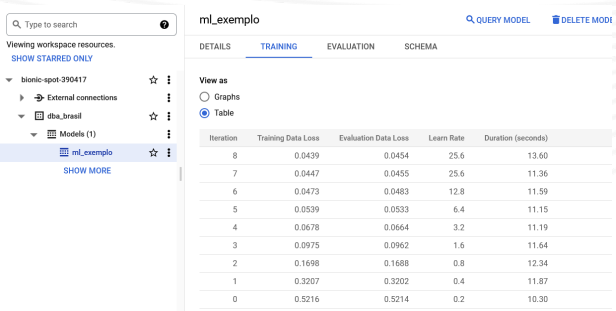
BigQuery ML

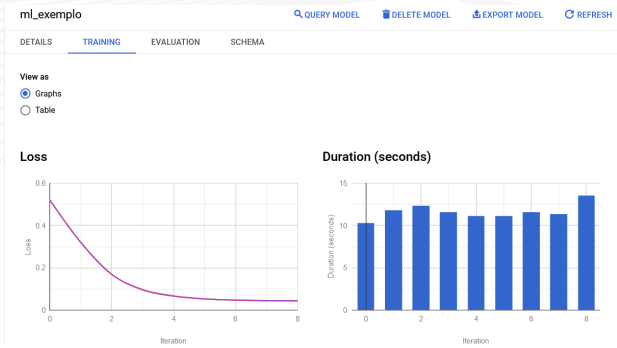
GOLDENGATEBR
Consultoria & Treinamentos

Existem muitos tipos de modelos integrados ao BQML:

- Regressão Linear
- Regressão Logística
- Serie Temporal
- K-means
- Árvores aumentadas
- PCA
- Fatoração de array
- Suporta importação de modelos do TensorFlow

```
CREATE MODEL `dba_brasil.ml_exemplo`
OPTIONS(model_type='logistic_reg') AS
SELECT
  IF(totals.transactions IS NULL, 0, 1) AS label,
  IFNULL(device.operatingSystem, "") AS os,
  device.isMobile AS is_mobile,
  IFNULL(geoNetwork.country, "") AS country,
  IFNULL(totals.pageviews, 0) AS pageviews
FROM
  `bigquery-public-
data.google_analytics_sample.ga_sessions_*`
WHERE
  _TABLE_SUFFIX BETWEEN '20160801' AND '20170630'
```

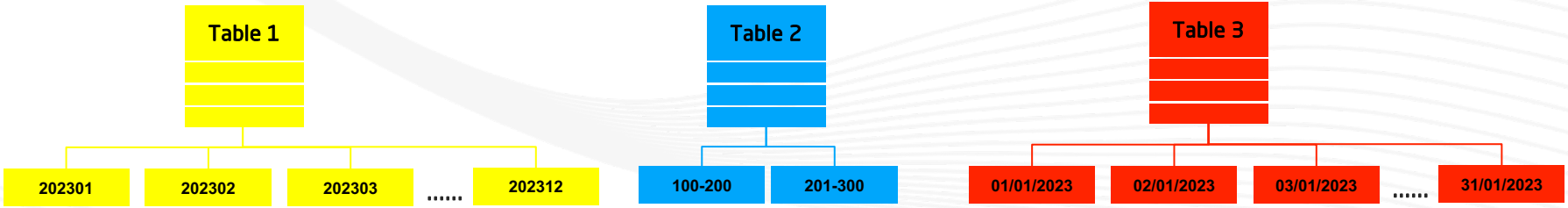




Particionamento

GOLDENGATEBR
Consultoria & Treinamentos

A tabela é dividida em segmentos/partições baseada em algum critério.



Tabelas podem ser particionadas por:

- Unidade de Tempo (Date - Timestamp - Datetime)
- Tempo de Ingestão (Dia, Hora)
- Integer (Range)

Observações:

- Melhora o desempenho das consultas.
- A chave do particionamento deve ser bem definida.
- Defina prazo de validade para as partições.
- Redução de custo. (FULL COLUMN SCAN).

```
CREATE TABLE
  mydataset.newtable (transaction_id INT64,
transaction_date DATE)
PARTITION BY transaction_date
AS (
  SELECT
    transaction_id, transaction_date
  FROM
    mydataset.mytable
);
```

```
bq mk \
   -t \
   --schema
'ts:TIMESTAMP,qtr:STRING,sales:FLOAT' \
   --time_partitioning_field ts \
   --time_partitioning_type HOUR \
   --time_partitioning_expiration 259200
\
   mydataset.mytable
```

Consulta curinga

GOLDENGATEBR
Consultoria & Treinamentos

Você pode especificar um intervalo de sufixos:

```
SELECT coluna from dataset.tabela
WHERE
_TABLE_SUFFIX BETWEEN '2001' AND '2003'
```

```
SELECT coluna from dataset.tabela
WHERE
_TABLE_SUFFIX = '2001'
```

```
SELECT coluna from dataset.tabela
WHERE
_TABLE_SUFFIX < '2001'
```

- As tabelas precisam ter nomes parecidos, como PrefixoSufixo
- Exemplo: PART é o prefixo, o ano é sufixo

Time Travel Query

GOLDENGATEBR
Consultoria & Treinamentos

- Permite execução de queries baseadas em um snaphot de até 7 dias no passado.
- Ideal para utilizar em recuperação de dados alterados ou removidos acidentalmente
- É necessário adicionar a cláusula FOR SYSTEM_TIME AS OF

```
SELECT *
FROM `<projeto>.<dataset>.<tabela>`
FOR SYSTEM_TIME AS OF
TIMESTAMP_SUB(CURRENT_TIMESTAMP, INTERVAL 2
HOURS)
```

```
create table
`<projeto>.<dataset>.<tabela_restaurada>`
as
SELECT *
FROM `<projeto>.<dataset>.<tabela>`
FOR SYSTEM_TIME AS OF
TIMESTAMP_SUB(CURRENT_TIMESTAMP, INTERVAL 2
HOURS)
```

Melhores Práticas no Bigquery

GOLDENGATEBR
Consultoria & Treinamentos

- Evite SELECT* , utilize somente as colunas necessárias nas queries.
- Utilize a opção PREVIEW para visualização dos dados (é grátis).
- Confira o quanto sua consulta será cobrada.
- Evite tabelas externas.
- Utilize Partições e/ou Clustering sempre que aplicável.
- Clausula LIMIT é sempre bem-vinda.

Preços do BigQuery

GOLDENGATEBR
Consultoria & Treinamentos

Processamento de consultas:

- **Preço sob demanda:** cobrado pelo número de bytes processados por cada consulta efetuada.
- **Preço Fixo:** Você compra slots, que são CPU's virtuais - uma capacidade de processamento dedicada que você pode utilizar para executar e processar suas consultas

Armazenamento:

- Preço sob dados armazenados no BigQuery

Operação	Preços	Detalhes
Consultas (sob demanda)		O primeiro terabyte (1 TB) por mês é gratuito.

Detalhes do preço

O preço de armazenamento tem como base a quantidade de dados armazenados nas suas tabelas quando eles são descompactados. O tamanho dos dados é calculado com base nos tipos de dados de cada coluna. Para uma explicação detalhada sobre como o tamanho dos dados é calculado, consulte Cálculo do tamanho dos dados.

Os preços de armazenamento são rateados por MB, por segundo. Por exemplo, ao armazenar:

- 100 MB pela metade de um mês, você paga US$ 0,001 (um décimo de centavo);
- 500 GB pela metade de um mês, você paga US$ 5;
- 1 TB por um mês completo, você paga US$ 20.

Preços de ingestão de dados

O BigQuery oferece dois modos de ingestão de dados:

- **Carregamento em lote**: carrega os dados de origem em uma ou mais tabelas do BigQuery em uma única operação.
- **Streaming**: faz streaming de dados de um registro de cada vez ou em lotes pequenos.

Para mais informações sobre qual modo escolher, consulte Introdução ao carregamento de dados.

Veja os seguintes dados: (), Mensal

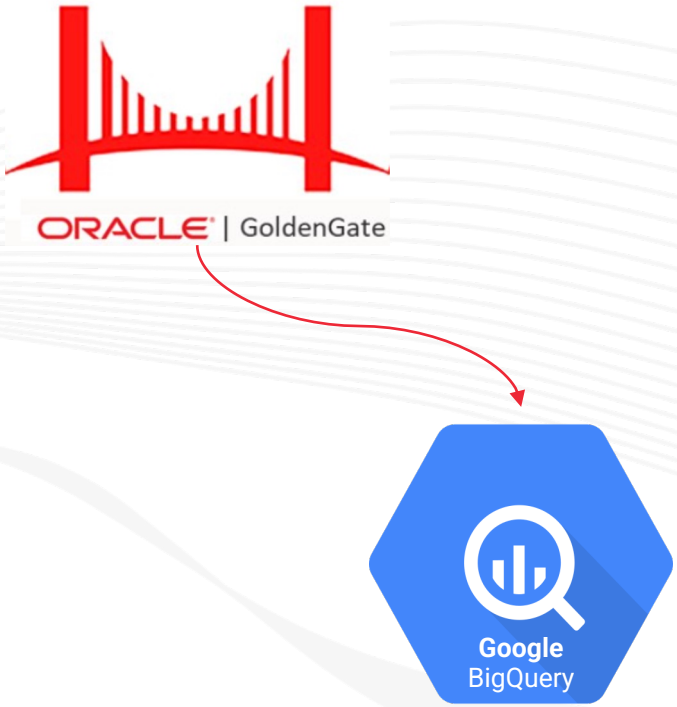
Operação	Preços	Detalhes
Carregamento em lote	Grátis usando o pool de slots compartilhado.	Os clientes podem escolher preços fixos para a capacidade garantida. Quando os dados são carregados no BigQuery, a cobrança é feita.
Inserções por streaming (`tabledata.insertAll`)		A cobrança é realizada conforme o número de linhas inseridas com sucesso. Cada linha é calculada usando um tamanho mínimo de 1 KB.
API BigQuery Storage Write		Os primeiros 2 TB por mês são gratuitos.

Ingestão de dados com

Oracle GoldenGate

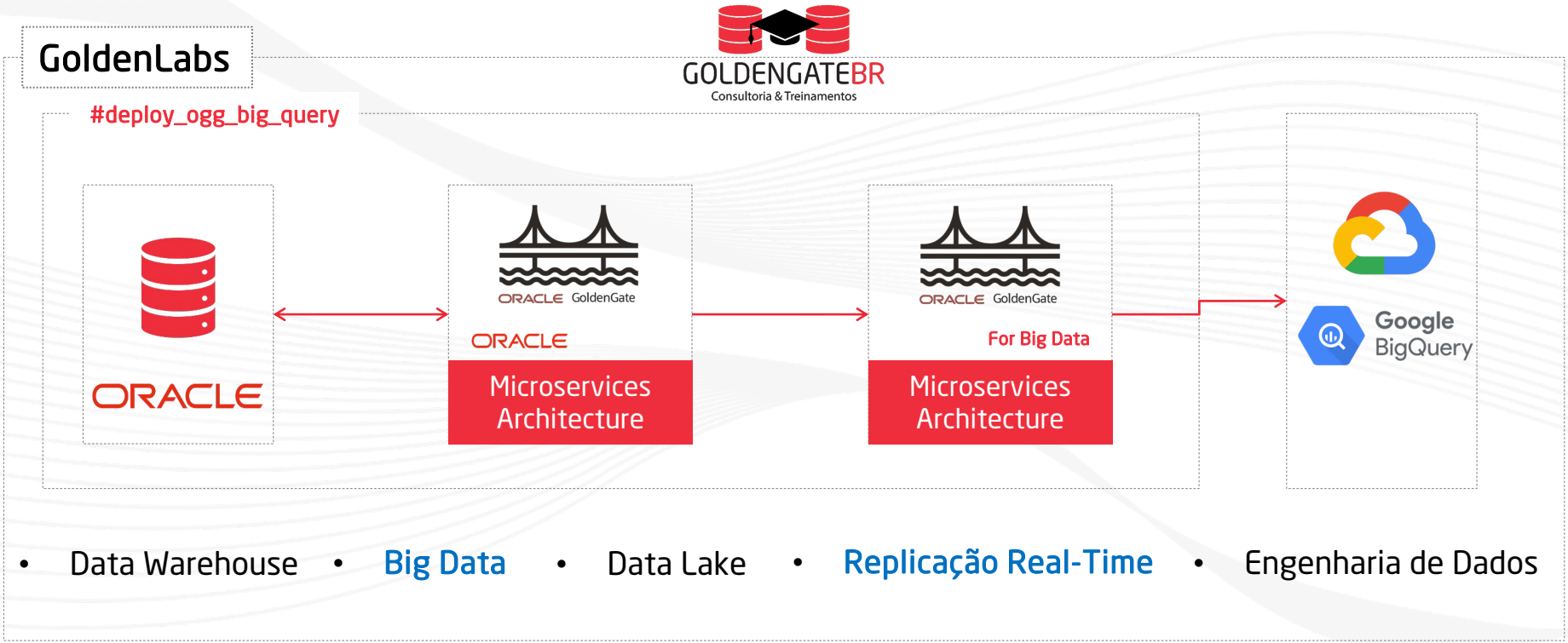
+

Google BigQuery



Ambiente de Replicação

GOLDENGATEBR
Consultoria & Treinamentos



- Data Warehouse
- Big Data
- Data Lake
- Replicação Real-Time
- Engenharia de Dados

Perguntas?

DBA BRASIL
DATA & CLOUD

Gilson Martins

Oracle ACE Pro

Rafael Milanez

Obrigado!