

Data Warehouse e Modelagem Dimensional de Dados em Tempos Modernos de Nuvem

Lourenço Taborda

Universidade de Fortaleza
09.11.2020 19h30

Data Warehouse e Modelagem Dimensional de Dados em Tempos Modernos de Nuvem

Lourenço Taborda

Universidade de Fortaleza
09.11.2020 19h30



Este trabalho está licenciado sob uma Licença Creative Commons Atribuição-Compartilhamento 4.0 Internacional. Para ver uma cópia desta licença, visite <http://creativecommons.org/licenses/by-sa/4.0/>.



LOURENÇO TABORDA



Certified Business
Intelligence Professional



Inovação com dados em nuvem

TRILHA

#TheDevConf
Oracle





Foto de Edgar Chaparro no Unsplash



Foto de Oscar Nord no Unsplash



Foto de Sebastian Bjune no Unsplash



EXPERIÊNCIA

Foto de Edgar Chaparro no Unsplash



ÚNICO

Foto de Oscar Nord no Unsplash



NÃO-RIVAL

Foto de Sebastian Bjune no Unsplash

Por que construímos Data Warehouses, Data Lakes e Lakehouses?



PORQUE DADOS
GERAM VANTAGEM
COMPETITIVA PARA
OS NEGÓCIOS.



DADOS SÃO FATORES
DE PRODUÇÃO DE
BENS E SERVIÇOS
DIGITAIS.



DADOS SÃO UM
BEM DE EXPERIÊNCIA,
INFUNGÍVEL E NÃO-
RIVAL

Por que construímos Data Warehouses, Data Lakes e Lakehouses?

“Information Systems Strategy Triangle é um framework simples para compreender o impacto dos sistemas de informação nas organizações.

Nas empresas de sucesso, a Estratégia de Negócio direciona a Estratégia Organizacional e a Estratégia da Informação.

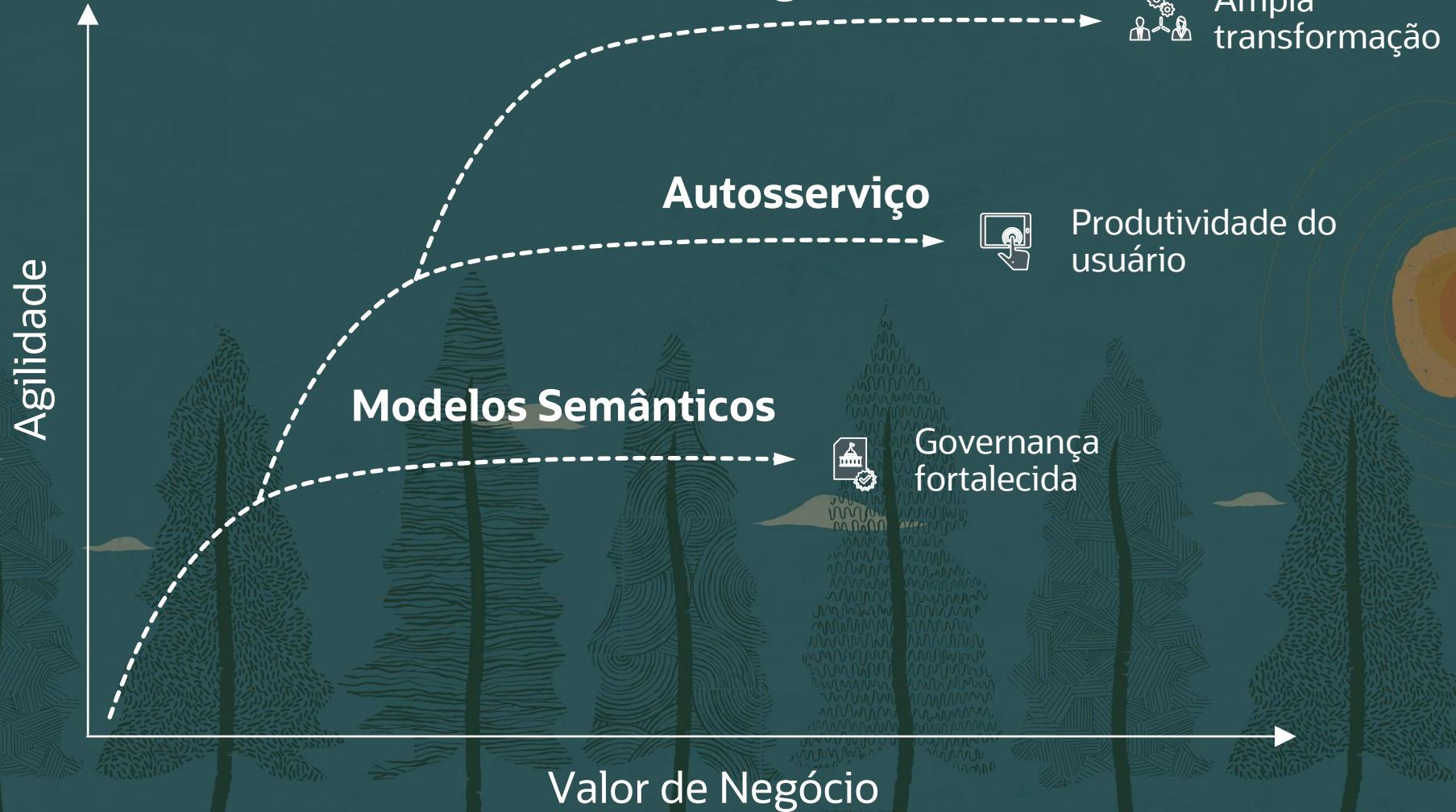
Mudanças em qualquer estratégia requer mudanças nas demais para manter a harmonia.”



Information Systems Strategy Triangle
Referência:

Keri Pearson & Carol Saunders, Managing and Using Information Systems A Strategic Approach (Hoboken: John Wiley & Sons, 2010)

Estratégia de Dados





O recipiente mágico entrega:

Escalabilidade

Desempenho

Transação ACID | Atomicidade | Consistência | Isolamento | Durabilidade

Formatos diversos | Estruturado | Semiestruturado | Não estruturado

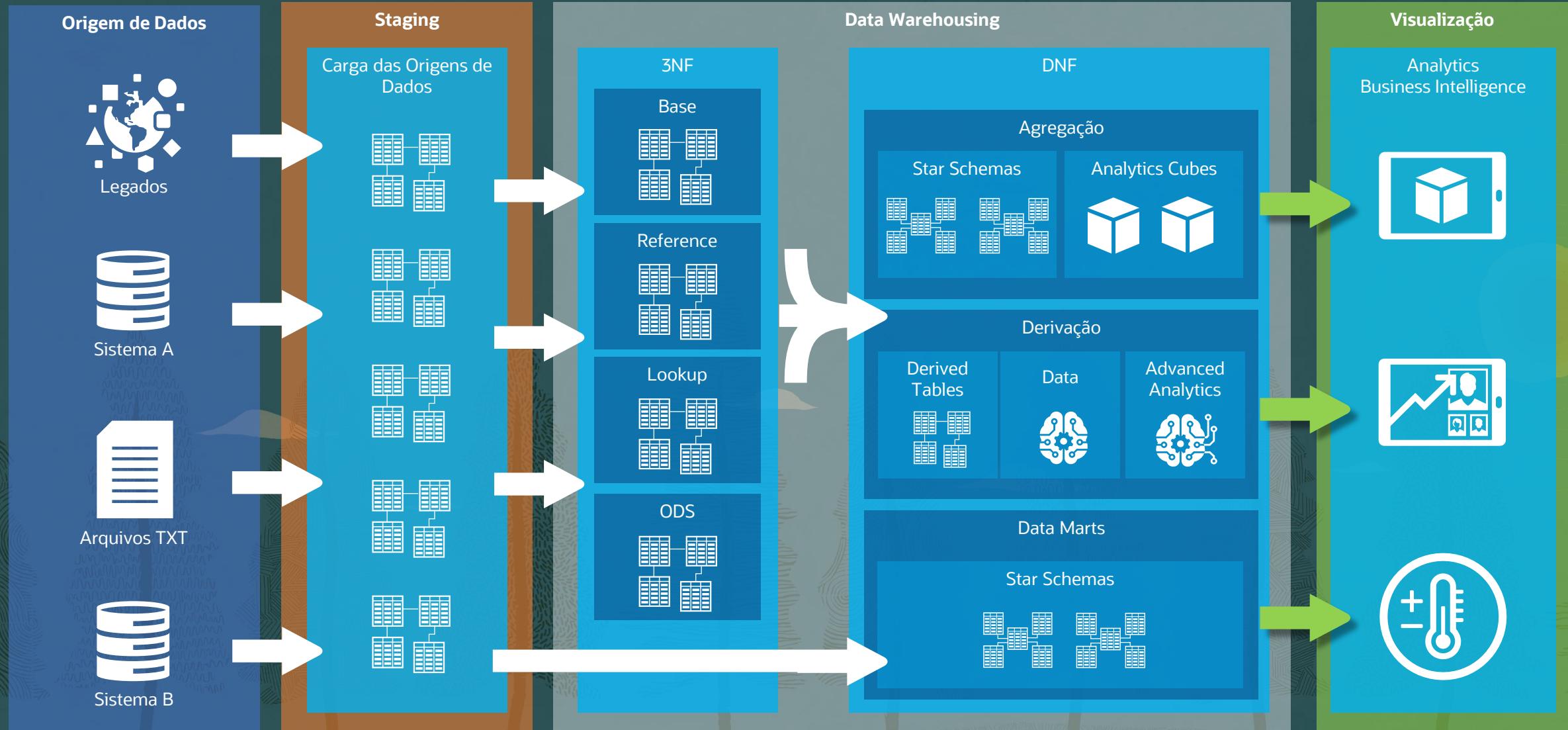
Cargas mistas | SQL para BI | Batch de ETL | Streaming | AI e ML

Acessibilidade

Referência: Learning Spark, 2nd Edition. O'Reilly Media. 2020. ISBN 9781492050049.



Arquitetura Clássica do Data Warehouse



A fama do Data Warehouse



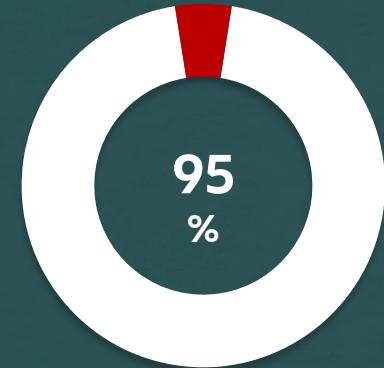
Manual



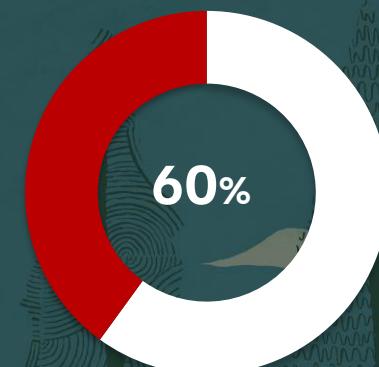
Complexo e caro



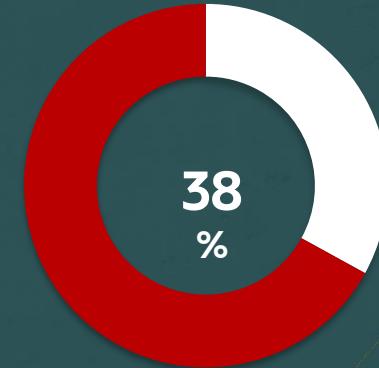
Lento



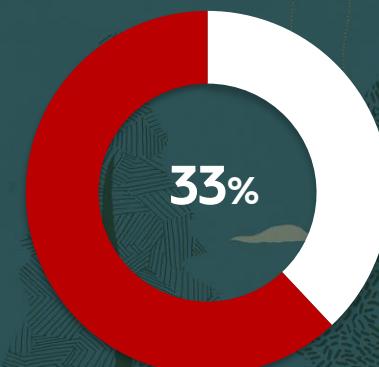
Requer extenso envolvimento manual



Muito complexo de gerenciar



Aquisição inicial e custos contínuos com manutenção

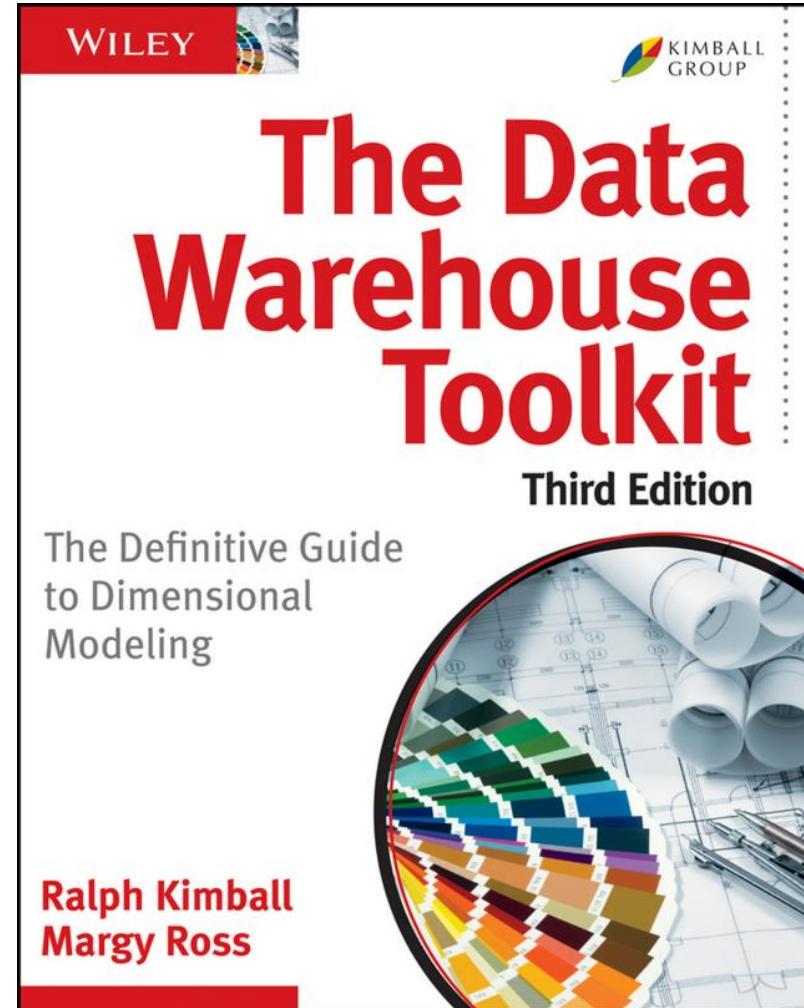


Lento e demorado de implementar

Fonte: Dimensional Research – The State of the Data Warehouse

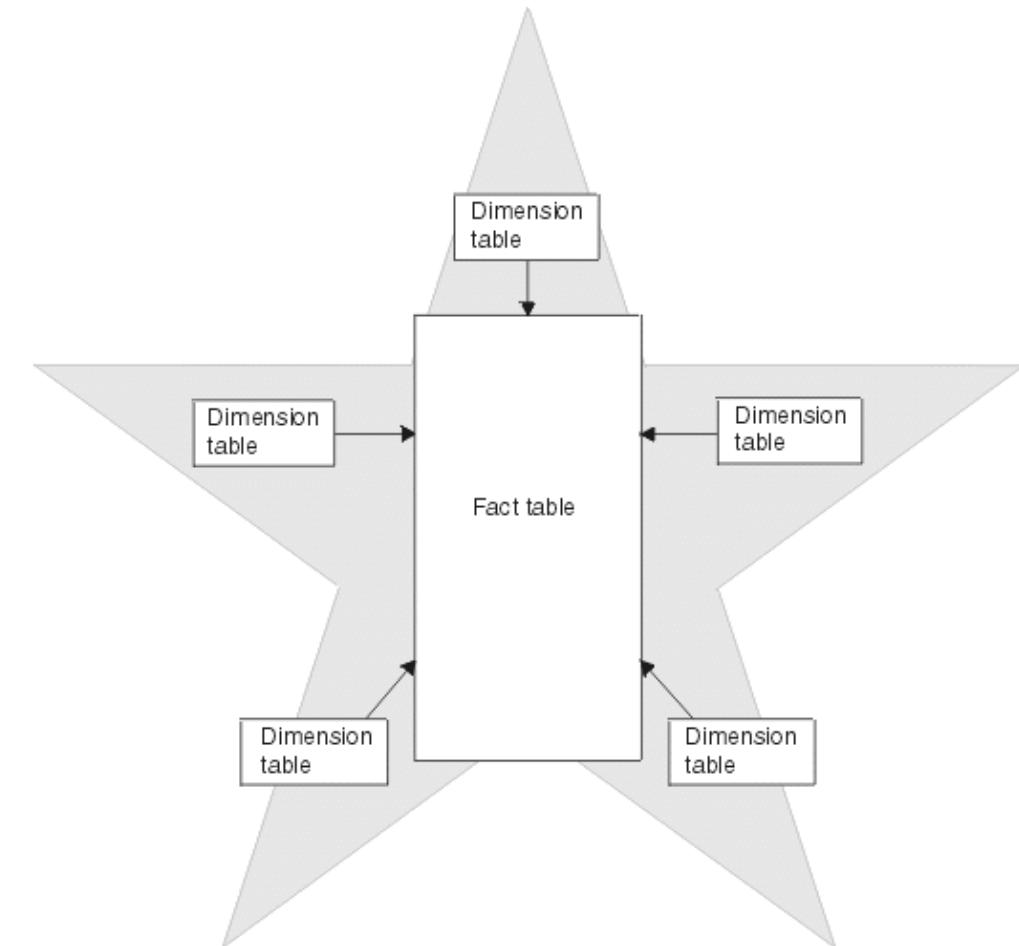
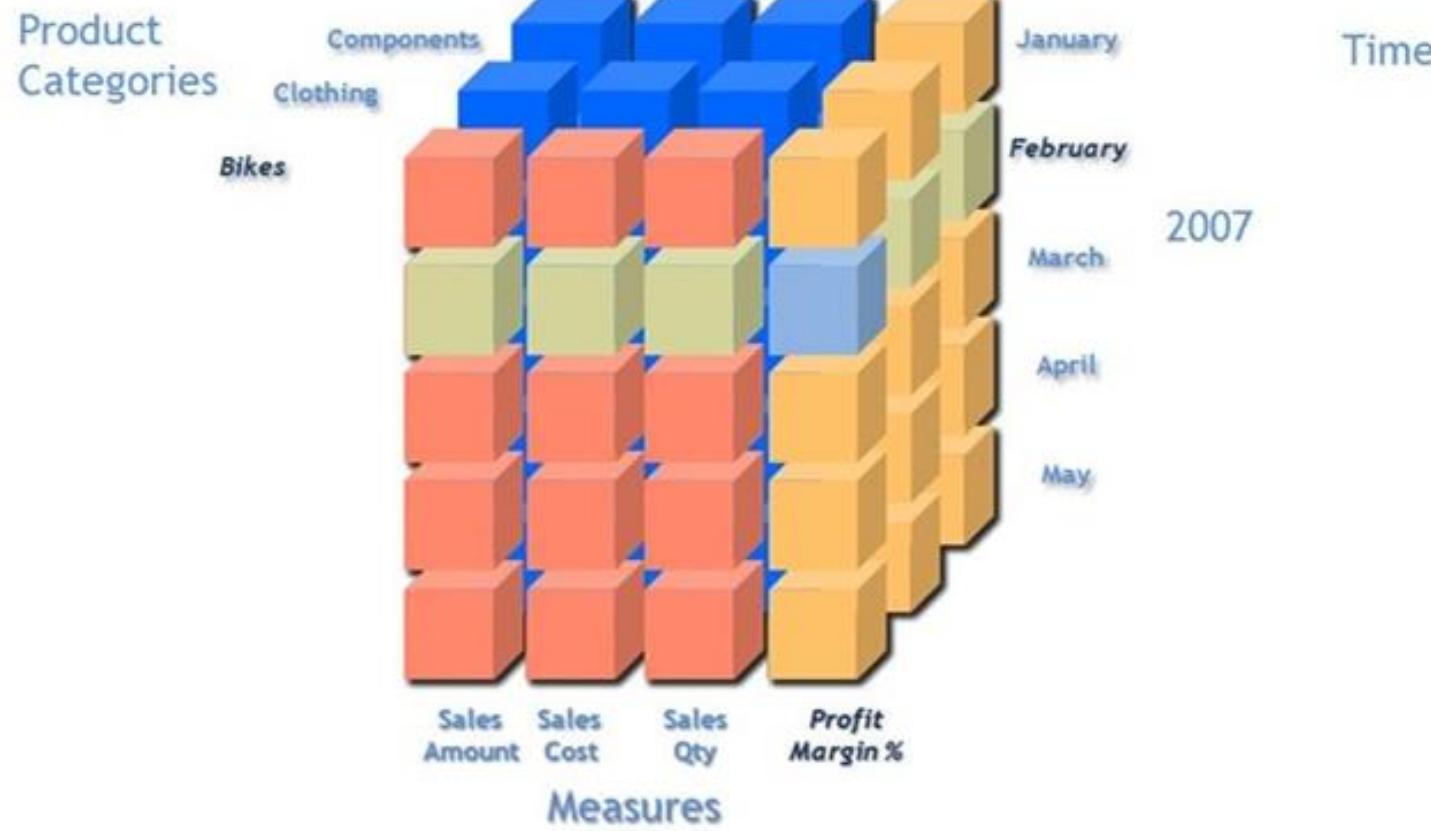
Modelagem Dimensional de Dados

Kimball, Ralph; Ross, Margy. The Data Warehouse Toolkit, The Definitive Guide to Dimensional Modeling, Third Edition, Wiley.

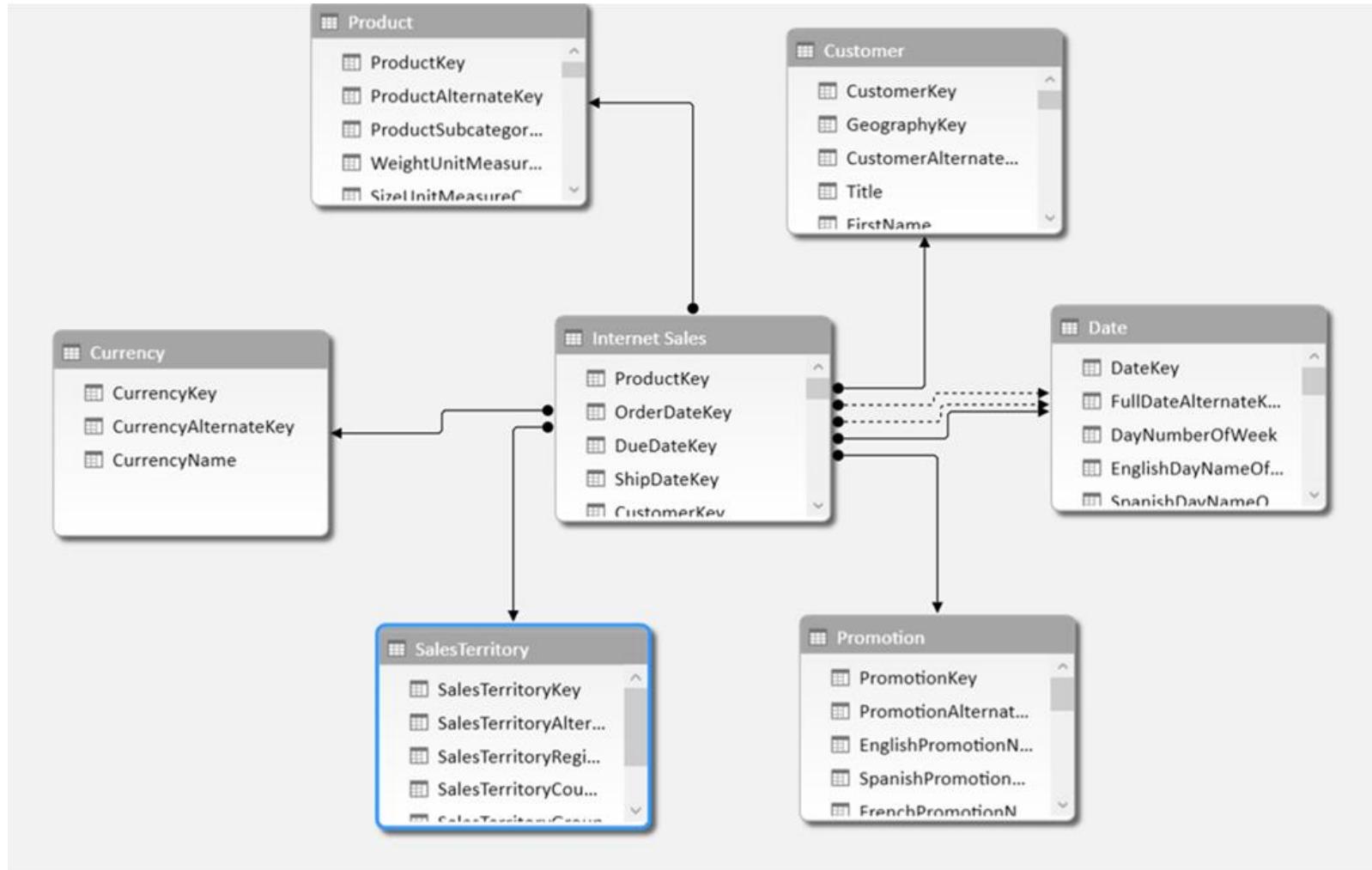


Modelagem Dimensional de Dados (Star Schema)

For Bikes show me the Profit Margin% for February



Modelagem Dimensional de Dados (Star Schema)



Processo de Modelagem Dimensional de Dados em 4 Passos

A Modelagem de Dados Dimensional requer 4 decisões:

- 1) Selecionar o processo de negócio.
- 2) Declarar o grão.
- 3) Identificar as dimensões.
- 4) Identificar os fatos.

Processo de Negócio: atividades operacionais executadas pela organização.

Grão: determina o que cada registro (linha na tabela) representa.

Dimensão: atributos descritivos (o que, quem, onde, quando, por que e como).

Fato: medições resultantes da execução do processo de negócio.



Prática de Modelagem Dimensional de Dados

Dim_Data_Vendas
Data_Vendas_SK
Data_Vendas_Dia
Data_Vendas_Mes_Vendas_SK
Data_Vendas_Mes_Vendas_ID
Data_Vendas_Trimestre_Vendas_SK
Data_Vendas_Trimestre_Vendas_ID
Data_Vendas_Ano_Venda_SK
Data_Vendas_Ano_Venda_ID
Trimestre_Fiscal_Vendas_SK
Trimestre_Fiscal_Vendas_Fiscal_ID
Ano_Fiscal_Venda_SK
Ano_Fiscal_Venda_ID

Fato_Item_Venda
Data_Vendas_SK
Hora_Vendas_SK
Loja_SK
Produto_SK
Vendedor_SK
Transacao_Venda_SK
Item_Venda_Quantidade
Item_Venda_Preco_Venda
Item_Venda_Custo
Item_Venda_Valor_Frete
Item_Venda_Valor_Tributo
Item_Venda_Margem_Lucro
Item_Venda_Valor_Lucro

Dim_Produto
Produto_SK
Produto_Codigo_PK
Produto_Nome
Produto_Descricao
Produto_Subcategoria_SK
Produto_Subcategoria_Codigo_ID
Produto_Subcategoria_Nome
Produto_Categoria_SK
Produto_Categoria_Codigo_ID
Produto_Categoria_Nome
Produto_Fabricante_SK
Produto_Fabricante_Codigo_ID
Produto_Fabricante_Nome

Dim_Hora_Vendas
Hora_Vendas_SK
Hora_Vendas_HHMM
Hora_Vendas_Periodo

Dim_Vendedor
Vendedor_SK
Vendedor_Codigo_PK
Vendedor_Nome
Vendedor_Cargo

Dim_Loja
Loja_SK
Loja_Codigo_PK
Loja_Nome
Loja_Tamanho_Metros
Loja_Bairro_SK
Loja_Bairro_Codigo_ID
Loja_Bairro_Nome
Loja_Cidade_SK
Loja_Cidade_Codigo_ID
Loja_Cidade_Nome
Loja_UF_SK
Loja_UF_Codigo_ID
Loja_UF_Nome
Loja_Pais_SK
Loja_Pais_Codigo_ID
Loja_Pais_Nome_Usual
Loja_Pais_Nome_Completo

Dim_Transacao_Venda
Transacao_Venda_SK
Transacao_Venda_Codigo_PK



Prática de Modelagem Dimensional de Dados

Dim_Loja
Loja_SK
Loja_Codigo_PK
Loja_Nome
Loja_Tamanho_Metros
Loja_Bairro_SK
Loja_Bairro_Codigo_ID
Loja_Bairro_Nome
Loja_Cidade_SK
Loja_Cidade_Codigo_ID
Loja_Cidade_Nome
Loja_UF_SK
Loja_UF_Codigo_ID
Loja_UF_Nome
Loja_Pais_SK
Loja_Pais_Codigo_ID
Loja_Pais_Nome_Usual
Loja_Pais_Nome_Completo

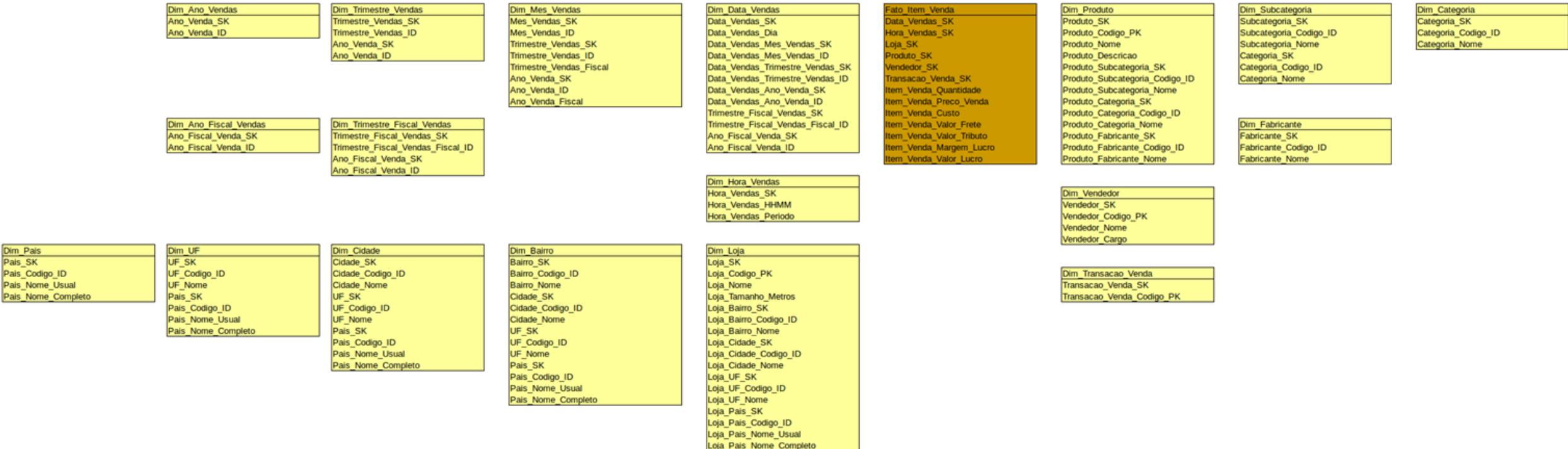
Dim_Data_Vendas
Data_Vendas_SK
Data_Vendas_Dia
Data_Vendas_Mes_Vendas_SK
Data_Vendas_Mes_Vendas_ID
Data_Vendas_Trimestre_Vendas_SK
Data_Vendas_Trimestre_Vendas_ID
Data_Vendas_Ano_Venda_SK
Data_Vendas_Ano_Venda_ID
Trimestre_Fiscal_Vendas_SK
Trimestre_Fiscal_Vendas_Fiscal_ID
Ano_Fiscal_Venda_SK
Ano_Fiscal_Venda_ID

Fato_Item_Venda
Data_Vendas_SK
Hora_Vendas_SK
Loja_SK
Produto_SK
Vendedor_SK
Transacao_Venda_SK
Item_Venda_Quantidade
Item_Venda_Preco_Venda
Item_Venda_Custo
Item_Venda_Valor_Frete
Item_Venda_Valor_Tributo
Item_Venda_Margem_Lucro
Item_Venda_Valor_Lucro

Dim_Hora_Vendas
Hora_Vendas_SK
Hora_Vendas_HHMM
Hora_Vendas_Periodo

Prática de Modelagem Dimensional de Dados

Famílias de Dimensão



Prática de Modelagem Dimensional de Dados

Famílias de Dimensão

Dim Ano Vendas
Ano_Venda_SK
Ano_Venda_ID

Dim Trimestre Vendas
Trimestre_Vendas_SK
Trimestre_Vendas_ID
Ano_Venda_SK
Ano_Venda_ID

Dim Mes Vendas
Mes_Vendas_SK
Mes_Vendas_ID
Trimestre_Vendas_SK
Trimestre_Vendas_ID
Trimestre_Vendas_Fiscal
Ano_Venda_SK
Ano_Venda_ID
Ano_Venda_Fiscal

Dim Data Vendas
Data_Vendas_SK
Data_Vendas_Dia
Data_Vendas_Mes_Vendas_SK
Data_Vendas_Mes_Vendas_ID
Data_Vendas_Trimestre_Vendas_SK
Data_Vendas_Trimestre_Vendas_ID
Data_Vendas_Ano_Venda_SK
Data_Vendas_Ano_Venda_ID
Trimestre_Fiscal_Vendas_SK
Trimestre_Fiscal_Vendas_Fiscal_ID
Ano_Fiscal_Venda_SK
Ano_Fiscal_Venda_ID

Dim Ano Fiscal Vendas
Ano_Fiscal_Venda_SK
Ano_Fiscal_Venda_ID

Dim Trimestre Fiscal Vendas
Trimestre_Fiscal_Vendas_SK
Trimestre_Fiscal_Vendas_Fiscal_ID
Ano_Fiscal_Venda_SK
Ano_Fiscal_Venda_ID

Dim Hora Vendas
Hora_Vendas_SK
Hora_Vendas_HHMM
Hora_Vendas_Periodo

Prática de Modelagem Dimensional de Dados

Famílias de Dimensão

Dim_Pais
Pais_SK
Pais_Codigo_ID
Pais_Nome_Usual
Pais_Nome_Completo

Dim_UF
UF_SK
UF_Codigo_ID
UF_Nome
Pais_SK
Pais_Codigo_ID
Pais_Nome_Usual
Pais_Nome_Completo

Dim_Cidade
Cidade_SK
Cidade_Codigo_ID
Cidade_Nome
UF_SK
UF_Codigo_ID
UF_Nome
Pais_SK
Pais_Codigo_ID
Pais_Nome_Usual
Pais_Nome_Completo

Dim_Bairro
Bairro_SK
Bairro_Codigo_ID
Bairro_Nome
Cidade_SK
Cidade_Codigo_ID
Cidade_Nome
UF_SK
UF_Codigo_ID
UF_Nome
Pais_SK
Pais_Codigo_ID
Pais_Nome_Usual
Pais_Nome_Completo

Dim_Loja
Loja_SK
Loja_Codigo_PK
Loja_Nome
Loja_Tamanho_Metros
Loja_Bairro_SK
Loja_Bairro_Codigo_ID
Loja_Bairro_Nome
Loja_Cidade_SK
Loja_Cidade_Codigo_ID
Loja_Cidade_Nome
Loja_UF_SK
Loja_UF_Codigo_ID
Loja_UF_Nome
Loja_Pais_SK
Loja_Pais_Codigo_ID
Loja_Pais_Nome_Usual
Loja_Pais_Nome_Completo

Prática de Modelagem Dimensional de Dados

Famílias de Dimensão

Dim_Produto
Produto_SK
Produto_Codigo_PK
Produto_Nome
Produto_Descricao
Produto_Subcategoria_SK
Produto_Subcategoria_Codigo_ID
Produto_Subcategoria_Nome
Produto_Categoria_SK
Produto_Categoria_Codigo_ID
Produto_Categoria_Nome
Produto_Fabricante_SK
Produto_Fabricante_Codigo_ID
Produto_Fabricante_Nome

Dim_Subcategoria
Subcategoria_SK
Subcategoria_Codigo_ID
Subcategoria_Nome
Categoria_SK
Categoria_Codigo_ID
Categoria_Nome

Dim_Categoria
Categoria_SK
Categoria_Codigo_ID
Categoria_Nome

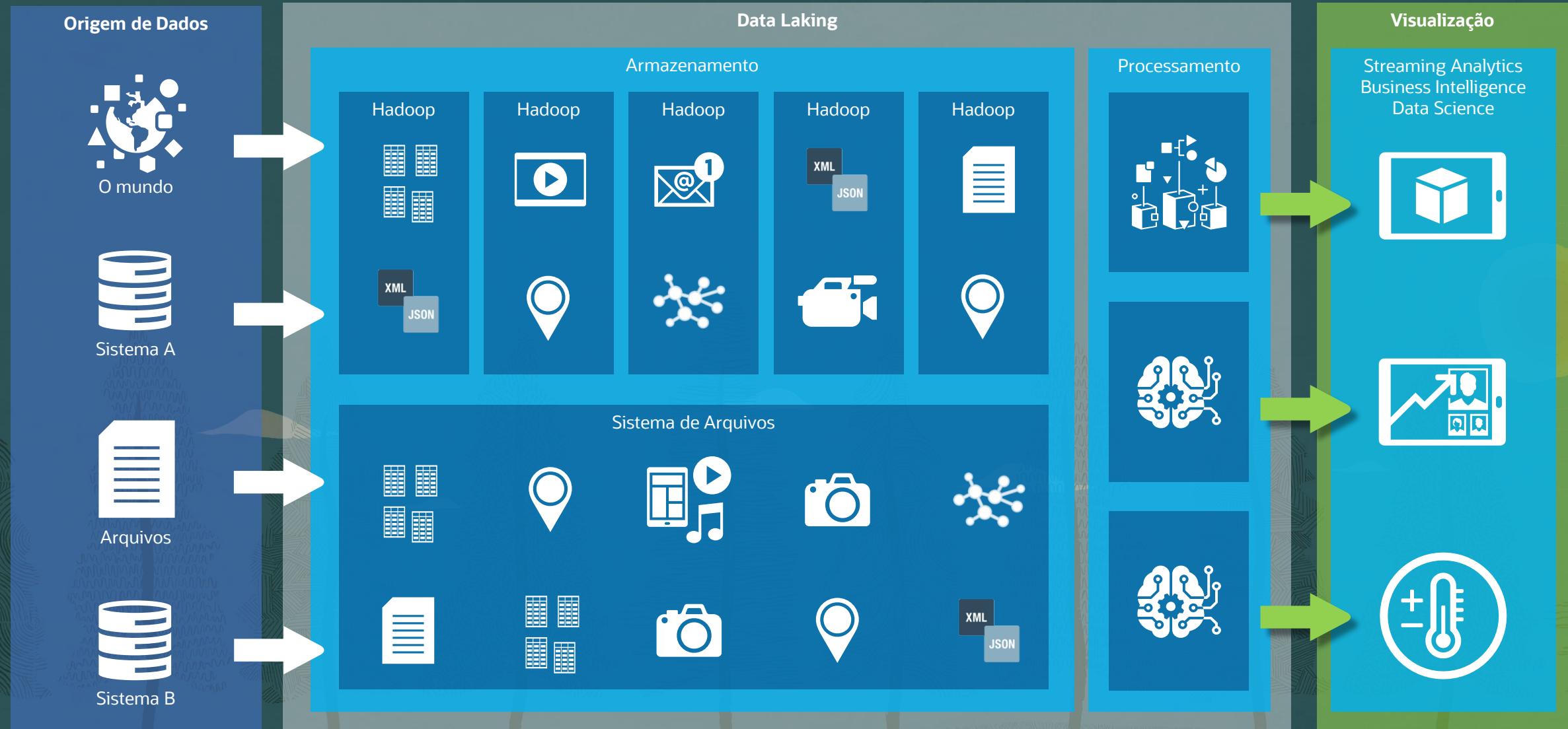
Dim_Fabricante
Fabricante_SK
Fabricante_Codigo_ID
Fabricante_Nome



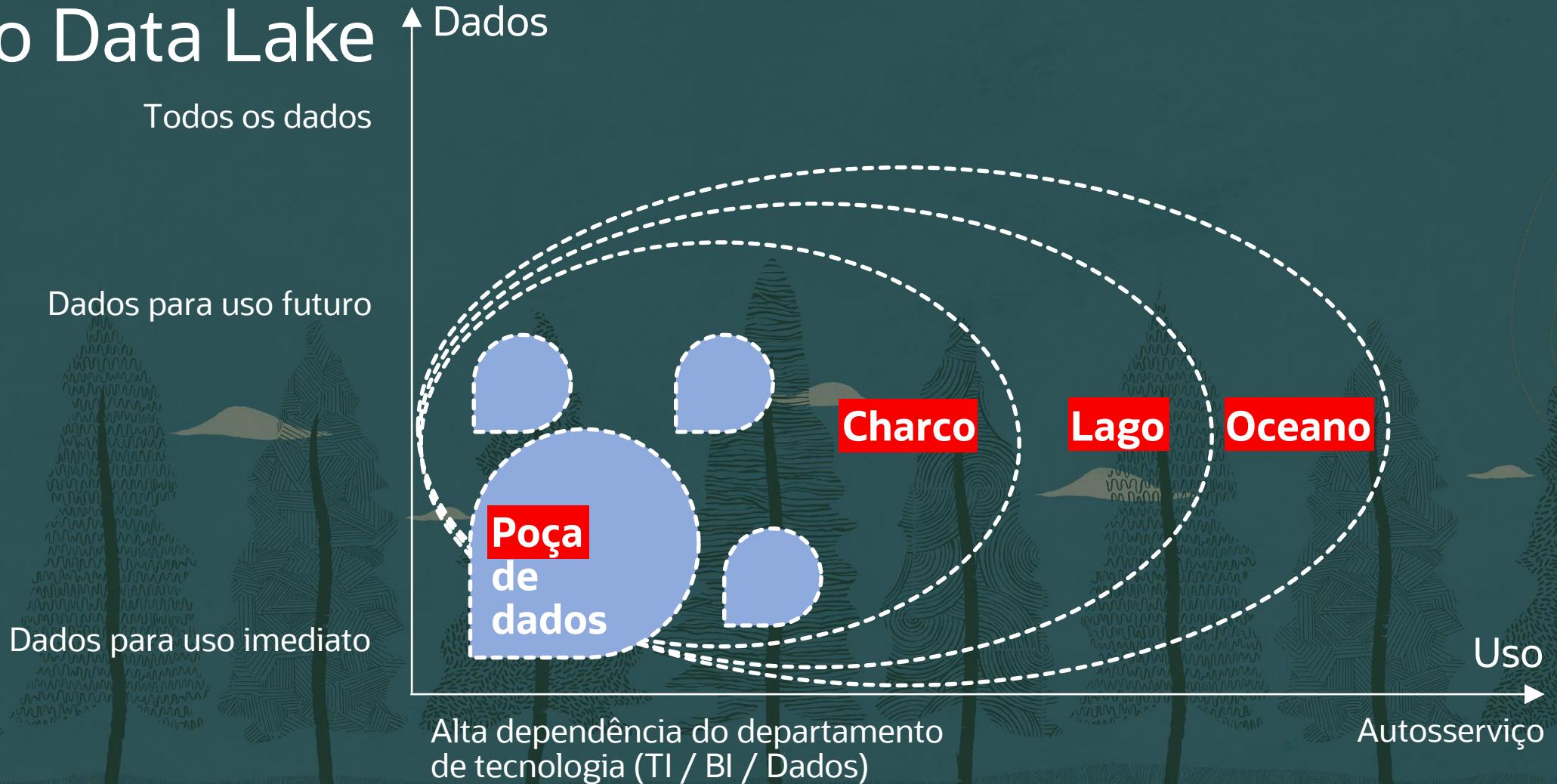


Foto de Felix M. Dorn no Unsplash

Arquitetura Clássica do Data Lake



Maturidade do Data Lake



A fama do Data Lake



Falta de atomicidade e isolamento transacional



Inconsistência de dados e qualidade de dados reduzida

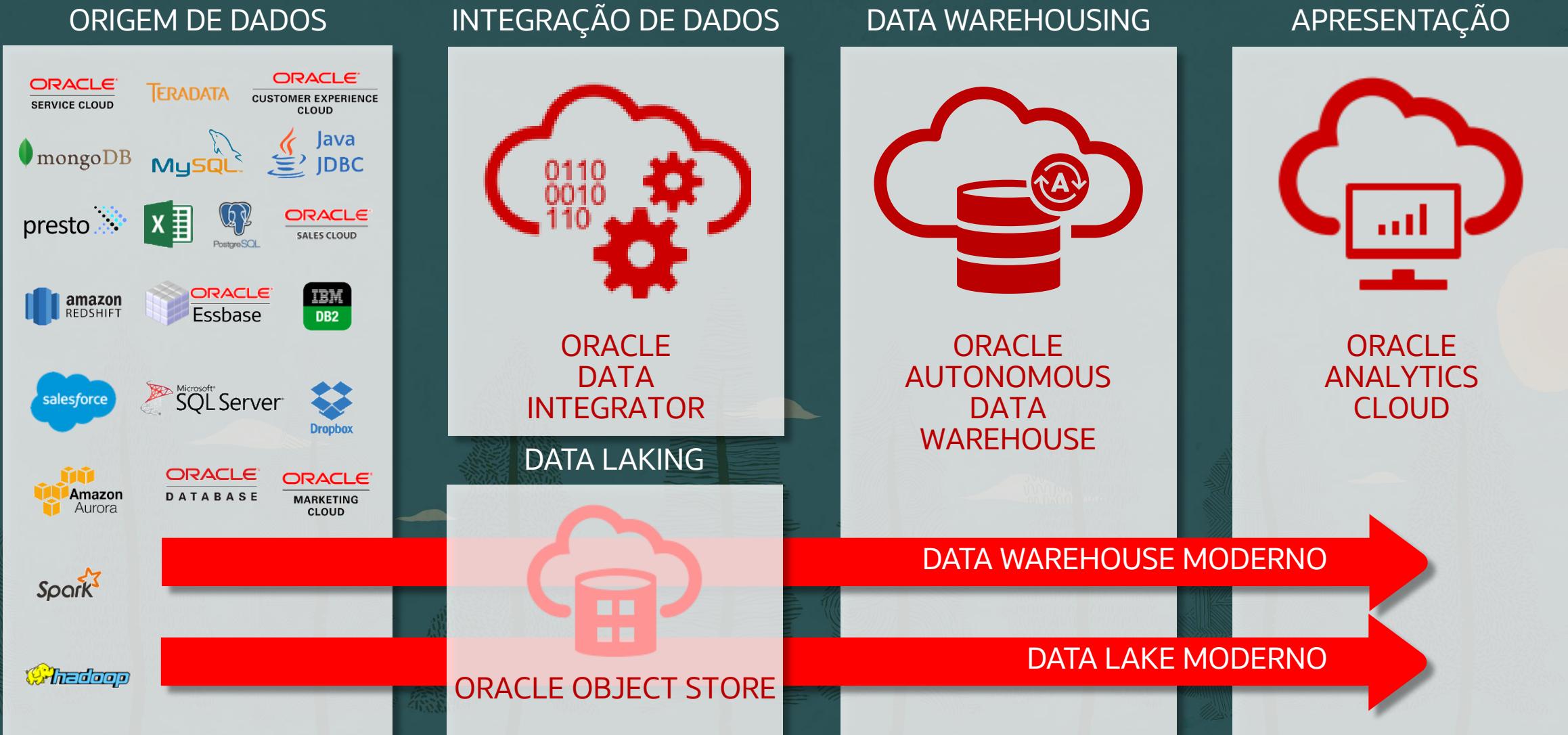


Caótico e complexo

Referência: Learning Spark, 2nd Edition, O'Reilly Media, 2020. ISBN 9781492050049.



Arquitetura de Solução Cloud Data Warehouse & Lake



Abordagem para Data Warehouse Moderno

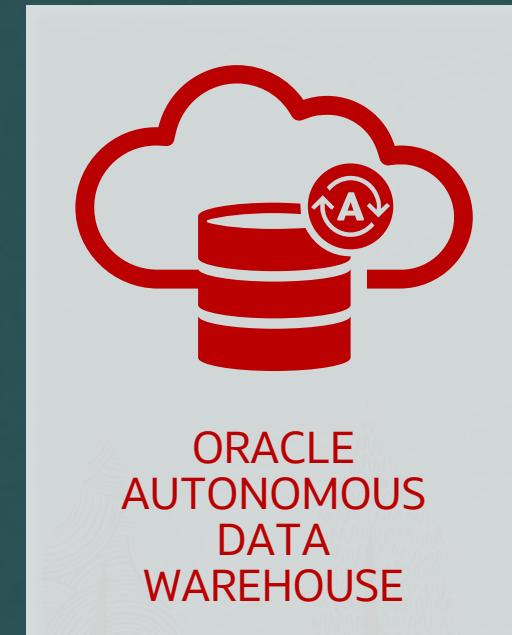
ORIGEM DE DADOS



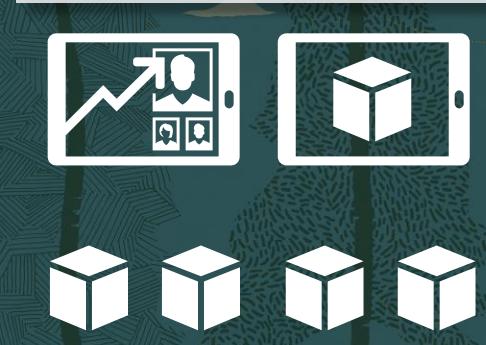
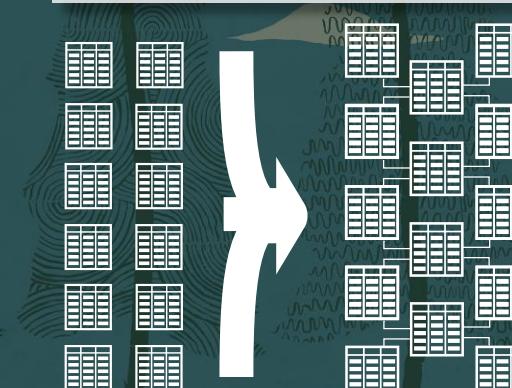
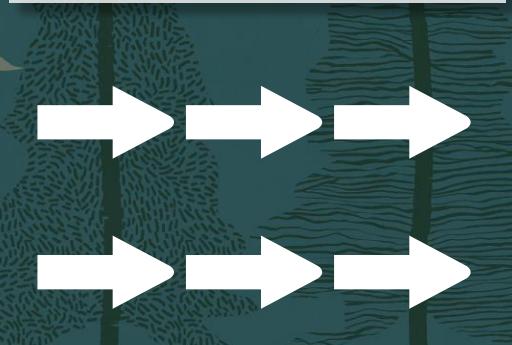
INTEGRAÇÃO DE DADOS



DATAWAREHOUSING



APRESENTAÇÃO



Abordagem para Data Lake Moderno

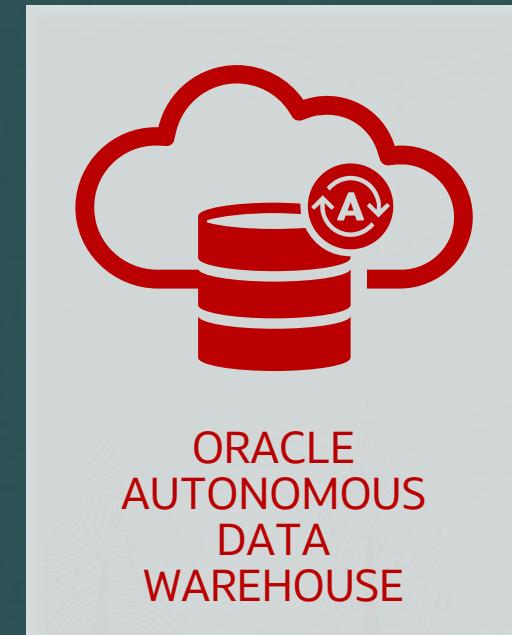
ORIGEM DE DADOS



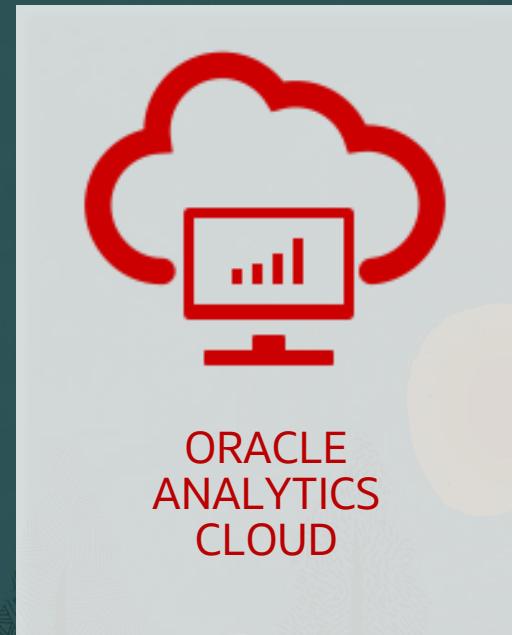
INTEGRAÇÃO DE DADOS



DATA LAKING



APRESENTAÇÃO



E se... eu quiser um Data Lake “mesmo”?





Foto de Evelyn Paris no Unsplash

Lakehouse: um novo paradigma que combina elementos de Data Lake e Data Warehouse.



TRANSAÇÃO
ACID



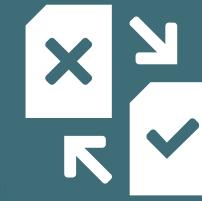
CONFORMIDADE
DE ESQUEMA



FORMATOS
DIVERSOS E
ABERTOS



CARGAS
MISTAS

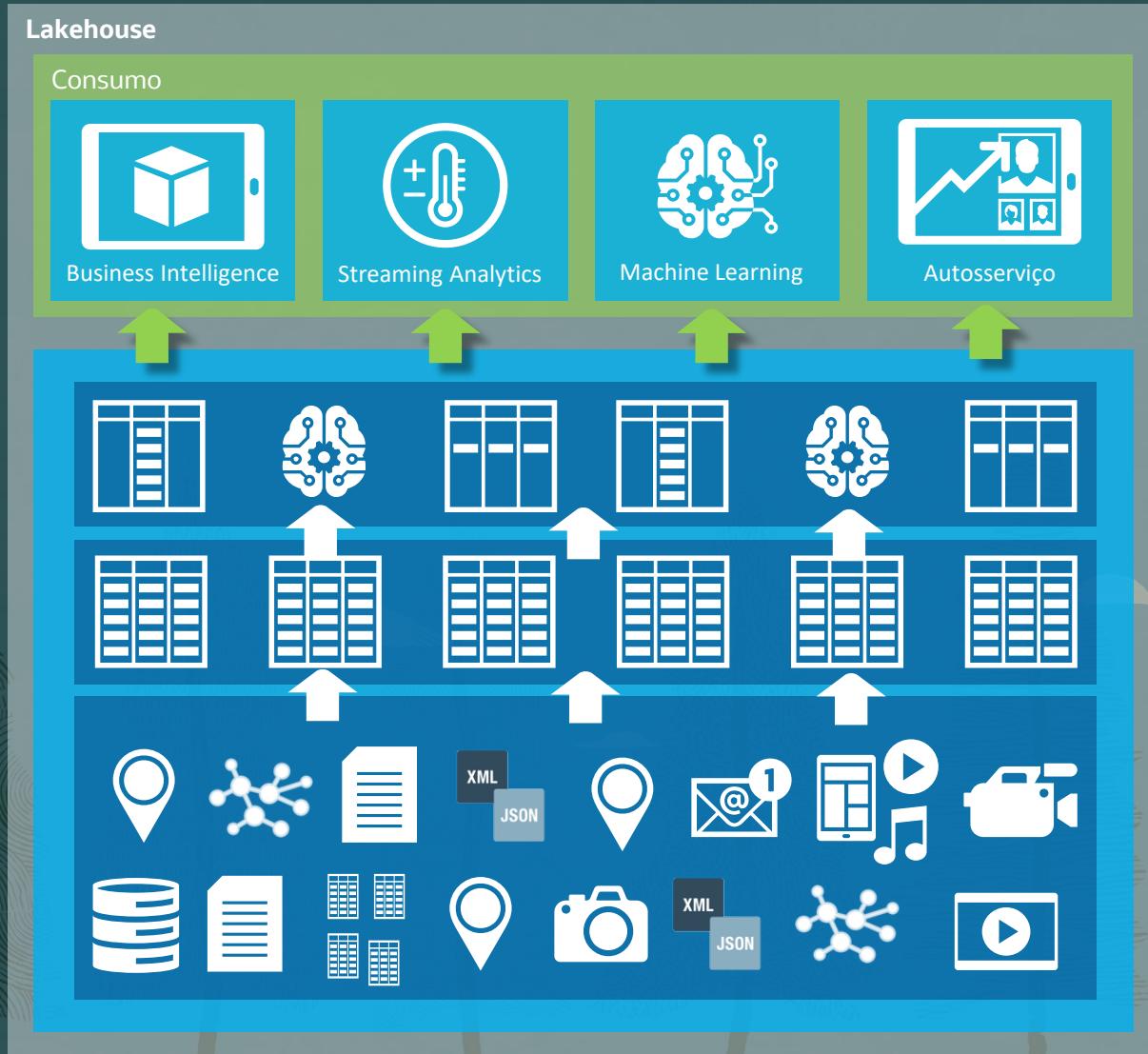


UPSERT &
DELETE
PARALELOS



GOVERNANÇA
DE DADOS

Arquitetura de Dados Lakehouse



Plataforma única para consumo

Motor de alto desempenho para consultas

Camada transacional estruturada

Data Lake para todos os dados

Projetos Lakehouse

APACHE HUDI

Hadoop Update Delete and Incremental

Focado em upserts e deletes em Chave-Valor

Combina formatos colunares e lineares

APACHE ICEBERG

Focado em propósito geral de armazenamento em tabelas únicas de grande tamanho

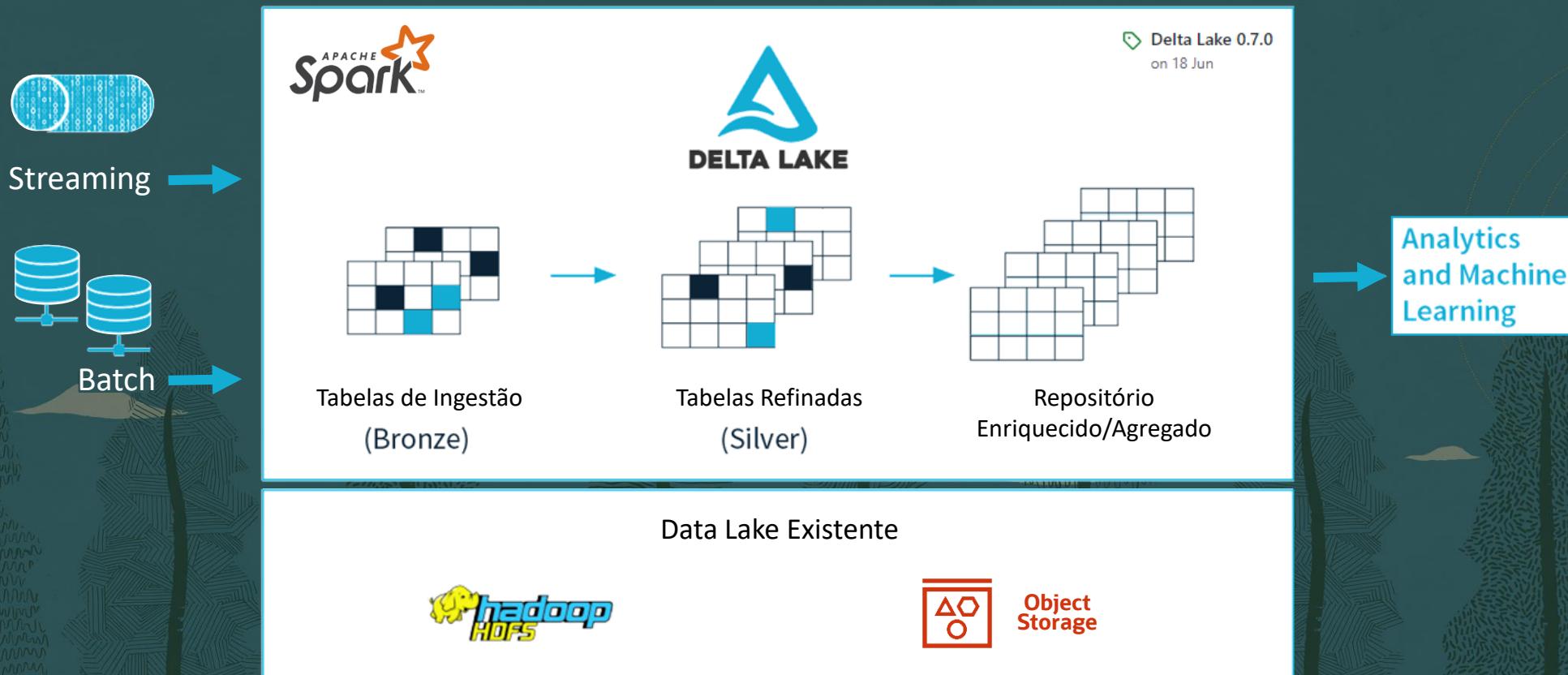
Evolução de esquema e particionamento, versionamento e isolamento serializado

DELTA LAKE

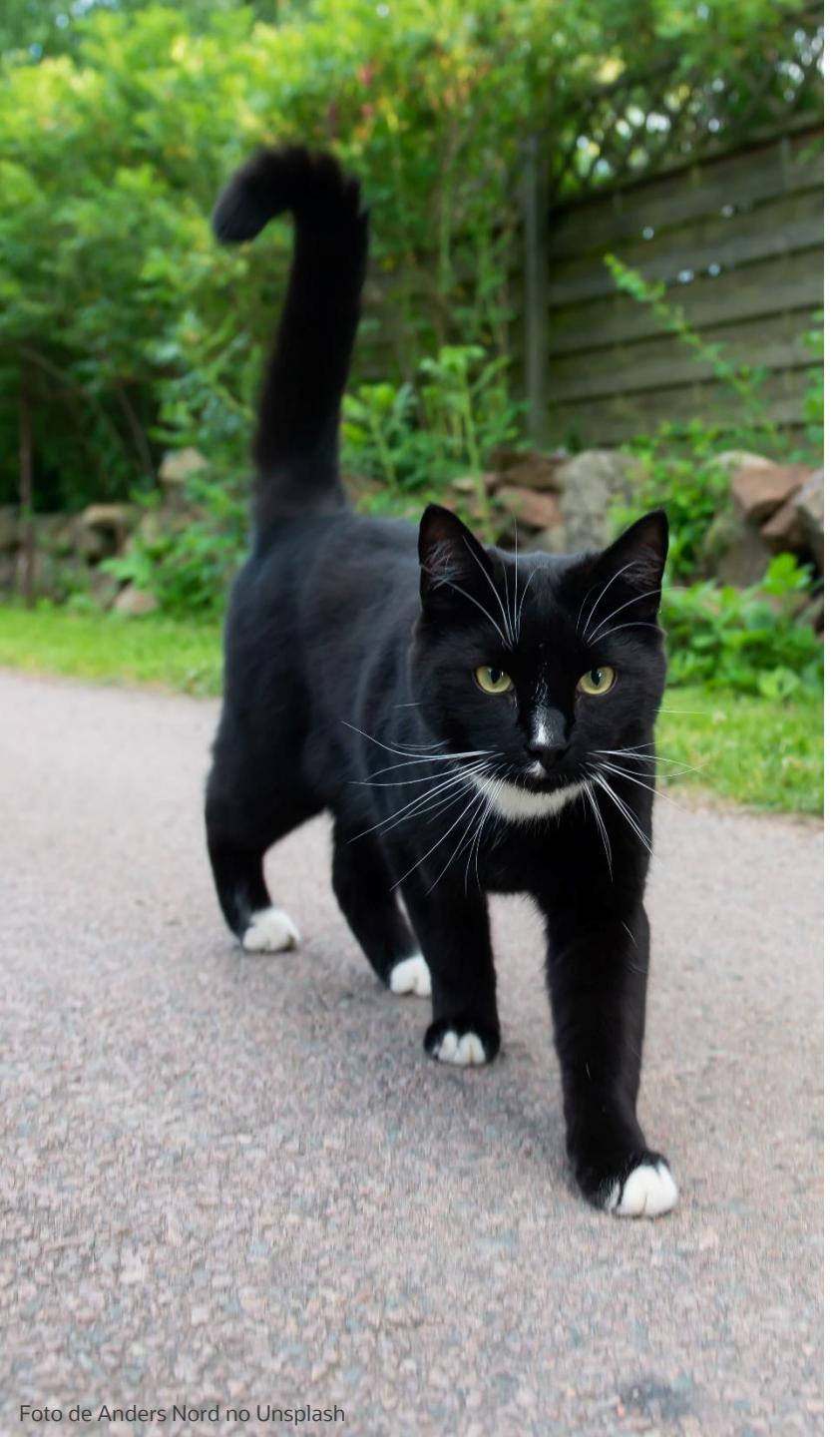
Mantido pela Linux Foundation e construído pelos criadores do Apache Spark

Formato aberto de armazenamento com suporte transacional e evolução de esquema

Arquitetura de Referência Delta Lake







Componentes do Data Warehouse moderno



INTEGRAÇÃO

Streaming,
batch data, on-premises e cloud



DATA WAREHOUSE

Autonomous,
self-driving,
self-securig,
self-repairing



DATA LAKE

Baseado em
Object storage e
integrado com o
Data Warehouse



ANALYTICS

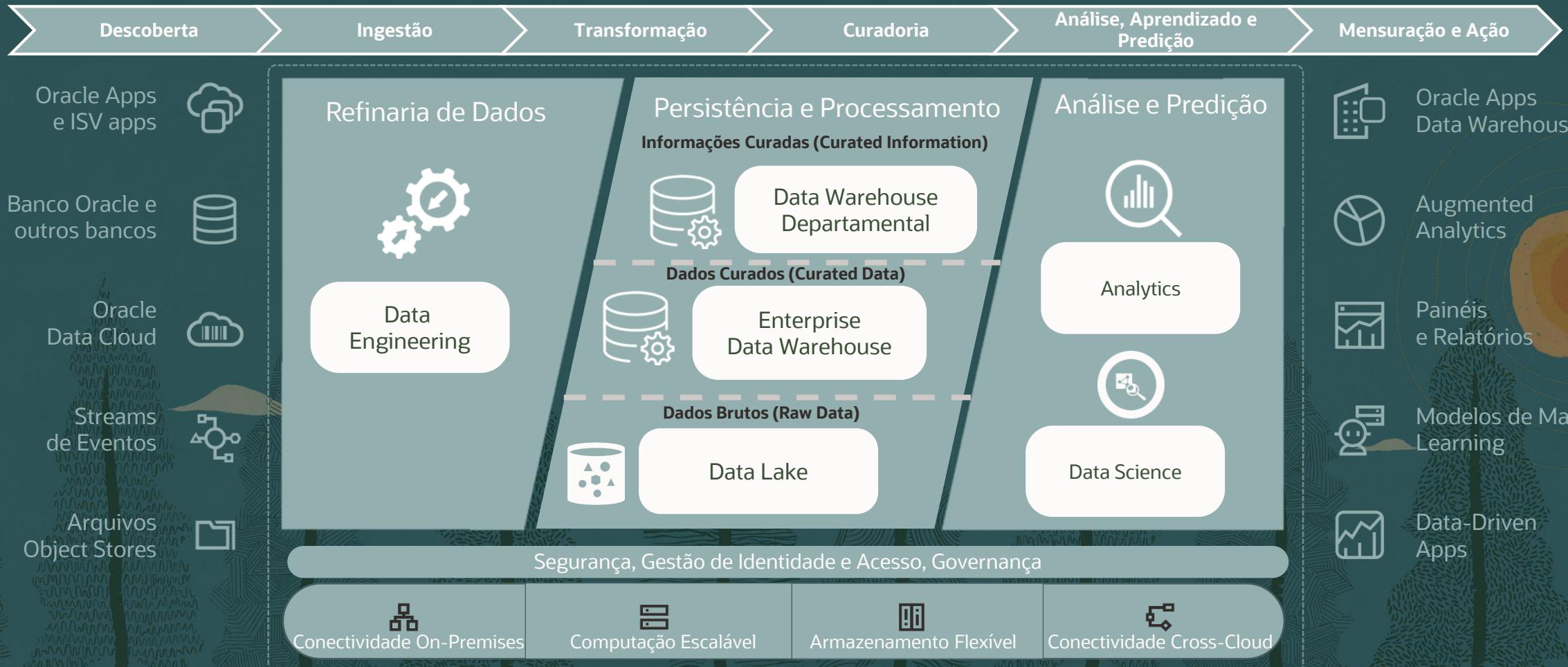
Visualização e
inteligênci
analític
baseadas
em Machine
Learning



DATA SCIENCE

Machine learning
de propósito geral
e in-database

Arquitetura de Solução do Data Warehouse moderno

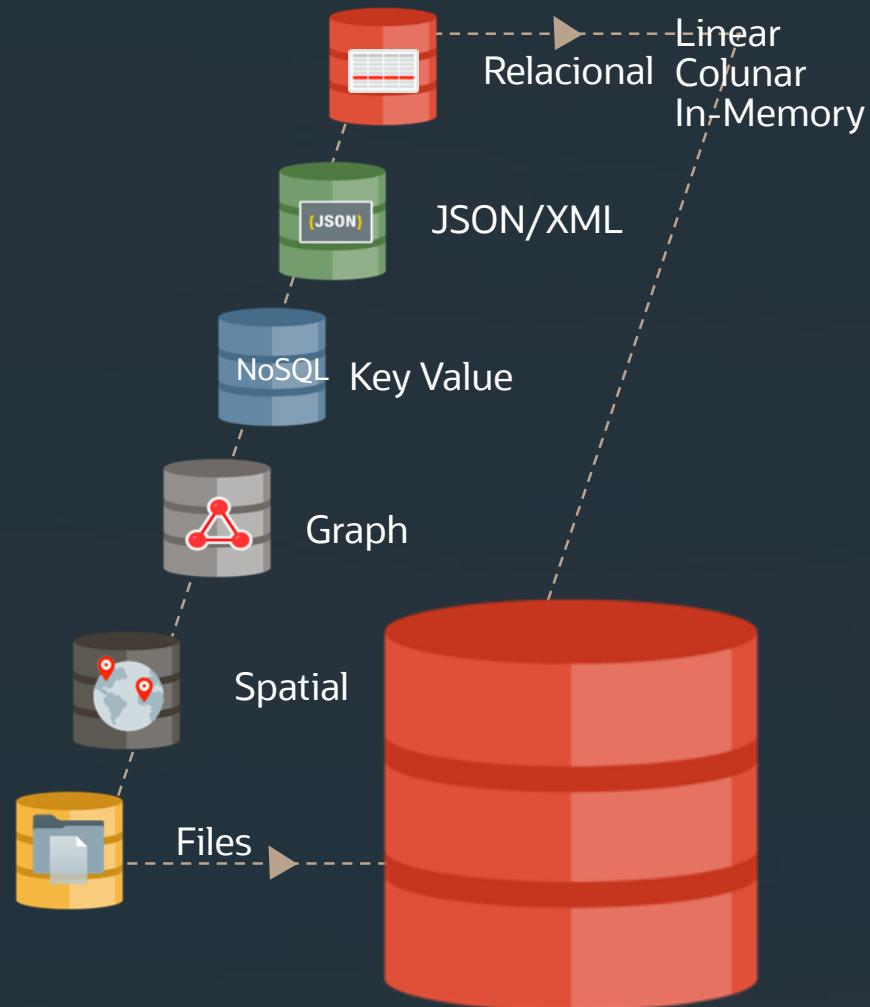


Gravidade
de dados...

Qual é a sua?



Oracle: Base de Dados Convergente



Resumindo...

Conheça a gravidade de dados do seu negócio e use todos os seus ativos de dados com as ferramentas maduras que você domina.

Data Warehouse e Data Lake são paradigmas complementares, relevantes ao seu negócio e permanecem com alta demanda pelo mercado.

O paradigma Lakehouse é o conjunto de práticas de sucesso conhecidas no manejo de dados e sua implementação pode ser aderente a diferentes tecnologias.

