



FUNDAMENTOS E TÉCNICAS EM CIÊNCIAS DE DADOS

PROF. JOSENALDE OLIVEIRA

josenalde.oliveira@ufrn.br

<https://github.com/josenalde/datascience>

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - UFRN

REFORÇANDO O CONCEITO DE NUVEM E SUAS ESTRUTURAS

O CAMINHO PARA A NUVEM

On-Premise: instalação própria



Database as a Service

Termo mais específico de uma **plataforma** com foco em uso dos dados e não na gerência/admin dos dados

<https://www.o3s.com.br/post/o-caminho-para-a-nuvem>

REFORÇANDO O CONCEITO DE NUVEM E SUAS ESTRUTURAS

Exemplos:

IaaS: Amazon Web Services (AWS),
Elastic Compute Cloud (EC2),
Microsoft Azure Virtual Machines,
Google Compute Engine etc.

Exemplos:

PaaS: Amazon Elastic Beanstalk,
Microsoft Azure App,
Google App Engine etc.

Exemplos:

SaaS: Google Workspace,
Microsoft Office 365,
Dropbox,
Salesforce etc.

RETORNANDO AO CICLO DO DADO TRANSFORMAÇÃO DE DADOS

Dados são produzidos com estrutura normalmente apropriada à persistência

Dados são produzidos com estrutura normalmente apropriada ao processamento

Não necessariamente adequados às operações de ANÁLISE DE DADOS, necessitando de **TRANSFORMAÇÃO** para o alvo da análise!



Desde uma conversão de formato:
html/xml ↔ csv ↔ json ↔ txt
Arrays ↔ dataframes
Transformar modelo relacional em
multi-dimensional (junção,
sumarização etc.); Rotinas de
limpeza, qualidade, remoção de
duplicados, mudança na codificação

A processos de **ETL** (Extração,
Transformação e Carga (Load)) em
DATA WAREHOUSES

Depósitos de dados – disponíveis para ferramentas de análise
É chamado **ARMAZENAMENTO ANALÍTICO**
Dado o foco em análise, também são chamados de Business Intelligence (BI)

ETL: EXTRACT, TRANSFORM, LOAD

Uso clássico: **carga de dados em DW**

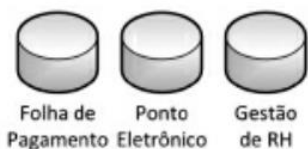
Mas também integração de dados e construção de **modelos analíticos**; Um ETL possui conexões com fontes de dados heterogêneas e geograficamente dispersas (CAD, GIS, Raster, DB corporativo, Web, GeoWeb, etc.)



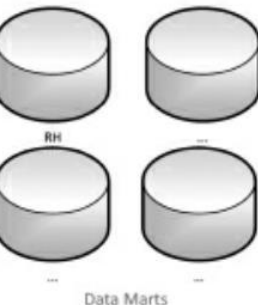
ETL

Sistemas de Origem

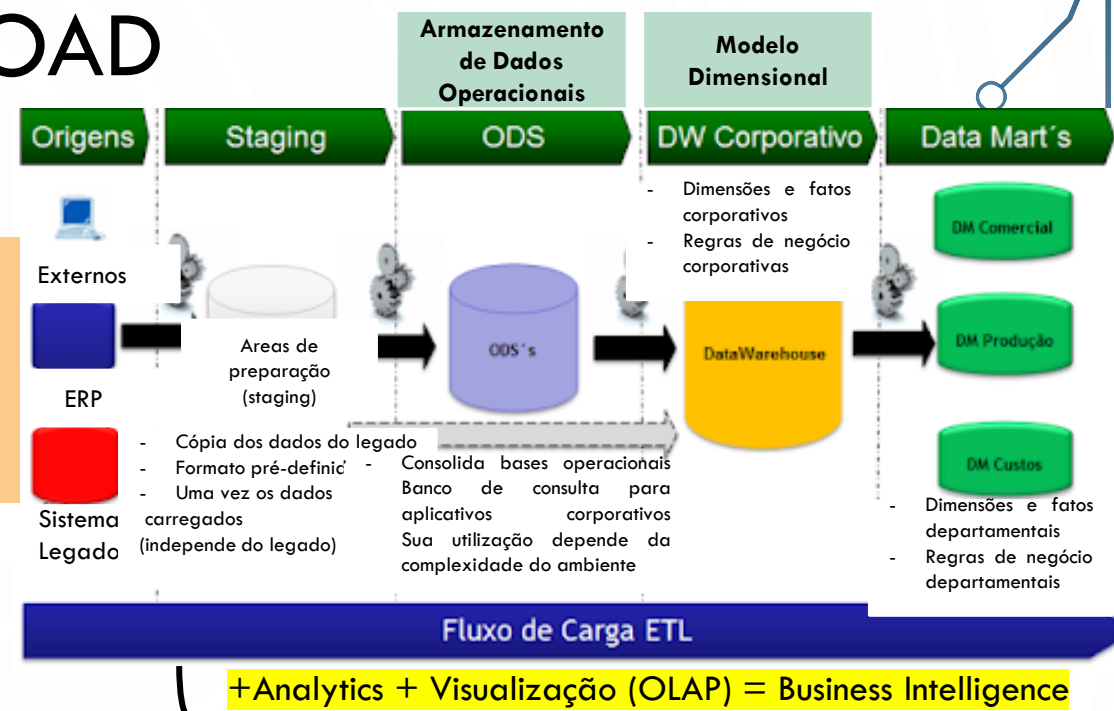
Data Warehouse



ETL



PIPELINE



Processos de extração, copiam dados das fontes (estruturadas/semi-estruturadas)

Staging: armazenamento temporário, pode ser em memória pode permanecer entre estágios ou ser apagado

ETL: Camada de integração e transformação

ODS é uma base de dados integrada, volátil, de valores correntes, e que contém somente dados detalhados.

Normalmente **são dados estruturados, bases relacionais**

Processo de ETL em lotes (intervalos)

Necessário tolerância às falhas (rede, discos, origem)

ETL: limpa, calcula, resume

CDC – CHANGED DATA CAPTURE

Carga em tempo real para Data Warehouses: com CDC você captura as mudanças de dados dos sistemas transacionais e já as entrega para o DW, em near real-time. É a letra E do ETL feita de uma forma *não-batch*. O CDC permite que o DW seja atualizado praticamente a todo tempo, e por isso essa capacidade habilita a disponibilização dos dados para suporte à decisão em tempo quase near real-time.

Exemplos:

<https://debezium.io/>

<https://www.oracle.com/br/integration/goldengate/>

EXEMPLOS DE SISTEMAS ETL



ORACLE
DATA INTEGRATOR

O Oracle Data Integrator (ODI) é uma **ferramenta ETL** integral para carregar dados em um grande **data warehouse**. A ferramenta foi projetada para grandes transferências de dados e transformações entre sistemas. ODI funciona melhor com bancos de dados Oracle e no Oracle Enterprise Manager.

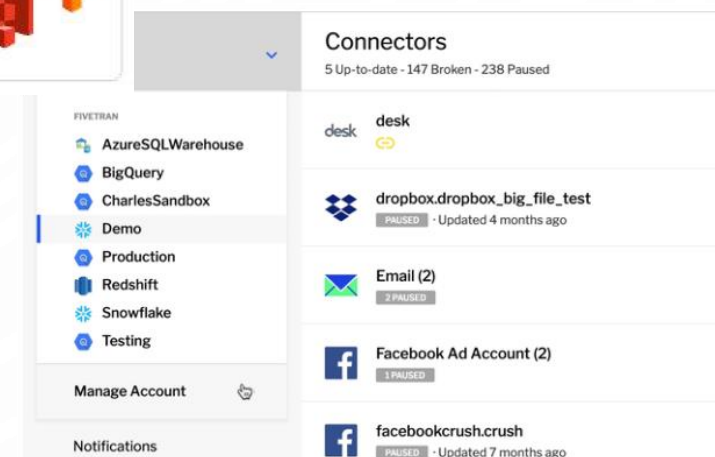
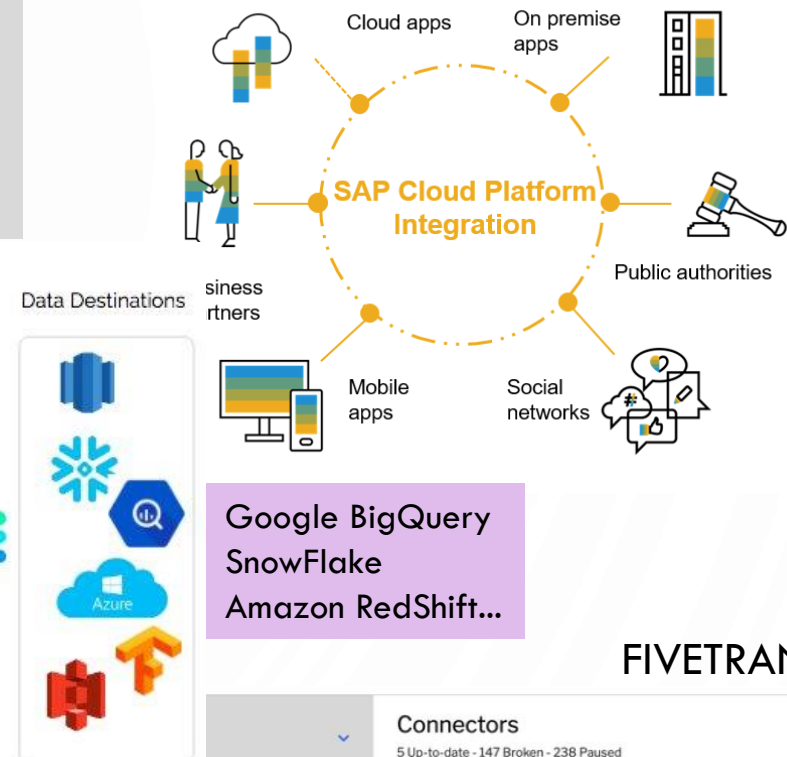
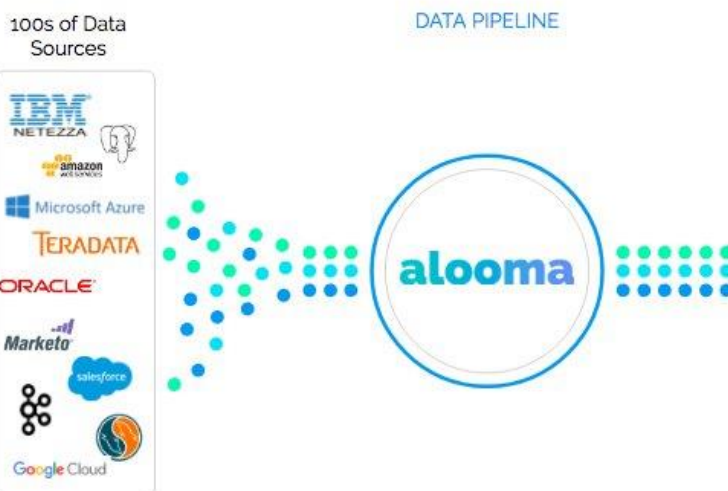
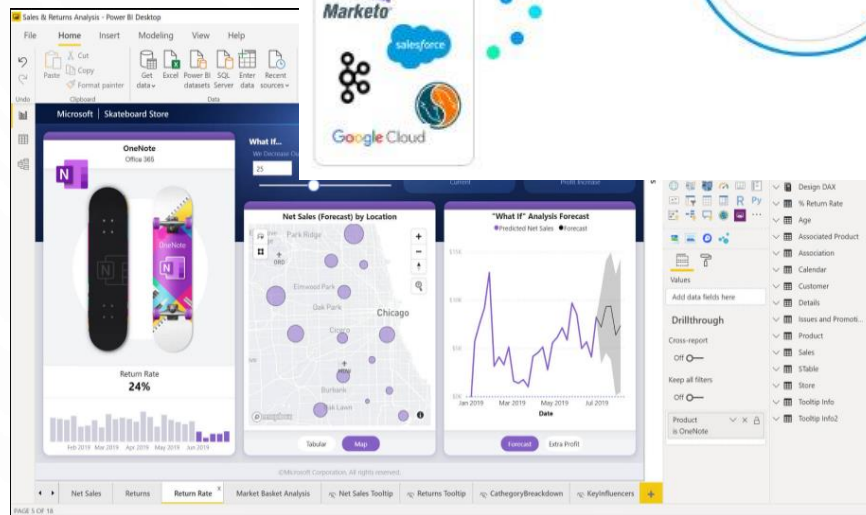
 **Informatica**
POWERCENTER


+80 conectores para bases
Azure Data Factory

 **Power BI**



Visualizações interativas,
personalizáveis, dashboards
Relatórios, análises



CONCEITO GERAL DE BUSINESS INTELLIGENCE (BI)

Ambiente de BI:

- Ambiente de negócios
- Empresas encontram em sua massa de dados informações sobre o seu negócio para:
 - Antecipar tendências
 - Lançar produtos
 - Conhecer melhor clientes
 - Concentra info para tomada de decisão

Ambiente de BI:

- Sistemas diferentes, plataformas diferentes, objetivos diferentes
- ETL faz parte do ambiente corporativo (no ambiente de BI)
- Normalmente construído sobre um DW

Desde a década de 90 surgem os DATA WAREHOUSES – estruturados a partir de **bases relacionais** na época

- a) contudo, adaptados à análise, com informações pré-calculadas e incluindo formatos semi-estruturados
- b) guarda informações históricas (3-10 anos), apenas consulta e inclui, não altera e exclui (isto é feito na origem)
 - Regras de negócio para descarte de modo a evitar ‘data swamp’ – **pântanos de dados**
- c) no DW a informação é sempre atualizada, mantendo um histórico, portanto não volátil (**Change data capture**)
- d) EXEMPLO: atualização de cadastro (normalmente perde o anterior)
- e) Data Warehouse usa o **MODELO MULTI-DIMENSIONAL – ARMAZENAMENTO ANALÍTICO**
- f) Este modelo não se preocupa com redundância (tal como o RELACIONAL), mas está focado na ANÁLISE

MODELO DIMENSIONAL

Fatos

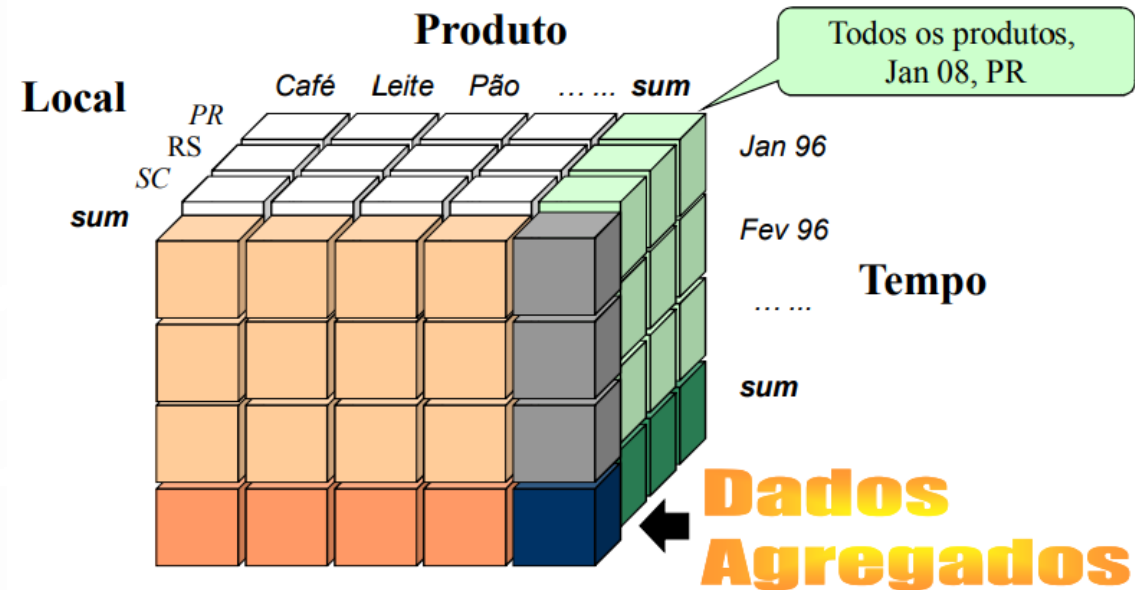
Elemento central do modelo. É a informação central, o tema ao qual se quer analisar

Dimensões

Um fato possui dimensões que são os diversos pontos de vista sobre o qual se quer analisar o fato

Medidas

Um fato possui medidas que são valores a serem analisados e pré-calculados, aditivos, semi-aditivos, não aditivos



Grão: nível de detalhe da informação gerencial
Menor: mais detalhe – células brancas acima

Exemplo 1: se a escolaridade de um cliente de determinada empresa muda, um novo registro é incluído, mantendo o histórico

Cliente	Escolaridade	Data da Alteração
José da Silva	Ensino Médio	21/09/2011

José da Silva	Ensino Médio	21/09/2011
José da Silva	Ensino Superior	30/12/2013

Tabela FATO com os dados do cliente

MODELO DIMENSIONAL

Exemplo 2: fato VENDAS com dimensões PRODUTO, VENDEDOR e FILIAL. Usuário pode analisar a venda sob qualquer destas dimensões: por PRODUTO, por VENDEDOR, por FILIAL. Uma dimensão TEMPO também é obrigatória pois o DW traz dados históricos. As medidas do FATO podem ser o valor da venda ou do lucro. O usuário pode ver o total e o lucro (medidas) das vendas (fato) por produto, vendedor ou filial (dimensões)

Fato:	Vendas
Dimensões:	Produto Vendedor Filial
Medidas:	Valor Lucro Quantidade

Um fato pode responder a diferentes perguntas, com a mesma medida vista por diferentes dimensões

Um fato também permite gerar uma gama de relatórios, indicadores (KPI), etc.

Fato: Vendas

Vendedores

Quem vendeu mais no Bimestre?
Quem vendeu mais no Ano?
Quem vendeu mais no Trimestre?

Produto

Qual produto apresentou mais margem?
Qual produto vendem mais na região Sul?

Clientes

Qual o nosso maior cliente?
Qual cliente está aumentando suas compras no Bimestre?



Google Data Studio

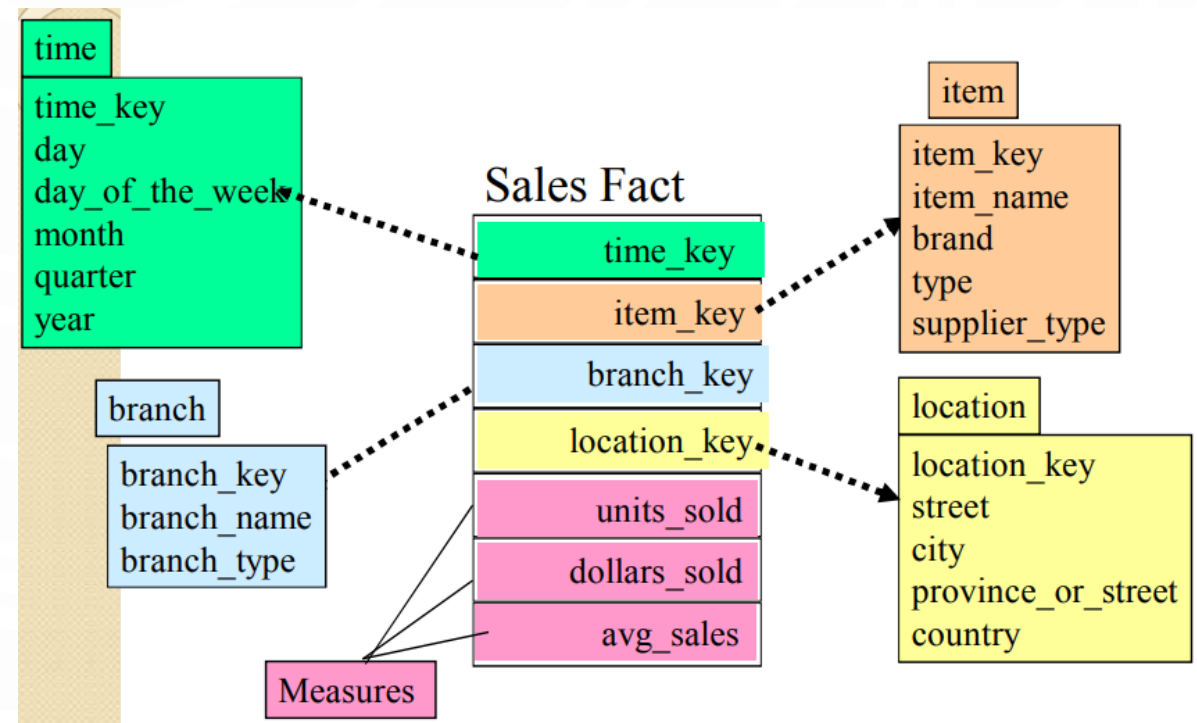
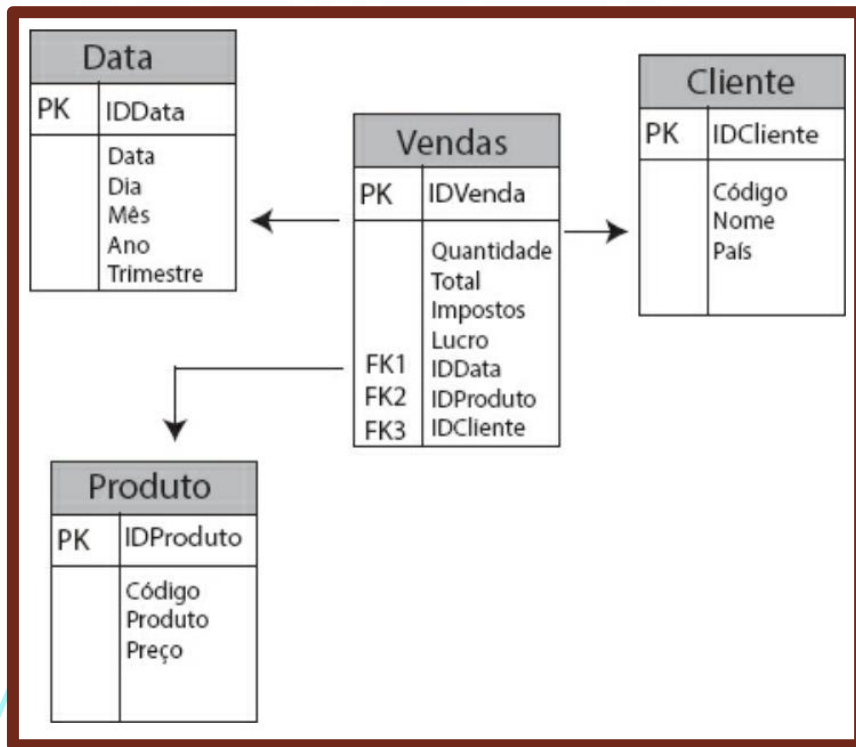
MODELO DIMENSIONAL - ESTRELA

MODELAGEM ESTRELA: FATO ao centro, com medidas já calculadas no processo de carga

IDCliente em Cliente e IDProduto em Produto são chamadas **chave substituta**

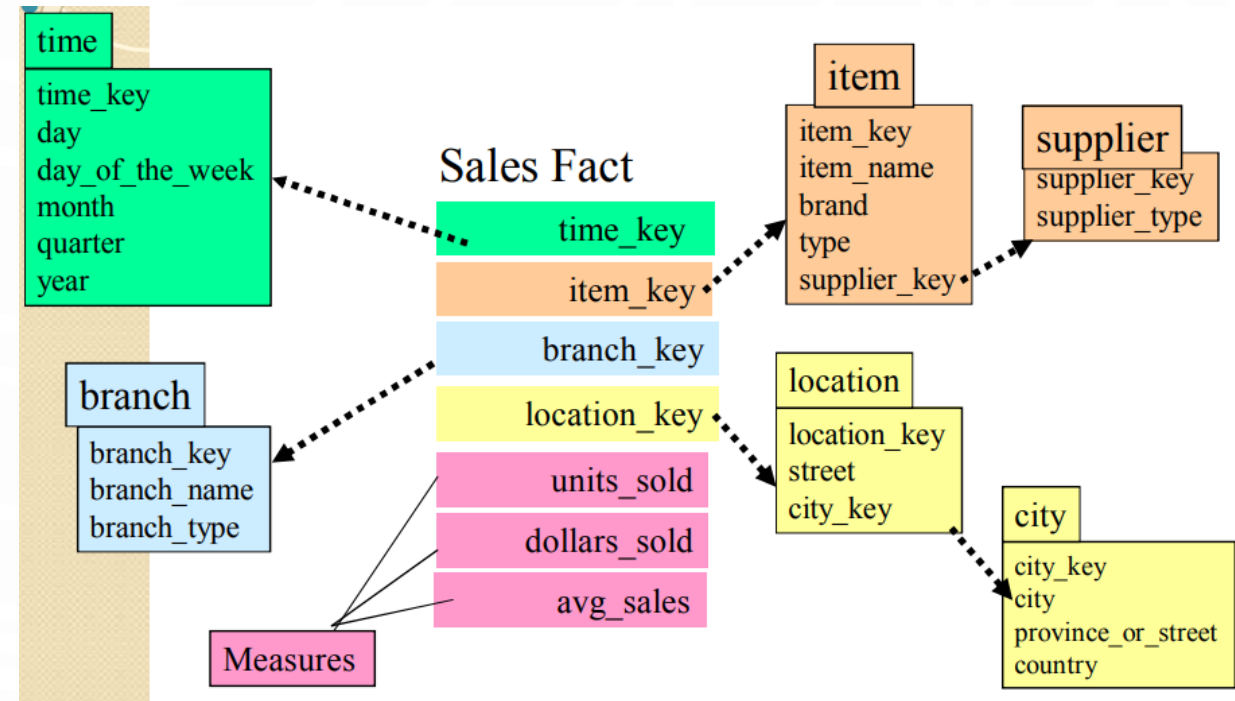
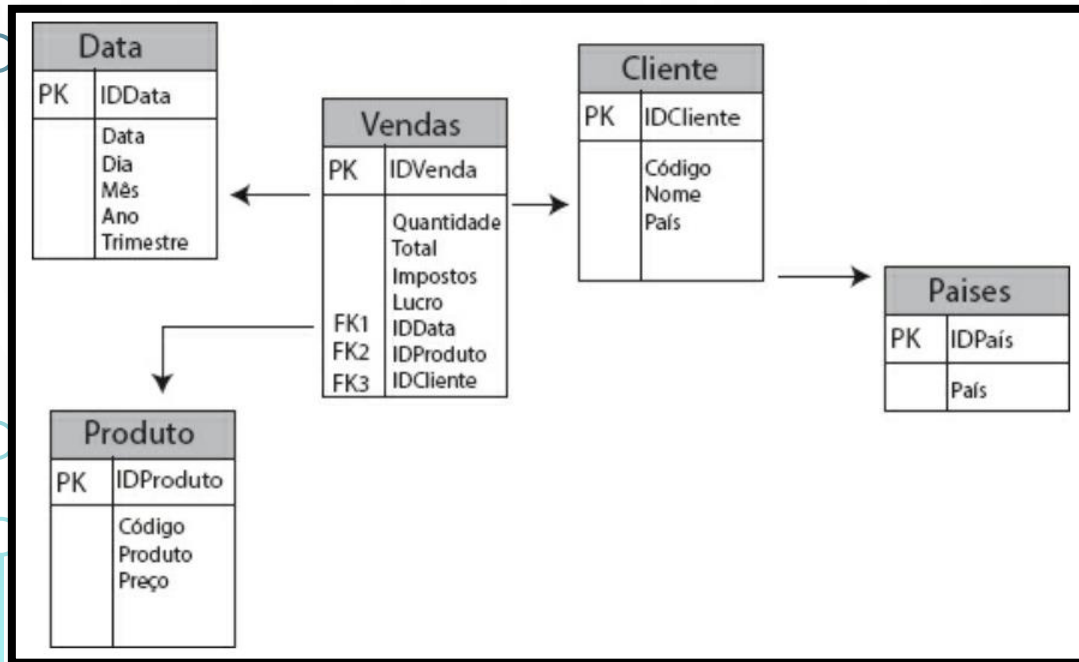
Não pode ser a PK como no relacional, pois o mesmo cliente pode ser inserido várias vezes na mesma dimensão CLIENTE e um produto várias vezes na dimensão PRODUTO

O campo **Código** é a chave primária da tabela de origem, para manter vínculo



MODELO DIMENSIONAL – FLOCO DE NEVE

MODELAGEM FLOCO DE NEVE: Similar à estrela, mas numa dimensão em que um atributo não varie com frequência, cria-se um relacionamento um-para-muitos. No exemplo a seguir, a cada atualização do cliente o país é repetido, logo criou-se uma dimensão PAÍSES ligada ao CLIENTE



PARA ESTE MODELO, SQL ADAPTADO É OPÇÃO...

Linguagens e modelos de consulta

MDX – Multidimensional Expressions

OLAP: Online Analytical Processing – DW normalmente roda sobre OLAP Server – apropriado à bancos multi-dimensionais

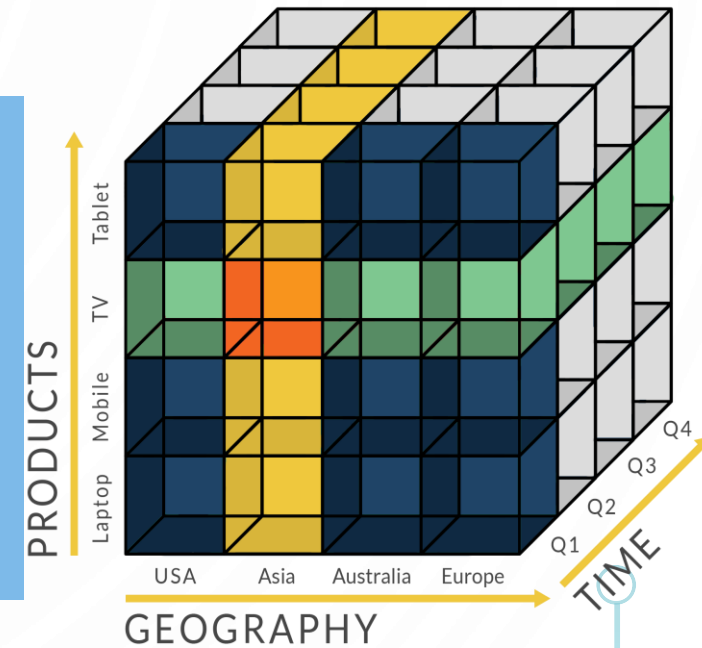
OLAP é multi-dimensional/analítico

OLTP (On-line Transactional Processing é relacional/transacional, clássico)

OLAP é um conceito de interface com o usuário que proporciona a capacidade de ter ideias sobre os dados, permitindo analisá-los profundamente em diversos ângulos. As funções básicas do OLAP são:

- Visualização multidimensional dos dados;
- Exploração;
- Rotação;
- Vários modos de visualização.

O OLAP e o DW são destinados a trabalharem juntos, enquanto o DW armazena as informações de forma eficiente, o OLAP deve recuperá-las com a mesma eficiência, porém com muita rapidez. As duas tecnologias se complementam, ao ponto de que um Data Warehouse para ser bem sucedido, já na sua concepção, deve levar em consideração o que se deseja apresentar na interface OLAP.

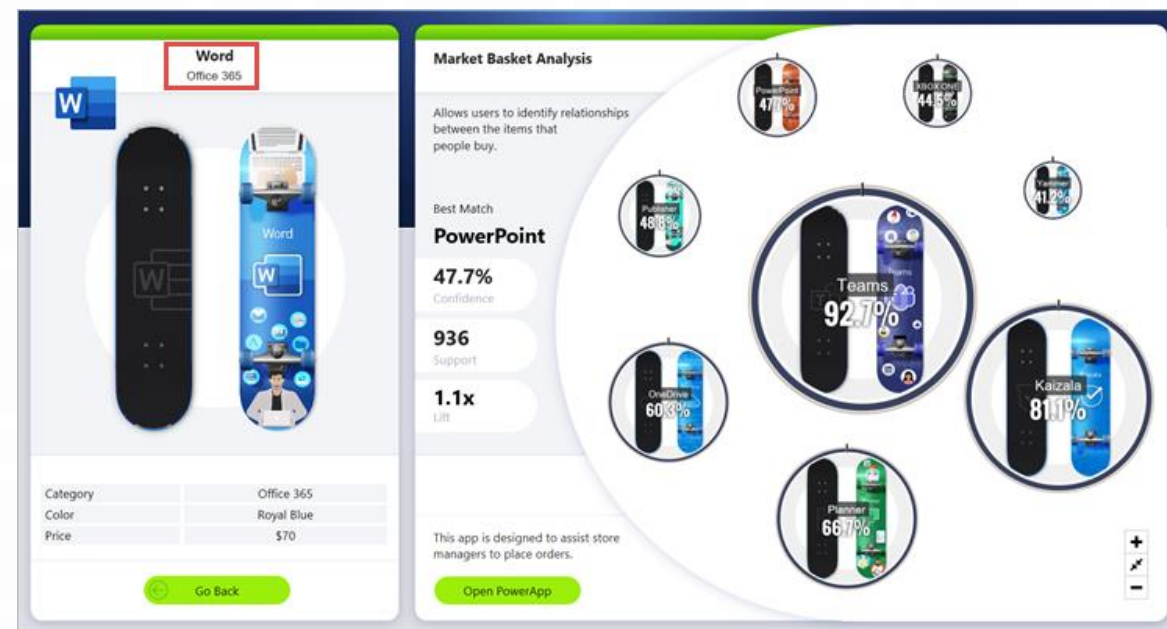
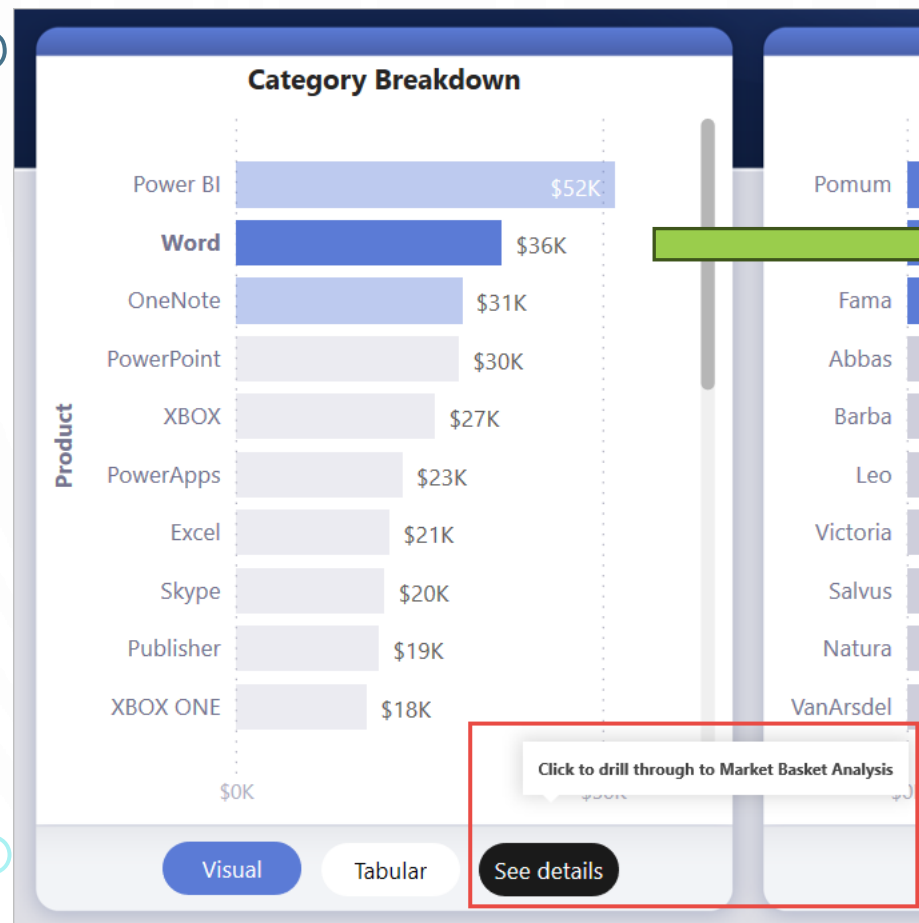


CUBOS normalmente representam um único FATO e permitem aumentar ou diminuir o nível de visualização

Operação DRILL DOWN: detalha sem alterar a medida – vendas por trimestre; ao clicar num trimestre, detalha os meses do trimestre

No power BI, Drill Through

Operação DRILL UP: operação inversa – o gráfico por meses retorna ao gráfico por trimestre



Botão de Drill Through no Power BI

DATA-WAREHOUSE E MAPREDUCE

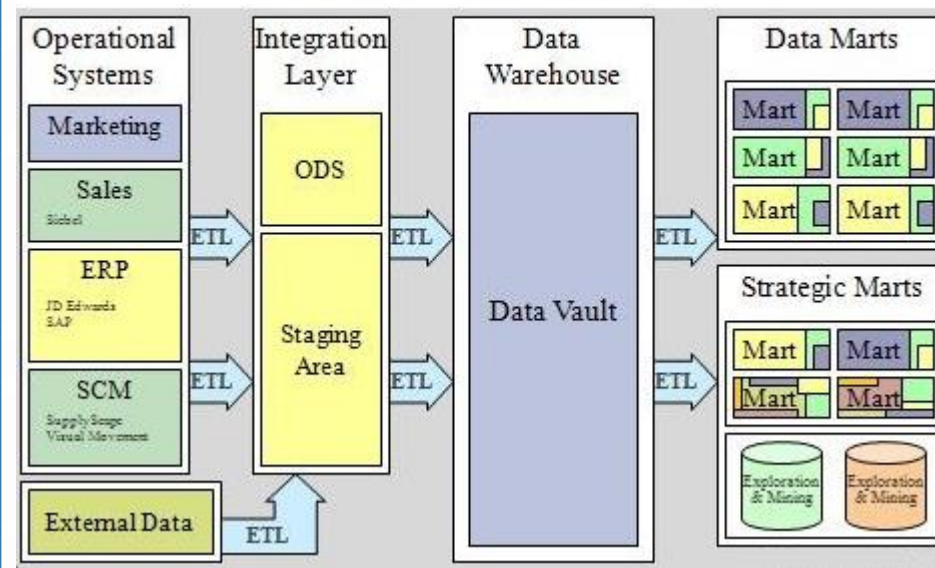
Data Warehouse: se consolidou como modelo para apoio à tomada de decisão. Na abordagem clássica, dados são Extraídos de modelos relacionais, transformados em modelos dimensionais e carregados no DW. Visto o modelo tradicional ser baseado em dado altamente estrutural, sua transformação é cara e de custo computacional

MapReduce: o MapReduce e sua implementação Hadoop foram destinados a resolver problemas específicos de análise de dados distribuído – requer infraestrutura, configuração, instalação e distribuição de software; problema a ser analisado modelado como chave-valor (de alguma forma). MapReduce é para grandes volumes de dados, que até às vezes não são suportados por DWs. Não substitui o DW tradicional, mas é alternativa para casos em que o DW não trata adequadamente

Atualidade: aplicações MapReduce são fontes de dados (ETL) para DATA MARTS, onde apenas a informação de maior valor, estruturada e tratada, é carregada para o DW.

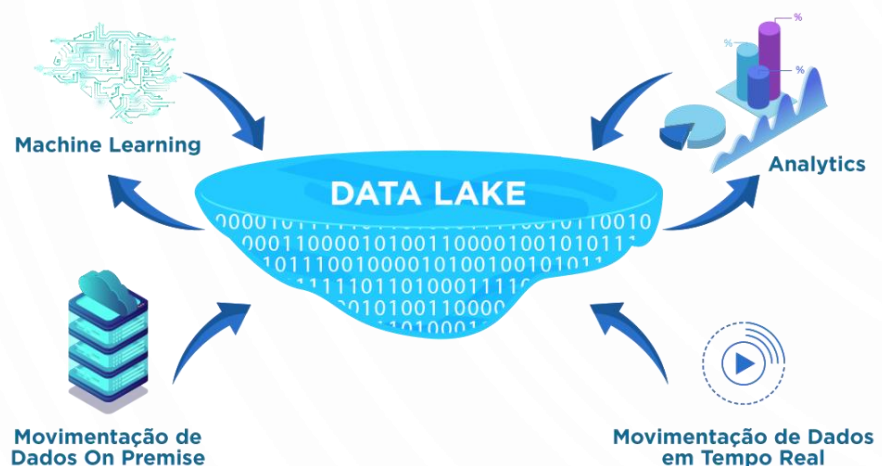
Pode incluir mapReduce, transição DW-DL

Data Warehouse



DATA-WAREHOUSE X DATA-LAKE

- **Data-Lake** (lago de dados): ideia de incorporar a modernidade dos tipos de dados não estruturados X DW surgidos no contexto relacional/dimensional – LEMBREMOS que o DW surge antes da WEB e do BIG DATA
- Esse vasto volume de dados introduziu **novos desafios** na captura, armazenamento, análise, pesquisa, compartilhamento, transferência, visualização, consulta, atualização e privacidade das informações



Argumentos:

- 1) estruturar/transformar na ingestão – possível perda de dados/impacto no desempenho
- 2) cálculos complexos em grande volume de dados recebidos -> desempenho
- 3) pode-se necessitar de dados perdidos durante a estruturação na ingestão
- 4) dificuldade de inferir, a priori, quais insights deseja-se obter dos dados

Ideia:

- 1) armazenar dado bruto, enriquecidos com metadados, descrevendo sua origem, tempo de ingestão, também podem ser particionados na ingestão
- 2) segurança normalmente na etapa de processamento e não nos dados brutos
- 3) coletar dados de diferentes fontes, usando protocolos diferentes, para armazenar dados relacionais e não relacionais, e oferecer API para análise e processamento. Flexível o suficiente para escalar para cima e para baixo.

“Data Lake pode ser definido como armazenamento centralizado, consolidado e persistente de **dados brutos (RAW), não modelados e não transformados de múltiplas fontes**, sem um esquema pré-definido explícito e sem metadados definidos externamente.”

DL: Armazenamento de dados não alterados

DW: Filtragem/Estruturação antes de armazenamento

DATA-WAREHOUSE X DATA-LAKE

DATA WAREHOUSE

VS

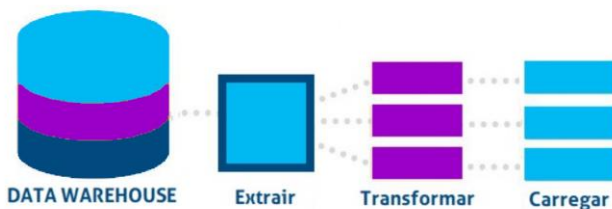
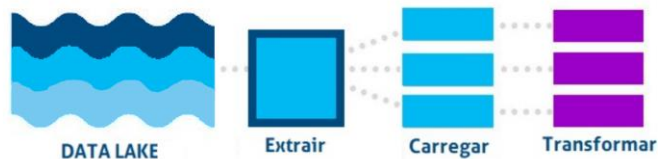
DATA LAKE

Data is processed and organized into a single schema before being put into the warehouse

Raw and unstructured data goes into a data lake

The analysis is done on the cleansed data in the warehouse

Data is selected and organized as and when needed



Data Lake

Razão para armazenar dado é indefinido

Dado é mantido raw (estruturado, semi-estruturado, não estruturado) até ser necessário

Usado por cientista de dados

Tecnologia emergente

Armazenamento de menor custo

Esquema definido após armazenamento, portanto flexível e ágil

Data Warehouse

Razão para armazenar é pré-definida

Dado é processado e está pronto para ser consultado, apenas info estruturada organizada em esquemas

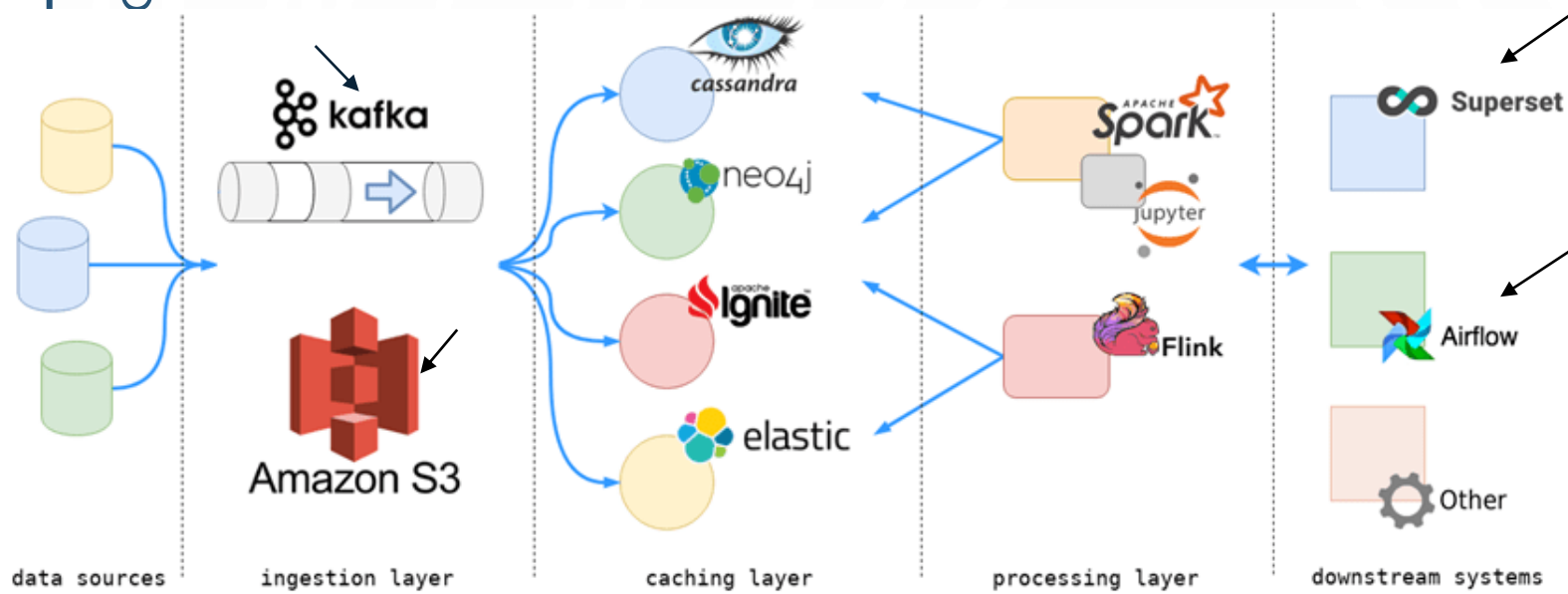
Usado por profissionais de negócios (organizacional)

Forte nível de maturidade

Armazenamento menos otimizado e com maior custo

Esquema definido previamente, provendo performance e segurança

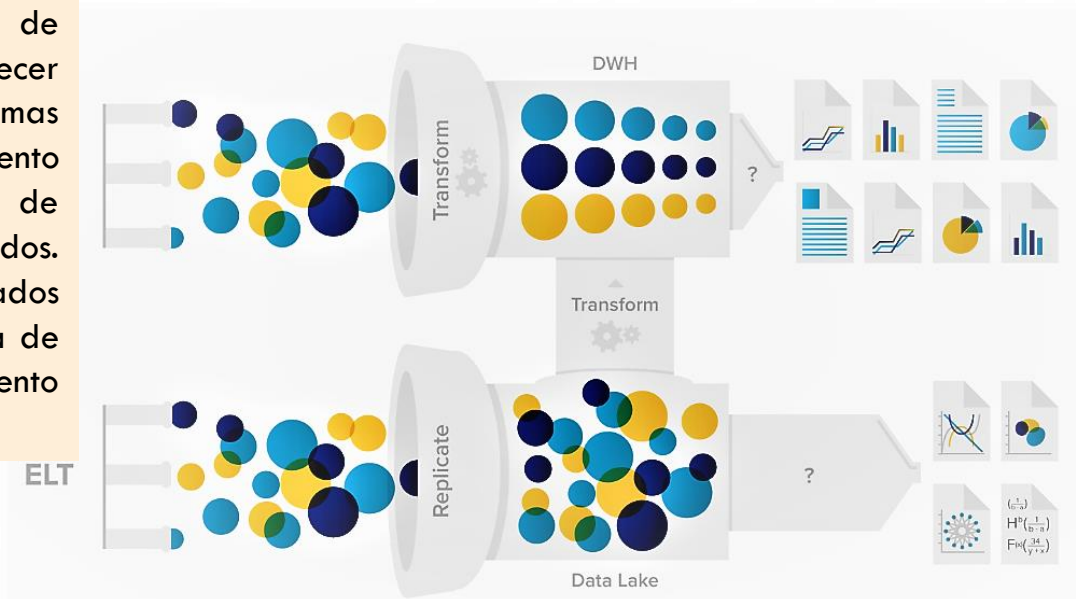
DATA-WAREHOUSE X DATA-LAKE



Ingestão: agir como coletor e armazenamento para dados brutos – pode haver enriquecimento com metadados e particionamento

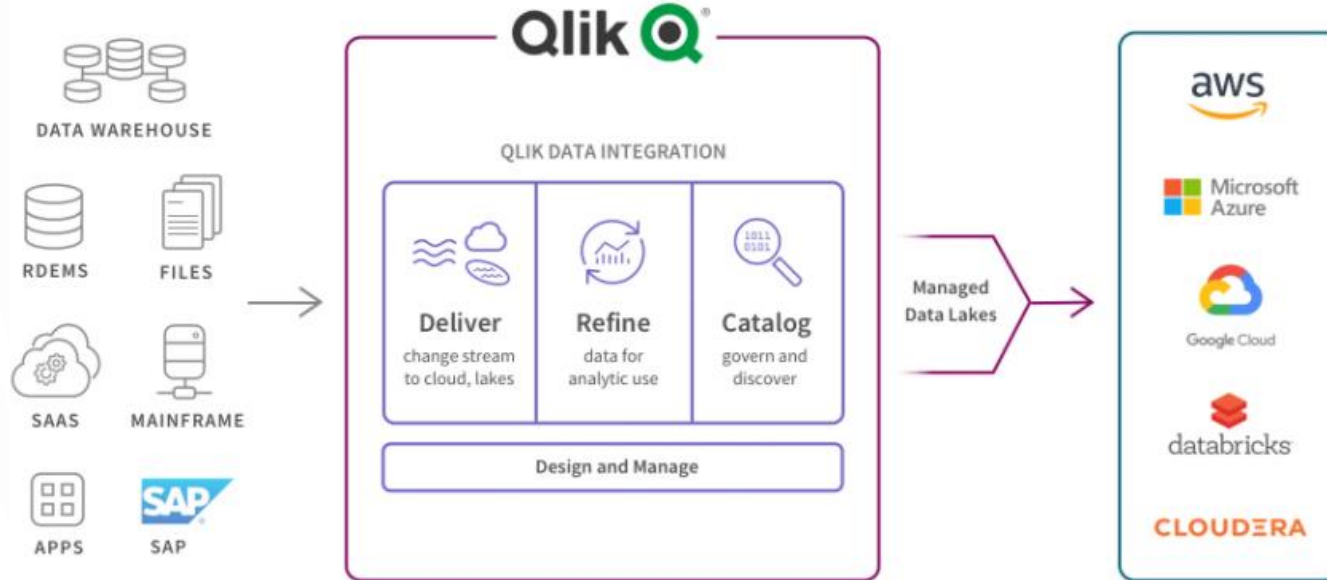
Caching (streaming): armazenar temporariamente ou permanentemente dados processados (ou pré-processados), relacionais ou não relacionais. Os dados armazenados aqui estão prontos para visualização / consumo por sistemas externos ou estão preparados para processamento adicional.

O objetivo da camada de processamento é oferecer uma ou mais plataformas para processamento distribuído e análise de grandes conjuntos de dados. Ele pode acessar dados armazenados na camada de ingestão e armazenamento em cache.



DATA-LAKE E ETL

<https://www.qlik.com/us/etl>



Compose™ gera o código ELT necessário no Apache Spark para padronizar e combinar streams de alterações em um único data lake pronto para analytics.

O Qlik Compose automatiza o projeto do warehouse, gerando código ETL, e aplicando atualizações rapidamente, enquanto aproveita as práticas recomendadas e os padrões de design comprovados. O Qlik Compose for Data Warehouses reduz drasticamente o tempo, o custo e o risco de projetos de BI, seja on-premise ou na nuvem.

Criando pipeline ETL com Spark e Python:

<https://towardsdatascience.com/create-your-first-etl-pipeline-in-apache-spark-and-python-ec3d12e2c169>