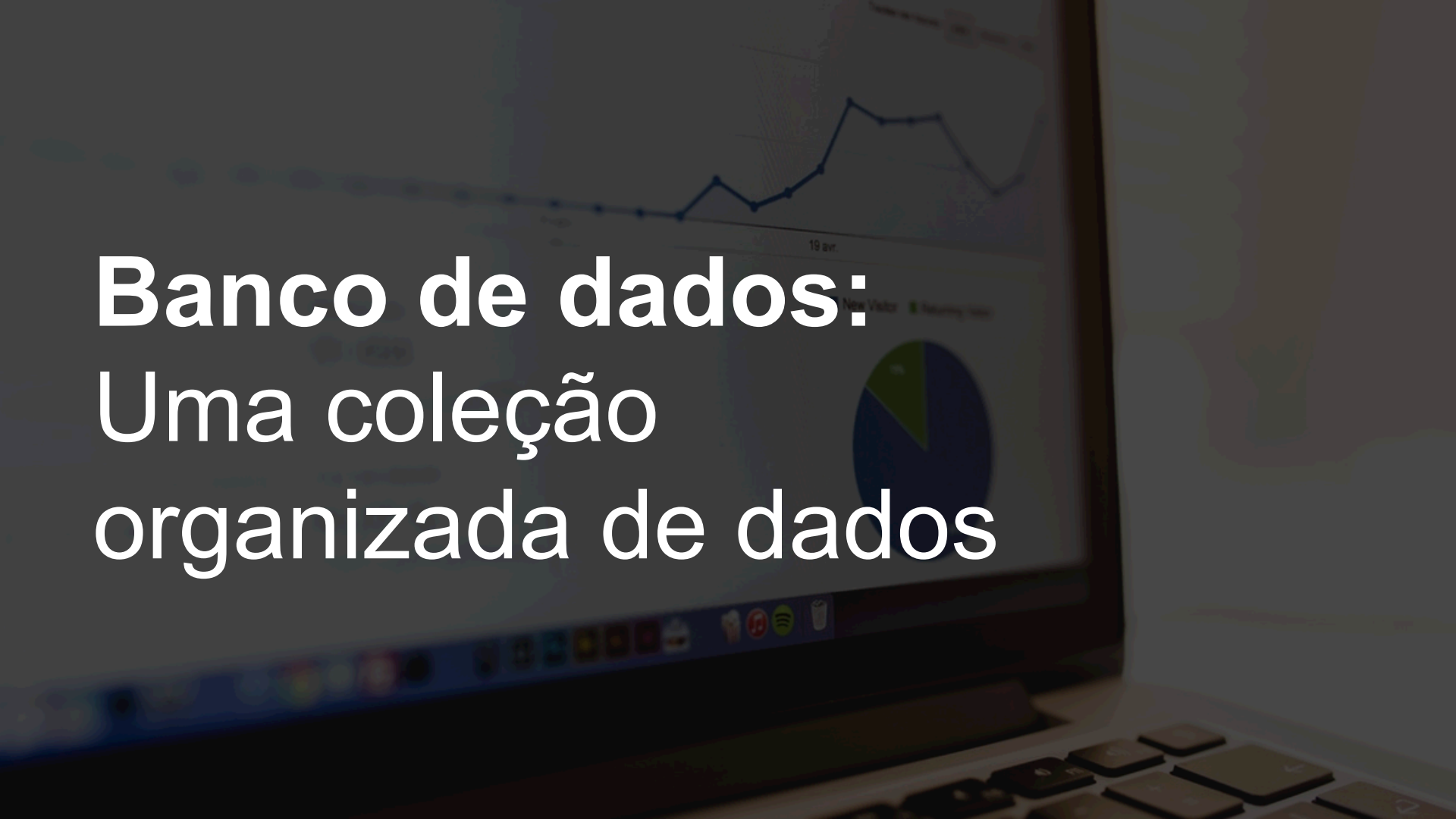
Banco de Dados

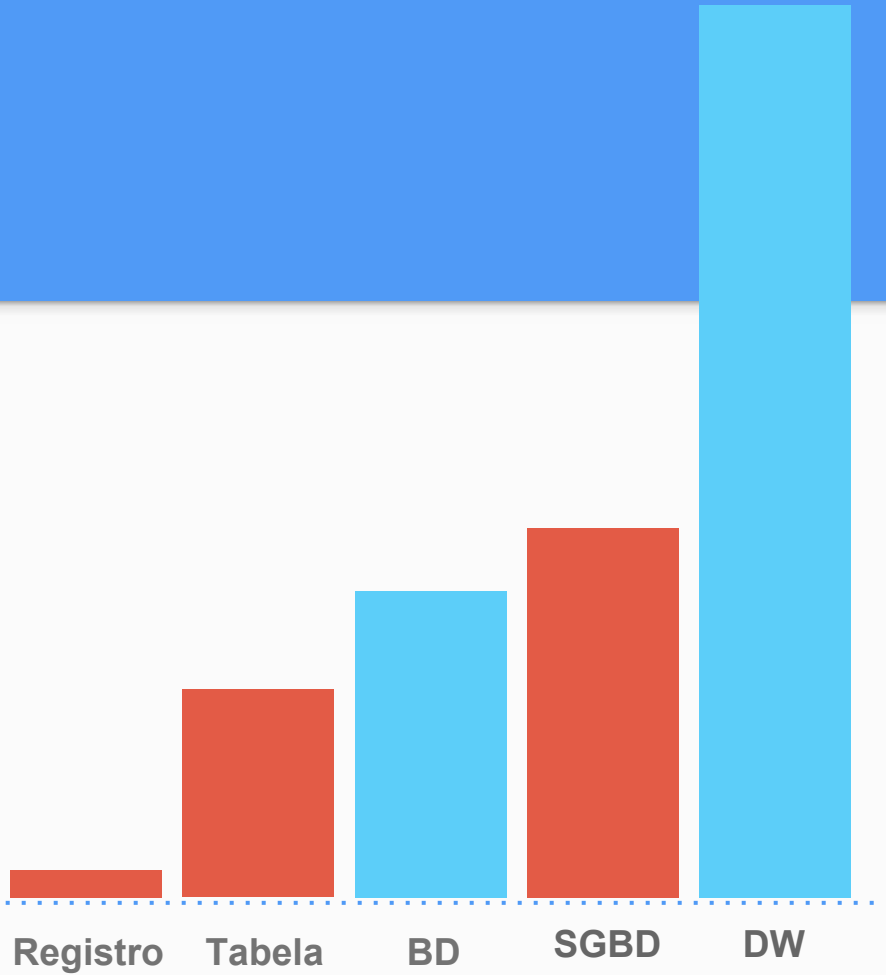
Do dado ao datalake

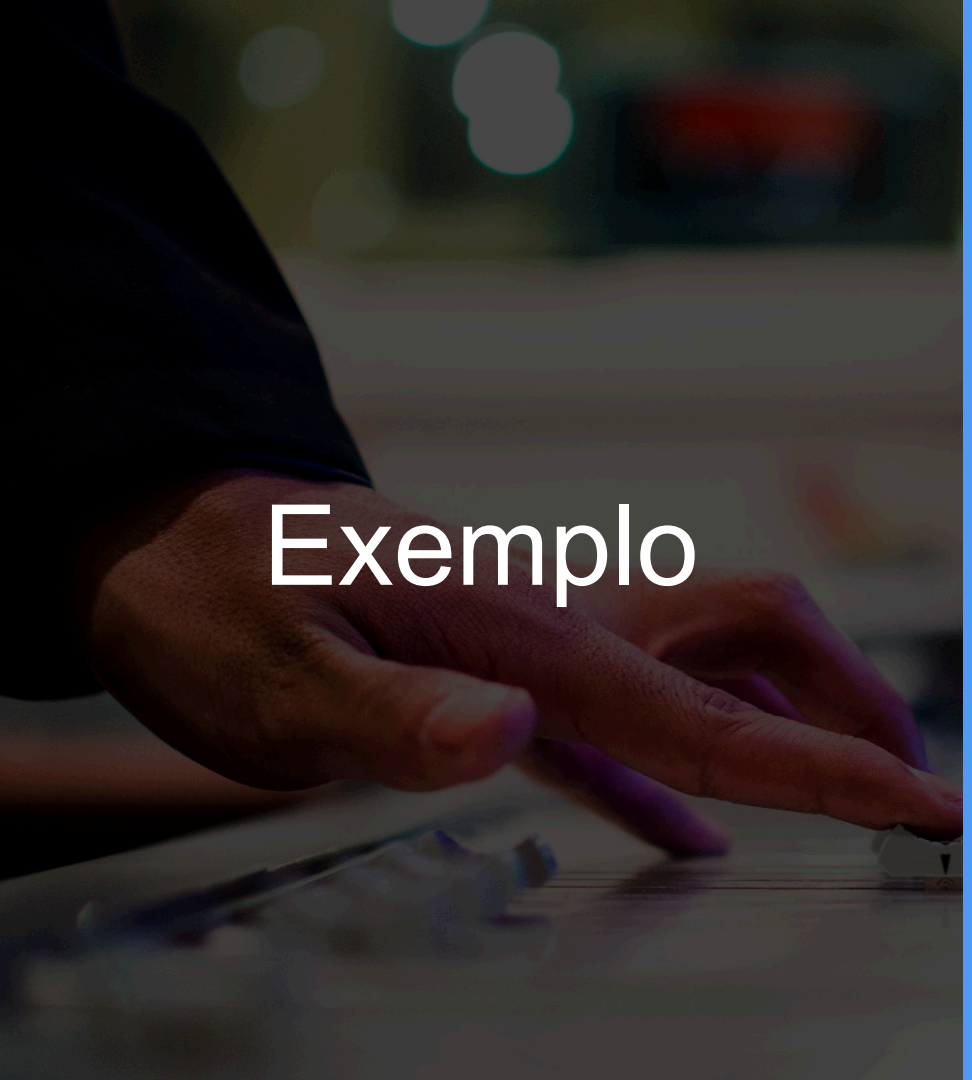


Banco de dados:
Uma coleção organizada de dados

O início

- Organização em arquivos
- Estrutura:
  - Variável
  - Registro
  - Tabela
  - Banco de dados
  - SGBD
  - Datawarehouse

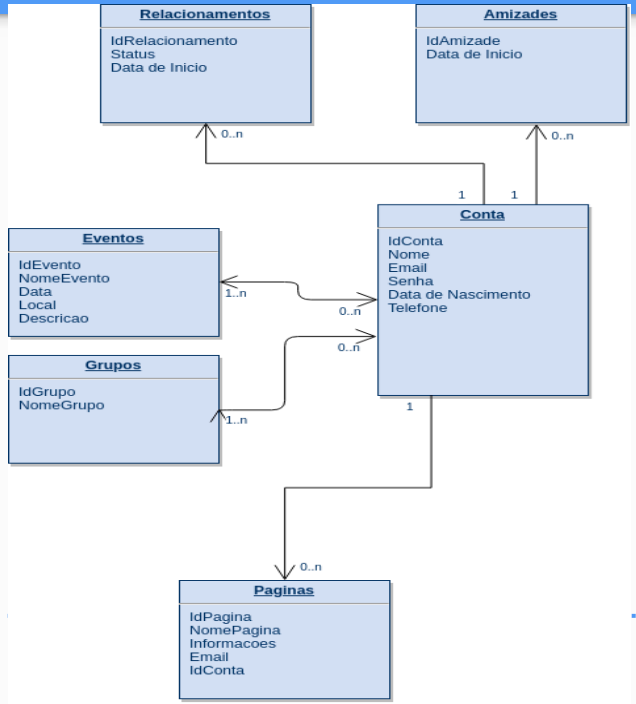




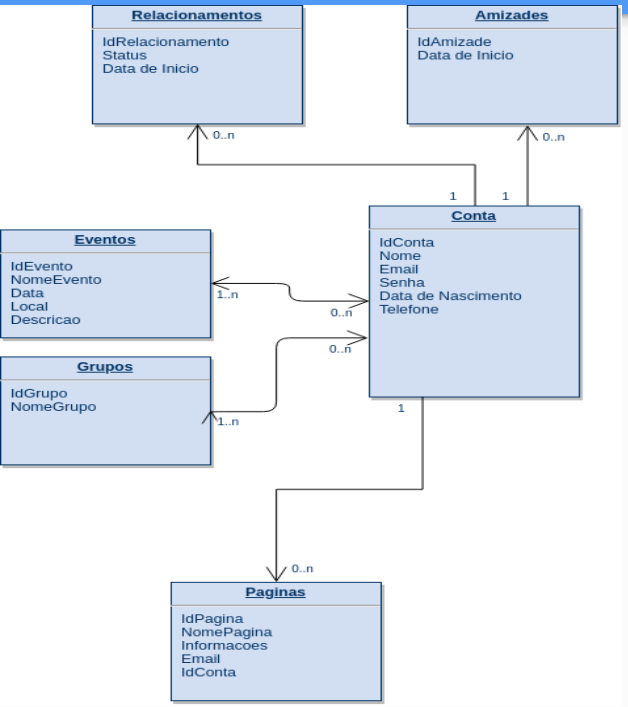
Exemplo

Vamos modelar o banco de dados do Facebook

Diagrama Entidade Relacionamento

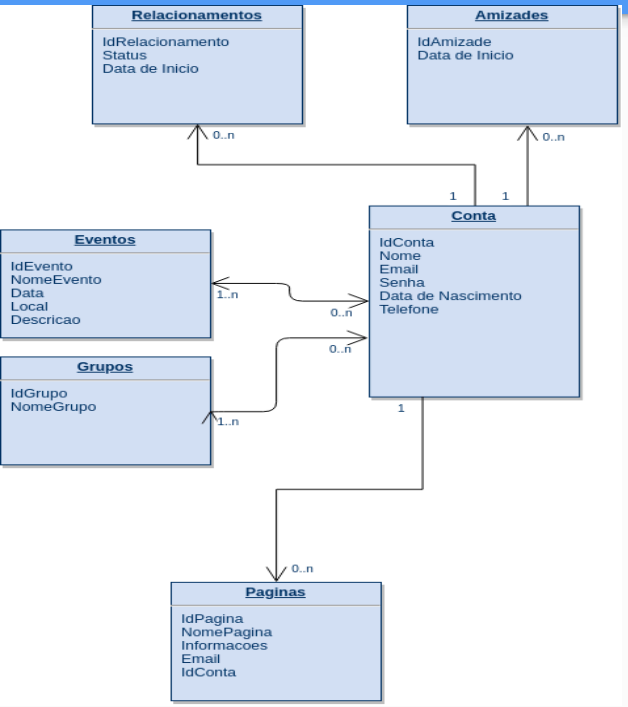


Chaves/Índices de busca



- Qual o campo que identifica unicamente minha conta?
  - Chave primária
- Qual o campo que identifica unicamente minha amizade?
  - Chave estrangeira
  - Chave composta
  - Surrogate
  - Índices de busca

Problemas de modelagem relacional



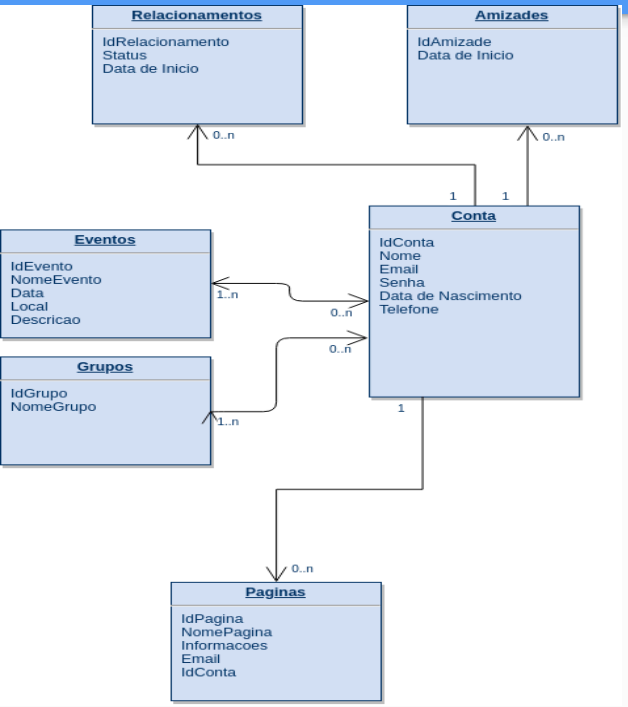
- Exemplo da tabela Amizade

idConta	amigo_idConta	data_inicio
130	157	2018-08-13
157	120	2018-08-14

Como descubro quais são os amigos da conta 157?

- Modelagem não relacional (NoSQL - Not only SQL)
- NoSQL -> GraphDatabase

Problemas de modelagem relacional



- Exemplo da tabela Eventos, coluna Descricao
  - Um campo texto puro (não estruturado)
  - Um campo JSON (semi estruturado)
    - Em alguns eventos eu posso colocar o número máximo de participantes ou o preço da entrada e em outros não

Como lidar com tabelas de estrutura variável?

- NoSQL -> Document Database

Datawarehouse

Como retirar relatórios e explorar os dados?

- O ‘Banco de Dados’ principal da empresa é conhecido como banco OLTP (On-line Transaction Processing)
- O Banco OLTP ele é otimizado para relatórios/operações de rotina. Ex:
  - Quero saber em quantos grupos ou páginas um determinado usuário está inscrito
  - De quantos Eventos em um mês, determinado usuário participou
- Percebam que com a criação de índices de busca e chaves conseguimos obter esses dados bastante diretamente

MAS

- **Se o gerente pede uma listagem de todas as pessoas que não estão em relacionamentos, que tenham mais do que 5 amigos, sendo que pelo menos 1 desses amigos frequentou a algum evento nos últimos 6 meses, na região de Ribeirão Preto?**
  - Não existe um conjunto de índices ‘pré-formatos’ para atender essa demanda
  - Imagine que ao fazer uma consulta dessas no banco em produção, você deixe o sistema da sua empresa 5 vezes mais lento por cerca de 2 dias!!!!!

O Que Fazer?

Datawarehouse

‘Cópia’ OLAP

- Podemos ‘duplicar’ o banco de dados da empresa para um modelo OLAP (On-line Analytical Processing)
- O Banco OLTP passa por um processo conhecido como ETL
- No Banco OLAP são criados ‘cubos de dados’ que visam refletir as mais diversas **dimensões** e **fatos**
  - **Dimensões:** Qualifica as informações provenientes das tabelas fatos
    - local, data, distâncias
  - **Fatos:** São as informações quantitativas dentro do DW. A fato armazena as medições necessárias para avaliar o assunto pretendido. Armazena o conteúdo histórico no DW, contendo longo período de tempo.
    - série temporal de todos os eventos que participei, contagem semanal de quantas novas amizades eu fiz

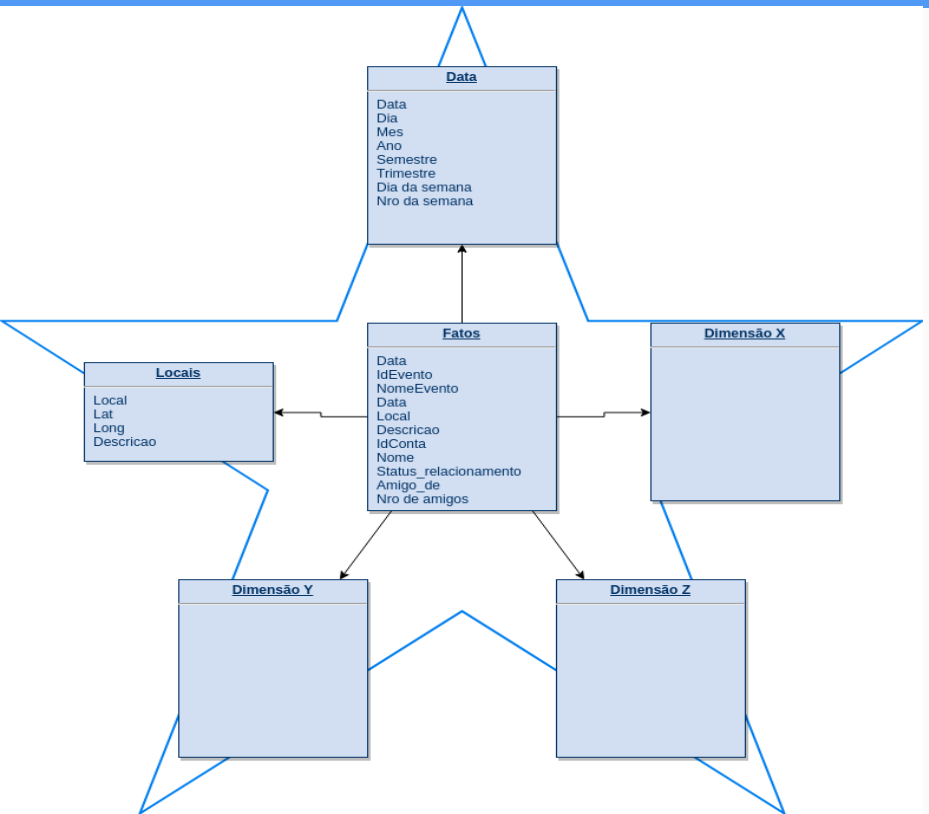
Datawarehouse

Sem ‘milagre’

- Geralmente as informações em um Datawarehouse são ‘somente leitura’, ou seja, não são apagadas e reprocessadas, a não ser que seja necessário (o processo de delete/insert é muito mais rápido que o de update - Indicadores) (EX: Receita Federal - CNPJs)
- A etapa de ETL é que fica responsável pela boa modelagem de dados que serão inseridos no Banco de dados OLAP
- Como o processo de ETL consiste basicamente de leitura sequencial do banco de dados OLTP, o mesmo praticamente não afeta o desempenho do banco em produção
- Geralmente a ferramenta ETL roda em OUTRA máquina, mas na MESMA rede do servidor OLTP.
- Os dados são armazenados no Datawarehouse, desnormalizados e em **formato estrela**

Datawarehouse

Esquema estrela



- E se agora eu quiser saber a evolução semanal do número de amizades de uma conta?
- Ou plotar em um mapa todos os eventos que as contas com mais de 100 amigos participaram

Datawarehouse

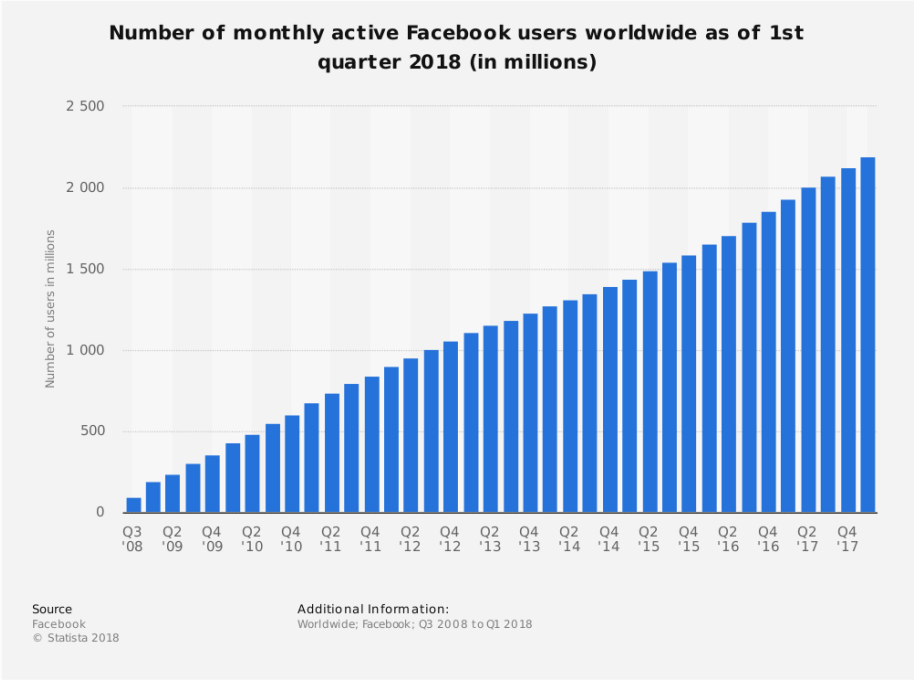
BI

- O Datawarehouse nada mais é que um banco preparado para gerar as mais diversas informações
- Possibilita conexão com diversas ferramentas de BI:
  - Qlikview
  - Tableu
  - Excel
  - Power BI
  - R/Shiny
  - Python/Pandas
- Essas ferramentas manipulam os dados, gerando relatórios e gráficos, possibilitando ao gestor insights em tomadas de decisão

Big Data

Complicando um pouquinho mais

- OK, sua empresa cresceu



- 5 V's
  - Velocidade
  - Volume
  - Variedade
  - Veracidade
  - Valor
- Os dados atingiram um patamar que apenas um servidor, por maior que seja o disco rígido, não suporta o volume de dados de uma tabela (Alguém já imaginou uma tabela com cerca de 3TBs?)
- Criou-se soluções DFS - Distributed File System (HADOOP)

Big Data

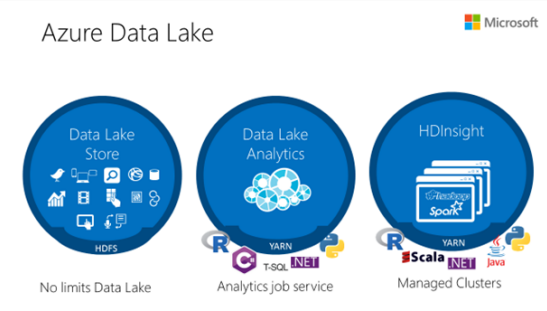
Complicando um pouquinho mais AINDA

- OK, até agora suas informações vêm de um banco de dados relacional, ou seja, ela vem completamente estruturada e podemos lidar com elas com as mesmas premissas que aprendemos com a mais simples tabela (chave, índices)
- **MAS** agora você precisa colocar informações das mais diversas:
  - Áudios do atendimento ao cliente
  - Tweets sobre a empresa (análise de sentimento)
  - Registros de sensores que monitoram a temperatura dos servidores
  - Analisar as fotos (rostos) que os usuários enviam para a plataforma
- Esse conjunto de dados não consegue ser armazenado em um simples Datawarehouse, por maior que seja (Armazenar arquivos binários como fotos e áudios em tabelas relacionais é uma péssima ideia)
- COMO JUNTAR TUDO ISSO e ainda assim ter um local de fácil acesso para encontrar e manipular esses dados??

A solução, quentinha do forno

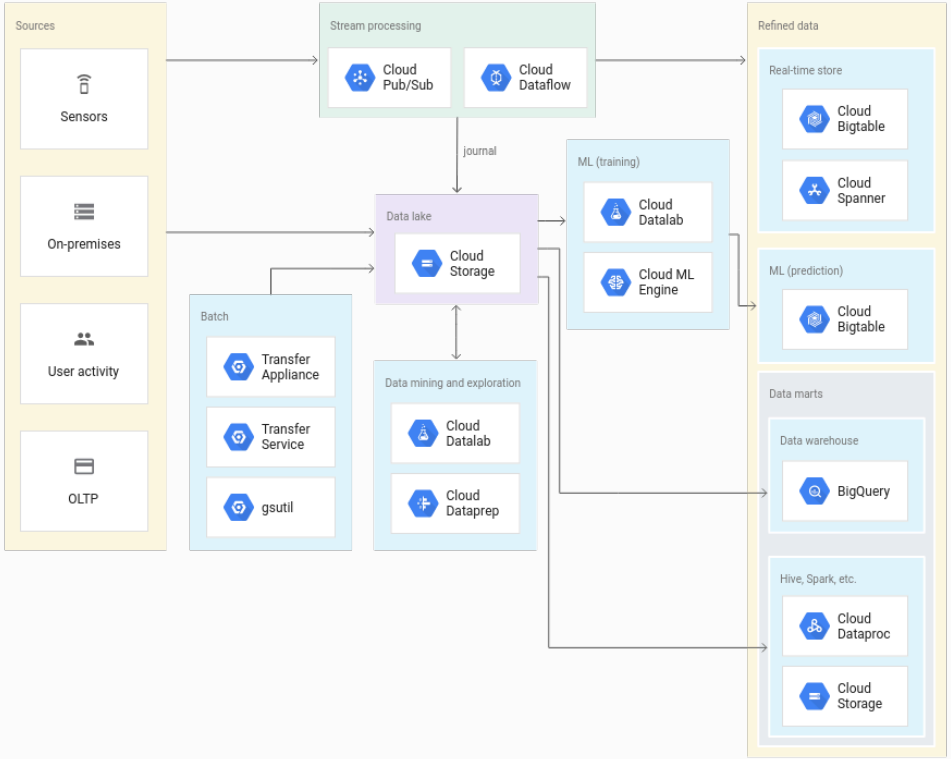
Será?

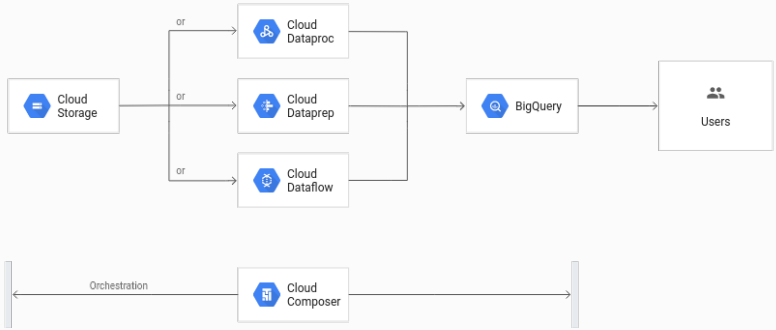
- DATALAKE, termo criado em 2015 para descrever um repositório de dados único dentro da empresa
- Repositório de baixo custo
  - US$ 0,026/Gb - Acessos frequentes
  - US$ 0,010/Gb - Acessos anuais
  - US$ 0,007/Gb - Menos de 1 acesso ao ano
- O Data Lake armazena dados brutos, sob qualquer forma do jeito que foram coletados na fonte de dados. Não há suposições sobre o esquema dos dados e cada fonte de dados pode usar qualquer esquema (Diferentemente do Datawarehouse)
- Grandes empresas como Google, Microsoft e Amazon provêm soluções de Datalake na nuvem



A solução, quentinha do forno

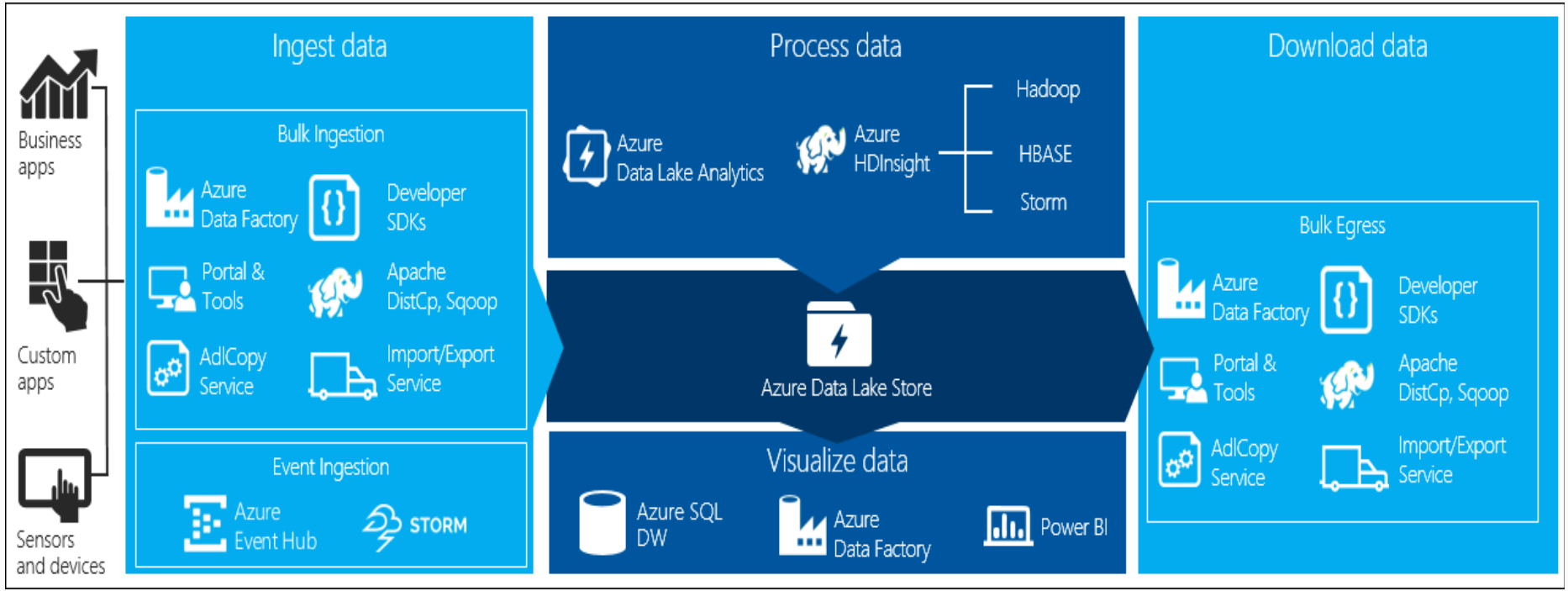
Funcionalidades do Google





A solução, quentinha do forno

Será?



Exemplos

Gedanken (www.gedanken.com.br)

- Velocidade - Dados com alta velocidade (Extract)
  - Revelagov (atualização bimestral ou trimestral dos dados)
- Variedade - Alta complexidade de processamento (Transform)
  - Cruzamento de mais de 300 bases
  - Uso de IA para detectar padrões (ou anomalias)
  - Hadoop, Spark
- Volume - Médio uso de dataware
  - Bases de dados em PostgreSQL (como dataware relacional)
  - Bases de dados em OrientDB (como dataware grafo)
  - Bases de dados em BigQuery (dataware documental)
  - Cerca de 40 TB (até julho de 2019) e crescendo

Exemplos

Criteo (https://www.criteo.com/)

- Velocidade - Dados com alta velocidade de input (Extract)
  - Cerca de 9 milhões de registros por segundo (imaginem no Excel)
  - 450 bilhões de mensagens por dia (2PB por dia)
  - A transação tem que ocorrer em até 100 ms
- Variedade - Alta complexidade de processamento (Transform)
  - Cruzamento de dados
  - Uso de IA para recomendação (Machine Learning/Visualização)
  - Redução e tratamento antes de entrar no Data Lake
- Volume - Alto uso de dataware (datacenters próprio)
  - 9 Data centers no mundo, cada um com cerca de 40 PB
  - 2 clusters Hadoop com 3000 máquinas
  - Apache Kafka (17 clusters) como Pub/Sub
  - Data Lake distribuído (EUA, Europa -França/Alemanha, China)

Exemplos

Facebook (https://www.facebook.com/)

- Velocidade - Dados com alta velocidade de input
  - 4PB por dia
  - 4 milhões de likes por minuto (2.5 bilhões dia)
  - 350 milhões de fotos por dia
  - 2 bilhões de pessoas visualizam as propagandas diariamente
- Variedade - Alta complexidade de processamento (Transform)
  - Cruzamento de dados
  - Machine Learning
  - Pattern and Face recognition
  - Redução e tratamento antes de entrar no Data Lake
- Volume - Alto uso de dataware
  - Cerca de 300 PB de dados
  - 15 Data Centers (cerca de 10000 servidores em cada - ~ 180000 servers)
  - MySQl, HBase, Cassandra, Haystack, Memcached, Hadoop

Os banco de dados já são uma realidade em qualquer empresa informatizada.
Ao administrador, cabe ter o conhecimento em como tirar valor da estrutura existente.