



Informações importantes

- Acesse a aula com o seu nome completo e com o mesmo e-mail utilizado no cadastro do ambiente de aprendizagem do IGTi. Entrou com os dados incorretos? É só sair da sala e entrar com os dados corrigidos!
- As presenças das Aulas Interativas são computadas através de uma enquete, que será realizada no ambiente de aprendizagem do IGTi. Para sua frequência ser computada, quando solicitado pelo professor, você deverá ir até a seção “Enquete de presença da Aula Interativa”, localizada na Área Acadêmica da disciplina, e responder a enquete. Essa seção estará logo abaixo da que contiver o link para a Aula Interativa em questão. A enquete ficará no ar por 10 minutos e sua nota estará disponível ao término desse tempo.
- Utilize o chat para interagir com os colegas durante a aula interativa! Em caso de dúvidas sobre o conteúdo, é só postá-las no Q&A que o tutor irá respondê-las! Lembre-se que você ainda pode esclarecer as suas dúvidas nos fóruns disponibilizados no Ambiente de Aprendizagem.

Informações importantes

- Ah! E se você não conseguir assistir a aula interativa, não se preocupe! Sua gravação ficará disponível no Área Acadêmica, juntamente com os slides utilizados pelo professor, em até 24 horas úteis após o término da aula. Você também poderá realizar a atividade de reposição para recuperar os pontos de presença!
- Se você precisar solicitar prorrogação e/ou 2ª oportunidade para entrega de atividades, saiba que isso é realizado somente mediante a apresentação de atestado médico ou de óbito de parentes de 1º grau.
- Para melhor experiência nas aulas interativas, sugerimos que você baixe o aplicativo do Zoom no seu computador.

Aplicações em ETL

SEGUNDA AULA INTERATIVA

PROF. RICARDO BRITO ALVES

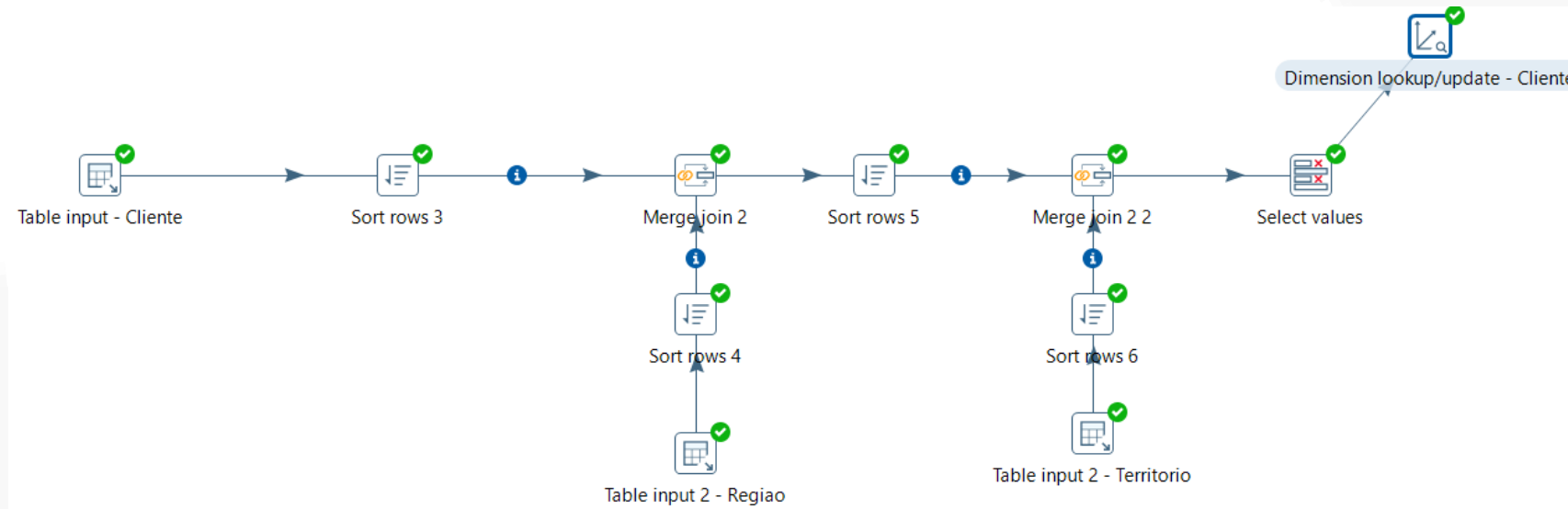
Nesta aula



- ☐ Desafio.
- ☐ Tópicos da Disciplina e Temas Interessantes.

Desafio

Desafio



Tópicos da disciplina e temas interessantes

Cientista de Dados



Cientistas de Dados *recebem uma enorme massa de dados (estruturados e não estruturados)* e usam suas habilidades em Matemática, Estatística, Ciência da Computação e Programação para *limpar, tratar e organizar os dados.*

Em seguida, eles aplicam suas capacidades analíticas – Machine Learning, Inteligência Artificial, conhecimento de negócio, ceticismo de suposições existentes – *para descobrir soluções para os desafios de negócios.*

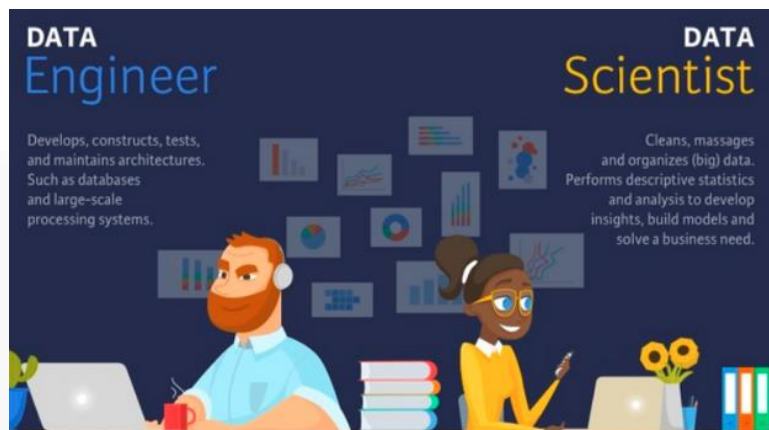


Engenheiro de Dados



Um Engenheiro de Dados é o profissional dedicado ao ***desenvolvimento, construção, teste e manutenção de arquiteturas, como um sistema de processamento em grande escala.***

O Engenheiro de Dados é responsável por criar o pipeline dos dados, ***desde a coleta até a entrega*** para análise ou para alimentar um produto ou serviço baseado em análise preditiva já em produção.



Engenheiro de Dados x DBA

*As atividades desempenhadas pelo Engenheiro de Dados englobam principalmente a já conhecida rotina de um DBA, porém acrescidas de muitas outras tarefas, tais como: **manutenção de sistemas de banco de dados relacionais e não-relacionais, ETL (Extração, Transformação e Carga), soluções de Data-Warehouse e Datamart, modelagem de dados e armazenamento em nuvem.***

Além disso, diversas tecnologias fazem parte do dia a dia desse profissional, tais como: Oracle, MSSQL, MySQL, PostgreSQL, Neo4J, MongoDB, Cassandra, Sqoop, HDFS, Hive e muitas outras.



Data Driven

O Data Driven se baseia no uso de ferramentas tecnológicas capazes de coletar e analisar dados diferentes da sua empresa.

Esses dados, por sua vez, podem ser compilados por meio de BI ou inteligência artificial e ajudam o gestor a ter uma ideia mais precisa do seu negócio, facilitando a tomada de decisão estratégica.

Data Driven



Assim, a gestão Data Driven é bem diferente dos modelos tradicionais, nos quais a tomada de decisão, geralmente, se baseava na intuição do dono ou nos “palpites” de especialistas, sem que houvesse dados reais para embasar essas atitudes.

A maior dificuldade dos gestores em implementar a gestão Data Driven é, justamente, compreender a sua importância. Afinal, muitos acreditam que a tomada de decisão estratégica não faz tanta diferença nos resultados finais.



Como adotar o Data Driven?

O principal objetivo do Data Driven **é entregar respostas mais precisas e confiáveis por meio de dados.**

Por isso, o primeiro passo para implementar essa ideia é modificar a postura dos gestores e colaboradores, de forma que eles **passem a reduzir os “achismos” e entendam o valor desses dados.**

- Transforme a sua cultura.
- Use boas soluções.
- Aprenda a entender os dados.
- Use os indicadores de performance.

Data Driven



Data Driven: como me tornar um profissional orientado por dados?

Engenheiro de dados:

É a base de todo o trabalho, aquele que opera por trás das cortinas. O engenheiro de dados é quem vai reunir todas as informações e criar uma estrutura tanto para armazenagem quanto para a apresentação dos dados coletados.

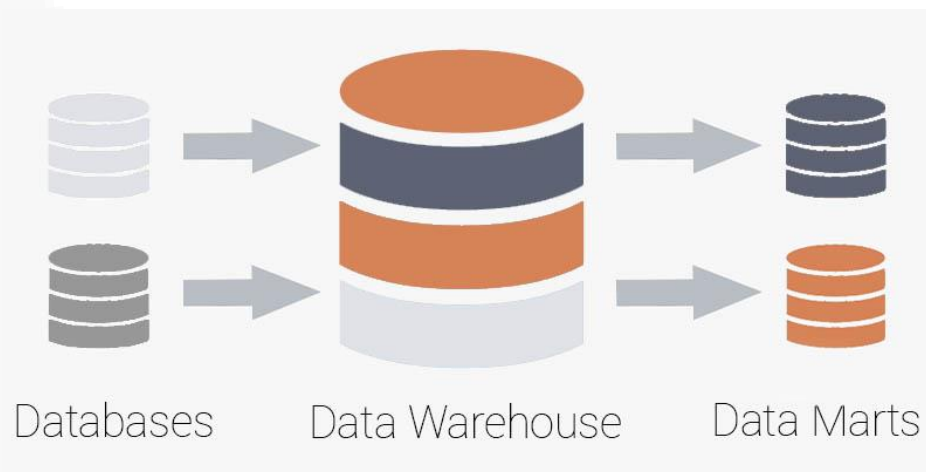
Arquiteto de dados:

Mais próxima dos DBAs (Administrador de Banco de Dados) que desejam ingressar no universo Data Driven. Além de conseguir administrar e montar um banco de dados, também espera que esse profissional seja capaz de ter uma visão sobre a estruturação em sincronismo com a organização dos dados.



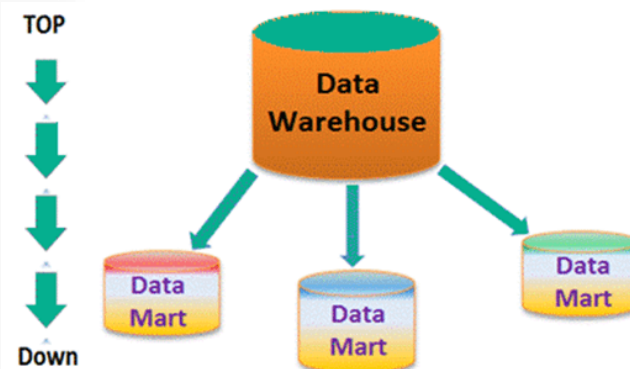
Data Warehouse (DW)

Data Warehouse é um ***depósito de dados digitais que serve para armazenar informações detalhadas relativamente a uma empresa, criando e organizando relatórios através de históricos***, que são depois usados pela empresa para ajudar a tomar decisões importantes com base nos fatos apresentados.



Data Mart

- Um Data Mart é uma **subdivisão ou subconjunto de um DW**. Os Data Marts são como pequenas fatias que armazenam subconjuntos de dados, normalmente organizados para um departamento ou um processo de negócio.
- Normalmente o Data Mart **é direcionado para uma linha de negócios** ou equipe, sendo que a sua informação costuma pertencer a um único departamento.



OLAP (Online Analytical Processing)



OLAP (Online Analytical Processing ou Processo Analítico em Tempo Real) é uma das ferramentas mais usadas para a exploração de um Data Warehouse. O OLAP possibilita alterar e analisar grandes quantidades de dados em várias perspectivas diferentes. As funções básicas do OLAP são:

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



Definição ETL



Extract – Extrair.

Transform – Transformar.

Load – Carregar.

Definição ELT

Extract – Extrair.

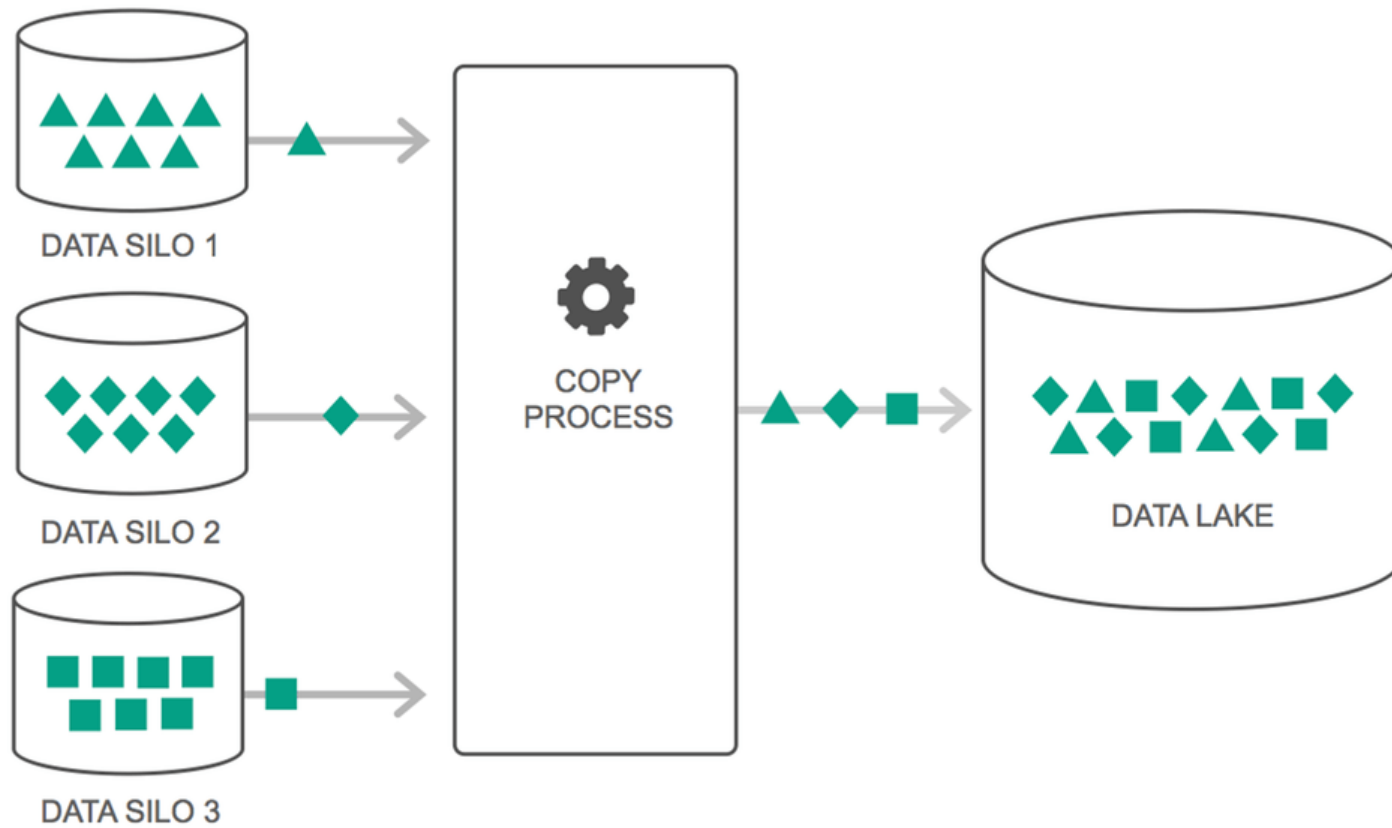
Load – Carregar.

Transform – Transformar.

O **ELT** é um processo de dados usado para replicar dados de uma fonte para um banco de dados de destino, sendo uma evolução ETL. Isso porque torna o processo de replicação de dados muito menos complexo, uma vez que o passo de transformação é realizado após os dados estarem no destino.



Data Lake



Data Lake



	Data Warehouse	Data Lake
Dados	<ul style="list-style-type: none">· Estruturados· Processados	<ul style="list-style-type: none">· Estruturados / Semi-estruturados / Não estruturados· Não processados (em estado bruto)
Processamento	<ul style="list-style-type: none">· Esquema de dados gerado no momento da escrita	<ul style="list-style-type: none">· Esquema de dados gerado no momento da leitura
Armazenamento	<ul style="list-style-type: none">· Alto custo para alto volume de dados	<ul style="list-style-type: none">· Criado para ser de baixo custo, independente do volume de dados
Agilidade	<ul style="list-style-type: none">· Pouco ágil, configuração fixa	<ul style="list-style-type: none">· Bastante ágil, pode ser configurado e reconfigurado conforme necessário
Segurança	<ul style="list-style-type: none">· Estratégias de segurança bastante maduras	<ul style="list-style-type: none">· Ainda precisa aperfeiçoar o modelo de segurança e acesso aos dados
Usuários	<ul style="list-style-type: none">· Analistas de Negócios	<ul style="list-style-type: none">· Cientistas e Analistas de Dados

Alguns SGBDs

IGTI



NoSQL

- NoSQL (originalmente se referindo a "no SQL": "não SQL" ou "não relacional", posteriormente estendido para ***Not Only SQL*** - Não Somente SQL) é um termo genérico que representa os bancos de dados não relacionais.
- Bancos de dados NoSQL são cada vez mais usados em ***Big Data*** e ***aplicações web de tempo real***.



New

IGTi

SQL

ScaleArc

AKIBAN

ScaleDB

SCHOONER

JustOneDB

NUO DB

SQLFire

TRANSATTICE

VoltDB

GENIE DB

H-Store

xeround

Clustrix

NewSQL



Os bancos de dados NewSQL buscam ***promover a mesma melhoria de desempenho e escalabilidade dos sistemas NoSQL, não abrindo mão dos benefícios dos bancos de dados tradicionais, da linguagem SQL e das propriedades ACID.***

A vantagem dos bancos de dados NewSQL é ***proporcionar consultas em tempo real, além de maior capacidade de processamento.*** Há um custo grande nos bancos NoSQL em não usar SQL, sendo exigido trabalho excessivo dos desenvolvedores para compensar sua ausência.

NewSQL



- ✓ **MemSQL:** como o próprio nome sugere, é operado em memória e é um sistema de banco de dados de alta escala por sua combinação de desempenho e compatibilidade com o SQL transacional e ACID na memória, adicionando uma interface relacional em uma camada de dados in-memory.
- ✓ **VoltDB:** projetado por vários pesquisadores de sistema de banco de dados bem conhecidos, esse banco oferece a velocidade e a alta escalabilidade dos bancos de dados NoSQL, mas com garantias ACID e sua latência em milissegundo e integração com Hadoop.
- ✓ **SQLFire:** servidor de banco de dados NewSQL da VMware, desenvolvido para escalar em plataformas nas nuvens e tomar as vantagens de infraestrutura virtualizadas.
- ✓ **MariaDB:** foi desenvolvido pelo criador do MySQL e é totalmente compatível com o MySQL. Também pode interagir com os bancos de dados NoSQL, como Cassandra e LevelDB.

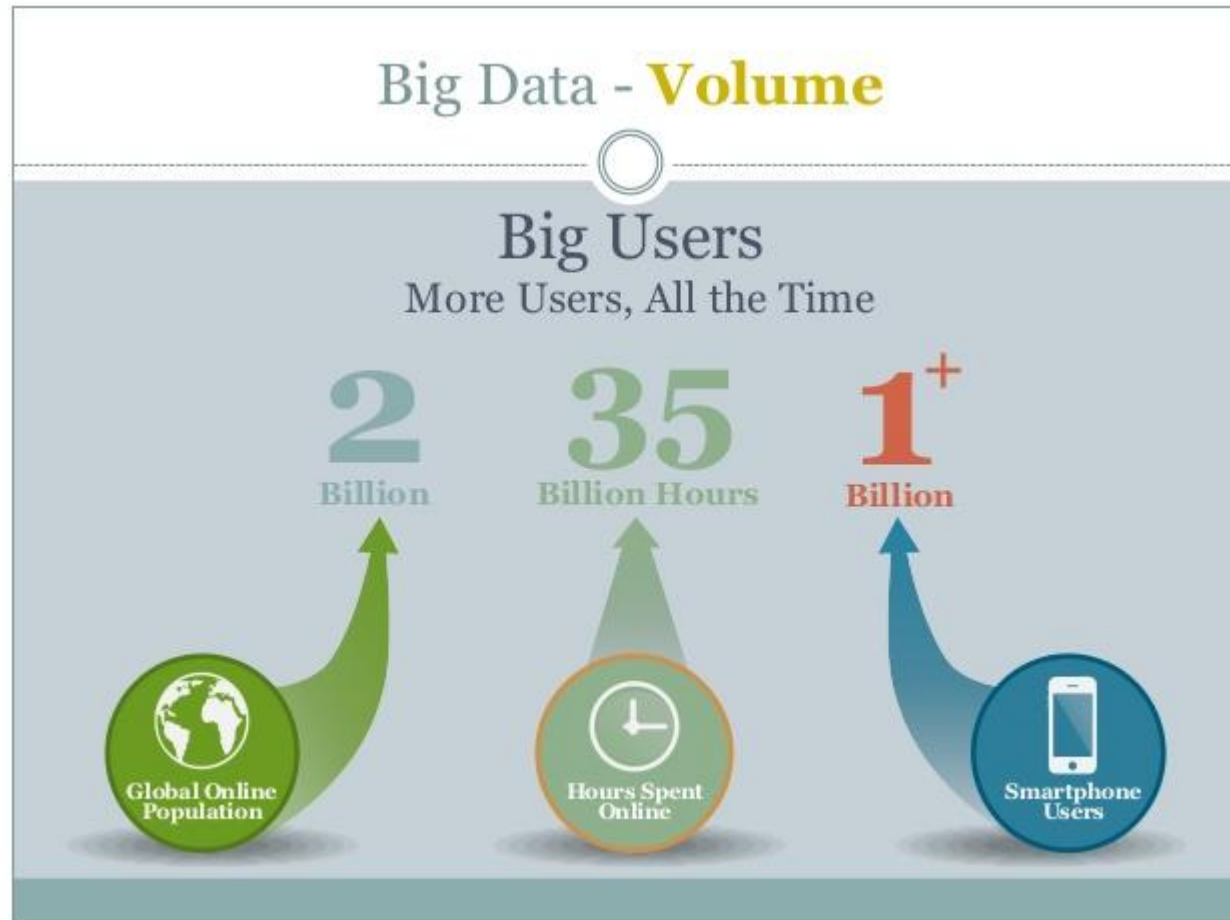
Por que NoSQL?

Hoje as empresas estão adotando NoSQL para um número crescente de casos de uso. A escolha que é impulsionada por quatro megatendências inter-relacionadas:

- Big Users.
- Big Data.
- Internet das Coisas.
- Cloud Computing.

Big Users

iGTi



Big Data



- Qual é o tamanho da web?



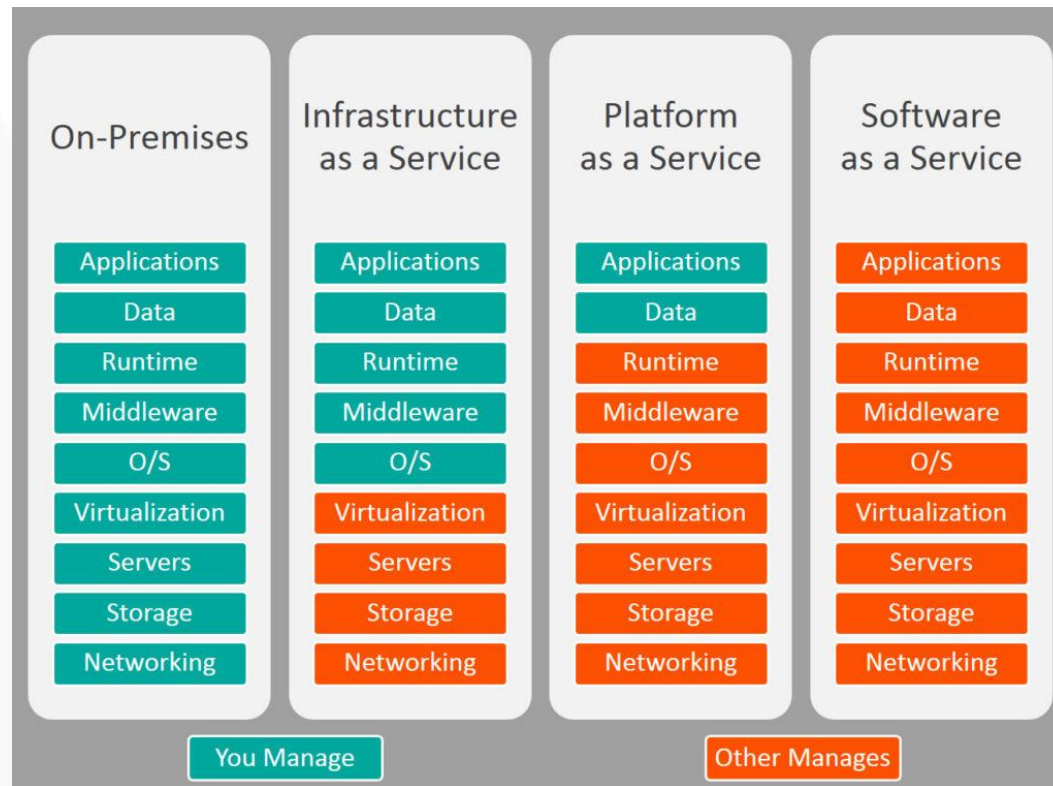
Fonte: Google.com

IoT - Internet of Things

IGTi



Cloud Computing



AWS – EC2



O Amazon **Elastic Compute Cloud** (Amazon EC2) oferece uma capacidade de computação escalável na Nuvem da Amazon Web Services (AWS).

O uso do Amazon EC2 **elimina a necessidade de investir em hardware inicialmente**, portanto, você pode desenvolver e implantar aplicativos com mais rapidez.

Você **pode usar o Amazon EC2 para executar o número de servidores virtuais que precisar**, configurar a segurança e a rede, e gerenciar o armazenamento.

O Amazon EC2 também permite a expansão ou a redução para gerenciar as alterações de requisitos ou picos de popularidade, reduzindo, assim, a sua necessidade de prever o tráfego do servidor.



AWS – S3



O Amazon ***Simple Storage Service*** é armazenamento para a Internet. Ele foi projetado para facilitar a computação de escala na web para os desenvolvedores.

O Amazon S3 tem uma interface simples de serviços da web que você pode usar para armazenar e recuperar qualquer quantidade de dados, a qualquer momento, em qualquer lugar da web.

Ela concede acesso a todos os desenvolvedores para a mesma infraestrutura altamente dimensionável, confiável, segura, rápida e econômica que a Amazon utiliza para rodar a sua própria rede global de sites da web.

O serviço visa maximizar os benefícios de escala e poder passar esses benefícios para os desenvolvedores.



AWS – RDS



O Amazon ***Relational Database Service*** (Amazon RDS) ***facilita a configuração, a operação e a escalabilidade de bancos de dados relacionais na nuvem.***

O serviço ***oferece capacidade econômica e redimensionável e automatiza tarefas demoradas de administração, como provisionamento de hardware, configuração de bancos de dados, aplicação de patches e backups.***

Dessa forma, você pode se concentrar na performance rápida, alta disponibilidade, segurança e conformidade que os aplicativos precisam.



BDaaS - Banco de Dados Como Serviço



Virtualização: permite que banco de dados seja instalado em uma máquina virtual.

DBaaS: fornece uma plataforma flexível escalável, sob demanda, que está orientada para o autosserviço e gerenciamento fácil, particularmente em termos de provisionamento de um negócio no próprio ambiente.



Ranking Banco de Dados



<https://db-engines.com/en/ranking>

Rank			DBMS	Database Model	Score		
Aug 2020	Jul 2020	Aug 2019			Aug 2020	Jul 2020	Aug 2019
1.	1.	1.	Oracle +	Relational, Multi-model i	1355.16	+14.90	+15.68
2.	2.	2.	MySQL +	Relational, Multi-model i	1261.57	-6.93	+7.89
3.	3.	3.	Microsoft SQL Server +	Relational, Multi-model i	1075.87	+16.15	-17.30
4.	4.	4.	PostgreSQL +	Relational, Multi-model i	536.77	+9.76	+55.43
5.	5.	5.	MongoDB +	Document, Multi-model i	443.56	+0.08	+38.99
6.	6.	6.	IBM Db2 +	Relational, Multi-model i	162.45	-0.72	-10.50
7.	↑ 8.	↑ 8.	Redis +	Key-value, Multi-model i	152.87	+2.83	+8.79
8.	↓ 7.	↓ 7.	Elasticsearch +	Search engine, Multi-model i	152.32	+0.73	+3.23
9.	9.	↑ 11.	SQLite +	Relational	126.82	-0.64	+4.10
10.	↑ 11.	↓ 9.	Microsoft Access	Relational	119.86	+3.32	-15.47
11.	↓ 10.	↓ 10.	Cassandra +	Wide column	119.84	-1.25	-5.37
12.	12.	↑ 13.	MariaDB +	Relational, Multi-model i	90.92	-0.21	+5.96
13.	13.	↓ 12.	Splunk	Search engine	89.91	+1.64	+4.03
14.	↑ 15.	↑ 15.	Teradata +	Relational, Multi-model i	76.78	+0.81	+0.14
15.	↓ 14.	↓ 14.	Hive	Relational	75.29	-1.14	-6.51

NoSQL



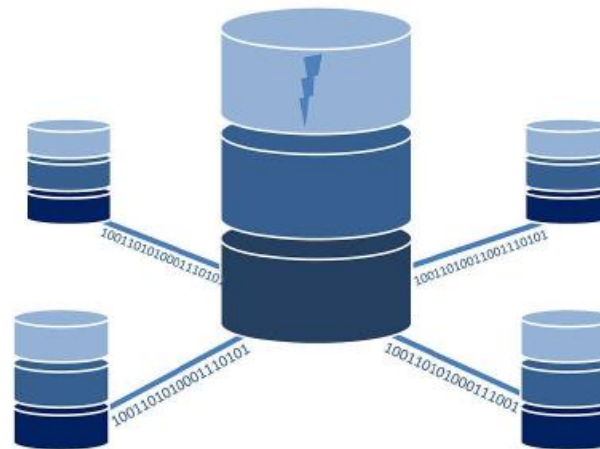
- NoSQL (originalmente se referindo a "no SQL": "não SQL" ou "não relacional", posteriormente estendido para Not Only SQL - Não Somente SQL) é um termo genérico que representa os bancos de dados não relacionais.
- Bancos de dados NoSQL são cada vez mais usados em big data e aplicações web de tempo real.



NoSQL – Banco de Dados de Documentos



- Armazenam chave/valor.
- O valor é um documento estruturado e indexado, com metadados.
- Valor (documento), pode ser consultado.
- JSON: JavaScript Object Notation.
 - Feito para troca de dados.
 - Mais compacto e legível que XML.



NoSQL - JSON



Estrutura nome/valor, entre aspas duplas



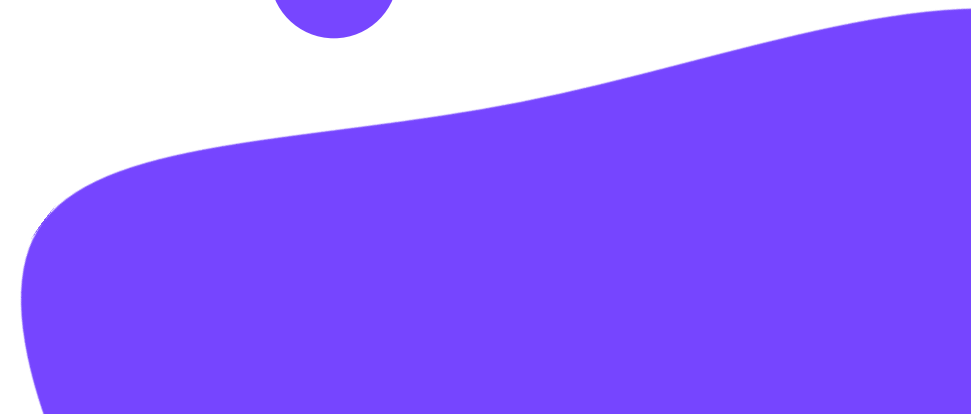
Dados separados por vírgula



Chaves separam objetos



Vetores entre colchetes



MongoDB



IGTI

- ✓ Open source.
- ✓ Multiplataforma.
- ✓ Escalável.



Relacional

Banco de Dados

Tabela

Linha

Coluna

MongoDB

Banco de Dados

Coleção

Documento

Campo

Ensaio no Pentaho



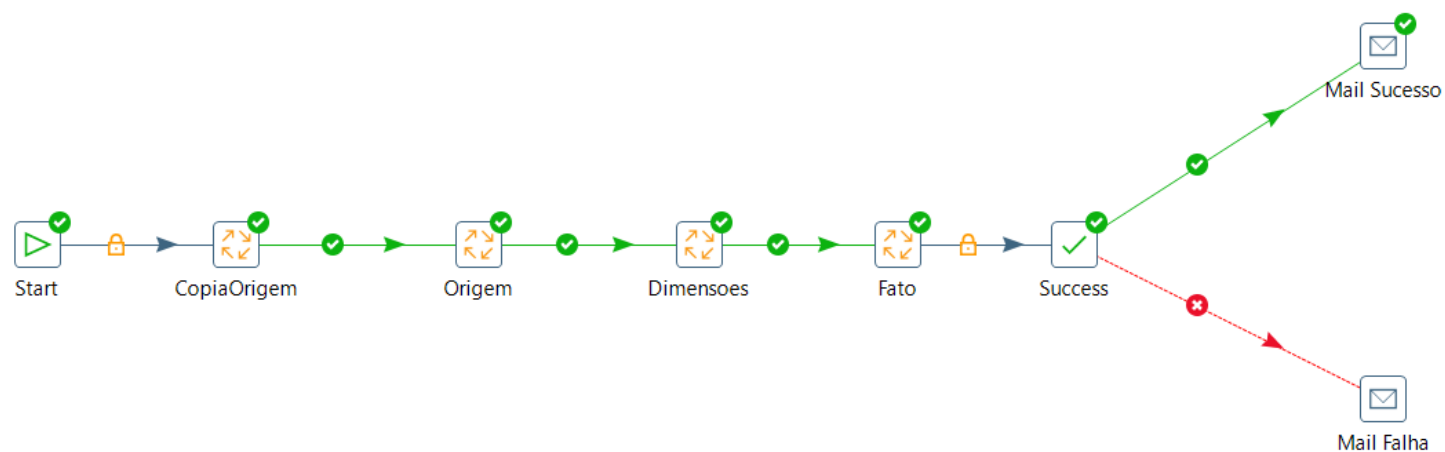
Trabalhando com a passagem de parâmetros e Jobs.



Ensaio no Pentaho



Trabalhando com a Jobs para o Desafio.



E-mail



Gmail



Início



Informações pessoais



Dados e personalização



Segurança



Pessoas e compartilhamento



Pagamentos e assinaturas

Bem-vindo, I

Gerencie suas informações, privacidade e seguran

Privacidade e personali- zação







Veja os dados na sua Conta do Google e escolha qual atividade será salva para personalizar sua experiência no Google



[Gerenciar seus dados e a personalização](#)

E-mail



-  Início
-  Informações pessoais
-  Dados e personalização
-  **Segurança**
-  Pessoas e compartilhamento
-  Pagamentos e assinaturas

 Encontrar um dispositivo perdido

[Gerenciar dispositivos](#)

[Gerenciar acesso](#)

Acesso a app menos seguro

Sua conta está vulnerável porque você permite que apps e dispositivos que usam tecnologias de login menos seguras a acessem. Para manter sua conta segura, o Google desativará essa configuração automaticamente se ela não estiver sendo usada. [Saiba mais](#)



 Ativado

[Desativar o acesso \(recomendado\)](#)

Fazer login em outros sites



Fazer login com o Google

Você usa sua Conta do Google para fazer login em 2 sites e apps

