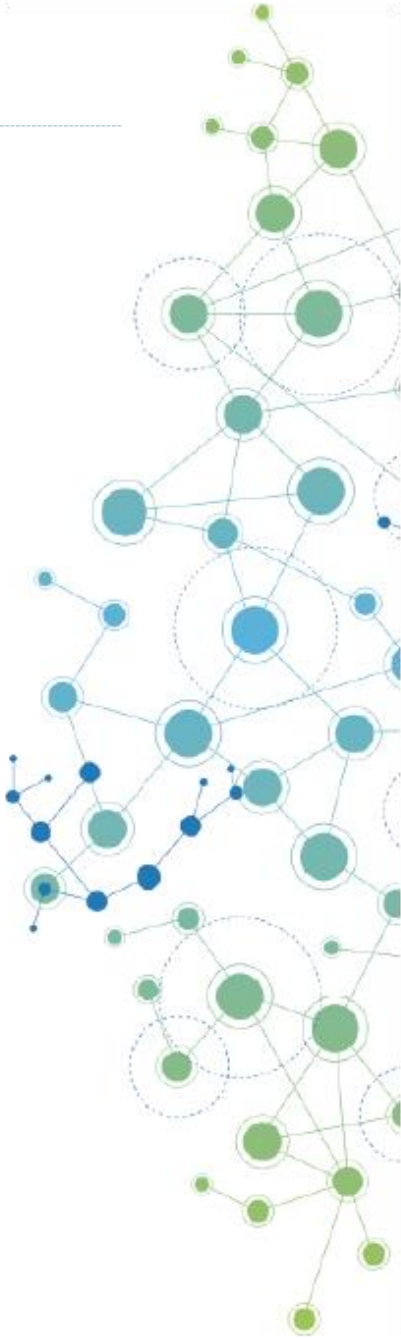
Preditiva.ai

Data Analytics
Fundamentos de Business Intelligence (BI)

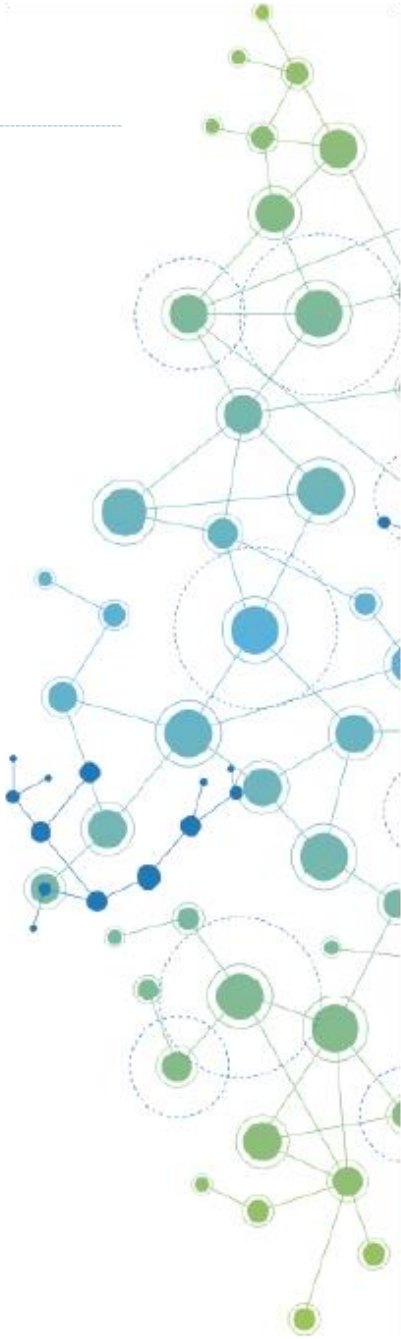
O que você verá nessa aula?

- ❑ Diferenças entre Planilhas Eletrônicas e Bancos de Dados
- ❑ O que é Data Integration, ETL e Ingestão de Dados?
- ❑ Diferenças entre Data Warehouses, Data Marts e Data Lakes
- ❑ Noções de Big Data e Computação Paralela



O que você verá nessa aula?

- ❑ Diferenças entre Planilhas Eletrônicas e Bancos de Dados
- ❑ O que é Data Integration, ETL e Ingestão de Dados?
- ❑ Diferenças entre Data Warehouses, Data Marts e Data Lakes
- ❑ Noções de Big Data e Computação Paralela



Fundamentos de Business Intelligence

Diferenças entre Planilhas Eletrônicas e Bancos de Dados

Preditiva.ai

Como é de se esperar, as planilhas eletrônicas como **Excel**, Google Sheets e tantos outros **não** são as formas mais adequadas de se armazenar dados. Embora seja excelentes softwares para dados de tamanho reduzido (até 100.000 linhas), suas funções são mais direcionadas para Analytics e não armazenamento de informação.

Portanto, quando estamos lidando com dados reais em empresas de médio a grande porte, temos que utilizar ferramentas mais **adequadas** para esse volume de informação.

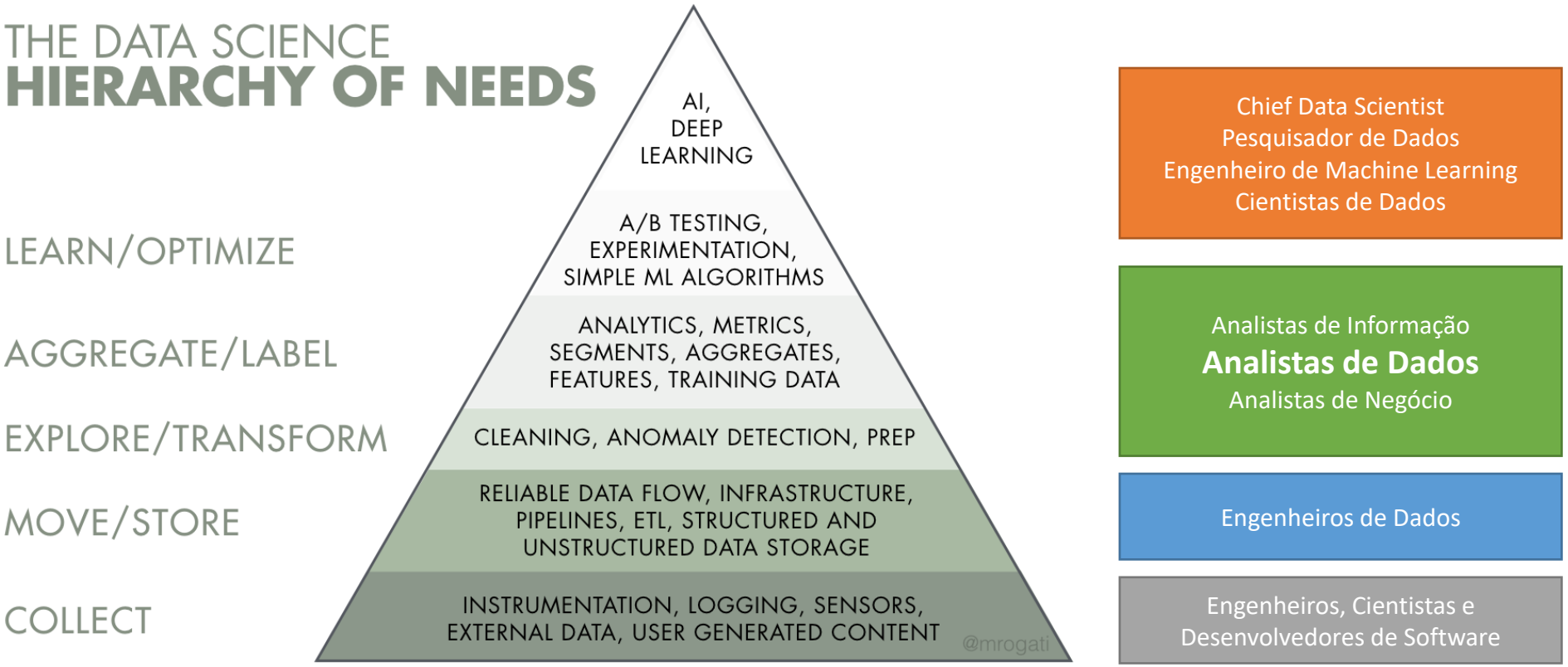
Neste contexto, surgem as **ferramentas de Banco de Dados** que iremos estudar ao longo do curso.

Fundamentos de Business Intelligence

Diferenças entre Planilhas Eletrônicas e Bancos de Dados

Preditiva.ai

Neste curso, estudaremos os **Banco de Dados na perspectiva de um Analista de Dados**. Seremos usuários da informação e não os profissionais que fazem a manutenção do banco (Analista de Banco de Dados, Arquiteto de Banco de Dados ou mesmo o **Engenheiro de Dados**). Relembre as funções em Dados abaixo:



Fundamentos de Business Intelligence

Três características de um Bancos de Dados

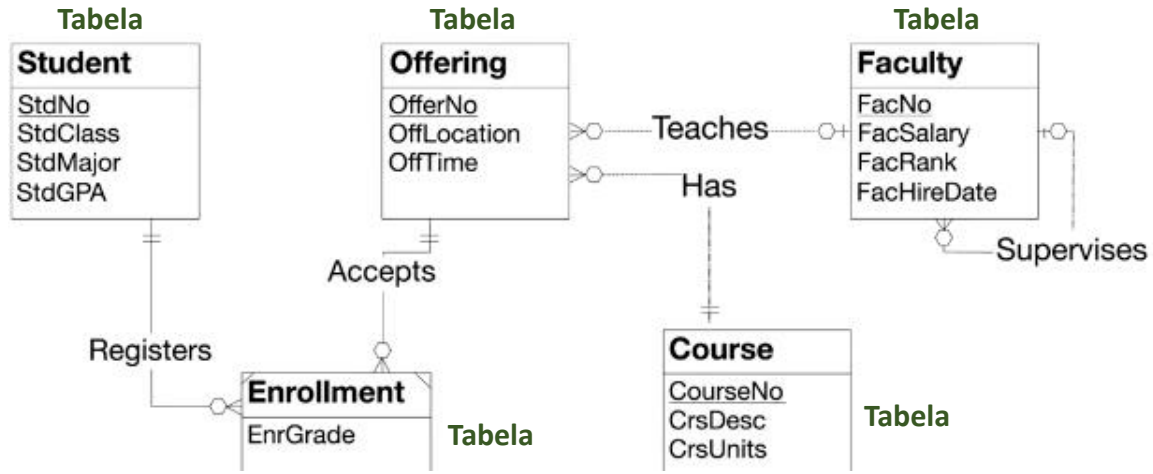
Preditiva.ai

1 **Persistência**: As informações guardadas no banco de dados "persistem" por muito tempo. Não só são armazenadas na memória RAM, como podem ser armazenadas em memória ROM (ex: Hard disks, Clusters, pen drives, DVD's ou mesmo na Cloud).

2 **Compartilhado**: O banco de dados é utilizado por uma infinidade de aplicações, como por exemplo os aplicativos de Desktop, aplicativos Web, Dispositivos móveis entre outros. Além disso, um banco de dados suporta acesso multi usuário de forma simultânea.

3 **Inter-relacionado**: O banco de dados cruza informações com uma série de "entidades". Uma entidade é uma unidade de negócio, como por exemplo Clientes, Produtos, Fornecedores e etc. Clientes se relacionam com os Produtos, por exemplo. As entidades são armazenadas em "tabelas" em um banco de dados. Portanto, podemos dizer **que um banco de dados é um conjunto de entidades/tabelas de informação armazenadas em conjunto.**

Exemplo de ERD (Entity Relationship Diagram), que é um tipo de modelagem dos dados em uma representação gráfica. Mostraremos com mais detalhes este e demais documentações do banco de dados quando falarmos de "Metadados" nas próximas aulas.



Fundamentos de Business Intelligence

Definição de DBMS

Preditiva.ai

Para acessar um Banco de Dados usamos um sistema chamado **DBMS** (Data Base Management System). Um DBMS é uma coleção de componentes para criar, utilizar e realizar a manutenção de um Banco de Dados.

Ao longo dos anos, muitos fornecedores lançaram soluções DBMS no mercado. Com isso, os DBMS evoluíram muito e conhecer todos as funcionalidades de cada plataforma exige muitos anos de estudo. Soluções atuais fornecem capacidade para grandes volumes de dados, acesso simultâneo para milhares de usuários, uma grande segurança dos dados e performance do banco de dados.

Alguns exemplos de fornecedores de DBMS:

IBM DB2

Microsoft SQL Server

teradata.

MicroStrategy

MySQL

SAP SYBASE

PostgreSQL

Fundamentos de Business Intelligence

A evolução trazida pelos Bancos de Dados

Preditiva.ai

Um banco de dados permite que os usuários acessem suas informações com facilidade, sem depender de linguagens de programação complexas e procedurais, como o C++ ou Java. Desta forma, não é necessário se preocupar com estruturas de LOOPs ou WHILE's bem comuns nas linguagens de programação.

O acesso não procedural ao banco de dados é realizado através de uma **Query (Consulta)**.

As **Queries (Consultas)** são instruções para resgatar informações do banco de dados ou mesmo para realizar operações adicionais como inserção, atualização ou remoção de informações dos bancos de dados.

Veja exemplos de consultas:

- Selecionar clientes acima de 18 anos que compraram o celular Samsung Galaxy S;
- Selecionar funcionários que ganham mais do que 5000 reais E estejam de férias;
- Selecionar operações financeiras em atraso a mais de 30 dias E que foram aprovadas após Jun 2018.

Fundamentos de Business Intelligence

A evolução trazida pelos Bancos de Dados

Preditiva.ai

Embora várias soluções DBMS suportam consultas através de ferramentas gráficas, a forma mais utilizada entre os profissionais é através de uma **Linguagem de Consulta Estruturada,** ou do inglês, **Structured Query Language – SQL** (lê-se “Siquou”).

Por exemplo, para clientes de um dos exemplos anteriores, poderíamos utilizar o SQL da seguinte forma:

```
SELECT
      Clientes
FROM
      Tabela_de_Clientes
WHERE
      Idade > 18
      and Produto = ‘Samsung Galaxy S’
```

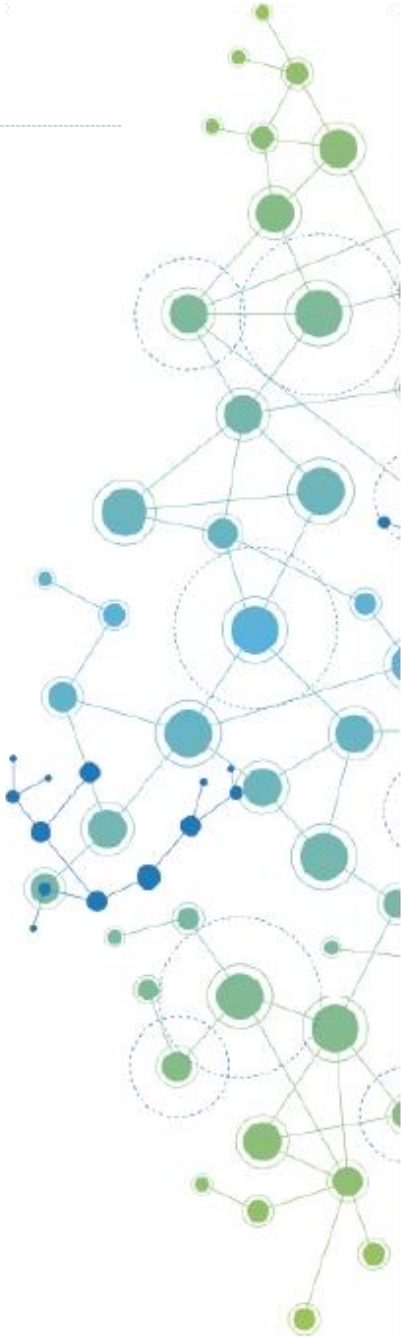
O SQL foi desenvolvido originalmente no início dos anos 70 pela IBM e hoje, após várias evoluções e padronizações, se tornou a linguagem padrão para banco de dados.

Muitos são os DBMS que usam SQL, entre eles:

Sybase, DB2, Microsoft SQL Server, Microsoft Access, MySQL, Oracle, PostgreSQL, SQLite, Teradata, entre outros.

O que você verá nessa aula?

- ❑ Diferenças entre Planilhas Eletrônicas e Bancos de Dados
- ❑ O que é Data Integration, ETL e Ingestão de Dados?
- ❑ Diferenças entre Data Warehouses, Data Marts e Data Lakes
- ❑ Noções de Big Data e Computação Paralela



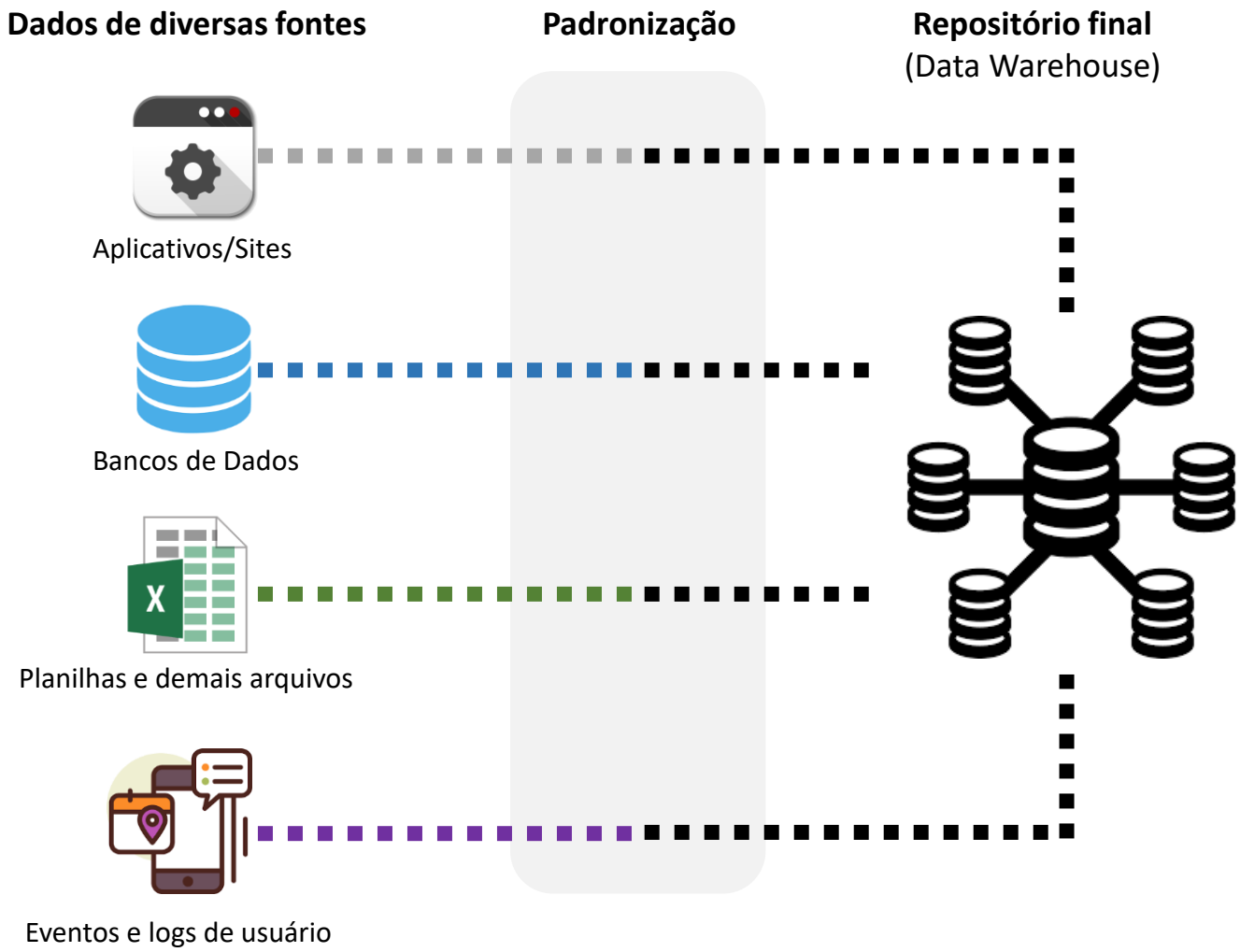
Fundamentos de Business Intelligence

O que é Data Integration, ETL e Ingestão de Dados?

Preditiva.ai

Agora que conhecemos o que são bancos de dados e como utilizá-los através de uma linguagem de consulta estruturada (SQL), precisamos entender **como e onde os dados são armazenados** para serem utilizados pelos Analistas de Dados da empresa.

Tudo começa com o processo de **Ingestão de Dados** (também conhecido como ***Data Integration***) que consiste em transportar os dados de diferentes fontes da empresa para um repositório único, geralmente um ambiente chamado **DW (*Data Warehouse*)** onde todos os usuários podem acessar as informações e realizar suas análises. Veja o processo de Ingestão no diagrama ao lado:

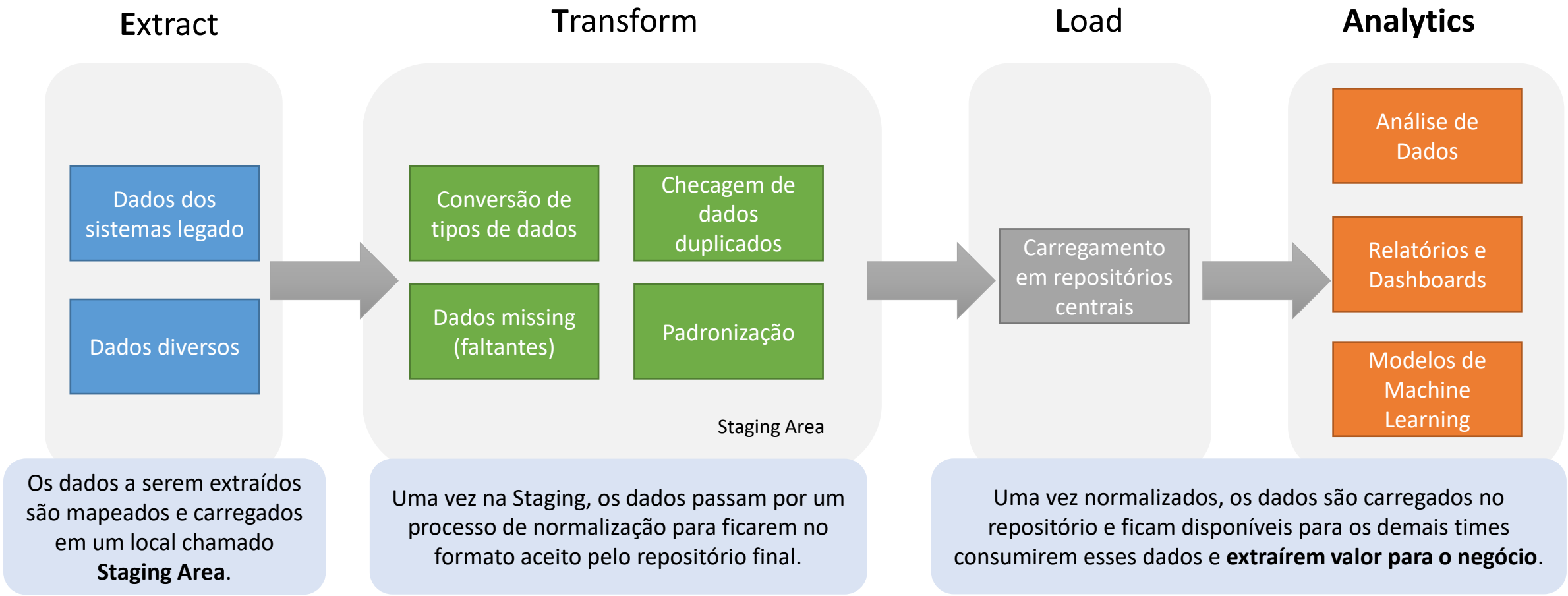


Fundamentos de Business Intelligence

O que é Data Integration, ETL e Ingestão de Dados?

Preditiva.ai

Para o processo de Ingestão de Dados acontecer, é necessário um procedimento de extração, transformação e carregamento desses dados no repositório final. Esse procedimento é chamado de **ETL (*Extract, Transform and Load*)** e geralmente é feito pelo time de Engenharia de Dados. Veja o procedimento simplificado abaixo:



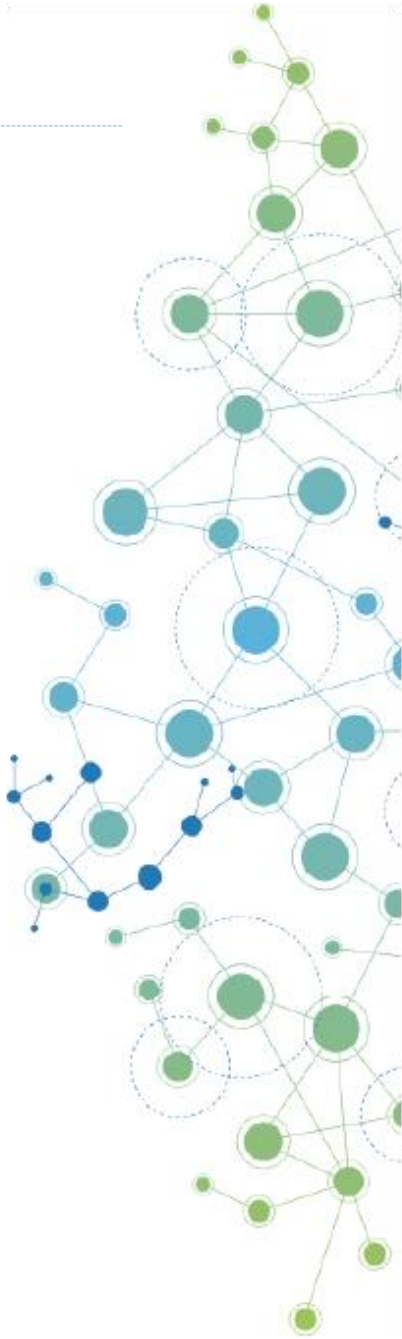
Os dados a serem extraídos são mapeados e carregados em um local chamado **Staging Area**.

Uma vez na Staging, os dados passam por um processo de normalização para ficarem no formato aceito pelo repositório final.

Uma vez normalizados, os dados são carregados no repositório e ficam disponíveis para os demais times consumirem esses dados e **extraírem valor para o negócio**.

O que você verá nessa aula?

- ❑ Diferenças entre Planilhas Eletrônicas e Bancos de Dados
- ❑ O que é Data Integration, ETL e Ingestão de Dados?
- ❑ Diferenças entre Data Warehouses, Data Marts e Data Lakes
- ❑ Noções de Big Data e Computação Paralela

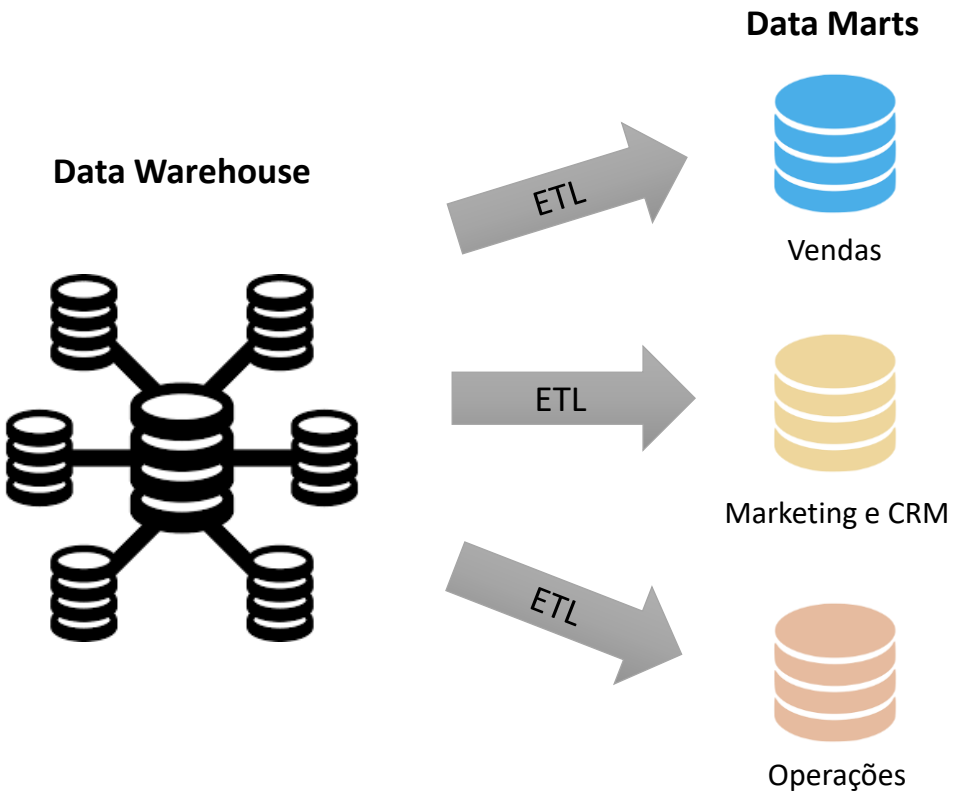


Fundamentos de Business Intelligence

Diferenças entre Data Warehouses, Data Marts e Data Lakes

Preditiva.ai

Geralmente pensamos como repositório final o Data Warehouse (DW), repositório este que armazena os dados transformados pelo procedimento de ETL. Acontece que como os dados do DW são muito diversificados, algumas áreas de negócio seriam mais produtivas se o time de Engenharia fornecesse um tipo de DW mais específico para o seu negócio. Surge então outro tipo de repositório conhecido como **Data Mart (DM)**, um subtipo de DW que contém apenas os dados necessários que determinada área precisa.



Principais vantagens do Data Mart:

- ✓ **Produtividade**: Como é mais específico e construído para as necessidades das áreas, seu uso é mais direto e economiza tempo dos analistas.
- ✓ **Performance**: Como muitas vezes os dados já estão processados do jeito que a área precisa, seu uso é mais rápido pois não existe muito processamento adicional.
- ✓ **Implantação**: Um DM pode ser implantado em questão de dias enquanto um DW demora alguns meses.

Fundamentos de Business Intelligence

Diferenças entre Data Warehouses, Data Marts e Data Lakes

Preditiva.ai

Nos últimos anos com o advento da **computação em nuvem** (em que servidores são oferecidos como serviços), cada vez mais empresas estão abandonando os servidores físicos internos (chamados de ***on-premises servers***) migrando sua infra-estrutura para servidores na nuvem (***cloud servers***). Os benefícios desta escolha são muitos. Veja:

	Servidores Físicos	Servidores na Nuvem
Instalação	**Complexa**: Depente de uma sala, refrigeração, no-break, hardware, equipe de TI dedicada etc.	**Simples**: Tudo é fornecido pelo provedor de serviços, bastando configurar os parâmetros do servidor.
Custo	**Alto**: O investimento para iniciar a operação é alto, pois depende de todos os itens de instalação.	**Baixo**: Os provedores só cobram pelo dimensionamento de servidor utilizado, barateando o custo.
Escalabilidade	**Baixa**: Sempre será necessário comprar mais servidores a medida que a empresa crescer.	**Alta**: Os serviços se auto ajustam sempre que for necessário, deixando a empresa livre de preocupações do crescimento da operação.

Principais provedores de Cloud Servers

amazon web services

Google Cloud Platform

Microsoft Azure

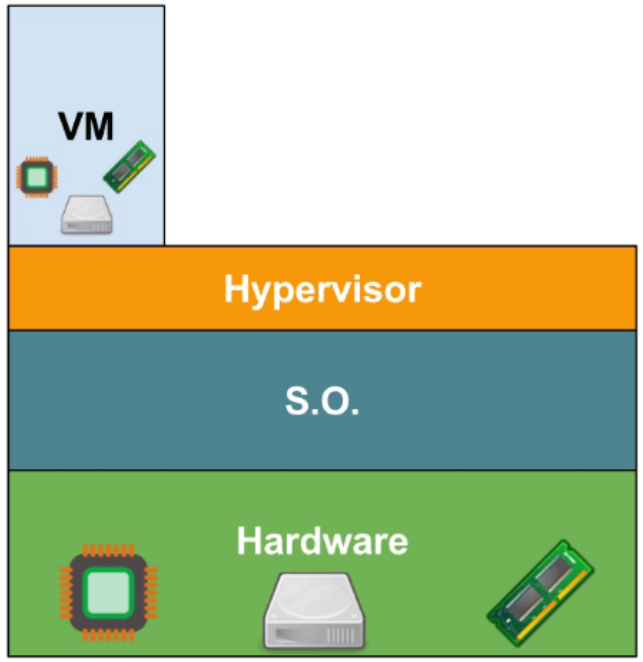
Fundamentos de Business Intelligence

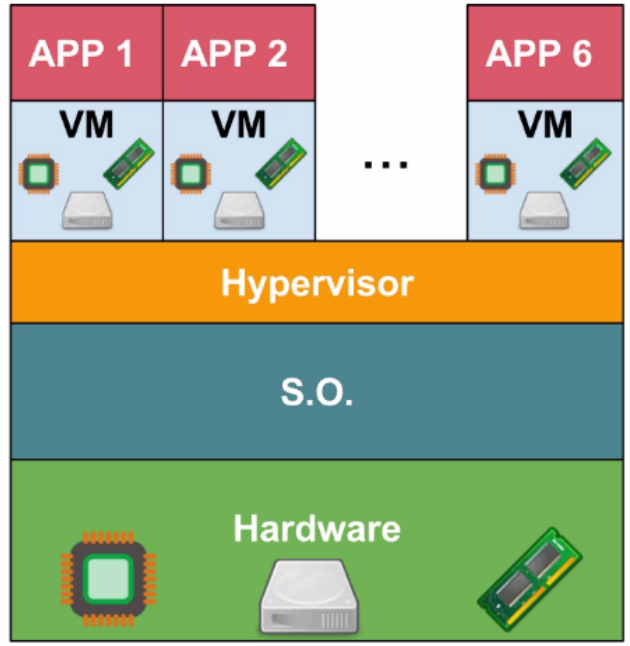
Virtualização

Preditiva.ai

Para fugir desses problemas de servidores ociosos e alto tempo e custo de subir e manter aplicações em servidores físicos, surgiu como solução a **Virtualização**. As Máquinas Virtuais (Virtual Machines) resolveram um grande problema de escala das empresas.

Antes: Servidores locais nas empresas





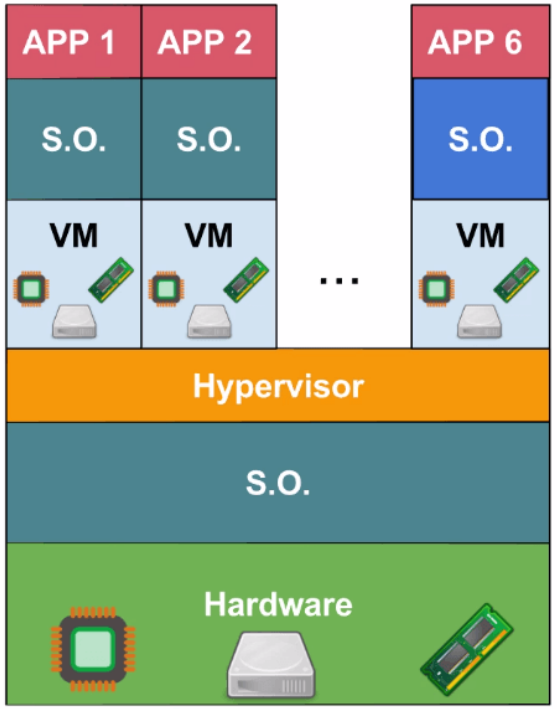
Depois: Servidores na Nuvem e Virtualizados

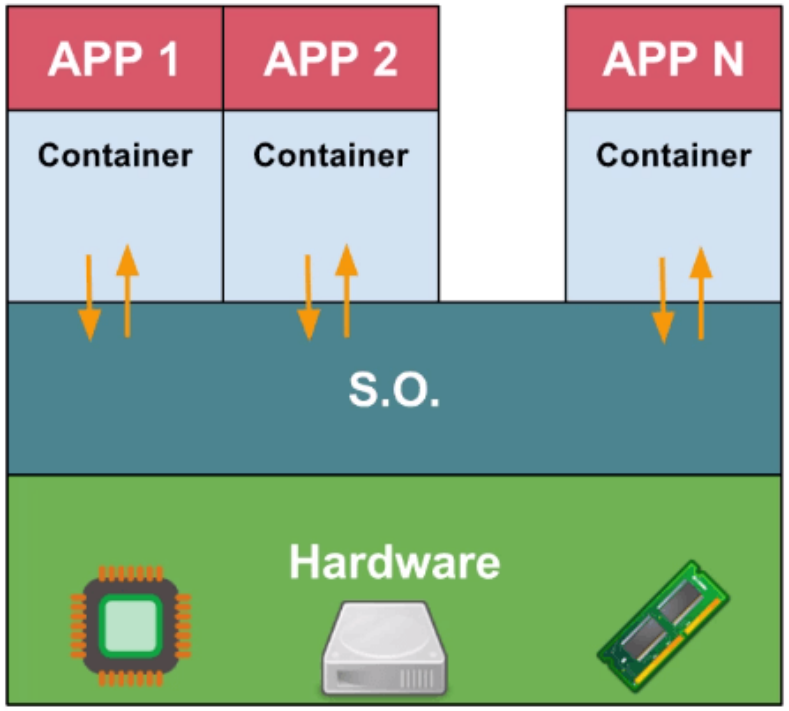
Fundamentos de Business Intelligence

Virtualização

Preditiva.ai

As VM’s são muito boas, mais ainda dão certo trabalho... É preciso inserir sistemas operacionais delas... ☹





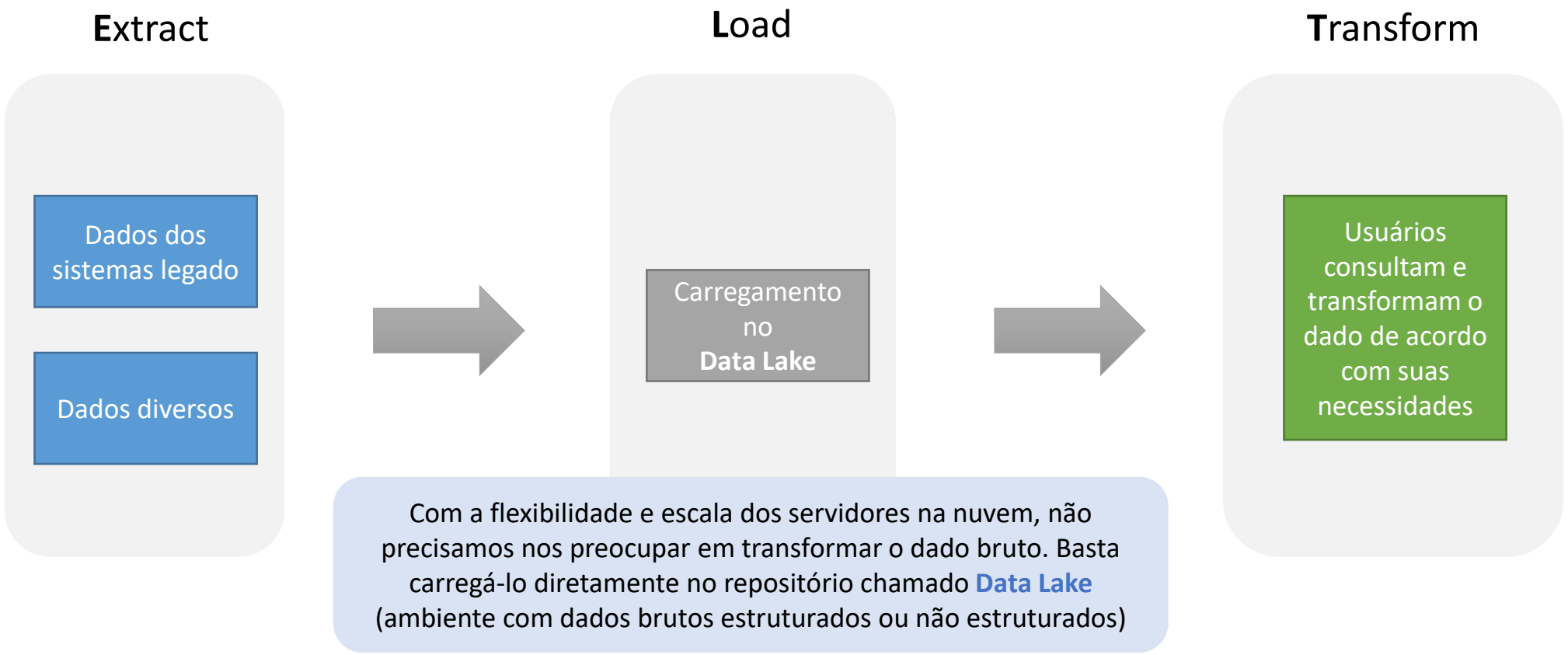
Solução: Containers !!! \0/

Fundamentos de Business Intelligence

Diferenças entre Data Warehouses, Data Marts e Data Lakes

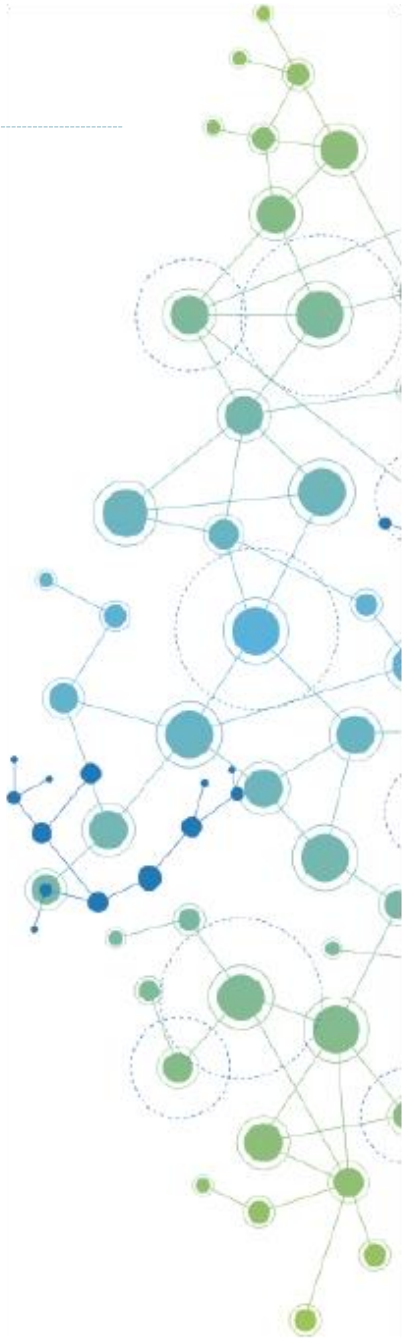
Preditiva.ai

Com essa flexibilidade trazida pela computação em nuvem, os processos de armazenamento de dados nas empresas ganharam novos horizontes. Unindo o custo de armazenamento mais baixo e a flexibilidade dos servidores em nuvem, o processo de Ingestão de Dados tem mudado em muitas empresas ao redor do mundo. Em vez de usar o procedimento de ETL habitual, faz mais sentido usar outro tipo de procedimento, **o ELT (Extract, Load and Transform).** Veja:



O que você verá nessa aula?

- ❑ Diferenças entre Planilhas Eletrônicas e Bancos de Dados
- ❑ O que é Data Integration, ETL e Ingestão de Dados?
- ❑ Diferenças entre Data Warehouses, Data Marts e Data Lakes
- ❑ Noções de Big Data e Computação Paralela



Fundamentos de Business Intelligence

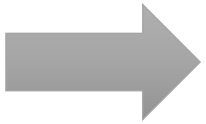
Noções de Big Data e Computação Paralela

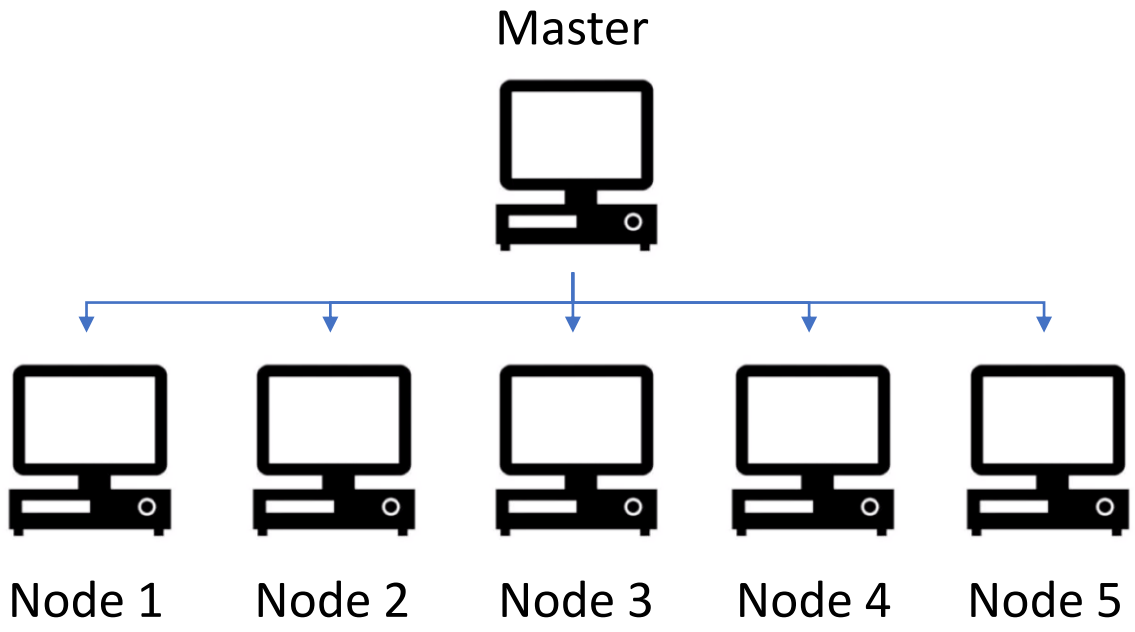
Preditiva.ai

Um conceito importante que todo Analista de Dados deve conhecer é o de **Computação Paralela** e as tecnologias derivadas para processamento de grandes volumes de dados (**Big Data**). Faremos uma introdução para as principais tecnologias: **o Hadoop e o Spark**. Para iniciar essa discussão, precisamos entender o que é Computação Paralela. Vejamos no diagrama abaixo:

Node (Nó)

Computação Paralela
(ou Computação Distribuída)





Quando usamos nosso computador para processar os dados, estamos usando os recursos individuais de uma máquina (Cores da CPU, Disco e Memória RAM)

Em **Computação Paralela**, usamos os recursos de múltiplos computadores (chamados de "nós") simultaneamente. Esses computadores são gerenciados por um nó chamado de **Master**. Nesta configuração, o conjunto de computadores é chamado de **Cluster**.

Fundamentos de Business Intelligence

Noções de Big Data e Computação Paralela

Preditiva.ai

O que é então o **Hadoop**? É um framework de computação paralela que permite processar grandes volumes de dados em um cluster de computadores usando o modelo de programação chamado **MapReduce**. Vejamos um exemplo de operação de soma de valor de compras por cliente usando o Hadoop:

Nome	Valor_Compras
Julia	200
Leandro	100
Maria	200
Maria	300
Leandro	300
Maria	100
Julia	100
João	300
Leandro	300
João	200
Leandro	200
Julia	300
Maria	200
João	300
João	100

Node 1

Nome	Valor_Compras
Julia	200
Leandro	100
Maria	200
Maria	300
Leandro	300

Node 2

Nome	Valor_Compras
Maria	100
Julia	100
João	300
Leandro	300
João	200

Node 3

Nome	Valor_Compras
Leandro	200
Julia	300
Maria	200
João	300
João	100

Map

(Etapa em que os nós mapeiam os valores)

- Julia: 200, Leandro: 100 , 300, Maria: 200 , 300
- Julia: 100, Leandro: 300, Maria: 100, João: 300 , 200
- Julia: 300, Leandro: 200, Maria: 200, João: 300 , 100

Merge

(Etapa em que os nós juntam os valores por nome)

- Julia: 200, Julia: 100, Julia: 300
- Leandro: 100 , 300, Leandro: 300, Leandro: 200
- Maria: 200 , 300, Maria: 100, Maria: 200
- João: 300 , 200, João: 300 , 100

Reduce

(Etapa em que os nós aplicam a função de agregação escolhida)

- Julia: 600
- Leandro: 900
- Maria: 800
- João: 900

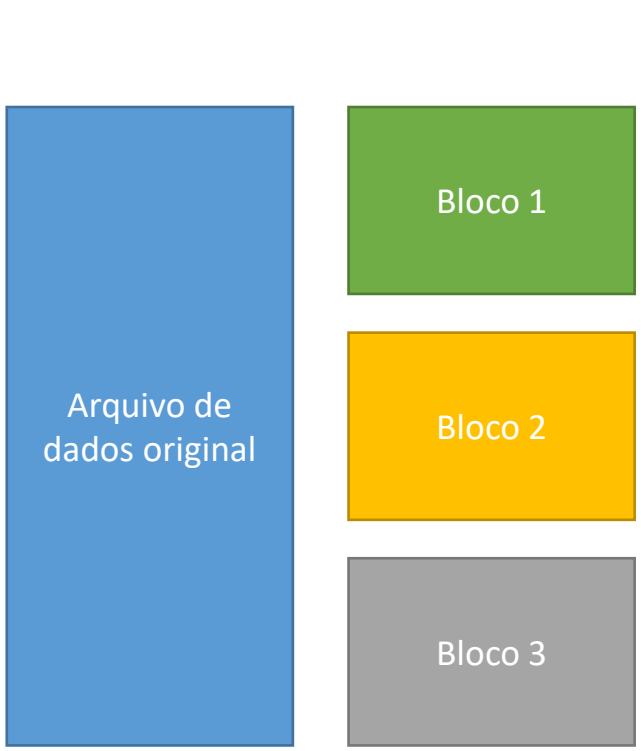
JOAO ESTEVAN LEONCIO DA SILVA BARBOSA - jestevan12@gmail.com - CPF: 134.982.877-70

Fundamentos de Business Intelligence

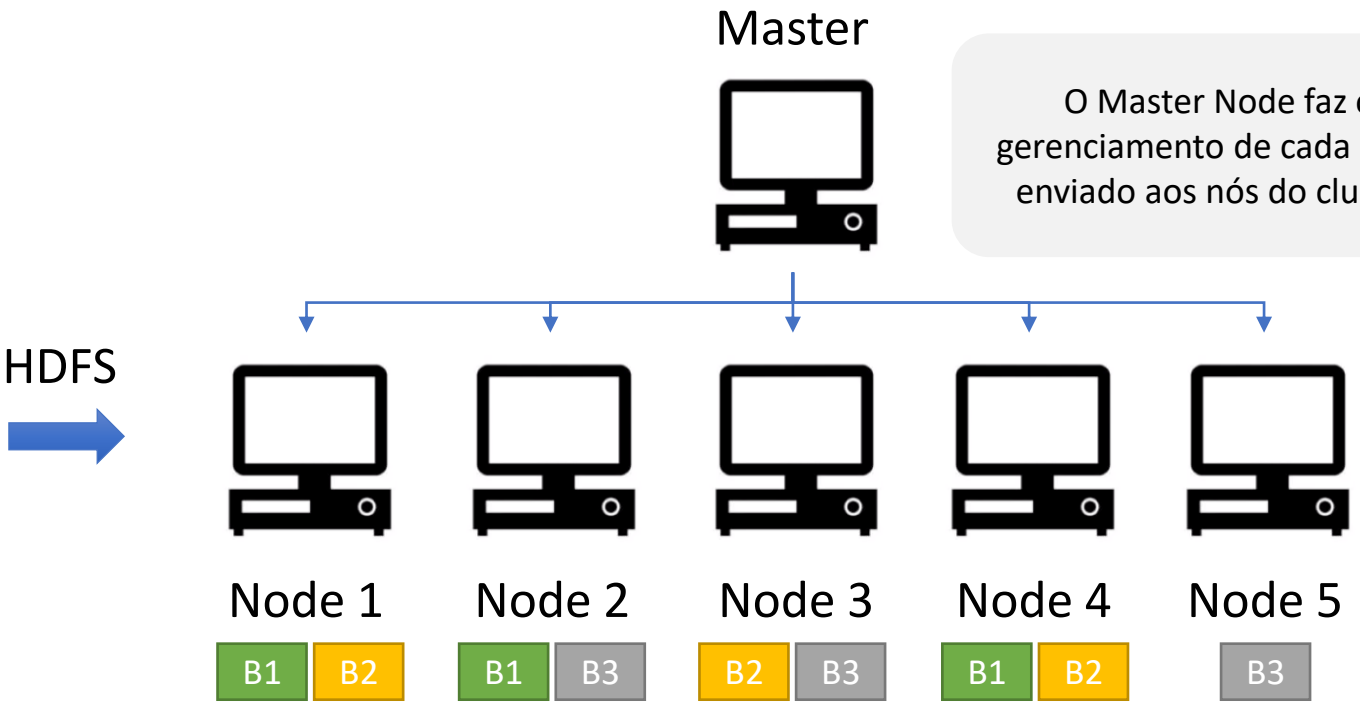
Noções de Big Data e Computação Paralela

Preditiva.ai

Além do modelo **MapReduce**, uma característica importante do Hadoop é seu sistema de arquivos chamado **HDFS (Hadoop Distributed File System)**. O HDFS é o responsável por distribuir um arquivo de dados no cluster. Veja como ele funciona:



HDFS



O Master Node faz o gerenciamento de cada bloco enviado aos nós do cluster.

O arquivo de dados é **separado** em pequenos blocos (geralmente de 128 MB cada) pelo HDFS. Quanto menor for o bloco, mais potencial de paralelização o Hadoop pode oferecer.

Com isso o HDFS distribui cada bloco de dados 3 vezes entre os nós do cluster. Desta forma, se uma das máquinas apresenta defeito, o **arquivo é restaurado** facilmente por uma de suas cópias nos outros nós do cluster.

Fundamentos de Business Intelligence

Noções de Big Data e Computação Paralela

Preditiva.ai

Embora o MapReduce tenha sido um divisor de águas no processamento de Big Data, ele **não é perfeito**. Um dos problemas é que ele **processa os dados em disco, o que é muito lento**. A cada etapa de Map e Reduce ele escreve dados em disco. Um outro problema conhecido é que o MapReduce só trabalha com dados no formato HDFS, o que acaba não sendo flexível em vários casos.

Uma tecnologia lançada em 2013 que endereça esses e outros problemas é o **Apache Spark**. Veja suas principais características:

- Spark é um framework de processamento de dados usando computação distribuída;
- Em vez de usar dados em disco, o **Spark armazena** a maioria de seus dados **em memória RAM**, o que promove processamento até 100x mais velozes do que o MapReduce do Hadoop.
- Spark pode trabalhar com diversos tipos de dados, entre eles:
  - Amazon S3
  - Cassandra
  - HDFS, entre outros.
- É possível usar o Spark com várias linguagens de programação, tais como : Java, Scala, Python, R e SQL.

Fundamentos de Business Intelligence

Noções de Big Data e Computação Paralela

Preditiva.ai

Caso você não trabalhe com linguagens de programação habituais, sem problemas. Ainda é possível trabalhar com computação distribuída usando linguagens mais amigáveis como o **SQL**. Uma das ferramentas que possibilita usar o Hadoop através do SQL é o **Apache Hive**.

https://hive.apache.org

presto

https://prestosql.io

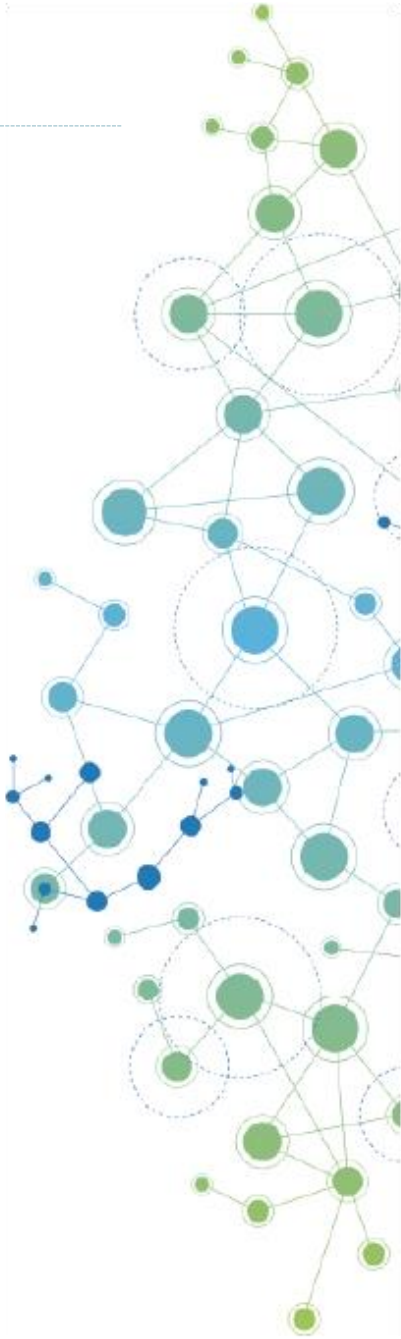
Outra ferramenta que possibilita tirar proveito da computação distribuída usando SQL é o **Presto SQL**. Esta ferramenta garante alta performance, versatilidade (roda on-premises ou em provedores de Cloud) e é compatível com a maioria das ferramentas de Visualização de Dados, tais como Tableau ou Power BI.

Revisão e Próximos Passos

Nesta seção passamos pela maioria dos conceitos, técnicas e ferramentas utilizadas no que chamamos de **Business Intelligence** (basicamente tomar melhor decisões de negócio usando dados e tecnologia adequada)**.**

Para isso, falamos sobre:

- ❑ A diferença de planilhas eletrônicas e **bancos de dados**;
- ❑ Os tipos de bancos de dados e ferramentas de acesso aos dados (**DBMS** através do SQL);
  - SQL é uma linguagem de consultas estrutura usada na maioria das atividades de análise de dados.
- ❑ O processo de **Ingestão de Dados** através de procedimentos **ETL** (Extract, Transform e Load);
- ❑ Os tipos de repositórios de Dados: Data Warehouse, Data Mart e **Data Lake**;
- ❑ Servidores on-premises (servidor físico interno na empresa) e **Servidores Cloud** (Amazon, Google e Microsoft);
- ❑ **Computação Paralela** e como o **Big Data** é processado usando ferramentas como o **Hadoop MapReduce, Spark** entre outros.



No próximo módulo vamos começar a estudar a principal linguagem para Analytics, o SQL. Até lá !

Preditiva.ai