

A primeira Escola presencial gratuita de Inteligência Artificial do Brasil



Apoio



Aula 01/2020: Introdução à Ciência de Dados, Big Data, Machine Learning e Inteligência Artificial

Professores: Andréa Longarini e Thiago Kuma

Objetivo da Aula

Explicar os conceitos de:

- Big Data
- Business Intelligence
- Ciência de Dados
- Machine Learning
- Inteligência Artificial
- Arquitetura de Soluções
- Serviços Cognitivos
- Engenharia de Dados
- Algoritmos e Aplicações

Após essa aula o aluno será capaz de entender a diferença entre as áreas e as atividades e para que serve cada algoritmo na Ciência de Dados.

Big Dada



Big Dada



Big Dada

Big Data é a coleta de um grande volume de dados, feito em alta velocidade em grande variedade.

3 Vs do Big Data:

- Volume - Um grande volume de dados (terabytes a petabytes)
- Velocidade - Dados sendo coletados em alta velocidade em tempo real
- Variedade - Vários formatos e várias origens de dados estruturados e não estruturados

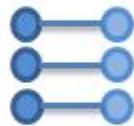
Temos + 2 Vs que foram adicionados posteriormente, pois não são mensuráveis.

- Veracidade: Os dados devem ser reais e refletirem a realidade
- Valor: Os dados devem apresentar valor, devem ser úteis para extrair insights

Big Dada



Chave Valor



Redis
DynamoDB

Família de Coluna



HBase
Cassandra

Documentos



MongoDB

Grafos

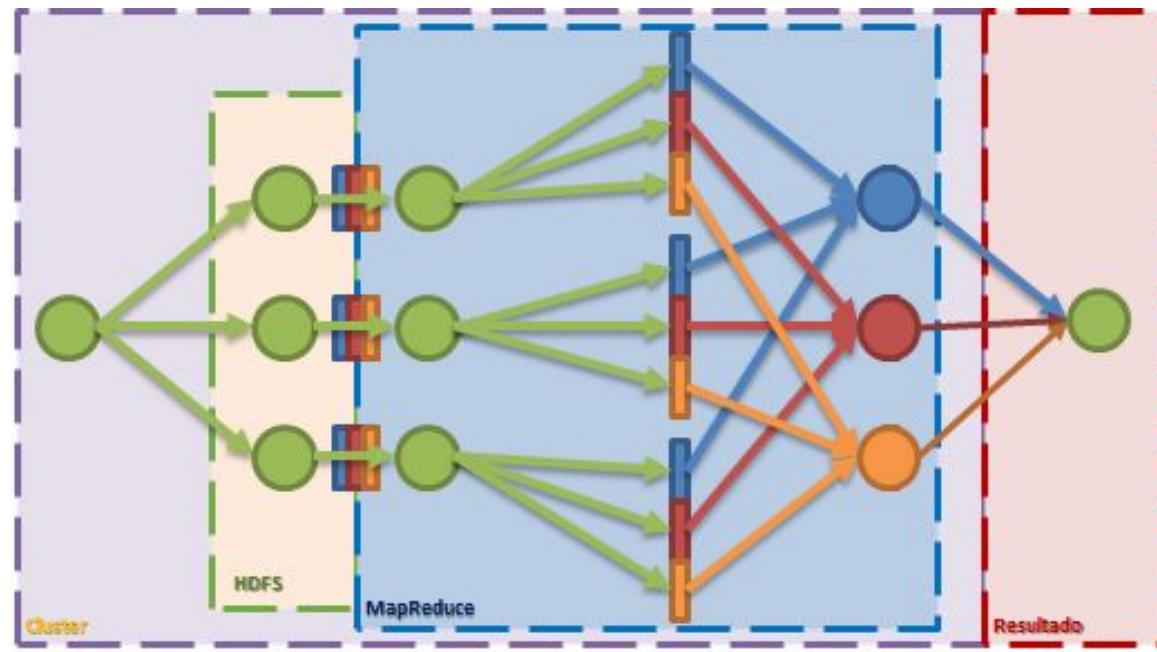


Neo4J

Big Dada

Para Armazenar e processar esse grande volume de dados foi necessário um sistema que utiliza dados distribuídos em vários computadores que chamamos de clusters. O Sistema é chamado de Hadoop que trabalha em disco e posteriormente foi lançado o Spark que trabalha em Disco e em memória. Resolve problema de processamentos de grandes volumes de dados.

Map e Reduce

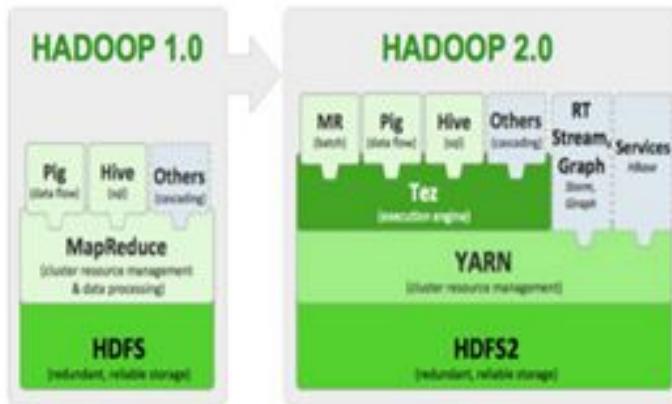


Big Dada

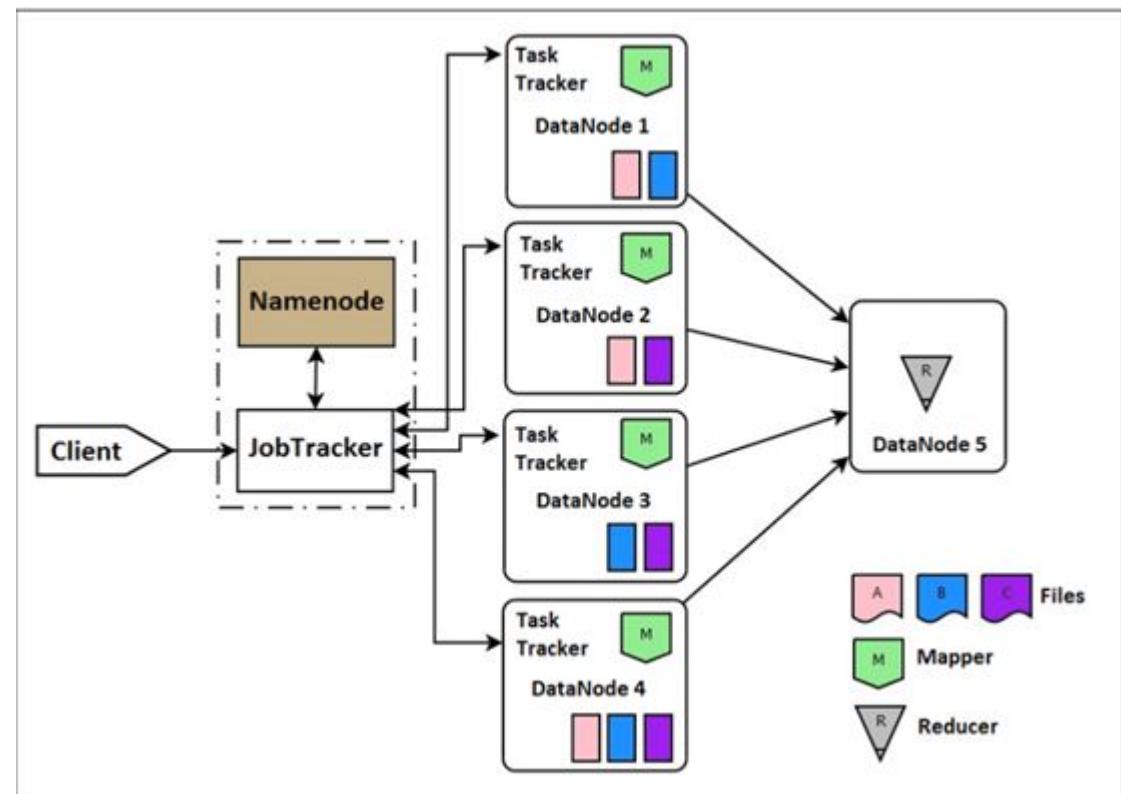
Arquitetura Haddop 1.0

Executa somente em Disco

Preocupação era Armazenar

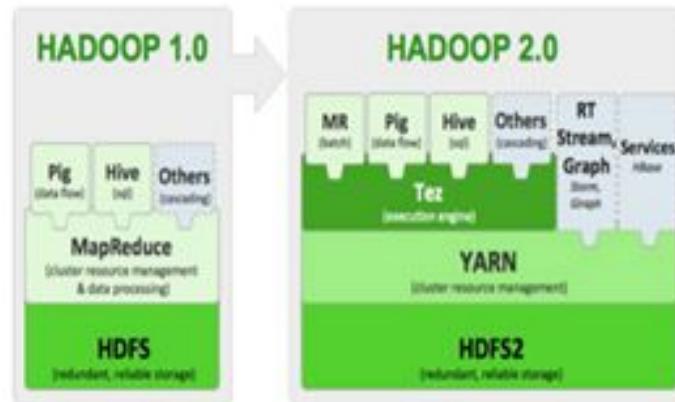


Fonte: INFINITE SCRIPT (2014).

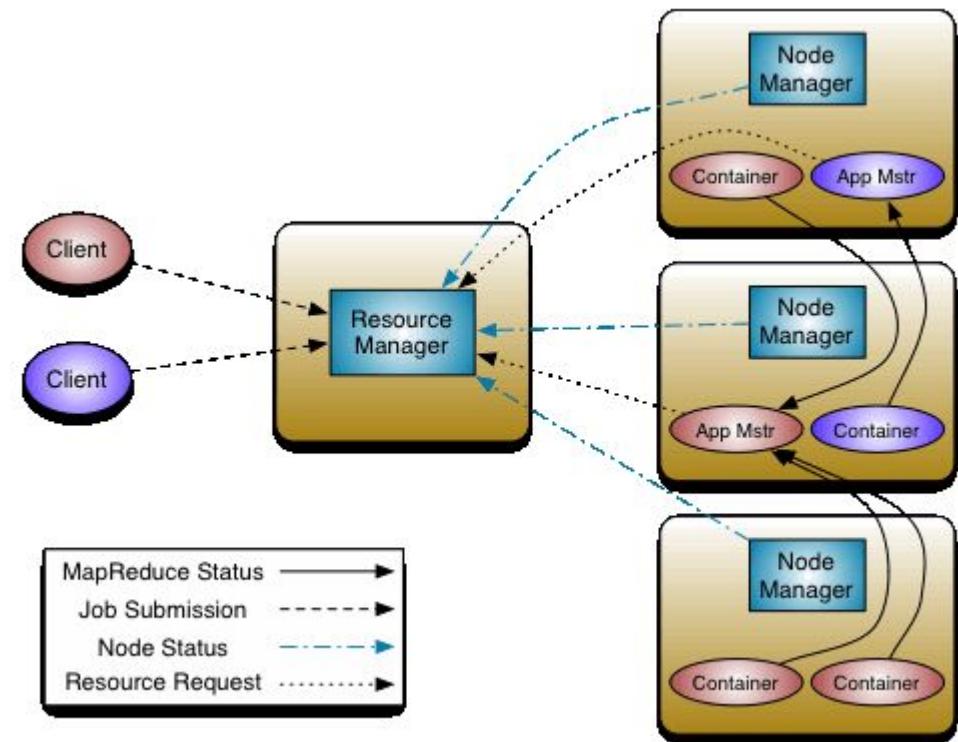


Big Dada

Arquitetura Haddop 2.0

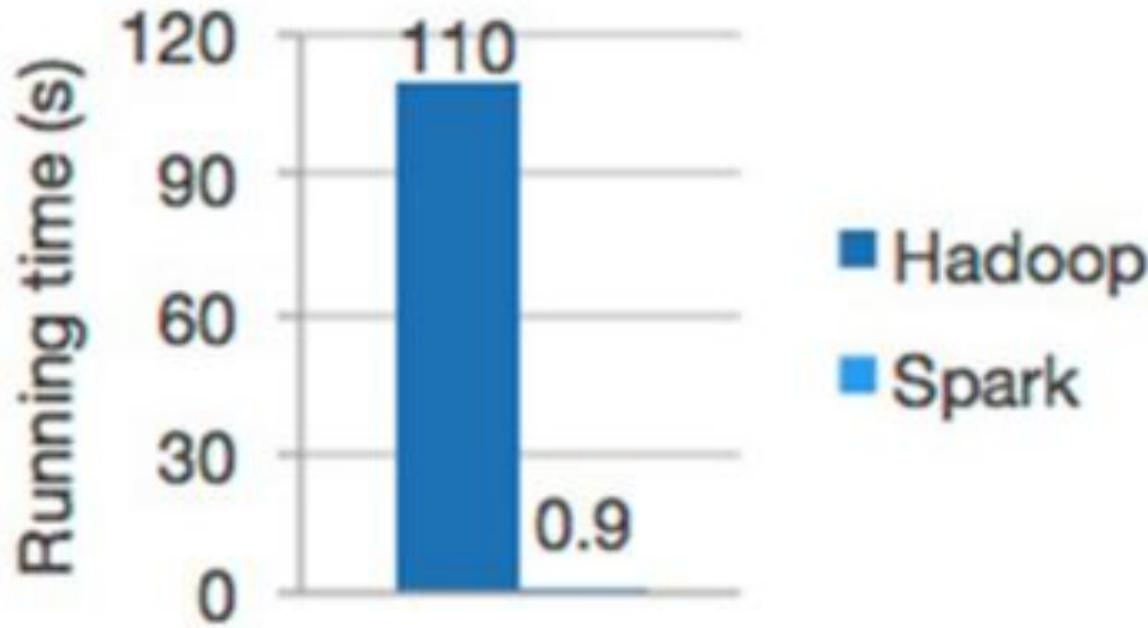


Fonte: INFINITE SCRIPT (2014).



Big Dada

Spark

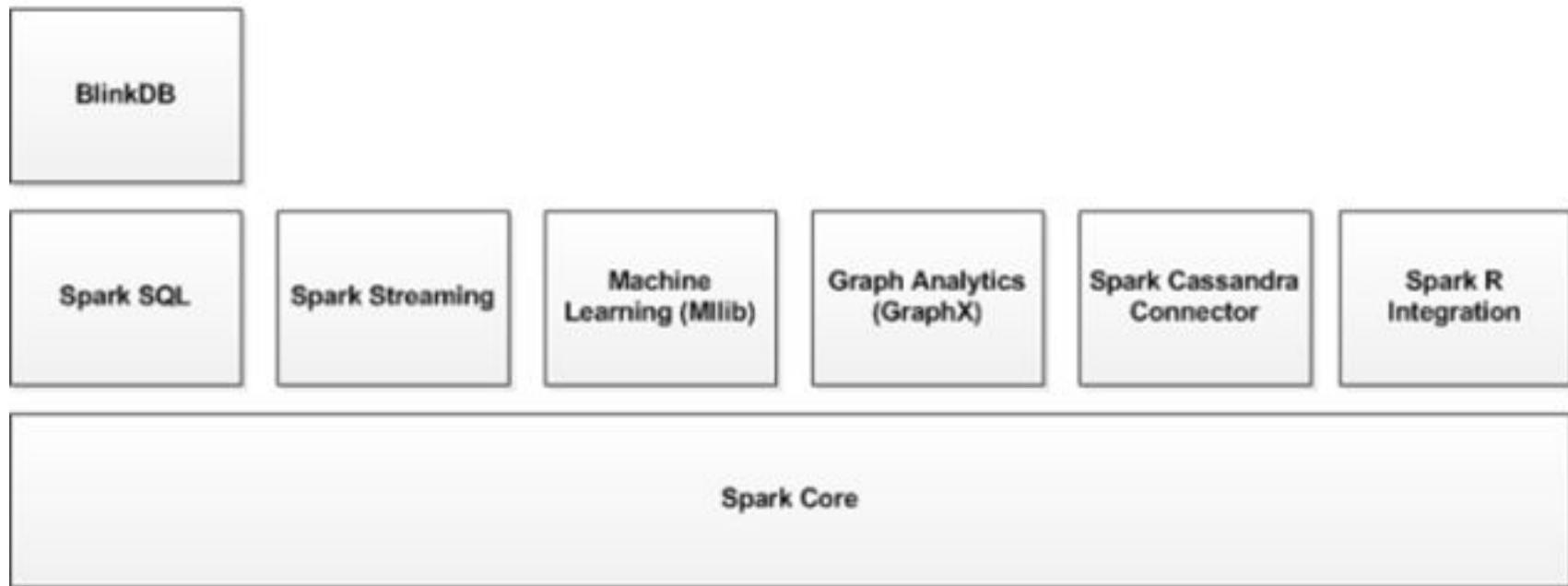


Executa em disco e em memória
Preocupação é Extrair valor

Big Dada

Spark

Spark Framework Ecosystem



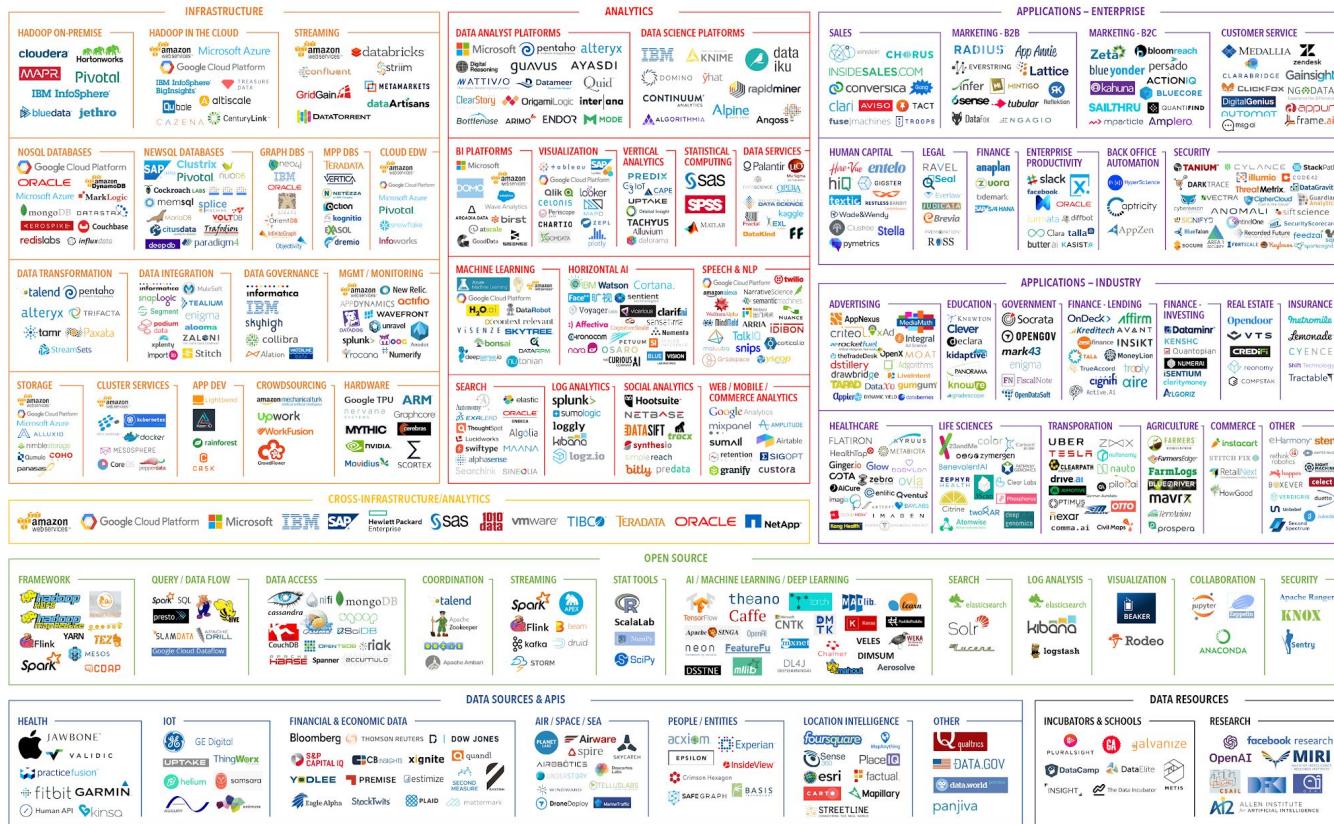
Big Data



Big Data

Ferramentas e Frameworks

BIG DATA LANDSCAPE 2017



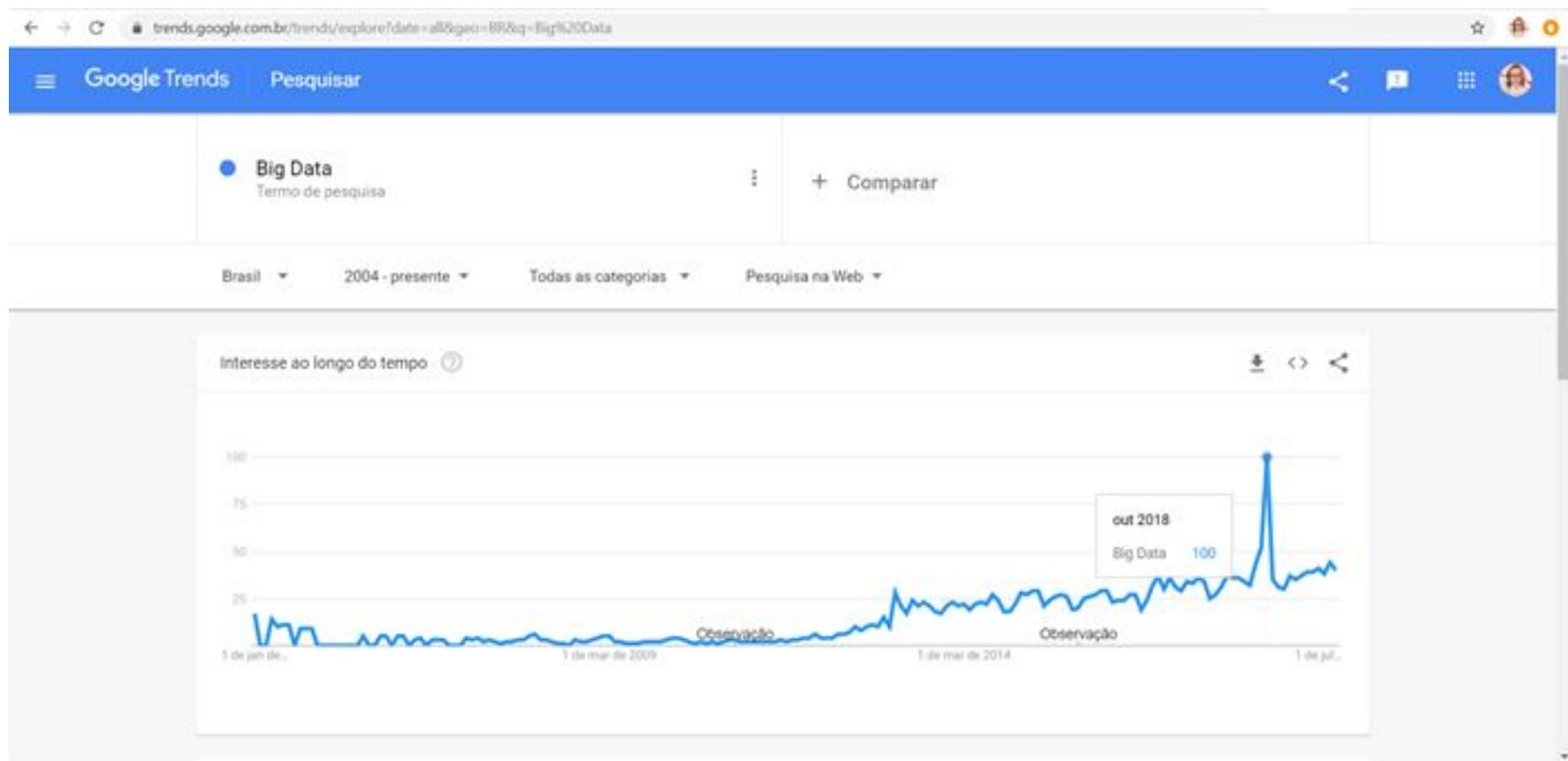
Last updated 4/5/2017

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap) mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Big Dada

Em Outubro de 2018 o Termo Big data atingiu o topo nas pesquisas do Google Trends



Business Intelligence

Inteligência nos Negócios, de acordo com os objetivos da empresa, o Analista de BI modela os dados para extrair as medidas/métrica para verificar o desempenho da empresa.

Trabalha junto à área de Negócios na identificação dos KPIs (Indicadores chaves de desempenho) para que seja possível medir a saúde da empresa através dos dados.

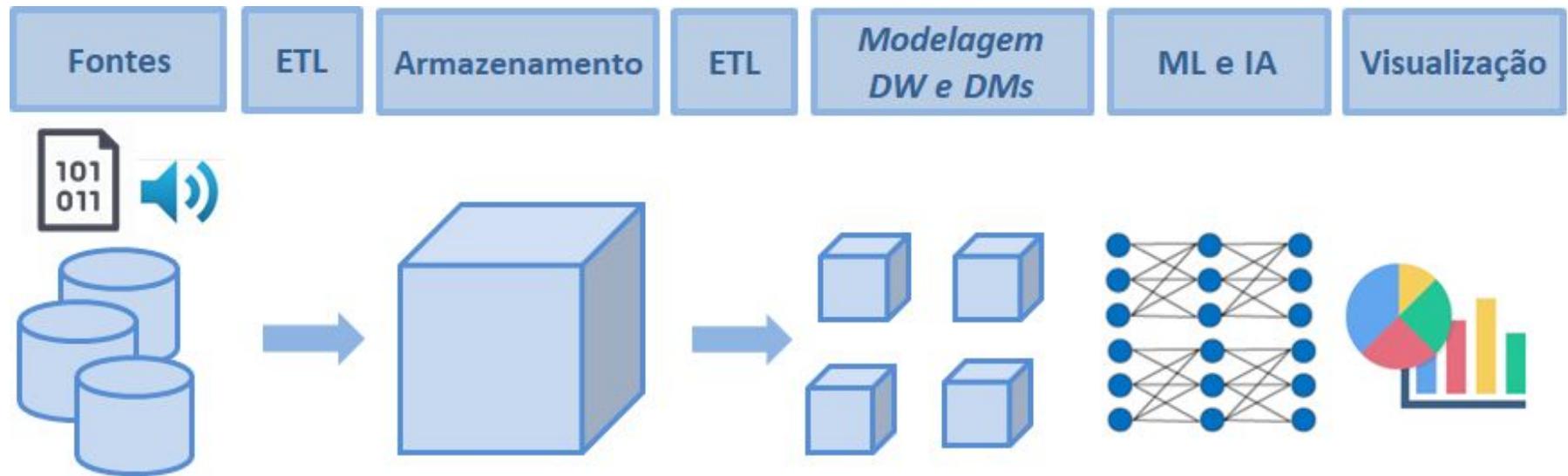
Quando percebe impacto financeiro nos dados, aciona as áreas de Negócios e Marketing para que seja feita uma mudança na estratégia.

Trabalha na Extração dos dados, tratamento, limpeza, dá qualidade aos dados.

Modela o Cubo de Dados (modelagem por assunto ou por métrica)

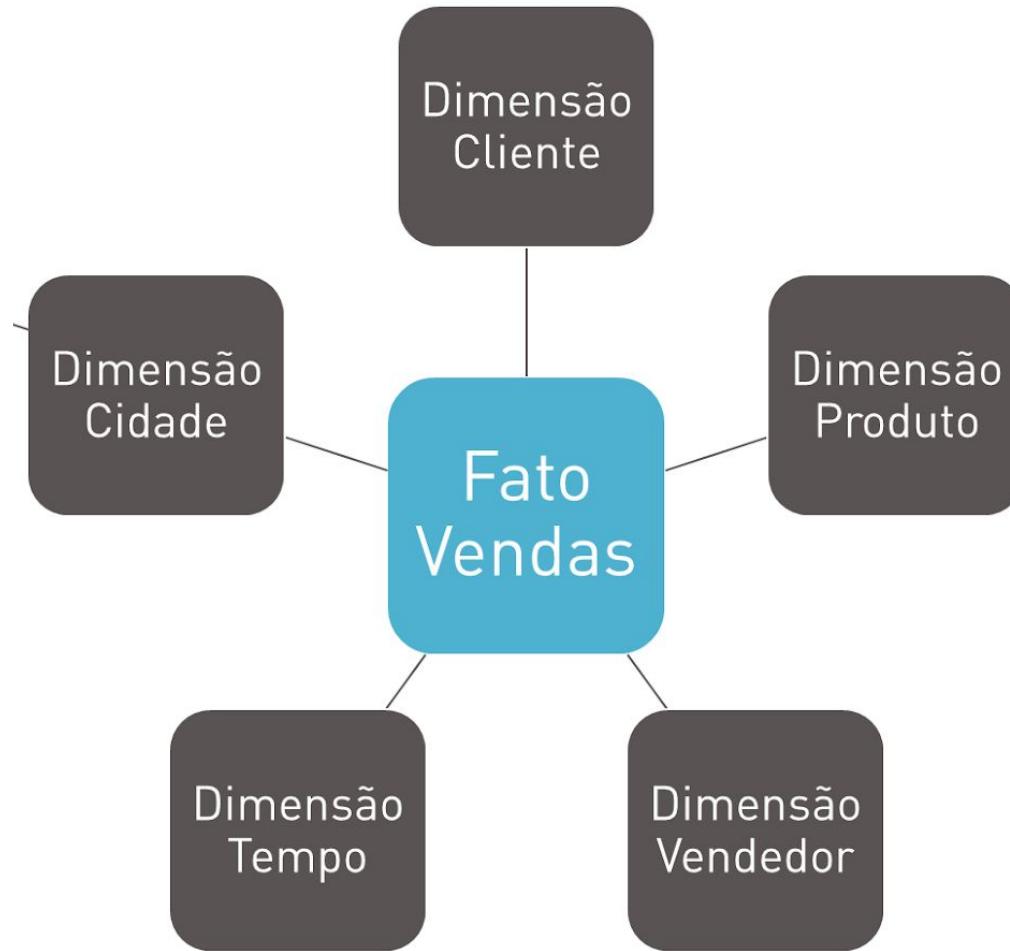
Prepara as visões, dashboards.

Business Intelligence

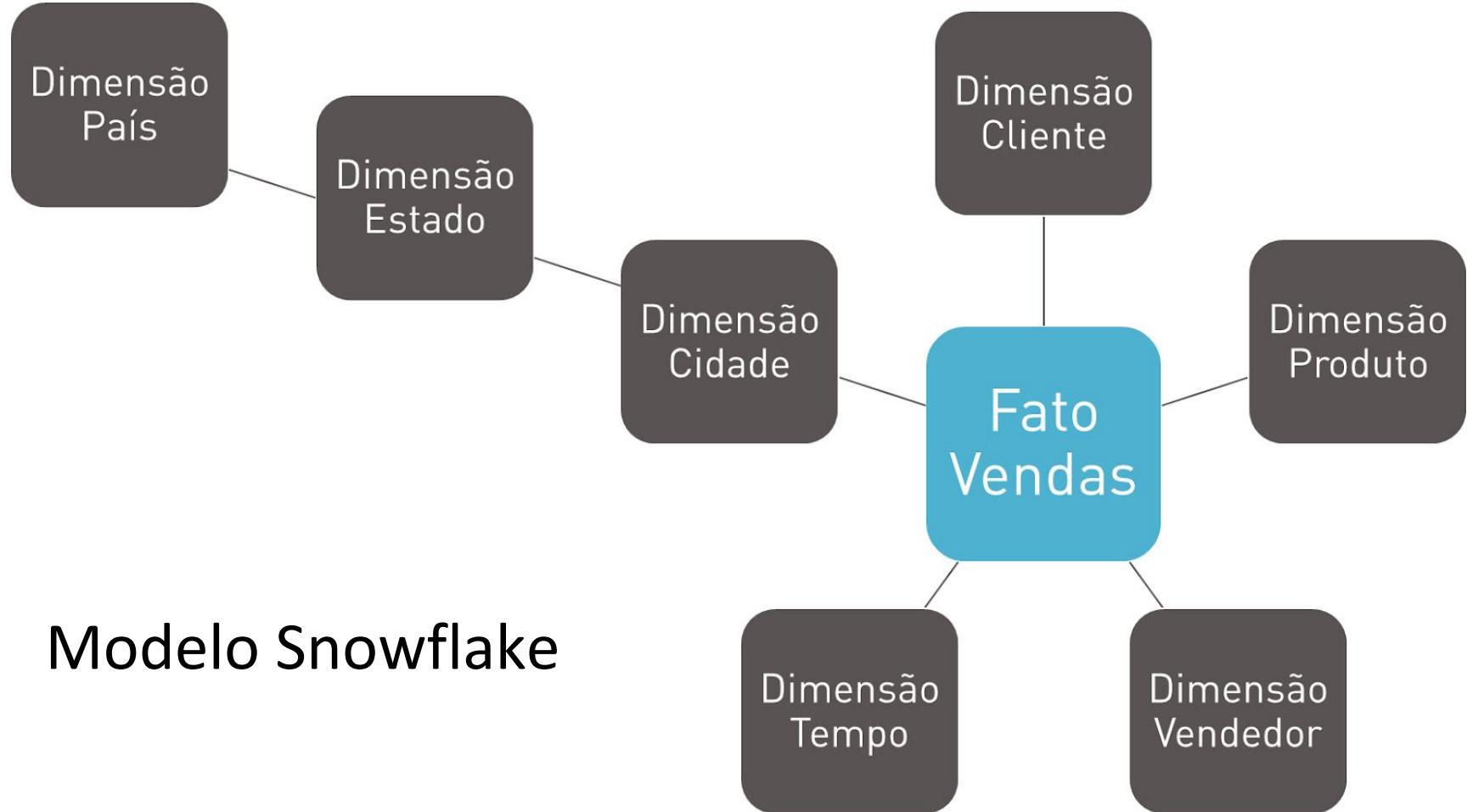


Business Intelligence

Modelo Star Schema



Business Intelligence



Modelo Snowflake

Business Intelligence

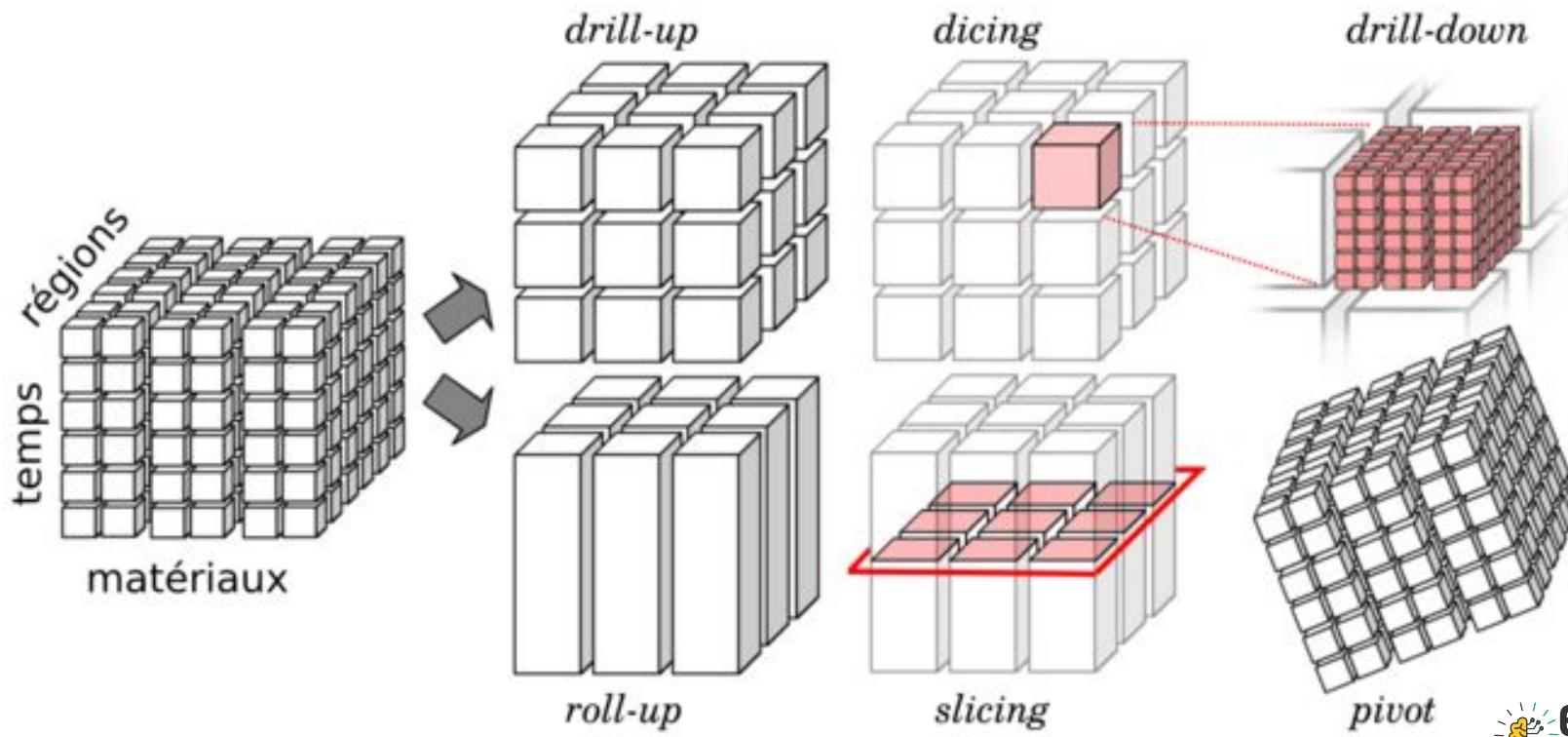
Drill Down - Navega de um nível menos detalhado para um nível mais detalhado

Drill up ou Roll up - Navega de um nível mais detalhado para um menos detalhado

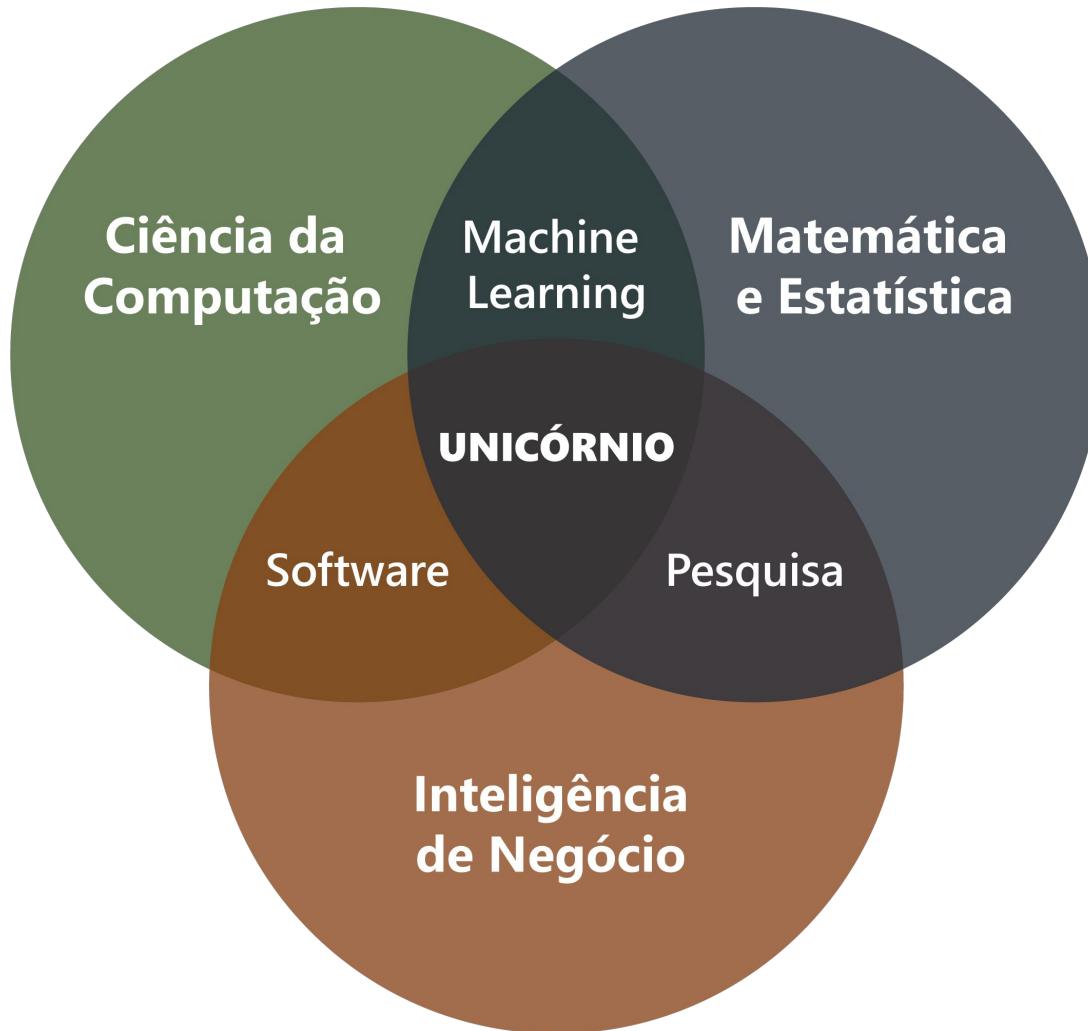
Slice - Seleciona uma faixa de valores, fatia o cubo baseado em uma dimensão

Dice - Seleciona uma faixa de valores baseado em várias dimensões

Drill Across - Cruza métricas que possuem dimensões em comum

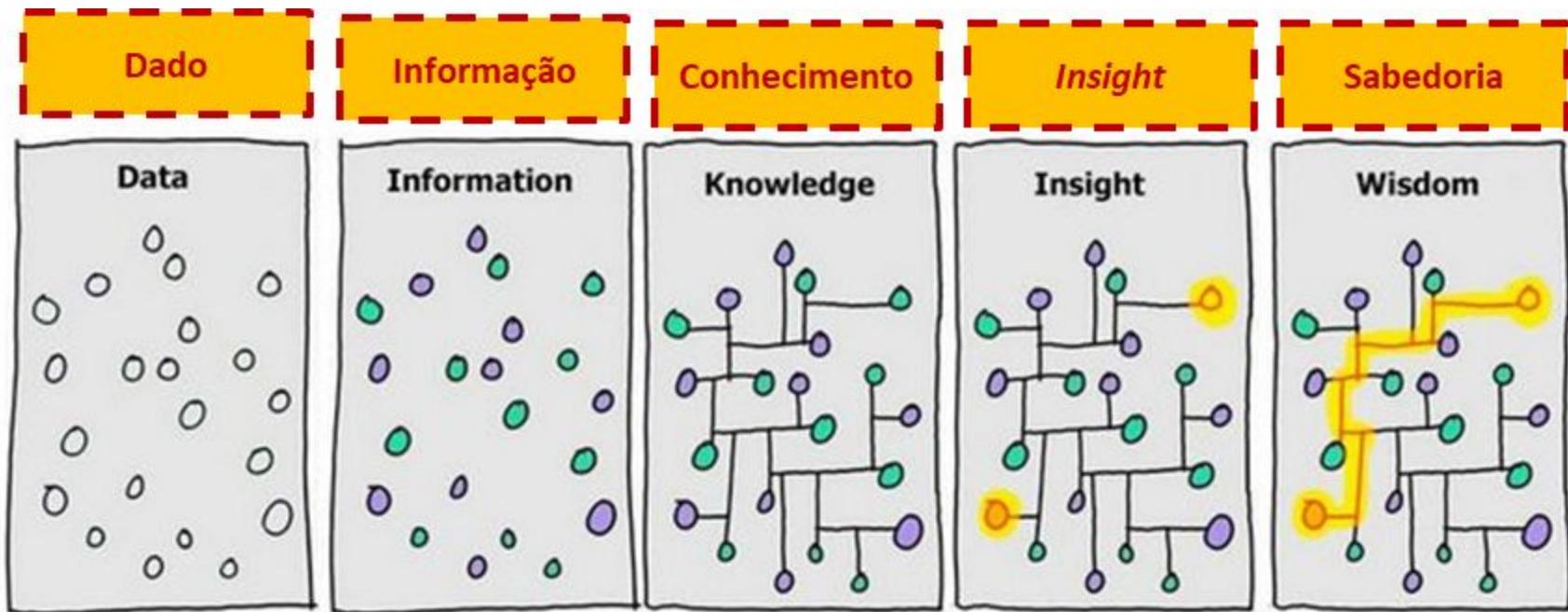


Ciência de Dados



Ciência de Dados

A Ciência de Dados trabalha na descoberta de valor através das informações coletadas, esse trabalho deve atender a expectativa das áreas de negócio para apoiar a tomada de decisão.



Machine Learning

No Machine Learning temos 3 tipos de Aprendizagem de Máquina:

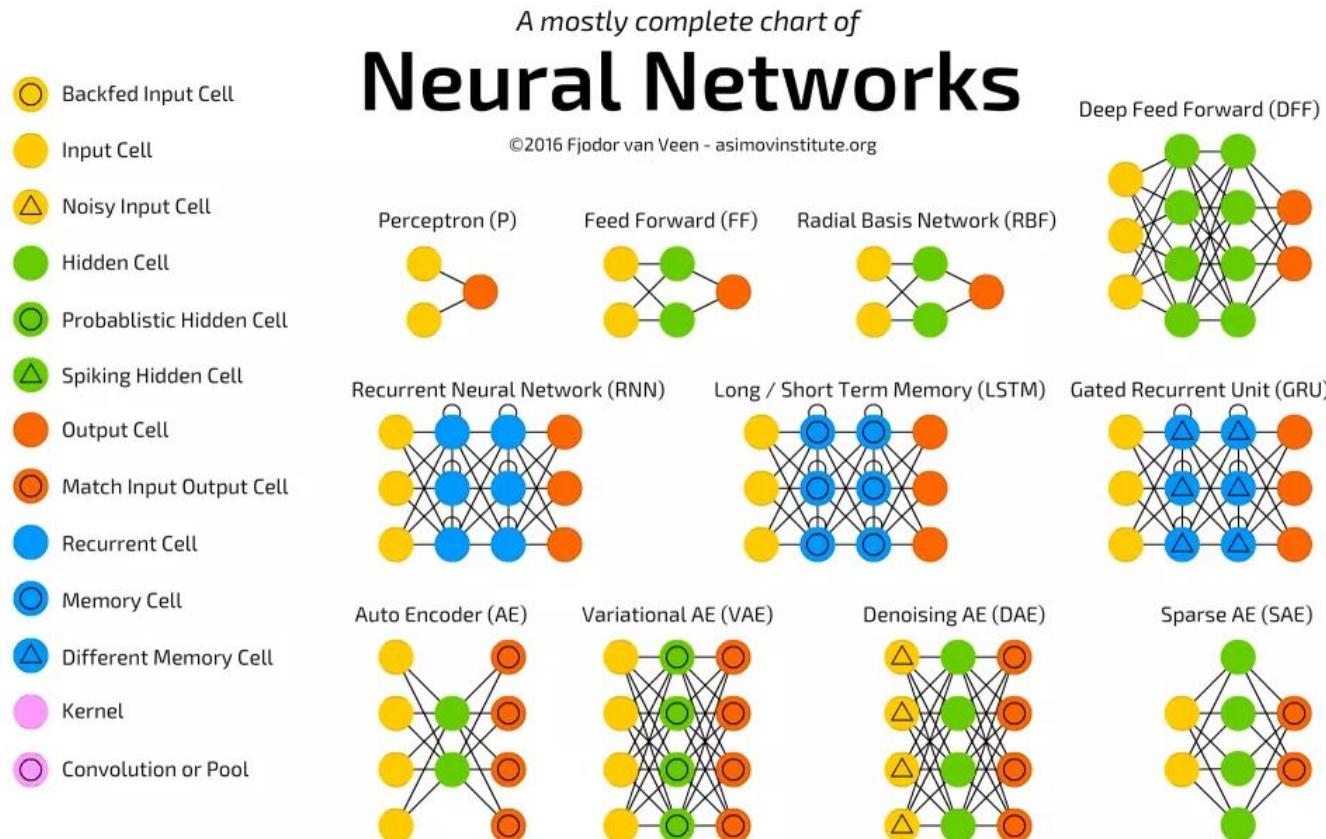
Aprendizagem de Máquina **Supervisionada**: A máquina aprende para fazer previsão baseada em dados históricos e com grupos ou alvos conhecidos. Quando a variável alvo ou classe é conhecida. Há presença de uma base de treinamento.

Aprendizagem de Máquina **Não Supervisionada**: A máquina faz previsão para identificar dados não classificados. Encontrar padrões em dados não rotulados ou de grupos desconhecidos.

Aprendizagem **por Reforço**: A máquina aprende sozinha e o algoritmo é penalizado de acordo com os erros e acertos.

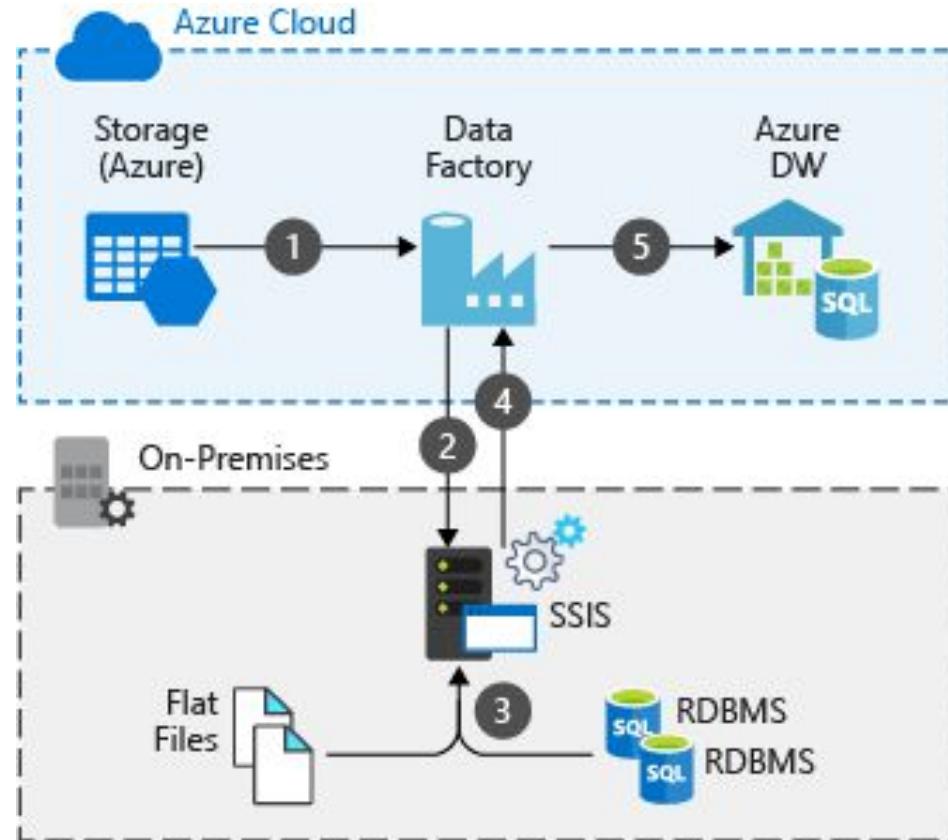
Inteligência Artificial

O Perceptron, neurônio mais simples de rede neural artificial(RNA) foi proposto em 1958 por Frank Rosenblatt, ele simula o neurônio humano, com uma entrada, um peso e uma saída. Os modelos foram evoluindo e hoje temos vários tipos de redes neurais.



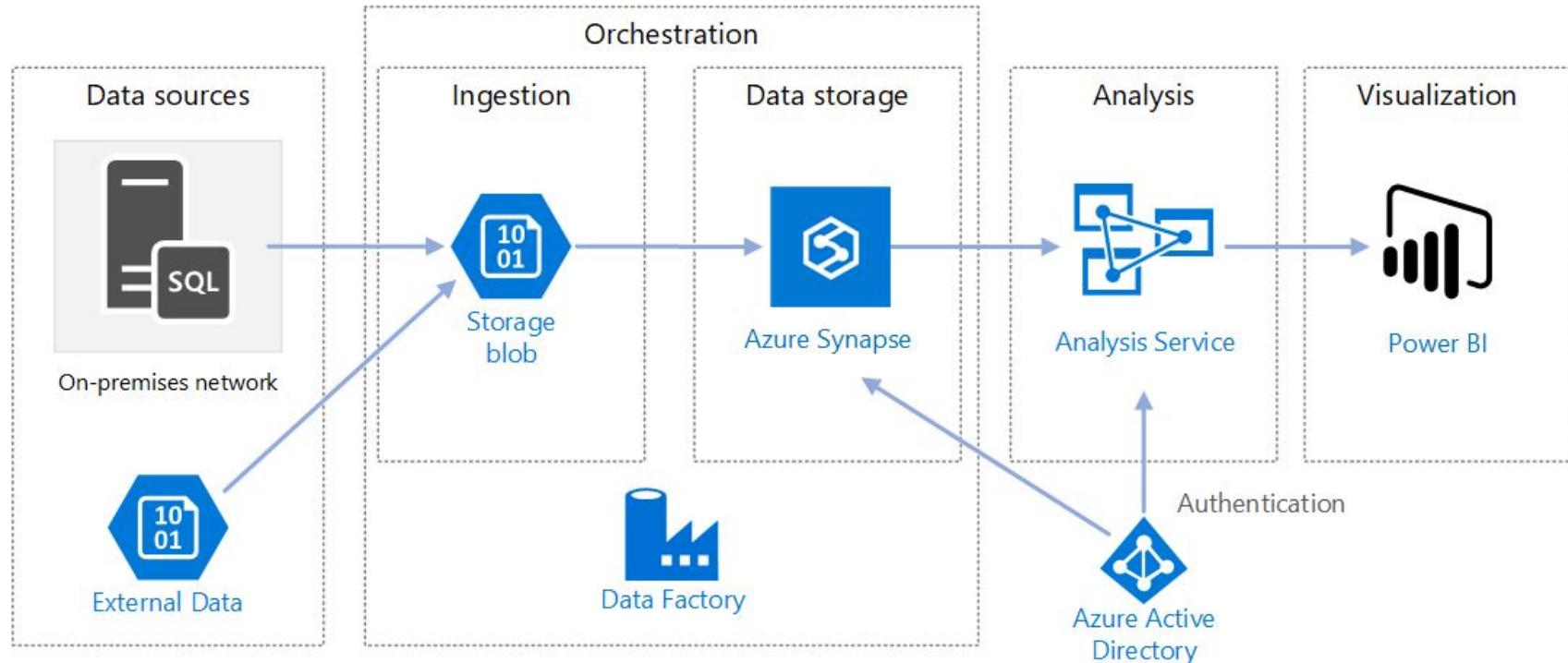
Engenharia de Dados

Engenheiro de dados é responsável por construir o pipeline de dados na empresa, ele acessa os dados na camada operacional e disponibiliza em uma camada separada e acompanha o processo até a extração de valor.



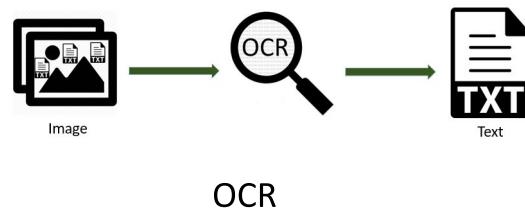
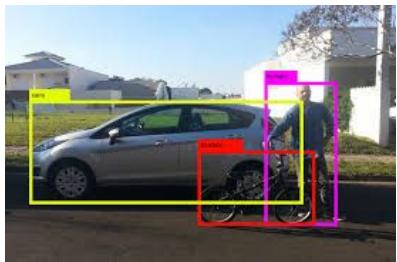
Arquitetura de Soluções

Um arquiteto de solução atua na construção de soluções para atender as necessidades do negócio, fazendo uso dos serviços e recursos de novas tecnologias ou já existentes na empresa, respeitando os padrões e integrações da empresa.

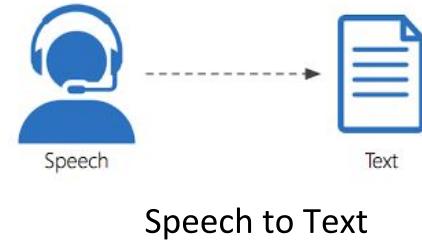


Serviços Cognitivos

Os Serviços em Nuvem já tem pacotes prontos de soluções de Inteligência Artificial, com uma porta e com as chaves de acesso conseguimos incorporar IA em Soluções de Software.

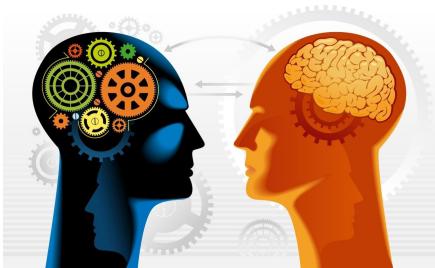


OCR



Speech to Text

Visão Computacional

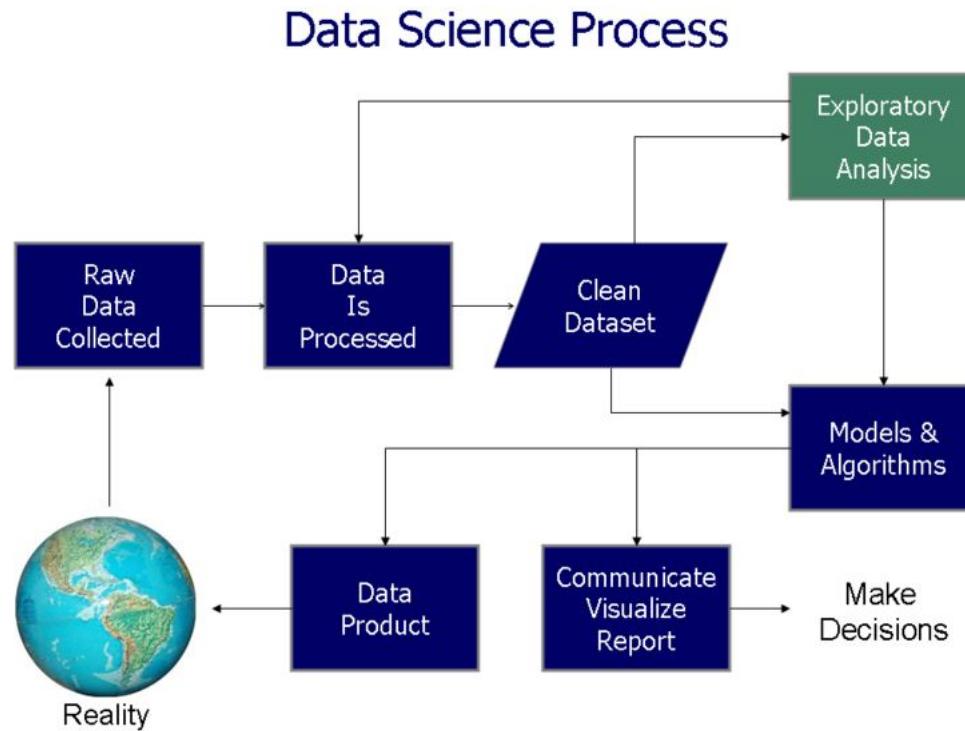


NLP



Machine Learning

Algoritmos - Inteligência Artificial



fonte: https://commons.wikimedia.org/wiki/File:Data_visualization_process_v1.png

Algoritmos - Inteligência Artificial

Introdução à Ciência de Dados, Big Data, Machine Learning e Inteligência Artificial

- Exploração dos Dados
 - Extração, Transformação e Carregamento
 - Linguagem Python
 - Linguagem R
 - Visualização de Dados
 - Seaborn
 - ggplot
 - Tableau
- Modelos
 - Algoritmos
 - Regressão
 - Classificação
 - Clusterização
 - Bibliotecas
 - Scikit-Learn
 - mlr

Algoritmos - Inteligência Artificial

Introdução à Ciência de Dados, Big Data, Machine Learning e Inteligência Artificial

- Exploração dos Dados

```
In [1]: import pandas as pd  
import numpy as np  
import seaborn as sns
```

Carregando dados

```
In [2]: iris = sns.load_dataset("iris")  
iris.__class__
```

```
Out[2]: pandas.core.frame.DataFrame
```

Visualizando dados

iris setosa



iris versicolor



iris virginica



```
In [3]: iris.head()
```

```
Out[3]:
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Algoritmos - Inteligência Artificial

Introdução à Ciência de Dados, Big Data, Machine Learning e Inteligência Artificial

- Exploração dos Dados

```
In [1]: import pandas as pd  
import numpy as np  
import seaborn as sns
```

Carregando dados

```
In [2]: iris = sns.load_dataset("iris")  
iris.__class__
```

```
Out[2]: pandas.core.frame.DataFrame
```

Visualizando dados

```
In [3]: iris.head()
```

```
Out[3]:
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

```
In [4]: iris.describe()
```

```
Out[4]:
```

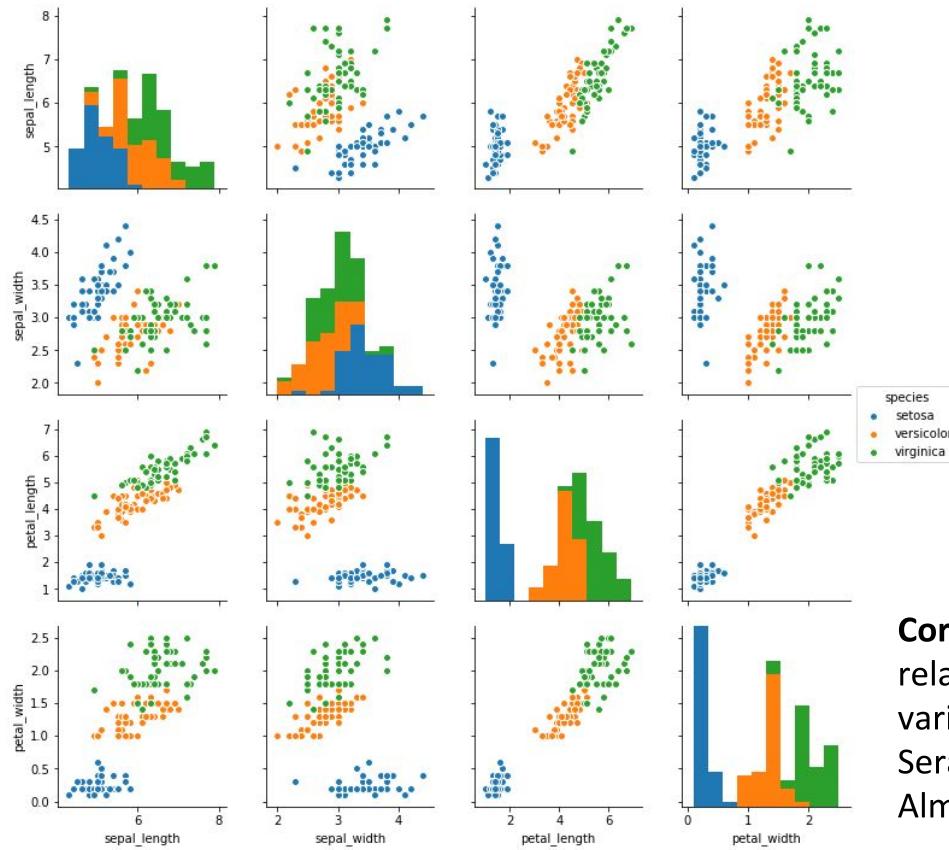
	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Algoritmos - Inteligência Artificial

Introdução à Ciência de Dados, Big Data, Machine Learning e Inteligência Artificial

- Exploração dos Dados

In [5]: g = sns.pairplot(iris, hue="species")



```
corr = iris.corr()
corr.style.background_gradient(cmap='coolwarm')
```

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1	-0.11757	0.871754	0.817941
sepal_width	-0.11757	1	-0.42844	-0.366126
petal_length	0.871754	-0.42844	1	0.962865
petal_width	0.817941	-0.366126	0.962865	1

Correlação, dependência ou associação é qualquer relação estatística (causal ou não causal) entre duas variáveis.

Será que existe relação entre intensidade do Sol e Almoço?

Algoritmos - Inteligência Artificial

Aprendizagem de Máquina **Supervisionada**

Quando a variável alvo ou classe é conhecida. Há presença de uma base de treinamento.

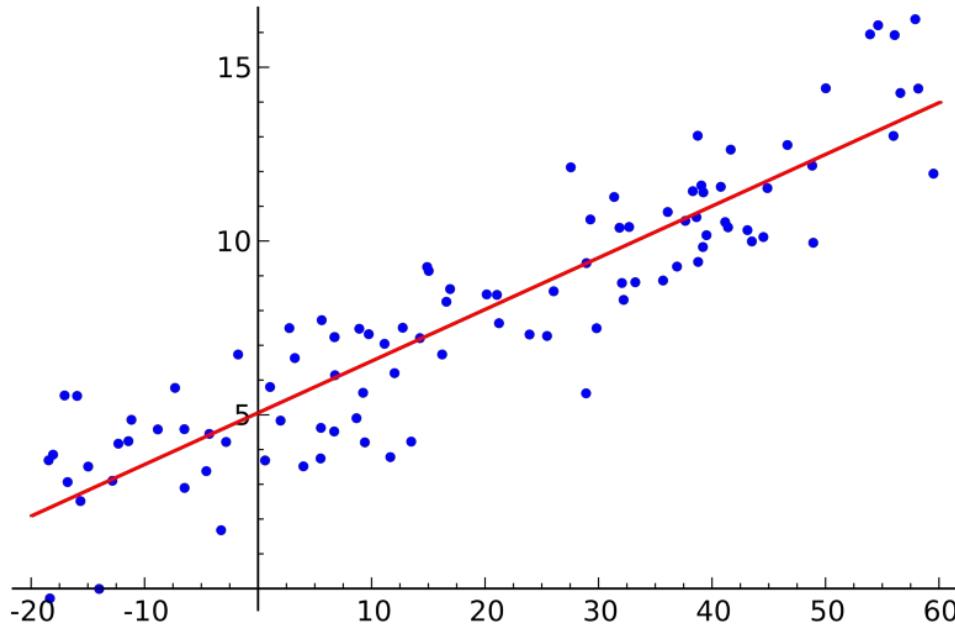
- **Regressão** - Previsão de valores Futuros (previsão de vendas)
- **Classificação** - Classifica os novos dados baseado nos dados Históricos, você já tem uma classificação e conforme os dados são inseridos são classificados pela máquina.
- **Árvore de Decisão** - Fluxo de condições até a melhor tomada de decisão. Muito utilizado para identificar o perfil de clientes. (venda de Seguros, aprovação de crédito)
- ...

Algoritmos - Inteligência Artificial

Aprendizagem de Máquina **Supervisionada**

Quando a variável alvo ou classe é conhecida. Há presença de uma base de treinamento.

- **Régressão** - Previsão de valores Futuros (previsão de vendas)



Algoritmos - Inteligência Artificial

Aprendizagem de Máquina **Supervisionada**

Quando a variável alvo ou classe é conhecida. Há presença de uma base de treinamento.

- **Regressão** - Previsão de valores Futuros (previsão de vendas)

Modelo simples de regressão

```
In [19]: x=iris[ ['petal_length' ] ].values  
y=iris[ ['petal_width' ] ].values  
  
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.5)  
  
from sklearn.linear_model import LinearRegression  
model = LinearRegression()  
model = model.fit(x_train, y_train)
```

```
In [26]: print(iris.head(2))  
print(iris.tail(2))  
  
predictions=model.predict([[1.4], [5.4]])  
  
print(predictions)
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
	sepal_length	sepal_width	petal_length	petal_width	species
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

```
[0.24929514]  
[1.87336908]
```

Algoritmos - Inteligência Artificial

Aprendizagem de Máquina **Supervisionada**

Quando a variável alvo ou classe é conhecida. Há presença de uma base de treinamento.

- **Árvore de Decisão** - Fluxo de condições até a melhor utilização para identificar o perfil de clientes. (vendas)

O x e y da questão

```
In [9]: x=iris[ ['sepal_length', 'sepal_width', 'petal_length', 'petal_width']
y=iris[ ['species'] ].values

from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.5)
```

```
In [10]: from sklearn import tree

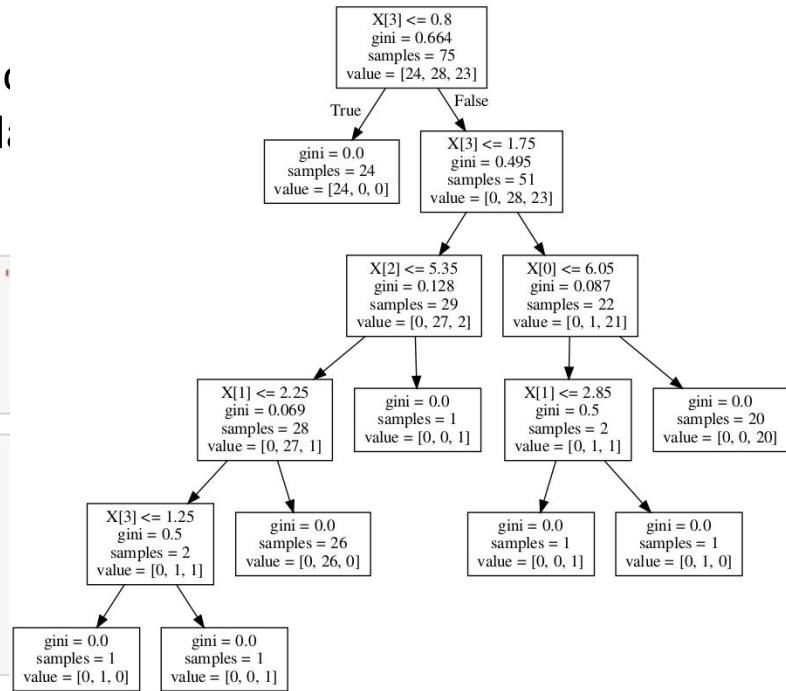
classifier=tree.DecisionTreeClassifier()
classifier.fit(x_train,y_train)
predictions=classifier.predict(x_test)

from sklearn.metrics import accuracy_score
print(accuracy_score(y_test,predictions))

0.9066666666666666
```

```
In [11]: classifier.predict([[4.6,3.1,1.5,0.2]])

Out[11]: array(['setosa'], dtype=object)
```



Algoritmos - Inteligência Artificial

Aprendizagem de Máquina Não Supervisionada:

Classifica os novos dados baseado nos dados Históricos, você já tem uma classificação e conforme os dados são inseridos são classificados pela máquina.

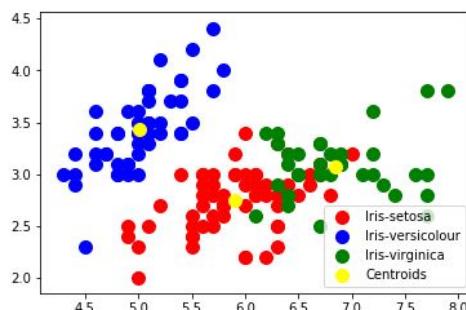
- Clusterização: Encontra padrões nos dados e agrupa de acordo com a similaridade e matriz de distância.

```
In [15]: kmeans = KMeans(n_clusters = 3, init = 'k-means++', max_iter = 300, n_init = 10, random_state = 0)
y_kmeans = kmeans.fit_predict(x)

In [16]: plt.scatter(x[y_kmeans == 0, 0], x[y_kmeans == 0, 1], s = 100, c = 'red', label = 'Iris-setosa')
plt.scatter(x[y_kmeans == 1, 0], x[y_kmeans == 1, 1], s = 100, c = 'blue', label = 'Iris-versicolour')
plt.scatter(x[y_kmeans == 2, 0], x[y_kmeans == 2, 1], s = 100, c = 'green', label = 'Iris-virginica')

plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:,1], s = 100, c = 'yellow', label = 'Centroids')
plt.legend()
```

Out[16]: <matplotlib.legend.Legend at 0x7ff230ba7da0>



Lembrando que não temos **PRÉVIAMENTE** os rótulos, estamos utilizando aqui para tornar mais clara a explicação.

Algoritmos - Inteligência Artificial

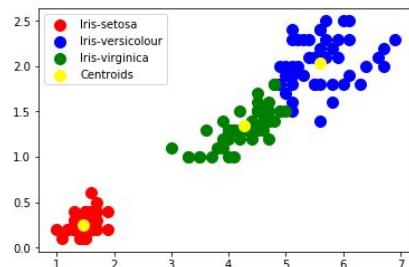
Aprendizagem de Máquina Não Supervisionada:

Classifica os novos dados baseado nos dados Históricos, você já tem uma classificação e conforme os dados são inseridos são classificados pela máquina.

- Clusterização: Encontra padrões nos dados e agrupa de acordo com a similaridade e matriz de distância.

Através da exploração dos dados...

```
In [17]: x=iris[ ['petal_length', 'petal_width' ] ].values  
y=iris[ ['species'] ].values  
  
In [18]: kmeans = KMeans(n_clusters = 3, init = 'k-means++', max_iter = 300, n_init = 10, random_state = 0)  
y_kmeans = kmeans.fit_predict(x)  
  
plt.scatter(x[y_kmeans == 0, 0], x[y_kmeans == 0, 1], s = 100, c = 'red', label = 'Iris-setosa')  
plt.scatter(x[y_kmeans == 1, 0], x[y_kmeans == 1, 1], s = 100, c = 'blue', label = 'Iris-versicolour')  
plt.scatter(x[y_kmeans == 2, 0], x[y_kmeans == 2, 1], s = 100, c = 'green', label = 'Iris-virginica')  
  
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:,1], s = 100, c = 'yellow', label = 'Centro')  
plt.legend()  
  
Out[18]: <matplotlib.legend.Legend at 0x7ff230b21ac8>
```



Lembrando que não temos **PRÉVIAMENTE** os rótulos, estamos utilizando aqui para tornar mais clara a explicação.

Algoritmos - Inteligência Artificial

Aprendizagem por Reforço:

Exemplos:

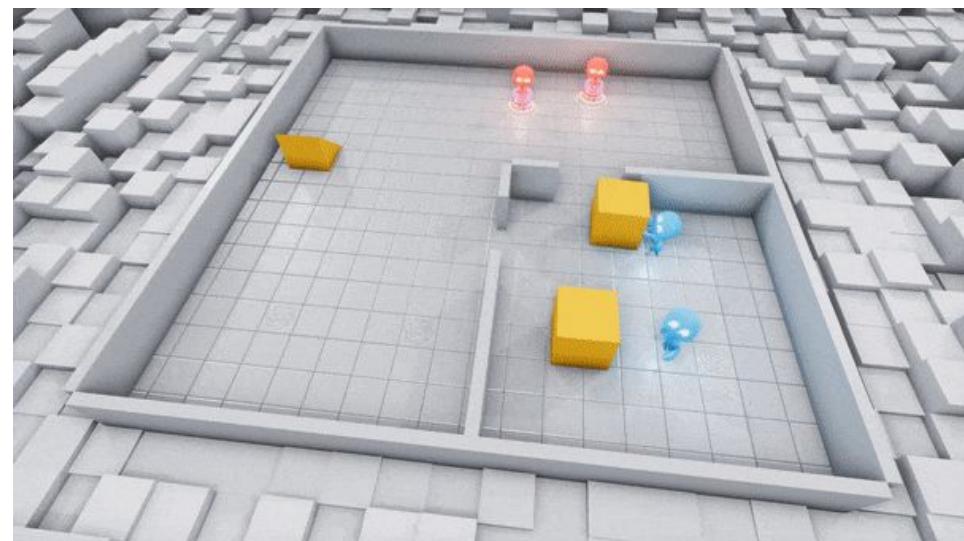
- O Algoritmo recebe pontuação para os erros e acertos.
- Carros Autônomos
- Robôs de Jogos

Algoritmos - Inteligência Artificial

Aprendizagem por Reforço/Deep Learning:

OpenAI Tried to Train AI Agents to Play Hide-And-Seek but Instead They Were Shocked by What They Learned

“... Initially, hiders and seekers learn to crudely run away and chase ... learn to use the tools at their disposal and intentionally modify their environment ... begin to construct secure shelters in which to hide by moving many boxes together or against walls and locking them in place... learn to move and use ramps to jump over obstacles ... hiders learn to bring the ramps to the edge of the play ...”

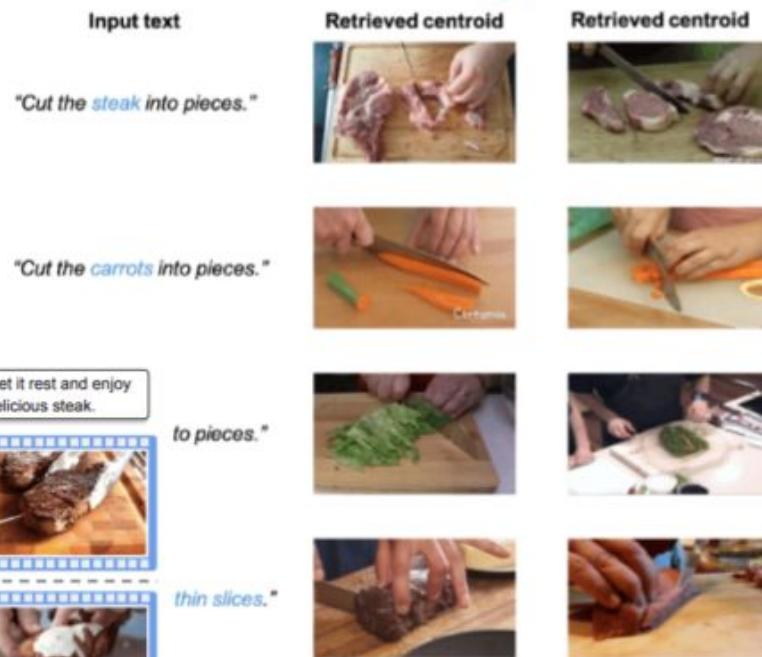


Algoritmos - Inteligência Artificial

Deep Learning:

What's Cooking? Google VideoBERT Predicts Recipes

... Model text and video representation... Given a few video frames that include a bowl of flour and cocoa powder, VideoBERT can speculate that the following video frames might involve baking a brownie or cupcake... Technically predicting missing word tokens or video tokens..."



Algoritmos - Inteligência Artificial

Deep Learning:

Researchers train AI to map a person's facial movements to any target headshot

“... “significant” mismatch between the face to be manipulated and the person doing the manipulating ... synthesize a reenacted face animated by the movement of a person (a “driver”) while preserving the face’s (target’s) appearance ...”



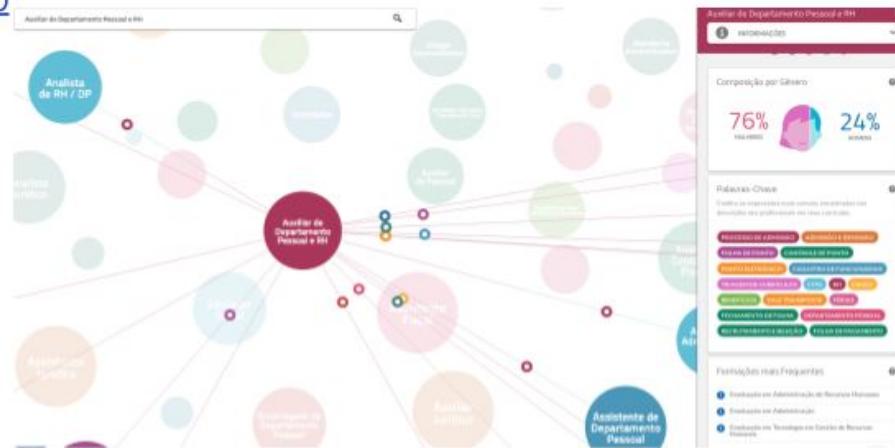
Algoritmos - Inteligência Artificial

Realidade mais próxima:



“... Processamento de 10 milhões de currículos, normalização dos nomes dos cargos ... Probabilidade dos passos da carreira ...”

<https://www.vagas.com.br/mapa-de-carreiras/cargos/auxiliar-de-departamento-pessoal-e-rh/0>



Algoritmos - Inteligência Artificial

Realidade mais próxima:

Sistema de Recomendação

"... Entendi o problema implementamos uma PoC e depois a solução final ... Dobrou a quantidade de candidaturas, gerando um negócio novo de publicidade ..."

<https://www.vagas.com.br/vagas/v1938992/programador>

Estagiário de TI - Desenvolvimento
v1945436
DIRECIONAL ENGENHARIA

Salário a combinar Belo Horizonte Estágio Expira em 10 dias

Programador
v1938992
Confidencial

Salário a combinar Belo Horizonte Técnico Expira em 6 dias

Atuar com desenvolvimento e customização de software para plataformas desktop, web e mobile.

Desenvolvimento, manutenção, evolução e correção de softwares.

Realizar análise de requisitos e levantamento de funcionalidades com clientes para desenvolvimento de projetos de software.

Realizar controle e versionamento de software, interagir com desenvolvedores, porte técnico e outros profissionais das filiais localizadas no exterior.

Realizar atendimento e suporte técnico para clientes via email, telefone, conexão remota e ferramenta própria de atendimento de chamados.

Requer prévio conhecimento em: Java, C#, PHP, HTML5, CSS3, JavaScript, jQuery,

Kona Container Car Guy
Delivering Paradise One Family At A Time Since 1999. Visit Our Website Now!

Kona Container/Car Guy OPEN >

Veja vagas similares

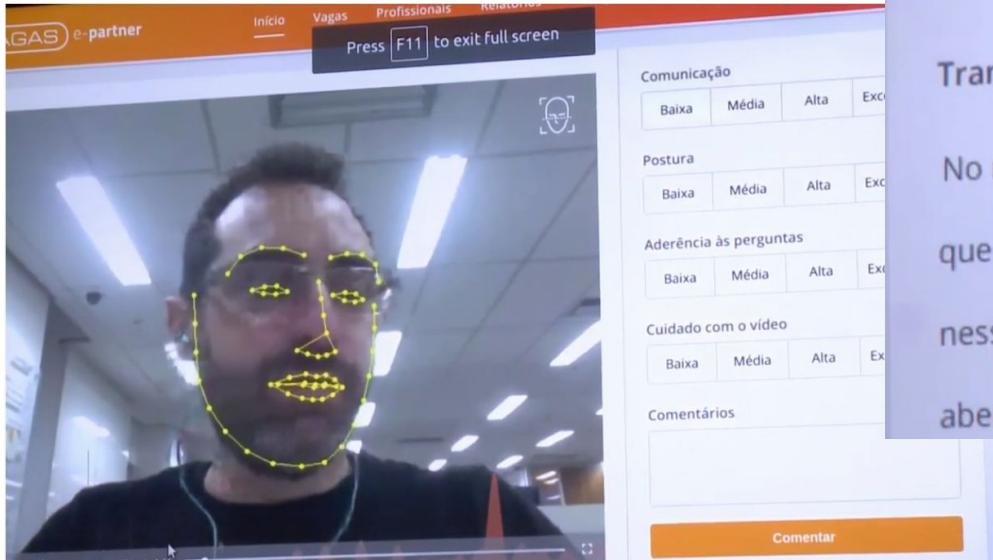
- * **Estagiário de TI - Desenvolvimento**
Estágio - DIRECIONAL ENGENHARIA
- * **Analista de Suporte Júnior**
Júnior/Trainee - Soloc Softwares
- * **Analista de Suporte Auxiliar/Operacional**
- HITSS
- * **Analista Desenvolvedor Júnior**
Júnior/Trainee
- * **Estágio em T.I. - Suporte Técnico**
Técnico - VInCI Energies

Algoritmos - Inteligência Artificial

Realidade mais próxima:

Video Entrevista

“... Processos com muitos candidatos ... Tempo e custo em agendar e a entrevista pessoalmente ... Assistir milhares de respostas ...”



melhores tento **observar** pego seletivo vagas mun empresa
vou nele **aberto** possuem contratações soft skills encontram necessárias

Transcrição da entrevista

No meu dia a dia eu pego um processo **seletivo** aberto e v que estão nele e eu tento **observar** todos os tipos e soft Skills q nesse processo Eu Tentei encontrar os melhores **candidatos** aberto para mim fazendo esse **processo** eu vou melhorar

Exercícios

- 1 - Quais são os 3 Vs do Big Data?
- 2 - Porque os outros 2 Vs não são considerados?
- 3 - O que são dados Distribuídos e para que servem?
- 4 - Qual o trabalho de um Analista de BI?
- 5 - Qual o trabalho de um Cientista de Dados?
- 6 - Qual o Trabalho de um Arquiteto de Soluções?
- 7 - Qual o Trabalho de um Engenheiro de Dados?
- 8 - Qual dessas três etapas utilizamos no código em aula?
- 9 - Qual dos tipos de algoritmos que não precisamos de rótulos?

Bibliografia

wikiwand.com/fr/Traitement_analytique_en_ligne

docs.microsoft.com

pt.wikipedia.org/wiki/Rede_neural_artificial

deeplearningbook.com.br/uma-breve-historia-das-redes-neurais-artificiais/

docs.microsoft.com/pt-br/azure/cognitive-services/welcome

Contatos

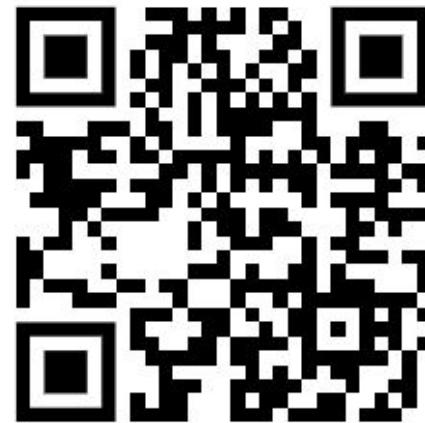


Andréa Longarini
Cientista de Dados
Profa. de Pós Graduação



<https://br.linkedin.com/in/andrea-longarini-2421325b>

Contatos



Thiago Kuma
Cientista de Dados
Professor



<https://www.linkedin.com/in/thiago-kuma/>