

Técnicas para o Processamento do Big Data

Capítulo 1. Introdução

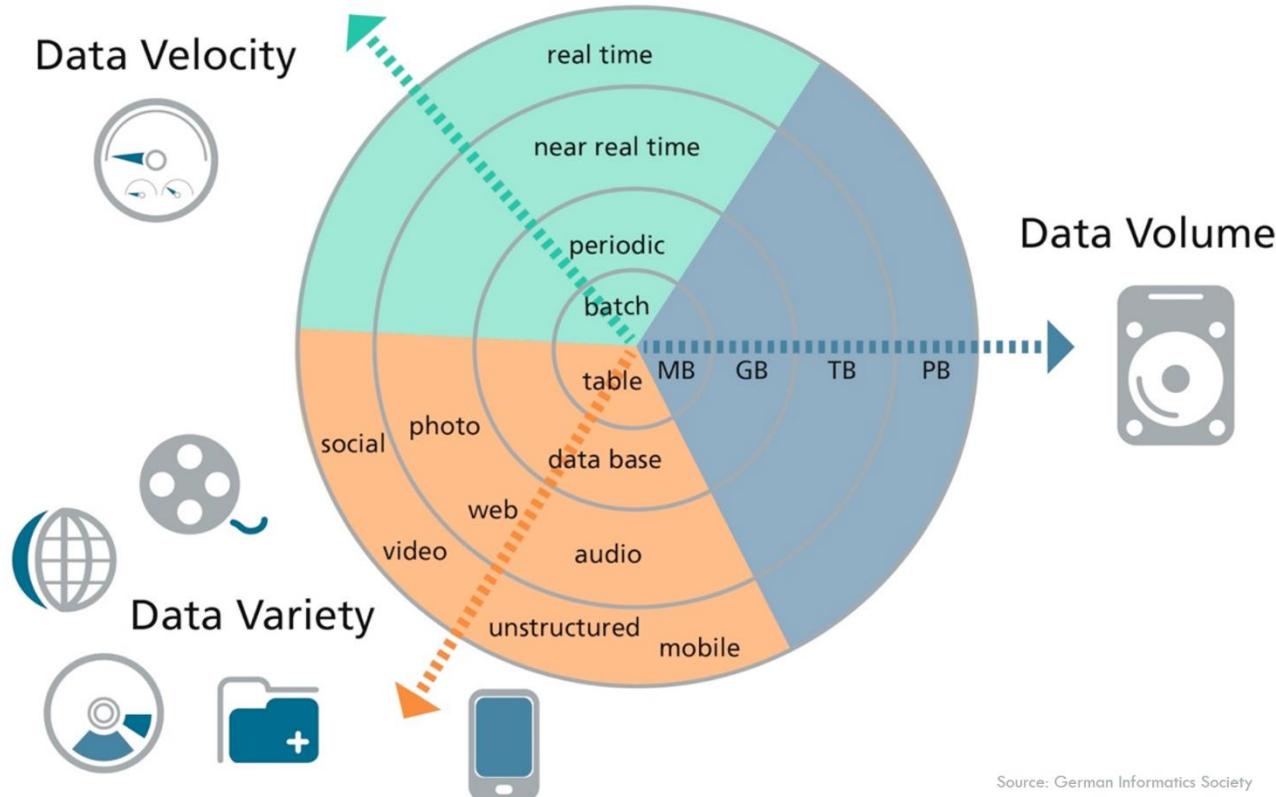
Prof. Túlio Philipe Vieira



Aula 1.1. Processamento do Big Data

- Por que Big Data?
- Por que processar o Big Data?
- O poder do processamento do Big Data através do machine learning.
- Etapas para aplicar o machine learning.

O Big Data



Source: German Informatics Society

O que fazer com todos esses dados?



- É a ciência que utiliza os dados para construir modelos que levam a uma melhor decisão. Essa decisão é capaz de adicionar valor às pessoas, companhias e instituições.

– Dimitris Bertsimas



Por que agora é tão importante?

- Encontrar o cliente ideal;



Por que agora é tão importante?

- Encontrar o cliente ideal;
- Otimizar o engajamento do consumidor;



Por que agora é tão importante?

- Encontrar o cliente ideal;
- Otimizar o engajamento do consumidor;
- Otimização do marketing e performance;



Por que agora é tão importante?

- Encontrar o cliente ideal;
- Otimizar o engajamento do consumidor;
- Otimização do marketing e performance;
- Redução dos custos;



Por que agora é tão importante?

- Encontrar o cliente ideal;
- Otimizar o engajamento do consumidor;
- Otimização do marketing e performance;
- Redução dos custos;
- Personalização de produtos em tempo real;



Por que agora é tão importante?

- Encontrar o cliente ideal;
- Otimizar o engajamento do consumidor;
- Otimização do marketing e performance;
- Redução dos custos;
- Personalização de produtos em tempo real;
- Pesquisa de mercado;



Por que agora é tão importante?

- Encontrar o cliente ideal;
- Otimizar o engajamento do consumidor;
- Otimização do marketing e performance;
- Redução dos custos;
- Personalização de produtos em tempo real;
- Pesquisa de mercado;
- Gerenciamento de reputação;

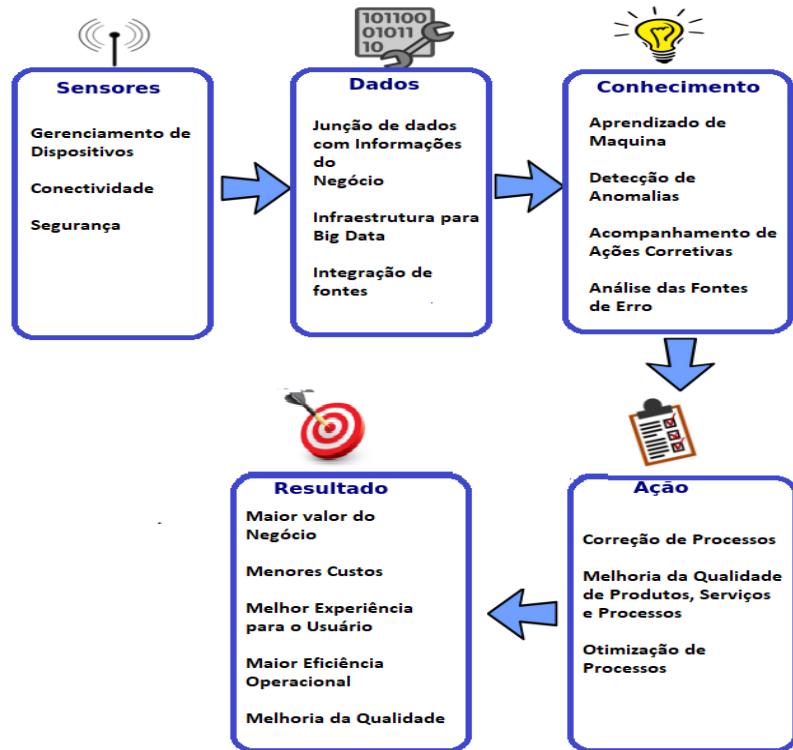


Por que agora é tão importante?

- Encontrar o cliente ideal;
- Otimizar o engajamento do consumidor;
- Otimização do marketing e performance;
- Redução dos custos;
- Personalização de produtos em tempo real;
- Pesquisa de mercado;
- Gerenciamento de reputação;
- Análise dos competidores.



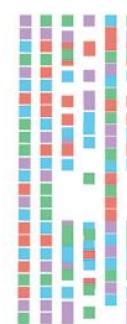
Importância do ML para geração de Valor



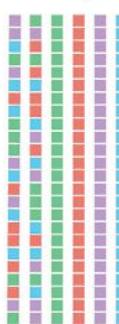
BIG DATA



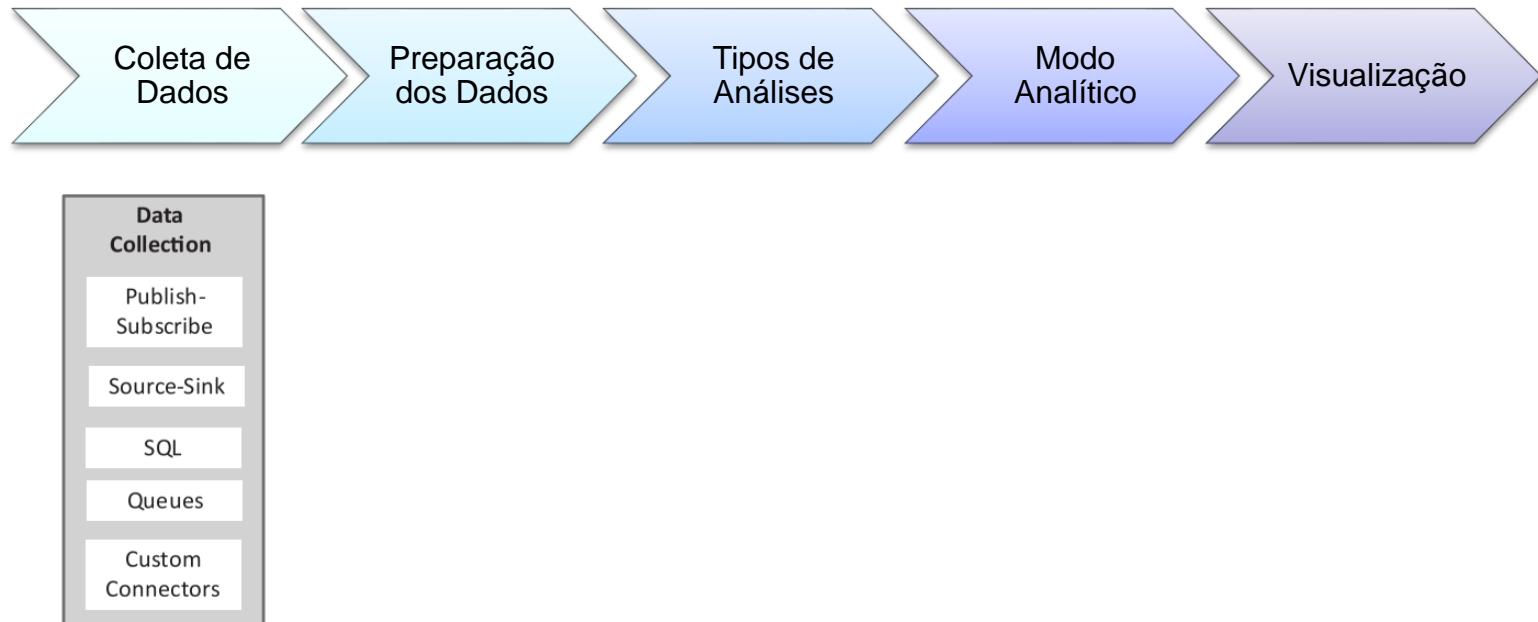
ANÁLISE



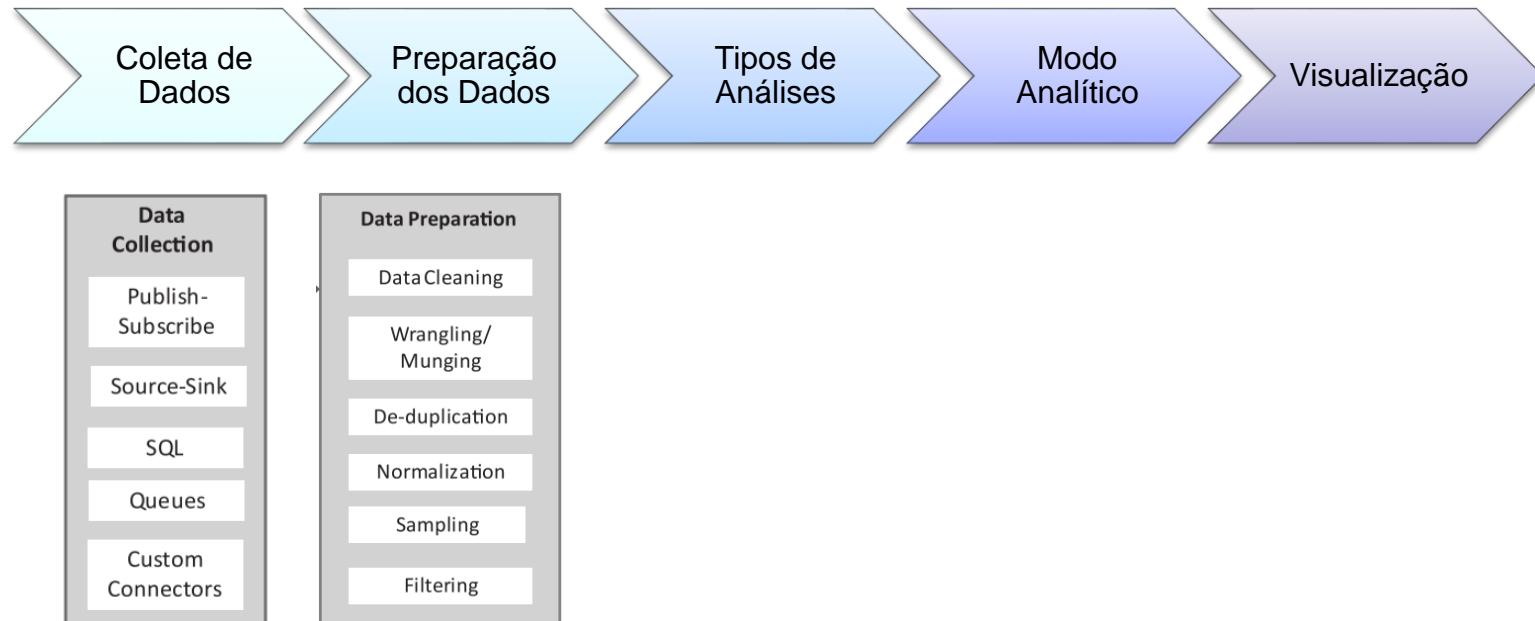
SOLUÇÃO



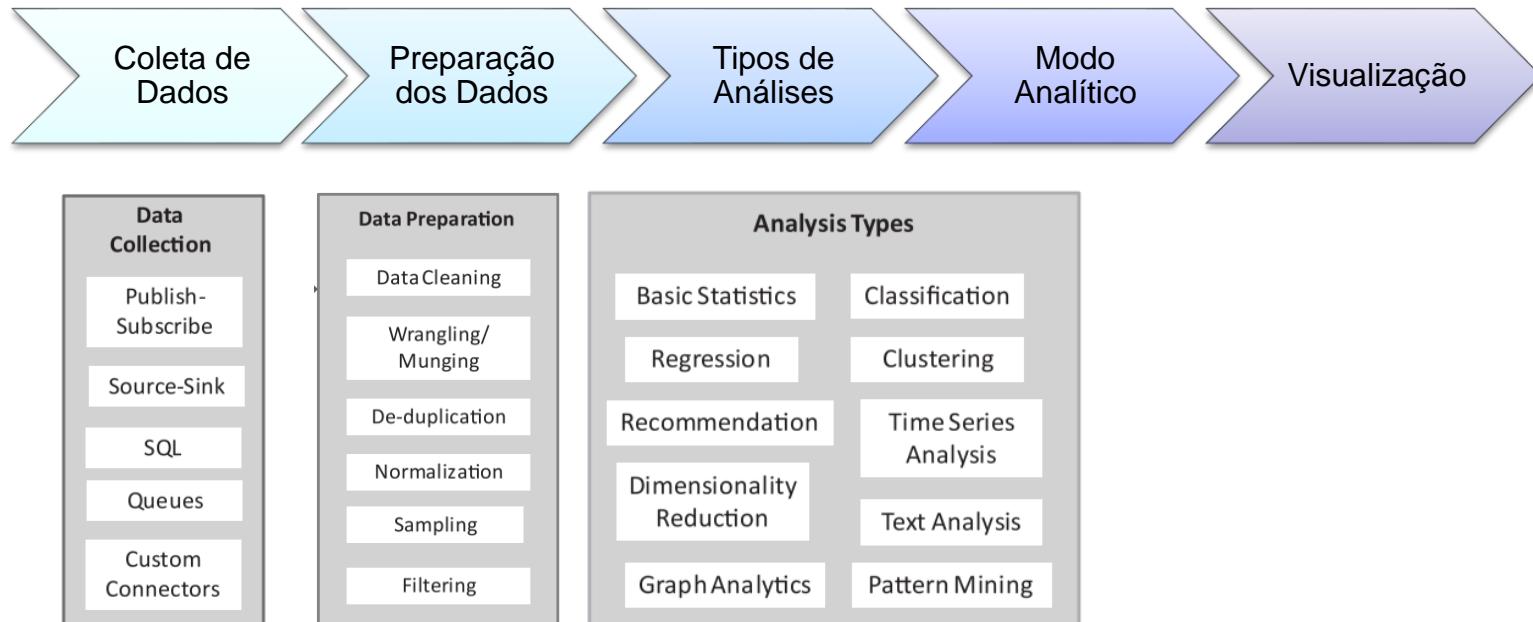
Etapas



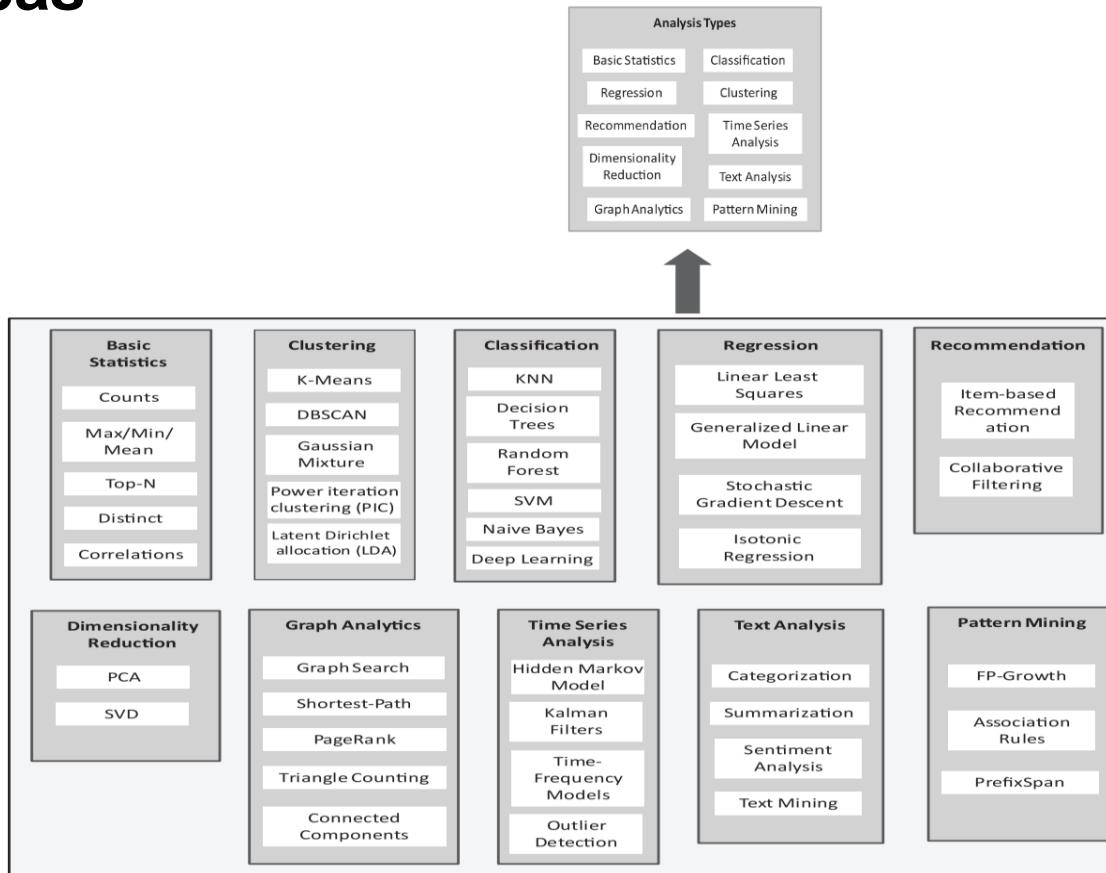
Etapas



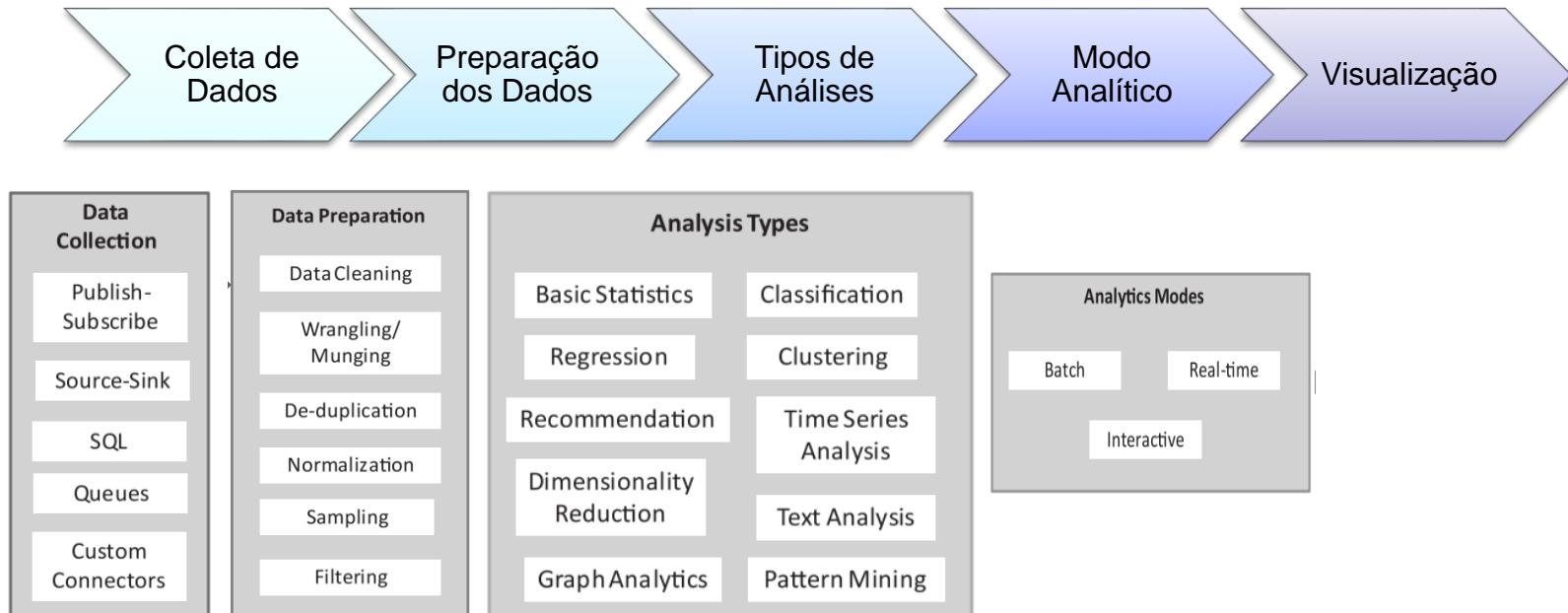
Etapas



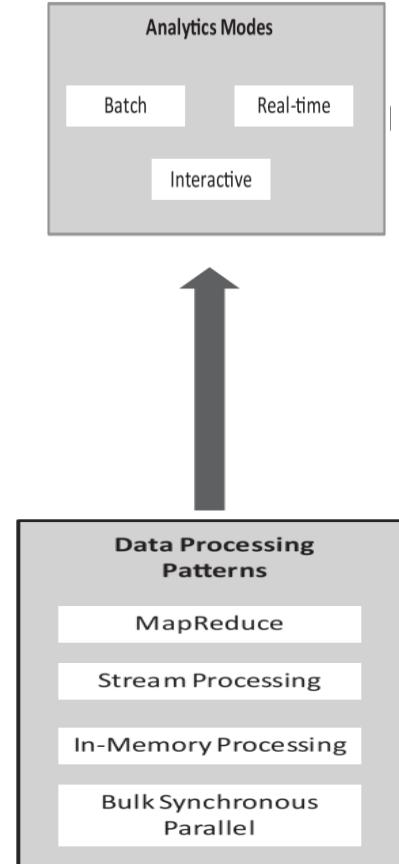
Etapas



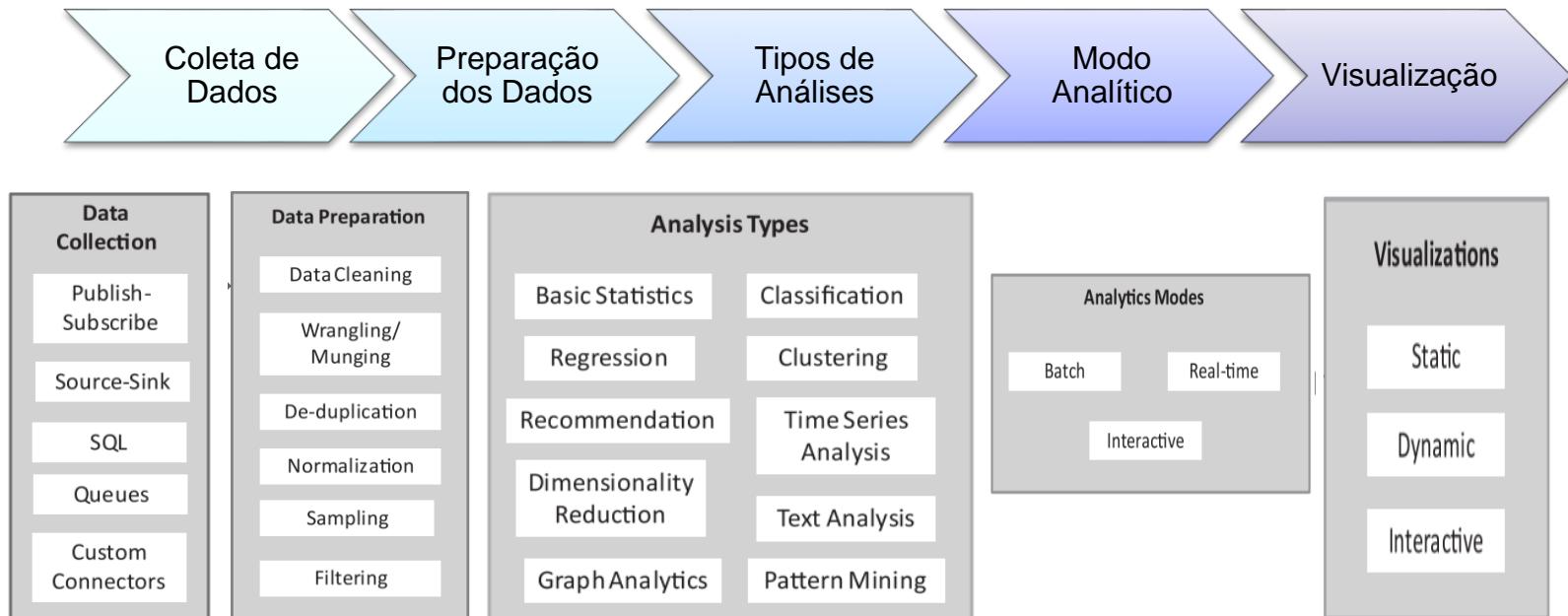
Etapas



Etapas



Etapas



- O por quê do Big Data.
- Porque processar o Big Data.
- O poder gerado pelo processamento do Big Data com ML.
- Etapas para aplicação do ML.

■ Próxima aula

- ❑ Pilha para processamento do Big Data.

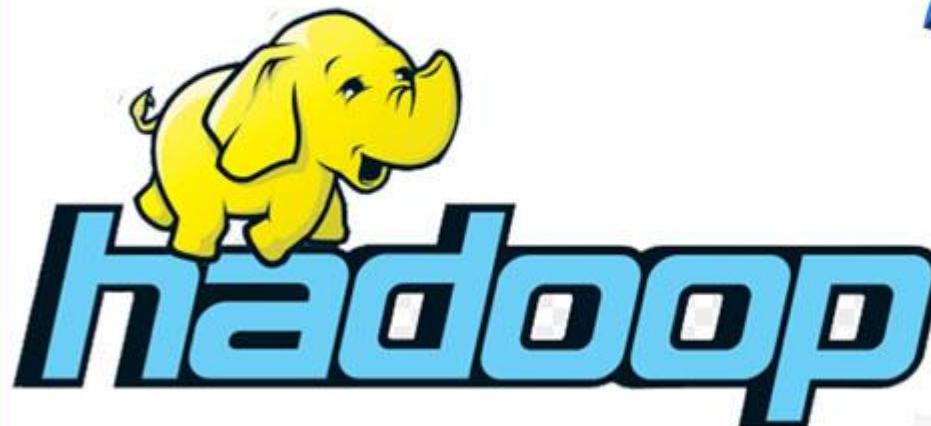


Aula 1.2. Pilha para processamento do Big Data

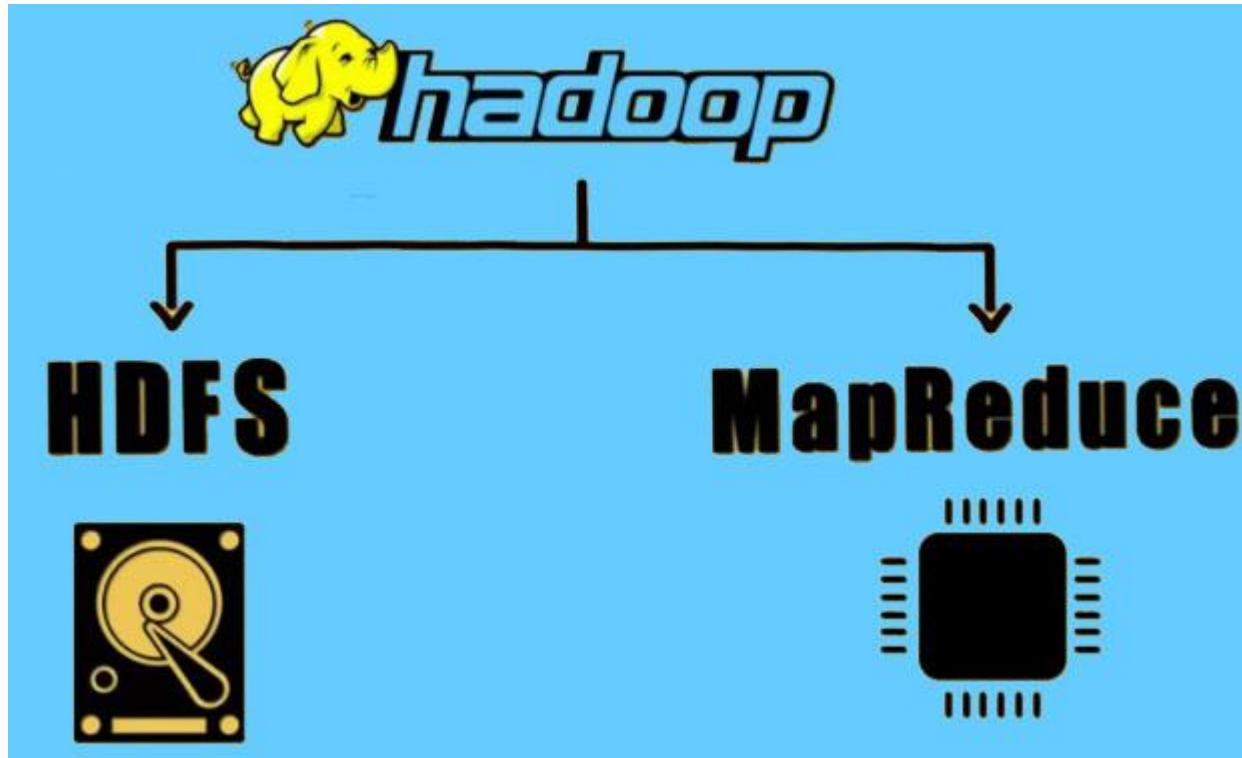
Nesta aula

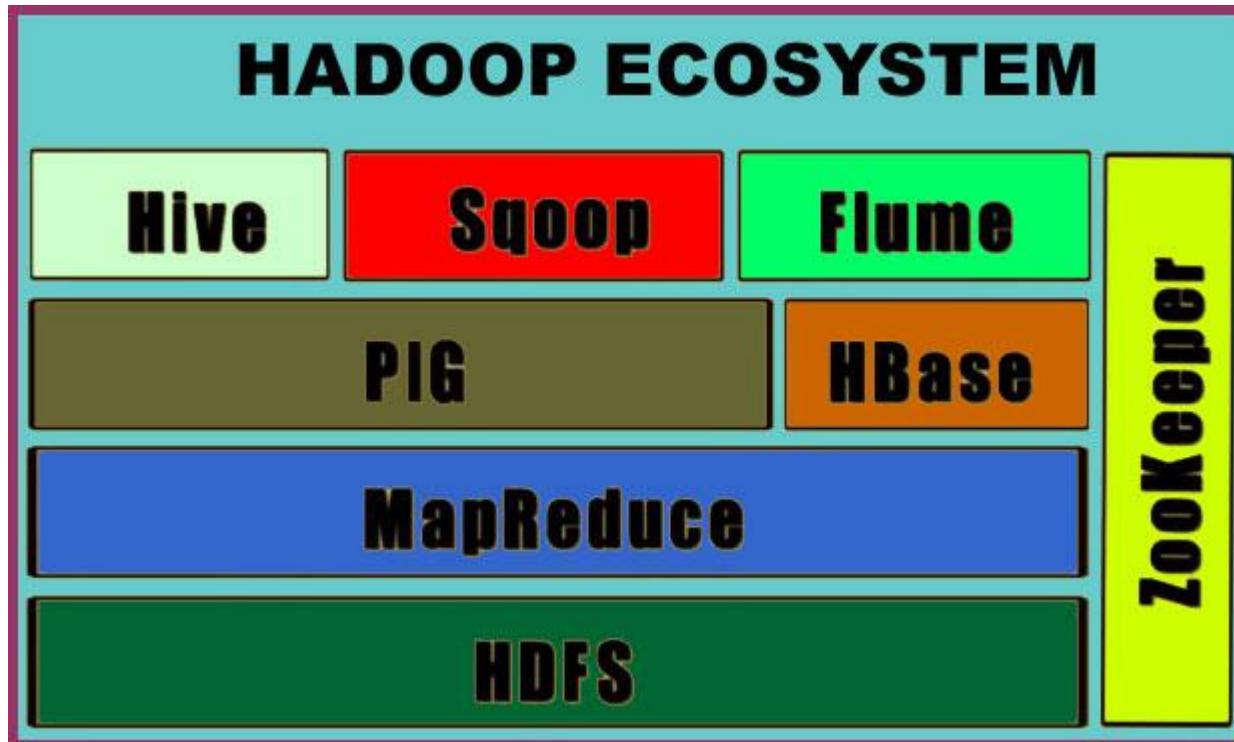
- Hadoop.
- Pilha para processamento.
- Mapeamento do fluxo de análise.

WHAT IS HADOOP?



Hadoop componentes

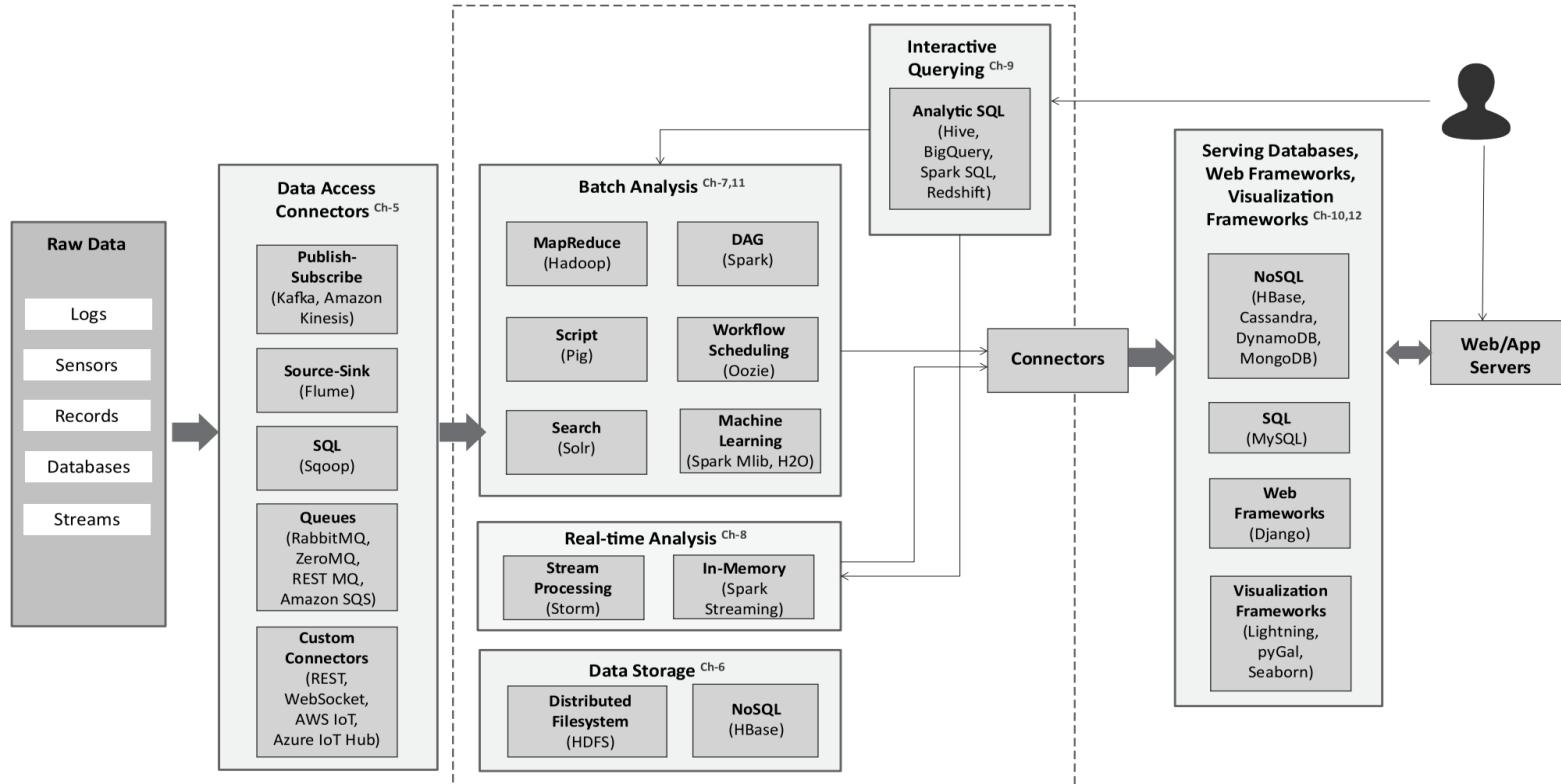




Pilha para processamento do Big Data



Ferramentas



Fonte: BAHGA e MADISSETTI (2016)

Mapeamento do fluxo de análise

Data Collection

Analysis Type	Framework (Mode)
Publish-Subscribe	Kafka, Kinesis
Source-Sink	Flume
SQL	Sqoop
Queues	SQS, RabbitMQ, ZeroMQ, RESTMQ
Custom Connectors	REST, WebSocket, MQTT

Data Preparation

Analysis Type	Framework
Data Cleaning	Open Refine
Data Wrangling	Open Refine DataWrangler
De-Duplication	Open Refine, Pig, Hive, Spark SQL
Normalization	MapReduce, Pig, Sampling, Hive, Spark SQL Filtering

Basic Statistics

Analysis Type	Framework (Mode)
Counts, Max, Min, Mean, Top-N, Distinct	Hadoop-MapReduce (Batch), Pig (Batch), Spark (Batch), Spark Streaming (Realtime), Spark SQL (Interactive), Hive (Integrative), Storm (Real-time)
Correlations	Hadoop-MapReduce (Batch), Spark Mlib (Batch)

Clustering

Analysis Type	Framework (Mode)
K-Means	Hadoop-MapReduce (Batch), Spark Mlib (Batch & Real-time) H2O (Batch)
DBSCAN	Spark (Batch)
Gaussian Mixture	Spark Mlib (Batch)
PIC	Spark Mlib (Batch)
LDA	Spark Mlib (Batch)

Classification

Analysis Type	Framework (Mode)
KNN	Spark Mlib (Batch , Realtime)
Decision Trees	Spark Mlib (Batch, Realtime)
Random Forest	Spark Mlib (Batch , Realtime), H2O (Batch)
SVM	Spark Mlib (Batch , Realtime)
Naïve Bayes	Spark Mlib (Batch, Realtime), H2O (Batch)
Deep Learning	H2O (Batch)

Regression

Analysis Type	Framework (Mode)
Linear Least Squares	Spark Mlib (Batch, Realtime)
Generalized Linear Model	H2O (Batch)
Stochastic Gradient Descent	Spark Mlib (Batch, Realtime)
Isotonic Regression	Spark Mlib (Batch, Realtime)

Fonte: BAHGA e MADISSETTI (2016)

Mapeamento do fluxo de análise

Graph Analytics

Analysis Type	Framework (Mode)
Graph Search	Spark GraphX (Batch)
Shortest-Path	Spark GraphX (Batch)
PageRank	Spark GraphX (Batch)
Triangle Counting	Spark GraphX (Batch)
Connected Components	Spark GraphX (Batch)

Time Series Analysis

Analysis Type	Framework (Mode)
Kalman Filter	Spark (Realtime)
Time Frequency Models	Spark (Realtime)

Dimensionality Reduction

Analysis Type	Framework (Mode)
SVD	Spark Mlib (Batch)
PCA	Spark Mlib (Batch), H2O (Batch)

Recommendation

Analysis Type	Framework (Mode)
Item-bases Recommendation	Spark Mlib (Batch)
Collaborative Filtering	Spark Mlib (Batch)

Text Analysis

Analysis Type	Framework (Mode)
Categorization	Hadoop-MapReduce (Batch), Storm (Realtime), Spark (Batch, Realtime)
Summarization	Spark (Batch)
Sentiment Analysis	Storm (Realtime), Spark (Batch, Realtime)
Text Mining	Storm (Realtime), Spark (Batch, Realtime)

Pattern Mining

Analysis Type	Framework (Mode)
FP-Growth	Spark Mlib (Batch)
Association Rules	Spark Mlib (Batch)
PrefixSpan	Spark Mlib (Batch)

Visualization

Analysis Type	Framework (Mode)
Web Frameworks	Django, Flask
SQL Databases	MySQL
NoSQL Databases	Hbase, DynamoDB, Cassandra, MongoDB
Visualization Frameworks	Lightning, pyGal, Seaborn

Fonte: BAHGA e MADISSETTI (2016)

Conclusão

- Hadoop.
- Pilha para processamento.
- Mapeamento do fluxo de análise.

■ Próxima aula

- Arquiteturas para o processamento do Big Data.



Aula 1.3. Arquiteturas de processamento

- ❑ Por que é necessário uma arquitetura?
- ❑ Tipos de arquiteturas.
- ❑ Diferenças entre arquiteturas.
- ❑ Qual arquitetura escolher?
- ❑ Aplicação x Arquitetura → Exemplos.

■ Por que é necessário uma arquitetura?

- Tratar grandes volumes de dados;



■ Por que é necessário uma arquitetura?

- Tratar grandes volumes de dados;
- Alta velocidade de processamento;



■ Por que é necessário uma arquitetura?

- Tratar grandes volumes de dados;
- Alta velocidade de processamento;
- Tolerância a falhas;



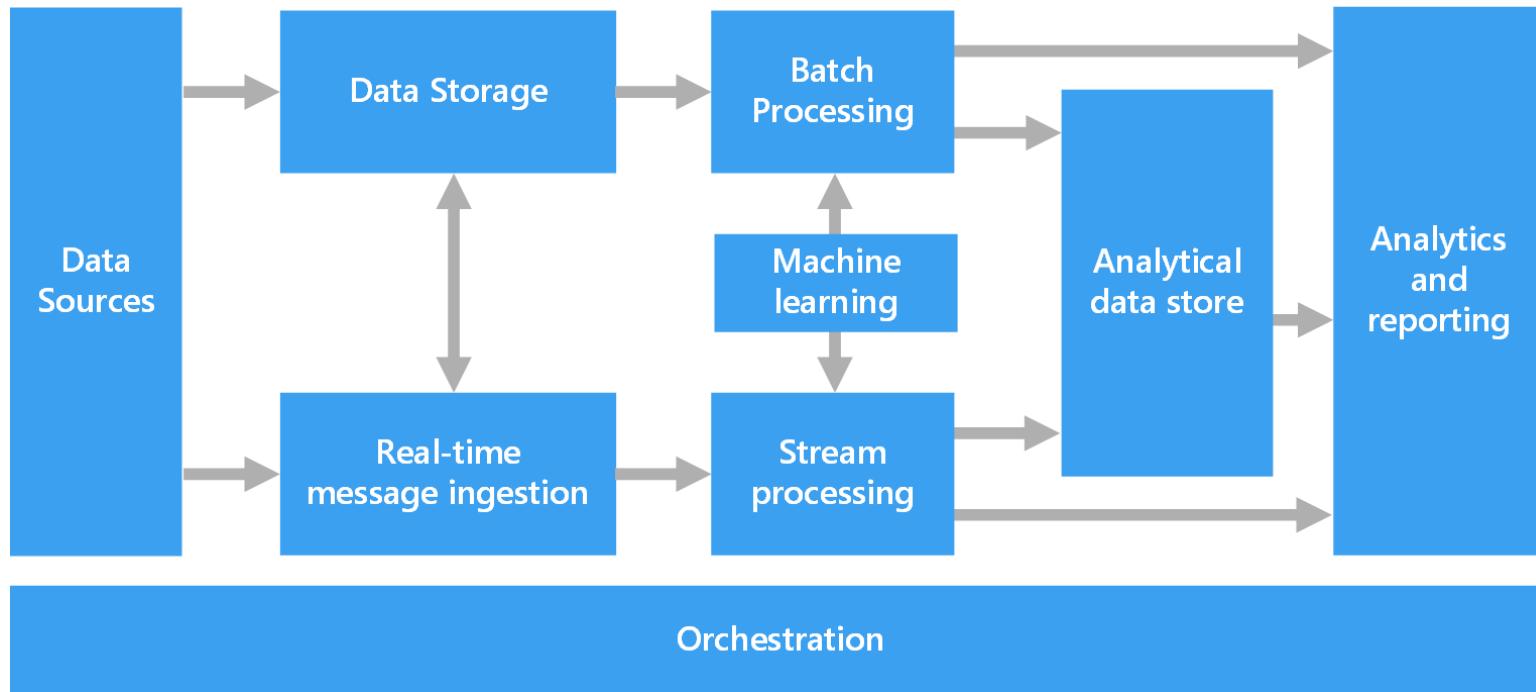
■ Por que é necessário uma arquitetura?

IGTI

- Tratar grandes volumes de dados;
- Alta velocidade de processamento;
- Tolerância a falhas;
- Escalabilidade.



O que é uma arquitetura para o Big Data?

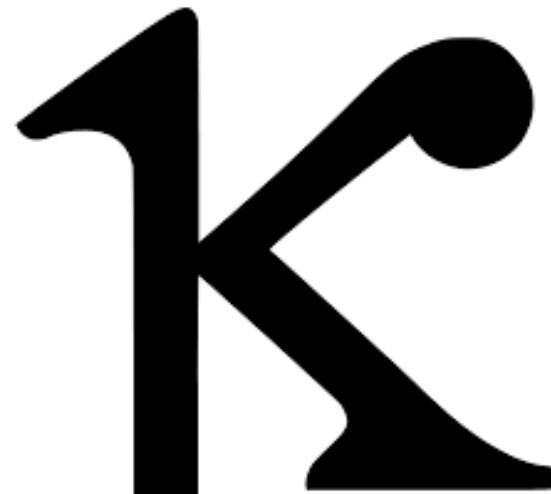


Fonte: <https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/>

Tipos de arquiteturas

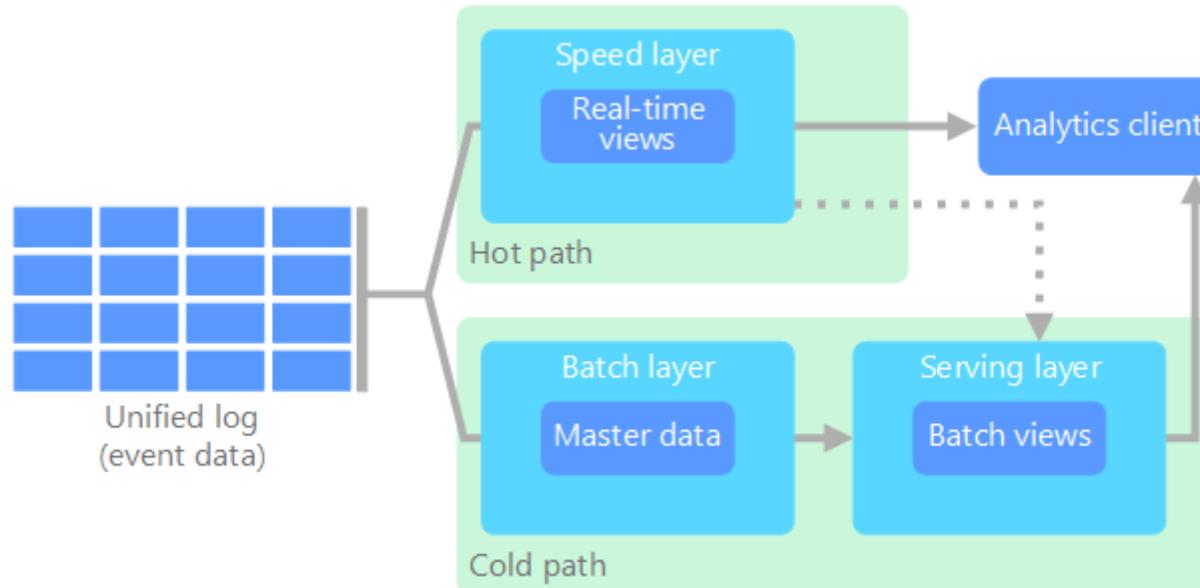
A large, bold, black Greek letter lambda (λ) is centered on the slide.

Lambda

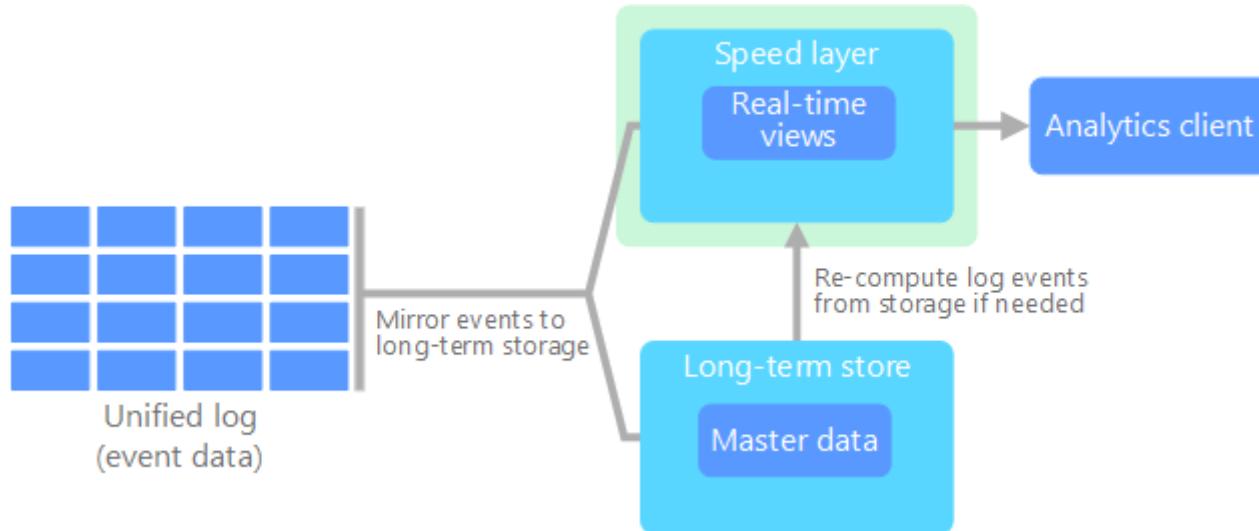
A large, bold, black Greek letter kappa (κ) is centered on the slide.

Kappa

Lambda



Fonte: <https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/>



Fonte: <https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/>

Qual arquitetura escolher?



- Tolerância a falhas;



Aplicação x Arquitetura

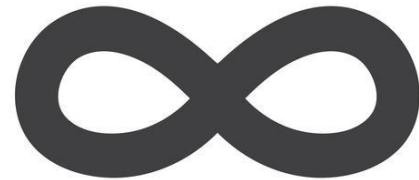
- Tolerância a falhas;
- Processamento de um grande volume de dados;



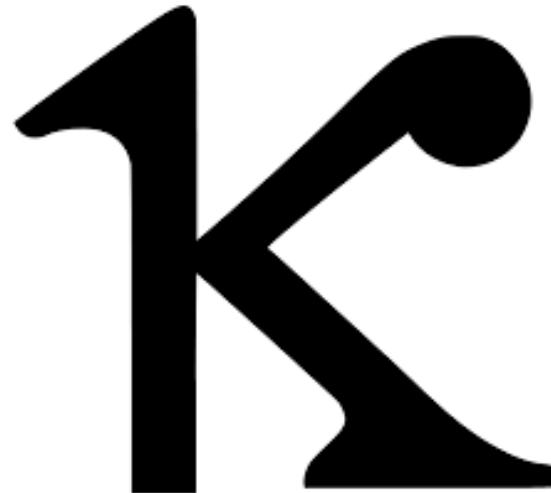
- Tolerância a falhas;
- Processamento de um grande volume de dados;
- Eficiência no processamento através de ML;



- Tolerância a falhas;
- Processamento de um grande volume de dados;
- Eficiência no processamento através de ML;
- Recursos “ilimitados”.



Qual arquitetura escolher?



Kappa

Aplicação x Arquitetura

- Recursos limitados;



Aplicação x Arquitetura

- Recursos limitados;
- Menor complexidade;

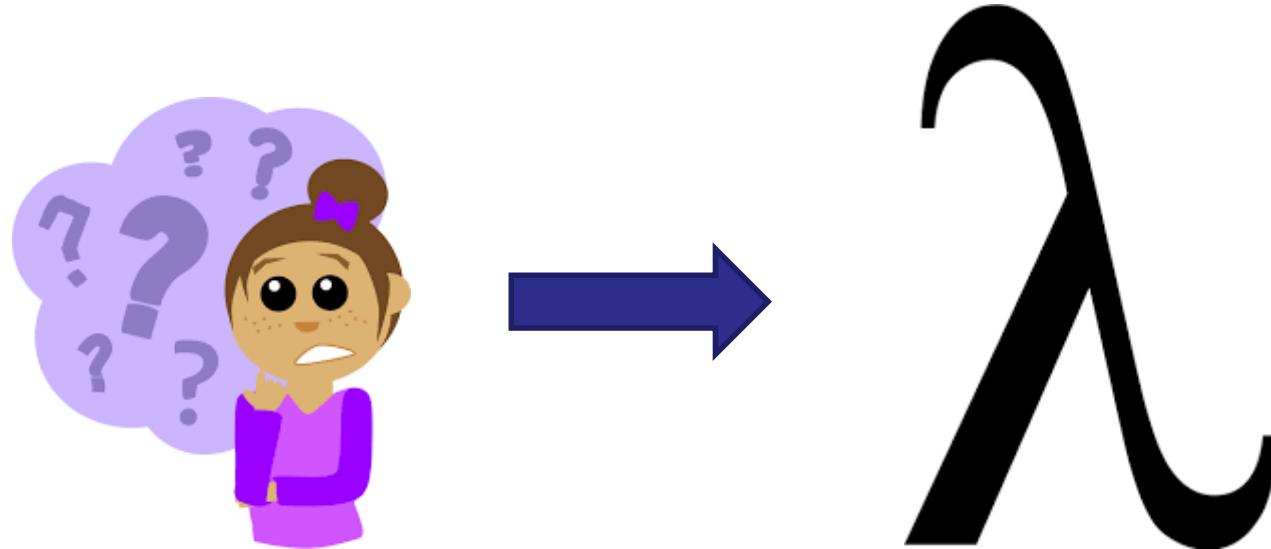


Aplicação x Arquitetura

- Recursos limitados;
- Menor complexidade;
- Velocidade é primordial.



Qual arquitetura escolher?



Lambda

Comparativo

Plataforma e Característica	Kappa	Lambda
Modo de Processamento	Tempo real	Lote e Tempo real
Reprocessamento	Somente quando o código for alterado	A cada ciclo
Acesso aos Dados	Algoritmos incrementais	Algoritmos incrementais e views
Confiabilidade	Tempo real pode variar	Lote é confiável e o tempo pode variar

Conclusão

- Tipos de arquiteturas.
- Diferenças entre arquiteturas.
- Como escolher a arquitetura.

- Exemplos de áreas que utilizam o processamento do Big Data.



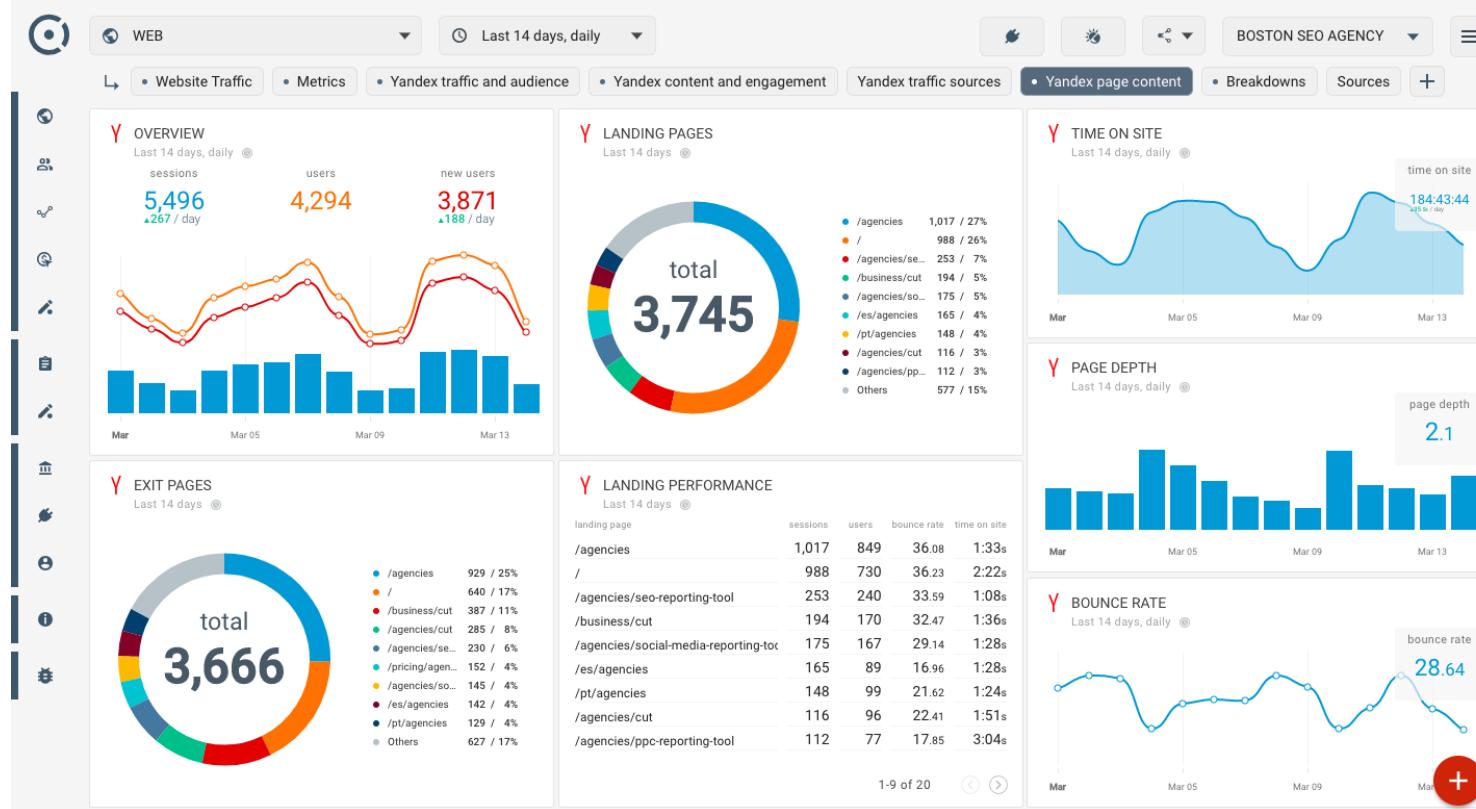
Aula 1.4. Aplicações que empregam o processamento do Big Data

Nesta aula

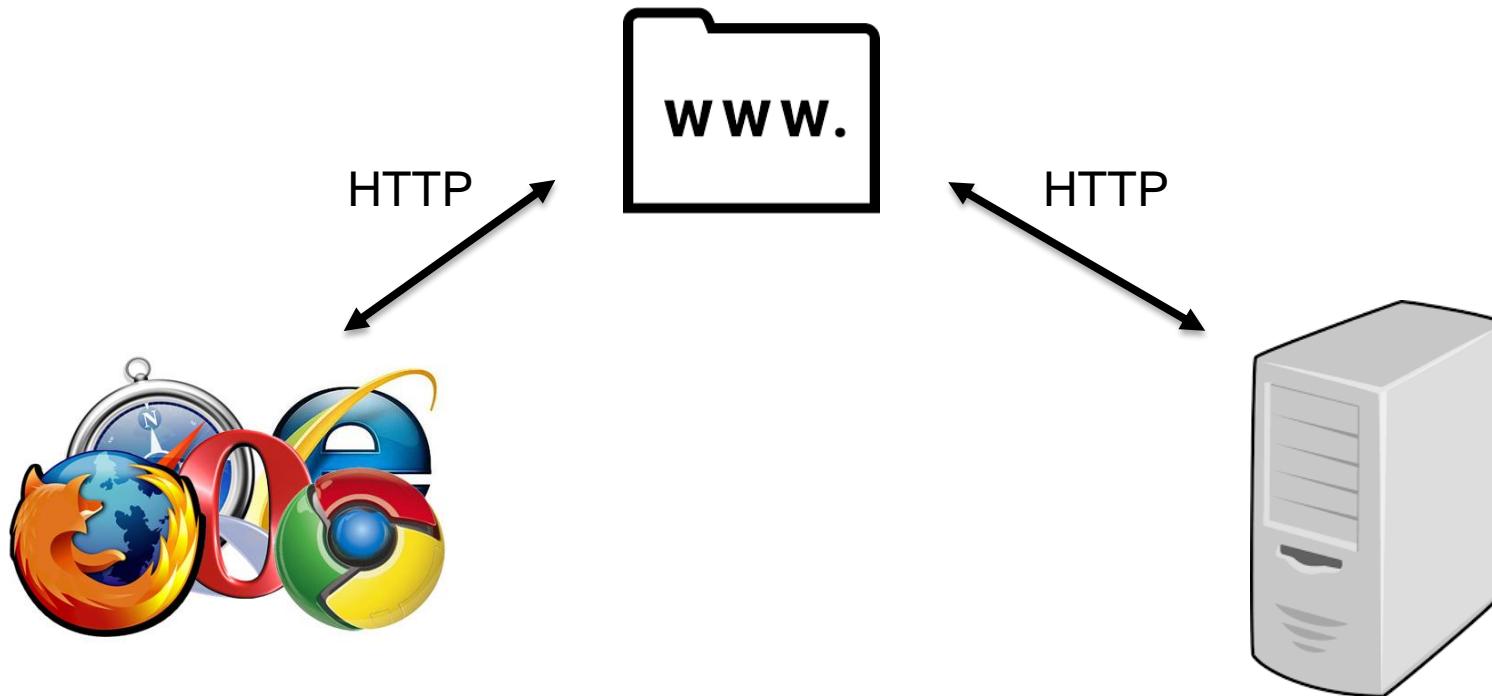
- Análise da web.
- Recomendação de conteúdo.
- Análise de risco.
- Detecção de fraudes.
- Healthcare.
- Internet das Coisas.

Análise da web

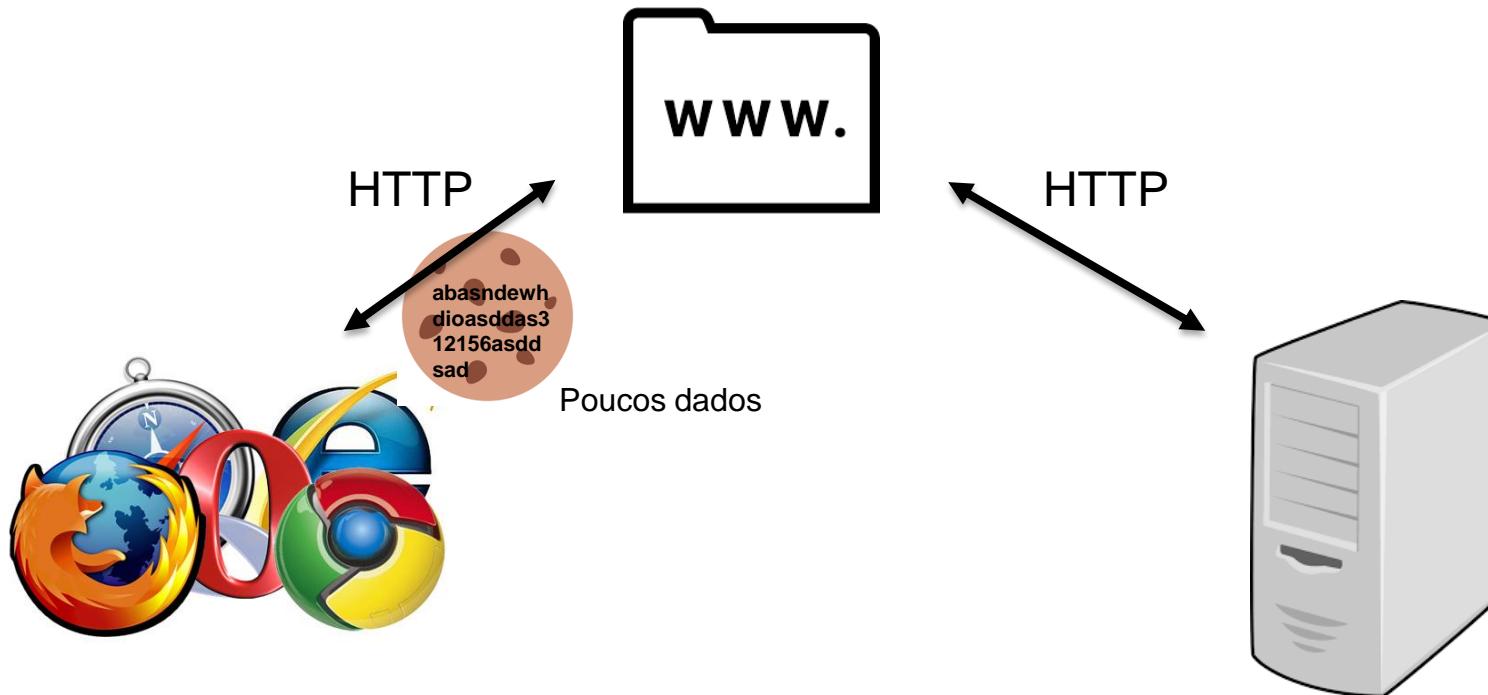
IGTI



Análise da web – Cookies



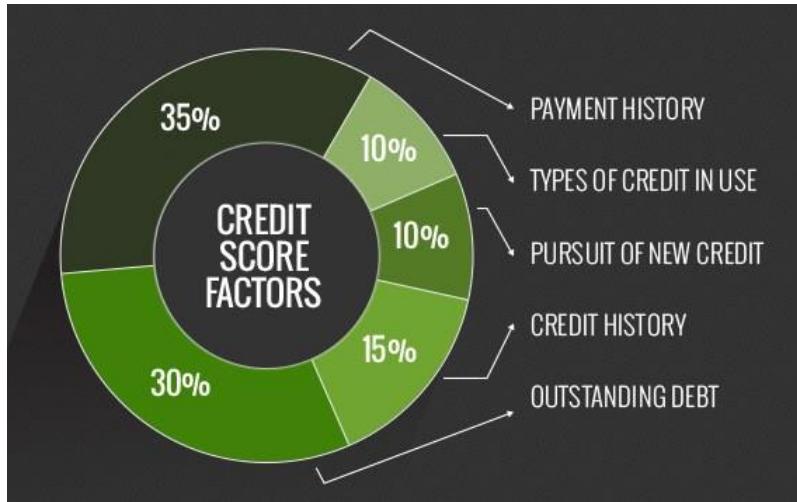
Análise da web – Cookies



Recomendação de conteúdo



Análise de risco



Detecção de golpes

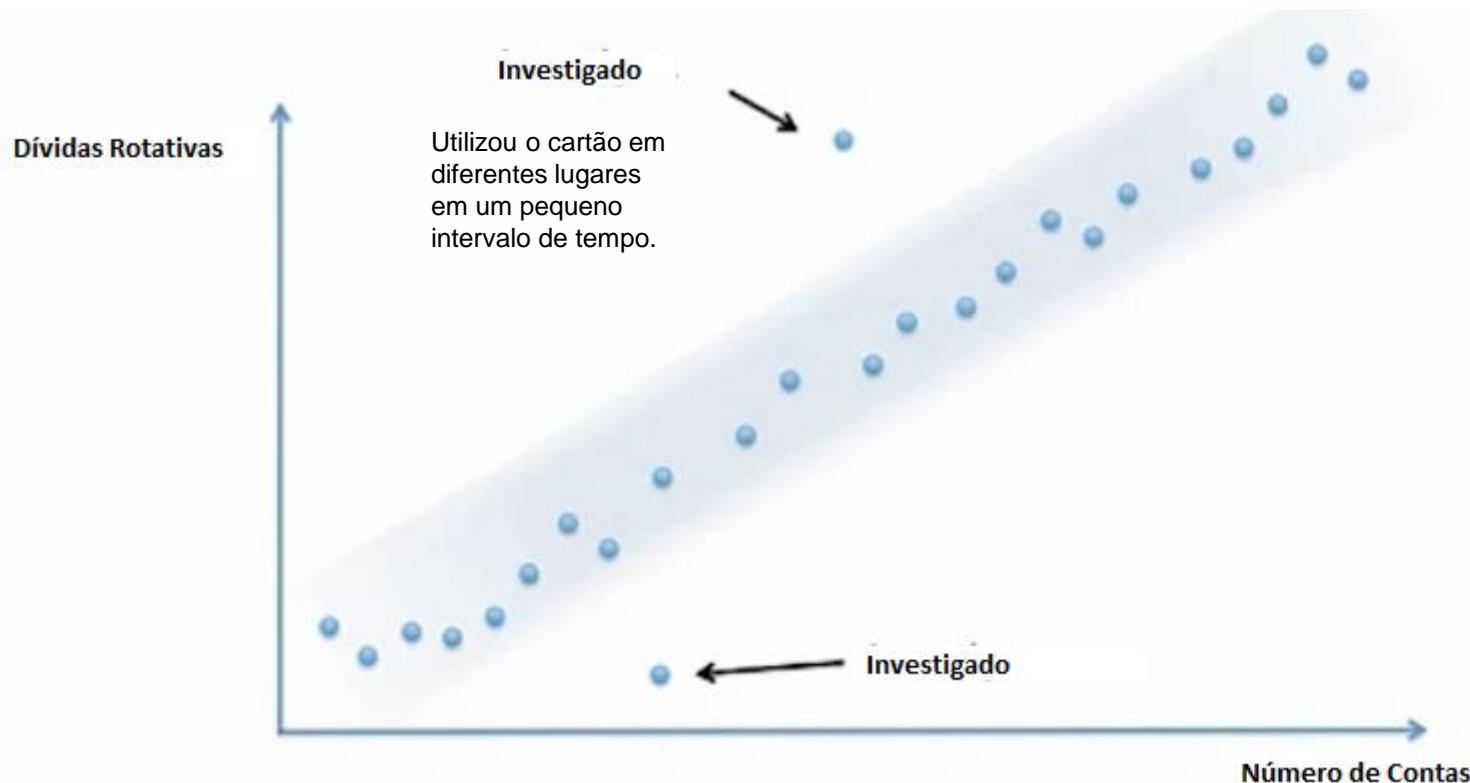
Novo golpe: Conta Fácil do Banco do Brasil
é usada por criminosos para roubar
créditos em dinheiro

13 de janeiro de 2017

- Cria várias identidades “falsas”;
- Rouba pouco de muitos.



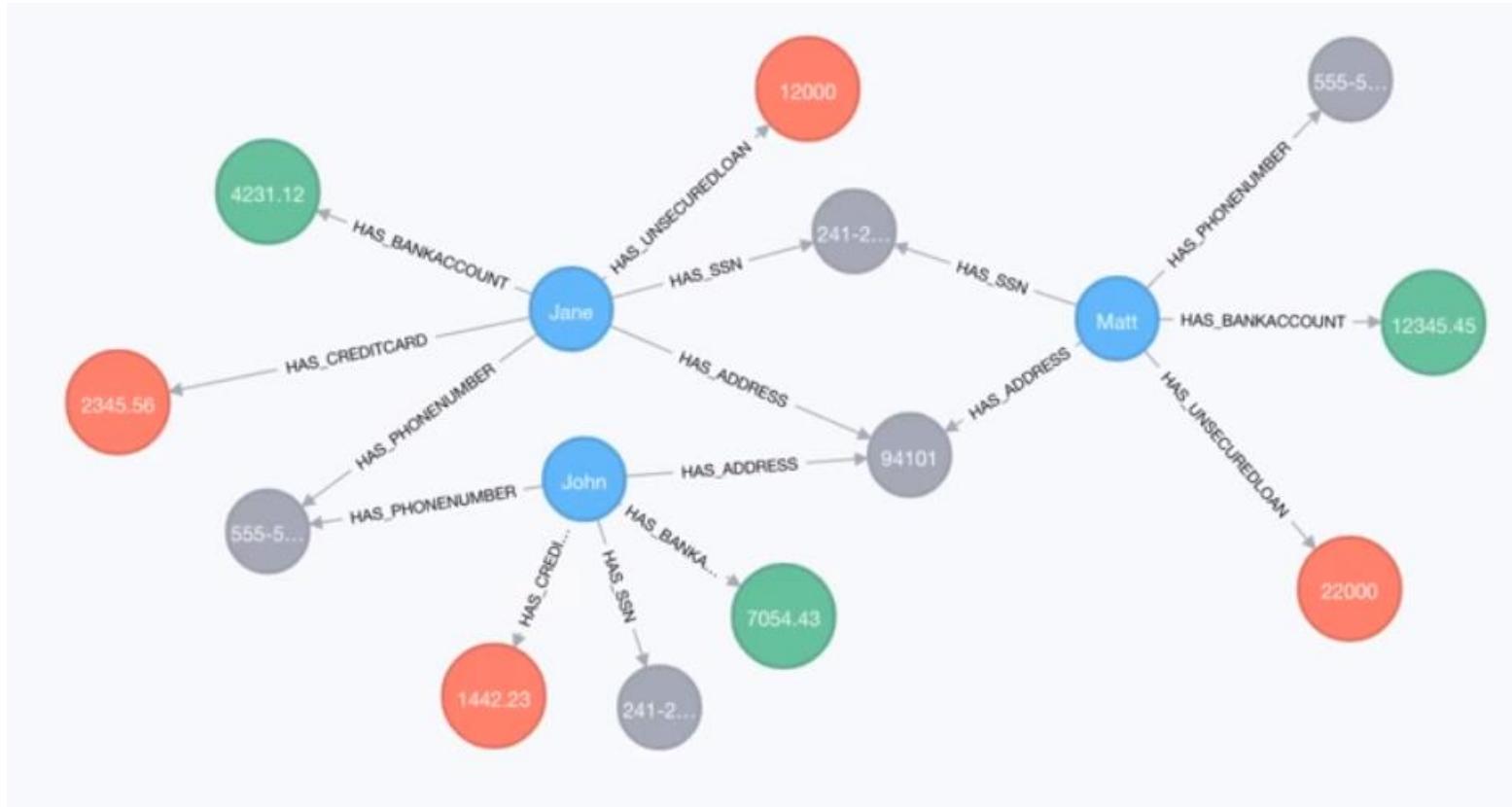
Detecção de golpes



Grafos na detecção de golpes



Grafos na detecção de golpes



Grafos na detecção de golpes

\$ MATCH (accountHolder:AccountHolder)-□->(contactInformation) WITH contactInformation, count(accountHolder) AS RingSize MATCH (contac...



FraudRing

		ContactType	RingSize	FinancialRisk
[MattSmith, JaneAppleseed, JohnDoe]		[Address]	3	34387
[MattSmith, JaneAppleseed]		[SSN]	2	29387
[JaneAppleseed, JohnDoe]		[PhoneNumber]	2	18046

Returned 3 rows in 1101 ms.

\$ MATCH (accountHolder:AccountHolder)-□->(contactInformation) WITH contactInformation, count(accountHolder) AS RingSize MATCH (contac...

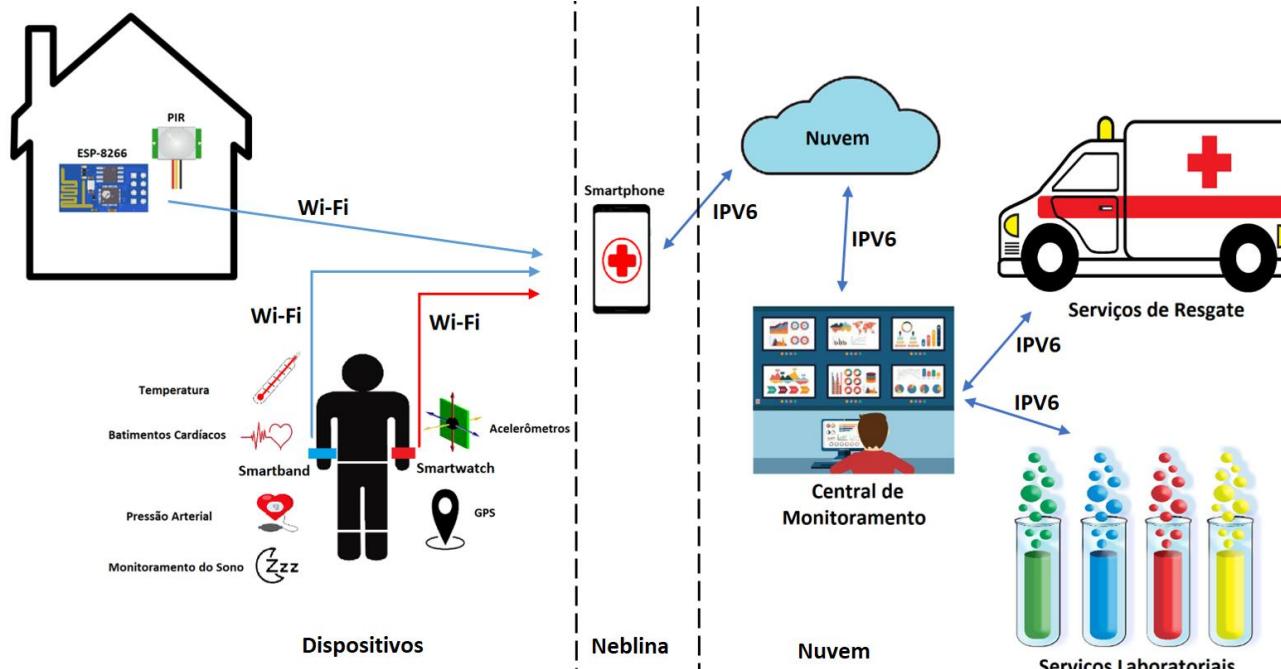


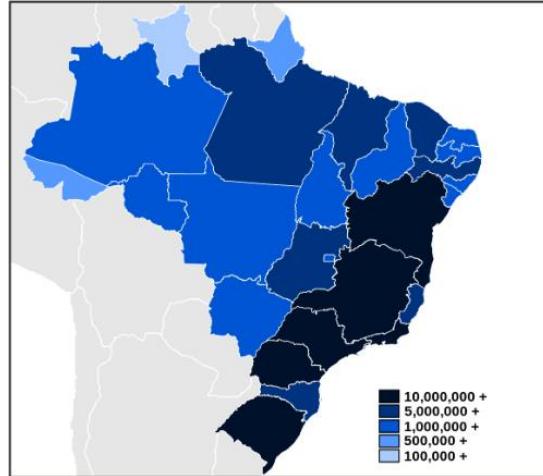
FraudRing

		ContactType	RingSize
[MattSmith, JaneAppleseed, JohnDoe]		[Address]	3
[MattSmith, JaneAppleseed]		[SSN]	2
[JaneAppleseed, JohnDoe]		[PhoneNumber]	2

Returned 3 rows in 1695 ms.

Healthcare



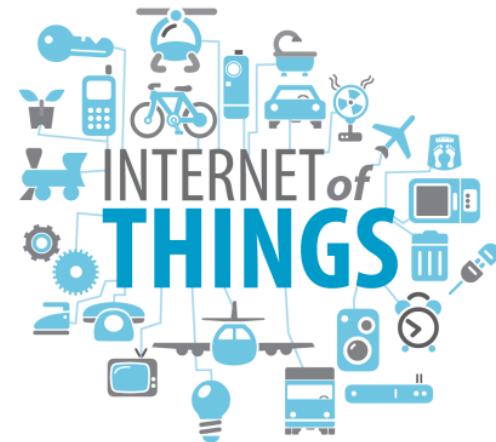


**HORA DE COMBATER
O MOSQUITO DA DENGUE!**



O que é a IoT?

- Internet das Coisas (IoT) é um conceito de computação que descreve um futuro em que objetos físicos do nosso cotidiano serão conectados à internet, o que permitirá a identificação, comunicação, integração e interação entre estes dispositivos.



Internet das Coisas



Conclusão

- Análise da web.
- Recomendação de conteúdo.
- Análise de risco.
- Detecção de fraudes.
- Healthcare.
- Internet das Coisas.

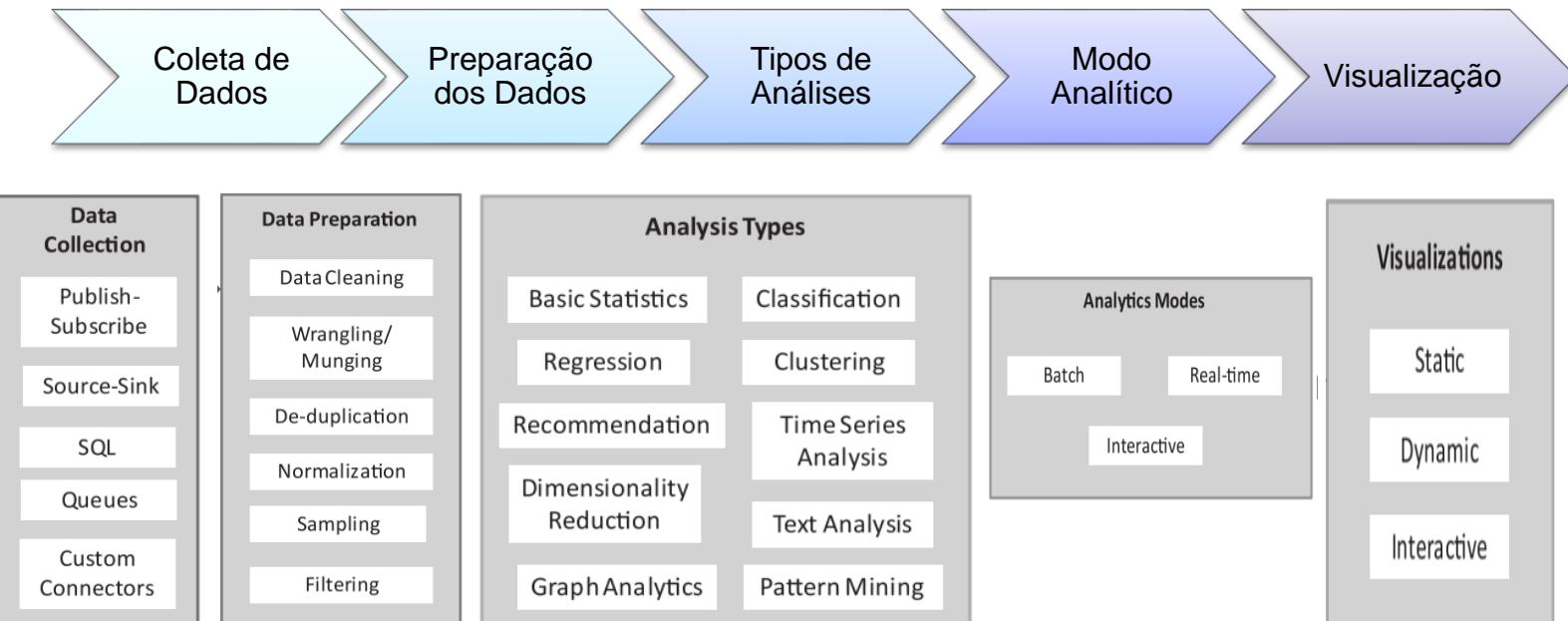
- Estudo de caso I – Aplicação do processamento do Big Data na análise do genoma.



Aula 1.5. Estudo de caso I – Aplicação do processamento do Big Data na análise do genoma

- ❑ Processamento do Big Data aplicado na análise do genoma.

Relembrando





Intel Science and Technology Center for Big Data



Intel Science and Technology Center for Big Data

- Gerador de Dataset para benchmark em análise do genoma;
- Gera 4 tipos diferentes de dado:
 - Array com expressões genéticas de vários pacientes;
 - Metadados de pacientes (idade, gênero, CEP) e resposta a drogas;
 - Metadados de genes (gene alvo, número de cromossomos, posição, comprimento da sequência e função);
 - Gene Ontology (mapeamento de cada gene).

Exemplo de cada um dos tipos

Patient meta-data

patientid	age	gender	zipcode	disease	drug response
0	41	0	7494	15	84.77
1	45	1	38617	6	62.4
2	51	1	62817	17	49.43
3	62	0	53316	18	25.88
4	23	1	49685	7	41.03
5	60	0	48726	8	23.35
6	77	0	99103	18	87.86
7	83	1	5210	18	55.05
8	56	1	7359	5	97.09
9	63	1	59483	17	15.05

Gene meta-data

geneid	target	position	length	function
0	-1	6.69E+08	175	633
1	-1	2.74E+09	974	7
2	-1	6.82E+08	260	909
3	-1	2.4E+09	930	28
4	1	2.01E+09	836	462
5	-1	1.64E+09	277	941
6	-1	2.6E+09	428	487
7	-1	1.02E+08	618	966
8	6	8.46E+08	635	328
9	3	2.77E+09	964	183

Micro-array data

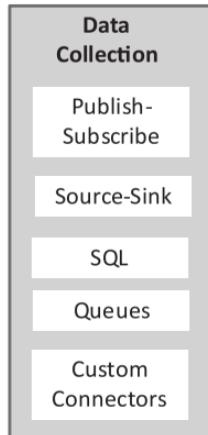
geneid	patientid	expression value
0	0	7.51
0	1	5.92
0	2	8.12
0	3	3.47
:	:	:
1	1	7.43
1	2	5.54
1	3	2.86
2	0	7.69
2	1	7.66
2	2	9.76
2	3	1.41

Gene Ontology data

geneid	goid	whether gene belongs to go
0	0	1
0	1	1
0	2	1
0	3	1
0	6	0
:	:	:
9	59993	1
9	59994	1
9	59995	1
9	59996	1

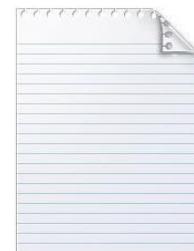
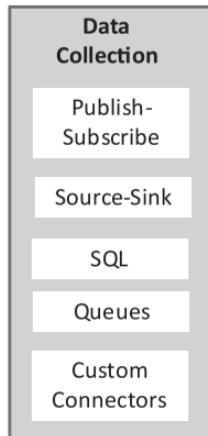
Coleta de dados

Coleta de
Dados

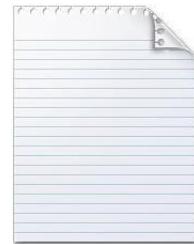
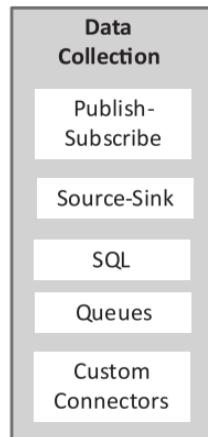


Coleta de dados

Coleta de
Dados



Coleta de dados

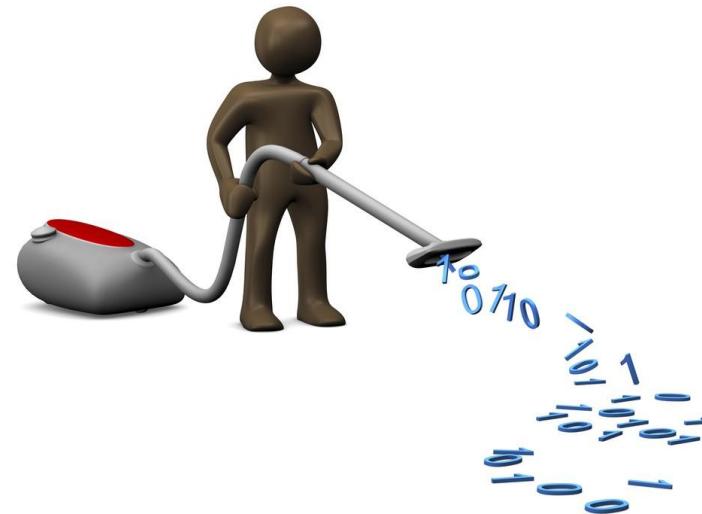
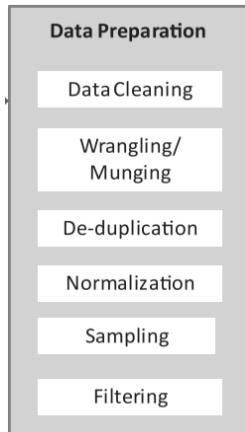


Conector SQL



Preparação dos dados

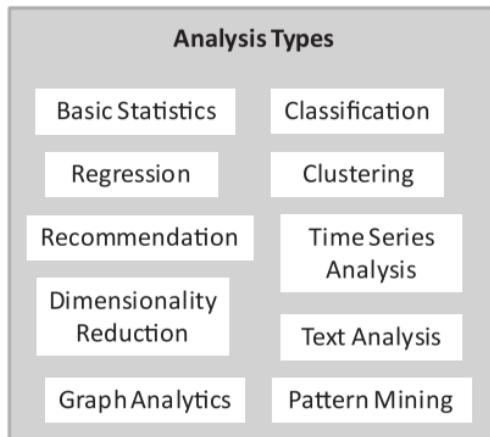
Preparação
dos Dados



Tipo de análise



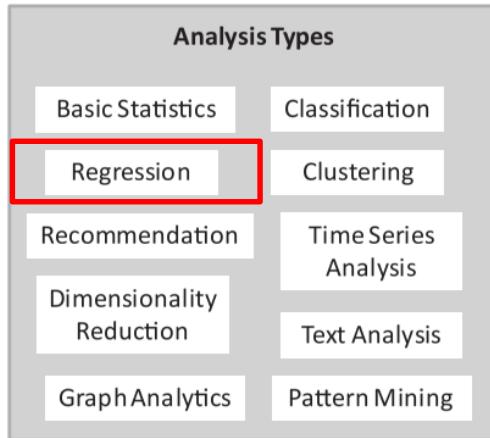
Tipos de
Análises



1. Prever a resposta a uma droga através da análise da expressão gênica.
2. Encontrar a correlação entre diferentes expressões genéticas.

Tipo de análise

Tipos de
Análises

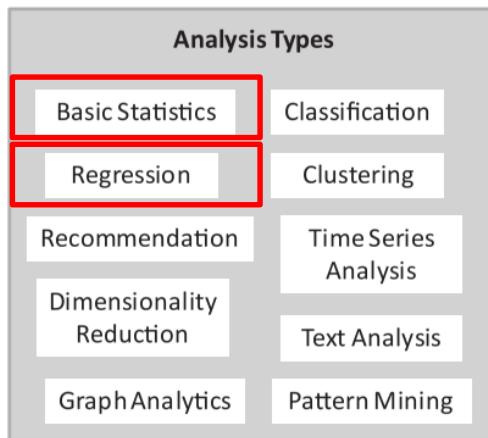


1. Prever a resposta a uma droga através da análise da expressão gênica.
2. Encontrar a correlação entre diferentes expressões genéticas.

Dependente: Resposta do paciente
Independente: Expressão genética

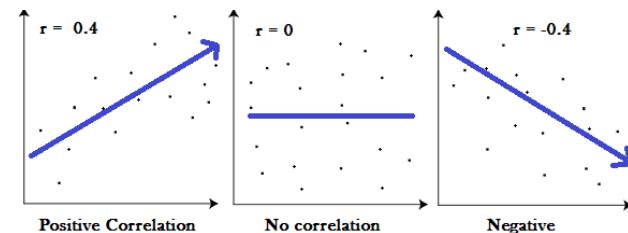
Tipo de análise

Tipos de
Análises

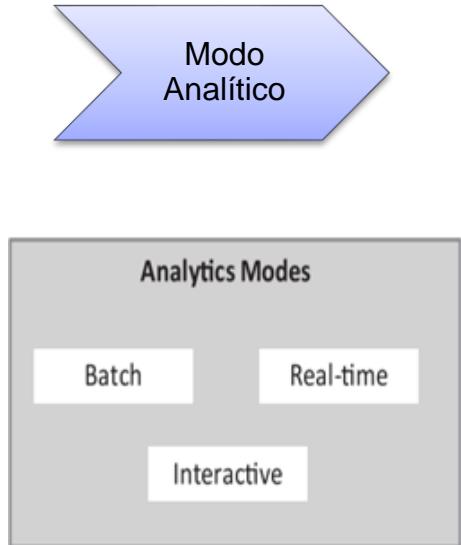


1. Prever a resposta a uma droga através da análise da expressão gênica.
2. Encontrar a correlação entre diferentes expressões genéticas.

Correlação



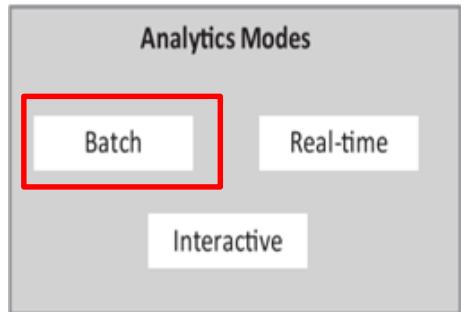
Modo analítico



Modo analítico



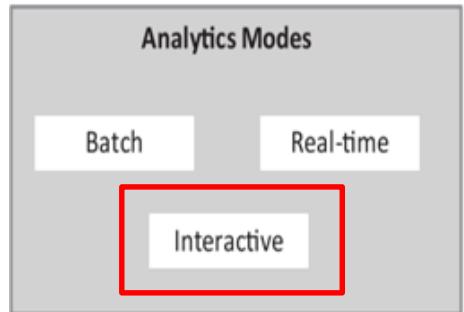
Dados já coletados → Batelada (Lote)



Modo analítico



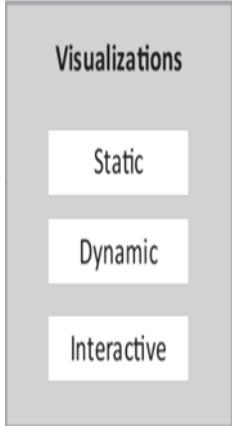
Dados já coletados → Batelada (Lote)



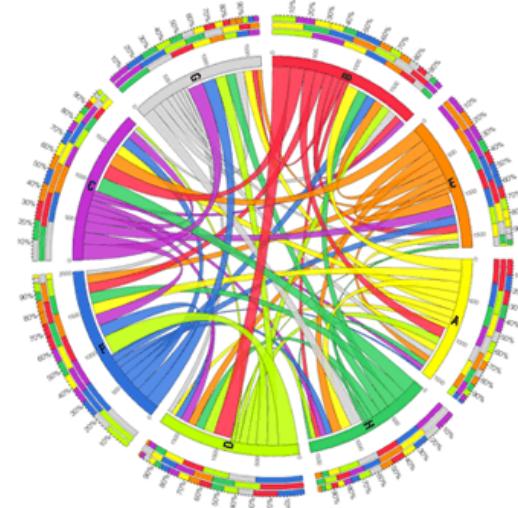
Interação com o usuário → Interativo

Visualização

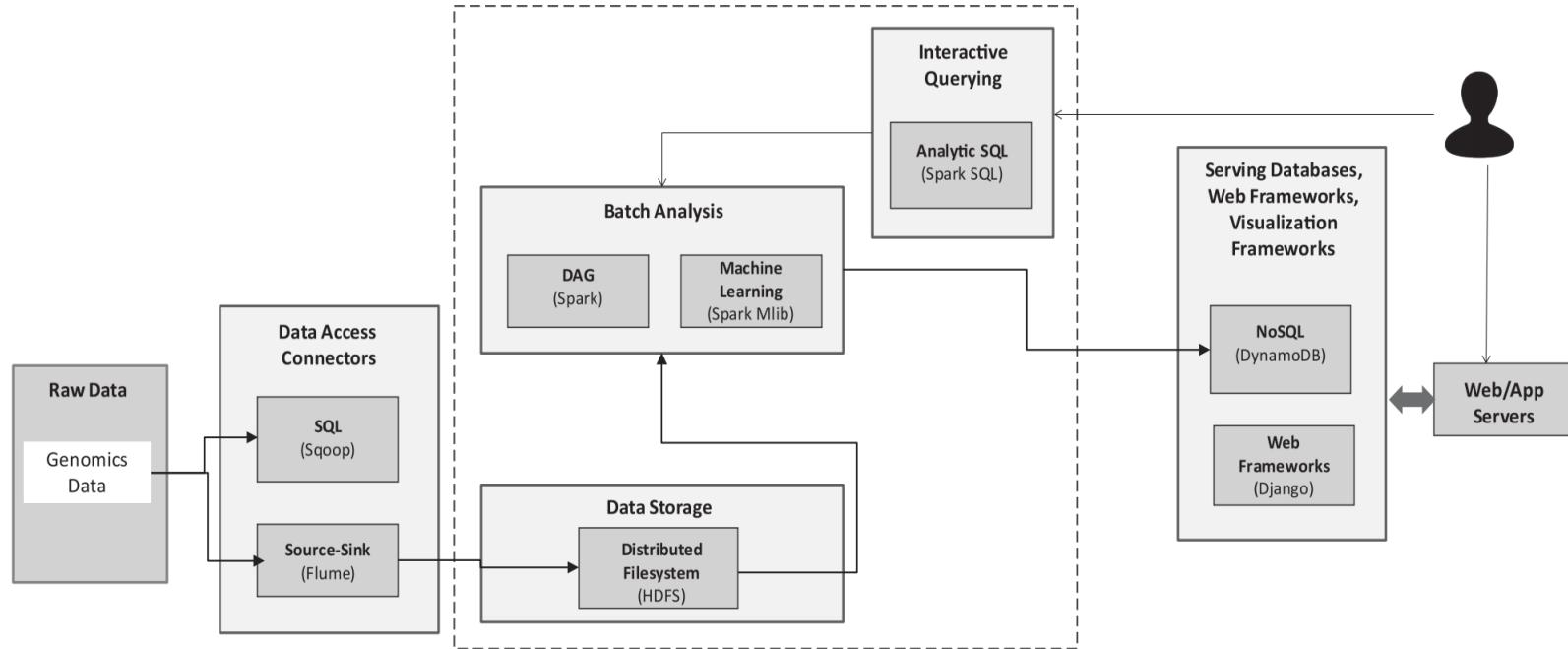
Visualização



	A	B	C	D	E	F	G	H
A	54	133	157	94	88	141	167	133
B	49	113	111	113	202	53	7	92
C	66	130	69	162	123	62	106	117
D	60	138	49	85	98	98	122	87
E	53	88	15	91	91	20	69	127
F	118	32	62	139	135	95	60	64
G	114	108	73	44	103	139	37	145
H	74	110	84	120	9	41	45	131



Ferramentas



Fonte: BAHGA e MADISSETTI (2016)

Passos para a regressão

Select genes with a particular set of functions and join gene meta-data with patient meta-data and microarray data

patientid	disease	geneid	exValue	drugResponse
0	14	0	701993.3	60.42
1	12	0	-16377.7	57.86
2	6	0	1296795	39.62
3	19	0	505206.7	24.83
4	10	0	732953.1	50.36
5	10	0	446293.6	12.63
6	8	0	-92641.8	12.24
7	14	0	-29881.8	73.71
8	7	0	115540.8	37.66
9	7	0	509479.7	62.43
0	14	9	1423755	60.42
1	12	9	1589411	57.86
2	6	9	7045.72	39.62
3	19	9	-65459.3	24.83
4	10	9	1373435	50.36
5	10	9	233481.5	12.63
6	8	9	832897.5	12.24
7	14	9	1258119	73.71
8	7	9	679142	37.66
9	7	9	1143324	62.43

Pivot

Pivot the table in previous step to get expression values for each type of gene for each patient

patientid	0	9
0	701993.3	1423755
1	-16377.7	1589411
2	1296795	7045.72
3	505206.7	-65459.3
4	732953.1	1373435
5	446293.6	233481.5
6	-92641.8	832897.5
7	-29881.8	1258119
8	115540.8	679142
9	509479.7	1143324

Select patient ID, disease and drugResponse from patient meta-data

patientid	disease	drugResponse
0	14	60.42
1	12	57.86
2	6	39.62
3	19	24.83
4	10	50.36
5	10	12.63
6	8	12.24
7	14	73.71
8	7	37.66
9	7	62.43

Join

patientid	disease	drugResponse	patientid	0	9
0	14	60.42	0	701993.3	1423755
1	12	57.86	1	-16377.7	1589411
2	6	39.62	2	1296795	7045.72
3	19	24.83	3	505206.7	-65459.3
4	10	50.36	4	732953.1	1373435
5	10	12.63	5	446293.6	233481.5
6	8	12.24	6	-92641.8	832897.5
7	14	73.71	7	-29881.8	1258119
8	7	37.66	8	115540.8	679142
9	7	62.43	9	509479.7	1143324

Use this table to train Linear Regression model to predict drug response (target variable is the patient drug response and the independent variables are gene expression values)

Fonte: BAHGA e MADISSETTI (2016)

Passos para a correlação

Select patients with some disease and join results with the micro-array table

patientid	disease	geneid	exValue
3	18	0	3.47
3	18	1	2.86
3	18	2	1.41
3	18	3	5.3
3	18	4	7.73
3	18	5	1.51
3	18	6	4.92
3	18	7	1.97
3	18	8	8.02
3	18	9	8.32
6	18	0	1.42
6	18	1	9.17
6	18	2	1.37
6	18	3	9.2
6	18	4	1.4
6	18	5	8.86
6	18	6	7.88
6	18	7	9.52
6	18	8	3.24
6	18	9	3.54

Pivot

Pivot the table to get the expression values for all genes for each patients

patientid	0	1	2	3	4	5	6	7	8	9
3	3.47	2.86	1.41	5.3	7.73	1.51	4.92	1.97	8.02	8.32
6	1.42	9.17	1.37	9.2	1.4	8.86	7.88	9.52	3.24	3.54
7	1.48	2.58	7.52	1.66	4.55	8.55	5.45	4.35	2.81	3.7



Compute the correlation between the expression levels of all pairs of genes

Correlation Matrix

	Gene-0	Gene-1	Gene-2	Gene-3	Gene-4	Gene-5	Gene-6	Gene-7	Gene-8	Gene-9
Gene-0	1	-0.48969	-0.47259	-0.04561	0.8799	-0.99993	-0.65789	-0.75961	0.994991	0.999993
Gene-1	-0.48969	1	-0.53696	0.893321	-0.84517	0.499756	0.978801	0.939038	-0.40008	-0.49297
Gene-2	-0.47259	-0.53696	1	-0.85881	0.002915	0.462357	-0.35279	-0.21418	-0.55832	-0.46928
Gene-3	-0.04561	0.893321	-0.85881	1	-0.5148	0.057171	0.782337	0.684345	0.054482	-0.04936
Gene-4	0.8799	-0.84517	0.002915	-0.5148	1	-0.88534	-0.93673	-0.97741	0.827993	0.881678
Gene-5	-0.99993	0.499756	0.462357	0.057171	-0.88534	1	0.666563	0.76709	-0.99377	-0.99997
Gene-6	-0.65789	0.978801	-0.35279	0.782337	-0.93673	0.666563	1	0.989549	-0.57931	-0.66071
Gene-7	-0.75961	0.939038	-0.21418	0.684345	-0.97741	0.76709	0.989549	1	-0.69079	-0.76205
Gene-8	0.994991	-0.40008	-0.55832	0.054482	0.827993	-0.99377	-0.57931	-0.69079	1	0.994608
Gene-9	0.999993	-0.49297	-0.46928	-0.04936	0.881678	-0.99997	-0.66071	-0.76205	0.994608	1

Fonte: BAHGA e MADISSETTI (2016)

Pilha do Big Data para a análise do genoma.

■ Próxima aula

- Estudo de caso II – Airbnb.



Aula 1.6. Estudo de caso II – Airbnb

- ❑ Como a Airbnb utilizou o Big Data e análise de dados.

Airbnb (Air, Bed and Breakfast)

IGTI



2008

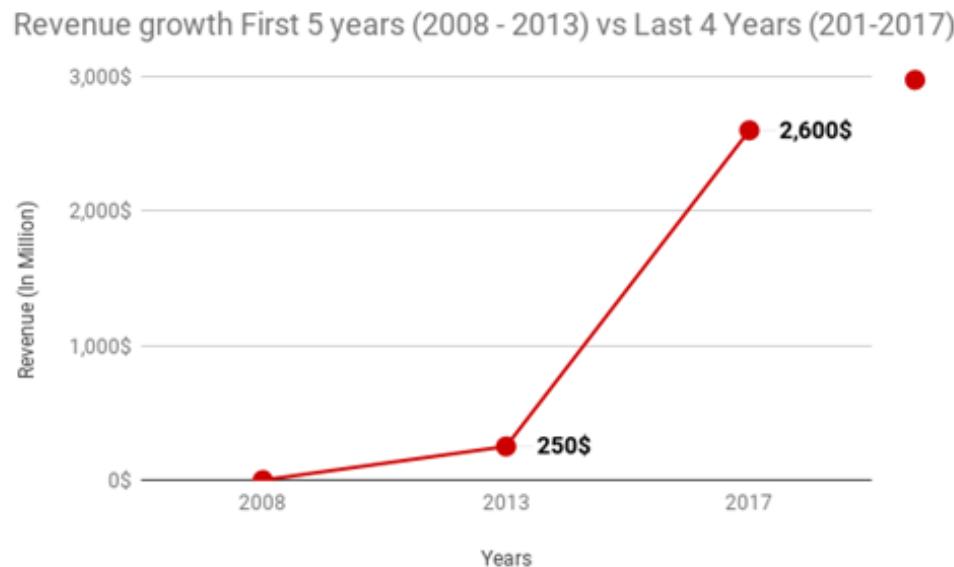


São Francisco, Califórnia

Nathan Blecharczyk, Brian Chesky e Joe Gebbia

Crescimento da Airbnb

2012:
Valor de Mercado
U\$ 1Bi



Como a Airbnb Conseguiu?

IGTI

- Em 2011, contratou o primeiro cientista de dados;
- De 2011 a 2016 cresceu 43000%;
- Como conectar os cientistas de dados a outras áreas?



Como a Airbnb Conseguiu?

- Como nós (Airbnb) caracterizamos os cientistas de dados?
- Como isso (análise de dados) está envolvido no processo de tomada de decisão?
- Como podemos escalar o nosso negócio para levarmos o Airbnb para todos os lugares do mundo?
- 4 Respostas.



Respostas

- Os dados não são números, são pessoas.



Respostas

- Parceria proativa vs coleta estatística reativa.



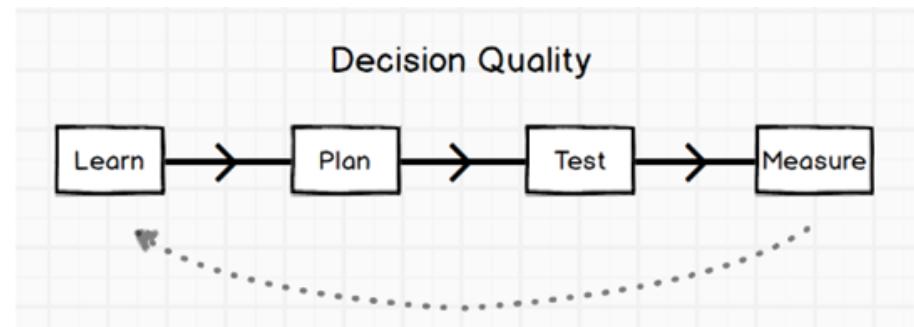
Respostas

- Decisões dirigidas pelos clientes.

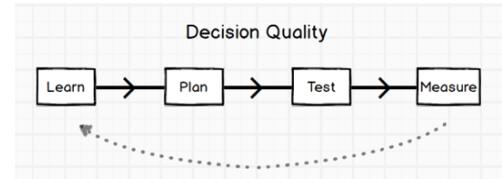


Respostas

- Sistema utilizado para tomada de decisão.



Respostas



- **Learn:** aprender sobre o contexto do problema, agrupando os resultados e pesquisas passadas com as oportunidades.
- **Plan:** transformar o aprendizado em um plano. Para isso, é utilizada uma análise preditiva avaliando os possíveis caminhos e os resultados (previstos) para cada uma dessas opções.
- **Test:** à medida que o planejamento evolui é criado um experimento controlado para avaliar a qualidade do plano escolhido.
- **Measure:** por último, os resultados dos experimentos são medidos identificando os impactos causados pelas estratégias escolhidas. Caso os resultados sejam satisfatórios, o plano é colocado em prática, caso contrário o ciclo recomeça.

Respostas

- Democratização da ciência de dados.

2011 – 1 escritório → 3 cientistas de dados

2012 – 10 escritórios



Conclusão

- ✓ Como a Airbnb utilizou a análise de dados para crescer.

■ Próxima aula

- Streaming de dados.



Técnicas para o Processamento do Big Data

Capítulo 2. Streaming de Dados

Prof. Túlio Philipe Vieira



Aula 2.1. Introdução

Nesta aula

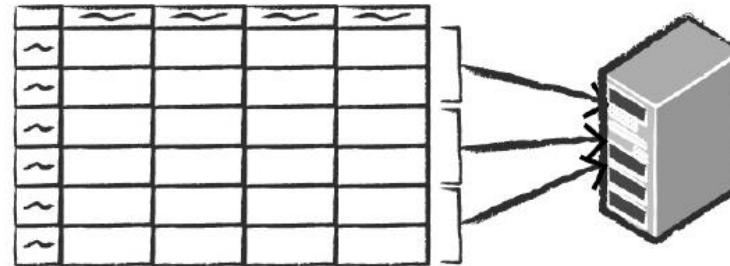
- O processamento estático/lote/batelada.
- O processamento em streaming/dinâmico.
- Aplicações do processamento em streaming.
- Diferenças entre o processamento estático e o streaming.

Modelo estático de processamento

Spreadsheet on
a single machine



Table or Data Frame
partitioned across servers
in a data center

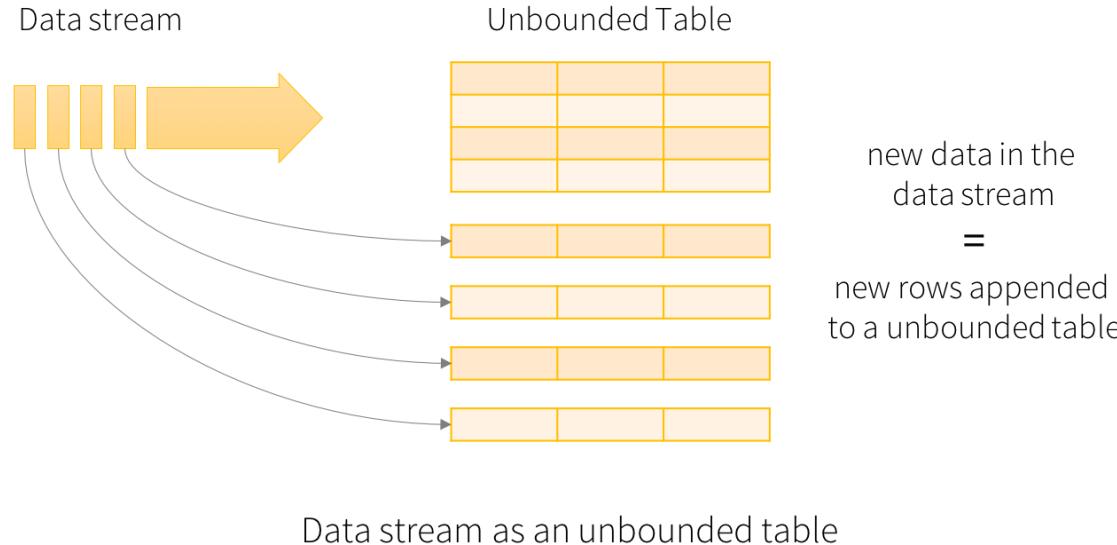


Fonte: Chambers, B. and Zaharia, M., 2018. *Spark: the definitive guide: big data processing made simple.* " O'Reilly Media, Inc. ".

Isso na prática...



Modelo construtivo do Structured Streaming

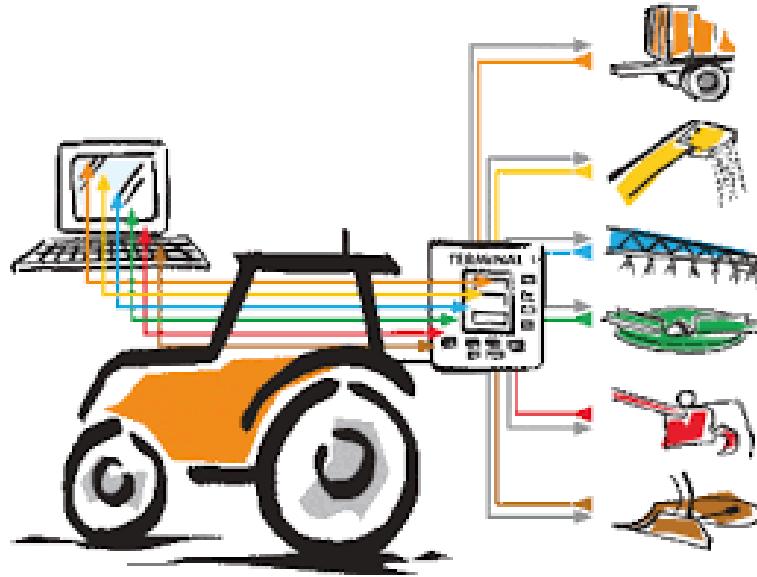


Fonte: SPARK, Apache. **Structured Streaming Programming Guide.**

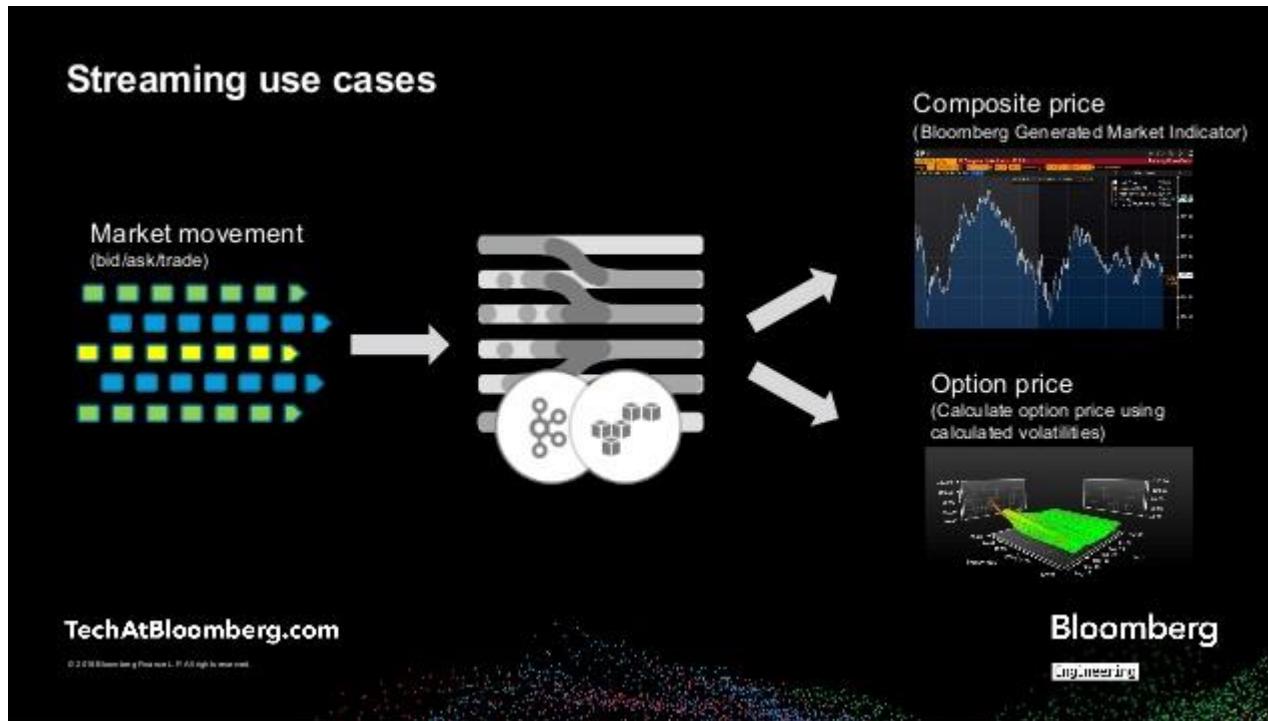
Disponível em: <<https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html>>.
Acesso em: 04 fev. 2020.



Aplicações do streaming



Aplicações do streaming



Aplicações do streaming



MMOs earn \$19.8B in 2016E, 60% of all digital PC game revenue

Asia wins out thanks to audience numbers, but North America and Europe have higher conversion and average spending.



Emerging markets drive free-to-play MMO revenue up by a CAGR¹ of 7.8% from 2016E to 2018E. Users in Asia and Eastern Europe flock to free-to-play MMOs because full-priced games are unaffordable for many. Players with low spending power are also less likely to pirate games that are free.

Pay-to-play MMO revenue levels out, earning \$2.7B in 2018E. Dedicated users continue to play older subscription-based titles, but fees turn off new players who have a variety of free-to-play options. In the West, players pay \$10-\$15 per month for MMO access, while in Asia, subscription MMOs tend to charge by the hour.

Worldwide PC MMO revenue Free-to-play vs. Pay-to-play

Year	Free-to-play (B)	Pay-to-play (B)
2015	\$16.9B	\$3.1B
2016E	\$17.1B	\$2.7B
2017E	\$17.4B	\$2.7B
2018E	\$18.1B	\$2.7B

1. Compound Annual Growth Rate. See slide 25 for definition.

Batch x Streaming

	Processamento em Lotes	Processamento em Streams
Escopo dos dados	Consultas ou processamento de todos ou a maioria do conjunto de dados	Consultas ou processamento de dados a cada nova atualização
Tamanho dos dados	Grandes lotes de dados	Registros individuais ou micro lotes
Performance	Latências de minutos a horas	Latências da ordem de segundos ou milissegundos
Análise	Dados analíticos complexos	Métricas mais simples, agregações e rotações

Conclusão

- O processamento estático/lote/batelada.
- O processamento em streaming/dinâmico.
- Aplicações do processamento em streaming.
- Diferenças entre o processamento estático e o streaming.

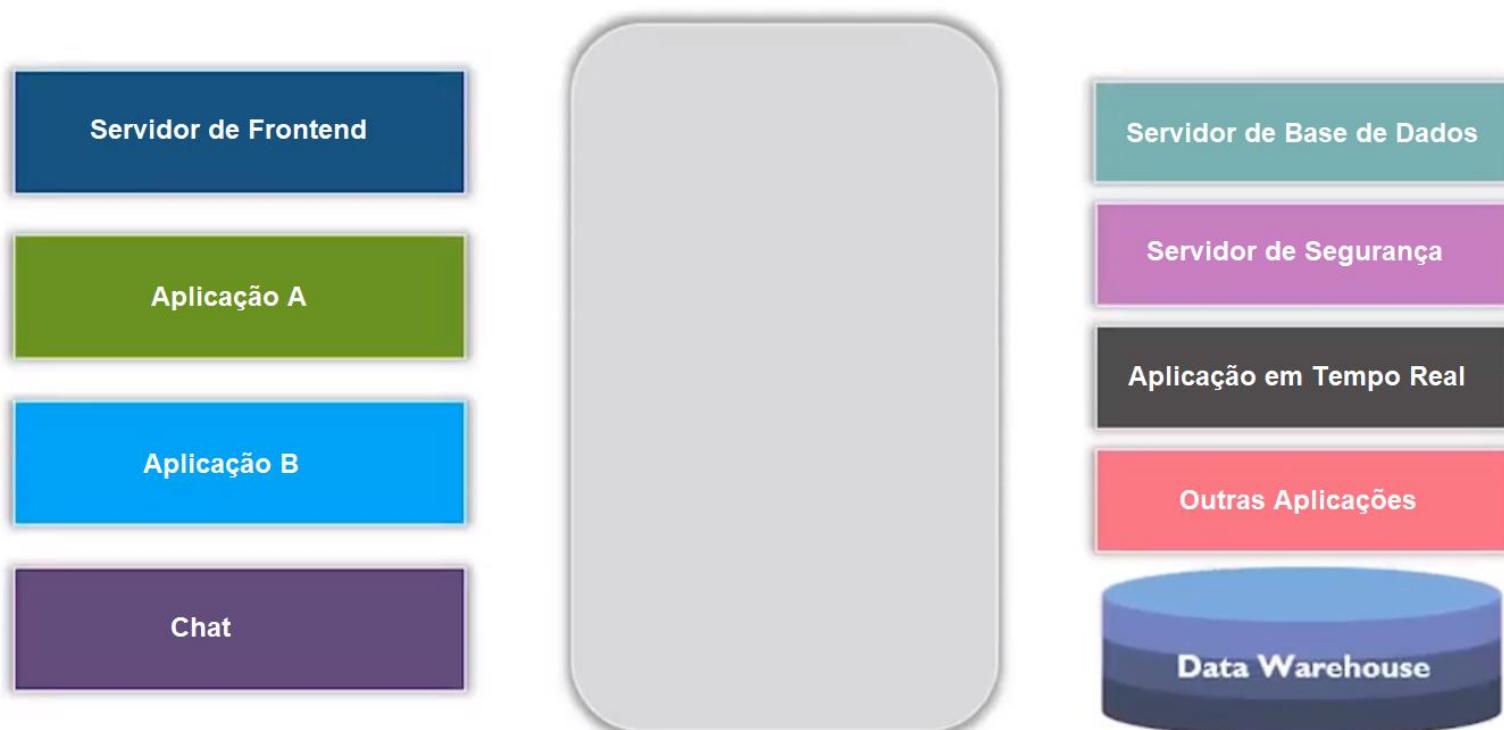
■ Próxima aula

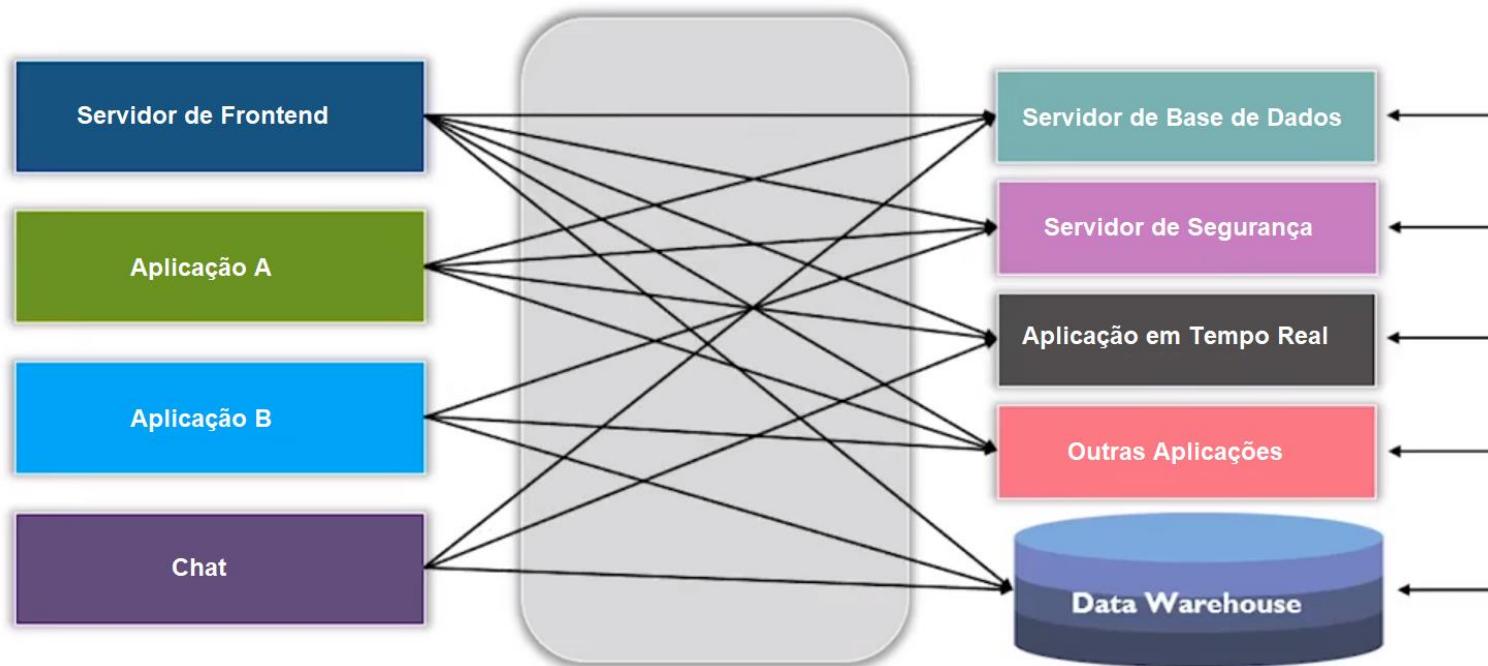
- ❑ Kafka, Flink e Amazon Kinesis.

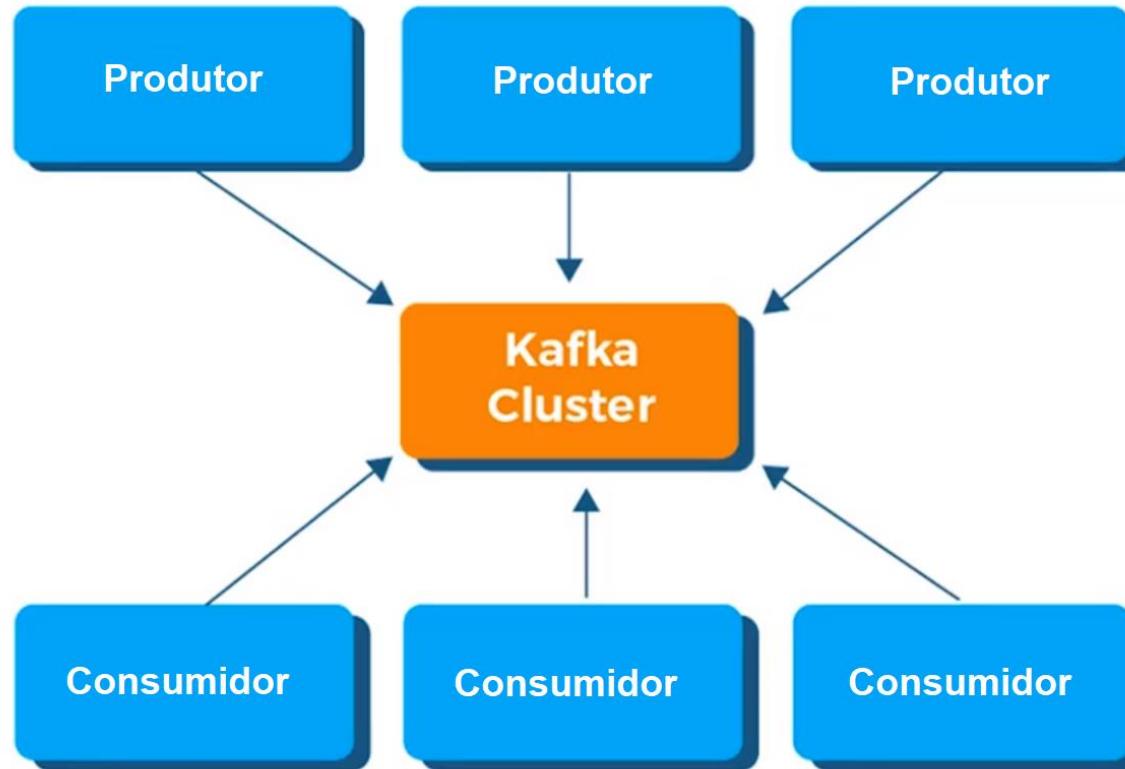


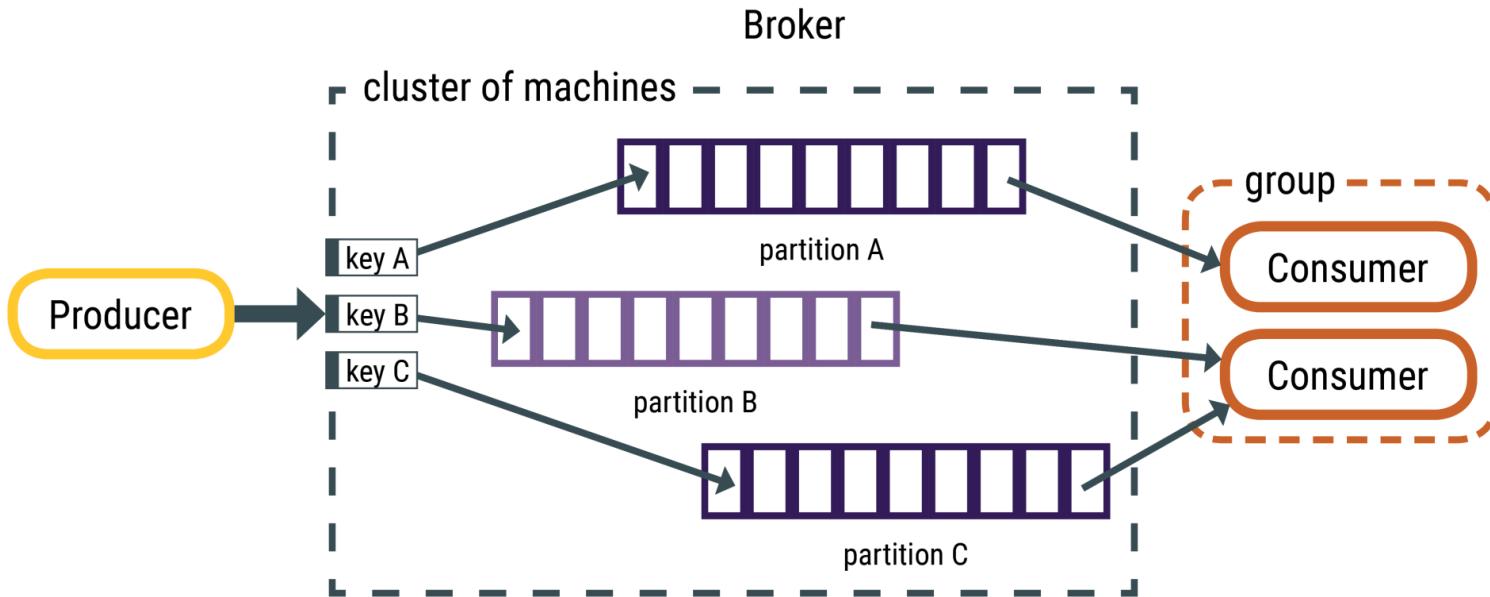
Aula 2.2. Kafka, Flink e Amazon Kinesis

- ❑ Kafka, Flink e Amazon Kinesis.

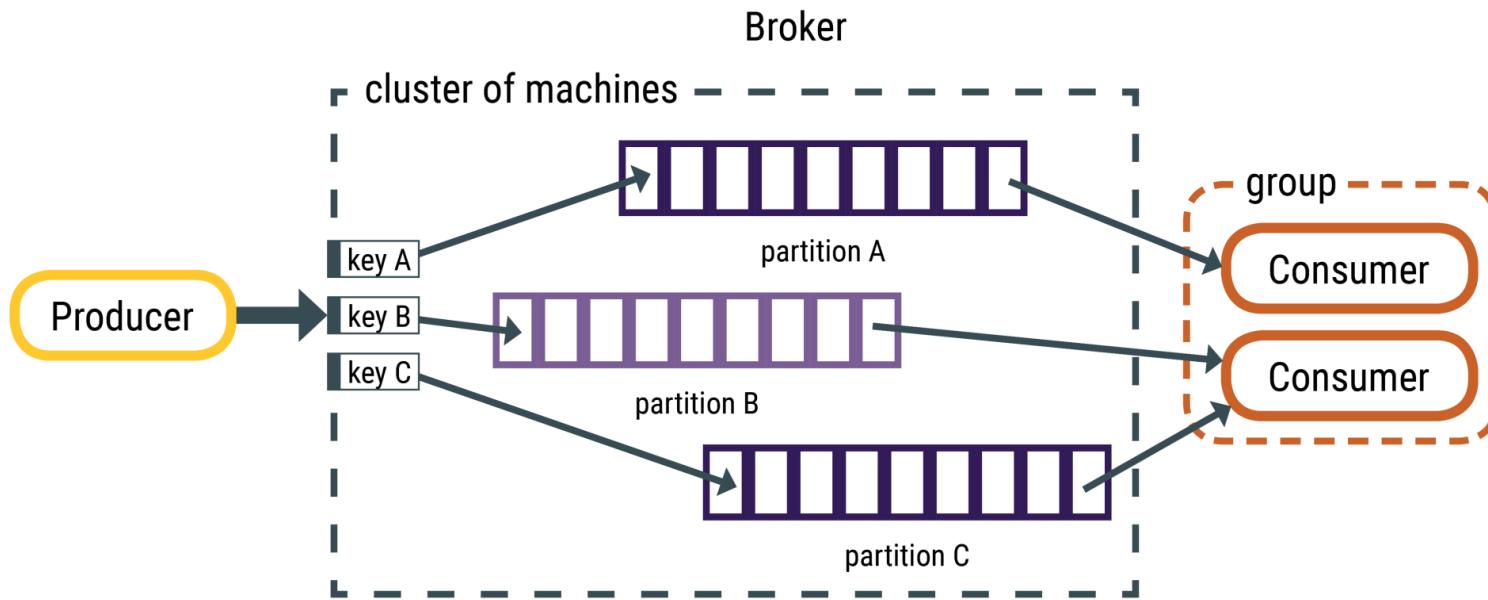




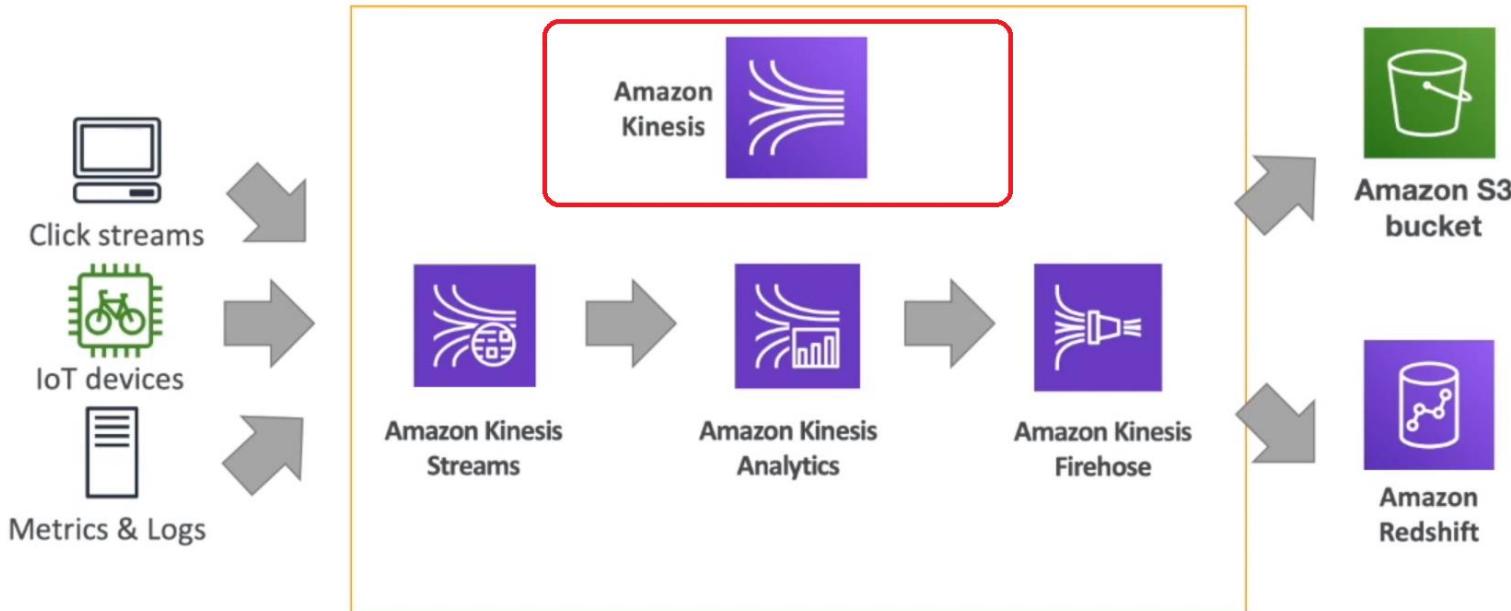




- *Producer Centric* (focado na garantia de entrega das mensagens entre Diversos broker);

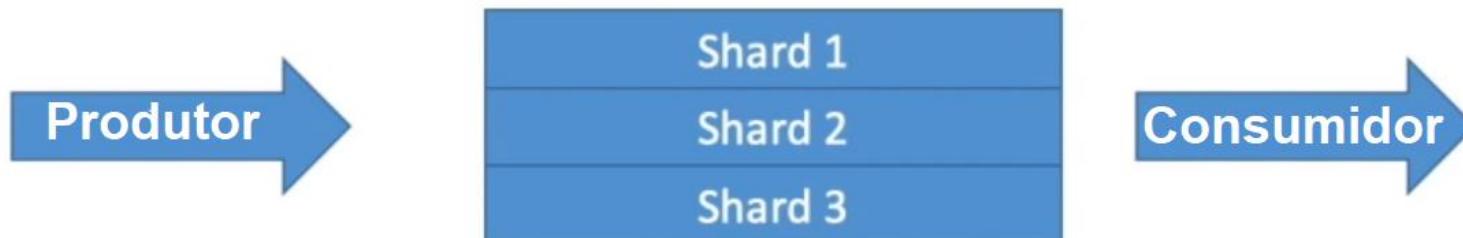


Amazon Kinesis – Arquitetura



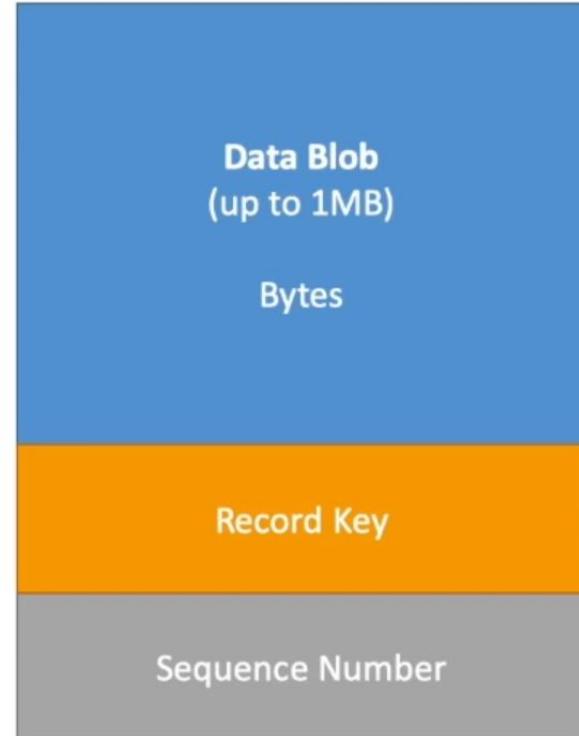
Amazon Kinesis Streaming

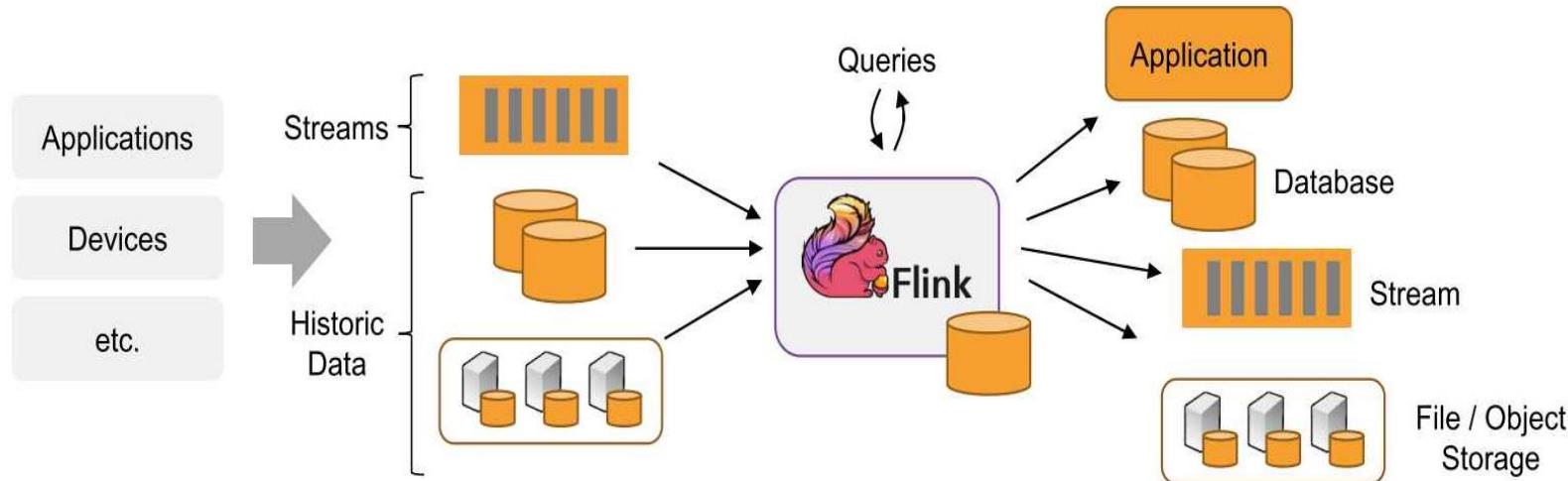
- Streams são divididos em Shards/Partições;
- São mantidos por 24h, mas pode ser aumentada para 7 dias;
- Replicar ou reprocessar os dados;
- Dados inseridos são imutáveis;
- Processamento em tempo real;

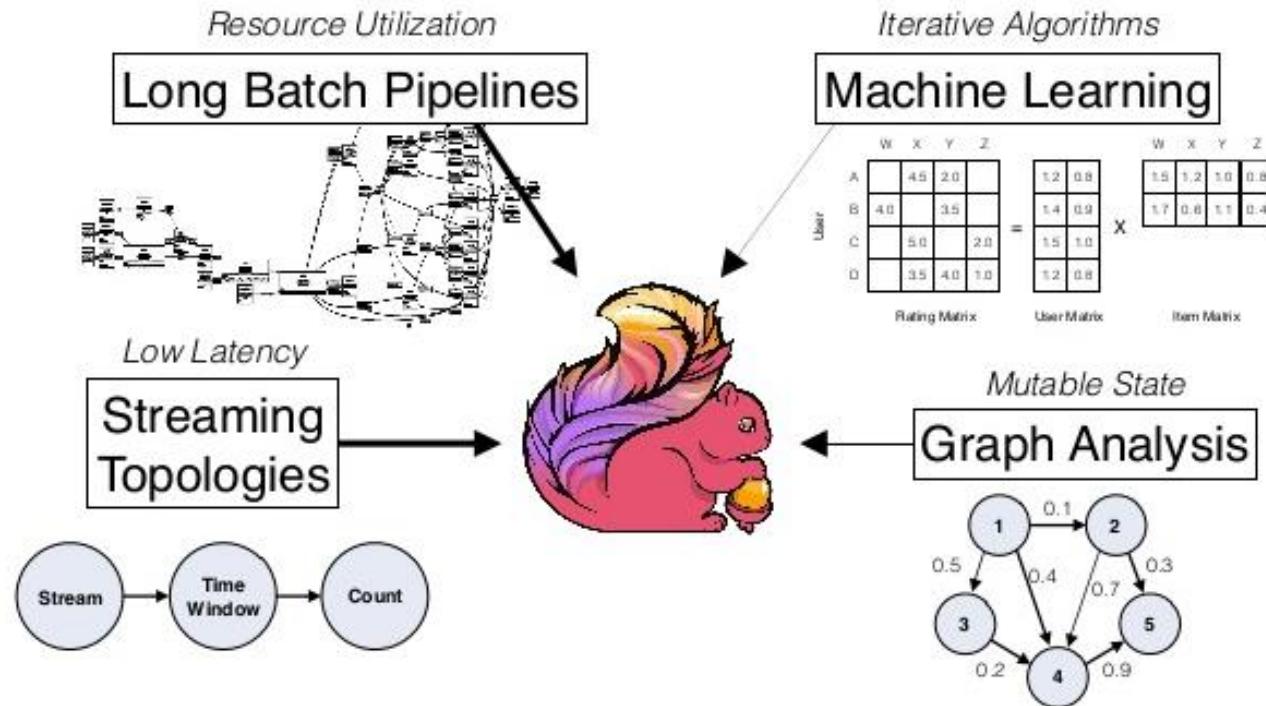


Kinesis – Shard/Partição

- Paga pela quantidade de “shards”;
- Data Blob:
 - Dado enviado.
- Record Key:
 - Agrupa os dados em um mesmo shard.
- Sequence number:
 - Identificador único para cada dado gravado no shard.







Kafka, Amazon Kinesis e Apache Flink

■ Próxima aula

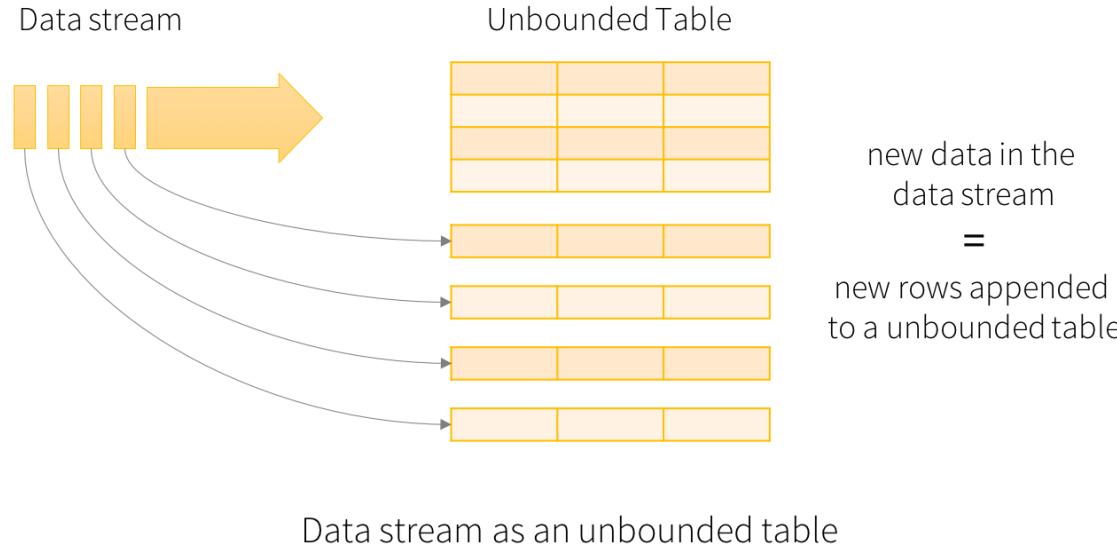
- ❑ Spark Streaming.



Aula 2.3. Spark Streaming

- Construir modelos utilizando o Spark Streaming.

Modelo construtivo do Structured Streaming



Fonte: SPARK, Apache. **Structured Streaming Programming Guide.**

Disponível em: <<https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html>>.
Acesso em: 04 fev. 2020.

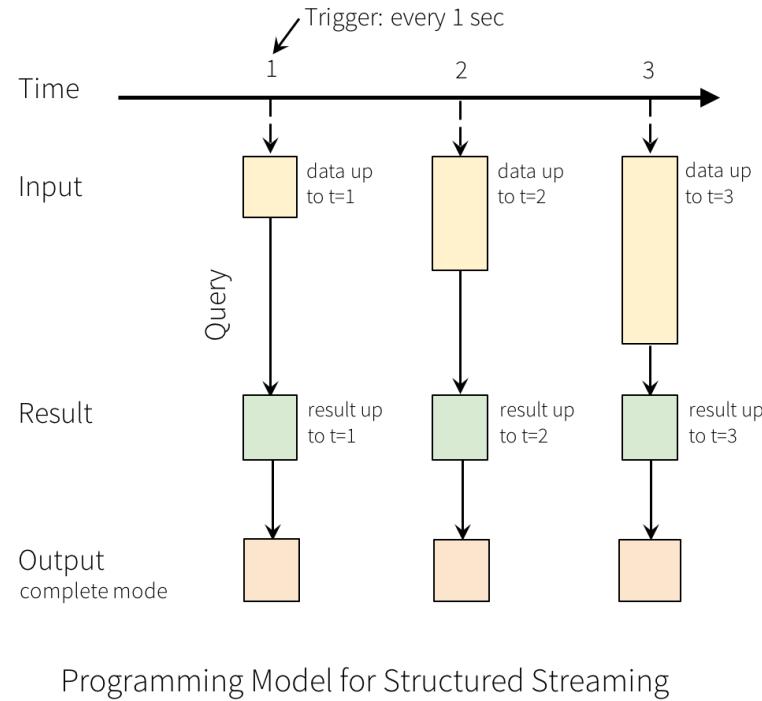
Temos que definir 3 diferentes etapas

1. De onde vamos receber os dados (source);
2. O que vamos fazer com os dados (query);
3. Para onde vamos enviar os resultados (sink).





Modelo de programação para Structured Streaming



Fonte: SPARK, Apache. **Structured Streaming Programming Guide.**

Disponível em: <<https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html>>.
Acesso em: 04 fev. 2020.

Podem ser definidos 3 modos de saída (Sink)

1. Modo Completo (Complete Mode);
2. Modo de Adição (Append Mode) - não pode ser utilizada com agregações;
1. Modo de Atualização (Updated Mode).

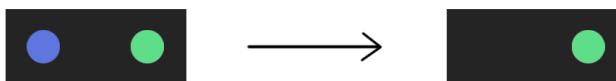
Complete Mode



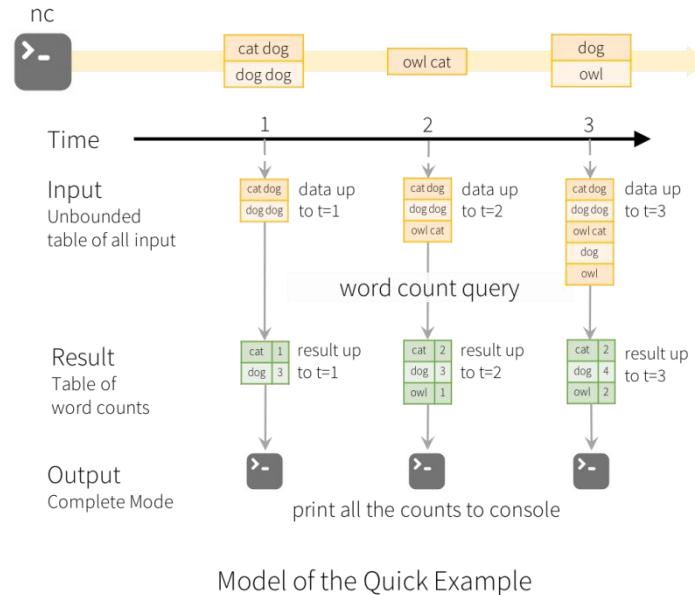
Update Mode



Append Mode



Como esse programa funciona



- O Structured Streaming mantém o mínimo possível do estado intermediário.

LET'S DO THIS

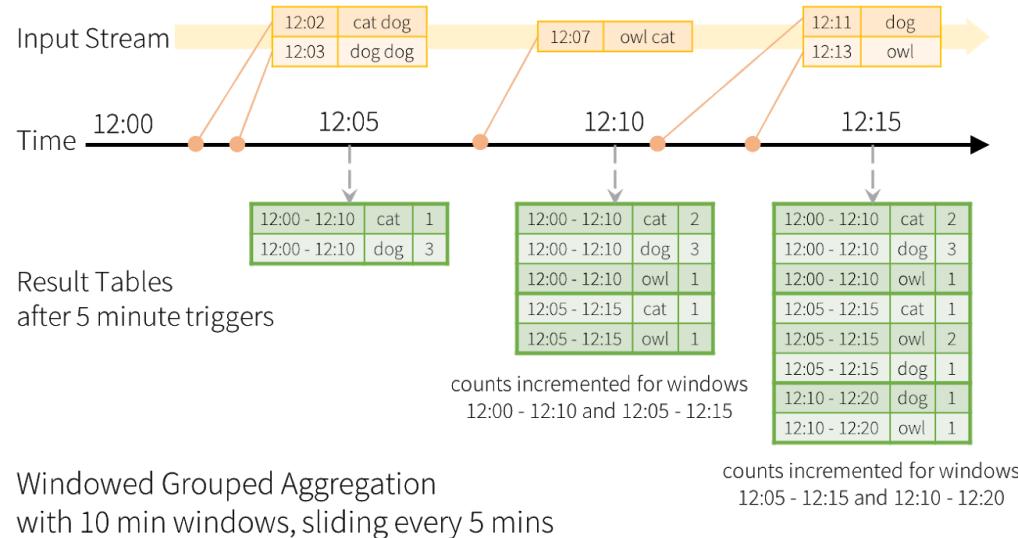
The text "LET'S DO THIS" is written in a bold, black, hand-drawn style font. The word "LET'S" is on the left, "DO" is in the center, and "THIS" is on the right. Above the text, there is a horizontal line of approximately 15 small yellow dots. Below the text, there is another horizontal line of approximately 15 small yellow dots.

Diferentes formatos de fonte de dados

- Arquivos (JSON, PARQUET, CSV, TEXT etc.);
- Socket (não garante tolerância a falhas);
- Kafka;
- Rate, usado para teste.



Janela de tempo sobre “Event-Time”





**KEEP
CALM
AND
LET'S
PRACTICE**

- Processamento utilizando o Spark Streaming

- ❑ Spark Streaming – Análise de sentimento via Twitter.

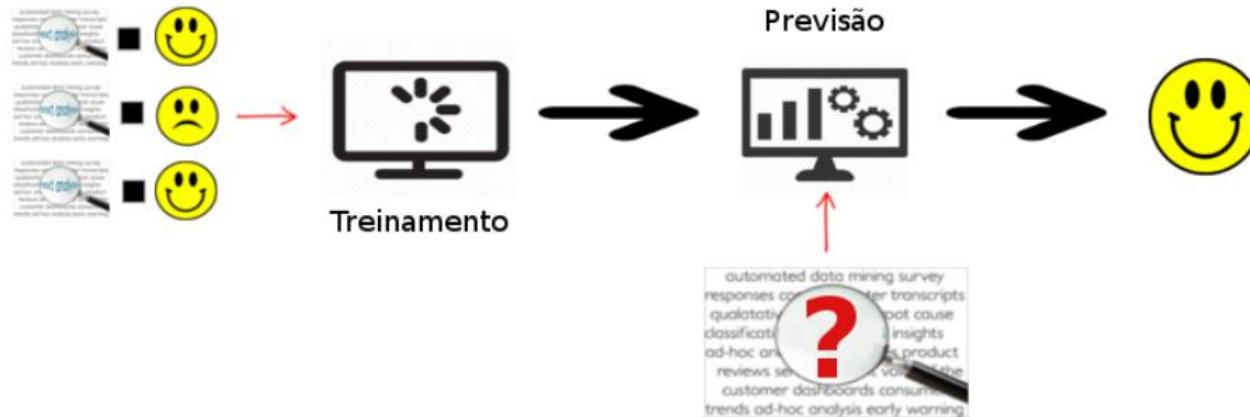


Aula 2.4. Análise de sentimento via Twitter

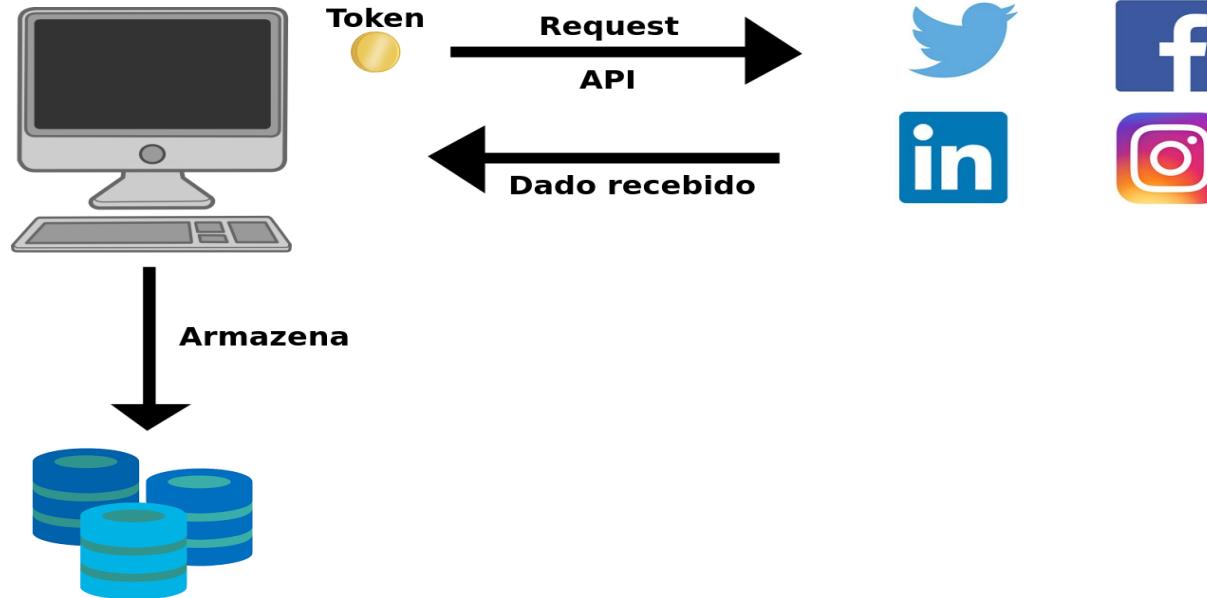
Nesta aula

- Aplicação da análise de sentimento via Twitter.

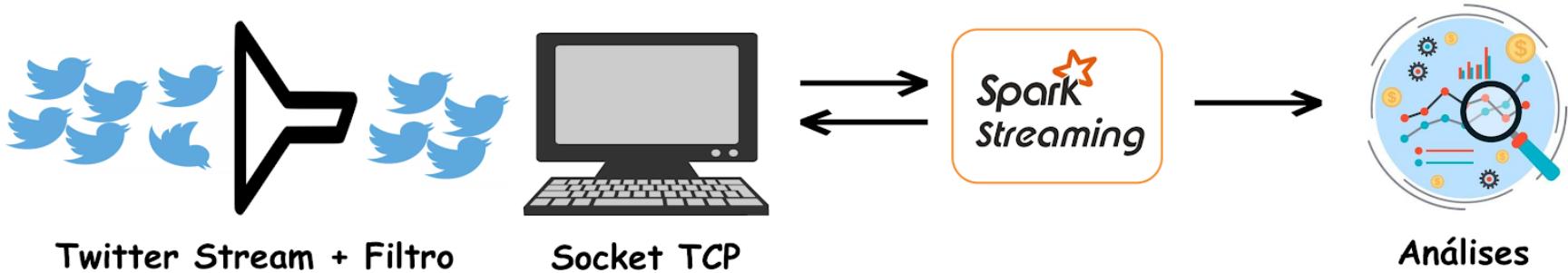
O que é análise de sentimento?



Utilizando o Streaming para o Twitter



Pipeline da aplicação



Isso na prática...

practice

practice

practice

practice

Conclusão

- ✓ Aplicação do processamento em streaming para análise de sentimento do Twitter.

■ Próxima aula

- Aplicação de streaming na análise de logs.



Aula 2.5. Análise de logs

Nesta aula

- ❑ Aplicação de streaming na análise de logs.

Aplicação

```
samplelog.log
1 #Software: Microsoft Internet Information Services X.X
2 #Version: X-
3 #Date: 2010-03-24 07:00:01-
4 #Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-
5 2010-03-24 07:00:01 ZZZZC941948879 RUFFLES 222.222.222.222 GET / - 80 - 220.181.7.113 HTTP/1.1
6 2010-03-24 07:00:23 ZZZZC941948879 RUFFLES 222.222.222.222 GET /2009/12/im_not_mean_im_just_ar-
7 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /terminal-blank.gif - 80 - 217.
8 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /grep-options.gif - 80 - 217.23
9 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /terminal-cat.gif - 80 - 217.23
10 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /terminal-pwd-cd.gif - 80 - 217.
11 2010-03-24 07:00:39 ZZZZC941948879 RUFFLES 222.222.222.222 GET /robots.txt - 80 - 95.55.207.95
12 2010-03-24 07:00:39 ZZZZC941948879 RUFFLES 222.222.222.222 GET /rss-short.xml - 80 - 173.45.23
13 2010-03-24 07:00:43 ZZZZC941948879 RUFFLES 222.222.222.222 GET /2009/08/22-things-you-dont-kno-
14 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /screen.css - 80 - 98.88.35.133
15 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/rss-header-red.gif - 80 -
16 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/logo.jpg - 80 - 98.88.35.1
17 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/input-emailsend.jpg - 80 -
18 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /images/cm-ebook-banner.gif - 8
19 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/bg.jpg - 80 - 98.88.35.133
20 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/bg-top.jpg - 80 - 98.88.35
21 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /21things/checkout-login.gif -
22 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/topnav-contact.jpg - 80 -
23 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /21things/portent-email-sub.gif
24 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /rss-header.jpg - 80 - 98.88.35
```

Aplicação





**KEEP
CALM
AND
LET'S
PRACTICE**

Conclusão

- Aplicação do processamento em streaming para análise de logs.

- Algoritmos de machine learning aplicados ao pré-processamento do Big Data.



Técnicas para o Processamento do Big Data

**Capítulo 3. Algoritmos de Machine Learning Aplicados
ao Pré-Processamento do Big Data**

Prof. Túlio Philipe Vieira



Aula 3.1. Detecção de anomalias

- ❑ Por que é importante identificar anomalias?
- ❑ O que são anomalias?
- ❑ Tipos de anomalias.
- ❑ Ferramentas para detecção de anomalias

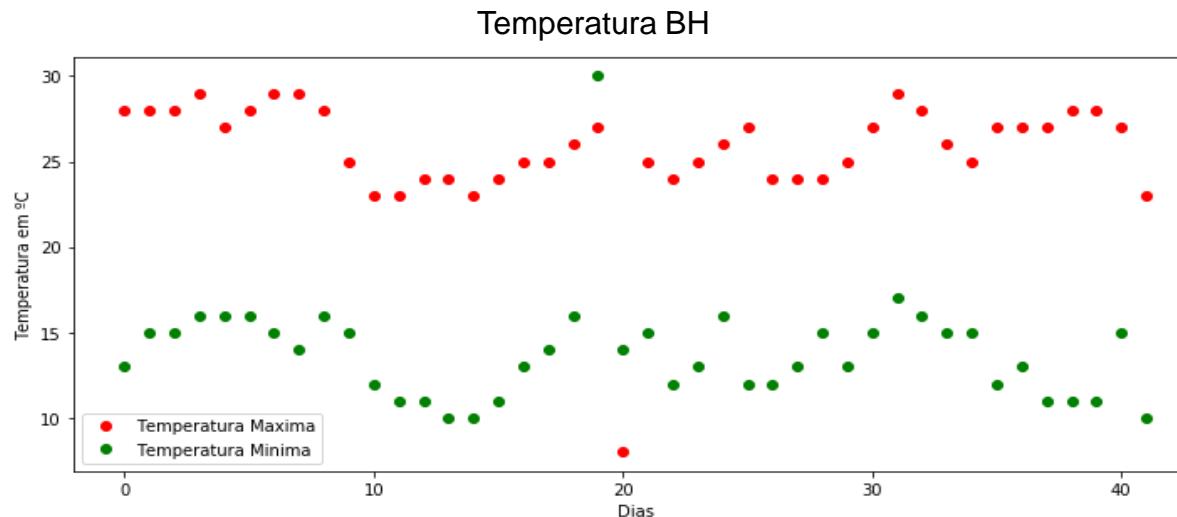
Importância em identificar anomalias

- Tornar a análise enviesada;
- Distorcer os padrões encontrados;
- Conduzir a uma análise equivocada;
- Colapso da aplicação.



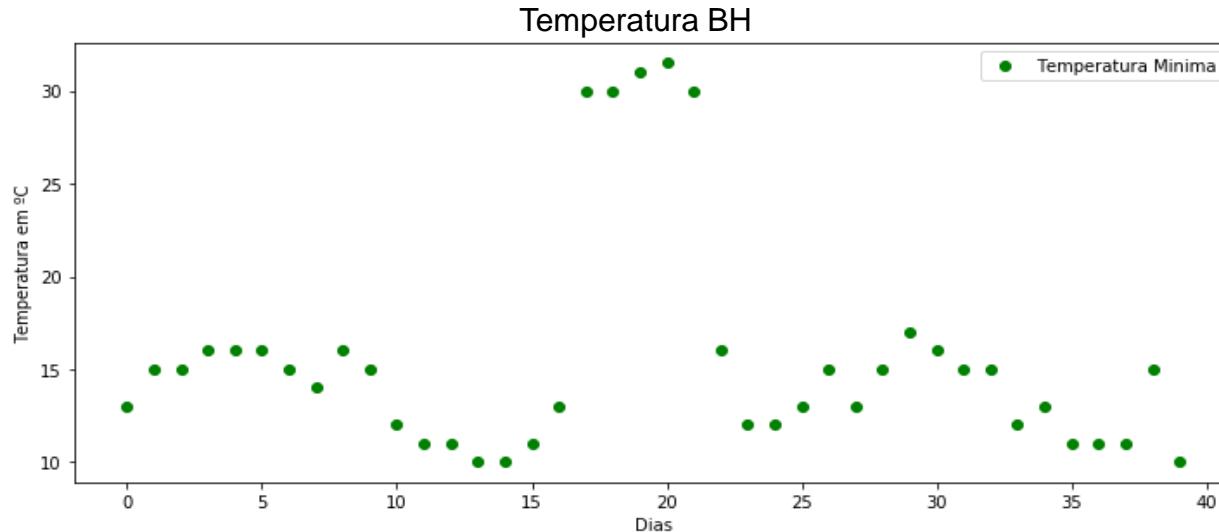
O que são anomalias?

- São dados ou conjunto de dados que não seguem o mesmo padrão ou a mesma estrutura do resto dos dados em um dataset.
- Anomalia de um ponto:
 - Apenas 1 ponto está fora do padrão.



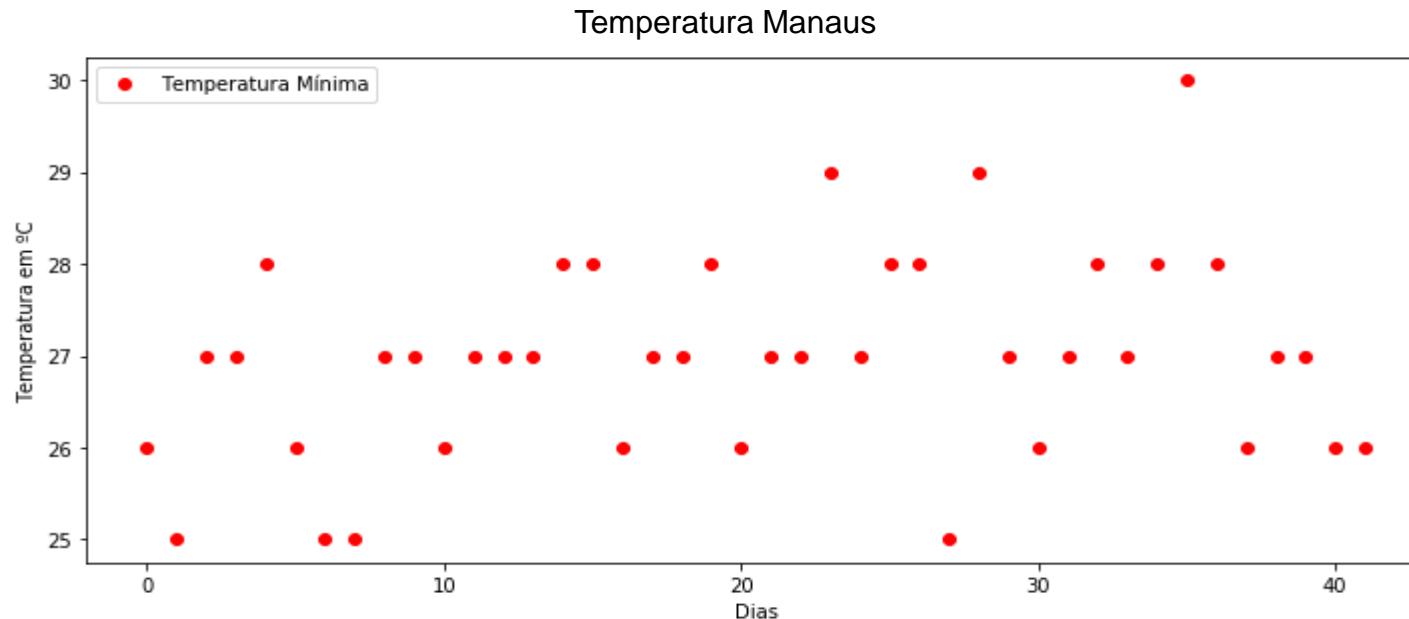
O que são anomalias?

- Anomalia coletiva.
 - Comportamento anormal de uma instância de dados.

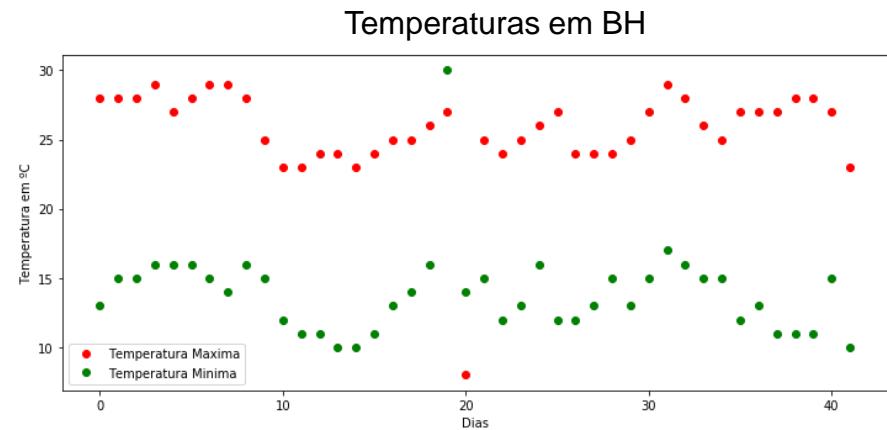
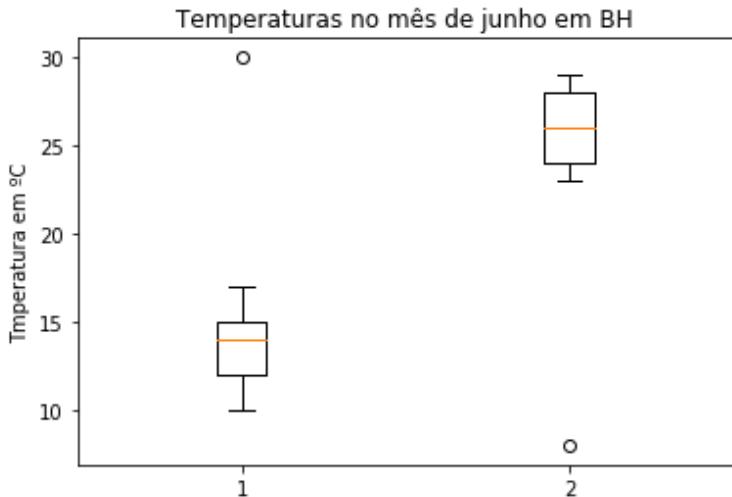


O que são anomalias?

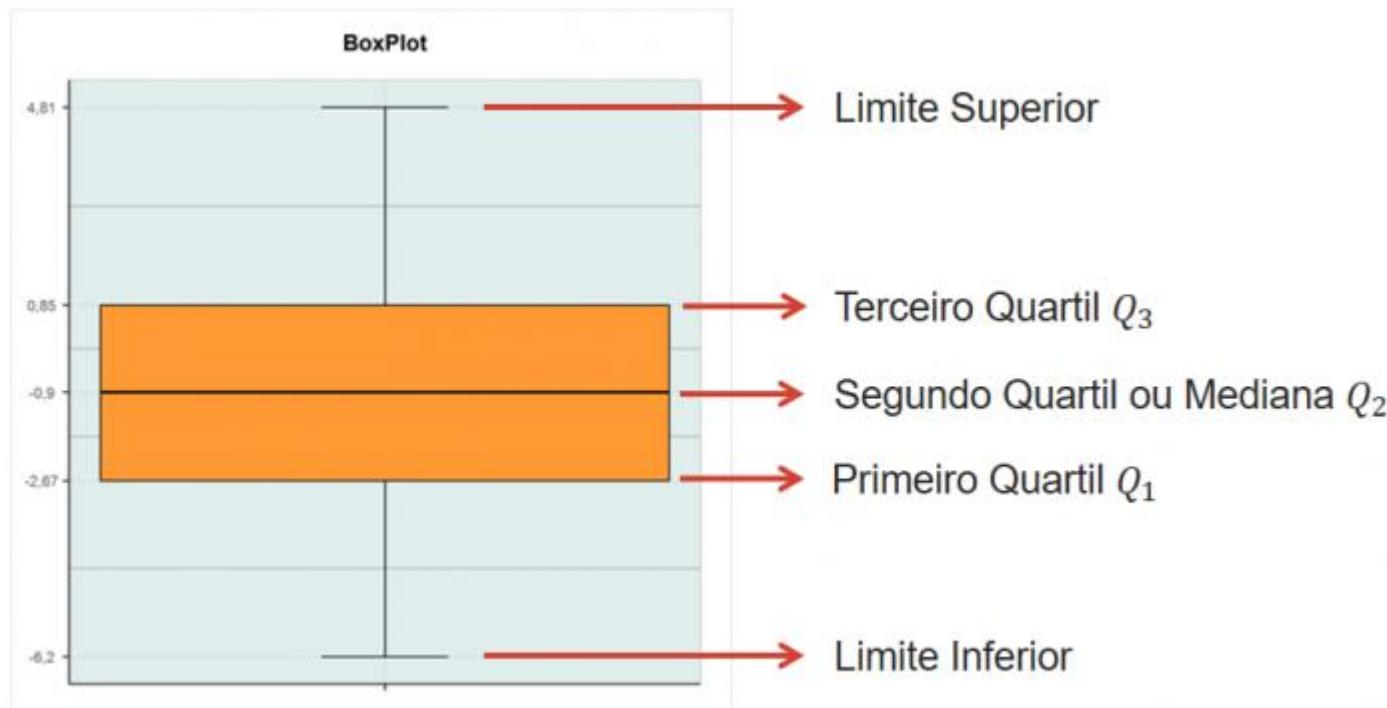
- Anomalia contextual.
 - Anomalia relacionada ao contexto dos dados.



Detecção de anomalias

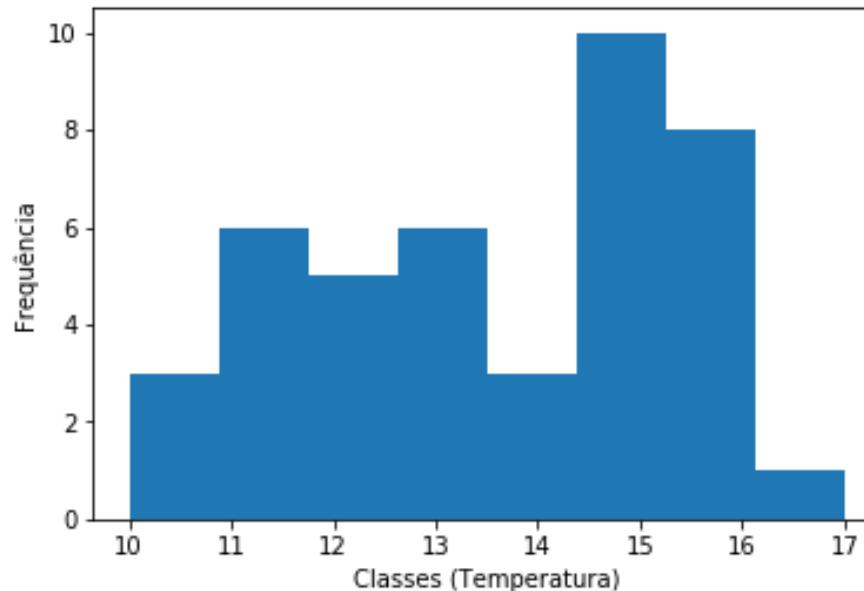


Boxplot



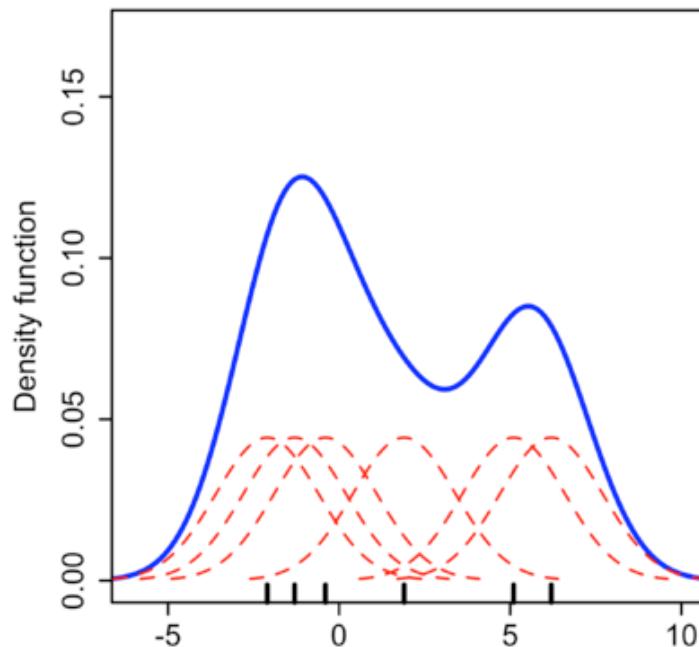
Histograma

- Utilizado para representar a distribuição dos dados.



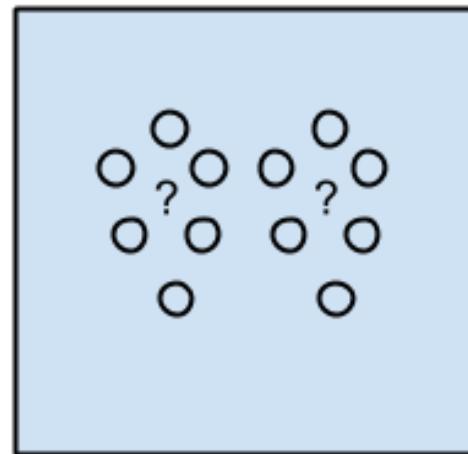
Plot de densidade (KDE)

- Utilizada para estimar a função densidade de probabilidade (PDE).



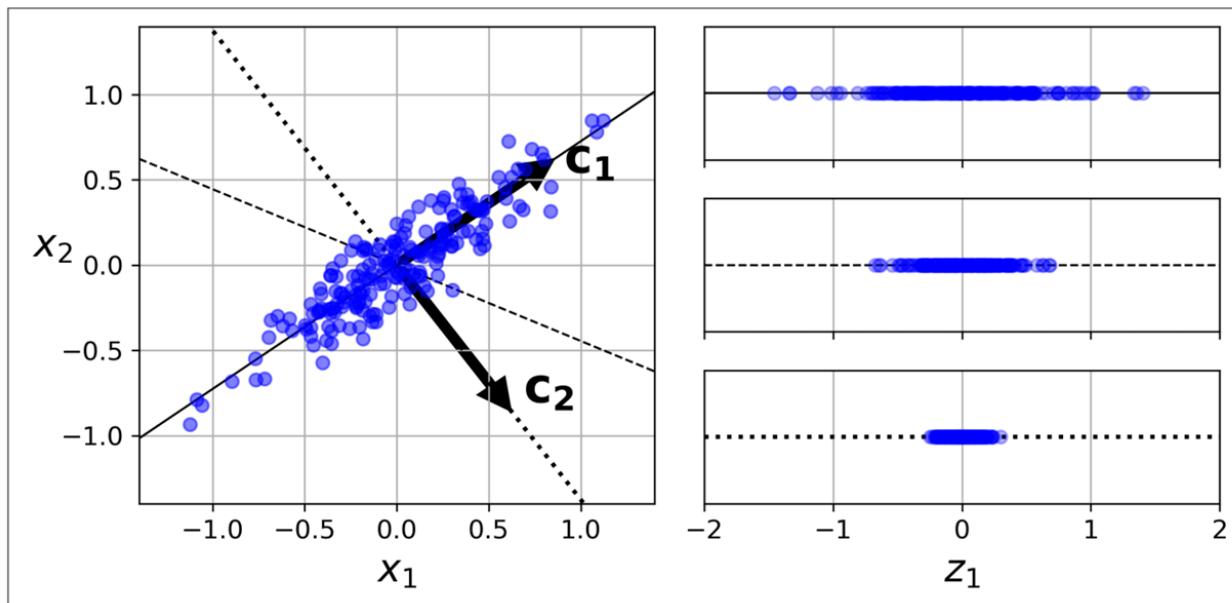
Algoritmos não supervisionados

- Não utilizam entrada x saída.



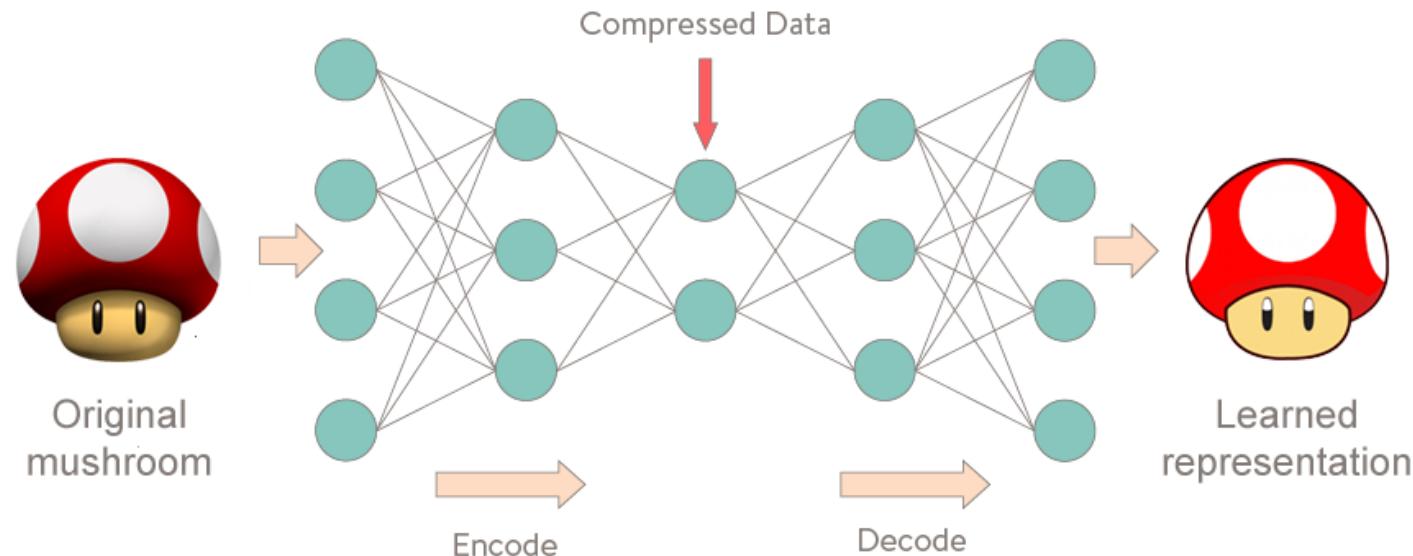
Unsupervised Learning
Algorithms

- Redução da dimensionalidade.

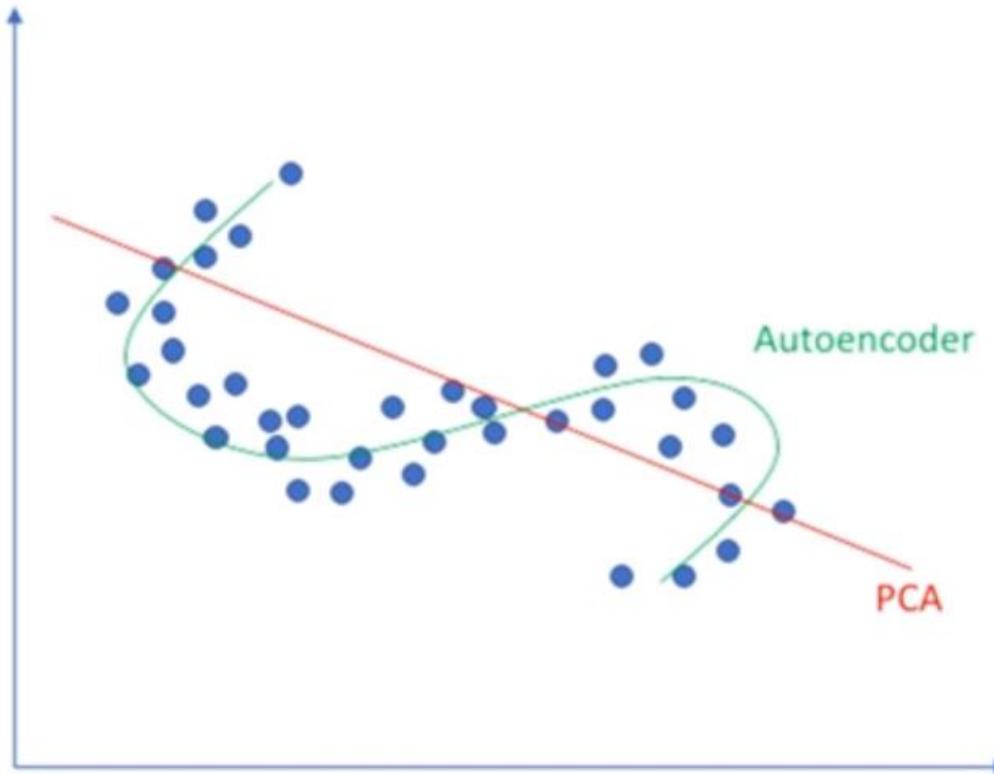


Autoencoder

- Redução da dimensionalidade.



PCA x Autoencoder



■ Próxima aula

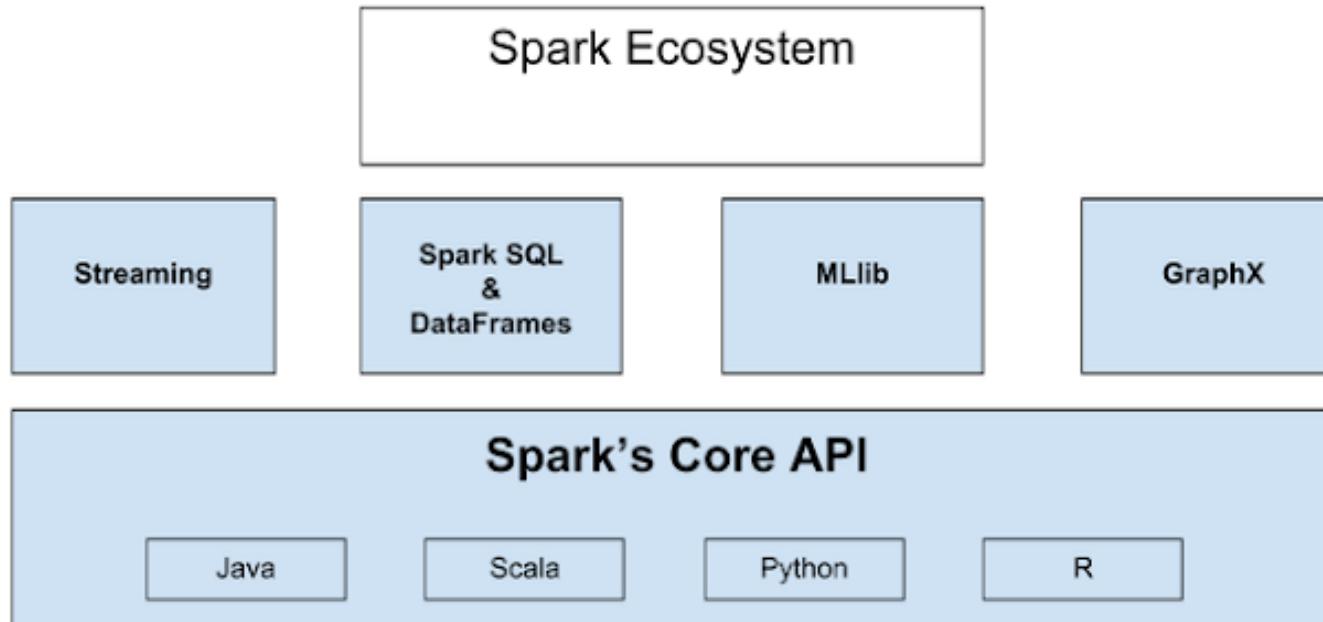
- Introdução ao Spark MLlib.



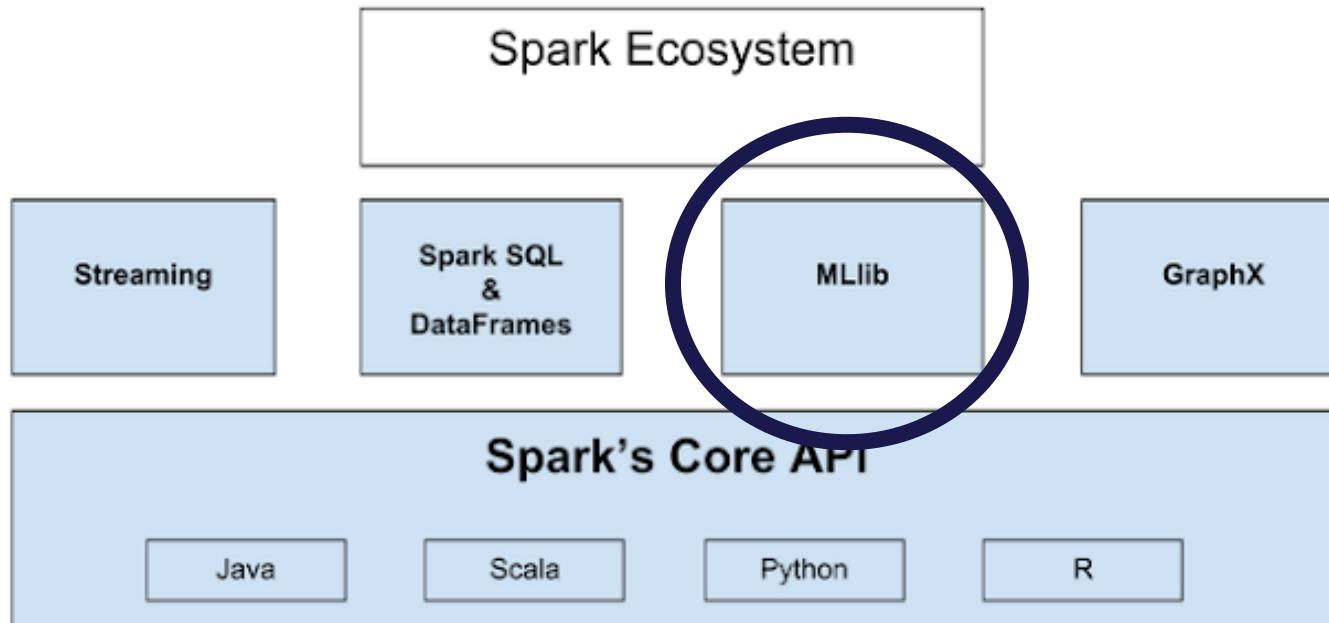
Aula 3.2. Introdução ao Spark MLlib

- O que é Spark MLlib?
- Como funciona o MLlib?
- Por que o MLlib?

Ecosistema Spark



Ecosistema Spark



O que é MLlib

- Alternativa ao MapReduce para algumas aplicações;
- Cluster para computação de baixa latência;
- Utilizada para processamento de um grande volume de dados;
- 100x mais rápida que o MapReduce;
- Utiliza o HDFS do Hadoop;
- Utiliza memória volátil.



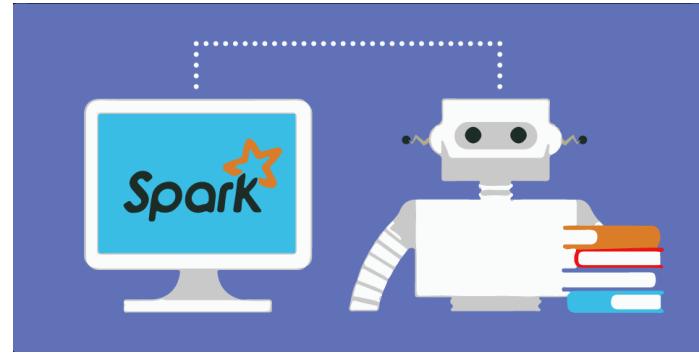
O que é MLlib?

- Biblioteca para machine learning do ecossistema Spark;
- Classificação;
- Regressão;
- Clusterização;
- Filtro colaborativo;
- Redução da dimensionalidade.



O que podemos fazer com Spark MLlib?

- Estatística básica;
- Extração de característica;
- Transformação nos dados;
- Otimização;
- Aplicação de vários algoritmos de ML.





Conclusão

- O que é Spark MLlib?
- Como funciona o MLlib?
- Por que o MLlib?

- ❑ MLlib para detecção de anomalias – Exemplo.



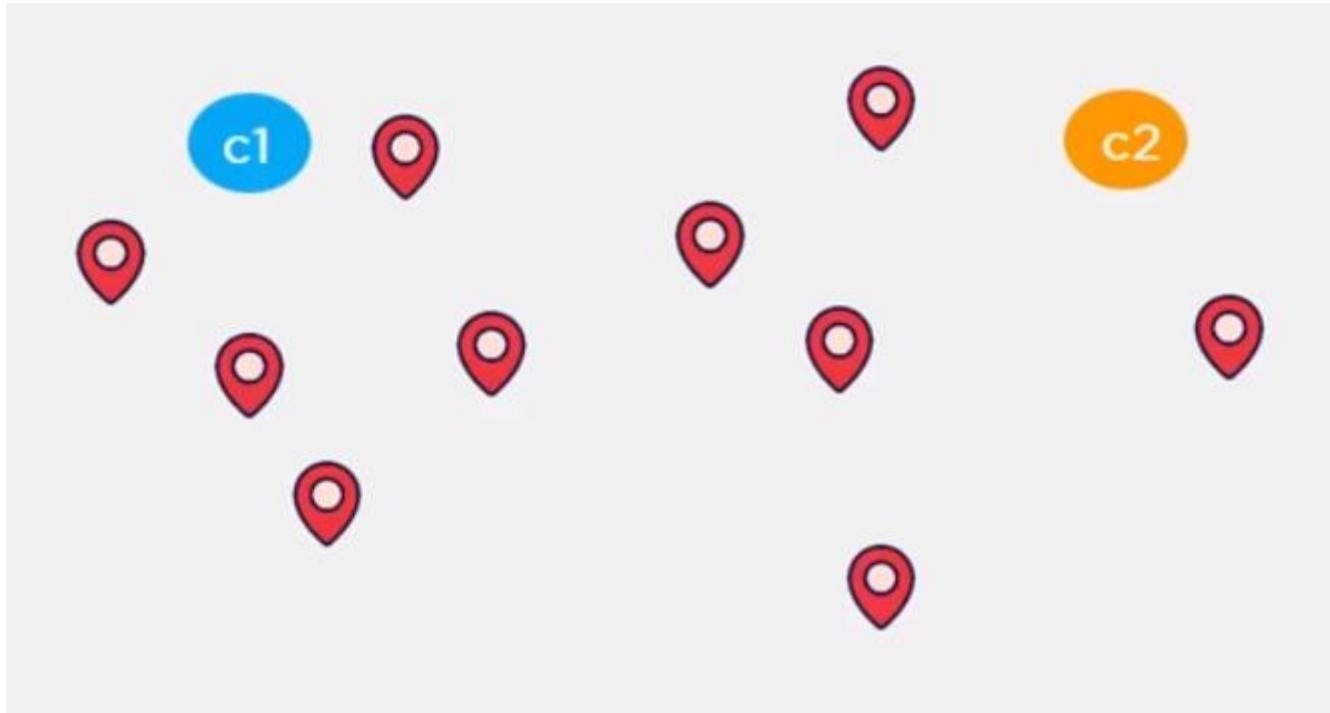
Aula 3.3. Exemplo – MLlib para detecção de anomalias

- Detecção de anomalias com o Spark.

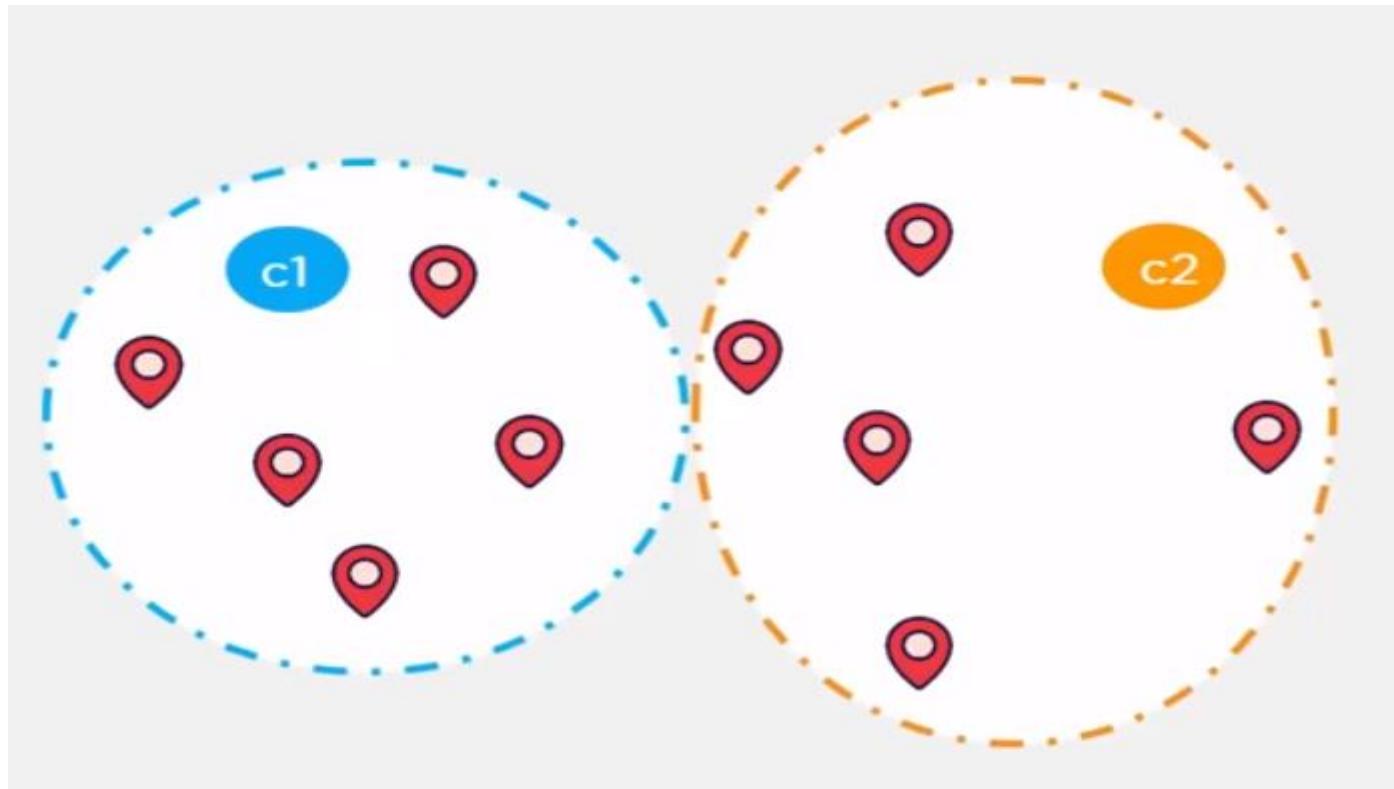


KEEP
CALM
AND
LET'S
PRACTICE

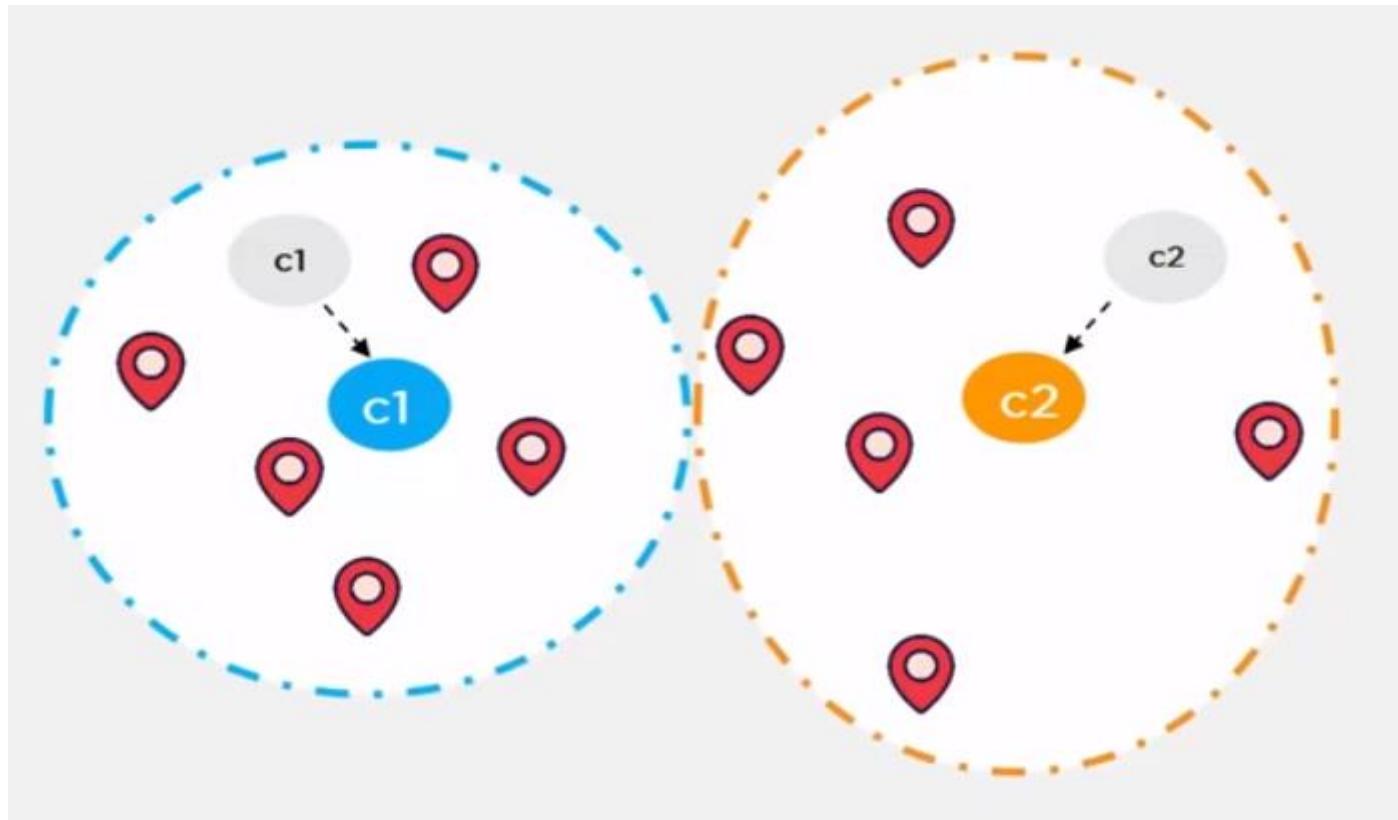
K-Means



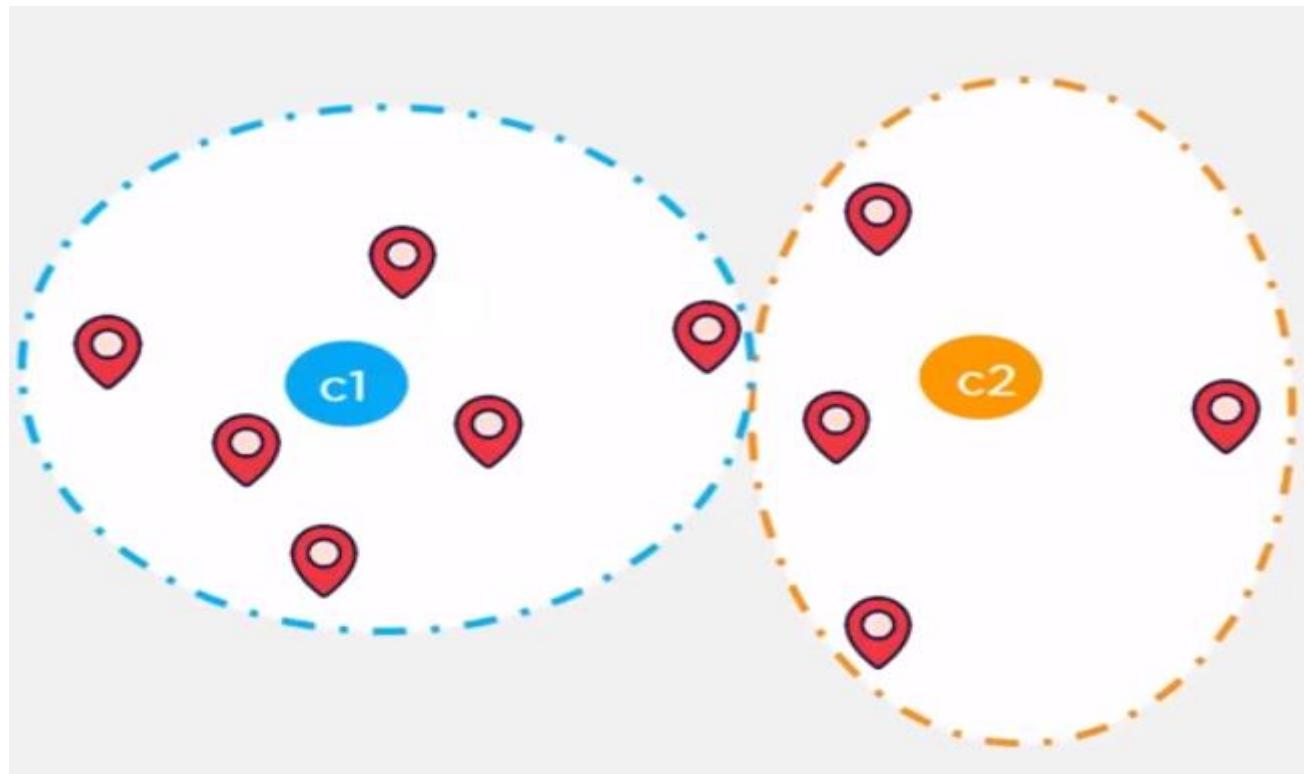
K-Means



K-Means



K-Means



Conclusão

- Detecção de anomalias com Spark.

■ Próxima aula

- Técnicas para preparação dos dados.

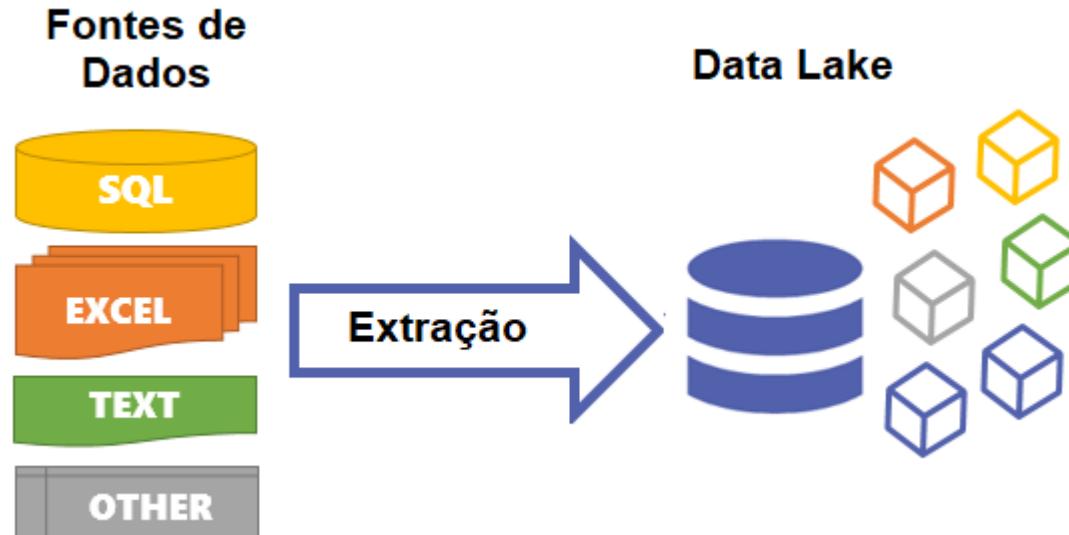


Aula 3.4. Técnicas para preparação dos dados

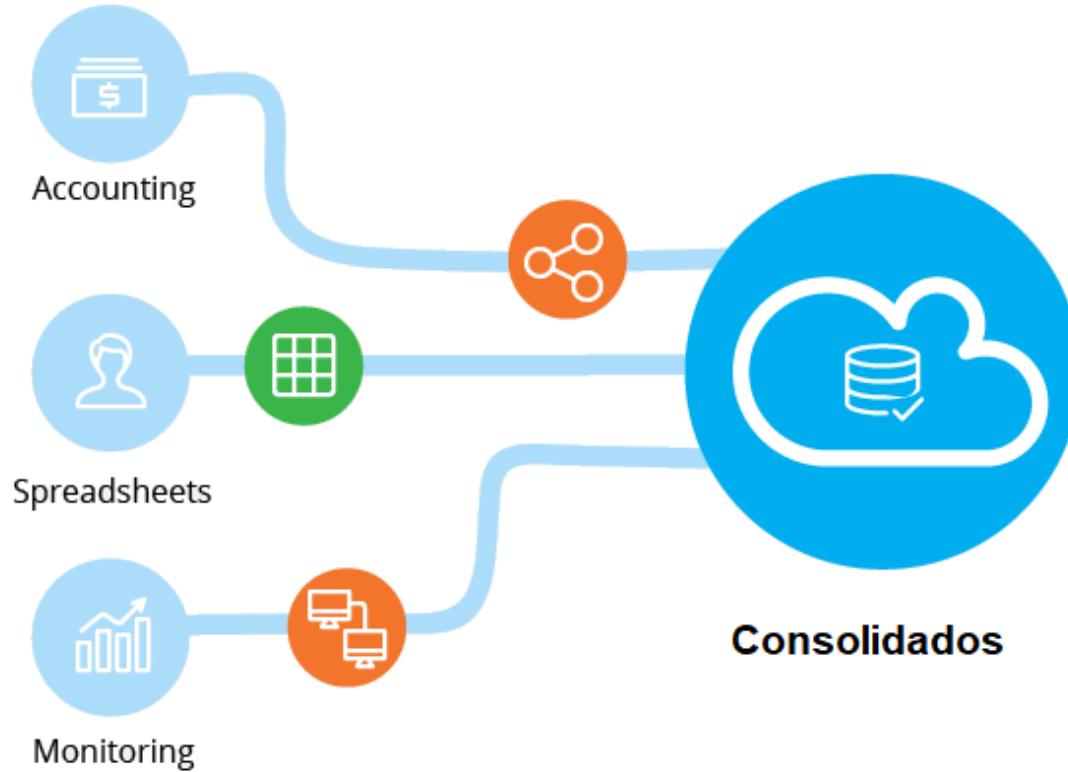
Nesta aula

- Aquisição de dados.
- Consolidação dos dados.
- Tratamento de dados faltosos.
- Tratamento de valores duplicados.
- Transformação dos dados.
- Correlações.
- Redução da dimensionalidade.

Aquisição de dados



Consolidação dos dados



O V de Veracidade

Min	Max	Media	Desvio Padrão
4.3	null	5.84	0.83
2.0	4.4	3.05	5000000
1500	7.9	1.20	0.25
0.1	2.5	?	0.75

Lidando com a veracidade

- Apenas retirar as linhas com NaN;

Min	Max	Media	Desvio Padrão
4.3	null	5.84	0.83
2.0	4.4	3.05	5000000
1500	7.9	1.20	0.25
0.1	2.5	?	0.75

Lidando com a veracidade

- Apenas retirar as linhas com NaN;
- Substituir por um valor fixo;

Min	Max	Media	Desvio Padrão
4.3	null	5.84	0.83
2.0	4.4	3.05	5000000
1500	7.9	1.20	0.25
0.1	2.5	?	0.75

Lidando com a veracidade

- Apenas retirar as linhas com NaN;
- Substituir por um valor fixo;
- Substituir pela média;

Min	Max	Media	Desvio Padrão
4.3	null	5.84	0.83
2.0	4.4	3.05	5000000
1500	7.9	1.20	0.25
0.1	2.5	?	0.75

Lidando com a veracidade

- Apenas retirar as linhas com NaN;
- Substituir por um valor fixo;
- Substituir pela média;
- Prever o valor faltoso;

Min	Max	Media	Desvio Padrão
4.3	null	5.84	0.83
2.0	4.4	3.05	5000000
1500	7.9	1.20	0.25
0.1	2.5	?	0.75

Valores duplicados

	id	first_name	last_name	email
▶	1	Carine	Schmitt	carine.schmitt@verizon.net
	4	Janine	Labrunе	janine.labrunе@aol.com
	6	Janine	Labrunе	janine.labrunе@aol.com
	2	Jean	King	jean.king@me.com
	12	Jean	King	jean.king@me.com
	5	Jonas	Bergulfsen	jonas.bergulfsen@mac.com
	10	Julie	Murphy	julie.murphy@yahoo.com
	11	Kwai	Lee	kwai.lee@google.com
	3	Peter	Ferguson	peter.ferguson@google.com
	9	Roland	Keitel	roland.keitel@yahoo.com
	14	Roland	Keitel	roland.keitel@yahoo.com
	7	Susan	Nelson	susan.nelson@comcast.net
	13	Susan	Nelson	susan.nelson@comcast.net
	8	Zbyszek	Piestrzeniewicz	zbyszek.piestrzeniewicz@att.net

Transformação dos dados

- Agrupar colunas;

Tempo (h)	Distância (km)
1	80
2	110
3	90



Velocidade Média (km/h)
80
110
90

Transformação dos dados

- Agrupar colunas;
- Transformação de unidades;

Velocidade Média (km/h)
80
110
90



Velocidade Média (m/s)
22,2
30,5
25

Transformação dos dados

- Agrupar colunas;
- Transformação de unidades;
- Dados categóricos em numéricos;

Categorias
Azul
Branco
Vermelho
Amarelo



Categorias
0
1
2
3

Transformação dos dados

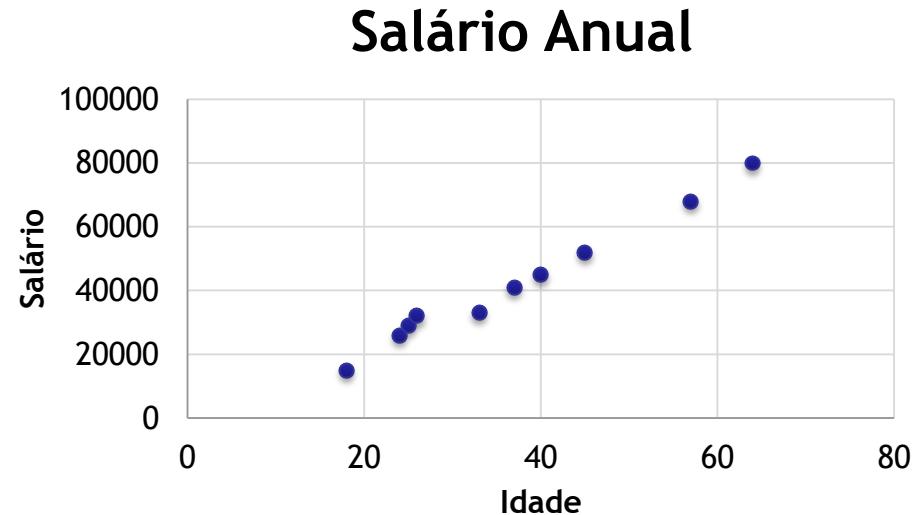
- Agrupar colunas;
- Transformação de unidades;
- Dados categóricos em numéricos;
- Transposição.

$$B_{3x2} = \begin{bmatrix} 1 & 5 \\ 7 & 3 \\ 8 & 2 \end{bmatrix}$$
$$B_{2x3}^t = \begin{bmatrix} 1 & 7 & 8 \\ 5 & 3 & 2 \end{bmatrix}$$

The diagram illustrates the concept of matrix transposition. It shows two matrices: B_{3x2} and B_{2x3}^t . The matrix B_{3x2} is a 3x2 matrix with elements [1, 5], [7, 3], and [8, 2]. The matrix B_{2x3}^t is a 2x3 matrix with elements [1, 7, 8] and [5, 3, 2]. A blue curved arrow points from B_{3x2} to B_{2x3}^t , indicating the transformation. A red curved arrow points from B_{2x3}^t back to B_{3x2} , indicating the inverse transformation.

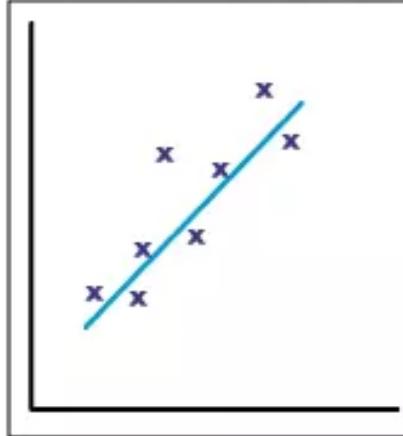
O que é correlação?

Medida	Idade	Salário Anual
1	18	R\$ 15000
2	25	R\$ 29000
3	57	R\$ 68000
4	45	R\$ 52000
5	26	R\$ 32000
6	64	R\$ 80000
7	37	R\$ 41000
8	40	R\$ 45000
9	24	R\$ 26000
10	33	R\$ 33000

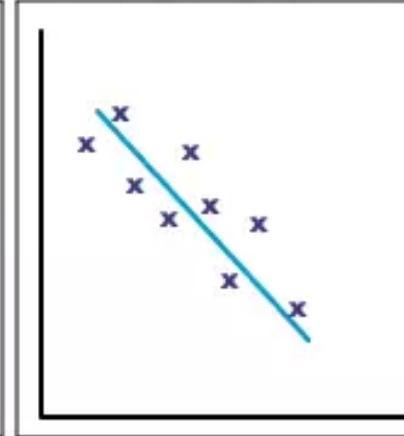


Possibilidades da correlação

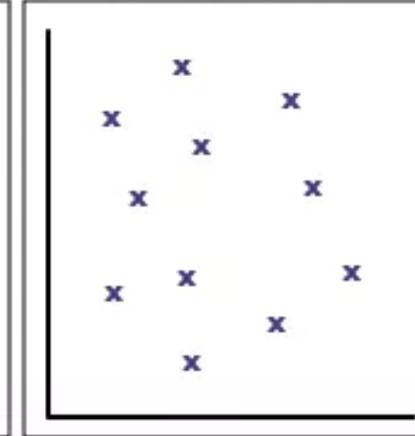
Correlação Positiva



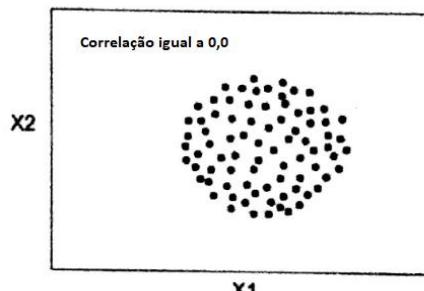
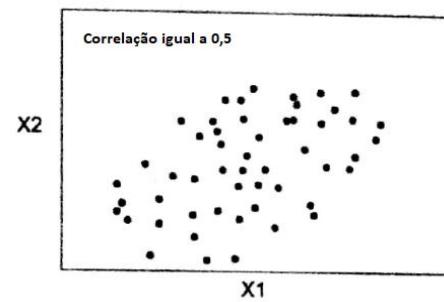
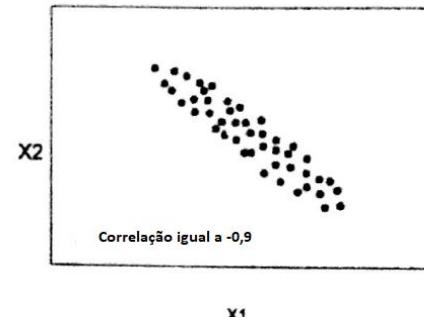
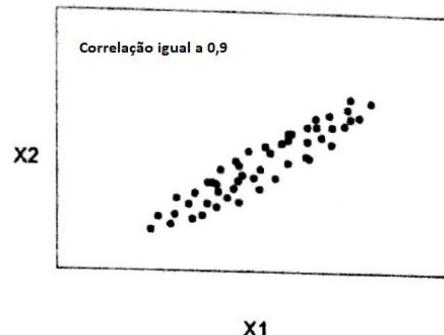
Correlação Negativa



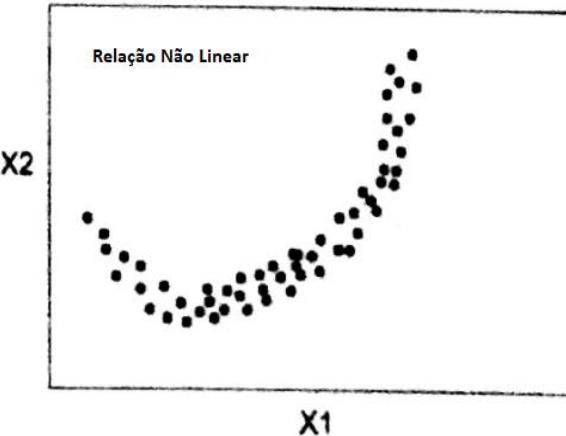
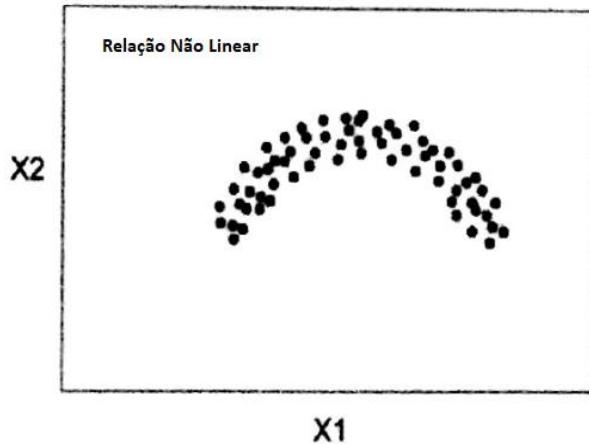
Sem Correlação



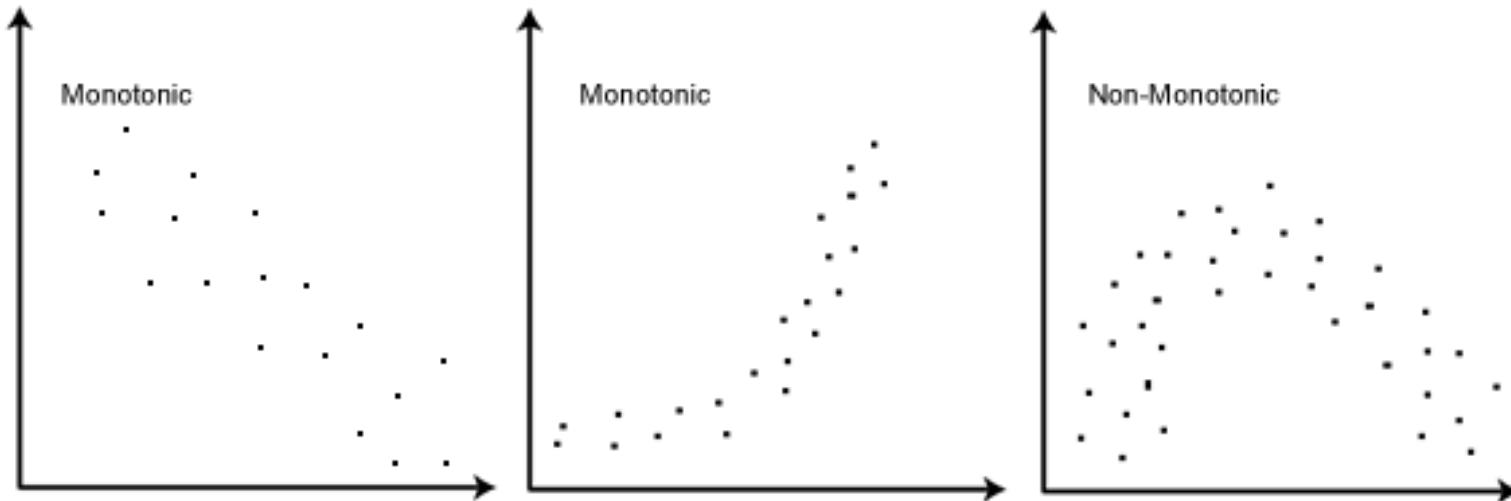
Possibilidades da correlação



Aspectos da correlação

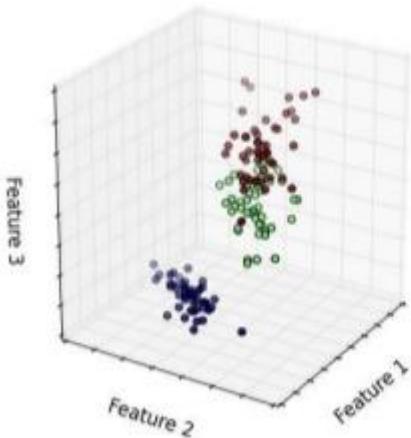


Correlação de Spearman

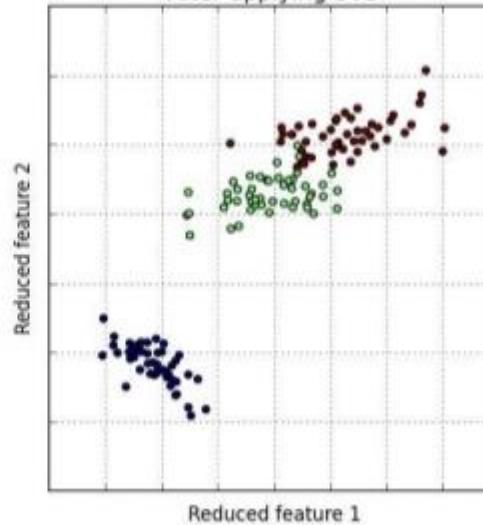


Redução da dimensionalidade

An example of SVD for dimensionality reduction on the Iris dataset
Before applying SVD



After applying SVD



Conclusão

- Aquisição de dados.
- Consolidação dos dados.
- Tratamento de dados faltosos.
- Tratamento de valores duplicados.
- Transformação dos dados.
- Correlações.
- Redução da dimensionalidade.

- Técnicas para preparação dos dados – Exemplo.



Aula 3.5. Exemplo – Preparação dos dados com MLlib

Nesta aula

- Aplicação para preparação dos dados.



Conclusão

- Aplicação utilizando Spark na preparação dos dados.

■ Próxima aula

- ❑ ML para o processamento do Big Data.



Técnicas para o Processamento do Big Data

**Capítulo 4. Algoritmos de Machine Learning Aplicados
ao Processamento do Big Data**

Prof. Túlio Philipe Vieira



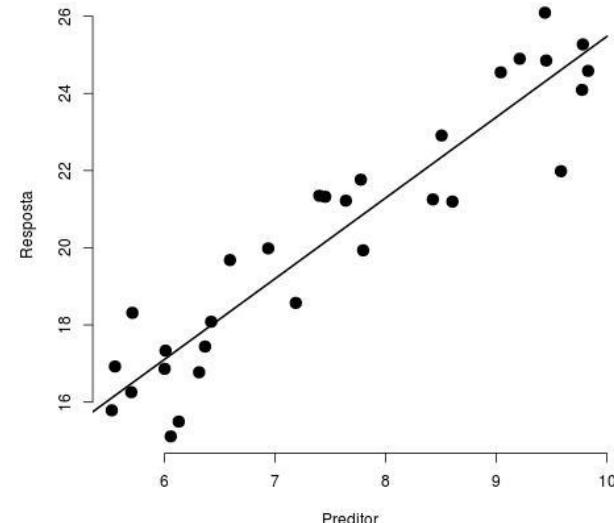
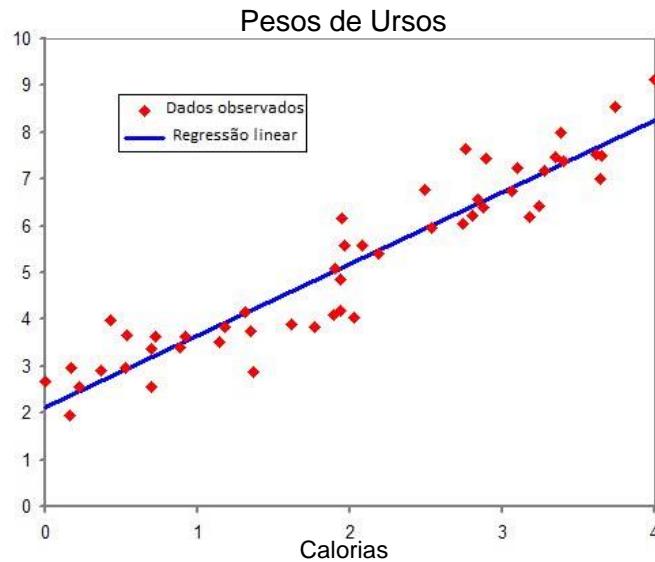
Aula 4.1.1. Algoritmos de ML para o processamento do Big Data (Parte I)

Nesta aula

- Análise de regressão.

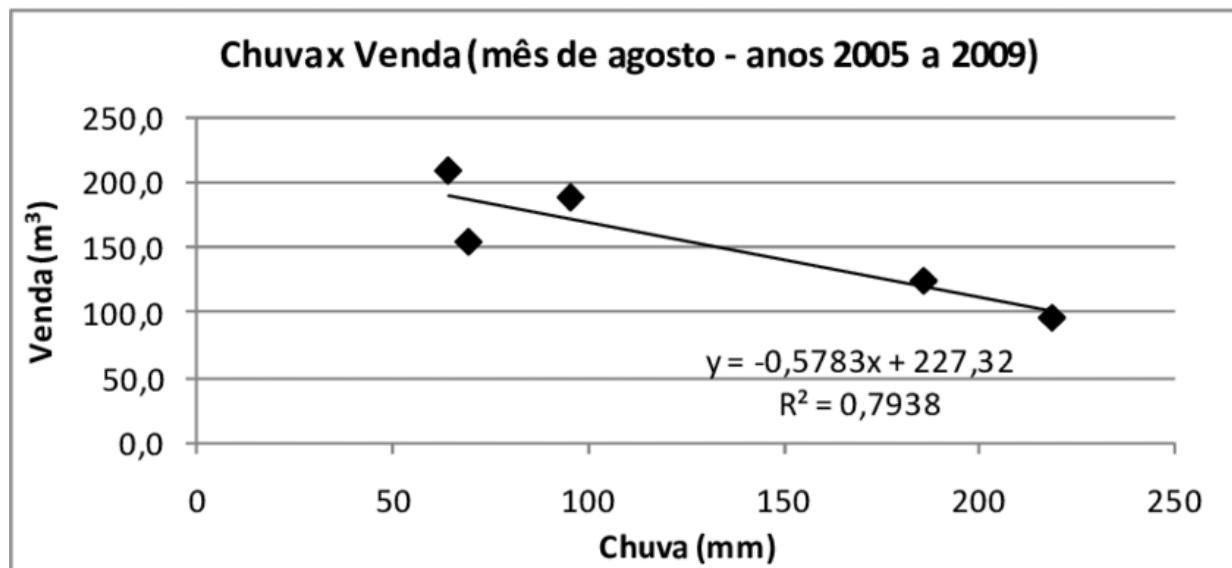
O que é regressão linear

- Encontrar a equação linear que representa a relação entre duas variáveis.
- Variáveis: **independente x dependente**.



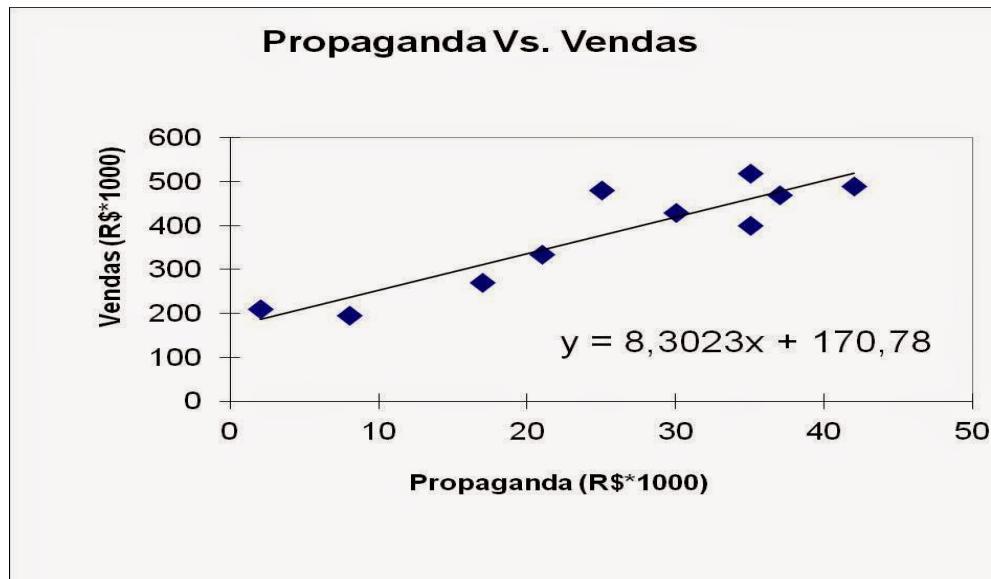
Onde encontramos regressão linear

- Determinar a relação entre duas variáveis (como a variável independente afeta a dependente).



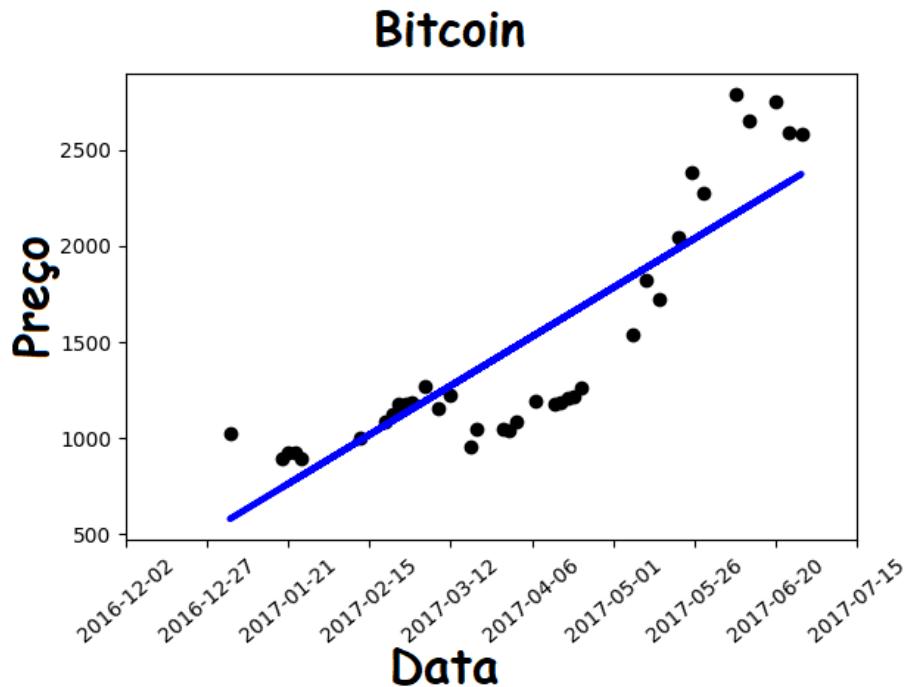
Onde encontramos regressão linear

- Prever um efeito.

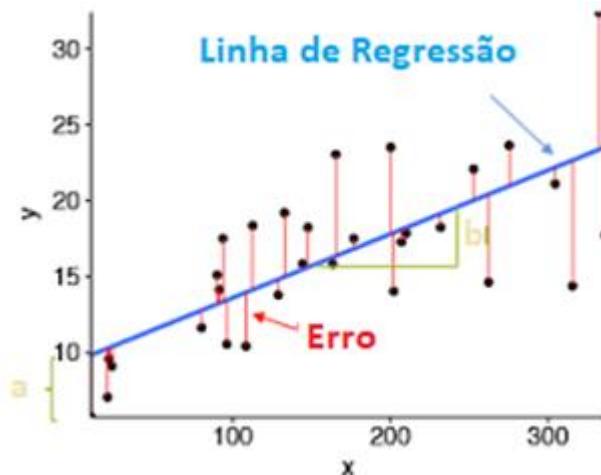


Onde encontrados regressão linear

- Prever tendências.



O algoritmo da regressão linear

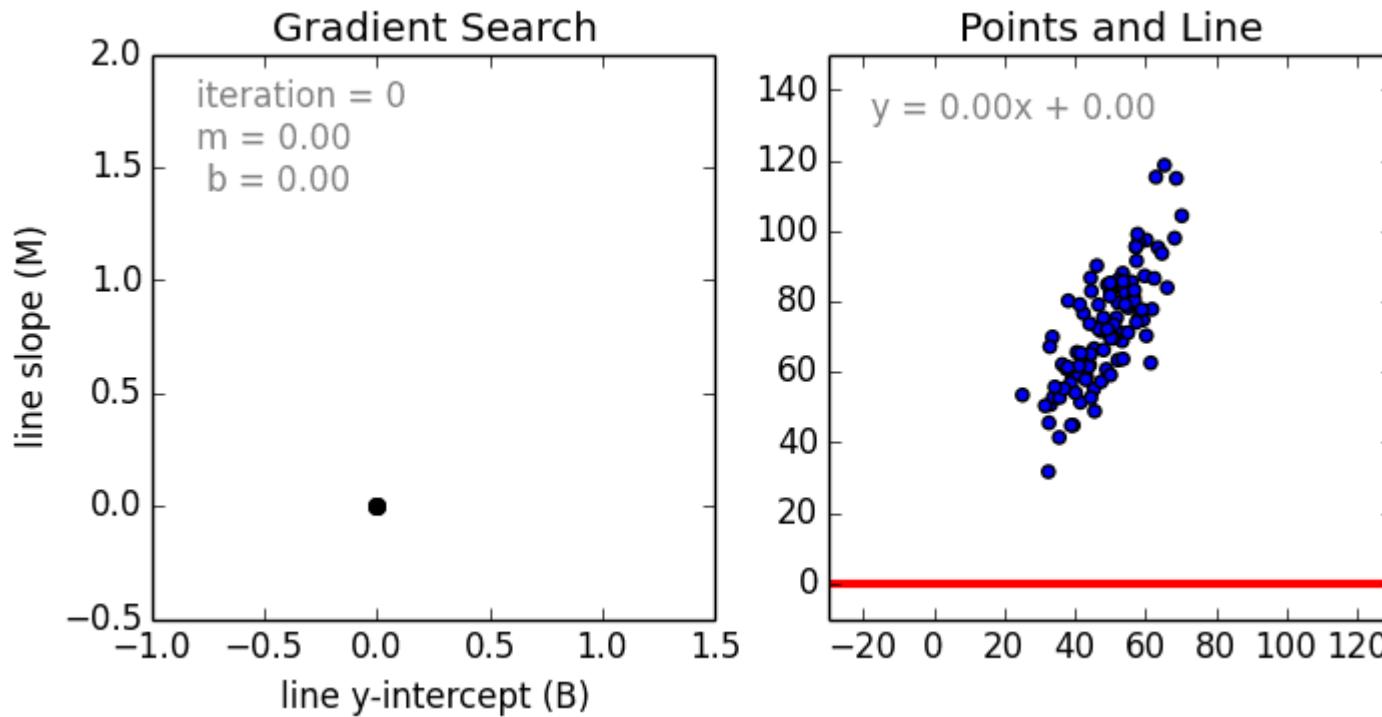


$$Y = ax + b$$

$$Erro = Y - Y'$$

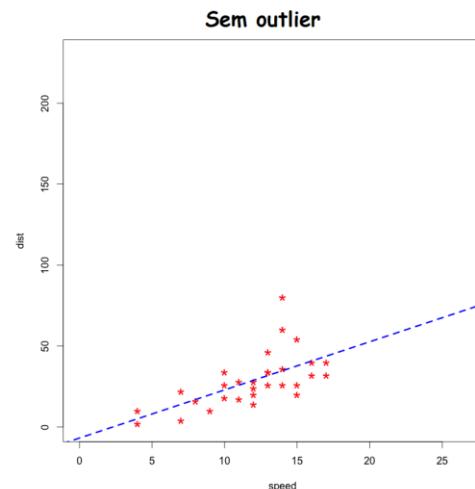
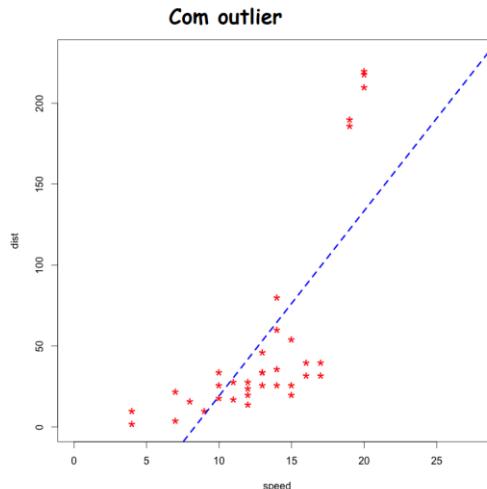
$$MSE = \frac{1}{n} \sum (Y_i - Y'_i)^2$$

Algoritmo da regressão linear



Características da regressão linear

- Dependentes da qualidade dos dados (sensível a outlier).
- Baixa complexidade computacional.
- Fácil compreensão.



■ Coeficiente de determinação

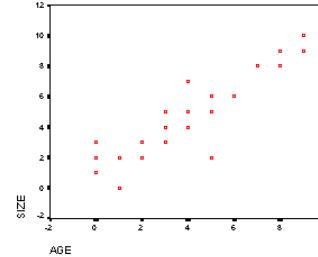
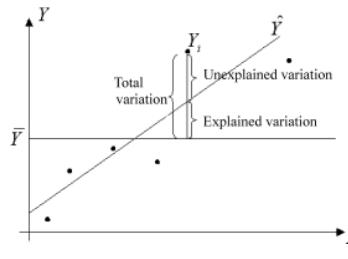
$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}}$$

Coeficiente de determinação

$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}}$$

$$r^2$$
$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Coeficiente de determinação x correlação



- Mede o quanto a variância de uma variável é explicada pela outra;
- Varia em 0 e 1;
- Não é simétrico;
- O que é explicado pela relação linear entre X e Y.
- Mede o grau de associação entre as variáveis (relação);
- Varia em -1 e 1;
- É simétrico;
- Dependência linear.

Conclusão

Regressão.

■ Próxima aula

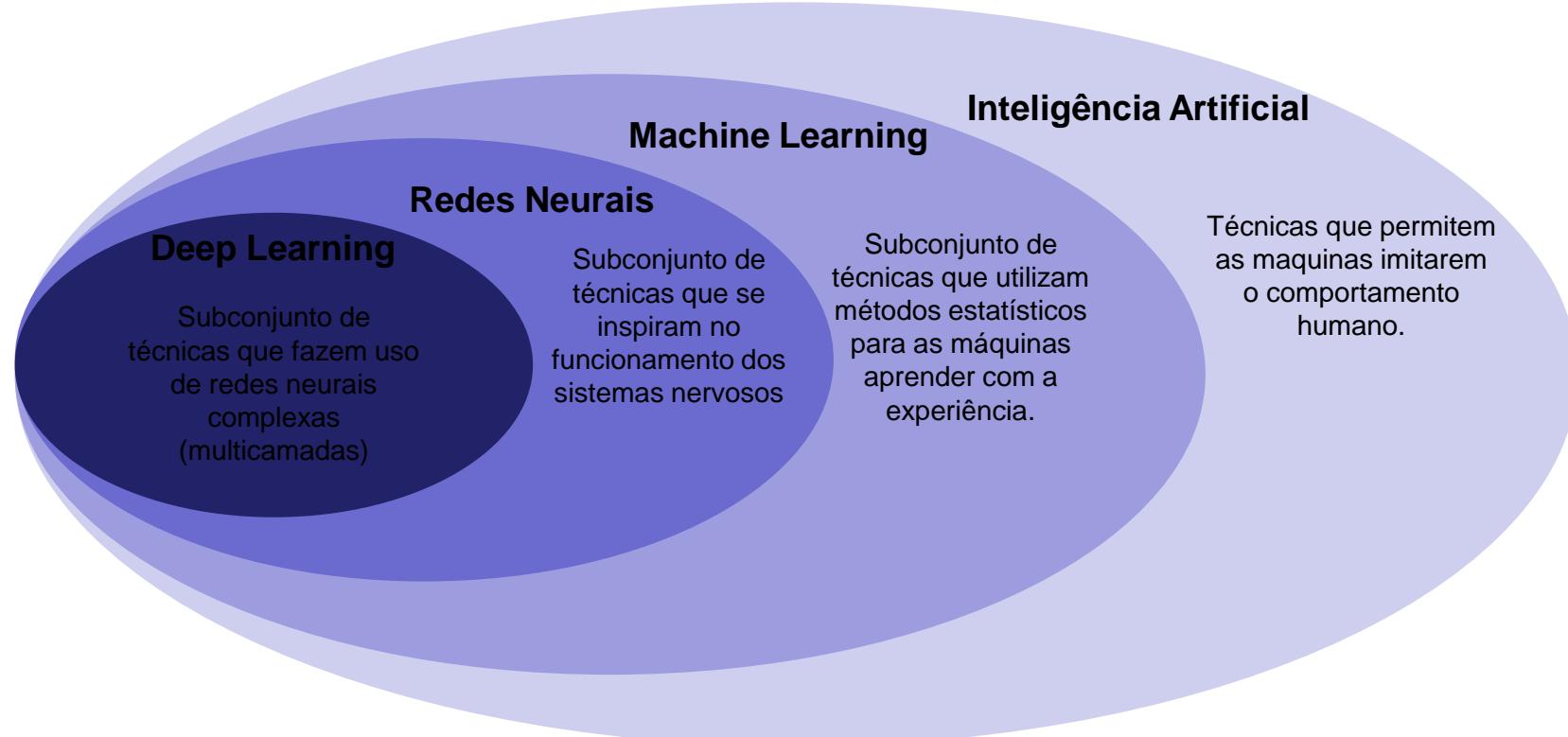
- Algoritmos para o processamento do Big Data (Parte II).



Aula 4.1.2. Algoritmos de ML para o processamento do Big Data (Parte II)

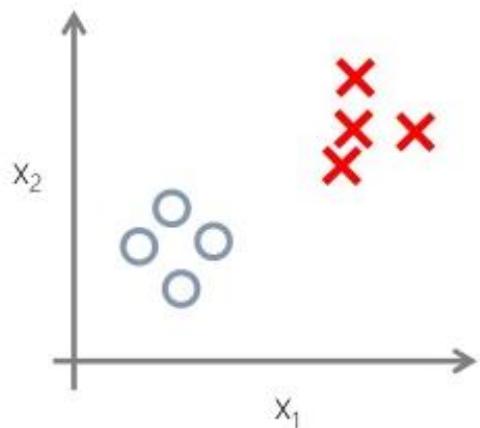
- Machine learning no processamento dos dados.
- KNN.
- Árvore de decisão.

O que é Machine Learning?

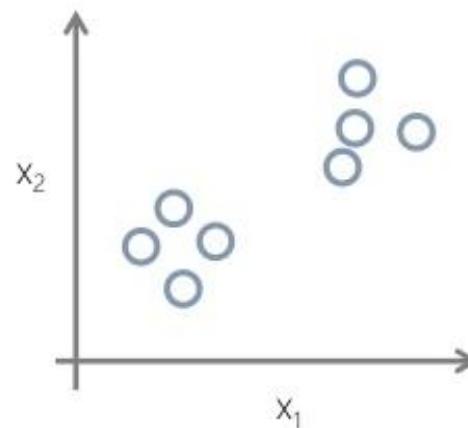


Supervisionado vs Não Supervisionado

Supervisionado



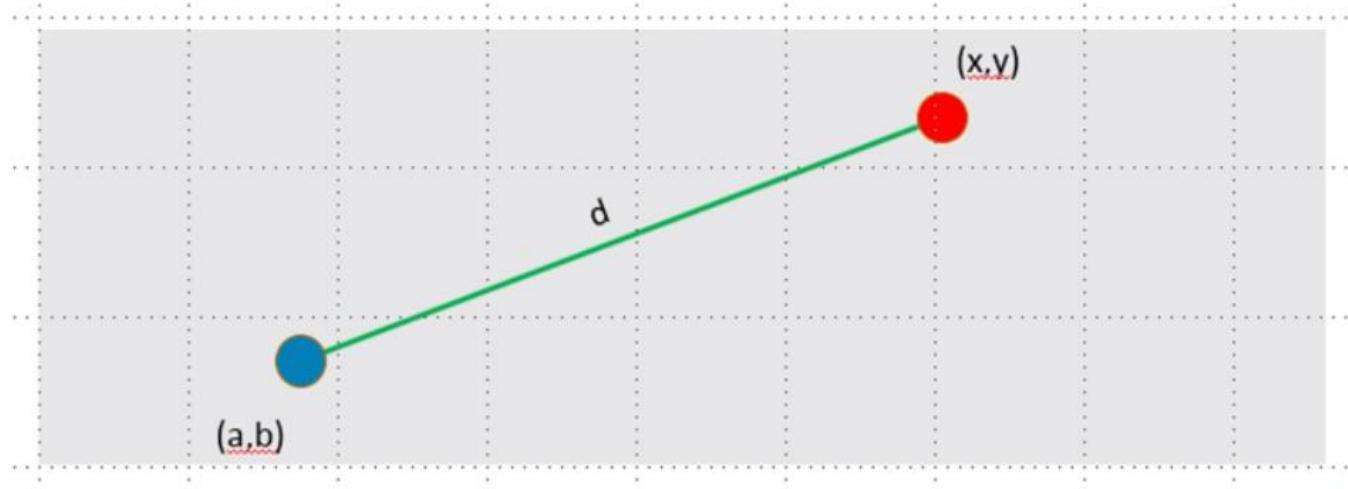
N Supervisionado



Peso	Altura	Classe
51	167	Sobrepeso
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Sobrepeso
58	169	Normal
57	173	Normal
55	170	Normal

- Distância euclidiana.

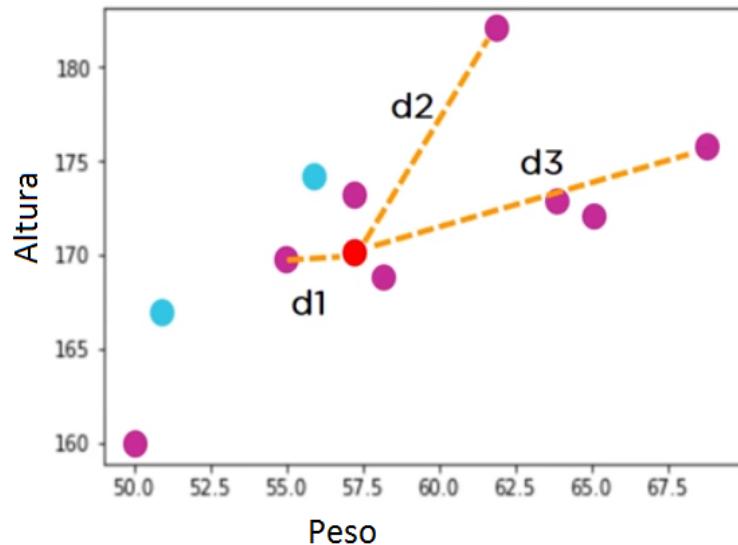
$$\text{dist}(d) = \sqrt{(x - a)^2 + (y - b)^2}$$



57

170

?



$$\text{dist}(d1) = \sqrt{(170-167)^2 + (57-51)^2} \approx 6.7$$

$$\text{dist}(d2) = \sqrt{(170-182)^2 + (57-62)^2} \approx 13$$

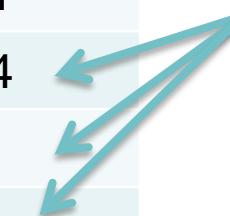
$$\text{dist}(d3) = \sqrt{(170-176)^2 + (57-69)^2} \approx 13.4$$

● Desconhecido (57,170)

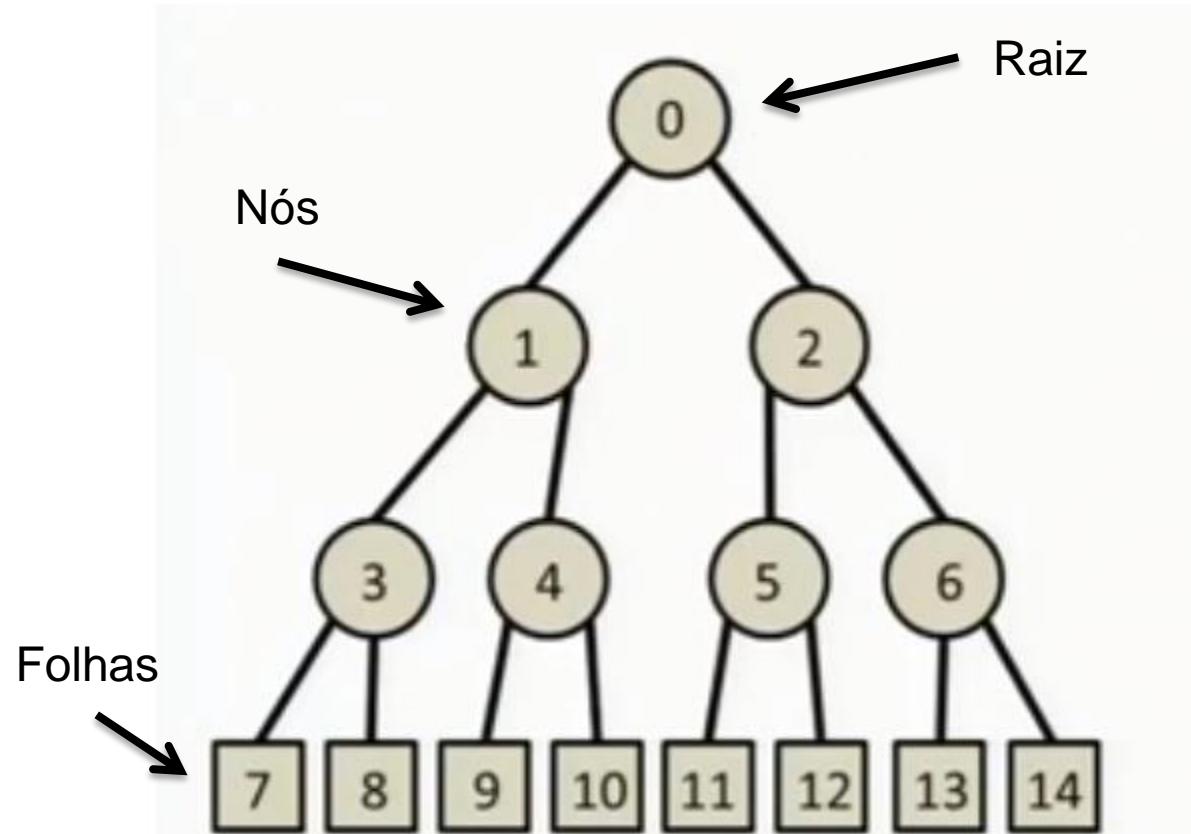
Peso	Altura	Classe	Distância
51	167	Sobrepeso	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Sobrepeso	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

Peso	Altura	Classe	Distância
51	167	Sobrepeso	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Sobrepeso	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

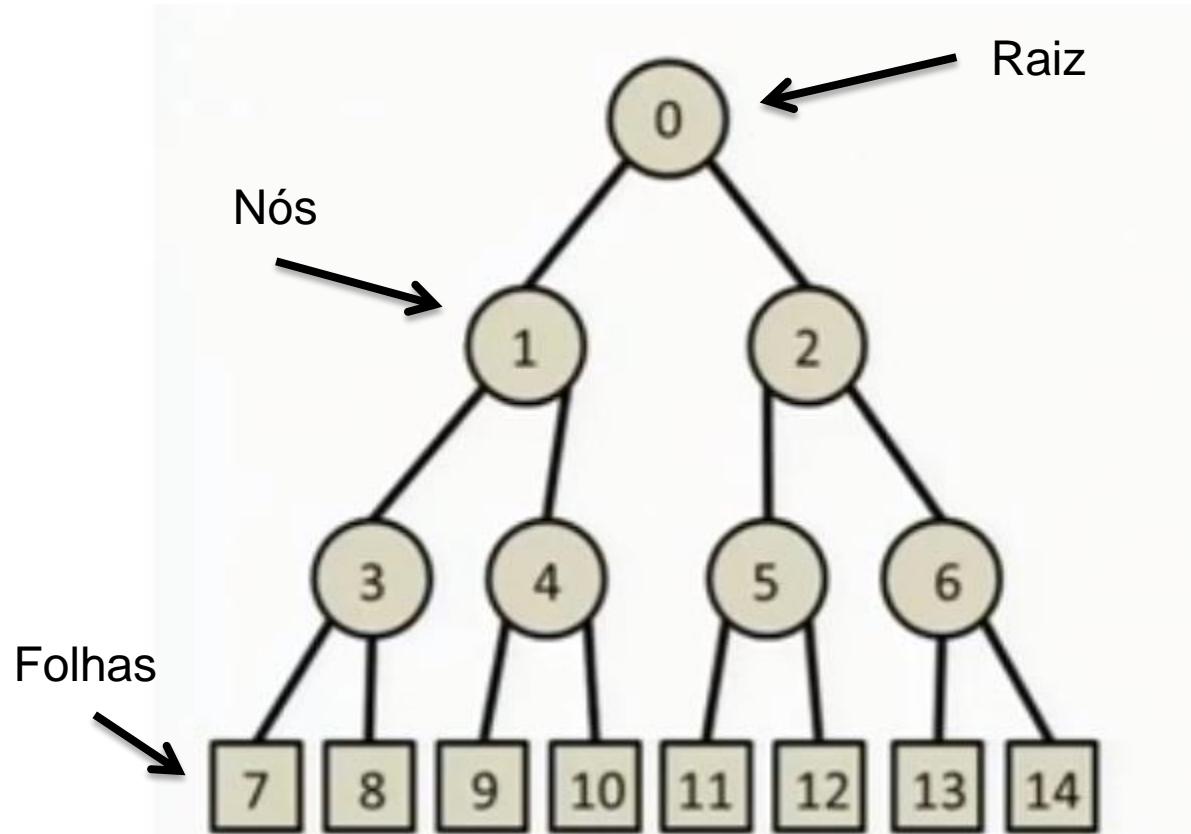
K=3



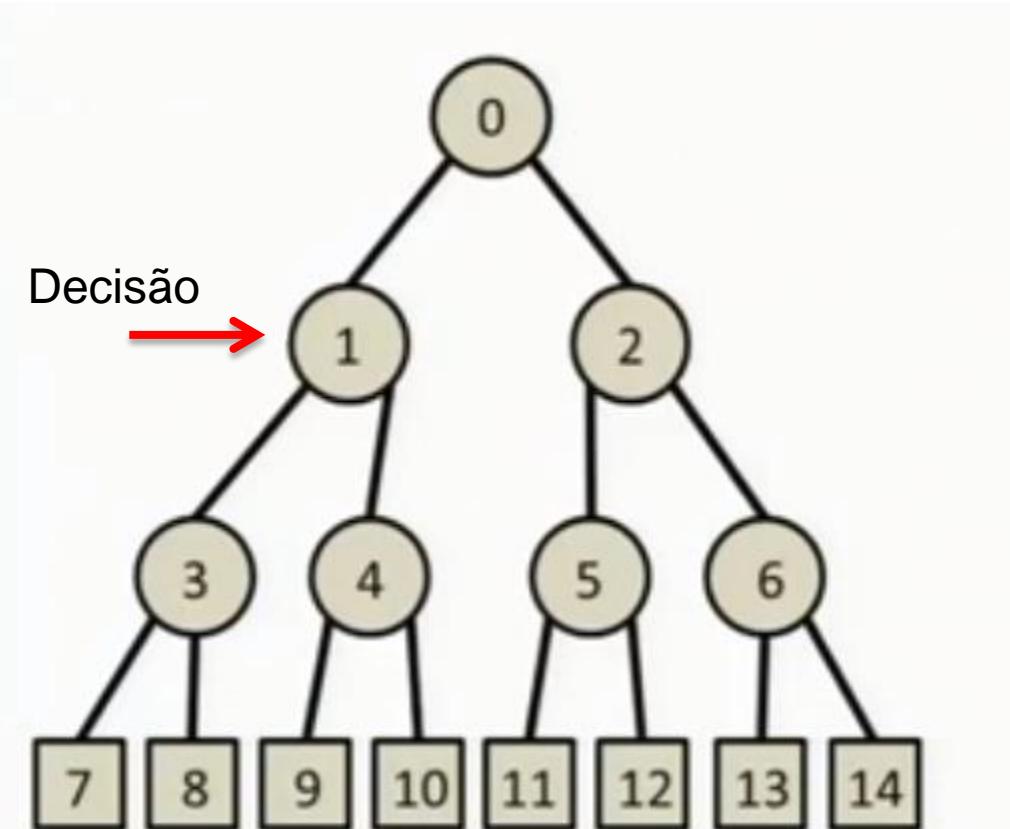
Árvore de decisão



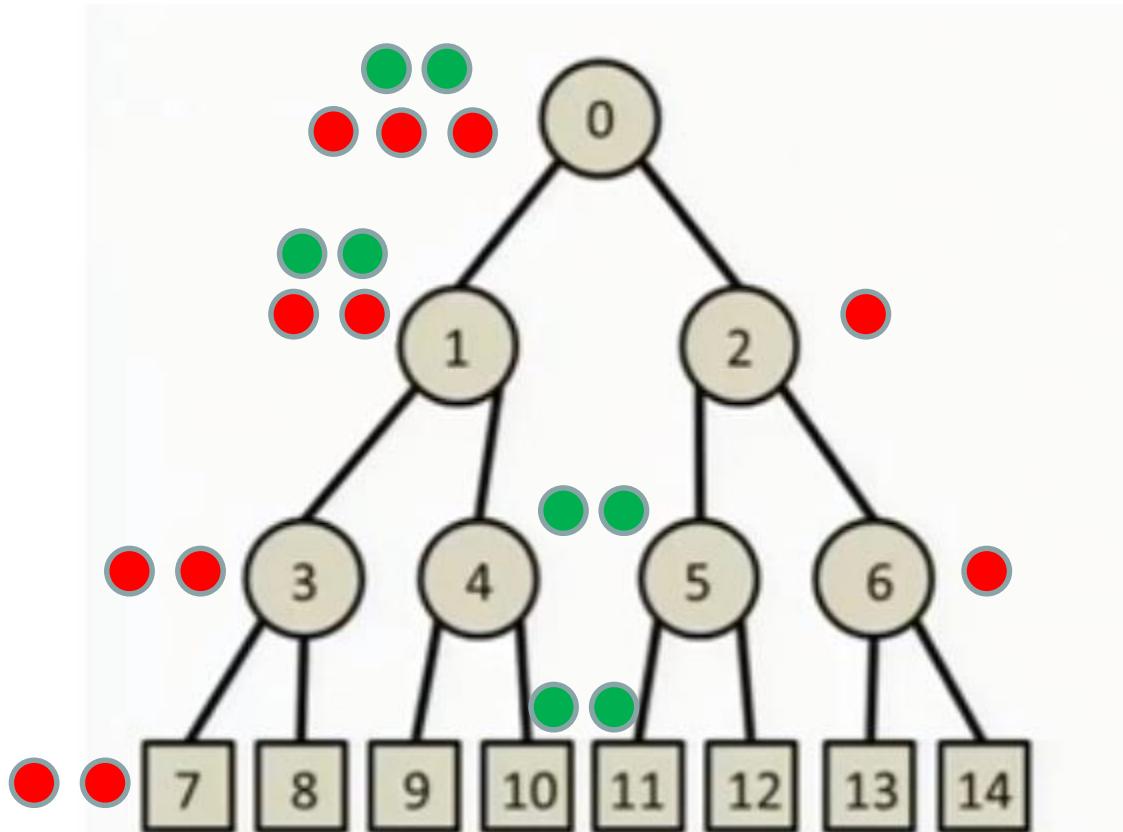
Árvore de decisão



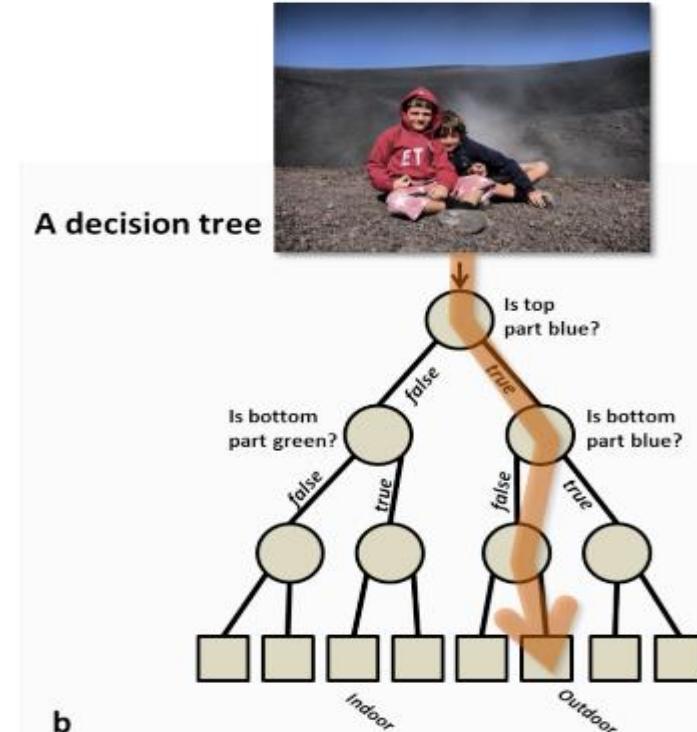
Árvore de decisão



Árvore de decisão

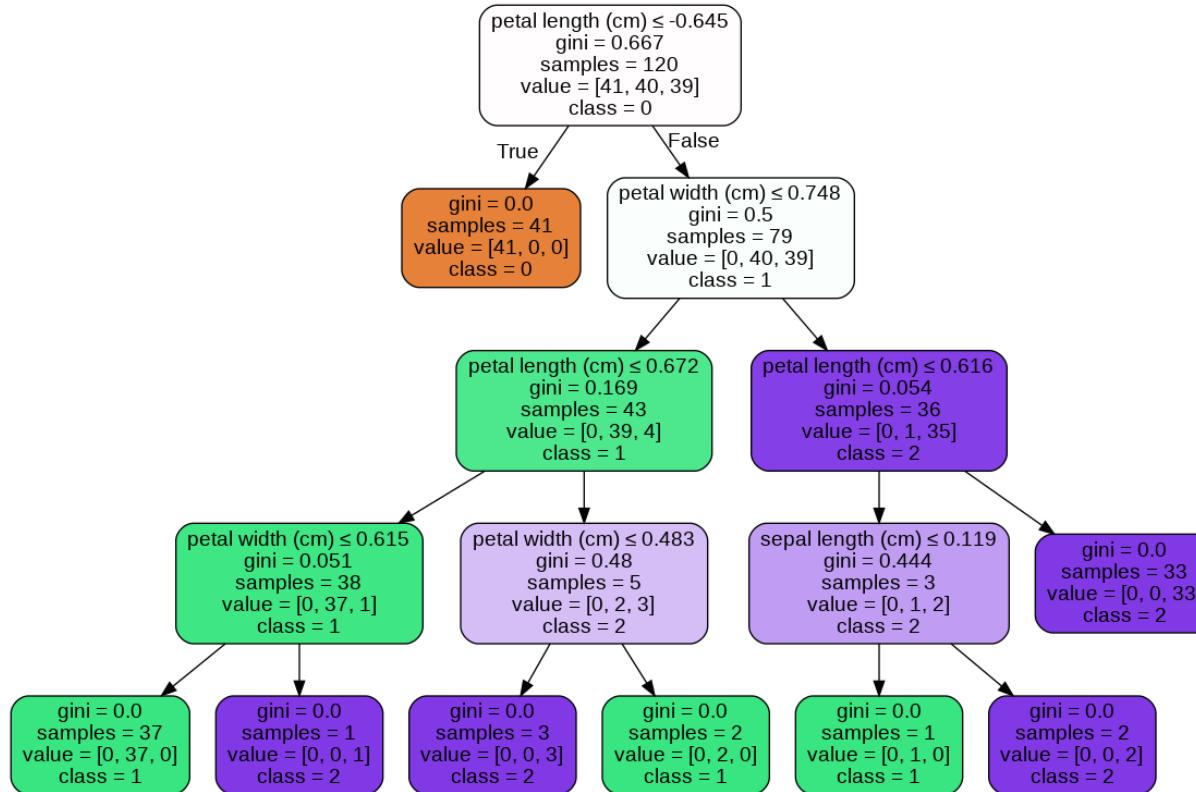


Árvore de decisão

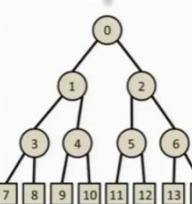
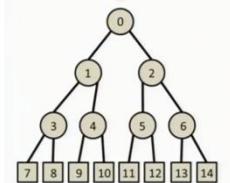
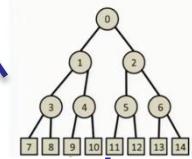
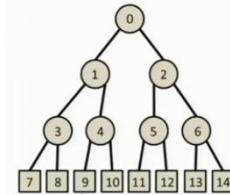
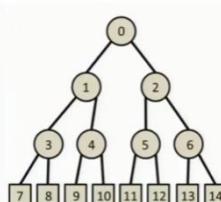
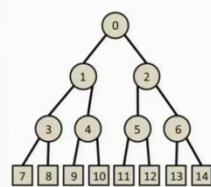
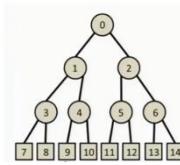
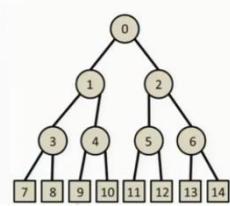


CRIMINISI, Antonio et al. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. **Foundations and Trends® in Computer Graphics and Vision**, v. 7, n. 2–3, p. 81-227, 2012.

Aplicação de árvore de decisão



Floresta randômica



Conclusão

Machine Learning.

KNN.

Árvore de decisão.

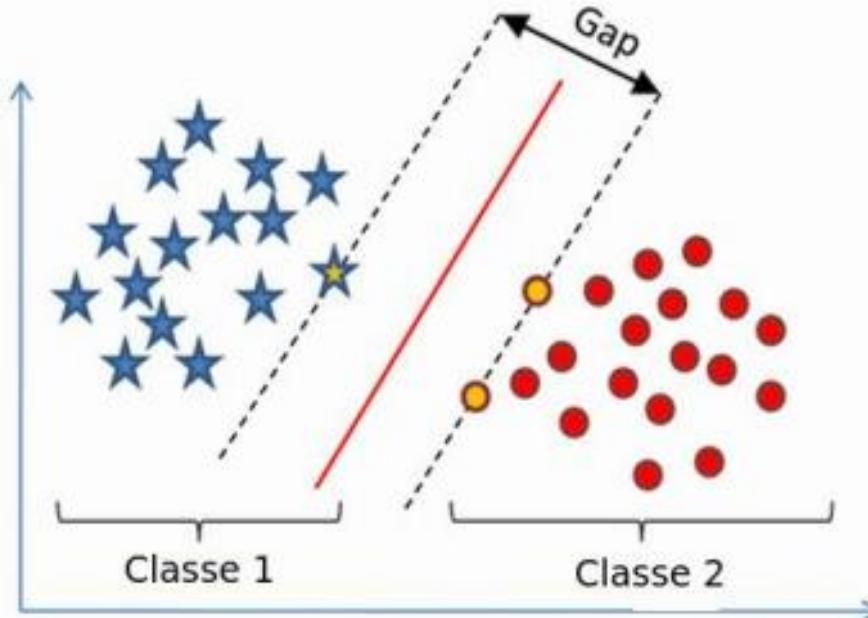
■ Próxima aula

- Algoritmos para o processamento do Big Data (Parte III).



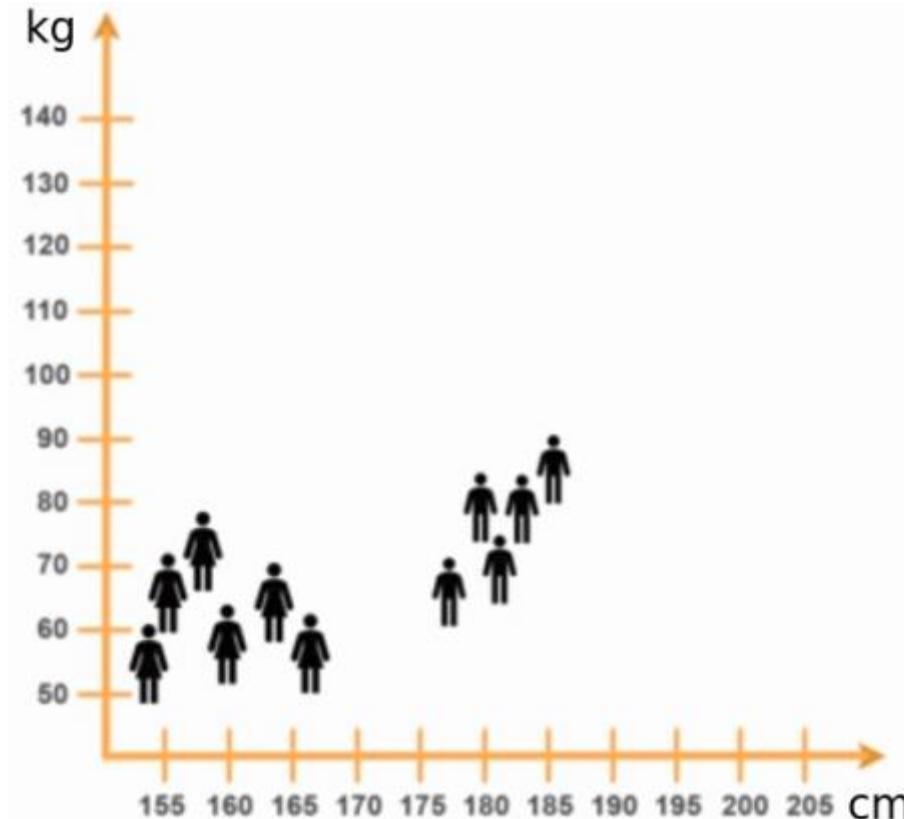
Aula 4.1.3. Algoritmos de ML para o processamento do Big Data (Parte III)

- SVM.
- Redes neurais artificiais.
- Deep learning.

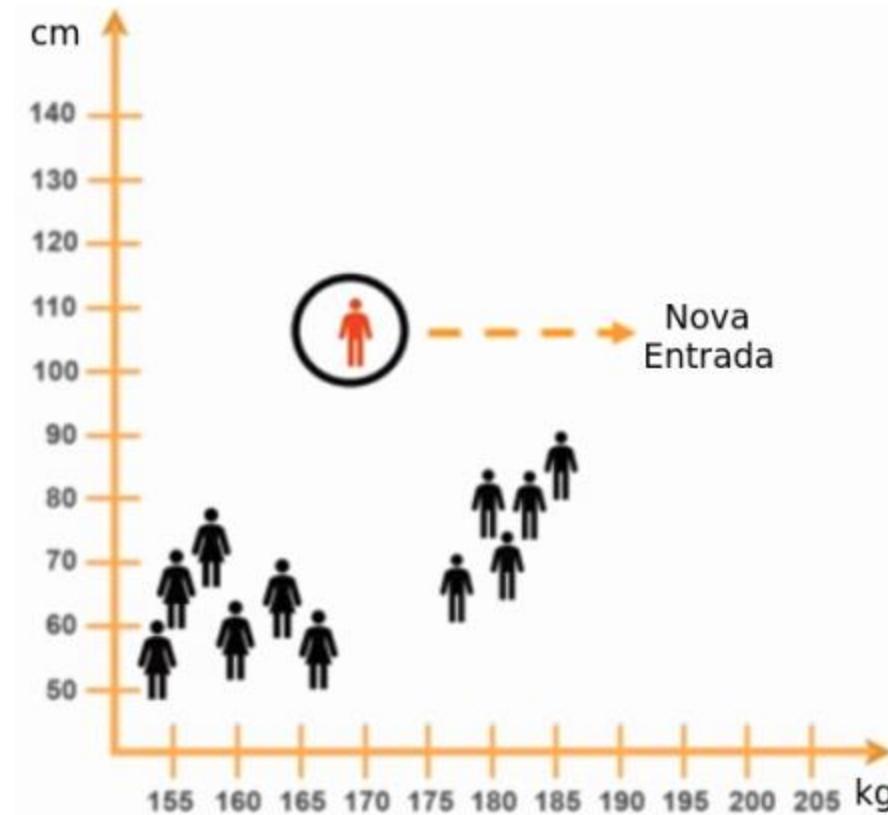


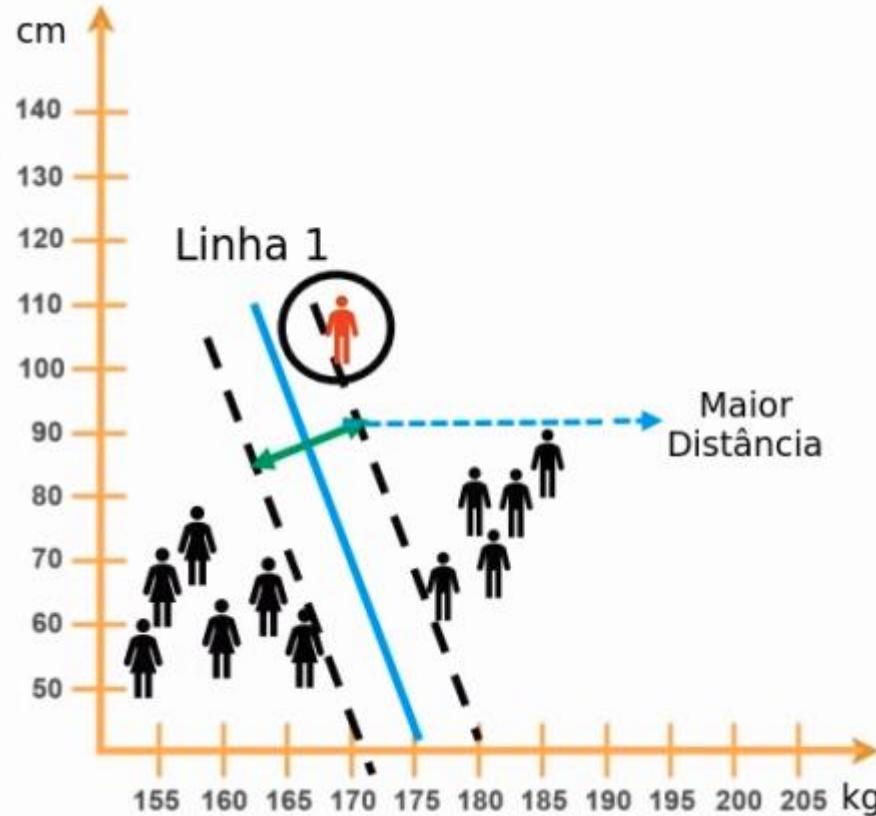
Entradas		Saída
Altura (cm)	Peso (kg)	Classificação
178	90	Homem
180	80	Homem
183	80	Homem
187	85	Homem
182	72	Homem
174	65	Mulher
174	88	Mulher
175	75	Mulher
180	65	Mulher
185	80	Mulher

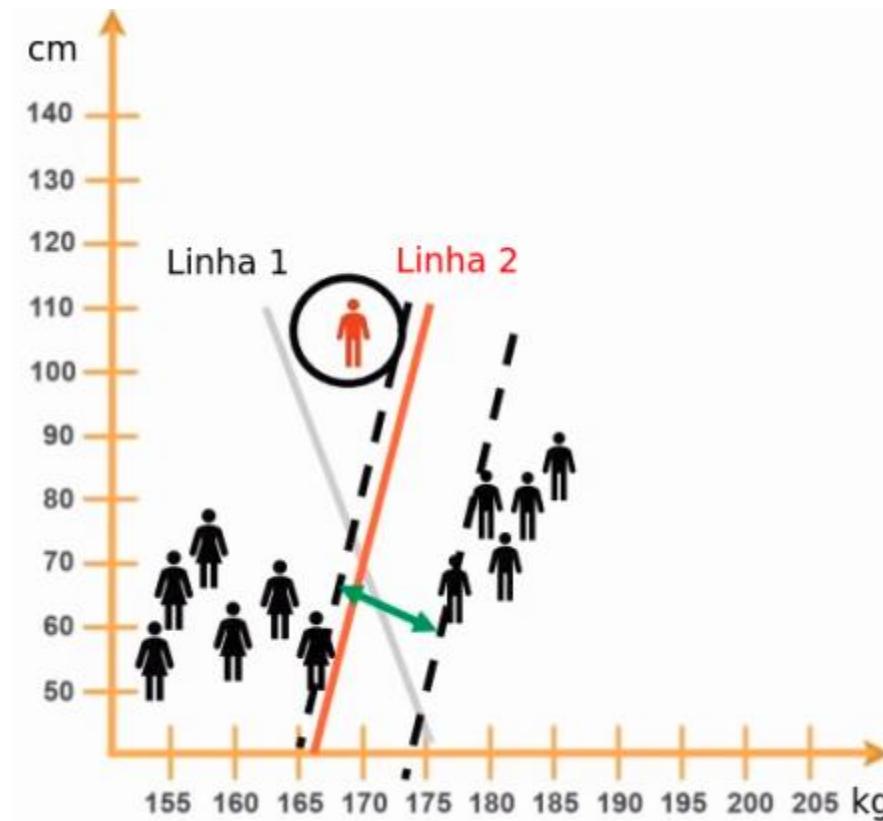
SVM

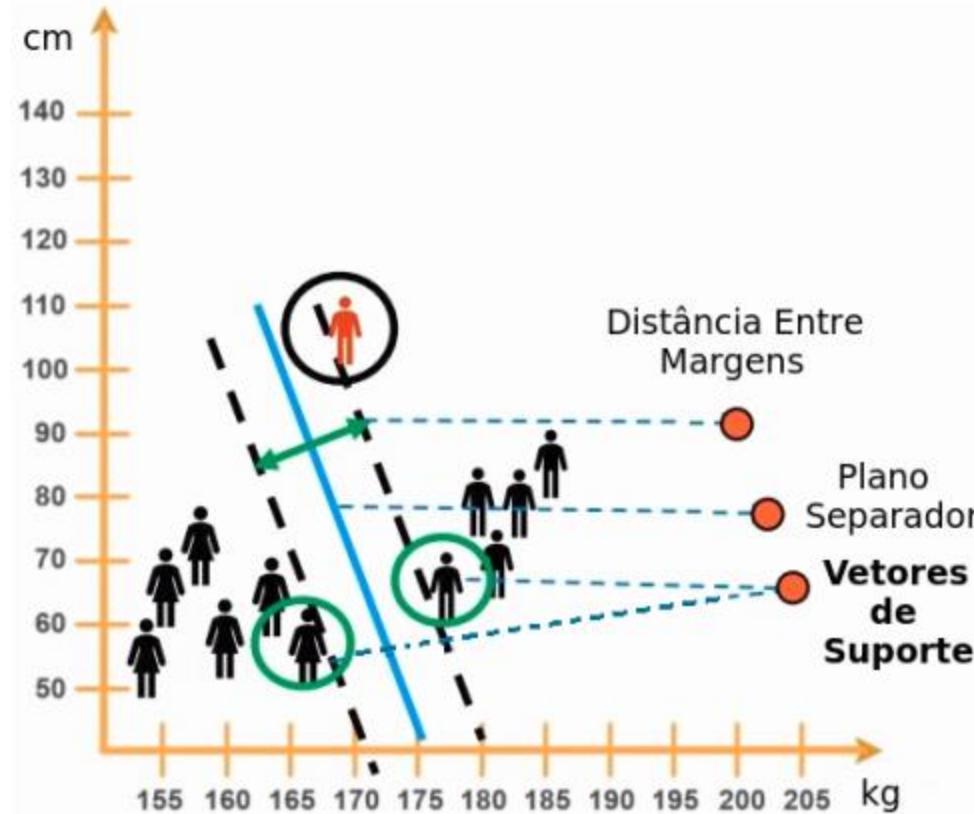


SVM

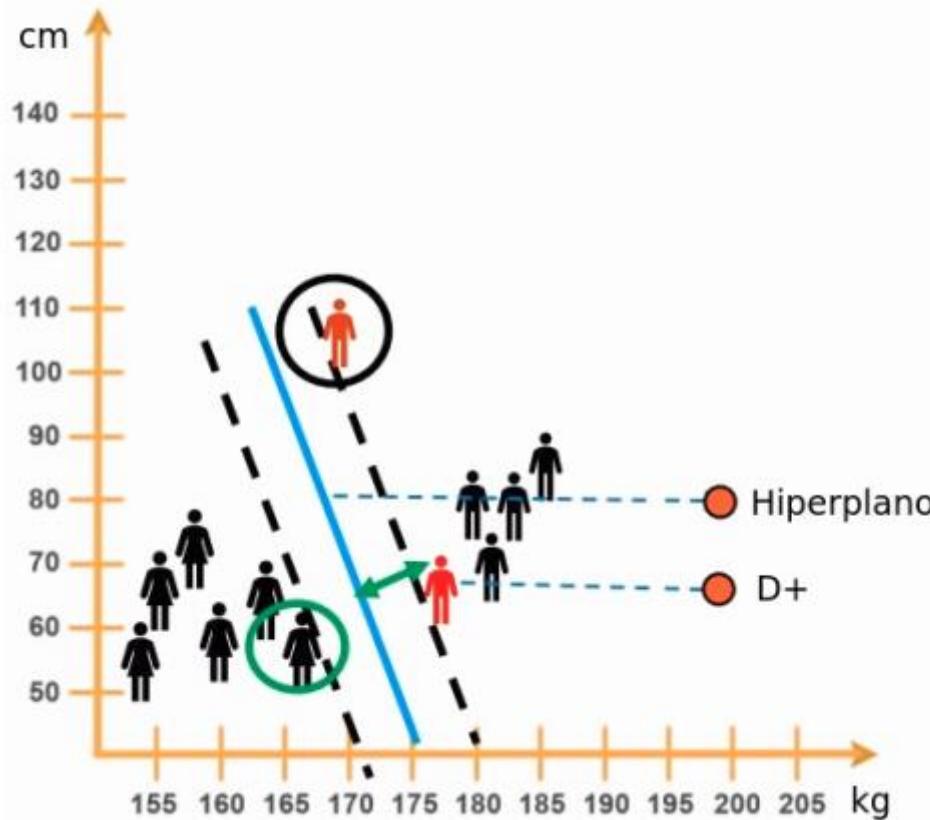




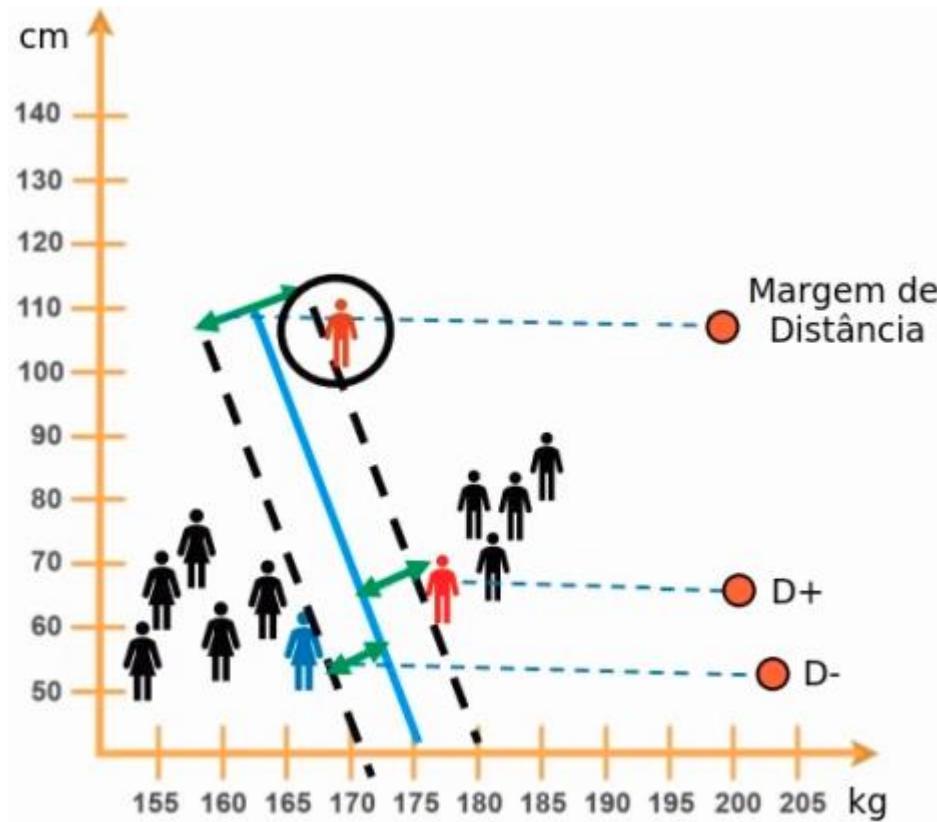


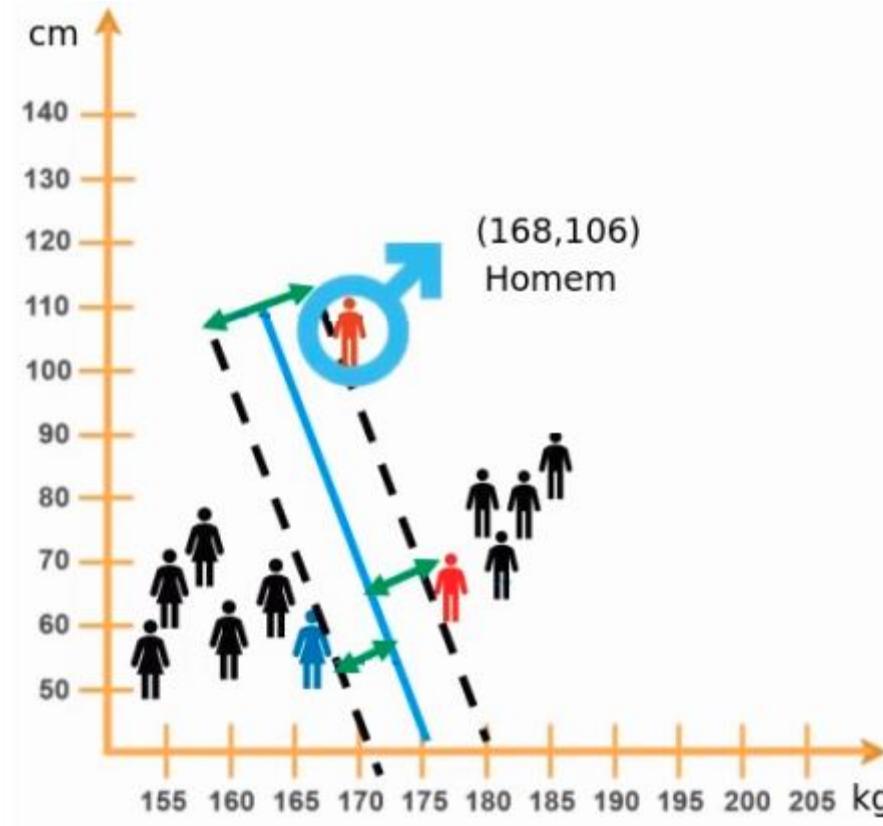


SVM



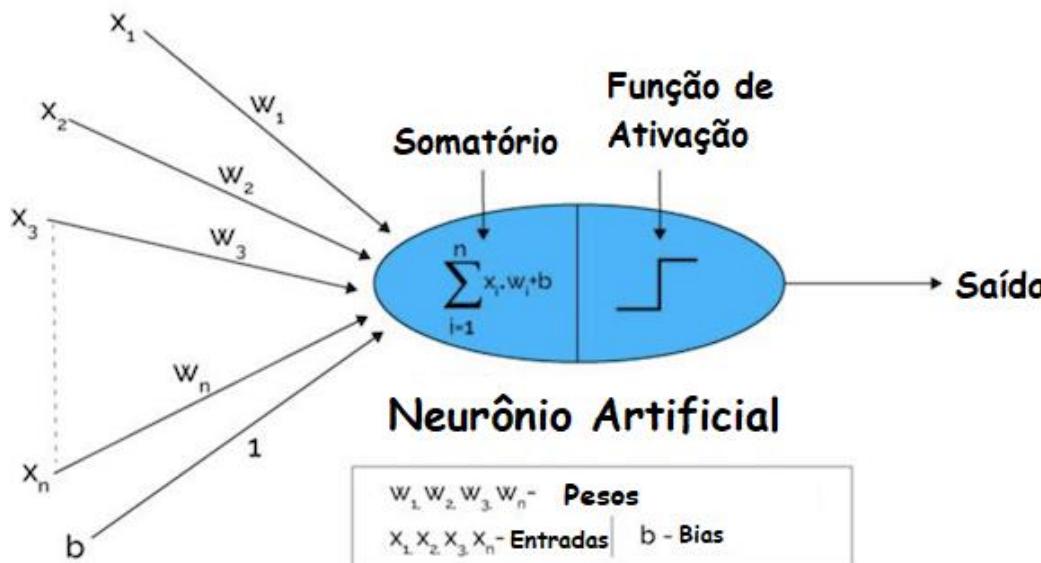
SVM



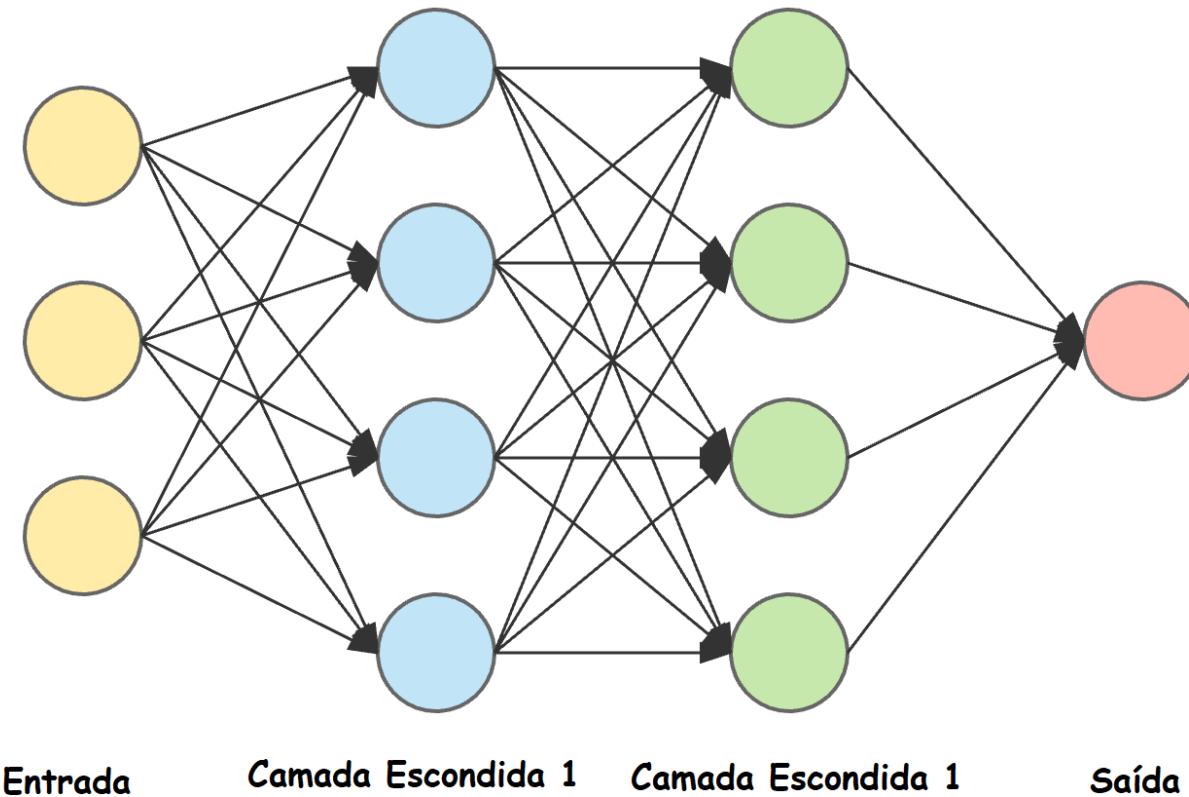


Redes neurais artificiais

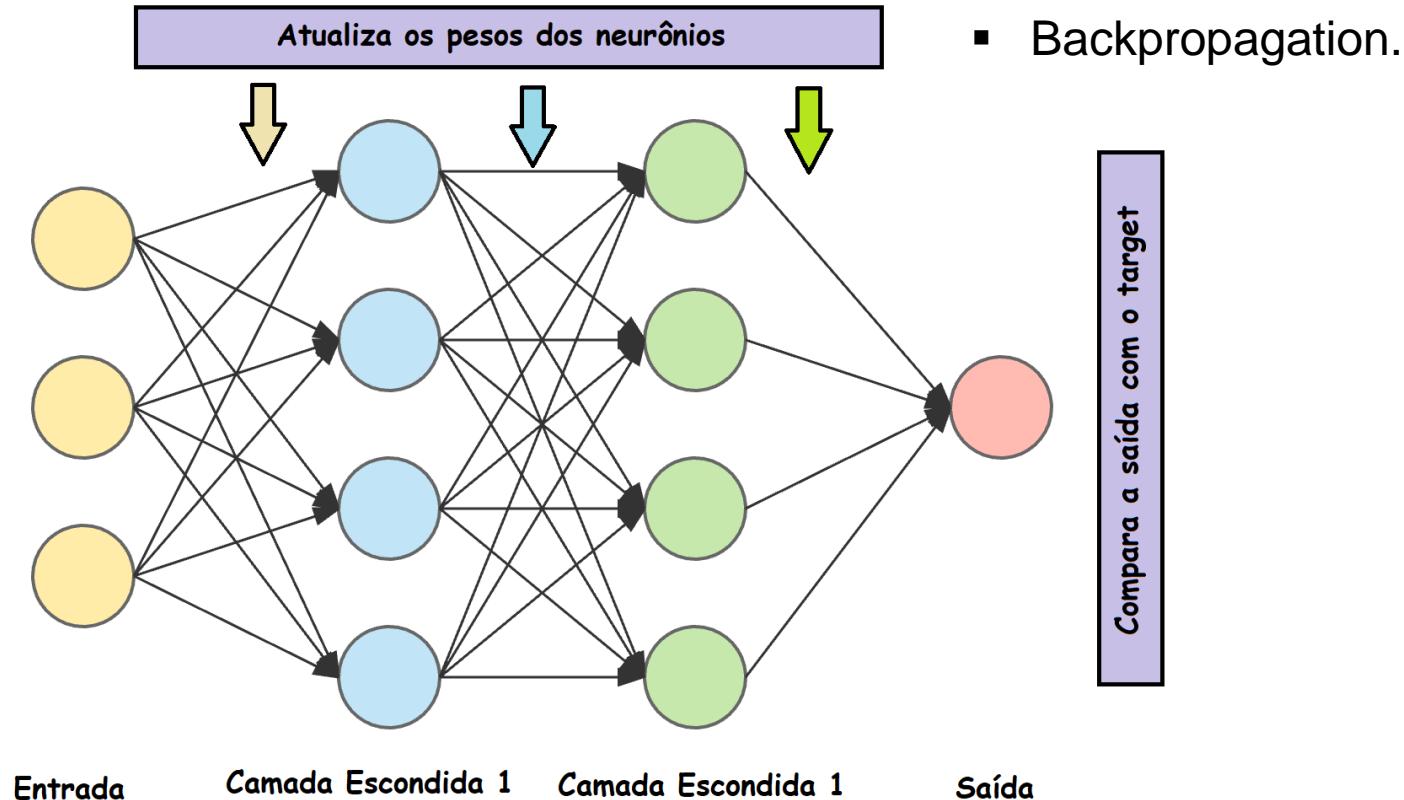
- Perceptron.



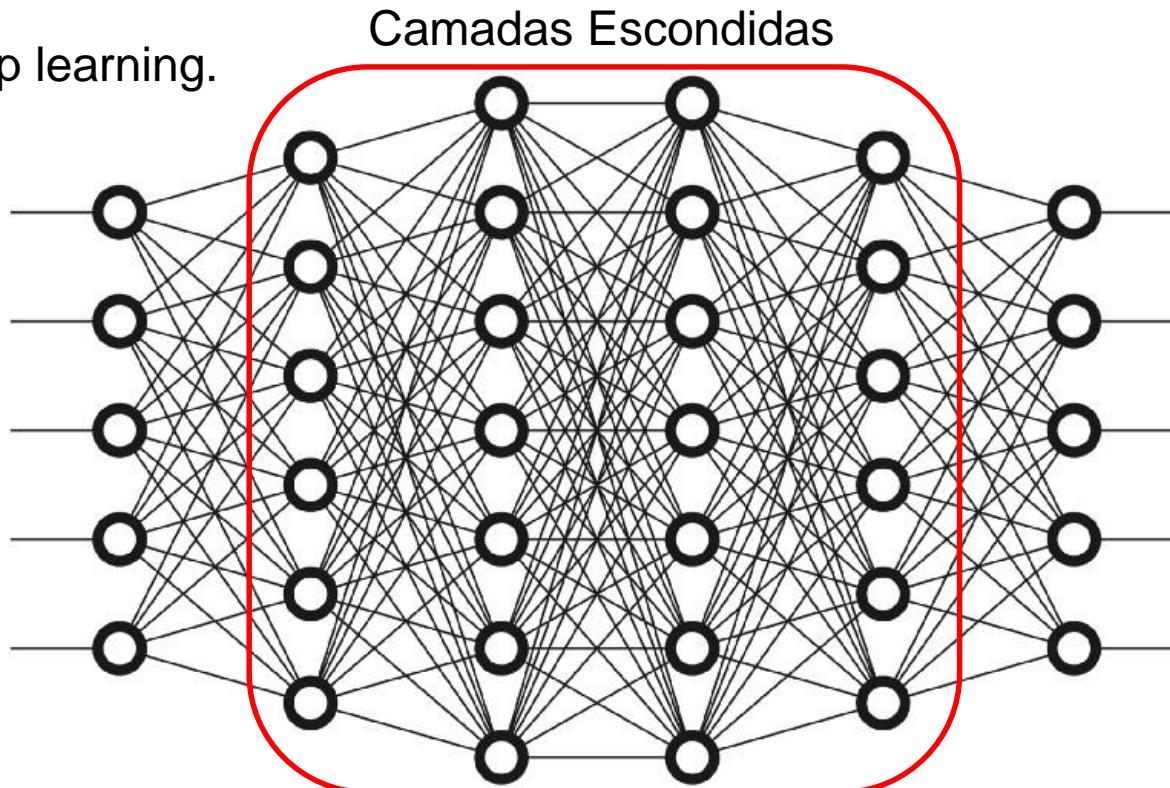
Redes neurais artificiais



Redes neurais artificiais



- Deep learning.



Tipos de redes neurais artificiais

- Perceptron Multicadas MLP;
- Redes Neurais Recorrentes (RNR);
- Redes Convolucionárias (CNN);
- Redes Generativas Adversárias (GAN).

Conclusão

- SVM.
- Redes neurais.
- Deep learning.

■ Próxima aula

- Aplicação de ML utilizando a MLlib.



Aula 4.2. MLlib para análise de regressão

Nesta aula

- Análise de regressão na previsão de valores.

the MORE
YOU PRACTICE
THE BETTER
YOU GET

Conclusão

- Aplicação utilizando análise através da regressão.

■ Próxima aula

- Aplicação de ML utilizando a MLlib.



Aula 4.3. MLlib: Árvore de decisão e floresta randômica

Nesta aula

- Aplicação de árvore de decisão e floresta randômica.



**KEEP
CALM
AND
LETS
PRACTICE**

- Aplicação utilizando MLlib para árvore de decisão e floresta randômica.

■ Próxima aula

- Aplicação de machine learning utilizando SVM.



Aula 4.4. Aplicação de machine learning utilizando SVM

Nesta aula

- Aplicação do SVM para classificação.



Conclusão

- Aplicação do SVM para classificação.

■ Próxima aula

- Sistemas de recomendação.



Técnicas para o Processamento do Big Data

Capítulo 5. Sistemas de Recomendação com MLlib

Prof. Túlio Philipe Vieira

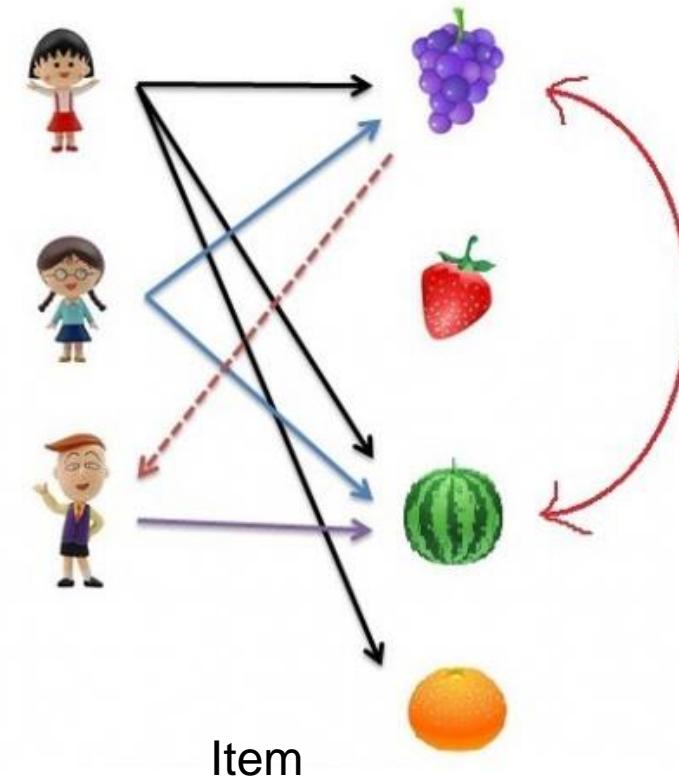
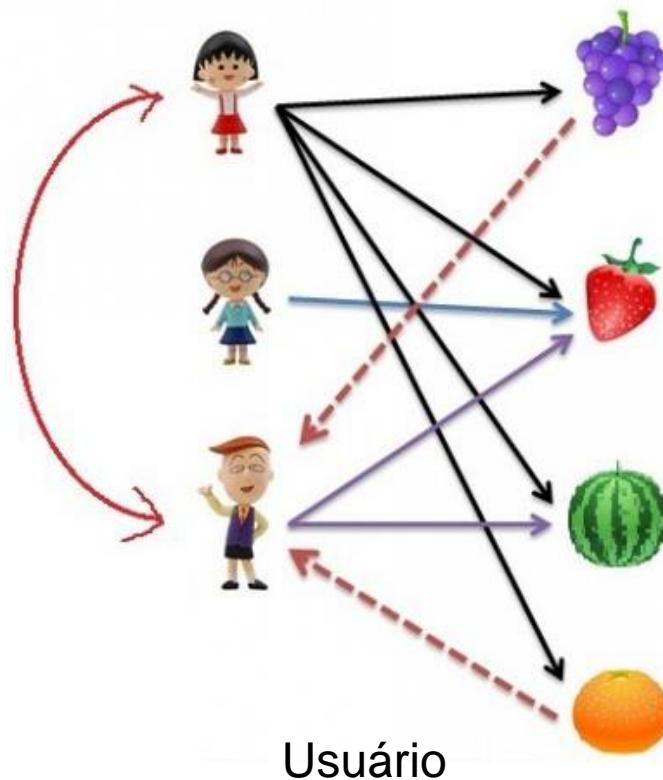


Aula 5.1. Sistemas de recomendação

Nesta aula

- O que são sistemas de recomendação?
- Estudo de caso da Spotify.

Sistemas de recomendação



Sistemas de recomendação



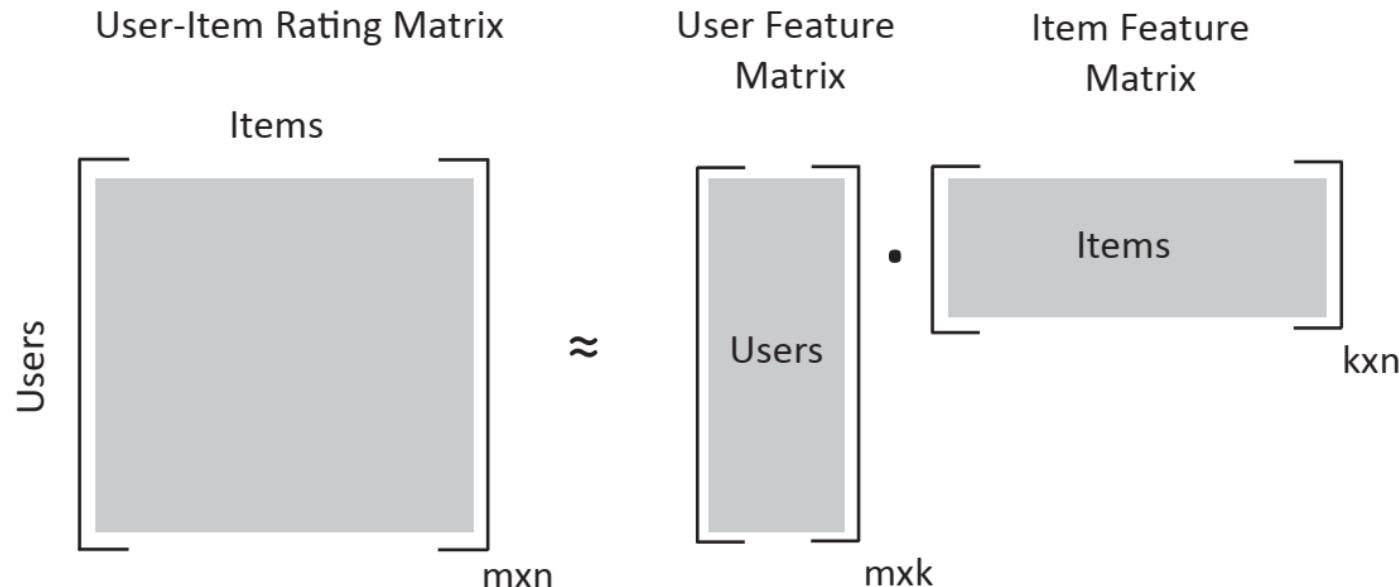
Richard
Mary
Steve

1
-1
-1

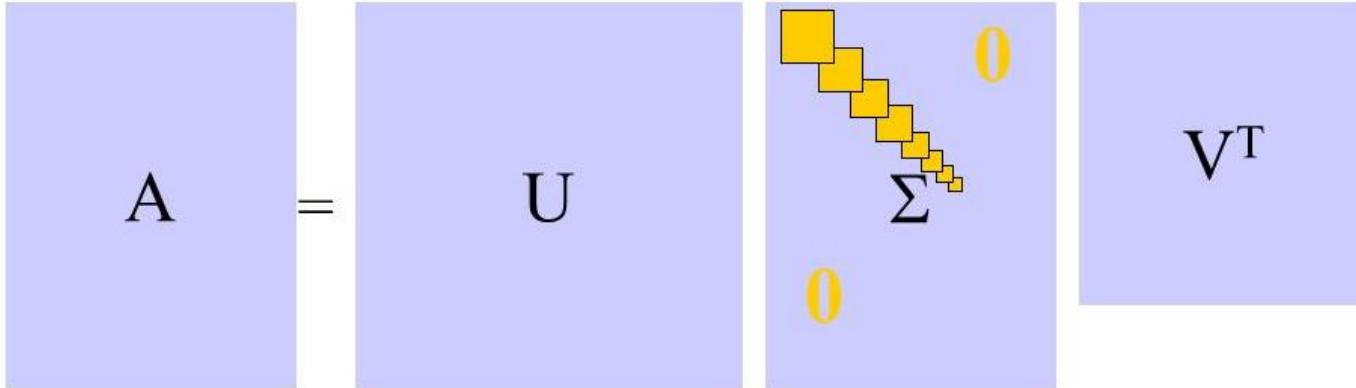
5	1	4
?	2	5
2	3	2



Alternating Least Squares (ALS)



Singular Value Decomposition (SVD)



$m \times n$

$m \times m$

$m \times n$

$n \times n$

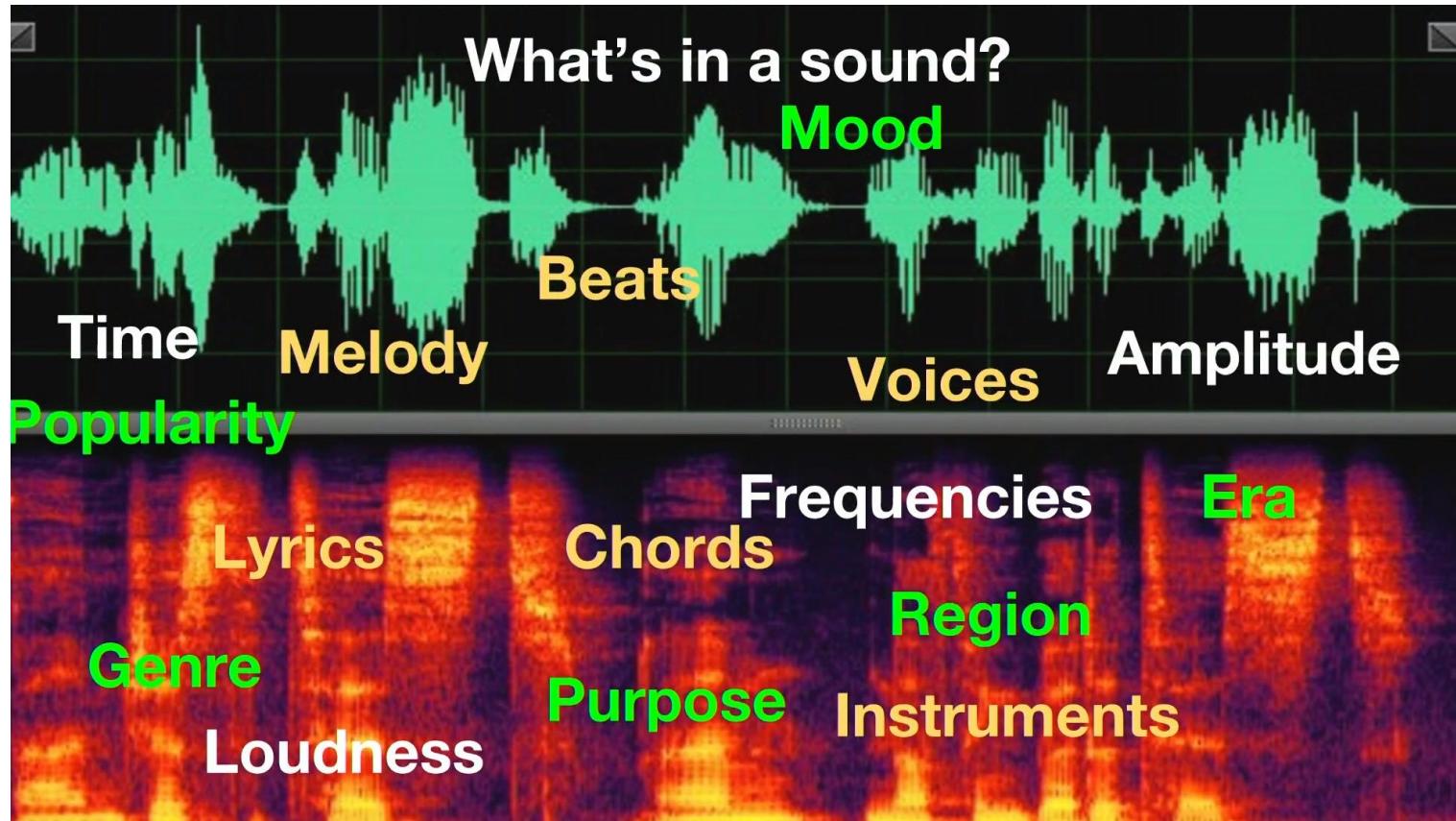
Original Matrix	Eigenvectors Matrix	Eigenvalues Matrix	Inverse of Eigenvectors Matrix
$\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$	$\begin{bmatrix} -1 & -1 \\ 2 & 1 \end{bmatrix}$	$\begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 \\ -2 & -1 \end{bmatrix}$

Estudo de caso: Spotify

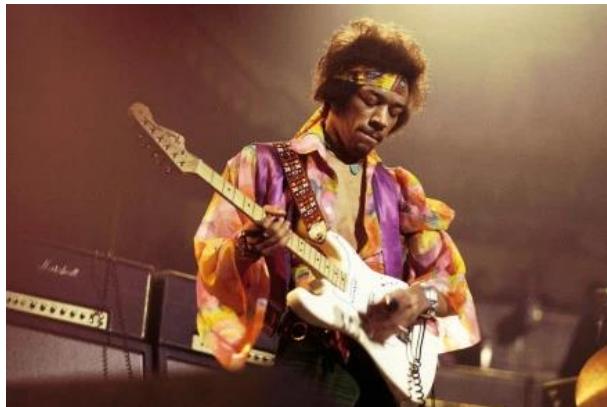
- 35 milhões de músicas;
- 217 milhões de usuários ativos ao mês;
- 100 milhões de usuários *Premium*;
- Valor de mercado: 27 bilhões de dólares.



O que é a música?



O que recomendar?



Conteúdo



Usuário



Como o que recomenda?

IGTI

The image shows four smartphones side-by-side, each displaying a different feature of a music streaming app:

- Radio:** Shows Ed Sheeran's profile with a "PLAY RADIO" button. Below it, a "In This Station" section lists artists like James Bay, One Direction, Ed Sheeran, James Arthur, and The Fray. At the bottom, a navigation bar includes Home, Browse, Search, Radio (highlighted), and Your Library.
- Daily Mix:** Shows a grid of "YOUR DAILY MIXES" including "Your Daily Mix 1" (Nicki Minaj, Pharrell Williams, Lil Wayne and others), "Daily Mix 2" (Earth, Wind & Fire, Al Green, Stevie Wonder and others), "Your Daily Mix 3" (Sorry by Drake), and "Your Daily Mix 4" (Maps by Imagine Dragons). Below the grid, a "Daily Mix 5" section lists Bob Marley & The Wailers, Madlyn Bailey, Colin & Caroline, Cameron Bedell, and Devin Penn, Ziggy Marley.
- This Is:** Shows Lianne La Havas's profile with a "SHUFFLE PLAY" button. Below it, a "Download" section lists songs like "Unstoppable" and "Green & Gold". A "What You Don't Do" section follows, and at the bottom is a "Tokyo" section.
- Recommended Songs:** Shows a "Summer Hits" playlist by Maroon 5 with a "PLAY" button. Below it, a "Recommended Songs" section based on the playlist lists "Talking Body" by Tove Lo and "Queen Of The Clouds". Further down, a "Photograph" section by Ed Sheeran (X Deluxe Edition) is shown, followed by "Want To Want Me" by Jason Derulo (Everything Is 4), "Black Magic" by Little Mix (Black Magic), and "The Summer" by Maroon 5 (V).

Como sugerem as músicas?

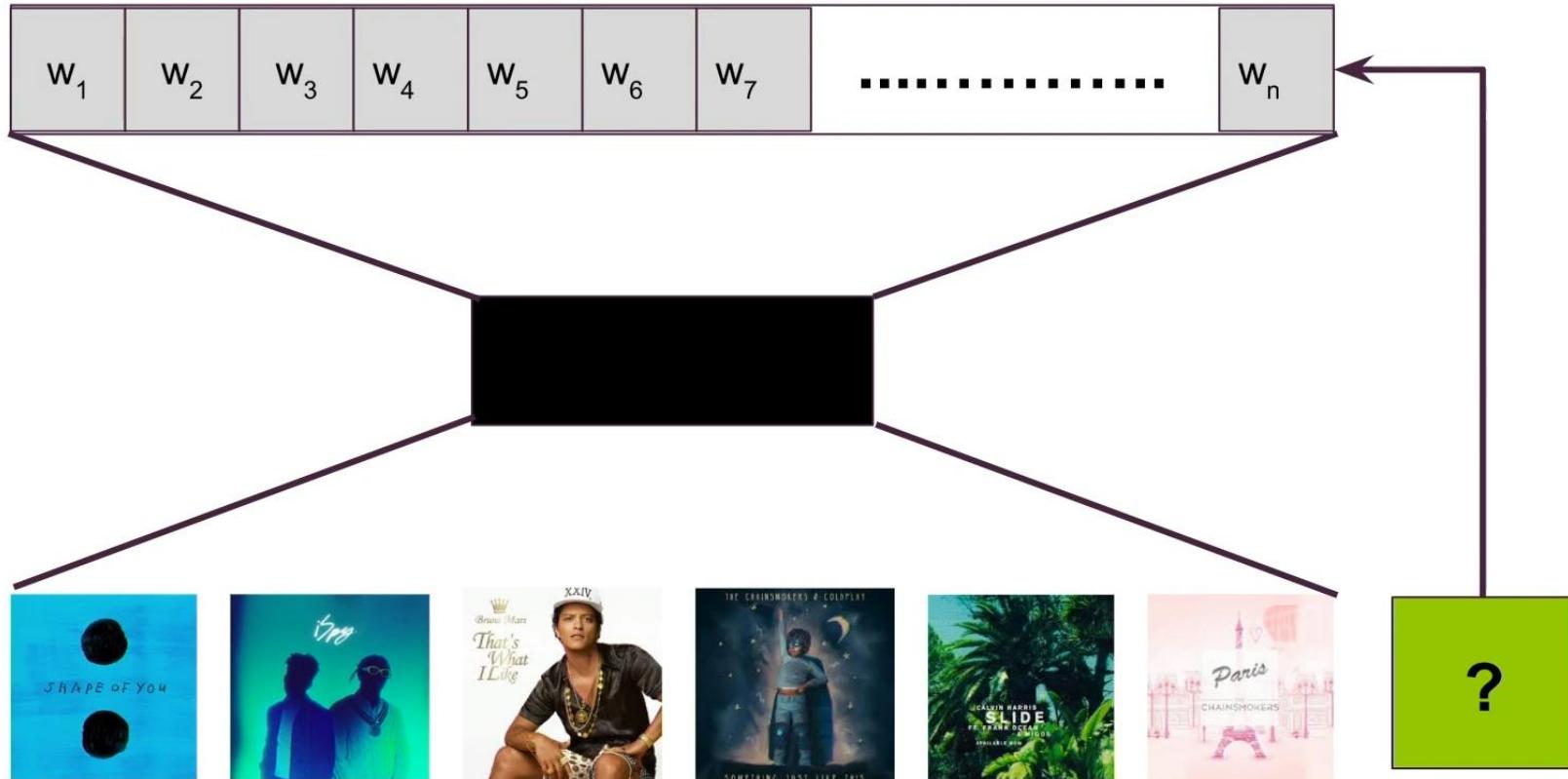
Document : Playlist →

The image shows a Spotify playlist page for 'Soft Pop Hits'. At the top, there's a thumbnail of a woman wearing headphones, the title 'Soft Pop Hits', a description 'Easy listening Pop from all your favorite artists!', and information 'Created by: Spotify • 246 songs, 15 hr 55 min'. Below this are buttons for 'PLAY', 'FOLLOW', and a three-dot menu. On the right, it shows 'FOLLOWERS 552,889'. A table below lists 10 songs with columns for 'SONG', 'ARTIST', 'ALBUM', and 'TIME'. The songs listed are: Take A Bow (Rihanna), Please Don't Go (Joel Adams), Up&Up (Coldplay), Peter Pan (Kelsea Ballerini), Send My Love (To Your New Lover) (Adele), Like I'm Gonna Lose You (Meghan Trainor, John Legend), Let It Go (James Bay), How to Save a Life (The Fray), Make It To Me (Sam Smith), and Remedy (Adele).

SONG	ARTIST	ALBUM	TIME
Take A Bow	Rihanna	Good Girl Gone Bad: Reloaded	3:49
Please Don't Go	Joel Adams	Please Don't Go	3:31
Up&Up	Coldplay	A Head Full Of Dreams	6:45
Peter Pan	Kelsea Ballerini	The First Time	3:20
Send My Love (To Your New Lover)	Adele	Send My Love (To Your New Lover)	3:43
Like I'm Gonna Lose You	Meghan Trainor, John Legend	Title (Deluxe)	3:45
Let It Go	James Bay	Chaos And The Calm	4:21
How to Save a Life	The Fray	How To Save A Life	4:23
Make It To Me	Sam Smith	In The Lonely Hour (Deluxe)	2:43
Remedy	Adele	25	4:05

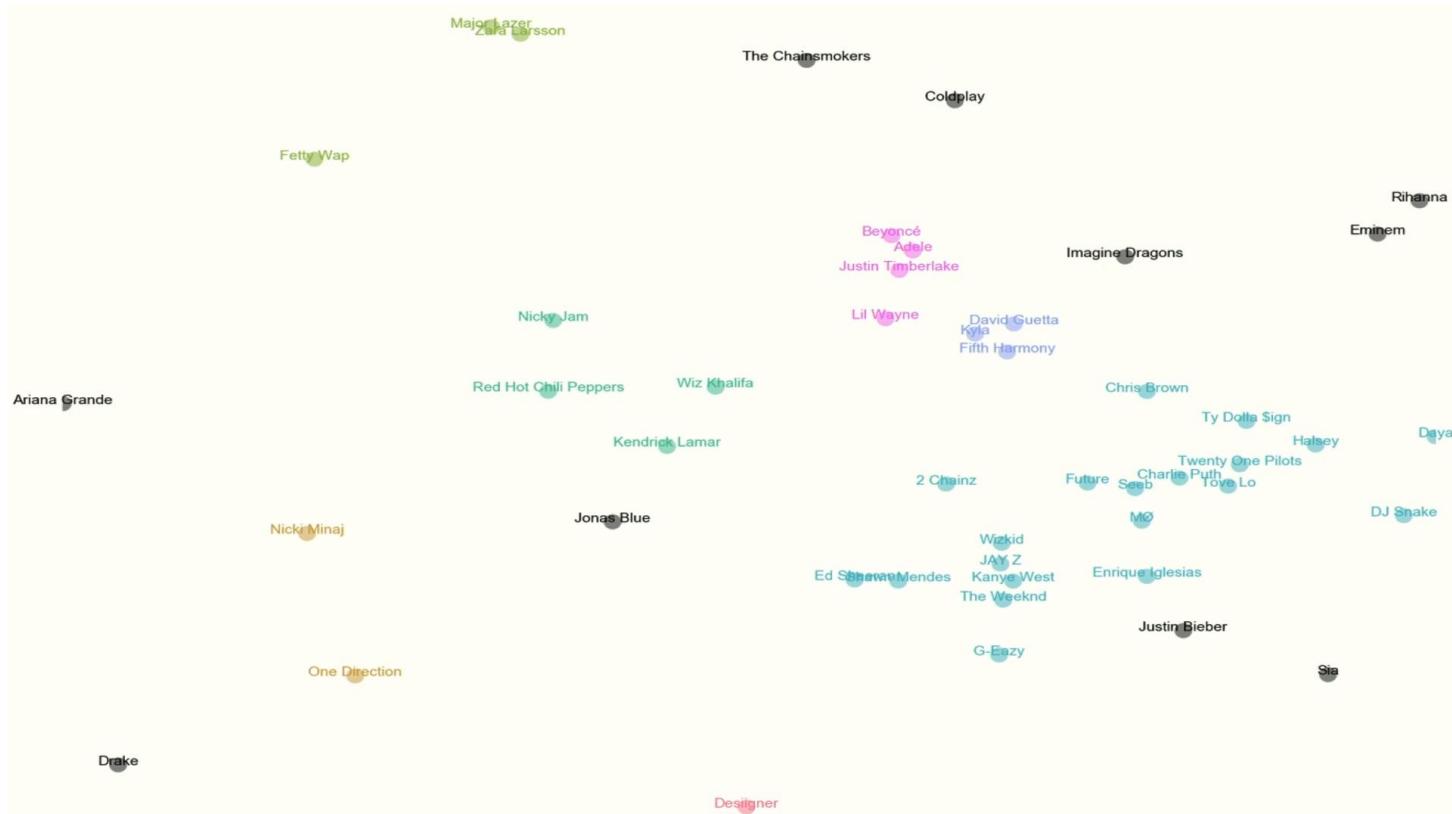
Word : Song →

Como sugerem as músicas?



Como sugerem as músicas?

IGTI



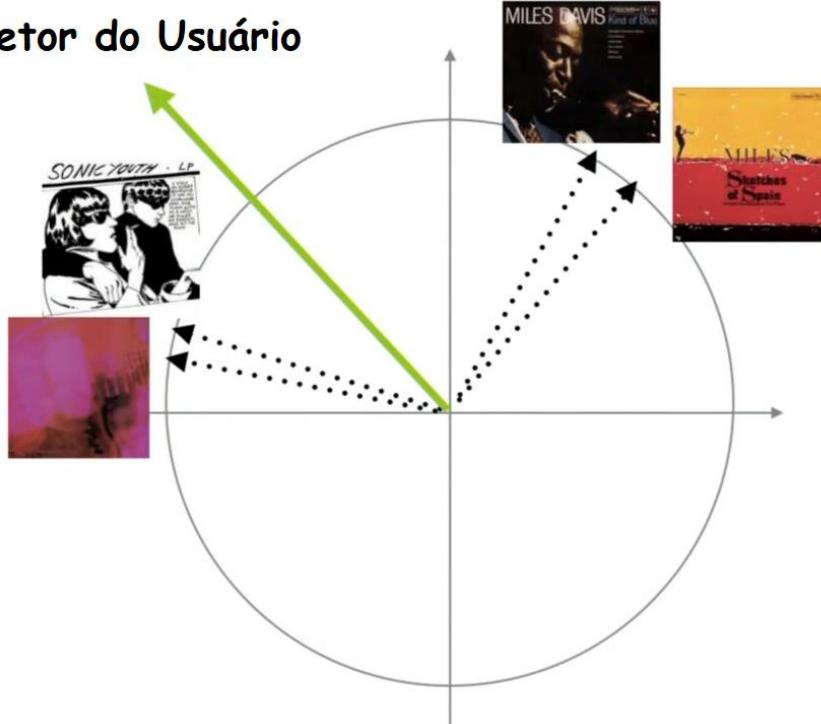
Como sugerem as músicas?

IGTI



Como sugerem as músicas?

Vetor do Usuário



- Similaridade;
- Preferências do usuário;
- Interações de usuários;
- Listas que mais ouve.

Conclusão

- O que são sistemas de recomendação.
- Estudo de caso da Spotify.

■ Próxima aula

- Aplicação de sistemas de recomendação.



Aula 5.2. Aplicação de sistemas de recomendação

Nesta aula

- Aplicação utilizando sistemas de recomendação.



Let's
Practice!

Conclusão

- ✓ Aplicação utilizando MLlib para sistemas de recomendação.

■ Próxima aula

- Teoria dos grafos.



Técnicas para o Processamento do Big Data

Capítulo 6. Teoria dos Grafos

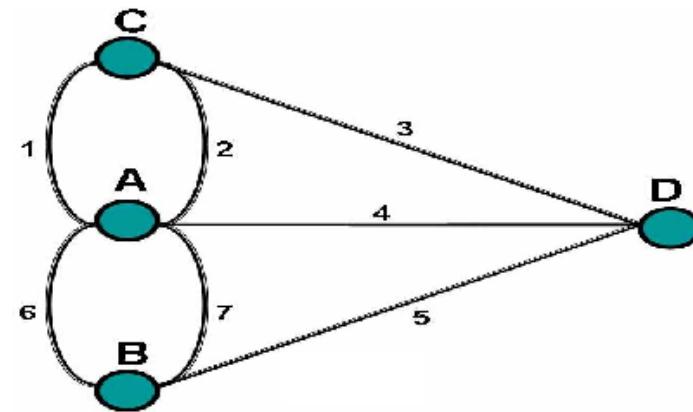
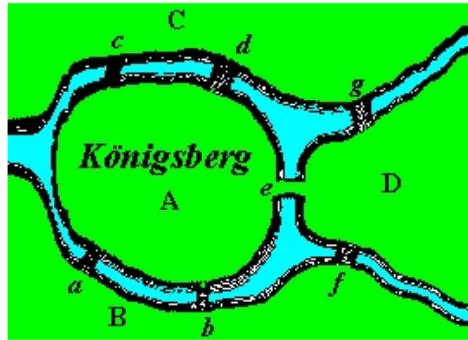
Prof. Túlio Philipe Vieira



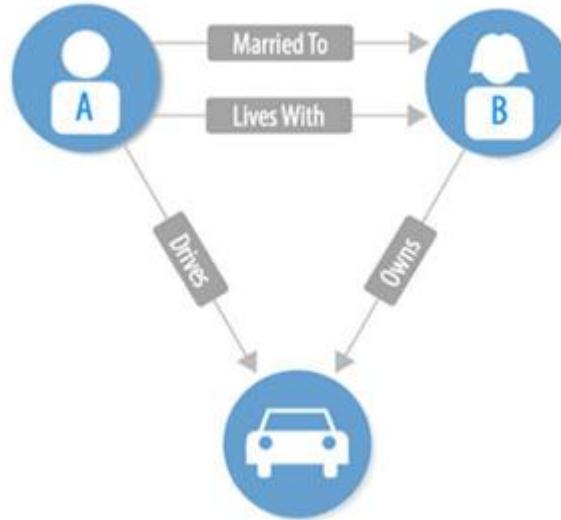
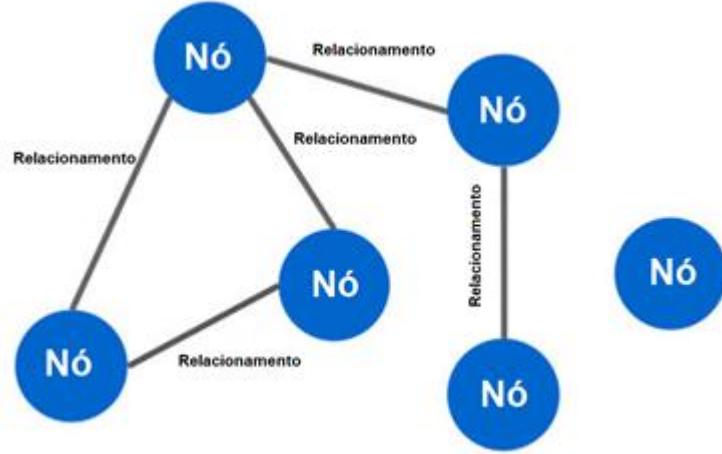
Aula 6.1. Introdução

- O que são grafos?
- Por que utilizar grafos?
- Onde encontramos grafos?

O que são grafos?

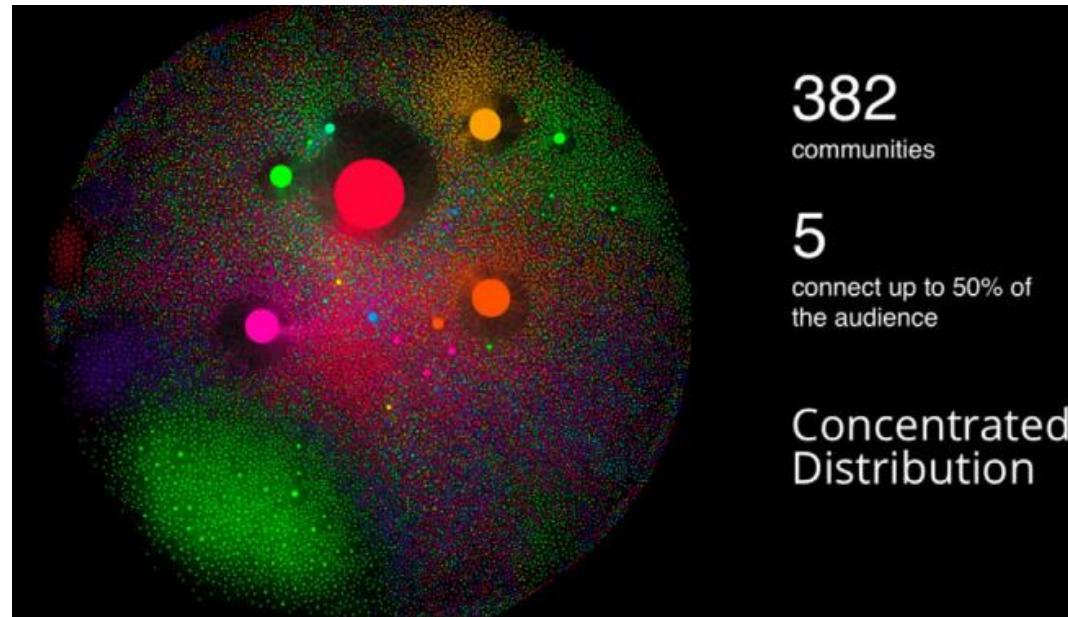


O que são grafos?

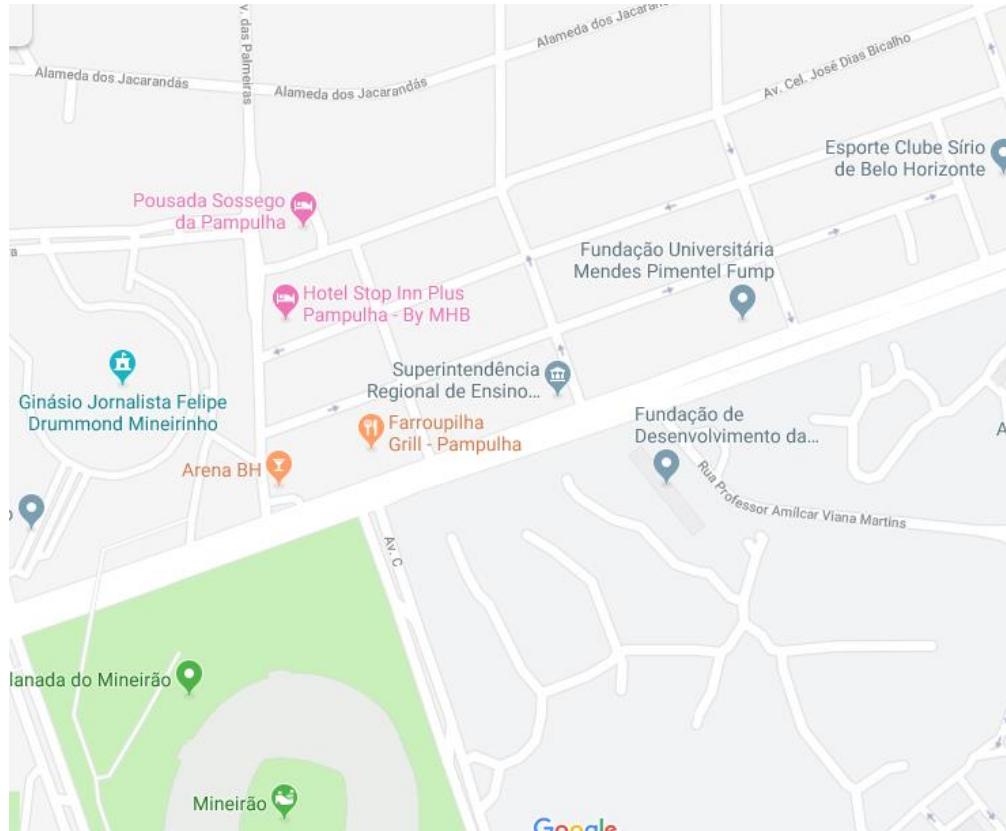


Por que utilizar grafos?

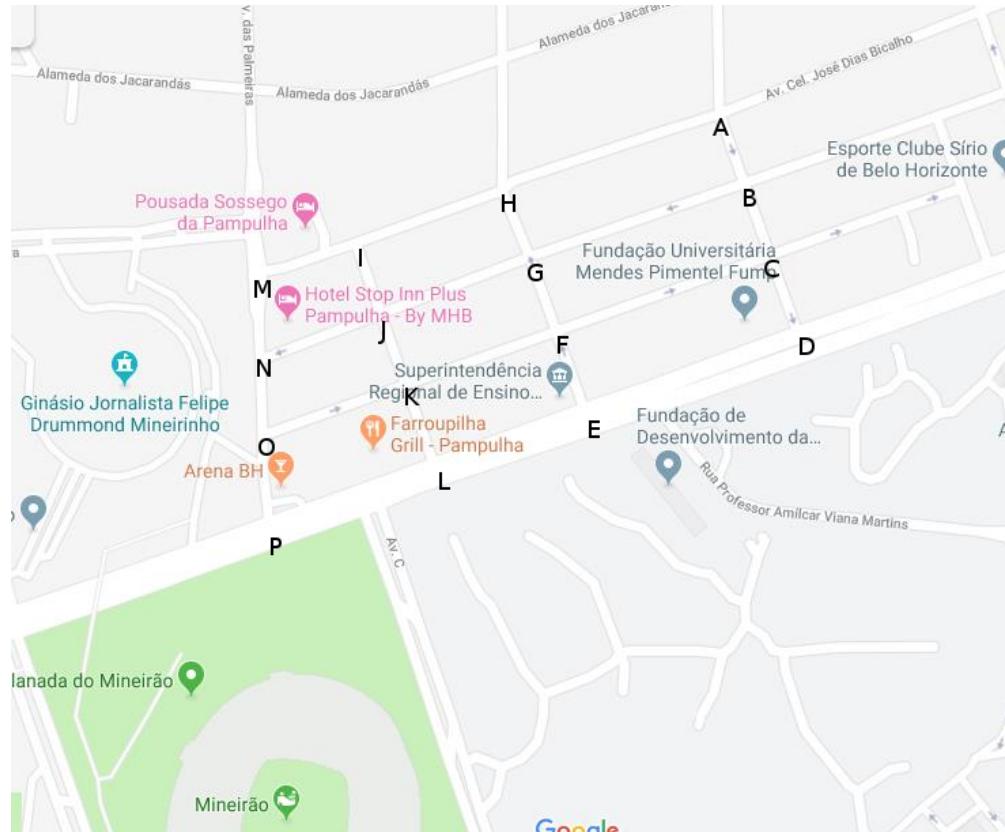
- Performance;
- Flexibilidade;
- Agilidade.



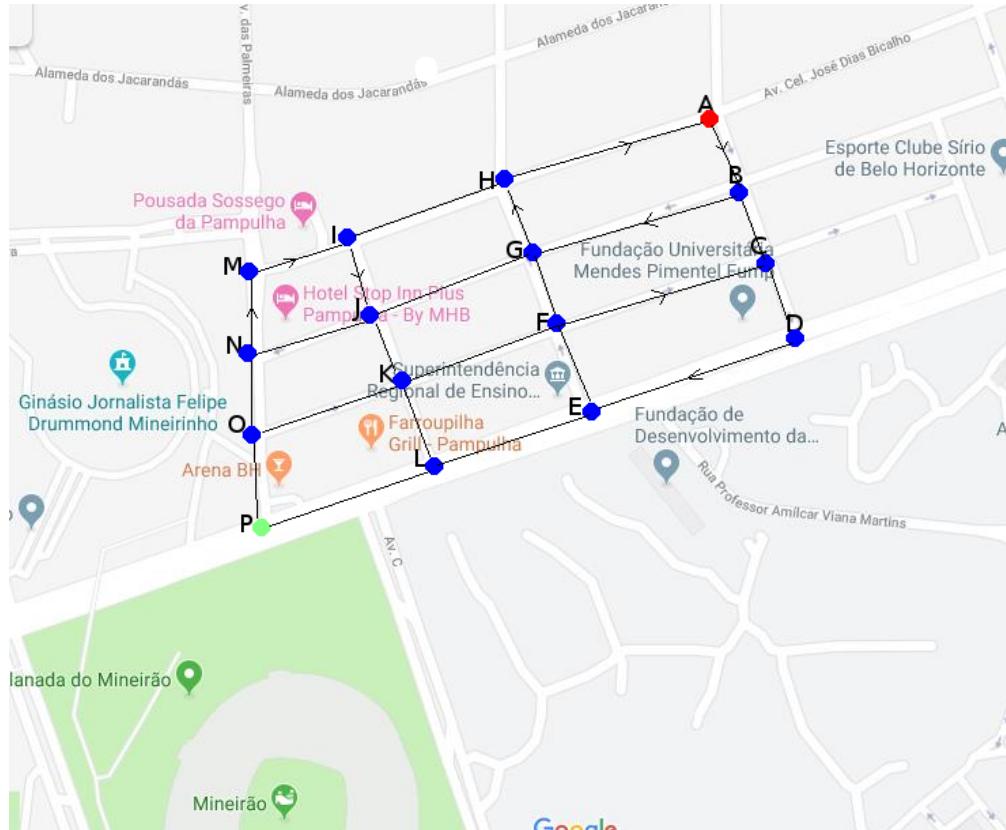
Grafos para geolocalização



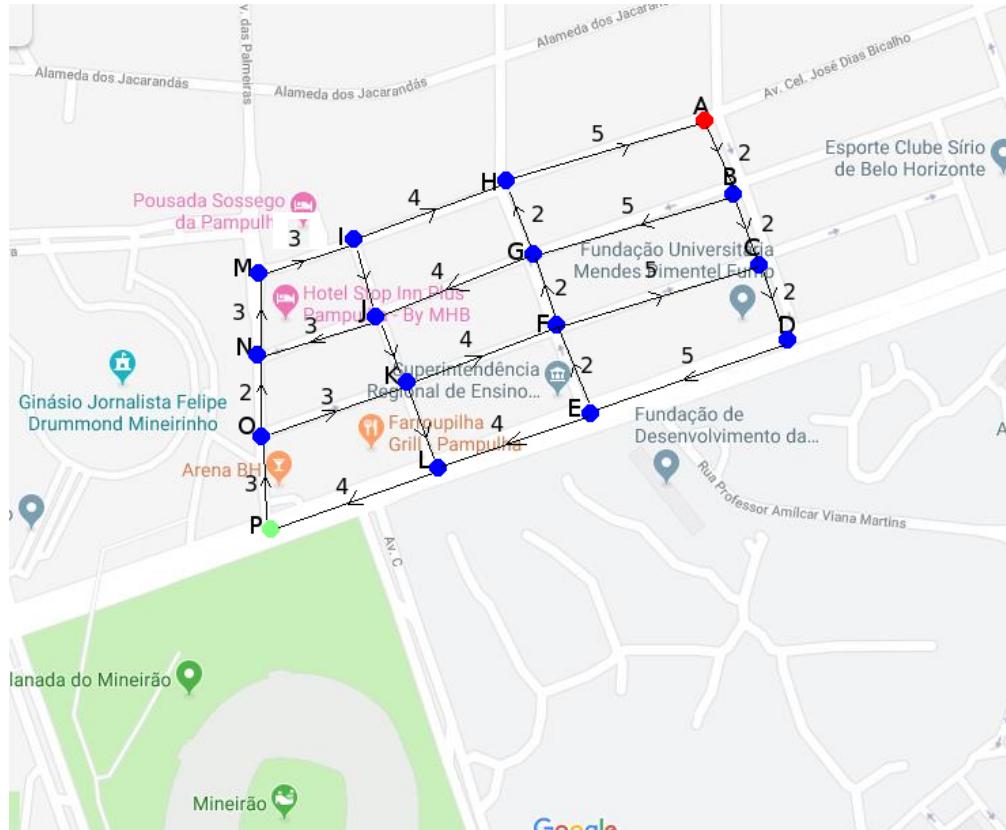
Grafos para geolocalização



Grafos para geolocalização



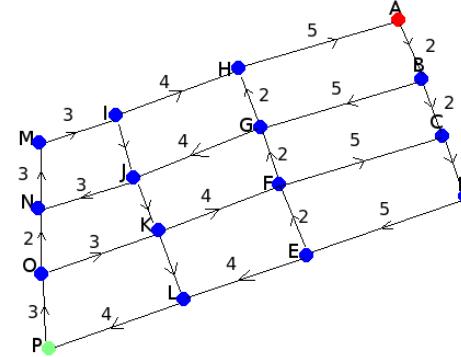
Grafos para geolocalização



Grafos para geolocalização

- Matriz de adjacências.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	-1	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
B	-1	-1	2	-1	-1	-1	5	-1	-1	-1	-1	-1	-1	-1	-1	-1
C	-1	-1	-1	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
D	-1	-1	-1	5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
E	-1	-1	-1	-1	-1	2	-1	-1	-1	-1	4	-1	-1	-1	-1	-1
F	-1	-1	5	-1	-1	-1	2	-1	-1	-1	-1	-1	-1	-1	-1	-1
G	-1	-1	-1	-1	-1	-1	2	-1	-1	4	-1	-1	-1	-1	-1	-1
H	5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
I	-1	-1	-1	-1	-1	-1	4	-1	2	-1	-1	-1	-1	-1	-1	-1
J	-1	-1	-1	-1	-1	-1	-1	-1	-1	2	-1	-1	3	-1	-1	-1
K	-1	-1	-1	-1	-1	4	-1	-1	-1	-1	2	-1	-1	-1	-1	-1
L	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	4
M	-1	-1	-1	-1	-1	-1	-1	3	-1	-1	-1	-1	-1	-1	-1	-1
N	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	3	-1	-1	-1	-1	-1
O	-1	-1	-1	-1	-1	-1	-1	-1	3	-1	-1	2	-1	-1	-1	-1
P	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	3	-1	-1	-1



Custo total:

$A \rightarrow P: 2+2+2+5+4+4$

$2+5+4+2+2+4$

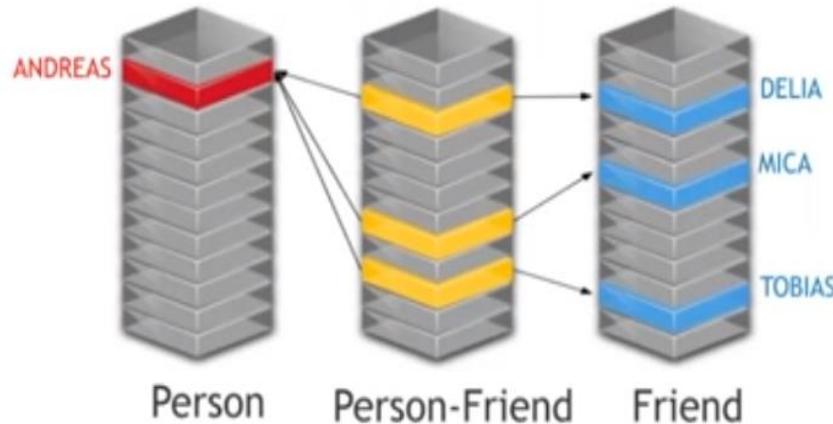
Processamento da linguagem natural

- O que é processamento da linguagem natural?
 - Tentar extrair o significado ou opinião contida em textos.

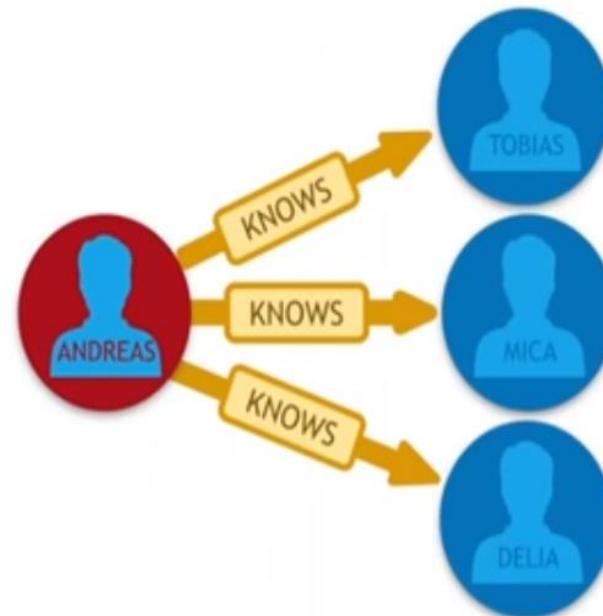


Processamento da linguagem natural

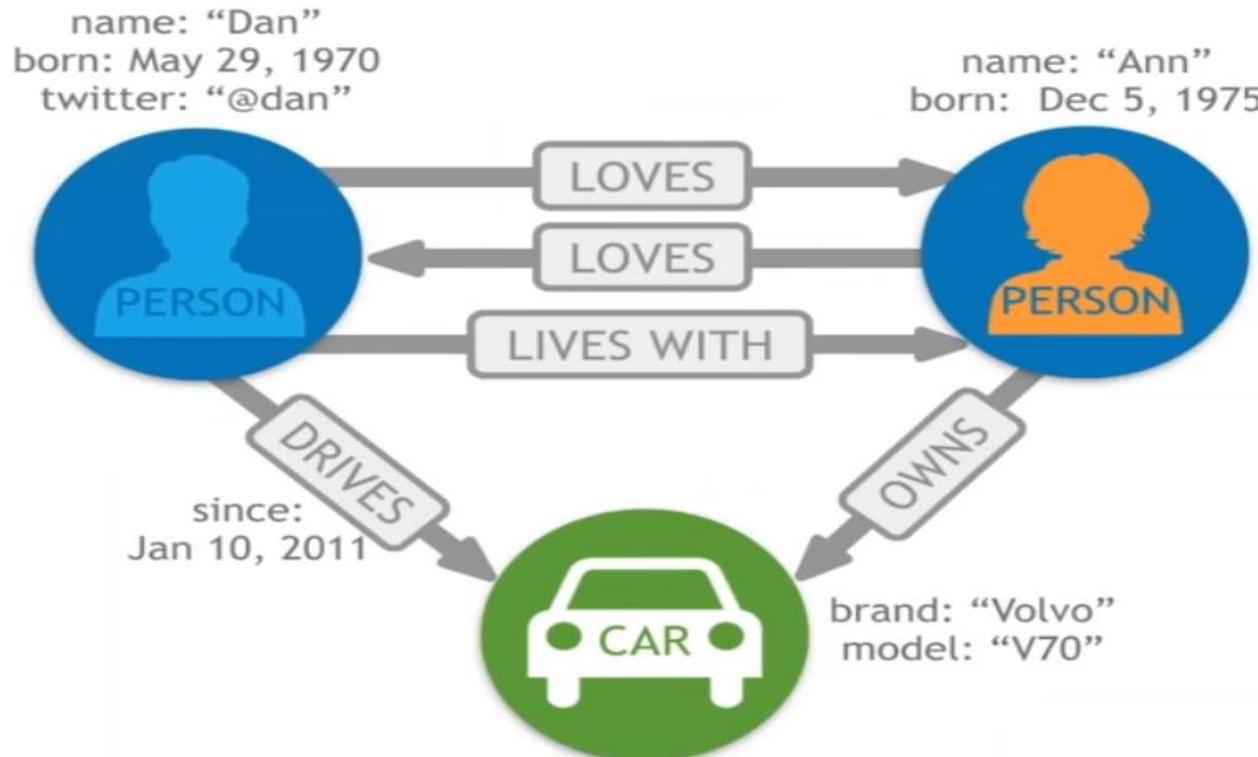
Relacional



Grafo

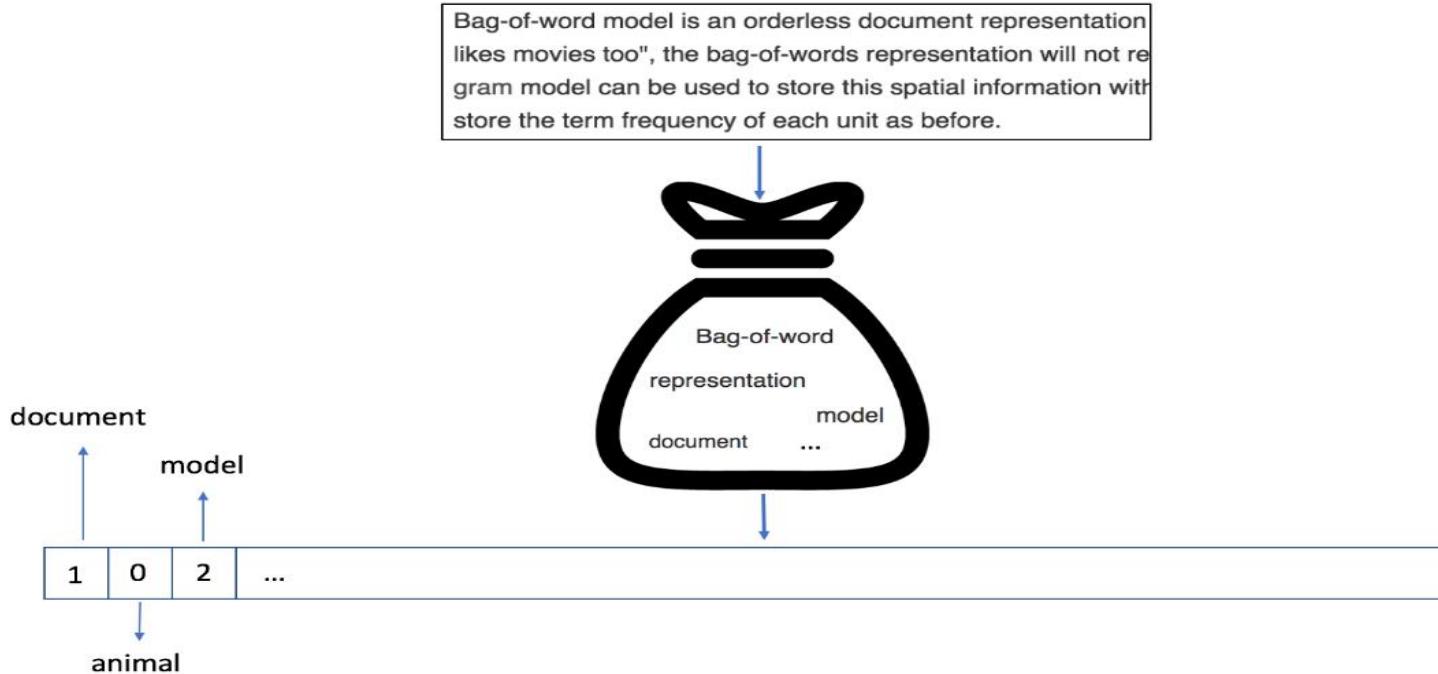


Processamento da linguagem natural



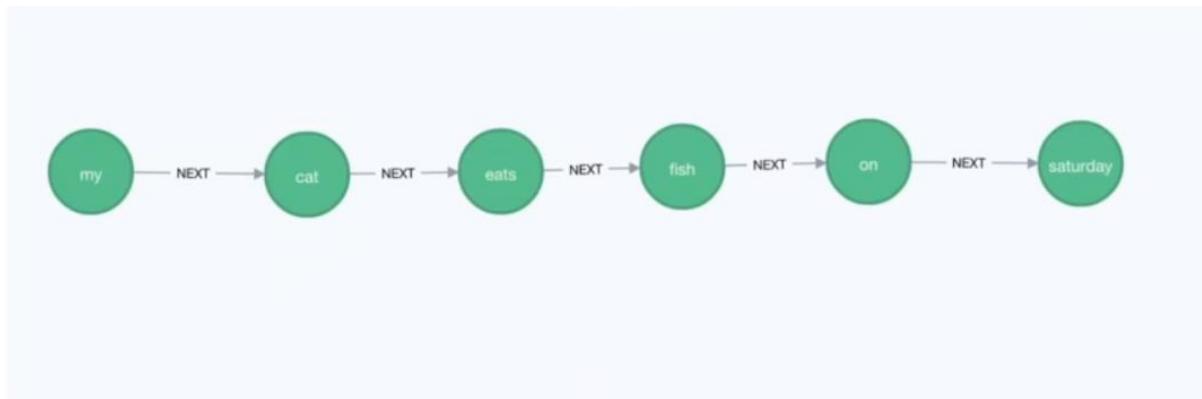
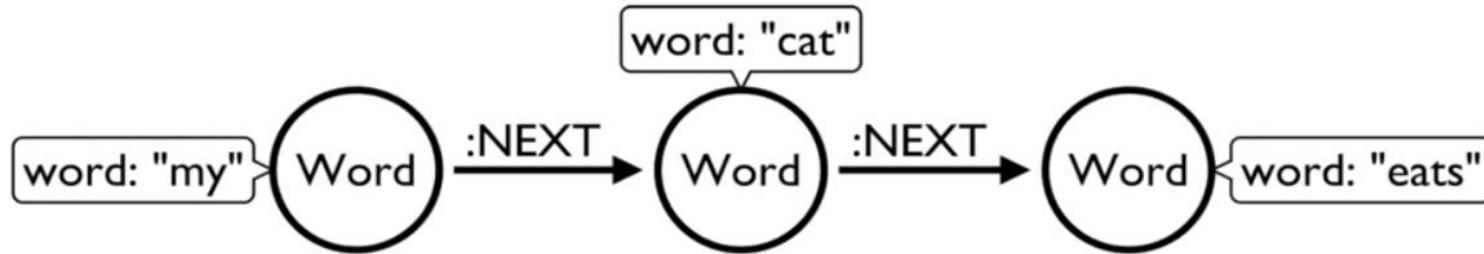
Processamento da linguagem natural

- Representando texto como vetor:



Processamento da linguagem natural

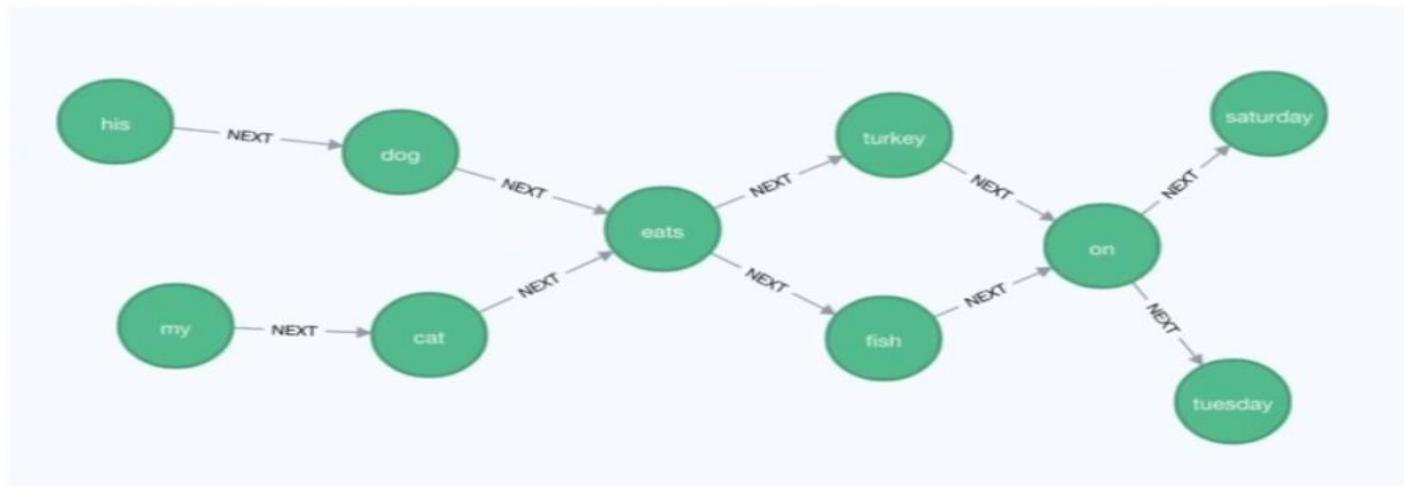
- Representando texto como grafo de adjacências:



Processamento da linguagem natural

- Representação de um texto completo:

```
1 WITH split(tolower("His dog eats turkey on tuesday"), " ") AS text
2 UNWIND range(0,size(text)-2) AS i
3 MERGE (w1:Word {name: text[i]})  
4 MERGE (w2:Word {name: text[i+1]})  
5 MERGE (w1)-[:NEXT]->(w2)
```



Processamento da linguagem natural

- Frequência das palavras no texto.

word	word_count
eats	4
on	4
dog	2
cat	2
turkey	2

- Frequência de associações no texto.

word_pair	count
[this, tv]	97
[tv, for]	69
[great, picture]	65
[bought, this]	63
[smart, tv]	60

Conclusão

- ✓ O que são grafos?
- ✓ Por que utilizar grafos?
- ✓ Onde encontramos grafos?

■ Próxima aula

- Definições.

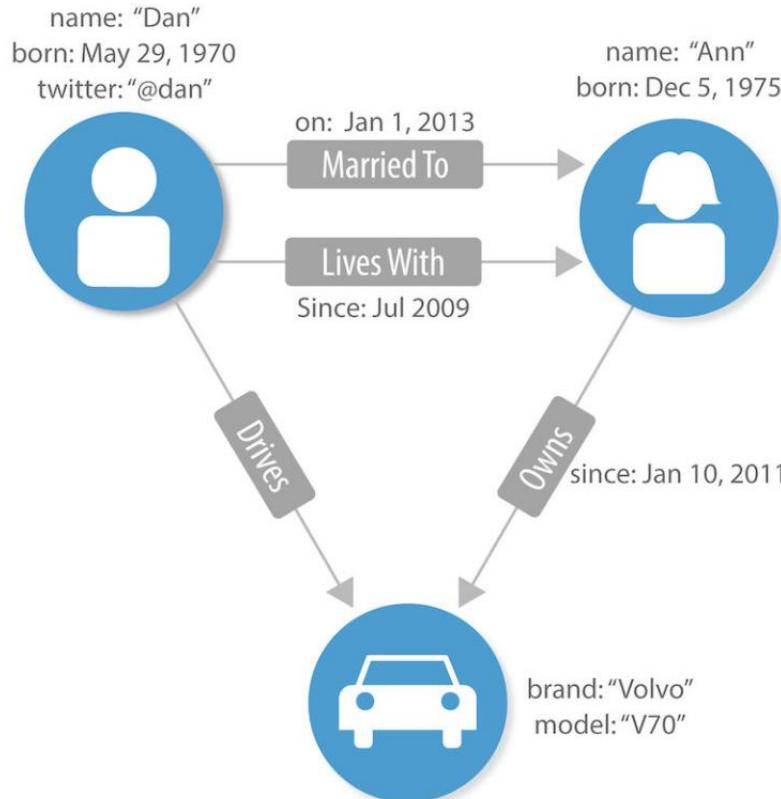


Aula 6.2.1. Conceitos e terminologias (Parte I)

Nesta aula

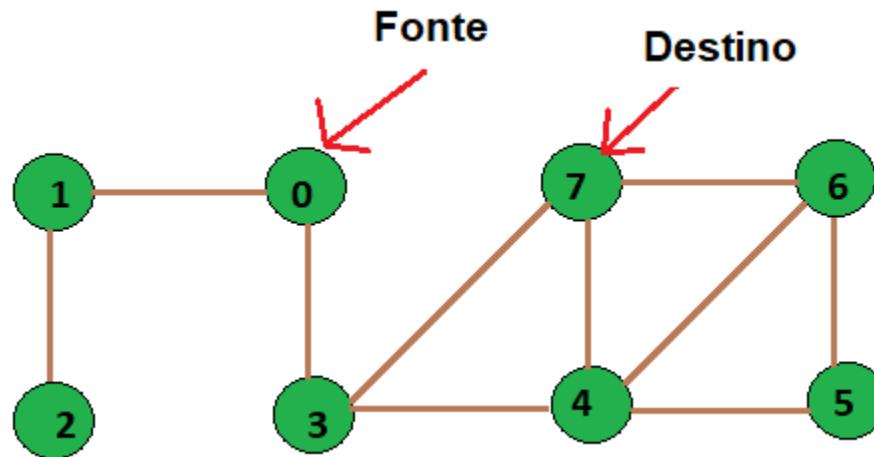
- Definições e terminologias.

Grafos rotulados

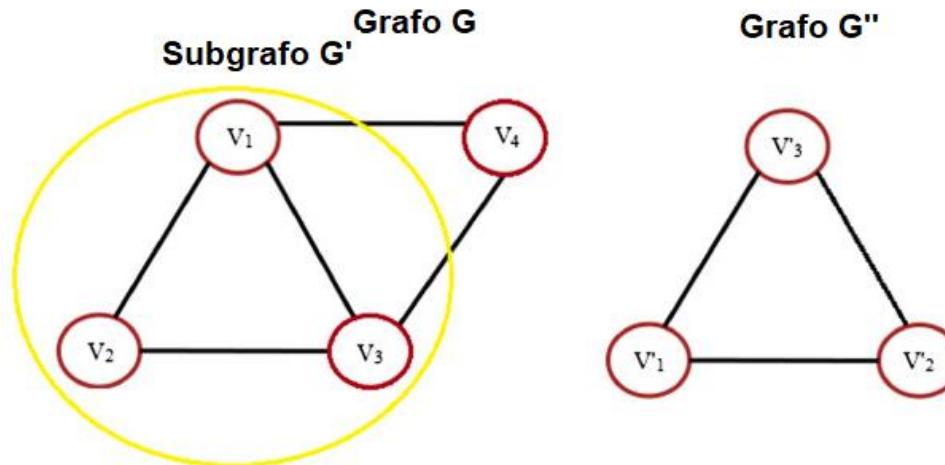


- Relacionamentos, propriedades e atributos.

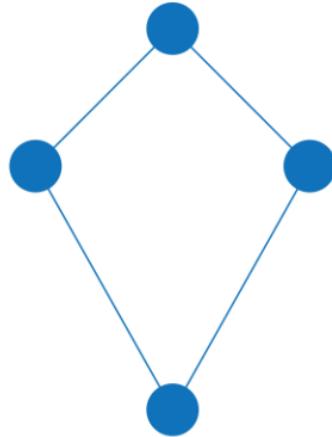
Caminhos



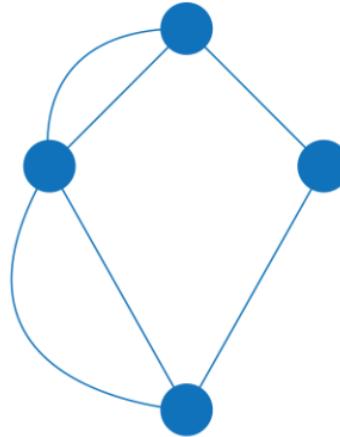
Grafo e subgrafo



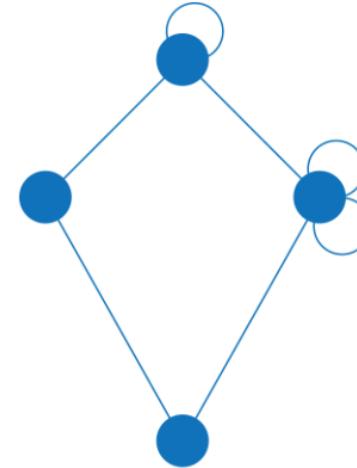
Grafos simples, múltiplos e pseudografo



Grafo Simples

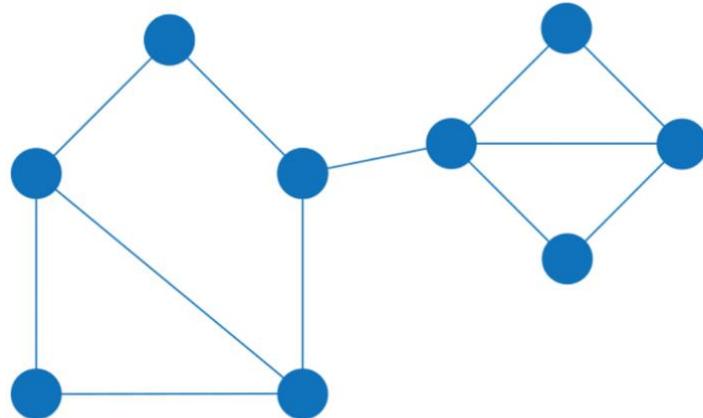


Grafo Múltiplo
(Multigrafo)

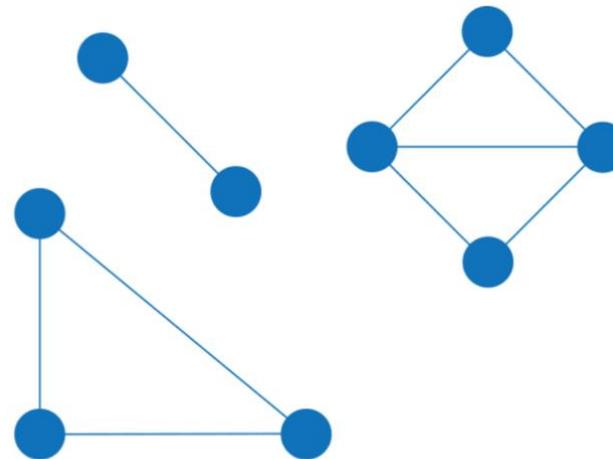


Pseudografo

Conecados e desconectados

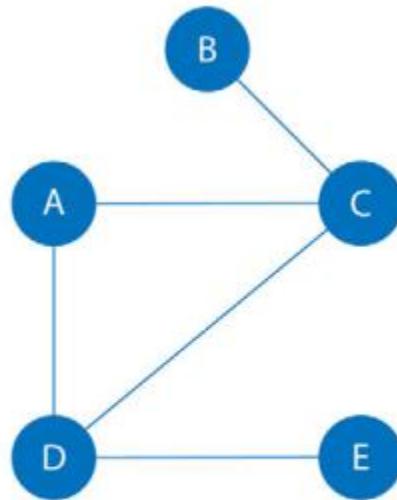


Conecado

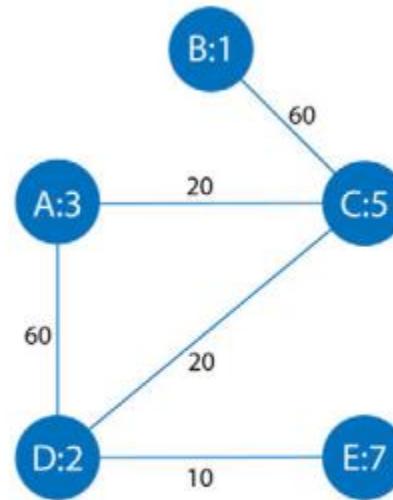


Desconectado

Pesos ou custos

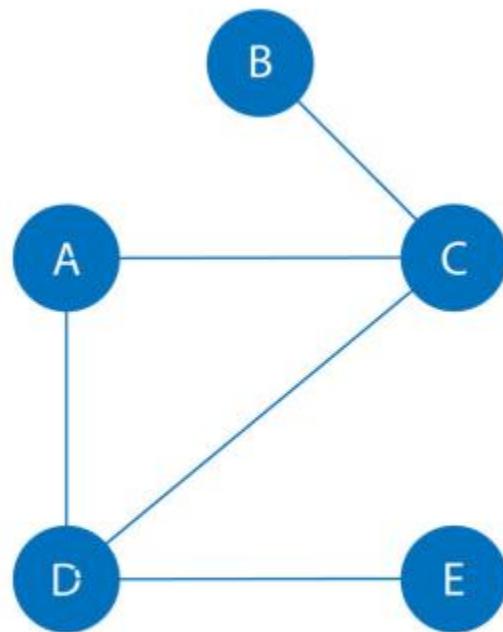


Sem pesos

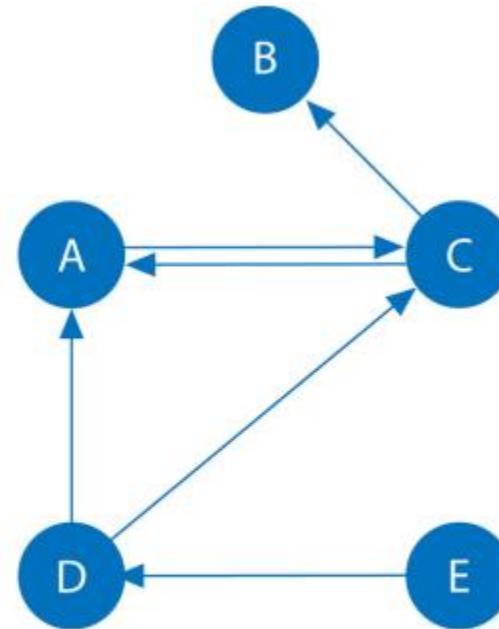


Com pesos

Direcionados e não direcionados

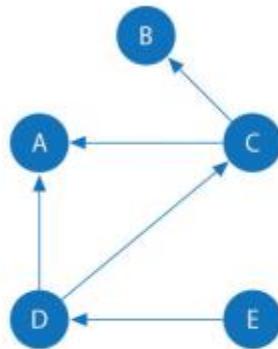


Não direcionado



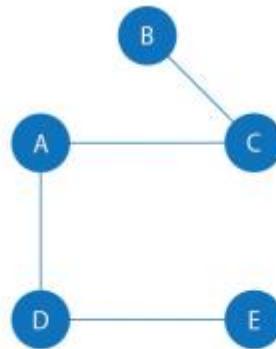
Direcionado

Cíclicos e acíclicos

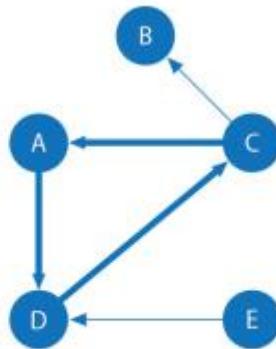


Direcionado

Acíclicos

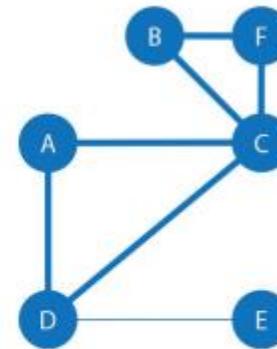


Não Direcionado

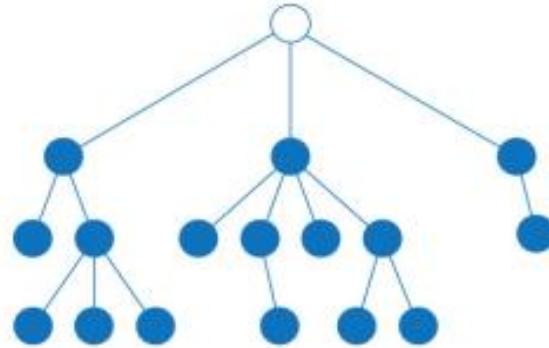


Direcionado

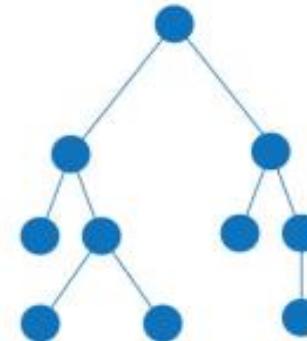
Cíclicos



Não Direcionado



Múltiplas



Binária

- ✓ Conceitos e terminologias.

■ Próxima aula

- Conceitos e terminologias (Parte II).



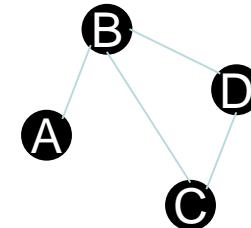
Aula 6.2.2. Conceitos e terminologias (Parte II)

Nesta aula

- Conceitos e terminologias (parte II).

Questão para aquecimento

- Como vocês já devem saber, um grafo é composto, basicamente, por vértices (nós) e arestas. Por exemplo, o grafo ao lado possui vértices A, B, C e D.
- As arestas são $S=\{(A,B), (B,C), (B,D), (C,D)\}$. Assim, esse grafo possui 4 vértices e 4 arestas. O vértice B, por exemplo, apresenta grau 3, pois possui 3 arestas incidentes. O mesmo vértice B, possui os vértices A, C e D como adjacentes. Podemos dizer que esse grafo possui a seguinte sequência de graus 1, 3, 2 e 2.

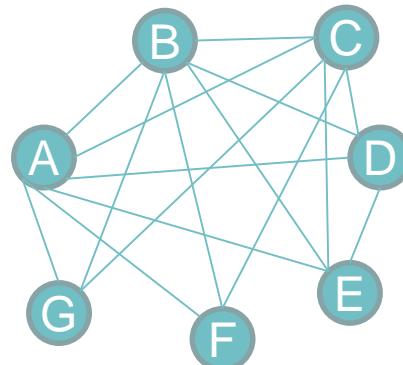


- Sendo assim, é possível construir um grafo com esta sequência de graus?

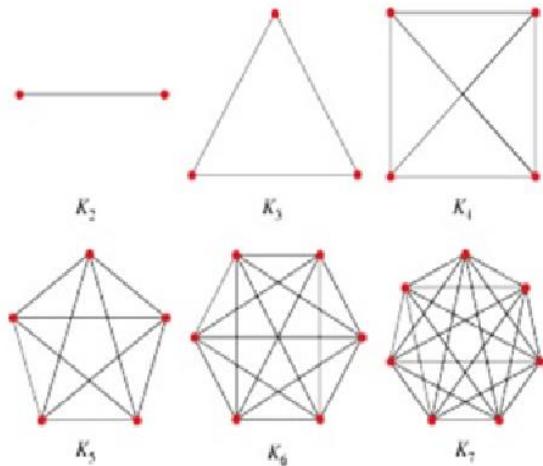
6, 6, 6, 4, 4, 2, 2.

Resposta

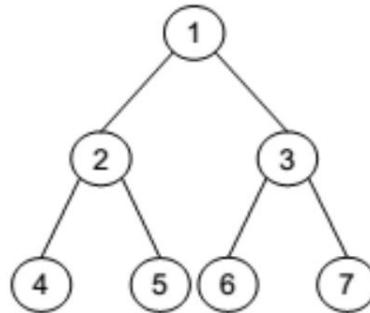
- NÃO. Uma vez que temos 3 vértices com grau 6, eles são adjacentes aos demais vértices. Assim, cada vértice do grafo deve ter, pelo menos, grau 3.



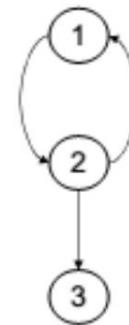
Grafos completos



Completo



Incompleto



- Máxima densidade:

$$\text{████████} = \frac{\text{█}(\text{█}-1)}{2}$$

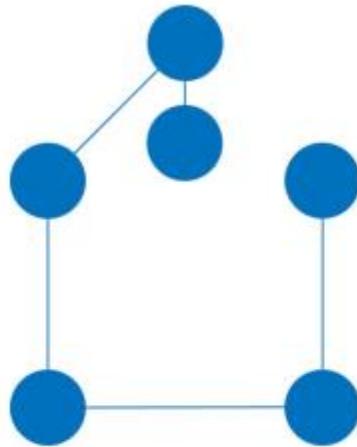
- Máxima densidade:

$$\text{████████} = \frac{\text{█}(\text{█}-1)}{2}$$

- Densidade real:

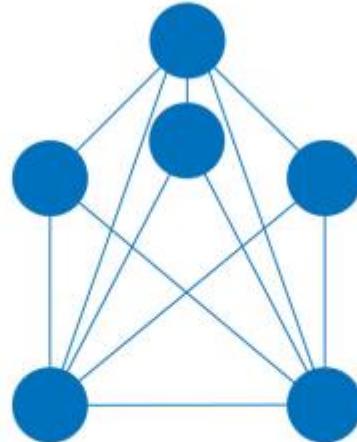
$$\text{█} = \frac{2(\text{█})}{\text{█}(\text{█}-1)}$$

Densidade



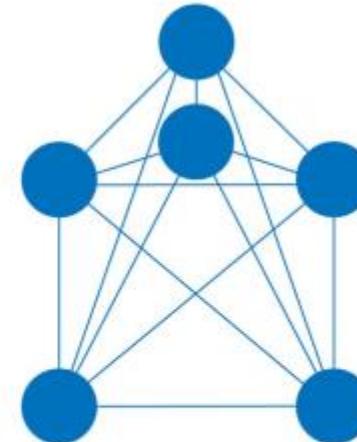
Sparse
Density = 0.3

$$D = \frac{2(5)}{6(6-1)}$$



Dense
Density = 0.8

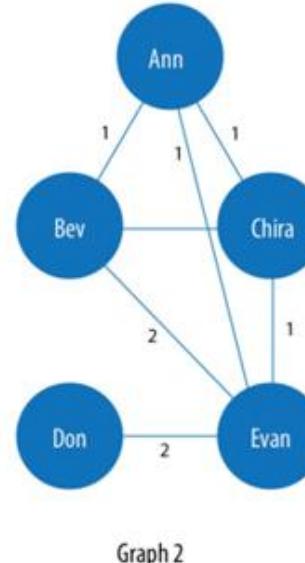
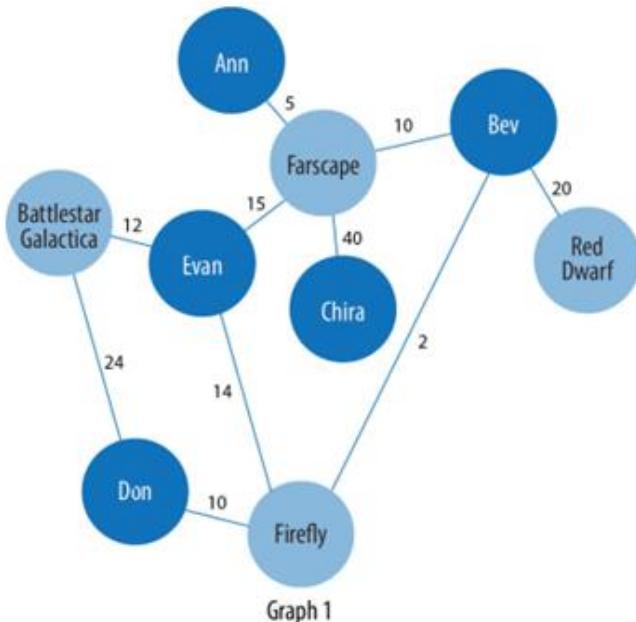
$$D = \frac{2(12)}{6(6-1)}$$



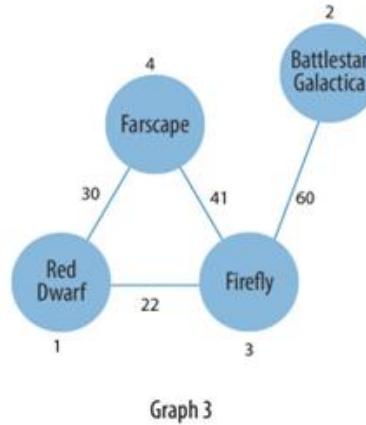
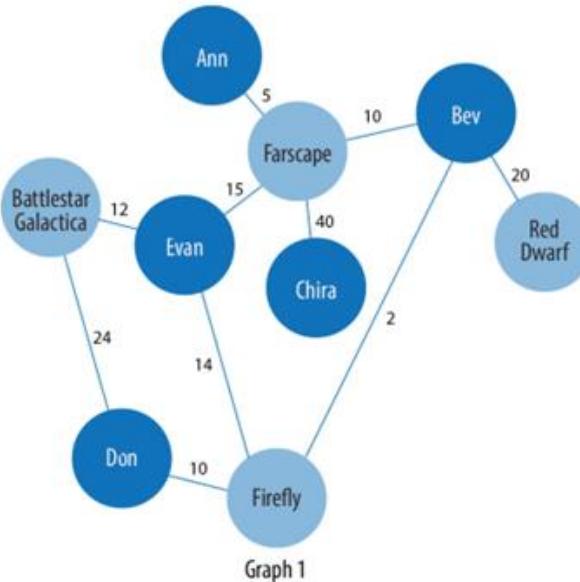
Complete (Clique)
Density = 1.0

$$D = \frac{2(15)}{6(6-1)}$$

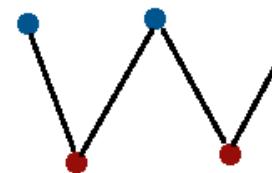
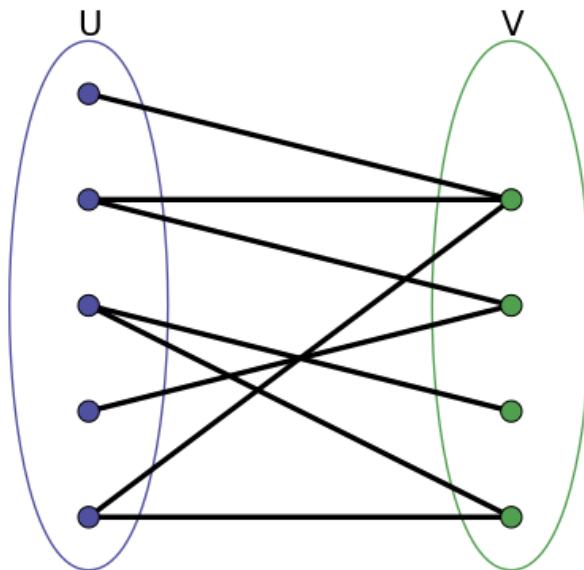
Monopartido, bipartido e K-partido



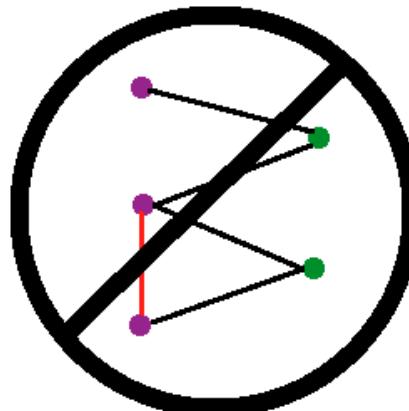
Monopartido, bipartido e K-partido



Bipartido

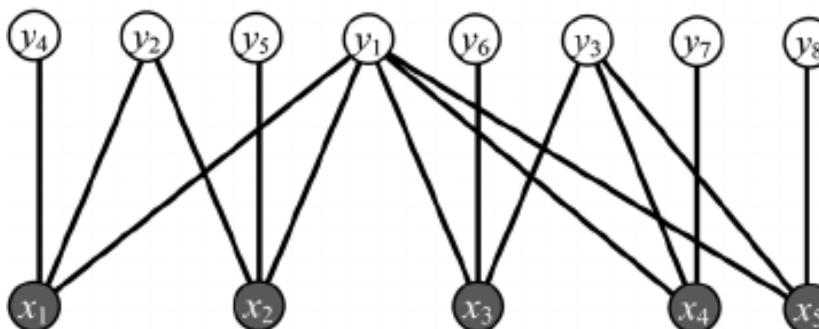
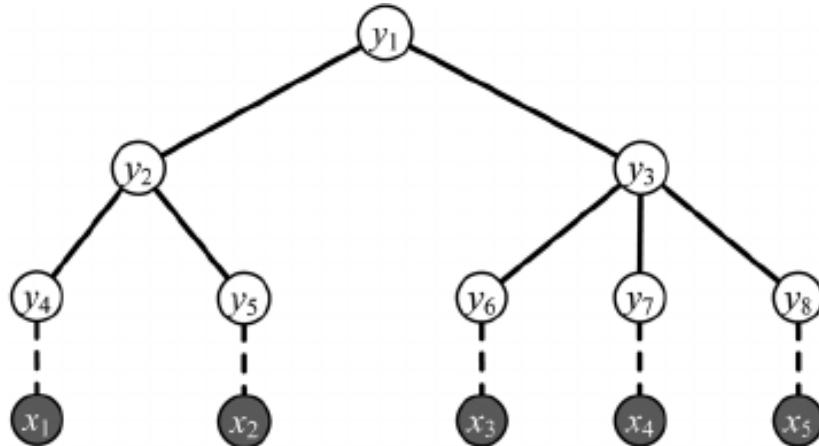


Bipartido



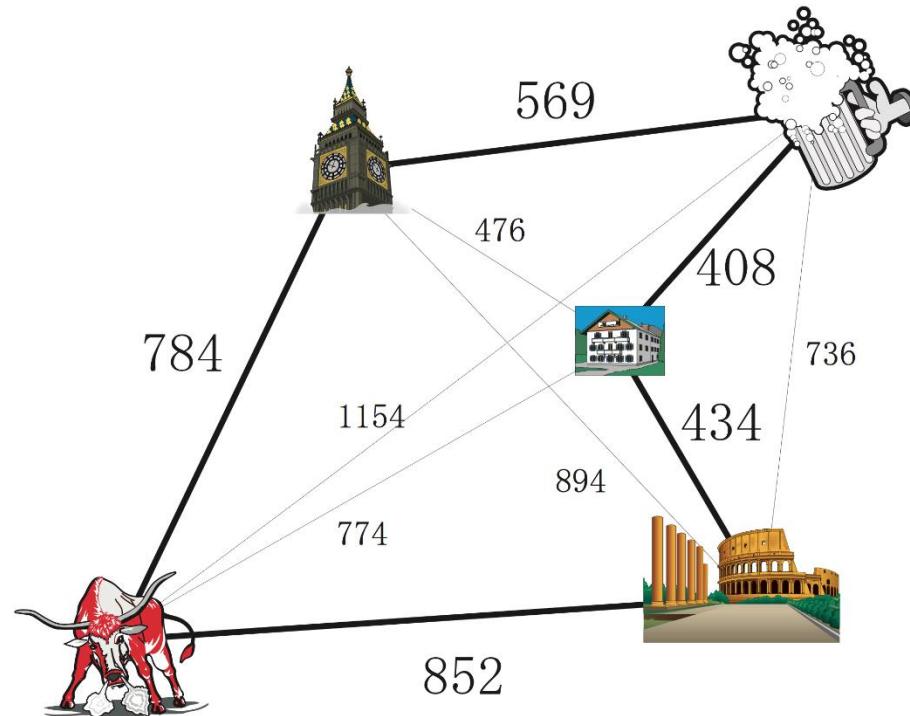
Não Bipartido

Bipartido

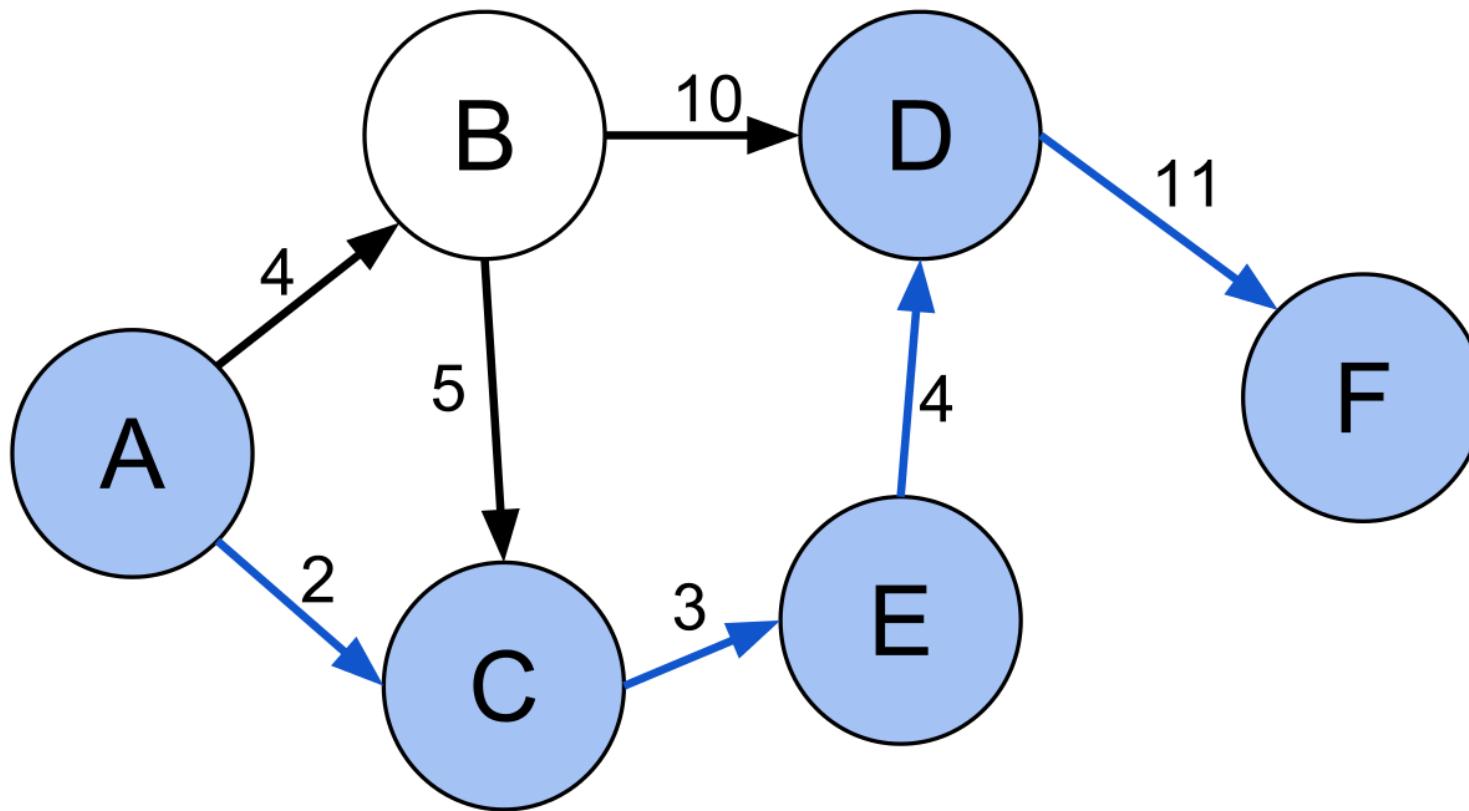


Algoritmos para encontrar caminho

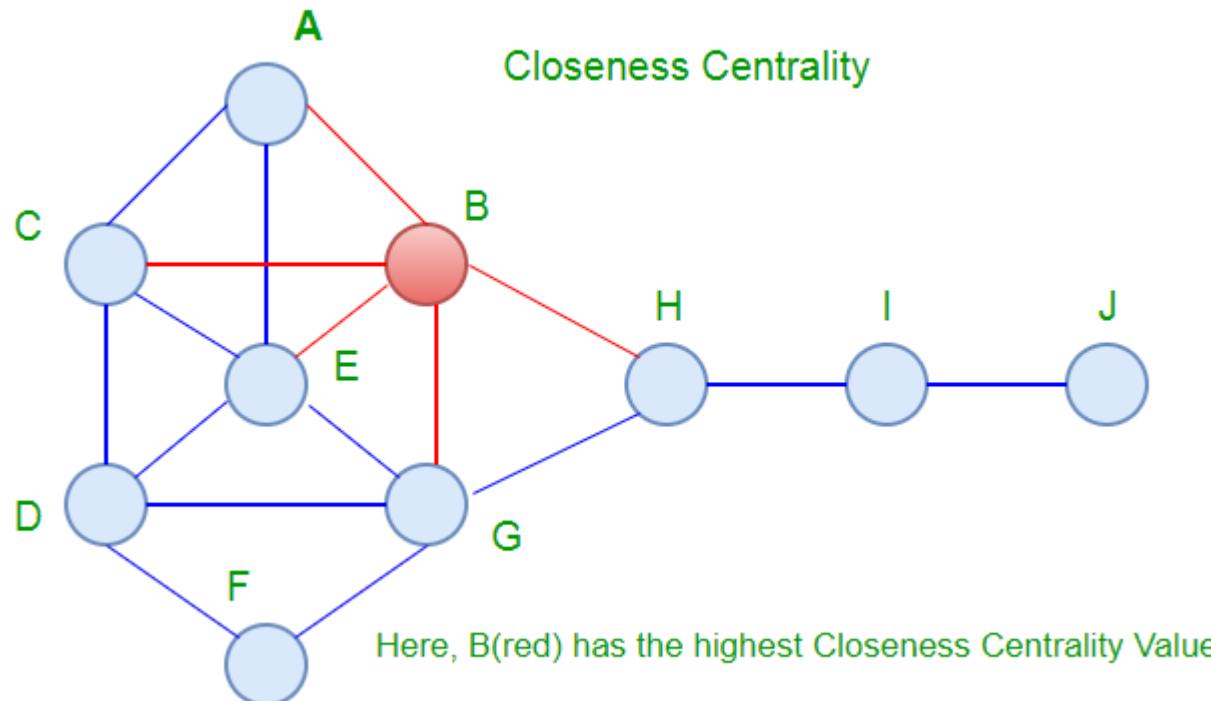
- Caixeiro viajante.



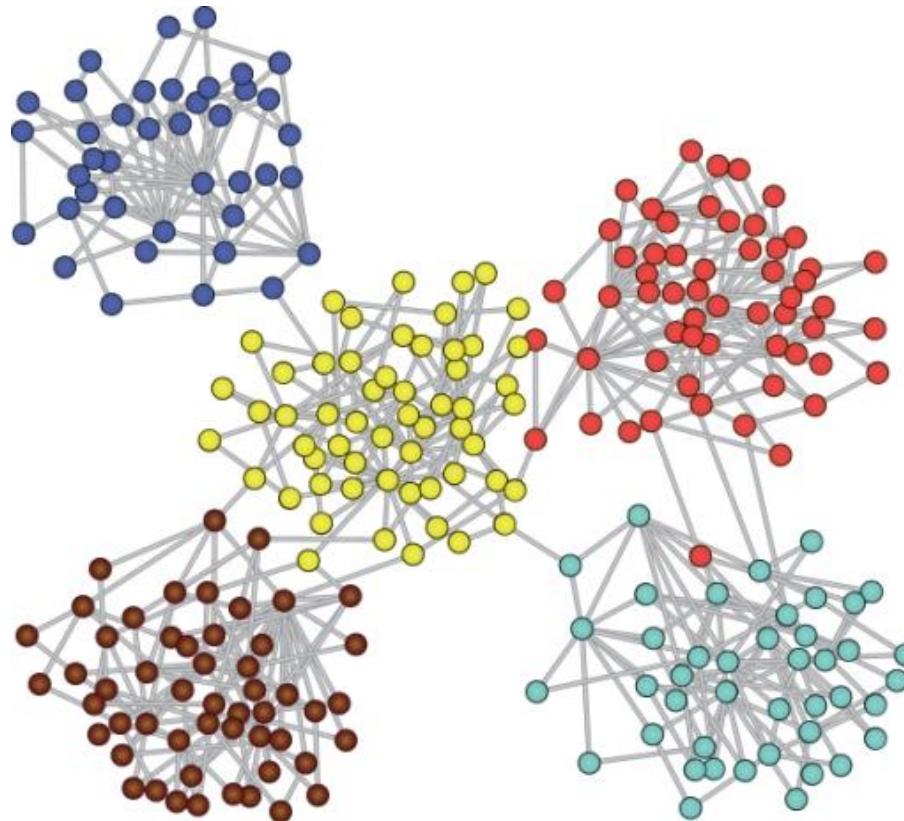
Algoritmos para encontrar menor caminho



Algoritmos de centralidade



Detecção de comunidades



Conceitos e definições.

■ Próxima aula

- Neo4j e Cypher.

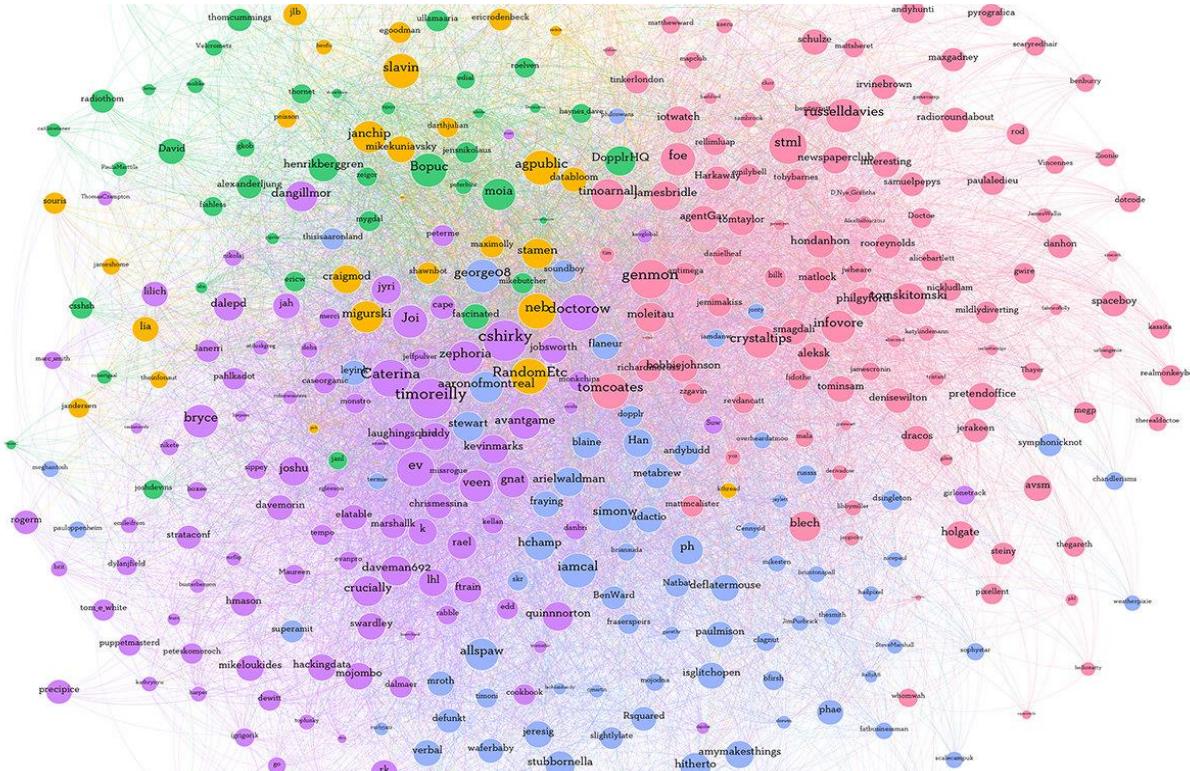


Aula 6.3.1. Neo4j e Cypher (Parte I)

Nesta aula

- O que é o Neo4j?
- O que é Cypher?
- Exemplos.

O que é Neo4j?



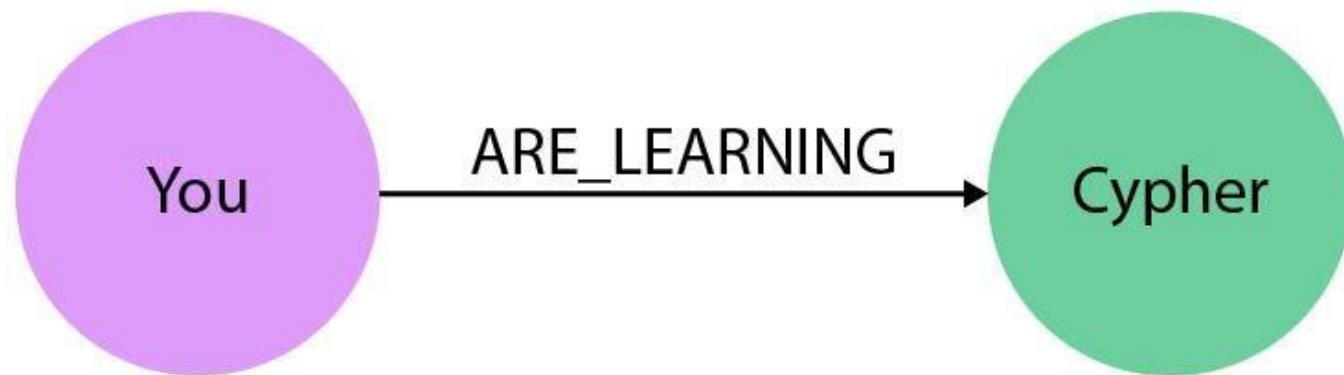
O que é Neo4j

- DB NoSQL;
- Código aberto;
- Grande comunidade;
- Interface amigável;
- Alta performance.



The #1 Database for Connected Data

O que é Cypher?



O que é Cypher?

Cypher Query Language



Utilizando a Cypher

- Nό:

```
()  
(matrix)  
(:Movie)  
(matrix:Movie)  
(matrix:Movie {title: "The Matrix"})  
(matrix:Movie {title: "The Matrix", released: 1997})
```

- Relacionamento:

```
-->  
-[role]->  
-[:ACTED_IN]->  
-[role:ACTED_IN]->  
-[role:ACTED_IN {roles: ["Neo"]}]-->
```

- Padrões:

```
(keanu:Person:Actor {name: "Keanu Reeves"} )  
-[role:ACTED_IN {roles: ["Neo"] } ]->  
(matrix:Movie {title: "The Matrix"} )
```

Criando um nó

```
CREATE (:Movie { title:"The Matrix",released:1997 })
```

```
+-----+  
| No data returned. |  
+-----+  
Nodes created: 1  
Properties set: 2  
Labels added: 1
```

Movie

title = 'The Matrix'
released = 1997

Retornando um valor

```
CREATE (p:Person { name:"Keanu Reeves", born:1964 })
RETURN p
```

```
+-----+
| p |
+-----+
| Node[1]{name:"Keanu Reeves",born:1964} |
+-----+
1 row
Nodes created: 1
Properties set: 2
Labels added: 1
```

Consultas

```
MATCH (m:Movie)  
RETURN m
```

Movie

title = 'The Matrix'
released = 1997

Movie

title = 'Forrest Gump'
released = 1994

```
MATCH (p:Person { name:"Keanu Reeves" })  
RETURN p
```

Person

name = 'Keanu Reeves'
born = 1964

Consultas

```
MATCH (p:Person { name:"Tom Hanks" })-[r:ACTED_IN]->(m:Movie)
RETURN m.title, r.roles
```

```
+-----+  
| m.title | r.roles |  
+-----+  
| "Forrest Gump" | ["Forrest"] |  
+-----+  
1 row
```

```
MATCH (m:Movie)
WHERE m.title = "The Matrix"
RETURN m
```

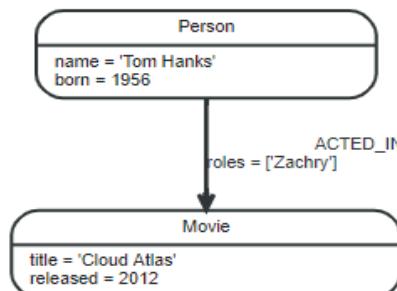
```
+-----+  
| m |  
+-----+  
| (:Movie {title: "The Matrix", released: 1997}) |  
+-----+  
1 row
```

Alterando valores

```
MERGE (m:Movie { title:"Cloud Atlas" })
ON CREATE SET m.released = 2012
RETURN m
```

```
+-----+
| m
+-----+
| Node[5]{title:"Cloud Atlas",released:2012} |
+-----+
1 row
```

```
MATCH (m:Movie { title:"Cloud Atlas" })
MATCH (p:Person { name:"Tom Hanks" })
MERGE (p)-[:ACTED_IN]->(m)
ON CREATE SET r.roles =['Zachry']
RETURN p,r,m
```



Conclusão

- ✓ O que é o Neo4j?
- ✓ O que é Cypher?
- ✓ Exemplos.

■ Próxima aula

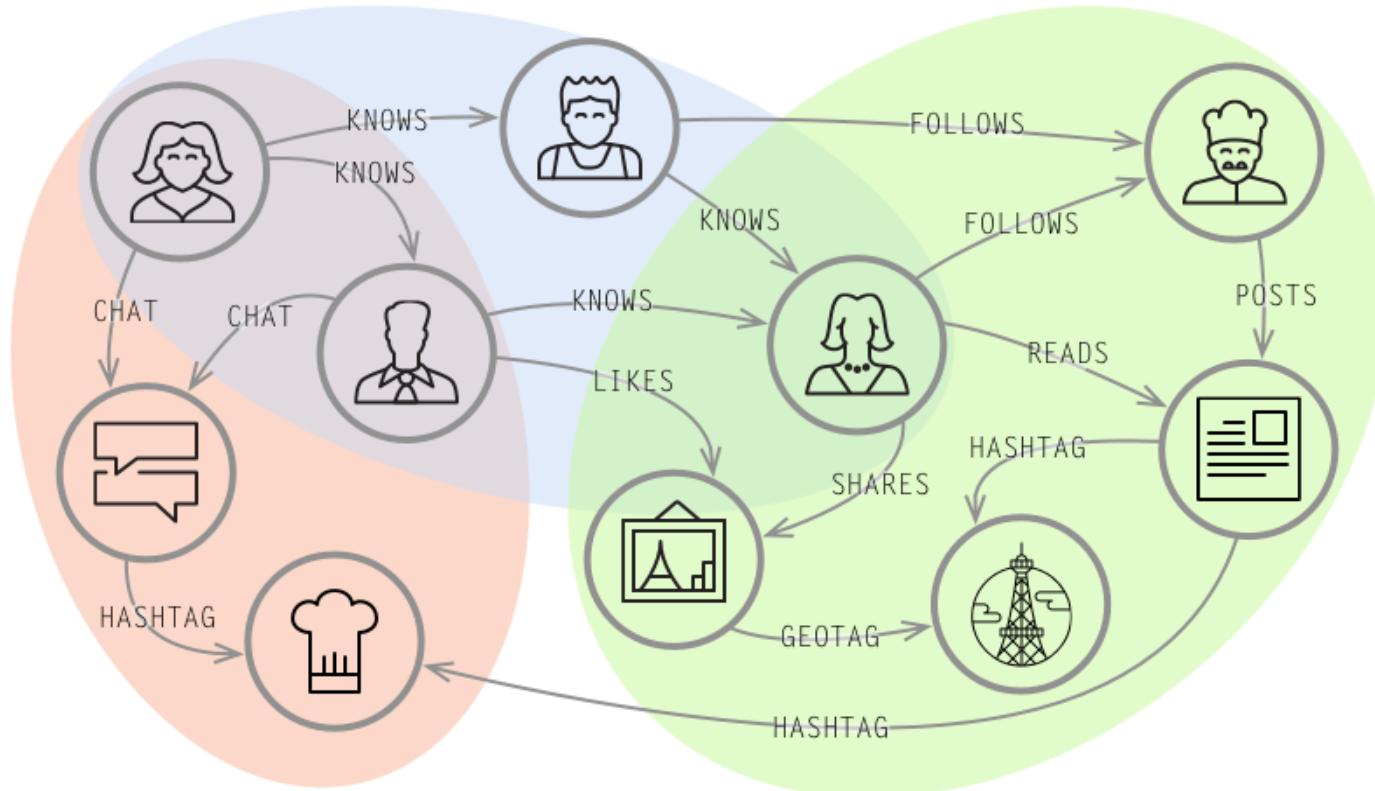
- Neo4j e Cypher (Parte II).



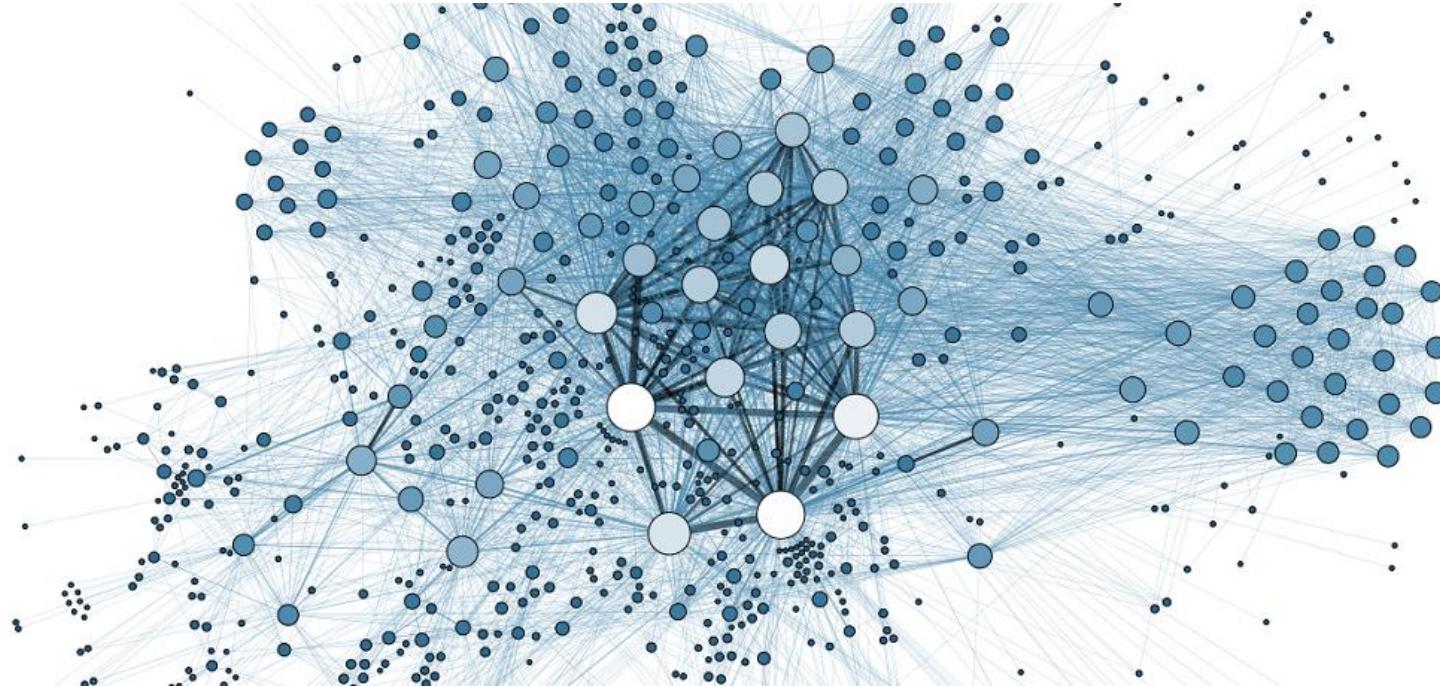
Aula 6.3.2. Neo4j e Cypher (Parte II)

- Neo4j e Cypher (Parte II).

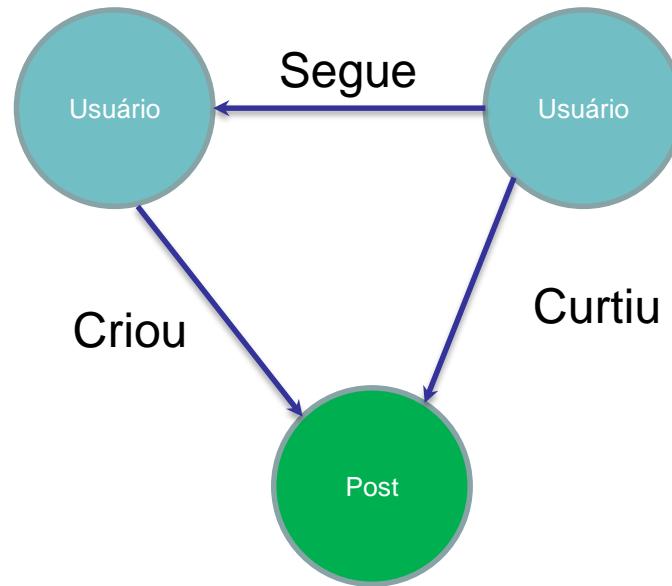
Grafos em redes sociais



Grafos em redes sociais



Grafos em redes sociais



- 2 nós;
- 3 relacionamentos.

- Criar o usuário:

Usuário: nome, e-mail.

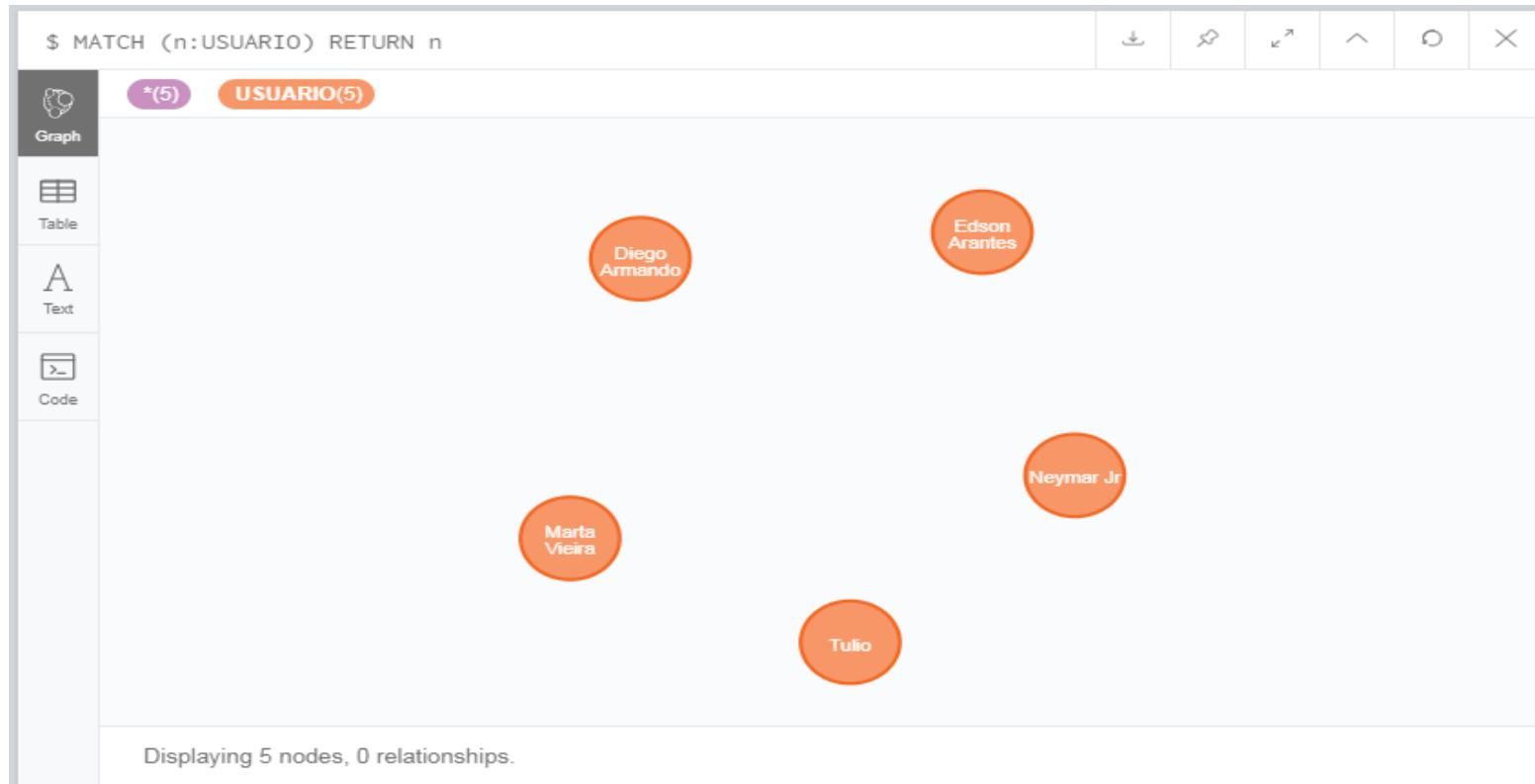
e-mail → único



```
CREATE CONSTRAINT ON (usuario:USUARIO) ASSERT usuario.email  
IS UNIQUE
```

```
CREATE (tulio:USUARIO {nome: "Tulio", email:tulio@example.com"}  
) RETURN tulio
```

Grafos em redes sociais



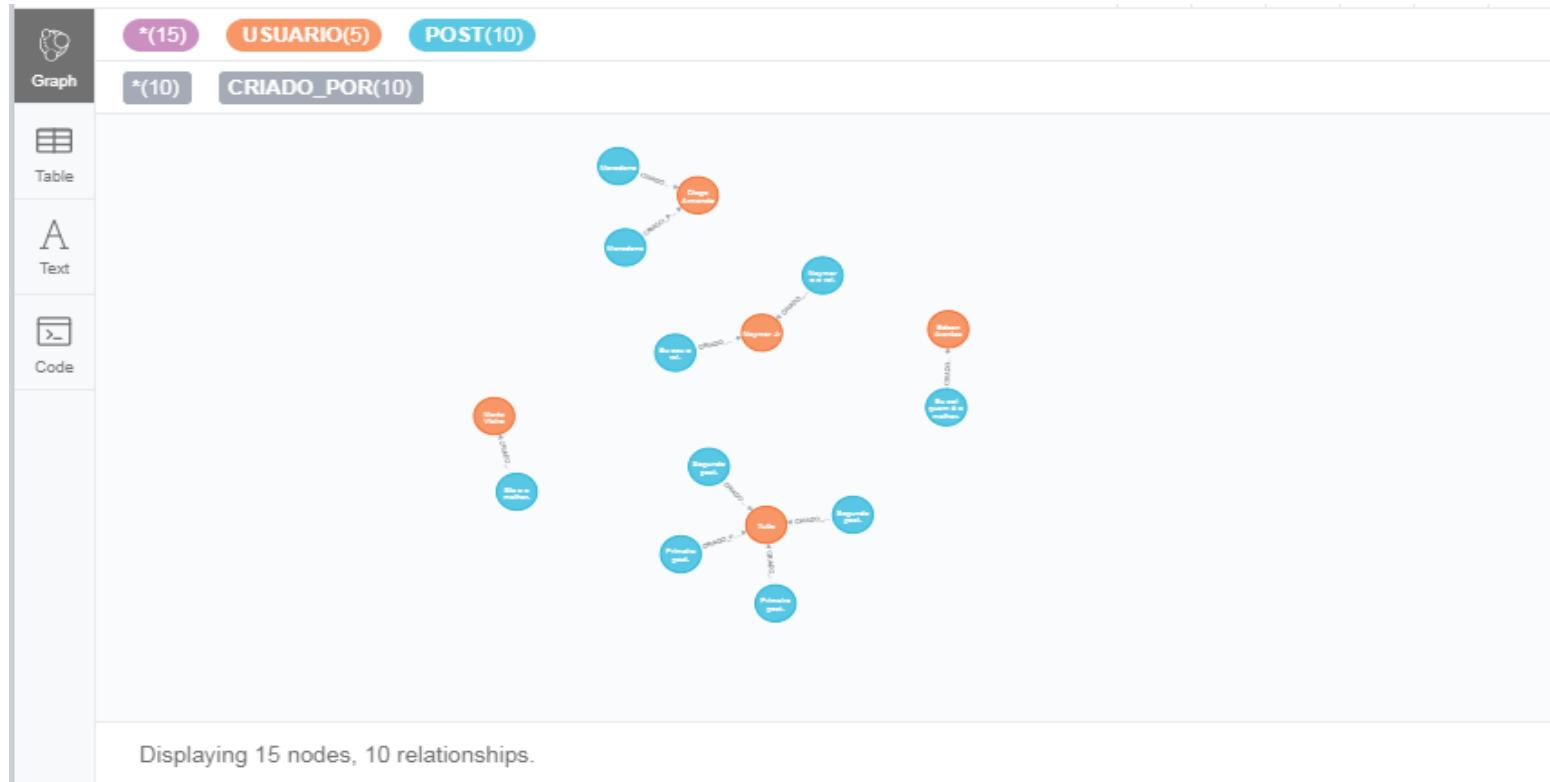
- Criar um post:

Post: título, conteúdo
não permite post anônimo



```
MATCH (tulio:USUARIO {email: "tulio@example.com"})
CREATE (tulio)-[r:CRIADO_POR]-(post:POST { titulo: "Primeiro post.",
conteudo: "Eu amo o Neo4j."})
}) RETURN tulio, r, post
```

Grafos em redes sociais



Grafos em redes sociais

- Seguindo usuários:



```
MATCH (tulio:USUARIO {email: "tulio@example.com"}), (pele:USUARIO {email: "pele@example.com"}) CREATE (pele)-[r:FOLLOWS]->(tulio)  
RETURN tulio, r, pele
```

Grafos em redes sociais

Graph

*(15) USUARIO(5) POST(10)

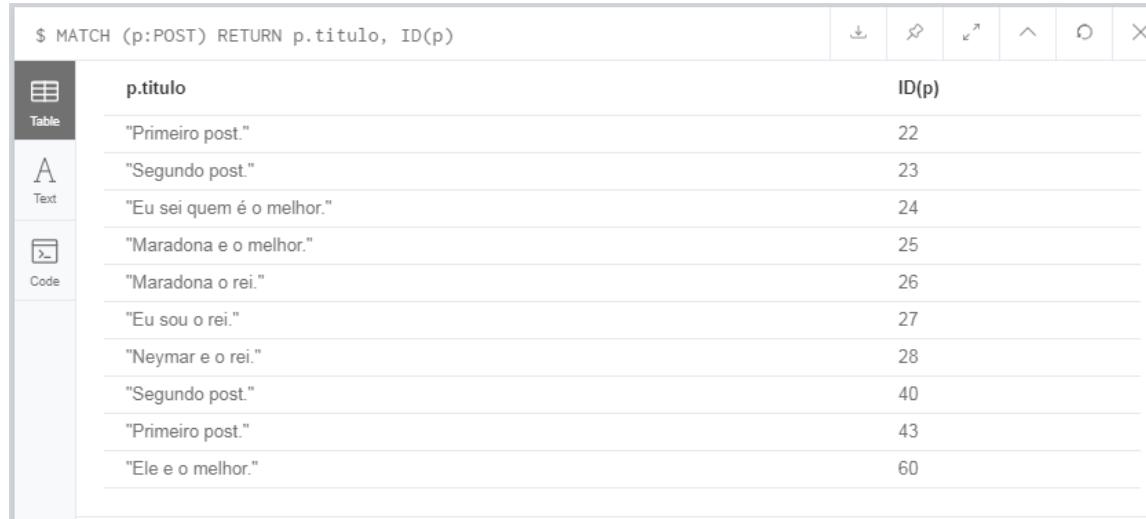
(19) FOLLOWS(9) CRIADO_POR(10)

Displaying 15 nodes, 19 relationships.

Grafos em redes sociais

- O Neo4j também possui um identificador único para cada nó.

```
MATCH (p:POST) RETURN p.titulo, ID(p)
```



The screenshot shows a Neo4j browser window with a query results table. On the left, there's a sidebar with icons for Table, Text, and Code, and a dropdown menu showing 'A'. The main area displays the following table:

\$ MATCH (p:POST) RETURN p.titulo, ID(p)	
p.titulo	ID(p)
"Primeiro post."	22
"Segundo post."	23
"Eu sei quem é o melhor."	24
"Maradona e o melhor."	25
"Maradona o rei."	26
"Eu sou o rei."	27
"Neymar e o rei."	28
"Segundo post."	40
"Primeiro post."	43
"Ele e o melhor."	60

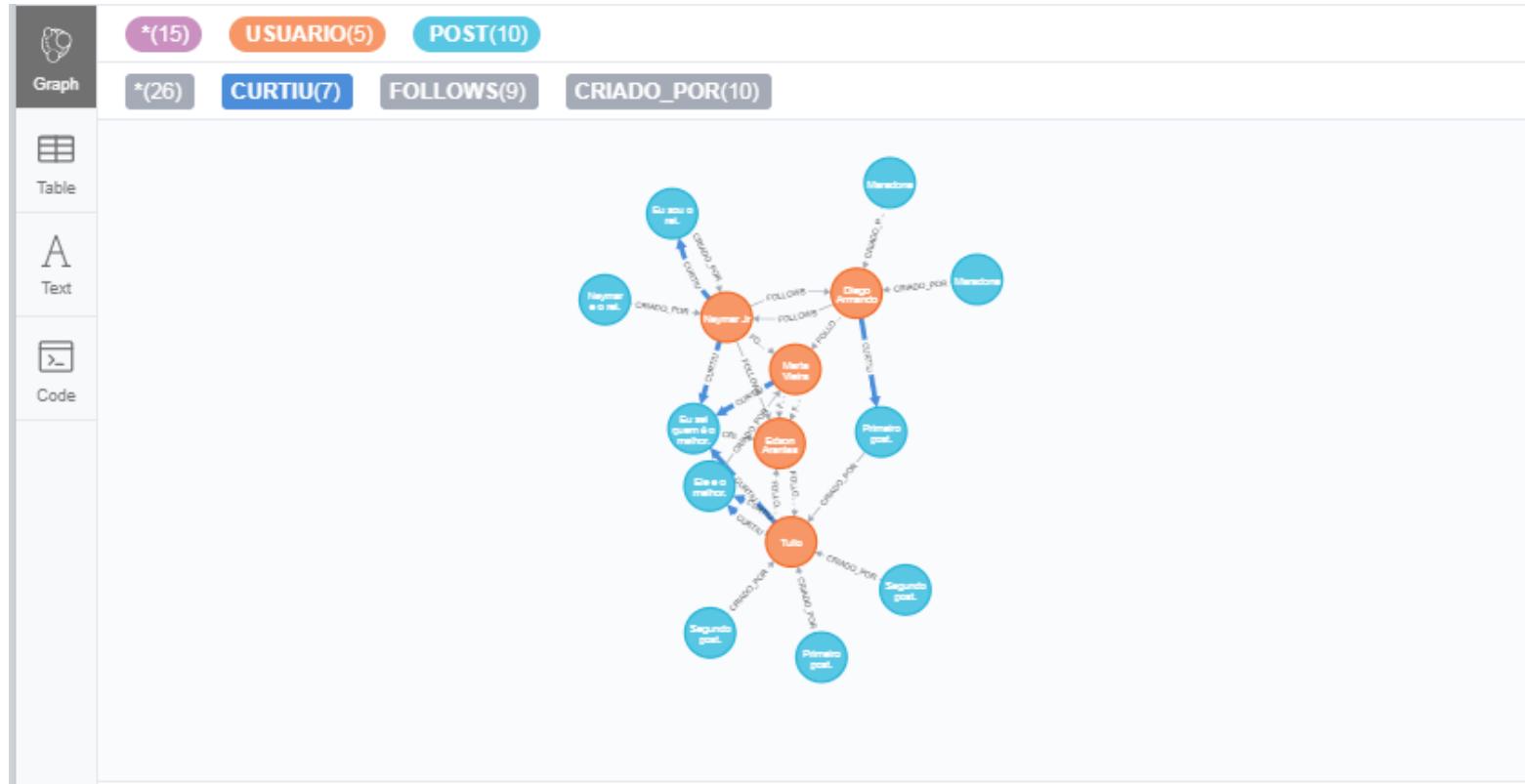
- Curtir um post:



```
MATCH (tulio:USUARIO {email: "tulio@example.com"}), (p:POST{titulo:"Ele e o melhor."}) CREATE (tulio)-[r:CURTIU]->(p) RETURN tulio,p.conteudo
```

```
MATCH (tulio:USUARIO {email: "tulio@example.com"}), (p:POST) WHERE ID(p)=60 CREATE (tulio)-[r:CURTIU]->(p) RETURN tulio,p.conteudo
```

Grafos em redes sociais



Grafos em redes sociais

- Obter todos os usuários:

```
MATCH (u:USUARIO) RETURN u{ .name, .email }
```

The screenshot shows a user interface for running MongoDB queries. On the left, there are tabs for 'Table' (selected), 'Text', and 'Code'. The main area contains a query: '\$ MATCH (u:USUARIO) RETURN u{ .name, .email }'. Below the query, three results are displayed as JSON objects:

- {
 "name": "Tulio",
 "email": "tulio@example.com"
}
- {
 "name": "Neymar Jr",
 "email": "neymar@example.com"
}
- {
 "name": "Edson Arantes",
 "email": "pele@example.com"
}

Grafos em redes sociais

- Obter o usuário e os posts relacionados.

```
MATCH (u:USUARIO {email:"tulio@example.com"})-<[:CRIADO_POR]-(p:POST)
RETURN u{ .nome, .email, posts: collect(p.titulo) }
```

The screenshot shows a Neo4j browser window with the following details:

- Query Bar:** \$ MATCH (u:USUARIO {email:"tulio@example.com"})-<[:CRIADO_POR]-(p:POST) ...
- Result Table:** A single row labeled "u" is displayed. The "posts" column contains the following JSON object:

```
{
  "posts": [
    "Segundo post.",
    "Segundo post.",
    "Primeiro post.",
    "Primeiro post."
  ],
  "nome": "Tulio",
  "email": "tulio@example.com"
}
```
- Toolbar:** Standard Neo4j browser icons for saving, sharing, and navigating.
- Bottom Status:** Started streaming 1 records after 2 ms and completed after 3 ms.

Grafos em redes sociais

- Obter recomendação de posts:

```
MATCH (u:USUARIO {email:"pele@example.com"})-[:FOLLOWERS]-> (:USUARIO)--(p:POST) RETURN p{id: ID(p), .titulo,.conteudo }
```

The screenshot shows the Neo4j browser interface with the following details:

- Left Sidebar:** Contains tabs for "Table", "Text", and "Code". The "Text" tab is selected.
- Top Bar:** Shows the query: \$ MATCH (u:USUARIO {email:"pele@example.com"})-[:FOLLOWERS]-> (:USUARIO)--(p:POST) RETURN p{id: ID(p), .titulo,.conteudo }.
- Result Area:** Displays three JSON results, each representing a post (p) returned by the query:

 - Post 1:

```
{  
    "titulo": "Ele é o melhor.",  
    "conteudo": "Pelé é o melhor de  
todos.",  
    "id": 60  
}
```
 - Post 2:

```
{  
    "titulo": "Eu sei quem é o melhor.",  
    "conteudo": "Sou o melhor de  
todos.",  
    "id": 24  
}
```
 - Post 3:

```
{  
    "titulo": "Ele é o melhor.",  
    "conteudo": "...",  
    "id": 10  
}
```

- Bottom Status:** Shows the message: Started streaming 7 records after 5 ms and completed after 32 ms.

Grafos em redes sociais

```
MATCH (u:USUARIO {email:"pele@example.com"})-[:FOLLOWERS]->  
(uf:USUARIO) RETURN uf.nome
```



ef.nome

"Tulio"

```
MATCH (u:USUARIO {email:"tulio@example.com"})-[:FOLLOWERS]->  
(uf:USUARIO) RETURN uf.nome
```



uf.nome

"Edson Arantes"

Grafos em redes sociais

```
MATCH (u:USUARIO {email:"tulio@example.com"})-[:CURTIU]->  
(p:POST) RETURN p.titulo
```

	Table
	Text

p.titulo

"Ele é o melhor."

"Eu sei quem é o melhor."

"Ele é o melhor."

```
MATCH (p:POST)- [:CRIADO_POR]-> (u:USUARIO  
{email:"tulio@example.com"})RETURN p.titulo
```

	Table
	Text

p.titulo

"Segundo post."

"Segundo post."

"Primeiro post."

"Primeiro post."

Conclusão

Neo4j e Cypher (Parte II).

■ Próxima aula

- Neo4j e Cypher (Parte III).

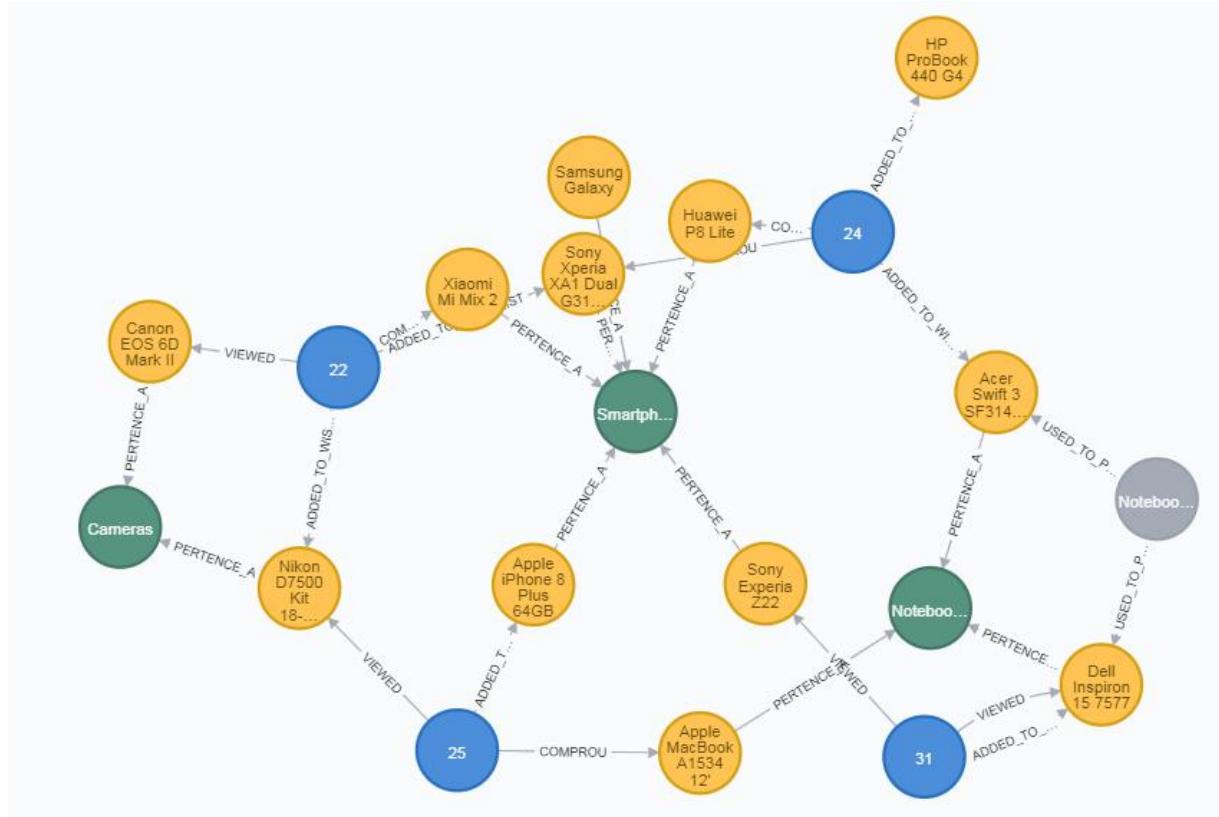


Aula 6.3.3. Neo4j e Cypher (Parte III)

Nesta aula

- Aplicação utilizando Neo4j e Cypher.

Recomendação de produtos



Recomendação de produtos



Conclusão

- ✓ Aplicação utilizando Neo4j e Cypher.

- Apache Spark Graph.

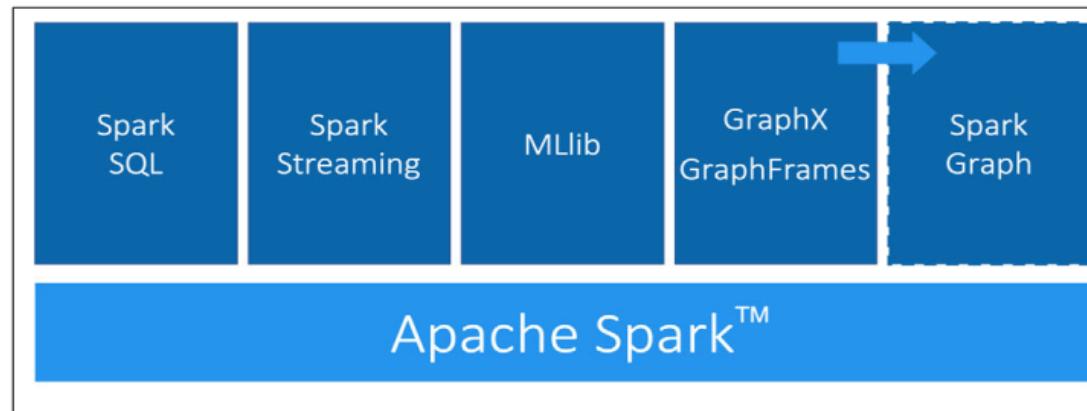


Aula 6.4.1. Apache Spark Graph (Parte I)

- Conhecendo o Apache Spark GraphX.

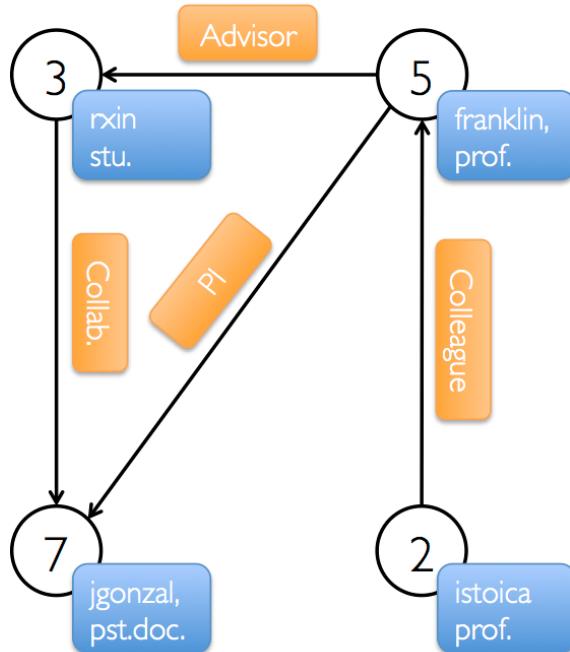
O que é o Spark Graph

- API do Apache Spark para computação paralela de grafos;
- Possui uma grande quantidade de algoritmos para processamento;
- Permite visualizar os dados como grafos ou coleções;
- Flexibilidade.



O que é o Spark Graph

Property Graph



Vertex Table

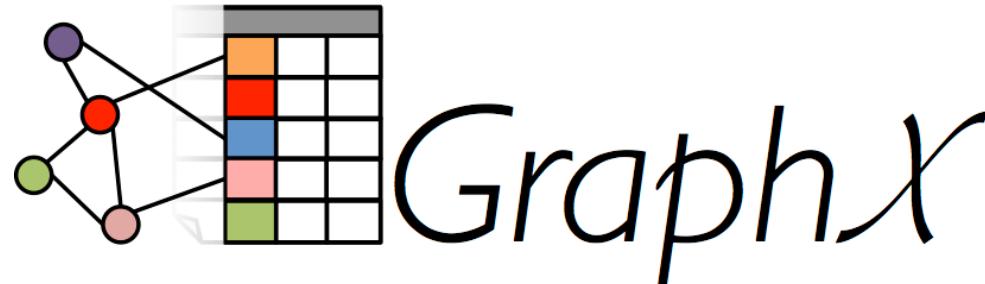
Id	Property (V)
3	(rxin, student)
7	(jgonzal, postdoc)
5	(franklin, professor)
2	(istoica, professor)

Edge Table

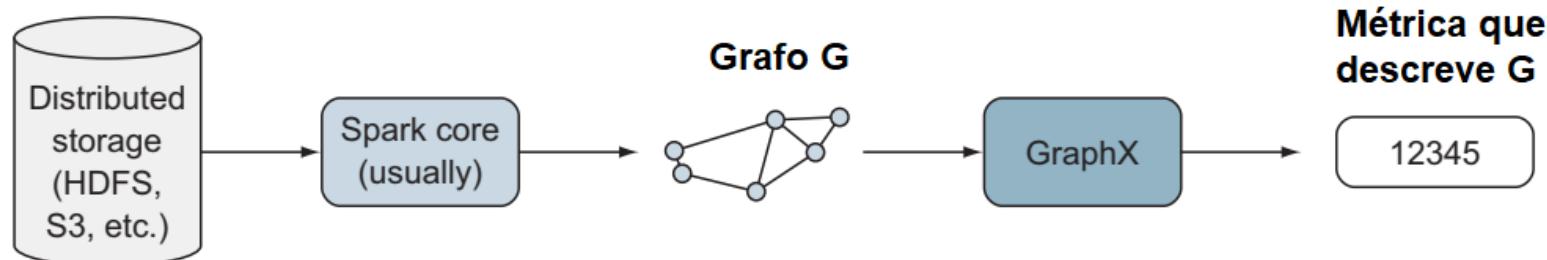
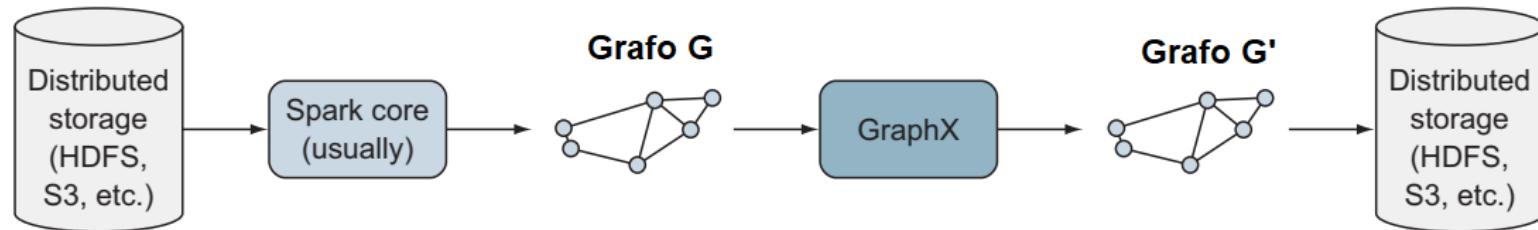
SrcId	DstId	Property (E)
3	7	Collaborator
5	3	Advisor
2	5	Colleague
5	7	PI

Como utilizar?

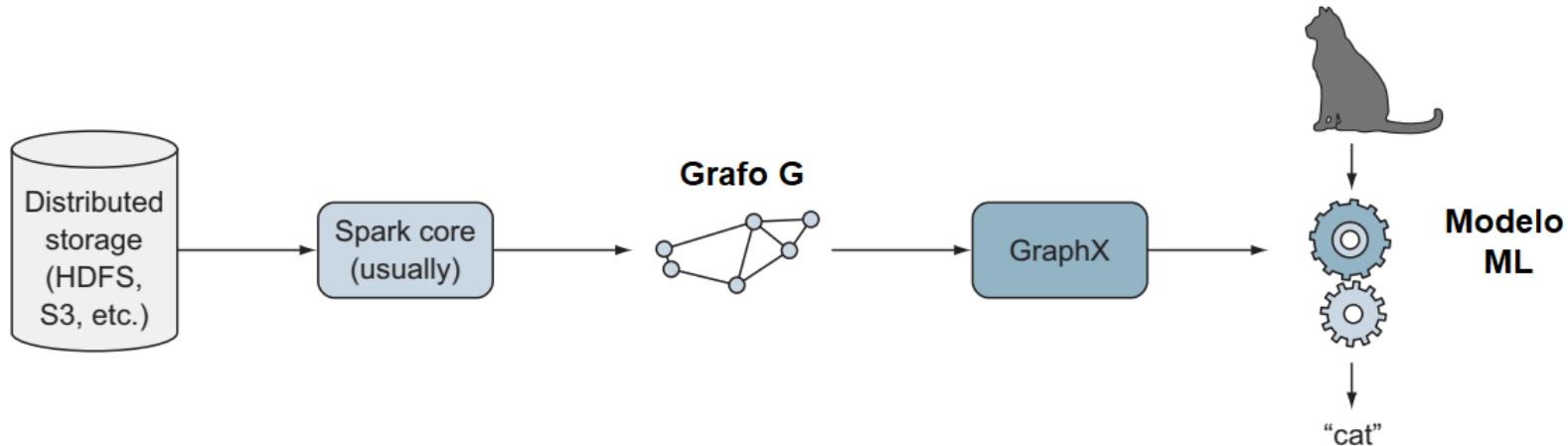
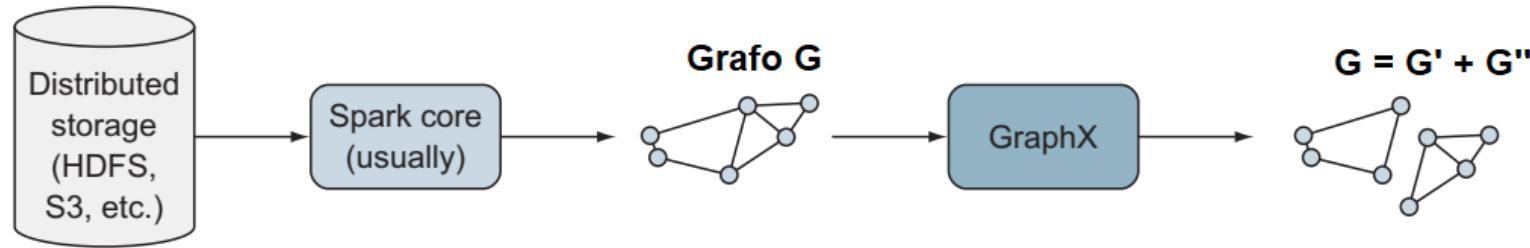
- Nós e relacionamentos podem vir de diferentes fontes (Parquet, JSON, CSV etc.).
- Consultas utilizam uma combinação entre API do PySpark e Spark SQL.



O que posso fazer com o Spark GraphX?



O que posso fazer com o Spark GraphX?



- ✓ Apache Spark Graph (Parte I).

■ Próxima aula

- Apache Spark Graph (Parte II).



Aula 6.4.2. Apache Spark Graph (Parte II)

- Aplicação em transporte.



**Let's
Practice!**

- ✓ Apache Spark Graph (Parte II).

■ Próxima aula

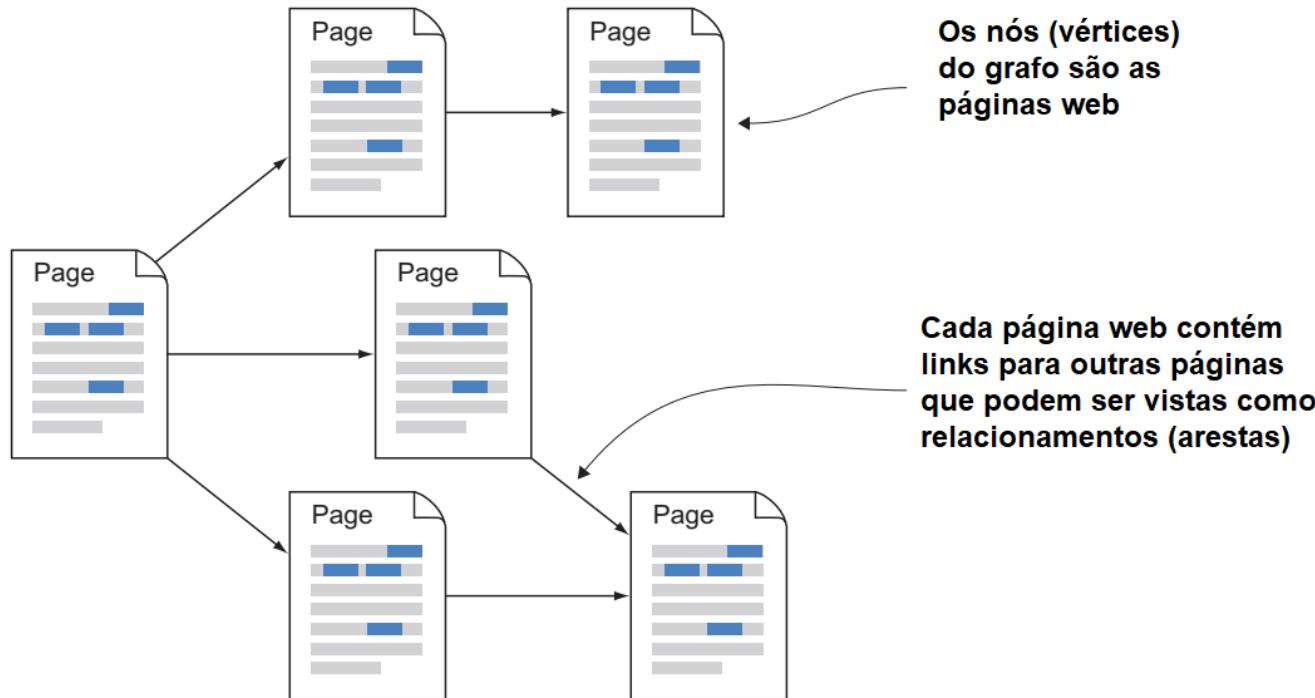
- Apache Spark Graph (Parte III).



Aula 6.4.3. Apache Spark Graph (Parte III)

- Aplicação utilizando o Apache Spark Graph para computar o PageRank.

PageRank

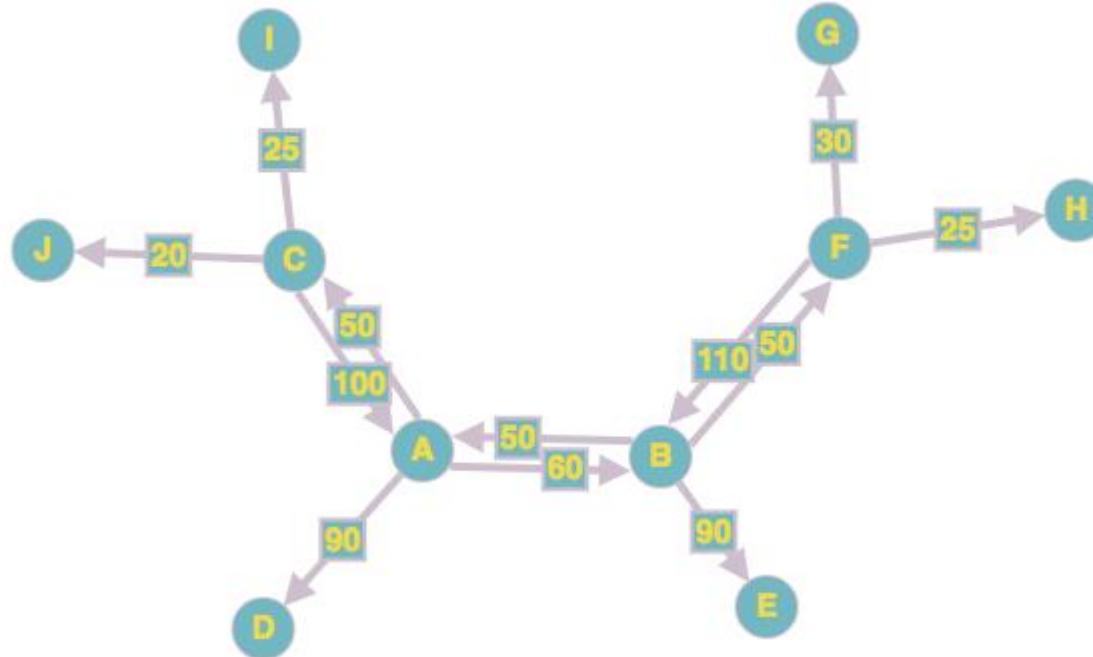


PageRank

EDGES		
src	dst	relationship
A	B	60
B	A	50
A	C	50
C	A	100
A	D	90
C	I	25
C	J	20
B	F	50
F	B	110
F	G	30
F	H	25
B	E	90

NODES		
id	name	total_seconds
A	ARON	350
B	BILL	360
C	CLAIR	195
D	DANIEL	90
E	ERIC	90
F	FRANK	215
G	GRAHAM	30
H	HENRY	25
I	INNA	25
J	JEN	20

PageRank





Conclusão

- ✓ Aplicação utilizando Apache Spark Graph.