Curso 03: Administração de Dados Complexos em Larga Escala

-- Advanced Analytics com o Apache Mahout --

Prof. Jose Fernando Rodrigues Junior

Objetivo: apresentar a solução de Aprendizado de Máquina sobre dados em larga escala Apache Mahout

Níveis da Análise de Dados

- **Análise de dados básica**: contagens, somas, médias, máximo, mínimo, e ordenação;
- **Análise de dados estatística**: distribuição de dados, ajuste de modelo, teste de hipóteses, métricas, etc;
- **Análise de dados avançada**: aprendizado de máquina, classificação, regressão, recomendação, clusterização, etc;
- **Aprendizado de máquina avançado**: arquiteturas de redes neurais visando inteligência artificial.

Níveis da Análise de Dados

- **Análise de dados básica**: [illegible]s, médias, máximo, mínimo, e ord[illegible]
- **Análise de dados estatística**: distribuição de dados, ajuste de modelo, t[illegible]s, métricas, etc;
- **Análise de dados avançada**: aprendizado de máquina, classificação, regres[illegible]o, clusterização, etc;
- **Aprendizado de máquina avançado**: arquiteturas de redes neurais visando inteligê[illegible]

Curso 03/11 - DW/OLAP

Curso 02/05

Curso 02/03 - Mahout

Curso 05/07/08/10

Machine Learning in a nutshell

- **Aprendizado de Máquina**: um ramo das técnicas de inteligência artificial que fornece ferramentas que permitem aos computadores melhorar sua análise com base em eventos anteriores;

- Aproveitam os dados históricos de **tentativas anteriores** de resolver uma tarefa para **melhorar o desempenho** de **futuras tentativas** de tarefas semelhantes;

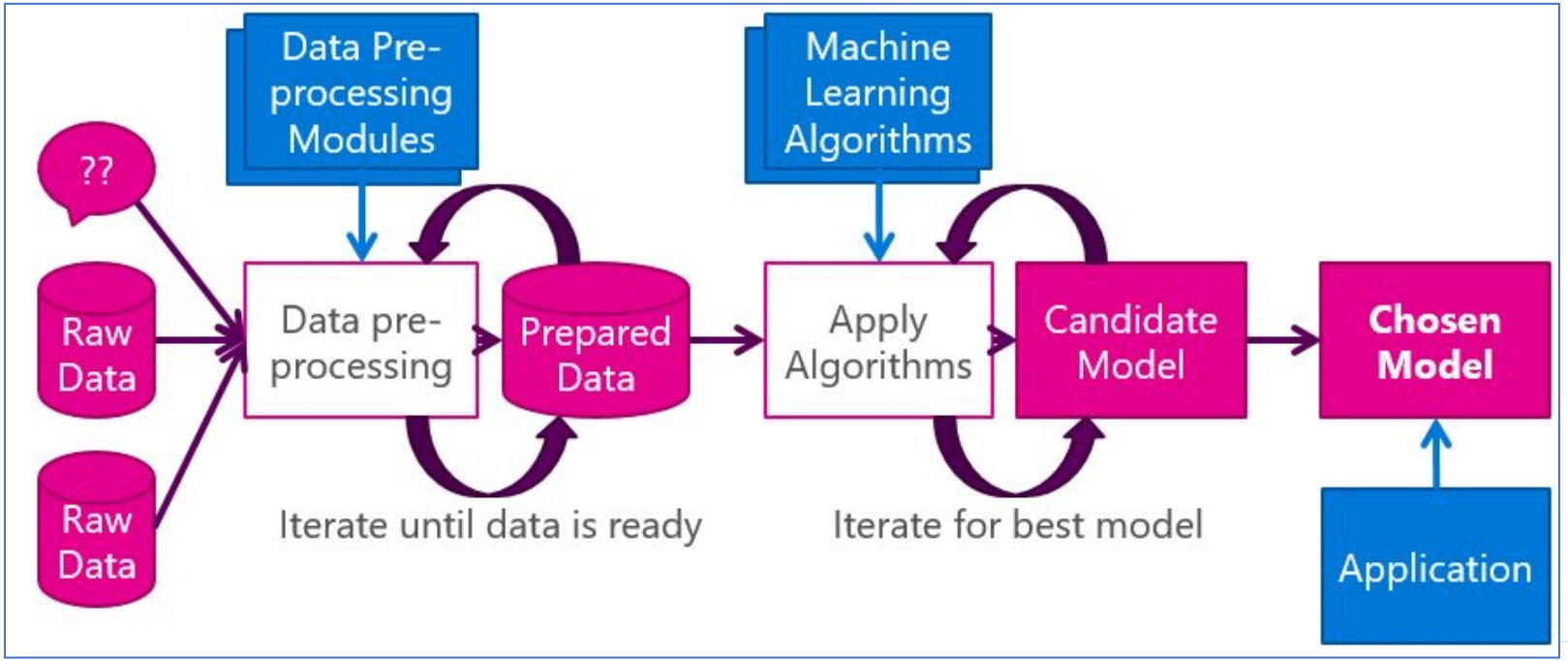
- Bibliotecas Mahout são implementadas em **Java MapReduce** e **executadas em seu cluster** como coleções de trabalhos MapReduce.

Machine Learning in a nutshell

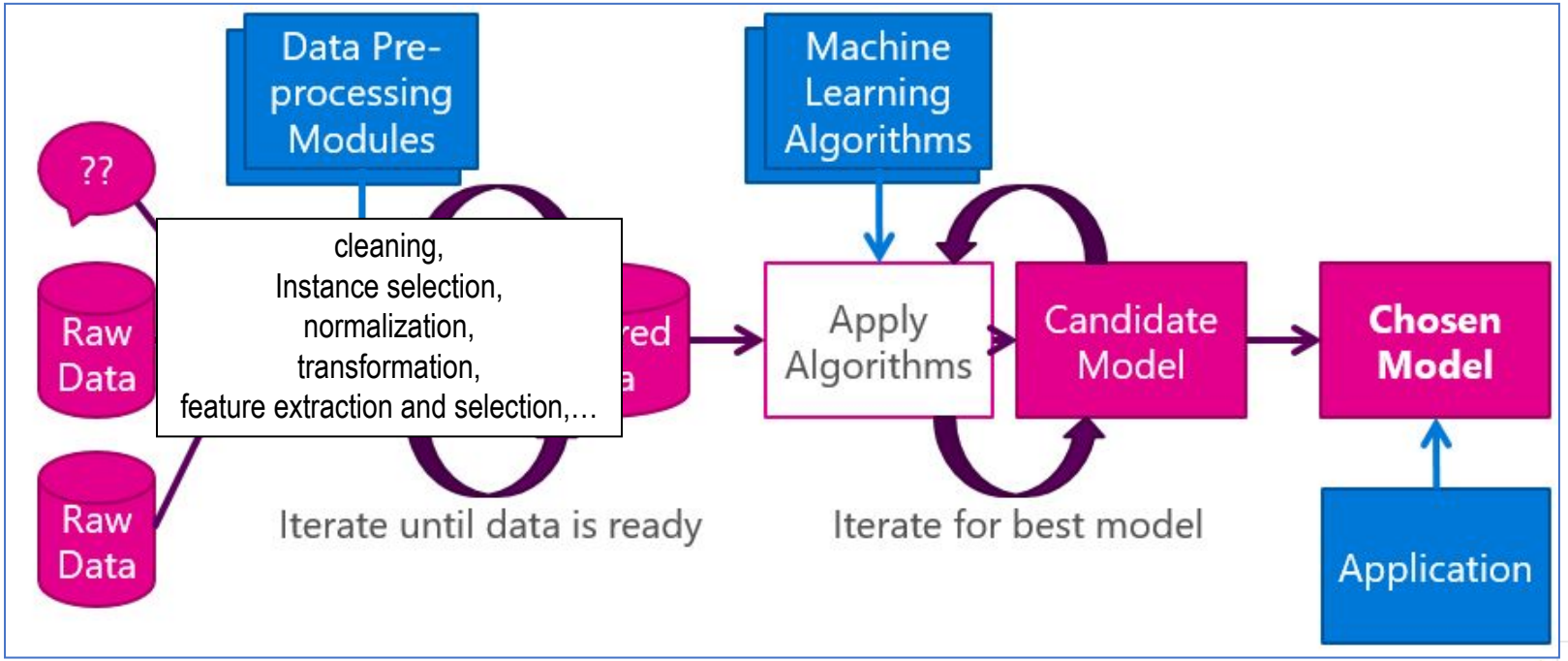
Tarefas comuns de Aprendizado de Máquina:

- **Recomendação** de produtos/amigos/pares;
- **Classificação** em tipos/grupos/posições;
- Encontrar **elementos semelhantes**;
- Encontrar **associações** em comportamentos/ações;
- Encontrar **assunto chave** em textos;
- Detectar **anomalias/fraudes/exceções**;
- **Ranquear resultados** de busca;
- Entre outras.

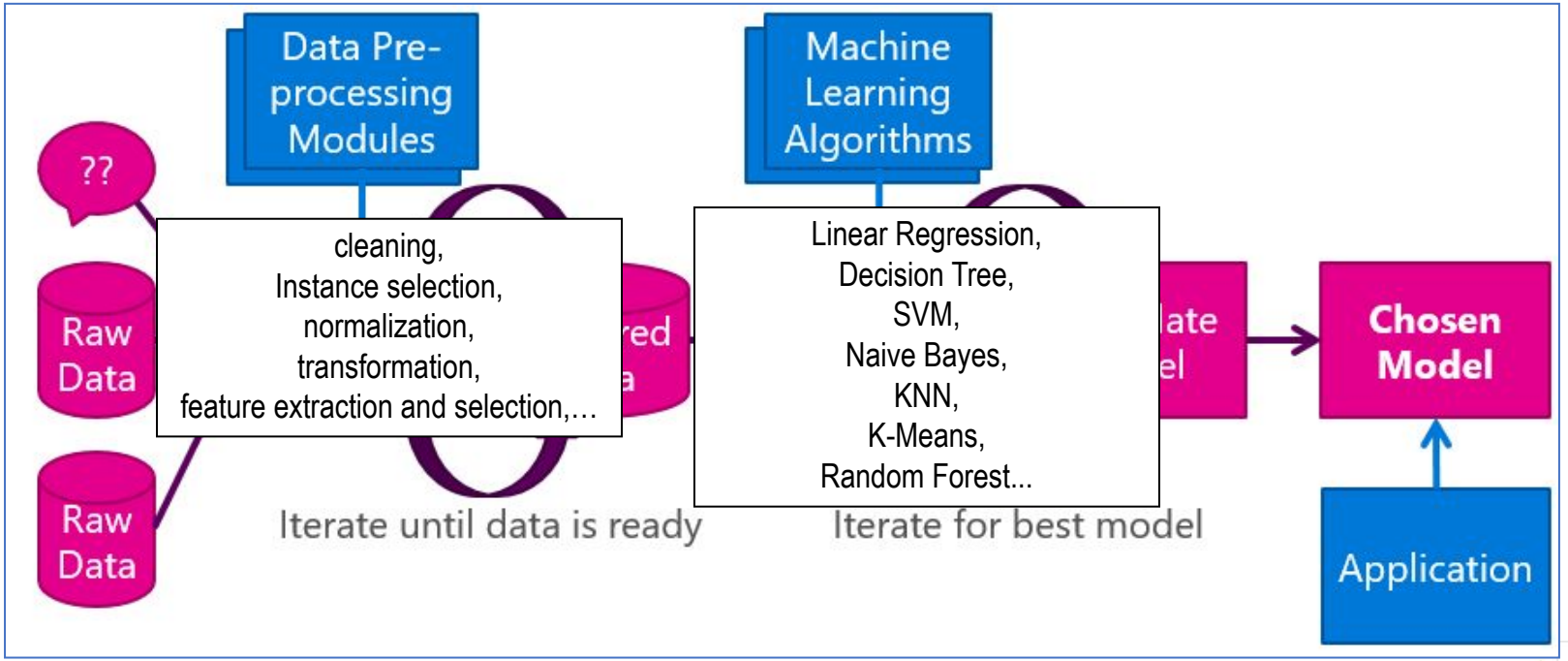
Machine Learning in a nutshell



Machine Learning in a nutshell



Machine Learning in a nutshell



Apache Mahout

-O Apache Mahout é uma **API Java** de algoritmos de *Machine Learning*

-Tem como **características**:

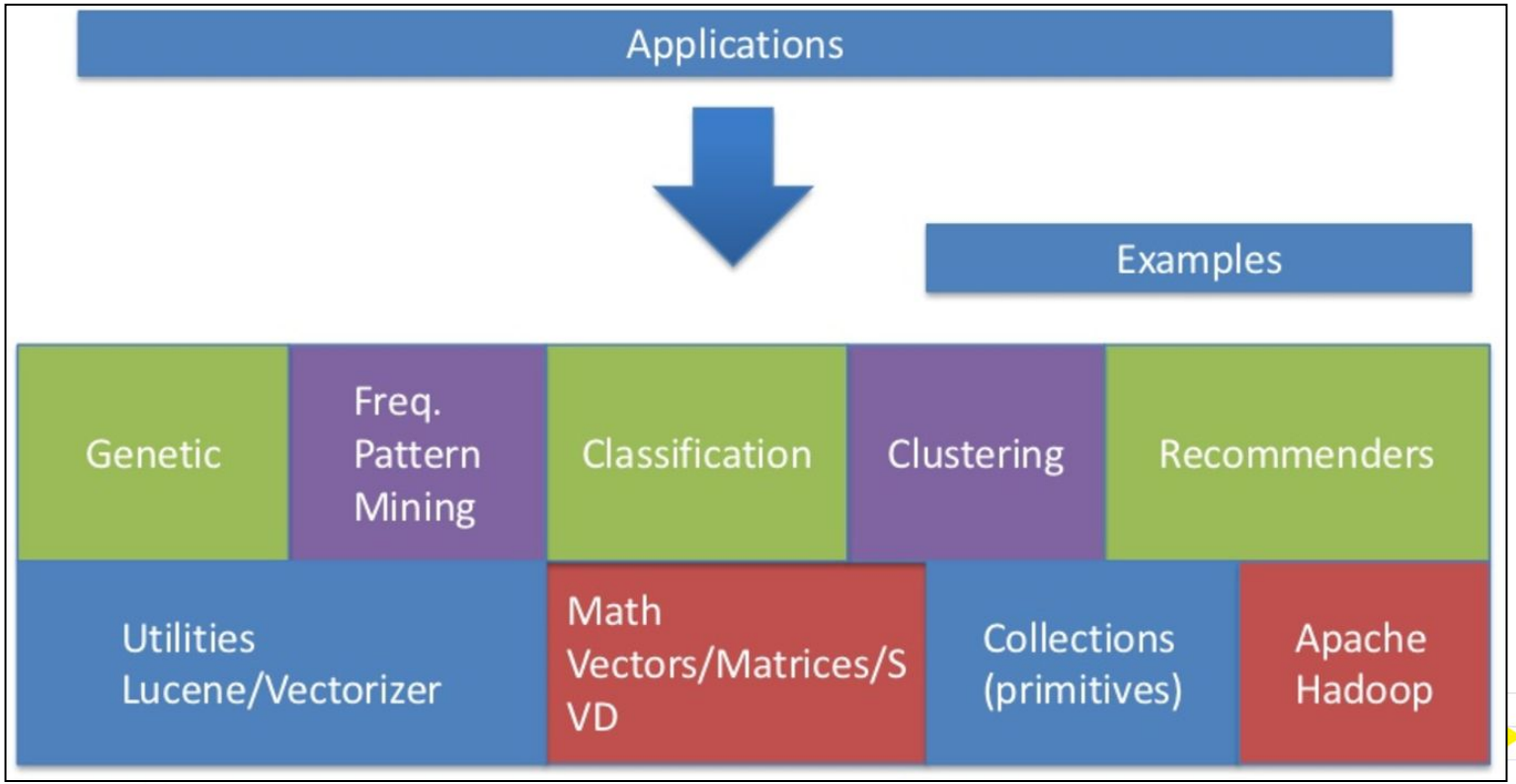
-escalabilidade;

-documentação extensa;

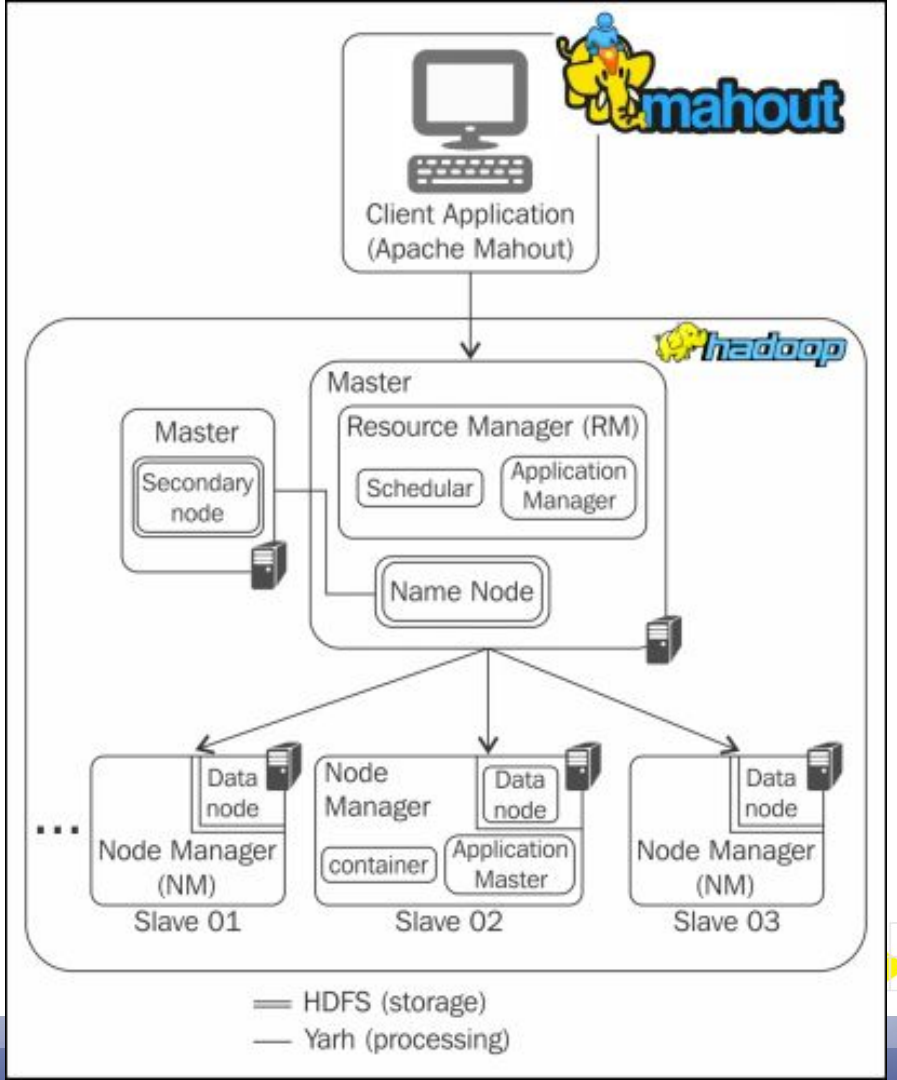
-uso aplicado.

-**Tipos de algoritmos**: detecção de agrupamento, classificação, alg. genéticos, recomendação, regras de associação (*frequent pattern matching*), entre outros.

Apache Mahout
Aplicações, algoritmos, bibliotecas, e arcabouço distribuído



Apache Mahout sobre a infraestrutura distribuída Hadoop



Apache Mahout

- Os algoritmos do Mahout **não necessariamente funcionam em MapReduce**;
- Isso, pois muitos algoritmos de Machine Learning **não são paralelizáveis** (pelo menos não diretamente);
- A partir de 2014, o projeto passou a priorizar Spark e generalidade com relação à infraestrutura de computação:
  - funcionamento independente do ecossistema Hadoop;
  - integração ao projeto H2O.ai (distributed in-memory ML), Apache Flink (stream data), além do Hadoop e do Spark;
  - suporte a programação de novos algoritmos de modo generalizado.

Apache Mahout Algoritmos

- **Recomendação**: item-based Collaborative Filtering, Matrix Factorization with Alternating Least Squares, e Matrix Factorization with Alternating Least Squares on Implicit Feedback;
- **Classification**: Naive Bayes, Complementary Naive Bayes, e Random Forest;
- **Clustering**: Canopy Clustering, k-Means Clustering, Fuzzy k-Means, Streaming k-Means, e Spectral Clustering;
- **Dimensionality Reduction**: Lanczos Algorithm, Stochastic SVD, e Principal Component Analysis;
- **Regression**: Ordinary Least Squares, Ridge Regression
- **Topic Models**: Latent Dirichlet Allocation;
- **Miscellaneous**: Frequent Pattern Matching, RowSimilarityJob, ConcatMatrices, e Colocations.

Apache Mahout Abrangência

O projeto Apache Mahout é imenso e pode ser usado a partir de diferentes perspectivas:

- usuário Hadoop de algoritmos existentes;
- usuário Spark de algoritmos existentes;
- desenvolvedor de novos algoritmos;
- integração com arcabouços como o H20;
- processamento em um único nó de processamento;
- entre outras possibilidades.

Aplicações de Aprendizado de Máquina Mahout

Clusterização	Classificação	Recomendação

Utilitários Mahout

Mahout Math

Hadoop Distributed File System (HDFS)

Fazendo Recomendações com o Apache Mahout Taste

Apache Mahout Taste

- Como visto, o Apache Mahout contém **dezenas de algoritmos** de Aprendizado de Máquina;
- Um de seus subprojetos de maior sucesso é o **Apache Mahout Taste**, um mecanismo de filtragem colaborativa flexível e rápido para Java;
- **Filtragem colaborativa**: um mecanismo que processa as preferências dos usuários para os itens ("gostos") e retorna as **preferências estimadas** para outros itens;
- **Exemplos**: próximos filmes da NetFlix; próximos livros da Amazon; próximos produtos no eBay.

⇒ Para saber mais: Mahout – Taste :: Part 1 – Introduction

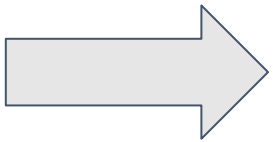
Apache Mahout Taste

Para fazer recomendações usando o Taste:

1) Dados, ou DataModel

Neste exemplo:

- **Usuário** refere-se a usuários de um serviço, por exemplo, de streaming;
- **Filme** refere-se ao identificador de um filme no catálogo de filmes;
- **Nota** refere-se ao quanto o usuário gostou de um filme, considerando notas entre 1 e 5.



Usuário	Filme	Nota
1	101	5
1	102	3
1	103	2,5
2	101	2
2	102	2,5
2	103	5
2	104	2
3	101	2,5
3	104	4
3	105	4,5
3	107	5
4	101	5
4	103	3
4	104	4,5
4	106	4

Apache Mahout Taste

Para fazer recomendações usando o Taste:

2) Uma medida de similaridade entre usuários e/ou items

Neste exemplo: a medida de similaridade **Coeficiente de Correlação de Pearson** (há muitas outras).

Suposição: se dois usuários são parecidos, supõe-se que os mesmos itens podem ser de interesse de ambos.

Similarity between users (2/3)

Pearson Correlation Coefficient

$$sim(u_a, u_b) = \frac{\sum_{i \in I}(r_{u_a,i} - \overline{r_{u_a}})(r_{u_b,i} - \overline{r_{u_b}})}{\sqrt{\sum_{i \in I}(r_{u_a,i} - \overline{r_{u_a}})^2}\sqrt{\sum_{i \in I}(r_{u_b,i} - \overline{r_{u_b}})^2}}$$

	Item1	Item2	Item3	Item4
Alice	5	3	4	4
User1	3	1	2	3
User2	4	3	4	3
User3	3	3	1	5
User4	1	5	5	2

sim=0.71

sim=-0.79

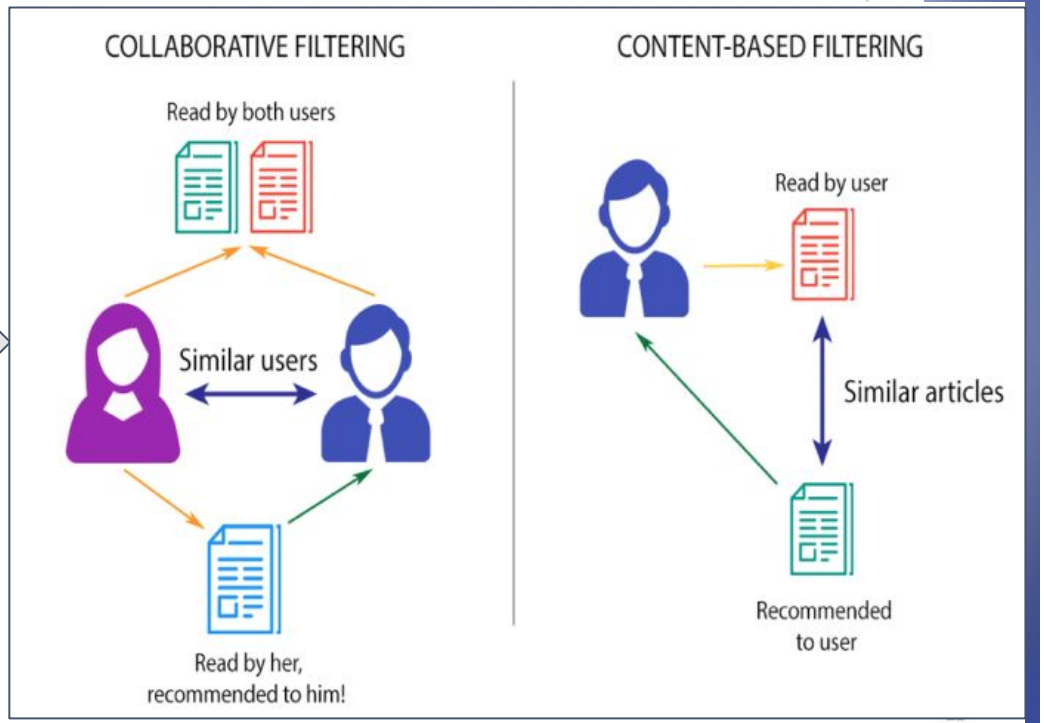
Apache Mahout Taste

Para fazer recomendações usando o Taste:

2) Uma medida de similaridade entre usuários e/ou items

Neste exemplo: a medida de similaridade Coeficiente de Correlação de Pearson (há muitas outras).

Suposição: se dois usuários são parecidos, supõe-se que os mesmos itens podem ser de interesse de ambos.



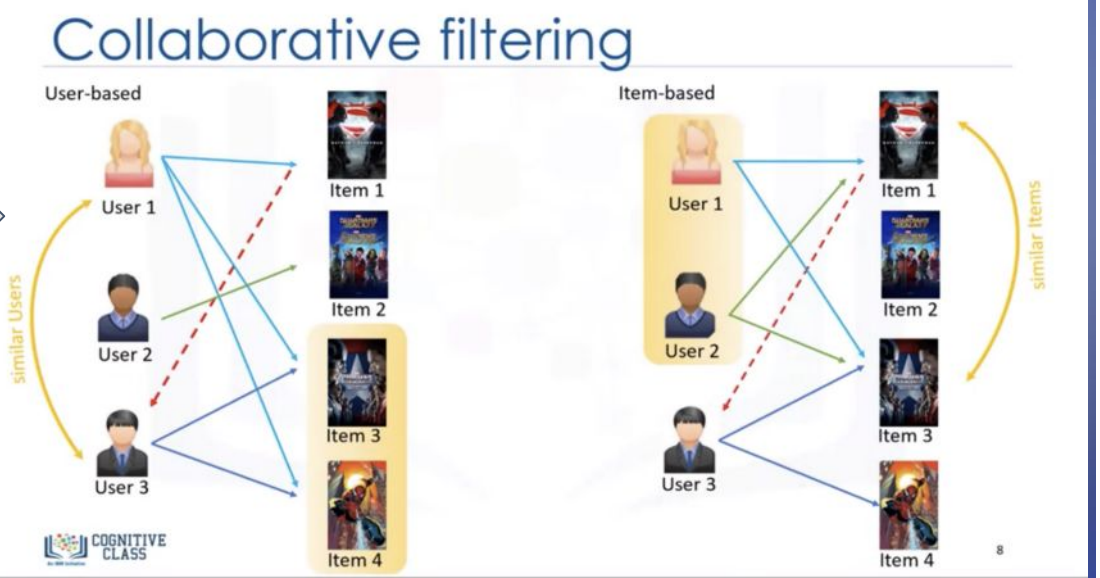
Apache Mahout Taste

Para fazer recomendações usando o Taste:

3) Uma vizinhança: quantos usuários parecidos serão considerados

⇒ Para saber mais: Collaborative Filtering

Neste exemplo, serão considerados os dois usuários mais parecidos ao usuário 3 com o objetivo de se computar as recomendações.



Apache Mahout Taste

Para fazer recomendações u

3) Uma viz

considera

Neste

conside

mais

com

computar as rec

Note que em um mecanismo como este, quanto mais dados, mais precisa e abrangente será a recomendação. É daí que vem o poder do Big Data.

Apache Mahout Taste

Exercício *Hands on* – criar um sistema de recomendação no Mahout:

Apache Mahout - Creating a User-Based Recommender in 5 minutes
https://github.com/felrukby/com.rukbysoft.examples.mahout

Passos:
1) Logar no GitHub;
2) Fazer download do projeto em https://github.com/felrukby/com.rukbysoft.examples.mahout;
3) Instalar o ambiente integrado de desenvolvimento Apache Netbeans;
4) Abrir o projeto com o Netbeans;
5) Testar as funcionalidades to Apache Mahout Taste.

Conclusões

-O ecossistema Hadoop oferece soluções com **custo reduzido** para o **processamento analítico** em larga escala;

-As ferramentas estão em **constante mudança**;

-Há muito espaço para **profissionais** em *Data Science* com **bom salário**; todavia, *Data Science* requer **estudo e dedicação**

Conclusões

-BigData tem muito a explorar em termos de dados **não estruturados**, como texto;

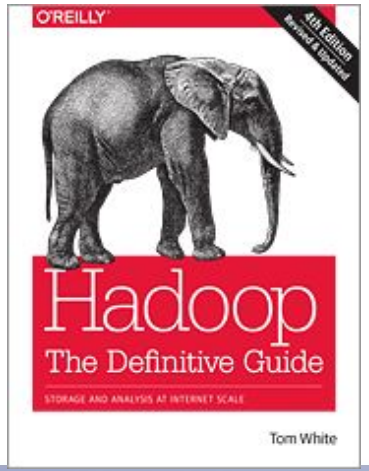
-O uso combinado de ***Deep Learning*** e processamento Big Data tem trazido ***breakthroughs*** na indústria;

-**Consolidação** do especialista em dados como um requisito de sobrevivência das corporações.

Referências

- Hadoop: The Definitive Guide, Storage and Analysis at Internet Scale, 4th Edition; By Tom White, O'Reilly Media, 2015

  http://shop.oreilly.com/product/0636920033448.do



Referências

- Analítica de Dados com Hadoop: Uma Introdução Para Cientistas de Dados, 1a. Edição; By Benjamin Bengfort, Jenny Kim. O'Reilly Media, 2016