Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Gabriel Yuri Silva Ribeiro

Plataformas de crawling em ambiente de computação em nuvem: Uma perspectiva prática

Curitiba
2022

Gabriel Yuri Silva Ribeiro

Plataformas de crawling em ambiente de computação em nuvem: Uma perspectiva prática

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Wagner Hugo Bonat

Curitiba
2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Plataformas de crawling em ambiente de computação em nuvem: Uma perspectiva prática

Gabriel Yuri Silva Ribeiro[1]
Pedro Augusto de Lima e Silva[2]
Luis C. E. Bona[3]

Abstract

Este trabalho tem como enfoque apresentar uma solução de engenharia de dados para um problema muito comum e ainda desafiador no mundo da tecnologia: A necessidade de obter dados de uma página web de maneira sistêmica.

Vários passos precisam ser executados de forma a desempenhar esta tarefa: Acessar o site de maneira automatizada, rastrear os elementos de interesse, condensar suas informações, salvar os dados de maneira segura e tratá-los para que tragam valor no fim do processo.

Desta forma, será apresentada uma solução que aborda principalmente três tópicos essenciais para uma solução moderna: Construção de um pipeline de dados orquestrado, infraestrutura como código para computação em cloud, e por fim, Integração contínua / deploy contínuo (CI/CD).

Seja pelo fato de precisarmos desempenhar uma série de tarefas em sequência com interdependência entre elas, manter a robustez de crawling em um site que pode (e irá) mudar ao longo do tempo ou realizar a mesma tarefa de maneira consistente dia após dia, se vê a necessidade de utilizar ferramentas de computação em nuvem, que nos garantem alta disponibilidade de recursos e nos propiciam flexibilidade para criar soluções.

Palavras-chave: Computação em nuvem, engenharia de dados, pipelines, crawling

Abstract

This paper focuses on presenting a data engineering solution to a very common and still challenging problem in the technology world: The necessity to fetch data from a web page in a systemic way.

Several steps need to be performed in order to accomplish this task: visiting the site in an automated way, tracking down the elements of interest, condensing its information to then save the data in a secure manner, to finally be able to process it in order to bring value at the end of the pipeline.

With this in mind, a solution will be presented that addresses three essential topics for a modern solution: Building an orchestrated data pipeline, infrastructure as code for cloud computing, and finally, Continuous integration / Continuous deployment (CI/CD).

Whether it is the fact that we need to perform a series of tasks in sequence with interdependency between them, robustness of crawling a site that can (and will) change over time, or performing change over time or perform the same task consistently day after task consistently day after day, there is a need to use cloud computing cloud computing tools, that guarantee high availability of resources and give us the flexibility to flexibility to create solutions.

Keywords: *Cloud computing, data engineering, pipelines, crawling*

1 Introdução

A crescente demanda por análises e modelos estatísticos tem evidenciado muito o protagonismo de estatísticos, analistas de dados e cientistas de dados. Mas para que toda essa cadeia de consumidores prospere, existe a urgente necessidade por dados confiáveis, facilmente trabalháveis e relevantes.

Para suprir essa demanda surgiu o engenheiro de dados. Essa profissão se compromete de arquitetar, construir e monitorar pipelines de dados, que têm como função atualizar datalakes ou datawarehouses com informações frescas para que os analistas montem seus dashboards e cientistas montem seus modelos e análises.

As ferramentas utilizadas por engenheiros de dados devem prezar não só pela eficiência em custos e tempos de execução, mas também de tudo confiabilidade e robustez. O fato de uma cadeia toda de consumidores basear suas análises, estudos, modelos e decisões em cima dos dados cria uma enorme pressão pela sua qualidade.

Atualmente, uma empresa pode se alimentar de diversas fontes de dados - sistemas internos da empresa, portais públicos, consultorias especializadas em inteligência, websites, Application Programming Interfaces (APIs) e muitos outros. Cada uma dessas

[1]Aluno do programa de Especialização em Data Science & Big Data, gyribeiro2014@gmail.com.
[2]Analista de dados da Gamersclub, pdrals16@gmail.com.
[3]Professor do Departamento de Informática - DInf/UFPR.

fontes pode fornecer dados em inúmeras formas e tamanhos, como .CSV, .JSON, .PARQUET, .AVRO e até .JPG.

Plataformas atuais de dados são capazes não só de lidar com dados tabulados e bem comportados, mas também com dados nas suas diversas formas, sendo também capaz de tratar dados de diversas fontes de informação em tabelas claras e bem estruturadas, capazes de alimentar elatórios, alertas, modelos e estudos.

A vida de analistas e cientistas é muito mais difícil quando não há dados frescos do dia anterior para trabalhar

Falar sobre como a literatura aborda esse tema em aplicações com intensidade de dados

- Parágrafo sobre robustez (reliability)
- Parágrafo sobre escalabilidade
- Parágrafo sobre capacidade de Manutenção

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

- Essa é uma lista de tópicos apenas para ilustrar como construir.
- Coloquei apenas do itens que é o suficiente para você entender.
  - Mas se a lista for hierárquica, então é só repetir o ambiente.
  - Bem fácil.

Fique atento aos seguintes símbolos

1. Para graus: as temperaturas usadas foram 25°C e 40°C.
2. Para números cardinais: as avaliações foram feitas no 7º e 15º dias.

Para referências bibliográficas, use `citet` para citação direta no texto e `citep` para citação ao final de parágrafos. Confira os exemplos.

Segundo **?**, blá blá blá. Tal coisa e coisa tal (**?**). Por outro lado, **?** indica que blá blá blá. Estudos dessa natureza já foram relatados (**??**).

2 Materiais e Métodos

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

2.1 O conjunto de dados

Para referências bibliográficas Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Table 1: Dicionário do conjunto de dados.

Variável	Descrição
Renda	(contínua) Renda mensal do cliente, em reais.
E.g.	R$ 10000, R$ 4500.
Dependentes	(discreta) Número de dependentes.
E.g.	0, 2, 5.
Ecivil	(categórica) Estado civil do cliente.
E.g.	casado, solteiro, divorciado, etc.

2.2 Limpeza e preparo dos dados

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

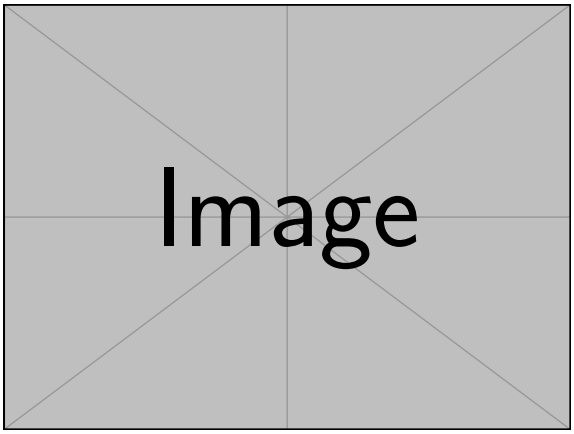

Figure 1: Legenda da figura. Exemplo: fluxograma de preparo dos dados considerando desde a extração do banco de dados, limpeza e imputação para emprego dos mesmos nos modelos.

2.3 Modelos empregados

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

2.3.1 Modelo A

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

2.3.2 Modelo B

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

2.4 Avaliação

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

2.5 Métricas

A *acurácia* dá a proporção de predições corretas.

$$\text{Acurácia: } A = \frac{TP + TN}{TP + FP + TN + FN}.$$

A *sensibilidade* mede a força do modelo prever um resultado positivo.

$$\text{Sensibilidade: } R = \frac{TP}{TP + FN}.$$

A *especificidade* mede a força do modelo prever um resultado negativo.

$$\text{Especificidade: } E = \frac{TN}{TN + FP}.$$

A *exatidão* mede a precisão de um resultado previsto como positivo.

$$\text{Exatidão: } P = \frac{TP}{TP + FP}.$$

O *F1* é a média harmônica entre precisão e a sensibilidade.

$$\text{F1: } F = 2 \cdot \frac{P \cdot R}{P + R}.$$

3 Resultados e Discussões

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

3.1 Ajuste dos modelos

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.

Table 2: Hiperparâmetros ajustados e tempo de execução para o ajuste dos modelos aos dados.

Modelo	Hiperparâmetros	Tempo (min)
SVR	39 vetores de suporte.	5
Random forest	$p = 10$ variáveis por árvore $m = 300$ árvores treinadas	15
Reg. Log. LASSO	$\alpha = 0.001$	1

3.2 Medidas de performance

Nulla in ipsum. Praesent eros nulla, congue vitae, euismod ut, commodo a, wisi. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nonummy magna non leo. Sed felis erat, ullamcorper in, dictum non, ultricies ut, lectus. Proin vel arcu a odio lobortis euismod. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Proin ut est. Aliquam odio. Pellentesque massa turpis, cursus eu, euismod nec, tempor congue, nulla. Duis viverra gravida mauris. Cras tincidunt. Curabitur eros ligula, varius ut, pulvinar in, cursus faucibus, augue.

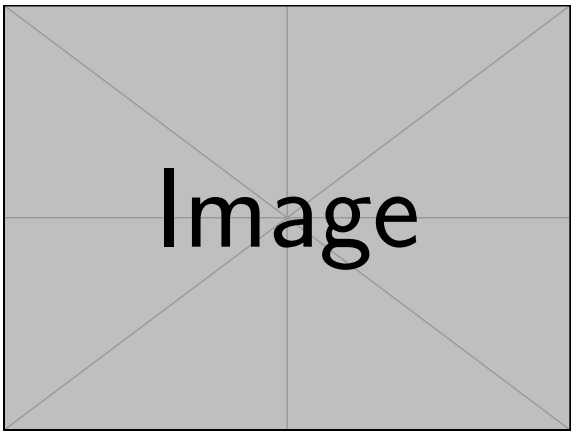

Figure 2: Legenda da figura. Exemplo: acurácia no conjunto de treino para os modelos treinados.

4 Conclusões

Nulla mattis luctus nulla. Duis commodo velit at leo. Aliquam vulputate magna et leo. Nam vestibulum ullamcorper leo. Vestibulum condimentum rutrum mauris. Donec id mauris. Morbi molestie justo et pede. Vivamus eget turpis sed nisl cursus tempor. Curabitur mollis sapien condimentum nunc. In wisi nisl, malesuada at, dignissim sit amet, lobortis in, odio. Aenean consequat arcu a ante. Pellentesque porta elit sit amet orci. Etiam at turpis nec elit ultricies imperdiet. Nulla facilisi. In hac habitasse platea dictumst. Suspendisse viverra aliquam risus. Nullam pede justo, molestie nonummy, scelerisque eu, facilisis vel, arcu.

Agradecimentos

Curabitur tellus magna, porttitor a, commodo a, commodo in, tortor. Donec interdum. Praesent scelerisque. Maecenas posuere sodales odio. Vivamus metus lacus, varius quis, imperdiet quis, rhoncus a, turpis. Etiam ligula arcu, elementum a, venenatis quis, sollicitudin sed, metus. Donec nunc pede, tincidunt in, venenatis vitae, faucibus vel, nibh. Pellentesque wisi. Nullam malesuada. Morbi ut tellus ut pede tincidunt porta. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam congue neque id dolor.