Artigo
Gerson Saito · Fev. 21, 2022 6min de leitura

Python e InterSystems IRIS© um Casamento Perfeito para Machine Learning.

A plataforma de dados InterSystems IRIS © agora apresenta suporte nativo completo para Python, liberando o poder da linguagem de programação mais popular do mundo para criar aplicativos de missão crítica com uso intensivo de dados.

https://www.intersystems.com/news-events/news/news-item/embedded-python-intersystems-iris-data-platform/

Essa decisão estratégica leva em consideração vários fatores:

- A grande escala no uso de Python dentro das corporações
- Python passou ser a um padrão de fato para desenvolvimento de modelos de Machine Learning (ML)
- A grande oferta e demanda de programadores Python.

A seguir iremos explorar esses pontos.

O crescimento do Python

Em vários artigos e publicações na internet é possível averiguar o crescimento explosivo do uso de Python nas grandes corporações.

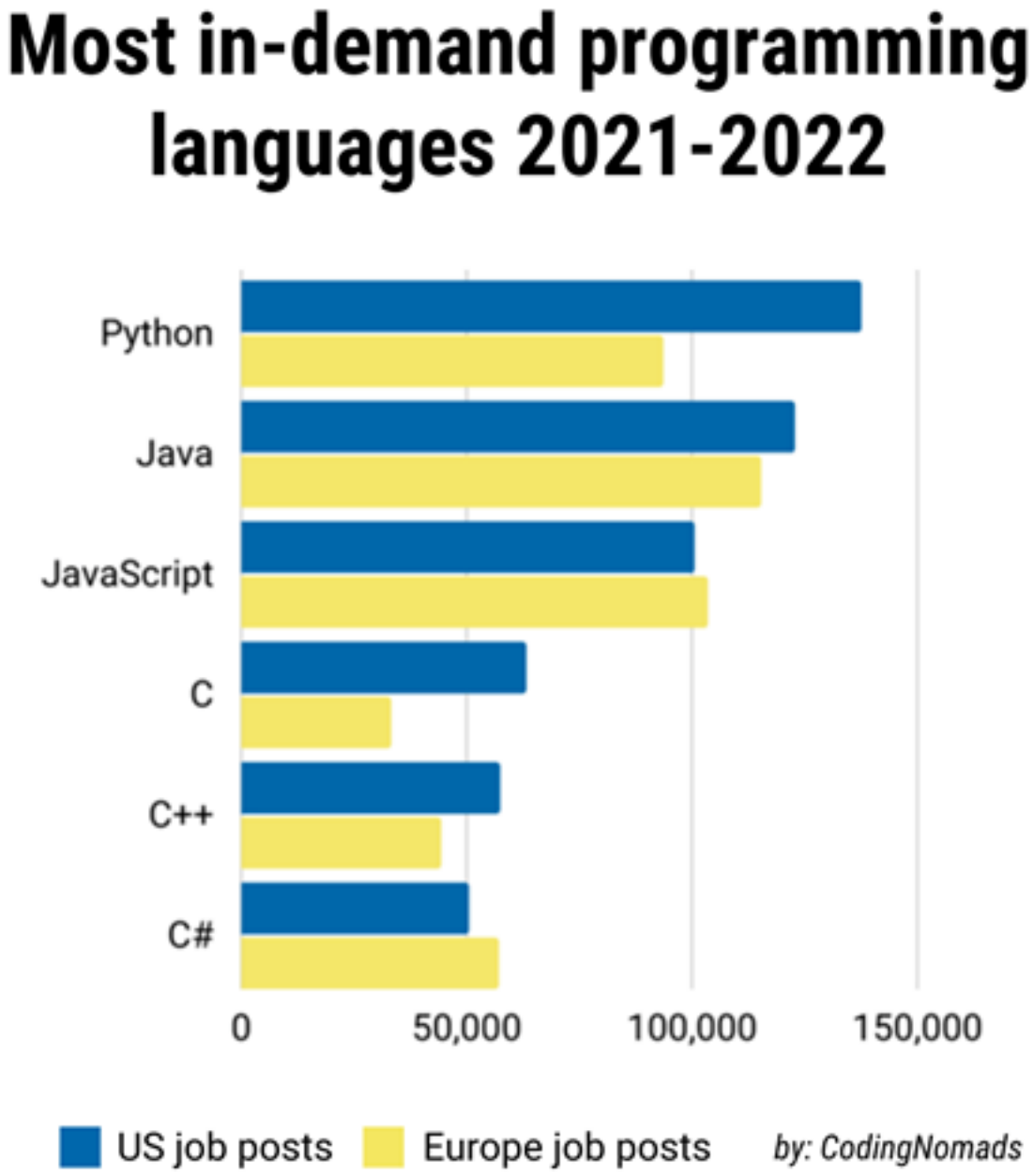
Para citar apenas algumas referências:

https://www.finextra.com/blogposting/21401/python-the-programming-and-development-language-of-the-future

https://www.siliconrepublic.com/careers/python-most-in-demand-coding-language-2022

https://canaltech.com.br/mercado/saiba-por-que-a-linguagem-python-sera-a-mais-exigida-no-mercado-em-2022-204935/

O gráfico abaixo mostra o aumento da demanda por programadores em Python.



Para os programadores independentes, Python oferece um grande ecossistema de desenvolvedores e " peer-support " (algo como suporte solidário) onde desenvolvedores menos experientes podem contar com a sabedoria e experiência de uma comunidade de milhões de desenvolvedores.

Além disso, existem milhares de bibliotecas de funções analíticas e de ML que aceleram o desenvolvimento de modelos de ML: TensorFlow, NumPy, Ramp, Matplotlib, e Pandas para citar apenas algumas. Esses são alguns dos motivos que tornaram Python tão popular entre os cientistas de dados, a ponto de superar o desenvolvimento de modelos em R.

Um novo padrão para Machine Learning

Os modelos analíticos e de Aprendizado de Máquina ou " Machine Learning " (ML) são baseados na criação de modelos de análise que precisam passar por um " treinamento " , que consiste em expor esse modelo a dados históricos e observar como o modelo de ML se comporta e se as conclusões do cientista da dados estão corretas ou não.

Porém, um dos pontos mais importantes do ML, que muitas pessoas esquecem ou não levam em consideração, é a qualidade dos dados usados na etapa de treinamento.

Um artigo recente trás esse ponto à luz:

https://www-forbes-com.cdn.ampproject.org/c/s/www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/amp/

Esse artigo foi comentado por um dos grandes pensadores da nossa comunidade de tecnologia no Brasil: Cezar Taurion.

https://www.linkedin.com/posts/ctaurionandrew-ng-launches-a-campaign-for-data-centric-activity-6881660104656199680-EPPN/

E do post acima eu extraio o seguinte trecho:

" (...)

Hoje gasta-se, pelo menos 80% do tempo na preparação e limpeza dos dados para um sistema de ML ser treinado. Modelos com esse conceito tendem a apresentar resultados ruins quando os dados mudam, por exemplo, quando entrando em produção.

(...) "

O tempo e o esforço que as corporações dispendem preparando os dados para serem ingeridos nos modelos de ML é muito grande e é " preconceituoso " . O pré-conceito é fazer os dados se adequarem ao modelo, quando o correto seria o inverso. Os dados acumulados representam a vida como ela é, ou seja, a realidade.

O uso de um " Smart Data Fabric " , onde os dados são tratados como cidadãos de primeira classe, permite que os dados sejam tratados, limpos e anonimizados sem perder a sua significância e a sua ligação com a realidade.

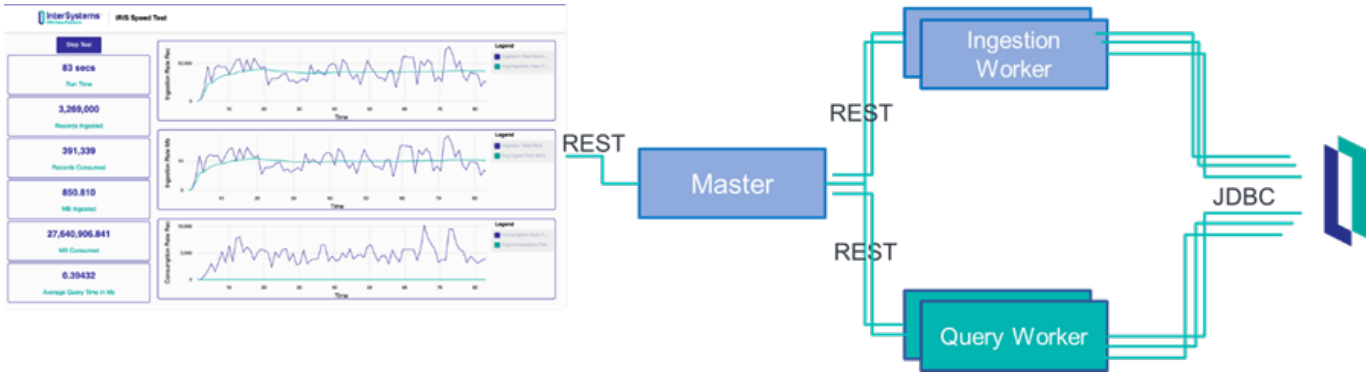
Um dos grandes gargalos de processamento de modelos analíticos ou de ML, é a tempo necessário para preparar os dados, que residem num sistema de armazenamento de dados transacional, e levá-los para o sistema de processamento analítico. Os modelos analíticos requerem uma grande massa de dados e mesmo o tempo de transferência do resultado de uma query entre os dois sistemas, pode ser significativo.

Um dos grandes apelos de se usar Python dentro do InterSystems IRIS© é conseguir rodar esses processos analíticos no servidor de banco de dados, eliminando a necessidade de transferir os dados do banco transacional para o servidor de processamento analítico.

A partir da versão 2021.2 do InterSystems IRIS©, Python é uma das linguagens de desenvolvimento que rodam nativamente dentro do InterSystems IRIS©, que também permite o desenvolvimento em Java, .Net e ObjectScript©. Os modelos de ML que já existem podem ser portados para rodar no InterSystems IRIS, com poucas ou nenhuma adaptação, agilizando o desenvolvimento e a entrada em produção dos modelos de ML.

InterSystems IRIS um banco de dados translítico ou HTAP

O InterSystems IRIS© é um banco translítico, termo usado pelo Forrester Group para descrever um banco de dados que é capaz de executar operações transacionais tradicionais concomitantemente com operações analíticas, HTAP do inglês *"Hybrid Transactional and Analytical Processing"* é outro termo usado para descrever esse tipo de servidor.



A figura acima mostra a arquitetura de um teste de ingestão transacional num ambiente translítico, esse teste de performance foi desenvolvido pela InterSystems e é auditado por uma consultoria externa para garantir a idoneidade dos resultados. Maiores detalhes do teste podem ser encontrados neste link:

https://www.intersystems.com/speedtest

Os resultados dos testes de performance HTAP mostram que o InterSystems IRIS© apresenta melhor performance, em comparação com outros bancos de dados, em um ambiente misto de ingestão de dados e tempo de resposta das queries.

Os testes mostram que o InterSystems IRIS © tem performance entre 60% e 4860% melhor na ingestão de novos registros, enquanto os tempos de resposta das queries analíticas variam entre 170% e 5600% mais rápidas.

No link acima, também é possível encontrar os procedimentos para executar o teste em seu próprio ambiente ou em uma nuvem como AWS.

A natureza híbrida do InterSystems IRIS, permite que ele seja extremamente performático em ambientes de ML, e agora com a possibilidade de executar código em Python nativamente na plataforma, isso permite que os cientistas de dados executem os processos no servidor de dados ao invés de levar os dados para o servidor de processamento analítico.

Além disso o InterSystems IRIS© suporta PMML (Predictive Model Markup Language) que permite que ele importe ou exporte modelos preditivos de Machine Learning, o que torna o InterSystems IRIS a plataforma ideal para a implementação de MLOps, ou Machine Learning Operations.

MLOps implementa ferramentas para o gerenciamento do ciclo de vida completo de um modelo de ML: treinar o modelo, conduzir análise dos erros para identificar os tipos de dados nos quais o algoritmo apresenta baixa performance, conseguir mais dados via nova coleta ou conseguir marcações mais consistentes para os dados que foram detectados como ambíguos. E finalmente o teste em produção no mundo real, onde os novos dados são usados para refinar continuamente o modelo. Uma pesquisa recente aponta que um terço das organizações não monitoram rotineiramente ou retreinam seus modelos de ML para melhorar a performance.

Novamente, o suporte ao Python dentro do InterSystems IRIS© alavanca e acelera este processo, tornando o InterSystems IRIS© a plataforma de dados ideal para projetos de ML que se alimentam de dados oriundos de um sistema transacional.

Espero que este breve artigo tenha despertado o seu interesse no uso de Python com o InterSystems

IRIS©, verdadeiramente um "*Casamento Perfeito*" !

#Analytics #Bancos de dados #Embedded Python #Artificial Intelligence (AI) #IntegratedML #Machine Learning (ML) #Python #InterSystems IRIS

URL de origem:https://pt.community.intersystems.com/post/python-e-intersystems-iris%C2%A9-um-casamento-perfeito-para-machine-learning