

ANÁLISE DE DESEMPENHO DO PYTHON NO SQL SERVER

AGENDA

- Descrição do problema
- Metodologia
- Análise
- Conclusões

DESCRIÇÃO DO PROBLEMA

- As vezes trazer os dados para processamento pode ser custoso
- Configurar o ambiente também pode ser complexo
- Diversos ativos intermediários
- Manter os dados perto do processamento pode ajudar

DESCRIÇÃO DO PROBLEMA

- Alguns bancos, como o SQL Server, estão incluindo módulos internos com ambientes de programação completos para processamentos de análise e Inteligência Artificial, como Python
- Permite que scripts Python sejam executados dentro do banco em um script T-SQL
- O banco tem um mecanismo na linguagem T-SQL para passar os dados para o Python que pode executar com os mesmos recursos da linguagem de um ambiente tradicional
- Porém, será que o desempenho é significativamente melhor?

METODOLOGIA

- Configurado um banco e um ambiente externo em máquinas de mesma configuração
- O banco consultado pelo ambiente externo é o mesmo que processou os scripts internos
- Foi executado o k-means do python em ambos os ambientes, com os mesmos parâmetros e sempre para k=20
- Foram realizadas 11 execuções, cada uma com 50 repetições
- Os dados foram coletados e analisados com relação a estatísticas, projeto fatorial 2kr e regressão da evolução da média com relação ao volume de registros

ANÁLISE: ESTATÍSTICAS

Média MSSQL Total	SEM MSSQL Total	Média Ext. Python	SEM Ext. Python	Média Ext. Total	SEM Ext. Total
2175.72	66.146825	0.004804	0.000059	0.025265	0.000330
2159.34	11.438339	0.005841	0.000074	0.045689	0.000965
2997.86	121.257372	0.006134	0.000111	0.082430	0.000952
3475.92	135.396255	0.007714	0.000092	0.171707	0.002071
3881.92	113.141684	0.015518	0.000538	0.301726	0.002303
5137.88	100.511393	0.016306	0.000115	0.572136	0.001617
7206.00	75.427689	0.022154	0.000156	1.167824	0.005852
11982.98	88.434435	0.040379	0.000604	2.358322	0.011923
22029.92	85.575845	0.061625	0.000942	4.620593	0.024592
46473.32	109.819159	0.170140	0.001823	9.521799	0.073421
110708.30	143.106215	0.293033	0.005766	18.544739	0.100246

ANÁLISE: ESTATÍSTICAS

Tamanho da Amostra	Razão das Médias Totais	Razão das Médias do Python
1000	86114.695321	25.696356
2000	47261.638887	29.909805
4000	36368.338310	163.448297
8000	20243.271353	188.973453
16000	12865.706386	117.887246
32000	8980.172447	184.804978
64000	6170.449176	224.412268
128000	5081.146381	236.838623
256000	4767.769222	311.439772
512000	4880.728799	251.496192
1024000	5969.795471	358.886295

Tabela 2: Razão entre as médias internas e externas.

ANÁLISE: PROJETO FATORIAL

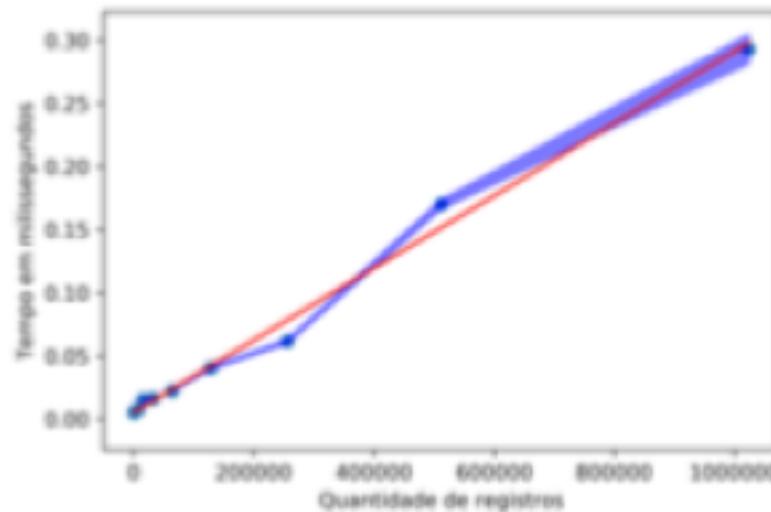
Fator	Proporção explicada
Tamanho da amostra (A)	0.0629278904135
Local de execução (B)	0.78594401553
Interação dos fatores (AB)	0.0628853486447
Erro experimental	0.0882427454119

Tabela 3: Resultado do Projeto 1 (Amostra máxima de 16.000).

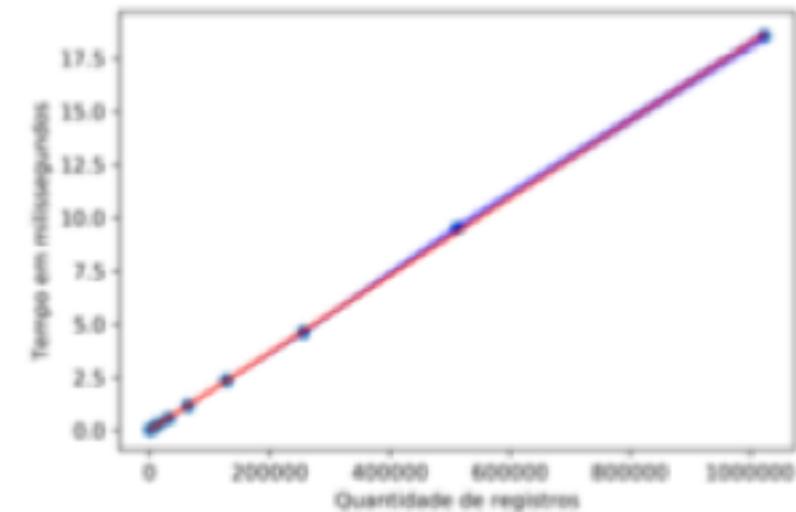
Fator	Proporção explicada
Tamanho da amostra (A)	0.325177950701
Local de execução (B)	0.349323866985
Interação dos fatores (AB)	0.325174787005
Erro experimental	0.000323395307937

Tabela 4: Resultado do Projeto 2 (Amostra máxima de 1.024.000).

ANÁLISE GRÁFICA E REGRESSÃO

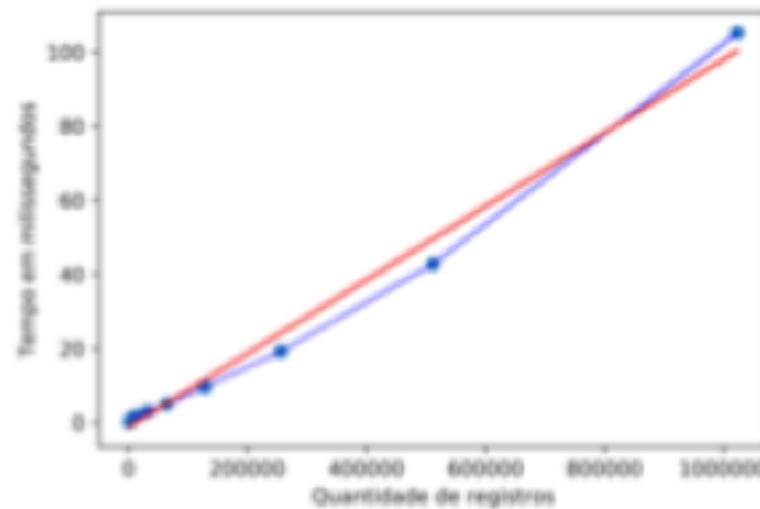


(a) Execução Externa apenas Python.

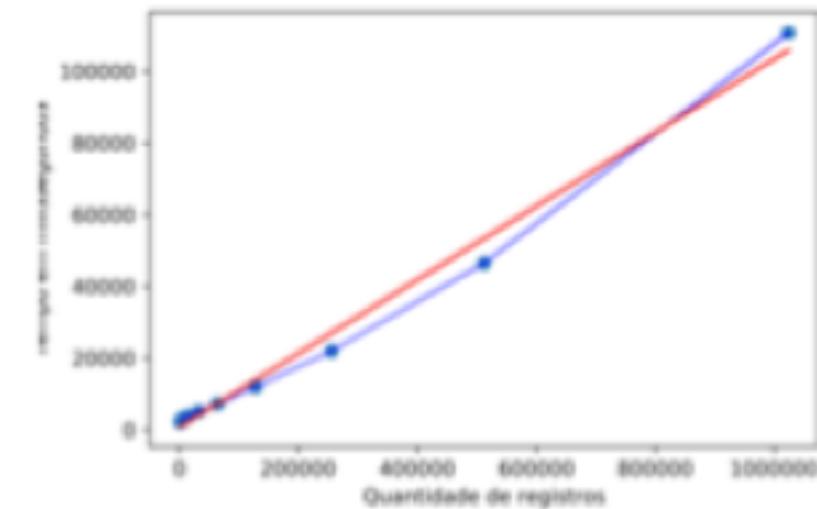


(b) Execução Externa Total.

ANÁLISE GRÁFICA E REGRESSÃO



(c) Execução Interna apenas Python.



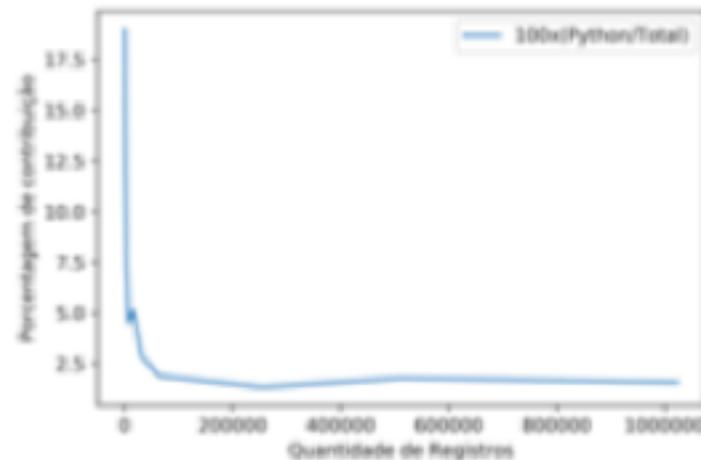
(d) Execução Interna Total.

ANÁLISE GRÁFICA E REGRESSÃO

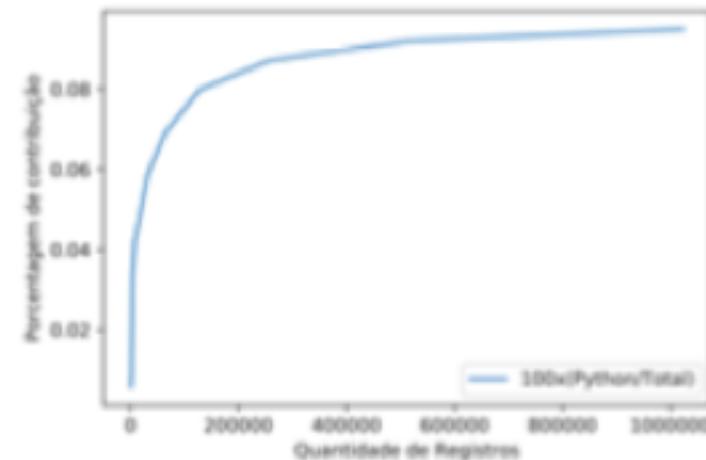
Regressão	Coeficientes	R^2	IC de 95%
Externa Python	$b_0=0.0053, b_1=2.8573e-07$	0.9917	$IC_0=(-0.0017, 0.0123),$ $IC_1=(2.6608e-07, 3.0538e-07)$
Externa Total	$b_0=0.0189, b_1=1.8175e-05$	0.9998	$IC_0=(-0.0410, 0.0787),$ $IC_1=(1.8007e-05, 1.8343e-05)$
Interna Python	$b_0=-1.2650, b_1=9.9267e-05$	0.9890	$IC_0=(-4.0693, 1.5393),$ $IC_1=(1.8007e-05, 1.8343e-05)$
Interna Total	$b_0=728.7589, b_1=0.1027$	0.9896	$IC_0=(-2097.5945, 3555.1123),$ $IC_1=(9.1400e-05, 0.0001)$

Tabela 5: R^2 e Intervalos de Confiança de 95% das Regressões.

ANÁLISE GRÁFICA: RAZÃO PYTHON VS TOTAL



(a) Externo: Proporção do tempo do Python com relação ao Total.



(b) Interno: Proporção do tempo do Python com relação ao Total.

CONCLUSÃO

- O ambiente externo teve um desempenho muito superior para esse experimento
- Entretanto, trazer para dentro de um banco relacional ferramentas para diminuir a movimentação dos dados é importante
- Evita mover grandes massas de dados
- Porém o desempenho geral ainda não parece justificável para migrar de um ambiente composto
- Outros tipos de análise poriam ser testadas