NoSQL at Twitter

Trabalho Prático

PÓS GRADUAÇÃO LATO SENSU – PUC MINAS

Marcus Thadeu Oliveira Campos | Banco de dados não relacionais | 23/12/2016

1. Objetivo

Utilizar os conhecimentos adquiridos na disciplina Banco de dados não relacionais ofertado pelo curso de Pós-graduação lato sensu da PUC Minas no período de outubro/16 até dezembro/16 para realizar um trabalho prático que permita:

- Coletar informações de redes sociais ou importar dados externos
- Armazenar ~1M de dados em um banco NoSQL
- Extrair informações do tipo :
    - Termos mais frequentes
    - Volume x dia
    - Volume x hora do dia

2. Metodologia

2.1 Importação da base de dados

A base utilizada neste trabalho foi extraída do portal https://www.kaggle.com que é uma plataforma de modelagem, previsão e analise de dados que estimula competições em que empresas e pesquisadores postam seus dados e produzem conhecimento e os resultados são avaliados, podendo ser premiados ou não. Com isso, a kaggle é uma das maiores e mais diversas comunidades de dados do mundo e uma pesquisa sobre dados em redes sociais me retornou uma coleção de dados extraída de postagens do Twitter no período de 14 de março de 2016 até 16 de março de 2016 em que ocorreu no Brasil o processo de impeachment da presidente Dilma. Estes dados foram coletados usando APIs de streaming e publicados por Moreno, C. (2016). DITRD-v1.0.0 - Dilma impeachment Twitter dados brutos [https://github.com/caiomsouza/TwitterRawData]. Madrid, Espanha: U-TAD, Programa de Certificado em Ciência de Dados. O download dos dados pode ser feito pelo link :
https://github.com/caiomsouza/TwitterRawData/releases/download/DITRD-v1.0.1/DITRD-v1.0.1.zip

2.2 Armazenamento no MongoDB

Os dados foram armazenados no MongoDB que é uma aplicação de código aberto, de alta performance, sem esquemas, orientado a documentos, diferente dos Bancos de dados tradicionais que seguem o modelo relacional. Esses bancos de dados também são chamados de Bancos NoSQL (Not Only SQL) tem como característica conter todas as informações importantes em um único documento, ser livre de esquemas, possuir identificadores únicos universais (UUID), possibilitar a consulta de documentos através de métodos avançados de agrupamento e filtragem (MapReduce) e também permitir redundância e inconsistência.

O arquivo contendo os twitters estava no formato JSON (JavaScript Object Notation), e foi carregado e processado com as técnicas de MapReduce.

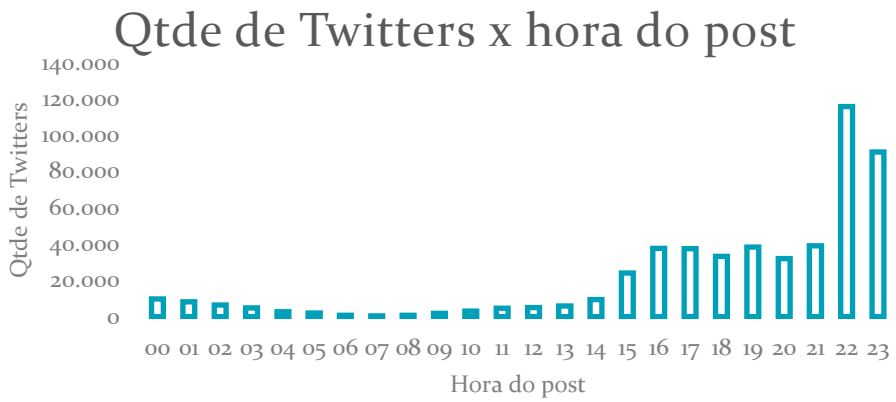
3. Análise das informações

O banco de dados analisado continha 514.116 twitters, nas quais extraímos as seguintes informações utilizando os scripts demonstrados em aula:

Volume x dia

Dia	Qtde de Twitters	%
Mar 16	412.452	80%
Mar 15	101.654	20%
Total Geral	**514.106**	**100%**

Volume x hora do dia



Termos mais frequentes

Palavras	Qtde de Citações
Dilma	123.291
Lula	85.469
Brasil	19.139
Moro	14.350
Brasília	5.861
PT	4.935
PF	4.866

4. Conclusão

Quando analisamos os horários dos posts vemos que o maior volume de postagens no twitter é entre 22h e 24h. Quando observamos o volume de posts nos dias 15 e 16 de março vemos o quanto o twitter é sensível para as questões do cotidiano. No dia 16 de março de 2016 uma conversa telefônica entre Dilma e Lula indicou que a nomeação do ex-presidente para assumir a Casa Civil teria a intenção de dar a ele foro privilegiado o que, consequentemente, faria as denúncias contra Lula saírem de Moro para o Supremo Tribunal Federal (STF), fazendo com que o "twitter" tivesse um volume de postagens 4x maior se comparado com o dia anterior. Já para análise das citações, vemos que os termos mais frequentes se correlacionam com o processo de impeachment que na época era aclamado pelo povo brasileiro.