UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

COORDENAÇÃO DE LICENCIATURA EM INFORMATICA

Especialização para Desenvolvimento em Ambiente Internet e Dispositivos Móveis

ONEIDE LUIZ SCHNEIDER

UM ESTUDO DA APLICABILIDADE DE FRAMEWORKS DE

DATA MINING EM APLICAÇÕES COMERCIAIS

FCO. BELTÃO

2014

ONEIDE LUIZ SCHNEIDER

UM ESTUDO DA APLICABILIDADE DE FRAMEWORKS DE

DATA MINING EM APLICAÇÕES COMERCIAIS

Projeto de pesquisa apresentada como requisito parcial à obtenção do título de Especialista, da Coordenação de Licenciatura em Informática, da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Me. Michel Albonico

FCO. BELTRÃO

2014

Resumo

Atualmente a quantidade de dados vem crescendo significativamente. As organizações/empresas vem se moldando em cima da desses dados. É preciso cada vez mais ferramentas melhores e mais eficazes para processar esses dados, pois o volume gerado aumenta a cada dia. O desafio é processar essa quantidade dados de forma correta, nunca eram coletadas tantas informações como são hoje em dia. Pela falta de estrutura desde computacional, hardware, software e armazenamento, que hoje temos disponíveis. Nesse trabalho o objetivo é mostrar alguns frameworks de Data Mining que estão disponíveis hoje, e testar quais são mais eficazes em bando de dados NoSQL.

**Palavras-chave:** Data Mining. Frameworks. NoSQL.

abstract

Currently the amount of data is growing significantly. Organizations / companies is shaping up of such data. We need increasingly better and more effective tools for processing the data, because the volume increases generated every day. The challenge is to process this much data correctly, never as much information as they are today were collected. Lack of structure from computational hardware, software and storage, today we have available. In this work the goal is to show some data mining frameworks that are available today, and test which are more effective in bunch of NoSQL database.

**Keywords:** Data Mining. Frameworks. NoSQL.

sumário

**1 INTRODUÇÃO …………………………………………………………………… 06**

**2 PROBLEMA DE PESQUISA …………………………………………………… 07**

**3 HIPOTESE ………………………………………………………………………... 07**

**4 OBJETIVOS ……………………………………………………………………… 07**

4.1 OBJETIVO GERAL **……………………………………………….…………… 07**

4.2 OBJETIVO ESPECÍFICO **…………………………………………………….. 07**

**5 JUSTIFICATIVA …………………………………………………………………. 08**

**6 METODOLOGIA …………………………………………………………………. 08**

**7 ESTRUTURA DO TRABALHO …..……………………………………………. 09**

**8 CRONOGRAMA …………………………………………………………………. 10**

**REFERÊNCIAS ….…………………………………………………………………. 11**

1. introdução

Atualmente a quantidade de dados vem crescendo rapidamente. De acordo com Hilbert (2011) existiam 300 exabytes de dados em 2007. Ressaltando que um exabyte equivale a 1 bilhão de gigabytes. Em 2013 já chegamos a marca de 1200 exabytes de dados.

Segundo Breternitz (2013) as empresas possuem grandes quantidades de dados armazenados. Esses dados podem servir para extrair informações mais precisas usando frameworks de data mining, do que relatórios das vendas, comissões de vendedores, etc.

Mayer-Schonberger e Cukier (2013) destacam que o mundo está mudando rapidamente e a quantidade de informações cresce mais rápido que a própria economia mundial, e poder de processamento dos computadores cresce nove vezes mais rápido. Isso leva mudança a todos. Dessa forma o processamento desses dados se torna uma tarefa difícil e especializada. Nestes casos a extração de dados não pode ser somente feita com SQL, portanto, são necessárias técnicas para extração de dados utilizando inteligência artificial (ex., mineração de dados).

Diante disso, iremos realizar testes com frameworks que permitem a mineração de dados em ambientes distribuídos, o que torna os ambientes mais escaláveis e por consequência, eles conseguem suportar o crescimento da quantidade de dados. Nesses testes é necessário realizar experimentos para verificar como esses frameworks (e.g., distribuídos) se comportam em relação aos algoritmos centralizados.

1. Problema de pesquisa

Dentro da área do Data Mining a principal atividade é analisar/processar grandes quantidades de dados. Onde o principal problema é saber qual é melhor forma de armazenar esses dados, onde salvar, o que pode ser feitos com eles, como posso torna-los uteis e se é possível utilizar essa estrutura em uma aplicação comercial. Para minerar dados é preciso uma grande quantidade de dados para se ter um resultado mais eficiente. Considerando que a quantidade de dados cresce constantemente os algoritmos de mineração são de grande custo computacional, nesse caso, acredita-se que os frameworks distribuídos permitem que esses algoritmos sejam executados em mais do que um computador e acredita-se que isso possa melhorar a performance de execução.

1. Hipótese

Para execução desse projeto será verificado se os frameworks distribuídos (e.g., Apache Mahout) serão mais eficientes do que os algoritmos não distribuídos, e verificar-se os frameworks distribuídos exercem algum tipo de influência nos resultados ou no tempo desses algoritmos.

1. objetivos

## objetivo geral

O presente trabalho tem por objetivo geral Identificar a influência dos frameworks de mineração de dados distribuídos em relação aos algoritmos de mineração de dados.

## objetivo específico

Como objetivos específicos, este trabalho pretende:

1. Criar um ambiente de trabalho onde possam ser realizados os procedimentos para os testes.
2. Executar algoritmos de data mining com frameworks distribuídos.
3. justificativa

A grande dificuldade que temos hoje é que as informações estão crescendo constantemente e cada vez mais é preciso criar novas soluções para poder processar de maneira correta e mais rápida essas informações. Para isso, podem ser utilizados algoritmos de mineração de dados. A mineração de dados consiste em um processo analítico que busca explorar de forma eficiente grandes quantidades de dados, como bases de bolsa de valores, organizações governantas, etc. Com isso mineração de dados se torna a cada dia mais importante, por isso precisamos que elas sejam mais precisas.

1. metodologia

O objetivo geral da pesquisa a ser realizada nesse trabalho é identificar qual framework se encaixa melhor em aplicações de comerciais. Para maior aprofundamento do assunto, incialmente será realizada uma pesquisa bibliográfica exploratória baseada em frameworks de mineração de dados, aprendizagem de máquina. A pesquisa bibliográfica busca melhorar as hipóteses de pesquisa, que até então são as seguintes:

1. Quanto a uso de frameworks de Data Mining:

* Principais no mercado hoje?
* Qual o principal função deles?
* Qual se enquadra melhor em Sistemas distribuídos?

1. Quanto a organizações conhecidas já usam frameworks de Data Mining:

* Quais as principais?
* Qual principal motivo para os utiliza-los?
* O que é Data Mining?
* Por que fazer mineração de dados?

1. Quanto a aplicações comerciais:

* Algum exemplo de aplicação que usa algum framework de Big Data?

1. Quanto a bancos de dados NoSQL:

* Como funcionam?
* Quais os principais no mercado?
* Quantos suportam os frameworks de Big Data?
* Quais são os mais indicados para o assunto proposto?

Todas as etapas acima serão executadas para confirmar a hipótese ou não. E durante o processo de execução serão executadas testes com os frameworks para verificar qual irá se comportar melhor e como irá se comportar.

Todos dos procedimentos serão feitos em ambiente de testes utilizando dados que estão disponíveis na internet, como bolsa de valores, organizações governamentais e dados de aplicações comerciais reais e alimentados manualmente por algoritmos criados para os testes para se chegar a um padrão e tentar entender como se chegou a esse padrão, para assim podermos afirmar qual se comportou melhor.

1. eSTRUTURA DO TRABALHO

Para a estrutura do trabalho pretende-se dividir em três partes. 1) Elaborar introdução e objetivos como a metodologia utiliza, 2) desenvolvimento teórico, 3) resultados e conclusões.

1. CRONOGRAMA

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividades** | **Meses(2014)** | | | | | | | | | | | |
| **Jan** | **Fev** | **Mar** | **Abr** | **Mai** | **Jun** | **Jul** | **Ago** | **Set** | **Out** | **Nov** | **Dez** |
| **Elaboração do Projeto de Pesquisa** | **X** |  |  |  |  |  |  |  |  |  |  |  |
| **Entrega Projeto de pesquisa** | **X** |  |  |  |  |  |  |  |  |  |  |  |
| **Levantamento Bibliográfico** |  | **X** | **X** |  |  |  |  |  |  |  |  |  |
| **Criação do ambiente de teste** |  |  | **X** | **X** |  |  |  |  |  |  |  |  |
| **Realização dos testes** |  |  |  | **X** | **X** |  |  |  |  |  |  |  |
| **Elaboração do Artigo** |  |  |  |  | **X** | **X** | **X** | **X** | **X** |  |  |  |
| **Entrega da Artigo** |  |  |  |  |  |  |  |  | **X** |  |  |  |
| **Apresentação do Artigo** |  |  |  |  |  |  |  |  |  |  |  | **X** |

ReferÊncias bibliogrÁficas

MAYER-CHONBERGER, Viktor; CUKIER, Kenneth. **Big Data:** como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana / Viktor Mayer-Schonberger, Kenneth Cukier ; tradução Paulo Polzonoff Junior. - 1. Ed. – Rio de Janeiro : Elsevier, 2013.

SADALAGE, Pramod J.; FOWLER, Martin. **NoSQL Essencial:** um Guia conciso para o Munda Emergente da Persistência Poliglota / Pramod J. Sadalage, Martin Fowler ; tradução Acauan Fernandes. - 1. Ed. - Rio de Janeiro : Novatec, 2013.

SILVA, Mauro M. de L. Big Data: um novo conceito gerando oportunidades e desafios. **Revista RETC**, - 13. Ed. - Jundiaí, out. 2013.

Disponível em: < http://revista-fatecjd.com.br/retc/index.php/RETC/article/view/74>. Acesso em: 28 jan. 2014.

HILBERT, Martin; LÓPEZ, Preiscilla. **The World’s Technological Capacity to Store, communicate, and Compute Information. Science,** honover, MD: Institute for Operations Research and the Management Sciences vol. 1, p.60-65, abril de 2011.

HILBERT, Martin.**How to measure the World’s Technological Capacity to Communicate, and Compute Information? International Journal of Communication.** Whoshinton, DC:USC Annenberg Press, p. 1042-55, 2012**.** Também disponível em < http://ijoc.org/index.php/ijoc/article/viewFile/1562/742/ >. Acesso em 29 jan. 2014.

Rodrigues, José dos Reis. **REVOLUÇÃO DO BIG DATA**. Disponível em:

< http://gm2x.com.br/blog/revolucao-big-data-voce-ja-esta-se-preparando/ >. Acesso em: 27 jan. 2014.