Análise de séries temporais a partir de um BD NoSQL

Solemar de Oliveira[1] ; Bianca Maria Pedrosa[2] ;
1-Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - Campus Bragança Paulista,
2-Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - Campus Campinas.

Objetivo

O objetivo deste trabalho é analisar dados de poluição atmosférica, causada por queimadas e poluentes industriais, e a sua correlação com dados de internações clínicas causadas por doenças respiratórias do município de Campinas. Esses dados são disponibilizados como dados abertos por órgãos públicos. Embora a ideia de coletar e analisar dados de séries temporais não seja nova, o volume, velocidade e variedade de dados contribuem para tornar a análise de séries temporais escaláveis um grande desafio. Constitui o contexto da pesquisa o estudo de tecnologias Big Data como o banco de dados NoSQL e R para correlacionar diferentes séries temporais e determinar como estes dados se relacionam.

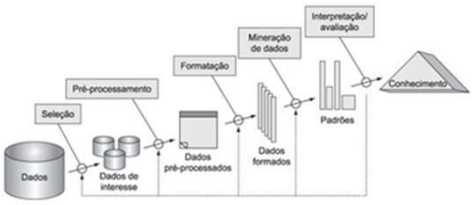
Introdução

Big Data é um conjunto de soluções tecnológicas para lidar com dados em volume, variedade e velocidade de processamento, gerando conhecimento. Para isto, reúne tecnologias inovadoras que combinam técnicas de processamento e armazenamento de dados distribuídos, ferramentas para análise estatística e os conhecimentos de uma área de domínio (1). O armazenamento e análise de séries temporais tem sido um desafio para a área de banco de dados, uma vez que estes dados são capturados de forma não estruturada, em velocidade e volumes típicos de sistemas big data (2). Nos últimos anos, os bancos de dados NoSQL se popularizaram, entre outros motivos, por suportar dados em formatos mais flexíveis que o modelo de dados relacional, tais como o modelo chave-valor, orientado a documentos e orientado a coluna, entre outros. Além da flexibilidade de modelagem, os bancos NoSQL possuem facilidades para escalar, isto é, podem crescer com a adição de nós de processamento, isolando o desenvolvedor dos aspectos do processamento distribuído (3).

Materiais e Métodos

A metodologia de desenvolvimento deste projeto segue o modelo de processo de descoberta do conhecimento, apresentada na Figura 1, que inclui nas etapas iniciais do processo, as operações de seleção e pré-processamento dos dados, do domínio de banco de dados, e nas etapas finais as atividades de mineração de dados e interpretação/avaliação, da área de ciência de dados. Os dados de poluição e internações hospitalares da região de Campinas foram obtidos a partir do DATASUS e CETESB. Esses dados foram preparados através de técnicas de limpeza e transformação dos dados, que são operações clássicas de um processo de migração e integração de dados. Nas próximas etapas do projeto, os dados serão modelados para armazenamento num banco NoSQL, provavelmente o HBASE (4), que tem facilidades de para integração com o sistema R, de análise estatística. Em seguida, será realizada a entrada de dados no Hbase através de ferramentas apropriadas. Para finalizar, serão utilizadas algumas funções no R para análise dos dados. As ferramentas de análise estatísticas serão estudadas e testadas para os dados selecionados. A integração do HBASE com o R será implementada e alguns testes em ambiente de processamento distribuído serão realizados.

Figura 1: Processo de descoberta do conhecimento



Fonte: DUNNING e FRIEDMAN, 2015

Resultados preliminaries

Os resultados preliminares obtidos a partir deste trabalho mostram que é possível identificar uma correlação fraca entre os dados de poluição e saúde. Entretanto, a análise estatística está apenas em estágio inicial. Até o momento foi utilizado o software R, para análise dos dados de um ano (2016-2017). Mas pretende-se escalar o banco de dados e para isto o Sistema de Computação Científica (SICC) do IFSP foi instalado e configurado com recursos de banco de dados escalável com hadoop e hbase, entre outras soluções tem Big Data.

Conclusão

A realização deste projeto está inovando o modo de lidar com séries temporais e desenvolvendo habilidades para análise de dados visando a geração de conhecimento dentro da perspectiva da ciência de dados, capaz de combinar banco de dados, estatística e inteligência artificial para decifrar complexos relacionamentos entre os dados.

Referências

(1) AYANKOYA, K; CALITZ, A; GREYLIG, J. 2014 **Intrinsic Relations between Data Science, Big Data, Business Analytics and Datafication**. Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014. p.192-198 .

(2) WLODARCZYK, T.W. 2012. **Overview of Time Series Storage and Processing in a Cloud Environment**, Proceedings of the 4th International Conference on Cloud Computing Technology and Science. p. 625-628.

(3) DUNNING, T; FRIEDMAN, E. 2015 **Time Series Databases New Ways to Store and Access Data**. Sebastopol, O'Reilly Media, 2a edição. 70 p.

(4) APACHE. **Hbase Reference Guide** <http://hbase.apache.org/book.html> Acesso em: 08 out 2017.