MODELO PARA ESTRUTURAÇÃO DE BASE DE DADOS DIGITAIS PARA APOIO À TOMADA DE DECISÃO EM RISCO AGRÍCOLA

Ricardo. A. Neves[1,2,*], Paulo. E. Cruvinel[1,2]

[1] *Embrapa Instrumentação, Rua XV de Novembro, 1452, 13560-970, São Carlos, SP*
[2] *Universidade Federal de São Carlos, Rodovia Washington Luís km 235, SP-310, 13565-905, São Carlos, SP*
[*] *Autor correspondente, e-mail: ricardo.alexandre@estudante.ufscar.br*

Resumo: Este artigo apresenta um modelo para estruturação de bases de dados digitais, para que se possa trabalhar em ambiente de Nuvem e atender às diferentes fontes de dados originadas de *Big Data*. O trabalho considera receber dados estruturados, semiestruturados e não estruturados, para uso na gestão de risco em agricultura. Sua concepção está estruturada como uma arquitetura que combina recursos de *Data Mart, Data Warehouse (NoSQL)* e de *Data Lake* para dar suporte adequado à tomada de decisão, por meio da descoberta do conhecimento e fazendo uso de algoritmos para mineração de dados por meio de recursos de aprendizado de máquina. A configuração apresentada atende aos diversos cenários que envolvem dados agrícolas, obtidos a partir de sensores em múltiplas modalidades.

Palavras-chave: *big data*, *data warehouse*, fusão de dados, gerenciamento de risco agrícola

COMPUTER ALGORITHM FOR DATA STRUCTURING USED IN AGRICULTURAL RISK MANAGEMENT

Abstract: This paper presents a model for digital databases structuring based on the use of a cloud environment and bigdata from different sources. Such model considers receiving structured, semi-structured and unstructured data for agricultural risk management. Its design was organized taken into account an architecture that combines Data Mart, Data Warehouse (NoSQL), and Data Lake capabilities for support decision making through knowledge discovery and the use of data mining. Besides, Machine Learning resources were used for the data mining. The configuration presented meets the various scenarios involving agricultural data, obtained from sensors in multiple modalities.

Keywords: big data, data warehouse, data fusion, agricultural risk management

1. Introdução

Por meio da utilização de tecnologias de *Business Intelligence* (*B.I.*), providas por processos sistematizados e automatizados para apoio ao processo de tomada de decisão, é notado que a quantidade, a escalabilidade e a heterogeneidade são características que compõe a realidade das bases de dados a nível mundial, nos mais variados segmentos, sejam estes governamentais, da produção industrial ou agrícola. Desta forma, se faz necessário uma escolha assertiva na adoção de métodos, ferramentas e, muitas vezes, da contratação de uma consultoria especializada para a obtenção de resultados de sucesso (ALEXANDRA MARIA IOANA; VLAD; RAMONA, 2016).

De acordo com Sosinsky (2011) há bons produtos comerciais, atualmente, que oferecem recursos para se trabalhar com *Big Data* no ambiente de Nuvem. Tais recursos, têm se mostrado como opção interessante, por meio do fornecimento de serviços elásticos, ou seja, por demanda, para projetos com diversidade de dados caracterizada por tipos, formatos, fontes, assim como a quantidade (volume). No ambiente em Nuvem são oferecidas arquiteturas distribuídas, paralelas ou seu uso integrado, cujas características principais são o alto desempenho, a alta disponibilidade, a redundância dos dados, a variedade de ferramentas automatizadas e integradas com a possibilidade de configuração customizada e, principalmente, com custo acessível para a maioria dos projetos.

No contexto de análise de dados, por meio de bancos de dados relacionais, aplicada ao auxílio às tomadas de decisões coorporativas, a partir de grandes massas de dados estruturados e divididas por assuntos, sendo muito comum a utilização de uma estrutura de *Data Warehouse (DW)*. Essa abordagem contempla a construção de modelo de dados, a partir de uma visão multidimensional, utilizando-se de estruturas dimensionais definidas como cubos de dados *OLAP* (*On-Line Analytical Processing*). Essas estruturas podem variar em diferentes arquiteturas, além de dispor de ferramentas *OLAP* utilizadas para operacionalizar os cubos em suas várias dimensões. No *DW* é utilizado um modelo dimensional, composto por uma tabela central, chamada tabela fato, junto a um conjunto de tabelas relacionadas chamadas de dimensões. A tabela fato contém o que será analisado e as dimensões armazenam as perspectivas de análise sobre tais fatos (HARINARAYAN; RAJARAMAN; ULLMAN, 1996; SINGH, 1998; KIMBALL; ROSS, 2013).

Para a construção do modelo de *DW* há diferentes abordagens que podem ser consideradas, visto que cada uma delas é sempre realizada diante de um conjunto de etapas a serem cumpridas pelo entregador. Porém, para todas as abordagens, existe a necessidade de conhecimentos específicos, nos quais o entregador tem que se adequar. Quando o entregador não tem total domínio desses conhecimentos, desenvolver um modelo de dados dessa natureza torna-se uma tarefa muito complicada e propensa a constantes erros e insatisfações, situações essas que podem ser verificadas diante da validação dos resultados obtidos, comparando-os com os objetivos propostos pelos requisitos fornecidos pelo demandante (INMON, 1997; SINGH, 1998; KIMBALL; ROSS, 2013).

As atualizações de um *DW* são feitas por meio de fontes externas e internas, utilizando de extração, transformação e carga *(ETL - Extract, Transform and Load).* Essas informações são inseridas no repositório de metadados, o qual também contém dados sobre o *DW*, dos *Data Marts* e das aplicações que os acessam. A parte final do processo fica a cargo do conjunto de programas que atuam na análise dos dados contidos no *DW* (INMON, 1997; KIMBALL; ROSS, 2013).

Big Data é tipicamente não estruturado (como texto, vídeo, áudio) ou semiestruturado (como *web-logs, e-mails, tweets*). Dados não estruturados podem ser entendidos como um documento de texto e não possuem estrutura explícita, porém dados semiestruturados podem envolver alguma estrutura. A partir da definição de *Big Data*, esta tecnologia se encaixa nas práticas existentes do *DW* corporativo. As novas tecnologias, como o *Hadoop*, desenvolvidas para grandes volumes de dados, tornam-se naturalmente parte do arsenal dos desenvolvedores de *ETL*, a partir de uma variedade de fontes de dados. Com esses avanços tecnológicos o *Big Data*, apesar de seus desafios, tornou-se outro tipo de conjunto de dados de origem para o *DW*, em sistemas de auxílio à tomada de decisão (HU,2015; JUKIĆ et al.,2015).

Diante dos vários cenários que envolvem as mais diversas metodologias para interpretação e análise de dados para tomada de decisão, atualmente a solução de *DW* tem sido utilizada a partir de fontes de *Big Data*. A integração de tais tecnologias deve ser cuidadosamente analisada, pois a estrutura de *DW* exige uma adaptação, quando o foco não é a utilização de dados puramente estruturados. Assim, quando o conjunto de dados envolvidos são semiestruturados e não estruturados, além do uso de bancos de dados relacionais ou transacionais, como nas abordagens de negócios tradicionais, a recomendação é a utilização de um banco de dados *NoSQL* (*Not Only SQL*), junto à estrutura do *DW* (CHEVALIER et al., 2015; ALEKSEEV et al., 2016).

Muitas abordagens têm sido apresentadas na literatura sobre tais adaptações. Porém, a abordagem de Solodovnikova e Niedrite (2018) se destacou por apresentar uma arquitetura, de forma que suas características se mostraram híbridas e consistentes. A arquitetura de *Big Data Warehouse* apresentada mostrou a manutenção das boas práticas da arquitetura tradicional do *DW*. Segundo Kimball e Ross (2013), essa arquitetura, também por trabalhar com os dados de origem de diversas fontes fazendo uso de *Wrappers*, pode atuar como um software *middleware* em *N* níveis de latência, na recepção das fontes de dados. Assim, a partir de uma estrutura definida como *Data Highway*, cada nível é alimentado com dados do nível anterior. No último nível, os dados são agregados e integrados para serem inseridos no repositório *Data Lake,* de forma que esse repositório facilite o pré-processamento e permita a integração dos dados para a carga, na estrutura de *Data Warehouse*. Logo, faz uso de um modelo de dados multidimensional estrela.

Segundo Zhou (2003) a mineração de dados possui uma natureza multidisciplinar e que, por esta razão, recebe diversas contribuições das comunidades de banco de dados, aprendizado de máquina, estatística, recuperação da informação, visualização de dados, computação distribuída e paralela, sendo que das três primeiras comunidades são recebidas as principais contribuições. A mineração de dados pode utilizar-se de técnicas de banco de dados, de aprendizado de máquina e estatística e, em cada abordagem, há aspectos importantes que devem ser observados, tais como: a eficiência, eficácia e validade, respectivamente. Dessa forma, esse autor considerou que o processo de mineração de dados é bem-sucedido quando todos esses aspectos forem levados em conta.

Por outro lado, a partir da disponibilidade de informações Wu e colaboradores afirmaram que quando as mesmas são de diferente natureza necessitam de uma etapa de integração para que possam ser utilizadas conjuntamente. Tal integração trata-se da fusão de informações de fontes heterogêneas, a partir de diferentes representações conceituais, contextuais e tipográficas. O conjunto de dados fundidos é diferente de um grande conjunto combinado e os pontos no conjunto de dados fundidos contêm atributos e metadados que podem não ter sido incluídos para os mesmos nos conjuntos dos dados originais. A integração é utilizada tanto na mineração de dados quanto na consolidação de dados de recursos não estruturados ou semiestruturados, o que se refere a representações textuais de conhecimento e que também podem ser aplicadas ao conteúdo *rich media* (WU et al., 2013).

Este trabalho apresenta um modelo para estruturação de bases de dados digitais para fins de gestão de risco agrícola.

2. Materiais e Métodos

Foi utilizado o conceito de arquitetura em nuvem de alto desempenho, distribuída e considerado o uso potencial de recursos do processamento paralelo.

As bases de dados envolvem tanto dados de origem privada (imagens de drones, de aeronaves tripuladas e de satélites) como dados de origem pública (verdade de campo, dados meteorológicos, previsão do tempo, solos e de outros sensores).

2.1. Descrição do modelo teórico

A elaboração do modelo teórico foi estabelecida considerando-se inicialmente a coleta dos dados de diversas fontes públicas que são providas pelo governo ou órgãos de controle, e de fontes privadas (a serem adquiridas), considerando os requisitos de qualidade de dados, de acordo com o contexto agrícola. Ao saber que os dados coletados são estruturados, semiestruturados e não estruturados, estes são armazenados em uma estrutura de *Data Lake* para facilitar a manipulação e integração no processo de carga na estrutura do *DW*. A estrutura de *DW* prevê a construção de repositórios para dados divididos por assunto, chamados de *Data Mart.* Cada *Data Mart* receberá do *Data Lake* os dados referentes ao assunto a ser tratado. Após esses repositórios serem populados, esses dados serão carregados em um repositório central, no *DW*, mantendo um histórico. Os dados mais específicos, de cada assunto, e recentes deverão ficar armazenados nos *Data Marts*, cujas cargas para o repositório central podem ser realizadas, a qualquer tempo, ou programadas, conforme interesse. Assim, estes dados devem ser submetidos a um processo de preparação, que consiste em trabalhar os dados nos algoritmos de mineração, baseados na abordagem de aprendizado de máquina, a fim de se conseguir resultados mais adequados, frente à necessidade do uso de *Big Data*, pois trata-se de uma análise diferenciada, de cunho mais complexo. Essa etapa de preparação é relevante, pois se os dados não forem preparados adequadamente a extração do conhecimento e análise do problema podem ficar comprometidas. O término da etapa de preparação dos dados dará início ao refinamento, utilizando-se de filtros específicos para trabalhar a qualidade dos dados, que deve ser pautada na seleção adequada dos requisitos de qualidade, diante dos requisitos a serem considerados. De posse dos dados filtrados, na etapa de qualidade dos dados estruturados, será aplicado um algoritmo que efetuará o processo de fusão dos dados para atuar na normalização dos dados processados. Neste momento, os dados estão organizados e seu armazenamento é feito em um vetor de dados para a utilização em um modelo de decisão.

2.2. Delineamento experimental

A operacionalização do modelo concebido é realizada tomando por base a arquitetura customizada apresentada na Figura 1.

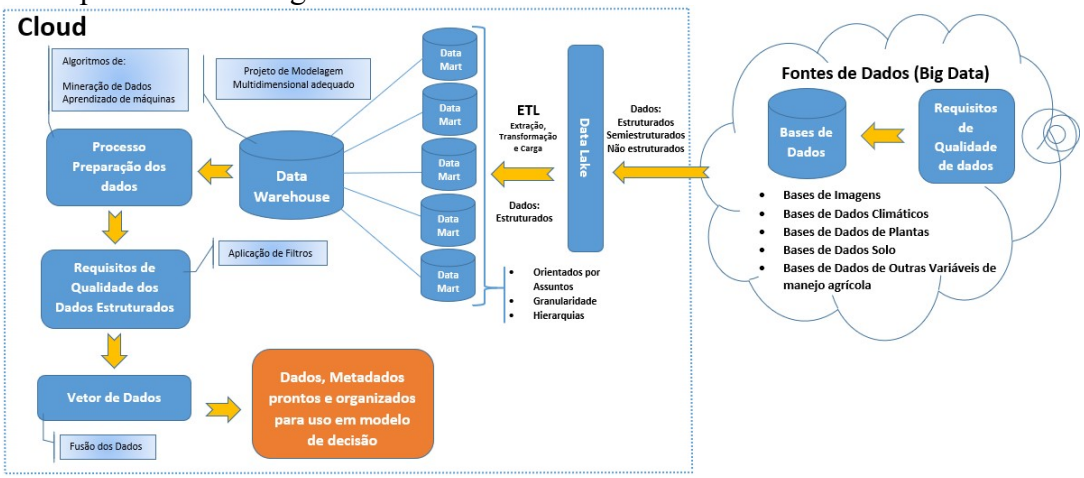

Figura 1. Arquitetura para estruturação de bases de dados digitais para operacionalização do modelo

De acordo com a Figura 1, o conjunto de fontes de dados representado pelas Fontes de Dados (*Big Data*) é composto por bases de dados para atuar na gestão de risco agrícola. Tais bases são carregadas em uma estrutura de *Data Lake*, definida por um repositório que recebe dados brutos de *Big Data* e que também, neste momento inicial, é utilizado para efetuar o tratamento da qualidade dos dados. A partir dessa primeira etapa, com os dados já tratados, faz-se uma seleção dos dados por assunto e sua carga é feita nos repositórios *Data Mart*. Os dados serão carregados no repositório de *DW,* por meio de um modelo de dados multidimensional, projetado para receber os dados para exploração. Com o *DW* alimentado com dados, os algoritmos de mineração de dados, com abordagem em aprendizado de máquinas, podem ser aplicados. De posse do resultado processado pelos algoritmos os dados são submetidos à etapa de filtros, para trabalhar a qualidade dos dados estruturados.

3. Resultados e Discussão

O resultado deste trabalho foi concebido a partir da configuração estabelecida pelo modelo para a organização de uma arquitetura estruturada e sua operação com diferentes bases de dados. A Figura 2 apresenta o modelo concebido utilizando o conceito de Diagrama *UML* (*Unified Modeling Language*), uma vez que essa linguagem facilita o entendimento do mesmo, oferecendo clareza e objetividade para o desenvolvimento da documentação de projetos associados à gestão de risco e envolve o uso de objetos, descrição de comportamentos e coesão das fases envolvidas no modelo.

Essas características podem ser notadas a partir da estrutura de armazenamento *Data Lake*, pois tal estrutura trabalha com dados brutos de *Big Data*, oferecidos por sensores de diversas naturezas conectados na WEB. Ao capturar informações online, referentes às verdades de campo, tem-se a possibilidade de alimentar sistemas de apoio à decisão em tempo real, com um mínimo de latência. Após essa coleta dinâmica dos dados aliada o uso agregado do *know-how* de processos de análise, por meio do *DW*, o qual agrega valor ao resultado final.

No contexto apresentado na Figura 2, tem-se bases de dados organizadas a partir do modelo customizado ao problema a ser tratado, que atendem, especificamente, requisitos estabelecidos por usuários para gerir riscos agrícolas.

Por fim, com o uso de algoritmo de Mineração de Dados, baseado na abordagem de Aprendizagem de Máquinas, é possível prever padrões de riscos agrícolas à priori, a partir da avaliação de series temporais e treinamentos com base nas variáveis envolvidas, tais como: imagens, dados climáticos, estado atual das plantas, qualidade de solo e outras variáveis de interesse para o manejo agrícola.

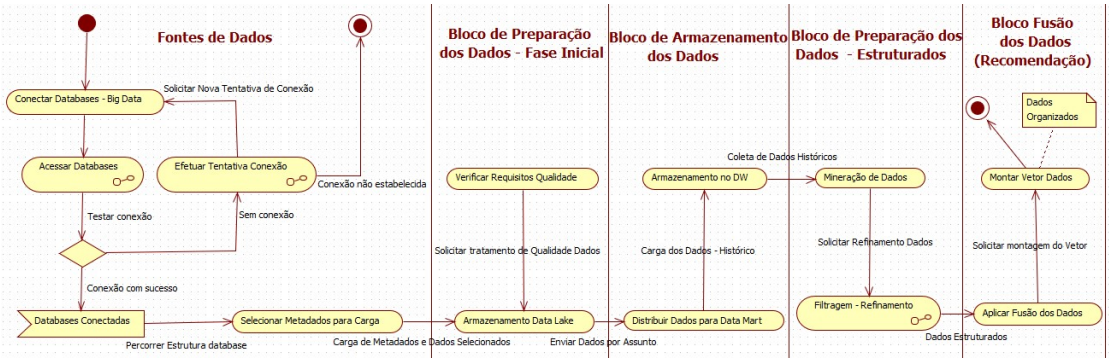

Figura 2. Diagrama UML para o Modelo de Estruturação das Bases de Dados Digitais

Este aspecto traz ao resultado oportunidades de análises inteligentes que podem ser disponibilizas via WEB, visando agregar valor ao processo de tomada de decisão no ambiente agrícola.

4. Conclusões

Este trabalho apresentou um modelo e sua arquitetura operacional para estruturação de bases de dados digitais para apoio à tomada de decisão, tendo como foco o uso de *Big Data* e arquiteturas planejadas para atuação em ambiente de Nuvem. Foram abordadas também estratégias, tais como: o uso de algoritmos de mineração, baseados na abordagem de aprendizagem de máquina para uso na etapa de preparação de dados, aplicação de filtros baseados em requisitos para uso na etapa de qualidade dos dados estruturados e, como recomendação, o uso da técnica de fusão de dados para auxílio à normalização dos dados e seus formatos. A validação da arquitetura para estruturação de bases de dados digitais utilizou Diagrama de Atividades *UML* como elemento principal, de forma a facilitar o entendimento e a sua implementação.

Agradecimentos

Este trabalho recebe apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Processo 17/19350-2, via convênio IBM Brasil e Embrapa, assim como da UFSCar junto ao Programa de Pós-Graduação em Ciência da Computação.

Referências

ALEXANDRA MARIA IOANA, F.; VLAD, D.; RAMONA, B. Data integration approaches using ETL. **Database Systems Journal**, n. 3, p. 19-27, 2016.

ALEKSEEV, A. et al. Efficient data management tools for the heterogeneous big data warehouse. **Physics of Particles and Nuclei Letters**, v. 13, n. 5, p. 689-692, 2016.

CHEVALIER, M. et al. Implementing multidimensional data warehouses into NoSQL. 2015.

HARINARAYAN, V.; RAJARAMAN, A.; ULLMAN, J. D. Implementing Data cubes Efficiently. In: ACM SIGMOD INTERNATIONAL CONFERENCE OF MANAGEMENT OF DATA, Montreal, Canada, p. 205 -216, 1996.

HU, P. The cooperative study between the hadoop big data platform and the traditional data warehouse. **Open Automation and Control Systems Journal**, v. 7, n. 1, p. 1144-1152, 2015.

JUKIĆ, N. et al. Augmenting Data Warehouses with Big Data. **Information Systems Management**, v. 32, n. 3, 2015.

KIMBALL, R.; ROSS, M. The data warehouse toolkit: **The definitive guide to dimensional modeling**. John Wiley & Sons, 2013.

SINGH, H. **Data Warehouse: concepts, technologies, implementations and management**. Saddle River: Prentice Hall PTR, 1998.

SOLODOVNIKOVA, D.; NIEDRITE, L. An Approach to Handle Big Data Warehouse Evolution, 2018.

SOSINSKY, B.: **Cloud Computing Bible**, 1st edn. Wiley, New York, 2011.

WU, H.; WU, H.; SENG, D.; FANG, X.; XU, H. **Application of information fusion technologies for multi-source data**. p. 560-564, 2013.

ZHOU, Z.-H. Three perspectives of data mining. **Artificial Intelligence Journal**, p.139–146, 2003.