MÉTODOS PARA MANIPULAÇÃO DE BANCO DE DADOS NÃO RELACIONAIS

Filipe Quina Pacheco[1]
Robson Leandro dos Santos[2]
Lucas Fagundes Teixeira[3]
Luiz Cláudio Afonso dos Santos[4]
Evaldo de Oliveira da Silva[5]

RESUMO

Este trabalho aborda os grupos de estudos realizados entre os anos de 2017 e 2018 pelos cursos graduação em Engenharia de Software e Sistemas de Informação do Centro de Ensino Superior de Juiz de Fora (CESJF). Os estudos feitos se relacionam com as áreas de Engenharia de Software e Banco de Dados a fim de enriquecer o conhecimento dos alunos acerca da manipulação de dados não relacionais. Inicialmente, é feita a fundamentação teórica envolvida na execução dos grupos de estudos. Em seguida, são descritas as técnicas utilizadas para manipulação de diferentes tipos de dados, em formato texto, binário e também dados semânticos. Por fim, são apresentados os resultados obtidos na realização dos grupos de estudos.

Palavras-chave: Engenharia de software. Banco de dados. Análise de dados. Web semântica. Dados não relacionais.

1 INTRODUÇÃO

Este artigo descreve as tecnologias e aplicações desenvolvidas para manipulação de diferentes formatos de dados que podem ser armazenados ou manipulados por aplicações computacionais. Baseia-se nos estudos sobre métodos e técnicas implementadas a partir dos trabalhos desenvolvidos por

[1] Discente do Curso de Sistemas de Informação do Centro de Ensino Superior de Juiz de Fora – CES/JF. E-mail: filipe.pacheco00@gmail.com
[2] Discente do Curso de Sistemas de Informação do Centro de Ensino Superior de Juiz de Fora – CES/JF. E-mail: robsousantos@hotmail.com
[3] Discente do Curso de Engenharia de Software do Centro de Ensino Superior de Juiz de Fora – CES/JF. E-mail: lucas_fagundes_teixeira@hotmail.com
[4] Discente do Curso de Engenharia de Software do Centro de Ensino Superior de Juiz de Fora – CES/JF. E-mail: luiz.santos89@yahoo.com.br
[5] Docente do Curso de Engenharia de Software e Sistemas de Informação do Centro de Ensino Superior de Juiz de Fora. E-mail: evaldosilva@cesjf.br

três grupos de estudos realizados no CESJF entre os anos de 2017 e 2018, com a participação de alunos das graduações em Engenharia de Software e Sistemas de Informação desta mesma instituição.

Nesta introdução será feita uma breve descrição acerca das propostas dos grupos de estudos. Tais propostas pretenderam enriquecer os conhecimentos abordados nos conteúdos das disciplinas de Engenharia de Software e Banco de Dados existentes nas grades curriculares dos cursos de graduação mencionados. Deste modo, os grupos de estudos que serviram de fundamentação para este artigo são citados resumidamente a seguir:

- Desenvolvimento de Aplicações para Ciência de Dados. A ciência de dados pode ser considerada uma área emergente que relaciona com a necessidade de integrar a necessidade de gestão empresarial, métodos estatísticos e tecnologias da informação, e que pode impactar em uma mudança de paradigma do desenvolvimento de aplicações de software. Deste modo, este grupo de estudos permitiu aos envolvidos a ampliação dos conhecimentos em computação no que diz respeito ao uso de novas técnicas e tecnologias para extração, transformação e carga de dados com base em rotinas de desenvolvimento utilizando a linguagem Python. As aplicações implementadas permitiram aprimorar o conhecimento em computação com um tema atual e de extrema necessidade para o desenvolvimento dos cursos envolvidos neste grupo de estudos.
- Gestão do Conhecimento de Base de Dados Corporativas. Este grupo de estudos teve como objetivo abordar as tecnologias que são utilizadas para manipular ontologias e formatos de dados que possam servir como representações semânticas. Uma vez que ontologias podem ser usadas como base de conhecimento na área da Computação, o grupo de estudos realizou as pesquisas bibliográficas com base nas áreas da Web Semântica e representação do conhecimento, com o objetivo de desenvolver rotinas de software para processar e visualizar ontologias representadas por linguagens computacionais. Foi proposto um protótipo

para permitir a geração de dados representados semanticamente a partir de ontologias armazenadas em bancos de dados.

- Serviços Web para manutenção de dados e documentos. Este foi um projeto de iniciação científica (PIC) criado a partir do interesse de pesquisar as tecnologias e métodos para manutenção de dados e documentos produzidos em organizações.

As propostas citadas se relacionam com a necessidade das organizações em manipular dados para sob várias perspectivas, tanto para extração de informações quanto para compartilhamento do conhecimento. Os grupos de estudos foram importantes para formação dos alunos envolvidos, e, principalmente, criou a oportunidade de aprenderem novas tecnologias ainda pouco discutidas dentro das organizações e setores de Tecnologia da Informação (TI).

Também é possível notar o grande crescimento do volume de dados gerados a partir de diferentes plataformas de aplicações com o objetivo de aumentar a inteligência competitiva, independentemente do tamanho das organizações (RODRIGUEZ e WANDERLEI FONTANA, 2005; STRAUSS, 2012; BARBOSA, 2006; MARINHEIRO e BERNARDINO, 2015).

A reutilização do conhecimento produzido dentro das organizações é uma área importante para melhoria e compartilhamento de informações. A conceituação dos dados permite a reutilização do conhecimento a partir de uma taxonomia de conceitos estruturados por meio de ontologias. Com base nesta proposta, foi necessária a pesquisa e implementação das rotinas para manipulação de dados semânticos seguindo as abordagens encontradas na área da Web Semântica e ontologias aplicadas à Ciência da Computação (CONEGLIAN et. al., 2017; W3C, 2017; ALMEIDA e BAX, 2003; GRUBER, 1993; GUARINO, 1998)

O banco de dados NoSQL (*Not Only* SQL) foi o tipo de tecnologia aplicada pelos grupos de estudos a fim de suprir a necessidade para o armazenamento de grandes volumes de dados extraídos de redes sociais,

documentos binários ou no formato texto. Atualmente, não somente os bancos de dados relacionais estão sendo utilizados como principal tecnologia para armazenamento de dados. Os bancos de dados NoSQL estão servindo como alternativa para armazenamento de dados (SOUSA, 2015).

Com base na discussão apresentada nesta introdução, este artigo apresenta a compilação dos métodos e tecnologias aplicadas pelos grupos de estudos citados anteriormente. Espera-se que este artigo sirva de fundamentação para novos trabalhos e de motivação para continuidade de novas aplicações nesta área.

O restante deste artigo irá descrever as técnicas utilizadas pelos grupos de estudos citados a fim de apresentar as diferentes tecnologias para manipulação de dados não relacionais. O texto está estruturado na forma que segue. A Seção 2 apresenta os conceitos e referências utilizados pelos grupos de estudos. A Seção 3 apresenta as técnicas e métodos aplicados durante o desenvolvimento dos estudos. A Seção 4 aborda os resultados e discussões acerca dos trabalhos alcançados. A Seção 5 descreve as considerações finais e trabalhos futuros.

2 REFERENCIAL TEÓRICO

Esta seção apresenta o referencial teórico e que serve para definição dos conceitos e tecnologias pesquisadas durante a realização dos grupos de estudos discutidos neste trabalho.

2.1 TECNOLOGIAS DE DESENVOLVIMENTO APLICADA À CIÊNCIA DE DADOS

A proposta do grupo de estudos referente ao Desenvolvimento de Aplicações para Ciência de Dados, foi a aplicação de tecnologias para a análise de dados extraídos a partir de formatos diferentes formatos, tanto no

formato de arquivo texto quanto no formato binário. Neste contexto, as pesquisas bibliográficas tiveram como requisitos a análise de tecnologias para aplicação de extração de dados a partir de dados não estruturados ou dados armazenados em documentos.

Dados não estruturados podem ser entendidos como: “qualquer dado não gerenciado pela forma padronizada oferecida pelos padrões estabelecidos pelos Sistemas Gerenciadores de Banco de Dados” (IMON, 2014). Assim, também é comum definir que os dados considerados não estruturados, que não são mantidos por um Sistema Gerenciador de Banco de Dados (SGBD), também são chamados de dados não-relacionais. Além disso, os SGBDs oferecem grande dificuldade para manter grandes volumes de dados não estruturados, principalmente quando se trata de dados oriundos da Web com o grande número de usuários manipulando vários tipos de informações (AMAZON, 2017).

Além das pesquisas realizadas para o armazenamento de dados não relacionais, também foram pesquisadas as tecnologias de desenvolvimento para a manipulação destes tipos de dados.

Incialmente, a linguagem Java foi escolhida devido ao conhecimento prévio dos alunos dos grupos de estudos por meio das disciplinas lecionadas nos cursos de graduação citados na Introdução deste trabalho. Porém, a linguagem Java possui limitações e a curva de aprendizado maior que outras linguagens (PUTANO, 2018; LO, LIN e WU, 2015). Como alternativa optou-se por pesquisar a linguagem Python, a qual tem sido amplamente utilizada em projetos de ciência de dados (BRUNNER e KIM, 2016).

Python é uma linguagem de programação de uso geral, e que tem liderado as iniciativas de Análise de Dados com base nas seguintes características: facilidade de aprendizado, pode ser implementada em diferentes plataformas, possui bibliotecas para uso de cálculos e funções estatísticas, diversas bibliotecas e módulos prontos para uso, sendo uma ferramenta totalmente gratuita (DATA SCIENCE ACADEMY, 2018).

Algumas diferenças podem ser vistas entre as linguagens Java e Python. O Java é uma linguagem que exige a definição dos tipos de dados, sendo considerada fortemente tipada. Por outro lado, o Python é dinamicamente tipada sendo uma linguagem não compilada. Além disso, a linguagem Java possui grande dificuldade para utilização dos ambientes de desenvolvimento, onde o desenvolvedor terá que configurar a JDK (*Java Development Kit*) e garantir que conhece profundamente o paradigma da orientação a objetos. Até mesmo o programa mais simples, por exemplo, "*Hello World*" requer uma estrutura, compilação e execução de classe Java. Com o Python, o processo é mais simplificado. Por exemplo, é possível encontrar em sistemas operacionais o ambiente de desenvolvimento do Python já instalado, sendo mais simplificada também a configuração deste ambiente, podendo utilizar apenas o *prompt* de linha de comando para a execução de linhas de código (UOPEOPLE, 2018).

2.2 ONTOLOGIA NA COMPUTAÇÃO COMO MECANISMO DE ARMAZENAMENTO DO CONHECIMENTO

Embora o termo ontologia seja amplamente utilizado na área de Ciência da Computação e subáreas, não existe um consenso sobre seu exato significado. Porém, o termo ontologia está relacionado à conceituação de entidades de um determinado domínio, permitindo que seja constituída uma coleção de classes de conceitos e o relacionamento entre as mesmas. Uma conceituação é uma visão abstrata e simplificada do mundo que se deseja representar (ALMEIDA e BAX, 2003). Gruber (1993) define uma ontologia como uma descrição de conceitos e relacionamentos que podem existir por meio de um agente ou uma comunidade de agentes. Esta definição está relacionada com o uso da ontologia como uma coleção de definições, usada inclusive na filosofia.

Na Ciência da Computação há um consenso do uso de ontologias para permitir que o conhecimento de um domínio seja compartilhado e reutilizado, a partir de vocabulários e coleções de conceitos. Desta forma, surge a necessidade de manipulação dos dados e informações armazenadas nestas estruturas.

Além das definições acima, o uso de ontologias em Inteligência Artificial (IA) como forma de compartilhamento do conhecimento permite o acesso a vocabulários para elaboração de perguntas e consultas, respeitando os conceitos e as relações existentes entre os mesmos. Em IA são desenvolvidos agentes de software capazes de acessar a ontologia, onde os conceitos e suas relações podem ser compartilhados permitindo a troca do conhecimento (GUARINO, 1998).

Os dados contidos em uma ontologia podem ser especificados computacionalmente por meio da linguagem OWL (*Ontology Web Language*). OWL é uma linguagem usada para descrever ontologias e foi projetada pelo *Web Ontology Working Group*, que é o grupo responsável pelo desenvolvimento de linguagens para especificação de ontologias para a Web Semântica (OWL, 2018).

A linguagem OWL permite a representação do conhecimento determinando conceitos, propriedades e suas relações definidas em um domínio de aplicação. Cada conceito em OWL é descrito como uma classe, e cada classe é descrita em um frame. De forma mais detalhada, as classes são definidas por atributos (*slots*) e estruturadas de acordo com o relacionamento com as subclasses constituindo um grafo ou uma taxonomia de conceitos.

2.3 WEB SEMÂNTICA E *LINKED DATA*

A Web Semântica é definida por Tim Bemers-Lee como uma extensão da Web obtida através da adição de semântica ao atual formato de representação de dados, tornando assim as informações existentes na web compreensíveis

tanto para humanos quanto para máquinas, facilitando o relacionamento e o acesso à informação (BERNERS-LEE, 2009).

A Web Semântica permite a criação de uma comunicação mais colaborativa entre homem e máquina por meio do uso de tecnologias utilizadas para oferecer significados aos dados processados pela web. Além disso, permite que softwares possam auxiliar de forma mais eficiente às atividades de busca de informação pelos humanos, atribuindo significados e conhecimento aos dados processados pelas aplicações de software. Assim torna-se possível organizar e compreender as relações entre os termos processados e minimizando as perdas durante pesquisas (W3C, 2017).

O modelo de web predominante atualmente proporciona uma publicação descontrolada de conteúdos e informações na rede nas quais em sua maioria as informações contidas nas publicações não são descritas de maneira correta, o que causa uma sobrecarga de informações descontextualizadas no retorno das buscas por conteúdo.

A proposta da Web Semântica também é reduzir ou eliminar os problemas estruturais do atual modelo de disponibilidade de dados na web sugerindo padrões onde as informações possam ser descritas e categorizadas, a fim de eliminar ambiguidades nas buscas por conceitos pelos “motores de busca” (ou *search engines*). Para que essa ideia se tornasse realidade, a Web Semântica passou a adotar o uso de ontologias e as tecnologias para representa-las computacionalmente, possibilitando o surgimento de outros conceitos da Web Semântica, como, por exemplo, os dados vinculados semanticamente (ou *Linked Data*).

Linked data é um conjunto de práticas para aprimoramento da web de dados na elaboração da Web Semântica para criar relacionamentos entre os dados disponíveis na rede de dados. A Web Semântica não consiste apenas em colocar dados na Web, trata-se de fazer links, para que uma pessoa ou máquina possa explorar a rede de dados. Assim, com os dados vinculados,

torna-se possível encontrar de forma mais facilitada os dados relacionados e com mesmos significados.

Na Web de dados este relacionamento é feito através dos links (ou hipertextos) em HTML que ligam os dados através de um documento HTML a outro. Por outro lado, na Web Semântica existe uma ligação entre os dados por meio de uma estrutura que define como o dado deve estar relacionado, que são feitas através do uso de URIs (*Uniform Resource Identifier* - identificador uniforme de recursos). Nestas estruturas são encontradas as descrições dos recursos, possibilitando assim a compreensão de seu significado descrito em OWL ou também em RDF (*Resource Description Framework*) (DAVIS, 2013). Para isto Tim Berners-Lee também sugere quatro regras a elaboração de projetos semânticos para a web, são elas:

- Uso de URIs.
- Uso de HTTP URIs para que as pessoas possam procurar esses nomes.
- Por meio de uma URI são fornecidas informações úteis usando os padrões (RDF).
- Inclusão de links para outros URIs para que possam descobrir mais coisas.

Para a implementação do acesso aos dados descritos em OWL ou RDF, foi pesquisado o *framework* Jena (APACHE JENA, 2017). O Jena é escrito na linguagem Java, sendo também amplamente utilizado pela comunidade acadêmica. Além dessa característica, o Jena possui vasto volume de material de pesquisa e exemplos de aplicações que o utilizam desde sua criação, o qual é fornecido pela empresa Apache. Devido a estas características e a familiaridade com a linguagem Java, o grupo de estudos sobre Gestão do Conhecimento de Base de Dados Corporativas, optou pelo *framework* Jena como plataforma de desenvolvimento do acesso aos dados em ontologias.

2.4 BANCO DE DADOS NOSQL

O MongoDB foi o banco de dados não relacional pesquisado nos grupos de estudos para servir como repositório de armazenamento dos diferentes tipos de dados trabalhados ao longo dos trabalhos desenvolvidos.

O MongoDB é um sistema gerenciador de banco de dados documental NoSQL, escrito em C++, sob licença GNU AGPL (*Affero General Public License*) versão 3.0, que armazena dados dentro de documentos no formato BSON (Binary JSON), uma versão "binária" do JSON (*JavaScript Object Notation*). Este banco de dados foi criado por Dwight Merriman (ex-Fundador do DoubleClick e CTO) e Eliot Horowitz (CTO 10gen e cofundador) baseados em suas experiências de dados em grande escala, alta disponibilidade e sistemas robustos. (MONGODB, 2018).

De acordo com Cross et. al. (2018) a capacidade de armazenar objetos no formato JavaScript nativamente do MongoDB, permite reduzir o tempo de processamento tornando a manipulação de dados mais rápida. Em vez de uma linguagem específica de domínio como o SQL, o MongoDB utiliza uma interface JavaScript simples para realizar consultas. Consultar um documento é tão simples como passar um objeto JavaScript que descreve parcialmente o destino da pesquisa.

3 METODOLOGIA

Esta seção reúne os métodos pesquisados para a manipulação dos diferentes formatos de dados pesquisados durante a realização dos grupos de estudos.

3.1 MANIPULAÇÃO DE DADOS A PARTIR DE ONTOLOGIAS

Os procedimentos apresentados nesta seção descrevem as técnicas de acesso às bases de conhecimento implementadas com base em ontologias especificadas na linguagem OWL. As técnicas foram aplicadas a partir da necessidade levantada pelo Grupo de Estudos sobre Gestão do Conhecimento de Base de Dados Corporativas.

Os métodos de acesso aos dados foram implementados por meio do *framework* Jena (APACHE JENA, 2017), e seguem descritos abaixo:

- Instanciação do modelo da ontologia
- Criação do modelo em memória
- Leitura das informações existentes em uma ontologia
- Acesso aos dados das classes e conceitos existentes em uma ontologia

3.1.1 Métodos de acesso às informações contidas em uma Ontologia

Dentre os principais métodos utilizados pelo *framework* Jena, para manipulação de ontologias, é necessário salvar o código escrito no formato OWL na memória. Para acesso à estrutura da ontologia utiliza-se o método `createOntologyModel(model)`, de acordo com o exemplo visto abaixo:

```
OntModel m = ModelFactory.createOntologyModel(model);
```

Para criar uma especificação de modelo em OWL de forma personalizada, é possível que o desenvolvedor crie uma nova especificação a fim de acessar os métodos de atribuição de valores ou para acessar as características de uma ontologia. Abaixo segue um exemplo de criação de uma especificação de ontologia (APACHE, 2017):

```
OntModelSpec s = new OntModelSpec( OntModelSpec.OWL_MEM );
s.setDocumentManager( myDocMgr );
OntModel m = ModelFactory.createOntologyModel(s);
```

A classe `OntModelSpec` é acessada como um tipo Enum pelo Java, independentemente do modelo de ontologia e o contexto a qual foi criado. Conforme a necessidade deste trabalho foi utilizada o objeto `OntModelSpec.OWL_MEM`, como forma de acessar toda a especificação do modelo descrito pela linguagem OWL. Desta forma, é possível armazenar e acessar o modelo da ontologia carregado na memória.

Além das instruções acima, existem métodos que tornam possível o acesso aos dados das entidades existentes nas ontologias. A partir de uma URI associada ao documento OWL lido, os seguintes métodos podem ser utilizados da seguinte forma:

```
read( String url )
read( Reader reader, String base )
read( InputStream reader, String base )
read( String url, String lang )
read( Reader reader, String base, String Lang )
read( InputStream reader, String base, String Lang )
```

Para que seja possível percorrer as classes de uma ontologia descrita em OWL, o *framework* Jena disponibiliza um método com base no padrão de projeto chamado *Iterator*, permitindo assim a leitura, escrita e exclusão das classes. Abaixo segue um exemplo da declaração do método `ExtendedIterator` que permite acesso às classes de uma ontologia:

```
public ExtendedIterator<OntClass> listClasses();
```

Ao instanciar o `ExtendedIterator` é possível acessar as informações de todas as classes por meio do método `listClasses()` da interface `OntModel` conforme visto abaixo:

```
ExtendedIterator classes = inf.listClasses();
```

Após instanciar a lista de classes pelo método anterior e com o auxílio de uma estrutura de repetição da própria linguagem Java, pode-se percorrer o modelo de uma ontologia usando método `next()` que é descrito pela interface `ExtendedIterator`, conforme visto no exemplo a seguir:

```
OntClass essaClasse = (OntClass) classes.next();
```

Um ponto a ser mencionado é a possibilidade de usar um método da interface `ExtendedIterator` que é o `hasNext()`, o qual retorna o valor `True` caso o `ExtendedIterator` possua resultados para serem percorridos. Esta estrutura pode ser vista conforme o exemplo abaixo:

```
      while (classes.hasNext()) {
      // aqui será feita a lógica de acordo com o contexto de
      cada caso.
}
```

3.2 MANIPULAÇÃO DE DADOS NO FORMATO EM RDF UTILIZANDO O MONGODB

A necessidade da manipulação de dados no formato em RDF foi levantada a partir dos estudos desenvolvidos pelo Grupo de Estudos sobre Gestão do Conhecimento de Base de Dados Corporativas.

Os documentos no formato RDF são utilizados como tecnologias para armazenamento da representação de dados gerados a partir dos modelos de ontologias. Os dados armazenados neste formato são armazenados em documentos com a extensão RDF.

Considerando que estes dados podem ser armazenados em arquivos no formato de texto, a integridade deste tipo de dado pode ser ameaçada, já que o

documento pode ser corrompido durante seu processo de manipulação dos dados. Além disso, o processamento e indexação das informações por meio de sistemas de arquivo podem ser dificultados, criando um obstáculo para recuperação de dados e informações sobre os dados no formato RDF. Com base nestas características, o gerenciamento de dados no formato RDF foi elaborado a partir do banco de dados MongoDB. É importante destacar que o MongoDB permite o armazenamento de dados no formato de documentos, permitindo o desenvolvimento de rotinas em diferentes ambientes de desenvolvimento para a recuperação de dados por meio deste formato.

O MongoDB tem sido utilizado como gerenciador de dados pelos autores destes trabalhos em grupos de pesquisa que tratam da extração de dados não estruturados, o que facilitou a escolha por esta tecnologia. Por meio do JSON o MongoDB suporta os seguintes tipos de dados:

- Null - Vazio.
- Boolean - True ou False.
- Number - Número com sinal que pode ter uma notação com E Exponencial.
- String - Uma sequência com um ou mais caracteres.
- Object - Array não ordenado com itens do tipo chave valor, onde todas as chaves devem ser Strings distintos no mesmo objeto.
- Array - Lista ordenada de qualquer tipo, inteira entre colchetes e com elementos separados por vírgula.

Porém o JSON não implementa uma padronização para o formato de data nem como trabalhar com dados binários. Por este motivo o MongoDB introduz o BSON (*Binary* JSON) que é uma extensão ao JSON que além dos formatos de dados suportados por este também comporta:

- MinKey , MaxKey , TimeStamp - Tipos utilizados internamente no MongoDB.
- BinData - array de bytes para dados binarios.
- ObjectId - Identificador unico de registros do MongDB.

- Date - Representação de data.
- Expressões regulares.

Com base na variedade de tipos de dados suportados pelo MongoDB, foi possível estabelecer a manipulação de dados no formato RDF de acordo com os seguintes requisitos:

- Armazenar estruturas contidas em ontologias RDF/XML.
- Permitir a revisão das estruturas contidas em uma Ontologia RDF/XML.
- Compartilhar o conhecimento armazenado na base de dados com outras aplicações a fim de mineração de dados para estudos de ciência de dados.

A partir das necessidades apresentadas foi criado um protótipo de aplicação em Java para manipulação de dados RDF armazenados no formato JSON pelo Mongo. As estruturas abaixo são aplicadas no protótipo para a compreensão de como os dados representados em RDF podem ser armazenados pelo MongoDB:

- Documento / Document - Estrutura onde são armazenadas as informações. Em um documento pode existir um valor simples ou uma lista de valores. Esta estrutura pode ser comparada as linhas e colunas contidas em uma tabela de uma base de dados relacional.
- Coleção / collection - Estrutura onde são agrupados os documentos criados. Esta estrutura pode ser comparada a uma tabela de uma base de dados relacional.
- Base de dados / database - Estrutura onde são armazenadas um conjunto de coleções. Estes se comparam aos bancos de dados relacionais.

3.3 MANIPULAÇÃO DE DOCUMENTOS UTILIZANDO O BANCO DE DADOS MONGODB

As técnicas de manipulação e armazenamento de dados descritas a seguir foram usados para a construção do protótipo da aplicação definida por meio do grupo de estudos chamado Serviços Web para manutenção de dados e documentos, tendo a contribuição do grupo referente ao Desenvolvimento de Aplicações para Ciência de Dados, o qual pode compartilhar os levantamentos feitos sobre o armazenamento de documentos no MongoDB.

Esta seção também caracteriza a proposta do trabalho por meio da implementação de uma aplicação móvel, e que também pode ser acessada via plataforma web, para manipulação de documentos utilizando o banco de dados MongoDB.

Como foi descrito no referencial teórico, é possível a criação de aplicações que tenham como propósito a manipulação de grandes volumes de dados. Também optou-se pelo banco de dados MongoDB como plataforma para armazenamento no formato BSON, uma vez que a proposta é a apresentação de uma solução que possa manipular dados no formato binário, definido como GridFS. No MongoDB o GridFS é uma especificação para armazenamento e recuperação de arquivos no formato BSON que excedam o limite de 16 megabytes (GRIDFS, 2018).

Visando flexibilizar a manipulação dos dados em documentos, o grupo de estudos sobre Serviços Web para manutenção de dados e documentos, realizou o desenvolvimento de um protótipo que permitiu o acesso às diversas funcionalidades tanto em aplicativos móveis, quanto em ambiente web desktop, seguindo assim a necessidade de acesso em multiplataforma.

Devido a necessidade de explorar o compartilhamento e integração de dados em documentos, o projeto deste protótipo iniciou-se a partir de estudos para criação de serviços web em PHP visando a manipulação de documentos. O PHP (*Hypertext Preprocessor*) é uma linguagem de programação para web

de script open-source, trabalha mesclado ao HTML (*Hypertext Markup Language*) e é executado no lado servidor, o que possibilita que o site seja dinâmico (PHP, 2018). Neste projeto, o serviço web para manipulação de dados de documentos é desenvolvido em PHP.

Para a aplicação da proposta deste trabalho utilizou-se a técnica de webview para construção de interface, a fim de facilitar o acesso tanto em smartphones quanto em ambiente desktop por meio de browsers (CHARLAND, 2018).

O nome dado ao protótipo citado neste trabalho é NoSQLMobile. As funcionalidades básicas utilizadas para implementação do NoSQLMobile são vistas abaixo:

- InserirCSV. Esta funcionalidade permite inserir um arquivo no formato texto, e que visa oferecer coleções de dados armazenados no banco de dados MongoDB para extração, transformação e carga de dados para geração de gráficos. Esta funcionalidade é um exemplo da manipulação de grandes volumes de dados, porém através de arquivos no formato texto.
- Inserir GridFS. Esta funcionalidade permite inserir documentos de qualquer tipo de formato.
- Collections. Serve para listar as coleções criadas para armazenar os dados de documentos armazenados no formato BSON.
- GridFS. Esta opção permite a consulta dos documentos manipulados, tanto pela funcionalidade de “InserirCSV” ou “Inserir GridFS”.
- Análise. Permite consultar os dados dos arquivos inseridos na funcionalidade “InserirCSV”, a fim de gerar o gráfico para análise dos dados armazenado no arquivo selecionado.

4 RESULTADOS E DISCUSSÃO

Esta seção apresenta um resumo dos resultados alcançados por cada grupo de estudos. Tais resultados estão registrados no Setor de Pesquisa e Extensão do CESJF e podem ser consultados publicamente.

Ao longo da realização dos grupos de estudos os alunos puderam aprofundar nas técnicas abordadas neste artigo, permitindo a geração dos resultados obtidos a partir de diferentes atividades e trabalhos acadêmicos descritos nas seções seguintes.

4.1 DESENVOLVIMENTO DE APLICAÇÕES PARA CIÊNCIA DE DADOS

Os alunos participantes desenvolveram o minicurso denominado “Tecnologias para BigData” na XXVI Semana de Informática e Engenharia de Software. Além disso produziram material para a apresentação da palestra “Grupos de Estudos em Engenharia de Software e Sistemas de Informação: O uso de tecnologia para Ciência de Dados”. A palestra teve objetivo de apresentar de forma prática as rotinas desenvolvidas pelo grupo de estudos a partir das soluções aplicadas em Python e MongoDB.

4.2 GESTÃO DO CONHECIMENTO DE BASE DE DADOS CORPORATIVAS

Com a realização deste grupo de estudos foram publicados 3 trabalhos acadêmicos, sendo 2 artigos no formato de Trabalho de Conclusão de Curso (TCC) e 1 resumo.

Os TCCs foram publicados no Caderno de Sistemas de Informação dos cursos de Engenharia de Software e Sistemas de Informação, v. 1, n. 2 (2017) disponível no link https://seer.cesjf.br/index.php/cesi/issue/view/68. Os títulos dos trabalhos foram os seguintes:

- Armazenamento e manipulação da representação semântica de dados utilizando as tecnologias Java e MongoDB, trabalho defendido pelo aluno Robson Leandro dos Santos.
- Desenvolvimento de uma ferramenta para anotação semântica utilizando o framework JENA baseado em ontologias de domínio, trabalho defendido pelo aluno Filipe Pacheco Quina.

Estes trabalhos de TCC apresentaram também a implementação de protótipos que mostraram a aplicação das técnicas pesquisadas.

O resumo foi publicado na Revista Anacleta, v. 3, n. 3 (2017) disponível no link https://seer.cesjf.br/index.php/ANL/issue/view/64/showToc, como resultado da apresentação do trabalho “Gestão do conhecimento de dados relacionais utilizando padrão de nomenclaturas e anotação semântica”, proferido no formato de palestra no III Seminário de Extensão e Pesquisa do CES/JF.

4.3 SERVIÇOS WEB PARA MANUTENÇÃO DE DADOS E DOCUMENTOS

A realização deste grupo de estudos permitiu a geração de um TCC no formato de artigo e aceito para publicação no Caderno de Sistemas de Informação. O TCC intitulado “Desenvolvimento de uma aplicação móvel para manipulação de documentos utilizando o banco de dados MongoDB” foi escrito pelo aluno Luiz Cláudio Afonso dos Santos.

O trabalho desenvolvido aplicou as técnicas e os requisitos descritos na Seção 3.3 deste trabalho. Apresenta ainda um protótipo desenvolvido a fim de verificar o uso de um banco de dados MongoDB. Foi desenvolvido um aplicativo mobile usando o webview Android com o objetivo de manipular documentos e dados não estruturados. Por fim, foram apresentados os resultados obtidos através da implementação da ferramenta que trata os dados para armazenamento no banco de dados e os analisa por meio da geração de gráficos e tabelas.

5 CONSIDERAÇÕES FINAIS

Os trabalhos desenvolvidos pelos grupos de estudos serviram para aprimorar o conhecimento absorvido nas disciplinas de Engenharia de Software e Banco de Dados. Além dos resultados discutidos na seção anterior, é possível evidenciar o enriquecimento dos currículos dos alunos tanto na participação de atividades acadêmicas durante o período na faculdade, quanto no aprendizado de tecnologias atuais. De forma geral, os alunos que participaram dos grupos de estudos se tornaram mais preparados profissionalmente, podendo colocar em prática um conjunto de tecnologias atuais e também aplicadas ao mercado.

Com a realização das propostas dos grupos de estudos, foi possível implementar várias técnicas e tecnologias pouco discutidas pelas disciplinas dos cursos de graduação. Os resultados apresentados neste artigo mostram que é possível aplicar diferentes tecnologias para manipulação de dados não relacionais.

Percebe-se também um crescimento grande do uso do banco de dados NoSQL por diferentes organizações que desejam manter diferentes tipos de dados armazenados. Os trabalhos apresentaram a manipulação de dados semânticos em formato RDF por meio do banco de dados MongoDB. Além dos dados no formato RDF, também foi possível a manipulação de dados no formato de documentos.

Como trabalhos futuros, espera-se o desenvolvimento de novos grupos de estudos que avancem com as tecnologias apresentadas neste trabalho a fim de criar rotinas que utilizem serviços web para leitura (ou extração), transformação e carga de dados tendo como origem as bases de dados nos formatos aqui mencionados.

ABSTRACT

This paper approaches academics study groups constituted between 2017 and 2018. these study groups were created from Software Engineering and Information System courses of Center of Learning Higher of Juiz de Fora. The studies concluded are related to the Software Engineering and Database areas in order to improve the knowledge of students regarding to manipulation of non-relational data. Futhermore, in this article is made a theorical fundamentation envolved on execution of studies realized by the groups. Following, the techinques used for manipulation of differents data types are described, including many formats, such as, text, binary and semantics. Finally, the results that were obtained are presented.

Palavras-chave: Software Engineering. Database Systems. Data Analysis. Semantic Web. Non Relational Data.

REFERÊNCIAS

ALMEIDA, M. B.; BAX, M. P. **Uma visão geral sobre ontologias:** pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. Ciência da Informação, Brasília DF, v. 32, n. 3, p. 7-20, 2003.

AMAZON. **O que é NoSQL.** Disponível em: https://aws.amazon.com/pt/nosql/.Acessoem 03 de Novembro de 2017.

APACHE. **Introduction to Apache Any23.** Disponível em: https://any23.apache.org/index.html. Acesso em 04 de novembro de 2017.

APACHE JENA. **Jena a Java RDF API and Toolkit.** Disponível em: https://www.w3.org/2001/sw/wiki/Apache_Jena. Acesso em: 04 de novembro de 2017.

BARBOSA, Ricardo Rodrigues. **Uso de fontes de informação para a inteligência competitiva:** um estudo da influência do porte das empresas sobre o comportamento informacional 10.5007/1518-2924.2006 v11nesp1p91. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 11, n. 1, p. 91-102, 2006.

BERNERS-LEE, Tim. **Linked Data (2009).** Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em 13 de Setembro de 2017.

BERNERS-LEE, Tim, James Hendler, and Ora Lassila. **"The semantic web."** Scientific American 284.5 (2001): 28-37.

BRUNNER, Robert J.; KIM, Edward J. **Teaching data science.** arXiv preprint arXiv:1604.07397, 2016.

CHARLAND, Andre; LEROUX, Brian. **Mobile application development: web vs. native.** Queue, v. 9, n. 4, p. 20, 2011.

CONEGLIAN, C. S., DIEGER, R., SEGUNDO, J. E. S. y CAPTREZ, M. (2017). **O papel estratégico da web semântica no contexto do big data**. Brasil: I Workshop de Informação, Dados e Tecnologia.
CROSS, Zach. POCHE, Aga. SANTIAGO, Daniel. SINGH, Divit. **Desenvolvendo aplicativos móveis com Node.js e MongoDB, parte 1:** Os métodos e resultados de uma equipe. Acelerando o tempo de retorno dos sistemas de engajamento. Disponível em: [https://www.ibm.com/developerworks/br/library/mo-nodejs-1/mo-nodejs-1-pdf.pdf]. Acesso em: 15/10/2018.

DATA SCIENCE ACADEMY. Python Fundamentos para Análise de Dados. Disponível em: https://www.datascienceacademy.com.br/course?courseid=python-fundamentos. Acesso em 02 dez 2018.

DAVIS, Ian; STEINER, Thomas; HORS, A. L. **Rdf 1.1 json alternate serialization (rdf/json).** W3C Working Group Note. W3C, v. 7, 2013.

GUARINO, Nicola. **Formal Ontology and Information Systems.** In: Proceedings of the First Int. Conference on Formal Ontology in Information Systems, Trento, Italy, Junho 1998.

GRIDFS. **MongoDB Documentation.** Disponível em: https://docs.mongodb.com/manual/core/gridfs/. Acesso em 03 dez 2018.
GRUBER, T. R. A Translation Approach to Portable Ontology Specifications. 1993. Disponível em <http://www-ksl.stanford.edu/KSL_Abstracts/KSL-92-71.html>. Acesso em 13 de Setembro de 2017.

IMON, W.H. **Untangling the Definition of Unstructured Data.** 2014. Disponível em : http://www.ibmbigdatahub.com/blog/untangling-definition-unstructured-data. Acesso em 08 de Novembro de 2017.

LO, Chieh-An; LIN, Yu-Tzu; WU, Cheng-Chih. **Which programming language should students learn first?** A comparison of Java and python. In: Learning and Teaching in Computing and Engineering (LaTiCE), 2015 International Conference on. IEEE, 2015. p. 225-226.

MARINHEIRO, Antonio; BERNARDINO, Jorge. **Experimental evaluation of open source business intelligence suites using OpenBRR.** IEEE Latin America Transactions, v. 13, n. 3, p. 810-817, 2015.

MONGODB. **Manual.** Disponível em: [https://docs.mongodb.com/manual/]. Acesso em: 10/10/2018.

PHP. **Manual do PHP.** Disponível em :http://php.net/manual/pt_BR/. Acesso em 22 de out de 2018.

PUTANO, BEN. **Java vs. Python:** Coding Battle Royale. Disponível em: https://stackify.com/java-vs-python/. Acesso em 02 dez 2018.
RDF4J. The Eclipse framework. Disponível em: http://rdf4j.org/about/. Acesso: 04 de novembro de 2017.

RODRIGUEZ Y RODRIGUEZ, Martius Vicente; WANDERLEI FONTANA, Edson. **Inteligência competitiva:** nível de uso e influência nas receitas nos

pequenos negócios exportadores. REAd-Revista Eletrônica de Administração, v. 11, n. 3, 2005.

SOUSA, Gonçalo da Cruz Pereira; PEREIRA, José Luís. **Document-Based databases:** estudo exploratório no âmbito das Bases de Dados NoSQL. In: 15ª Conferência da Associação Portuguesa de Sistemas de Informação. CAPSI 2015. Associação Portuguesa de Sistemas de Informação (APSI), 2015. p. 191-207.

STRAUSS, Luisa Mariele et al. **Inteligência competitiva, empresarial, estratégica ou de negócios?** Um olhar a partir da Administração de Empresas. FACEF Pesquisa-Desenvolvimento e Gestão, v. 14, n. 2, 2012.

OWL. **Web Ontology Language (OWL).** Disponível em : https://www.w3.org/OWL/. Acesso em 03 dez 2018.

UOPEOPLE. **Java vs Python:** How to Choose Which Programming Language to Study. Disponível em: https://www.uopeople.edu/blog/java-vs-python-how-to-choose-which-programming-language-to-study/. Acesso em 02 dez 2018.
W3C. Semantic Web. Disponível em: https://www.w3.org/standards/semanticweb/. Acesso em 09 de novembro de 2017.