UM ESTUDO EXPLORATÓRIO ACERCA DE COMO O “DADO” PODERÁ TRANSFORMAR A SAÚDE POR MEIO DA TECNOLOGIA DE BIG DATA

Gabriel Alan Madureiro GONÇALVES, Alex Sandro Romeo de Souza POLETTO

alang.von@gmail.com, apoletto@femanet.com.br

RESUMO: A Área da Saúde, por muitos anos, continua procurando novas soluções para epidemias e pandemias, a busca por métodos e tratamentos eficientes e adequados, pelo controle de pacientes e pela descoberta de novos conhecimentos. Com a evolução das tecnologias em geral e do poder de processamento de dados estruturados e não estruturados, Big Data se tornou um referencial para a descoberta de conhecimentos e padrões que ainda não haviam sido descobertos. Com o uso de ferramentas conhecidas como as do ecossistema Hadoop, este trabalho apresenta algumas tecnologias e métodos que podem ser utilizados a favor da Saúde.

PALAVRAS-CHAVE: Big Data, Saúde, Dados, Hadoop, Data Analytics

ABSTRACT: Health Area, for many years, continues looking for new solutions to epidemics and pandemics, the search for adequate methods and treatments, patient control and the discovery of new knowledge. With the evolution of technologies in general and the power of structures and unstructured data processing, Big Data became a benchmark for the discovery of knowledge and standards that had not yet been discovered. With the use of popular tools as those of the Hadoop ecosystem, this paper presents some technologies and methods that can be used in favor of Health.

KEYWORDS: Big Data; Health, Data; Hadoop; Data Analytics.

1. INTRODUÇÃO

A Área da Saúde, uma crítica área do conhecimento humano, tem sido um grande alvo de pesquisas e investimentos, sejam eles por universidades e hospitais ou pela iniciativa privada, como grandes centros farmacêuticos. Diariamente, diversas pesquisas em busca de novos tratamentos e curas são realizadas todos os dias, onde quão grande a quantidade de informações precisas e catalogadas,

mais benefícios chegarão ao ambiente hospitalar.

De forma semelhante, a evolução dos computadores e demais dispositivos desencadeou um crescimento exponencial no volume de dados em todo o mundo, e este crescimento tem tomado proporções cada vez maiores e espantosas. Em 2013 a IBM afirmou que a população mundial gerava, por dia, 2,5 quintilhões de dados e que destes dados, 90% deles foram gerados somente nos dois anos anteriores. Tais dados são gerados por meio de diversos dispositivos, um termo muito relevante neste assunto é IoT (*Internet of Things)*, onde diversos dispositivos conversam entre si e trocam dados constantemente. Grande parte dos dados também são criados através das Redes Sociais, segundo Torres (2017) as redes sociais online são hoje um dos principais ambientes de debate, discussão e troca de informação entre as pessoas; Tais informações são dados como textos, fotos e vídeos.

Com tamanha quantidade de dados, os estudos em buscas de novos conhecimentos através deles, tornou-se uma importante área em TI (Tecnologia da Informação), é comum atualmente, temas como *Data Science, Data Analytics, Machine Learning,* BI *(Business Intelligence)* e *Big Data* estarem em diversas palestras, disciplinas no Ensino Superior ou até mesmo requisitos no mercado de trabalho.

Entretanto, apesar de grandes avanços na qualidade e quantidade de dados gerados, as empresas ainda não haviam notado a importância da busca pela Informação. Um estudo da EMC de 2012 afirmava que de todo a quantia de dados, na época 643 exabytes, apenas 3% destes dados eram utilizados, deixando muitos dados sem qualquer análise. Segundo uma pesquisa do IDC, o crescimento de grandes dados e de soluções analíticas para 2018, será de 33% em infraestrutura de nuvem, 29% em software, e 29% no setor de serviços, entretanto, 70% das empresas, de acordo com o estudo, não tem capacidade necessária para atender às demandas de crescimento para maximizar os benefícios em nível empresarial.

Diante desse fato, pesquisadores consideram que estamos vivenciando o início de uma nova revolução industrial, na qual os dados passam a ser elementos chaves desta mudança. Podemos concluir, portanto, que esse é o momento ideal para criarmos

oportunidades a partir dos dados (Marquesone, 2016).

Atualmente há uma vasta quantidade de ferramentas de *Big Data* disponíveis, grande parte sendo *open-source* (código aberto) tais como o ecossistema Hadoop (Conjunto de Ferramentas para processamento paralelo de grandes quantidades de dados armazenados em disco) e o Apache Spark (Ferramenta para processamento de grandes quantidades de dados em memória). Ambas ferramentas, são algumas das mais utilizadas no mundo, e seus resultados se caracterizam pela busca de padrões em grandes quantidades de dados, seja em computadores sofisticados ou em mais simples (paralelismo).

A Saúde se beneficiou dos dados em alguns casos a respeito da transmissão de doenças, como no famoso acontecimento em Lahore no Paquistão em 2011, quando um grande surto de Dengue ocorreu, infectando aproximadamente 16.000 pessoas e com 352 mortes; Após isto, o governo do Paquistão utilizou ferramentas disponibilizadas pelo Google, com algoritmos capazes de detectar surtos de dengue e antes que os mosquitos pudessem se proliferar por Lahore, agentes do governo limpavam todo o local, impedindo novas infecções. No ano seguinte, com o uso destas ferramentas o número de infecções desceu para 234 casos confirmado e nenhuma morte. Atualmente existem diversas implementações de algoritmos de mineração de dados na área da Saúde, como aplicativos para a detecção de epidemias, entretanto, ainda há muitos dados que não estão analisados.

Este artigo tem como objetivo apresentar técnicas de detecção de padrões e informações relevantes mediante o uso de dados médicos e até mesmo das redes sociais através do desenvolvimento de uma aplicação utilizando as tecnologias de Big Data e Data Mining e então mostrar a importância e a vantagem de utilizar de dados que muitas vezes não são tratados com tanta relevância, para a descoberta de conhecimento, especialmente à área da saúde.

2. ECOSSISTEMA HADOOP

Para diversas aplicações ou pesquisas referentes a Big Data, Hadoop é uma palavra muito frequente entre artigos e demais conteúdos, quase sempre, unida de palavras-chave como *escalonamento, processamento paralelo, sistema de*

arquivos, *tolerância a falhas* e *map-reduce*.

Criado por Doug Cutting e Mike Cafarella, o framework, que antes era parte integrante do projeto Apache Nutch, foi lançado oficialmente em 2006, passando a se chamar Hadoop.

Hadoop foi desenvolvido baseando-se em duas publicações feitas pelo Google:

- Um sistema de arquivos distribuído chamado Google File System (GFS);
- Um novo método de programação distribuída chamado MapReduce.

Hadoop se trata de um *framework* utilizado para processamento de dados em larga escala, contendo alguns atributos importantes como suporte à escalabilidade, gerenciamento de arquivos tolerante a falhas e uma nova metodologia de processamento para um menor tráfego entre os nós de um cluster. Além disso, Hadoop não se trata de apenas uma aplicação, mas sim de um conjunto de aplicações, denominado, Ecossistema Hadoop, cada uma contendo uma função ou suporte a um determinado problema.

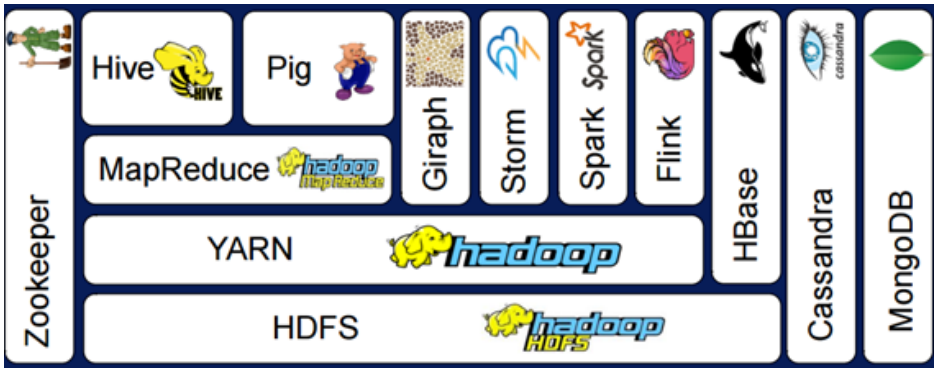

Figura 01. Título da Figura (Kieit)

Baseando-se nos artigos publicados pelo Google, os criadores do Hadoop desenvolveram uma versão *open-source* baseada nos dois modelos, onde surgiram o HDFS (*Hadoop Distributed File System)* e o Hadoop MapReduce, sendo estes um dos primeiros módulos do Ecossistema Hadoop, os quais serão abordados com mais profundidade a seguir, além de alguns outros módulos que foram utilizados para esta pesquisa.

3. HDFS - HADOOP DISTRIBUTED FILE SYSTEM

A partir do artigo do GFS surgiu o HDFS, que significa *"Sistema de Arquivos Distribuído Hadoop"*, que trouxe diversos benefícios ao grande armazenamento de dados, devido a alguns mecanismos que proporcionavam tolerância a falhas; A possibilidade de escalar a infraestrutura de modo simples e transparente, não sendo necessário conhecer toda a complexidade da aplicação; Uma melhor metodologia para processamento de dados que diminuíam drasticamente o tráfego de dados entre os nós do cluster.

Quando algum determinado conjunto de dados entra no HDFS, é gerado um bloco (geralmente com 64 MB em cada bloco) que é replicado para os *data nodes* do cluster. Data nodes são os nós do cluster onde os dados são armazenados e utilizados posteriormente em algum tipo de processamento, entretanto, tais dados não são visíveis a um usuário a ponto de serem identificáveis quando dentro de um *Data node*, necessitando-se então de um outro tipo de nó, os *Namenodes*.

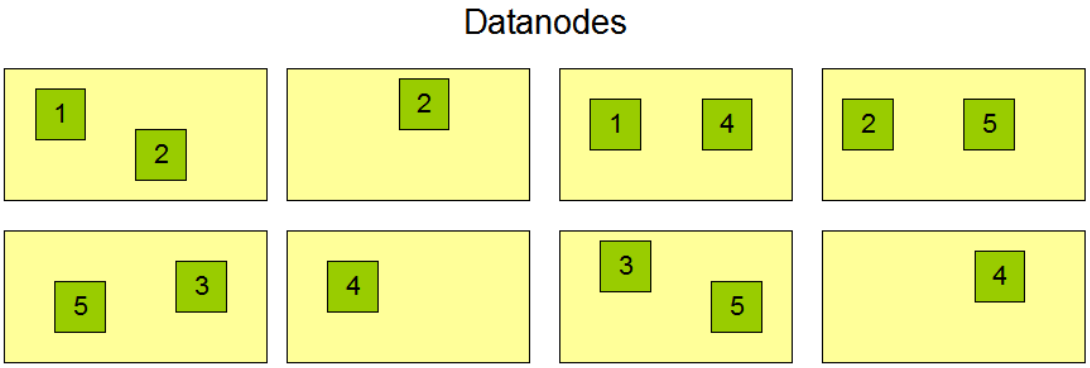

Figura 02. Ilustração do Armazenamento em Data Nodes (Apache Hadoop)

Namenodes são os nós que comandam todo o armazenamento de dados no HDFS, estes possuem os Metadados que indicam para um processamento, onde estão localizados cada informação para que possa buscar os dados corretos com êxito, caso este nó falhe, todos os dados do HDFS estarão comprometidos, a menos que, exista um segundo Namenode em funcionamento.

4. HADOOP MAP-REDUCE

Hadoop Map Reduce é uma das principais partes do framework Hadoop, sendo este um módulo responsável pelo processamento de dados. Inspirado também pelo artigo da Google, esta foi a implementação do algoritmo de mapeamento e redução, possuindo características importantes como o processamento paralelo. Uma característica interessante é que, igualmente em boa parte dos módulos do *framework* Hadoop, seu funcionamento não necessita de programação em como o processamento será realizado, mas em qual será a lógica do processamento, deixando configurações de infraestrutura e de mais complexidades de baixo nível em função do próprio Hadoop, que realiza isso de forma totalmente transparente ao analista de dados.

O algoritmo de Map-Reduce tem sua base em, como no próprio nome, duas etapas principais, o mapeamento e a redução, e a partir dos dados gerados após este processamento, um leque de possibilidades será aberto para uma futura análise, já que, apesar de relevar alguma informação e padrões, não traz tantas respostas ao problema solicitado, porém pode ser utilizado como dados ideias para um segundo processamento, assim, obtendo informação útil. A seguir, a Figura 03 apresenta cada etapa do processo de Map-Reduce:

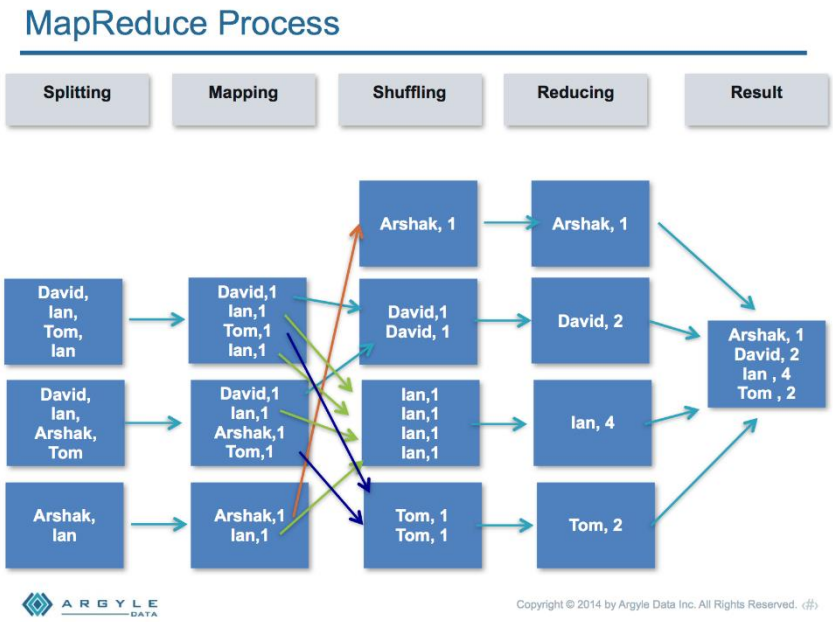

Figura 03. Etapas do Processo Map-Reduce (Fraud & Technology Wire)

Divisão (*Splitting)* é o momento em que os dados são lidos e divididos para que, em seguida, possam ser enviados às tarefas de mapeamento.

Mapeamento (*Mapping)* é a etapa onde os dados são obtidos, e ocorre o primeiro processamento que "mapeia" os dados em uma tupla de chave-valor, então, quando uma palavra é lida, automaticamente um grupo chave-valor é gerado, sendo a palavra o valor e o número um a chave e então a próxima palavra é processada até o final do arquivo.

Ordenação (Shuffling) é a etapa em que após geradas as tuplas por cada tarefa de mapeamento, é realizado um agrupamento, unindo todos os valores processados pela etapa anterior, que possuam a mesma chave. Após isso, são enviadas a etapa de Redução, cada uma das suas estruturas, com seus dados ordenados.

Redução (*Reducing)* é a etapa que recebe os valores da etapa anterior como entrada já ordenados e separados em blocos por cada tipos de chaves encontradas e em cada repetição de tuplas, o valor de cada chave é somado com as demais repetições encontradas, assim, ao final do processamento, é salvo no HDFS o resultado de todo o processamento em um arquivo de texto, onde são exibidas todas as chaves encontradas no arquivo e seus respectivos valores de repetição.

Assim como dito anteriormente, são etapas praticamente imperceptíveis ao programador, sendo necessário apenas o desenvolvimento da lógica do problema, como local do arquivo de entrada, etapas do processo, e local de saída, após isso, o Hadoop toma conta de todo o decorrer do processamento, mostrando cada etapa em *log* para o usuário.

5. APACHE SPARK

Com o crescimento da quantidade de dados para armazenamento, Hadoop surgiu para a necessidade de suportar toda esta necessidade, além de permitir realizar processamento paralelo, diminuindo o custo para escalar a infraestrutura para suportar o problema. No entanto, foi criado em uma época com padrões e valores diferentes da década atual, permitindo a aquisição de memória de armazenamento com mais facilidade que memórias RAM. Com esse contexto, o Apache Hadoop foi criado com foco em disco, algo que em algum tempo, causaria uma certa lentidão; então foi desenvolvido em 2012, uma ferramenta que pudesse dar

suporte a esta lentidão, uma extensão ao modelo MapReduce, conhecida como Apache Spark.

Apache Spark, assim como Hadoop, é uma ferramenta utilizada para processar grandes conjuntos de dados, realizando isso de maneira distribuída, porém com uma performance muito maior, pois esta foi desenvolvida com foco em processamento em memória, diferente do framework Hadoop onde havia foco apenas em disco. Com este grande diferencial, o Apache Spark pode atingir uma performance até cem vezes maior do que o Hadoop, segundo o site da própria ferramenta.

Uma característica interessante é também a possibilidade de utilizar três linguagens para sua programação que são Java, Scala e Python, além de que a ferramenta fornece mais componentes integrados em si mesma como o Spark SQL *(componente utilizado em processamento de dados estruturados, permitindo realizar consultas SQL*) e Spark Streaming *(componente utilizado na criação de fluxos de processamento tolerante a falhas sobre dados em tempo real/streaming*).

Spark tem como conceito principal os RDDs, que são como blocos ou tabelas com dados de diversos tipos, que são armazenados em partições diferentes. Estes possuem como característica a imutabilidade, não podendo ser alterados, mas sim sofrendo uma transformação e gerando um novo RDD.

A palavra RDD vem de *"Resilient Distributed Datasets"*, sigla que possuem os seguintes conceitos:

- Resilient: Os dados na memória podem ser recuperados caso haja algum problema durante o processamento.
- Distributed: São "distribuídos" e armazenados na memória por todo o cluster da aplicação.
- Datasets: Dados iniciais podem ser criados a partir de algum arquivo ou serem criados programaticamente

Estes objetos possuem dois tipos de operações:

- **Transformação**: São operações que não retornam valores, mas sim um novo RDD, não havendo assim algum processamento relevante, mas apenas o retorno de um novo RDD com novas instruções.

  A seguir há algumas funções que se caracterizam como de transformação:

- Filter
- flatMap
- groupByKey
- Map
- reduceByKey

- **Ação**: São operações onde há realmente um processamento com algum retorno. Assim que uma função de ação é chamada, todos os processamentos descritos por funções de transformação são executados e o valor é retornado.

  A seguir há algumas funções que se caracterizam como de ação:

  - Collect
  - Count
  - First
  - Reduce
  - Take

6. MÓDULOS DE COLETA E ANÁLISE DE DADOS

Para que possa haver processamento de dados, é necessário existir dados previamente selecionados para serem tratados. Pensando nisto, a seguir serão abordadas duas ferramentas de grande importância para a coleta e a análise de dados: Apache Flume e Hive.

Criado em 2011 pela Cloudera e tornando-se um projeto da Apache Software Foundation no ano seguinte, Flume é um sistema de grande importância para o mundo de Big Data, realizando a coleta e armazenamento centralizado de grandes quantidades de dados, tanto como de diversas fontes. O Apache Flume permite que diversos tipos de arquivos possam ser transportados para dentro do HDFS ou um banco de dados, como por exemplo: Logs, dados gerados em redes sociais, e-mails entre outras fontes (Gomes).

Tem como objetivo escutar dados criados em eventos e armazená-los na maioria das vezes ao HDFS. Nele são criados os agentes que se constituem em três partes: Source, relacionada a parte do agente onde os dados são recebidos; Channel, onde os dados anteriormente recebidos são armazenados para o Sink; Sink, onde os dados são enviados para seu destino.

Após a coleta dos dados, é possível realizar uma análise nos dados coletados e para isto, o módulo Hive se torna uma aplicação ideal. Apache Hive é uma aplicação do ecossistema Hadoop para *data warehousing*. Com Hive é possível aplicar uma estrutura à grandes quantias de dados não estruturados, como exemplo, os arquivos JSON.

A linguagem utilizada em suas consultas é chamada HiveQL (derivada da linguagem SQL) e permite que consultas muito similares às SQL possam ser realizadas em dados não-estruturados (sem tipos pré-definidos).

Dados podem ser lidos de uma variedade de formatos, de arquivos não-estruturados com texto separado por virgula ou espaçamento, arquivos não-estruturados JSON e tabelas estruturadas do banco HBase, um banco de dados distribuído pertencente ao ecossistema Hadoop. (MAPR)

7. ESTUDO DE CASO

A busca por conhecimento na área da saúde, unida com ferramentas de manipulação e processamento de dados é possível com algumas das ferramentas mostradas anteriormente. É evidente a busca por diversas empresas por assuntos relacionados a BI (*Business Intelligence*), termo referente à utilização de dados em busca de informação para a tomada de decisões de uma determinada corporação, entretanto, o foco deste artigo se dá à área da Saúde, portanto algumas variáveis terão de ser alteradas.

Um dos assuntos muito recorrentes a respeito de BI é análise de sentimento, termo que diz respeito a avaliação de uma empresa, produto e outros assuntos, através da coleta de dados, com alguma parte relevante oriunda das redes sociais. Como exemplo de implementação, há o artigo *Utilizando Analise de Sentimentos para Definição da Homofilia Política dos Usuários do Twitter durante a Eleição Presidencial Americana de 2016* publicado na PUC Minas.

Para realizar tal análise, é necessário utilizar uma ferramenta chamada Apache Flume, através desta ferramenta é possível coletar dados de redes sociais ou outras fontes, apenas necessitando assim de algumas informações como chaves e *tokens* de acesso para que a máquina comece a receber dados destas fontes. A escolha para algumas implementações durante o estudo foi da rede social Twitter onde alguns *tweets* foram coletados para um possível processamento posteriormente. Para que os dados possam vir com maior valor, é possível realizar uma configuração em que somente sejam coletados dados com palavras-específicas, assim, são definidas palavras relacionadas ao assunto, no caso da área da saúde, há exemplos de palavras como: ‘Gripe’,’Resfriado’,’Febre’,’Dor de cabeça’. Além de algumas variantes

destas palavras que ocorrem com muita normalidade em meios informais como: ‘gripe’,’resfriado’,’febree’,’doooor’.

Os dados coletados são recebidos na máquina como um arquivo no formato JSON (*JavaScript Object Notation*), então assim que estão já dentro do HDFS é possível que estes dados possam ser manipulados pelo apache Hive, uma outra aplicação utilizada neste contexto de análise de sentimento pertencente também ao ecossistema Hadoop.

Após isso é definido um arquivo com tuplas chave-valor onde a algumas chaves (no caso palavras) e números relacionados a pontuação destas chaves são escritos, os quais serão importantíssimos para o próximo processamento. Além disto, são definidas também as chaves específicas onde deverá haver algum resultado.

A partir deste ponto, é realizado em processamento através do Apache Spark com o algoritmo específico de análise de sentimentos, com a indicação de arquivos de amostra, pontuações e chaves de retorno. Após iniciado, será executada uma leitura extensiva por todo o arquivo de amostra coletado pelo Flume, e cada *tweet* será processado, buscando relações com as chaves de resposta, e as chaves referentes a palavras de pontuação, assim que uma palavra de pontuação é encontrada, juntamente com outra chave de resposta, sua pontuação é calculada com a atual da chave em questão e o processamento é continuado até o fim do arquivo.

Assim que concluído, é possível visualizar tais resultados no Apache Hive, onde haverá as chaves e seus valores finais, indicando as chaves mais bem e mal pontuadas, é claro, que os critérios de avaliação devem ser feitos pelo analista de dados, para que não haja um desentendimento perante os resultados obtidos.

Por outro lado, é possível também realizar processamentos mais simples através do algoritmo de MapReduce. Os dados podem ser coletados de fontes mais específicas que as dos algoritmos de Análise de Sentimento, como por exemplo: relatórios, laudos, arquivos de *log*, entre outros. Como seu processamento é mais simples que o anteriormente demonstrado, serão necessários conhecimentos em Estatística para que haja um entendimento concreto dos resultados.

Após coletada uma quantidade definida de conjuntos de dados, poderá ser iniciado o algoritmo de Mapeamento e Redução, assim percorrendo todo o

arquivo e seus respectivos conjuntos de dados, criando chaves e valores fixos e em seguida reduzindo-os para que o resultado possa ser gerado e visualizado no HDFS. Após isso, é necessário fazer uma limpeza nas chaves que não são úteis para a análise e em seguida, com as chaves corretas e seus resultados processados, poderá ser feito algum tipo de cálculo estatístico para a descoberta de conhecimento, tanto como, tomada de decisões, como, por exemplo, aquisição mais focada em algum tipo de medicamento em alguma parte de ano, devido a ocorrência de um padrão de doenças.

8. CONSIDERAÇÕES FINAIS

Conclui-se que através de tecnologias de processamento de dados é possível realizar buscas por informações e padrões em dados (recebidos de diversas fontes e formatos) referentes à saúde como busca por epidemias, padrões em laudos, entre outros, utilizando de várias diferentes formas e tecnologias de Big Data no processamento de dados.

É evidente que através de tecnologias de manipulação e processamento dados de diferentes fontes e variados assuntos possam ser obtidas através de dados que não possuam tanto valor a priori. A abordagem de conceitos a respeito de ferramentas é importante para que haja compreensão do funcionamento e do algoritmo utilizado para processamento de dados, além da necessidade de enfatizar a importância do estudo da estatística, que é um atributo essencialmente importante para um cientista de dados.

Ainda há muita informação a ser processada em diversas áreas ainda não exploradas, e para que conhecimentos possam ser encontrados, a informação a respeito de tecnologias de Big Data e a constante criação de novas tecnologias, poderão abrir cada vez mais, novas portas para solucionar problemas de qualquer área do conhecimento humano, e com grande certeza, a Saúde.

9. REFERÊNCIAS

ANDRADE, Tiago Pedroso da Cruz de. MapReduce – Conceitos e Aplicações. Disponível em <http://www.ic.unicamp.br/~cortes/mo601/trabalho_mo601/tiago_cruz_map_reduce/relatorio.pdf>. Acesso em 31/07/18.

APACHE HADOOP. HDFS Architecture Guide. Disponível em <https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html>. Acesso em 31 jul 2018.

EMC. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the

Far East. Disponível em <https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf>. Acesso em 11 de jun 2018.

FRAUD & TECHNOLOGY WIRE. An “F-Leven” Guide to Hadoop: MapReduce. Disponível em <https://www.fraudtechwire.com/an-f-level-guide-to-hadoop-mapreduce/>. Acesso em 31 de jul 2018.

GARCIA, Marco. O que é Hadoop? Disponível em <https://www.linkedin.com/pulse/o-que-%C3%A9-hadoop-marco-garcia>. Acesso em 30 de jul 2018.

GOMES, Eduardo. Introdução ao Apache Flume. Disponível em <https://www.concrete.com.br/2016/08/09/introducao-ao-apache-flume/>. Acesso em 20 de jul 2018.

IBM. 2.5 Quintillion bytes of data created every day. How does cpg retail manage it? Disponível em <https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/>. Acesso em 11 de jun 2018.

INFOQUEUE. Big Data com Apache Spark. Disponível em <https://www.infoq.com/br/articles/apache-spark-introduction>. Acesso em 31 de jul 2018.

KIETEC. Big Data e o Ecossistema Hadoop. Disponível em <http://www.kieit.com/big-data-e-o-ecossistema-hadoop/>. Acesso em 31 de jul 2018.

MAPR. Apache Hive. Disponível em <https://mapr.com/products/product-overview/apache-hive/>. Acesso em 29 de jul.2018.

MARQUES-NETO, Humberto Torres; CAETANO, Josemar Alves; LIMA, Helder Seixas; SANTOS, Mateus Freira dos. Utilizando Análise de Sentimentos para Definição da Homofilia Política dos Usuários do Twitter durante a Eleição Presidencial Americana de 2016. XXXVII Congresso da Sociedade Brasileira de Computação, 2017, São Paulo.

MARQUESONE, ROSANGELA DE FÁTIMA PEREIRA. Big Data, Técnicas e tecnologias para extração de valor dos dados. Casa do Código, 2016.

MEDIUM, O FUTURO DA MEDICINA. O papel de Big Data na luta contra o Zika. Disponível em <https://medium.com/futuro-da-medicina/o-papel-do-big-data-na-luta-contra-o-zika-bdc295d55d87>. Acesso em 30 de jul.2017.

PUBLISH. Tendências em Big Data e Análises: Hoje é digital, amanhã cognitivo. Disponível em <https://publish.com.br/artigo/tendencias-em-big-data-e-analises-hoje-e-digital-amanha-cognitivo/>. Acesso em 31 de jul.2018.

PUBLISH. Tendências em Big Data e Análises: Hoje é digital, amanhã cognitivo. Disponível em <https://publish.com.br/artigo/tendencias-em-big-data-e-analises-hoje-e-digital-amanha-cognitivo/>. Acesso em 31 de jul.2018.

RELVAS, Carlos Eduardo Martins. Apache Spark. Disponível em <https://https://www.ime.usp.br/~gold/cursos/2015/MAC5742/reports/ApacheSpark.pdf>. Acesso em 31 de jul 2018.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S. Sistemas de Bancos de Dados. 5. ed. Tradução de Daniel Vieira. Rio de Janeiro: Editora Elsevier, 2006.