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

Alex Sandro Romeo de Souza POLETTO, Gabriel Alan Madureiro GONÇALVES

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

RESUMO: com a crescente exponencial dos dados em escala mundial, grande parte deles sendo não-estruturados, percebeu-se a possibilidade de extração de informação útil, como padrões, direções para negócios, estatísticas e até mesmo benefícios para a área da Saúde, tudo isso com as tecnologias referentes a Big Data. O objetivo deste trabalho é analisar as tecnologias e soluções para demonstrar como o uso dos dados, juntos à Big Data, podem trazer evoluções na área da saúde.

PALAVRAS-CHAVE: Big Data; Dados; NoSQL; Saúde.

ABSTRACT: with the exponential increasing of data on a global scale, most of them being unstructured, it was noticed the possibility of extracting useful information, such as "standarts", business directions, statistics and even benefits for the health area, using Technologies referring to Big Data. The main objective of this work is to analyze the Technologies and solutions, showing how the use of data, working with Big Data, can bring Evolution on the Health area.

KEYWORDS: Big Data; Data; NoSQL; Health.

1. INTRODUÇÃO

Quando o termo Big Data é mencionado, logo vem em mente o sentido de algo grande, pesado e de técnicas específicas para ser manipulado. Podemos definir Big Data como uma grande explosão de dados, estes em grande dimensão, em alta variedade e que se expande em altíssima velocidade. Com os avanços em processamento e armazenamento de dados, a quantidade de dados circulando entre máquinas e internet aumentou exponencialmente, afirma o artigo *Big Data: Ferramentas e Aplicabilidade*. Hoje com a IoT (*Internet of Things),* onde constantemente máquinas conversam entre si, transmitindo dados oriundos da utilização de usuários como mensagens de texto, ações em aplicativos; ou dados criados pelas próprias máquinas como dados de sensores, arquivos de log, entre outros.

Com essa grande massa de dados, com diversos formatos e grande parte dela sendo não estruturada (não necessitando suportar todas as propriedades ACID como nos Bancos de Dados Relacionais), foi possível extrair informações importantes para diversas áreas do conhecimento humano como Acadêmica ou Corporativa, como estatísticas, padrões, avaliação de produtos e serviços.

Para que tais dados (não estruturados) possam ser armazenados e analisados utiliza-se um modelo de armazenamento alternativo, o NoSQL. Por meio deste tipo de solução, dados antes não aceitos por Banco de Dados Relacionais, podem ser armazenados e consultados.

A área da Saúde constantemente utiliza de tecnologias para analisar e criar soluções como curas, tratamentos, catalogar informações sobre doenças e sintomas. Com a ajuda desta grande imensidão de dados, esta área pode ser beneficiada de inúmeras formas.

Esta pesquisa tem como objetivo mostrar como a utilização dos dados, em conjunto com aplicações de Big Data, pode trazer benefícios para a área da Saúde; explanando com alguns exemplos, como tais dados podem ser armazenados, tratados e unidos com as ferramentas adequadas.

Big Data, assim como a Saúde, é um tema de grande importância global, sendo estudado e utilizado pelas grandes empresas de Tecnologia da Informação, assim como pequenas empresas de diferentes ramos; por meio dos dados, essas empresas e instituições tem conseguido tomar decisões inteligentes nos seus negócios; unindo tal tecnologia com a

área da Saúde, atualmente de suma importância para a sociedade, a Saúde poderá se beneficiar de uma imensa fonte de informação.

2. DADOS

Dados podem ser considerados como registros ou documentações de nossa realidade, seja ela do mundo físico, sensorial e semelhantes. Determinados parâmetros de observação e medição são utilizados em nosso mundo, assim, levando coisas de nosso mundo como tempo, sons, imagens de forma abstrata para as máquinas. (MOHAMMED; WAGNER MEIRA,2014).

Estes podem ter dois meios de criação; existem dados criados por pessoas e dados criados por máquinas. Dados criados por nós, seres humanos, em sua maior parte contêm uma propriedade intelectual; refletem a interação humana com a natureza, outras pessoas, pensamentos e ações. Grande parte deles atualmente podem ser obtidos por meio das redes sociais, onde constantemente são armazenados diversos tipos de pensamentos, e momentos, criados ou que sofreram interação humana; também podem ser criados em sites de e-commerce, como avaliação de produtos entre outros.

Dados gerados por máquinas são aqueles produzidos por meio de aplicações, operações de sistema, onde não há interação humana. Com o Advento da IoT (*Internet of Things*), máquinas começaram a conversar entre si, constantemente criando dados vindo de sensores, por meio de registros de log, streaming de vídeo, etc.

Há algumas décadas, grande parte dos dados existentes no mundo eram estruturados, este tipo de dado é o que consegue suportar todas as propriedades ACID de um Banco de Dados Relacional que são Atomicidade, Consistência, Isolamento e Durabilidade. Hoje com inúmeros dispositivos conectados entre si e com a Internet, uma grande massa de dados é composta por não estruturados, como também uma parte composta por dados semi-estruturados.

Dados semi-estruturados são dados que contêm uma estrutura pré-definida, mas não tão rígida como os dados estruturados, *como* arquivos XML, JSON. Dados não estruturados geralmente são constituídos de imagens, vídeos, músicas e formatos de textos específicos; estes não podem ser facilmente armazenados em banco de dados relacionais. Com as redes sociais, o tamanho dos dados não estruturados superou a quantidade de dados

estruturados, hoje, aproximadamente 80% dos dados disponíveis em uma corporação são não estruturados, afirma a BSA em *Qual é o "x" da questão com relação a dados?*. Apesar de não poderem ser organizados em SGBDs Relacionais, utilizando de ferramentas corretas e eficazes podem ser retirados desses dados informações muito importantes para qualquer área do conhecimento humano.

3. BIG DATA E DATA MINING

Com o aperfeiçoamento da capacidade de processar e armazenar dados, unidos com os avanços de comunicação via web, vários tipos de mídias sociais e dispositivos móveis, uma grande explosão de dados se iniciou. Big Data é o termo que descreve o imenso volume de dados – estruturados e não estruturados – que impactam os negócios no dia a dia (SAS). Um infográfico da Vouchercloud afirma que 90% de todos os dados disponíveis existentes foram criados apenas nos últimos dois anos, grande parte deles sendo não estruturados.

O desafio para as ferramentas de Big Data é entre outros a manipulação de dados semiestruturados e não estruturados no intuito de extrair valor destes através de correlações e outros processamentos de análise e então compreendê-los para que tragam valor ao determinado meio aplicável (GALDINO,2016).

No universo de Big Data e seu grande objetivo, há algumas informações referentes aos dados e suas características com maior destaque: Volume, Variedade e Velocidade. Estes atributos provocam a dificuldade de armazenar os dados em SGBDs Relacionais.

Volume é sem dúvida, uma das mais visíveis características de Big Data; os dados estão disponíveis, porém em uma quantidade tão grande, que se tornou um desafio armazená-los e processá-los. Para obter-se a dimensão deste volume, há algumas estatísticas de grandes empresas:

Em 2014, aproximadamente 3,3 bilhões de buscas eram realizadas no Google diariamente. Foi contabilizado em 2016 pelo Facebook uma média de 1.13 bilhão de usuários, 2.7 bilhões de "likes" diariamente. Segundo a consultora EMC, em 2013 haviam 4.4 Zettabytes de dados em todo o mundo, a estimativa é que essa quantidade atinja 44 Zettabytes em 2020, um crescimento exponencial em termos de volume de dados.

Variedade está relacionada à estrutura do dado. Armazenar dados estruturados em tabelas de bancos de dados relacionais e realizar consultas sobre os determinados campos é possível; todavia, a dificuldade atual é que a maioria dos dados são não estruturados. Armazenar um vídeo, foto, música ou outro tipo de arquivo específico em um SGBD Relacional é uma tarefa complicada, o que impulsionou o desenvolvimento de uma série de soluções para o ambiente de Big Data.

Por fim, a Velocidade completa as três principais características. Os dados que hoje necessitam ser armazenados e minerados trafegam em velocidade constante, como exemplo há casos como: Ações na bolsa de valores, que constantemente necessitam de atualizações em seus valores; registros de log. Quanto mais preciso o tempo em que o dado é criado ou obtido, mais valor ele terá para uma determinada necessidade, como a quantidade de usuários utilizando determinada aplicação, estoque de produtos de uma empresa, tráfego aéreo.

Há dados de grandes serviços como mídias sociais, que constantemente, por meio de seus usuários, geram dados com altíssima velocidade. A cada 1 minuto são criados pelos usuários do Facebook, aproximadamente, 510.000 comentários e são feitos 136.000 uploads de fotos e vídeos.

Com todas estas características chega-se ao grande desafio: Como minerar e extrair informação útil a partir desta imensa explosão de dados. Empresas costumavam a armazenar seus dados para consultas comuns, com a evolução do processamento de dados, começou-se a armazenar mais dados, mas não os analisar efetivamente.

Tecnologias de Big Data visam suprir todas essas características e necessidades que antes eram consideradas difíceis ou inviáveis, descobrindo e proporcionando informação minerada útil. Para a execução de uma aplicação de Big Data, devem existir alguns processos específicos.

O primeiro passo é definir qual é a resposta que se quer obter por meio da aplicação, qual é o objetivo, quais são os tipos de informações que devem ser extraídas dos dados. Em seguida, são definidas as fontes de onde os dados serão obtidos e como os mesmos serão armazenados.

Em seguida começa a fase de processamento onde são aplicados diversos algoritmos de análise de dados como Mineração de Dados (Data Mining), métodos estatísticos,

fundamentos matemáticos, entre outros. Um exemplo de algoritmo de mineração de dados é o algoritmo Apriori.

Proposto em 1994 pela equipe de pesquisa do Projeto QUEST da IBM, o algoritmo Apriori é capaz de detectar padrões de associação, como por exemplo, a porcentagem de transações que um ou mais itens têm em comum.

COD	Itens Comprados
1	Achocolatado Leite Pão
2	Pão Presunto Café
3	Refrigerante Pão
4	Ovos Pão
5	Café Leite

Figura 1 - Dados separados por transação

Primeiramente são armazenados os dados em formato de transação, podendo haver repetições de itens em transações diferentes, como é possível ver o com o item "pão", que se repete por 4 vezes nas transações do exemplo.

Item	Frequência
Achocolatado	20%
Leite	40%
Pão	80%
Presunto	20%
Refrigerante	20%
Ovos	20%
Café	40%

Figura 2 – Primeira Análise de Frequência

Item	Frequência
Leite	40%
Pão	80%
Café	40%

Figura 3 – Itens aceitáveis selecionados

Então é realizada a análise de frequência de cada item existente nas transações disponíveis, como é possível observar na Figura 2. Após isso, é estipulado uma frequência mínima para que o processo possa seguir em execução, no caso acima, foi estipulado no mínimo 40% de frequência necessária. Os Itens que não atingirem esta margem de ocorrência serão descartados.

Itens	Frequência
Leite Pão	33,33%
Leite Café	33,33%
Pão Café	0,00%

Figura 4 – Frequência de Grupo

Criam-se pares de itens, e da mesma maneira feita anteriormente, são verificadas as frequências em que cada par aparece nas transações disponíveis, este passo pode ser repetido de acordo com a banco de dados existente. Em seguida valores com frequência inferior à mínima são removidos, no caso, foi utilizado 30% de frequência mínima.

O resultado desta mineração aponta que aproximadamente 1 terço das pessoas que compram o item “leite”, também podem comprar o item “pão” ou “café”. Este é apenas um tipo de algoritmo que pode ser utilizado em aplicações de Big Data.

Então chega-se à última fase de uma aplicação, há a visualização dos dados tratados e minerados, que de uma forma simples de ser entendida, mostra qual foi o resultado de todo armazenamento e análise da base de dados obtida.

O uso de tecnologias de Big Data pode trazer muitos obstáculos, mas quando uma aplicação é executada com êxito, de forma rápida e eficiente, grandes resultados que antes eram muito difícil de serem encontrados, podem ser obtidos através de um grande mar de dados antes não analisados.

4. NOSQL

Bancos de Dados Relacionais sempre estiveram presentes há anos armazenando dados estruturados. SQL (*Structured Query Language)*, os dados são guardados em tabelas, junto com seus campos, onde há uma estrutura pré-definida. Com o crescimento exponencial dos dados, junto com sua grande velocidade e variedade, os SGDBs Relacionais começaram a ter obstáculo para armazenar tais dados, os quais não supriam as propriedades ACID (Atomicidade, Consistência, Isolamento e Durabilidade). Utilizando os conhecimentos do livro “Sistemas de Banco de Dados”, de Elmasri e Navathe, obtêm-se os conceitos a seguir.

Atomicidade é a propriedade de um SGDB Relacional que garante que as transações sejam “atômicas”, ou seja, ou as transações serão totalmente executadas ou não serão executadas. Caso a transação falhe, como um problema de execução, resultando em uma transação incompleta, técnicas de restauração serão utilizadas para desfazer a transação com erro.

A Consistência indica que para que uma transação ocorra com sucesso, é necessário que nenhuma regra do banco de dados seja violada, ou seja, os dados deverão satisfazer todas restrições especificadas pelo esquema.

Isolamento é a regra que diz que cada transação deve ser executada isoladamente das outras, por mais que por um curto período de tempo entre outra transação. Assim, é possível prevenir que uma transação não sofra interferência de outra transação concorrente.

Durabilidade garante que o resultado de toda transação realizada com êxito deverá ser mantido no banco de dados, mesmo na ocorrência de falhas.

Com Big Data, grande maioria dos dados disponíveis se tornaram não estruturados, e armazená-los em SGBD Relacionais se tornou algo de muita dificuldade. Para avançar sobre tais obstáculos, foram então criados os Modelos NoSQL.

Derivado de *Not only SQL*, "não somente SQL", NoSQL representa os novos modelos de armazenamento de dados, desenvolvidos para que possam dar suporte as necessidades existentes no contexto de Big Data. Diferente do banco de dados relacional, em que o foco principal é voltado à integridade dos dados, os modelos existentes em NoSQL tendem a sacrificar uma ou mais propriedades ACID, para assim oferecer maior desempenho e escalabilidade às soluções que lidam com grande volume de dados (MARQUESONE, 2016).

NoSQL não contêm apenas um modelo de armazenamento de dados, mas sim um específico para cada necessidade de uma aplicação de Big Data. Uma empresa varejista pode ter necessidades de manipulação de dados relacionados a seu estoque diferente de uma empresa de turismo ao receber opiniões sobre os serviços prestados. Algumas podem precisar de leituras constantes, outras de gravações constantes no banco.

Cada modelo em NoSQL pode ter sua classificação de acordo com seu tipo de armazenamento, existem três grandes modelos que podem ser observados: Orientado a Chave-Valor, Orientado a Documentos e Orientado a Grafos.

Banco de dados Orientado a chave-valor tem uma estrutura simples de ser compreendida, neste modelo, os dados são gravados nas "chaves", que funcionam como identificadores dos dados gravados no campo chamado "valor". A chave é geralmente constituída de um campo do tipo String enquanto o valor pode ser de diversos tipos de dados, sem precisar

de uma predefinição de tipo de dado, o esquema. Este modelo de armazenamento pode ser muito eficiente para aplicações onde são realizadas muitas leituras e escritas de dados, devido à simples estrutura de armazenamento, que possibilita um alto desempenho ao realizar consultas.

Banco de dados Orientado a Documentos oferecem grande flexibilidade em gerenciamento dos dados, permitindo a criação de índices sobre valores, possibilitando consultas mais precisas. Documentos podem ser definidos como estruturas flexíveis, que por meio de dados semiestruturados como XML, não necessitam de um esquema rígido, podendo ocorrer variações de estrutura em cada documento. Um funcionário de um departamento de uma fábrica pode não ter um campo de ensino superior completo em seu registro, porém, um funcionário de outro departamento pode conter este campo. Este tipo de banco de Dados contém alta disponibilidade, trabalhando com replicação de dados em cluster, o que permite que ele fique disponível mesmo se um de seus servidores tenha algum problema.

Bancos de Dados orientado a grafos trabalham com o relacionamento dos dados. Além da informação sobre os dados armazenados, são armazenadas informações relacionadas a ligação entre tais dados. Essa solução é utilizada para meios como, redes sociais na busca de amigos em comum ou relacionamento entre produtos. Seu formato possui estrutura definida na teoria dos grafos, diferente dos demais, para armazenar os dados e seus relacionamentos. É um modelo eficiente para aplicações que necessitam trabalhar com caminhos e relacionamentos, como rotas, conexões, entre outros.

Bancos de Dados NoSQL, não descartam a existência e importância dos bancos de dados relacionais, mas são extremamente importantes para aplicações de Big Data, que precisam de escalabilidade e desempenho com grandes volumes de dados.

5. SOLUÇÕES PARA SAÚDE

Saúde é sem dúvidas uma das áreas de maior importância do conhecimento humano, sempre em constante evolução e busca de técnicas, hábitos, tratamentos e curas para que a expectativa de vida possa ser prolongada.

Big Data pode contribuir de maneira bastante efetiva nesta área, como por exemplo, na prevenção de doenças e pandemias. Utilizando-se de uma grande base de dados, é

possível que infecções e mortes não ocorram, se estes dados forem armazenados e utilizados corretamente.

Em 2011, em Lahore, Paquistão, um grande surto de Dengue ocorreu, infectando aproximadamente 16.000 pessoas e com 352 mortes. Para conter casos parecidos posteriormente, o governo Paquistanês utilizou ferramentas do Google, para a criação de algoritmos que fossem capazes de detectar precocemente surtos de gripe e dengue. Como resultado, no ano seguinte, o número de infecções desceu para 234 casos confirmado e nenhuma morte.

Existem ferramentas interessantes, que por meio de mapas e uma grande base de dados fornecem informação a respeito de onde e que tipo de doença está se espalhando por determinado local. Existem aplicações gratuitas que podem apresentar informações relacionadas a doenças como o Google Flu Trends e também o SickWeather.

Hoje, o controle e prevenção de vírus como o Zika são geralmente detectados por sistemas de vigilância tradicionais—por meio de consultas médicas e em departamentos de saúde, que não correlacionam os dados captados e, por conseguinte, não geram inteligência para um melhor controle da doença (MEDIUM, 2016). Utilizando o poder dos dados, este processo pode ser acelerado e problemas podem ser contidos.

De uma perspectiva diferente, mas de igual importância, temos a Medicina de Precisão. Esta é uma abordagem crescente para a prevenção de doenças e tratamentos, considerando as variações individuais das pessoas com genes, meio ambiente e estilo de vida.

A partir do uso de tecnologias de Big Data e sua imensidão constante de dados, unidos à mineração, é possível extrair informações importantes por meio de genes, tecnologias de análise biomédica e de grandes conjuntos de dados disponíveis na internet.

No último mandato presidencial estadunidense, o presidente Barack Obama anunciou que seriam investidos 215 bilhões de dólares para o objetivo de construir uma grande base de dados, com informações genéticas, informações médicas e outras informações individuais de mais de um milhão de americanos.

Com esta grande base de dados, pesquisadores buscam entender melhor como cada doença ocorre, como se propaga, quais os tipos de pessoas mais vulneráveis a se contaminar. Aplicando algoritmos inteligentes neste tipo de base, é possível detectar

padrões, que se analisados corretamente, resultarão em respostas importantes, tanto para a parte de doenças, como para o desenvolvimento de curas e tratamentos.

Há também, outros meios de unir Big Data com a área da saúde como Monitoramento de pacientes em tempo real, Extração de informação por meio de imagens médicas. Todas estas áreas específicas de atuação necessitam de tecnologias eficientes e ágeis para que a Saúde possa evoluir, tanto em prevenção de doenças e controle de pacientes como para a obtenção de uma enorme quantidade de informação minerada, oriunda de grandes bases de dados.

6. CONSIDERAÇÕES FINAIS

Big Data se tornou um tema altamente debatido, devido ao crescimento exponencial na quantidade de dados disponíveis hoje em máquinas e redes. Dados não estruturados, antes ignorados por grande parte das grandes e pequenas empresas, passaram a ter um valor importantíssimo para negócios, perceptível com a criação de novos cargos como o Cientista de Dados.

Diversas áreas sociais começaram a adotar tais técnicas de Big Data, em busca de respostas objetivas, que por meio delas, trouxeram direções para negócios, um grande aumento de informação para catalogar e padrões antes jamais percebidos. A área da Saúde foi e tem sido beneficiada pelo “dado”, unindo a grande quantidade de dados existentes, equipamentos com sensores inteligentes, avanços na genética com Big Data. Hoje a medicina de precisão, prontuários eletrônicos e outras ferramentas da Saúde, contribuem para que cada vez mais, haja informação e desenvolvimento de maior qualidade e velocidade, utilizando o dado.

De modo geral foram apresentados tópicos importantes relacionados a Big Data, demonstrando o que são e como funcionam os dados; a grandeza das Tecnologias de Big Data e Data Mining e quais são seus obstáculos e objetivos; como funciona um banco de dados NoSQL e suas ramificações de modelos de armazenamento e como utilizando destas tecnologias avanços podem ser obtidos na área da saúde.

O artigo teve como finalidade demonstrar como a utilização do dado e tecnologias de Big Data funcionam e favorecem a área da Saúde, juntamente com algumas de suas tecnologias de armazenamento e suas características.

Um futuro trabalho pode ser realizado abordando, como tais tecnologias funcionam de maneira prática, com o desenvolvimento de uma pequena aplicação para demonstrar como os dados são armazenados, tratados e exibidos, passo a passo.

REFERÊNCIAS BIBLIOGRÁFICAS

ANA CAROLINA LORENA; ANDRÉ C. P. L. F. DE CARVALHO. Uma Introdução às Support Vector Machines. Revista de Informática Teórica e Aplicada, UFRGS, 2007.

BSA. Qual é o "x" da questão com relação a dados? Disponível em <https:// http://data.bsa.org/wp-content/uploads/2015/10/BSADataStudy_br.pdf>. Acesso em 30 de jul.2017.

DATA SCIENCE ACADEMY. Big Data Fundamentos, 2017.

DIAS PORTO CHIAVEGATTO, ALEXANDRE FILHO. Uso de Big Data em Saúde no Brasil: perspectivas para um futuro próximo.

ELMASRI, RAMEZ; NAVATHE, SHAMKANT B. Sistemas de Banco de Dados. Addison-Wesley, 4a. edição em português.

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

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.

MOHAMMED J. ZAKI; WAGNER MEIRA, JR. Data Mining and Analysis, Fundamental Concepts and Alghoritms. 32 Avenue of the Americas, New York, NY 10013-2473, USA, 2014.

GALDINO, NATANAEL. Big Data: Ferramentas e Aplicabilidade.

Software & Soluções de Analytics. O que é Big Data? Disponível em <https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html>