Tecnologia Big Data na Área da Saúde

Alex Sandro Romeo de Souza POLETTO, Flávio Henrique ALVES

Fundação Educacional do Município de Assis, Instituto Municipal de Ensino Superior de Assis, São Paulo-SP, Brasil

apoletto@femanet.com.br, fhalves@live.com

Abstract: The term Big Data has been gaining momentum in all areas of science in recent years, mobile equipment has been drastically updating the world scenario, making any user can communicate and share their data, for example through social networks. This communication has an impact in all sectors, and is no different for the health area, nowadays a reality, it assists in decision making, making it possible to draw efficient strategies. Big Data is a term that represents mechanisms and tools to support the management of large data mass, this article will use three points in a pipe line that will allow to use mechanisms for decision making insides that can help to analyze the scenario of a holistic way within the Health Area being them, Big Data System, Health Data and Analysis Algorithms.

Keywords: Algorithms; Analyze; Big Data; Data; Health.

Resumo: O termo Big Data vem ganhando proporção em todas áreas das ciências nos últimos anos, os equipamentos moveis vem atualizando drasticamente o cenário mundial, fazendo com que qualquer usuário possa se comunicar e compartilhar seus dados, por exemplo através de redes sociais. Essa comunicação tem impacto em todos os setores, e não é diferente para a área da saúde, hoje realidade, auxilia nas tomadas de decisões, possibilitando traçar estratégia eficientes. Big Data é um termo que representa mecanismos e ferramenta para apoio a gestão de grandes massas de dados, o presente artigo utilizara três pontos em uma *pipe line* que irá permitir utilizar mecanismos para tomada de decisão criando *insides* que ajuda a analisar o cenário de uma maneira holística dentro da Área da Saúde sendo eles, Sistema de Big Data, Dados de Saúde e Algoritmos de Analise.

Palavras-chave: Algoritmo; Analise; Big Data; Dados; Saúde.

1. INTRODUÇÃO

Para LAUDON e LAUDON (1999, p. 10), “conhecimento é o conjunto de ferramentas conceituais e categorias usadas pelos seres humanos para criar, colecionar, armazenar e compartilhar a informação”. As informações são criadas a partir da transformação dos dados, através da aplicação do conhecimento humano.

O avanço tecnológico entre suas ambições, busca a implementação deste conhecimento nas maquinas para que elas nos auxiliem na tomada de decisão e nas ações cotidianas. Nos últimos anos temos observado inúmeras revoluções tecnológicas e culturais na sociedade (AMARAL, 2011), dentre elas o crescimento exponencial na potência computacional (BOLLIER, 2010), a onipresença da tecnologia de informação e a explosiva disponibilidade de dados atitudinais/comportamentais e transacionais (LOHR, 2012).

Este fenômeno tecnológico é conhecido como Big Data, termo cunhado em 2008. Trata-se de um conceito associado a ferramentas gerenciais (SMOLAN, 2012) e que impulsiona a dinâmica mercadológica a uma velocidade exorbitante (ARTHUR, 2013). De fato, pode-se apontar que “a coleta e análise de dados em larga escala tem se tornando uma nova fronteira da diferenciação competitiva” (BUGHIN; LIVINGSTON; MARWAHA, 2011, p. 1).

A grande questão está em como todas estas tecnologias que o Big Data possui, podem alavancar a tomada de decisão na área da Saúde. Hoje, a comunidade médica vem adotando tecnologias para facilitar o dia a dia e salvar vidas, porém, estão apenas no início, com relação ao que essas tecnologias têm a lhes oferecer.

2. BIG DATA

Big Data traz consigo uma complexidade em seus processos para tratar os dados de todas as formas, sua arquitetura e tão complexa quanto se definição assim como para Zikopoulos et al (2012) diz que Big Data se caracteriza por quatro aspectos inicialmente: volume, velocidade, variedade, veracidade e não podemos deixar de listar o valor, que foi adotado posteriormente. Essa definição complementa o que é

explicado em Aurélio (2017), definindo que dados simplesmente são símbolos quantitativos e qualitativos, que possam ser utilizados para o processo de uma informação.

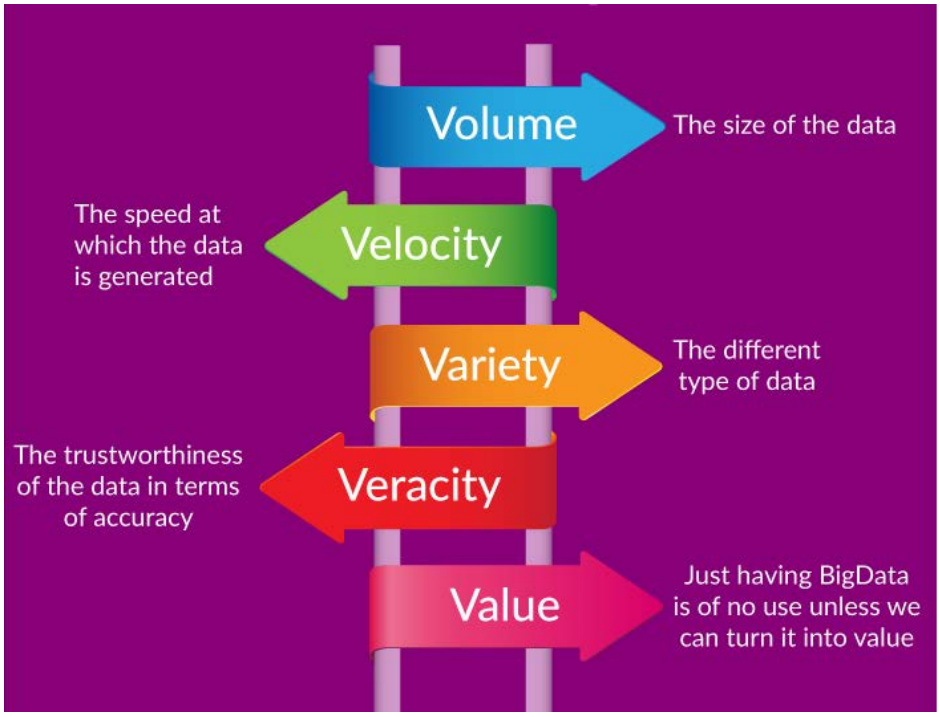

Figura 1: 5Vs do Big Data (NAVDEEP, 2017).

É necessário que fique bem defino o conceito de Big Data, para que possa ser utilizado na sua forma com mais eficaz, assim é necessário compreender os 5Vs que o compõe como na Figura 1. **Volume**, refere-se a grande quantidade de dados gerado initerruptamente nos dispositivos seja ele desde um computador pessoal até uma colisor de matéria, esses dados podem ser armazenados e analisados a qualquer momento utilizando uma técnica de sistemas distribuídos. **Velocidade**, refere-se a rapidez com que os dados são gerados, a frequência de novos dados que trafegam na internet é extremamente elevada, é possível analisa-la sem quer ela seja armazenada em uma base de dados. **Variedade**, refere-se a quantos modos é possível se encontrar um dado e pode estar em uma estrutura tabelar assim como é visto em um banco de dados relacional, porem 80% deles não tem estrutura o que leva a dificuldade na identificação e mineração deles. **Veracidade**, refere-se à confiabilidade que se pode ter nesses dados, sabendo da grande variedade é necessário que se tenha cuidado quanto a qual dado traz a informação confiável. **Valor**, refere-se a o que esse dado vai proporcionar quando for extraído,

este é o processo mais importante dos 5Vs, pois é onde será dado o pontapé para a tomada de decisão com as informações extraídas.

3. BIG DATA E SAÚDE

A área da Saúde a cada dia busca o conhecimento e a aplicação de novas tecnologias que possam auxilia-los nas tomadas de decisões cada vez mais rápidas, eficientes e com menor tempo de execução. Tudo isso porque na era que estamos vivendo denominada “Era da Informação” a comunidade, gera um grande volume de dados que só tende a aumentar a cada minuto que passa, fazendo com que obrigatoriamente, esses dados ocupem um único espaço e sejam organizados com a máxima praticidade possível de maneira que ao acessar essas informações obtenha-se um resultado imediato. O Big Data nos permite a utilização de diversas tecnologias de gerenciamento dessa massiva geração de dados como, aplicação de Data Warehouse, Data Mart, Data Mining, Web Data Mining, BI (Business Intelligence) e Cloud Computing. A necessidade de utilizar estas ferramentas está nos dados que são gerados pela era da internet e não são estruturados, visto que estão no formato de vídeos, textos, *logs*, históricos de localização e imagens (LOHR, 2012), dados especialmente ricos em relação aos mais concretos, tais como o perfil sócio demográfico ou comportamental (CHEN; CHIANG; STOREY, 2012).

Outro desafio é que, enquanto nos estudos de medicina tradicional os dados provem basicamente de uma única fonte, na era do Big Data as informações estão dispersas sendo necessário recorrer a inúmeras bases e depois padronizar e vincular as variáveis.

4. BANCO DE DADOS RELACIONAL E NÃO RELACIONAL

Dos Bancos de Dados o grande sucesso do modelo relacional destaca-se a padronização de conceitos, sua base formal e a facilidade de uso da linguagem SQL (Structured Query Language), linguagem padrão para consultas e manipulação de dados relacionais, levando o modelo SQL a ser aderido por toda comunidade

tecnológica, com o avanço e a necessidade de tratar dados surgiram outros modelos que poderiam melhorar aplicação que tratam dados variados.

Para o gerenciamento do Big Data pode se obter um aproveitamento considerável fazendo a utilização de Bancos de Dados não Relacionais ou Pós-Relacionais, definidos como Bancos de Dados NoSQL. Eles surgiram como uma solução para os problemas de armazenamento de escalabilidade, paralelismo e gerenciamento de grandes volumes de dados não estruturados. Em geral, os sistemas NoSQL tem as seguintes características (HECHT e MUHAMMAD), se baseiam em um modelo de dados não relacional, eles dependem de processamento distribuído, alta disponibilidade e escalabilidade são as principais preocupações, alguns são sem esquema e tem a capacidade de lidar com dados estruturados e não estruturados.

Para tratar dados na escala de Volume, Variedade e Velocidade é necessário a utilização do Big Data, e suas ferramentas que utilizam modelos mais apropriados como os Sistemas Gerenciadores de Bancos de Dados NoSQL, projetados para tratar imensos volumes de dados estruturados e não estruturados.

Existem diversos modelos, um deles é o colunar, como o Big Table, usados internamente pelo Google, bem como o modelo Key/value, como DynamoDB da Amazon, como também o modelo apoiado em documentos baseado no conceito proposto pelo Lotus Notes da IBM e aplicado em softwares como MongoDB.

Esta diversidade de alternativas demanda conhecimento e estudo sobre as tecnologias existentes com o intuito de aprimora-las para que seja possível tratar os dados não estruturados de uma maneira eficaz, considerando o Volume, Variedade e Velocidade com que esses dados são acessados.

5. SISTEMA DE BIG DATA

Este segmento existe há mais de seis anos no mundo sendo que no Brasil começa aparecer agora com maior destaque (DOURADO, J.). Nesse período, no mundo, ocorreram grandes avanços nas empresas que se utilizam dessas tecnologias. É o caso de empresas como Google, Yahoo e Apple, dentre outras, que armazenam

grandes volumes de dados e realização o processamento intensivo dessas informações.

O Hadoop vem de uma iniciativa da empresa Yahoo, é uma biblioteca que permite o processamento distribuído de uma quantidade massiva de dados em um *clusters* de computadores onde maquinas são escravizadas e possibilitam a utilização de seus processadores e memoria que auxiliam o armazenamento e tratamento de dados. A biblioteca é capaz de detectar e tratar falhas na camada de aplicação, garantindo a confiabilidade necessária para manipulação de dados em *cluster* de computadores, levando em consideração que este método é propenso a falhas.

Hadoop é composto de tecnologias que auxiliam o controle de todo processo de tratamento de dados, o projeto inclui um Sistema de arquivos distribuídos (HDFS), que fornece acesso aos dados do aplicativo, Hadoop YARN que realiza agendamento de trabalhos e gerenciamento de recursos de *cluster*, e por fim o Map Reduce, sistema baseado em YARN para processamento paralelo de grandes conjuntos de dados. (HADOOP, 2017).

5.1. DADOS DE SAÚDE

Para tratar os dados é necessário que se utilize de uma ferramenta que suporte a velocidade, variedade e volume que o Big Data traz consigo, no mercado existem diversas ferramentas que oferecem tais funcionalidades, para essa pesquisa foi utilizado um framework que se destaca pela sua velocidade, facilidade de uso e analise sofisticada, é o Spark, que traz um aglomerado de funcionalidades permitindo aplicação em *clusters*, executando 100 vezes mais rápido em memória e até 10 vezes mais rápido em disco**,** de fácil desenvolvimento em Java, Scala ou Python (SPARK, 2017).

A caráter introdutório foi utilizado a API PySpark que possibilita iniciantes a manipulação de Big Data em uma curva de aprendizagem que apresenta os estágios necessário para o entendimento dos processos complexo da extração e manipulação dos dados (PYSPARK, 2017).

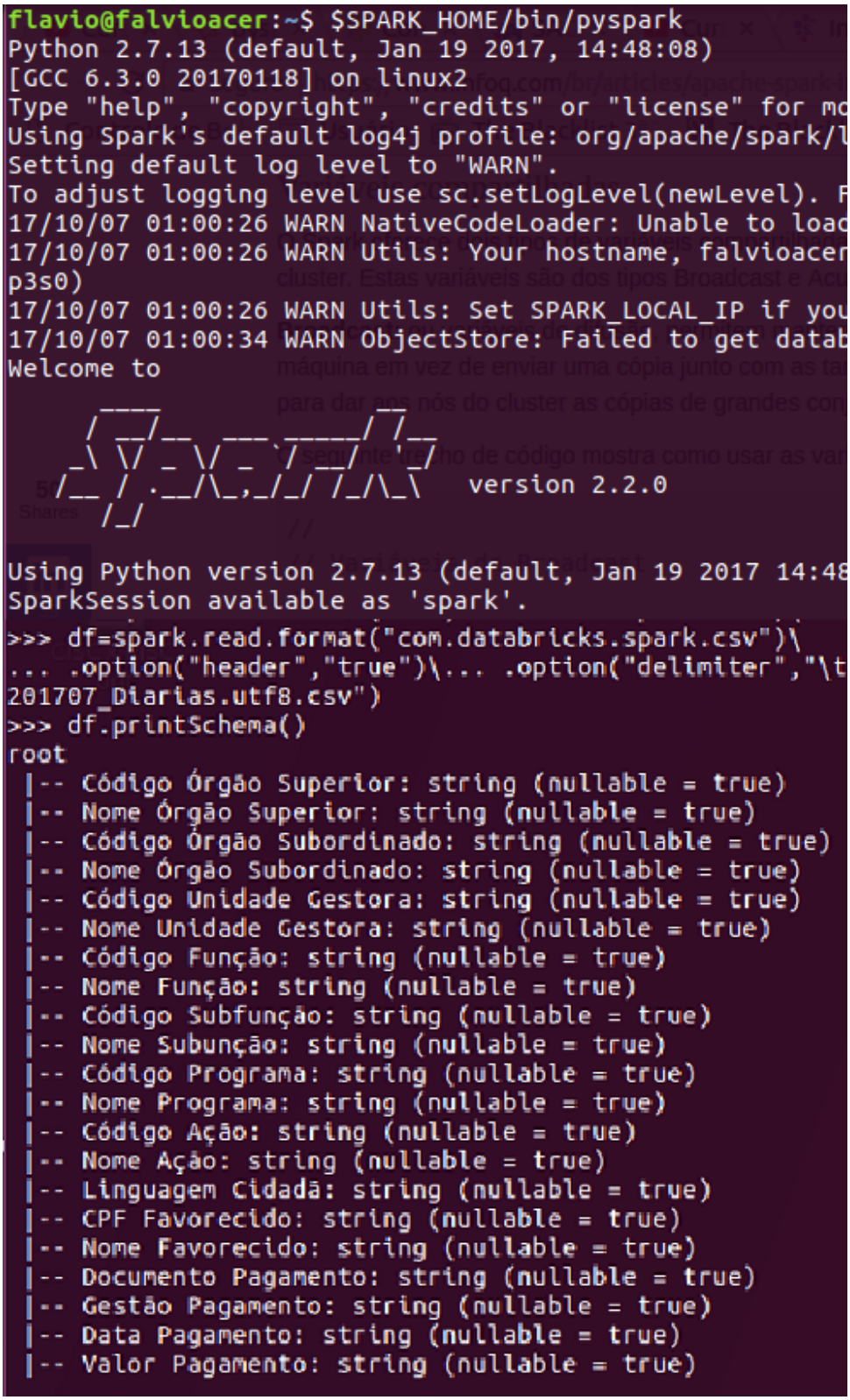

Figura 2: Arquivo csv carregado na API PySpark.

É possível manipular arquivos extraídos da rede com log, json e xml entre outras extensões, na Figura 2, é apresentado o ambiente PySpark carregado com um arquivo CSV, que possibilita tratar os dados com comando em SQL/PL, na

linguagem Pynton. O arquivo CSV, contem informações de gastos públicos, neta aplicação foi extraído informações remetentes a gastos do setor de Saúde Publica.

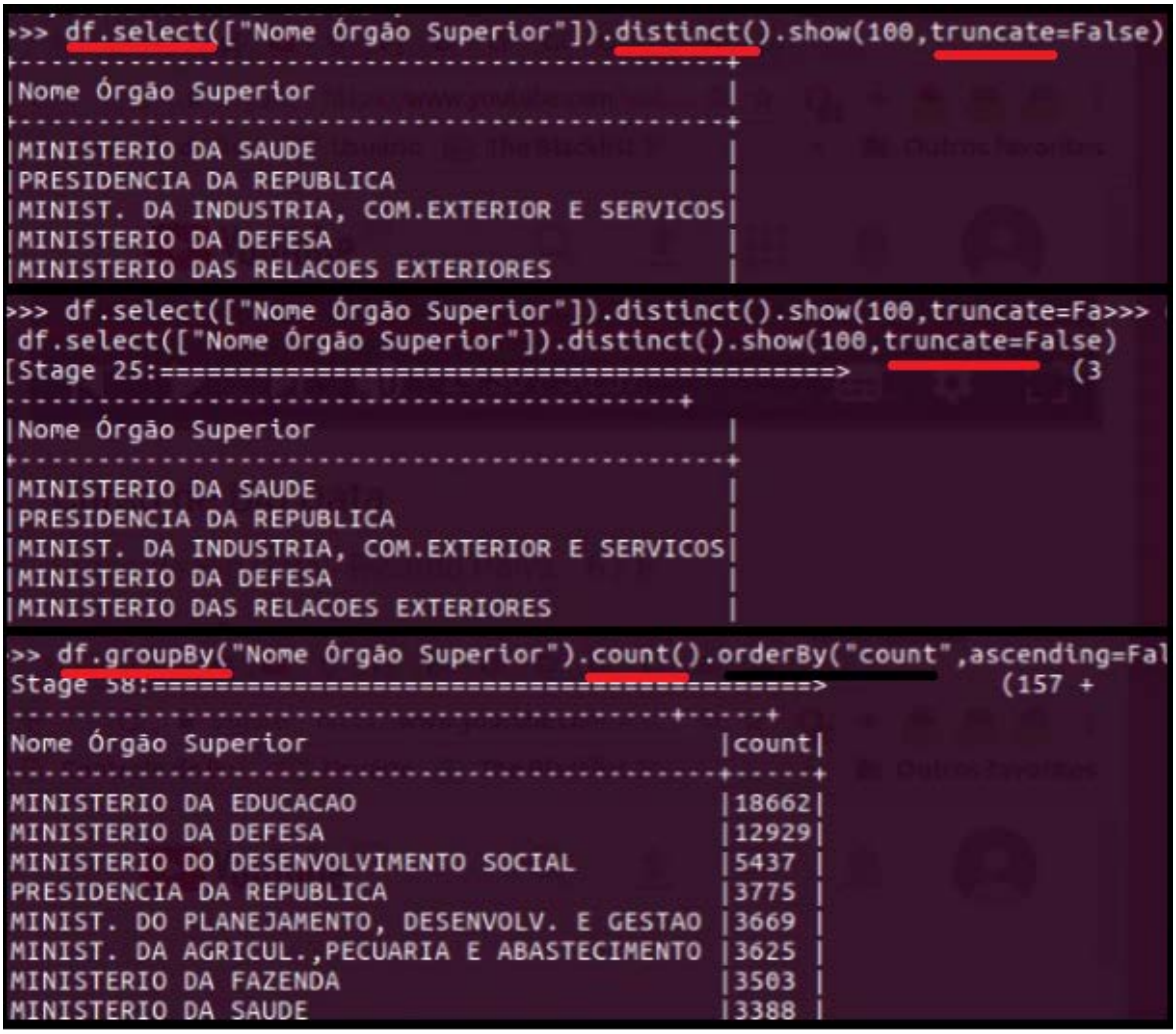

Figura 3: Tratamento de dados extensão csv utilizando Data Frame e SQL Context.

Com contexto básico de SQL é possível modelar e extrair os dados de Big Data na forma deseja, na Figura 3, presenta instruções SQL como, Select, Distinct, Truncate, GroupBy, Count e OrderBy esta e demais instruções são aceitas pela API, baseado em Data Frame que é uma coleção de dados distribuídos e organizados em forma de colunas nomeada muito similar a um banco de dados relacional.

Prosseguindo com a manipulação dos dados foi utilizado outra linguagem e outro ambiente, que tem forte influência no meio acadêmico, cientifico e profissional. A Linguagem R fornece uma base flexível e poderoso para a computação e ambiente para análise estatística e produção de gráficos, um projeto GNU com iniciativa Open Source (R STUDIO, 2017).

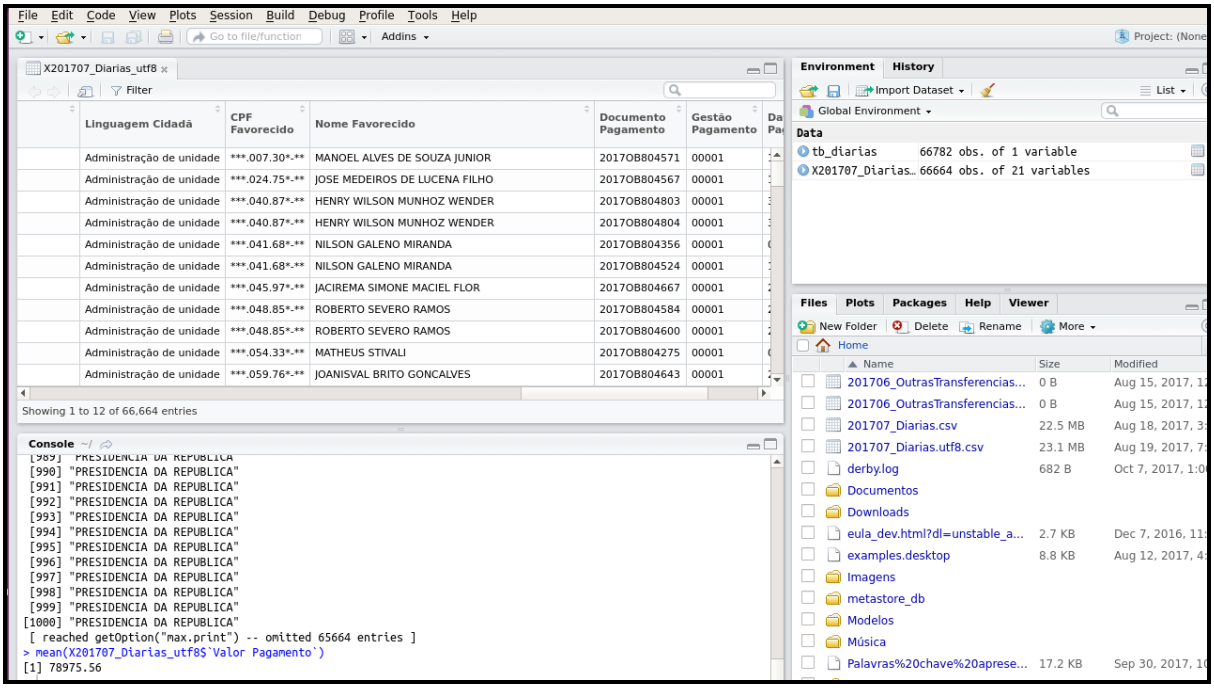

Figura 4: Arquivo Dados de Saúde, índice de surto de Dengue, extensão csv carregado no ambiente R Studio.

R Studio é um ambiente avanço em relação ao ambiente PySpark, a ferramenta oferece recursos avançado e intuitivos que contribuem para a manipulação de dados a Figura 4, mostra o ambiente com o arquivo csv carregado.

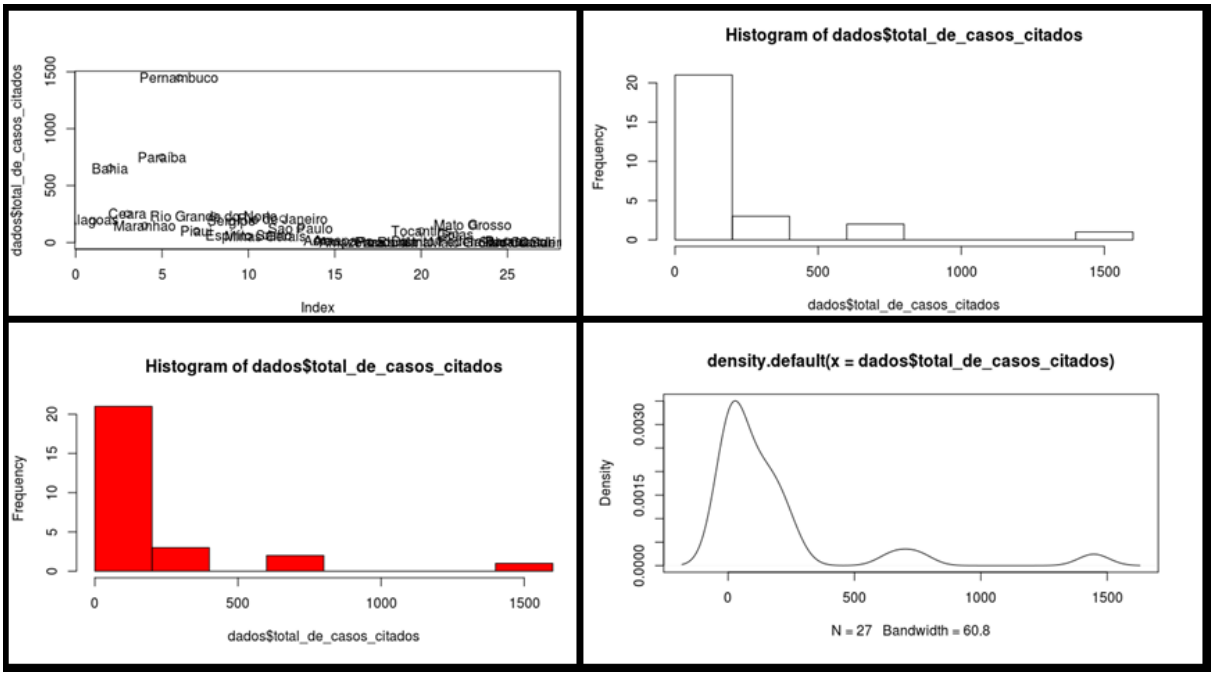

Figura 5: Exemplo Criação de Gráficos com Linguagem R no ambiente R Studio.

Com a linguagem R aliada ao seu ambiente é possível a criação de gráficos que podem ilustrar o tratamento de dados, assim como a figura 5, levando a analise intuitiva disponibilizando material pra insides, tomada de decisão e uma análise detalhada.

6. ALGORITMOS DE ANALISE

Uma área do Big Data na Saúde que ganha conhecimento e expansão é o analise preditiva de dados, que constitui de traçar um padrão com dados existentes para prever ações futura, a analise se apropria da técnica de Fenotipagem de Algoritmo, Fenotipagem é um termo utilizado comumente na área da medicina que se correlaciona com o genótipo, pode ser definida com um conjunto de características que constituem a manifestação de um genótipo ou seja parte de um conjunto de informações que compõe um indivíduo em nosso contexto um grupo de vírus ou doença. Essa técnica é aplicada para saber se um indivíduo tem as características pertencentes a uma doença sendo assim correndo risco de estar ou ter a doença avaliada.

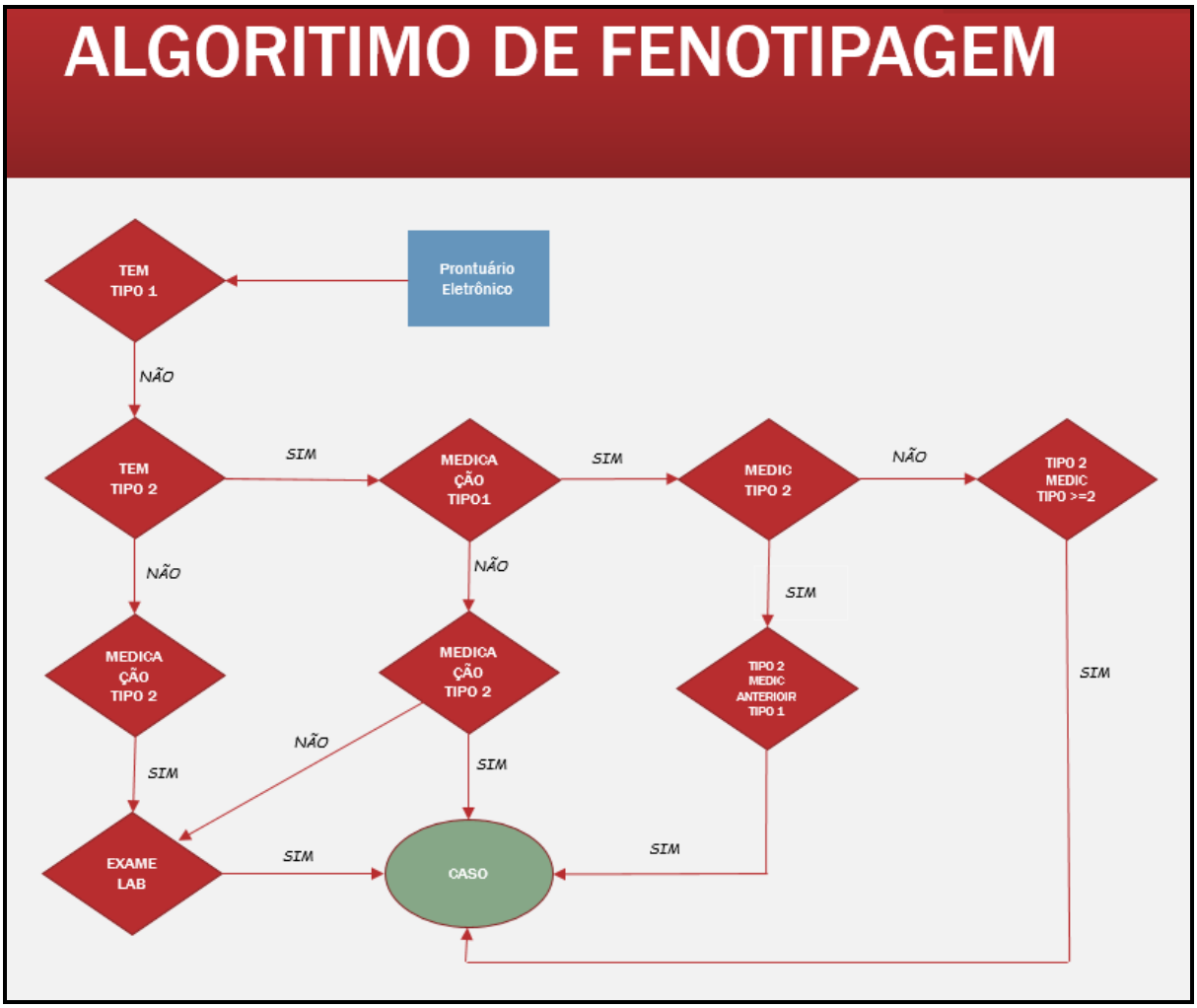

Figura 6: Analise de reconhecimento para diagnosticar um indivíduo com Diabetes utilizando Algoritmo Monolítico de Fenotipagem.

Para enquadrar um indivíduo em um grupo de risco é necessário que seja feito uma análise minuciosa de toda base de dados referente a ele, várias condições devem ser impostas para que ele seja adicionado ao quadro, na Figura 6, o algoritmo monolítico ilustras todas as condições que os dados do paciente têm que passar, a análise é realizada a partir de uma base de dados existente, um Prontuário Eletrônico, para saber se ele vai ou não se enquadrar no caso de diabetes tipo 2.

7. CONSIDERAÇÕES FINAIS

Todo esse ecossistema big data, trará importantes ganhos em termos de tempo, confiabilidade, custos, organização e qualidade, na área da saúde. Ao descobrir novas associações e pela compreensão dos padrões e tendências nos dados, a

análise de Big Data tem o potencial para melhorar o atendimento prestado ao usuário (SILVA, 2016). A potencialidade da análise de big data está apenas iniciando e já é uma realidade concreta na área da saúde, apesar existirem alguns entraves metodológicas e problemas de privacidade, a era big data traz imensas oportunidades para o avanço do conhecimento em saúde (CHIAVEGATTO FILHO).

Diante da gama de possibilidade o artigo seleciona algumas técnicas e ferramentas de coleta de dados que podem contribuir para evolução da medicina, o que estes dados que estão espalhados de diversas formas, podem nos trazer. Sendo acessados da maneira correta e com a velocidade suficiente para analisarmos as situações cotidianas.

REFERÊNCIA BIBLIOGRAFICA

AMARAL, S. DO. **Marketing da informação: abordagem inovadora para entender o mercado e o negócio da informação**. Ciência da Informação, v. 40, n. 1, p. 85–98, 2011.

AURÉLIO, Dicionário. **Dicionário Aurélio Português Online 2017.** Disponível em < http://www.dicionariodoaurelio.com > Acesso em 26 set 2017.

ARTHUR, L. **Big Data marketing: engage your customers more effectively and drive value** (Google eBook). 1. ed. New Jersey: Wiley, 2013. p. 208.

BOLLIER, David et al. **The promise and peril of big data**. Washington, DC: Aspen Institute, Communications and Society Program, 2010.

BUGHIN, J.; LIVINGSTON, J.; MARWAHA, S. **Seizing the potential of "big data"**. McKinsey Quarterly, v. 1, n. October, p. 103–109, 2011.

DOURADO, Joana. Semantix – **Treinamentos**. Disponível em: <http:// www.semantix.com.br>. Acesso em: 09 outubro de 2016.

CHEN, H.; CHIANG, R. H. L.; STOREY, V. C. **Business intelligence and analytics: from Big Data to big impact**. Mis Quarterly, v. 36, n. 4, p. 1165–1188, 2012.

CHIAVEGATTO FILHO, A. D. P. **Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. Epidemiologia e Serviços de Saúde**, Brasília, DF, v. 24, n. 2, p. 220-221, 2015.

HADOOP. **Bem-Vindo ao Apache Hadoop**. Disponível em < http://hadoop.apache.org/ > Acesso em 15 Out 2017.

LAUDON, Kenneth C.; LAUDON, Jane Price. **Sistemas de informação.** 4. ed. LTC: Rio de Janeiro,1999.

LOHR, S. **The age of Big Data**. The New York Times, p. 1–5, 2012.

NAVDEEP. **Why we Need Modern Big Data Intergration Platform**. Disponível em < https://www.xenonstack.com/blog/why-we-need-modern-bigdata-integration-platform > Acesso em 05 abr 2017.

PYSPARK. **Documentação PySpark 2.1.0. Modulo pyspark.sql.** Disponível em < http://spark.apache.org/docs/2.1.0/api/python/pyspark.sql.html > Acesso em 25 Out 2017.

R STUDIO. **Why RStudio**. Disponível em < https://www.rstudio.com/about/ > Acesso em 01 Nov 2017.

SILVA, F. A. B. **Big Data e Nuvens Computacionais: Aplicações em Saúde Pública e Genômica**. Journal of health Informatics. v. 8, n. 2 (2016).

SPARK. **Apache Spark, Computador de cluster rápido**. Disponível em < https://spark.apache.org/ > Acesso em 23 Out 2017.

ZIKOPOULOS, P; DE ROOS, D; PARASURAMAN, K; DEUTSCH, T; GILES, J; CORRIGAN, D. **Harness the power of Big Data- The IBM Big Data Platform. Emeryville: McGraw-Hill Osborne Media**, 2012.