Fatec Cruzeiro Prof. Waldomiro May
Revista Científica H-TEC Humanidades & Tecnologia

USO DE BIG DATA E IMPLEMENTAÇÃO DO PROCESSO DE EDA: UM CASE DE DADOS CLIMÁTICOS PARA GESTÃO DE AGRICULTURA

Autores
Anderson Henrique Potye de Paula[1]
José Walmir Goncalves Duque[2]

Resumo
Com o surgimento da IoT, dispositivos com diversos sensores têm a habilidade de comunicar-se com outros dispositivos ou objetos, resultando em um grande volume de dados – isso trouxe a necessidade de soluções de tecnologia da informação focadas no desempenho de armazenamento, consultas e, posteriormente, análises complexas: Big Data. A área de gestão de agricultura é um exemplo importante dessa nova realidade. Em face ao exposto, este trabalho tem como objetivo propor a infraestrutura para Big Data para a captura e armazenamento de dados meteorológicos bem como aplicação, em termos de software, do processo de EDA, a fim de preparar tal contexto para futuras análises em Data Science. Tais análises podem permitir a obtenção de conhecimentos científicos e insights de oportunidades em diferentes campos de negócios, em particular, na gestão da agricultura digital. Para atingir o objetivo deste trabalho, na primeira etapa, foi proposta uma infraestrutura de Big Data bem como dos sensores mais viáveis para uma futura implementação em termos de hardware. Na segunda etapa foi feita a implementação da infraestrutura para Big Data em termos de software bem como a aplicação do processo de EDA para fins de captura e armazenamento de dados meteorológicos. Por fim, percebeu-se que o protótipo proposto é viável para a aplicação do processo de EDA com dados meteorológicos. Após toda a extração de dados e EDA, os dados foram armazenados para análises futuras. Entende-se, por fim, que os dados armazenados podem de fato contribuir para a gestão da agricultura.
Palavras-chaves: 1. Big Data; 2. EDA; 3. IoT; 4. Data Science; 5. Gestão da agricultura

USE OF BIG DATA AND IMPLEMENTATION OF THE EDA PROCESS: A CASE OF CLIMATE DATA FOR AGRICULTURE MANAGEMENT

Abstract
With the emergence of IoT, devices with multiple sensors have the ability to communicate with other devices or objects, resulting in a large amount of data - this has brought the need for information technology solutions focused on storage performance, queries and, later complex analyzes: Big Data. The area of agricultural management is an important example of this new reality. In view of the above, this work aims to propose the infrastructure for Big Data for the capture and storage of meteorological data as well as software application of the EDA process, in order to prepare such context for future analyzes in Data Science. Such analyzes may enable the achievement of scientific knowledge and insights of opportunities in different fields of business, in particular, in the management of digital agriculture. In order to achieve the objective of this work, in the first stage a Big Data infrastructure was proposed as well as the most viable sensors for a future implementation in terms of hardware. In the second stage, the infrastructure

[1] Graduado em engenharia da computação pelo Unisal – Lorena. Email: walmir.duque@fatec.sp.gov.br
[2] Mestrado em Engenharia Eletrônica e Computação pelo ITA e docente na Fatec Prof. Waldomiro May. Email: walmir.duque@fatec.sp.gov.br

was implemented for Big Data in terms of software as well as the application of the EDA process for the purpose of capturing and storing meteorological data. Finally it was noticed that the proposed prototype is feasible for the application of the EDA process with meteorological data. After all data extraction and EDA data were stored for future analysis. It is understood, finally, that stored data can actually contribute to the management of agriculture.
Keywords: *1. Big Data; 2. EDA; 3. IoT; 4. Data Science; 5. Agriculture Management*

INTRODUÇÃO

A condição climática mundial vem mudando drasticamente como, por exemplo, as temperaturas se elevarem ou caírem em um padrão fora do normal – isso pode ser muito ruim para algumas áreas como a saúde, indústrias e agricultura. Tais setores investem amplamente em tecnologia para gerenciar, tratar, bem como fazer a análise de dados, muitas vezes, em tempo real, de forma a minimizar danos ou prejuízos potencialmente causados por tais mudanças climáticas.

A humanidade descobriu diversas formas de poder usufruir da internet, procurando sempre usá-la a seu favor no cotidiano, mas um assunto que tem se expandido em importância na atualidade é a "internet das coisas" – do inglês Internet of Things (IoT) – basicamente, trata-se de dispositivos que provem a comunicação entre "coisas", pessoas, animais, objetos etc. Tal facilidade de interação entre meios tem auxiliado a humanidade em diversas tarefas, como, por exemplo, na agricultura: um agricultor não precisa verificar constantemente suas plantações já que com sensores ele consegue medir a temperatura local, umidade, velocidade do vento, calcular uma média de como ao ambiente ao redor funciona e assim trabalhar de forma mais eficaz de acordo com as condições climáticas.

Portanto, dado o crescimento vertiginoso de dados gerados pela sociedade, inclusive pela IoT, outra importante vertente tecnológica surge com grande força: Big Data. Em todo o mundo, milhares de empresas, centros científicos, instituições estão usando Big Data para encontrar padrões onde quer que estejam os dados massivos. Big Data, por exemplo, tem oferecido suporte para prever condições meteorológicas, com quantidades massivas de dados.

Esta ferramenta pode ser usada de várias maneiras.

O maior problema em relação à grande quantidade de dados é a sua manipulação, principalmente no que se diz respeito ao armazenamento, quais os procedimentos necessários para que se consiga uma organização adequada dados, pois, se não há ferramentas que otimizem todos esses processos, pode-se tonar inviável essa aplicação.

"A utilidade do Big Data está no tratamento desse volume de dados, que provém de diversas fontes e que necessitam de um processamento de alta velocidade, na busca por um valor." (TAURION, 2013).

A principal função do Big Data é organizar os dados, manipulá-los e auxiliá-los na extração de informações significativas com o objetivo de solucionar problemas de onde tais dados foram extraídos, portanto, são utilizadas ferramentas específicas que em sua maioria requerem um grande poder de processamento. Devido ao seu grande fluxo de dados, é necessário o estudo de bancos de dados para verificar qual melhor meio de armazenamento para se obter o acesso e manipulação mais eficientes e efetivos.

A fim de poder realizar uma extração de informações mais significativas, é necessário aplicar o processo de EDA (Exploratory Data Analysis, em português, Análise Exploratória de Dados), para que se tenha uma melhor visualização e entendimento desses dados, permitindo uma melhor interpretação e, assim, uma eficaz preparação para futuras análises em Data Science.

Atualmente o conceito de agricultura digital é mais recente e mais amplo do que agricultura de precisão, pois envolve a coleta de dados no campo usando uma variedade de sensores e o processamento desses dados por meio do uso de modelos físicos e agronômicos para previsões e tomadas de decisão. O conceito de agricultura de precisão tem mais foco no georreferenciamento.

Em outras palavras, a Agricultura Digital permite criar simulações computacionais de como diferentes culturas agrícolas se comportam em diferentes condições, usando os dados coletados para identificar padrões e conhecimentos importantes para a tomada de decisão sobre qual variedade plantar e onde, e com que quantidade de insumos, por exemplo.

Percebe-se que o processo de EDA é extremamente importante para o entendimento dos dados, bem como para a preparação, tornando uma futura análise de dados mais efetiva, e, também, esse processo torna possível a utilização de dados de

outras áreas para tomar decisões mais assertivas em uma área específica como, por exemplo, o uso de dados climáticos na agricultura.

Em face ao exposto, este trabalho tem como objetivo propor a infraestrutura para Big Data para a captura e armazenamento de dados meteorológicos bem como aplicação, em termos de software, do processo de EDA, a fim de preparar tal contexto para futuras análises em Data Science. Tais análises podem permitir a obtenção de conhecimentos científicos e insights de oportunidades em diferentes campos de negócios, em particular, na gestão da agricultura digital.

Para atendimento do objetivo geral, apresenta-se a seguir os seguintes objetivos específicos:

- Implantar a infraestrutura de software adequada para viabilizar o processo de captura e armazenamento de dados meteorológicos, em particular, Hadoop, banco de dados não relacional Hive;

- Aplicar o processo de EDA para o contexto do problema, preparando dados para futuras análises, utilizando as melhores práticas em termos de métodos, técnicas e ferramentas.

2 FUNDAMENTOS DE SENSORES, IOT, BIG DATA E EDA

2.1 Sensores

Sensores são dispositivos capazes de captar ações ou estímulos externos e responder em consequência. Estes dispositivos recebem as grandezas como temperatura, umidade e pressão atmosférica e as transformam em grandezas elétricas, em suma, são ferramentas que permitem obter informação do meio para posterior interação com ele.

O sinal de saída de um sensor normalmente deve ser manipulado antes de ser utilizado por um sistema de controle, pois, nem sempre tais dados estarão adequados aos níveis de tensão ou corrente requerido para tal sistema. Para isso, é necessário utilizar um circuito de condicionamento de sinal.

Os sensores podem ser analógicos ou digitais: os sensores analógicos podem assumir em seu sinal de saída, ao qualquer valor de tensão ao longo do tempo, dentro de sua faixa de operação; por outro lado, os sensores digitais podem apenas assumir dois

valores distintos de tensão, que são tratados como zero ou um lógico (WENDDLING, 2010).

2.2 Internet das Coisas (IoT)

A origem do nome Internet das Coisas é atribuída a Kevin Ashton. Internet das Coisas foi o nome de uma apresentação feita por ele, em 1999, na empresa Procter & Gamble (P&G). Mais tarde, em 2009, em um artigo publicado por meio do RFID Journal, ele fez referência à apresentação e cita o que é tido por muitos como a definição de IoT.

> …Se tivéssemos computadores que soubessem de tudo o que há para saber sobre coisas, usando dados que foram colhidos, sem qualquer interação humana, seríamos capazes de monitorar e mensurar tudo, reduzindo o desperdício, as perdas e o custo. Gostaríamos de saber quando as coisas precisarão de substituição, reparação ou atualização, e se eles estão na vanguarda ou se tornaram obsoletas. (ASHTON, 2009)

O conceito IoT é atribuído à possibilidade de conexão de qualquer coisa à internet. As coisas, ou objetos conectados, podem ou não possuir interações entre si, compartilhando informações, gerando e armazenando dados, tudo de uma forma rápida e segura. A IoT atualmente gera uma grande massa de dados pois existem bilhões de dispositivos conectados à rede, em crescimento vertiginoso.

2.3 Hive

Antes de falar sobre o Hive é importante explicar sobre o conceito de Data Warehouse. Inmon (2005) define Data Warehouse como uma coleção de dados integrados, orientados por assuntos, não voláteis e variáveis com o tempo, na qual oferece suporte ao processo de tomada de decisões. Uma arquitetura deste tipo armazena, de forma centralizada, os dados granulares de uma determinada empresa e permite uma análise elaborada destas informações que são coletadas de diferentes fontes.

De acordo com Kimball e Ross (2013), um dos principais objetivos de um Data Warehouse é facilitar o acesso à informação contida nos dados armazenados, de forma que não apenas os desenvolvedores sejam capazes de interpretar, mas também usuários com uma visão voltada para o negócio e não somente para aspectos técnicos de

implementação. Este tipo de solução estrutura os dados para auxiliar a realização de consultas e análises.

O Apache Hive é uma ferramenta open-source para Data Warehousing construída no topo da arquitetura Hadoop (THUSOO et al., 2009). Este projeto foi desenvolvido pela equipe do Facebook para atender as necessidades de análise do grande volume de informações geradas diariamente pelos usuários desta rede social. A motivação encontrada para a criação deste projeto se deu pelo crescimento exponencial da quantidade de dados processados pelas aplicações de BI (Business Intelligence), tornando as soluções tradicionais para Data Warehouse inviáveis, tanto no aspecto financeiro, como computacional.

Além das aplicações de BI utilizadas internamente pelo Facebook, muitas funcionalidades providas pela empresa utilizam processos de análise de dados. Até o ano de 2009 a arquitetura para este requisito era composta por um Data Warehouse que utilizava uma versão comercial de um banco de dados relacional. Entre os anos de 2007 e 2009, a quantidade de dados armazenados cresceu de forma absurda, passando de 17 terabytes para 700 terabytes. Alguns processos de análise chegavam a demorar dias, quando executados nestas condições.

Segundo Thusoo et al. (2009), para resolver essa importante questão do grande volume de dados, a equipe do Facebook decidiu adotar o Hadoop como solução. A escalabilidade linear, capacidade de processamento distribuído e habilidade de ser executado em clusters compostos por hardwares de baixo custo motivaram a migração de toda antiga infraestrutura para esta plataforma. O tempo de execução dos processos de análise que antes podiam levar dias foi reduzido para apenas algumas horas.

No entanto, a solução com o Hadoop exigiu que os usuários desenvolvessem programas MapReduce para realizar qualquer tipo de análise, até mesmo pequenas tarefas, como por exemplo, contagem de linhas e cálculos de médias aritméticas. Essa situação prejudicava a produtividade da equipe, pois nem todos eram familiarizados com esse paradigma e, em muitos casos, era necessário apenas realizar análises que seriam facilmente resolvidas com o uso de uma linguagem de consulta SQL por exemplo. De acordo com Thusoo et al. (2009), o Hive foi desenvolvido para facilitar este processo introduzindo os conceitos de tabelas, colunas e um subconjunto da linguagem SQL ao universo Hadoop, mantendo todas as vantagens oferecidas por esta arquitetura.

2.4 Big Data

Big Data é um conceito atribuído a um grande volume de dados, não necessariamente estruturados, os quais são gerados a todo momento. Nos tempos atuais, com o avanço constante da tecnologia e da informação, existem trilhões de bytes gerados por dia, provenientes de inúmeros locais, processados e transformados em informações que possam ser utilizadas de alguma forma útil, como por exemplo, um agricultor que pode conseguir antecipar algum problema, como uma doença específica na sua lavoura, ou um ter mais eficácia no controle de suas plantações, seja na colheita, ou quando seja necessário mais irrigação dependendo das condições climáticas, assim conseguindo trabalhar com maior qualidade e produtividade.

O objetivo do Big Data é melhorar os processos de trabalho dos seus utilizadores, por permitir interpretações rápidas e valiosas sobre as tendências de mercado, comportamento de consumo e oportunidades potenciais, portanto, o Big Data começa a ser considerado por cada vez mais empresas que procuram um melhor posicionamento no mercado.

Para uma melhor compreensão, o Big Data é descrito em 5 pilares fundamentais, conhecidos como os 5 V's, sendo eles:

• Volume

Segundo Marr (2015), o volume é a enorme quantidade de dados produzidos diariamente pelas empresas, por exemplo. A geração de dados é tão grande e complexa que não pode mais ser salva ou analisada usando métodos convencionais de processamento de dados.

O Big Data possibilita o armazenamento de fontes variadas tais como transações comerciais, redes sociais e dados automáticos transmitidos por máquinas ou sensores. Tudo o que acontecer em uma empresa ou organização fica registrado em uma infraestrutura integrada. Em outros tempos essa era uma realidade impossível: armazenar uma quantidade tão vasta de dados, transformar em informações e usar tais informações para tomar decisões estratégicas.

• Velocidade

Marr (2015) também cita o conceito de velocidade com que os dados são gerados, analisados e reprocessados: hoje isso é possível em uma fração de segundo, conhecido como tempo real. Como se pode calcular, os dados úteis das organizações são disponibilizados a uma incrível velocidade, caso contrário, não faria sentido serem utilizados, porque perderiam a validade. Felizmente, o Big Data oferece recursos tecnológicos para lidar com a possibilidade de captura e armazenamento de dados, em tempo real, proveniente de sensores RFID, sensores industriais, dispositivos móveis e medidores inteligentes para lidar com a velocidade do mundo informatizado.

• Variedade

Marr (2015) também se refere ao conceito de Variedade que trata da diversidade de tipos de dados e fontes de dados: é sabido que 80% dos dados no mundo atual não são estruturados e, à primeira vista, não mostram nenhuma indicação de relacionamentos. Graças ao Big Data e seus algoritmos, os dados podem ser classificados de maneira estruturada e examinados quanto a relacionamentos. Os dados nem sempre compreendem apenas conjuntos de dados convencionais, mas também imagens, vídeos e gravações de fala. Este é um dos pilares do Big Data que se refere aos diferentes formatos e estruturas em que os dados são gerados. Há importantes dados numéricos tais como áudios e vídeos que precisam ser organizados e estudados de acordo com os objetivos de uso de uma organização. A variedade de dados de um determinado assunto em análise permite à empresa ter uma noção mais aprofundada dele.

• Veracidade

Marr (2015) também define a veracidade como autenticidade e credibilidade dos dados. O Big Data envolve o trabalho com todos os graus de qualidade, já que o fator Volume pode resultar em uma falta de qualidade.

O pilar da veracidade refere-se aos dados que resultam da dinâmica humana, em particular os que são registrados nos motores de pesquisa, navegadores browser ou interação de redes sociais. Os dados catalogados são entendidos como interações reais, válidos para o Big Data.

• Valor

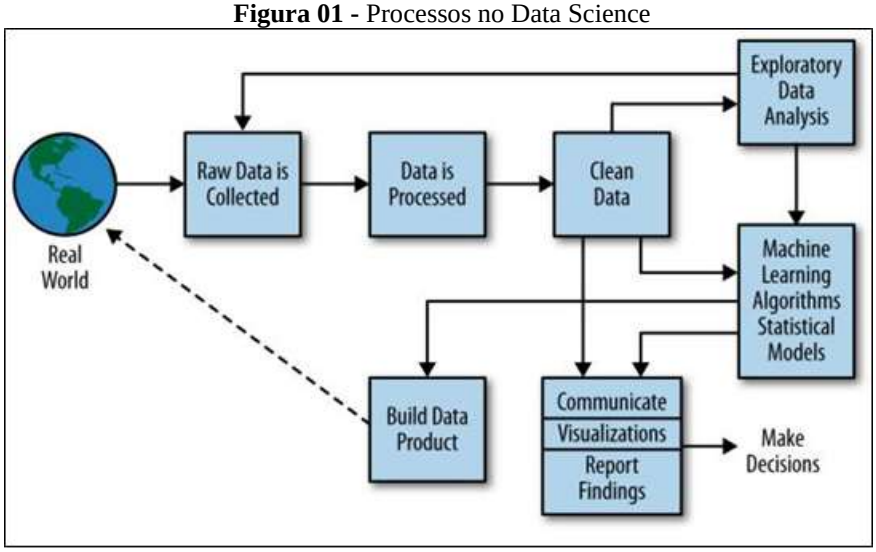
O Big Data tem a capacidade de distinguir dados e informações que apresentam maior valor para o negócio, como a página com mais visitas, a página que gera mais interação ou o canal com maior efetividade em conversão de vendas. Marr (2015) diz que apenas ter acesso a Big Data não é bom a menos que os dados possam ser transformados em valor para o negócio, ou seja, transformação dos dados em informações valiosas e significativas para tomadas de decisão.

2.5 EDA (Exploratory Data Analysis)

EDA, uma sigla em inglês para *Exploratory Data Analysis*, em português, Análise Exploratória de Dados – trata-se de uma atitude, um estado de flexibilidade, olhar para aquelas coisas que acreditamos que não estão lá, bem como aqueles que acreditamos estar lá. (TUKEY, 1986 apud O'Neil, 2014, tradução nossa).

Dentro dos processos da área de Data Science,o EDA é um dos processos mais importantes pois essa é a etapa na qual o analista consegue um entendimento básico dos seus dados: o EDA permite interpretar e entender as relações existentes entre as variáveis analisadas e, a partir disso, identificar se os dados seguem algum modelo conhecido que permita estudar o fenômeno sob análise ou, então, se é necessária a sugestão de um novo modelo. A Figura 01 a seguir mostra o processo de Data Science como um todo.

Figura 01 - Processos no Data Science



Fonte: O'Neil (2014)

Em suma, o EDA tem como finalidade examinar os dados antes da aplicação de qualquer técnica estatística, no Data Science. Desta forma, o analista consegue um

entendimento básico de seus dados e das relações existentes entre as variáveis analisadas, fazendo todo o tratamento prévio necessário para a posterior aplicação da técnica de Data Science em si.

2.6 Pandas

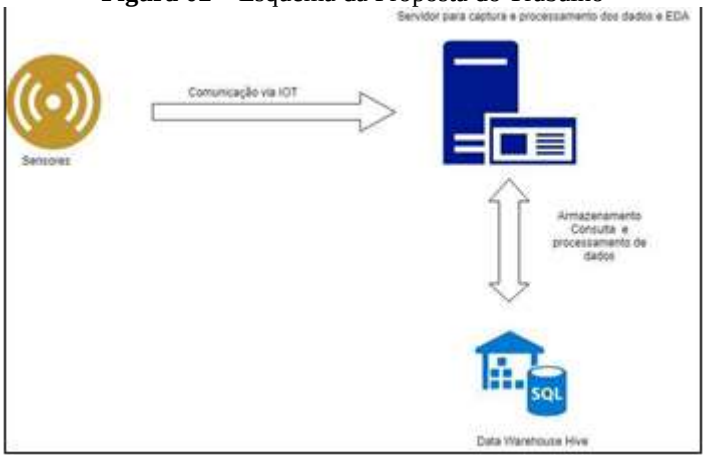
Segundo Souza (2017), Pandas é uma biblioteca da linguagem Python que oferece alta performance para suas estruturas de dados e ferramentas de análise e manipulação de dados. Pandas foi criada por Wes McKinney e atualmente conta com suporte de algumas organizações como Anaconda Inc. e o Paris-Saclay Center for Data Science.

Alguns dos principais recursos no qual Pandas oferece é o tratamento de dados nulos, redimensionamento de Dataframes (e suas features), leitura de dados de diversos formatos de entrada como XML, CSV, dentre outras.

3 MATERIAIS E MÉTODOS

Este trabalho foi realizado em duas etapas: a primeira etapa constitui-se da proposta da infraestrutura de Big Data bem como dos sensores mais viáveis para uma futura implementação em termos de hardware. A segunda etapa constitui-se da implementação da infraestrutura para Big Data em termos de software bem como a aplicação do processo de EDA para fins de captura e armazenamento de dados meteorológicos. A Figura 02 a seguir mostra um resumo esquemático da proposta para realização deste projeto.

Figura 02 – Esquema da Proposta do Trabalho



Fonte: Os Autores

3.1 Equipamentos Propostos

Nesta seção apresenta-se a descrição técnica de hardware e software da infraestrutura de Big Data proposta para esse protótipo. Em particular, apresentam-se os sensores mais viáveis para uma futura implementação, dada a gama de diversidade presente no mercado.

3.1.1 Sensores

Foram propostos sensores de alta precisão – a seguir a lista bem como sua descrição técnica:

- Sensor de umidade: HDC1080 - Texas Instruments;
- Sensor de pressão atmosférica: PX2780-60B5V - Omega Engineerings;
- Sensor de temperatura: TMP468 - Texas Instruments;
- Sensor de radiação solar: MS-802 – EKO;
- Sensor de direção e velocidade do vento:MeteoWind2 - MB Control & System PVT. Ltd;
- Sensor de precipitação: Rain[e] Lambretch Meteo Ltd.

3.1.2 Computadores

Para captação processamento e armazenamento dos dados, recomenda-se a utilização de computadores desktop, para instalação dos sistemas operacionais necessários ou até mesmo virtualizar o ambiente para se trabalhar utilizando um sistema em uma máquina virtual, aproveitando o desempenho da máquina física.

Propõe-se o uso de uma máquina desktop com essas especificações para realizar a instalação do sistema e prepará-lo para captar os dados, bem como processar e armazená-los. A seguir configuração mínima básica para esse protótipo:

- Processador: Intel® Core™ i5- 4590 CPU @ 3.30GHz 3.30GHz;
- Memória RAM: 8,00 GB;
- HD:100 GB.

Evidentemente um computador com configuração de um servidor seria o ideal para melhor desempenho, mas, foi estudada máquina com configuração mínima para viabilizar esse protótipo.

3.1.3 Softwares utilizados no Processo de EDA

A seguir apresentam-se os softwares utilizados na infraestrutura de Big Data necessária para captação, armazenamento e processamento dos dados, no processo de EDA, de forma a plena utilização em futuras análises por Data Science. Para a montagem do protótipo de infraestrutura de Big Data utilizou-se os seguintes softwares:

- Oracle Virtualbox 5.2.15;
- Sistema Operacional CentOS 6.10(Cloudera);
- Anaconda;
- Python 3.7;
- Hive.

3.1.4 Processo de EDA

A partir da especificação da infraestrutura de Big Data, em termos de hardware e software, foram aplicadas técnicas de visualização e tratamento dos dados - uma primeira análise foi realizada e foram removidos dados não necessários, a partir da sua interpretação. Ao final do EDA, os dados tratados foram armazenados no Hive para futuras análises.

No presente script feito na linguagem Python, é necessária a importação de algumas bibliotecas, nas quais aumentam a eficácia na criação do script. A seguir apresentam-se as bibliotecas que foram utilizadas – foi atribuído um nome temporário a cada lib para poder montar o script de forma mais prática:

- import pandas as pd
- import re
- import seaborn as sns
- import numpy as np
- import scipy as sp

- from pyhive import hive as db
- import matplotlib.pyplot as plt

3.1.4.1 Visualização dos Dados

Nessa parte do processo de EDA, começa a ser realizada a visualização dos dados: pode-se perceber alguma falha em algum ponto do dataset em análise como, por exemplo, dados que estão nulos ou corrompidos sendo mostrados como “NaN”.

O dataset que foi utilizado para aplicar o processo de EDA no projeto é proveniente do site Kaggle, disponibilizado pela organização “PROPPG/PPG em Informática - Doutorado e Mestrado”. Trata-se de dados de dados de clima de 122 estações da região Sudeste do Brasil,

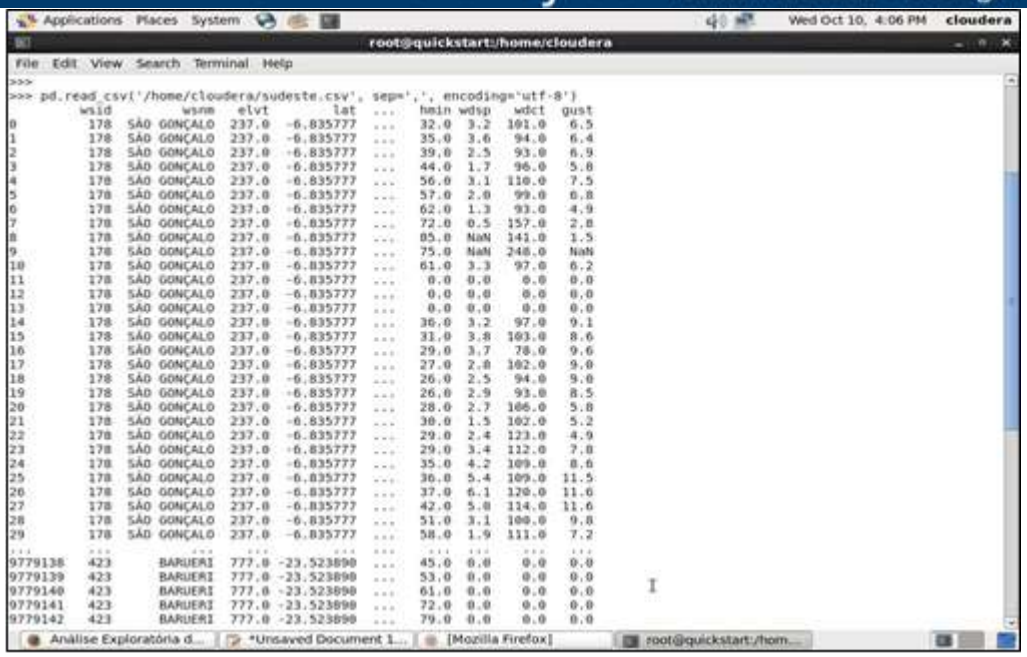
que incluem os estados de São Paulo, Rio de Janeiro, Minas Gerais e Espírito Santo. A base de dados foi acessada em Ago.2018 e está disponível no link a seguir: https://www.kaggle.com/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region

A seguir é apresentado código em Python utilizado para leitura do dataset e armazenamento em um dataframe:

```
>>>df = pd.read_csv('/home/cloudera/sudeste.csv', sep=',', encoding='utf-8')
```

Através desse comando é possível visualizar os dados presentes no dataset, como mostra a Figura 03 a seguir:

Figura 03 - Visualização dos dados em Python



Fonte: Os Autores

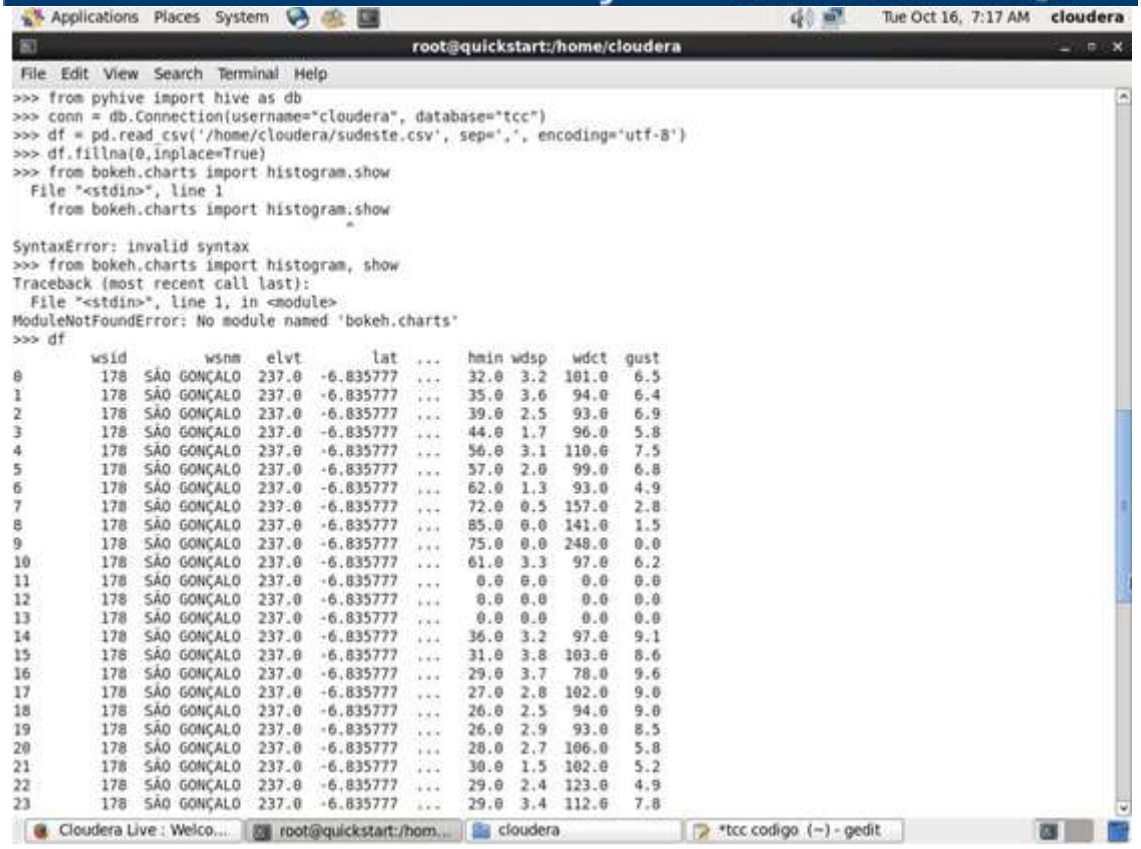
3.1.4.2. Tratamento dos Dados

Como visualizado no dataframe, alguns dados estão corrompidos ou nulos (“NaN”) e, para realizar seu tratamento, utilizou-se o método de transformar os dados corrompidos em zero por meio do comando:

```
df.fillna(0,inplace=True)
```

Uma vez aplicado esse comando, pode-se visualizar o resultado desse pré-processamento, como mostra a Figura 04 a seguir:

Figura 04 - Visualização pós tratamento dos dados em Python



Fonte: Os Autores

Após esse primeiro tratamento dos dados, faz-se importante remover as colunas que não são de interesse para futuras análises. O comando utilizado para tal é:

```
data.drop(['elvt','lat','lon','inme','date','yr','mo','da','hr'], axis = 1, inplace = True)
```

Após a remoção dos dados desnecessários, carrega-se novamente o dataframe para visualização do resultado de mais esse pré-processamento, conforme mostra a Figura 05, a seguir:

Figura 05- Dados após a exclusão de colunas não importantes

```
>>> df.drop(['elvt','lat','lon','inme','date','yr','mo','da','hr'], axis = 1, inplace = True)
>>> df
         wsid         wsnm         city prov                 mdct  prcp  ...  hmdy  hmax  hmin  wdsp  wdct  gust
0         178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 00:00:00   0.0  ...  35.0  58.0  32.0   3.2 101.0   6.5
1         178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 01:00:00   0.0  ...  39.0  39.0  35.0   3.6  94.0   6.4
2         178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 02:00:00   0.0  ...  44.0  44.0  39.0   2.5  93.0   6.9
3         178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 03:00:00   0.0  ...  58.0  58.0  44.0   1.7  96.0   5.8
4         178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 04:00:00   0.0  ...  57.0  58.0  56.0   3.1 110.0   7.5
5         178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 05:00:00   0.0  ...  62.0  62.0  57.0   2.0  99.0   6.8
6         178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 06:00:00   0.0  ...  72.0  72.0  62.0   1.3  93.0   4.9
7         178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 07:00:00   0.0  ...  86.0  89.0  72.0   0.5 157.0   2.8
8         178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 08:00:00   0.0  ...  93.0  94.0  85.0   0.0 141.0   1.5
9         178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 09:00:00   0.0  ...  75.0  94.0  75.0   0.0 248.0   0.0
10        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 10:00:00   0.0  ...  61.0  76.0  61.0   3.3  97.0   6.2
11        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 11:00:00   0.0  ...   0.0   0.0   0.0   0.0   0.0   0.0
12        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 12:00:00   0.0  ...   0.0   0.0   0.0   0.0   0.0   0.0
13        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 13:00:00   0.0  ...   0.0   0.0   0.0   0.0   0.0   0.0
14        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 14:00:00   0.0  ...  36.0  42.0  36.0   3.2  97.0   9.1
15        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 15:00:00   0.0  ...  32.0  37.0  31.0   3.8 103.0   8.6
16        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 16:00:00   0.0  ...  31.0  34.0  29.0   3.7  78.0   9.6
17        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 17:00:00   0.0  ...  29.0  31.0  27.0   2.8 102.0   9.0
18        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 18:00:00   0.0  ...  27.0  29.0  26.0   2.5  94.0   9.0
19        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 19:00:00   0.0  ...  28.0  30.0  26.0   2.9  93.0   8.5
20        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 20:00:00   0.0  ...  30.0  33.0  28.0   2.7 106.0   5.8
21        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 21:00:00   0.0  ...  40.0  42.0  30.0   1.5 102.0   5.2
22        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 22:00:00   0.0  ...  29.0  42.0  29.0   2.4 123.0   4.9
23        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-06 23:00:00   0.0  ...  35.0  35.0  29.0   3.4 112.0   7.8
24        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-07 00:00:00   0.0  ...  40.0  40.0  35.0   4.2 109.0   8.6
25        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-07 01:00:00   0.0  ...  37.0  41.0  36.0   5.4 109.0  11.5
26        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-07 02:00:00   0.0  ...  42.0  42.0  37.0   6.1 120.0  11.6
27        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-07 03:00:00   0.0  ...  51.0  51.0  42.0   5.0 114.0  11.6
28        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-07 04:00:00   0.0  ...  58.0  58.0  51.0   3.1 100.0   9.8
29        178  SÃO GONÇALO  São Gonçalo   RJ  2007-11-07 05:00:00   0.0  ...  59.0  59.0  58.0   1.9 111.0   7.2
...       ...          ...          ...  ...                  ...   ...  ...   ...   ...   ...   ...   ...   ...
9779138   423      BARUERI      Barueri   SP  2016-09-29 18:00:00   0.0  ...  53.0  55.0  45.0   0.0   0.0   0.0
9779139   423      BARUERI      Barueri   SP  2016-09-29 19:00:00   0.0  ...  61.0  62.0  53.0   0.0   0.0   0.0
```

Fonte: Os Autores

3.1.4.3 Processo de EDA

Após o pré-processamento de tratamento de dados e sua carga no dataframe, é possível então a realização da análise exploratória dos dados, analisando os dados com interesse em constituir um futuro modelo de análises, em particular análise preditiva. Na Figura, a seguir, apresenta-se o código utilizado para analisar a possibilidade de eliminação de medições de temperaturas muito extremas, pois podem impactar de forma ruidosa o cálculo de médias:

Figura 06 - Código em Python com as funções definidas

```
>>>cidadetemperaturamax = df.groupby('wsnm').temp.max().sort_values()
>>>cidadetemperaturamin = df.groupby('wsnm').temp.min().sort_values()
>>>cidadeumidademax = df.groupby('wsnm').hmdy.max().sort_values()
>>>cidadechuvamaxemmili = df.groupby('wsnm').prcp.max().sort_values()
>>>cidadechuvaminemmili = df.groupby('wsnm').prcp.min().sort_values()
```

Fonte: Os Autores

Após a realização do processo de EDA pode ser gerado um novo arquivo .csv para facilitar inserções nas bases de dados e também para futuras análises.

```
df.to_csv(tcc, sep=',', encoding='utf-8')
```

Após esse comando, o dataframe é exportado e salvo em um .csv de forma a facilitar sua carga no banco de dados. Após isso, é preciso um comando para gerar as colunas para inserção no banco de dados – isso ajudará na criação de tabelas:

```
Head -1 tcc.csv
```

Para evitar que no momento de realizar o carregamento do .csv no banco de dados os dados apareçam como nulos, é necessário fazer um arquivo .csv apenas com os valores e sem o cabeçalho. Isso pode ser conseguido por meio do comando:

```
Sed ‘1d’ tcc.csv > tcc_clean.csv
```

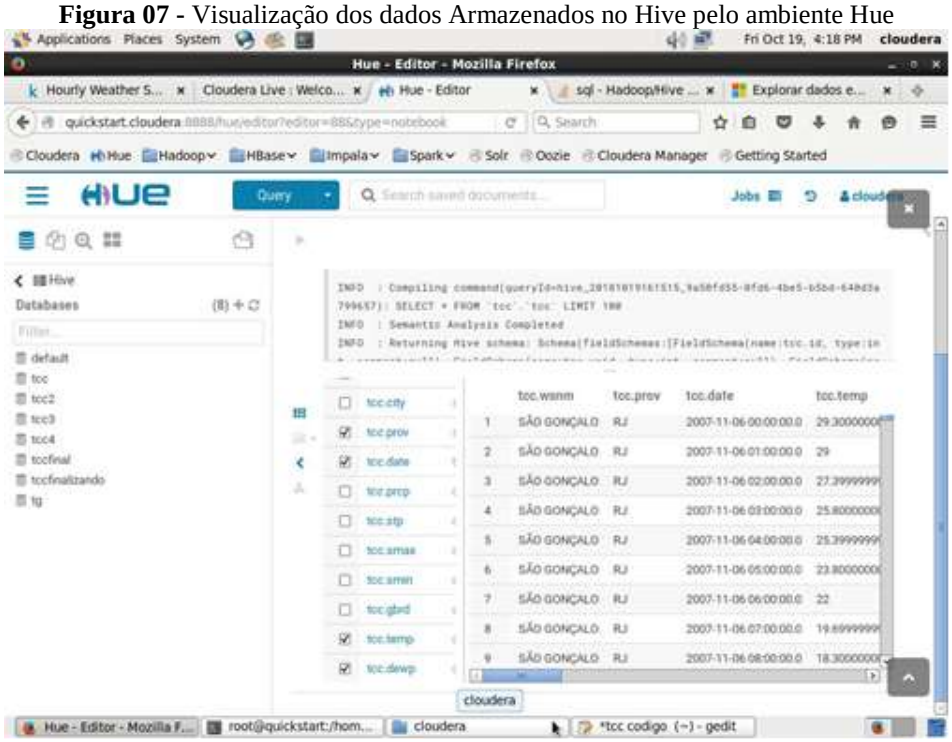
Após isso, faz-se necessário apenas realizar o carregamento desse .csv já preparado para inserção no Hive no qual é feito a partir do comando.

```
LOAD DATA local INPATH '/home/cloudera/tcc_clean.csv' OVERWRITE INTO TABLE tg;
```

Por fim, já no ambiente Hive, usando uma interface gráfica chamada Hue, é possível observar no servidor Hive a tabela já com todos os dados inseridos, como mostra a Figura 07. Nessa Figura também foi utilizado um comando SQL para demonstrar o uso desse ambiente, portanto já um banco de dados de Big Data, o próprio Hive.

```
SELECT * FROM tcc limit 10;
```

Figura 07 - Visualização dos dados Armazenados no Hive pelo ambiente Hue



Fonte: Os Autores

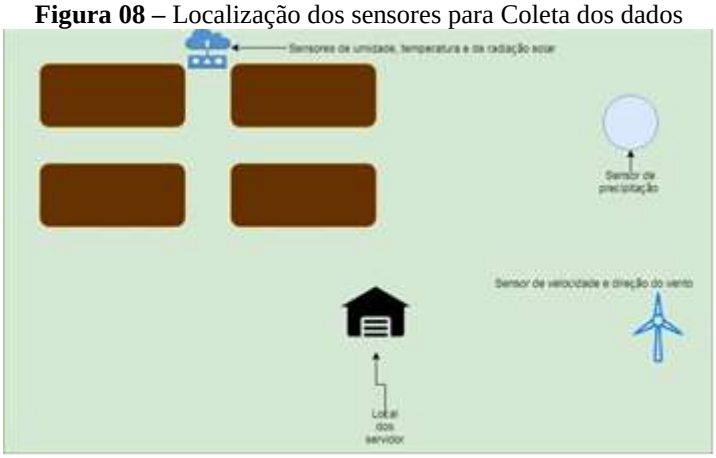
Percebe-se, ao final do EDA, a eliminação de dados não relevantes para análises futuras usando algoritmos de Data Science. Também no EDA foi possível entender os dados e manter somente aqueles que podem ser, de fato, úteis para Gestão de Agricultura. Em particular, a remoção de temperaturas extremas do dataset, tanto máximas como mínimas, já podem melhorar significativamente as análises estatísticas.

4 RESULTADOS

Após a especificação da proposta da infraestrutura de Big Data bem como dos sensores mais viáveis para uma futura implementação e aplicação do processo de EDA, percebeu-se a possibilidade de uso efetivo de dados meteorológicos para um agricultor pode ter uma decisão mais assertiva em sua gestão.

Neste trabalho, adotou-se um dataset que contém os dados que os sensores especificados certamente poderiam captar. Na Figura 08 a seguir apresenta-se como poderia ser feita a implantação desses sensores:

Figura 08 – Localização dos sensores para Coleta dos dados



Fonte: Os Autores

A partir da aplicação do processo de EDA, pela infraestrutura de Big Data proposta, pode-se perceber a versatilidade para a exploração dos dados, por meio de visualização dos dados inclusive, podendo assim realizar a elaboração de um modelo mais efetivo para análise.

Nas Figuras 09, 10 e 11, a seguir, respectivamente, apresentam-se os dados considerados como mais relevantes para a futura elaboração de um modelo de análise de

dados meteorológicos, para gestão de agricultura, sendo eles Temperatura (°C), Umidade (%) e Precipitação de chuva (mm).

Figura 09 – Temperatura

Figura 10 – Umidade

Figura 11 – Precipitação

Fonte: Os Autores

Dados climáticos são de grande importância para a gestão de agricultura, pois a partir deles o agricultor pode tomar decisões mais assertivas na sua cultura: por exemplo, sabendo qual época ou qual local é mais viável para o cultivo de determinada plantação, se em determinado período de tempo é necessário que ele faça a irrigação em maior ou menor quantidade, de acordo com a precipitação de chuva, a aplicação ou não de defensivos agrícolas, dentre outras decisões.

CONSIDERAÇÕES FINAIS

A IoT vem sendo utilizada em muitos projetos devido ao aumento no número de dispositivos e periféricos conectados à internet, com isso grande volume de dados e, portanto, podem ser transformados em informações significativas, assim, surge a necessidade de preparar esses dados e armazená-los de forma propícia para futuras análises.

Com o advento do Big Data, atualmente há solução tecnológica viável para tal volume de dados, para qualquer área de aplicação.

Após a obtenção dos dados necessários e de seu tratamento, foi possível realizar o processo de EDA para, assim, preparar os dados e armazená-los para análises futuras,

levando em consideração os meios estão sendo utilizados para captar informações e com isso gerar um futuro modelo para futuras análises.

Entende-se que a plataforma Hadoop propicia um ambiente completo para a utilização da linguagem Python e suas bibliotecas de EDA, além de promover a integração perfeita com o ambiente do Hive, de forma muito prática.

A biblioteca Pandas se mostrou muito eficiente, em termos de consumo de tempo e espaço, e disponibilizou excelentes recursos para a captura e tratamento dos dados, também permitindo a exportação desses dados para importação no Hive.

O Hive, por sua vez, se mostrou muito prático, em particular utilizando-se do ambiente Hue. A utilização de comandos SQL em um ambiente de Big Data, de fato, favorece a praticidade e aproveitamento de conhecimento técnico de profissionais de banco de dados, já familiarizados com a sintaxe dessa linguagem de consulta em bancos de dados e ambientes de Data Warehouse.

Entende-se que os dados obtidos para análise, o dataset de dados climáticos, de fato, são representativos daquilo que os sensores propostos para captura poderiam fazer.

Por fim, compreende-se que os dados preparados pelo EDA podem sim contribuir para a gestão de agricultura: temperatura, umidade e precipitação são, de fato, os dados mais importantes e podem trazer uma grande eficiência e eficácia nas agriculturas, sejam elas quais forem. A eliminação de dados inconsistentes bem como o tratamento das temperaturas extremas, que podem atrapalhar um cálculo de médias, de fato mostram a efetividade desse EDA.

REFERÊNCIAS

ASHTON, Kevin. The "Internet of Things" Thing, 2009. Disponível em: <http://www.itrco.jp/libraries/RFIDjournal-That%20Internet%20of%20Things%20Thing.pdf>. Acesso em: 25 abr. 2018.

KIMBALL, R.; ROSS, M. The Data Warehouse Toolkit: The Definitive Guide to DimensionalModeling. [S.l.]: Wiley, 2013. ISBN 9781118732281.

INMON, W. Building the Data Warehouse. [S.l.]: Wiley, 2005. (Timely, practical, reliable). ISBN 9780764599446.

MARR, Bernard. Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance, Wiley 2015.

O'NEIL, Cathy; SCHUTT, Rachel, Doing Data Science, O'Reilly 2014.

SOUZA, Lucas Andrade Moreira. Aplicação de Aprendizado de Máquina para Predição de Prioridade em Gestão de Incidentes. UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO, 2017.

TAURION, C. Big Data. Brasport.2013

THUSOO, A. et al. Hive: A warehousing solution over a map-reduce framework. Proc. VLDB Endow., VLDB Endowment, v. 2, n. 2, p. 1626–1629, aug 2009. ISSN 2150-8097.

WENDDLING, Marcelo. Sensores. Colégio Técnico Industrial de Guaratinguetá.Guaratinguetá, 2010.