# ANÁLISE DA QUALIDADE DO AR UTILIZANDO FERRAMENTAS DE BIG DATA

Vitor Henrique Mendes (IC) e Leandro Augusto da Silva (Orientador)

**Apoio:PIVIC Mackenzie**

**RESUMO**

A poluição do ar é resultante da emissão de gases tóxicos, os quais alteram a estrutura original da atmosfera e segundo a Organização Mundial de Saúde, causa a morte de mais de 7 milhões de pessoas a cada ano. Nesse contexto, essa pesquisa teve como objetivo apresentar um estudo sobre o monóxido de carbono, um dos principais poluentes urbanos, através de um ambiente computacional capaz de coletar em tempo real, armazenar e processar grande quantidade de dados utilizando sensores de Internet das Coisas, ferramentas de *Big Data open source* como *Apache Hadoop* e *Apache Spark* além da entrega de valor dos dados através *dashboards* de visualização de dados.

**ABSTRACT**

Air pollution is caused by the emission of toxic gases, which change original atmosphere structure and according Word Health Organization, causes more than 7 million people die by year. In this context, this research aims to do a study on carbon monoxide, one of the main urban pollutants, using a computational environment capable of collecting real time, storing and processing large amounts of data using Internet of Things sensors, open source Big Data tools like Apache Hadoop and Apache Spark, in addition to the delivery of data value through data visualization panels.

**Keywords:** *Big Data*, *Apache Spark*, Visualização de dados

## INTRODUÇÃO

A poluição do ar é caracterizada pela presença de quaisquer substâncias que, dependendo da sua concentração, possa se tornar prejudicial à saúde e ao meio ambiente.

Dentre os vários poluentes do ar, o Monóxido de Carbono (CO) se destaca negativamente no âmbito urbano, principalmente por ser um dos principais agentes causadores de doenças respiratórias, chegando a levar a óbito aproximadamente 7 milhões de pessoas a cada ano (UNITED NATIONS, 2015). O CO é um gás incolor e sem cheiro, resultante da queima incompleta de combustíveis de origem orgânica, tais como combustíveis fósseis e biomassa. É encontrado em altas concentrações nas grandes cidades, tendo como principal emissor os veículos automotores (CETESB, 2016).

Neste cenário surge a Internet das Coisas ou IoT (*Internet of Things*), a qual pode ser compreendida como uma área de pesquisa utilizada para descrever o conceito de uma rede composta de diferentes dispositivos inteligentes interligados, coletando e enviando dados para rede e comunicando-se entre si através de um mesmo padrão (OLIVEIRA; SILVA, 2018). Sensores de monitoramento tem sido amplamente utilizados para realizar coleta de dados em tempo real possibilitando uma maior rapidez na tomada de decisão (BORGES, 2017).

Atualmente muitas empresas investem em dispositivos para IoT, uma vez que a análise e cruzamento de dados podem ajudar a melhorar de forma relevante não somente os negócios, mas também a vida humana. Estima-se que até 2020 haverá aproximadamente 50 bilhões de dispositivos conectados ao redor do mundo (LIMA, 2017). A rápida produção de dados advindos de sensores, em diferentes formatos, em grande variedade e volume caracterizam a necessidade de um paradigma diferenciado para o armazenamento e processamento dos dados, constituindo dessa forma o termo *Big Data* (MADDEN, 2012).

Atualmente vivemos na era dos dados. Não é trivial estimar o volume total de dados armazenados no universo eletrônico, mas uma pesquisa realizada revelou que em 2013 haviam aproximadamente 4.4 *zettabytes* de dados armazenados na rede, e que em 2020 haveriam 44 zettabytes, o equivalente à um milhão de terabytes (WHITE, 2015). O desenvolvimento de uma aplicação em Internet das Coisas necessita de uma arquitetura de *Big Data* para a o armazenamento e análise (*Big Data Analytics*), do entendimento da produção de sensores e a ligação dos mesmo em uma rede única (BORGES, 2017).

Uma das soluções mais conhecidas de *Big Data* é o *Apache Hadoop*, um framework construído pela *Apache* o qual permite processamento distribuído em grandes conjuntos de dados através clusters que utilizam hardware de baixo custo (APACHE, 2015). Um dos primeiros componentes do projeto *Hadoop* foi o *Hadoop Distributed File System* (HDFS), um sistema de arquivos distribuídos altamente tolerante a falhas, capaz de armazenar enorme quantidade de dados. Para processar esses dados era necessário utilizar um modelo de

programação distribuída, nesse propósito surgiu o MapReduce, o qual permitia realizar transformações em lote de dados armazenados pelo HDFS usando computação paralela (WHITE, 2015). Conforme o *Hadoop* foi sendo utilizado em larga escala, algumas demandas de performance e negócios foram surgindo, as quais o *MapReduce* não era capaz de tratar, como por exemplo análise de dados em tempo real.

A evolução tecnológica para suprir tais demandas é apresentada pelo *Apache Spark*, um motor unificado proposto para oferecer suporte à processamento de dados distribuídos, os quais podem ser persistidos na memória *RAM*, tornando-o de 10 à 100 vezes mais rápido do que o *MapReduce*, além de oferecer bibliotecas de alto nível com suporte a streaming de dados, *machine learning*, processamento de grafos distribuídos e suporte a sintaxe SQL (*Structured Query Language*) (DATABRICKS,2016).

O objetivo deste trabalho consiste na utilização de uma arquitetura que possibilita coletar dados de CO e armazenar em um ambiente *Big Data*, tudo isso através de sensores IoT e ferramentas do ecossistema *Hadoop*, como o HDFS e o *Apache Spark*. De forma mais específica, propõe-se aqui uma análise da qualidade do ar na região metropolitana de São Paulo e na cidade de Mogi das Cruzes, situada a 45 km da capital, através de uma arquitetura flexível e escalável. Os resultados serão amplamente apresentados com uso de técnicas de visualização de dados.

# 1 REFERENCIAL TEÓRICO

## 1.1 Padrões de Qualidade do Ar

O avanço dos conhecimentos técnicos e científicos, levaram a União Europeia e os Estados Unidos a revisar os valores dos padrões de qualidade adotados como referência. Em 2005, a Organização Mundial de Saúde (OMS) lançou um documento com a revisão de todos os valores de referência para cada poluente atmosférico tendo como objetivo a proteção de saúde da população (CETESB, 2016). Baseado nas diretrizes da OMS, em 2008 São Paulo iniciou um processo de revisão dos padrões de qualidade do ar, que culminou no Decreto Estadual nº 59113 de 23/04/2013. Este decreto estabelece que o controle da qualidade do ar no Estado de São Paulo deverá ser mediado através dos seguintes critérios (CETESB, 2016):

Metas Intermediárias – (MI) estabelecidas como valores temporários a serem cumpridos em etapas, visando à melhoria gradativa da qualidade do ar no Estado de São Paulo, baseada na busca pela redução das emissões de fontes fixas e móveis, em linha com os princípios do desenvolvimento sustentável.

Padrões Finais (PF) – Padrões determinados pelo melhor conhecimento científico para que a saúde da população seja preservada ao máximo em relação aos danos causados pela poluição atmosférica.

Os padrões de qualidade do ar para o CO, estabelecidos no DE nº 59113/2013 estão apresentados na tabela 1. O valor vigente está sublinhado.

| Poluente | Tempo de Amostragem | M1 (ug/m³) | M2 (ug/m³) | M3 (ug/m³) | PF |
|---|---|---|---|---|---|
| Monóxido de Carbono (CO) | 8 horas | - | - | - | 9 ppm |

**Tabela 1. Padrões Estaduais de Qualidade do CO (CETESB, 2016)**

Os padrões finais são instituídos sem etapas intermediárias quando não for estabelecida nenhuma meta intermediária, como foi o caso do monóxido de carbono. Para o resto dos poluentes, os padrões finais passam a ser considerados a partir do prazo de duração do MI3. O índice de qualidade do ar (IQA), é uma ferramenta matemática construída para simplificar o processo de divulgação da qualidade do ar. É calculado um índice para cada poluente medido, que é um valor adimensional. Dependendo do índice obtido, o ar recebe uma qualificação relacionada a qualidade (CETESB, 2016).

A tabela 2 a seguir mostra o índice de qualidade do ar para cada poluente medido.

| Qualidade | Índice | MP10 24 horas | MP2,5 24 horas | O3 8 horas | CO 8 horas | NO2 1 horas | SO2 24 horas |
|---|---|---|---|---|---|---|---|
| N1– Boa | 0–40 | 0–50 | 0–25 | 0–100 | 0–9 | 0–200 | 0–20 |
| N2- Moderada | 41–80 | >50-100 | >25-50 | >100-130 | >9-11 | >200-240 | >20-40 |
| N3-Ruim | 81-120 | >100-150 | >50-75 | >130-160 | >11-13 | >240-320 | >40-365 |
| N4 – Muito Ruim | 121-200 | >150-250 | >75-125 | >160-200 | >13-15 | >320-1130 | >365-800 |
| N5 – Péssima | >200 | >250 | >125 | >200 | >15 | >1130 | >800 |

**Tabela 2. Estrutura do Índice de Qualidade do Ar (CETESB, 2016)**

Para efeitos de divulgação, utiliza-se o índice mais elevado, ou seja, mesmo que a qualidade do ar de uma estação seja avaliada por todos os poluentes monitorados, a classificação é determinada pelo pior caso, isto é, o maior índice (CETESB, 2016).

### 2.2 Internet das Coisas

A Internet das Coisas (IoT) segundo Atzori (2010, apud OLIVEIRA; SILVA, 2018) é um paradigma que abrange uma realidade onde objetos físicos embarcados com dispositivos interconectados através de uma rede sem fio trocam informações por meio da internet, moldando uma rede de objetos inteligentes capazes de realizar diversos processamentos, capturar variáveis ambientais e reagir a estímulos externos. De acordo com (TRAN; MARGINEAN; KARZEL, 2016), o objetivo da Internet das Coisas é interligar redes de objetos, a fim de realizar a troca de dados/informações para o cumprimento de tarefas para seus

usuários. A chave para possibilitar essa comunicação entre todas coisas é a adoção de um padrão, porém sua construção e implantação em escala global não é algo tão trivial. Para poder conectar bilhões de dispositivos inteligentes à Internet, precisa -se de uma arquitetura flexível. O modelo mais básico para IoT é constituído de três camadas (SANTOS, 2017):

Camada de Percepção, representação dos objetos físicos, os quais utilizam dispositivos como sensores por exemplo, para a coleta de dados, Camada de Rede, abstração das tecnologias, padrões de comunicação e serviços de gerenciamento, Camada de Aplicação, responsável por prover serviços aos usuários/clientes.

Essa arquitetura abrange e envolve a Internet e as redes de comunicação, necessitando de identificação única de objetos, sensores e capacidade de conexão, como suporte para o desenvolvimento independente de serviços e aplicações (BORGES, 2017).

### *2.3 Big Data*

O conceito de *Big Data* está relacionado a informações que não podem ser processadas e analisadas utilizando ferramentas ou processos tradicionais. Cada vez mais, empresas e organizações vem enfrentando desafios relacionado a esse tema. Eles possuem acesso a informação, mas não sabem como extrair valor, uma vez que os dados estão em um formato semiestruturado ou não estruturado (ZIKOPOULOS et al, 2013).

Conhecidas como propriedades do *Big Data*, os 5 Vs descritos a seguir, envolvem completamente o ecossistema do modelo denominado *Big Data Ecosystem*, o qual por sua vez abrange *Big Data Infrastructure, Big Data Analytics, Big Data Management* e *Big Data Security* (DEMCHENKO; LAAT; MEMBREY, 2014, apud BORGES et al, 2017).

Variedade: além dos dados estruturados, dados semiestruturados e não estruturados, são geradas muitas informações por diversos dispositivos tais como sensores IoT*, smart phones* ou qualquer outro objeto inteligente que possui acesso a rede através da internet. Volume, estima-se que o volume de dados disponível em nosso universo digital é de aproximadamente 2,7 *zetabytes.* Velocidade, os fluxos de dados são gerados em uma velocidade e quantidade muito grande de forma contínua. Veracidade, os dados podem assumir diversas formas e possuir várias interpretações. E finalmente Valor, ou seja possibilidade de extração de respostas aplicáveis ao negócio.

### **2.4 Ecossistema *Apache Hadoop***

Uma das soluções de *Big Data* mais conhecidas no mercado é o *Apache Hadoop*, criado por *Doug Cutting* em 2005, inspirado no *Google File System* (GFS) e no paradigma de programação paralela *MapReduce* (ZIKOPOULOS et al, 2013). Diferentemente do seu início, hoje o projeto *Hadoop* se tornou uma plataforma de *Big Data* a qual possui diversos produtos

interligáveis para as mais diversas tarefas de *data analytics, data ingestion, data mining* e *data science.* O diferencial dessa plataforma se da na capacidade de tratar volumes massivos de dados estruturados, semi-estruturados e não estruturados de forma escalável, confiável, flexiva e com baixo custo através de *clusters* formados por *hardwares* comuns (GODOI, 2016).

Escalabilidade: Há um tratamento de dados local para cada nó em um *cluster*, além de permitir armazenar, gerenciar, processar e analisar dados em uma escada de petabytes.

Confiabilidade: *Clusters* que utilizam *hardware* de grande porte são propensos a falhas de nós individuais. Uma das principais características do *Hadoop* é sua alta tolerância a falhas, uma vez que uma máquina é derrubada, o processamento e gerenciamento dos dados são passados para as restantes.

Flexibilidade: Ao contrário dos bancos de dados tradicionais, muitas das ferramentas do *Hadoop* não requerem esquemas estruturados criados antes do armazenamento dos dados. Pode-se armazena-los em qualquer formato ou estrutura.

Baixo Custo: *Hadoop* é uma plataforma *open source* e executado em máquinas de *hardware* de baixo custo, trazendo assim um conceito de escalabilidade horizontal.

A tabela 3 mostra algumas das principais ferramentas do ecossistema *Hadoop*.

| **Camada** | **Ferramenta** | **Descrição** |
|---|---|---|
| Análise | *Spark* | Ferramenta para processamento distribuído de grandes bases de dados, substituindo o MapReduce, chegando até a ser 100x mais rápido. |
| Análise | *Hive* | Ferramenta para análise de grandes bases de dados utilizando sintaxe SQL e jobs MapReduce. |
| Armazenamento | *HBase* | Banco de Dados não relacional, projetado para trabalhar com Big Data. |
| Armazenamento | *HDFS* | Sistema de arquivos distribuído tolerante a falhas e capaz de armazenar dados na ordem de petabytes. |
| Streaming | *Kafka* | Gerenciador de fluxo de dados em tempo real |

**Tabela 3. Principais Ferramentas do Ecossistema *Hadoop* (GODOI, 2016).**

### 2.5 *Hadoop Distributed File System*

Segundo Coulouris  e colaboradores (2013, apud BORGES, 2017), um sistema de arquivos distribuído permite as aplicações acessarem e armazenarem informações remotas como se estivessem em um sistema de arquivos local, possibilitando aos usuários acesso a partir de qualquer máquina em rede. Neste contexto, *Big Data*, o qual envolve um enorme volume de dados, a utilização de um mecanismo para o armazenamento dos mesmos através de várias máquinas se torna muito necessário. Nesse sentido, uma das principais camadas do *Hadoop* é o *Hadoop Distributed File Sysem* (HDFS), um poderoso sistema de arquivos distribuído (BORGES, 2017).

O HDFS foi projetado para oferecer uma armazenagem redundante para *Big Data*, mantendo os dados distribuído em máquinas de baixo custo ligadas através de um *cluster.* Ao serem armazenados, os dados são divididos em blocos de 128MB cada, replicados e distribuídos nas máquinas (nó) do *cluster,* possibilitando assim a recuperação de arquivos caso algum nó se torne inacessível (BENGFORT; KIM, 2016).

### 2.6 *Apache Spark*

Uma das primeiras maneiras de se processar e manipular dados armazenados no HDFS*,* era através do *MapReduce*, um algoritmo de programação distribuído orientado a grandes volumes de dados, no qual o usuário especifica uma função *mapper* por meio de um conjunto de pares {chave,valor}, gerando uma saída intermediária, e também define uma função *reduce*, a qual une os resultados da saída intermediária associada a uma mesma chave (DEAN e GHEMAWAT, 2008, apud BORGES, 2017).  No entanto, à media que o *Hadoop* passou a ser utilizado em larga escala, outras demandas importantes foram surgindo, a qual o *MapReduce* já não era mais suficiente. Isso porque para ser altamente tolerante a falhas e conseguir realizar tarefas paralelas era necessário um modelo de extração de dados baseado em frequente escrita intermediária no HDFS, fazendo com que as aplicações perdessem muita performance uma vez que leitura e escrita em disco são fatores computacionalmente lentos (BENGFORT; KIM, 2016).

Nesse contexto, o *Apache Spark* foi desenvolvido para oferecer uma API (*Application Programming Language*) para programação distribuída, semelhante ao *MapReduce*, porém projetada e otimizada para consultas interativas e algoritmos iterativos através dos RDD'S (*Resilient Distributed Datasets*). Uma das principais características do *Spark* é a possibilidade de armazenar os grandes volumes de dados *in-memory*, ou seja, na memória RAM, diminuindo assim consideravelmente o tempo de processamento das aplicações (KARAU et al., 2015).

Outra característica que destaca essa ferramenta, é a possibilidade de utilizar diferentes modelos de programação para processar os dados, como por exemplo o uso de sintaxe SQL, programação funcional ou até mesmo o próprio *MapReduce.* O *Spark* é uma ferramenta altamente flexível, a qual permite o acesso a maioria dos banco de dados ou sistemas de arquivos existentes atualmente, além disso é possível acessar suas funções através de diversas linguagens de programação, como Java, Scala, *Python* e R. Além disso, o foco do *Spark* é totalmente processamento e não a armazenagem de dados, isso o faz uma ferramenta focada em soluções de processamento distribuído (BENGFORT; KIM, 2016).

A tabela 4 a seguir mostra os principais módulos de processamento disponíveis no *Spark*.

| **Módulo** | **Definição** |
|---|---|
| *Spark SQL* | Responsável por converter os dados em coleções distribuídas organizadas em colunas, semelhantes a um banco de dados tradicional. |
| *Spark Streaming* | Permite o processamento e a manipulação de *streams* de informações em tempo real. |
| *Spark MLlib* | Permite a criação de modelos de *Machine Learning* distribuído. |
| *Spark GraphX* | Coleção de algoritmos e ferramentas para criar, manipular e executar processamento paralelo em grafos. |

**Tabela 4.** Módulos **do *Apache Spark* (BENGFORT; KIM, 2016).**

## 3. METODOLOGIA

A metodologia completa deste trabalho consiste em um processo com as seguintes etapas: coleta de dados, armazenamento dos dados, processamento paralelo e distribuído de análise de dados e, por fim, aplicação de métodos de visualização.

### 3.1 Coleta

Para a coleta, utilizou-se uma placa analógica do tipo ESP8266 com cobertura de 91 metros, comunicação TCP e UDP e tensão de 3.3 Vdc, a qual tem dedicada na mesma um sensor de CO e de wireless, possibilitando conexão ¬com a rede. No total foram utilizadas duas placas, sendo que uma foi colocada na área metropolitana de São Paulo e o outra na cidade de Mogi das Cruzes, interior de São Paulo. Ambos os sensores de cada placa capturam a concentração de CO no ar em um intervalo de 10 minutos, 24 horas por dia e os envia através da internet para uma API responsável por receber os dados e repassá-los a um servidor de banco de dados NoSQL (Not Only SQL).

### 3.2 Armazenamento

Os dados são armazenados em uma sequência de timestamp (concatenação de informações do dia e horário) e concentração de monóxido de carbono (CO). As

concentrações de CO relacionadas a cada cidade são persistidas em formato JSON (*Javascript Object Notation*) e replicadas no HDFS através do rhdfs, um pacote capaz de realizar integração entre o R e o *Hadoop Distributed File System* (REVOLUTION ANALYTICS, 2014) evitando assim eventuais perdas de informação no processo, por exemplo caso um arquivo seja corrompido e se torne inacessível.

### 3.3 Processamento

Para este projeto, utilizou-se um *notebook* com 16 GB de memória *RAM* e 4 núcleos de processamento. Em uma aplicação *Spark* há 2 processos principais distintos, sendo eles o driver e o executor respectivamente. O driver é criado assim que o usuário inicia uma aplicação *Spark*, ele é responsável por solicitar recursos ao gerenciador do *cluster*, além disso, tem a função de realizar a tradução do código escrito pelo usuário para o *Spark*. Já o executor por sua vez, tem o papel de realizar todas as operações requisitadas pelo processo driver (KARAU et al., 2015).

A tabela 5 mostra as configurações do *Spark* utilizadas neste projeto. Para que as tarefas pudessem ser paralelizadas, utilizou-se 3 núcleos de processamento para o executor, restando assim apenas 1 núcleo para o driver. Para a memória *RAM*, optou-se por dividir a quantidade total entre os 2 processos.

| - | Memória RAM | Núcleos |
|---|---|---|
| *Executor* | 8 GB | 3 |
| *Driver* | 8 GB | 1 |

**Tabela 5. Configuração *Apache Spark***

O processamento de dados é iniciado com a leitura dos mesmos dados no HDFS através do *Spark* e transformados em um *SchemaRDD* (Dataset Resiliente Distribuído), o qual são particionados, distribuídos e persistidos na memória RAM em formato tabular, com cada coluna contendo seu *schema*. A linguagem de programação utilizada para esta manipulação de dados e aplicação de métodos estatísticos foi o R, juntamente com a interface *RStudio* (RSTUDIO, 2017), ambos integrados com o *Spark* através do pacote *SparkR* (APACHE SPARK, 2017), o qual possibilita o uso de computação distribuída.

Para fins de análises, de forma que possam ser interpretadas e comparadas com outros sensores que mensuram o mesmo tipo de poluente, foi necessário a conversão dos dados de concentração do CO coletados pelos sensores usados neste trabalho para a escala ppm (partícula por milhão). Essa escala é também utilizada pela Companhia Ambiental do Estado de São Paulo (CETESB). Com a adoção deste padrão, pode-se neste trabalho fazer comparações e outras definições de análises. Além disso, os horários foram arredondados para que fosse possível obter-se informação de hora em hora. Logo 14 horas e 10 minutos por exemplo, se tornam 14 horas e assim por diante.

Inicialmente os dados foram agrupados por mês, dia e horário e calculou-se a média móvel, o qual baseia-se na média de CO das 8 horas anteriores. Para algumas análises foi necessário sumarizar a média móvel por dia. A maior média móvel registrada no dia é usada como métrica para definir a qualidade do ar do mesmo (CETESB, 2016). Todas as análises realizadas neste trabalho utilizaram o conceito da maior média móvel diária e serão referidas daqui em diante como: concentração de CO, níveis de CO e termos correlatos. Para cada dia atribui-se o índice de qualidade do ar (IQA) utilizado pela CETESB, o qual baseia-se na regra mencionada anteriormente, seguindo a escala mostrada na tabela 6.

| **Bom** | **Moderado** | **Ruim** | **Muito Ruim** | **Péssimo** |
|---|---|---|---|---|
| Entre 0 e 9 ppm | Entre 9 e 11 ppm | Entre 11 e 13 ppm | Entre 13 e 15 ppm | Maior que 15 ppm |

**Tabela 6. Classificação da qualidade dor ar (CETESB, 2016)**

### 3.4 Visualização

A aplicação do *dashboard* foi feita utilizando os pacotes *RShiny* (RSTUDIO, 2014) e *ShinyDashboard* (RSTUDIO, 2014), ambos são *frameworks* os quais tornam fácil construir aplicações web interativas por meio da linguagem R.

A visualização dos dados e dos resultados obtidos foram feitas através dos pacotes *plotly* (PLOTLY, 2016), *highcharter* (KUNST, 2017) e principalmente *ggplot2,* um pacote para produzir gráficos estatísticos ou de dados, utilizando a gramática dos gráficos, a qual cada parte dos gráficos é independente uma da outra, facilitando assim a construção do mesmo (HICKHAM; SIEVERT, 2016).

Utilizou-se para a análise, dados coletados entre junho e setembro do ano de 2018. Além disso, histogramas e *boxplots* foram usados para realizar a comparação entre as distribuições de concentrações de monóxido de carbono entre as duas cidades. Estudos comparando os níveis de CO entre os meses e em seguida entre os dias da semana foram realizados através de gráficos de dispersão e barras, obtendo assim um maior entendimento sobre o comportamento do poluente .

A figura 1 ilustra a arquitetura utilizada para este projeto.

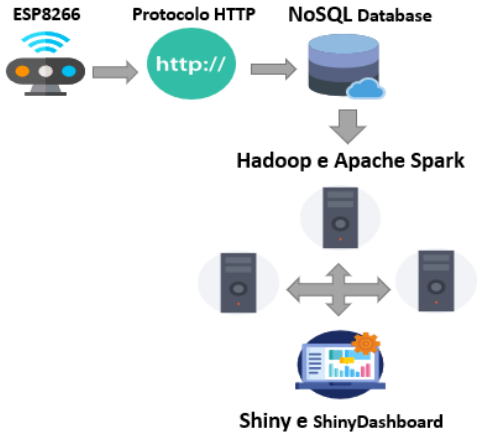


**Figura 1. Arquitetura do Projeto**

## 4 RESULTADO E DISCUSSÃO

Os resultados serão apresentados nesta seção, por cidade, realizando comparações envolvendo as concentrações de CO entre os meses e seus respectivos dias da semana. Adicionalmente ao longo das discussões serão realizados breves comparativos entre ambas cidades.

### 4.1. São Paulo

Observar-se na figura 2 que 73% dos dias analisados na capital, tiveram a qualidade do ar classificada como boa, 23.8% como moderado e 3.3% como ruins.

| qualidade / month | Boa | Moderado | Ruim | Totals |
|---|---|---|---|---|
| Agosto | 77.4% | 16.1% | 6.5% | 100.0% |
| Julho | 61.3% | 38.7% | | 100.0% |
| Junho | 76.7% | 20.0% | 3.3% | 100.0% |
| Setembro | 76.7% | 20.0% | 3.3% | 100.0% |
| Totals | 73.0% | 23.8% | 3.3% | 100.0% |

**Figura 2. Porcentual de qualidade por mês em São Paulo**.

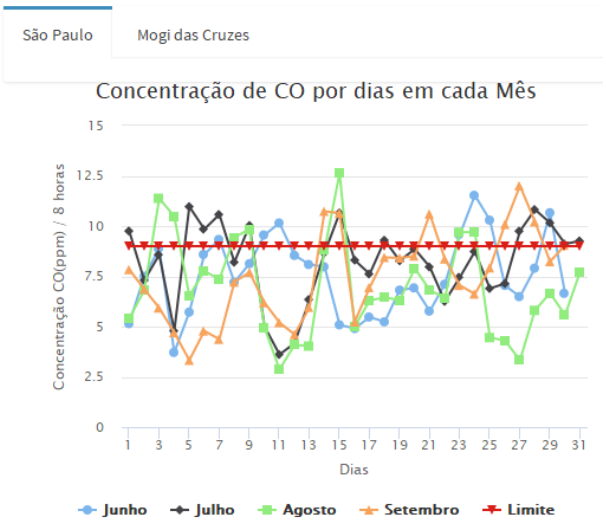


**Figura 3. Concentração de CO em São Paulo**

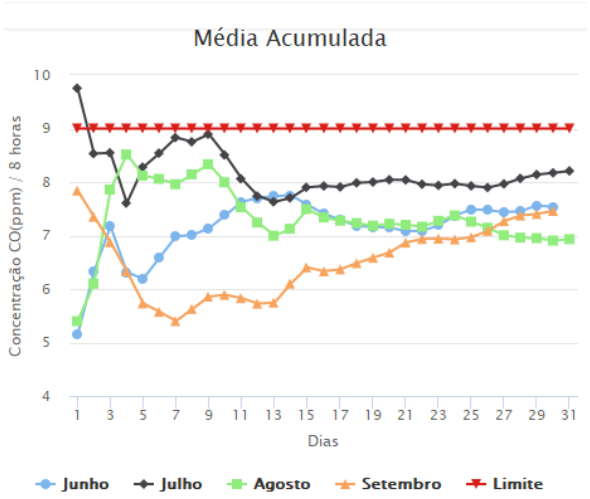


**Figura 4. Média Acumulada da Concentração de CO em São Paulo**

Nota-se de acordo com a figura 3, que julho foi o pior mês em relação a concentração do poluente estudado, foram quase 40% dos dias fora do padrão considerado ideal, ou seja, acima de 9 ppm. Esse fator destaca-se ainda mais quando observamos a figura 4, onde é mostrado a média acumulada da concentração de CO ao longo de cada mês. Em contrapartida, no mês setembro, observou-se que em quase todo momento, a concentração de CO foi mantida a menor dentre todos os meses durante quase 25 dias seguidos, embora tenha terminado a segunda menor.

Ainda de acordo com a figura 4, percebe-se que por mais que cada mês tenha um comportamento particular, é possível notar um certo grau de similaridade entre eles. Por exemplo, julho e setembro, foram caracterizados por um início de mês com altas concentrações de CO, por outro lado, junho e agosto por sua vez, mostraram um comportamento oposto, sendo assim caracterizados por um início de mês com baixas concentrações de CO. Adicionalmente, notou-se que a partir do dia 15 (praticamente metade do mês), as concentrações de CO em todos os meses começaram a se estabilizar, indicando

assim que por mais altas ou baixas que sejam a variação de CO no mês, ela tende a se normalizar.

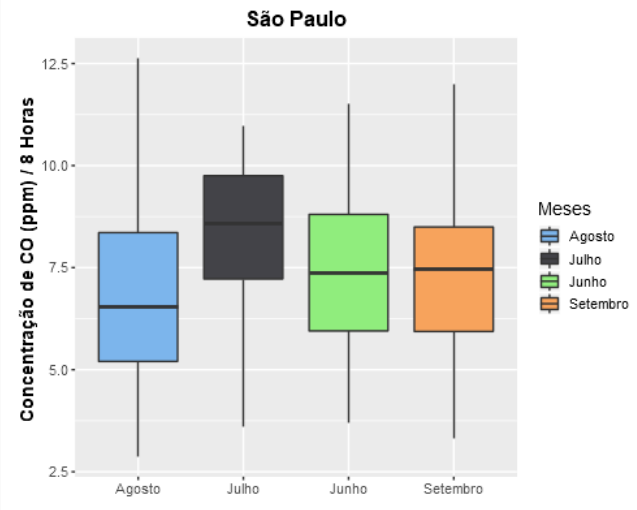


**Figura 5. Distribuição Concentração de CO em São Paulo**

Uma outra maneira de visualizar as concentrações de CO para cada mês é a partir do gráfico *boxplot*, conforme mostra a figura 5. Neste é possível visualizar a distribuição dos dados, o centro e algumas estatísticas como o primeiro, segundo (mediana) e terceiro quartil, bem como os valores discrepantes do conjunto *(outliers)*. A haste inferior se estende desde o menor valor do conjunto até o primeiro quartil (25% dos dados). Similarmente a haste superior se estende desde o terceiro quartil (75% dos dados) até o maior valor do conjunto. A barra horizontal no meio da caixa refere-se a mediana (50% dos dados (PORTAL ACTION, 2015).

Conforme é ilustrado na figura 5, notou-se que mesmo que agosto possua a maior concentração de CO registrada entre todos os meses, é julho quem demonstrou ser o pior mês dentre todos, conforme citado anteriormente, pois 50% de seus dias registraram até 8.207 ppm e 75% até 9.752 ppm, mostrando assim que neste mês as concentrações de CO captadas foram bem acima dos demais. Além disso percebeu-se que a variação do CO nos meses de junho e setembro foram similares.


**Figura 6. Máxima Concentração de CO em São Paulo por dias da semana**

A figura 6 apresenta a concentração de CO para cada dia da semana durante todo o mês. Foi verificado em todos os meses, que alguns dias da semana tais como, quinta, sexta, sábado ultrapassaram o limite de CO pelo menos uma vez, revelando assim a existência de um possível padrão.

### 4.2. Mogi das Cruzes

Em Mogi das Cruzes o cenário é bem diferente, como ilustra a figura 7. Durante os 4 meses estudados, apenas em 3 dias a concentração de CO excedeu ao limite, correspondendo a 2.4 % em relação ao total.

| qualidade / month | Boa | Moderado | Ruim | Totals |
|---|---|---|---|---|
| Agosto | 100.0% | | | 100.0% |
| Julho | 100.0% | | | 100.0% |
| Junho | 100.0% | | | 100.0% |
| Setembro | 90.0% | 3.3% | 6.7% | 100.0% |
| Totals | 97.5% | 0.8% | 1.6% | 100.0% |

**Figura 7. Porcentual de qualidade por mês em Mogi das Cruzes**

Um dos fatores que colaboram para o resultado é o fato da frota de veículos em Mogi das Cruzes ser muito menor se comparado a capital, são 238.919 ao total, o equivalente a 0.026% da frota de São Paulo (DENATRAN, 2018). Além disso, notou-se que em setembro, houve algum evento especial o qual impactou diretamente na concentração de CO, quebrando assim os padrões até então observados nesta cidade.

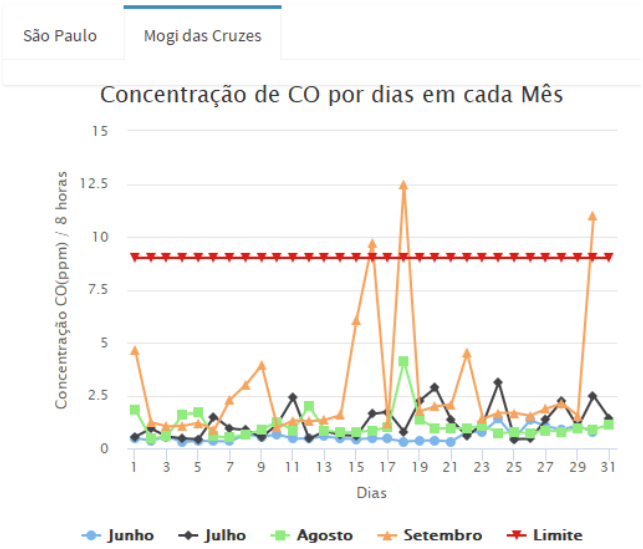


**Figura 8. Concentração de CO em Mogi das Cruzes**

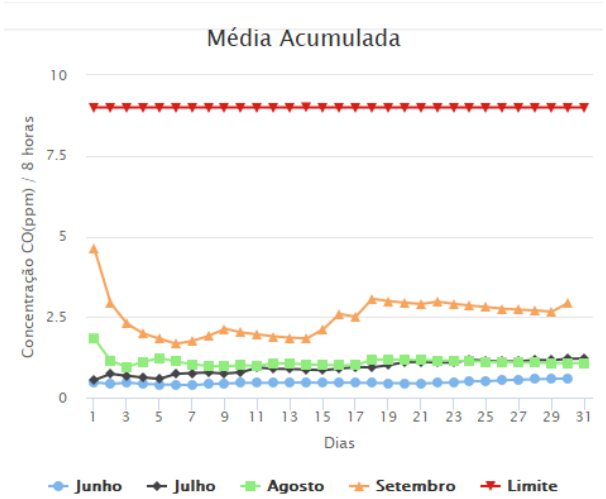


**Figura 9. Média Acumulada da Concentração de CO em Mogi das Cruzes**

Diferentemente de São Paulo onde a concentração de CO se estabiliza em média perto do final do mês, em Mogi das Cruzes esse fenômeno ocorre desde os primeiros dias, indicando que raramente há aumentos ou diminuições bruscas no valor do CO, com exceção de setembro, como observa-se na figura 8 e 9.

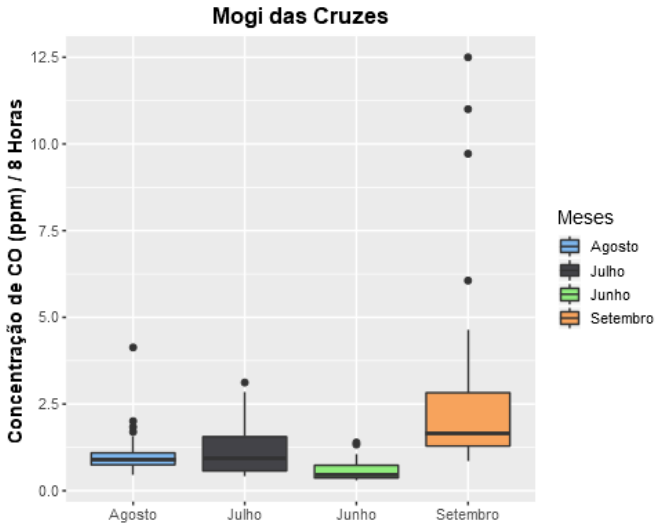


**Figura 10. Distribuição Concentração de CO em Mogi das Cruzes**

Em geral a concentração de CO ao longo dos meses em Mogi se manteve bem abaixo do limite, mesmo em setembro, onde três dias foram classificados com qualidade não ideal, a distribuição de CO se manteve em sua maioria com valor baixo, conforme nota-se na figura 10. Percebe-se também que os meses de julho e setembro possuem uma distribuição mais alongada se comparado a agosto e junho, indicando que nesses meses houveram maior variabilidade de valores da concentração de CO. Além disso, foi verificado em todos os meses, a presença de *outliers,* observações com valores muito distantes de seus respectivos conjuntos, no gráfico *boxplot*, eles são representados em um formato circular, acima da haste superior ou inferior.


**Figura 11. Máxima Concentração de CO em Mogi das Cruzes por dias da semana**

Apesar de quase nenhum dia ter obtido altas concentrações de CO em Mogi a ponto de exceder o limite, observou-se que os dias da semana em que os valores mais se acentuaram

foram quase os mesmos de São Paulo, indicando a existência de um possível padrão em comum nas cidades de médio e grande porte, conforme é ilustrado na figura 11.

## 5 CONSIDERAÇÕES FINAIS

Pôde-se perceber na análise que o atual cenário em relação ao poluente monóxido de carbono é bem otimista em ambas cidades. Mesmo em uma metrópole como São Paulo, não houveram indícios de dias classificados como “muito ruim” ou até mesmo “péssimo”. Ainda há muito a ser estudado em relação ao CO, já que o mesmo muda dependendo de elementos climáticos como a umidade por exemplo. Conforme citado acima, uma possível extensão deste trabalho seria a coleta de diferentes fontes de dados como poluentes ou indicadores climáticos, tais como: temperatura, umidade do ar etc.., para um melhor entendimento sobre comportamento do poluente estudado. Tudo isso viabilizado pela arquitetura de *Big Data*, uma vez que as ferramentas utilizadas neste projeto também suportam coleta de dados de diferentes fontes em tempo real. Espera-se que essa pesquisa possa apresentar diversas contribuições, algumas das quais podemos destacar: alertar sobre os riscos à saúde em relação ao CO e incentivar o uso e estudo de ferramentas de *Big Data* e IoT, as quais vem cada vez mais sendo requisitadas não só no mercado de trabalho.

## REFERÊNCIAS

APACHE SPARK. SparkR (R on Spark). 2017. Disponível em: <https://spark.apache.org/docs/2.3.2/sparkr.html>. Acesso em: 12 jul. 2018.

APACHE. Apache Hadoop: The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.. 2015. Disponível em: <http://hadoop.apache.org/>. Acesso em: 26 maio 2018.

BENGFORT, B.; KIM, J. Processamento em memória com Spark. Analítica de dados com Hadoop. São Paulo: Novatec, 2016. p. 99-105.

BENGFORT, B.; KIM, J. Um sistema operacional para big data: HDFS. Analítica de dados com Hadoop. São Paulo: Novatec, 2016. p. 41-42.

BORGES, M. A et al. An Architecture for the Internet of Things and the Use of Big Data Techniques in the Analysis of Carbon Monoxide. In: INTERNATIONAL CONFERENCE ON INFORMATION REUSE AND INTEGRATION, 200., 2017, San Diego, Ca, USA: IEEE, 2017. p. 184 - 191.

BORGES, M. A. et al. Uma Arquitetura para Internet das Coisas para Análise da Concentração de Monóxido de Carbono na Grande São Paulo por meio de Técnicas de Big Data. 2017. 105 f. Dissertação (Mestrado) - Curso de Mestre em Engenharia Elétrica e Computação, Universidade Presbiteriana Mackenzie, São Paulo, 2017.

CANARY, V. P. A Tomada de Decisão no Contexto do Big Data: Estudo de Caso Único. 2013. 74 f. TCC (Graduação) - Curso de Administração, Escola de Administração, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

CETESB. Padrões de Qualidade do Ar. 2016. Disponível em: <https://cetesb.sp.gov.br/ar/padroes-de-qualidade-do-ar/>. Acesso em: 18 abr. 2018.

CETESB. Poluentes. 2016. Disponível em: <https://cetesb.sp.gov.br/ar/poluentes/>. Acesso em: 24 abr. 2018.

CETESB. Relatório aponta melhoria da qualidade do ar ao longo dos anos na RMSP. 2018. Disponível em: <https://cetesb.sp.gov.br/blog/2018/05/11/relatorio-aponta-melhoria-da-qualidade-do-ar-ao-longo-dos-anos-na-rmsp/>. Acesso em: 21 ago. 2018.

CETESB. Relatório de qualidade do ar no estado de São Paulo 2008. São Paulo: Littera Conteúdos Editoriais, 2008. 340 p.120. Disponível em: <https://cetesb.sp.gov.br/ar/publicacoes-relatorios/>. Acesso em: 29 jun. 2018.

DATABRICKS. Apache Spark is a lightning-fast unified analytics engine for big data and machine learning. It was originally developed at UC Berkeley in 2009. 2016. Disponível em: <https://databricks.com/spark/about>. Acesso em: 23 jun. 2018.

DETRAN. Frota de Veículos - 2018: Frota Nacional (Setembro 2018). 2018. Documento 01 - Frota por UF e Tipo de Veículo. Disponível em: <http://www.denatran.gov.br/estatistica/635-frota-2018>. Acesso em: 05 set. 2018.

DETRAN. Frota de Veículos em SP - por tipo de veículo. Disponível em: <https://www.detran.sp.gov.br/wps/wcm/connect/portaldetran/detran/detran/estatisticastransito/sa-frotaveiculos/d28760f7-8f21-429f-b039-0547c8c46ed1>. Acesso em: 20 ago. 2018.

ECYCLE. Monóxido de carbono: o que é, fontes emissoras e como evitar intoxicação: Intoxicação por monóxido de carbono. 2015. Disponível em: <https://www.ecycle.com.br/2350-monoxido-de-carbono>. Acesso em: 08 mar. 2018.

ECYCLE. O que é poluição do ar? Conheça causas e tipos. Disponível em: <https://www.ecycle.com.br/2949-poluicao-do-ar-atmosferica>. Acesso em: 12 mar. 2018.

GODOI, D. Apache Hadoop: Tudo o que você precisa saber. 2016. Disponível em: <https://www.cetax.com.br/apache-hadoop-tudo-o-que-voce-precisa-saber/>. Acesso em: 17 jun. 2018.

HICKHAM, H.; SIEVERT, C. Ggplot2 Elegant graphics for data analysis. Houston, Texas: Springer, 2016. 2 v.

KARAU, H. et al. Introduction to Data Analysis with Spark: What Is Apache Spark?. Learning Spark : Lightning-Fast Data Analysis. Sebastopol: O'Reilly, 2015. p. 01-04.

KARAU, H. et al. Running on a Cluster. Learning Spark : Lightning-Fast Data Analysis. Sebastopol: O'Reilly, 2015. p. 118-119.

KUNST, J. HighCharter. 2017. Disponível em: <jkunst.com/highcharter/>. Acesso em: 24 jul. 2018.

LIMA, J. 5 aplicações inteligentes da Internet das Coisas. 2017. Disponível em: <https://www.impacta.edu.br/blog/5-aplicacoes-inteligentes-da-internet-das-coisas/>. Acesso em: 09 maio 2018.

MADDEN, S. From Database to Big Data. In: INTERNATIONAL CONFERENCE ON INFORMATION REUSE AND INTEGRATION, 200., 2012, Piscataway, Nj, USA: IEEE, 2012. v. 16, p. 4 - 6.

OLIVEIRA, J. L. S; SILVA, R. O. A internet das coisas (IOT) com enfoque na saúde. Tecnologia em Projeção, Taguatinga, v. 8, n. 1, p.77-78, 2017. Disponível em: <http://revista.faculdadeprojecao.edu.br/index.php/Projecao4/article/viewFile/824/726>. Acesso em: 13 abr. 2018.

PLOTLY. Plotly R Open Source Graphing Library. 2016. Disponível em: <https://plot.ly/r/>. Acesso em: 24 jul. 2018.

PORTAL ACTION. Boxplot. 2015. Disponível em: <http://www.portalaction.com.br/estatistica-basica/31-boxplot>. Acesso em: 11 jul. 2018.

REVOLUTION ANALYTICS. User>rhdfs>Home. 2014. Disponível em: <https://github.com/RevolutionAnalytics/RHadoop/wiki/user>rhdfs>Home>. Acesso em: 20 jul. 2018.

RSTUDIO. Interact. Analyze. Communicate. 2014. Disponível em: <https://shiny.rstudio.com/>. Acesso em: 10 maio 2018.

RSTUDIO. RStudio Open source and enterprise-ready professional software for R. 2017. Disponível em: <https://www.rstudio.com/>. Acesso em: 03 mar. 2018.

RSTUDIO. ShinyDashboard makes it easy to use Shiny to create dashboards like these. 2014. Disponível em: < https://rstudio.github.io/shinydashboard/>. Acesso em: 14 maio 2018.

SANTOS, B. P. Internet das Coisas: da Teoria à Prática. 2017. 50 f. Dissertação (Mestrado) - Departamento de Ciência da Computação, Internet das Coisas: da Teoria à Prática, Belo Horizonte, 2017.

TRAN, T.; MARGINEAN, H.; KARZEL, D. Uma arquitetura de referência para a Internet das Coisas - Parte 1. 2016. Disponível em: <https://www.infoq.com/br/articles/internet-of-things-reference-architecture>. Acesso em: 15 jun. 2018.

United Nations. New UN health agency report cites measures to reduce risks from climate pollutants. 2015. Disponível em: <https://news.un.org/en/story/2015/10/513392-new-un-health-agency-report-cites-measures-reduce-risks-climate-pollutants#.Vi4xwn6rQ2y>. Acesso em: 08 abr. 2018.

WHITE, T. MapReduce. Hadoop The Definitive Guide 4th Edition. 4. ed. California: O' Reilly, 2015. Cap. 2, p. 40.

WHITE, T. Meet Hadoop: Data!. In: WHITE, T. Hadoop The Definitive Guide 4th Edition. 4. ed. California: O' Reilly, 2015. p. 23.

WHITE, T. The Hadoop Distributed Filesystem: The Design of HDFS. Hadoop The Definitive Guide 4th Edition. 4. ed. California: O' Reilly, 2015. Cap. 3, p. 64.

ZIKOPOULOS, P. et al. Big Data: From the Business perspective: What is Big Data?Hint: You are part of it every Day. Understanding Big Data. New York: Mc Graw Hill, 2013. p. 03.

ZIKOPOULOS, P. et al. Big Data: From the Technology perspective: All about Hadoop: The Big Data Lingo Chapter. Understanding Big Data. New York: Mc Graw Hill, 2013. p. 54.

**Contatos:** 31517900@mackenzista.com.br , leandroagusto.silva@mackenzie.br