Tackling Air Pollution with Big Data Analytics

*Full Name: Ayush Singh*

*GWID: G29283616*

*Course: CSCI 6441 Big Data and Analytics*

*Semester: 3rd Semester*

*Department: Department of Computer Science*

*The George Washington University, Washington DC, 20037, USA*

*Email:* [*ayushsingh@gwu.edu*](mailto:ayushsingh@gwu.edu)

1. **Introduction**

One of the most important elements that distinguishes our planet earth from the others in our solar system and make it habitable is the atmosphere. The gases that are present in the earth’s atmosphere are vital for our planet as they are the reason why earth can sustain human life on it. But if we look closer, our planet is not the only planet in our solar system that has an atmosphere. Venus, also known as the earth’s twin, has an atmosphere. But it is 90% denser when compared to earth’s and full of toxic gases. The surface temperature on the Venus can rise to 864 degrees, enough to melt lead. “But some 3 Billion years ago, things might have been different for this planet” [1]. When scientists at NASA’s Goddard Institute for Space Studies applied a computing-model used originally for climate predictions on earth and applied it to Venus, the results seemed to indicate that Venus had a shallow ocean and a habitable surface temperature, slightly cooler than earth’s for possibly 2 billion years of the planet’s early history. “But some 700 to 750 million years ago, a mysterious event triggered the release of CO2 from the rocks on the planet. This release of CO2 lead to a non-stop greenhouse effect that drastically changing the climate on the planet.” [2]

What would we gain by talking about another planet’s history? Although it has an atmosphere, it doesn’t mean that it is just like earth. Our objective here is to learn that a planet “that has very similar characteristics to earth and possibly had hospitable temperature and water in liquid form in the past changed because of CO2, which is considered one of the major air pollutants in our atmosphere” [3]. It is also considered to be one of the greenhouse gases that is responsible for global warming. The mysterious event that triggered the release of high amounts of CO2 in the atmosphere of the Venus helped expedite the Greenhouse Effect on that planet, “which did not let the gas get absorbed back in the rocks and is responsible for now scorching hot temperatures on that planet” [2]. “And levels of CO2, along with the other six criteria pollutants in our atmosphere like Carbon Monoxide, Lead, Nitrogen Oxides” [6] have been increasing year after year.

As an answer to this problem, “many research groups are also trying to conduct studies over longer timeframes and larger locations to pinpoint and track the development of major pollutants in the atmosphere which can cause serious and potentially life-threatening diseases in humans like asthma” [4]. These studies are of immense importance as we endeavor to track the developments of these pollutants in a confined area and over a time period to find solutions powered by the technology of today.

1. **Background**

Often a discussion as important as safeguarding the environment is usually followed by the thoughts of melting polar caps or cutting down of evergreen forests. But something that we are forgetting, is so much closer then that. “The World Health Organization has estimated that nearly 92% of the population on this planet does not have access to clean air. Even more shocking in terms of world economy, The World Bank has estimated that in 2013, we lost $225 billion in productivity to air pollution” [6]. Of course, a strong action must be taken against this global crisis if we want to save the planet for future generations. And the hotspots focused in this article is on cities, where majority of the urban population lives and works. The latest and the most vulnerable victims of this major epidemic is children, affecting their health and growth. “The CO2 level is not the only thing that affects the perceived state of wellbeing. The other factors influencing the air quality are temperature and relative humidity” [11]. But for the sake of this term paper we will stick with our decided 6 major air pollutants as outlined by CDC.

As we are going to talk more about this endeavor further into the term paper, this discussion paves way how Big Data and its processing can take a giant leap towards studying and ultimately working towards finding a solution to this problem. We are going to examine some approaches that are related to AQ betterment cause and how they are making a difference. Two of the companies that are already taking its first steps to tackle this problem. Intel and Bosch are currently working on a joint project that focuses on monitoring the air quality in a confined area as small as a neighborhood or as larger as a city. “They are driving the initiative to make our cities cleaner and better for the future generation” [7]. The technology’s name is “Micro Climate Monitoring Systems (MCMS)” [7]. Intel is making it possible for the cities their air quality (AQ) data from all parts in real time. Governing bodies then can actively generate the data analysis and use those insights that can range from being a neighborhood specific to city specific insights and initiate efforts to change the rules and regulations that can result in transforming the air quality (AQ) of that particular region and lead to bigger environmental changes.

There are various solutions that the two companies are joining hands to provide to the cities. These are as follows:

1. *Reduce Pollution and Operational Costs:* “This will be done by releasing abatement systems based on the data that clearly shows their effect on air quality” [7].
2. *Covering 10x-20x more locations:* “Intel powered MCMS devices has the capability of encompassing more geographical area then traditional air quality monitors” [7].
3. *Early Warning Signals:* “Signals can be sent out to construction sites based on the AQ readings. This can lead to improving working conditions and improving efficiency of the construction workers and residents alike” [7].
4. *Intel helps collect emission data:* “This can lead to government officials make critical decisions of redirecting peak hour traffic from a particular path or lift /lower bridges for ferries and such water bound vehicles” [7].
5. *Air Quality Alerts:* “Air quality alerts can be generated over in real time and can be sent to the drivers who may use the provided information to avoid poor AQ areas” [7].
6. *Tailored Notification to Neighborhoods:* “Allows the MCMS technology to provide the AQ updates to a homeowner’s association or to the local municipality as they happen” [7].

There are even more advantages to the Intel-Bosch collaboration project, but all of these use IoT connected devices that can stream real-time unstructured or semi structured data to the servers and that data can be analyzed upon to give real time information to the citizens, so that the citizens can take the course of action as necessary.

We are also going to examine another research paper “where researchers use two of the best big data tools and their programming models to check AQ simulation and guiding the next generations of researchers who continue on this path” [4]. The programing models are of Hadoop MapReduce and Spark Language. The authors of that paper are trying to establish the best programing model possible for the other researchers to take.

The 2 use cases include:

1. *“Calculating the eight-hour rolling average of pollutants in a given restricted area” [4].*
2. *“identifying clusters of sensors showing similar patterns in pollutant concentration over multiple years in the state of Texas” [4].*

The dataset that they are using for this analysis is air pollution data collected over 15 years by over 175 monitors across the state for a variety of pollutants [4]. We will be studying their findings and examining which model works better and will also be looking at the Intel-Bosch collaboration on the MCMS technology and critiquing after we are done examining it.

1. **Body**

In this section of our paper, we will delve deeper into the technologies and the type of data analysis methods that have already been tried and tested by the researchers and corporations alike. Some of these analysis methods are created very recently with the development of the world-wide used resources like the internet and others are just based on basic powerful concepts of data analysis and data’s life cycle in general.

***Intel and Bosch Air Quality Simulators and Analytics as a Service:***

First, we are going to continue our conversation on the product that was created by Intel Corporation and Bosch to monitor the AQ around a particular area and over a period of time. The Micro Climate Monitoring System (MCMS) is being hailed by the company as “a complete package of sensor, software and services offered to the customer. It is also being advertised as the solution that provides the right people with the smart, contextual and relevant information and does not acts as just another source of data and end-to-end win that brings together information gathering, economic advantage, precision telemetry, and user/customer experiences” [9]. The system has 3 most important components given below:

1. Sensors to sense the amount of air pollutants in the atmosphere. The Bosch MCMS device has the ability to sense “Carbon Monoxide, Sulphur Dioxide, Nitrogen Dioxide, ground level Ozone, Particulate Matter” [9] and other environmental sensing capabilities.
2. Air Quality Management Edge Software.
3. Intel Cloud Infrastructure for cloud analytics using AaaS (Analytics as a Service).

Intel and Bosch together also sell their system by saying that applying their systems, MCMS can generate dynamic alerts and notifications from the officials and can lead the environment monitoring for smart cities. “From a wide coverage to actionable intelligence to near real time telemetry” [9], MCMS is designed to better the face of AQ monitoring.

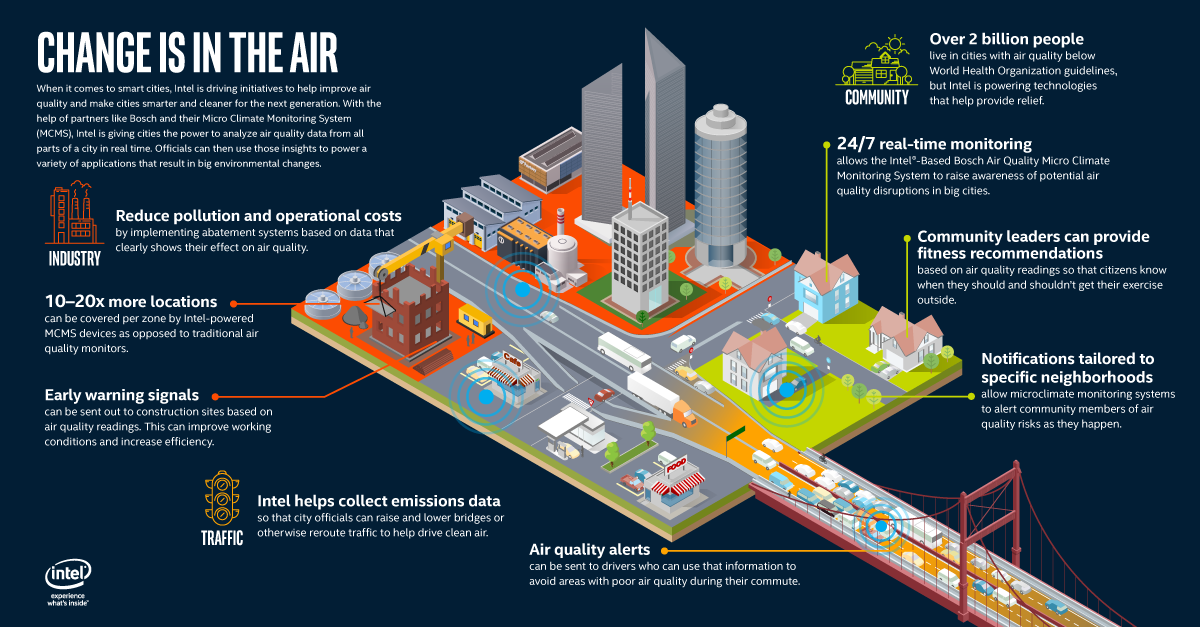


Fig. 1. “Change is In the Air. Intel’s Infographic to promote the MCMS technology, using it in a virtual, restricted space and highlighting it’s use cases” [7].

Digging deeper into their sophisticated setup will reveal that part of it is managed by Intel Cloud Infrastructure using Analytics as a Service platform. Medium to big organizations such as Intel already have multiple private cloud environments and the only thing holding these organizations back from using these environments is the ability to perform analytics on the cloud. Using cloud infrastructure to perform Big Data Analytics makes a lot of sense because it is decreases CAPEX and OPEX of a company. Rather than buying infrastructure in-house, it’s better to use cloud infrastructure and basically outsourcing your analytics needs. It’s also important to use cloud infrastructure because a lot of times, data to be examined can have an internal or external origin. “Analytics as a Service addresses user’s data needs, from data cleaning and preprocessing to analytics to management and data usage” [8]. Intel’s AaaS framework can encompasses a lot of capabilities:

1. “Capturing and extracting structured and unstructured data from trusted sources” [8]. In the case of MCMS, that data source is primarily the sensors that are being used to examine air quality where large streams of unstructured data can pour in. The framework also prioritizes the “most critical data first and is responsible for the time durations as to how long that data must be kept in the systems” [8].
2. “Managing and controlling under comprehensive policy and governance guidelines” [8]. Data management is a very important part of the process as in addition to overlooking the quality of the data may also be responsible for standards and practices for data security. Usually is the case when we have to clean and standardize the data according to a particular industry or a global enterprise or even a domain and Intel’s AaaS can surely deliver.
3. “Choosing the right type of data to be used and right type of filters to be chosen to perform integration, transformation, analysis and visualization” [8]. In other words, overlooking the whole data lifecycle from the inception to the appropriate usage for the Air pollution use case and delivering the right the right information at the right time so the notifications can be pushed to the IoT connected end user devices like mobile phones in real time.

An extended use of Intel’s AaaS can also be seen when deployed on other cloud service types. Calculating the correct combination of services can really benefit any cause. The basic cloud services for AaaS is Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). “These basic cloud service types give Intel to run Software like Rackspace Hadoop, Amazon Elastic MapReduce and Google Big Query Services” [8]. “The MCMS can also be integrated into enterprise or smart city digital platforms like SAP HANA or GE Predix” [8].

As a result, the MCMS puts itself right ahead of the competition from the traditional AQ sensors. The technology is documented as a:

1. “High accuracy gas measurement systems in PPB (Polybrominated Biphenyls) levels, that conform to global (WHO, CPCB, EPA) AQI measurement guidelines” [8].
2. “Intelligent edge to edge management, which allows Over-the-air (OTA) capabilities reducing physical interventions for items like upgrades, patches, configuring units and more” [8].
3. “Easy integration with APIs including AQI as per local body guidelines” [8].
4. “Deep engineering and R&D to create analytics and algorithms to derive valuable AQ telemetric data” [8].
5. “End-to-end system security with encrypted binary code, data communication protocols, software and cloud storage and with API management (Intel Security Solutions, Bosch IoT Suite, etc)” [8].

All of these features along with the cloud infrastructure analytics platform makes the technology usable and responsive to the air pollutants and makes monitoring possible and provides visualizations.

*Critique:*

We heard about how the technology works and we know that the last step and probably the most important one is to push the notifications to end user’s devices. We can easily deduce that the whole system depends upon the condition of every device being connected to the IoT in order to get that information. People might have to sign up for these services and then only might be able to receive these updates. We also need to dig up the critical points about this technology that people should consider before buying. And the most important part of it is to figure out what will happen upon implementation.

Although the cloud analytics is an amazing way for an entity to outsource the task to a third-party company and keep paying for a continued service over an agreed period of time, it can have its disadvantages too.

1. Dependency: The governing committee has to accept the fact that by employing 2 third-party corporations and implementing their technology either indoors or outdoors basically means relinquishing control of the data used from these sensors to those corporations. If the governing committee loses control of the data, it basically means that the standards and practices of this technology is in the hands of those corporations. Data will be acquired, transformed and loaded from the sensors into the AQ edge system by the corporations itself. The committee will not be able to play any active role on where the data is stored, how it is examined and how it is analyzed upon. The governing committee also would not have any control over how the system is managed and changes that are being made to the system. This gives rise to another key factor that whenever the system or the sensors would need maintenance, the two corporations must be involved and asked to send trained personals to trouble shoot the components. It also limits the ‘open-source’ ideology of the people as only people from those 2 corporations get the rights to update the system, potentially holding back other ideas to upgrade the technology. The 2 corporations will get a monopolized control over the AQ monitoring of that particular area which can grow further into antitrust issues for the governing committee.
2. Risk: All cloud computing services are generally accessed from remote servers in the closed locations of the corporations. These servers are the primary source of computation and analytics for this kind of data. If a group of malicious intended attackers decide to attack the corporation, penetrate through their safety firewalls and get access to the data or the algorithms running behind the software, the whole operation can be compromised and the integrity of data could be lost. The hackers could gain the access over the system and stop the service. In an event of a natural calamity, if the servers stop or fail to work, the whole implementation can become useless and can potentially takes days to resolve which can lead to an unmonitored increase in the volume of pollutants ppm in the air. Events like a natural disaster and such can also result in potential data losses if the servers are damaged severely.
3. Migration: If a governing committee decides to shift from a traditional way of monitoring the air quality with this new technology, there must be decisive planning and agreements must be drafted as to how to migrate from a legacy system to the new and updated system. This can potentially result into a large downtime again and the rise in the quantity of the air pollutants in the air can go unchecked.
4. Connectivity: All the cloud services need an active and state of the art connectivity with wide bandwidth to accommodate data and transfer it from the sensors all the way to the processing servers, because we are dealing with big data which can be unstructured / semi structured, the sensors always need to be connected to the internet provided by the service providers. This may include another third-party service provider to come into the picture and disrupt the control ratio of the technology. A disconnected sensor means that no work can be essentially done and the system is sitting idle, doing nothing. Keeping up this kind of service would inevitably require the all of the entities involved to rigorously work and fulfill their service level agreements with the governing committee and avoid the disconnection of the system at any cost. This can again give rise to the security concerns and risks that we mentioned in the above point with another company including the monopoly of the sector and antitrust issues.

***University of Houston’s study of Big Data Models for AQ Simulations:***

Until now in our term paper, we saw how a technology that was developed to monitor the Air Quality at any given particular area over a period of time would fair over with their governing committee and we also provided a critique regarding the sensor technology and the underlying system. As practical and for-the-field this system is, now it’s time to open the academic discussions for Big Data. The study that we are going to consider for this section is the University of Houston’s study and comparison of 2 data models that they ran for their Air Quality Simulations. “They used Hadoop MapReduce for one simulation and Spark Programming model for other simulation. Their aim with this study is to guide future code developers interested in this kind of analyses.

The also generated 2 use cases with regards to their AQ simulators which include:

* 1. “Calculating the eight-hour rolling average of pollutants in a given restricted area” [4].
  2. “Identifying clusters of sensors showing similar patterns in pollutant concentration over multiple years in the state of Texas” [4].

The study can be deemed as a very close relative to the AQ sensor application implemented by Intel and Bosch and this was one of the primary reasons why this study was chosen to be reviewed and critiqued in this term paper.

One can argue that most of the studies that are being conducted regarding the topic of air pollution are being done because of the standards and practices established in consensus with the “Environmental Protection Agency’s (EPA) process of establishing National Ambient Air Quality Standards (NAAQS) for six criteria pollutants implicated in photo chemical smog” [4]. The US Congress also chimed in and created “an array of acts to study Air Quality, establish reasonable guidelines to for regulation and implement policies to protect Air health” [4]. This resulted in a big network of raw and publicly available information about the air pollutants. In recent days although, the effects of medical diseases like asthma from these air pollutants is becoming a major motivation to get our peers in medicine involved in this kind of studies. Recent studies have also shown that there are some signs that diseases that we didn’t even expected to be “related to the effects of air pollution are Alzheimer’s disease, sickle cell anemia and other heart conditions” [4].

We can basically make very minimal changes for the analyses for Air Quality Simulations and more or less obtain the results of a similar type of Big Data analyses done for any other purpose, using it for our Air Quality simulators. “We can compare large datasets in a slightly different manner and use the large available datasets from India and China” [4], because they are the most populous countries and their AQ has been lagging behind several years now and “model the health effects for an individual under different types of exposures from air pollutants” [4]. Needless to say, it’s going to need a lot of parallel and high-performance computing.

Let’s talk a little bit about our models that we are going to use. The first model that we are going to use is MapReduce. “MapReduce was found by researchers at Google in 2004” [10] and it proved to be a very famous and widely used programming paradigm to this date. It is an integral part of Apache Hadoop system version 2 where the other integral systems that work with MapReduce are HDFS and YARN to deliver the Hadoop functionality. Structurally, it has two main components, Mapper and the Reducer, we can also end up ending multiple sections like partitioner and combiner depending upon the size of the dataset that we are using. The key value pairs are passed on to the Mapper and it applies the Map function and gives out intermediate key-values to the distributed processors which calculate their required part. All of the results come back to the Reducer to be phased as a Final Result.

The other programming paradigm that we are going to talk about is Apache Spark. It is basically a distributed cluster framework and facilitates a large range of the distributed computations. “We can also end using the same dataset multiple times across parallel operations” [10]. Spark also does provide users with a rich Application Programming Interface API in Python, Java and Scala. It mainly relies on the concepts of RDDs or Resilient Distributed Datasets and uses them to full extent. It uses the transformation and the actions from RDDs where a transformation returns a pointer to the new RDD and action returns result to the driver program. Together, they can perform complex analysis on datasets.

Moving on to their implementation, the researchers did the following tasks:

1. Calculating the eight hour long rolling average of pollutants in a restricted region, e.g. all sensors located in the city of Houston.
2. Identifying clusters of sensors showing similar pattern in pollutant concentration over the years in the state of Texas.

The researchers conducted 2 versions of these tasks, one was conducted with a Map Reduce with a variable range of producers and the other one with Spark with a variable range of cores used for the calculations. Both of the models ran for a single pollutant (O3) and then for multiple pollutants. “The data was captured by nearly 179 AQ sensors which provided measurement of various pollutants” [4]. Every sensor was programmed to take a reading every five minutes and record it as a data entry into the database. “A data entry consists of a time stamp (year, month, day, hour, minute), location identification (region id, site id, sensor id), and the measurement itself (parameter id, parameter name, value), as well as a flag indicating whether the value provided is considered reliable or unreliable (e.g.

during a calibration cycle)” [4].

The results of a MapReduce version of the code is shown below. The first histogram is for a single pollutant Ozone and the second histogram is for multiple pollutants for a ranging number of reducers for the same code. The following figure calculates the average execution of time of the MapReduce code for both single pollutants and multiple pollutants.



Fig. 2. Average execution time for calculating eight-hour rolling average with HDFS as data source using Hadoop MapReduce [4]

“The Spark model that was calculated for the single pollutant version is presented in the figure below. Results are shown for various numbers of executors and cores per executor, as well as for the default setting of Spark for both parameters. The latter are represented in the figure as not set” [4]. Not set basically means that the number of executors that were executing were not set manually. Executor refers to the process that was launched to run an application on a worker node.

A screenshot of a cell phone

Description automatically generated

Fig. 3. “Average execution time for calculating eight-hour rolling average for single pollutant (O3) concentration with HDFS as data source using Spark” [4].

*Critique:*

We know that MR jobs basically runs a bit faster as we increase the number of reducers from 1 to 5. But no significant efficiency can be detected as the number of reducers are increased from 5 to 10. “It is plausible that more values are aggregated for some keys than others and there cannot be any further distribution among the reduce instances, which results in approximately the same computational time” [4].

It can be seen that a “not set” configuration for the Spark model performs well. But as we see through the chart provided that performance increases as the number of cores and number of executors are increased and specially with 8 cores, we are able to spend only approximately 100 seconds for both single and multiple pollutants.

This makes it crystal clear that given a particular number of cores and a particular number of executors, Apache Spark Model outperforms MapReduce Model for multiple pollutants being examined at once. “To be precise, the Spark model did 21% better in the case of a single ozone pollutant and 25% better for the multiple pollutants than the MapReduce model. It is also found that Spark’s MLlib, which is the machine learning library performs better than the Mahout Library” [4].

Further digging into Spark also suggested that Spark has it’s fair share of problems and the performance benefits were not as significant as originally expected by the researchers.

1. **Conclusion and Future Work**

In this term paper, we saw and critiqued two very different approaches. The first one is the practical approach that was more industry driven.

Top tech companies that drive innovation such as Intel and Bosch are coming up with new sensors that we can deploy around any specific area which can detect the AQI of that area and give us recommendations as habitants of that area to help us monitor and better the AQI of that area. Since sensors generate unstructured data streams that can be analyzed upon to provide those recommendations, Big Data can help us identify many problems that we overlook, as humans, since they are creating or intensifying air pollution and drive us to change how we conduct and live our life in general and reduce and in some cases, prevent air pollution.

And the second one, which is personally interesting to me is the academic approach which delves deeper into a study by the University.

The study that we considered is the University of Houston’s study and comparison of 2 data models that they ran for their Air Quality Simulations. “They used Hadoop MapReduce for one simulation and Spark Programming model for other simulation. Their aim with this study is to guide future code developers interested in this kind of analyses. It gave us an insight on which running model might perform better in some cases but is not applicable for all the cases where an AQ simulator can be of use. Every approach has its limitation and one might outperform the previously acclaimed as “better” approach for another scenario. “Performance of the system like this can also be evaluated by comparing the prediction results of the Air Quality Index (AQI) with those of a traditional SVM algorithm” [12].

For future work, it can be phased as a two-way street. For practical AQ simulations implemented by the industry, the corporations must consider reducing the amount of dependency on other third-party companies that will also reduce the communication needed with other companies in case of an emergency. We can also keep developing our hardware sensors to detect more complex pollutants and become faster, because the technology changes every day and corporations should make it an objective of theirs that their sensors are cheaper to buy but still don’t compromise with quality. The academic field can also see a hike in the productivity of the analysis of the big data streaming in. No big data technology or programming paradigm is perfect, but optimizations can still be made by tweaking the preexisting models or playing with new technologies all together.

Whether studied upon by the subject matter experts and researchers at a university or implemented commercially by our peers in the industry, AQ simulators and technologies alike can be very helpful to detect and track the growth of pollutants in the air so that necessary action can be promptly taken by the governing officials to better the quality of the air around us and beyond.

1. **References**

# [M. J. Way](https://agupubs.onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Way%2C+M+J); [Anthony D. Del Genio](https://agupubs.onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=del+Genio%2C+Anthony+D); [Nancy Y. Kiang](https://agupubs.onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Kiang%2C+Nancy+Y); [Linda E. Sohl](https://agupubs.onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Sohl%2C+Linda+E); [David H. Grinspoon](https://agupubs.onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Grinspoon%2C+David+H); [Igor Aleinov](https://agupubs.onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Aleinov%2C+Igor); [Maxwell Kelley](https://agupubs.onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Kelley%2C+Maxwell); [Thomas Clune](https://agupubs.onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Clune%2C+Thomas), *“Was Venus the first habitable world of our solar system”* , 2016 August 11th, Geophysical Research Letters, Volume 43, Issue 16.

# CNN article: Venus Habitat Scene, [online], Available: <https://www.cnn.com/2019/09/20/world/venus-habitability-scn/index.html>

# National Geographic Article: Air Pollution Explained, [online], Available: <https://www.nationalgeographic.com/environment/global-warming/pollution/>

1. Haripriya Ayyalasomayajula ; Edgar Gabriel ; Peggy Lindner ; Daniel Price,

*“Air Quality Simulations Using Big Data Programming Models”,* 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService).

1. Centers for Disease Controls and Prevention: Air Pollution, [online], Available: <https://www.cdc.gov/air/pollutants.htm>
2. Intel IoT: Why cities need to get smart about pollution, [online], Available: <https://blogs.intel.com/iot/2017/06/14/why-cities-need-to-get-smart-about-pollution/#gs.i60uo2>
3. Intel and Bosch Help Cities Monitor Air Quality, [online], Available: <https://www.intel.com/content/www/us/en/internet-of-things/bosch-smart-city-infographic.html>
4. Intel Article: Big Data Cloud Technologies Brief, [online], Available: <https://www.intel.com/content/dam/www/public/us/en/documents/product-briefs/big-data-cloud-technologies-brief.pdf>
5. Intel Document: Bosch MCMS explained with questions, [online], Available: <https://www.intel.com/content/dam/www/public/us/en/documents/faqs/bosch-mcms-faq.pdf>
6. Dean Jeffrey; Ghemawat Sanjay; *"Mapreduce: Simplified data processing on large clusters",* Proceedings of the 2004 6th Conference on Symposium on Operating Systems Design & Implementation.
7. Judith Molka-Danielsen ; [Per Engelseth](https://ieeexplore.ieee.org/author/37086245883); [Veronika Olešnaníková](https://ieeexplore.ieee.org/author/37085623243); [Peter Šarafín](https://ieeexplore.ieee.org/author/37085704563); [Róbert Žalman](https://ieeexplore.ieee.org/author/37085672058), *“Big Data Analytics for Air Quality Monitoring at a Logistics Shipping Base via Autonomous Wireless Sensor Network Technologies”,* [2017 5th International Conference on Enterprise Systems (ES)](https://ieeexplore.ieee.org/xpl/conhome/8119322/proceeding)
8. Z. Ghaemi; A. Alimohammadi; M. Farnaghi, *“LaSVM-based big data learning system for dynamic prediction of air pollution in Tehran”,* Environ Monit Assess (2018) 190:300

Review:

Content

Well-written paper.  You took a challenging approach to the topic and did well with it.

Good opening about Venus's past, and covering both a commercial monitoring solution and a performance comparison of processing on varying configurations make an interesting dual-focus paper.

Good points about drawbacks of using a semi-commercial solution for AQI data collection/analytics.  This situation definitely shows the tension between commercial and community interests.

The limitations of improvement with Spark are a very solid illustration of the reality of these big data platforms and configurations.  Sometimes big data platforms provide limited improvement because of data characteristics or algorithmic limitations.

Nice use of graphics.

Sources

Good set of references drawn from commercial, government, and academic sources

Formatting, Grammar, Organization

It appears there are words omitted in a few sentences.  Example:  ”Intel is making it possible for the cities their air quality (AQ) data from all parts in real time.”

Section 3 needs a descriptive title instead of "Body"

Suggestions

In addition to running a spelling/grammar check, it can be helpful to have someone else read your paper to help you spot where words may have been omitted.