

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/336083177>

Air pollution prediction through internet of things technology and big data analytics

Article in *International Journal of Computational Intelligence Studies* · January 2019

DOI: 10.1504/IJCISTUDIES.2019.10024282

CITATION

1

READS

408

3 authors, including:



Yousef Farhaoui

Université Moulay Ismail, Faculty of sciences and Technics, Morocco

118 PUBLICATIONS 477 CITATIONS

[SEE PROFILE](#)



B. Aksasse

Université Moulay Ismail

81 PUBLICATIONS 528 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Moulay Ismail University [View project](#)



Image Indexation [View project](#)

Air pollution prediction through internet of things technology and big data analytics

Safae Sossi Alaoui*, Brahim Aksasse and
Yousef Farhaoui

Department of Computer Science,
Faculty of Sciences and Techniques,
Moulay Ismail University,
M2I Laboratory, ASIA Team,
B509 Boutalamine, 52000 Errachidia, Morocco
Email: sossialaouisafae@gmail.com
Email: baksasse@yahoo.com
Email: youseffarhaoui@gmail.com

*Corresponding author

Abstract: Air pollution is one of the biggest and serious challenges facing our planet nowadays. In fact, the need to develop models to predict this issue is considered so crucial. Indeed, our work aimed at building an accurate model to predict air quality of US country by using a dataset collected from connected devices of internet of things (IoT), namely from wireless sensor networks (WSN). Therefore, the huge amount of data captured by these sensors (approximately 1.4 million observations) brings about a highly complex data that necessitates new form of advanced analytic; it is about big data analytics. In this paper, we examine the possibility to make a fusion between the two new concepts big data and internet of things; in the context of predicting air pollution that occurs when harmful substances; like NO₂, SO₂, CO and O₃, are introduced into Earth's atmosphere.

Keywords: internet of things; IoT; wireless sensor networks; WSNs; air pollution; air quality index; AQI; big data analytics; Apache Spark.

Reference to this paper should be made as follows: Sossi Alaoui, S., Aksasse, B. and Farhaoui, Y. (2019) 'Air pollution prediction through internet of things technology and big data analytics', *Int. J. Computational Intelligence Studies*, Vol. 8, No. 3, pp.177–191.

Biographical notes: Safae Sossi Alaoui is a PhD student who is preparing her thesis within the Department of Computer Science in the Faculty of Sciences and Techniques in Errachidia, Morocco. The thesis topic is related to machine learning, big data and decision making. Furthermore, she is an Engineer of state in telecommunications and information technology. She graduated from the national institute of the posts and telecommunications (INPT) in 2015.

Brahim Aksasse is a Full Professor in the Department of Computer Science at the Faculty of Science and Technology Errachidia, Morocco. He graduated from Fez University in 2000. He spent three years as a Postdoctoral Fellow at the University of Bordeaux1 France (2001–2004). He is the Head of the Systems Analysis and Applied Informatics research team. His research works concern signal modelling, image filtering, image indexing and multidimensional spectral analysis.

Yousef Farhaoui is a Professor at the Department of Computer Science in Faculty of Sciences and Techniques, Moulay Ismail University, Morocco. He received his PhD degree in Computer Security from the University IBN Zohr. His research interest includes computer security, big data, data mining, data warehousing, data fusion, etc.

This paper is a revised and expanded version of a paper entitled 'Air pollution prediction through internet of things technology and big data analytics' presented at ISCSA2017: Special Issue on: 'Computational Intelligence and Applications', Errachidia. Morocco, 26–28 October 2017

1 Introduction

Internet of things (IoT) is a technology concept that currently used to highlight the rapidly growing network that encompasses everything connected to the internet. These connected objects include several devices such as; embedded sensors, smartphones, actuators and wearable's; capable of collecting and exchanging data between each other.

As a new report, Cisco Systems predicts there will be more than 50 billion devices connected to the internet by 2020 (RCR Wireless News, 2016). Hence, the amount of data that is going to be created by the IoT is going to require new advanced analytic techniques; principally a different processing approach called big data. Indeed, big data is a paradigm which reflects the changing world we live in, it describes a holistic information management strategy that includes and integrates structured and unstructured data that is so large and difficult to uncover its insights and meaning by using relational database engines.

Big data is arriving from multiple sources depending on its applications namely; public sector services, healthcare contributions, insurance services, industrialised and natural resources, transportation services, banking sectors and so on (Intellipaat Blog, 2016). In fact, in this work, we focus on one of the most serious environment issues confronting our civilisation nowadays; it's about air pollution which refers to the presence of any chemical, physical or biological agent that changes the natural characteristics of the atmosphere. Some of the common sources of air pollution are household combustion devices, motor vehicles, industrial facilities and forest fires.

The chemical compounds that lower the air quality are usually referred to as air pollutants which principally include carbon monoxide, ozone, nitrogen dioxide and sulphur dioxide. These pollutants can lead to more serious symptoms and conditions affecting human health such as the respiratory and inflammatory systems and at elevated levels they can cause heart disease and cancer (Wikipedia, 2017).

The rest of the paper is organised as follows; Section 2 briefly reviews what has been written about big data on IoT. Section 3 proceeds to outline the fundamental concepts of big data. Section 4 presents the concepts of IoT. Section 5 describes the functionalities of air pollution monitors. Section 6 highlights the methodology followed and the tool used. Section 7 presents the different results obtained. The last section concludes our paper.

2 Related works

In recent years, research works involving big data on IoT is quite relevant.

First of all, Aly et al. (2015) offered a general survey of the different aspects of big data on IoT including applications architecture, technologies, techniques, challenges and future directions. In fact, according to Aly et al. (2015) there exist already developed applications of IoT in different fields like transportation, smart environments domain, health care domain, food sustainability, and futuristic applications, by using myriad of technologies that solve IoT data management for instance big data, cloud computing, semantic sensor web, data fusion techniques, and middleware. On the other hand, IoT faces many challenges namely architecture, environment innovation, technical, hardware, privacy and security challenges as well as standard, business, development strategies and data processing challenges.

Souza and Amazonas (2015) suggested an outline algorithm using big data processing and IoT architecture namely the Hadoop framework and Mahout K-means algorithm implementation. This algorithm runs integrated with the IoT architecture implemented by the LinkSmart middleware. Its scalability is guaranteed by the use of the technology big data allowing physical objects and sensors to be connected directly to the middleware. In fact, it offers a high scalability allowing the creation of clusters with hundreds or thousands of instances Hadoop that can be connected seamlessly in the client and LinkSmart applications. The object-oriented structured programming allows integrating other implementations in the extended LinkSmart middleware.

Sun et al. (2016) proposed the concept of “smart and connected communities (SCC)”, which designed a unified framework integrating smart cities and beyond. This article suggests an IoT architecture, and choose best IoT enabling technologies, and IoT services, applications, and standards in order to achieve the purpose of this article whose vision is to ameliorate livability, preservation, revitalisation, and sustainability, of a small town, different from so-called smart cities as well as to shed light on the opportunities and challenges of applying IoT and big data analytics to culture preservation and revitalisation of SCC. As a case study, TreSight was presented as an integration of IoT and big data analytics for smart tourism and sustainable cultural heritage in the city of Trento from Italy.

Ochoa et al. (2017) worked with ten articles that contributed in the current knowledge in several aspects related to the design, implementation and use of IoT-enabled cyber-physical systems (CPS). Indeed, the research community has recognised the complexity of CPS in dealing with the two intertwined concepts namely IoT and big data which are involved in the new generations of collaborative solutions; especially those based on devices heterogeneity and ad hoc interactions.

3 Fundamentals of big data

3.1 The 10 Vs model of big data

Understanding and effectively communicating a concept often requires first building a simple model (The 42 V's of Big Data and Data Science, no date). In fact, big data is a term commonly based on the three fundamental V's model which include volume,

velocity and variety, but it can be extended to ten V's; which describe more its specific characteristics and properties; gathering them under Table 1 (MapR, no date):

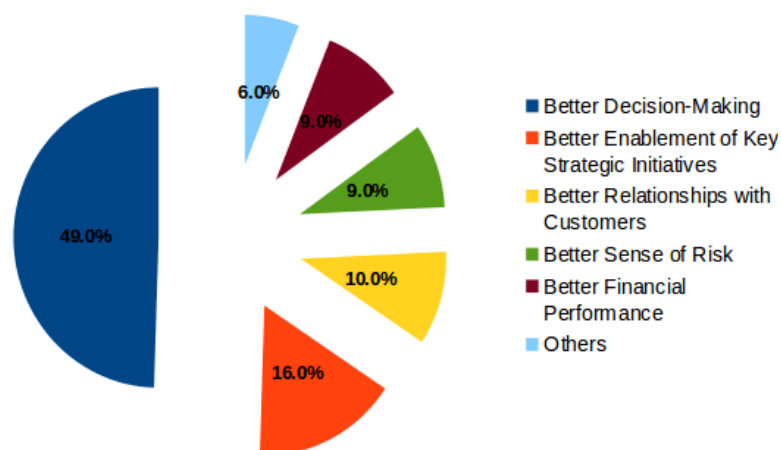
Table 1 Description of 10 V's model of big data

<i>10 V's model</i>	<i>Definition</i>
Volume	Size of data
Velocity	Speed, in which data is generated, produced, created, or refreshed.
Variety	Different types of data; structured semi structured and mostly unstructured data.
Variability	dynamic, evolving, spatiotemporal data, time series, seasonal, and any other type of non-static behaviour in the data sources, customers, objects of study, etc.
Veracity	Data accuracy
Validity	Data quality, governance, master data management (MDM) on massive, diverse, distributed, heterogeneous, 'unclean' data collections.
Vulnerability	Bringing new security concerns
Volatility	How long does the data need to be kept for before it is considered irrelevant, historic, or not useful any longer.
Visualisation	Representing data in graphic form.
Value	Extracting useful information from data.

3.2 Big data analytics

Big data analytics is the process of collecting, organising and analysing large sets of data to discover hidden pattern, unknown correlations, market trends, customer preferences and other useful information that can help organisations make more-informed business decisions (SearchBusinessAnalytics, no date).

Figure 1 Key benefits of big data analytics (see online version for colours)



Big data analytics is considered as a key competitive resource for many companies. So, according to the 'Peer-Research Big Data Analytics Survey', 74% of the respondents have agreed that big data analytics is adding value to their organisation by offering; better decision making (49%), better enablement of key strategic initiatives (16%), better

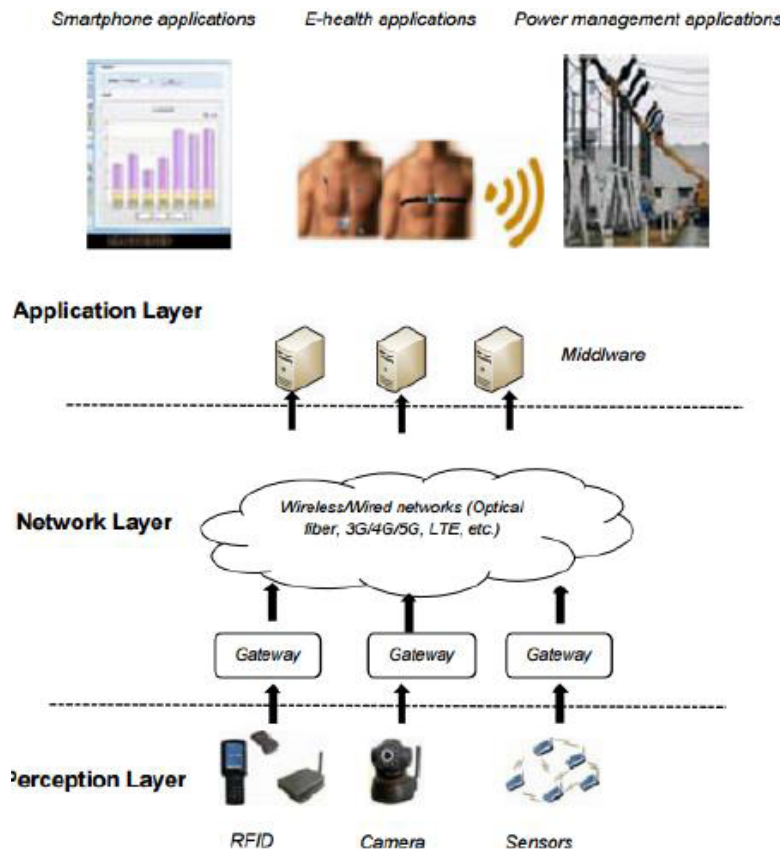
relationships with customers (10%), better sense of risk (9%), better financial performance (9%) and others (6%), as shown in Figure 1 (Edureka Blog, 2015).

4 IoT concepts

4.1 The three layer architecture of IoT

The most well-known architecture of IoT is generally divided into three layers: the perception layer, the network layer and the application layer. Figure 2 illustrates the component of each layer.

Figure 2 Three-layer architecture of IoT (see online version for colours)



4.1.1 The perception layer

The perception layer is the core layer as well as the information origin of IoT. The purpose of this layer is to identify objects and to gather information by the technologies of sensors, wireless sensors network (WSN), radio frequency identification (RFID) system, tags and reader-writers, camera, intelligent terminals, global position system (GPS), electronic data interface (EDI), objects, and so on. (Wu et al., 2010)

4.1.2 The network layer

The network layer, also called transport layer, is like the neural network and brain of IoT, its principal function is transmitting and processing information obtained from perception layer. The network layer contains a convergence network of communication and internet network, network management centre, information centre and intelligent processing centre, etc. (Wu et al., 2010)

4.1.3 The application layer

The application layer, additionally called service layer, contains two sub-layers; data management sub-layer and application service sub-layer. The first sub-layer serves to process complex data as well as uncertain information; also it provides directory service by using service oriented architecture (SOA), cloud computing technologies, and so on. The second sub-layer converts information into content and gives good user interface for both; higher level enterprise application and end users (Jia et al., 2012).

4.2 IoT technologies

Today, the concept of IoT has evolved due to the integration of different technologies, including RFID, WSNs and RFID sensor networks (RSN).

RFID system is based on a reading device named a reader and one or more tags, this system employs radio waves to read and capture information saved on a tag attached to an object, RFID might identify objects wirelessly without line-of-sight (Atzori et al., 2010).

WSN aims at sensing as well as monitoring the environment. The system works for periods changing autonomously from weeks to years. The sensor network consists of large number of sensor nodes that can be used on the ground, in the air, in vehicle, inside building, ..., etc. (Atzori et al., 2010).

RSN is an integration of RFID with WSN which serves the identification and location of an object and also provides information concerning the condition of the object carrying the sensors enabled RFID tag (Atzori et al., 2010).

Table 2 Comparison between RFID systems, wireless sensor networks, and RFID sensor networks

	<i>RFID</i>	<i>WSN</i>	<i>RSN</i>
Processing	No	Yes	Yes
Sensing	No	Yes	Yes
Communication	Asymmetric	Peer-to-peer	Asymmetric
Range (m)	10	100	3
Power	Harvested	Battery	Harvested
Lifetime	Indefinite	< 3 years	Indefinite
Size	Very small	Small	Small
Standard	ISO18000	IEEE 802.15.4	None

Source: Atzori et al. (2010)

5 Air pollution monitors

Air pollution is a real public health and environmental issue that can lead to more serious effects including global warming, acid rain, and the deterioration of the ozone layer. Table 3 names the four common pollutants, their abbreviation, and their definitions.

Table 3 Definition of major air pollutants

<i>Pollutant</i>	<i>Abbreviation</i>	<i>Definition</i>
Nitrogen dioxide	NO ₂	A reddish-brown gas; with a strong smell at high levels; which hails from the burning of fossil fuels.
Sulphur dioxide	SO ₂	A toxic gas that cannot be seen or smelled at low levels but can have a 'rotten egg' smell at high levels. It is released naturally by volcanic activity.
Carbon monoxide	CO	A colourless, odourless, and tasteless gas that comes from the burning of fossil fuels, mostly in cars.
Ozone	O ₃	A pale blue gas with a distinctively pungent smell. It can be found in two places; in the Earth's upper atmosphere and at ground level.

Source: Major Air Pollutants (no date)

5.1 What is the AQI?

The air quality index (AQI) can be defined as a number used by government agencies to report daily air quality in order to communicate to the public how clean or unhealthy the air is (Air Quality Index, no date).

Figure 3 Air quality index levels (see online version for colours)

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health warnings of emergency conditions. The entire population is more likely to be affected.
Hazardous	301 to 500	Health alert: everyone may experience more serious health effects

As shown in Figure 3, each AQI category has assigned to a specific colour and the corresponding health warnings.

In fact, knowing what the colour codes mean may help people protect their health during air quality levels associated with low, moderate, high and very high health risks.

5.2 *Monitoring air pollution*

In environmental monitoring system, it is necessary to make the periodic updates in order to react immediately against disaster; indeed, when air pollution is higher than critical level, the system will start raising the alarm. The risky level threshold changes depending on the area like a school, a factory, or an apartment (Jain and Vijaygopalan, 2010).

The types of sensors required for measuring air pollution are different, for the case of our study, the Clean Air Act; a USA federal law designed to control air pollution on a national level; requires every state to install a network of air monitoring stations for criteria pollutants, respecting criteria set by the Office of Air Quality Planning and Standards (OAQPS) for their location and operation. The monitoring stations in this network are named the State and Local Air Monitoring Stations (SLAMS). The states is obliged to provide OAQPS with an annual summary of monitoring results at each SLAMS monitor, and detailed results must be accessible to OAQPS upon request. To get more prompt and detailed information about air quality in strategic locations across the nation, OAQPS established an additional network of monitors: the National Air Monitoring Stations (NAMS) which are part of the SLAMS network, must not only meet more stringent monitor sitting, equipment type, and quality assurance criteria, but also must submit detailed quarterly and annual monitoring results to OAQPS (US EPA, no date).

6 **Methodology**

6.1 *Dataset description*

In this paper, we had used a dataset taken from a website named Kaggle (Kaggle: Your Home for Data Science, no date) which provides online datasets for data scientists and aims at discovering and seamlessly analysing open data. Our dataset (US Pollution Data, no date) deals with air pollution in the USA country and it has been well documented by the US Environmental Protection Agency (EPA). Hence, it describes the measurements of the four major pollutants namely; nitrogen dioxide, sulphur dioxide, carbon monoxide and ozone; for every day from 2000 to 2016. It contains a number of observations totalled to over 1.4 million as well as a total of 29 attributes shown on Tables 4 and 5.

Each of the four pollutants (NO₂, SO₂, CO and O₃) has five specific attributes described in Table 5.

The last attribute we had created is based on Figure 3, it called 'airQuality', it has six values: {1, 2, 3, 4, 5, 6} for successively the following classes {good, moderate, unhealthy-for-sensitive-groups, unhealthy, very-unhealthy, hazardous}.

Table 4 Description of attributes

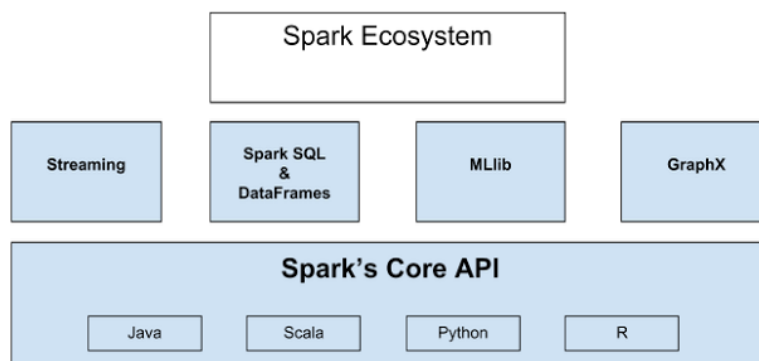
<i>Name of attribute</i>	<i>Details of attribute</i>
State code	The code allocated by US EPA to each state
County code	The code of counties in a specific state allocated by US EPA
Site num.	The site number in a specific county allocated by US EPA
Address	Address of the monitoring site
State	State of monitoring site
County	County of monitoring site
City	City of the monitoring site
Date local	Date of monitoring

Table 5 Description of attributes for NO₂

<i>Name of attribute</i>	<i>Details of attribute</i>
NO ₂ units	The units measured for NO ₂
NO ₂ mean	The arithmetic mean of concentration of NO ₂ within a given day
NO ₂ AQI	The calculated AQI of NO ₂ within a given day
NO ₂ 1st max value	The maximum value obtained for NO ₂ concentration in a given day
NO ₂ 1st max hour	The hour when the maximum NO ₂ concentration was recorded in a given day

6.2 Technologies used

Apache Spark is an open source processing framework as well as a fast and general-purpose cluster computing system. It supports rapid application development for big data; it enables high-level APIs in different programming languages; Java, Scala, Python and R, and an optimised engine that supports general execution graphs. Additionally, it provides a rich set of higher-level devices like Spark SQL for SQL and structured data processing, MLlib for machine learning (ML), GraphX for graph processing, and Spark streaming (Overview – Spark 2.2.0 Documentation, no date).

Figure 4 Spark ecosystem (see online version for colours)

6.3 Selected algorithm

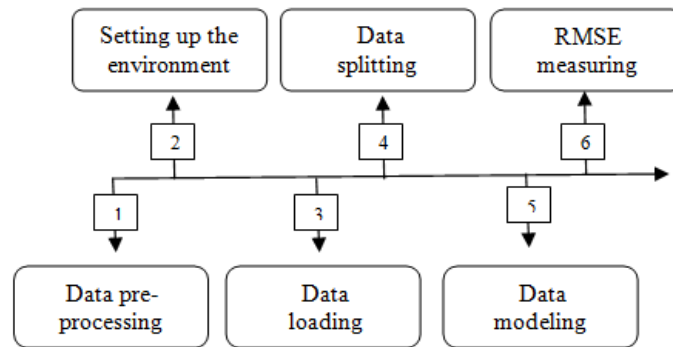
Spark MLlib (MLlib: Main Guide – Spark 2.2.0 Documentation, no date) can be defined as Spark’s ML library, it offers many tools like: ML algorithms, featurisation, pipelines, persistence and utilities. It aims at making practical ML scalable and easy. In this paper, we had chosen:

- Gradient-boosted trees (GBTs): a popular classification algorithm among others (Sossi Alaoui et al., 2017) as well as a regression technique using ensembles of decision trees. GBTs iteratively train decision trees so as to minimise a loss function. Like decision trees, GBTs deal with categorical features, reach out to the multiclass classification setting, don’t demand feature scaling, and can capture nonlinearities and feature interactions.
- ML pipelines: a set of tools which provide users to construct, evaluate, and tune ML pipelines.

6.4 Basic followed steps

In this work, we had adopted a simple and effective strategy for analysing our data as well as creating a model based on GBTs classifier. Figure 5 serves to outline the six steps followed.

Figure 5 Basic steps



Step 1 Data pre-processing

It consists on preparing data to be more easily and effectively processed for the purpose of the user; because data is generally incomplete, noisy and inconsistent.

Step 2 Setting up the environment

In this section, we had created an account in Databricks (no date) which is a unified analytics platform; founded by the creators of Apache Spark; its goal is to help clients with cloud-based big data processing using Spark.

Step 3 Data loading

Data loading is the act of copying and loading the dataset from the source file to DBFS (Databricks File System, no date) (Databricks File System) which is a distributed file system installed on Spark Clusters in Databricks.

Step 4 Data splitting

Data splitting is the process of dividing available data into two parts, mostly for cross-validation purposes. One portion; named training data; is utilised to build a predictive model and the other; called test data; to evaluate the model's performance.

Step 5 Data modelling

Data modelling consists on building a ML model so as, not only to predict air quality in the future but also to make recommendations regarding air pollution effects on human health.

Step 6 Root-mean-square error (RMSE) measuring

RMSE is the well-known metric that measures the differences between the values predicted by the model and the values really observed.

7 Results

After developing our model based on GBTs as well as ML pipelines and using Python in Spark, we obtained the results bellow:

First, we transform our data to a DataFrame; called df; which is a distributed collection of data organised into named columns. Then, we used the count() function which returns the number of elements in a dataset or the resulting output of an RDD operation. Figure 6 outlines the number 1,048,575 of rows of our dataset.

Figure 6 Rows number of dataset (see online version for colours)

```
print "Our dataset has %d rows." % df.count()

> (1) Spark Jobs
Our dataset has 1048575 rows.
```

To see a sample of our dataset, we call display() for the DataFrame 'df' as shown in Figure 7.

In Figure 8, the function printSchema() prints out the schema for our Spark DataFrame in a tree format.

To split the dataset randomly into 70% for training data and 30% for test data (Figure 9), we used the function RandomSplit() as shown in Figure 10.

To view the results easier like in Figure 11, we limited the columns displayed to

- Airquality: the true value of air quality.
- Prediction: our predicted value of air quality using our model.
- Featurecols: the other feature columns for our dataset.

Figure 7 Dataset visualisation

```
display(df)
```

► (1) Spark Jobs

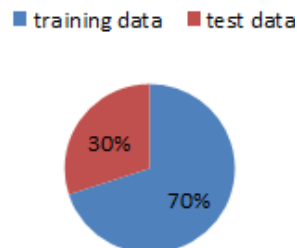
StateCode	CountyCode	SiteNum	NO2AQI	O3AQI	SO2AQI	COAQI	airQuality
4	13	3002	46	34	13	0	1
4	13	3002	46	34	13	25	1
4	13	3002	46	34	0	0	1
4	13	3002	46	34	0	25	1
4	13	3002	34	27	4	0	1
4	13	3002	34	27	4	26	1
4	13	3002	34	27	0	0	1

Figure 8 Types of attributes

```
df.printSchema()

► df: pyspark.sql.dataframe.DataFrame = [StateCode: double, CountyCode:
double ... 6 more fields]

root
 |-- StateCode: double (nullable = true)
 |-- CountyCode: double (nullable = true)
 |-- SiteNum: double (nullable = true)
 |-- NO2AQI: double (nullable = true)
 |-- O3AQI: double (nullable = true)
 |-- SO2AQI: double (nullable = true)
 |-- COAQI: double (nullable = true)
 |-- airQuality: double (nullable = true)
```


Figure 9 Data splitting percentage (see online version for colours)


Finally, computing evaluation metrics is important for understanding the quality of predictions. In our case we had used RMSE to evaluate our predictions. In general, for good predictive model RMSE values should be less than 0.3 (Veerasamy et al., 2011). indeed, our RMSE (Figure 12) is equal to $0.13 < 0.3$ as a result, we can say that our model is accurate.

Figure 10 Data splitting (see online version for colours)

```
# Split the dataset randomly into 70% for training and 30% for testing.
train, test = df.randomSplit([0.7, 0.3])
print "We have %d training examples and %d test examples." % (train.count(), test.count())
```

► (2) Spark Jobs

►  train: pyspark.sql.dataframe.DataFrame = [StateCode: double, CountyCode: double ... 6 more fields]

►  test: pyspark.sql.dataframe.DataFrame = [StateCode: double, CountyCode: double ... 6 more fields]

We have 734052 training examples and 314523 test examples.

Figure 11 Predictions (see online version for colours)

```
display(predictions.select("airQuality", "prediction", *featuresCols))
```

airQuality	prediction	StateCode	CountyCode	SiteNum	NO2AQI	O3AQI	SO2AQI	COAQI
1	1.007263230921091	4	13	3002	1	48	0	0
1	1.007491368877246	4	13	3002	1	48	0	7
1	1.0000949866670232	4	13	3002	3	32	0	0
1	0.999773853342812	4	13	3002	3	32	1	9
1	1.0010394459116573	4	13	3002	3	39	0	10
1	1.0004593060588285	4	13	3002	3	39	1	0
1	1.0005595942105392	4	13	3002	3	47	10	0
2	1.9929548733431994	4	13	3002	3	71	0	7

Figure 12 RMSE value (see online version for colours)

```
print "RMSE on our test set: %g" % rmse
```

► (1) Spark Jobs

RMSE on our test set: 0.131808

8 Conclusions

This paper has concentrated of making a fusion between two new concepts; big data and IoT; when dealing with environmental issues namely Air pollution. Indeed, working with US pollution dataset and using Spark technology in Databricks platform, we managed to build an accurate model capable of making good predictions for air quality, which can help to the better understanding of negative effects produced by air pollution in our life and making more efforts to prevent, control and reduce this issue as soon as possible.

References

- Air Quality Index (no date) *A Guide to Air Quality and Your Health* [online] https://www.airnow.gov/index.cfm?action=aqi_brochure.index (accessed 28 August 2017).
- Aly, H., Elmogy, M. and Barakat, S. (2015) 'Big data on internet of things: applications, architecture, technologies, techniques, and future directions', *International Journal of Computer Science Engineering*, Vol. 4, No. 6, pp.300–313.
- Atzori, L., Iera, A. and Morabito, G. (2010) 'The internet of things: a survey', *Computer Networks*, Vol. 54, No. 15, pp.2787–2805, DOI: 10.1016/j.comnet.2010.05.010.
- Databricks (no date) *Databricks – Making Big Data Simple* [online] <https://databricks.com/> (accessed 13 October 2017).
- Databricks File System (DBFS) (no date) *Databricks Documentation* [online] <https://docs.databricks.com/user-guide/dbfs-databricks-file-system.html> (accessed 13 October 2017).
- Eureka Blog (2015) *10 Reasons Why Big Data Analytics is the Best Career Move* | Eureka.co, 8 January [online] <https://www.eureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move> (accessed 3 August 2017).
- Intellipaat Blog (2016) *7 Examples of Big Data Use cases in Real Life*, 13 July [online] <https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/> (accessed 26 July 2017).
- Jain, P.C. and Vijaygopalan, K.P. (2010) 'RFID and wireless sensor networks', *Proceedings of ASCNT-2010*, CDAC, Noida, India, pp.1–11.
- Jia, X. et al. (2012) 'RFID technology and its applications in internet of things (IoT)', in *Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference*, IEEE, pp.1282–1285 [online] <http://ieeexplore.ieee.org/abstract/document/6201508/> (accessed 21 August 2017).
- Kaggle: Your Home for Data Science (no date) [online] <https://www.kaggle.com/> (accessed 27 July 2017).
- Major Air Pollutants (no date) [online] <https://www.infoplease.com/science-health/environment/major-air-pollutants> (accessed 24 August 2017).
- MapR (no date) *Top 10 Big Data Challenges – A Serious Look at 10 Big Data V's* [online] <https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/> (accessed 3 August 2017).
- Ochoa, S.F., Fortino, G. and Di Fatta, G. (2017) 'Cyber-physical systems, internet of things and big data', *Future Generation Computer Systems*, Vol. 75, pp.82–84, DOI: 10.1016/j.future.2017.05.040.
- RCR Wireless News (2016) *50B IoT Devices Connected by 2020 – beyond the Hype and into Reality*, 28 June [online] <http://www.rcrwireless.com/20160628/opinion/reality-check-50b-iot-devices-connected-2020-beyond-hype-reality-tag10> (accessed 26 July 2017).
- SearchBusinessAnalytics (no date) What is big data analytics? – Definition from WhatIs.com [online] <http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics> (accessed 3 August 2017).
- Sossi Alaoui, S., Farhaoui, Y. and Aksasse, B. (2017) 'A comparative study of the four well-known classification algorithms in data mining', in *Advanced Information Technology, Services and Systems, International Conference on Advanced Information Technology, Services and Systems*, Springer, Cham (Lecture Notes in Networks and Systems), pp.362–373, DOI: 10.1007/978-3-319-69137-4_32.
- Souza, A.M.C. and Amazonas, J.R.A. (2015) 'An outlier detect algorithm using big data processing and internet of things architecture', *Procedia Computer Science*, Vol. 52, pp.1010–1015, DOI: 10.1016/j.procs.2015.05.095.
- MLlib: Main Guide – Spark 2.2.0 Documentation (no date) [online] <https://spark.apache.org/docs/latest/ml-guide.html> (accessed 13 October 2017).

- Overview – Spark 2.2.0 Documentation* (no date) [online] <https://spark.apache.org/docs/latest/> (accessed 28 August 2017).
- Sun, Y. et al. (2016) 'Internet of things and big data analytics for smart and connected communities', *IEEE Access*, Vol. 4, pp.766–773, DOI: 10.1109/ACCESS.2016.2529723.
- The 42 V's of Big Data and Data Science* (no date) [online] <http://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html> (accessed 3 August 2017).
- US EPA, O. (no date) Air Pollution Monitoring [online] <https://www3.epa.gov/airquality/montring.html#criteria> (accessed 30 August 2017).
- US Pollution Data (no date) [online] <https://www.kaggle.com/sogun3/uspollution> (accessed 27 July 2017).
- Veerasamy, R. et al. (2011) 'Validation of QSAR models-strategies and importance', *International Journal of Drug Design & Discovery*, Vol. 2, No. 3, pp.511–519.
- Wikipedia (2017) *Air Pollution* [online] https://en.wikipedia.org/w/index.php?title=Air_pollution&oldid=792269331 (accessed 25 July 2017).
- Wu, M. et al. (2010) 'Research on the architecture of internet of things', in *Advanced Computer Theory and Engineering (ICACTE)*, 2010 3rd International Conference, IEEE, pp.V5–484 [online] <http://ieeexplore.ieee.org/abstract/document/5579493/> (accessed 15 August 2017).