International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 6 June 2021, pp: 12-17 www.ijaem.net ISSN: 2395-5252

Weather Data Analysis Using Pyspark

Girish Prabu Sd, Jaya Surya S, Monish Prakash S
[1] msc Data science III Year
[2]DEPARTMENT OF DATA SCIENCE
COIMBATORE INSTITUTE OF TECHNOLOGY

Guided by:
Dr.M.SujithraM.C.A, M.Phil., Ph.D.
Dr.P.VelvadivuM.C.A., M.Phil. Ph.D.

Submitted: 25-05-2021 Revised: 31-05-2021 Accepted: 03-06-2021

ABSTRACT: The weatheris very important factor in many perspectives such as business, tourism, etc.This weather data influence success of many operations. For example, temperature and rainfall in one area is not same for other area. So it is important to analyze the climatic data in different regions, for the purpose of better decision making and precaution measures. This project helps to analyze the various weather condition data of various region. Weather data are radically increasing in volume and complexity. So the concept of big data emerges here. Hadoop, Hive, Spark, Machine leaning techniques, MapReducewhich are big data concepts are used here to analyze the weather data for the purpose of better understanding and decision-making.The main Objective is toanalyze the climatic data using Hadoop to extract significant knowledge for the purpose of better decision-making.

Keywords: Big data – Weather data–Machine learning – clustering–Hadoop–Hive–Spark–MapReduce.

I. INTRODUCTION

Weather is a mix of events that occurs in our atmosphere, weather at different part of place differs from each other and changes at time to time, every activity either economical or agricultural or sports based active completely depends on the weather during that day at that particular place, in other terms the ups and downs of particular types of stocks depends on how good the weather is, another example is weather decides the team winning a cricket match like if due is too high in atmosphere then it's hard to hold the ball, which makes the bowler too loose his grip and bowl at wrong side and makes the match more favorable to the batsman, another example is from agriculture sector, consider a farmer is planting banana and ready for harvesting but on certain one day a huge storm ruins all his 10 months effort in a day so, it's necessary to analysis the weather data in that particular region, read them to the extinct and make use of the insights for better decision making.

1.1 Hadoop

Analyzing the weather data of every region isn’t the simple task that can be done just by making use of some machine learning algorithms and fitting simply into the model, to tell in brief, it impossible to build a machine learning model using python, where the record size if of around 3 million. This kind of huge volume data is called as big data. Big Data is also **data** but with a **huge size**. Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time. In short, such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

1.1.1. So, how specifically big data differs from traditional data?

1) **Volume –** As the name **“BIG DATA**” clearly tells that the data we dealing is with enamors volume. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with Big Data.

2) **Variety-** Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured

data poses certain issues for storage, mining and analyzing data.
3) **Velocity-** Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors,Mobile,devices, etc. The flow of data is massive and continuous.
4) **Variability –** This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

1.1.2. So, how big data is been processed?

Processing big data, not every platform can do it, there are specific application that are built to process big data like Xplenty, Apache Hadoop, CDH (**Cloudera** Distribution for **Hadoop**), **Cassandra,** Knime, Datawrapper, MongoDB, Lumify are some of major platform used for processing big data.
Here in weather data analysis for big data analysis Apache Hadoop is used for Data cleaning and model fitting. So, what is apache Hadoop? Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

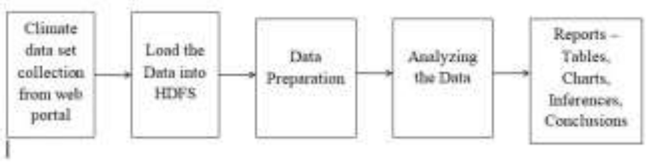
1.1.3 PYSPARK

Apache **Spark** is a **data** processing framework that can quickly perform processing tasks on very **large data** sets, and can also distribute **data** processing tasks across multiple computers, either on its own or in tandem with other distributed computing tools. Pyspark is python version of sparks, the reason pyspark is famous the front end is connected through python which makes us to access every python libraries in spark. Python is known for its machine learning libraries among data scientist community, by embedding python with spark (pyspark) then huge sea of libraries are accessible view sparks which makes analytics on big data much easier.

1.2 PowerBI

Microsoft Power BI is **used to** find insights within an organization's data. **Power BI** can help connect disparate data sets, transform and clean the data into a data model and create charts or graphs to provide visuals of the data. ... This **Power BI** preview shows the reporting and dashboard capabilities that **Power BI** offers. PowerBi is considered as one of the best analytical visualization tool and here for getting further insights after data cleaning and fitting the model the processed data is extracted and connected with powerBi and interpreted.

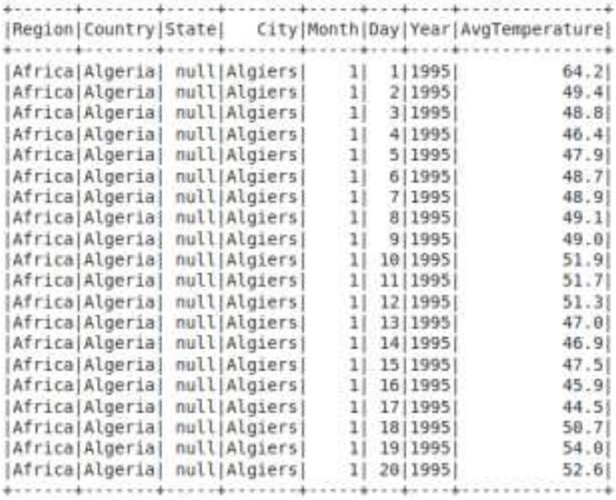
2. Road Map


Climate data set collection from web portal
Load the Data into HDFS
Data Preparation
Analyzing the Data
Reports – Tables, Charts, Inferences, Conclusions

The weather data set for cities across 120 countries from 1995 to 2020 for each data has been collected from the reputed medium (links will be shared in reference part), once the data is loaded it's mandatory to clean the data I.e, check for null value and fill it with mean and mode statistics, eliminating duplicate rows and encode the categorical features and perform appropriate feature engineering and once data preparation is done k means clustering model is fitted for grouping similar cities based on their temperature and dig down deep for gaining major insights and finally all the predicted and processed data is connected in power bi for visualization.

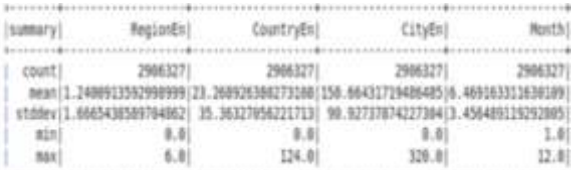
II. DATA PREPROCESSING.

2.1. Dataset Description.
1) Over 120 countries we have collected the weather data. The dataset contains data from the year 1995 till 2020.
2) The data set contains over 2.9 million rows and 8 columns.
3) For each and every day we have collected the average weather temperature of the particular day.
Fig: Just of weather data.

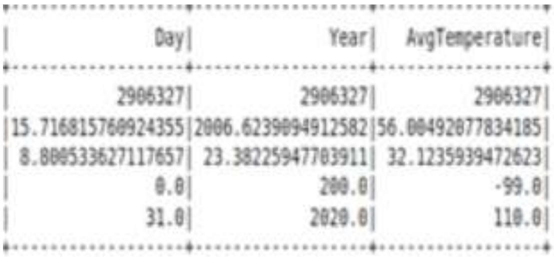


```
+------+-------+-----+-------+-----+---+----+--------------+
|Region|Country|State|   City|Month|Day|Year|AvgTemperature|
+------+-------+-----+-------+-----+---+----+--------------+
|Africa|Algeria| null|Algiers|    1|  1|1995|          64.2|
|Africa|Algeria| null|Algiers|    1|  2|1995|          49.4|
|Africa|Algeria| null|Algiers|    1|  3|1995|          48.8|
|Africa|Algeria| null|Algiers|    1|  4|1995|          46.4|
|Africa|Algeria| null|Algiers|    1|  5|1995|          47.9|
|Africa|Algeria| null|Algiers|    1|  6|1995|          48.7|
|Africa|Algeria| null|Algiers|    1|  7|1995|          48.9|
|Africa|Algeria| null|Algiers|    1|  8|1995|          49.1|
|Africa|Algeria| null|Algiers|    1|  9|1995|          49.0|
|Africa|Algeria| null|Algiers|    1| 10|1995|          51.9|
|Africa|Algeria| null|Algiers|    1| 11|1995|          51.7|
|Africa|Algeria| null|Algiers|    1| 12|1995|          51.3|
|Africa|Algeria| null|Algiers|    1| 13|1995|          47.0|
|Africa|Algeria| null|Algiers|    1| 14|1995|          46.9|
|Africa|Algeria| null|Algiers|    1| 15|1995|          47.5|
|Africa|Algeria| null|Algiers|    1| 16|1995|          45.9|
|Africa|Algeria| null|Algiers|    1| 17|1995|          44.5|
|Africa|Algeria| null|Algiers|    1| 18|1995|          50.7|
|Africa|Algeria| null|Algiers|    1| 19|1995|          54.0|
|Africa|Algeria| null|Algiers|    1| 20|1995|          52.6|
+------+-------+-----+-------+-----+---+----+--------------+
```

Fig: Weather data Description



```
+-------+------------------+------------------+------------------+-----------------+
|summary|          RegionEn|         CountryEn|            CityEn|            Month|
+-------+------------------+------------------+------------------+-----------------+
|  count|           2906327|           2906327|           2906327|          2906327|
|   mean|1.2400913592998999|23.268926384273108|150.66431719486485|6.469163311630189|
| stddev|1.6665438589704862| 35.36327056221713| 90.92737874227304|3.456489115292005|
|    min|               0.0|               0.0|               0.0|              1.0|
|    max|               6.0|             124.0|             320.0|             12.0|
```

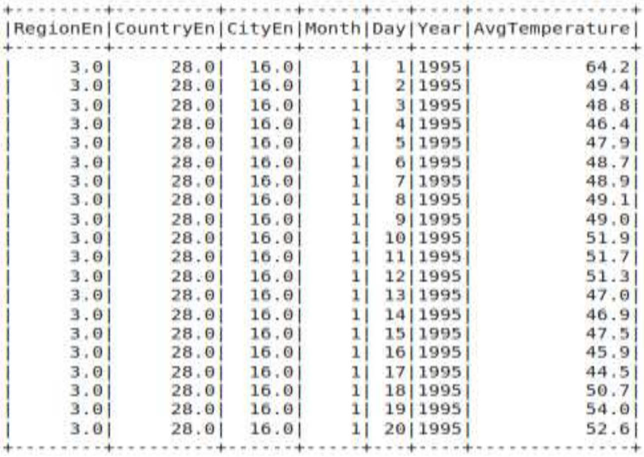


```
+------------------+------------------+-----------------+
|               Day|              Year|   AvgTemperature|
+------------------+------------------+-----------------+
|           2906327|           2906327|          2906327|
|15.716815760924355|2006.6239094912582|56.00492077834185|
| 8.800533627117657| 23.38225947703911| 32.1235939472623|
|               0.0|             200.0|            -99.0|
|              31.0|            2020.0|            110.0|
+------------------+------------------+-----------------+
```

Interpretation:
The above weather data consist of 8 features and 2906327 records, among them 3 features are categorical and it is necessary to encode them before moving further.

2.2 Data pre-processing and cleaning.

Fig: Feature encoding

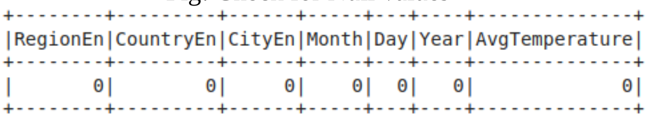


```
+--------+---------+------+-----+---+----+--------------+
|RegionEn|CountryEn|CityEn|Month|Day|Year|AvgTemperature|
+--------+---------+------+-----+---+----+--------------+
|     3.0|     28.0|  16.0|    1|  1|1995|          64.2|
|     3.0|     28.0|  16.0|    1|  2|1995|          49.4|
|     3.0|     28.0|  16.0|    1|  3|1995|          48.8|
|     3.0|     28.0|  16.0|    1|  4|1995|          46.4|
|     3.0|     28.0|  16.0|    1|  5|1995|          47.9|
|     3.0|     28.0|  16.0|    1|  6|1995|          48.7|
|     3.0|     28.0|  16.0|    1|  7|1995|          48.9|
|     3.0|     28.0|  16.0|    1|  8|1995|          49.1|
|     3.0|     28.0|  16.0|    1|  9|1995|          49.0|
|     3.0|     28.0|  16.0|    1| 10|1995|          51.9|
|     3.0|     28.0|  16.0|    1| 11|1995|          51.7|
|     3.0|     28.0|  16.0|    1| 12|1995|          51.3|
|     3.0|     28.0|  16.0|    1| 13|1995|          47.0|
|     3.0|     28.0|  16.0|    1| 14|1995|          46.9|
|     3.0|     28.0|  16.0|    1| 15|1995|          47.5|
|     3.0|     28.0|  16.0|    1| 16|1995|          45.9|
|     3.0|     28.0|  16.0|    1| 17|1995|          44.5|
|     3.0|     28.0|  16.0|    1| 18|1995|          50.7|
|     3.0|     28.0|  16.0|    1| 19|1995|          54.0|
|     3.0|     28.0|  16.0|    1| 20|1995|          52.6|
+--------+---------+------+-----+---+----+--------------+
```

Interpretation:

Three features Region, country and city are of categorical non-numeric type, so before fitting the model it is necessary to convert them to binomial or other numerical discrete value and thanks to pysparks , a spark query was built and the resulted encoded was displayed in above figure.

Fig: Check for Null values

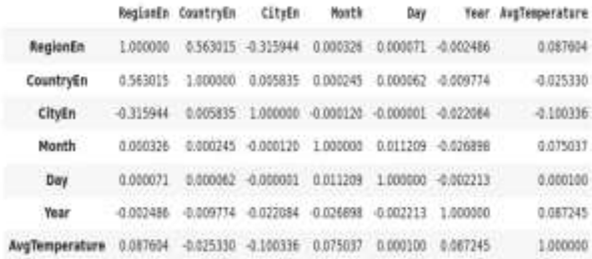


```
+--------+---------+------+-----+---+----+--------------+
|RegionEn|CountryEn|CityEn|Month|Day|Year|AvgTemperature|
+--------+---------+------+-----+---+----+--------------+
|       0|        0|     0|    0|  0|   0|             0|
+--------+---------+------+-----+---+----+--------------+
```

Interpretation:

With the help of spark query null values are checked in 2906327 X 8 values and luckily resulted with 0 null values.

Fig: correlation matrix.



	RegionEn	CountryEn	CityEn	Month	Day	Year	AvgTemperature
RegionEn	1.000000	0.563015	-0.315944	0.000326	0.000071	-0.002486	0.087604
CountryEn	0.563015	1.000000	0.005835	0.000245	0.000062	-0.009774	-0.025330
CityEn	-0.315944	0.005835	1.000000	-0.000120	-0.000001	-0.022084	-0.100336
Month	0.000326	0.000245	-0.000120	1.000000	0.011209	-0.026898	0.075037
Day	0.000071	0.000062	-0.000001	0.011209	1.000000	-0.002213	0.000100
Year	-0.002486	-0.009774	-0.022084	-0.026898	-0.002213	1.000000	0.087245
AvgTemperature	0.087604	-0.025330	-0.100336	0.075037	0.000100	0.087245	1.000000

Interpretation:

From the above correlation matrix, we are clearly able to see the strongly associated features with avgtemperature (City) and weakly associated features with Avgtemperature(day) and now we can further move to model fitting.

III. K-MEANS CLUSTERING.

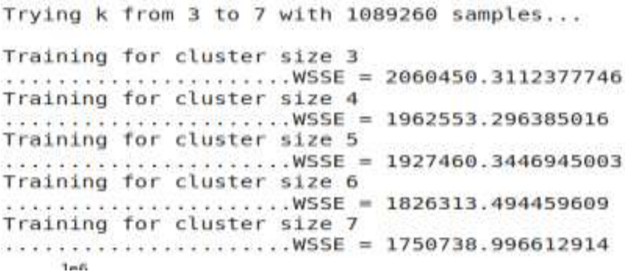
K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed appropriately. The main idea is to define k centers, one for each cluster.

3.1 Elbow Method:

The KElbowVisualizer implements the "elbow" method to help data scientists select the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, then the "elbow" (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. In the visualizer "elbow" will be annotated with a dashed line.

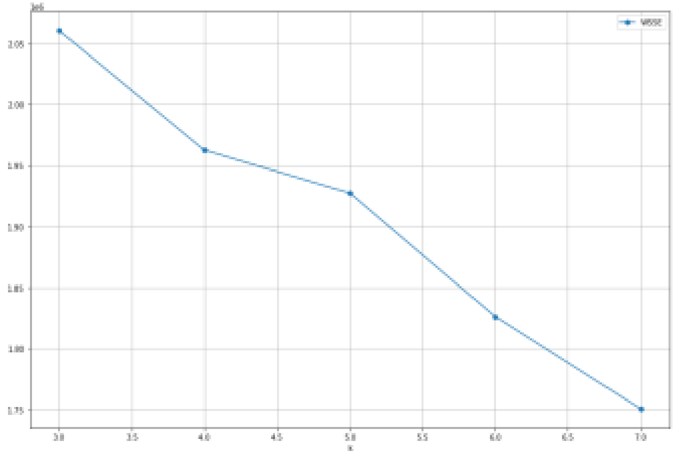
To put it simple Elbow method is an iterative approach to tune the hyper parameter (K in Kmeans) in other words to decide the number of clusters to form for our weather data.

FIG: weather data Elbow method



```
Trying k from 3 to 7 with 1089260 samples...

Training for cluster size 3
......................WSSE = 2060450.3112377746
Training for cluster size 4
......................WSSE = 1962553.296385016
Training for cluster size 5
......................WSSE = 1927460.3446945003
Training for cluster size 6
......................WSSE = 1826313.494459609
Training for cluster size 7
......................WSSE = 1750738.996612914
    1e6
```

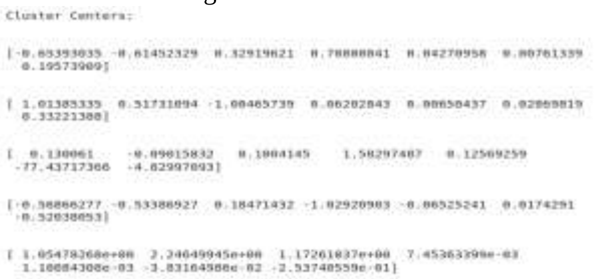


Interpretation:

For the weather data across 120 countries to figure out the perfect number of clusters elbow method is performed and displayed above, from the graph we could state that K=5 looks similar to the elbow and we can further continue to build our K-means model with k=5.

3.2 K-means Model

Fig: Cluster centres



```
Cluster Centers:

[-0.65393035 -0.61452329  0.32919621  0.78888841  0.84270958  0.00761339
  0.19573909]

[ 1.01383335  0.51731094 -1.00465739  0.06282843  0.00658437  0.02869819
  0.33221388]

[  0.130061    -0.09815832   0.1004145    1.58297487   0.12569259
 -77.43717366  -4.82997093]

[-0.38866277 -0.53386927  0.18471432 -1.02920903 -0.06525241  0.0174291
 -0.52038053]

[ 1.05478268e+00  2.24649945e+00  1.17261837e+00  7.45363399e-03
  1.10084308e-03 -3.83164988e-02 -2.53740559e-01]
```

Interpretation

Kmeans clustering model has been fitted for the weather data, the above figure represents the centroid of 8 features in 5 clusters and the visualization part will be shown in later part.

IV. CONCLUSION

The weather is the import factor from ancient times, most of the events from A-Z are completely depends on how the weather is, Big data has become a common term now a days with the development of IoT the 4 v's are evolving rapidly so, tools like Hadoop is gaining popularities nowadays and here the weather data across 120 countries and day-wise data for 20 years (around 3 million records) are loaded in spark environment and the data is cleaned, removed duplicate values and null values then proper statistical techniques are implemented to extract the appropriate features and for the finalized data K means Clustering model has been fitted and the similar cities with similar weather conditions has been clustered and finally the in-depth insights are gained by visualizing the resultant weather data using power bi.

REFERENCE

[1]. Radhika Y, Shashi M. Atmospheric temperature prediction using support vector machines. International Journal of Computer Theory and Engineering. 2009 Apr; 1(1):1793– 8201.

[2]. Brenner I.S. (1986). "Biases in MOS (Model Output Statistics) Forecasts of Maximum and Minimum Temperatures in Phoenix, Arizona", Weather and Forecasting, Vol. 1(3), 226–229.

[3]. Brunet M., Sigro J., Jones P.D., Saladie O., Aguilar, Moberg A., Della-Marta P.M., Lister D., Walter A. (2007). "Annual and Seasonal changes in the distribution of daily maximum and minimum temperature data in temperatureextreme indices thoughout the 1901–2005 period over mainland Spain.", Geophysical Research Abstracts 9, 07167.

[4]. Driscoll D.M. (1988). “A Comparison of Temperature and Precipitation forecasts issued by Telecasters and National Weather Service”, Weather and Forecasting, Vol.3(4), 285–295.

[5]. Gyakum J.R. (1986). “Experiments in Temperature and Precipitation Forecasting in Illinois”, Weather and Forecasting, Vol. 1(1), 77–88.

[6]. Reading C. (2004). “Student Description of Variation While Working with Weather Data”, Statistics Education Research Journal, Vol. 3(2), 84–105.

[7]. Stone, Weaver (2002), “Daily maximum and minimum temperature trends in a climate model”, Geophysical Research Letters, Vol. 29(9), 70–71.

[8]. Serra C., Burgueno A., Lana X. (2001). “Analysis of Maximum and Minimum daily temperatures recorded at Fabra Observatory Barcelona NE Spain in the period 1917–1998”, International Journal of Climatology, Vol. 21, 617–636.

[9]. S.Chakraborty and N.K.Nagwani ,“Performance evaluation of incremental K-means clustering algorithm ”, IFRSA International Journal of Data Warehousing & Mining (IIJDWM), vol.1, 2011,pp-54-59.

[10]. Chakraborty and Nagwani, S. and N.K.. , “Analysis and Study of Incremental K-Means Clustering Algorithm”, accepted by International conference on high performance architecture and grid computing (HPAGC) sponsored by Springer Germany, Punjab (India), 2011.

[11]. Chouksey P., Chauhan A., “A Review of Weather Data Analytics using Big Data”, IJARCCE, ISSN: 2278-1021 Volume-06, Issue01, Page No (365- 368), January, 2017.

[12]. Riyaz P.A., Surekha M.V., “Leveraging MapReduce With Hadoop for Weather Data Analytics” IOSR Journal of Computer Engineering, Volume 17, Issue 03.