THE OVERVIEW: BIG DATA ANALYTICS EXPENDING WITH PYTHON TO BE ACCOMPLISH BY USING MACHINE LEARNING

Author 1: Vishwa Priya V Research Scholar,
Department of Computer science Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India.
Email- vishwapriya13@gmail.com
Author 2: Dr. R. Renuga Devi Assistant Professor,
Department of Computer science Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India.
Email- nicrdevi@gmail.com

ABSTRACT

In a real-time environment, big data analytics occupies many fields in its manipulation. In a wide-area network, major data can be stored, operated, controlled, managed, and maintained through big data. Python API can run machine learning algorithms to develop big data analytics. In various organizations, the entire information is retrieved from data warehousing. Machine learning helps to increase the security level, high accuracy, prediction, suitable invention, and which generates the applications via big data analytics. Machine learning is the best way to handle various approaches to evaluating system domain software for the customer's needs. When python introduces big data analytics and machine learning, python gives the best outcomes throughout data visualization, data analysis, and summarizing the data from the dashboard after being compared with existing methods. Machine learning patterns are combined with real-world environments. When machine learning, big data analytics, and python link to improve different kinds of information can easily be handled by different types of people with secured electronic data to be distributed with the best accuracy and utilized with the least possible time.

Keywords: *Big data, Big Data Analytics, Machine Learning, Python.*

I) INTRODUCTION

In the historical period, they have written their data on palm leaves, manuscripts, handbooks, etc. for future reference to know their life history. These histories are known as traditional data from the historical periods of data are implemented.

In real-time, big data can store the information in a huge set of data which is to be taken for backup that will be used on the internet and it will help the customer's needs. It will be virtually machine learning to operate the process of big data applications and the tools to perform them. Big data, software tools are more helpful to access in the real-time process by using the application interface. In big data analytics, it is mainly used in machine learning, visualization, analysis, process, and extraction of the database.

Information from the data is extremely complex in businesses and they can be used in advanced analytic techniques can be used here.

II) LITERATURE SURVEY

S. No	Author	Work	Method & Tools	Discussion & Drawbacks
1.	Pittaras, N., Papadakis, G.,	Geo-Sensor: Semantifying Change and Event Detection over Big Data	Supervised learning, CNN, HDFS, Ground Center Point (GCP), and Detection controls.	RDF plays a vital role in geo-sensors for visualization and as a remote sensor, but it is not easy to access and handle all the tools to manipulate it [1].
2.	N Alex; I Daniel, Ben T S.,	Systematic Review of the Literature on Big Data in the Transportation Domain: Concepts and Applications	Survey about DOT (Dept., of Transport) about traffic system.	The transport sector is managing the data by roadside sensors for the alert messages to prevention for environmental security needs, but it is difficult to maintain in urban areas [2].
3.	Gan, Y., Zhang, Y., Hu, K., Cheng, D.,	Seer: Leveraging Big Data to Navigate the Complexity of Performance Debugging in Cloud Micro-Services	Deep Neural Network, CNN, LSTM, softmax	End-to-End services with Qos with deep knowledge application use for limited users [3].
4.	A Anju and K Shyma.,	Security and Clustering Of Big Data in Map-Reduce Framework: A Survey	Survey on K-means, KNN, DBSCAN clustering, using Map Reduce	Privacy, security using cryptography with proper usage of authorized information using Map Reduce framework but time complexity is more in it [4].
5.	V.M.P.karan, P.Devatharini	Building a Big Data provenance with its applications for Smart cities	A smart city survey of big data analytics	In various sectors, big data plays a vital role in innovations [5].

III) CHARACTERISTICS OF BIG DATA ANALYTICS

Big Data Analytics means, Big→huge, Data→information, Analytics→ statistical & logical reasoning. Big data analytics means “Huge set of information to be analysis with the proper manner to be implemented for the given dataset” is known as big data analytics.

i) BIG DATA

Big data provides functions such as storing, retrieving, updating, and enhancing the data in a massive set of Variations to enlarge Volume by streaming the data with the peak of Velocity to be performed. It is a huge size with a more complex data set that can be handled here. Big data uses different kinds of V’s to improve the level of analysis.

Big data is not a new invention. It is related to before the civilization period itself. It is related to daily life. For example: From the Paleolithic period, which means the old Stone Age period, people handling the big data concept starts there itself. When humans started their life by building shelter, collecting their food, different styles of clothes, agriculture, shipping goods, exchanging a product, storage, trading, medicine to be provided by prediction manner, warehousing, etc. these are some of them related to real-time data; this could be the best example for big data. This information will be followed by the next generation to implement and follow as well to carry it forward. This information is given by sculptures, treasure hunts, caves diagrams, and so on. These are the huge sets of data; those data that are real-time are called historical data. The historical period of data is a reference to traditional data to be implemented in a real-time environment; nowadays this information is preserved through an excavation research center via the cloud network. So, further information to the public can. They can access the data easily, gaining knowledge about human evolution and mortalities by using the search engine.

Traditional data which is maintained till now by storing in a particular area is called data warehousing. Data warehousing will store a huge size of data with the help of big data. For example, overall organization details will be placed here.

ii) BIG DATA ANALYTICAL

In real-time, surviving our life nowadays become so easy and smooth compared with the golden period because they cultivate all on their own but currently, our technology plays a vital role in human life called Artificial Intelligence (AI); Machine learning, cloud computing, Internet of Things (IoT), etc.

Big Data Analytics (BDA) is used to analyze a huge set of data by using advanced analytical techniques to be pursuing. Here, they used 3 various data types called structured, semi-structured, and unstructured data to perform the calculations. These 3 various data types can be performed from different sources from a different set of sizes. Hereby these sizes usually begin from terabyte to zettabyte.

BDA techniques include testing the data, analyzing and integrating the data; Data Mining (DM) will extract the patterns from the enormous datasets. Machine Learning (ML) is used to implement data analysis; Natural Language Processing (NLP) is used to implement the toolto-be analysis method to help the human natural language; Statistics: is mainly used to survey, experiment, and maintain for future use. Here, some additional techniques are included, namely spatial analysis, predictive analysis, network analysis, and numerous analyses will be performed by BDA.

iii) Applications in BDA

In BDA, it is used in various sectors begins from small-scale industries to large-scale industries. Examples are given below,

In *Banking*, fraud can be detected and exchanged data in a secured manner. In *Communication*, they can interchange the information (or) data sent person to person in various sectors through different ways. In *Media*, data can be passed through network topology and communication channels. For example, video size, mail size, radiofrequency, etc. In *Entertainment* through websites, DTH satellites, services can be amazon prime, amazon web services, etc. On the *Education* side, the overall institution details will be maintained by data warehousing and overall student details will be maintained by data mart. In the *Government* sector, the current status of the population and financial market analysis will be maintained here. In *Agriculture* sector will handle the climatic changes, soil, frame management, cropping, livestock, land, insurance, water facility, storage information, fertilizer, and chemical seeds.

IoT (Internet of Things) is more used, placed in the transport sector, smart grid, and so on. In *Transport* they provide traffic system properly, the traffic pattern is unique; collisions will be avoided; road safety will be in a proper manner by using an alarm system, traffic jams will be reduced. *Insurance* will be generated through digital channels, social media, and real-time monitoring by the organization. *Health care* providers will maintain the electronic health report for patient details for future needs; they will visualize the data, perform predictive analytics, and the clinical report will be generated using BDA.

BDA plays a vital role in various sectors, nowadays BDA is the most important supporting factor in vast industries. Big data is used in various devices such as data from black boxes is used in airplanes, helicopters, etc. Big data using in social media by Facebook, Twitter, etc. Big data in the stock market is used for the buyer and the seller to interchange massive sets of documents sent through proper email id, organization id, etc., by these ways BDA is more popular in various sectors.

IV) MANIPULATING ML METHODS

i) Machine Learning

Machine learning (ML) is one of the most important criteria for the students to know how to write the algorithms with their practices and knowledge by using the given data. ML is a segment of Artificial Intelligence [7]. ML helps to strengthen and develop the Data Science sectors. In ML the most important word is called training data; this data gives proper decision-making as well as helps to predict precisely by using different kinds of applications. In ML, basic knowledge of awareness about algorithms helps to identify the problem in various sources of learning; then they can rectify the problem to get better results for the given data. Break the procedures into hunks. Begin with the sample dataset by using existing data and authorization with the right-handed application, then start writing a code for the process to be performed.

ii) ML methods designed for BDA

ML method is used to construct the models for the data toward training by various algorithms to make proper decision making. The ML approach is mainly used for scientific, mathematical, and statistics. In ML, there are 3 approaches of learning used in Big Data Analytics: Supervised learning, unsupervised learning, and Semi-Supervised Learning.

a) *Supervised Learning (SL):* In SL, it follows a set of rules and procedures that should be manipulated by both the input (t) and output (v) data functions. The function called f(t) will be mapped using the SL method equation called by $v = f(t)$ will be performed. For example, in a company, a manager will be supervising the entire team activity is called SL. In SL, the learning phase provides the training data that will generate the report to track the progress of the work to be done and follow the steps by using certain simulations. Types of SL are Classification and Regression. In SL, the working process should take the raw input data that could be well training data then it sends to the procedural activity which will be applied for the further process that will be executed output for the given data. In SL, Generates only labeled data by proper input and output pattern through training data; In SL, Classes are known and well defined, real-time learning methods can also handle in offline mode, high-level of accuracy with trustworthy results; simple methods can be generated easily in SL method.

b) *Unsupervised Learning (UL):* In UL it takes only inputs from the dataset and will determine the structure of the data. UL gives only input data for the current situation and no consistent data will be provided for the output; this is such an exciting factor in UL. For example a baby and a baby family cat. The baby will recognize cat characteristics like nose, legs, tail, etc. where they are not taught by anyone, but they learn from mind data by observing is called Unsupervised Learning. In UL, the working process generates the raw input data will analyze the procedure which will process through execution for output. Types of UL are Clustering and Association. In UL, mainly concentrate on self-learning method, the data are unlabelled, the classes are unknown, less accuracy but trustworthy results, learning method can be implemented in a real-time environment; data output is not available; processing will be a little bit complicated in UL method.

c) *Semi-Supervised Learning (SSL*): SSL lies between SL and UL. It reduces the human workload. It will handle both known and unknown data. They contain 3 types of supposition is continuity will find the nearest value and replace the new label value; the cluster will divide and split the label value; the manifold will generate the estimated value in SSL [11].

iii) Techniques in ML used for BDA

In ML there are various approaches are:

Regression will predict the value by using strategy to analyze with the help of statistical format. *Clustering* forces to divide the data for the performance determined with the same group of data, predict the unknown data, and collect the uncharacterized pattern from the data. *Classification* is used to merge the output toward the class. When it comes to input, the label will be classified into two types of prediction; they are: Binary prediction and Multi-class prediction. The *Decision Tree* spirit helps to take proper decision making. It should be in a procedural format to execute the process to choose suitable decisions for the current situation. It will execute branch by branch and output will be performed.

Support Vector Machines (SVM) in this technique mainly focus on the hyperplane. The hyperplane is used to identify the approximate value which holds the proper dataset. It gives an optimal solution. In SVM they calculate the margin value by using different viewpoints. A binary approach is the best way to handle SVM. *Naïve Bayes* is used to predict the data in the dataset and improve the accuracy, reduce the cost, reduce the time complexity.

K- Nearest Neighbor (KNN) new data desire to identify and point to the unknown new data to be predicted and rectified by using KNN. It can be classified by examining the data to determine the distance and identify the nearest neighbor through elected for the label. *Gradient Boosting Classification (GBC)* helps to predict the error rate of the problem. GBC will merge all uncertain learning approaches combined with a perfect predictive approach etc., these are some of the algorithms used in various sectors.

iv) ML process in BDA

ML will be overwhelmed by the previous knowledge to face the current situation and take the clever decision-making for the BDA protection to secure the data from robust. ML using BDA will reduce the cost factor, reduce the time complexity, predict components to be improved, services are based on patterns, and high accuracy compared with existing methods.

Phases for the BDA techniques:

1. Describe the problematic situation with a proper explanation. Find the internal and external data sources.
2. Accumulate and choose the proper information. Training data helps to get the proper production results.
3. Fabricate the mathematical formula. Estimate the formula. Implement all formulas together.
4. Manipulate the formulas for the resolved situation by using the above steps.

V) IMPLEMENT THROUGH PYTHON API

Python is simple to approach the object-oriented programming (OOP) concepts with the help of the Application Programming Interface (API). In python, they can access many libraries using ML. ML algorithms are used to develop a python framework to reduce the time complexity as well they can predict the unknown and unlabelled classes by using various BDA techniques. BDA helps to enlarge the various V's concept to expand the methods for the given customer's information.

i)Python

Python means "huge size of snake it will be squashing large size of living thing it is also familiar snake known as ANACONDA". Python in a computer term "High-level programming language it can be generated by anaconda distribution with pip package by data science".

Python helps to create the data, retrieve the information via possible ways, and analyze the given raw information to get the proper output for the customer's need. Python will overhaul all OOP concepts and it is an open-source platform so all customers can easily learn the procedure; easy to install, packages are more to explore in various applications under a single language is called python. Programs will be saved by the ".py" extension using API. Python is a bridge to BDA using HDFS (Hadoop Distributed File System).

HDFS

It contains a large set of data to handle the Map-reduce concept for a framework to reduce the work process for the customers.

Step 1: *Mapping:* it takes raw data as input to format the given data to be splitting as various input data for a read to map the midway key will combine and reduce the repeated data as per the process of the mapping.

Step 2: *Reducing:* it shuffle, sort, and merge the data to reduce and write HDFS output for the data.

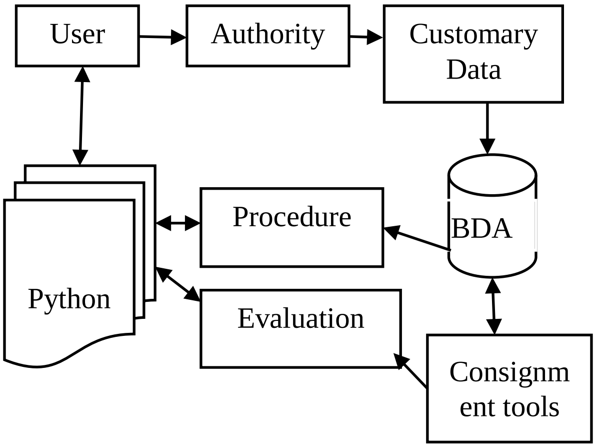
ii) Merge ML, BDA, Python API

BDA←→ML ←→ Python ←→BDA

Big data will analyze, extract and it contains a huge set of V's. In BDA contains structure, unstructured and semi-structured are implemented from the ML algorithms (MLA). ML can predict unknown data by using various algorithms. Big data tools contain Hadoop, Mangodb, HBase, Hive, Pig, etc. extracts by using various ML algorithms. Python using ML tools to be directed and implemented use in NumPy, pandas, scikit, etc. by writing coding for the client's expectation.

BDA implements raw data to glance for patterns to support and make convincing decision formulating for the organization. ML study begins from training the data to make proper decision making with the help of real-time prediction learned by the algorithms.

VI) ARCHITECTURE FOR BDA EMBEDDED ML IMPORTING PYTHON API



Discussion

The user expectation will be easy to learn about working process and model should be advanced transformation compared with existing methods. Users want to select the software based on client expectations: which software, platform, application, and algorithms for the client’s belief [14].

User request details are sent to the authorized person as well as a programming team. The maintenance team will search the related data in customary data. Customary data will filter and send the necessary information to BDA. BDA will send read and write process can be done with a huge set of database using ML streaming procedure. In streaming, it will train data with proper processing techniques to be established with the help of customary data. The procedure will manipulate prediction, GCC, SVM, Naïve Bayes, Neural networks, etc. The ML stream executes the MLA and sends the process to python (API).

BDA holds the HDFS to execute the consignment tools handle with classification, clustering, regression, association rules, language known also using here; some of them are R, weka, etc. these tools can be batch by ML tools. Batching it may contain cyclic or non-cyclic with upgraded processing with ML tools and methods. In evaluation generates the dashboard concept is established by various sections are data vision, data investigation, and data review. After using various tools it will move to the next stage.

In python, it will handle both the procedural and evaluation processes to implement in their work progress for the user need. The overall process will execute for visualization process in API.

VII) CONCLUSION

BDA will be more secure, extract the features, store the huge size of data, and retrieve various kinds of data coupled with many motions to be performed here. BDA contains different types of V’s to be performed here; some of them are volume, velocity, etc [12]. BDA will be enlarged in statistics, numeric, analysis, classification, etc. are coming under 3 approaches of learning are SL, SSL, and UL. This learning is an import from ML. ML constructs a procedure to reduce the complexity, reduce time complexity, increase accuracy, prediction, decision making, and many other ways to improve the quality of the product by MLA. In python, they use many library functions that are imported by MLA. So, here conclude by using BDA entrenched with MLA introduction with Python is the best solution maker to get the best quality fabricator growth of development in several zones.

VIII) REFERENCE

1. Pittaras, N., Papadakis, G., Stamoulis, G., Argyriou, G., Taniskidou, E. K., Thanos, E., Koubarakis, (2019), "GeoSensor: Semantifying Change and Event Detection over Big Data" GeoSensor. Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing - SAC '19. doi:10.1145/3297280.3297504.
2. Neilson, Alex; Indratmo, ; Daniel, Ben; Tjandra, Stevanus, (2019), "Systematic Review of the Literature on Big Data in the Transportation Domain: Concepts and Applications". Big Data Research, published in Elsevier, volume 17, Pages 35-44, doi:10.1016/j.bdr.2019.03.001.
3. Gan, Y., Zhang, Y., Hu, K., Cheng, D., He, Y., Pancholi, M., & Delimitrou, C. (2019). "Seer: Leveraging Big Data to Navigate the Complexity of Performance Debugging in Cloud Microservices" Seer. Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS '19. doi:10.1145/3297858.3304004.
4. A Anju, K Shyma,(2018) "Security and Clustering Of Big Data in Map Reduce Framework: A Survey" International Journal of Advance Research, Ideas and Innovations in Technology,ISSN: 2454-132X ,(Volume 4, Issue 1).
5. V.M. P. karan, P.Devatharini, L.D. kumar, K.Prasanth,(2018), "Building a Big data provenance with its applications for Smart cities" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 , Volume: 05 Issue: 03, p-ISSN: 2395-0072.
6. M. M. Najafabadi, F. Villanustre, T. M. K. Goftaar, N.Seliya, R. Wald, and E. M. Magic, (2015), "Deep learning applications and challenges in big data analytics", Journal of Big Data, vol. 2, pp. 1- 21, doi: 10.1186/s40537-014-0007-7.
7. L. C. R. D. Sharma, (2020) "Python Tools for Big Data Analytics", International Journal of Science and Research (IJSR) ISSN: 2319-7064 Research Gate, Volume 9 Issue 5, pp-597- 602, doi: 10.21275/SR20507222308.
8. K. S Divya, P.Bhargavi , S. Jyothi, (2018) , "Machine Learning Algorithms in Big data Analytics" International Journal of Computer Sciences and Engineering , Volume-6, Issue-1 E-ISSN: 2347-2693, pp- 63-70, doi:10.26438/ijcse/v6i1.6370.
9. Jorge A. Delgado, Nicholas M. Short, Daniel P. Roberts, Bruce Vandenberg (2019) "Big Data Analysis for Sustainable Agriculture on a Geospatial Cloud Framework" Frontiers in Sustainable Food Systems, ISSN: 2571-581X, doi:10.3389/fsufs.2019.00054.

10. A. Belle, R. Thiagarajan, S.M.R. Soroushmehr, F. Navidi, DA.Beard, (2015), “Big Data Analytics in Healthcare”, Hindawi Publishing Corporation BioMed Research International Volume 2015, Article ID 370194, 16 pages, doi.org/10.1155/2015/370194+
11. G. Zhang, S.-X. Ou, Y.-H. Huang, and C.-R. Wang, (2015), "Semi-supervised learning methods for large scale healthcare data analysis," International Journal of Computers in Healthcare, vol. 2, pp. 98-110.
12. A. Sanghi, R. Sood, J.Haritsa, S Tirthapura (2018) , “Scalable and Dynamic Regeneration of Big Data Volumes” Published in Proceedings of the 21st International Conference on Extending Database Technology (EDBT), ISSN:2367-2005 on Open Proceedings doi:10.5441/002/edit.2018.27.
13. J L. LUSK, (2017), “Consumer Research with Big Data: Applications from The Food Demand Survey (Foods)” Published by Oxford University of the Agricultural and Applied Economics Association., Volume 99, Issue 2, pp- 303-320, doi:org/10.1093/ajae/aaw110.
14. A Baldominos, Esperanza Albacete, Yago Saez and Pedro Isasi ,(2014) “A Scalable Machine Learning Online Service for Big Data Real-Time Analysis” IEEE Symposium on Computational Intelligence in Big Data (CIBD)**,** doi: 10.1109/CIBD.2014.7011537.