IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Application Of Cloud Computing For Data Wrangling In Python Programming And Machine Learning

[1]Mohd Shahnawaz, [2]Arun Saini, [3] Anup Kumar, [4]Kuldeep Chauhan

[1]Assistant Professor, [2] Assistant Professor, [3]Assiatant Professor, [4]Assiatant Professor
[1, 2, 3, 4] Dept. of computer science,
[1, 2, 3, 4] Shobhit University Gangoh, Saharanpur, India.

Abstract: Python is an interpretable and scriptable programming language that can be used for both learning and practical applications. Guido van Rossum created the potent high-level language python. It is an interpretable programming language that is object-oriented. The primary Python programming software tools for data wrangling, cloud computing, and machine learning approaches will be introduced in this presentation. In summary, this paper will begin with an introduction to Python programming and data wrangling. It will also include an overview of cloud computing, machine learning, and data wrangling. Finally, it will discuss popular packages used in the data wrangling and machine learning fields, including NumPy, SciPy, Tensor Flow, Keras, Matplotlib, and others. We will then go on to demonstrate the value of Python in developing cloud computing and data wrangling apps.

***Index Terms* -** Machine learning · Data Wrangling- Tools · Languages · Python, Cloud computing

1. Introduction to data wrangling

Data wrangling, also known as data munging, is an essential step in the data analytics pipeline that entails preparing raw data for analysis by cleaning, organizing, and enriching it. As part of this procedure, the data is cleaned by eliminating or fixing errors, inconsistencies, and duplication [1]. Additionally, it entails organizing the data, frequently putting it in a tabular format that makes analytical applications easier to deal with another crucial stage is enriching the data, which involves adding new information to increase its analytical value and having it verified to guarantee its quality and accuracy. Analysts and data scientists can obtain important insights more quickly and precisely by using data wrangling to make raw data more understandable and accessible [2].

1.1 How Data Wrangling Works?

The act of transforming unprocessed data into a format suitable for analysis is known as data wrangling, and it involves multiple crucial processes [2]. This shift is essential for revealing insightful information that affects strategic planning and decision-making. This is a thorough explanation of data wrangling's operation:

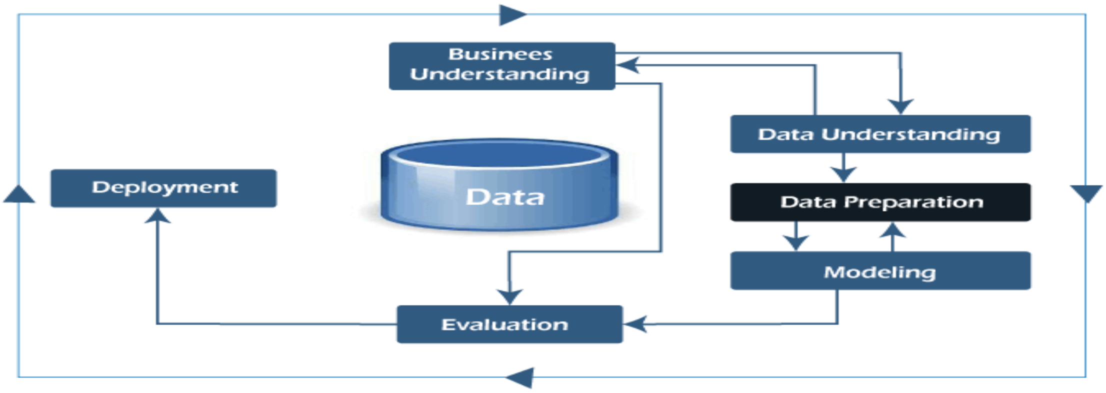

Figure 1. Process and operation of data wrangling.

(a) Collection

Gathering raw data from diverse sources is the initial stage in the data wrangling process. Databases, files, external APIs, web scraping, and numerous other data streams are examples of these sources. The information gathered may be semi-structured (JSON, XML files), unstructured (text documents, photos), or structured (SQL databases, for example) [3].

(b) **Cleaning**

After the data is gathered, the cleaning procedure starts. Errors, inconsistencies, and duplicates that could distort analysis results are eliminated in this step. Cleaning may entail: deleting information that isn't useful or adds nothing to the analysis. Fixing data mistakes, like misspelling so in accurate values [4, 5]. Addressing missing values by deleting them, assigning them to other data points, or making statistical approximations of them. Recognizing and fixing discrepancies, like disparities in date or currency forms [4].

(c) Structuring

After cleaning, data needs to be structured or restructured into a more analysis-friendly format. This often means converting unstructured or semi-structured data into a structured form, like a table in a database or a CSV file [5]. This step may involve:

i. Parsing data into structured fields.
ii. Normalizing data to ensure consistent formats and units.
iii. Transforming data, such as converting text to lowercase, to prepare for analysis.

(d) Enriching

Data enrichment involves adding context or new information to the dataset to make it more valuable for analysis. This can include:

Merging data from multiple sources to develop a more comprehensive dataset. Creating new variables or features that can provide additional insights when analyzed [1, 3].

(e) Validating

Validation ensures the data's accuracy and quality after it has been cleaned, structured, and enriched. This step may involve:

Data integrity checks, such as ensuring foreign keys in a database match. Quality assurance testing to ensure the data meets predefined standards and rules [4].

(f) Storing

The final wrangled data is then stored in a data repository, such as a database or a data warehouse, making it accessible for analysis and reporting. This storage not only secures the data but also organizes it in a way that is efficient for querying and analysis [6].

(i) Documentation

Documentation is critical throughout the data wrangling process. It records what was done to the data, including the transformations and decisions. This documentation is invaluable for reproducibility, auditing, and understanding the data analysis process [4, 5, 6].

1.2 Introduction to Python Language

Python is a widely used programming language that offers several unique features and advantages compared to languages like Java and C++. Our Python tutorial thoroughly explains Python basics and advanced concepts, starting with installation, conditional statements, loops, built-in data structures [2] , Object-Oriented Programming, Generators, Exception Handling, Python RegEx, and many other concepts. This tutorial is designed for beginners and working professionals.

In the late 1980s, Guido van Rossum dreamed of developing Python. The first version of Python 0.9.0 was released in 1991. Since its release, Python started gaining popularity [7]. According to reports, Python is now the most popular programming language among developers because of its high demands in the tech realm.

(a)It is used Python Programming language

- Web development (server-side),
- Software development,
- Mathematics,
- System scripting.

(b)What can Python do?

- Python can be used on a server to create web applications.
- Python can be used alongside software to create workflows.
- Python can connect to database systems. It can also read and modify files.
- Python can be used to handle big data and perform complex mathematics.
- Python can be used for rapid prototyping, or for production-ready software development.

1.3 Introduction to Machine learning

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of Machine Learning [6].
A subset of artificial intelligence known as machine learning focuses primarily on the creation of algorithms that enable a computer to independently learn from data and previous experiences. Arthur Samuel first used the term "machine learning" in 1959 [8]. It could be summarized as follows:
Without being explicitly programmed, machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things [9].Machine learning algorithms create a mathematical model that, without being explicitly programmed, aids in making predictions or decisions with the assistance of sample historical data, or training data. For the purpose of developing predictive models, machine learning brings together statistics and computer science [5]. Algorithms that learn from historical data are either constructed or utilized in machine learning. The performance will rise in proportion to the quantity of information we provide.

1.3.1 How does Machine Learning work?

A machine learning system builds prediction models, learns from previous data, and predicts the output of new data whenever it receives it. The amount of data helps to build a better model that accurately predicts the output, which in turn affects the accuracy of the predicted output [10].

Let's say we have a complex problem in which we need to make predictions. Instead of writing code, we just need to feed the data to generic algorithms, which build the logic based on the data and predict the output. Our perspective on the issue has changed as a result of machine learning [10].

The Machine Learning algorithm's operation is depicted in the following block diagram:

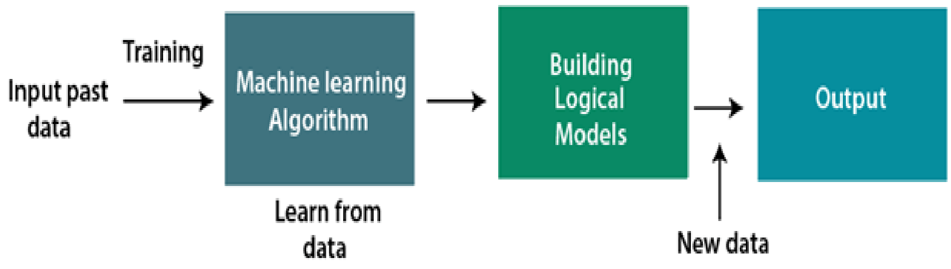

Figure 2. Basic data processing in ML.

1.4 Introduction to Cloud Computing

Cloud Computing is the delivery of computing services such as servers, storage, databases, networking, software, analytics, intelligence, and more, over the Cloud (Internet).Cloud Computing provides an alternative to the on-premises datacenter. With an on-premises datacenter, we have to manage everything, such as purchasing and installing hardware, virtualization, installing the operating system, and any other required applications, setting up the network, configuring the firewall, and setting up storage for data. After doing all the set-up, we become responsible for maintaining it through its entire lifecycle [11][12].But if we choose Cloud Computing, a cloud vendor is responsible for the hardware purchase and maintenance. They also provide a wide variety of software and platform as a service. We can take any required services on rent. The cloud computing services will be charged based on usage.

1.4.1 Advantages of cloud computing

- Cost: It reduces the huge capital costs of buying hardware and software.
- Speed: Resources can be accessed in minutes, typically within a few clicks.
- Scalability: We can increase or decrease the requirement of resources according to the business requirements.
- Productivity: While using cloud computing, we put less operational effort. We do not need to apply patching, as well as no need to maintain hardware and software. So, in this way, the IT team can be more productive and focus on achieving business goals.

- Reliability: Backup and recovery of data are less expensive and very fast for business continuity.
- Security: Many cloud vendors offer a broad set of policies, technologies, and controls that strengthen our data security.

1.4.2 Types of Cloud Computing

- **Public Cloud:** The cloud resources that are owned and operated by a third-party cloud service provider are termed as public clouds. It delivers computing resources such as servers, software, and storage over the internet.
- **Private Cloud:** The cloud computing resources that are exclusively used inside a single business or organization are termed as a private cloud. A private cloud may physically be located on the company’s on-site datacenter or hosted by a third-party service provider.
- **Hybrid Cloud:** It is the combination of public and private clouds, which is bounded together by technology that allows data applications to be shared between them. Hybrid cloud provides flexibility and more deployment options to the business.

1.4.2 Types of Cloud Services

Infrastructure as a Service (IaaS): In IaaS, we can rent IT infrastructures like servers and virtual machines (VMs) [11] , storage, networks, operating systems from a cloud service vendor. We can create VM running Windows or Linux and install anything we want on it. Using IaaS, we don’t need to care about the hardware or virtualization software, but other than that, we do have to manage everything else [13]. Using IaaS, we get maximum flexibility, but still, we need to put more effort into maintenance. Platform as a Service (PaaS): This service provides an on-demand environment for developing, testing, delivering, and managing software applications. The developer is responsible for the application, and the PaaS vendor provides the ability to deploy and run it. Using PaaS, the flexibility gets reduce, but the management of the environment is taken care of by the cloud vendors. Software as a Service (SaaS): It provides a centrally hosted and managed software services to the end-users. It delivers software over the internet, on-demand, and typically on a subscription basis. E.g., Microsoft One Drive, Dropbox, WordPress, Office 365, and Amazon Kindle. SaaS is used to minimize the operational cost to the maximum extent [5] [6].

2 Objective of study

1. To identify the features of Python Programming.
2. To investigate python modules for Data Wrangling like NumPy which is used for matrix and vector.
3. To focus on python modules for Machine learning like Tensor flow numerical computations for machine learning, Key areas for neural networks and machine learning.

3 Related work

3.1 Basic features of data wrangling.

Any analyses a business performs will ultimately be constrained by the data that informs them. If data is incomplete, unreliable, or faulty, then analyses will be to diminishing the value of any critical insights gleaned [11]. Data wrangling seeks to remove that risk by ensuring data is in a reliable state before it's analyzed and leveraged. This makes it a critical part of the analytical process. It's important to note that data wrangling can be time-consuming and taxing on resources, particularly when done manually. This is why many organizations institute policies and best practices that help employees streamline the data cleanup process—for example, requiring that data include certain information or be in a specific format before it's uploaded to a database [12]. For this reason, it's vital to understand the steps of the data wrangling process and the negative outcomes associated with incorrect or faulty data.

3.1 Basic features of Python

Python provides many useful features which make it popular and valuable from the other programming languages. It supports object-oriented programming, procedural programming approaches and provides dynamic memory allocation [11].

We have listed below a few essential features.

a. **Easy to Learn and Use --**Python is easy to learn as compared to other programming languages. Its syntax is straightforward and much the same as the English language. There is no use of the semicolon or curly-bracket, the indentation defines the code block. It is the recommended programming language for beginners.

b. **Expressive Language ---**Python can perform complex tasks using a few lines of code. A simple example, the hello world program you simply type print ("Hello World"). It will take only one line to execute, while Java or C takes multiple lines.

c. **Interpreted Language --** Python is an interpreted language; it means the Python program is executed one line at a time. The advantage of being interpreted language, it makes debugging easy and portable.

d. **Cross-platform Language --** Python can run equally on different platforms such as Windows, Linux, UNIX, and Macintosh, etc. So, we can say that Python is a portable language. It enables programmers to develop the software for several competing platforms by writing a program only once.

e. **Free and Open Source ---** Python is freely available for everyone. It is freely available on its official website www.python.org. It has a large community across the world that is dedicatedly working towards make new python modules and functions. Anyone can contribute to the Python community. The open-source means, "Anyone can download its source code without paying any penny."

f. **Object-Oriented Language ---** Python supports object-oriented language and concepts of classes and objects come into existence. It supports inheritance, polymorphism, and encapsulation, etc. The object-oriented procedure helps to programmer to write reusable code and develop applications in less code.

g. **Extensible- --**It implies that other languages such as C/C++ can be used to compile the code and thus it can be used further in our Python code. It converts the program into byte code, and any platform can use that byte code.

h. **Large Standard Library --** It provides a vast range of libraries for the various fields such as machine learning, web developer, and also for the scripting. There are various machine learning libraries, such as Tensor flow, Pandas, NumPy, Keras, and Pytorch, etc. Django, flask, pyramids are the popular framework for Python web development.

i. **GUI Programming Support --** Graphical User Interface is used for the developing Desktop application. PyQT5, Tkinter, Kivy are the libraries which are used for developing the web application.

j. **Integrated --** It can be easily integrated with languages like C, C++, and JAVA, etc. Python runs code line by line like C, C++ Java. It makes easy to debug the code.

k. **Embeddable** -- The code of the other programming language can use in the Python source code. We can use Python source code in another programming language as well. It can embed other language into our code.

l. **Dynamic Memory Allocation --**In Python, we don't need to specify the data-type of the variable. When we assign some value to the variable, it automatically allocates the memory to the variable at run time. Suppose we are assigned integer value 15 to x, then we don't need to write int x = 15. Just write x = 15.

4. Basic features of Machine Learning

i. **Feature Creation:** Feature creation is finding the most useful variables to be used in a predictive model. The process is subjective, and it requires human creativity and intervention [6]. The new features are created by mixing existing features using addition, subtraction, and ration, and these new features have great flexibility [9].

ii. **Transformations:** The transformation step of feature engineering involves adjusting the predictor variable to improve the accuracy and performance of the model. For example, it ensures that the model is flexible to take input of the variety of data; it ensures that all the variables are on the same scale, making the model easier to understand [5]. It improves the model's accuracy and ensures that all the features are within the acceptable range to avoid any computational error.

iii. **Feature Extraction:** Feature extraction is an automated feature engineering process that generates new variables by extracting them from the raw data. The main aim of this step is to reduce the volume of data so that it can be easily used and managed for data modelling. Feature extraction methods include cluster analysis, text analytics, edge detection algorithms, and principal components analysis (PCA) [6] [7].

iv. **Feature Selection:** While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant. If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model. Hence it is very important to identify and select the most appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning. "Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features" [12] [6] [8].

5. Basic features of Cloud Computing

5.1. Resources Pooling

Resource pooling is one of the essential features of cloud computing. Resource pooling means that a cloud service provider can share resources among multiple clients, each providing a different set of services according to their needs. It is a multi-client strategy that can be applied to data storage, processing and bandwidth-delivered services. The administration process of allocating resources in real-time does not conflict with the client's experience [4] [6].

5.2. on-Demand Self-Service

It is one of the important and essential features of cloud computing. This enables the client to continuously monitor server uptime, capabilities and allocated network storage. This is a fundamental feature of cloud computing, and a customer can also control the computing capabilities according to their needs [5].

5.3. Easy Maintenance

This is one of the best cloud features. Servers are easily maintained, and downtime is minimal or sometimes zero. Cloud computing powered resources often undergo several updates to optimize their capabilities and potential. Updates are more viable with devices and perform faster than previous versions [4] [3].

5.4. Scalability and Rapid Elasticity

A key feature and advantage of cloud computing is its rapid scalability. This cloud feature enables cost-effective handling of workloads that require a large number of servers but only for a short period. Many customers have workloads that can be run very cost-effectively due to the rapid scalability of cloud computing [4].

5.5. Economical

This cloud feature helps in reducing the IT expenditure of the organizations. In cloud computing, clients need to pay the administration for the space used by them. There is no cover-up or additional charges that need to be paid. Administration is economical, and more often than not, some space is allocated for free.

5.6. Measured and Reporting Service

Reporting Services is one of the many cloud features that make it the best choice for organizations. The measurement and reporting service is helpful for both cloud providers and their customers. This enables both the provider and the customer to monitor and report which services have been used and for what purposes. It helps in monitoring billing and ensuring optimum utilization of resources [9].

5.7. Security

Data security is one of the best features of cloud computing. Cloud services make a copy of the stored data to prevent any kind of data loss. If one server loses data by any chance, the copied version is restored from the other server. This feature comes in handy when multiple users are working on a particular file in real-time, and one file suddenly gets corrupted [8].

5.8. Automation

Automation is an essential feature of cloud computing. The ability of cloud computing to automatically install, configure and maintain a cloud service is known as automation in cloud computing. In simple words, it is the process of making the most of the technology and minimizing the manual effort. However, achieving automation in a cloud ecosystem is not that easy. This requires the installation and deployment of virtual machines, servers, and large storage. On successful deployment, these resources also require constant maintenance [6].

5.9. Resilience

Resilience in cloud computing means the ability of a service to quickly recover from any disruption. The resilience of a cloud is measured by how fast its servers, databases and network systems restart and recover from any loss or damage. Availability is another key feature of cloud computing. Since cloud services can be accessed remotely, there are no geographic restrictions or limits on the use of cloud resources [5].

5.10. Large Network Access

A big part of the cloud's characteristics is its ubiquity. The client can access cloud data or transfer data to the cloud from any location with a device and internet connection. These capabilities are available everywhere in the organization and are achieved with the help of internet. Cloud providers deliver that large network access by monitoring and guaranteeing measurements that reflect how clients access cloud resources and data: latency, access times, data throughput, and more [12].

6. Conclusion

In this paper we have presented usage of python as a tool in various research areas like Data Wrangling, Machine learning and Cloud Computing. Along with Python language, there are many other languages are used for Data wrangling, Machine learning using cloud computing and for developing IoT devices like Java, C++ etc. But right now most of the developers use python scripting language than Java, C++. Because of its easy syntax, secure coding, and its simplicity. When it comes to robust and performance, developers choose Python. With respect to the future work there is still huge space for this language to serve other upcoming research areas because of its features like simplicity, extensive library, inbuilt and extensible In future we will propose python as a powerful tool which is used by many research communities.

Acknowledgement

We would like to thank Department of Computer Science & Engineering, Shobhit University Gangoh. For encouragement and support in writing this paper.

References

1. Cline Don, Yueh Simon and Chapman Bruce, Stankov Boba, Al Gasiewski, and Masters Dallas, Elder Kelly, Richard Kelly, Painter Thomas H., Miller Steve, Katzberg Steve, Mahrt Larry, (2009), NASA Cold Land Processes Experiment (CLPX 2002/03): Airborne Remote Sensing.
2. A. Fatima, N. Nazir, and M. G. Khan, "Data Cleaning In Data Warehouse: A Survey of Data Pre-processing Tech- niques and Tools," Int. J. Inf. Technol. Comput. Sci., vol. 9, no. 3, pp. 50–61, 2017.
3. Kandel Sean, Paepcke Andreas, Hellersteiny Joseph and Heer Jeffrey (2011), Wrangler: Interactive Visual Specifi- cation of Data Transformation Scripts, ACM Human Fac- tors in Computing Systems (CHI) ACM 978-1-4503- 0267-8/11/05.
4. Chaudhuri. S and Dayal. U (1997), An overview of data warehousing and OLAP technology. In SIGMOD Record
5. (2001) "Potter's Wheel: An Interactive Data Cleaning Sys- tem", Proceedings of the 27th VLDB Conference.
6. Ahuja.S, Roth.M, Gangadharaiah R, Schwarz.P and Bas- tidas.R, (2016), "Using Machine Learning to Accelerate Data Wrangling", IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016, Barcelona, Spain, pp. 343-349.doi:10.1109/ICDMW.2016.0055.
7. Data wrangling platform (2017) publication, www.trifacta.com.[Online] Available: https://www.trifacta.com/products/architecture//. [Accessed on: 01 May 2017].
8. Norman D.A, (2013), Text book on "The Design of Eve- ryday Things, Basic Books", [Accessed on: 12 April 2017].
9. Jordan, M.I., Mitchell, and T.M.: Machine learning: trends, perspectives, and prospects. Science 349(6245), 255–260 (2015)
10. Le Cun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436– 444 (2015)
11. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics - state- of-the-art, future challenges and research direc-tions. BMC Bioinform. 15(S6), I1 (2014)
12. Wolfram, S.: Mathematica: A System for Doing Mathematics by Computer. Addi-son Wesley Longman Publishing Co., Inc., Boston (1991)
13. Engblom, S., Lukarski, D.: Fast MATLAB compatible sparse assembly on multicore computers. Parallel Comput. 56, 1–17 (2016) https://www.researchgate.net/publication/330513589_Internet_of_Things_IOT