International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Automating Machine Learning Workflows with Cloud-Based Pipelines

[1]Pradeep Etikani, [2]Vijaya Venkata Sri Rama Bhaskar, [3]Savita Nuguri, [4]Rahul Saoji, [5]Krishnateja Shiva

Submitted: 05/11/2023 **Revised**: 18/12/2023 **Accepted**: 07/01/2024

Abstract: The paper is aimed at discussing cloud-based pipelines for automating machine learning processes. The paper also discusses how these types of systems overcome fundamental issues, which are associated with ML processes such as, including inefficiency, scalability problems and convolutions of collaborating among other similar systems. Cloud-based pipelines use distributed computation and storage to automate the whole ML pipeline right from data processing to model deploying. The study identifies advantages including more efficient process organization, managing resources as well as better integration of employees. Techniques that have been examined are automated data pipeline creation, large-scale model building and training and methods on service deployment and maintenance. Major findings show that the use of this framework leads to a reduction of the time required for accomplishing ML projects and enhancement in the quality of the models developed, in addition to facilitating effective replication of experiments.

Keywords: *inefficiency, accomplishing, organization, enhancement*

Introduction

Automation of ML pipelines, which is a new trend in artificial intelligence, is another problem that has received significant attention recently. As the complexity and the size of ML projects enhance, the basic methods of creating and implementing models encounter essential issues. Such challenges include an increased amount of time that is taken in performing the work manually, disparity which results from low reliability, and challenges in the distribution of resources throughout the process. Cloud-based pipelines seem to address these problems as an effective means of automating the ML processes within a flexible platform. Cloud based ML pipelines take advantage of distributed processing and storage models in the entire process of model development. It ranges from data curation and feature transformation to model training and inference these pipelines help in eliminating redundant work and in efficient utilization of computational resources. Besides, it accelerates the development process and improves the modularity and replicability of the models created with ML. There are many advantages of cloud pipelines for exercising ML automation.

Literature Review

Cloud-Based Machine Learning Platforms

According to García *et al.* 2020; Cloud-based machine learning platforms have become powerful tools for developing and deploying the ML models. These platforms help to organize a comprehensive environment in which different elements of the ML process are located, as tools for handling data, preparing them, training models, and deploying the results. It plans to provide the end-users with pre-built algorithms, auto-selection of the best model, and convenient APIs for data scientists and developers. Another significant strength of cloud-based ML platforms is in the execution of big data processing and training of the models.Moreover, it is observed that these pointed platforms afford version control, collaboration, and reproducibility features that would further assist teams to work more efficiently on the ML projects (García *et al.* 2020). It should be noted that the management of infrastructures is another problem solved by cloud-based ML platforms. It hides elements such as the installation and management of environments necessary for Machine Learning, so the data scientist doesn't have to bother with system configurations. This abstraction is helpful in accelerating the development as well as in the democratizing process of the ML to such an extent that everyone can use it, and it does not require special experience and time for infrastructure and technology management.

[1]*Independent Researcher,USA.*
[2]*Independent Researcher,USA*
[3]*Independent Researcher,USA.*
[4]*Independent Researcher,USA.*
[5]*Independent Researcher,USA*

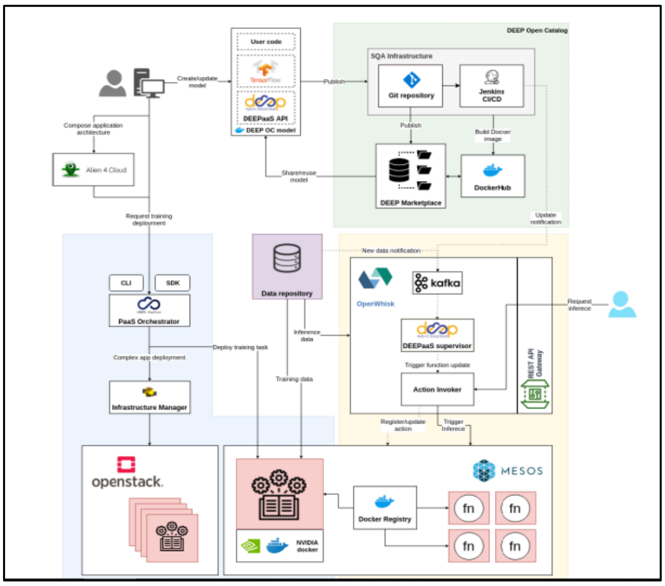

Figure 1: Deep detailed architecture

(Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8950411)

Automated Feature Engineering and Selection

According to Xin *et al.* 2021; Feature engineering and feature selection are now considered as catalytic steps in any automated ML pipelines. These processes include synthesizing, operating, and selecting from raw data to convert them into features without the reliance on human intervention and at the same time expose other features that may have been unnoticed by human analysts. ML pipelines in the cloud commonly include feature engineering automation tools that can support different data formats and forms. These tools perform several operations including, encoding nominal variables, management of missing values, and generating interaction terms to name but a few (Xin *et al.* 2021). These activities have to be done manually and it takes time, but with automated ML pipelines, the amount of time spent on data preparation is dramatically decreased, as this task usually occupies a large portion of a data scientist's working time. Analysis routines for feature extraction enable one to select the most relevant features for an ML task at hand, thus improving a model's learning efficiency and decreasing computational workloads.

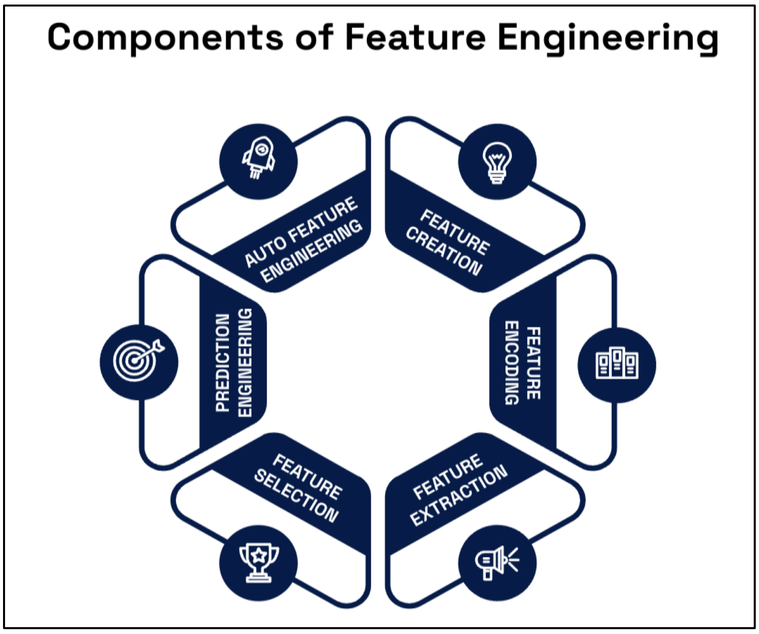

Figure 2: Components of feature engineering

(Source: https://miro.medium.com/v2/resize:fit:1400/0*hwaCbMptUC0fgqs9)

Scalable Model Training and Deployment

According to Bartezzaghi *et al.* 2022; The cloud-based ML pipelines stand out as very effective when it comes to the aspect of scaling. It is possible to implement an adaptive method for managing computational resources based on the specifics of the task in the field of ML to organize effective use of resources available in a cloud environment (Bartezzaghi *et al.* 2022). In training of models, cloud platforms can scale the workload across many machines as opposed to loading large and complex models into one machine. The distributed approach not only accelerates the training of the model but also enables the experimenting with different architecture of models and their parameters. Regarding deployment, the pipelines based on clouds allow models to be smoothly transferred to the production environment. It usually offers containerization platforms and serverless platforms that can be used to orchestrate and deploy ML models as and when the need arises. Moreover, it often consists of the features for observation of model efficacy, as well as checks on model drift and handling of model changes, thereby preventing long-term declines in the functionality of ML solutions.

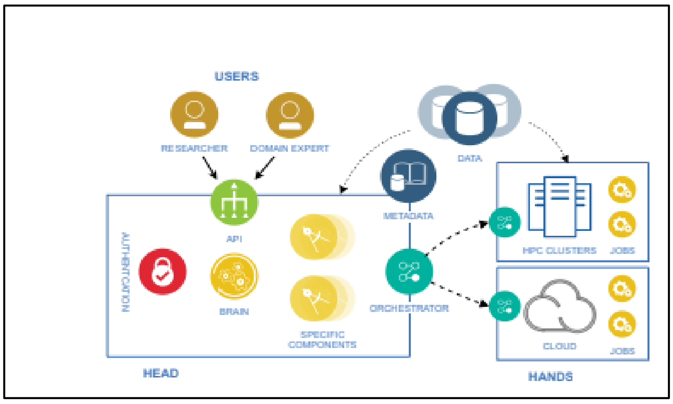

Figure 3: OCL architecture

(Source: https://drive.google.com/file/d/1fXFSqBpAavrWSbjbLhEBwE6jaIRXt2Xy/view)

Methods

Data Pipeline Design

The case of employing cloud environments for implementing ML solutions, proper data pipeline design is critical to the automation of various tasks in the ML life cycle. At the first stage, data acquisition is done, where raw data is accumulated from multiple sources and placed in data lakes or warehouses in the form of data marts. Subsequently, the procedures of data preparation are applied to preprocess, transform, and normalize the data (Bustamante *et al.* 2023). This usually entails dealing with cases of missing values; converting categorical data and normalizing the numerical data. Cloud-based pipelines allow for the development of the concept of microservices that is present in data processing pipelines. All these components can be easily linked to achieve complex data transformation in a given sequence. There are management tools which coordinate these pipelines to ensure that data is processed in a procedure as and when scheduled.

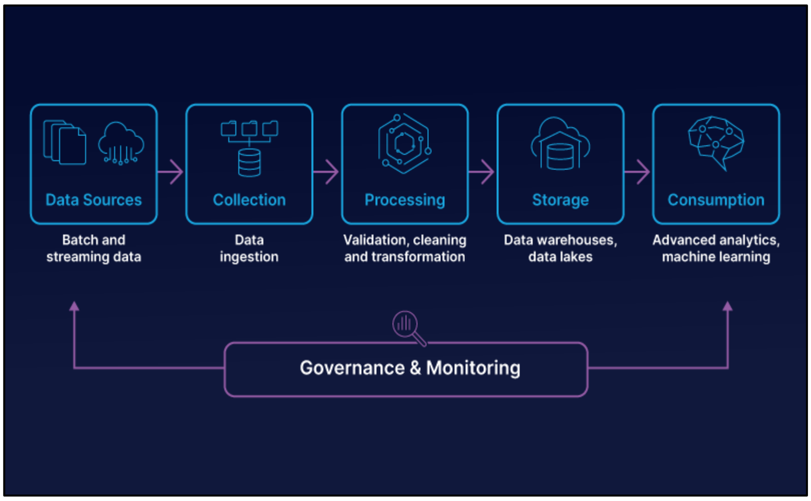

Fig 4: Data pipeline management

(Source: https://www.striim.com/wp-content/uploads/2022/06/image1-2.png)

Model Development and Training

Cloud-based ML pipelines in particular describe the model development and training in such a way that it is free and efficient. The first one includes choosing correct algorithms regarding the type of a problem and the specifics of its data. Most of the cloud providers have arrays of default algorithms, and, in many cases, possibility to upload custom models. To decide what kind of machine learning algorithm should be used and what values of its hyperparameters should be optimal, AutoML is utilized (Chowdhury *et al.* 2020). Bearing the same purpose in the search, these tools perform tests on a number of model architectures and configurations to determine the optimal combination for each task. Large data and complicated models require the usage of distributed training that can divide the training set and training parameters across multiple devices for efficiency's sake. Changes in Model code, Hyperparameters, and Training datasets are well Version control that is a part of the Development process.

Deployment and Monitoring

The Cloud-based ML pipelines help in solving the deployment problem where tenant setups come as a big disadvantage when moving from development to production environment. Containerisation technologies like docker are typically used to deploy models, bundles all the model related information and its dependencies into a single stack for different platforms. Automatic change management systems are put in place to control the release of new models or updated ones. The systems include capability for canary releases or A/B testing to determine how well the model is performing before it is widely released (Xin *et al.* 2021). Based on the load balancing and auto scaling the features are used to maintain and control the flow of traffic depending upon the extent of the traffic received. Another thing that has to be taken into account is the monitoring of the models that are live, after that have been deployed.

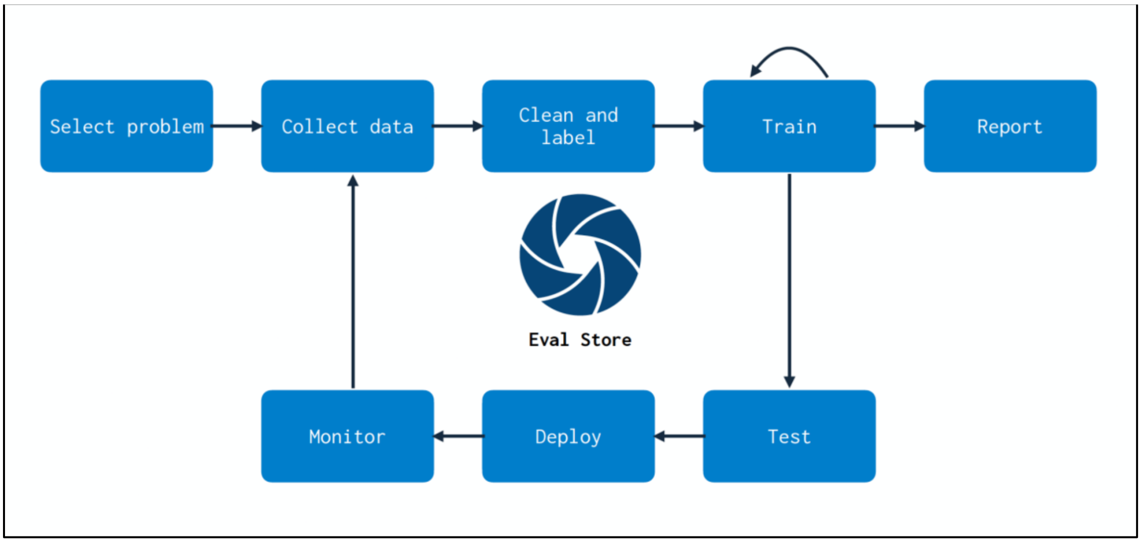

Fig 5: Deployment and monitoring

(Source: https://fullstackdeeplearning.com/spring2021/lecture-11-notes-media/image8.png)

Results

Improved Efficiency in ML Workflows

The use of pipeline automation coupled with cloud has had a great impact on the enhancement of machine learning. Some of the tasks in the traditional ML techniques are time-consuming and delicate with most of them requiring manual input, while with the help of the cloud solutions, data preparation, training, and model deployment time is significantly cut (Spjuth *et al.* 2021). Currently, data scientists do not have to spend much time on data cleaning and feature engineering since these are performed most times by the data science tools. It lets them spend more time on such activities as interpretation of models and their application in business. The time taken to develop work has been reduced significantly and thus fast iteration and experimentation including deploying models has been made. Also, through these efficient and optimized processes, the levels of errors that are associated with human beings have been brought down, hence enhancing the quality of ML projects. The usage of computerized methods in logging and tracking of experiments have helped in making it easier to come across successful techniques to be utilized and this has added efficiency to the system.

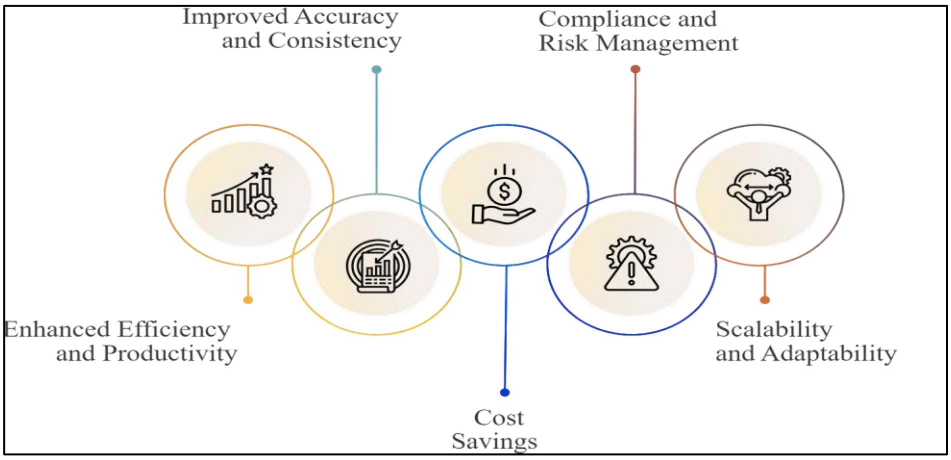
Improved Accuracy and Consistency
Compliance and Risk Management
Enhanced Efficiency and Productivity
Scalability and Adaptability
Cost Savings

Figure 6: Benefits of workflow automation

(Source: https://www.speridian.com/wp-content/uploads/2023)

Scalability and Resource Optimization

The set up and progression of serious, cloud-based ML pipelines have been demonstrated to be highly scalable and economically efficient. These systems can balance the computing requirements and allocate more computational resources when needed and decrease, when there is lesser demand for it across the cloud. As for the intensive tasks such as model training, working with large datasets or handling numerous users' requests, the pipelines can immediately allocate resources to complete the tasks as fast as possible (Goh *et al.* 2021). On the other hand, with satisfaction low, then fewer apparatus and accouterments are used to reduce dominant costs. This dynamic scaling has empowered organizations to deal with projects with different extents and intricacies while receiving no huge initial investment in gadgets.

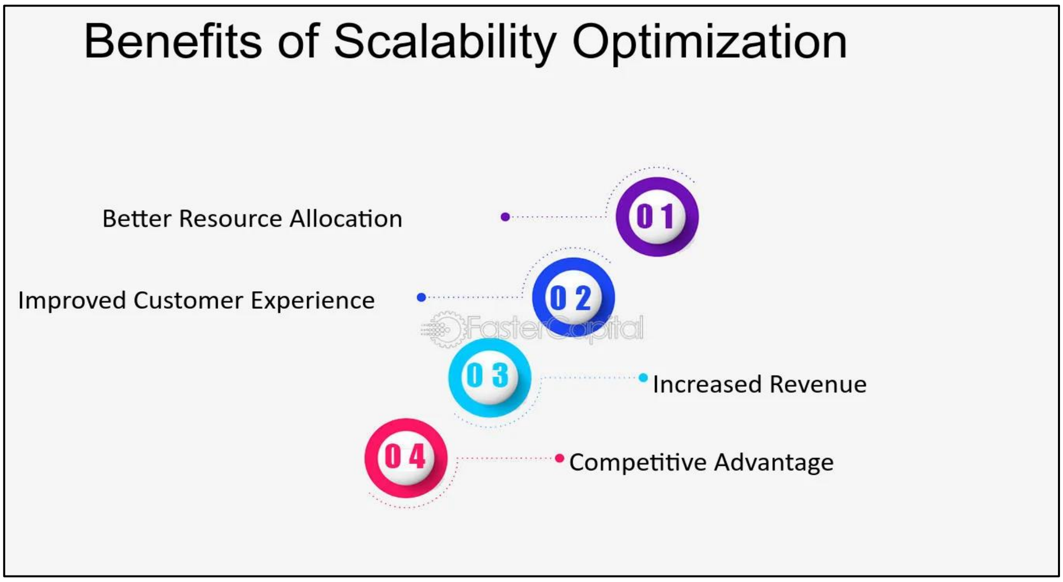
Benefits of Scalability Optimization
Better Resource Allocation
01
Improved Customer Experience
02
FasterCapital
03
Increased Revenue
04
Competitive Advantage

Fig 7: Scalability optimization

(Source: https://fastercapital.com/i/Scalability-optimization--Optimizing-Sca)

Enhanced Collaboration and Reproducibility

Deliveries of cloud-centered pipelines to ML have brought about better communication among developers and improved the repeatability of ML trials. These platforms create central working spaces that make it possible for data scientists, engineers, and the rest to collaborate. The use of version control systems in these pipelines enables the team users to view alteration of data and different versions of code and models (Mani *et al.* 2021). This makes it easy to work on project shares, roll back to previous state if there is a necessity, and comprehend development of ML models at a given time. Standard operating procedures have also advanced tremendously in terms of reproducibility of experiments. Most of the activities that are performed while working on the ML workflow, including data preprocessing and the transformation process, the parameters of the chosen model, and the metrics used for the model's assessment, are logged in detail, thus allowing for easy reproducibility of the experiment.

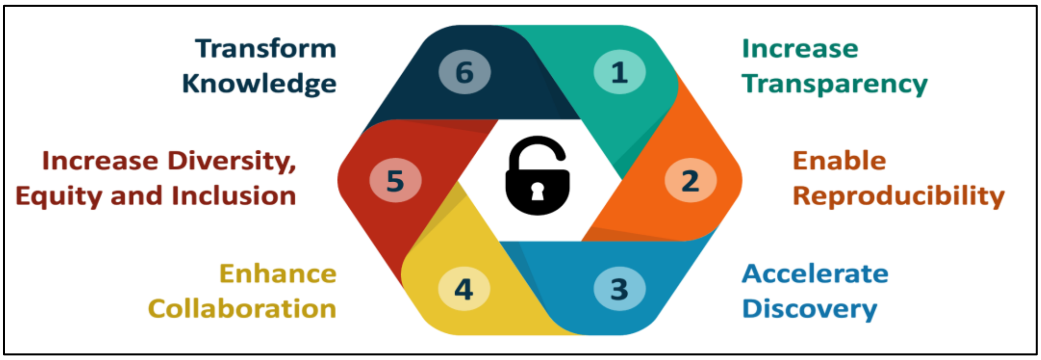

Fig 8: Data and reproducibility

(Source: https://learning.nceas.ucsb.edu/2022-04-arctic/images/open-science.png)

Discussion

Machine learning pipelines can now be hosted on the cloud and automated and this has greatly helped in the advancement of the field of artificial intelligence and data science. These systems have managed to overcome most of the problems that are characteristic of classic ML tasks, including, for example, low performance, non-scalability, and problems with teamwork. Nevertheless, there are some difficulties when introducing these automated business process flows (Alarcon *et al.* 2022). The concerns that organizations typically experience when it comes to the storage and transfer of data include the security of the data and privacy. Legal requirements like GDPR or CCPA are the additional challenges in the implementation of the cloud-based ML pipelines. Continuing challenges are also related to the high rate of development of technologies used in the cloud computing environment and ML. To carry out their activities effectively, teams need to keep refreshing their skills and learn how to use the available tools and applications (Colonnelli *et al.* 2021). This constant evolution leads to integration problems in cases where one is trying to transfer from one cloud service to another or try to integrate multiple cloud service providers. Nevertheless, there are key advantages of using cloud-based environments to build the ML pipelines.

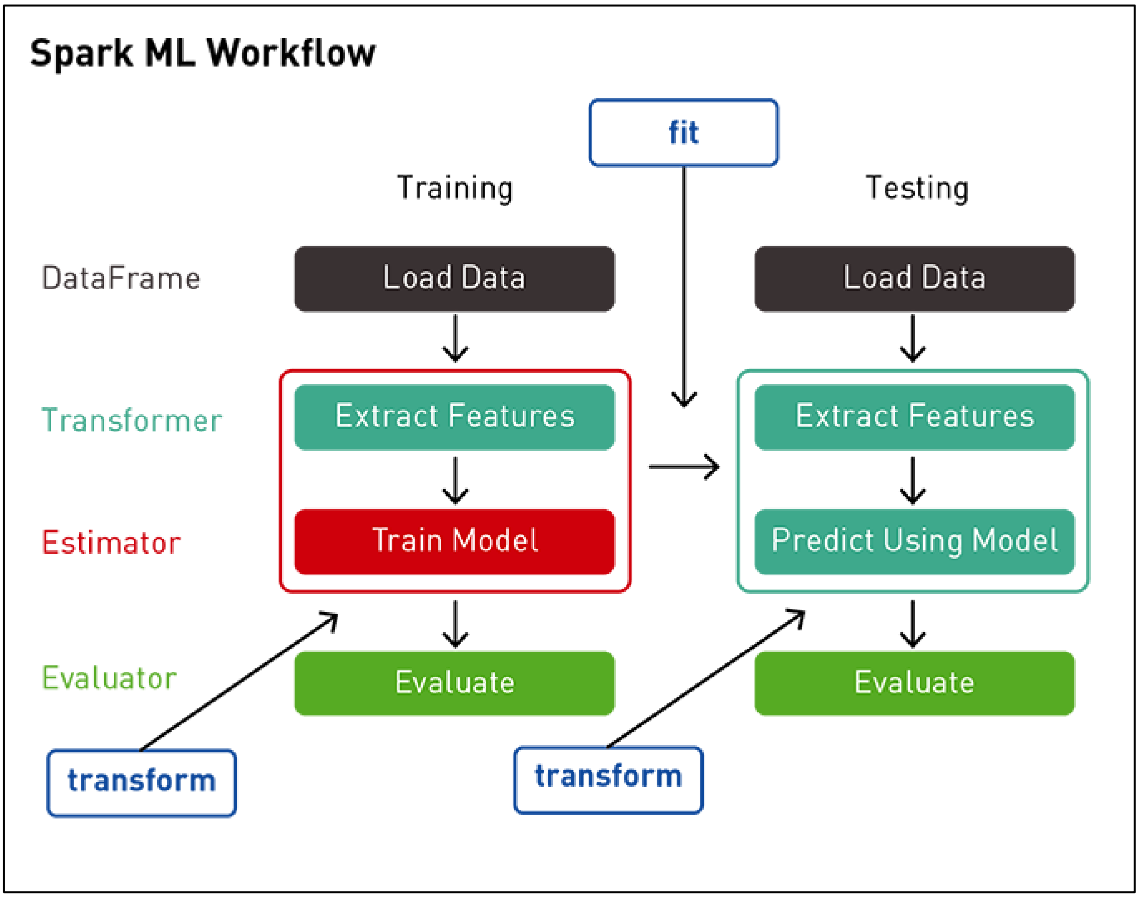

Fig 9: Pipeline in machine learning

(Source: https://miro.medium.com/v2/resize:fit:680/1*c4SNMDj18FHQakGS6Gmgsg.png)

Future Directions

Further evolution of automation and intelligence of cloud-based ML pipelines would be most probably observed in the future. Some of the current trends include improvement of AutoML involving advanced technology that will allow creation of excellent models regardless of the identity of the developer (Quaranta *et al.* 2021). The explainable AI techniques will be integrated into automated business processes to improve the interpretability of models. Direct connection of the concept of edge computing with cloud pipelines will enhance the capability of processing data nearer to the source.

Conclusion

In conclusion, cloud-based pipelines have drastically changed the ways that machine learning processes were implemented. These systems have made it easy to enhance efficiency, scalability and teamwork in ML activities due to automation of several crucial procedures. Learned that organizations implementing the aforementioned technologies enjoyed better development cycles and optimized resources while the models had better quality. Despite the difficulties related to data security and the imperative to constantly learn from the data, the advantage of using cloud-based ML pipelines is obvious. There are clear indications that as the field progresses there will be more incorporation of enhanced automation and AI into the ML processes.

Reference List

Journals

[1] Alarcon, M.L., Oruche, R., Pandey, A. and Calyam, P., 2022. Cloud-based data pipeline orchestration platform for COVID-19 evidence-based analytics. In *Novel AI and Data Science Advancements for Sustainability in the Era of COVID-19* (pp. 159-180). Academic Press.

[2] Bartezzaghi, A., Giurgiu, I., Marchiori, C., Rigotti, M., Sebastian, R. and Malossi, C., 2022, June. Design of a Cloud-Based Data Platform for Standardized Machine Learning Workflows with Applications to Transport Infrastructure. In *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)* (pp. 764-769). IEEE.

[3] Bustamante, A.L., Patricio, M.A., Berlanga, A. and Molina, J.M., 2023. Seamless transition from machine learning on the cloud to industrial edge devices with thinger. io. *IEEE Internet of Things Journal*, *10*(18), pp.16548-16563.

[4] Chowdhury, K., Lamacchia, D., Frenk Feldman, V., Mallik, A., Rahman, I. and Alam, Z., 2020, November. A Cloud–Based Smart Engineering and Predictive Computation System for Pipeline Design and Operation Cost Reduction. In *Abu Dhabi International Petroleum Exhibition and Conference* (p. D012S116R200). SPE.

[5] Colonnelli, I., Cantalupo, B., Spampinato, C., Pennisi, M. and Aldinucci, M., 2021. Bringing AI pipelines onto cloud-HPC: setting a baseline for accuracy of COVID-19 AI diagnosis. *arXiv preprint arXiv:2108.01033*.

[6] García, Á.L., De Lucas, J.M., Antonacci, M., Zu Castell, W., David, M., Hardt, M., Iglesias, L.L., Moltó, G., Plociennik, M., Tran, V. and Alic, A.S., 2020. A cloud-based framework for machine learning workloads and applications. *IEEE access*, *8*, pp.18681-18692.

[7] Goh, P.J., Hoe, Z.Y., Low, C.Y., Koh, C.T., Mohammad, U., Lee, K. and Tan, C.F., 2021, November. Conceptual design of cloud-based data pipeline for smart factory. In *Symposium on Intelligent Manufacturing and Mechatronics* (pp. 29-39). Singapore: Springer Nature Singapore.

[8] Mani, D.R., Maynard, M., Kothadia, R., Krug, K., Christianson, K.E., Heiman, D., Clauser, K.R., Birger, C., Getz, G. and Carr, S.A., 2021. PANOPLY: a cloud-based platform for automated and reproducible proteogenomic data analysis. *Nature methods*, *18*(6), pp.580-582.

[9] Quaranta, L., Calefato, F. and Lanubile, F., 2021. A taxonomy of tools for reproducible machine learning experiments. *AIxIA 2021*.

[10] Spjuth, O., Frid, J. and Hellander, A., 2021. The machine learning life cycle and the cloud: implications for drug discovery. *Expert opinion on drug discovery*, *16*(9), pp.1071-1079.

[11] Xin, D., Miao, H., Parameswaran, A. and Polyzotis, N., 2021, June. Production machine learning pipelines: Empirical analysis and optimization opportunities. In *Proceedings of the 2021 International Conference on Management of Data* (pp. 2639-2652).

[12] Xin, D., Wu, E.Y., Lee, D.J.L., Salehi, N. and Parameswaran, A., 2021, May. Whither automl? understanding the role of automation in machine learning workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

[13] Chenchala, P. K., Choppadandi, A., Kaur, J., Nakra, V., & Pandian, P. K. G. (2020). Predictive Maintenance and Resource Optimization in Inventory Identification Tool Using ML. International Journal of Open Publication and Exploration, 8(2), 43-50. https://ijope.com/index.php/home/article/view/127

[14] Predictive Maintenance and Resource Optimization in Inventory Identification Tool Using ML. International Journal of Open Publication and

Exploration, 8(2), 43-50. https://ijope.com/index.php/home/article/view/127

[15] Fadnavis, N. S., Patil, G. B., Padyana, U. K., Rai, H. P., & Ogeti, P. (2020). Machine learning applications in climate modeling and weather forecasting. NeuroQuantology, 18(6), 135-145. https://doi.org/10.48047/nq.2020.18.6.NQ20194

[16] Tilala, Mitul, and Abhip Dilip Chawda. "Evaluation of Compliance Requirements for Annual Reports in Pharmaceutical Industries." NeuroQuantology 18, no. 11 (November 2020): 138-145. https://doi.org/10.48047/nq.2020.18.11.NQ20244.

[17] AI-Driven Customer Relationship Management in PK Salon Management System. (2019). International Journal of Open Publication and Exploration, ISSN: 3006-2853, 7(2), 28-35. https://ijope.com/index.php/home/article/view/128

[18] Mitul Tilala, Abhip Dilip Chawda, Abhishek Pandurang Benke, Akshay Agarwal. (2022). Regulatory Intelligence: Leveraging Data Analytics for Regulatory Decision-Making. International Journal of Multidisciplinary Innovation and Research Methodology, ISSN: 2960-2068, 1(1), 78–83. Retrieved from https://ijmirm.com/index.php/ijmirm/article/view/77

[19] Tilala, Mitul, and Abhip Dilip Chawda. "Evaluation of Compliance Requirements for Annual Reports in Pharmaceutical Industries." NeuroQuantology 18, no. 11 (November 2020): 138-145. https://doi.org/10.48047/nq.2020.18.11.NQ20244.

[20] Kamuni, Navin, Suresh Dodda, Venkata Sai Mahesh Vuppalapati, Jyothi Swaroop Arlagadda, and Preetham Vemasani. "Advancements in Reinforcement Learning Techniques for Robotics." Journal of Basic Science and Engineering 19, no. 1 (2022): 101-111. ISSN: 1005-0930.

[21] Narukulla, Narendra, Joel Lopes, Venudhar Rao Hajari, Nitin Prasad, and Hemanth Swamy. "Real-Time Data Processing and Predictive Analytics Using Cloud-Based Machine Learning." Tuijin Jishu/Journal of Propulsion Technology 42, no. 4 (2021): 91-102.

[22] Nitin Prasad. (2022). Security Challenges and Solutions in Cloud-Based Artificial Intelligence and Machine Learning Systems. International Journal on Recent and Innovation Trends in Computing and Communication, 10(12), 286–292. Retrieved from https://www.ijritcc.org/index.php/ijritcc/article/view/10750

[23] Big Data Analytics using Machine Learning Techniques on Cloud Platforms. (2019). International Journal of Business Management and Visuals, ISSN: 3006-2705, 2(2), 54-58. https://ijbmv.com/index.php/home/article/view/76

[24] Shah, J., Prasad, N., Narukulla, N., Hajari, V. R., & Paripati, L. (2019). Big Data Analytics using Machine Learning Techniques on Cloud Platforms. International Journal of Business Management and Visuals, 2(2), 54-58. https://ijbmv.com/index.php/home/article/view/76