Canonical

MLOps explained: choose the right stack to build your end-to-end solutions

Executive summary

Generative AI and large language models (LLMs) are not just a hot topic, but a great opportunity for enterprises to tackle some of the most challenging problems they have. Generative AI offers organisations a chance to automate repetitive tasks, improve processes and move away from traditional approaches to performing different activities.

This is all exciting but the tooling selection is not that easy to navigate. Having the right machine learning tools unblocks teams that are starting their journey. Once you identify use cases and have the data at hand, building an environment where professionals can initially experiment and then scale is crucial. Among the key factors, they need to consider the underlying hardware, the chosen cloud environment and the machine learning tools available. Open source is widely adopted in the AI ecosystem. However, challenges such as security, user management or tool integration are still barriers to adoption.

This whitepaper presents a toolkit for organisations that want to assess their AI readiness. It walks you through the entire stack, from the hardware layer to the application layer. It covers key factors to consider when building a solution, as well as suggested solutions for different parts of the stack. Organisations that are looking to run production-grade environments with enterprise support or managed services will also find this paper useful.

The machine learning toolkit suggested by Canonical includes:

- Hardware and software that is already tested and validated on the market
- Machine learning tools such as Charmed Kubeflow
- Container solutions such as MicroK8s
- Cloud computing and different scenarios
- Production-grade solutions that can be rolled out within an enterprise

Having a solution that covers both all layers, from the OS to the machine learning tools, as well as the machine learning lifecycle defines the success of the projects and the speed of initiative delivery.

Contents

AI/ML trends and challenges

AI/ML is transforming the world at a rapid pace. It's gone from being a researched-focused practice to a discipline that delivers production-grade projects. Enterprises are growing their AI budgets, and are open to investing both in infrastructure and talent to accelerate their initiatives. Still, there are multiple challenges that companies need to address. These can be grouped into four main categories:

- **People:** there is a skills gap on the market that makes hiring difficult. AI projects also have multiple stakeholders that can slow down the project delivery.

- **Operations:** scaling up a project is difficult, mainly because of the associated operational maintenance costs.

- **Technology:** there is a fast-changing landscape of new tools, frameworks and libraries that teams need to consider. Identifying the right choices is often overwhelming.

- **Data:** both the continuous growth of data volume, as well as the need to deal with highly sensitive data are aspects that keep teams up at night.

The tangible benefits of AI/ML

AI/ML is without a doubt here to stay. Enterprises are often attracted to it mainly because of the benefits that it brings, as well as the high promises of the possible return on investment. Though it has been used for a long time for research purposes, nowadays there is a wider spectrum of industries which consider it to do business as usual, such as DNA sequencing in life sciences for example, or field map analysis in oil and gas.

AI/ML addresses pressing issues for companies. Task automation is one of the most common benefits; activities such as data entry or email filtering don't need to be a burden anymore. This reduces the probability of human error and increases productivity. Furthermore, AI/ML often brings cost savings to enterprises, by enabling professionals to automate processes and tackle complex problems.

More recently, AI/ML has been used to tackle the labour shortage in some areas and to improve sustainability. For example, in the telco industry, self-organising networks are often used to optimise power usage and reduce its consumption.

In the upcoming chapters, we will outline how to overcome common challenges and benefit from AI/ML with the right tools. From choosing the right hardware for your compute to suitable machine-learning tools for your use case, we will go through all the layers of an end-to-end machine learning solution. We will cover how enterprises can roll out open source tooling and benefit from security patching, support and upgrades.

Compute power for machine learning

Machine learning is known for requiring extended compute resources. This is the main reason AI/ML projects are considered high-cost. Deep learning, in particular, has been often hindered due to computational limitations.

Graphical Processing Units (GPUs) have been a huge accelerator for machine learning adoption, given that they increase compute power. Whilst the chip shortage remains a challenge, GPUs enable professionals to run models with massive numbers of parameters quickly and efficiently. They enable parallel processing of training tasks, task distribution over clusters of processors and simultaneous compute operations. Compared to Compute Processing Units (CPUs), they accelerate project delivery and can be optimised for targeted tasks to complete computations more quickly.

When choosing a GPU, there are specific factors that are worth considering:

- **Unit interconnection:** this is directly related to scalability and often projects will require an increasing amount of compute power, so considering it from the very beginning is crucial.

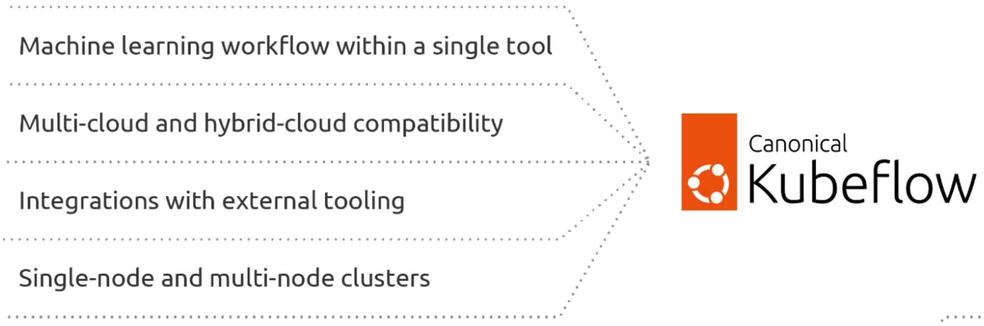
- **Supported software:** supported machine learning libraries and frameworks make a difference when choosing the hardware layer, enabling professionals to start quickly without having to build custom integrations.

- **Licensing:** depending on the hardware need, there are various requirements and constraints that need to be taken into account. Often, CUDA licensing represents a key factor when making a choice.

NVIDIA DGX

NVIDIA DGX systems are a leading example of infrastructure that is purpose-built for enterprise AI use cases. They feature pioneering NVIDIA Tensor Core GPUs, which vastly outperform traditional CPUs for ML workloads, alongside advanced networking and storage capabilities. NVIDIA DGX systems provide enterprises with proven infrastructure solutions on which they can build and scale their AI infrastructure. Powered by clusters of NVIDIA DGX A100 or DGX H100 systems and NVIDIA InfiniBand networking, DGX configurations easily meet the compute requirements of AI projects, while also providing record-breaking performance. DGX systems work seamlessly alongside MLOps solutions certified through the NVIDIA DGX-Ready Software program. Leading MLOps platforms, such as Canonical's Charmed Kubeflow, are tested and optimised to work on clusters of DGX systems, ensuring that users can get the most out of their AI infrastructure without worrying about manually integrating and configuring their MLOps software.

Read more about running AI at scale with Canonical and NVIDIA

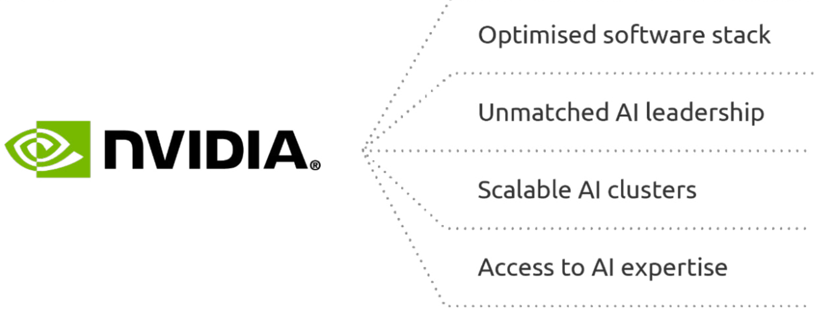


Machine learning workflow within a single tool

Multi-cloud and hybrid-cloud compatibility

Integrations with external tooling

Single-node and multi-node clusters



Optimised software stack

Unmatched AI leadership

Scalable AI clusters

Access to AI expertise

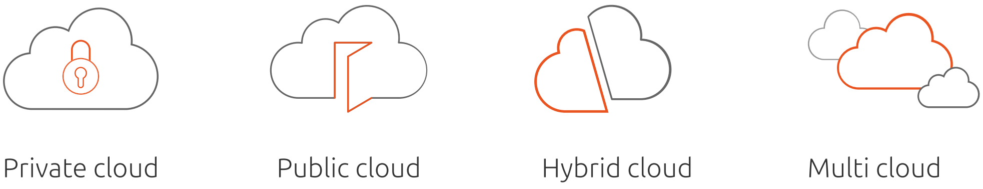
Clouds for machine learning

Creating effective machine learning models is highly correlated with the amount of data and the model training and deployment needs a project has. Cloud computing seems like a natural fit for AI, allowing solution architects to do away with the burden of provisioning, spinning down or turning on large-scale clusters with little downtime. While the experimentation phase requires fewer resources, building enterprise-ready models requires serious computing resources. In order to overcome this barrier, companies take different approaches which have pros and cons.

Often, organisations will:

- Build their own infrastructure (or private cloud)
- Benefit from the public cloud computing power
- Choose a hybrid cloud
- Build a multi-cloud scenario

Types of cloud deployment:



ML on private clouds

A private cloud is a cloud computing environment that is exclusively dedicated to a single entity or a service. It runs on the organisation’s premises or in an external data centre. It is managed by the organisation’s operations team or a managed service provider. Solutions such as Charmed Openstack are a great example of a cloud platform that can be used to build and deploy machine learning applications. On one hand, it has computing capabilities that enable building, managing and testing AI initiatives. It combines the power of open source tooling with the experience of a mature product that is supported by a large community. On the other hand, OpenStack has not been the first option that enterprises choose for running AI projects, due to the configuration challenges that it brings.

The last couple of years saw the emergence of a couple of interesting projects from the OpenStack community that are designed for machine learning. Magnum and Ironic are among them. The OpenInfra Foundation, the biggest open source community that supports infrastructure software, is also investing in machine learning projects to help abstract compute storage and network resources using virtualisation. This divides the hardware resources, abstracting away the physical resources from the software systems running on them.

Private clouds such as Charmed Openstack are a handy solution for enterprises when it comes to AI/ML. Charmed OpenStack is a cost-effective platform that provides enterprise-grade infrastructure for machine learning workloads. It reduces the initial investment and also enables teams to start initiatives faster. Security and governance have always been a priority for machine learning projects, so private clouds are a natural fit. Data protection regulations that have been developed around the globe encourage enterprises to choose private clouds as part of their machine-learning toolkit.

Charmed Openstack is well integrated with features and solutions form NVIDIA, the AI leader in the market. NVIDIA enables various types of accelerator devices in Charmed OpenStack, including virtual GPUs (vGPUs) or data processing units (DPUs) to maximise the usage of the cloud. This enables professionals to iterate faster.

Although private clouds have their advantages and provide flexibility to meet compliance requirements and lower costs, using private clouds exclusively can prove problematic. For example, it takes longer to train models given that it takes a long time to set up a private cloud compared to the out-of-the-box solutions offered by public clouds. Scaling a private cloud and associated storage can also be challenging. MLOps platforms that are available as a service are another option for vendors. Canonical works closely with both public and private clouds to offer a well-integrated solution for machine learning workloads, which addresses different layers of the stack, depending on the end use case.

Robust storage solutions are needed to process workflows in a quick manner. Given the large data sets often used in these projects, organisations also need to consider requirements such as low latency or connection to specific GPU clusters. New storage concepts are appearing on the market to cover this need, which is why OpenStack is partnering with vendors like NVIDIA. The inner and outer ring of storage is a new concept that divides the high speed and low latency. On the first one mentioned, from the throughput and high availability from the second one. These are just some examples of how private clouds address the needs of large datasets, leading to performance improvements and increased workload portability.

However, private clouds are not the only option that organisations have. When private clouds reach their maximum capacity, public clouds stand as a suitable option.

ML on public clouds

Public clouds are cloud computing platforms offered by third-party providers like Google, Amazon and AWS. They have different pricing models and offer different types of instances. This diversity in instances is beneficial to companies, as their compute needs vary depending on the amount of data and model training requirements.

Machine learning projects are usually run on public clouds because of the low initial cost. There is no need to purchase hardware since public cloud vendors offer multiple instance types with powerful computing power, including GPUs or DGXes. This leads to time efficiency from various perspectives since the time to set up the environment is close to none. Each of the large public clouds developed its own machine learning platform; SageMaker, VertexAI and Neptune are tools of choice. They are optimised for the infrastructure layer underneath but also offer an out-of-the-box solution that requires little deployment time.

Depending on the enterprise's AI readiness, professionals are looking for solutions that cover different steps of the machine learning lifecycle. Public clouds can cover all the stages, from data cleansing to model serving, enabling professionals to focus less on compatibility issues between tools, and more on the coding side of the job. Public clouds also offer enterprise support, ensuring professionals' work continuity.

However, public clouds are not always the best fit. While their cost seems lower in the initial phase, it grows exponentially as machine learning projects scale. This effect is also accelerated by data mobility concerns since any change translates into extra costs and carries risks that might impact model performance. Furthermore, there are security concerns about using public networks for different data sources, with regulations emerging all around the globe to tackle this issue. Cloud-based machine learning models are often exposed to public networks, which are vulnerable to different attacks such as denial of service. Many of these threats are non-existent when deployed behind a private cloud's firewall.

Public cloud marketplaces: an easy alternative

Public clouds make it easy to set up a new environment for machine learning. They also encourage third parties to share their applications as an out-of-the-box solution, which enables enterprises to quickly get started on AI. The Charmed Kubeflow Appliance in the AWS Marketplace is a good example of such a solution. It is a perfect match for the public cloud, enabling data scientists and machine learning engineers to set up an MLOps environment in minutes, reducing time spent on the setup phase. Users can get familiar with the tool before running AI at scale.

Charmed Kubeflow on AWS:

- Easy to install
- Easy to use
- Full capabilities of Charmed Kubeflow
- Easy to maintain
- Managed or support services
- Community-driven
- Open-source toolkit
- Closer to data sources

ML on hybrid clouds

Both public and private clouds have pros and cons for running AI. While public clouds allow users to quickly get started, from a cost perspective, on-prem solutions seem more attractive. Hybrid clouds are the middle ground.

Hybrid clouds unlock the value of data. Data is spread across different clouds and data complexity continues to increase. Thus, monolithic data infrastructures can become ineffective and organisations may need to rethink their infrastructure. Hybrid clouds give the necessary flexibility and accessibility for such a need. They offer the data foundation needed to scale and operationalise AI, enabling models to be better fed with data, regardless of where it is stored, leading to better accuracy.

At the same time, considering that AI adoption has been constrained by costs, hybrid clouds take this burden away. Training and running models often result in hidden costs because of the difficulty to predict the price attached to them. Hybrid clouds are a good alternative to carefully allocate investments since they provide an easy way to optimise spending and avoid any overhead of migrating workloads from one place to another. Expensive tasks can be done on-prem, letting leaders better control and predict spending.

However, hybrid clouds are often difficult to set up initially. Their implementation requires tools that are compatible with both public and private clouds. At the same time, there are security questions that arise, mainly because of the complexity of these types of solutions. Organisations need to check their security systems before loading any highly sensitive data into a hybrid cloud. Lastly, integrating a hybrid cloud into existing infrastructure can be challenging. Solution architects need to consider the robustness and reliability of both the cloud and existing infrastructure before making any change.

Multi-cloud scenarios

Multi cloud scenarios include two or more public cloud services. When compared to hybrid clouds, the main difference is that hybrid cloud users own and manage a private cloud resource as part of their cloud infrastructure.This is usually hosted in-house, in on-premises data centres, or on dedicated servers in third-party data centres. The private cloud then syncs with public cloud workloads to create an overall business solution.

Generative AI is just one of the drivers accelerating multi-cloud adoption, due to the extremely large datasets. Multi clouds are great because organisations can use the strengths of different clouds to optimise their machine learning projects. It mitigates the risk of vendor lock-in, promoting flexibility across the clouds. Multi-clouds ensure that applications can always run at the lowest possible cost. They result in improved stability, reducing the risk of service disruptions.

Like hybrid clouds, interoperability represents one of the biggest challenges. Implementation carries a large overhead and the long-term maintenance should not be underestimated. Multi-cloud environments need extra management, which is directly proportional to the number of service providers used. Different clouds have various proprietary tools and APIs, so seamless integrations are difficult.

How to choose your cloud

Machine learning projects are raising the bar when considering cloud providers. The obvious choices are private and public clouds. Hybrid and multi-cloud environments are more likely to increase in adoption in the future but are still challenging to adopt for most organisations. How do you determine which cloud is right for your use case?

The most important considerations are:

- **Enterprise AI readiness:** An organisation’s needs will vary depending on their AI maturity stage. While companies that are just getting started can easily use an appliance from a public cloud marketplace, they might need to consider other scenarios to scale up their initiatives.

- **Security and compliance requirements:** Different markets and industries have different compliance requirements. When building a project, taking those into account from an early stage is crucial. Compliance may dictate not only the environment but also the tooling. For example, highly sensitive data should not be trained on public clouds.

- **Existing infrastructure:** Machine learning might be a new initiative within the company, but considering the existing infrastructure is crucial. If an organisation is already investing heavily on the public cloud, starting there makes sense. Similarly, if there is a spare public cloud that could be used, professionals should not hesitate to follow that path.

- **Team size:** While experimentation on a workstation does not require more than a Jupyter Notebook and a GPU, larger teams need user management for projects to be developed in a reproducible manner. This enables initiative continuity, improving user efficiency, as well as portability.

- **Dataset size:** The larger the dataset is, the more computing power is needed. Usually, computer vision projects are the most intensive ones; they entail large volumes of data and also a variable number of data sources that require further processing, such as videos.

- **Long term vision:** Aligning a new initiative with the long term vision of the company is crucial. It can accelerate the project or bring it down to a halt. For example, for organisations that are trying to migrate from one cloud to another, considering the migration from the start will accelerate the process and unblock budgets.

- **Technologies and service providers:** Before settling for any providers, ensure that the platform and preferred technologies implemented on the service align with the business environment and support your cloud objectives. Key considerations include the provider's services, standards, and architectures - these should suit your workloads and management preferences. Many providers offer migration paths, so changes are less of a burden. In addition, the roadmap of the cloud providers plays an important role.

- **Reliability and performance:** Checking the performance against SLAs is a good measure for reliability. Most of these stats are publicly available. While cloud downtime is inevitable, what really matters is how the provider deals with it. Ensuring that clouds have an established, effective process for dealing with planned and unplanned downtimes is key. Data recovery is also a crucial aspect to consider. Understanding the provider's data recovery provisions and assessing their ability to support data preservation expectations is also important. This includes everything from the criticality of data to scheduling and integrity checks, as well as backup and restore capabilities.

- **Cost:** All service providers have unique pricing models. They compute cloud costs differently, making it virtually impossible to make a side-by-side comparison. To get around these limitations when comparing costs, organisations first need to consider their requirements, then decide which pricing model suits the project and long term vision. In most cases, long-term consumption timelines are better priced. Additionally, different organisations offer different cost assessments when scaling up or down. Some also have hidden charges, so you should carefully scrutinise the contract before making your decision.

Cloud providers, regardless of their type, ensure certain capabilities for machine learning. However, containers make machine learning deployments simpler and portable. Solutions can run on any of the scenarios presented above, ensuring that professionals can focus on modelling rather than dealing with the infrastructure layers.

Kubernetes for machine learning

Kubernetes is an open source cloud-native platform designed for container orchestration. It is used for automating software deployment, scaling and management. Kubernetes optimises application development on the cloud, giving users a platform to schedule and run containers on clusters of physical or virtual machines (VMs).

Container platforms such as Kubernetes are valuable for machine learning due to different reasons, such as:

- **Reproducibility and portability:** they encapsulate the entire stack, ensuring consistency in deployment and portability of machine learning models across different environments.

- **Dependency isolation:** dependencies can be efficiently isolated, preventing conflicts and simplifying any dependency, This makes it easy to work with different library versions.

- **Scalability and resource management:** containers enable efficient resource allocation and scaling of ML workloads, improving performance and reducing costs.

For enterprises that already have a private cloud deployed, Kubernetes is also an option to simplify and automate machine learning workloads. Kubernetes can be used as an extension of Charmed OpenStack, Canonical's private cloud solution. This enables enterprises to kickstart AI projects on the existing infrastructure.

Kubernetes acts as an orchestrator that automates the deployment, scaling, and management of containerised applications. Multiple tools such as Kubeflow and MLFlow run on top of it, ensuring a seamless experience for data scientists and machine learning engineers. By combining a container orchestration platform with a machine learning application, users ensure cost reductions and faster project delivery.

MicroK8s

Learn more about MLOps on highly sensitive data using MicroK8s

MicroK8s is a low-ops distribution of Kubernetes, developed by Canonical. It is a production-grade platform that offers high-availability Kubernetes clusters with minimum effort. It is often used for the deployment of machine learning tools, since it is fully open source and requires little resources to be deployed. Capabilities such as strict confinement are especially attractive for use cases that work with highly sensitive data. Strict confinement provides complete isolation, up to a minimal access level to the host resources, including restrictions in regards to file access, networks, processes, or any other system resource. Strict confinement uses security features in the Linux kernel, including AppArmor, seccomp and namespaces, to prevent applications and services from accessing the wider system.

Machine learning tooling

Machine learning operations (MLOps) is shortly defined as DevOps for machine learning. It is a set of practices that simplify the processes and automate machine learning workloads. Most MLOps solutions are cloud native, running on Kubernetes.

Open source MLOps is an alternative that gives access to the source code, can be tried without paying and allows for software contributions. AI/ML benefited from upstream open source communities right from the start. This led to an open source MLOps landscape that produced solutions focused on various categories such as:

- **End-to-end platforms:** Kubeflow, MetaFlow or MLFlow are a few examples of end-to-end MLOps platforms. Rather than covering a certain problem, they offer a solution with a suite of features. Amongst Kubeflow's components, there's Jupyter Notebooks, Tensorflow job operator and simplified containerisation. MLFlow's capabilities centre around tracking, project and modelling.

- **Development and deployment:** MLRun, ZenML or SeldonCore are just some of the examples that fall into this category. Their main purpose is to automate tasks and perform deployments on multiple environments.

- **Data:** Being the core of any AI project, this category itself can be further split into:

  - **Validation:** Hadoop and Spark are the leaders when it comes to this kind of activity. Their main goal is to perform data quality checks, automating this process. Duplicates, inconsistencies or missing data are just some of the capabilities these tools can handle.

  - **Exploration:** Jupyter Notebook is the main example of this sub-category. It automates the data analysis phase, providing an easy way to visualise data easily, track the changes and execute the code as you go.
  - **Versioning:** Pachyderm or DVC are just two of the examples that cover this functionality. Knowing that ML models require multiple experiments, a versioning control tool is required, in order to keep track of the evolution.

- **Testing:** Flyte is a good candidate to ensure these features. It is the last step in the machine learning lifecycle. Long term, It ensures the reproducibility of tasks and easily flags any problems that might arise in the process
- **Monitoring:** Tools such as Prometheus and Grafana are known on the market for covering these needs. They integrate with other tools. Their main goal is to ensure that models work as expected when deployed and that any change performed in the model does not have a negative impact.
- **The scheduling system:** Volcano is the recommended open source tool for this kind of job. Optimised for intensive computing workloads, it is a Kubernetes native batch scheduler. In AI initiatives, jobs are scheduled entirely to avoid failure and allow for faster training.

End-to-end MLOps platform	Data	Development and deployment	Testing	Monitoring
Kubeflow mlflow METAFLOW	Pachyderm hadoop jupyter APACHE Spark	ZenML SELDON BENTOML MLRun	Flyte	Prometheus Grafana

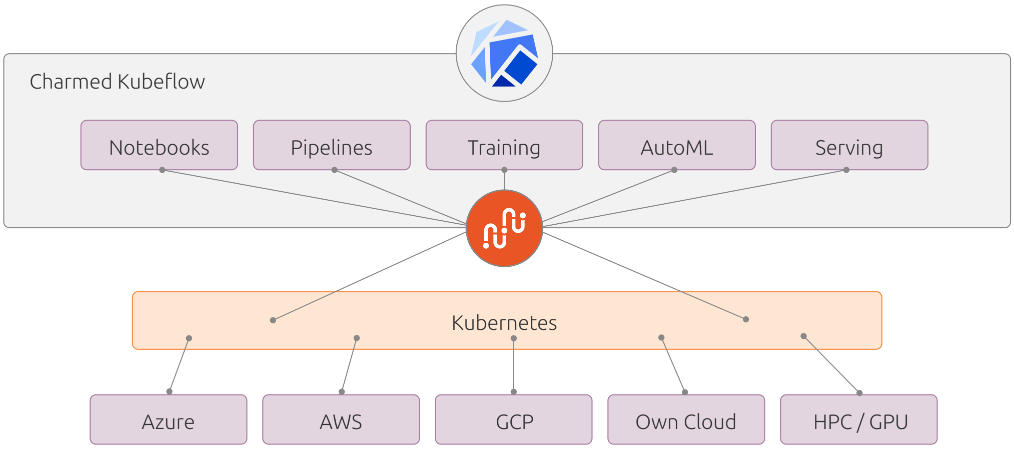
Enterprise support for open source tooling

While open source tools are ideal for several reasons, when it comes to enterprise-grade solutions, having experts that offer support, ensure security and provide guidance accelerates machine learning projects and allows data scientists to focus on modelling rather than debugging. Thus, many companies distribute open source tooling, by providing patches, upgrades, and updates on top, in order to protect the data and model. They provide support, answering issues in a timely manner, and often contribute to open source initiatives.

Charmed Kubeflow

Charmed Kubeflow is a production-grade, end-to-end open-source MLOps platform that translates steps in the data science workflow into Kubernetes jobs. It is one of the official distributions of the Kubeflow upstream project.

Using it, data scientists and machine learning engineers benefit from having ML deployments that are simple, portable and scalable. It has a wide range of capabilities, from experimentation using Notebooks to training using the Kubeflow Pipelines or tuning using Katib.



Charmed Kubeflow allows faster project delivery, enables reproducibility and uses the hardware to its fullest potential. With the ability to run on any cloud, the MLOps platform is compatible with both public clouds, such as AWS or Azure, as well as private clouds. Furthermore, it is compatible with legacy HPC clusters, as well as high-end AI-dedicated hardware, such as NVIDIA's GPUs or DGX.

Charmed Kubeflow is a suite of tools that covers all the steps of the machine learning lifecycle. It helps users build ML models, analyse model performance, tune hyperparameters, manage computer power, deploy models to production and finally monitor them. Because it can run on hybrid cloud environments, it's ideal for organisations that don't want to be locked in with one cloud vendor and gives them the flexibility to go with the cloud of choice for each use case. This also enables them to address compliance and data regulations more easily.

The MLOps platform covers the entire machine learning lifecycle, often defined as a suite of tools. Charmed Kubeflow is a modular solution and is able to decompose into different applications, such that professionals can run AI at scale or at the edge.

Charmed MLFlow

Charmed MLFlow is an open source platform, used for managing machine learning workflows. It is a distribution of the upstream MLFlow project. MLFlow has four primary functions that include experiment tracking, model registry, model management and code reproducibility. Like Charmed Kubeflow, Charmed MLFlow runs on any environment, including hybrid or multi-cloud scenarios, and on any Kubernetes.

With over 10 million downloads in 2022, MLFlow evolved over time and now has four main components:

- **MLFlow tracking** is the heart of MLFlow, used primarily for experiment tracking. It consists of an API and user interface that records data about machine learning experiments, allowing data scientists to query them. The UI is helpful to visualise intermediary results for each run, as well as the experiment as a whole.
- **MLFlow projects** are used for code packaging within data science projects. They

enable code reusability, making it easier to reproduce experiments.

- **MLFlow models** offer a unit for packaging models. They build a standard that enables users to reuse machine learning models.

- **MLFlow model** registry is used for model management, providing a centralised place to look after models and their lifecycle.

MLFlow Components:

Tracking	Projects	Models	Model Registry
Record and query experiments: code, data, config. and results	Package data science code in a format that enables reproducible runs on any platform	Deploy machine learning models in diverse serving environments	Store, annotate and manage models in a central repository

Kubeflow vs MLFlow

Both Kubeflow and MLFlow are open source solutions designed for the machine learning landscape. They received massive support from industry leaders, as well as a thriving community whose contributions are making a difference in the development of the projects. The main purpose of both Kubeflow and MLFlow is to create a collaborative environment for data scientists and machine learning engineers, to develop and deploy machine learning models in a scalable, portable and reproducible manner.

However, comparing Kubeflow and MLFlow is like comparing apples to oranges. From the very beginning, they were designed for different purposes. The projects evolved over time and now have overlapping features. But most importantly, they have different strengths. On the one hand, Kubeflow is proficient when it comes to machine learning workflow automation, using pipelines, as well as model development. On the other hand, MLFlow is great for experiment tracking and model registry. From a user perspective, MLFlow requires fewer resources and is easier to deploy and use by beginners, whereas Kubeflow is a heavier solution, ideal for scaling up machine learning projects.

Overall, Kubeflow and MLFlow should not be compared on a one-to-one basis. Kubeflow allows users to use Kubernetes for machine learning in a proper way and MLFlow is an agnostic platform that can be used with anything, from VSCode to JupyterLab, from SageMake to Kubeflow. The best approach is to integrate Kubeflow and MLFlow and use them together - provided the layer underneath is Kubernetes. Charmed Kubeflow and Charmed MLFlow, for instance, are integrated, providing the best of both worlds.

Canonical MLOps

The landscape of open source ML tooling is evolving quickly, with new frameworks, libraries and even models available. Many of them are already integrated in bigger projects such as Charmed Kubeflow, including Jupyter

Notebooks, NVIDIA Triton Inference Server or KServe. However, keeping up with the latest changes can be challenging.

Ideally, when building an architecture, it is crucial to have tools that cover the entire machine learning lifecycle. Using fewer tools creates less friction and a lower chance to bump into compatibility problems. Integrations are essential and using proven solutions will help you accelerate project delivery.

Canonical MLOps offers an end-to-end fully open source solution that includes integrations between tools such as Charmed Kubeflow, Charmed MLFlow or Charmed Spark. It enables professionals to focus on modelling, rather than tool compatibility. Charmed Kubeflow is the foundation of a growing ecosystem that is integrated with other tools, depending on the user's needs and validated on different platforms, including any CNCF-compliant K8s distribution and any cloud environment.

Charmed Kubeflow is orchestrated with Juju, an open source orchestration engine for software operators that enables the deployment, integration and lifecycle management of applications at any scale, on any infrastructure. It enables professionals to deploy and use only the components that they need from the bundle. The composability of Canonical's MLOps tooling is crucial when running machine learning in different environments since the needs to run at the edge and at scale are different. Whereas Kubeflow has about 30 components, at the edge, it is enough to deploy three: Istio, Seldon and MicroK8s.

How to manage the stack

Managing an MLOps stack at an enterprise level is challenging, especially when you consider all the different parts of the stack that you need to maintain. Companies struggle with various aspects, including tool integrations, version dependencies and user management. The machine learning lifecycle brings an additional layer of complexity, because of the different stages and the tasks that need to be fulfilled, such as data ingestion, feature store, model training or model deployment. Furthermore, security concerns and compliance requirements are often a blocker that enterprises face, especially in industries that are highly regulated like healthcare or the public sector.

Open source platforms such as Canonical MLOps are a suitable solution for enterprises which look for a solution to use for running machine learning projects in production. It offers a modular MLOps platform that is easy to deploy, supporting tasks such as scale in/out, upgrades and updates. Canonical MLOps has Charmed Kubeflow as the foundation of an end-to-end platform, where different tools such as Charmed MLFlow or Charmed Spark are plugged in and out depending on the use case.

Solutions offered as part of Canonical MLOps benefit from 24/7 enterprise support as part of Canonical's Ubuntu Pro subscription. Ubuntu Pro is a comprehensive subscription that ensures security patching and compliance with standards like FIPS or DISA-STIG. Enterprise support can be purchased to get phone and ticketing support, with response times from 1h for severity 1 to 12h for severity 4. Canonical MLOps with Ubuntu Pro + Support is suitable for organisations that have professionals who can manage the infrastructure used by data scientists. They need to have an understanding of MLOps, as well as the needs of their business users. The deployment of the platform is carried out by Canonical.

Managed Canonical MLOps

Since machine learning is a new area for many organisations and there is a shortage of specialised professionals, Canonical MLOps also provides managed services. In this case, Canonical's specialists fully manage the MLOps infrastructure to give enterprises the time and freedom to focus on building models and seeing them through to production. Managed MLOPs include 24/7 monitoring as well as high availability. Patching updates, upgrades or user management are just some of the activities that Canonical engineers will ensure are running smoothly.

Conclusion

The machine learning toolkit presented in this guide is composed mainly of open source solutions that cover the entire stack. It enables professionals to run the entire machine learning lifecycle, allowing them to automate workloads, develop machine learning models and deploy them. It addresses pressing problems that the industry faces, including security patching, upgrades or user management.

As AI/ML continues to mature and companies gain a better understanding of their AI readiness, there is going to be a shift in the market. Production-grade models will be the new standard and their outcomes will need to be better documented. AI is going to be more embedded within team activities, with specialised squads that develop models, manage the infrastructure underneath and liaisons who translate the business needs into technical projects. Seamless experiences as well as modular solutions will benefit from wider adoption since professionals will look for flexibility. Professionals can plug tools in and out depending on their use cases.

Canonical MLOps is the solution to run AI at scale, providing the underlying layer to deliver machine learning projects faster. It offers the same capabilities in any cloud environment and runs on any Kubernetes distribution, giving organisations the flexibility to build solutions based on their use case. Canonical can be your partner for the full stack.

Contact us today to build your MLOps solution

Read more about Canonical MLOps

- A guide to MLOps
- Getting Started with AI
- AI on Ubuntu

Canonical Limited, Registered in Isle of Man, Company number 110334C, Registered Office: 2nd Floor, Clarendon House, Victoria Street, Douglas IM1 2LN, Isle of Man, VAT Registration: GB 003 2322 47

Canonical