Arrikto

Whitepaper

Rok: Data Management for Data Science

At Arrikto, we are building software to empower faster and easier collaboration for data scientists working in the same or different clouds, in the same or different locations in a secure and efficient manner.

We live in the big data era. Each day we create about 2.5 quintillion bytes of data. The huge growth of data led companies to adopt machine learning techniques to process all this data for marketing, HR, sales, business intelligence and more. In addition, biotechnology, cybersecurity, autonomous cars, and many more industries rely on machine learning and artificial intelligence approaches to gain insights, make more accurate predictions, boost productivity, and grow their business.

Whether you use data science for weather forecasting, machine learning for medical image recognition, artificial intelligence for virtual personal assistants or medical research, you have one thing in common: you handle an enormous amount of data across multiple clouds and diverse locations.

Any machine learning workflow begins with data. Even before starting to use the data to train your model, you have to understand it, filter it, clean it, augment it, or pre-process it in some way. A dataset may include non-ASCII characters, duplicate records, it may miss some values or some headers in a CSV file, it may suffer from a combination of the above problems, which the data scientist has to fix before they can have a useful dataset.

Changes in data are common, and maintaining the correct and relevant test data when the underlying model changes is not an easy task. Again, data reconstruction may be necessary.

All this data manipulation and dataset changes require an easy way to track, version, and distribute your data, as well as keeping each dataset version associated with the corresponding code. Data Scientists, Data Engineers, and DevOps Engineers need a convenient way to discover the right piece of data anywhere, anytime, and connect it to their environment of choice, so they can manipulate it and produce new versions to be shared with their colleagues.

Arrikto has created two software products to make data-driven collaboration seamless:

Rok   Rok Registry

Use Rok and Rok Registry to share whole data science environments (code + libs + data) with hundreds of collaborators, across any location, on-prem or in the cloud.

Rok can snapshot, version, distribute, and clone your full environment along with its associated datasets. It is infrastructure-agnostic, so you can continue using your laptop, any public cloud, or your existing on-prem virtualization or container platform.

Rok Registry is the single pane of glass where you search, discover, and share your datasets and environments with other users. Users can create private or public groups on Rok Registry and can define fine-grained Access Control Lists. The Registry gives you full control of your data over individual users, locations, and devices, thus ensuring your sensitive data remains secure.

Data Scientists, Data Engineers and DevOps use Arrikto to iterate faster and easier, creating new collaboration workflows among teams, at global scale.

	Packaging	**Collaboration**
Code	Git	GitHub
Code + Libs	Docker	Docker Hub
Code + Libs + Data	**Rok**	**Rok Registry**

Figure 1: Rok and Rok Registry enable versioning and packaging of code, libs, and data and collaboration at global scale

Manage your data like code

Rok enables you to maintain many different, immutable versions of your whole environment along with its datasets. Every time you make a change to your data, you can take an instant snapshot of your current environment, and add user-friendly metadata. This way, you can roll back to any point in time, and instantly recreate your complete environment exactly the way it was.

By taking a snapshot of an individual dataset or a whole environment, you create a File or a group of Files on Rok. Each File has versions, and it may contain any type of data. Rok can handle multi-TB snapshots with ease.

Rok tracks the changes in Files and their versions, so when you need to train or test a model with an older dataset, you can simply restore an older version of the corresponding File, without having to move data around. The ability to go back and forth between data versions simplifies testing and training processes, and makes iterations blazing fast.

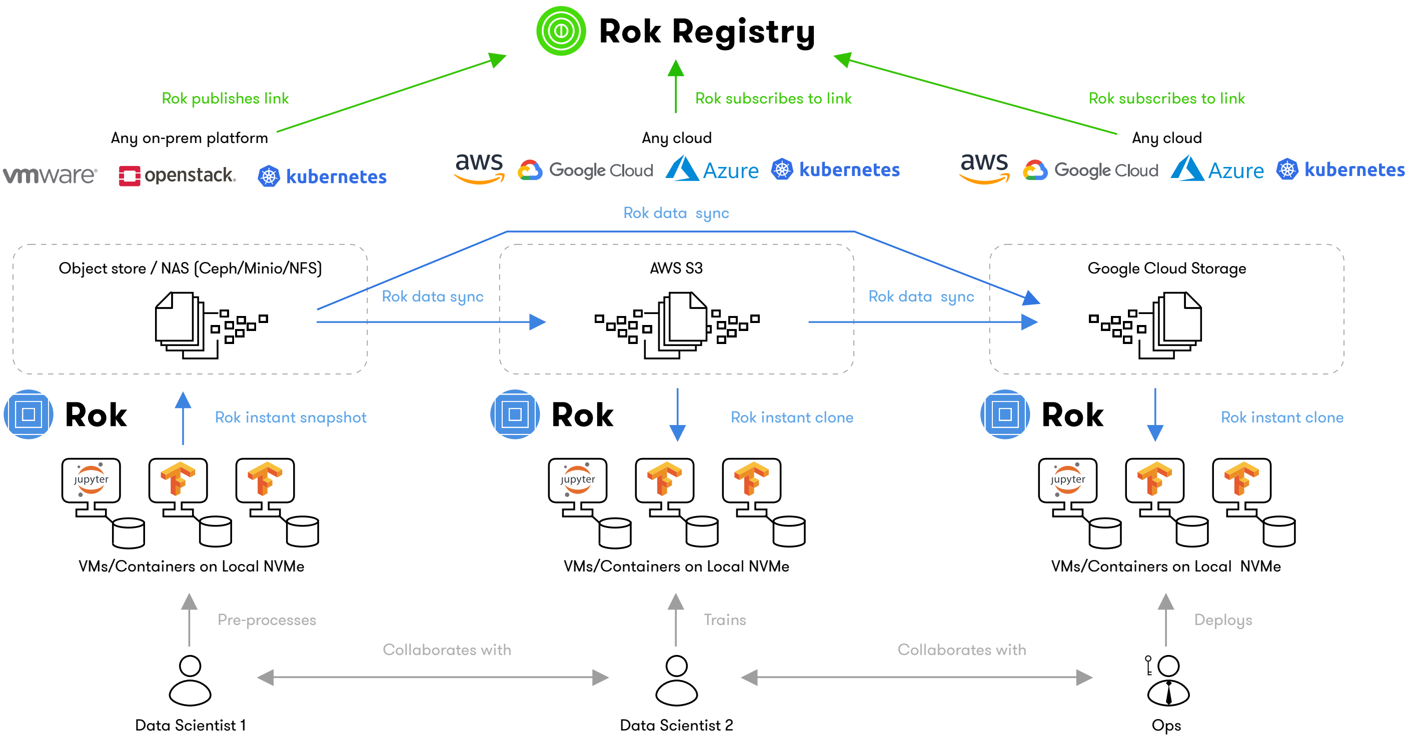

Figure 2: Easy collaboration among your team, across your infrastructure anywhere in the world

Reproducibility

For a colleague to be able to reproduce the results of your machine learning process and build on your work, they need a consistent copy of your code, libraries, data, and model, in the same place at the same time. Rok enables you to package everything together, so you or a colleague can deploy your whole machine learning environment to anyplatform, anywhere in the world. Instantly.

It doesn't matter whether you train and test your model on your laptop, on VMs, containers, a public cloud, or on-prem. You are able to reproduce your whole environment instantly.

Rok works at the infrastructure layer and doesn't impose any software or data format restrictions. Rok exposes file and block storage interfaces, so you can access it easily from your favorite data science tool chain.

Data Privacy and Security

Your data is shared over encrypted, point-to-point connections, which you can opt to run over your private network. Your data never crosses or gets stored at a central point. Moreover, you have full control over participating locations, devices, and users you are sharing your data with.

You can group the members of your organization into teams and define permissions for each team. Your organization and your teams are controlled by one ore more administrators, who can add and remove members, and set permissions for them.

Rok encrypts all data at rest. This way, your sensitive data remain safe during their whole lifecycle.

Deduplication

Rok assumes that you will be creating new versions of your full environment manually, whenever it is convenient, or automatically, at specific intervals; it can handle one version every few minutes with ease.

Since each snapshot of an individual dataset or whole environment is a File on Rok, and can be up to many TB in size, Rok detects the parts that have remained unchanged between snapshots and only stores them once, in a process called deduplication. This way, Rok makes the most efficient use of the underlying storage capacity, and keeps storage consumption low, even when it has to maintain a few hundred versions of multi-TB-sized environments.

Collaboration - Your data anywhere

Rok Registry is the single pane of glass where you search for, discover, and share environments with others in private or public groups.

Rok Registry is **not** a centralized store for your data. Unlike how GitHub stores your code, or Docker Hub stores your images, Rok Registry only contains *references* to your data. The actual data remains stored on Rok, at whatever location you choose: On your laptop, on one or more cloud providers, or on-prem. Thus, ensuring data privacy.

Rok Registry connects all Rok installations into the first decentralized network for sharing entire environments. Whenever you create new versions, they are shared directly among participating Rok installations over encrypted, point-to-point connections, over public or private networks.

You now get a single pane of glass to define policies, and a decentralized way to enforce them. With Rok and Rok Registry, you eliminate the bottlenecks of a centralized storage location and the headache of a single point of failure, while harnessing the power and efficiency of peer-to-peer data exchange.

Rok and Rok Registry make your whole environment reproducible and available anywhere, on any infrastructure, opening up a whole new way for teams to collaborate and iterate, faster and easier than ever.

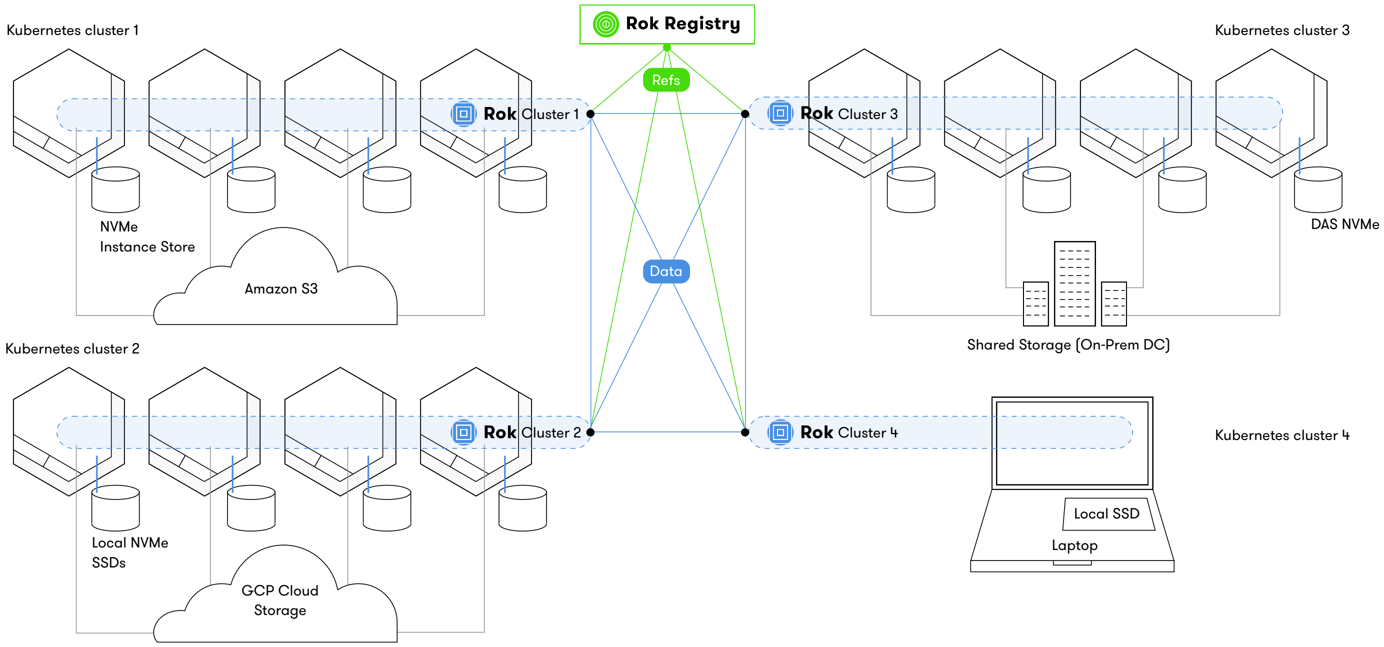

Figure 3: Rok and Rok Registry enable you to share your entire data science environment among locations over a decentralized network

Savings

Rok now makes it viable to run over super-fast, cost-efficient, ephemeral, local NVMe or traditional SSD storage without compromising on intelligent data services. You are now getting top performance in just a fraction of the cost of shared storage.

You can either take full advantage of the I/O bandwidth and latency of local NVMe to boost your performance, or consolidate your hardware for extra cost reduction.

Make the most out of your storage capacity by utilizing Rok's deduplication. Store only what has changed and get rid of duplicate data, resulting in up to 70% capacity savings depending on use case.

Arrikto

Arrikto Inc.
3505 El Camino Real, Palo Alto, CA 94306
www.arrikto.com

rev. 20181002-2