

# HOPSWORKS IN HEALTHCARE

**User-Friendly, Secure Machine Learning and Big Data**

**on low-cost commodity hardware**

# Hopsworks for Healthcare

Digitalization is sweeping through Healthcare creating a need to manage the huge waves of data coming from medical devices, DNA sequencing, and healthcare monitoring. Hopsworks is an open-source horizontally scalable platform for the storage and processing of sensitive healthcare data, with a unique multi-tenant security model and support for the most advanced machine and deep learning software.



Hopsworks comes with a user-friendly web interface for running, managing, and accessing data and programs and supports secure 2-factor authentication as well as integration with authentication systems, such as LDAP, ActiveDirectory, OpenAuth-2. Internally, all communications between services is encrypted with TLS/SSL and data can be encrypted at rest, providing Enterprise-grade security for your data. Hopsworks is designed to enable UI-driven, secure collaboration between Data Owners, who are responsible (according to GDPR) for managing access to sensitive data, and researchers who wish to process that data. By default, researchers are prevented from importing data into or exporting data from sandboxed projects - projects are a mechanism to help ensure that sensitive medical data can be stored in a shared cluster, but only made accessible to those individuals who need the data when they need it.

The screenshot shows the Hopsworks web interface. At the top, there is a search bar and a user profile for 'admin@hopsworks.ai'. Below this is a table of experiments with columns for Name, Metric, User, Start, End, State, and Actions. One experiment, 'fashion mnist autoML', is circled in red. A red arrow points from this row to a detailed view of the experiment. This view includes a 'Function' field set to 'grid\_search', a 'Direction' of 'max', and an 'Optimization key' of 'accuracy'. It also shows 'Experiment Results' with a table of hyperparameters and outputs.

Hyperparameters	Outputs					
dropout	learning_rate	loss	logfile	confusion_matrix	illustration	accuracy
0.6	0.0005	0.5995				0.6005
0.6	0.0003	0.5997				0.6003
0.6	0.0001	0.5999				0.6001
0.5	0.001	0.499				0.501
0.5	0.0005	0.4995				0.5005

# Dynamic Role-based Access Control

Multi-Tenant Studies with Projects, with secure Dataset sharing

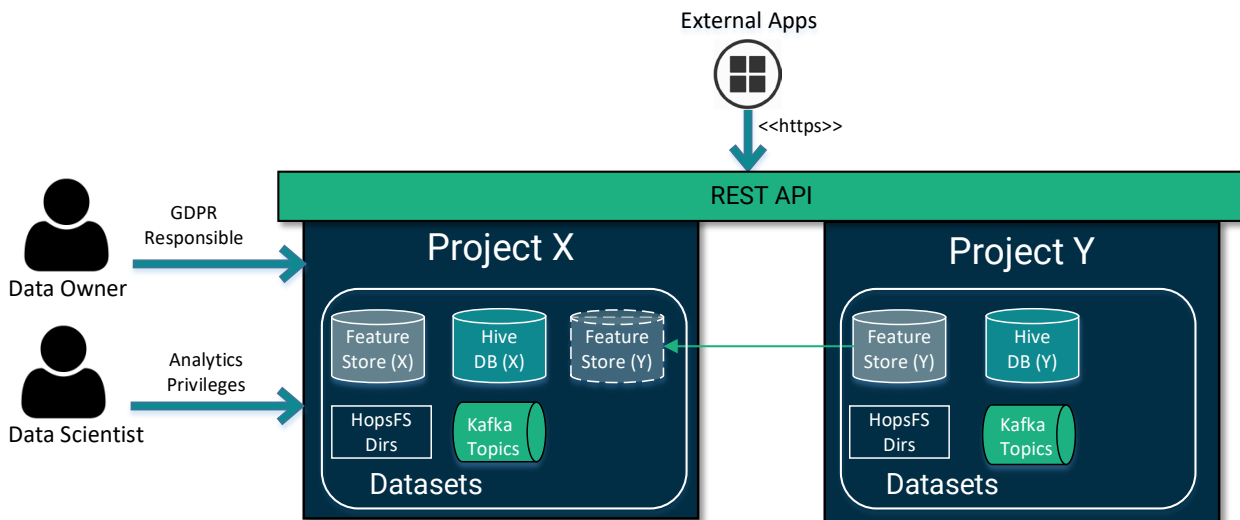


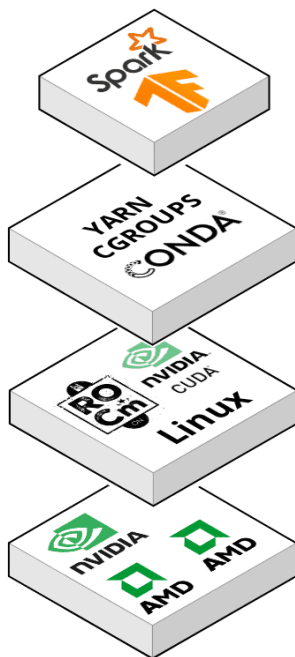
Figure 2: Projects and Datasets are first-class entities. Files, databases, feature-stores, and Kafka topics can be shared between projects.

Hopsworks provides a new GDPR-compliant security model for managing sensitive data in a shared data platform - a key requirement for medical data. Hopsworks’ security model is built around Projects, which are sandboxes for data, and enable a data owner to allow researchers process a dataset without that researcher being able to copy that data outside of the project or cross-link it with other data sources (outside of the project in question). This key capability is known in the security community as *Dynamic Roles*.

Projects contain datasets, users, and programs (code). Sensitive datasets can be sandboxed inside a project, and users can be assigned roles that prevent them from exporting data from the project. In Hopsworks, sharing data does not involve copying data. Datasets can still be securely shared between projects, without the need for duplicating the dataset. In Hopsworks, thanks to a unified metadata layer, a Dataset is not just a directory in HopsFS, but also a Feature Store, a Hive databases, or a Kafka topic. That is, databases, feature stores, and Kafka topics are all multi-tenant - they are private to a project, but can also be securely shared between projects. Hopsworks implements its project-based multi-tenancy security model by internally using X.509 certificates for user authentication, with a new certificate created for every user in a project. Hopsworks also provides role-based access control within projects, with pre-defined “Data Owner” and ”Data Scientist” roles, provided for GDPR compliance (“Data Owners” are responsible for the data and access to the data, while ”Data Scientists” are processors of the data).

# Commodity Hardware for Storage and Compute

**Storing large volumes of data and processing that data with lots of compute and GPUs (Graphical Processing Units) can be an expensive undertaking. Hopsworks is typically installed on commodity hardware and even commodity GPUs can be used for low cost Deep Learning.**



When companies and institutes purchase medical devices, such as NGS sequencing machines, additional money needs to be set aside to for the software and hardware costs, that can be significant if a traditional Enterprise IT company is involved. Hopsworks, however, is designed to work on on-premises hardware, and, in particular, commodity hardware. This can mean savings of 90% or more when purchasing PBs of commodity hard-disks compared to Enterprise storage applications. Similarly, when high-end GPUs are purchased for deep learning, the price can be significant. However, Hopsworks supports both Nvidia and AMD GPUs. For researchers, Nvidia commodity GPUs can be used for deep learning, while AMD have both lower cost Enterprise GPUs (MI50/MI60) and commodity GPUs (Radeon VII) that work seamlessly with TensorFlow. For example, switching out NVidia V100 GPUs for Radeon VII GPUs would result in savings of about 90%.

Commodity hardware does not mean less reliability, as Hopsworks and its underlying services provide high availability using software. So, if a host or hard disk or GPU dies, the Hopsworks platform will continue running as before, and the failed device can be replaced in the platform. Everything from the backend database (MySQL Cluster) to the filesystem (HopsFS) to the resource manager (YARN) to Kafka, Elasticsearch are all high availability services.

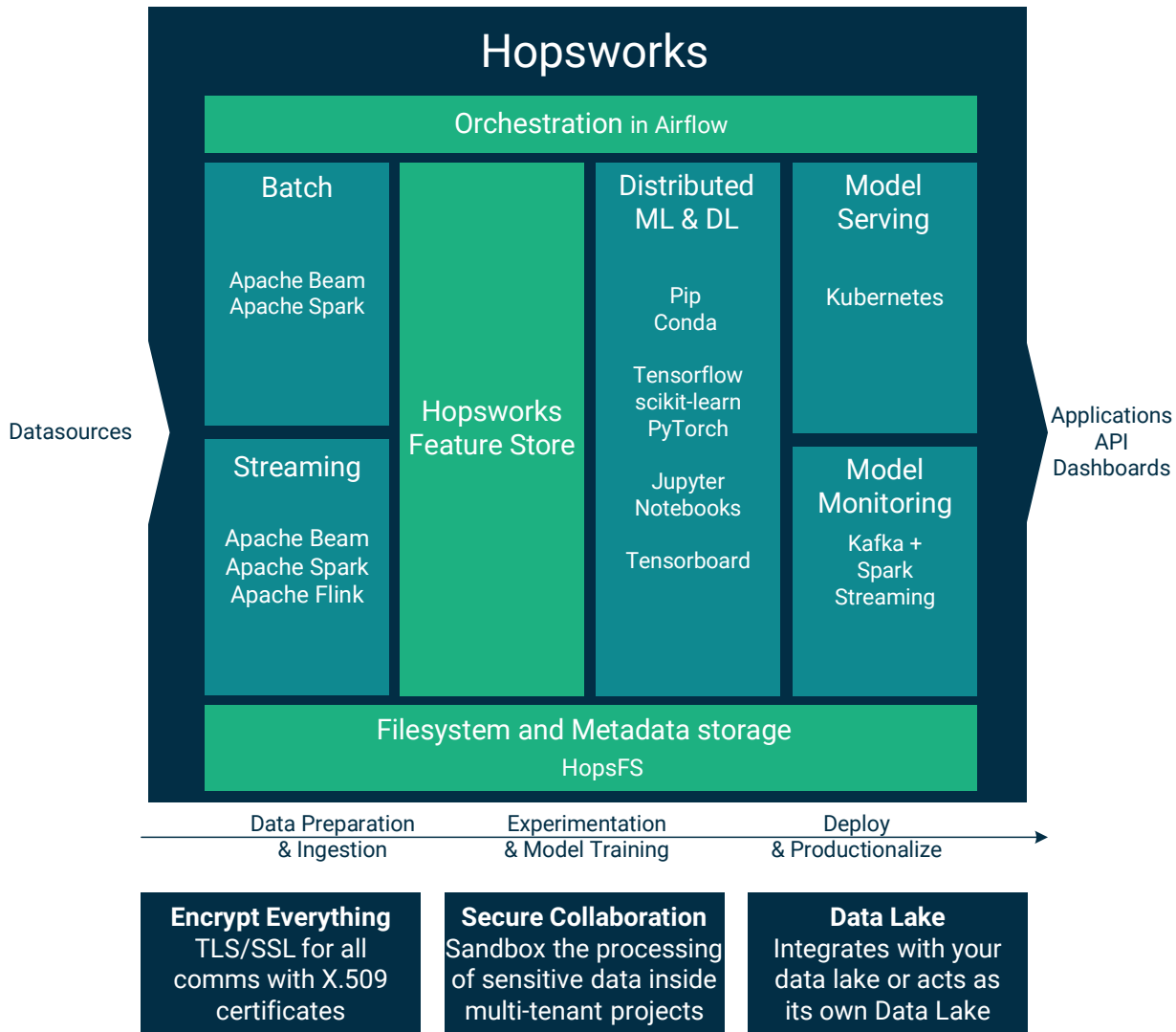


Figure 1: Data Lake, Analytics, and Machine Learning in a Single Platform

# Use Case: Hopsworks at Karolinska

Hopsworks is deployed at Karolinska in Stockholm, Scandinavia’s largest University Hospital, where it is used to store and process vast amounts of Next-Generation Sequencing data. The platform is configured using commodity hardware and designed to store PBs of data. The project-based multi-tenancy model enables data owners to sandbox sensitive research data and give selected researchers the ability to process the data, while ensuring that those researchers cannot cross-link data with other sources or export the data from the project’s sandbox. The [platform also enables Deep Learning](#), to extract patterns from genomic data.

LOGICAL CLOCKS

the creators of

HOPSWORKS

**Reach out to us!**

[www.logicalclocks.com](http://www.logicalclocks.com)

[info@logicalclocks.com](mailto:info@logicalclocks.com)

**.. or visit us at one of our offices:**

**Silicon Valley Office**

470 Ramona St  
Palo Alto,  
94301 California  
USA

**Stockholm Office**

Box 1263,  
Isafjordsgatan 22  
164 40, Kista  
Sweden

**UK Office**

IDEALondon,  
69 Wilson St  
EC2A2BB, London  
UK