A LOGICAL CLOCKS WHITE PAPER

HOPSWORKS IN HEALTHCARE

User-Friendly, Secure Machine Learning and Big Data

on low-cost commodity hardware

V.02

LOGICAL CLOCKS

HOPSWORKS
DATA INTENSIVE AI

Hopsworks is an open-source Enterprise platform for designing and operating machine learning (ML) pipelines at scale. You can write end-to-end ML pipelines entirely in Python and all pipeline stages can be easily scaled out to handle more data and progress faster. Hopsworks supports popular open-source frameworks for data engineering and data science, including ScikitLearn, Spark, Beam/Flink, TensorFlow, PyTorch. Hopsworks makes it easier for Data Scientists to write production-ready code, by supporting a Feature Store to ensure data quality and clean training data for ML models, and also by making Jupyter notebooks first-class citizens in the platform. Notebooks can be used to write production code that is run directly in ML pipelines.

Airflow can be used to orchestrate and operate the different stages in ML pipelines, while Hopsworks also provides support for HopsFS, the world's most scalable HDFS-compatible filesystem, with unique support for small files and high throughput.

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 health-care 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.

DYNAMIC ROLE-BASED ACCESS CONTROL

Multi-Tenant Studies with Projects, with secure Dataset sharing.

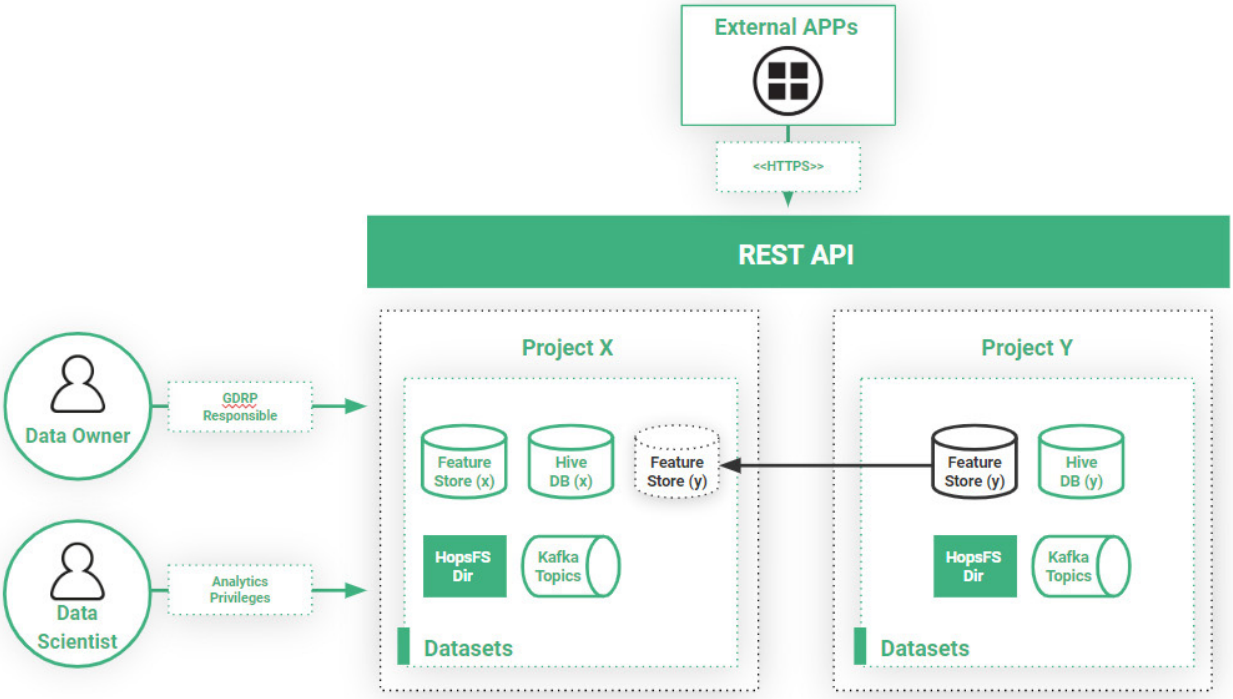

Figure 2:
Projects and Datasets are first-class entities. Files, databases, feature-stores can be shared between projects.

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 multitenant - 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 onpremises hardware, and, in particular, commodity hardware.

This can mean savings of 90% or more when purchasing PBs of commodi-

ty 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 availbility 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.

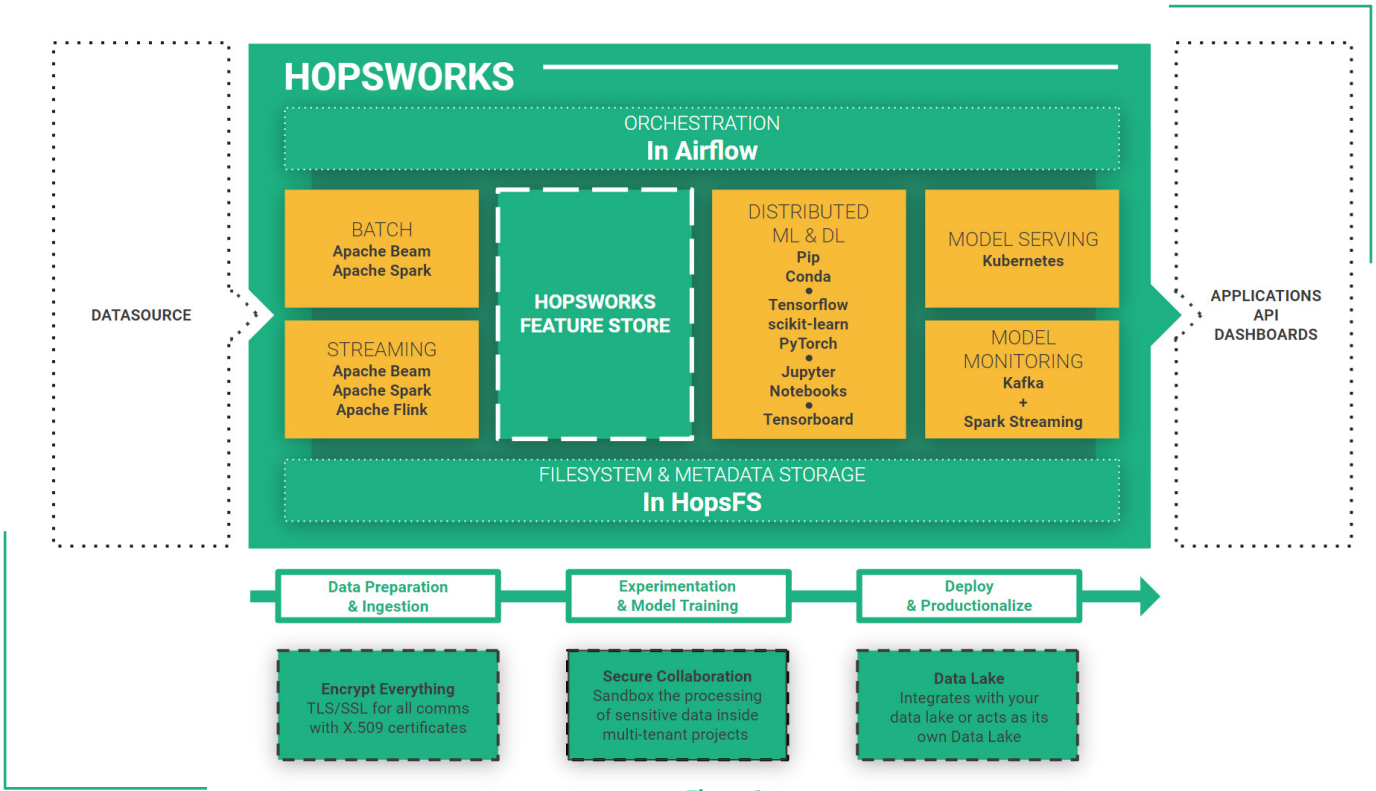

Figure 1
Data Lake, Analytics, and Machine Learning in a Single Platform

Hopsworks at Karolinska

Hopsworks is deployed at Karolinska institutet 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.**

Sequencing machines have generated 800+ TBs of next-generation sequencing data, requiring both low-cost storage and secure large-scale processing by researchers.

The organisation utilizes large-scale processing on Apache Spark and deep learning on TensorFlow to analyze these scale sensitive datasets to identify novel viruses, perform large cohort studies, and identify genetic mutations causing diseases. However:

Karolinska Institutet deployed Hopsworks to provide it with a secure and scalable platform to manage genomic data and perform secure research studies. Hopsworks is built around projects, providing a GDPR-compliant environment that enables secure collaboration between researchers on medical studies within a shared cluster.

Hopsworks is optimized for commodity hardware and runs on any data center. Clusters can be easily expanded by adding capacity, when needed enabling a low cost solution for up to PBs of data. Similarly, Hopsworks supports commodity or enterprise GPUs that can be used for deep learning.

Hopsworks' user-friendly web interface enables researchers to run, manage and access data and programs without software administration knowledge and skills.

www.logicalclocks.com
www.hopsworks.ai

Stockholm Office
Box 1263,
Isafjordsgatan 22
164 40, Kista
Sweden

Silicon Valley Office
470 Ramona St
Palo Alto,
94301 California
USA

UK Office
IDEALondon,
69 Wilson St
EC2A2BB, London
UK