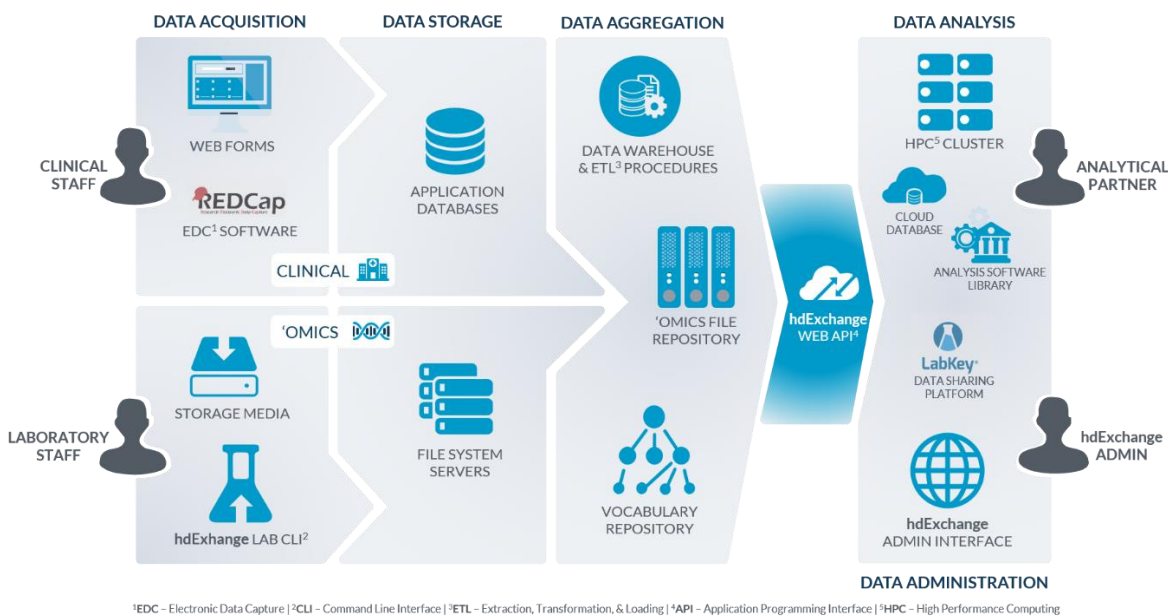


hiiData.Exchange Platform Overview

hiiData.Exchange (hdExchange) is a comprehensive platform for the acquisition, management, integration, analysis, and sharing of scientific data assets, particularly the Big Data associated with 'omics research. The **hdExchange Platform** sits at the center of the 'omics Big Data enterprise, making integrative 'omics research possible.

APPROACH

The **hdExchange Platform** is an innovative response to the unique challenges presented by 'omics Big Data. The **hdExchange** approach is based upon the core principle of *bringing the analysis to the data*. The Platform exposes an entire remotely accessible ecosystem of infrastructure, tools, and expertise in support of integrative 'omics Big Data analysis. The data and infrastructure exist in one place coupled with a robust platform to support remote data management and analysis. This makes it possible for top experts in diverse 'omics fields to collaborate in a shared space, leveraging shared resources to analyze the massive amounts of data involved.



PLATFORM

The **hdExchange Platform** is built on a foundation of shared storage and compute infrastructure comprising Petabytes (PB) of storage servers and hundreds of high performance compute nodes. This High Performance Computing (HPC) cluster environment exists to allow analytical partners to remotely install and run analytical pipelines on integrated clinical phenotypic and 'omics data. Layered on top of this infrastructure are the **hdExchange** components dedicated to data management and access. These components support each step of the data lifecycle: facilitating 'omics data transfer (**hdExchange** Lab CLI); transforming and warehousing clinical phenotypic data (Clinical Data Warehouse); structuring, cataloging, and storing 'omics data ('Omics Data Repository); exposing external resources like controlled vocabularies (Controlled Vocabulary Repository); and integrating all of these assets together with a single-point-of-contact for analytical partners (**hdExchange** API). A browser-based administrative user interface (**hdExchange** Admin Interface) is built on top of the **hdExchange** API to facilitate administration and curation of the data. Cloud database integration and a library of analysis software are also available in the HPC cluster environment to support development of analytical pipelines. A dedicated team with expertise spanning data engineering and bioinformatics is extended along with the Platform to support its utilization. Access to all data assets is secured through an integrated, token-based security model with data encryption in place both in transit and at rest. All identifiers included in the exposed data are dynamically masked to ensure research integrity and subject privacy. This combination of infrastructure, platform components, and expert support comprise the **hdExchange Platform** and the **hdExchange** approach to comprehensively supporting integrative 'omics research.

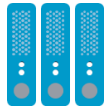
hiiData.Exchange Platform Overview

KEY COMPONENTS



Clinical Data Warehouse

- A combination of data models, data warehousing infrastructure, and ETL (Extraction, Transformation, Loading) procedures for housing complex clinical data from diverse sources.
- Leveraged by the **hdExchange API** in the integration of clinical and 'omics data served on the platform.



'Omics Data Repository

- Curated file system and centralized infrastructure dedicated to cataloging and managing 'omics Big Data.
- Includes over 2 PBs of storage arrays integrated within platform infrastructure components via a secure high-speed connection.



High Performance Computing (HPC) Cluster

- Advanced computing resource comprising hundreds of compute nodes, thousands of cores, and over a dozen TB of memory.
- Allows remote high-throughput, parallelized processing required for the complex and custom analytical software pipelines used for the analysis of 'Omics Big Data.



hdExchange API

- Primary mechanism by which investigators may programmatically access data assets for analysis within bioinformatics pipelines. Hides complexities of backend data management providing single point of contact for access to data assets.
- Secure web Application Programming Interface (API) that accepts requests for clinical datasets, 'omics file collections, and associated metadata and (given the requisite permissions) transfers the data/files/metadata to the configured environment.



Administrative User Interface

- Secure, web based user interface for platform administration.
- Allows data administrators to create and manage projects, users/groups, permissions, clinical datasets, 'omics file collections, and data releases.



Laboratory Transfer CLI

- Command Line Interface (CLI) utility for conducting transfer of 'omics files from the laboratory to the data repository.
- Supports implementation of custom rules and validation logic for file details and structure, presents these rules in straightforward way to users, enforces them via validation, and ensure data integrity via pre-transfer index generation.



Controlled Vocabulary Repository

- Managed repository maintaining and exposing controlled vocabularies via the **hdExchange API**.
- Allows investigators to directly access and integrate vocabularies within analytical pipelines.
- Examples include: RxNorm, SNOMED, InChI, ICD, CTCAE, and MedDRA

HIGHLIGHTS

- **Quantity of Data Managed:**
 - 'Omics: >500 TB of 'omics data
 - Clinical: >1k variables for >8k participants
- **Sources of Data:**
 - J laboratories with 28 total analytes
 - Genomics, proteomics, metabolomics, microbiome & more
- **Analytical Partners:**
 - J institutions with >50 platform users
- **Data Releases:**
 - >67 curated data releases
- **Research Supported:**
 - Major research networks, including TEDDY

ADDITIONAL FEATURES

- Additional **hdExchange Platform** features include:
- Data Sharing Platform (LabKey Server)
 - Data release management
 - Data asset cataloging and versioning
 - Dynamic identifier masking
 - Data access and user activity logging
 - Resource availability and utilization reporting
 - Cloud database integration

