The smart tips guide to high performance data and AI architecture

Find out how to get your data ready for precision medicine and experience new records for speed and scale through tips that are based on real-world use cases of high-performance genomics and imaging.

IBM

Finding a needle in a haystack

Every healthcare and life sciences organization strives to be data-driven, resulting in groundbreaking research to support the development of new diagnostics and therapies.

But, what does it take to get there?

To become data-driven, organizations need to implement a holistic approach that help them:

– Deal with data that is growing fast—in volume, variety and complexity—siloed in systems that needs to be integrated.

– Automate the entire workflow to be able to analyze the data and make sense of it, while supporting rapidly evolving frameworks and applications.

From an IT perspective, this requires new records for speed, scalability, flexibility and advanced data management capabilities that help enable collaboration, reduce costs and accelerate insights.

This guide lays out the critical functions your IT architecture needs in order to fully exploit the value of data and turn massive data sets into medically actionable insights. Having a powerful high performance data and AI (HPDA) architecture in place can keep your organization at the forefront of transforming biomedical research and advancing precision medicine.

Smart tip 1:

Create a Data Hub to manage the ocean of data

The five critical functions your HPDA Data Hub should have:

Landing data

Data flows out of every instrument—faster and bigger each day—creating massive quantities of mostly unstructured data that needs to be stored and processed fast. Store and analyze data coming at tremendous speed and volume from various sources with a solution that provides extreme scalability and performance, helping you avoid a situation that data preparation is becoming a bottleneck.

Tiering data

Automate data placement by moving data to the optimal storage tier as quickly as possible, depending if it requires high-performance storage or needs to be archived for long-term retention. This helps reduce capital and operating expenses with a common pool of storage for modern workloads with access to both file and object storage.

Sharing data

Empower geographically distributed organizations with a single global namespace that allows researchers to share, access and analyze data with minimal time delay, in a secure manner for maximum collaboration and productivity. This allows data to be available to anyone that needs it—anytime, anywhere.

Peering data

Replicate your current data to allow continued access to it from anywhere—remotely, from the cloud, from other sites—without having to make large copies of data. This strategy helps you avoid the inefficiencies and high costs that are associated with data duplication.

Cataloging data

Track different data types and the full lifecycle management of data that is growing fast and is siloed in many locations and projects. Label and classify massive data collections so you are able to find the exact data that you are looking for, quickly and easily. With a simple dashboard, you can efficiently use the data for analysis.

Smart tip 2:

Create an Orchestrator to put guard rails around the jungle of applications

The five critical functions your HPDA Orchestrator should have:

Parallel computing
Deploy intelligent workload and policy-driven resource management to optimize computing clusters across the data center, on premises and in the cloud, for faster time to insights and better outcomes. This helps enable hundreds of concurrent users to run millions of jobs in parallel on thousands of nodes with no downtime and no disruption to users or applications.

Platform as a Service
Enhance the high-performance computing infrastructure with analytics frameworks that are adaptable to key healthcare and life sciences workloads (genomics, imaging, clinical), such as using SPARK as a platform to analyze data and also generate quarries to integrate data.

Cloud computing
Take on-premises workloads and burst them into the cloud elastically so you can handle unpredictable requirements and avoid bottlenecks that might cause some jobs to run slower. Benefit from multi-cloud workload orchestration to achieve faster insights and costs savings.

Pipelining
Build a genomics or imaging analysis pipeline based on a variety of different applications and integrate them into complex workflows you can manage efficiently with a shared compute infrastructure for faster time to results.

App Center
Manage your environment with a graphical user interface tool that provides a flexible, easy-to-use interface for cluster users and administrators. This tool allows you to see in real time the status of your applications and interact with intuitive, self-documenting standardized interfaces.

Smart tip 3:

Make the Data Hub and Orchestrator work together: Data Hub + Orchestration = high performance, low cost, ease of use, collaboration

Multi-cloud orchestration by workload

Place data where it makes the most sense for your organization to run its applications and services by making the computing resource data-aware. When a compute workload is dispatched to another location such as different site or public cloud, the Orchestrator works with the Data Hub to move the data in sync.

Zero downtime

Prevent the IT infrastructure from getting overloaded and potentially crashing by making the computing resource IO-aware. Just before the infrastructure gets overwhelmed, the Data Hub knows to detect this trend and works with the Orchestrator to balance and control the dispatching of workloads to maximum performance and ensure business continuity.

Take the next step—Check out these available resources

Watch this webinar and learn how these tips are helping world-class organizations to advance precision medicine: Webinar: Experience New Records for Speed and Scale: High Performance Genomics & Imaging

Meet our reference architecture for high-performance data and AI in healthcare and life sciences based on software-defined infrastructure: Video: High Performance Data Architecture for Healthcare

Start for free!

Try IBM® Spectrum Scale™: experience the extreme scalability and reduced storage costs today with a free 30-day trial

Try IBM Spectrum™ LSF: ensure that the right resources are automatically allocated to the right jobs, to reduce costs and maximize performance:

For more information and case studies visit:

Data storage for healthcare

ibm.com/it-infrastructure/storage/healthcare-industry