# NEW STORAGE ARCHITECTURE FOR THE ERA OF MODERN INTELLIGENCE

## Open Letter to the Storage Industry

September 12, 2018

Powered by analytics and AI, the era of modern intelligence has presented the storage industry with a unique opportunity. Data is the new currency and our opportunity is to be its steward. Yet historically we have actively held enterprises back from making progress with data. Legacy architectures, like data silos and data lakes, are built to lock data away, and can't do the one thing required to realize data's full value — share.

Data lake is dying. It was built on the obsolete premise that all unstructured data is meant to be stored. A new storage standard is needed in the post-data lake era. Modern intelligence requires an architecture designed not only to store data, but to share and deliver data. **We call this new architecture a data hub.**

The importance of putting data to work is easy to put into perspective. A recent study conducted by Baidu showed its dataset needed to increase by 10 million times in order to lower its language model's error rate from 4.5 to 3.4 percent[1]. That's 10,000,000x more data for one percent of progress! A luminary in the field of AI, Professor Andrew Ng from Stanford University, noted "data, not software, is the defensible barrier (competitive edge) for many businesses"[2] and enterprises must "unify their data warehouses."[3]

This emphatic call to unify data takes direct aim at the problem. Data is stuck in a complex sprawl of silos, and the storage industry is largely to blame for it. When an industry is so focused on developing technologies to store things, it naturally creates silos. But in today's data-first world, silos are counter-productive. Data is out of reach from modern applications that can drive insights and innovation.

It's time to rethink storage. A data hub is designed on first principles not only to store data, but to unify and deliver data. Unifying data means that the same data can be accessed by multiple applications at the same time with full data integrity. Delivering data means each application has the full performance of data access that it requires, at the speed of today's business. Data hub shatters legacy infrastructure barriers where applications are given their own silos and replicated datasets.

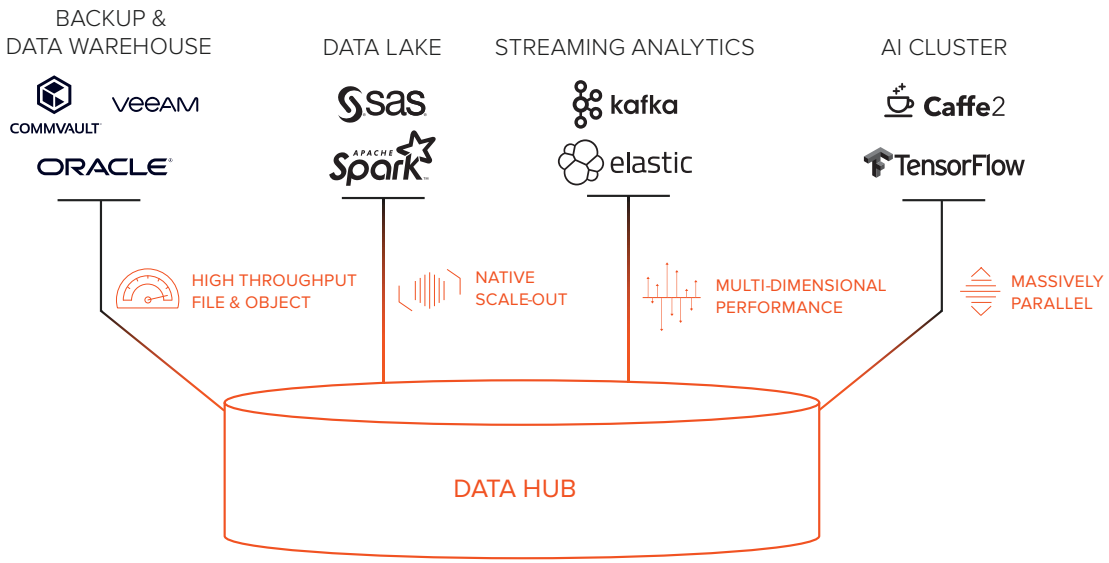
Data hub is a data-centric architecture for storage that powers data analytics and AI. Its architecture is built on four foundational elements:

- High throughput for file and object store
- Native scale-out design
- Multi-dimensional performance
- Massively parallel architecture

1 Deep Learning Scaling is Predictable, Empirically: https://arxiv.org/pdf/1712.00409.pdf
2 https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now
3 Nuts and Bolts of Applying Deep Learning, https://www.youtube.com/watch?reload=9&v=5PrvLq6_xm8

There are four classes of silos in the world of modern analytics: data warehouse, data lake, streaming analytics, and AI clusters. A data warehouse requires massive throughput. Data lakes deliver scale-out architecture for storage. Streaming analytics go beyond batched jobs in a data lake, requiring storage to deliver multi-dimensional performance regardless of data size (small or large) or I/O type (random or sequential). And AI clusters, powered by tens of thousands of GPU cores, require storage to also be massively parallel, servicing thousands of clients and billions of objects without data bottleneck.

Then there is cloud. Applications are increasingly cloud-native, architected on the premise that infrastructure is disaggregated and storage is limitless. The de facto standard for cloud storage is object.




A data hub must have all four qualities. All are essential to unifying data. A data hub may have other features, like snapshots and replication, but if any of the four features are missing from a storage platform, it isn't built for today's challenges and tomorrow's possibilities. For example, if a storage system delivers high throughput file and is natively scale-out, but needs another system with S3 object support for cloud-native workloads, then the unification of data is broken, and the velocity of data is crippled. It is not a data hub.

In this era, it's better to share and deliver data than to lock it away in silos, and systems built to share data are fundamentally different than those built to store data. Now is the time for the storage industry — Pure included — to deliver a new and modern architecture. We look forward to the rest of the industry embracing this opportunity as well.