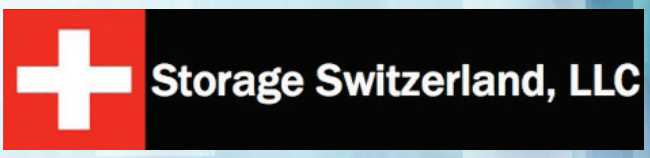
DEALING WITH THE AI AND ANALYTICS DATA EXPLOSION - MAPR BRIEFING NOTE

by George Crump



Accuracy and response time defines the success or failure of an AI or analytics project. The faster and more accurate the response the more trusted the system is. The more data provided to the project the more accurate it becomes. The problem is that data is outstripping compute; organizations need a way to store this data cost-effectively, yet still, have it reasonably accessible. The solution is often “add another cluster” which introduces more data movement and associated costs. What’s needed is a single cluster approach with a Global Namespace to provide a single view and point of access for all data available to AI and Analytics efforts.

Like most projects, AI and analytics projects have a subset of data that is the most valuable and most active. AI and analytics projects are adding new, continuously read data to the active tier. Additionally, like other projects, AI and analytics also have a larger set of data that is less active. However, unlike traditional projects, AI queries the less active subset more often than traditional data sets, and it requires a relatively rapid response to those queries.

Organizations need a multi-tier storage architecture that transparently moves data between various data centers and clouds while using different protocols to transmit and store the data on file systems like NFS, HDFS, and S3 / object storage. The hardware to fulfill these needs exists. Flash storage as the active tier, for example, can provide high-performance response times. High capacity hard disk drives can provide cost-effective but responsive storage, and the cloud can provide a deep archive that reduces data center floor space but still provides some responsiveness to queries.

The software is the missing element. There are storage systems that provide some intelligent data movement, but most require vendor specific hardware, don't scale to meet the AI and analytics demands and are too expensive for these projects. Additionally, most on-premises storage systems either entirely ignore the cloud or use it only for a mirrored copy of data; they can't archive older data to it to free up on-premises capacity.

> The primary value of a software-only HA and DR solution is the flexibility that it provides.

THE MAPR DATA PLATFORM

MapR provides an industry-leading data platform for AI and Analytics. The data platform is known for providing scalable, high performance for AI and Analytic workloads, so enterprises can analyze data and integrate the results into their businesses processes to increase revenue, reduce costs and mitigate risks. In the latest release of the data platform, MapR extends its data lifecycle capabilities to incorporate object storage and cloud storage.

With this new release, MapR customers can establish three tiers of access; frequently accessed, infrequently accessed and rarely accessed. Automatically moving data between these tiers based on policy, enables customers to deliver high performance and cost-effective storage, managed by a single solution without investing IT time to drive the process.

The frequently accessed tier is performance optimized. It typically uses flash storage for high-performance response to queries and searches. It even optimizes data protection for performance, using replication instead of a capacity savings technology like erasure coding. While erasure coding does consume more compute, it requires less capacity overhead. Moreover, with MapR's intelligent tiering, the high-performance tier is likely smaller than its competitors are.

The infrequently accessed tier typically consists of high capacity disk drives. It also uses erasure coding to optimize capacity utilization further while still being responsive. The primary purpose of the infrequent access tier is to provide acceptable response times while off-loading data from the all-flash frequently accessed tier to reduce costs.

The rarely accessed tier leverages object storage and can store data either on an on-premises object storage system or to the cloud. The goal is to reduce costs and leverage any cloud like Amazon, Azure or Google Cloud to reduce data center floor space.

The movement between tiers is a policy based waterfall model that moves cold data to each tier as the tier above it fills. Recalls are transparent and automatic based on access but also have the option to schedule a recall if a known data need occurs.

A crucial part of including the cloud in a data storage model is security. In this release, MapR also adds encryption for data that is in-flight and at-rest.

STORAGESWISS TAKE

AI and Analytics fueled in large part by IoT are driving the next wave of data center growth. AI, in particular, has a widespread appeal that most organizations want. The major stumbling block is data storage. Organizations that can meet the needs of storing the vast quantities of data required in a cost-effective, yet highly responsive manner are the ones that have a competitive advantage in our AI-driven future. The MapR Data platform is a strong contender to be the platform that resolves the cost-effective, highly responsive paradox.

ABOUT US

Storage Switzerland, LLC

Storage Switzerland is an analyst firm focused on the storage, virtualization and cloud marketplaces. Our goal is to educate IT Professionals on the various technologies and techniques available to help their applications scale further, perform better and be better protected. The results of this research can be found in the articles, videos, webinars, product analysis and case studies on our website storageswiss.com

George Crump is President and Founder of Storage Switzerland. With over 25 years of experience designing storage solutions for data centers across the US, he has seen the birth of such technologies as RAID, NAS and SAN. Prior to founding Storage Switzerland he was CTO at one the nation's largest storage integrators where he was in charge of technology testing, integration and product selection.