Whitepaper

FusionData with OceanStor Pacific: Solution to Serve as Secondary Storage and Data Lake Platform

Trademarks and Permissions

and other Futurewei trademarks are trademarks of Futurewei Technologies. Huawei trademarks are trademarks of Huawei Technologies Co., Ltd. All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure the accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

FUTUREWEI TECHNOLOGIES, INC.

Boston Research Center

Address: 111 Speen Street, Suite 114
Framingham, MA 01701
United States of America

Website: http://www.futurewei.com/

Contents

1 INTRODUCTION

In the past years, the concept of "secondary storage" caught people's attention. This secondary storage is not the same old concept where data are stored in drives whereas primary data are stored in DRAM. It is a newly invented term to contrast itself to today's enterprise storage (termed as primary storage). The main character of primary storage is that data stored there are for production usage. For example, if production database blocks or files are served by a set of storage arrays. Then these arrays are called primary storage. In contrast, the backups and extra copies of data are called "secondary storage" in this context.

The rising of the secondary storage concept is due to several reasons. First, even though many forms of new storage arrays are developed, they are not being trusted enough to be accepted into production usage yet. Therefore, the vendors are aiming to the backup, business intelligence, and big data use cases, where the availability requirements are usually lower than the production storage environment. Second, the backup and data analytic applications are looking for new ways to lower the TCO. Naturally, they are looking into alternatives that can provide similar services as the "primary storage" but with a lower cost.

The secondary storage concept is often related to software-defined storage (SDS). Since an SDS has the advantage of platform-independent, it can leverage its flexibility and the latest progression in cost reduction technologies. For example, when a cloud is providing a lower cost in some scenarios, an SDS system can be quickly deployed on the cloud to provide secondary storage.

2 SECONDARY STORAGE

In this chapter, we survey several companies that are growing in the secondary storage market. We use two products on the market to illustrate the pros and cons.
Secondary storage can be used to generate dev/test environment, data migration, etc. The term secondary storage refers more to the business model than to what kind of storage platform a vendor can provide.

Theoretically, a dev/test environment is not limited to dev and test. One can create a production environment in the same fashion as long as the storage system can sustain the IOPS, latency, and availability requirements.

2.1 COHESITY SOLUTION

As pointed in [1], the data protection issue of the big data environment is often ignored, while the big data environment is gradually replacing some of the traditional database functions. Cohesity provides data protection solutions for big data and NoSQL environments, such as MongoDB, Cassandra, Hadoop, Hbase, Cloudera, and Hortonworks [2].

Cohesity combines the concept of data protection (aka backup/restore), data usage (create test/dev environment), and data management (getting insight from data). In other words, positioned as secondary storage, Cohesity solutions are trying to help customers deal with different workloads in a non-production environment.

In this whitepaper, since we are discussing the big data ecosystem, we will focus on the dev/test environment building. Cohesity provides a fully functional storage system with a distributed file system called SpanFS [3]. This file system can provide NFS/SMB/S3 interface for hosts. Therefore, a dev/test environment can be created by exposing a writable point-in-time copy to the hosts.

2.2 RUBRIK SOLUTION

Rubrik has a data backup and recovery platform that can backup data on-premises, at the edge, and in the cloud. Rubrik also provides a data management system to manage the data saved in the system. Data are stored in a distributed file system called Atlas [4]. This cloud-scale file system uses Erasure Coding (EC) to provide high availability with high space efficiency.

Rubrik also supports NoSQL data protection with Rubrik Mosaic [5]. It can create backup copies of NoSQL databases (e.g., MongoDB, Cassandra, etc) and create dev/test environments.

2.3 OTHER VENDORS AND TRENDS

Many other vendors in the industry provide backup and restore features for big data and NoSQL environments. Gartner includes secondary storage vendors in its magic quadrant for data center backup and recovery solutions because many secondary storage vendors are famous for their backup and recovery capabilities. According to the magic quadrant diagram shared by Veeam [6], Veeam is the leader in this category while Cohesity and Rubrik are catching up in this category.

Figure 1. Magic Quadrant for Data Center Backup and Recovery Solutions

Source: Gartner (July 2020)

There are two types of vendors in this magic quadrant. One type contains software companies. Vendors in this type provide backup and recovery software, and users must provision storage from other vendors. They work with secondary storage vendors to backup enterprise data.

In contrast, vendors who serve as secondary storage companies (e.g., Cohesity and Rubrik) provide not only the backup and recovery software but also the storage platforms.

Vendors in this chart can partner with FusionData to provide value for customers. For example, customers can use their existing backup software to ship data to FusionData. Secondary storage vendors can provide an extra layer of protection for FusionData.

3 FUSIONDATA WITH OCEANSTOR PACIFIC: SECONDARY STORAGE AND DATA LAKE PLATFORM

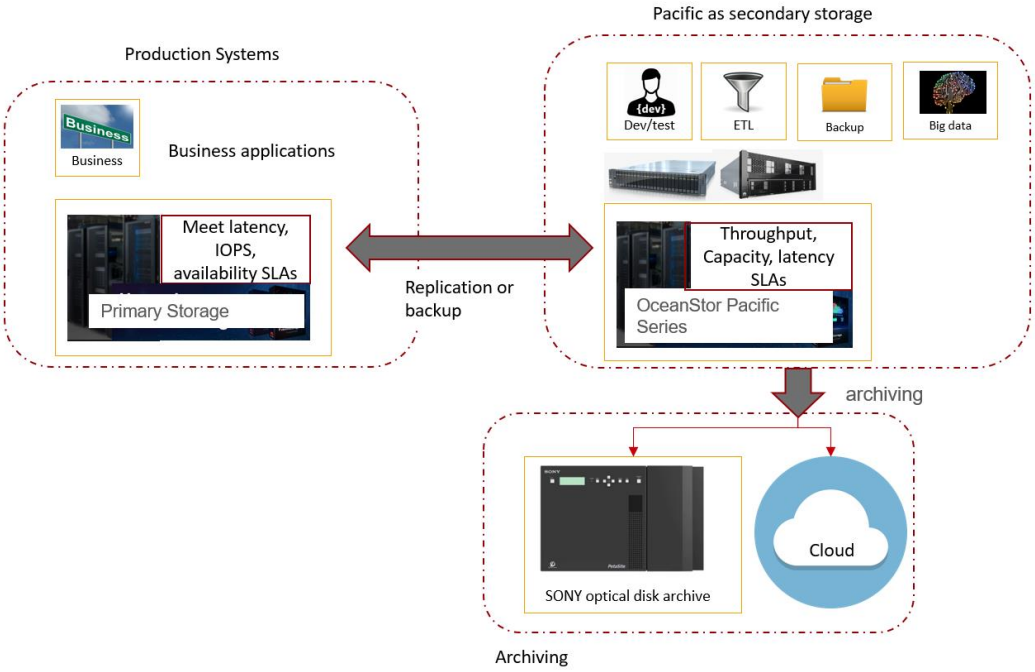
FusionData is Huawei's on-prem data analytics platform. It combines data analytics software (such as Spark, Hive, FusionInsight, etc) and Huawei's hardware to form a solution for both secondary storage and big data platform.

3.1 OCEANSTOR PACIFIC AS SECONDARY STORAGE AND DATA LAKE STORAGE

Huawei provides data storage for a massive amount of data via OceanStor Pacific Series [7]. OceanStor Pacific Series provides block, file, HDFS, and object services via high-density, high-performance nodes. Although the platform is capable of being used as primary storage for production applications, it is also suitable to serve the purpose of secondary storage as needed.

Due to its support of features like SSD/HDD mixture, dedupe, compression, erasure coding, snapshot, remote replication, and high-density nodes, the OceanStor pacific series takes the total cost of ownership into accounts. Therefore, it overlaps with the goals of secondary storage. While customers use other high-end systems, such as OceanStor Dorado, as primary storage, the Pacific series can serve as one-size-fits-all secondary storage to provide the rest of the storage services, including Backup/Data Ingestion(ETL)/Big data analytics/BI/Dev/Test.

Pacific is multi-protocol and can store massive data, so it already has the foundation of the characteristics of secondary storage. If the primary storage is OceanStor, replication links may be used. Otherwise, backup software (e.g., Veeam) can be used to back up point-in-time copies. The following diagram illustrates the Pacific series solution.



The production data is backed up or replicated to OceanStor pacific series via backup software or replication links. From this point on, all actions are happening on the pacific series as secondary storage.

Backup copies are saved in the Pacific, and they can be restored on demand. At the same time, the Pacific can serve as a data lake. Data can be processed for ETL, and queries can be executed with the HDFS interface. It can also push backups to long-term retention media or public clouds.

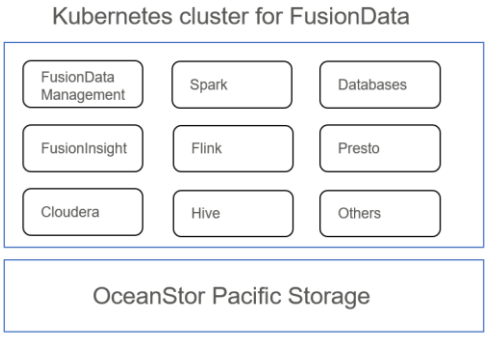
3.2 Requirements for Cloud-native FusionData

The software layer of FusionData combines several powerful data analytic tools: (1) the Cloudera/FusionInsight bundle with includes Spark, Hive, etc. (2) the optional relational DB component, and (3) a distributed query engine to query data from multiple data sources.

With the powerful capabilities comes the challenge of making the solution easy to use. For customers who wish to use their own data analytics suites, FusionData solutions should provide a way to allow customers to choose and pick what software to install. Luckily, with the progress of container orchestration platforms such as Kubernetes, the flexibility of software installation is greatly improved. On the other hand, with the progress of the open-source movement, the choice of software that is available is greatly increased.

As programs packed in containers are downloadable to multiple locations and platforms, being scheduled in Kubernetes clusters, the boundaries between compute/storage/on-

prem/public clouds are gradually blurred. While the community of Kubernetes is growing, a new generation of users is familiar with the cloud-native form of DevOps. The software layer of FusionData should allow the flexible configuration but at the same time ensuring the Enterprise quality that customers want.



Once FusionData's software layer is cloud-native, user's jobs and tasks can be seamlessly managed in a DevOps way and be relocated to anywhere the Kubernetes platform is supported.

The software layer will become customizable by users. Even though the software packages are pre-built and released, users can pick and mix with the other tools and applications that are useful. Kubernetes provides a flexible way for both vendors and users to configure their computing environment.

3.3 SUMMARY

FusionData with OceanStor Pacific can meet the demands in the industry to cover the customer needs for secondary storage integrated with data protection capabilities, and also cover the needs for a data lake and data analytics platform. At the same time, customers are switching to cloud-native applications and data analytic platforms. The requirements from users show that FusionData will be more flexible and easier to use with a Kubernetes-based software layer.

4 REFERENCES

[1] C. Bertrand and M. Leone, "Time for Prime Time: Effective Data Management," ESG whitepaper, 2019.

[2] Cohesity, "Integrated backup and recovery for NoSQL and Hadoop," [Online]. Available: https://www.cohesity.com/solutions/backup-and-recovery/nosql-and-hadoop/.

[3] Cohesity Whitepaper, "Cohesity SpanFS and SnapTree".

[4] A. Gee, "Introducing Atlas, Rubrik’s Cloud-Scale File System," Rubrik, 21 July 2015. [Online]. Available: https://www.rubrik.com/en/blog/architecture/15/7/introducing-atlas-rubriks-cloud-scale-file-system.

[5] Rubrik, "Rubrik Mosaic: Speed, data mobility and storage efficiency for NoSQL data protection," Rubrik, [Online]. Available: https://www.rubrik.com/content/dam/rubrik/en/resources/data-sheet/Rubrik-Mosaic-NoSQL-data-protection-data-sheet.pdf.

[6] Gartner, "2020 Gartner Magic Quadrant for Data Center Backup and Recovery Solutions, Veeam named a Leader for the 4th time," [Online]. Available: https://www.veeam.com/2020-gartner-magic-quadrant.html.

[7] Huawei, "OceanStor Pacific Series," [Online]. Available: https://e.huawei.com/us/products/storage/distributed-storage/oceanstor-pacific-series.