MEMORY
BIG DATA GROUP
AT SANDISK®

CASE STUDY

SanDisk's Memory Big Data Group Deploys InfiniFlash™ to Fish the Data Lake for Better Business Data Analysis

Solution Focus

- Big Data Analytics
- Cloudera/Hadoop
- Semiconductor/Data storage industry

Summary of Benefits

- Ability to store billions of rows of data in the data lake
- Sub-second performance
- 50X cost reduction compared to traditional IT data solutions

Product

- InfiniFlash™ System

"If companies don't tap into their data to be able to fine-tune their competitive edge in the marketplace, they will not be able to differentiate themselves from the market leaders. This is crucial for enterprises."

Janet George, Fellow and Chief Data Scientist, Memory Big Data, SanDisk

SanDisk®
a Western Digital brand

Summary

To harness the power of enterprise-wide data, SanDisk deployed InfiniFlash™ System and Cloudera Enterprise to implement a data lake and garner analytical insights for better business decisions.

The Challenge

Like many organizations, SanDisk wanted to harvest business insight from the copious amounts of data that was gathered throughout the organization on a daily basis. The Company needed to develop a Big Data strategy that a) included a data-centric platform architecture; and b) could unify data from siloes across the organization—data from disparate sources that were often not accessible except through an application or query language.

In addition, the strategy needed to be able to address the compatibility and management of the growth of Big Data into the future, while avoiding data stagnation, and expensive deduplication and redundant data storage. "What data scientists did many years ago, before flash memory solutions, was to use a very traditional storage mechanism, like a SAN or network file system," said Janet George, Fellow and Chief Data Scientist, Memory Big Data Group at SanDisk. "These legacy systems are not distributed; they are not scalable. They cannot handle very high amounts of data, like in the petabytes. They usually handle small amounts of data. We don't want to analyze small samples of data, which is what we used to do before machine learning. Now we can do analysis with very large sample sets."

While one priority was to unite the data to develop deep business insights, it was also imperative that the team preserve data lineage and unlock inaccessible data from traditional database systems. In addition, the team was exploring ways to encourage the organization to embrace first order and second order advanced analytics. Last, the team needed to plan for data growth and manage the complexity that was sure to arise from the growth scale of Big Data.

"This is similar to what other enterprises face. They need to combine structured, unstructured, and semi-structured data—including regulatory data, product data, government data, and partner data," George told us. "They need to think about how they are going to unite this data. In addition, every enterprise has policies about their data. They want their data to be secure. Our architecture for our data platform has to adhere to those enterprise data policies, such as security."

The SanDisk Memory team needed a solution that could enable them to draw intelligence from the large amount of data they were collecting in a timely and secure fashion.

> *“We picked InfiniFlash based on requirements. We could have selected any hardware we wanted. InfiniFlash is very competitive with respect to pricing, storage, performance and analytics. When you are extracting intelligence from data you first need to combine these data—structured, unstructured, and semi-structured. This is the use case for which we have deployed InfiniFlash as a data lake. This is the power of InfiniFlash.”*
>
> **Janet George, Fellow and Chief Data Scientist, Memory Big Data, SanDisk**

The Solution

To determine the best Big Data analytics solution for SanDisk, the team evaluated several options in the marketplace—Hadoop, MapR, Cloudera, and Hortonworks. Apache Hadoop is an open source software project that enables distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines with a very high degree of fault tolerance. Value is generated by the combination of fault tolerance with replication. Hadoop is cost-efficient for very large data sets and offers both scalability and storage flexibility. In addition, a self-healing capability ensures that when nodes fail due to replication, the system will run on other nodes.

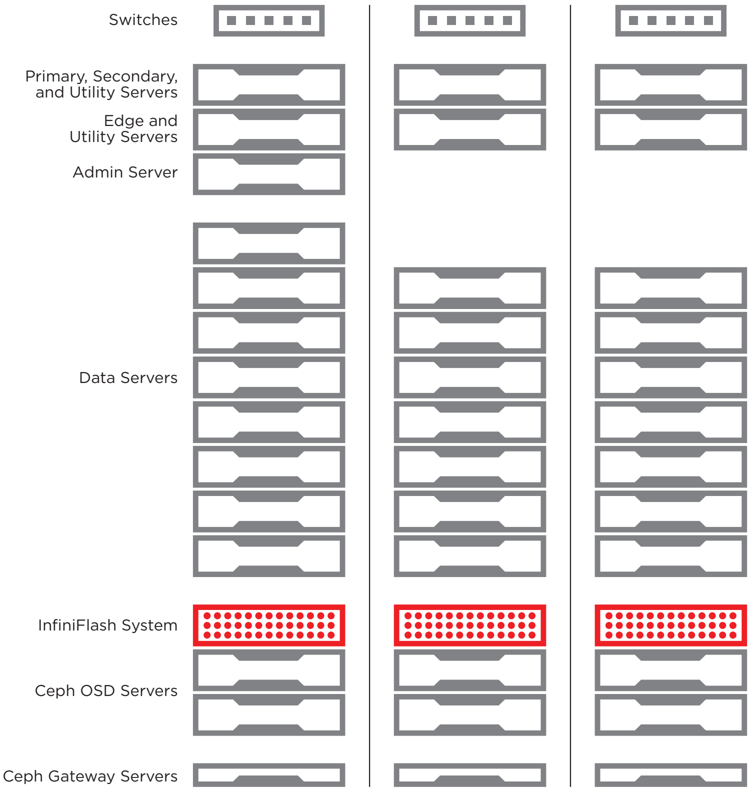
The key was to determine the appropriate enterprise data platform strategy and data-centric architecture for the long term flow of the data. Based on requirements, SanDisk chose to implement Cloudera Enterprise, which combines a high-performance system based on Apache Hadoop with enterprise tools to improve manageability and security at scale.

Embracing a Data Lake

“We learned through our implementation that embracing a data lake and the right storage solution for uniting all your Big Data sources is critical,” remarked George. “Although we initially did not implement a data lake, we eventually adopted the concept to unite our data and realize better security governance.”

The team evaluated every storage solution in the marketplace and eventually selected the InfiniFlash System from SanDisk because the system met their requirements for cost, performance, flexibility, distribution, and object storage. “We picked InfiniFlash based on requirements,” explained George. “We could have selected any hardware we wanted. InfiniFlash is very competitive with respect to pricing, storage, performance, and analytics. When you are extracting intelligence you first need to combine these data—structured, unstructured, and semi-structured. This is the use case for which we have deployed InfiniFlash as a data lake. This is the power of InfiniFlash.”

System Overview



The SanDisk team deployed three of SanDisk's 256TB InfiniFlash System units to serve as its data lake. The InfiniFlash™ System provides petabyte-scale capacity, high density, and high performance for OpenStack and Ceph environments, delivering breakthrough economics for customers with Big Data storage requirements, such as a data lake. The InfiniFlash System scales easily as each unit may connect up to eight servers, providing very low latency, extreme IOPS, and sustained throughput. Each system can be configured with up to 64 8TB hot-swappable cards delivering up to half a petabyte (512 TB) of raw flash storage in a 3U enclosure and up to 6PB in a single rack. The Ceph cluster software provides high availability, load balancing, data replication, and storage services required by OpenStack solutions.

"Deploying InfiniFlash was game-changing and transformational for us in our journey towards connecting the enterprise data for SanDisk," George told us. "We were able to pick and choose the different disparate data sources to unite into our data lake." InfiniFlash provided massive capacity in a highly scalable, high-performance, cost-effective architecture.

The team uses a Ceph Object Gateway to populate the data lake and pull data from there into Cloudera. "This is how we use InfiniFlash as a data lake within our enterprise data platform architecture," said George. "We use the object store mechanism within InfiniFlash to store all this data. So you can think of us ingesting data from many different data sources and putting them into InfiniFlash, which is our data store, and then within that data store we can access the data, transform the data, and load the data. Then, finally, that data will flow through to the Hadoop platform, where we can do data science-related activities like machine learning."

> *"Deploying InfiniFlash was game-changing and transformational for us in our journey towards connecting the enterprise data for SanDisk. We were able to pick and choose the different disparate data sources to unite into our data lake."*
>
> **Janet George, Fellow and Chief Data Scientist, Memory Big Data, SanDisk**

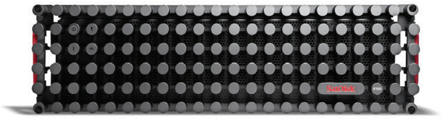
The Result

The team realized a number of positive results immediately. First, the InfiniFlash architecture provided increased data availability and decreased the effort needed to deploy algorithms for analysis. "It had been taking us up to eight hours to pull data out of the traditional tools," explained George. "We used to have to decompress files before we could pull the data. However, once we ingested the data into the data lake, the speed at which we could access that data was unprecedented within SanDisk."

Second, the team was able to retain the full lineage of any data wrangling and cleansing within the data lake. Users still have access to both raw and transformed data, details regarding the source of the data, and can execute complementary reporting (i.e., what happened) and predictive (i.e., what will happen) analyses.

The benefits of the data lake were also quickly realized. Storing data in a data lake is far more cost-effective than storing data in Hadoop, which has a replication factor of three. Although Hadoop is effective for data that are frequently used or for more frequent processing of data, a data lake is better for data that need to be found and retrieved on an as-needed basis. Cataloguing of the data in the data lake provides the ability to map data across different sources, the efficient tracking of cold, warm, and hot data, and the visibility of the data to different users. The data lake also provides a unified, central repository for all kinds and types of data sets—including structured, semi-structured, and unstructured data and internal, external, and partner data—that reside seamlessly without conflict.

"I am very happy to report that last week we had literally 2.8 billion rows of data—literally petabytes of data—flowing through our data platform and through our InfiniFlash data lake," said George. "That was a milestone for me because this much data flowing with sub-second performance, sub-second benefits in transforming the data, manipulating the data, looking at the data, querying the data—this is very much a record for us."



InfiniFlash™ System

InfiniFlash offers a number of unique benefits when used as a data lake. The architecture provides a rapid, automated ingest framework with low latency and the ability to prioritize advanced analytics seamlessly. “InfiniFlash enables us to work with as much data as available. With InfiniFlash we are able to handle petabytes of data in a distributed fashion. Data can be worldwide in different geographic locations. Large data sets alleviate problems with anomalies that can be seen with smaller data sets,” explained George. “More data allows you to build more accurate models. Overall, a weak assumption coupled with complex algorithms is far less efficient than using more data with simpler algorithms. Essentially, more data beats better algorithms. Without flash and a system like InfiniFlash it simply would not be possible to obtain true data analysis that allows us to make reliable business decisions.”

Economics of the Solution

In past years, when companies wanted to conduct data analysis, they were constrained by their IT architecture. Databases and platforms were more expensive, and memory was not as inexpensive as today. “We did not have InfiniFlash and economies of scale many years ago where we could build such high flash memory systems,” said George. “Today, because of technology advances like InfiniFlash we are able to tackle Big Data. With InfiniFlash we can come up with an enterprise architecture that allows the insights of data to flow and enable third order and fourth order level of analysis.”

The InfiniFlash System delivers the performance of an all-flash array with the economics of an HDD-based system. According to George, a system that used to cost $50 million can now be matched by a system for less than $1 million. “We put the entire infrastructure in place in a co-location without any major help from the IT department. The data scientists and Hadoop developers and high performance computing team were involved. Next year, with some additional budget, I can add nodes to it and literally double my infrastructure. That is the disruption that is happening in the industry today.”

Outlook

In the future, George anticipates that there will be a more emphasis on scale and performance. “Right now, enterprises are using data for analysis and reporting because their infrastructure does not provide for extraction of intelligence from the data. Where we are going is a place where the infrastructure is going to enable you to extract intelligence from the data. Also, if you are going to process petabytes of data worldwide, you are going to need an infrastructure that enables you to do that. That is the use case. You want to extract deep intelligence. That is the future.”

George is excited about what is in store for data science and the analysis of data. “We can look at things that we couldn’t look at before. We can look at the topology of data and derive results based on shapes and correlations. Machine learning can help us point to root cause issues. We often chase one problem and start to do analysis at scale and then look at the data from a multi-dimensional standpoint. We discover things from the data that we didn’t know we should be looking at in terms of correlations and root cause. The livelihood of companies will depend on using intelligence from data, for competing in the marketplace, and in the future retaining their leadership position.”

More frequently than ever, companies are thinking about their enterprise data strategy. “Enterprises are in discussions about the data lake, confirmed George. “InfiniFlash is part of this strategy. A traditional data storage infrastructure will not necessarily help them. However, if they don’t tap into their data to be able to fine-tune their competitive edge in the marketplace, they will not be able to differentiate themselves from the market leaders. This is crucial for enterprises.”

Contact information

datacentersales@sandisk.com

Western Digital Technologies, Inc.
951 SanDisk Drive
Milpitas, CA 95035-7933, USA
T: 1-800-578-6007

Western Digital Technologies, Inc. is the seller of record and licensee in the Americas of SanDisk® products.

SanDisk Europe, Middle East, Africa
Unit 100, Airside Business Park
Swords, County Dublin, Ireland
T: 1-800-578-6007

SanDisk Asia Pacific
Suite C, D, E, 23/F, No. 918 Middle
Huahai Road, Jiu Shi Renaissance Building
Shanghai, 20031, P.R. China
T: 1-800-578-6007

For more information, please visit:
www.sandisk.com/infiniflash

SanDisk®
a Western Digital brand

At SanDisk, we’re expanding the possibilities of data storage. For more than 25 years, SanDisk’s ideas have helped transform the industry, delivering next generation storage solutions for consumers and businesses around the globe.