NetApp®

Datasheet

Run Hadoop Analytics on Existing NFS Systems Without Moving Data

The simple yet powerful NetApp NFS Connector for Hadoop brings big data analytics to existing data

KEY BENEFITS

- Run analytics natively on data stored in NFSv3 without creating additional infrastructure.
- Combine Hadoop capabilities with enterprise-level data management.
- Benefit from a lower-cost and more efficient way to run analytics.
- Install and configure quickly and easily
- Use only one copy of data for Hadoop and still have high availability
- Compatible with leading open source analytical tools

The Challenge

Managing enterprise data alongside data analytics

In today's world, most enterprises receive data in large or varied amounts, or very quickly. Using this big data is essential for gathering insights and for operational efficiency. This new data is also in addition to the large and diverse data that an enterprise already has on its existing NFS-based storage.

To manage and gather insights from both sources of data, businesses have turned to Apache Hadoop, an open-source ecosystem of products and technologies that can ingest, store, and analyze big data. However, a significant challenge in leveraging analytics frameworks such as Apache Hadoop is the need to create a cluster with dedicated infrastructure to ingest all the data. This can create another data silo and increase the time to results. Managing these data silos can be difficult, inefficient, and costly.

The Solution

NetApp® NFS Connector for Hadoop enables big data analytics to run on NFSv3 data. The NFS connector allows users to analyze NFS data without moving the data into the analytics cluster, saving both time and space. Without the need to copy and manage data across different silos, IT administrators and operations can support Apache Hadoop analytics without additional storage hardware, and data workflows are simplified. The NFS connector is ideal for running a proof of concept or pilot for Apache Hadoop. For QA testing, new Apache Hadoop environments can be quickly started by leveraging NetApp FlexClone® technology.

Runs Analytics Natively on Existing Storage

With NetApp NFS Connector for Hadoop, users can immediately analyze data on existing NFS-based storage, such as NetApp FAS storage arrays. A dedicated cluster is not necessary to analyze semi-structured and unstructured data such as text files, log files, source code, and images. In fact, the NetApp NFS Connector allows NFS storage to be used as: (1) a secondary file system, in which Apache Hadoop uses HDFS for its primary storage and uses NFS for secondary storage; and (2) a primary file system, in which Apache Hadoop runs entirely on NFS storage. Users can read and write data between HDFS storage and NFS storage, enabling easy data sharing between storage that runs either file system.

Figure 1 shows how the NFS connector fits into the Apache framework. Both NFS and HDFS can sit alongside each other in the same cluster, or Hadoop

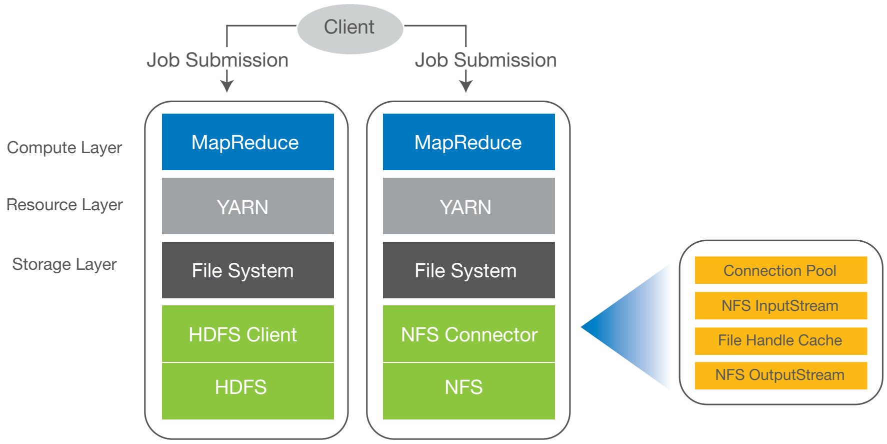

Figure 1) With NetApp NFS Connector, both NFS and HDFS can sit alongside each other on the same cluster, or NFS can replace HDFS so that Hadoop analytics can be performed on NFS data. It also enables users to read and write data between HDFS and NFS.

analytics can run on a cluster that has only NFS. And this happens without any modification to applications or jobs.

Offers a Single Copy of Hadoop Data with High Availability

NetApp NFS Connector for Hadoop decouples storage from compute, thereby allowing several optimizations. First, it allows analytics on data stored on other file systems, such as NFS. Second, it improves storage efficiency by leveraging existing technologies such as NetApp RAID DP® and SnapMirror® for data protection (rather than the three copies used by HDFS). Third, it allows the use of all the data management features of the NetApp Data ONTAP® operating system, such as deduplication, FlexClone volumes, NetApp Snapshot® copies, and high availability.

Open Implementation Promotes No Lock-In

The connector works with any NFS-based storage, and it has no proprietary features designed specifically for clustered Data ONTAP. It is fully open source and is hosted on GitHub, and NetApp plans to contribute the code to mainline Hadoop. NetApp has been an innovator in NFS, pioneering NFS standards to advance file-based storage access in UNIX® and Linux® environments. NetApp was also the first to market with pNFS support for NAS. Our storage solutions come tested against the leading NFS RFC standards.

Implementation Is Easy

Installation of the NetApp NFS Connector for Hadoop is as simple as changing an entry in a configuration file. Users make no change in applications written for Apache Hadoop. For Apache Hadoop, the change is made in `core-site.xml`; for Apache HBase, the change is made to `hbase-site.xml`, and similar changes are made to other projects. This simple change allows applications and tools such as MapReduce, HBase, and Spark to access data stored in NFS and perform analytics on that data.

Supports Key Open-Source Projects—Apache Hadoop, Apache Spark, and Tachyon

The NetApp NFS Connector for Hadoop works specifically with MapReduce for the compute or processing part of the Apache Hadoop framework. There is also support for other Apache projects, such as Apache HBase (columnar database) and Apache Spark (another processing engine compatible with Hadoop). The NFS connector also works with Tachyon, an in-memory file system that can run with Apache Hadoop and Apache Spark. For high performance, analytics applications can use Tachyon for caching a working set of data closer to the compute, thereby allowing many more workloads to use NFS data.

How to Get NFS Connector and More Information

You can download the NetApp NFS Connector at the github repository and the technical report here. https://github.com/NetApp/NetApp-Hadoop-NFS-Connector

About NetApp

Leading organizations worldwide count on NetApp for software, systems and services to manage and store their data. Customers value our teamwork, expertise and passion for helping them succeed now and into the future.

www.netapp.com

NetApp®

Follow us on: