Solution overview

NetApp Solutions

NetApp
October 20, 2023

Table of Contents

TR-4657: NetApp hybrid cloud data solutions - Spark and Hadoop based on customer use cases

Karthikeyan Nagalingam and Sathish Thyagarajan, NetApp

This document describes hybrid cloud data solutions using NetApp AFF and FAS storage systems, NetApp Cloud Volumes ONTAP, NetApp connected storage, and NetApp FlexClone technology for Spark and Hadoop. These solution architectures allow customers to choose an appropriate data protection solution for their environment. NetApp designed these solutions based on interaction with customers and their business use-cases. This document provides the following detailed information:

- Why we need data protection for Spark and Hadoop environments and customer challenges.
- The data fabric powered by NetApp vision and its building blocks and services.
- How these building blocks can be used to architect flexible data protection workflows.
- The pros and cons of several architectures based on real-world customer use cases. Each use case provides the following components:
  - Customer scenarios
  - Requirements and challenges
  - Solutions
  - Summary of the solutions

Why Hadoop data protection?

In a Hadoop and Spark environment, the following concerns must be addressed:

- **Software or human failures.** Human error in software updates while carrying out Hadoop data operations can lead to faulty behavior that can cause unexpected results from the job. In such case, we need to protect the data to avoid failures or unreasonable outcomes. For example, as the result of a poorly executed software update to a traffic signal analysis application, a new feature that fails to properly analyze traffic signal data in the form of plain text. The software still analyzes JSON and other non- text file formats, resulting in the real-time traffic control analytics system producing prediction results that are missing data points. This situation can cause faulty outputs that might lead to accidents at the traffic signals. Data protection can address this issue by providing the capability to quickly roll back to the previous working application version.
- **Size and scale.** The size of the analytics data grows day by day due to the ever-increasing numbers of data sources and volume. Social media, mobile apps, data analytics, and cloud computing platforms are the main sources of data in the current big data market, which is increasing very rapidly, and therefore the data needs to be protected to ensure accurate data operations.
- **Hadoop's native data protection.** Hadoop has a native command to protect the data, but this command does not provide consistency of data during backup. It only supports directory-level backup. The snapshots created by Hadoop are read-only and cannot be used to reuse the backup data directly.

Data protection challenges for Hadoop and Spark customers

A common challenge for Hadoop and Spark customers is to reduce the backup time and increase backup reliability without negatively affecting performance at the production cluster during data protection.

Customers also need to minimize recovery point objective (RPO) and recovery time objective (RTO) downtime and control their on-premises and cloud-based disaster recovery sites for optimal business continuity. This control typically comes from having enterprise-level management tools.

The Hadoop and Spark environments are complicated because not only is the data volume huge and growing, but the rate this data arrives is increasing. This scenario makes it difficult to rapidly create efficient, up-to-date DevTest and QA environments from the source data. NetApp recognizes these challenges and offers the solutions presented in this paper.

Next: Data fabric powered by NetApp for big data architecture.

Copyright information

Trademark information