HDFS Connector for OCI Object Storage Unlocks Performance for Big Data

August 2024, Version 1.0

Public

Disclaimer

This document in any form, software or printed matter, contains proprietary information that is the exclusive property of Oracle. Your access to and use of this confidential material is subject to the terms and conditions of your Oracle software license and service agreement, which has been executed and with which you agree to comply. This document and information contained herein may not be disclosed, copied, reproduced or distributed to anyone outside Oracle without prior written consent of Oracle. This document is not part of your license agreement nor can it be incorporated into any contractual agreement with Oracle or its subsidiaries or affiliates.

This document is for informational purposes only and is intended solely to assist you in planning for the implementation and upgrade of the product features described. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described in this document remains at the sole discretion of Oracle. Due to the nature of the product architecture, it may not be possible to safely include all features described in this document without risking significant destabilization of the code.

Table of Contents

Introduction

Oracle Cloud Infrastructure (OCI) Object Storage is a high-performance, cost-effective, scalable replacement for on-premises or self-managed Hadoop distributed file system (HDFS) storage. To make the replacement work, OCI provides an HDFS connector to bridge the gap between HDFS applications and OCI Object Storage. We recently made significant performance improvements in our HDFS connector and then stress-tested those improvements with one of the largest Hadoop installations in the world. The latest version of the connector is up to 10 times faster than last year's version and outperformed our customer's very large on-premises Hadoop cluster by 40%.

This paper explains how OCI Object Storage and the HDFS connector can be used as a replacement for a self-managed HDFS cluster, and provides performance comparisons with the previous versions of the connector and with a very large on-premises Hadoop cluster.

Background

Hadoop is a distributed compute and storage system that delivers massive aggregate throughput and allows parallel processing of many data sets at once. It's the cornerstone of the big data movement. The standard storage interface to Hadoop is the Hadoop distributed file system, and a long list of big data tools have grown on top of this interface, from streaming data processing systems like Flink to distributed database systems like Pinot.

Many companies who use Hadoop or HDFS-based applications build and manage a large cluster of machines, often in their own data centers. This setup has the advantage of performance but the disadvantages of managing a large cluster throughout its lifecycle, such as budget renewals, capacity planning, breaks and fixes, technology refreshes, and version control. Even when companies build Hadoop clusters in the cloud, they might still rely on many manually configured volumes for the underlying data storage, which comes with many of the same disadvantages.

OCI Object Storage overcomes all these disadvantages: It has unlimited expansion, requires no maintenance, delivers 11 nines of durability, and doesn't require you to preplan or provision capacity. It's also generally the least expensive option per GB stored. Additionally, Object Storage delivers the throughput required to match a large on-premises Hadoop cluster and is compatible with the existing HDFS applications.

OCI offers an HDFS connector that bridges the gap between unmodified Hadoop applications and the Object Storage service. Many customers have used this connector with success for their big data applications. Recently, one of the largest Hadoop installations in the world decided to move many hundreds of nodes of HDFS storage into Object Storage, and we decided to take a closer look to ensure that we were up to the challenge.

Improvements and Stress-Testing the Results

Object Storage is not only a convenient, low-cost storage option, but like HDFS, it offers a distributed storage system that can deliver multiple terabytes per second of aggregate bandwidth. However, the previous version of the HDFS connector wasn't taking full advantage of all that aggregate bandwidth. The HDFS connector client runs in each Compute node and makes Object Storage look like HDFS. Hadoop ecosystem applications—such as Map/Reduce, Flink, and Pinot—can communicate with the HDFS connector with the standard HDFS commands, such as `dfsPut`, `dfsGet`, and `dfsCP`. The connector translates those commands to perform the equivalent operations on Object Storage.

Figure 1 shows the architecture for the example deployment of the HDFS connector in OCI.

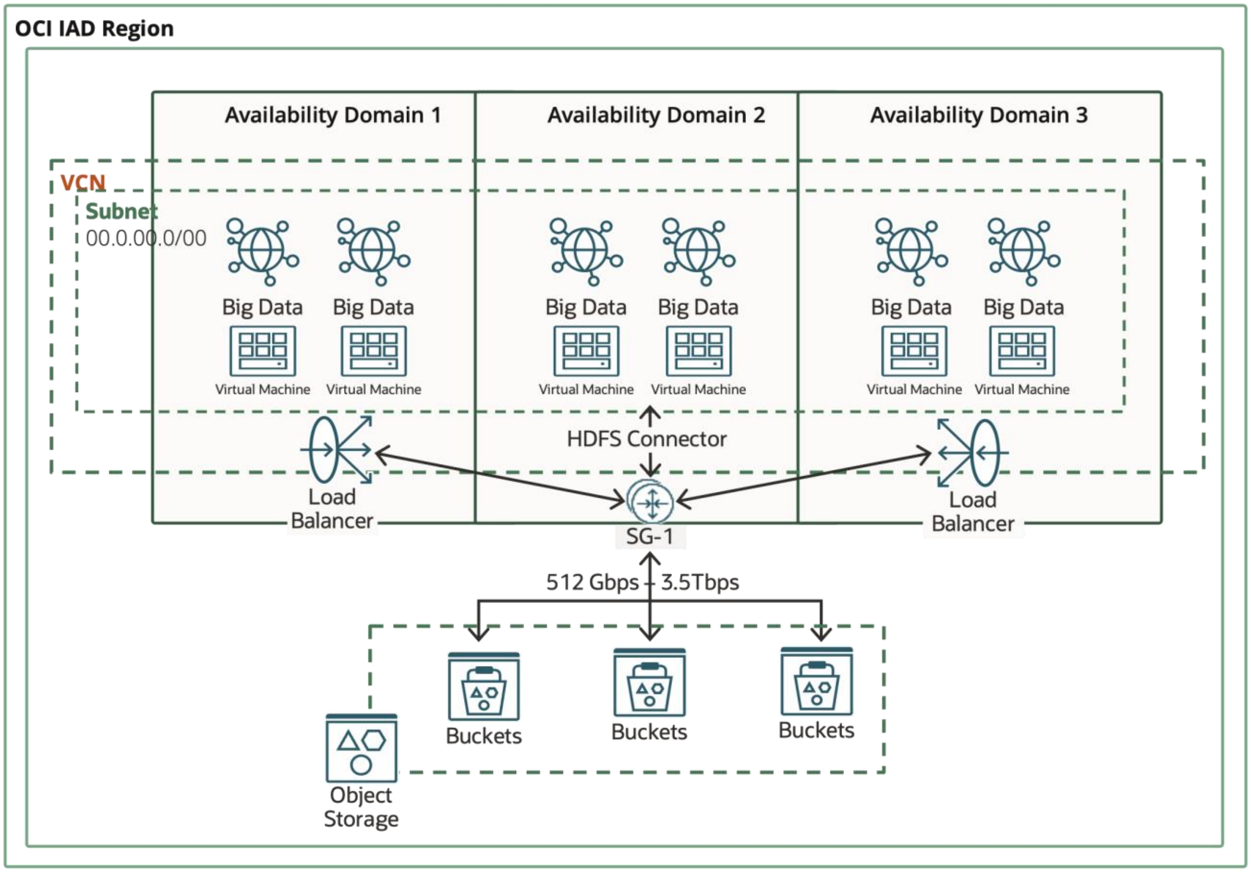

Figure 1. Reference Architecture for the Example Deployment of the HDFS Connector in OCI

Working with the customer, we ran an HDFS benchmark on their local Hadoop environment and then ran it again on Object Storage. With those benchmark targets recorded, we started optimizing the HDFS connector to make maximum use of the distributed power of our storage platform. The latest version of the connector (starting with version 3.3.4.1.2.0) is now about 10 times faster than before. More importantly, the combination of the improved HDFS connector and Object Storage is an average of 40% faster than our customer's on-premises HDFS cluster.

These improvements are available to any OCI user and include the following features:

- **Writing large files with parallel MD5 checksum processing**, which improves PUT time on an 80-GB file by 75% over the previous version and is 48% faster than the on-premises benchmark.
- **Retrieving large files with parallel range-reads**, which improves GET time on an 80-GB file by 800% over the previous version and is 36% faster than the on-premises benchmark.
- **Better memory management**, which allows a `dfs copy` operation on an 80-GB file to be 10 times faster than the previous version and nearly 56% faster than the on-premises benchmark.
- **Improvements in directory listing**, which speed up small file processing. For a directory of 21,000 small files, `dfsGET`, `dfsPUT`, and `dfsMOVE` all improved by 25 times versus the previous version.

For more information, read the [full release notes on GitHub].

ORACLE

Results

The following graph shows the tests in the benchmark (`dfsGET`, `dfsPUT`, `dfsCP`, and `dfsCount`) for four different file size configurations (80 GB, 10 GB, 100 MB, and 21,000 small 100-KB files) spread over many directories. For each combination of command and file size, we measured the number of seconds needed to complete the task, for example, how many seconds were needed to get an 80 GB file and count 21,000 small files. We ran this benchmark on the on-premises HDFS cluster and again on Object Storage using the HDFS connector. The length of the bars in the graph indicates how much faster OCI Object Storage completed the task compared to the on-premises cluster. The color of the bars indicates what kind of test they measure; for example, GET tests are green.

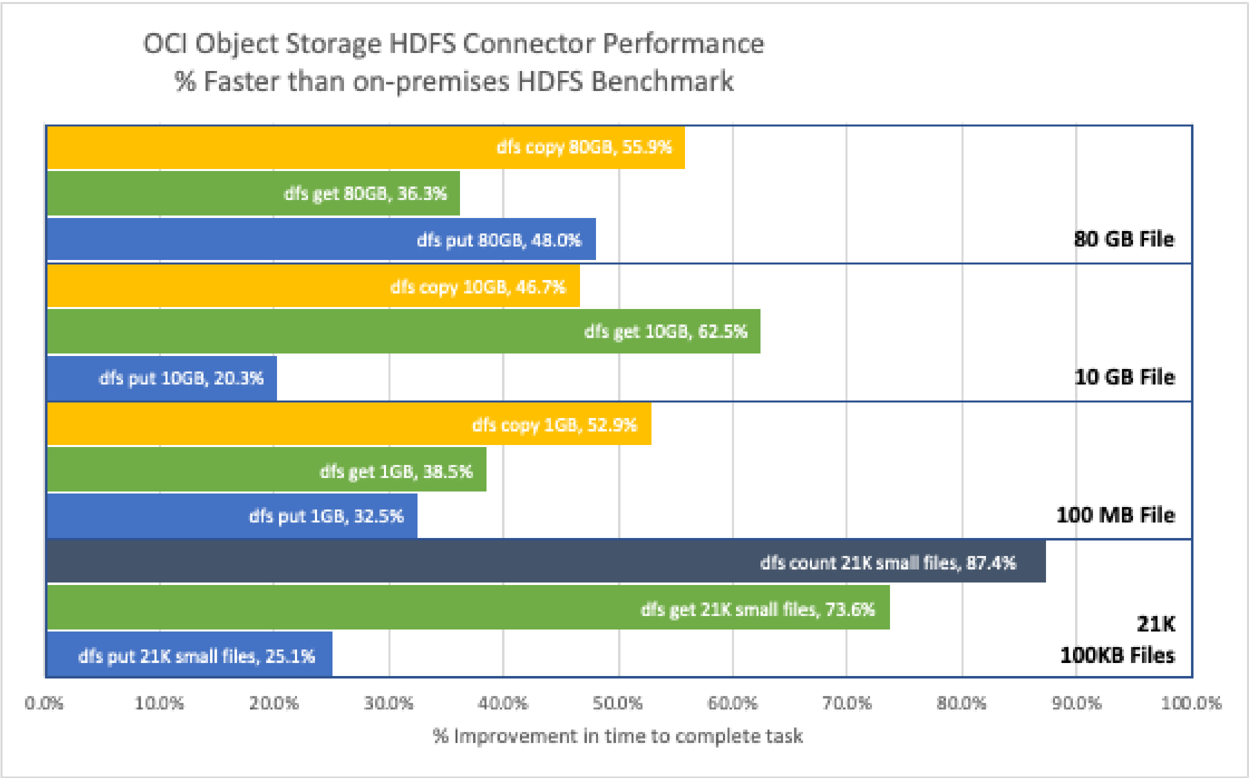

Figure 2. OCI Object Storage HDFS Connector Performance Compared to Performance of On-Premises HDFS Clusters

The following table shows the performance improvement percentages displayed in the preceding graph.

File Size	dfs copy	dfs get	dfs put	dfs count
80 GB	55.9%	36.3 %	48.0%	Not applicable
10 GB	46.7%	62.5%	20.3%	Not applicable
100 MB	52.9%	38.5%	32.5%	Not applicable
100 KB (21,000 files)	Not applicable	73.6%	25.1%	87.4%

Conclusion

Performance is obviously a minimum threshold for migrating a big data environment to Object Storage, but customers like it for its unlimited scalability, easy management, low cost, and high security. Analytics departments are eager to focus their attention and budgets on building the next great solution and not on managing a massive fleet of servers and disks, doing capacity planning, and coordinating technology refreshes.

Individual results vary, but the improved HDFS connector has shown in tests to be faster than the previous version, and in our largest regions, may offer improved performance compared to on-premises infrastructure.

The OCI Object Storage HDFS connector is available from GitHub and is automatically installed on each Compute instance in the Oracle Big Data Service. For installation and usage information, see the official documentation.