# HORTONWORKS HDP WITH ECS REFERENCE ARCHITECTURE

## ABSTRACT

This white paper describes a reference architecture for using Hortonworks HDP with EMC ECS.

August, 2015

REDEFINE

EMC WHITE

EMC²

To learn more about how EMC products, services, and solutions can help solve your business and IT challenges, contact your local representative or authorized reseller, visit www.emc.com, or explore and compare products in the EMC Store
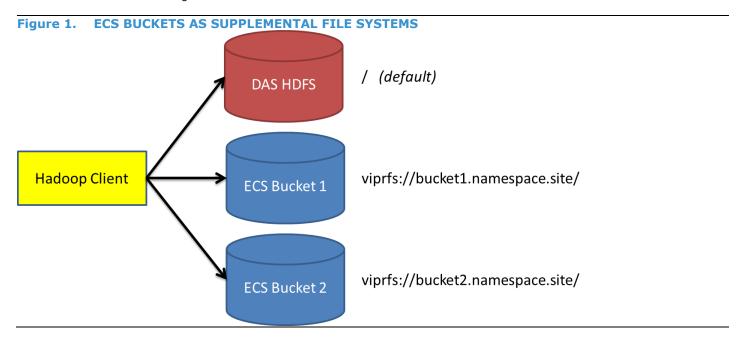
## TABLE OF CONTENTS

# PRODUCT VERSIONS

This document applies to following product versions.

**Table 1     PRODUCT VERSIONS**

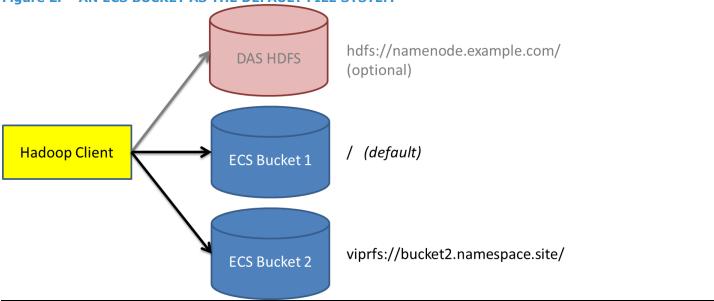| PRODUCT | VERSION |
|---|---|
| Hortonworks HDP | 2.2.6.0-2800 |
| Ambari | 2.0.1 |
| CentOS | 6.6 |
| ECS | 2.0.1.0-427.6d6535a |
| ViPRFS Client | 1.2.0.0-hadoop-2.3 |

# DEPLOYMENT OPTIONS

ECS buckets can be used as supplemental or auxiliary file systems. In this scenario, fs.defaultFS in core-site.xml remains as the default and references the HDFS NameNode. Applications can access data in any ECS bucket by using fully-qualified paths in the format "viprfs://bucket.namespace.site/". Jobs and applications may need to be reconfigured or updated to use a fully-qualified path. This is the recommended configuration and is what this document describes.

**Figure 1.     ECS BUCKETS AS SUPPLEMENTAL FILE SYSTEMS**



An ECS bucket can be used as the default file system in Hadoop, in which all Hadoop paths without an explicit URI prefix such as "hdfs://" refer to objects in a single ECS bucket. In this scenario, fs.defaultFS in core-site.xml is set to "viprfs://bucket.namespace.site/". Data on the traditional HDFS NameNode and DataNodes will be accessible by using the fully-qualified URI such as "hdfs://namenode.example.com/". Some applications that require the Apache HDFS implementation will not work in this Hadoop cluster.

**Figure 2.    AN ECS BUCKET AS THE DEFAULT FILE SYSTEM**

DAS HDFS — hdfs://namenode.example.com/ (optional)

Hadoop Client

ECS Bucket 1 — /  (default)

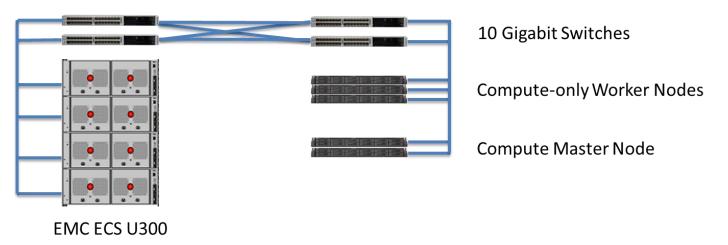ECS Bucket 2 — viprfs://bucket2.namespace.site/

In this reference architecture document, it is assumed that any DAS HDFS usage is minimal. If DAS HDFS usage is significant, then refer to a DAS-orientied reference architecture document.

## CLUSTER TOPOLOGY

The diagram below shows a simplified topology diagram for a cluster.

All ECS, worker nodes, and master nodes have dual 10-Gigabit NICs for maximum performance and redundancy.

**Figure 3.    Cluster Toplogy Diagram**



10 Gigabit Switches

Compute-only Worker Nodes

Compute Master Node

EMC ECS U300

Caption – Add caption. Delete row if not needed.

# CLUSTER COMPONENT LIST

**Table 2    Cluster Component List**

| COMPONENT | CONFIGURATION | NOTES | QUANTITY |
|---|---|---|---|
| Compute-only Worker Node | | 3 is the recommended quantity unless the workload is primarily in-memory, in which case it can be increased. | 3 |
| Form Factor | Rack-mount or Blade | | |
| CPU | 2 sockets<br>6+ cores/socket<br>2+ GhZ | | |
| Memory | 256 GiB | | |
| Disks (OS, applications) | 2 SATA/SAS disks | | |
| Disks (Intermediate data) | 8-12 SATA/SAS disks<br>*or* 1-2 SSDs | Consider the amount of intermediate data that may be produced during job execution. SSDs are good options for blade servers. | |
| Compute Master Node | | | 2 |
| CPU | 2 sockets<br>6+ cores/socket<br>2+ GhZ | | |
| Memory | 128 GiB | | |
| Disks | 2 SATA/SAS disks | | |
| Ethernet Switch for Compute Nodes | 16 port 10-Gigabit Ethernet | | 2 |
| EMC ECS U300 | (4) Nodes<br>(15) drives per node<br>(2) 10 GbE Arista 7124 switch | | 1 |

Caption – Add caption. Delete row if not needed.


# REFERENCES

Hortonworks HDP with ECS Configuration and Best Practices Guide