HFTP Guide

Table of contents

1 Introduction

HFTP is a Hadoop filesystem implementation that lets you read data from a remote Hadoop HDFS cluster. The reads are done via HTTP, and data is sourced from DataNodes. HFTP is a read-only filesystem, and will throw exceptions if you try to use it to write data or modify the filesystem state.

HFTP is primarily useful if you have multiple HDFS clusters with different versions and you need to move data from one to another. HFTP is wire-compatible even between different versions of HDFS. For example, you can do things like: `hadoop distcp -i hftp://sourceFS:50070/src hdfs://destFS:50070/dest`. Note that HFTP is read-only so the destination must be an HDFS filesystem. (Also, in this example, the `distcp` should be run using the configuraton of the new filesystem.)

An extension, HSFTP, uses HTTPS by default. This means that data will be encrypted in transit.

2 Implementation

The code for HFTP lives in the Java class `org.apache.hadoop.hdfs.HftpFileSystem`. Likewise, HSFTP is implemented in `org.apache.hadoop.hdfs.HsftpFileSystem`.

3 Configuration Options

Name	Description
dfs.hftp.https.port	the HTTPS port on the remote cluster. If not set, HFTP will fall back on `dfs.https.port`.
hdfs.service.host_**ip:port**	Specifies the service name (for the security subsystem) associated with the HFTP filesystem running at **ip:port.**