ALIBABA CLOUD

阿里云

文件存储HDFS
快速入门

文档版本：20220629

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.开通文件存储HDFS版服务

本文介绍如何开通**文件存储HDFS版**服务。

背景信息

在使用阿里云**文件存储HDFS版**服务之前，请确保您已经注册了阿里云账号并完成实名认证。如果您还没有创建阿里云账号，系统会在您开通**文件存储HDFS版**时提示您注册账号。

操作步骤

1. 登录阿里云官网。
2. 将鼠标移至**产品**，单击**文件存储HDFS版**，打开**文件存储HDFS版产品详情**页面。
3. 在文件存储HDFS版产品详情页，单击**立即开通**。
4. 开通服务后，在**文件存储HDFS版产品详情**页面单击**管理控制台**直接进入**文件存储HDFS版**管理控制台界面。

   您也可以单击位于官网首页右上方菜单栏的**控制台**，进入阿里云管理控制台首页。选择**产品与服务**页签，在**全部产品与服务**区域，单击产品与服务列表右侧的，选择**存储与CDN > 文件存储HDFS版**进入**文件存储HDFS版**管理控制台界面。

2.快速入门

本文将指导您快速部署和使用阿里云**文件存储HDFS版**。您需要先创建**文件存储HDFS版**文件系统并完成挂载操作。部署成功后，您就可以像在Hadoop分布式文件系统（Hadoop Distributed File System）中一样管理和访问数据。

前提条件

- 已开通文件存储HDFS版服务。具体操作，请参见开通文件存储HDFS版服务。
- 已购买ECS实例。更多信息，请参见选购ECS实例。

  本文使用的ECS实例地域在华东1（杭州）地域。
- 已为ECS实例安装JDK，且JDK版本不低于1.8。
- 已为ECS实例安装Hadoop客户端，建议您使用的Hadoop版本不低于2.7.2。

  本文使用的Hadoop版本为Apache Hadoop 2.7.2。

步骤一：创建文件系统

您可以通过**文件存储HDFS版**控制台创建文件系统或调用CreateFileSystem创建文件系统。

1. 登录文件存储HDFS版控制台。
2. 在顶部菜单栏选择要创建文件系统实例的区域。例如**华东1（杭州）**。
3. 在**概览**页面，单击**创建文件系统**。

   > **说明**
   > - 单个文件系统容量上限为1 EiB。
   > - 每个账号在单个地域内最多可以创建3个文件系统。

4. 在**创建文件系统**页面，配置如下必要参数。其他参数请您根据实际业务需求选择或选用默认配置。

参数	说明
可用区	下拉选择相应的可用区。
文件系统名称	输入想要创建的文件系统的名称。文件系统命名规则如下： ○ 全局唯一且不能为空字符串。 ○ 长度为6~64个字符。 ○ 支持英文字母，可包含数字、下划线（_）和短划线（-）。
协议类型	选择**HDFS**协议。
存储类型	选择**标准型**。

参数	说明
文件系统容量（单位 GB）	输入您预期要配置的文件系统容量，防止使用超出预期的空间容量。配置的文件系统容量不用作计费依据。 **说明** ○ 只允许输入正整数来设置文件系统容量。 ○ 文件系统创建后，您还可以修改该文件系统容量。更多信息，请参见修改文件系统。
吞吐模式	选择**标准吞吐**或**预置吞吐**，预置吞吐取值范围为1~1024 MiB/s。请根据需求选择。

5. 单击**确定**，完成文件系统的创建。

   如果新创建的文件系统未在列表中显示，请刷新此页面。

> **说明** 初始情况下，每个阿里云账号都会自动生成一个VPC默认权限组，允许同一VPC网络下的任何IP地址通过该挂载点访问文件系统。您也可以根据业务场景创建权限组。具体操作，请参见创建权限组和规则。

步骤二：创建挂载点

挂载点是**文件存储HDFS版**文件系统在网络环境中的连接点，通过挂载点实现数据传输。

1. 返回HDFS控制台，选择**文件系统**。
2. 单击刚创建的文件系统，选择**挂载点**页签。
3. 单击**添加挂载点**。
4. 在**添加挂载点**页面，配置如下参数。

参数	说明
挂载点类型	选择VPC。
权限组	根据需求选择权限组。 初始情况下，每个账号都会自动生成一个VPC默认权限组，允许同一VPC网络下的任何IP地址通过该挂载点访问文件系统。您也可以根据业务场景创建权限组。具体操作，请参见创建权限组和规则。
VPC网络ID	选择已创建的VPC网络。如果还未创建，请前往VPC控制台创建。
VPC网络交换机ID	选择VPC网络下创建的交换机。

5. 配置完成后，单击**确定**。

步骤三：挂载文件系统

在使用文件系统前，您还需要通过挂载点挂载**文件存储HDFS版**文件系统。

1. 连接ECS实例。连接方式，请参见连接ECS实例。
2. 配置core-site.xml。
   i. 执行如下命令打开*core-site.xml*文件。

   ```
   vim /usr/local/hadoop-x.y.z/etc/hadoop/core-site.xml
   ```

   其中，`x.y.z`为Hadoop版本号，请根据实际替换。

   ii. 在*core-site.xml*文件中，配置如下信息。

   ```
   <property>
        <name>fs.defaultFS</name>
        <value>dfs://f-xxxxxxxx.cn-xxxxx.dfs.aliyuncs.com:10290</value>
   </property>
   <property>
        <name>fs.dfs.impl</name>
        <value>com.alibaba.dfs.DistributedFileSystem</value>
   </property>
   <property>
        <name>fs.AbstractFileSystem.dfs.impl</name>
        <value>com.alibaba.dfs.DFS</value>
   </property>
   ```

   其中，`f-xxxxxxxx.cn-xxxxx.dfs.aliyuncs.com`为文件存储HDFS版挂载点地址，请根据实际情况替换。

   > 注意 如果fs.defaultFS属性的<value>值中包含`hdfs://`，请将其相应替换为`dfs://`。

   iii. （可选）调整core-site.xml配置，优化集群性能（例如，io.file.buffer.size和dfs.connection.count等），示例如下。更多有关文件存储HDFS版性能优化方法，请参见性能优化最佳实践。

   ```
   <property>
        <name>io.file.buffer.size</name>
        <value>4194304</value>
        <description>To achieve high throughput, no less than 1MB, no more than 8MB</d
   escription>
   </property>
   <property>
        <name>dfs.connection.count</name>
        <value>1</value>
        <description>If multi threads in the same process will read/write to DFS, set
   to count of threads</description>
   </property>
   ```

   iv. 将*core-site.xml*文件同步到所有依赖hadoop-common的节点上。
3. 部署文件存储HDFS版Java SDK。
   i. 下载最新的文件存储HDFS版Java SDK。

ii. 将下载的文件存储HDFS版Java SDK部署到Hadoop生态系统组件的CLASSPATH路径下（推荐部署在hadoop-common-x.y.x.jar目录中）。

> 说明 Hadoop版本不低于2.7.2。

4. 使用hadoop fs命令行工具，执行 `hadoop fs -ls /` 命令验证部署，如下图所示。

```
[hadoop@iZ5wf05xt7fvxpnkx15oy2Z ~/hadoop-2.7.2]$ bin/hadoop fs -ls /
Found 12 items
drw-r-----T   - hadoop       hadoop          75498848 1970-01-01 08:00 /MR
drwxr-----T   - alicloud-dfs alicloud-dfs    75498848 1970-01-01 08:00 /benchmarks
drw-r-----T   - hadoop       hadoop          75498848 1970-01-01 08:00 /hadoop
drwxr-----T   - hadoop       hadoop          75498848 1970-01-01 08:00 /tcpds
drw-r-----T   - alicloud-dfs alicloud-dfs    75498848 1970-01-01 08:00 /tmp
```

如果未报错，则部署成功。

步骤四：上传下载数据

部署成功后，您可以在ECS上把HDFS文件系统当做Hadoop分布式文件系统来访问和使用。

常见问题

- **什么是文件存储HDFS版？适用什么场景？**

  文件存储HDFS版是面向阿里云ECS实例及容器服务等计算资源的文件存储服务。文件存储HDFS版兼容了标准的Hadoopfs协议接口，使您无需对现有大数据分析应用做任何修改，即可使用具备无限容量及性能扩展、单一命名空间、高可靠和高可用的分布式文件系统。相比自建HDFS存储，使用文件存储HDFS版服务可以大量节约维护成本，降低数据安全风险。

  文件存储HDFS版适用于互联网行业、金融行业等有大数据计算与存储分析需求的行业客户进行海量数据存储和离线计算的业务场景，充分满足以Hadoop为代表的分布式计算业务类型对分布式存储性能、容量和可靠性的多方面要求。

- **开通文件存储HDFS版服务后，就开始计费吗？**

  仅开通文件存储HDFS版服务，不会产生费用。当您在文件存储HDFS版中写入文件数据产生实际存储容量，则开始计费。

- **文件系统用于计费的计费存储量是如何计算的？**

  按每小时计费存储量的最大值（峰值）计费。

  - 计费存储量=MAX[核算存储量,实际存储量]
  - 实际存储量是指文件系统中所有文件大小的总和（不含目录），包括文件空洞。
  - 核算存储量是指5 MiB×Inode（包括文件和目录）数量获得的存储量。Inode数量可以通过控制台和容量监控获得。