数据加速器 GooseFS

快速入门

产品文档

腾讯云

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100。

快速入门

最近更新时间：2022-07-29 18:08:05

本文档主要提供 GooseFS 快速部署、调试的相关指引，提供在本地机器上部署 GooseFS，并将对象存储（Cloud Object Storage，COS）作为远端存储的步骤指引，具体步骤如下：

前提条件

在使用 GooseFS 之前，您还需要准备以下工作：

1. 在 COS 服务上创建一个存储桶以作为远端存储，操作指引请参见 控制台快速入门。
2. 安装 Java 8 或者更高的版本。
3. 安装 SSH，确保能通过 SSH 连接到 LocalHost，并远程登录。

下载并配置 GooseFS

1. 从官方仓库下载 GooseFS 安装包到本地。官方仓库下载链接：goosefs-1.3.0-bin.tar.gz。
2. 执行如下命令，对安装包进行解压。

```
tar -zxvf goosefs-1.3.0-bin.tar.gz
cd goosefs-1.3.0
```

   解压后，得到 goosefs-1.2.0，即 GooseFS 的主目录。下文将以 ${GOOSEFS_HOME} 代指该目录的绝对路径。
3. 在 ${GOOSEFS_HOME}/conf 的目录下，创建 conf/goosefs-site.properties 的配置文件，可以使用内置的配置模板：

```
$ cp conf/goosefs-site.properties.template conf/goosefs-site.properties
```

4. 在配置文件 conf/goosefs-site.properties 中，将 goosefs.master.hostname 设置为localhost：

```
$ echo "goosefs.master.hostname=localhost">> conf/goosefs-site.properties
```

启用 GooseFS

1. 启用 GooseFS 前，检查系统环境，确保 GooseFS 可以在本地环境中正确运行：

```
$ goosefs validateEnv local
```

2. 启用 GooseFS 前，执行如下命令，对 GooseFS 进行格式化。该命令将清除 GooseFS 的日志和 worker 存储目录下的内容：

```
$ goosefs format
```

3. 执行如下命令，启用 GooseFS。在系统默认配置下，GooseFS 会启动一个 Master 和一个 Worker。

```
$ ./bin/goosefs-start.sh local SudoMount
```

该命令执行完毕后，可以访问 http://localhost:9201 和 http://localhost:9204，分别查看 Master 和 Worker 的运行状态。

使用 GooseFS 挂载 COS（COSN）或腾讯云 HDFS（CHDFS）

如果 GooseFS 需要挂载 COS（COSN）或腾讯云 HDFS（CHDFS）到 GooseFS 的根路径上，则需要先在 conf/core-site.xml 配置中指定 COSN 或 CHDFS 的必需配置项，其中包括但不限于： fs.cosn.impl 、 fs.AbstractFileSystem.cosn.impl 以及 fs.cosn.userinfo.secretId 和 fs.cosn.userinfo.secretKey 等，如下所示：

```
<!-- COSN related configurations -->
<property>
<name>fs.cosn.impl</name>
<value>org.apache.hadoop.fs.CosFileSystem</value>
</property>

<property>
<name>fs.AbstractFileSystem.cosn.impl</name>
<value>com.qcloud.cos.goosefs.CosN</value>
</property>

<property>
<name>fs.cosn.userinfo.secretId</name>
<value></value>
```

```
</property>


<property>
<name>fs.cosn.userinfo.secretKey</name>
<value></value>
</property>


<property>
<name>fs.cosn.bucket.region</name>
<value></value>
</property>


<!-- CHDFS related configurations -->
<property>
<name>fs.AbstractFileSystem.ofs.impl</name>
<value>com.qcloud.chdfs.fs.CHDFSDelegateFSAdapter</value>
</property>


<property>
<name>fs.ofs.impl</name>
<value>com.qcloud.chdfs.fs.CHDFSHadoopFileSystemAdapter</value>
</property>


<property>
<name>fs.ofs.tmp.cache.dir</name>
<value>/data/chdfs_tmp_cache</value>
</property>


<!--appId-->
<property>
<name>fs.ofs.user.appid</name>
```

```
<value>1250000000</value>
</property>
```

> 说明：
>
> - COSN 的完整配置可参考：Hadoop 工具。
> - CHDFS 的完整配置可参考： 挂载 CHDFS。

下面将介绍一下如何通过创建 Namespace 来挂载 COS 或 CHDFS 的方法和步骤。

1. 创建一个命名空间 namespace 并挂载 COS：

```
$ goosefs ns create myNamespace cosn://bucketName-1250000000/3TB \
--secret fs.cosn.userinfo.secretId=AKXXXXXXXXXX \
--secret fs.cosn.userinfo.secretKey=XXXXXXXXXX \
--attribute fs.cosn.bucket.region=ap-xxx \
```

> 注意：
>
> - 创建挂载 COSN 的 namespace 时，必须使用 --secret 参数指定访问密钥，并且使用 --attribute 指定 Hadoop-COS（COSN）所有必选参数，具体的必选参数可参考 Hadoop 工具。
> - 创建 Namespace 时，如果没有指定读写策略（rPolicy/wPolicy），默认会使用配置文件中指定的 read/write type，或使用默认值（CACHE/CACHE_THROUGH）。

同理，也可以创建一个命名空间 namespace 用于挂载腾讯云 HDFS：

```
goosefs ns create MyNamespaceCHDFS ofs://xxxxx-xxxx.chdfs.ap-guangzhou.myqcloud.com/3TB \
--attribute fs.ofs.user.appid=1250000000
--attribute fs.ofs.tmp.cache.dir=/tmp/chdfs
```

2. 创建成功后，可以通过 ls 命令列出集群中创建的所有 namespace：

```
$ goosefs ns ls
namespace mountPoint ufsPath creationTime wPolicy rPolicy TTL ttlAction
myNamespace /myNamespace cosn://bucketName-125xxxxxx/3TB 03-11-2021 11:43:06:239
CACHE_THROUGH CACHE -1 DELETE
```

```
myNamespaceCHDFS /myNamespaceCHDFS ofs://xxxxx-xxxx.chdfs.ap-guangzhou.myqcloud.com/3TB 03-11-2021 11:45:12:336 CACHE_THROUGH CACHE -1 DELETE
```

3. 执行如下命令，指定 namespace 的详细信息。

```
$ goosefs ns stat myNamespace

NamespaceStatus{name=myNamespace, path=/myNamespace, ttlTime=-1, ttlAction=DELETE, ufsPath=cosn://bucketName-125xxxxxx/3TB, creationTimeMs=1615434186076, lastModificationTimeMs=1615436308143, lastAccessTimeMs=1615436308143, persistenceState=PERSISTED, mountPoint=true, mountId=4948824396519771065, acl=user::rwx,group::rwx,other::rwx, defaultAcl=, owner=user1, group=user1, mode=511, writePolicy=CACHE_THROUGH, readPolicy=CACHE}
```

元数据中记录的信息包括如下内容：

序号	参数	描述
1	name	namespace 的名字
2	path	namespace 在 GooseFS 中的路径
3	ttlTime	namespace 下目录和文件的 ttl 周期
4	ttlAction	namespace 下目录和文件的 ttl 处理动作，有两种处理动作：FREE 和 DELETE，默认是 FREE
5	ufsPath	namespace 在 ufs 上的挂载路径
6	creationTimeMs	namespace 的创建时间，单位是毫秒
7	lastModificationTimeMs	namespace 下目录和文件的最后修改时间，单位是毫秒
8	persistenceState	namespace 的持久化状态
9	mountPoint	namespace 是否是一个挂载点，始终为 true
10	mountId	namespace 挂载点 ID
11	acl	namespace 的访问控制列表
12	defaultAcl	namespace 的默认访问控制列表
13	owner	namespace 的 owner

序号	参数	描述
14	group	namespace 的 owner 所属的 group
15	mode	namespace 的 POSIX 权限
16	writePolicy	namespace 的 写策略
17	readPolicy	namespace 的 读策略

使用 GooseFS 预热 Table 中的数据

1. GooseFS 支持将 Hive Table 中的数据预热到 GooseFS 中，在预热之前需要先将相关的 DB 关联到 GooseFS 上，相关命令如下：

```
$ goosefs table attachdb --db test_db hive thrift://
172.16.16.22:7004 test_for_demo
```

> ⚠ 注意：
> 命令中的 thrift 需要填写实际的 Hive Metastore 的地址。

2. 添加完 DB 后，可以通过 ls 命令查看当前关联的 DB 和 Table 的信息：

```
$ goosefs table ls test_db web_page

OWNER: hadoop
DBNAME.TABLENAME: testdb.web_page (
wp_web_page_sk bigint,
wp_web_page_id string,
wp_rec_start_date string,
wp_rec_end_date string,
wp_creation_date_sk bigint,
wp_access_date_sk bigint,
wp_autogen_flag string,
wp_customer_sk bigint,
wp_url string,
wp_type string,
wp_char_count int,
```

```
wp_link_count int,
wp_image_count int,
wp_max_ad_count int,
)
PARTITIONED BY (
)
LOCATION (
gfs://172.16.16.22:9200/myNamespace/3000/web_page
)
PARTITION LIST (
{
partitionName: web_page
location: gfs://172.16.16.22:9200/myNamespace/3000/web_page
}
)
```

3. 通过 load 命令预热 Table 中的数据：

```
$ goosefs table load test_db web_page
Asynchronous job submitted successfully, jobId: 1615966078836
```

预热 Table 中的数据是一个异步任务，因此会返回一个任务 ID。可以通过 goosefs job stat <Job Id> 命令查看预热作业的执行进度。当状态为 "COMPLETED" 后，则整个预热过程完成。

使用 GooseFS 进行文件上传和下载操作

1. GooseFS 支持绝大部分文件系统操作命令，可以通过以下命令来查询当前支持的命令列表：

```
$ goosefs fs
```

2. 可以通过 ls 命令列出 GooseFS 中的文件，以下示例展示如何列出根目录下的所有文件：

```
$ goosefs fs ls /
```

3. 可以通过 copyFromLocal 命令将数据从本地拷贝到 GooseFS 中：

```
$ goosefs fs copyFromLocal LICENSE /LICENSE
Copied LICENSE to /LICENSE
$ goosefs fs ls /LICENSE
-rw-r--r-- hadoop supergroup 20798 NOT_PERSISTED 03-26-2021 16:49:37:215 0% /LICENSE
```

4. 可以通过 cat 命令查看文件内容：

```
$ goosefs fs cat /LICENSE
Apache License
Version 2.0, January 2004
http://www.apache.org/licenses/
TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND DISTRIBUTION
...
```

5. GooseFS 默认使用本地磁盘作为底层文件系统，默认文件系统路径为 ./underFSStorage，可以通过 persist 命令将文件持久化存储到本地文件系统中：

```
$ goosefs fs persist /LICENSE
persisted file /LICENSE with size 26847
```

使用 GooseFS 加速文件上传和下载操作

1. 检查文件存储状态，确认文件是否已被缓存。文件状态 PERSISTED 代表文件已在内存中，文件状态 NOT_PERSISTED 则代表文件不在内存中：

```
$ goosefs fs ls /data/cos/sample_tweets_150m.csv
-r-x------ staff staff 157046046 NOT_PERSISTED 01-09-2018 16:35:01:002 0% /data/cos/sampl
e_tweets_150m.csv
```

2. 统计文件中有多少单词 “tencent”，并计算操作耗时：

```
$ time goosefs fs cat /data/s3/sample_tweets_150m.csv | grep-c tencent
889
real 0m22.857s
user 0m7.557s
sys 0m1.181s
```

3. 将该数据缓存到内存中可以有效提升查询速度，详细示例如下：

```
$ goosefs fs ls /data/cos/sample_tweets_150m.csv
-r-x------ staff staff 157046046
ED 01-09-2018 16:35:01:002 0% /data/cos/sample_tweets_150m.csv
$ time goosefs fs cat /data/s3/sample_tweets_150m.csv | grep-c tencent
889
real 0m1.917s
user 0m2.306s
sys 0m0.243s
```

可见，系统处理延迟从1.181s减少到了0.243s，得到了10倍的提升。

关闭 GooseFS

通过如下命令可以关闭 GooseFS：

```
$ ./bin/goosefs-stop.sh local
```