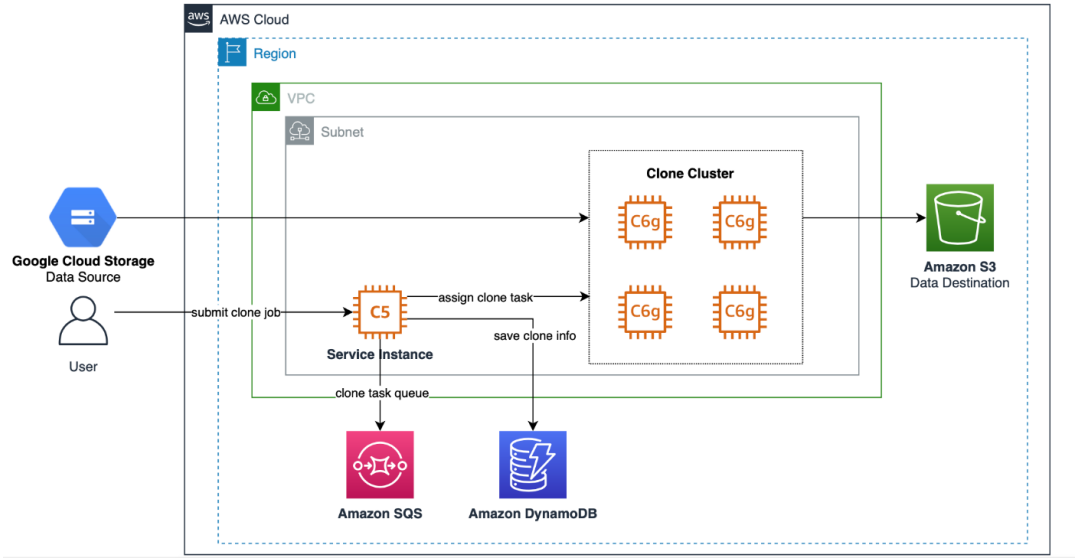
数据传输服务使用手册

本服务用于将Google GCS及其他云厂商（Rclone支持的云厂商）存储服务的数据传输到AWS S3上。可以让您批量启动和管理传输EC2实例。

架构设计

基于Google GCS为数据源的架构图如下：



组件作用：

Amazon SQS：传输任务下发后存储与排序。

Amazon DynamoDB：系统在初始化时会自动创建三张表，用于存储传输设备、任务状态表与配置。

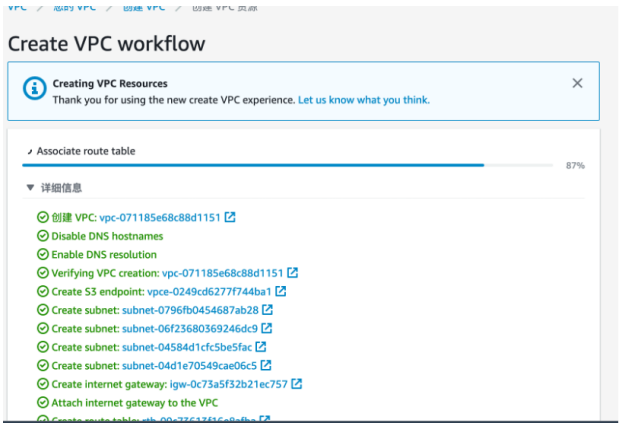
部署目录

1. 创建VPC
2. 修改子网配置
3. 创建部署所需的IAM role
4. 创建SQS Queue
5. 验证邮箱
6. 传输EC2实例的AMI制作
7. 部署传输服务
8. 传输任务配置
9. 启动传输集群
10. 添加传输任务
11. 查看传输实例
12. 查看传输任务
13. 关闭传输集群
14. 清除资源

1. 创建VPC

1. 进入AWS 控制台，搜索栏中输入“VPC”，点击“创建VPC”按钮， 选择vpc与子网一同创建。

2. 使用默认配置即可，点击确认。



2. 修改子网配置

1. 选中要修改的子网，点击操作“编辑子网”

2. 勾选，自动分配IPv4

3. 创建部署所需的IAM ROLE

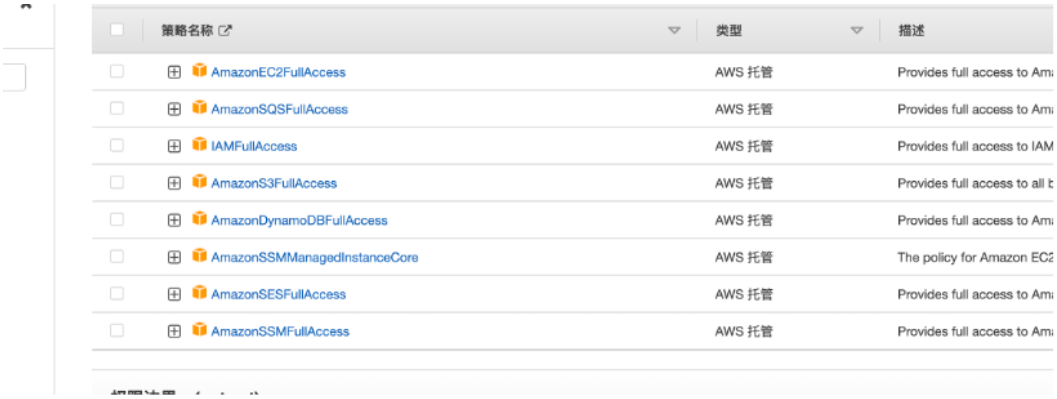
1. 打开AWS console 进入 IAM service

2. 点击左侧“角色”将页面切换到角色，点击右上角的“创建角色”按钮

3. 添加权限，此处需要添加S3的权限和AmazonSSMManagedInstanceCore的权限

4. 为角色命名:Rclone-EC2-Instance-Role，并点击右下角创建按钮(注意此角色名称必须为：Rclone-EC2-Instance-Role)

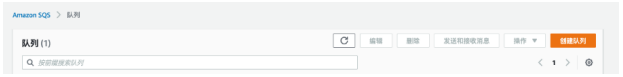
5. 参考本章节的第1-4步，创建另一个名为Rclone-ec2-controller-role,并赋予以下权限。



4. 创建SQS QUEUE

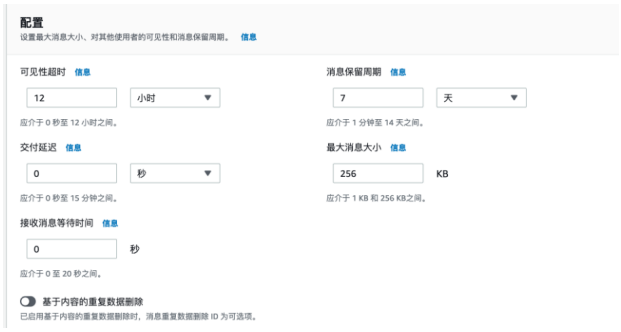
创建一个SQS queue用来传输copy数据的任务的消息。

1. 登录AWS管理控制台，将区域切换到us-east-1.
2. 在上方的服务搜索栏中输入SQS，点击并进入
3. 点击页面右上角的“创建队列”按钮

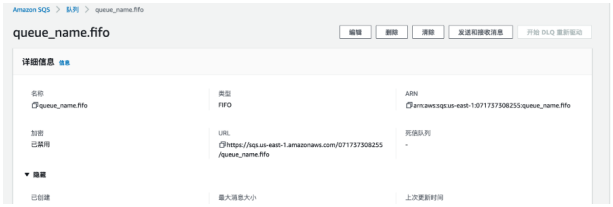


4. 选择FIFO

5. 配置消息的各类属性



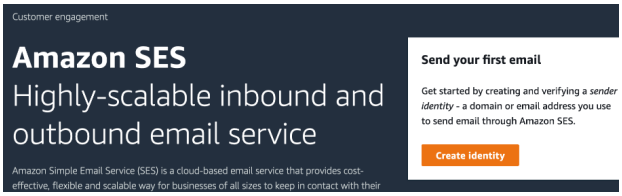
6. 配置完成后，点击页面右下角橙色按钮“创建队列”
7. 队列创建完成之后，会生成URL，点击复制按钮即可，我们会在后面的应用页面中使用到该URL



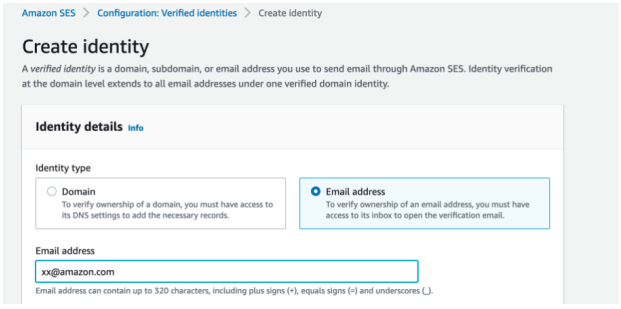
5. 验证邮箱

邮箱用来接收数据copy过程中的异常告警，需要您提供两个邮箱，一个From ，一个To（From和To可以为同一个邮箱地址），两个邮箱均需要在对应的区域的**Amazon Simple Email Service**服务中进行验证，以下为验证步骤：

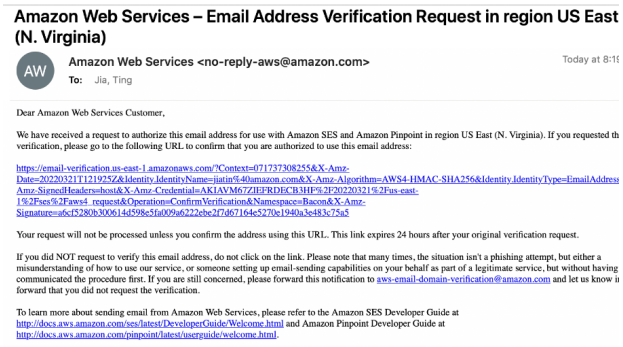
1. 进入AWS管理控制台，将区域切换到us-east-1
2. 在上方的搜索栏中输入SES，点击进入**Amazon Simple Email Service**
3. 点击右侧的橙色按钮“Create Identity”



4. 选择Email Address



5. 点击右下方的 “Create identity”按钮
6. 在您填写的邮箱中，会收到一封邮件，格式如下图：



Amazon Web Services – Email Address Verification Request in region US East (N. Virginia)

AW **Amazon Web Services <no-reply-aws@amazon.com>** Today at 8:1

To: Jia, Ting

Dear Amazon Web Services Customer,

We have received a request to authorize this email address for use with Amazon SES and Amazon Pinpoint in region US East (N. Virginia). If you requested th verification, please go to the following URL to confirm that you are authorized to use this email address:

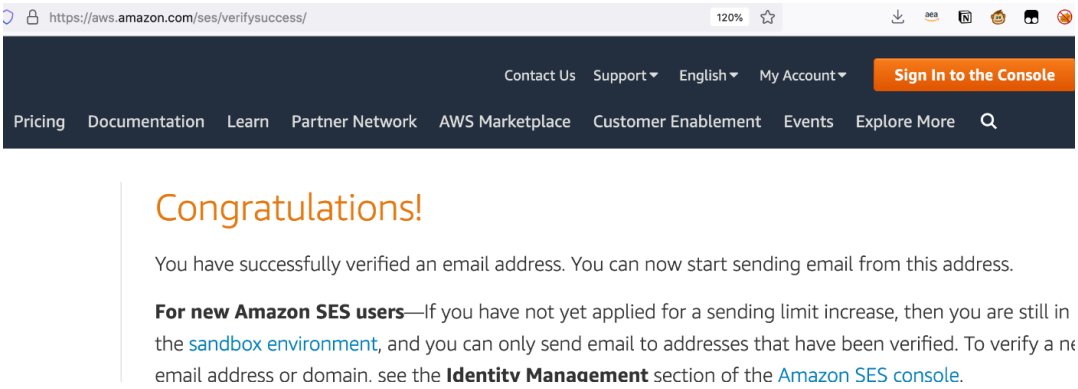
https://email-verification.us-east-1.amazonaws.com/?Context=071737308255&X-Amz-Date=20220321T121925Z&Identity.IdentityName=jiatin%40amazon.com&X-Amz-Algorithm=AWS4-HMAC-SHA256&Identity.IdentityType=EmailAddress Amz-SignedHeaders=host&X-Amz-Credential=AKIAVM67ZIEFRDECB3HF%2F20220321%2Fus-east-1%2Fses%2Faws4_request&Operation=ConfirmVerification&Namespace=Bacon&X-Amz-Signature=a6cf5280b300614d598e5fa009a6222ebe2f7d67164e5270e1940a3e483c75a5

Your request will not be processed unless you confirm the address using this URL. This link expires 24 hours after your original verification request.

If you did NOT request to verify this email address, do not click on the link. Please note that many times, the situation isn't a phishing attempt, but either a misunderstanding of how to use our service, or someone setting up email-sending capabilities on your behalf as part of a legitimate service, but without having communicated the procedure first. If you are still concerned, please forward this notification to aws-email-domain-verification@amazon.com and let us know i forward that you did not request the verification.

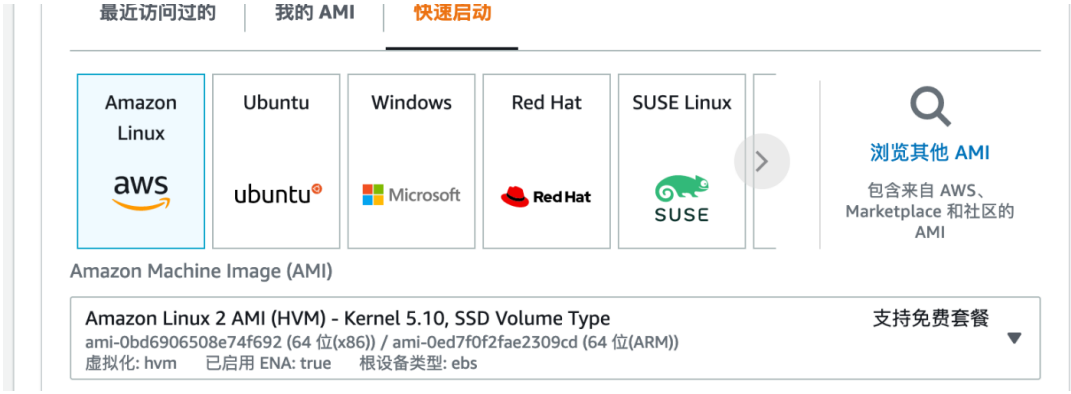
To learn more about sending email from Amazon Web Services, please refer to the Amazon SES Developer Guide at http://docs.aws.amazon.com/ses/latest/DeveloperGuide/Welcome.html and Amazon Pinpoint Developer Guide at http://docs.aws.amazon.com/pinpoint/latest/userguide/welcome.html.

7. 点击邮件中的 https://email-verification开头的长链接，进行验证即可，点开链接后无需其他操作，页面如下图：

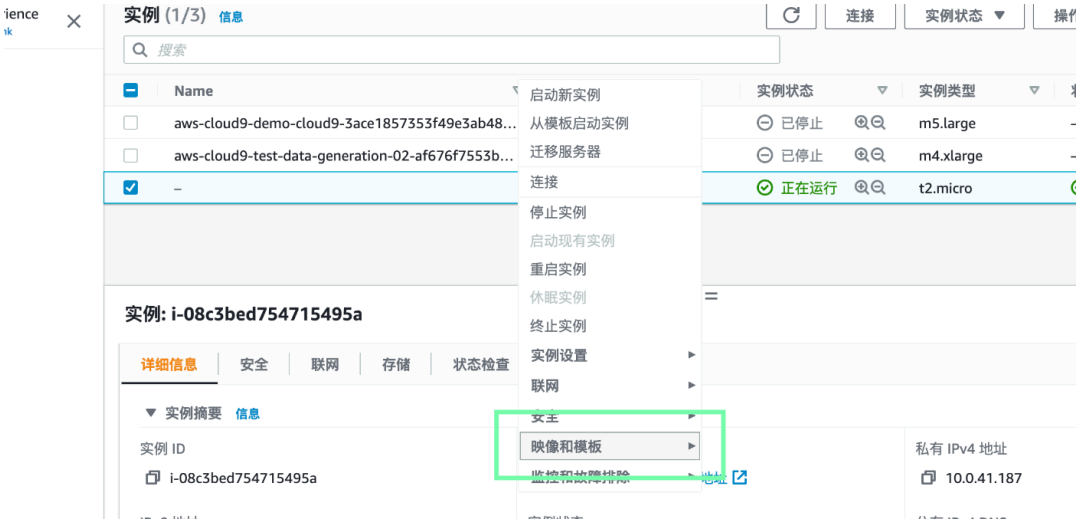


6. 传输EC2实例的AMI制作

制作的AMI用来启动传输数据时用的EC2实例。
1.使用系统默认的AMI启动机器



2.将GCP的密钥放到启动的机器中，可将密钥存放于用户在机器上创建的文件夹（示例/root/auth-key/）下。
3.生成AMI（可参考文档)

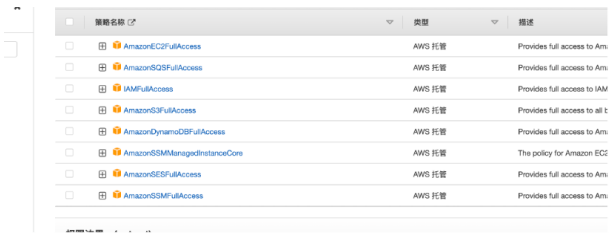


在AMI中查询刚刚生成的AMI的ID，之后会在传输机器的config中配置，请看下图位置。

7. 部署传输服务

如果您没有用于通过SSH登录EC2实例的密钥对，请参考文档进行创建。

1. 进入AWS管理控制台，搜索EC2并进入，在“实例”页面中点击“启动新实例”按钮， 点击右侧“Amazon Linux2 AMI”对应的“选择”按钮



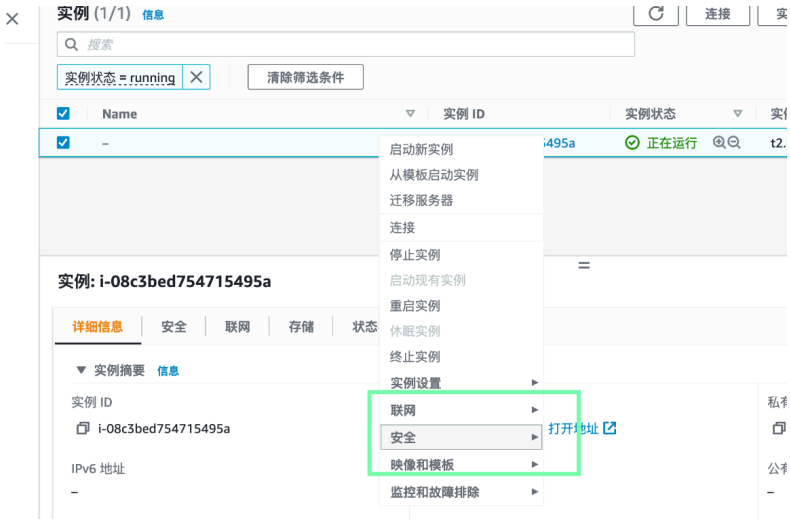
2. 选择机型

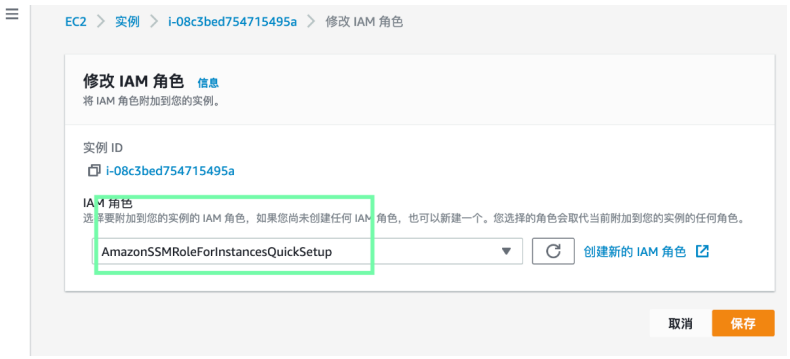
3. 在“配置实例详细信息”的步骤中，选择我们在第1步和第2步创建的VPC与修改过后的公有子网

4. 在“添加存储”中，调整硬盘大小，调到40GB即可

5. 在“配置安全组”步骤中，创建一个新的安全讲，设置安全组规则，将可访问IP限制为您自己的IP即可，如下图

6. 选择您的密钥对（Key Pair），启动实例。
7. 设置机器的权限，我们将第3步创建好的Rclone-ec2-controller-role角色赋予这个机器





8. 连接到机器，选择您创建好的实例，点击右上角第一个“连接”按钮

9. 选择“会话管理器”， 并点击“连接”进入

9. 查询是否有默认的java安装，如过无，则执行安装。

```
java -version
sudo yum install java-1.8.0-openjdk.x86_64
```

10. 从我们您给我们put代码的S3桶拉取代码包

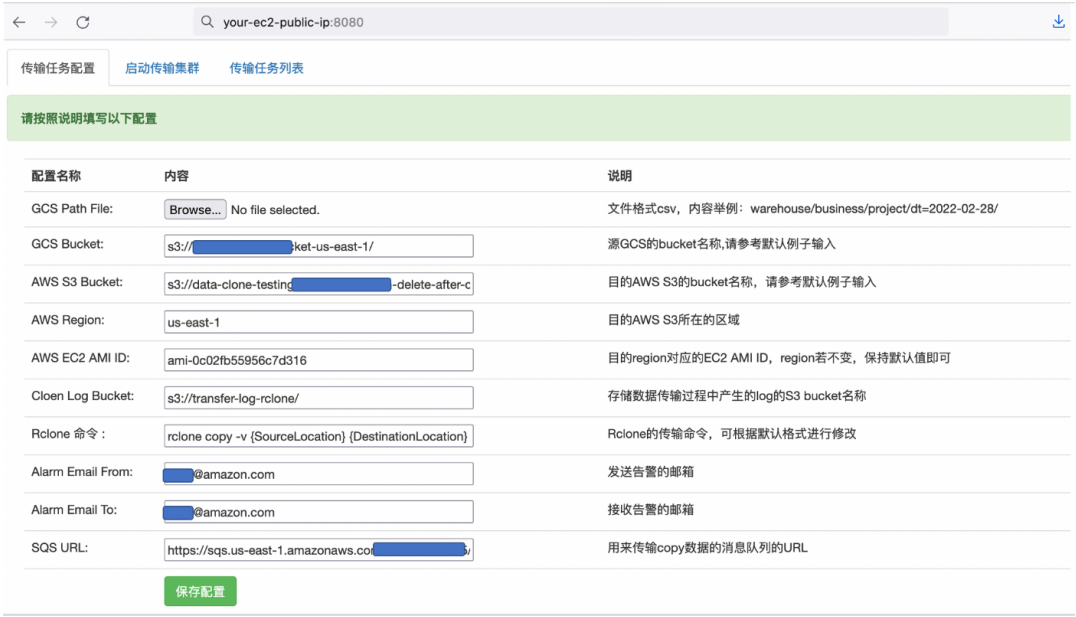
```
aws s3 cp [s3存储路径] [本地路径]
```

11. 启动程序

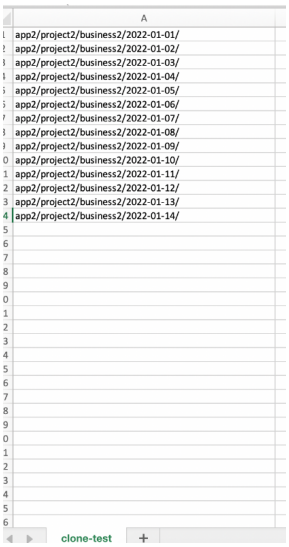
```
java -jar ***.jar > transfer.log &
```

8. 传输任务配置

1. 打开「EC2的外网IP」:8080 web页面，如下图，按照“说明”一列进行配置。



2. 其中“GCS Path File ”上传的文件需要是**csv文件，文件中只有一列，即您需要copy的文件路径**，最好按照大小进行拆分为天级别的文件路径， **注意文件中不需要包含bucket的名字，如果您的某个存储路径为gs://your-bucket-name/app1/project1/date=20220501/file.gz+'工作表1'!D3, 那么放入csv的内容为：app1/project1/date=20220501/ ，每个路径对应一个具体的传输任务，保存配置后会存储到DynamoDB对应的表中。文档格式亦可参考：**demo.csv ，**也可以参考下方csv文件截图：**



3. 其中“AWS EC2 AMI ID”为第6步创建的AMI的ID。
4. “Alarm Email From”和“Alarm Email To” 为第5步验证过的邮箱地址。
5. “SQS URL”为第4步创建的SQS的队列的URL，在AWS管理控制台SQS队列的页面上即可以复制。
6. “Rclone命令”默认使用copy命令，如果需要添加传输参数，请参考附录中的Rclone文档参考，并在此修改，注意{SourceLocation}、{DestinationLocation}与rclone -v需保持默认的格式，示例如下。

   rclone copy -v --transfers 8 --checkers 4 —s3-chunk-size 10M {SourceLocation} {DestinationLocation}

7. 所有信息填写好后， 点击“保存配置”即可。

9. 启动传输集群

1. 在web页面中点击“启动传输集群”

1. 填写机器数量和机器类型，**注意**机器类型需要严格和AWS EC2的机型保持一致， **点击“启动集群”**。

以下是验证过的两种机型上的传输效果对比，以供参考：

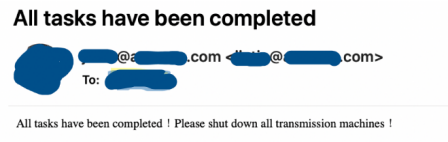
	A	B	C	D
1	机型	价格	网络性能	传输性能
2	C5.large	0.097 USD	最高可达10GB	最高300Mib/s，通常在传输400GB数据后会降低至85Mib/s
3	C6gn.large	0.0986 USD	最高可达25GB	最高300Mib/s，通常在传输600GB数据后会在200Mib/s上下波动。最低可至30Mib/s，这种情况比较少。

2. 等待3-5分钟，上一步启动的EC2实例需要进行初始化。
3. “Rclone传输配置”可以在已有rclone的服务器上配置好后直接复制过来（可通过rclone config show查看）。
4. “Rclone传输配置”格式参考如下：

```
[gcp]
type = google cloud storage
client_id = your-clienid
client_secret = your-client-secret
project_number = your-number
service_account_file = /file-path/gcs.json  # GCS证书
object_acl = authenticatedRead
bucket_acl = authenticatedRead

[s3]
type = s3
provider = AWS
env_auth = true
region = us-east-1
location_constraint = us-east-1
```

5. 其他Rclone支持的云厂商对应的配置可以参考Rclone文档。
6. 点击“安装插件”， 大约1分钟左右，插件安装完后，右侧EC2实例列表中的“插件安装状态”一列会变为“Done”。
7. 此时在第7步上传的文件路径生成的传输任务已经在传输的SQS队列中，因此**数据自动开始传输**。
8. 所有任务传输完成后您填写的To邮箱会收到以下邮件：



10. 添加传输任务

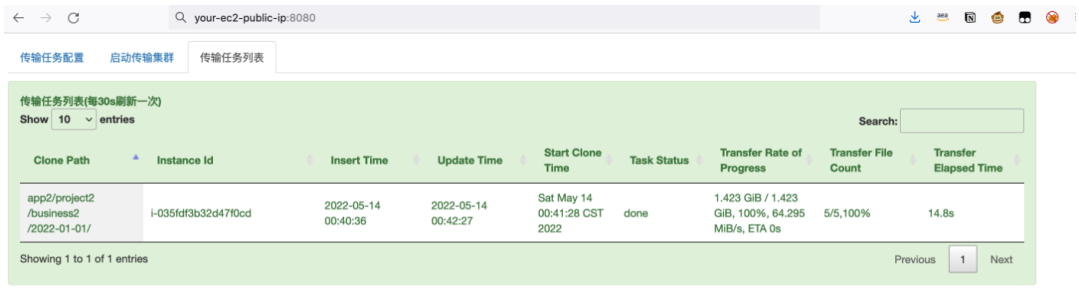
如果您的一批传输任务已经完成，需要添加新的传输任务，则将文件路径放在一个新的文件中，参考第8步，直接上传新的文件，点击“保存配置”即可，其他信息如果没有需要可不进行修改（第一次写入的其他配置信息已保存）。

11. 查看传输实例

页面右侧可以看到您当前启动的所有用来传输数据的实例。

12. 查看传输任务

在web页面中，点击“传输任务列表”，进行查看；最后的三列如果数据还未传输完成，显示为“N/A”。



13. 关闭传输集群

点击以下页面中的“关闭集群”的按钮即可，**注意：关闭集群意味着所有用来传输数据的实例都将被Terminate（终止），请务必确认所有实例上的传输任务已完成（所有任务完成后会有邮件通知）。**

14. 清除资源

需要手动删除包括VPC、IAM role、SQS、SES与所创建的AMI。

DynamoDB中的消息task列表的内容是传输过程中的完成状态，如不需要直接删除表即可。

附录

1. AWS system manager介绍
2. Rclone文档
3. DynamoDB介绍