Hyperdrive Notebooks
AWS Deployment Guide

Table of Contents

Introduction to Hypergiant Hyperdrive Notebooks

Welcome to the AWS deployment guide for Hyperdrive Notebooks from Hypergiant. This document will walk you through everything you need to know to deploy and operate the Hyperdrive Notebook solution.

Hyperdrive Notebooks is a data science lab in a single EC2 instance. It's powered by the open source JupyterHub project, which is running in a containerized Kubernetes environment locally on the host instance.

The internal product name for Hyperdrive Notebooks is "Firefly". You may see references to Firefly in various places within the solution and the supporting resources.

Use Cases

Hyperdrive Notebooks would be a good fit for data science teams looking for Jupyter Notebook hosting infrastructure in an easily-deployable package. The solution builds on the Zero to JupyterHub project, without requiring the operator to deal with some of the complex details of managing a Kubernetes cluster. Hyperdrive Notebooks can be deployed in about 10 minutes, making it an ideal choice for standing up data science lab infrastructure quickly. And because the solution runs in a single EC2 instance, it's easy to stop the instance when the lab is not in use and change the instance type when more or less resources are required.

Typical Deployment

A common deployment of Hyperdrive Notebooks involves the EC2 instance for the host and a set of supporting resources that enable access over HTTPS. These supporting resources include an Application Load Balancer (ALB) where TLS connections are terminated and a DNS ALIAS record pointing to the ALB from a Route 53 hosted zone.

List of Deployed Resources

- 1 EC2 Instance - The Hyperdrive Notebooks host
- 2 Security Groups - One for the host and one for the ALB
- 1 Application Load Balancer - To enable HTTPS access
- 1 ALB Target Group - Pointing to the host
- 1 ALB Listener - Configured for HTTPS and using an existing ACM certificate
- 1 Route 53 Record Set - Added for DNS resolution in an existed hosted zone

Deployment Options

As a single EC2 instance solution, Hyperdrive Notebooks is designed to run in a specific availability zone within a region.

When deploying, it's important to select an instance type that will provide sufficient CPU and memory for all the Jupyter notebooks that might be running concurrently. The instance type can be specified in the CloudFormation parameters during deployment.

Time Required to Deploy

Hyperdrive Notebooks can be deployed in 10 minutes or less. This assumes the operator has an existing Route 53 hosted zone and an ACM certificate ready to use for the deployment.

Supported AWS Regions

The AMI for Hyperdrive Notebooks is available in the following AWS regions:

- us-east-1
- us-east-2
- us-west-1
- us-west-2

If you are interested in running Hyperdrive Notebooks in another region, please contact us at support@hypergiant.com.

Prerequisites and Requirements

The following section details the prerequisites and requirements for deploying Hyperdrive Notebooks.

Technical Prerequisites

An AWS account is required to deploy Hyperdrive Notebooks. You should have "Power User" access to the AWS account, specifically the IAM permissions to deploy a CloudFormation template and create the resources listed in the "Deployed Resources" section above.

Skills and Knowledge Required

Basic familiarity with running CloudFormation templates to deploy solutions in AWS is required. Once deployed, an administrator will be expected to understand how to work with JupyterHub to provision users and notebook servers.

Preparing the AWS Environment

There are 4 prerequisite resources that must be set up prior to deploying: an EC2 SSH key pair, a VPC with Internet access, a Route 53 hosted zone, and an ACM certificate.

You can use an existing EC2 SSH key pair or create a new one. This will be used for administrators to SSH into the Hyperdrive Notebooks host instance.

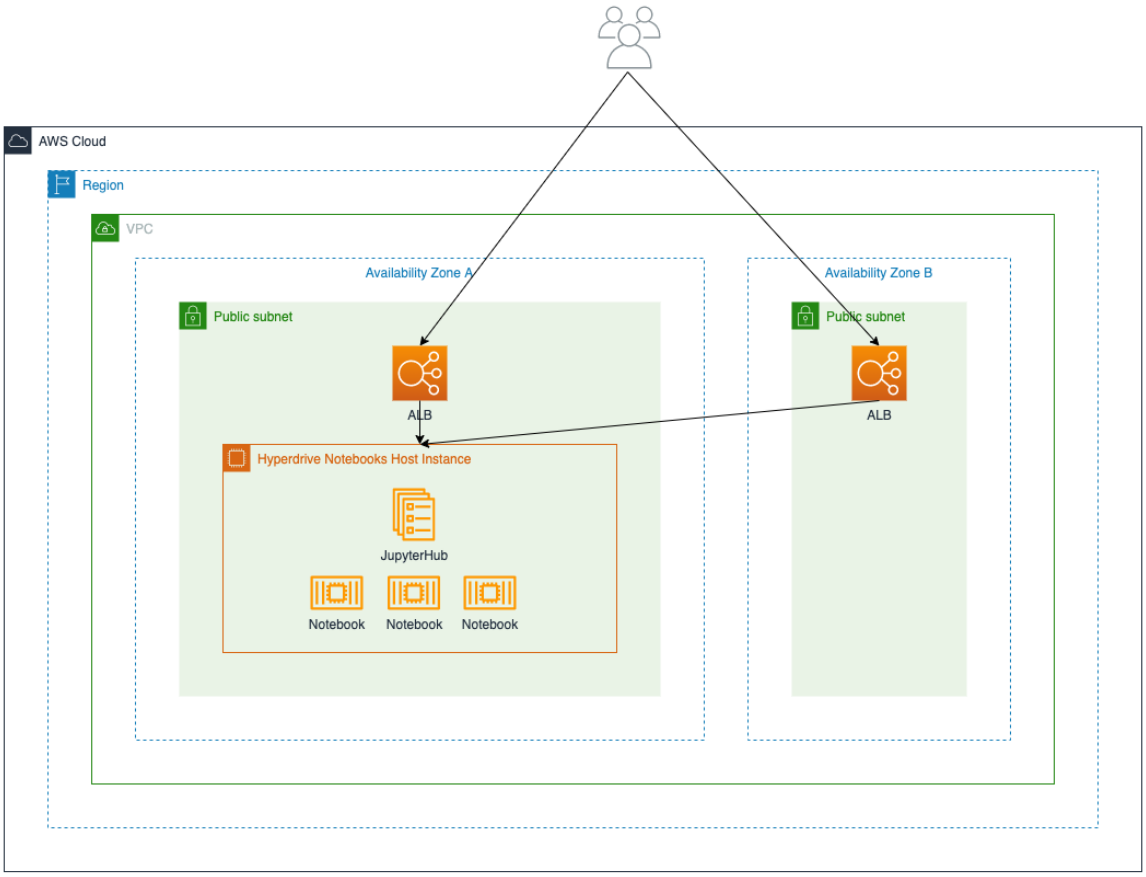
The VPC used for the Hyperdrive Notebooks deployment should be created before creating the CloudFormation stack. You can use the Default VPC or a different VPC created in your AWS account. The VPC must have at least two public subnets (in two different availability zones) and an Internet Gateway attached. The ALB will be provisioned into the two or more public subnets. The host instance will be provisioned into the first of the public subnets.

A Route 53 hosted zone corresponding to the desired FQDN domain for the deployed Hyperdrive Notebooks solution should be available in the AWS account. For example, if you wanted to use the FQDN “notebooks.dslab.example.com”, you should have a hosted zone for the domain “dslab.example.com” already created. An ALIAS record will be added to this zone as part of the CloudFormation deployment.

HTTPS access to Hyperdrive Notebooks requires an AWS Certificate Manager certificate to be available for use in the AWS account. This certificate must have a subject name that matches the FQDN you plan to use for your Hyperdrive Notebooks deployment. For example, if you plan to use “notebooks.dslab.example.com”, you would need to request and validate a certificate for “notebooks.dslab.example.com” or the wildcard version “*.dslab.example.com”. This certificate will be used in the ALB which terminates TLS connections to your deployment. Make sure the certificate is created in the same AWS region where you plan to deploy Hyperdrive Notebooks.

Architecture Diagram

The following diagram illustrates the components of a Hyperdrive Notebooks deployment.



Security

The following section details important security information about your Hyperdrive Notebooks instance.

Do Not Use the AWS Root User

When deploying Hyperdrive Notebooks, do not use the AWS Root user account. Deployments should only be done with IAM users or roles.

Least Privilege Approach to IAM Policies

No IAM policies are created or required to deploy Hyperdrive Notebooks.

Publicly Accessible Resources

No S3 buckets are created or required to deploy Hyperdrive Notebooks. The Route 53 hosted zone is typically a public DNS zone. The Hyperdrive Notebooks JupyterHub web interface is accessible over public HTTPS but protected by authentication.

IAM Users & Policies Created

No additional IAM users or policies are created when deploying Hyperdrive Notebooks.

Encryption Keys Created

No encryption keys are created when deploying Hyperdrive Notebooks.

Managing Secrets

An initial ‘firefly’ admin user password is configured at the end of the CloudFormation deployment process, using the “Sign Up” page. This password can be changed from the control panel web interface.

Locations of Sensitive Data

Docker volumes are located on the host instance at /var/lib/docker/volumes/. These contain the contents of the notebook servers, JupyterHub metadata and configuration files for the embedded kubernetes cluster.

Data Encryption Configurations

The Hyperdrive Notebooks host instance is provisioned with an encrypted EBS root volume, providing encryption at rest.

The Application Load Balancer is configured to terminate TLS connections over HTTPS using an existing ACM certificate specified during deployment. This provides encryption in flight.

Network Security Configurations

Two EC2 security groups are provisioned as part of the Hyperdrive Notebooks solution.

The first security group is for the Application Load Balancer and allows incoming HTTPS connections from the public internet. After deployment, the source CIDR range for this ingress rule can be adjusted to lock down access to a particular client network.

The second security group is for the Hyperdrive Notebooks host instance. This security group allows HTTP port 8090 connections from the ALB security group and public SSH access.

Costs & Licensing

This section details the cost and licensing considerations for your deployment of Hyperdrive Notebooks.

AWS Service Costs

The largest cost component of running Hyperdrive Notebooks is the EC2 instance that hosts the solution. This cost will vary based on the instance type selected during deployment. The host instance can be stopped to save compute costs while not in use.

Other costs include standard EC2 data transfer charges for traffic to the Internet and between availability zones, the ALB costs, and the Route 53 hosted zone costs.

Hyperdrive Notebooks Licensing Model

Hyperdrive Notebooks is an open source product licensed under the Apache 2 license. More detail are available in the LICENSE file in the project's Github repo: https://github.com/gohypergiant/firefly

Sizing Considerations

For production deployments, select an instance type that has adequate CPU and memory to host all the notebooks that will be used concurrently. Each concurrent user will likely require 1-2 CPUs and 2-6 GB of memory, but exact usage requirements will vary based on the notebook workloads being hosted. You can always change the instance type after deployment by stopping the Hyperdrive Notebooks host instance, setting the new instance type, and restarting the host instance.

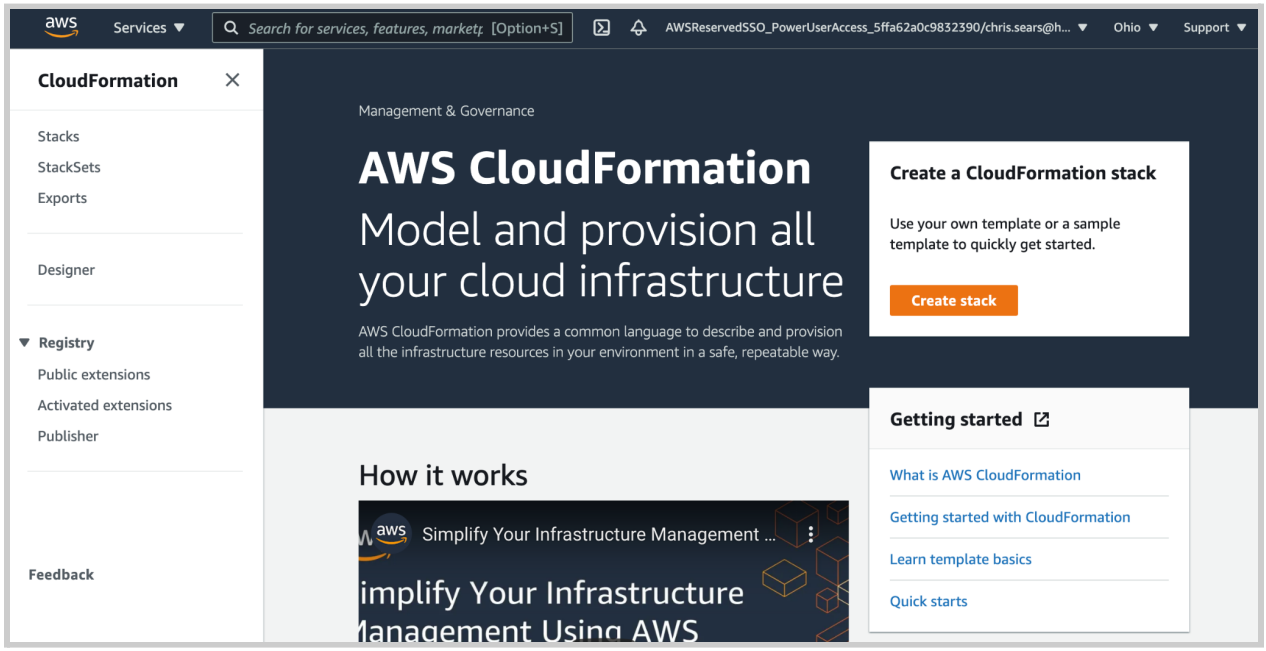
Deployment Steps, Testing & Troubleshooting

This section covers the specific steps required to deploy Hyperdrive Notebooks, along with guidance on testing the deployment and troubleshooting common issues.

Deployment Process

1. Log into the AWS console. Make sure the selected region is where you wish to deploy Hyperdrive Notebooks.

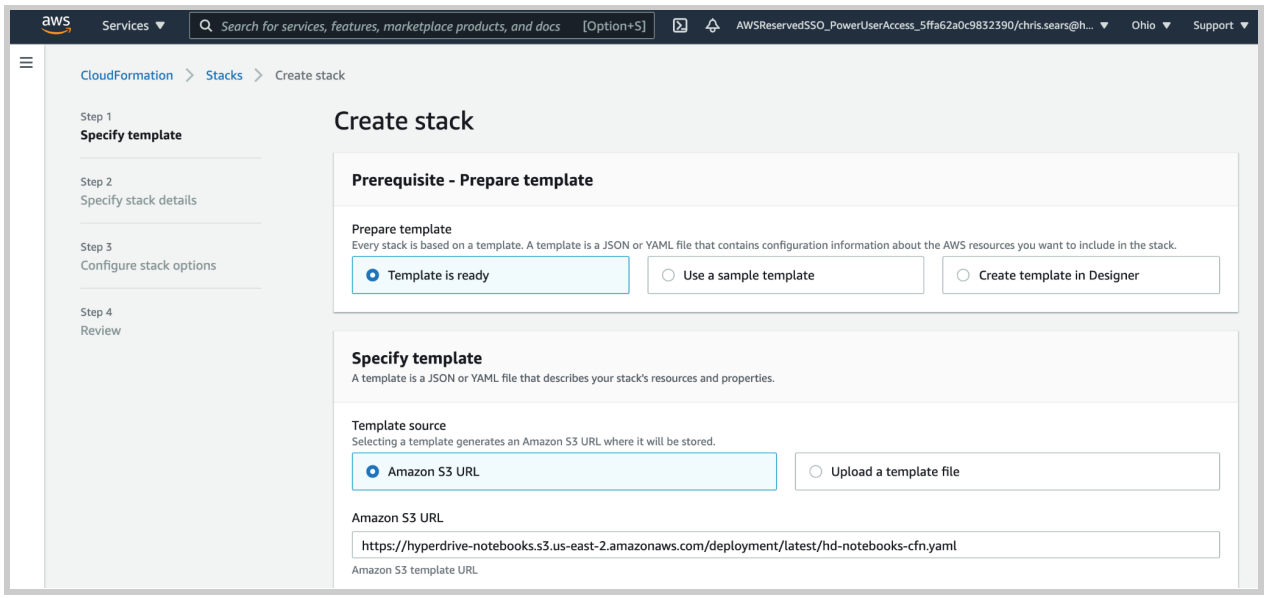
2. Navigate to the CloudFormation service in the AWS console or click this link: https://console.aws.amazon.com/cloudformation



3. Click the “Create stack” button to start the stack deployment wizard.

4. On the first screen, choose or enter the following:

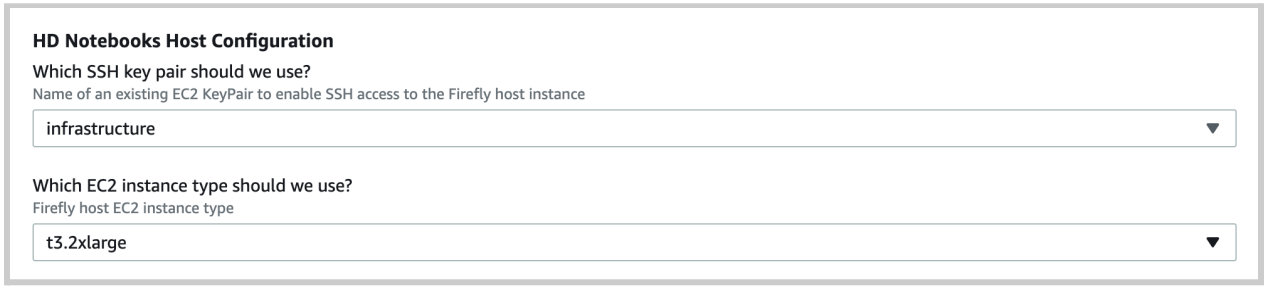
   Prepare template: Template is ready
   Template source: Amazon S3 URL
   Amazon S3 URL:
   https://hyperdrive-notebooks.s3.us-east-2.amazonaws.com/deployment/latest/hd-notebooks-cfn.yaml



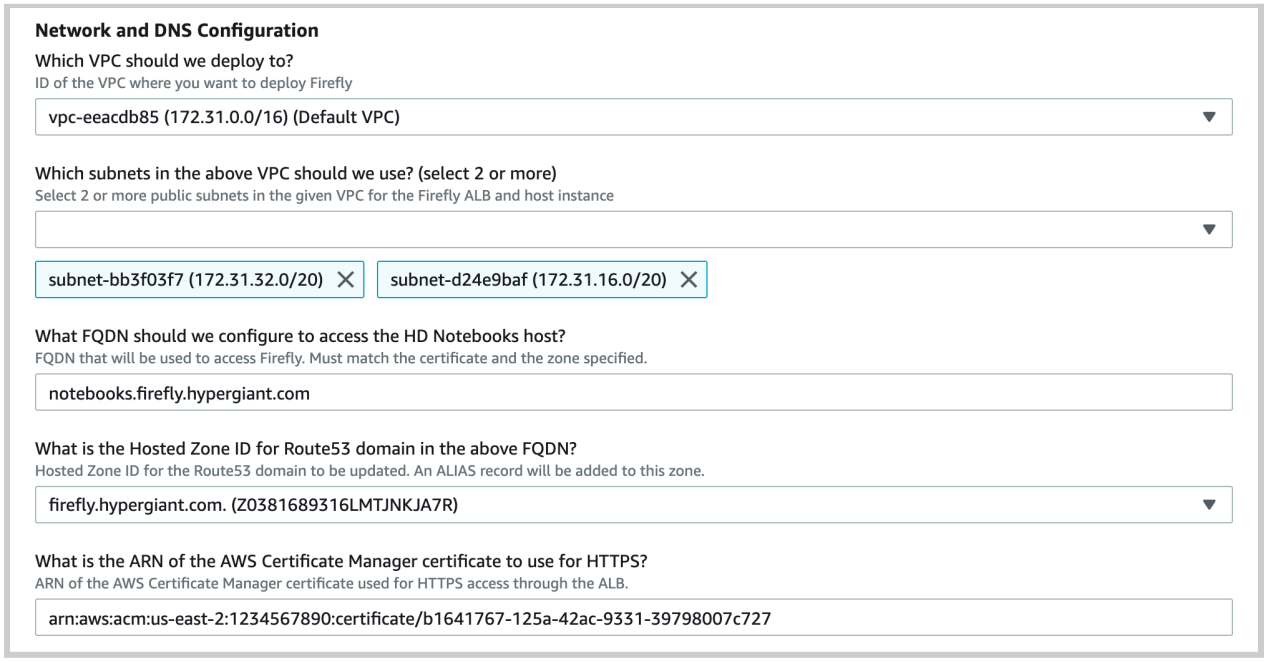
5. Click the orange “Next” button in the bottom right corner.

6. On the second page, give the stack a name. This can be any name you want.

7. In the Parameters section, select values for the Host Configuration.



8. Under Network and DNS Configuration, select or enter values for each field.



9. Under the Additional Configuration section, leave the release branch “stable” selected unless you are testing the development release.

10. Click the orange “Next” button.

11. On the third page, Configure Stack Options, leave all the default values and click the orange “Next” button.

12. On the final page, you may review the parameter values, which should look similar to the screenshot below.

Parameters (8)

Search parameters

Key	Value	Resolved value
HostedZoneId	Z00105853JCLFNRJS1T3Z	-
HostnameWithDomain	notebooks.firefly.hypergiant.com	-
InstanceType	t3.large	-
KeyName	infrastructure	-
PublicSubnets	subnet-bb3f03f7,subnet-d24e9baf	-
ReleaseBranch	dev	-
TLSCertificateARN	arn:aws:acm:us-east-2:1234567890:certificate/b1641767-125a-42ac-9331-39798007c727	-
VPC	vpc-eeacdb85	-

13. Click the orange “Create stack” button in the bottom right corner.

14. On the next page you will be able to monitor the progress of CloudFormation as it provisions and configures the resources for the deployment. Click the circular arrow icon to refresh the list of events.

hyperdrive-notebooks

Delete | Update | Stack actions | Create stack

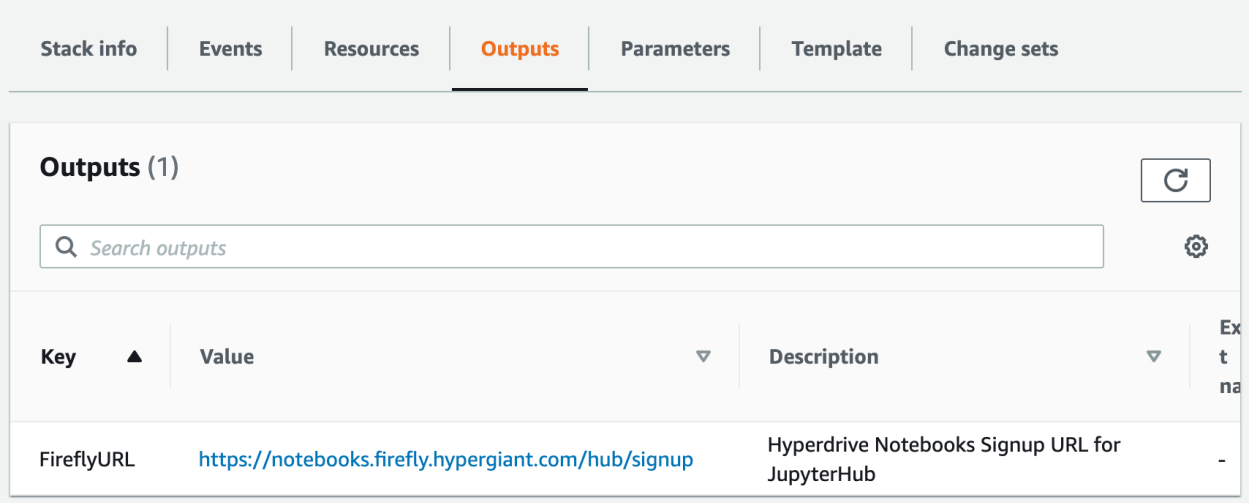
Stack info | Events | Resources | Outputs | Parameters | Template | Change sets

Events (1)

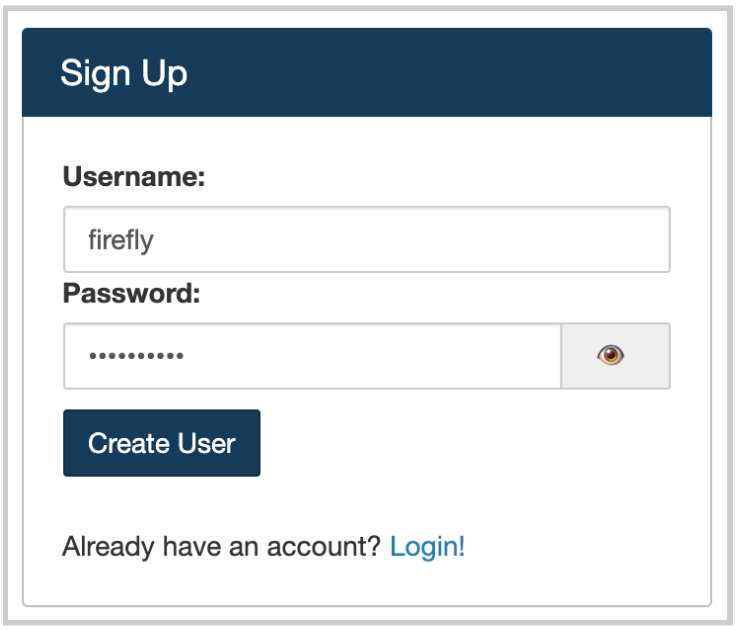
Search events

Timestamp	Logical ID	Status	Status reason
2021-10-18 10:39:37 UTC-0400	hyperdrive-notebooks	CREATE_IN_PROGRESS	User Initiated

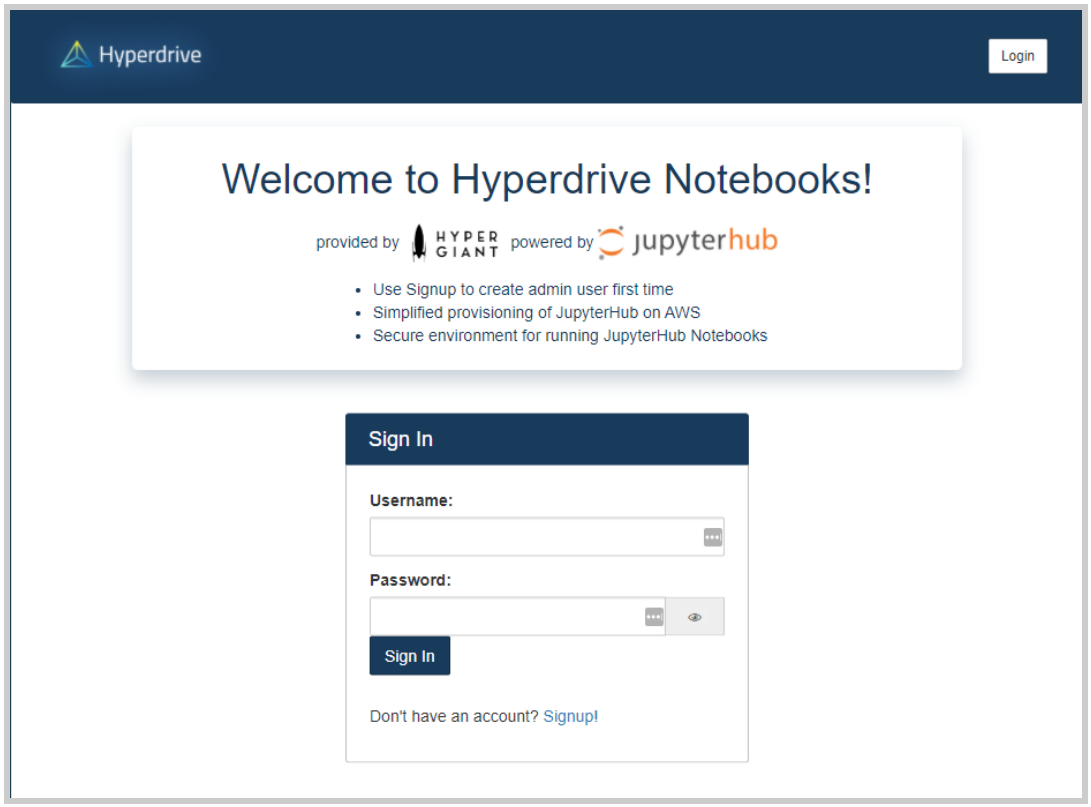
15. When the stack has completed (after about 5 minutes), click on the Outputs tab and find the FireflyURL. Use this URL to access the Sign Up page of your Hyperdrive Notebooks deployment. It is important to complete the next step shortly after the CloudFormation stack completes. Failure to do so would leave your deployment vulnerable.



16. On the Sign Up page, set a password for the initial admin user. For User Name, enter “firefly”, and enter a password at least 10 characters long. Click the “Create User” button.



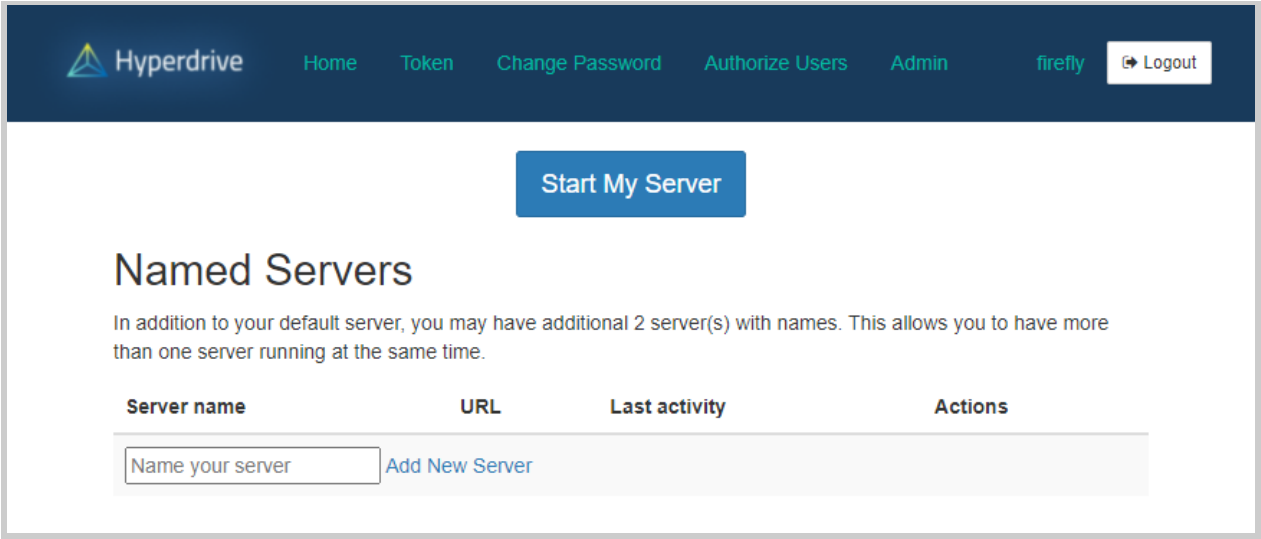
17. Click the “Login!” link to go to the main sign in page for your deployment. Enter “firefly” and the initial password you selected in the previous step and click the “Sign in” button.



You have now successfully deployed Hyperdrive Notebooks!

Testing Your Hyperdrive Notebooks Deployment

Once you are signed into your Hyperdrive Notebooks deployment, click the “Start My Server” to spin up a Jupyter Notebook server.



You will be redirected to a loading page while your server container launches, and then you will be redirected again to the JupyterLab web interface. Your notebook is now ready to use!

Troubleshooting Common Issues

If the CloudFormation stack fails when creating the “ALBListener” resource and the status reason mentions the certificate was not found, please make sure the ACM certificate was created in the same AWS region where you are deploying Hyperdrive Notebooks. If it was created in a different region, please create a new certificate in the correct region and try redeploying the CloudFormation template.

Monitoring the Deployment

You can monitor the overall health of your Hyperdrive Notebooks deployment using any monitoring tool that can check HTTPS endpoints. Configure the check to request the “/hub/login” page of your deployment and confirm it returns a status 200. For example, if your deployment FQDN was “notebooks.example.com”, configure the check to request “https://notebooks.example.com/hub/login”.

For OS-level resource monitoring, the Amazon CloudWatch agent can be installed on the host instance. Directions for installing the CloudWatch agent are available here: https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/install-CloudWatch-Agent-on-EC2-Instance.html

Backup & Recovery

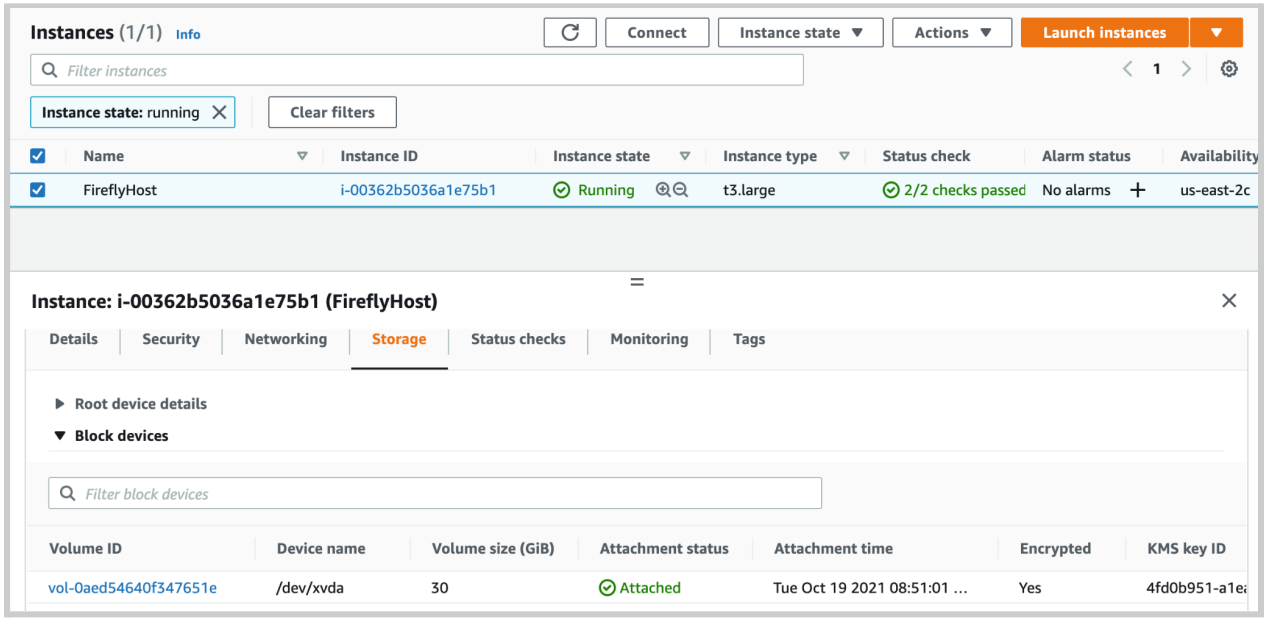
This section covers backup and recovery procedures for your Hyperdrive Notebooks deployment.

Data Stores & Configuration to Protect

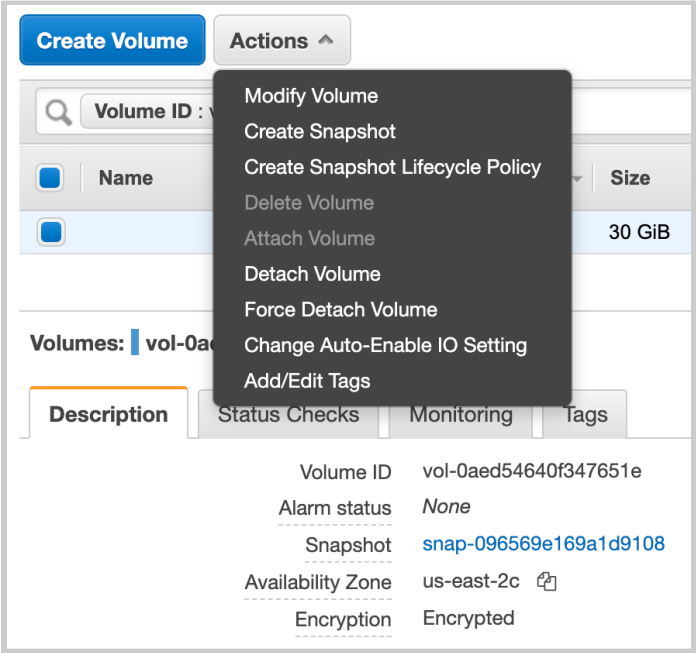
All the data and configuration for your deployment is stored on the root EBS volume of the host instance.

Backup Procedure

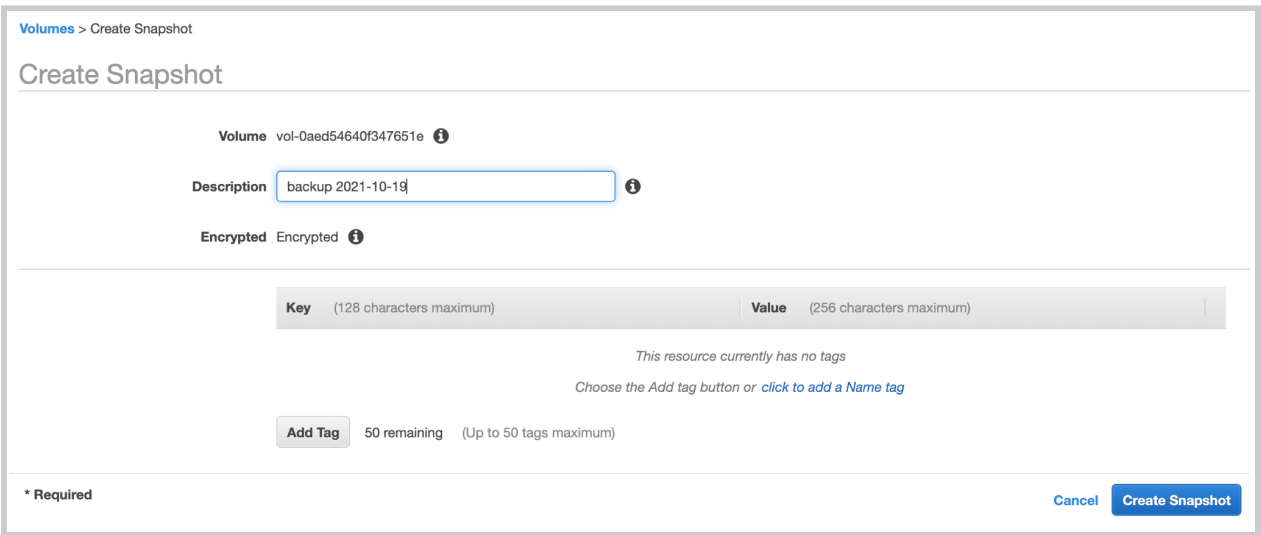
To backup your host instance, you will need to take a snapshot of the root EBS volume. To do so, launch the AWS console, navigate to the EC2 service page, then view your running instances. Select “FireflyHost” from the list of instances and click the “Storage” tab in the details pane, as shown in the screenshot below. Scroll down to the “Block devices” section and click on the Volume ID link for device /dev/xvda.



This should pull up the EBS volume page. Click the “Actions” dropdown menu at the top of the page and choose “Create Snapshot”.



On the Create Snapshot page, give the snapshot a description and click the blue “Create Snapshot” button.



Optionally, you can use the EBS Lifecycle Manager to create a Snapshot Lifecycle Policy to automate the creation of EBS snapshots on a regular basis.

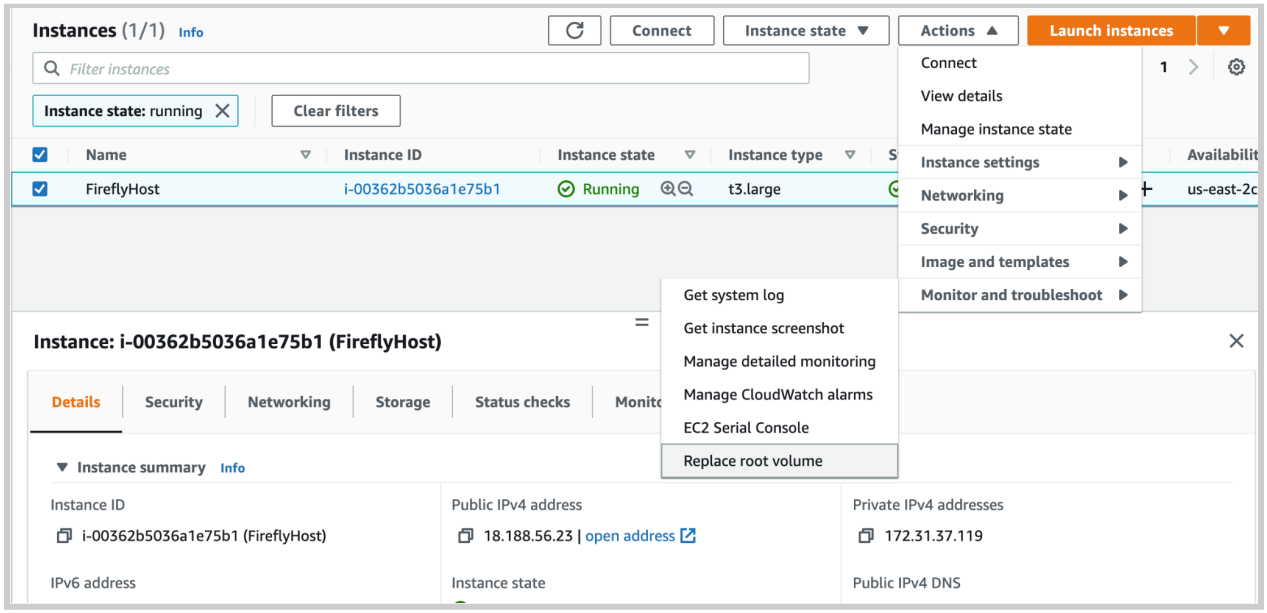
Restore Procedure

To restore a backup, you will need to create a new volume from the backup snapshot and replace the root volume with the restored volume on your Hyperdrive Notebooks host instance.

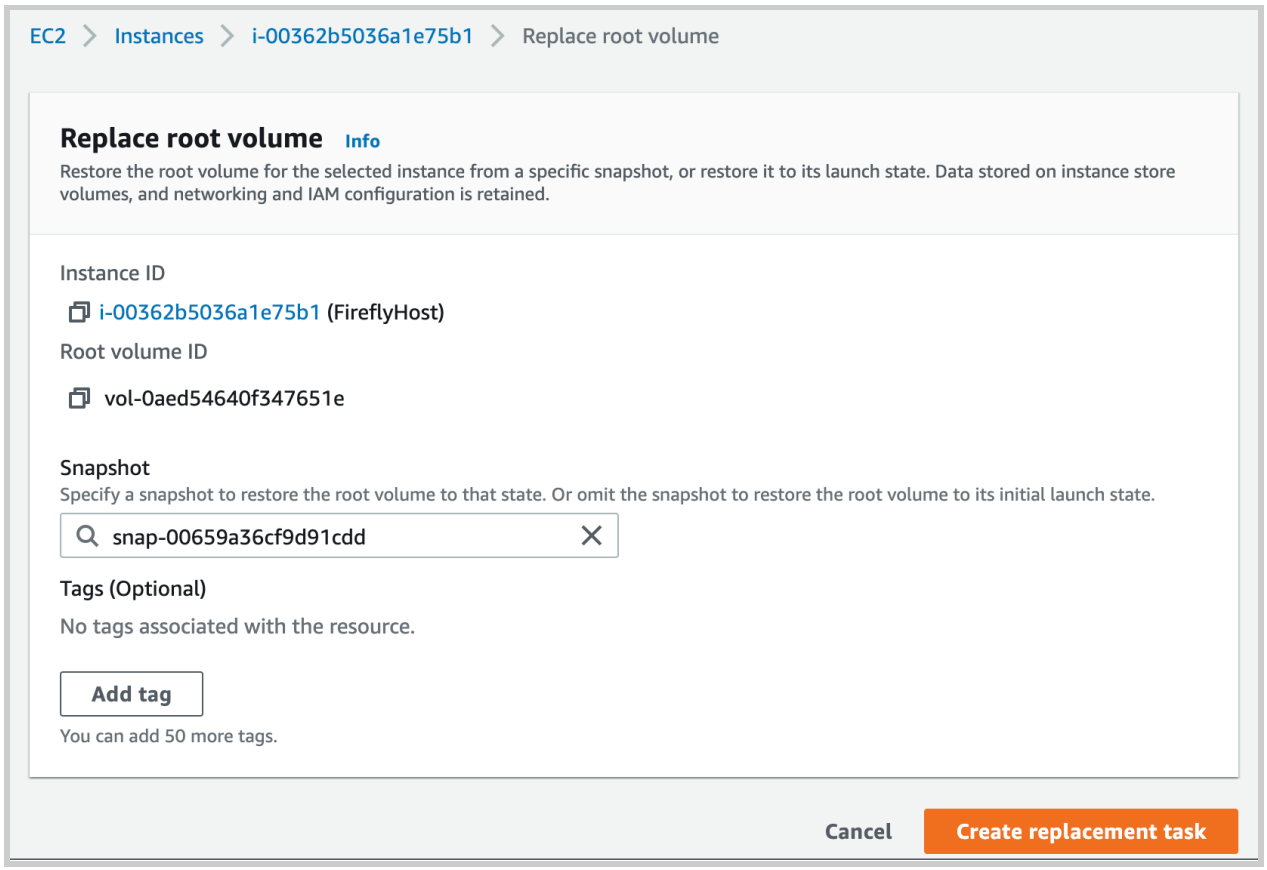
1. Navigate to the EC2 service in the AWS console.

2. From the Instances list, select FireflyHost.

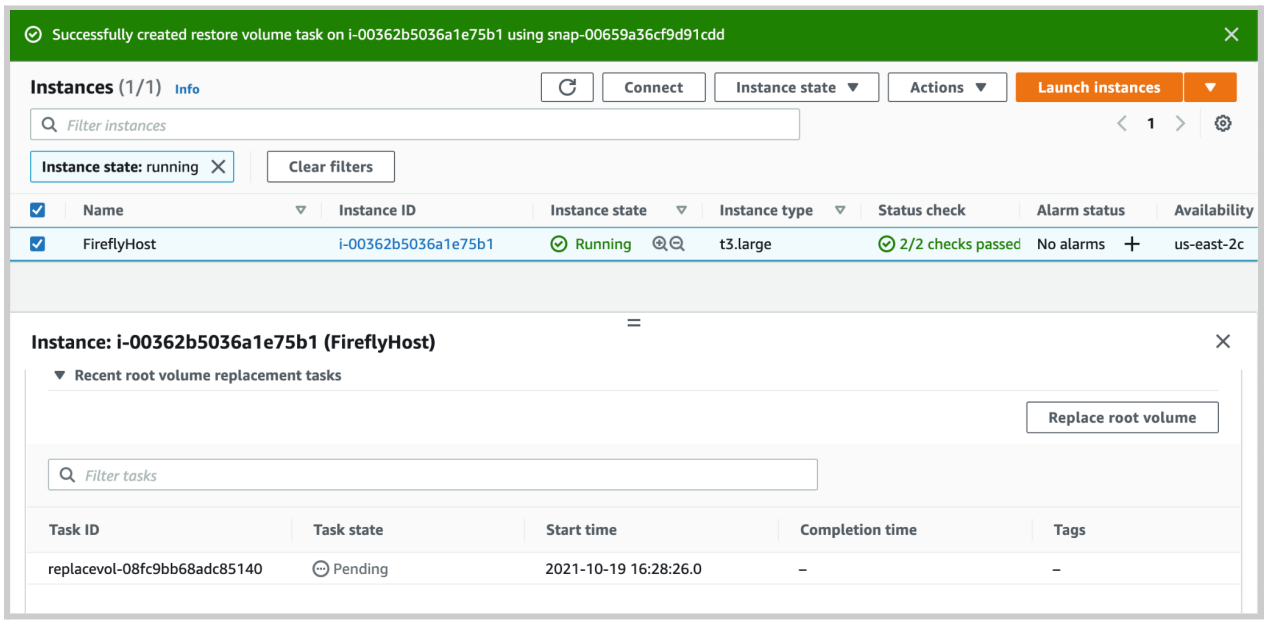
3. In the “Actions” dropdown menu, choose “Monitor and troubleshoot” > “Replace root volume”.



4. On the “Replace root volume” page, select the snapshot you wish to restore and click “Create replacement task”. Note that the replacement task will overwrite the existing root volume and reboot the host instance.



5. The replacement task will be created and can be monitored on the “Storage” tab of the instance details pane for the FireflyHost instance.



6. When the task state changes to “Successful”, you should be ready to log into your Hyperdrive Notebooks instance to verify that the data was correctly restored.

Routine Maintenance

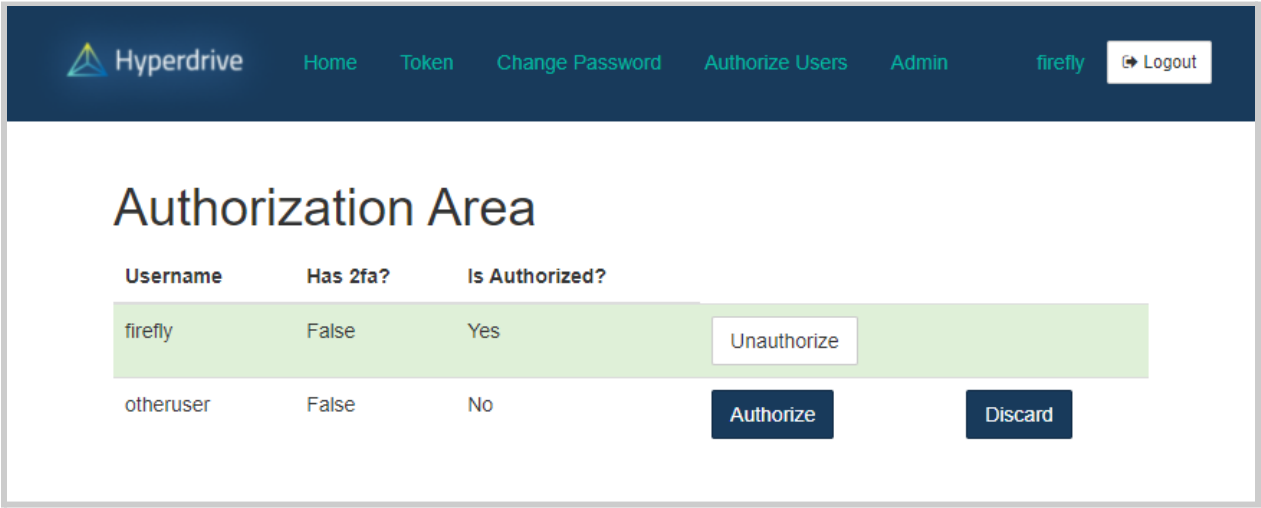
This section details routine operational procedures for your Hyperdrive Notebooks deployment.

Adding Additional Users

Adding additional users is a two-step process.

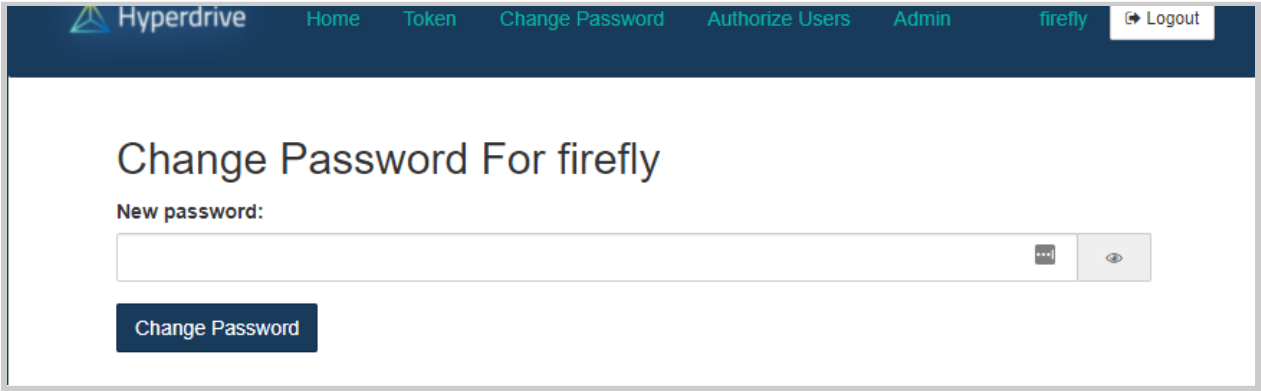
- First, the new user must sign up (or you can sign them up) at https://<your-fqdn>/hub/signup, which involves setting a username and password.

- Second, an administrator must authorize the user from the Authorization Area page, which is available by clicking the “Authorize Users” link in the navigation bar. To do so, click the “Authorize” button for each user you wish to authorize. The new user can now sign in to https://<your-fqdn> using the username and password they specified.



Rotating Keys & Credentials

JupyterHub users can change their passwords in the Hyperdrive Notebooks web interface by clicking on the “Change Password” link in the control panel, or by visiting https://<your-fqdn>/hub/change-password.



Accessing the Host Instance via SSH

Navigate to the EC2 page in the AWS Console. Select the FireflyHost instance from the list of EC2 instances. Click the “Connect” button. On the “Connect to instance” page, select the “SSH Client”. Follow the instructions to SSH to the host instance.

Once connected, you should be able to sudo to root and run kubernetes troubleshooting commands with kubectl.

```
[ec2-user@ip-172-31-40-5 ~]$ sudo su -
Last login: Wed Oct 20 16:04:28 UTC 2021 on pts/0
[root@ip-172-31-40-5 ~]# kubectl get pods --all-namespaces
NAMESPACE     NAME                                                      READY   STATUS    RESTARTS   AGE
kube-system   metrics-server-86cbb8457f-jw5hl                           1/1     Running   0          3h27m
kube-system   local-path-provisioner-5ff76fc89d-6m9pw                   1/1     Running   0          3h27m
kube-system   coredns-7448499f4d-vcm9b                                  1/1     Running   0          3h27m
firefly       continuous-image-puller-lm8j9                             1/1     Running   0          3h27m
firefly       firefly-nginx-ingress-controller-default-backend-65775f9ccnhhsw   1/1     Running   0          3h27m
firefly       svclb-firefly-nginx-ingress-controller-vcwfn              2/2     Running   0          3h27m
firefly       proxy-69d4fb8b5b-h75fq                                    1/1     Running   0          3h27m
firefly       user-scheduler-7d9dfdb6d-hlpz8                            1/1     Running   0          3h27m
firefly       user-scheduler-7d9dfdb6d-4zm2n                            1/1     Running   0          3h27m
firefly       firefly-nginx-ingress-controller-9d9ffd56b-w47hd          1/1     Running   0          3h27m
firefly       hub-7957b89679-k575t                                      1/1     Running   0          3h27m
firefly       jupyter-firefly                                           1/1     Running   0          3h23m
firefly       jupyter-firefly--demo                                     1/1     Running   0          3h22m
[root@ip-172-31-40-5 ~]#
```

Patching & Upgrades

- Update the base instance according to this guide: [Update instance software on your Amazon Linux instance - Amazon Elastic Compute Cloud]

- Upgrading k3d versions can be done via instructions on the k3d website: [k3d]

- Deploy a new version of the helm chart by doing a helm upgrade.
    - Connect to host via SSH

- Verify your kubeconfig works by executing `kubectl get pods --all-namespaces` (Hint: you may need to run `sudo su -` first)
- Navigate to the Firefly directory: `cd /home/ec2-user/firefly`
- Update the repo: `git pull`
- Upgrade your deployment: `helm upgrade firefly ./helm/firefly`
    - If you need to override or update any local values, you can use the helm guide found here: Helm Upgrade

Managing Licenses

Hyperdrive Notebooks is itself an open source product and is built on other open source products. There are no software licenses to manage.

Managing AWS Service Limits

Deploying Hyperdrive Notebooks is unlikely to significantly impact your AWS service limits, since the main resources being deployed are a single EC2 instance and an ALB. If you wish to review your current account service limits or wish to request a limit increase, please visit the Service Quotas page in the AWS console: https://console.aws.amazon.com/servicequotas

Emergency Maintenance

The two most likely scenarios that would lead to emergency loss of access or functionality to Firefly are an out-of-memory condition and a full disk.

Out-of-memory issues manifest by slow performance or the instance becoming unresponsive when issuing commands. If you are trying to use a remote connection utility such as SSH, this will likely also timeout and fail. These conditions are typically resolved by rebooting the machine. If you find that this happens frequently, consider taking a snapshot of the machine and restoring it to a larger instance size.

Full disk issues can manifest as a machine suddenly becoming unresponsive, as it is no longer able to allocate space for temporary files it needs to operate. For a disk full issue, please follow the instructions provided by AWS on expanding the root EBS volume and then expanding the filesystem on the host instance: Request modifications to your EBS volumes - Amazon Elastic Compute Cloud

Responding to Fault Conditions

Most faults can be identified and resolved by using standard Kubernetes troubleshooting methods. Typical issues include but are not limited to:

- Incorrect image names causing image pull errors
- Image pull secrets not being present or properly configured if using private images
- Containers failing to start correctly, placing the pod in a CrashLoopBackOff state

For troubleshooting these issues, we strongly recommend referring to the official Kubernetes documentation on troubleshooting (Troubleshoot Applications) and using k9s for interacting with the cluster (derailed/k9s: 🐶 Kubernetes CLI To Manage Your Clusters In Style!).

Recovery Procedures

Many fault conditions can be resolved by restarting the host instance. To do so, navigate to the EC2 service page in the AWS console, select the FireflyHost instance, then from the “Instance state” dropdown menu, choose “Reboot instance”.

Support Resources

This section details the process for getting support on your Hyperdrive Notebooks deployment.

Support Tiers Available

Hypergiant offers two support models for Hyperdrive Notebook customers: free community support available through public channels and paid commercial support available through private channels.

Accessing Technical Support

For community support, please open a Github issue on the main Hyperdrive Notebooks repo: https://github.com/gohypergiant/firefly/issues

For commercial support, please contact the Hypergiant support team at support@hypergiant.com.

Support Tier SLAs

Community support responses are provided on a best effort basis.

The response SLA for commercial support is 4 hours.