Task8

Task 8: AI Model Deployment & MLOps

We have used ECS to host the Docker task and added an ALB (Application Load Balancer) to the publicly exposed Docker service. This is all achieved using a CloudFormation template and ECS Fargate.

The Docker service takes a voice sound file and converts it into text.

The famous Docker image used for this is `onerahmet/openai-whisper-asr-webservice:latest`.

Dockerfile & Kubernetes YAML Files

CloudFormation File

```
AWSTemplateFormatVersion: '2010-09-09'

Description: Deploy Whisper ASR API to ECS Fargate
Parameters:
  WhisperModel:
    Type: String
    Default: tiny
    AllowedValues: [tiny, base, small, medium, large]
    Description: Whisper model to use

Resources:
  WhisperVPC:
    Type: AWS::EC2::VPC
    Properties:
      CidrBlock: 10.0.0.0/16
      EnableDnsSupport: true
      EnableDnsHostnames: true
      Tags: [{ Key: Name, Value: WhisperVPC }]

  WhisperSubnet1:
```

```yaml
    Type: AWS::EC2::Subnet
    Properties:
      VpcId: !Ref WhisperVPC
      CidrBlock: 10.0.1.0/24
      AvailabilityZone: !Select [0, !GetAZs '']
      MapPublicIpOnLaunch: true

  WhisperSubnet2:
    Type: AWS::EC2::Subnet
    Properties:
      VpcId: !Ref WhisperVPC
      CidrBlock: 10.0.2.0/24
      AvailabilityZone: !Select [1, !GetAZs '']
      MapPublicIpOnLaunch: true

  WhisperInternetGateway:
    Type: AWS::EC2::InternetGateway

  WhisperAttachGateway:
    Type: AWS::EC2::VPCGatewayAttachment
    Properties:
      VpcId: !Ref WhisperVPC
      InternetGatewayId: !Ref WhisperInternetGateway

  WhisperRouteTable:
    Type: AWS::EC2::RouteTable
    Properties:
      VpcId: !Ref WhisperVPC

  WhisperRoute:
    Type: AWS::EC2::Route
    DependsOn: WhisperAttachGateway
    Properties:
      RouteTableId: !Ref WhisperRouteTable
      DestinationCidrBlock: 0.0.0.0/0
      GatewayId: !Ref WhisperInternetGateway

  WhisperSubnetRouteTableAssoc1:
    Type: AWS::EC2::SubnetRouteTableAssociation
    Properties:
```

```
      SubnetId: !Ref WhisperSubnet1
      RouteTableId: !Ref WhisperRouteTable

  WhisperSubnetRouteTableAssoc2:
    Type: AWS::EC2::SubnetRouteTableAssociation
    Properties:
      SubnetId: !Ref WhisperSubnet2
      RouteTableId: !Ref WhisperRouteTable

  WhisperSecurityGroup:
    Type: AWS::EC2::SecurityGroup
    Properties:
      GroupDescription: Allow HTTP access
      VpcId: !Ref WhisperVPC
      SecurityGroupIngress:
        - IpProtocol: tcp
          FromPort: 9000
          ToPort: 9000
          CidrIp: 0.0.0.0/0

  WhisperCluster:
    Type: AWS::ECS::Cluster

  WhisperTaskExecutionRole:
    Type: AWS::IAM::Role
    Properties:
      AssumeRolePolicyDocument:
        Statement:
          - Effect: Allow
            Principal:
              Service: ecs-tasks.amazonaws.com
            Action: sts:AssumeRole
      ManagedPolicyArns:
        - arn:aws:iam::aws:policy/service-
role/AmazonECSTaskExecutionRolePolicy

  WhisperTaskDefinition:
    Type: AWS::ECS::TaskDefinition
    Properties:
      Family: whisper-task
```

```
      RequiresCompatibilities: [FARGATE]
      Cpu: 512
      Memory: 1024
      NetworkMode: awsvpc
      ExecutionRoleArn: !GetAtt WhisperTaskExecutionRole.Arn
      ContainerDefinitions:
        - Name: whisper
          Image: onerahmet/openai-whisper-asr-
webservice:latest
          PortMappings:
            - ContainerPort: 9000
          Environment:
            - Name: ASR_MODEL
              Value: !Ref WhisperModel

  WhisperService:
    Type: AWS::ECS::Service
    DependsOn: WhisperALBListener
    Properties:
      Cluster: !Ref WhisperCluster
      LaunchType: FARGATE
      DesiredCount: 1
      NetworkConfiguration:
        AwsvpcConfiguration:
          AssignPublicIp: ENABLED
          SecurityGroups: [!Ref WhisperSecurityGroup

          Subnets: [!Ref WhisperSubnet1, !Ref WhisperSubnet2]
      TaskDefinition: !Ref WhisperTaskDefinition
      LoadBalancers:
        - ContainerName: whisper
          ContainerPort: 9000
          TargetGroupArn: !Ref WhisperTargetGroup

  WhisperALB:
    Type: AWS::ElasticLoadBalancingV2::LoadBalancer
    Properties:
      Name: whisper-alb
      Subnets: [!Ref WhisperSubnet1, !Ref WhisperSubnet2]
      SecurityGroups: [!Ref WhisperSecurityGroup]
```

```
      Scheme: internet-facing
      Type: application

  WhisperTargetGroup:
    Type: AWS::ElasticLoadBalancingV2::TargetGroup
    Properties:
      Port: 9000
      Protocol: HTTP
      VpcId: !Ref WhisperVPC
      TargetType: ip
      HealthCheckPath: /docs

  WhisperALBListener:
    Type: AWS::ElasticLoadBalancingV2::Listener
    Properties:
      LoadBalancerArn: !Ref WhisperALB
      Port: 9000
      Protocol: HTTP
      DefaultActions:
        - Type: forward
          TargetGroupArn: !Ref WhisperTargetGroup

Outputs:
  WhisperAPIURL:
    Description: Whisper REST API URL
    Value: !Join ["", ["http://", !GetAtt WhisperALB.DNSName,
":9000"]]
```

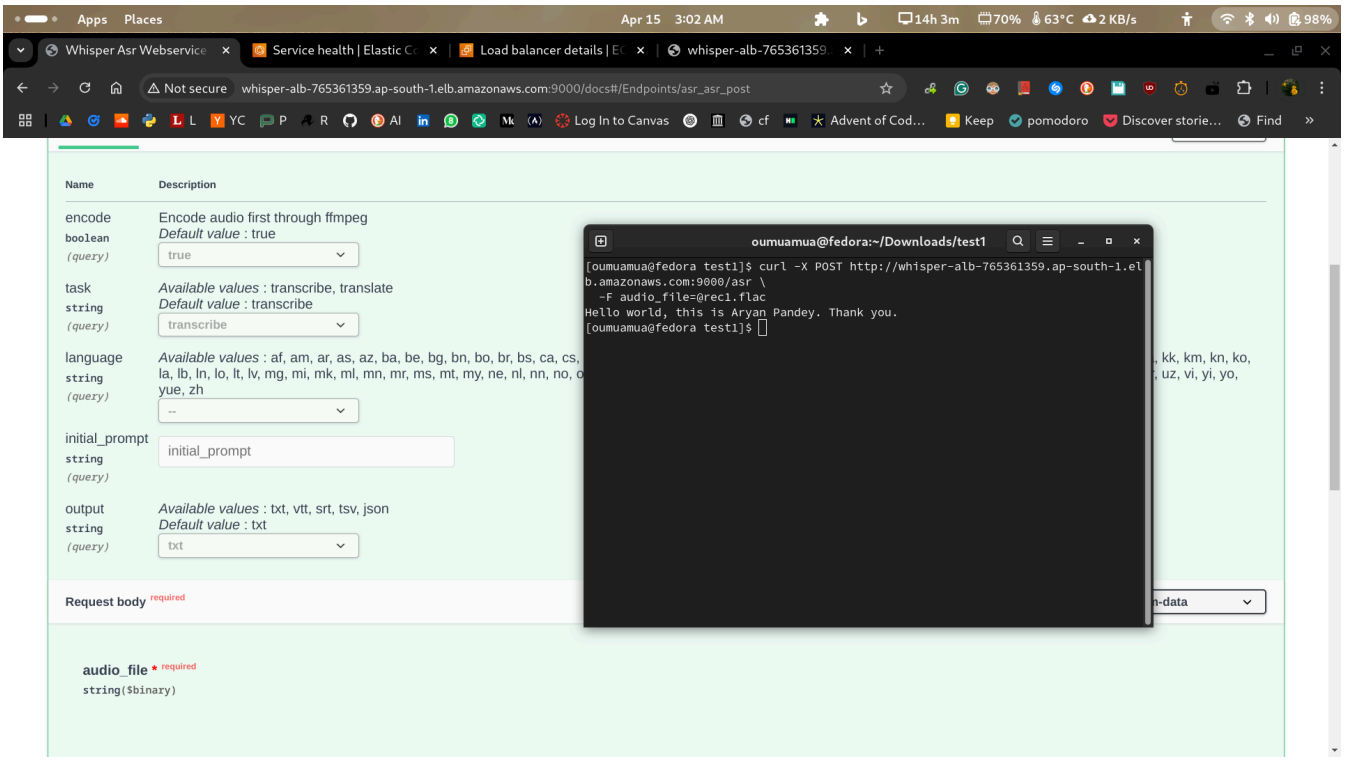
Steps to Deploy the Model

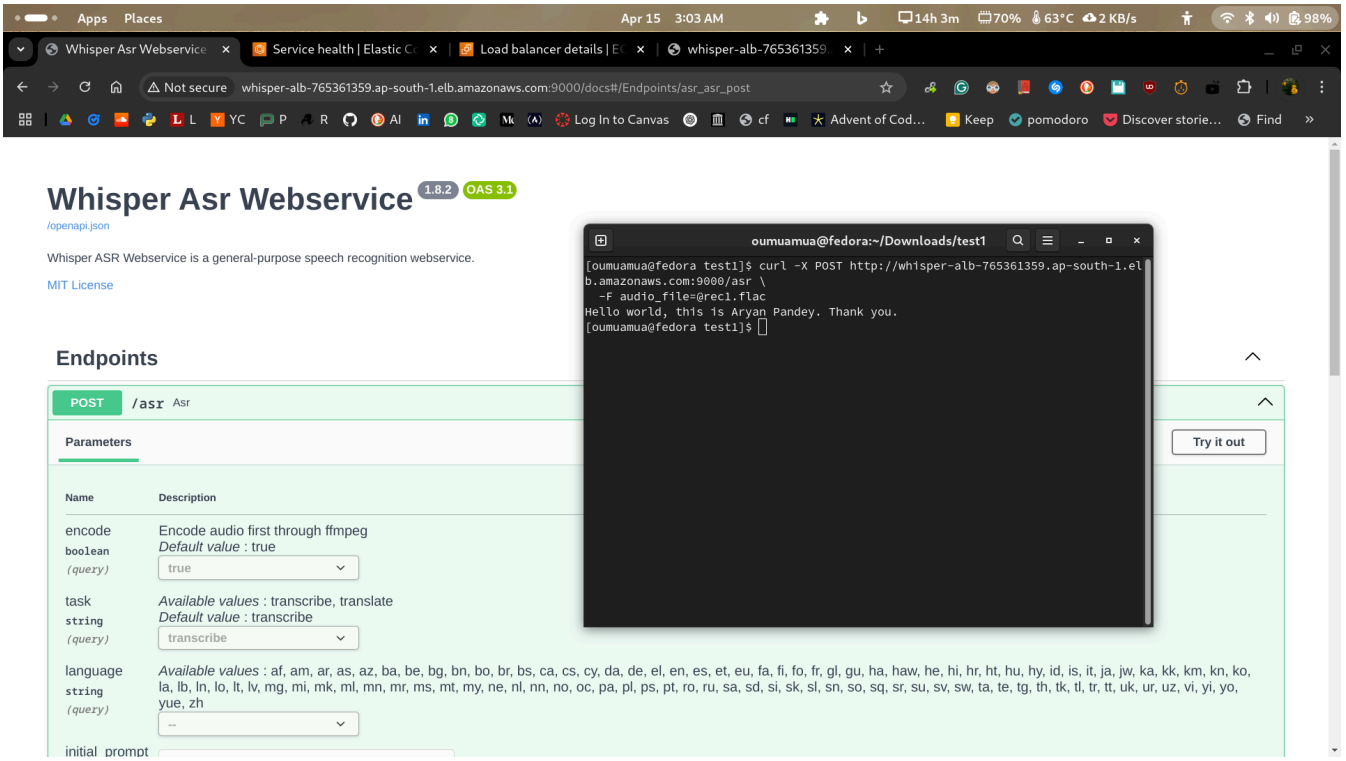
Deploy this using the AWS CLI deploy command:

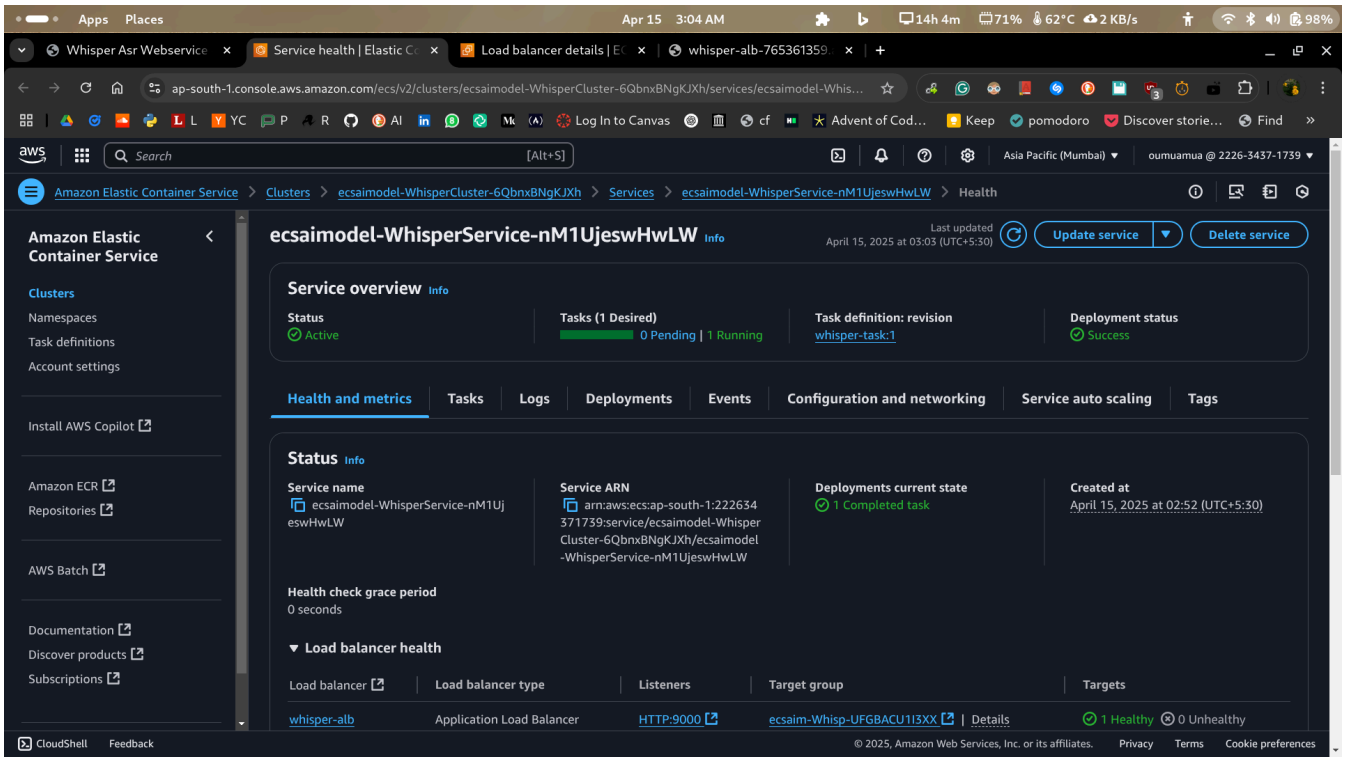
```
#!/bin/bash

aws cloudformation deploy --region ap-south-1 \
  --template-file ./main.yaml \
  --stack-name ecsaimodel \
  --tags madeFromCLI=yeah anotherTagForAllStackResources=okay
\
  --capabilities CAPABILITY_NAMED_IAM
```

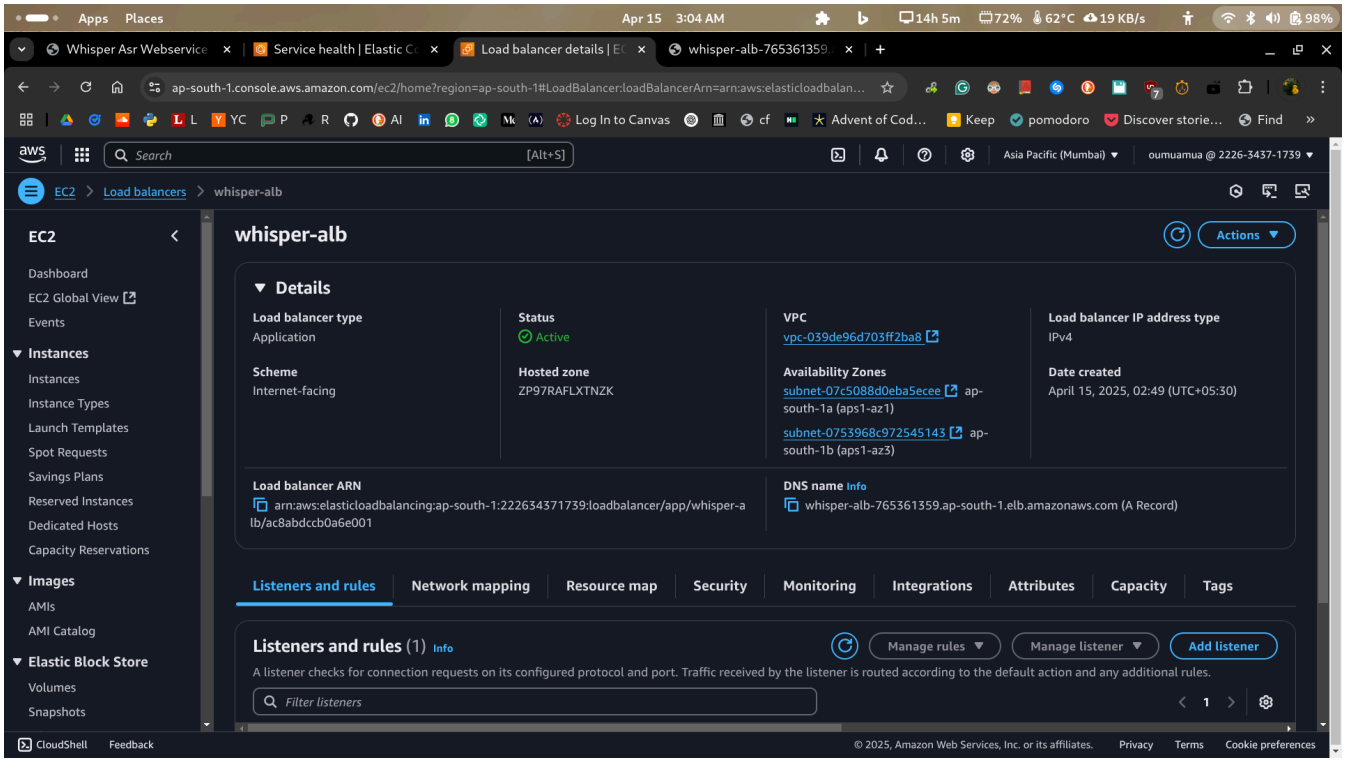
```
# --no-execute-changeset
```

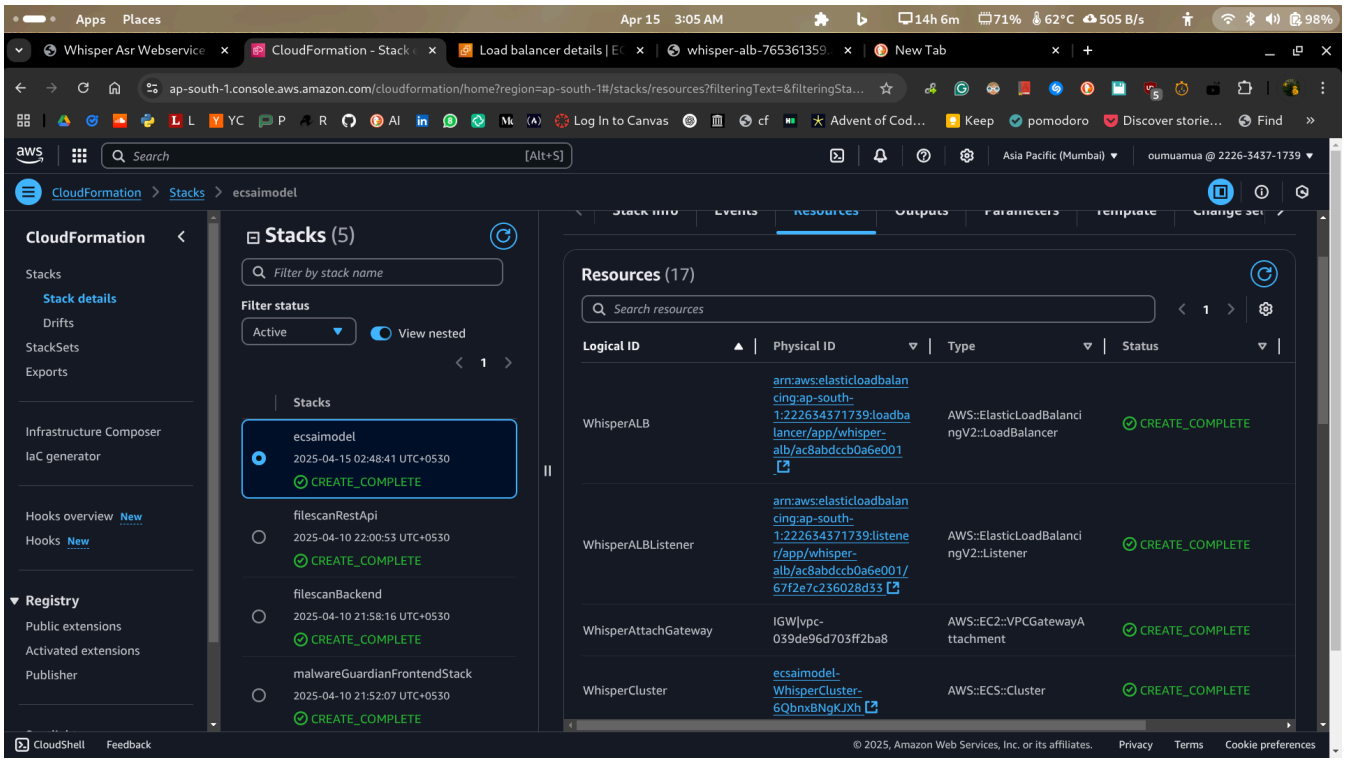
Screenshot of the model running on ECS

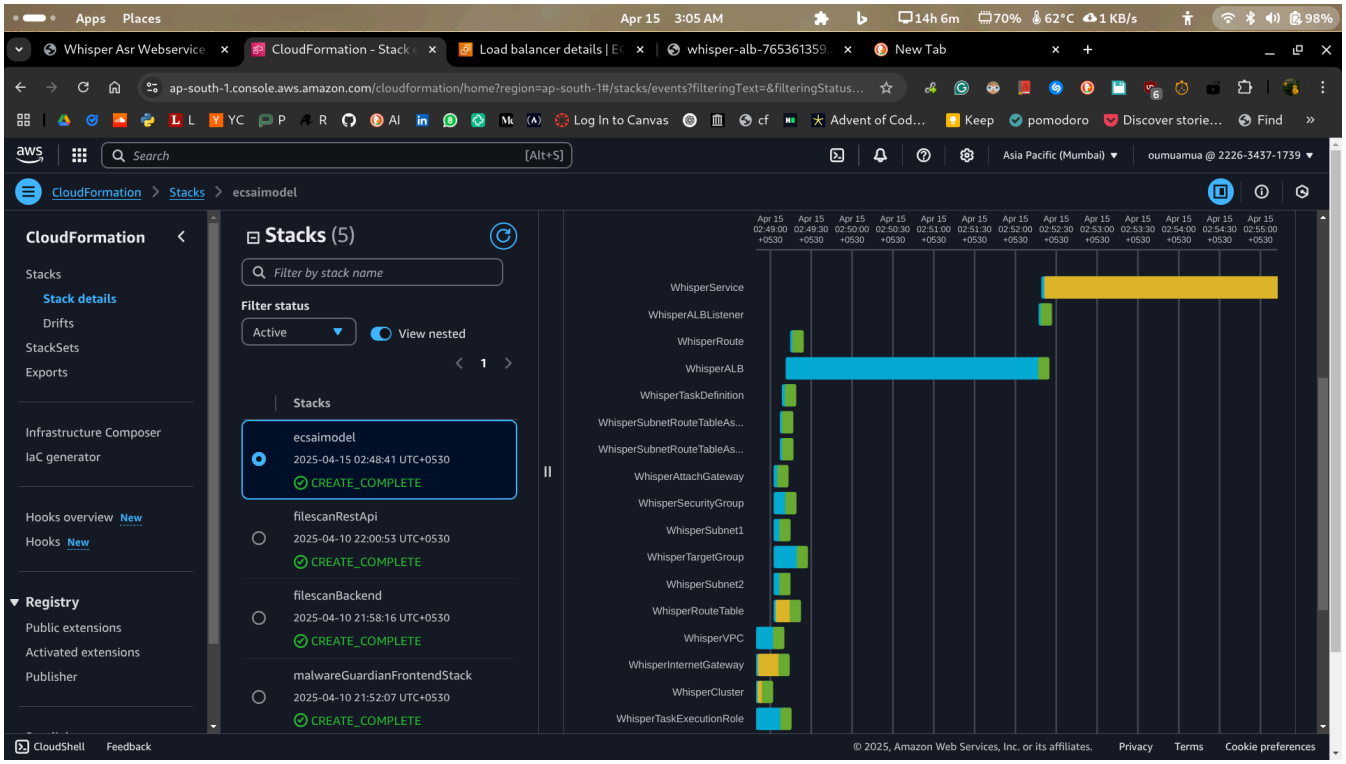





Apps Places Apr 15 3:04 AM 14h 4m 71% 62°C 2 KB/s 98%
Whisper Asr Webservice
Service health | Elastic C
Load balancer details | E
whisper-alb-765361359
ap-south-1.console.aws.amazon.com/ecs/v2/clusters/ecsaimodel-WhisperCluster-6QbnxBNgKJXh/services/ecsaimodel-Whis...
Log In to Canvas
Advent of Cod...
Keep
pomodoro
Discover storie...
Find
Search [Alt+S]
Asia Pacific (Mumbai)
oumuamua @ 2226-3437-1739
Amazon Elastic Container Service > Clusters > ecsaimodel-WhisperCluster-6QbnxBNgKJXh > Services > ecsaimodel-WhisperService-nM1UjeswHwLW > Health
Amazon Elastic Container Service
Clusters
Namespaces
Task definitions
Account settings
Install AWS Copilot
Amazon ECR
Repositories
AWS Batch
Documentation
Discover products
Subscriptions
ecsaimodel-WhisperService-nM1UjeswHwLW Info
Last updated April 15, 2025 at 03:03 (UTC+5:30)
Update service
Delete service
Service overview Info
Status
Active
Tasks (1 Desired)
0 Pending | 1 Running
Task definition: revision
whisper-task:1
Deployment status
Success
Health and metrics
Tasks
Logs
Deployments
Events
Configuration and networking
Service auto scaling
Tags
Status Info
Service name
ecsaimodel-WhisperService-nM1UjeswHwLW
Service ARN
arn:aws:ecs:ap-south-1:222634371739:service/ecsaimodel-WhisperCluster-6QbnxBNgKJXh/ecsaimodel-WhisperService-nM1UjeswHwLW
Deployments current state
1 Completed task
Created at
April 15, 2025 at 02:52 (UTC+5:30)
Health check grace period
0 seconds
Load balancer health
Load balancer
Load balancer type
Listeners
Target group
Targets
whisper-alb
Application Load Balancer
HTTP:9000
ecsaim-Whisp-UFGBACU1I3XX | Details
1 Healthy 0 Unhealthy
CloudShell Feedback
© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences


Apps Places Apr 15 3:04 AM 14h 5m 72% 62°C 19 KB/s 98%
Whisper Asr Webservice
Service health | Elastic C
Load balancer details | E
whisper-alb-765361359
ap-south-1.console.aws.amazon.com/ec2/home?region=ap-south-1#LoadBalancer:loadBalancerArn=arn:aws:elasticloadbalan...
Log In to Canvas
Advent of Cod...
Keep
pomodoro
Discover storie...
Find
Search [Alt+S]
Asia Pacific (Mumbai)
oumuamua @ 2226-3437-1739
EC2 > Load balancers > whisper-alb
EC2
Dashboard
EC2 Global View
Events
Instances
Instances
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances
Dedicated Hosts
Capacity Reservations
Images
AMIs
AMI Catalog
Elastic Block Store
Volumes
Snapshots
whisper-alb
Actions
Details
Load balancer type
Application
Status
Active
VPC
vpc-039de96d703ff2ba8
Load balancer IP address type
IPv4
Scheme
Internet-facing
Hosted zone
ZP97RAFLXTNZK
Availability Zones
subnet-07c5088d0eba5ecee ap-south-1a (aps1-az1)
subnet-0753968c972545143 ap-south-1b (aps1-az3)
Date created
April 15, 2025, 02:49 (UTC+05:30)
Load balancer ARN
arn:aws:elasticloadbalancing:ap-south-1:222634371739:loadbalancer/app/whisper-alb/ac8abdccb0a6e001
DNS name Info
whisper-alb-765361359.ap-south-1.elb.amazonaws.com (A Record)
Listeners and rules
Network mapping
Resource map
Security
Monitoring
Integrations
Attributes
Capacity
Tags
Listeners and rules (1) Info
Manage rules
Manage listener
Add listener
A listener checks for connection requests on its configured protocol and port. Traffic received by the listener is routed according to the default action and any additional rules.
Filter listeners
CloudShell Feedback
© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences


Apps Places Apr 15 3:05 AM 14h 6m 71% 62°C 505 B/s 98%
Whisper Asr Webservice | CloudFormation - Stack | Load balancer details | whisper-alb-765361359 | New Tab
ap-south-1.console.aws.amazon.com/cloudformation/home?region=ap-south-1#/stacks/resources?filteringText=&filteringSta...
Search [Alt+S] Asia Pacific (Mumbai) oumuamua @ 2226-3437-1739
CloudFormation > Stacks > ecsaimodel
CloudFormation
Stacks
Stack details
Drifts
StackSets
Exports
Infrastructure Composer
IaC generator
Hooks overview New
Hooks New
Registry
Public extensions
Activated extensions
Publisher
Stacks (5)
Filter by stack name
Filter status
Active
View nested
Stacks
ecsaimodel
2025-04-15 02:48:41 UTC+0530
CREATE_COMPLETE
filescanRestApi
2025-04-10 22:00:53 UTC+0530
CREATE_COMPLETE
filescanBackend
2025-04-10 21:58:16 UTC+0530
CREATE_COMPLETE
malwareGuardianFrontendStack
2025-04-10 21:52:07 UTC+0530
CREATE_COMPLETE
Resources (17)
Search resources
Logical ID
Physical ID
Type
Status
WhisperALB
arn:aws:elasticloadbalancing:ap-south-1:222634371739:loadbalancer/app/whisper-alb/ac8abdccb0a6e001
AWS::ElasticLoadBalancingV2::LoadBalancer
CREATE_COMPLETE
WhisperALBListener
arn:aws:elasticloadbalancing:ap-south-1:222634371739:listener/app/whisper-alb/ac8abdccb0a6e001/67f2e7c236028d33
AWS::ElasticLoadBalancingV2::Listener
CREATE_COMPLETE
WhisperAttachGateway
IGW|vpc-039de96d703ff2ba8
AWS::EC2::VPCGatewayAttachment
CREATE_COMPLETE
WhisperCluster
ecsaimodel-WhisperCluster-6QbnxBNgKJXh
AWS::ECS::Cluster
CREATE_COMPLETE
CloudShell Feedback
© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences


Apps Places Apr 15 3:05 AM 14h 6m 70% 62°C 1 KB/s 98%
Whisper Asr Webservice | CloudFormation - Stack | Load balancer details | whisper-alb-765361359 | New Tab
ap-south-1.console.aws.amazon.com/cloudformation/home?region=ap-south-1#/stacks/events?filteringText=&filteringStatus...
Search [Alt+S] Asia Pacific (Mumbai) oumuamua @ 2226-3437-1739
CloudFormation > Stacks > ecsaimodel
CloudFormation
Stacks
Stack details
Drifts
StackSets
Exports
Infrastructure Composer
IaC generator
Hooks overview New
Hooks New
Registry
Public extensions
Activated extensions
Publisher
Stacks (5)
Filter by stack name
Filter status
Active
View nested
Stacks
ecsaimodel
2025-04-15 02:48:41 UTC+0530
CREATE_COMPLETE
filescanRestApi
2025-04-10 22:00:53 UTC+0530
CREATE_COMPLETE
filescanBackend
2025-04-10 21:58:16 UTC+0530
CREATE_COMPLETE
malwareGuardianFrontendStack
2025-04-10 21:52:07 UTC+0530
CREATE_COMPLETE
WhisperService
WhisperALBListener
WhisperRoute
WhisperALB
WhisperTaskDefinition
WhisperSubnetRouteTableAs...
WhisperSubnetRouteTableAs...
WhisperAttachGateway
WhisperSecurityGroup
WhisperSubnet1
WhisperTargetGroup
WhisperSubnet2
WhisperRouteTable
WhisperVPC
WhisperInternetGateway
WhisperCluster
WhisperTaskExecutionRole
CloudShell Feedback
© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences