

快速部署数据探索及可视化平台

Apache Superset

部署指南

易珂

版本: v1.2.0

最后更新时间: 2022 年 09 月



Copyright (c) 2022 by Amazon.com, Inc. or its affiliates.

目 录

目 录	2
背景.....	3
解决方案描述.....	3
系统架构.....	3
架构图.....	3
部署服务.....	5
部署说明.....	5
部署前提.....	5
登陆到您的账户.....	6
快速部署.....	6
启动 AMAZON CLOUDFORMATION 堆栈.....	7
指定堆栈详细信息.....	7
配置堆栈选项.....	8
审核.....	8
开始使用.....	9
登录 APACHE SUPERSET 控制台	9
登录 CLOUDWATCH 仪表板.....	9
对接现有数据.....	10
FAQ	13
版本发布说明	14
附录.....	14
CLOUDFORMATION 堆栈参数.....	14
源代码地址.....	15

背景

本部署指南详细描述了如何在云上使用 Amazon CloudFormation 模板部署基于 Amazon Elastic Container Service (ECS) , Cloud Map, Elastic File System (EFS) 等服务构建的 Apache Superset, 实现 Business Intelligence (BI) 的快速构建以帮助 IT 人员, 数据分析师, 业务决策者聚焦业务本身, 加速企业实现数据驱动认知到数据驱动决策的转变。

解决方案描述

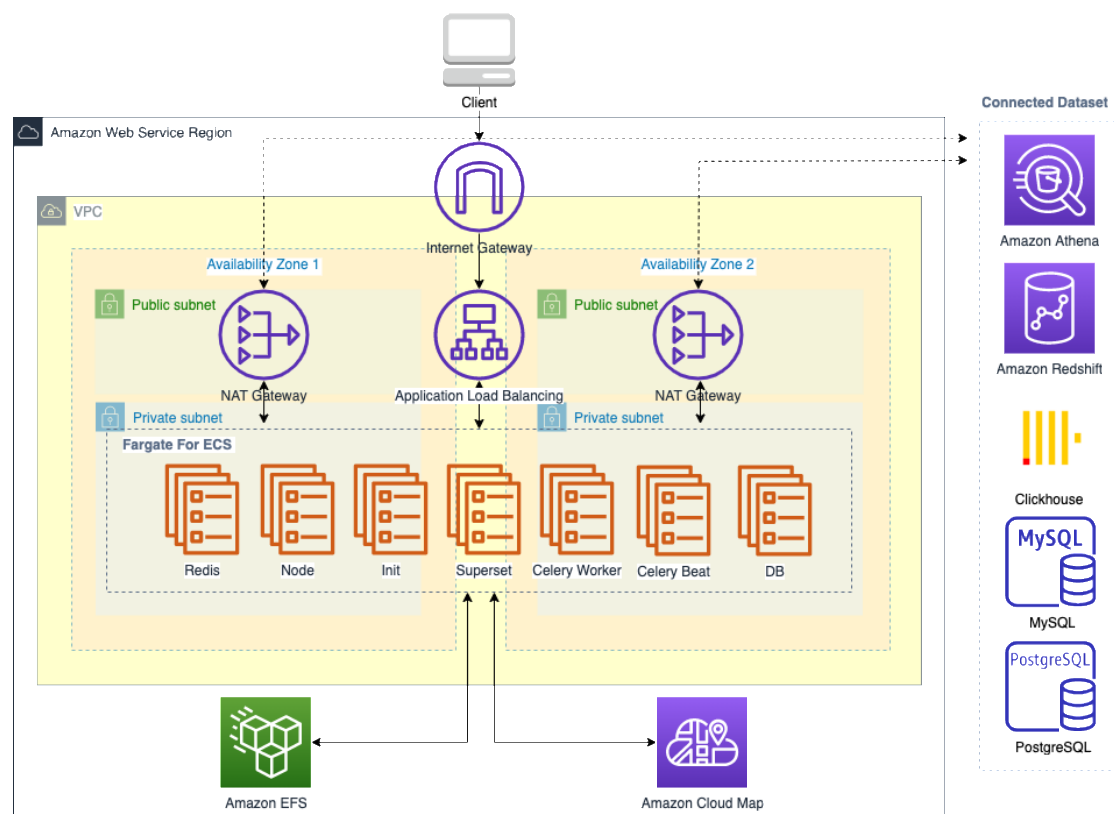
该解决方案利用 Amazon CloudFormation 实现 Apache Superset 平台的自动部署, 其中 Apache Superset 是一个开源的基于主流云原生技术构建的数据可视分析平台 Business Intelligence (BI) , 给用户提供轻量, 直观, 可定制的操作界面来对接各种数据源, 实现数据的查询, 编排和可视化。通过结合 Amazon Elastic Container Service (Amazon ECS), Cloud Map 等托管服务, 我们可以将 Apache Superset 快速构建至托管的容器集群, 而无需安装、操作和扩展额外的容器编排和集群管理基础设施。

系统架构

架构图

Apache Superset 各功能模块采用松耦合的方式独立开发迭代, 模块间的通信通过 Celery 构建的分布式消息队列实现, 对于容器技术如 Docker, Kubernetes 的支持相对完善。在亚马逊云平台运行的 Apache Superset 业务功能通过 Amazon ECS 进行托管, 利用 ECS Fargate 实现资源的调度, 服务的健康检查, 各 ECS 服务本身通过 Cloud Map 创建的私有 DNS 进行服务发现, 寻址和通信, 各 ECS 服务自身的数据存

储共享通过 Amazon EFS 实现，以获取更好的可用性，灵活性和低成本。网络规划我们遵循 Amazon 最佳实践，用户上行 inbound 访问流量通过 Application Load Balancer 对接到 ECS 集群的 Superset Service，Superset Service 的下行 outbound 流量如连接外部数据源，获取示例数据等则通过 NAT Gateway 实现，结合 VPC 安全组实现网络流量的端口控制（如 Superset 默认的端口 8088），整体方案的软件架构如下图所示：



相较于 Apache 社区版本，运行在亚马逊云上的 Apache Superset 存在如下优势：

- 核心模块（Superset，Cache，Database）高可用；
- 业务数据（元数据，查询数据，交互数据）持久化；
- 平台资源弹性伸缩，用户无需关心底层资源调度；
- 预装 SQL，PostgreSQL，Redshift，Athena，ClickHouse 数据源驱动，创建完毕即可连接已有数据
- 预装时序预测算法，基于导入数据实现未来趋势预测
- 可视化看板实时监控亚马逊云服务各项指标以及应用综合指标

部署服务

该解决方案可在由西云数据运营的 Amazon（宁夏）区域或由光环新网运营的 Amazon（北京）区域中部署，也可在海外地域实现部署，具体支持区域参见[这里](#)。

整个方案由以下几个模块组成：

- Amazon VPC，构建安全灵活可靠的虚拟网络，实现业务数据的双向流动；
- Amazon Elastic Container Service（ECS），利用无服务计算引擎 Fargate 实现 Apache Superset 各核心组件的负载隔离，资源调度；
- Amazon Elastic File System（EFS），提供可扩展，弹性，安全的 NAS 存储业务元数据，查询缓存等数据；
- Amazon Cloud Map，实现 Superset 各核心组件的服务发现，位置查询

待方案成功部署完毕后，用户可在 CloudFormation 的 Output 选项中查看 Apache Superset 的控制台地址，登陆之后即可开始连接现有数据源进行数据的查询，分析和可视化。方案本身预装 SQL，PostgreSQL，SQL Server，Oracle，ElasticSearch，Redshift，Athena，Clickhouse 数据源驱动，用户无需额外操作。

除此之外，该解决方案会根据亚马逊最佳实践创建一个 Amazon Virtual Private Cloud（Amazon VPC），作为 Amazon ECS Fargate 容器集群运行的网络基础设施，为整体方案提供安全可靠，灵活可配的虚拟联网环境。

部署说明

部署前提

- 确保您 Application Load Balancer（ALB）的 80/443 端口已正常备案并打开（仅限中国大陆用户）
- 确保您服务的配额足够，该方案所用到的服务以及配额如下：

所用服务	服务配额
VPCs	<1>
Nat Gateway	<2>
Elastic IP addresses	<2>
Security groups	<1>
Identity and Access Management (IAM) roles	<14>
Application Load Balancers	<1>
Elastic Container Service (ECS) Cluster	<1>
Elastic Container Service (ECS) Service	<7>
Elastic File System (EFS) AccessPoint	<3>
Elastic File System (EFS) FileSystem	<3>
Cloud Map ServiceDiscovery	<7>

登陆到您的账户

使用具有必需权限的 IAM 用户角色，通过 <https://console.amazonaws.cn/> 登录到您的 Amazon 账户。

快速部署

下文步骤主要针对在运行在由西云数据运营的 Amazon（宁夏）区域或由光环新网运营的 Amazon（北京）区域中部署该解决方案，针对海外区域的部署流程则完全相同，在此不再赘述。

启动 Amazon CloudFormation 堆栈

单击下面的按钮将打开 Amazon 管理控制台（如果还没登录会先跳转到登录页面进行登录）并进入 Amazon CloudFormation 启动界面，点击【下一步】进行启动参数配置。

北京（部署到已有 VPC）



北京（部署到新建 VPC）



宁夏（部署到已有 VPC）



宁夏（部署到新建 VPC）



CloudFormation > 堆栈 > 创建堆栈

第 1 步
指定模板

第 2 步
指定堆栈详细信息

第 3 步
配置堆栈选项

第 4 步
审核

创建堆栈

先决条件 - 准备模板

准备模板
每个堆栈都基于一个模板。模板是一个 JSON 或 YAML 文件，其中包含有关您希望在堆栈中包含的 亚马逊科技 资源的配置信息。

☒ 模板已就绪 ☐ 使用示例模板 ☐ 在设计器中创建模板

指定模板
模板是一个 JSON 或 YAML 文件，该文件描述了堆栈的资源属性。

模板源
选择模板会生成一个 Amazon S3 URL，在那里它将被存储。

☒ Amazon S3 URL ☐ 上传模板文件

Amazon S3 URL

Amazon S3 模板 URL

S3 URL: <https://superset-template.s3.cn-north-1.amazonaws.com.cn/superset-on-aws.main.template.yaml> [在设计器中查看](#)

取消 [下一步](#)

指定堆栈详细信息

在此填入需要创建堆栈名称，登录 Apache Superset 控制台的用户名和密码。其中“Pre-populate example dashboard”用于配置是否要获取官方内置示例数据集以及仪表板，“Install Prophet library”用于配置是否要安装 Prophet 软件包以实现数据的在线预测功能，“Elastic Container Cluster Name”用于配置平台所运行的 Elastic Container Service 集群的名称。其他配置采用默认值即可，之后点击【下一

步】到堆栈选项。

Account Configuration, set account info

Initial Superset User Name

Superset user name.

supersetuser

Initial Superset User Password

Superset user password. Strict password policy is recommended.

Pre-populate Example Dashboard

Enables or disables pre-populating official dashboard example. no = DON'T pre-populate official dashboard example; yes = pre-populate official dashboard example

no

Install Prophet library

Enables or disables Prophet library installation for Forecasting Analytics. no = DON'T install Prophet for Forecasting Analytics; yes = install Prophet to enable Forecasting Analytics

no

Elastic Container Cluster Name

Name for the ECS Cluster to use.

supersetOnAWS

下列参数中的 QSS3BucketRegion 默认为北京（cn-north-1），如果您在宁夏部署，需要修改成 cn-northwest-1。

其他参数

QSS3BucketName

S3 bucket name for the Quick Start assets. This string can include numbers, lowercase letters, uppercase letters, and hyphens (-). It cannot start or end with a hyphen (-).

aws-cn-quickstart

QSS3BucketRegion

Region where the Quick Start S3 bucket (QSS3BucketName) is hosted. When using your own bucket, you must specify this value.

cn-north-1

QSS3KeyPrefix

S3 key prefix for the Quick Start assets. Quick Start key prefix can include numbers, lowercase letters, uppercase letters, hyphens (-), periods (.) and forward slash (/).

quickstart-superset-on-aws/

配置堆栈选项

该步骤按照默认值并点击【下一步】进入审核。

审核

勾选如下复选框，然后点击【创建堆栈】。

功能

The following resource(s) require capabilities: [IAM::CloudFormation::Stack]

此模板包含 Identity and Access Management (IAM) 资源。检查您想要创建每一个这些资源，他们拥有所需的最低权限。此外，他们有自定义名称检查自定义名称在您的 亚马逊云科技 账户中是否是唯一的。 [了解更多。](#)

对于此模板，Amazon CloudFormation 可能需要一个无法识别的功能: CAPABILITY_AUTO_EXPAND。检查这些资源的功能。 [了解更多。](#)

☒ 我确认，Amazon CloudFormation 可能创建具有自定义名称的 IAM 资源。

☒ 我确认，Amazon CloudFormation 可能需要以下功能: CAPABILITY_AUTO_EXPAND

取消 上一步 创建更改集 创建堆栈

等待约 30 分钟，堆栈创建完成后您可以看到堆栈状态显示为 CREATE_COMPLETE，
点击【资源】标签栏可显示所创建的具体服务。

开始使用

登录 Apache Superset 控制台

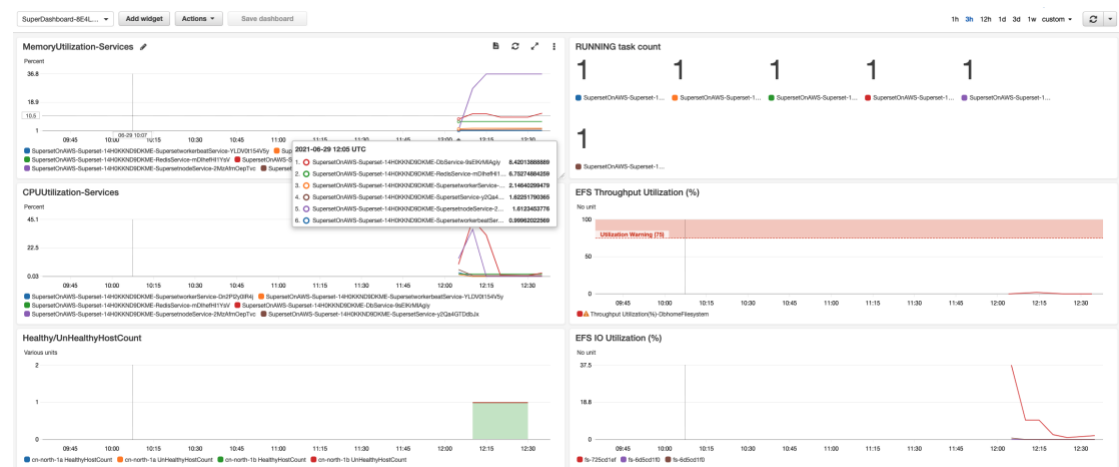
在 Amazon CloudFormation 界面的【输出】选项中找到对应的控制台地址，输入配置的用户名和密码进行登录。



登录进入 Dashboards，可以看到预置的示例仪表板。

登录 CloudWatch 仪表板

跳转到 CloudWatch 控制台，查看创建完毕的仪表板，上半部分显示信息包括 Elastic Container Service (ECS) 集群的 CPU，内存使用情况，ECS 各任务的执行状态，Elastic File System (EFS) 的吞吐量，IO 使用率以及 Application Load Balancer (ALB) 后端节点的健康状态；



下半部分显示信息包括 Elastic Container Service (ECS) 集群的全部执行日志以及过滤所得到的告警日志。

```

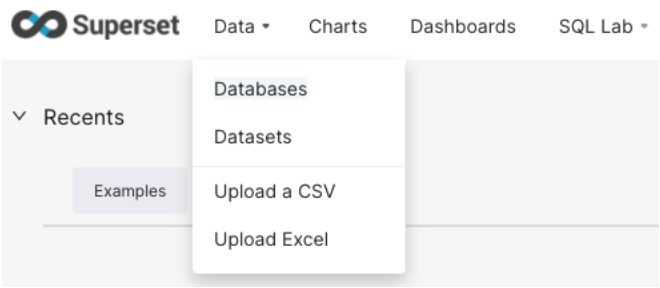
Log group: /docker-compose/supersetOnAWS-15d933e0-d8d2-11eb-89b5-0649e25261ec
# | @timestamp | @message |
+---+
x1 | 2021-06-29T12:35:53.831Z | 192.168.13.222 - - [29/Jun/2021:12:35:53] "GET /superset/superset/stats HTTP/1.1" 200 -
x2 | 2021-06-29T12:35:53.831Z | INFO:werkzeug:192.168.13.222 - - [29/Jun/2021:12:35:53] "[304] GET /superset/superset/stats HTTP/1.1" 304 -
x3 | 2021-06-29T12:35:53.832Z | DEBUG:superset.stats_logger: [304] GET /superset/superset/stats HTTP/1.1" 304 -
x4 | 2021-06-29T12:35:53.832Z | INFO:werkzeug:192.168.13.222 - - [29/Jun/2021:12:35:53] "[304] GET /api/v1/charts HTTP/1.1" 200 -
x5 | 2021-06-29T12:35:53.832Z | 192.168.13.222 - - [29/Jun/2021:12:35:53] "[304] GET /api/v1/charts HTTP/1.1" 200 -
x6 | 2021-06-29T12:35:53.867Z | DEBUG:superset.stats_logger: [304] GET /api/v1/charts HTTP/1.1" 200 -
x7 | 2021-06-29T12:35:53.867Z | DEBUG:superset.stats_logger: [304] GET /api/v1/charts HTTP/1.1" 200 -
x8 | 2021-06-29T12:35:53.772Z | INFO:werkzeug:192.168.13.222 - - [29/Jun/2021:12:35:53] "[304] GET /api/v1/charts HTTP/1.1" 200 -
x9 | 2021-06-29T12:35:53.772Z | 192.168.13.222 - - [29/Jun/2021:12:35:53] "[304] GET /api/v1/charts HTTP/1.1" 200 -
x10 | 2021-06-29T12:35:53.772Z | DEBUG:superset.stats_logger: [304] GET /api/v1/charts HTTP/1.1" 200 -

Log group: Filter with exceptions
# | @timestamp | @message |
+---+
x1 | 2021-06-29T12:12:55.727Z | Error from chokidar C:\app\superset-frontend\node_modules\antd\node_modules\react-design\icons\lib\icons\ExceptionOutline: Error: ENOSPC: System limit for number of file watchers reached, watch C:\app\superset-frontend\node_modules\antd\node_modules\react-design\icons\lib\icons\ExceptionOutline
x2 | 2021-06-29T12:12:55.727Z | Error from chokidar C:\app\superset-frontend\node_modules\antd\node_modules\react-design\icons\lib\icons: Error: ENOSPC: System limit for number of file watchers reached, watch C:\app\superset-frontend\node_modules\antd\node_modules\react-design\icons\lib\icons\ExceptionOutline
x3 | 2021-06-29T12:12:55.653Z | Error from chokidar C:\app\superset-frontend\node_modules\antd\node_modules\react-design\icons\lib\icons: Error: ENOSPC: System limit for number of file watchers reached, watch C:\app\superset-frontend\node_modules\antd\node_modules\react-design\icons\lib\icons\ExceptionOutline
x4 | 2021-06-29T12:12:55.633Z | Error from chokidar C:\app\superset-frontend\node_modules\antd\node_modules\react-design\icons\lib\icons: Error: ENOSPC: System limit for number of file watchers reached, watch C:\app\superset-frontend\node_modules\antd\node_modules\react-design\icons\lib\icons\ExceptionOutline
x5 | 2021-06-29T12:12:55.495Z | Error from chokidar C:\app\superset-frontend\node_modules\antd\lib\util: Error: ENOSPC: System limit for number of file watchers reached, watch C:\app\superset-frontend\node_modules\antd\lib\util\onReachableException.js
x6 | 2021-06-29T12:12:55.495Z | Error from chokidar C:\app\superset-frontend\node_modules\antd\lib\util: Error: ENOSPC: System limit for number of file watchers reached, watch C:\app\superset-frontend\node_modules\antd\lib\util\onReachableException.js
x7 | 2021-06-29T12:12:55.436Z | Error from chokidar C:\app\superset-frontend\node_modules\antd\lib\util: Error: ENOSPC: System limit for number of file watchers reached, watch C:\app\superset-frontend\node_modules\antd\lib\util\onReachableException.js
x8 | 2021-06-29T12:12:55.436Z | Error from chokidar C:\app\superset-frontend\node_modules\antd\lib\util: Error: ENOSPC: System limit for number of file watchers reached, watch C:\app\superset-frontend\node_modules\antd\lib\util\onReachableException.js
x9 | 2021-06-29T12:12:55.365Z | c:\websock-progress\696 building 11576/11571 modules 1 active C:\app\superset-frontend\node_modules\antd\lib\util\onReachableException.js
x10 | 2021-06-29T12:12:55.365Z | c:\websock-progress\696 building 11576/11571 modules 24 active C:\app\superset-frontend\node_modules\antd\lib\util\onReachableException.js

```

对接现有数据

点击【Data】，下拉框中选择 Databases 开始连接已有数据源。Apache Superset 目前支持的数据源包括 SQL, PostgreSQL, Redshift, Athena, ClickHouse 等多种主流数据库，数据仓库，支持列表可参考[这里](#)。该方案已经预置 SQL, PostgreSQL, SQL Server, Oracle, ElasticSearch, Redshift, Athena, Clickhouse 数据源驱动。



输入对应数据源的 URL，如 RedShift 格式为 redshift+psycopg2://<userName>:<DBPassword>@<RedShift End Point>:5439/<Database Name>，更多其他数据源的 URL 格式可以参考[这里](#)

Edit database

CONNECTION*

PERFORMANCE

SQL LAB SETTINGS

SECURITY

EXTRA

DATABASE NAME*

redshift

SQLALCHEMY URI*

redshift+psycopg2://demo:XXXXXXXXXX@utility-cluster.cqyc5ijfg8d.cn-northwe:

TEST CONNECTION

Refer to the [SQLAlchemy docs](#) for more information on how to structure your URI.

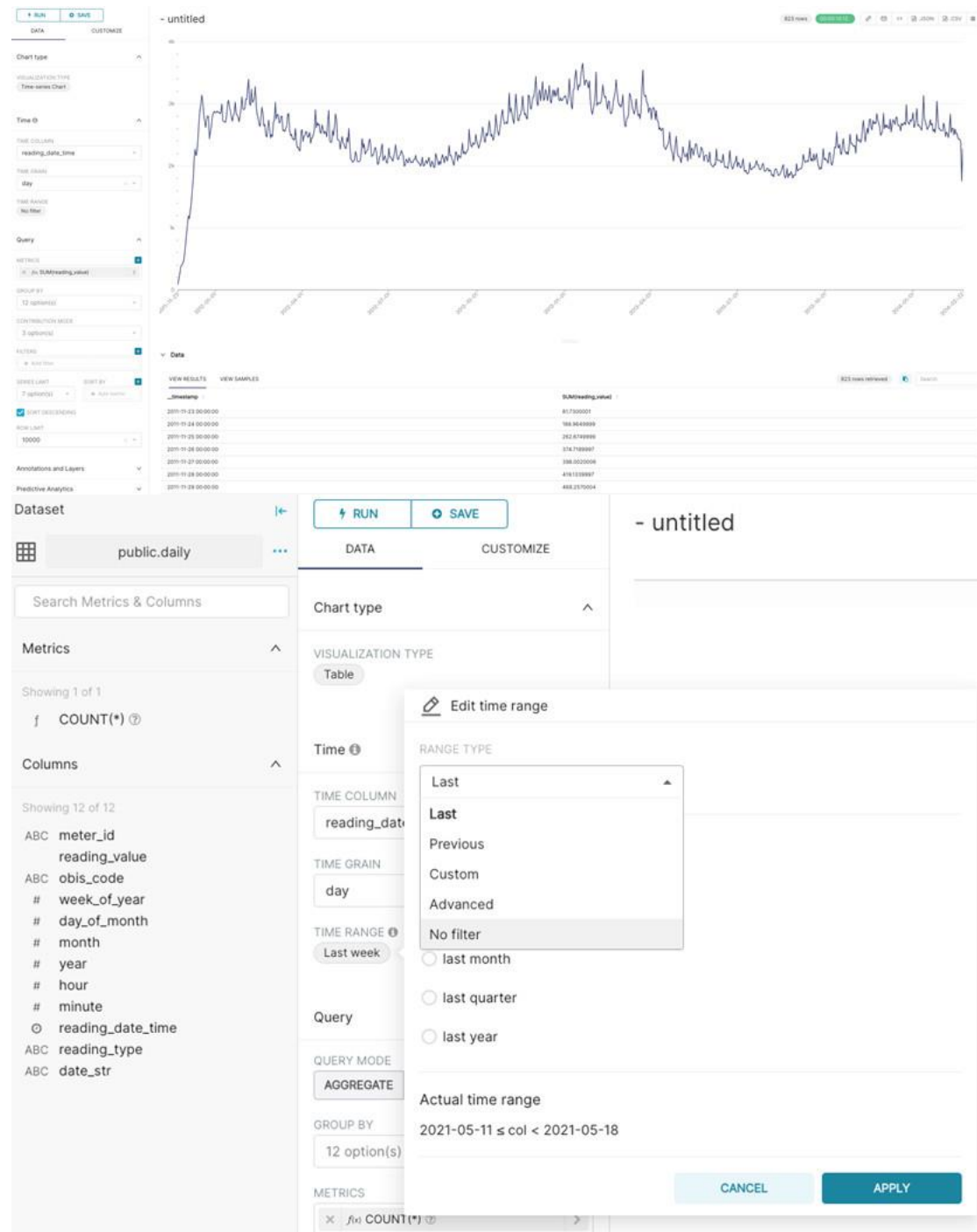
CANCEL

SAVE

连接成功后，点击 Datasets 查看对应的数据集是否正常显示，这里的示例数据是 "daily"，来源于[智能电表数据分析平台](#)。

Data	Databases	Datasets	Saved queries	Query history
Owner: All	Database: All	Schema: All	Type: All	Q Search
Name	Type	Source	Schema	
daily	Physical	redshift	public	

点击进去进行编辑，VISUALIZATION TYPE 中选择 Time-series Chart，TIME RANGE 选择 No filter，Query 中 COLUMN 为要显示的数值，AGGREGATE 选择 SUM 操作，点击 RUN 可以看到数据按照时间序列绘制出来，同时在 Data 一栏展示了数据样本，注意上述配置需要根据实际数据进行修改。需要注意横向时间轴的最后时间为 2014-02-22。



如果在创建 Superset 的选项我们选择了 Install Prophet library 为 yes，我们可以进一步体验 Superset 内置的时间序列预测功能。点击 Predictive Analytics，勾选 ENABLE FORECAST 选项，其他选项默认，重新执行 RUN。

Predictive Analytics ^

☒ ENABLE FORECAST

FORECAST PERIODS

10

CONFIDENCE INTERVAL ⓘ

0.8

YEARLY SEASONALITY

3 option(s) ▾

WEEKLY SEASONALITY

3 option(s) ▾

DAILY SEASONALITY

3 option(s) ▾

可以看到 Superset 在之前的数据基础上对未来 10 天的数据走势进行了预测（此时横向时间轴的最后时间为 2014-03-04）并绘制出了数据的置信区间，其中紫色圆点为原始的时间序列离散点，紫色实线为使用时间序列拟合所得到的取值，实线周围的浅色区域则为数据的置信区间，即合理的上界和下界。在该操作中我们需要输入的是包含时间戳和值的原始数据，需要预测的时间序列长度，得到的输出为未来时间序列趋势以及对应的置信区间。



FAQ

1. 该解决方案代码是否开源，我们可以进行自定义修改以适配，集成我们自己的软件系统吗？

方案基于 Apache 2.0 协议开源，源代码地址：<https://github.com/aws-quickstart/quickstart-aws-superset>，整体方案基于无服务器架构思想设计，用户无需运维底层基础资源，同时无缝对接 Amazon CloudTrail，CloudWatch，Compliance 等服务，实现方案的合规可审计。

2. 如何了解该方案所包含使用的亚马逊服务状态，底层资源使用，应用日志信息？
方案成功启动后，您可以登录到 EC2 控制台中的负载均衡器一栏，查看活动流数量等基本 CloudWatch 信息。



您可以进一步登陆 Elastic Container Service (ECS) 控制台查看对应的容器服务是否为 Active 状态。

任务名称	状态	服务类型	任务定义	预期任务数	正在运行的任务数	启动类型	平台版本
SupersetOnNewVPC-Superset-35TWC4YHSNID-...	ACTIVE	REPLICA	supersetOnAWS-medi34	1	1	FARGATE	1.4.0
SupersetOnNewVPC-Superset-35TWC4YHSNID-...	ACTIVE	REPLICA	supersetOnAWS-superset31	1	1	FARGATE	1.4.0
SupersetOnNewVPC-Superset-35TWC4YHSNID-...	ACTIVE	REPLICA	supersetOnAWS-superset-node30	1	1	FARGATE	1.4.0
supersetIntService	ACTIVE	REPLICA	supersetOnAWS-superset-int31	0	0	FARGATE	1.4.0
SupersetOnNewVPC-Superset-35TWC4YHSNID-...	ACTIVE	REPLICA	supersetOnAWS-dbi31	1	1	FARGATE	1.4.0
SupersetOnNewVPC-Superset-35TWC4YHSNID-...	ACTIVE	REPLICA	supersetOnAWS-superset-worker30	1	1	FARGATE	1.4.0
SupersetOnNewVPC-Superset-35TWC4YHSNID-...	ACTIVE	REPLICA	supersetOnAWS-superset-worker-be...	1	1	FARGATE	1.4.0

针对底层资源使用和应用日志，您可以登录到 CloudWatch 控制台，查看该方案所创建的控制面板 SupersetDashboard 来集中监控 Apache Superset 应用相关的 CPU，内存，网络，亚马逊服务指标，容器日志等信息。

版本发布说明

日期	版本	发布说明
2021 年 6 月	1.0.0	首次发布。 支持 Apache Superset 核心模块的高可用，业务数据的持久化，计算资源的弹性伸缩，预置数据库驱动（RedShift, Athena, ClickHouse 等）以及时间序列算法。
2022 年 2 月	1.1.0	新增特性包括可配置的 UI 访问端口，用户密码强度检测，BigQuery 数据连接支持等。
2022 年 9 月	1.2.0	支持水平条形图，改善数据加载时间。

附录

CloudFormation 堆栈参数

参数名称	参数默认值	参数描述
基础配置		
Initial Superset User Name	None	Superset 控制台用户名
Initial Superset User Password	None	Superset 控制台用户密码
Pre-populate example dashboard	Yes	是否要获取官方内置示例数据集以及仪表板
Install Prophet library	No	是否要安装 Prophet 软件包以实现数据的在线预测功能

ClusterName	supersetOnAWS	配置业务运行的 Elastic Container Service 集群名称
网络配置		
Amazon Fargate VPC CIDR Block	192.168.10.0/24	Amazon ECS 使用的 VPC CIDR (无类别域间路由)
Public Subnet for the first Availability Zone	192.168.11.0/24	第一个可用区的公开子网的 CIDR (无类别域间路由)
Public Subnet for the second Availability Zone	192.168.11.0/24	第二个可用区的公开子网的 CIDR (无类别域间路由)
Private Subnet for the first Availability Zone	192.168.20.0/24	第一个可用区的子网的 CIDR (无类别域间路由)
Private Subnet for the second Availability Zone	192.168.21.0/24	第二个可用区的公开子网的 CIDR (无类别域间路由)
其他配置		
QSS3BucketName	aws-cn-quickstart	该方案所在 S3 桶的名称
QSS3BucketRegion	cn-north-1	该方案所在 S3 桶的区域
QSS3KeyPrefix	quickstart-superset-on-aws/	该方案所在 S3 桶的前缀

源代码地址

您可以访问我们的 [GitHub 地址](#) 来下载源代码进行部署，同时欢迎您提交 PR 和建议。欢迎您登陆[官方网站](#)了解更多方案