使用Glue进行大数据建设

> ✊ 经验没有压缩算法。 —— 安迪 · 贾西

一、什么是Glue

> AWS Glue 是用于提取、转换和加载 (ETL) 操作的无服务器数据准备服务。它使数据工程师、数据分析师、数据科学家以及 ETL 开发人员能够轻松地提取、清理、丰富、规范化和加载数据。AWS Glue 将开始分析数据所需的时间从数月缩短到几分钟。它为您提供了直观和基于代码的界面，使数据准备过程变得简单轻松。数据工程师和 ETL 开发人员只需单击几下鼠标，就可以使用 AWS Glue Studio 创建、运行和监控 ETL 作业。数据分析师和数据科学家可以使用 AWS Glue DataBrew 直观地清理和规范化数据，而无需编写代码。

二、WorkShop场景介绍

> 一般公司都会使用mysql 作为业务数据库，所以我们以从mysql8 为例，看看如何使用glue 抽取数据,并进行简单的ETL进行业务表数据建模，将建好的中间表分别存在S3和数仓产品redshift
>
> 接下来我们会分别使用s3和redshift作为数据源，对接分析工具和BI工具进行大数据的分析和报表的制作

三、准备工作

1. mysql

- 保存访问的终端节点，备用。为了减少对线上业务的影响，建议使用从库的终端节点
- 在ec2 的页面创建 容许对3306端口流量进出的安全组，配置给MySQL服务
- 在vpc页面创建对mysql可以访问的终端节点

> 说明，由于glue 是aws的公有服务，默认是没有权限访问您的vpc。要访问运行在您vpc的服务，必须创建一个终端节点，建立对您vpc的私有链接供glue使用，才能访问您的rds 数据库服务。

- 在vpc页面创建对mysql可以访问的终端节点
- 创建样例数据，我们将以同步如下数据并做简单etl构建数仓模型

SQL

```sql
use demo;

CREATE TABLE IF NOT EXISTS `order` (
    order_id INT AUTO_INCREMENT NOT NULL,
    user_mail varchar(20) NOT NULL,
    status char(10) NOT NULL,
    good_count INT NOT NULL,
    city varchar(20) NOT NULL,
    amount FLOAT NOT NULL,
    create_time datetime NOT NULL,
    update_time datetime NOT NULL,
    PRIMARY KEY (`order_id`)
);
```

2. AWS Secrets Manager

- 用来保存您的数据库账号和密码，防止密码明文硬编码在代码里
- 请使用它创建你mysql 和数仓产品redshift的账号密码托管
- 在vpc页面创建对aws secrets manager可以访问的终端节点
- 在ec2 的页面创建 容许对443端口流量进出的安全组

3. Redshift

- 在集群页面找到jdbc链接字符串备用
- 配置redshift的安全组，容许 5439（redshift的默认端口） 的流量访问

4. S3

- 在vpc页面创建对S3可以访问的终端节点

四、动手实验

1. Clone 代码

Plain Text

```
1  git clone git@github.com:tingxin/glue_workshop.gi
```

2. 使用 AWS Secrets Manager 创建密码托管

a. 根据导航一步一步就能创建成功

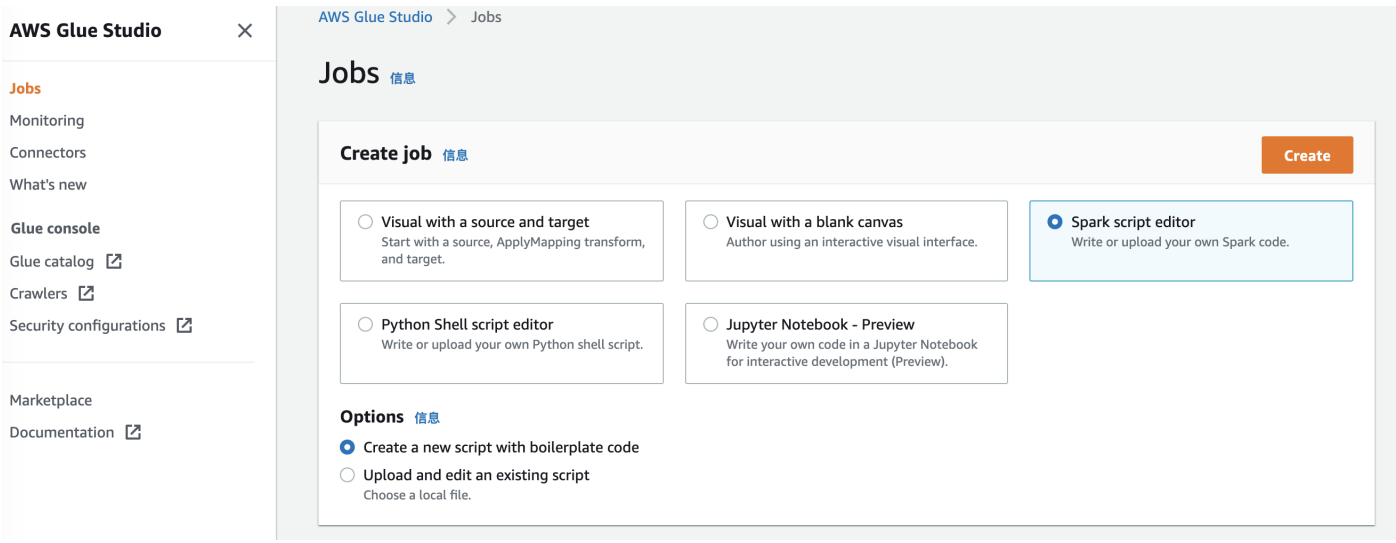
b. 回到到代码库，修改secret.py 的代码TODO 部分

c. 把secret.py上传到s3,并保存s3地址备用

3. 创建dwd 任务

a. 打开 glue studio

b. 创建sparky 脚本

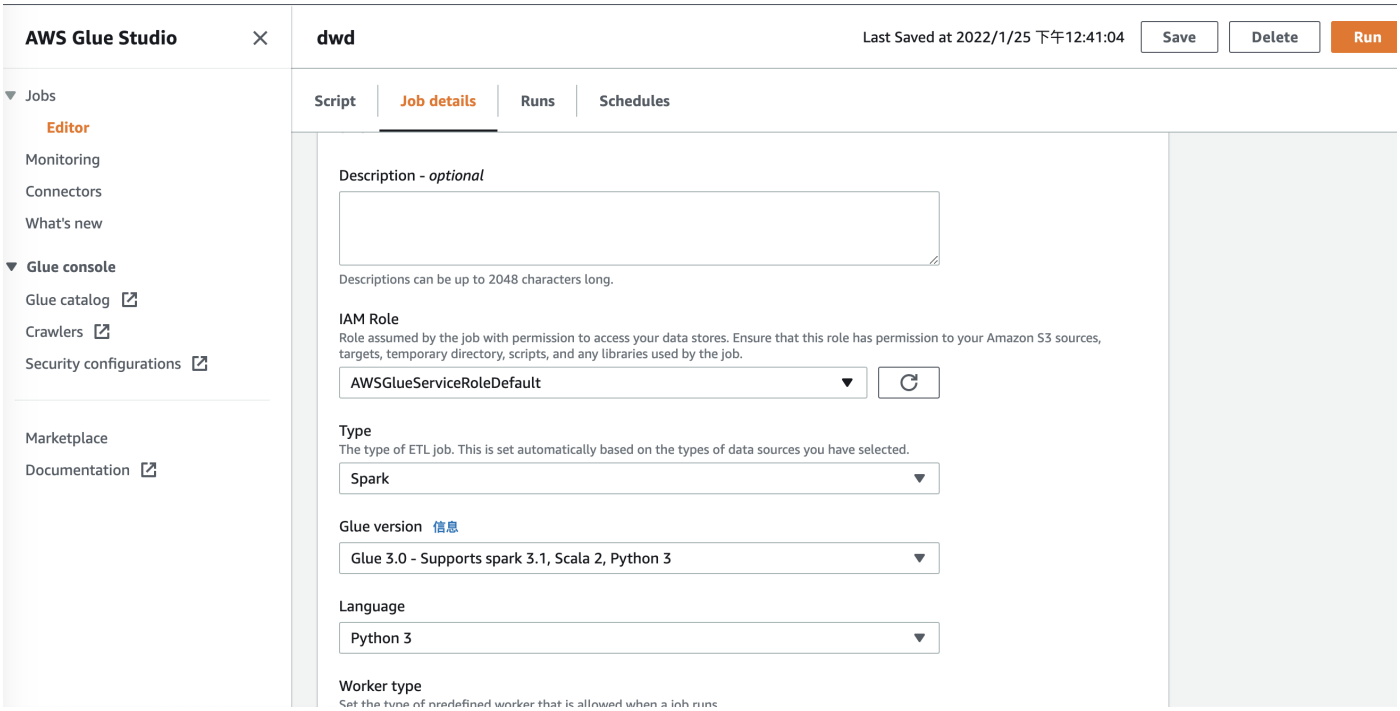


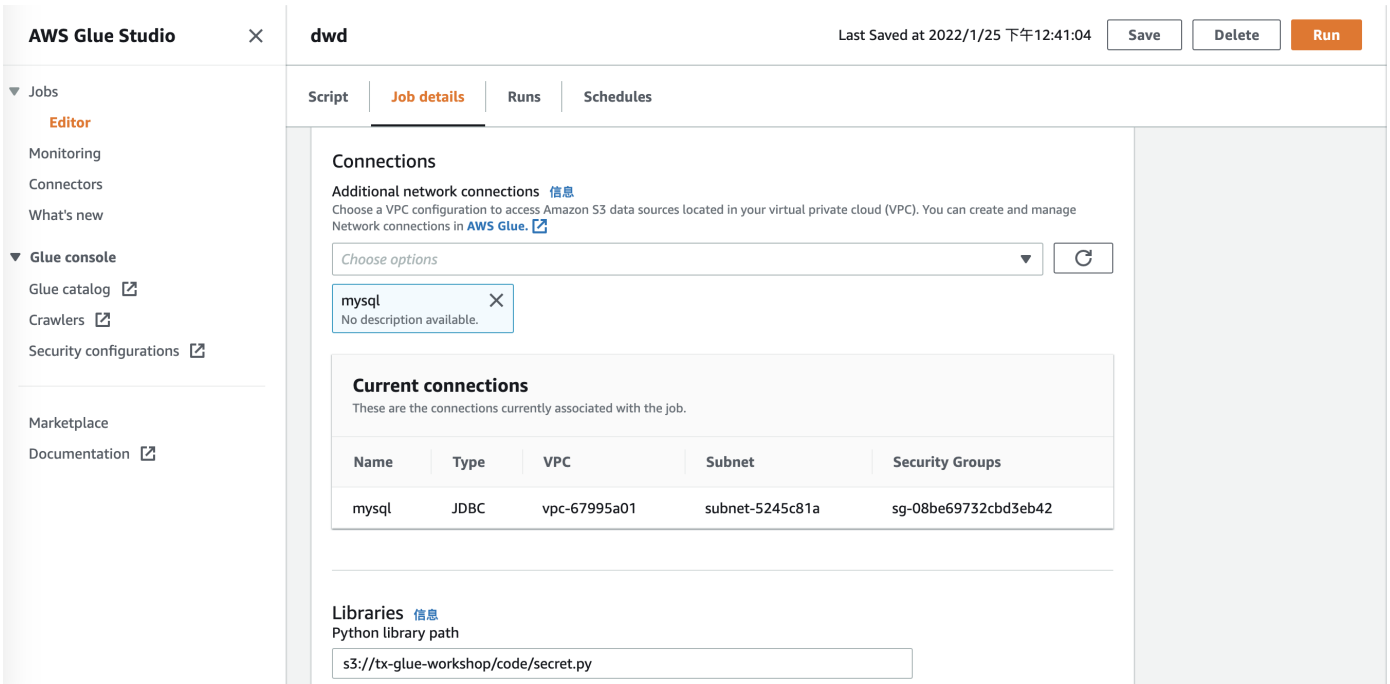
c. 将dwd.py脚本拷贝到代码编辑框

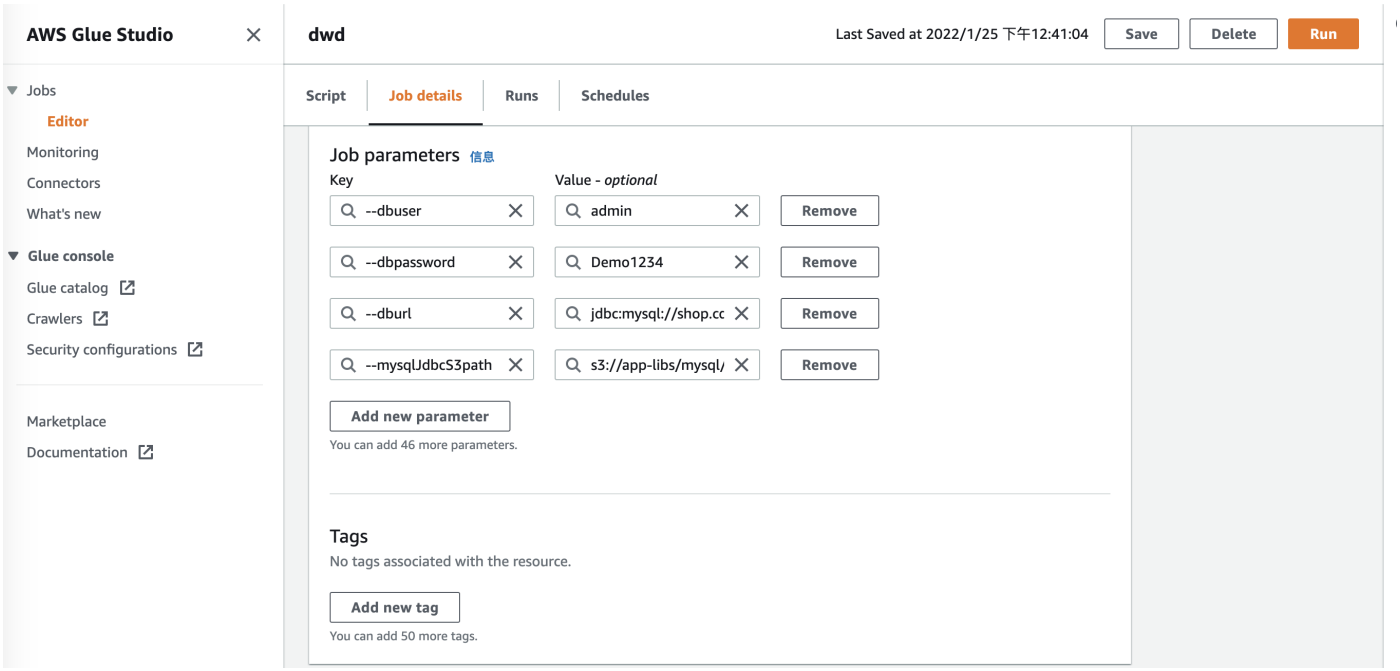
d. 将代码中 # TODO 部分的地方修改成您的代码

e. 新开一个浏览器，我们需要配置mysql8的链接信息，glue根据这个信息查找mysql的子网，安全组信息

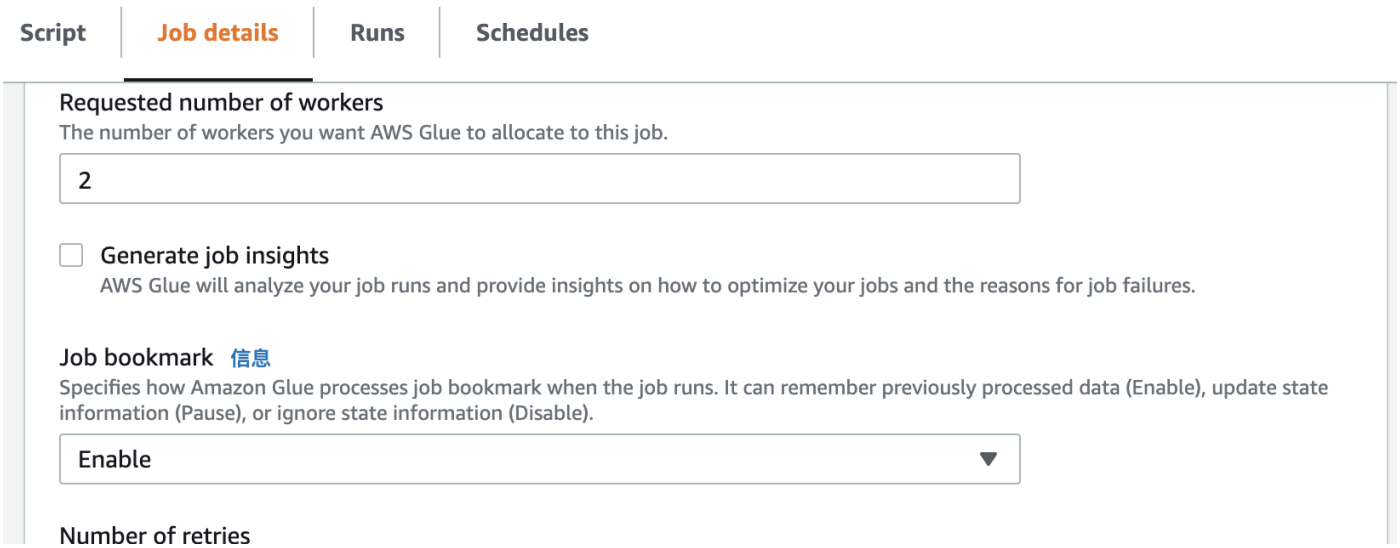
f. 浏览器切换回到glue 代码编辑框，切换到job detail 界面,按如图配置







g. 如果要使用书签功能进行增量同步，请仔细阅读dwd.py文件中的注释部分，并在job detal 中如下图设置 Job bookmark 为enable



有几个地方需要你注意：

role	需要在IAM role 配置对rds,s3, secret manager，redshift的访问权限
libraryes	需要您secret.py的s3地址填入，如果有多个依赖库，可以用逗号隔开

h. 点击保存，保存代码，点击允许运行任务

i. 详细参考 https://docs.aws.amazon.com/zh_cn/glue/latest/dg/add-job.html

4. 创建dws 任务

同dwd任务，不再赘述

5. 使用 workflow 进行任务的编排调度

> 在 AWS Glue 中，可以使用工作流程创建和可视化涉及多个爬网程序、作业和触发器的复杂的提取、转换和加载 (ETL) 活动。每个工作流都管理其所有任务和爬网程序的执行和监控。当工作流运行每个组件时，它会记录执行进度和状态。这将为您提供大型任务的概览和每个步骤的详细信息。AWS Glue 控制台以图表形式呈现工作流。
>
> https://docs.aws.amazon.com/zh_cn/glue/latest/dg/workflows_overview.html

6. 通过glue 自动进行源数据发现

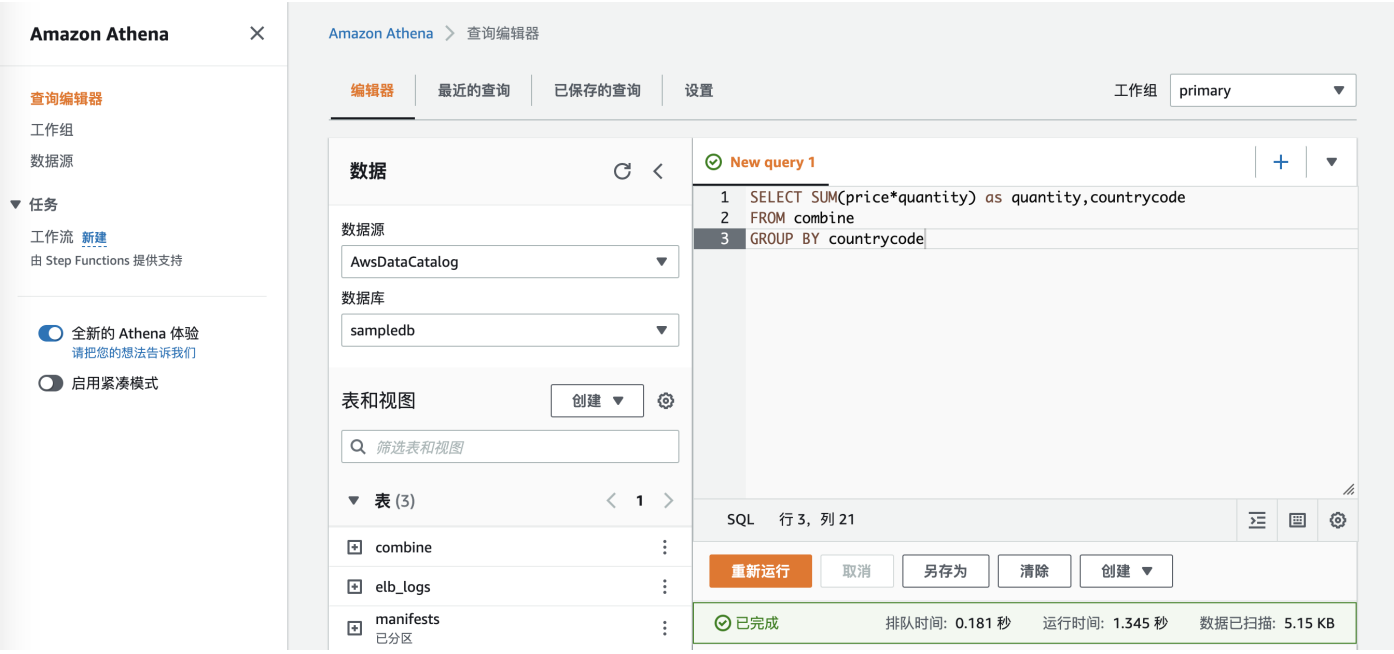
详细参考 https://docs.aws.amazon.com/zh_cn/glue/latest/dg/add-crawler.html

1. 根据向导一步一步创建爬网程序
2. 运行爬网程序，大概要一到两分钟出结果
3. 点击上图导航栏的数据库，表，查看是否已经生成了表的 schema

7. 通过athena 自动进行源数据发现

> Amazon Athena 是一种交互式查询服务，让您能够轻松使用标准 SQL 直接分析 Amazon Simple Storage Service (Amazon S3) 中的数据。只需在 AWS Management Console 中执行几项操作，即可将 Athena 指向 Amazon S3 中存储的数据，并开始使用标准 SQL 运行临时查询，然后在几秒钟内获得结果。
>
> Athena 没有服务器，因此您无需设置或管理任何基础设施，且只需为您运行的查询付费。Athena 可自动扩展（并行执行查询），因此，即使在数据集很大、查询很复杂的情况下也能很快获得结果。

1. 打开athena 在数据面板，可以查看通过爬网程序生成的数据数据表
2. 可以傻瓜式的进行查询，无需文档



8. 通过quicksight 进行看板制作

> Amazon QuickSight 是一项快速的业务分析服务，用于构建可视化、执行临时分析以及从数据中快速获取业务洞察。Amazon QuickSight 无缝地发现 AWS 数据源，使企业能够扩展到数十万用户，并通过使用 Amazon QuickSight Super-fast, Parallel, In-Memory, Calculation Engine (SPICE) 提供快速响应的查询性能。

这里有比较详尽的教程：

https://docs.aws.amazon.com/zh_cn/quicksight/latest/user/quickstart-createanalysis.html

五、注意点

1. 输出到 Redshift 的表无需手动创建，代码可以默认帮你创建

六、常见问题

1. 连接redshift timeout

A. 检查安全组

2. 用glue增量更新,书签不起作用

在某些情况中，您可能已启用了 AWS Glue 作业书签，但 AWS Glue 作业却重新处理了之前运行中已处理过的数据。发生这种情况可能是由于以下原因：

缺少作业提交 – AWS Glue ETL 脚本末尾的 job.commit() 语句用于更新作业书签的状态。如果未包括该语句，则作业将重新处理之前处理过的文件和新文件。请确保在您的用户脚本中所有通向作业完成的代码路径中，都会执行该作业提交语句。

缺少转换上下文 – 转换上下文在 GlueContext 中是可选参数，但作业标签需要该参数才能正常工作，请确认在创建 DynamicFrame 时包括了转换上下文参数。请参阅以下代码示例：

Python

```
sample_dynF=glueContext.create_dynamic_frame_from_catalog(database,
table_name,
transformation_ctx=”sample_dynF”)
```

JDBC 源 – 在使用 JDBC 连接访问关系数据库时，作业书签要求源表要么包含主键列，要么包含单调递增或递减值的列（需要在源选项中指定）。作业书签可以仅捕获新增行。此行为不适用于 S3 上存储的源表。

最后修改时间 – 为确定处理存储在 S3 上的哪些文件，作业书签将检查对象的最后修改时间，而不是文件名。如果自作业最后一次运行以来更改过输入对象，则当作业再次运行时，将重新处理这些对象。