ALIBABA CLOUD

阿里云

# Databricks 数据洞察

## 管理集群

文档版本：20201026

阿里云

# 法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

# 通用约定

| 格式 | 说明 | 样例 |
| --- | --- | --- |
| 危险 | 该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。 | 危险<br>重置操作将丢失用户配置数据。 |
| 警告 | 该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。 | 警告<br>重启操作将导致业务中断，恢复业务时间约十分钟。 |
| 注意 | 用于警示信息、补充说明等，是用户必须了解的内容。 | 注意<br>权重设置为0，该服务器不会再接受新请求。 |
| 说明 | 用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。 | 说明<br>您也可以通过按Ctrl+A选中全部文件。 |
| > | 多级菜单递进。 | 单击**设置**> **网络**> **设置网络类型**。 |
| **粗体** | 表示按键、菜单、页面名称等UI元素。 | 在**结果确认**页面，单击**确定**。 |
| `Courier字体` | 命令或代码。 | 执行 `cd /d C:/window` 命令，进入Windows系统文件夹。 |
| *斜体* | 表示参数、变量。 | `bae log list --instanceid`<br>*Instance_ID* |
| [] 或者 [a\|b] | 表示可选项，至多选择一个。 | `ipconfig [-all\|-t]` |
| {} 或者 {a\|b} | 表示必选项，至多选择一个。 | `switch {active\|stand}` |

# 目录

# 1.创建集群

本节介绍如何使用Databricks数据洞察控制台创建集群。

## 前提条件

已注册阿里云账号，并完成实名认证。详情请参见阿里云账号注册流程。

## 操作步骤

1. 使用阿里云账号登录Databricks数据洞察控制台。
2. 在Databricks数据洞察控制台页面，选择所在的地域（Region）。创建的集群将会在对应的地域内，一旦创建后不能修改。
3. 在左侧导航栏中，单击**集群**。
4. 在**集群管理**页面，单击**创建集群**。
5. 设置基础信息。

| 参数 | 描述 |
| --- | --- |
| 集群名称 | 集群的名字。长度限制为1~64个字符，仅可使用中文、字母、数字、连接号（-）和下划线（_）。 |
| Knox账号 | 为了更好的安全性，Web UI访问（如Zeppelin Notebook、Spark UI、Ganglia UI）需要Knox账号和密码，来保障您的账号安全。若无RAM子账号，请前往RAM进行创建https://ram.console.aliyun.com/users/new |
| Knox密码 | 两次确认Knox密码，登录Web UI时候使用，请您牢记。 |
| Databricks Runtime版本 | Databricks Runtime的版本信息，版本号与Databricks官方保持一致，包含Scala和Spark的版本。版本详情请参见Databricks Runtime版本说明。 |
| Python版本 | 默认版本为Python 3。 |
| 付费类型 | 目前支持的付费类型为按量付费。<br>即根据实际使用的小时数来支付费用，每小时计费一次。适合短期的测试任务或是灵活的动态任务。 |
| 可用区 | 可用区为在同一地域下的不同物理区域，可用区之间内网互通。<br>一般选择默认的可用区即可，亦可选择与已购阿里云产品部署在同一个可用区。 |

| 参数 | 描述 |
| --- | --- |
| ECS实例 | 由Master和Worker两种类型的节点组成：<br>○ Master节点：主要负责集群资源管理和作业调度。默认节点个数为1。<br>○ Worker节点：集群的计算节点，主要负责作业的执行。最小节点数量为3。 |

6. 设置高级信息。高级信息包括如下两方面：
   - Spark设置

| 参数 | 描述 |
| --- | --- |
| Spark配置 | 输入Spark的配置信息。配置的属性值将会更改到*spark-defaults.conf*文件中。支持的配置列表为*spark.apache.org/docs/latest/configuration.html#spark-properties* |
| 环境变量 | 您可以自定义Spark执行的环境变量。配置的属性将会更新到*spark-env.sh*中。 |

   - 服务目录

| 参数 | 描述 |
| --- | --- |
| 类型 | 包括以下两种类型：<br>■ 默认值<br>■ 自定义 |
| OSS路径 | 该目录用来存放集群服务组件的临时文件等。<br>该目录会作为产品的根目录来使用。当用户有多个集群时，不需要为每个集群单独指定服务目录。不同Region需要有不同的服务目录，产品会为每个集群在服务目录下创建子目录，即 *oss://${specified-bucket-or-dir}/ddi-${clusterid}/*。 |

7. 阅读并勾选服务条款。
8. 单击创建。集群创建需要时间，当状态更新为空闲时表示创建成功，请您耐心等待。

## 问题反馈

您在使用阿里云Databricks数据洞察过程中有任何疑问，欢迎用钉钉扫描下面的二维码加入钉钉群进行反馈。

Databricks数据洞察产品群

# 2.查看集群列表信息

本文介绍如何查看已创建集群的详情。

## 前提条件

已创建集群，详情请参见创建集群。

## 操作步骤

1. 使用阿里云账号登录Databricks数据洞察控制台。
2. 在Databricks数据洞察控制台页面，选择所在的地域（Region）。 创建的集群将会在对应的地域内，一旦创建后不能修改。
3. 在左侧导航栏中，单击集群。 集群管理页面展示您所拥有的所有集群的基本信息，以及各集群支持的操作。

| 参数 | 说明 |
| --- | --- |
| 集群ID/名称 | 集群ID是产品自动分配的集群唯一标识；名称是用户在集群创建时自定义的集群名称。 |
| 集群类型 | 集群的付费类型。 |
| 状态 | 集群的状态：<br>○ 初始化中：集群正在构建，包括两个阶段：一是物理ECS机器的创建；二是集群服务的启动，稍等片刻即可达到运行中的状态。<br>○ 空闲：集群目前没有作业运行。<br>○ 运行中：集群处于正常运行状态。<br>○ 构建失败：创建过程中遇到异常，已经创建的ECS机器会自动回滚，在集群列表页面单击状态右边的问号，可以查看异常明细。<br>○ 终止中：目前集群处于终止状态。<br>○ 终止失败：终止集群时失败。<br>○ 已终止：集群已终止。 已终止的群集无法运行笔记本或作业。<br>○ 异常：表示集群异常。 |
| 创建时间/运行时间 | 集群创建的时间以及运行的时长。 |
| 付费类型 | 集群的付费类型。 |

| 参数 | 说明 |
| --- | --- |
| 操作 | 支持的集群操作：<br>○ 详情：进入集群的详情页，查看集群创建后的详细信息。<br>展示已创建集群的详细信息，包括集群信息、网络信息、软件信息和主机信息四部分。<br>。<br>○ Spark UI：Apache Spark history server提供的Web UI。您可以在此界面查看Spark作业的运行信息。<br>○ Ganglia监控：用来监控集群内节点的运行状况。<br>○ Notebook：进入集群对应的DataInsight Notebook页面，Notebook相关操作请参见管理Notebook。<br>○ 释放：释放当前集群，详情请参见释放集群。 |

# 3.集群Web UI

Databricks数据洞察集群提供了多个Web UI的访问入口，包括Notebook、Spark UI、Yarn UI和Ganglia监控。

## 使用概述

用户在集群详情页面单击Web UI的链接，会跳转到Knox账号的验证页面。输入Knox账号和密码即可登录到相应的Web UI页面。



## Web UI登录

首次登录Web UI时，用户可能会在浏览器看到如下告警。用户可以根据浏览器告警的提示，进行操作。通常情况下，用户会看到下面两种告警。

### 告警提示一

单击“高级”按钮，展开隐藏详情后，会出现急需前往链接。单击该链接即可访问Web UI。



### 告警提示二

出现“您的连接不是私密连接”的告警提示，且点开高级按钮，没有继续访问链接时，请在当前页面直接键盘盲输入11个字符：thisisunsafe

# 4.释放集群

当集群不再使用时，您可以随时进行释放，以节约成本。

## 背景信息

待释放集群的状态必须是创建中、运行中或空闲中，其他状态不支持释放。

## 操作步骤

1. 使用阿里云账号登录Databricks数据洞察控制台。
2. 在Databricks数据洞察控制台页面，选择所在的地域（Region）。 创建的集群将会在对应的地域内，一旦创建后不能修改。
3. 在左侧导航栏中，单击**集群**。
4. 设置释放。
   - 在**集群管理**页面，单击待释放集群所在行的**释放**。
   - 单击待释放集群的集群ID，在**集群基础信息**页面，单击**集群操作 > 释放**。
5. 在弹出的**集群管理-释放**对话框中，单击**释放**。

# 5.访问外部数据源

本文介绍如何在Databricks数据洞察实现访问外部数据源的需求。

## 背景信息

Databricks数据洞察为了满足您在计算任务里访问您在阿里云上已有的数据，支持通过添加外部数据源的方式，打通您现有其他类型集群的网络。目前支持的数据源类型有三种：Aliyun EMR HDFS，Aliyun EMR Kafka 和 Aliyun ECS。

## 绑定数据源

绑定数据源的本质是打通集群间的网络，即将数据源集群所在VPC与目标Databricks数据洞察集群所在VPC的网络打通。数据源绑定之后，您可以在Notebook或Spark作业里直接访问相应的集群数据。

> 说明 对于数据源绑定场景，如果多个数据源共用一个VPC下的交换机，打通其中一个数据源意味着相同交换机下的所有数据源一并打通。因此，只能打通同一Region下的数据源。

1. 在Databricks数据洞察控制台，进入集群详情页面。
2. 点击详情页面**数据源**标签，在**添加数据源**弹窗选择要添加的数据源类型。
3. 在所选类型的数据源列表里勾选希望绑定的EMR集群或ECS实例（支持复选）。
4. 建议补充数据源**描述**信息，便于辨识已绑定数据源实例。
5. 点击下一步，**确认安全组和交换机信息**。

> 说明 对于Aliyun EMR HDFS和Aliyun EMR Kafka类型数据源，目前支持各自添加一个集群 。Aliyun ECS类型可以多选，如果是自建集群（如Kafka或HDFS），只需要选择集群中的一个实例即可。

## 数据源访问说明

对于Aliyun EMR HDFS集群，数据源打通之后您可以通过以下方式访问集群数据。

对于HA集群，默认使用emr-cluster作为hostname。

```
sc.textFile("hdfs://emr-cluster/tmp/user0/airline_statistic_usa.csv").count()
```

对于非HA集群，请直接使用EMR HDFS集群namenode的ip访问。

```
sc.textFile("hdfs://192.168.xxx.xxx:9000/tmp/user0/airline_statistic_usa.csv").count()
```

> 说明 对于Aliyun EMR Kafka集群，支持通过ip或者hostename访问。

## 解绑数据源

解绑数据源本质是将数据源所在VPC与目标Databricks数据洞察集群VPC网络隔离。如果多个数据源共用一个交换机，解绑操作会使得当前Databricks数据洞察集群无法继续访问该交换机下所有数据源集群。

1. 在Databricks数据洞察控制台，进入集群详情页面。
2. 点击详情页面**数据源**标签。

3. 在已绑定数据源列表里选择要解绑的交换机，点击解绑即可。

# 6.RAM访问控制

## 6.1. OSS访问服务

首次使用Databricks数据洞察服务创建集群时，需要使用主账号为Databricks数据洞察服务授权名为AliyunDDIAccessingOSSRole的系统默认角色。同时需要您创建一个系统目录存储Bucket。

### 背景信息

关于角色详细信息，具体可以参见RAM角色概览。

- 通过授予AliyunDDIAccessingOSSRole角色，您创建的Databricks数据洞察集群可以以免AK的方式访问阿里云OSS资源，详细信息请参见基于MetaService免AccessKey访问阿里云资源。 **注意** 首次使用Databricks数据洞察服务时，必须用主账号完成默认角色授权和Bucket创建，否则子账号和主账号不能使用Databricks数据洞察。

### 角色授权流程

1. 首次使用Databricks数据洞察服务创建集群时，会弹出如下提示：

2. 单击**点击前往RAM进行授权**。单击**同意授权**，将默认角色AliyunDDIAccessingOSSRole授予给Databricks数据洞察服务。

3. 完成以上授权后，您需要刷新Databricks数据洞察控制台，然后即可进行相关操作。如果您想查看AliyunDDIAccessingOSSRole相关的详细策略信息，您可登录RAM的控制台查看。

### AliyunDDIAccessingOSSRole权限内容

默认角色AliyunDDIAccessingOSSRole包含系统权限策略为AliyunDDIAccessingOSSRolePolicy，OSS相关权限内容如下。

```
"Action": [
  "oss:GetObject",
  "oss:ListObjects",
  "oss:PutObject",
  "oss:DeleteObject",
  "oss:ListBuckets",
  "oss:AbortMultipartUpload",
  "oss:ListMultipartUploads"
]
```

### 系统目录Bucket创建

1. 使用主账号首次创建集群，并完成必填信息填写。
2. 单击创建按钮，弹出创建OSS Bucket对话框。

3. 单击Bucket名称复制图标。
4. 单击OSS控制台，跳转到OSS控制台。
5. 单击创建bucket。

6. 粘贴Bucket名称。
7. 选择区域。
8. 单击确定。
9. 返回集群创建页面，单击已完成Bucket。

> 说明
>
> 此Bucket为系统目录Bucket，不建议存放数据，您可以再创建一个Bucket来读写数据。

# 6.2. 服务关联角色

本文介绍Databricks数据洞察服务关联角色AliyunServiceRoleForDDI以及如何删除该角色。

## 背景信息

Databricks数据洞察服务关联角色AliyunServiceRoleForDDI是Databricks数据洞察在某些情况下，为了完成自身的某个功能，需要获取其他云服务的访问权限而提供的RAM角色。更多关于服务关联角色的信息请参见服务关联角色。

## AliyunServiceRoleForDDI应用场景

Databricks数据洞察集群创建及数据源绑定功能需要访问云服务器ECS、专有网络VPC等云服务的资源时，需要通过服务关联角色AliyunServiceRoleForDDI获取访问权限。

## AliyunServiceRoleForDDI权限说明

AliyunServiceRoleForDDI具备以下云服务的访问权限：

```
"Action": [
    "vpc:DescribeVSwitches",
    "ecs:CreateNetworkInterface",
    "ecs:DeleteNetworkInterface",
    "ecs:DescribeNetworkInterfaces",
    "ecs:CreateNetworkInterfacePermission",
    "ecs:DeleteNetworkInterfacePermission",
    "ecs:CreateSecurityGroup",
    "ecs:AuthorizeSecurityGroup",
    "ecs:RevokeSecurityGroup",
    "ecs:AuthorizeSecurityGroupEgress"
  ]
```

## 删除AliyunServiceRoleForDDI

如果您需要删除AliyunServiceRoleForDDI服务关联角色，需要先释放依赖这个服务关联角色的Databricks数据洞察集群。

具体操作步骤如下：

1. 登录RAM控制台，在左侧导航栏中单击**RAM角色管理**。
2. 在**RAM角色管理**页面的搜索框中，输入**AliyunServiceRoleForDDI**，自动搜索到名称为AliyunServiceRoleForDDI的RAM角色。
3. 在右侧**操作**列，单击**删除**。
4. 在**删除RAM角色**对话框，单击**确定**。
   i. 如果当前账号下存在关联的Databricks数据洞察集群，则需先释放集群后才能删除AliyunServiceRoleForDDI，否则提示删除失败。
   ii. 如果当前账号下已释放所有Databricks数据洞察集群，则可直接删除AliyunServiceRoleForDDI。

## 常见问题

为什么我的RAM用户无法自动创建Databricks数据洞察服务关联角色AliyunServiceRoleForDDI?

您需要拥有指定的权限，才能自动创建或删除AliyunServiceRoleForDDI。因此，在RAM用户无法自动创建AliyunServiceRoleForDDI时，您需为其添加以下权限策略。

AliyunServiceRoleForDDI时，您需为其添加以下权限策略。

```
{
  "Statement": [
    {
      "Action": [
        "ram:CreateServiceLinkedRole"
      ],
      "Resource": "acs:ram:*:主账号ID:role/*",
      "Effect": "Allow",
      "Condition": {
        "StringEquals": {
          "ram:ServiceName": [
            "ddi.aliyuncs.com"
          ]
        }
      }
    }
  ],
  "Version": "1"
}
```

**说明** 请将 `主账号ID` 替换为您实际的阿里云账号（主账号）ID。

### 相关文档

- 服务关联角色

# 6.3. 为RAM用户授权

为确保RAM用户能正常使用Databricks 数据洞察控制台的功能，您需要使用云账号登录访问控制RAM（Resource Access Management），授予RAM用户相应的权限。

## 背景信息

访问控制RAM是阿里云提供的资源访问控制服务，更多详情请参见什么是访问控制。以下举例访问控制RAM的典型场景：

- 用户：如果您购买了多台Databricks 数据洞察集群实例，您的组织里有多个用户（如运维、开发或数据分析）需要使用这些实例，您可以创建一个策略允许部分用户使用这些实例。避免了将同一个AccessKey泄露给多人的风险。
- 用户组：您可以创建多个用户组，并授予不同权限策略，授权过程与授权用户过程相同，可以起到批量管理的效果

## 权限策略

权限策略分为**系统策略**和**自定义策略**。

- 系统策略：阿里云提供多种具有不同管理目的的默认权限策略。Databricks 数据洞察经常使用的系统策略：
  - AliyunDDIFullAccess：管理Databricks 数据洞察的权限，主要包括对Databricks 数据洞察的所有资源的所有操作权限。
  - AliyunDDIDevelopAccess：Databricks 数据洞察开发者权限，与AliyunDDIFullAccess策略相比，不授予集群的创建和释放等操作权限。
- 自定义策略：需要您精准地设计权限策略，适用于熟悉阿里云各种云服务API以及具有精细化控制需求的用户。您可以参见权限策略语法和结构创建自定义策略。

系统策略默认仅为RAM用户提供查看Bucket和Object列表权限，RAM用户无法编辑Bucket和Object。如需更多OSS权限策略请参见OSS数据权限隔离

**授权建议：**主账号可为RAM用户授予AliyunDDIDevelopAccess权限，基本可满足日常开发需求；

## 授权RAM用户

执行以下步骤在访问控制RAM控制台授权RAM用户Databricks 数据洞察相关权限。

1. 使用云账号登录RAM控制台。
2. 单击左侧导航栏的**人员管理 > 用户**。
3. 单击待授权RAM用户所在行的**添加权限**。
4. 单击需要授予RAM用户的权限策略，单击**确定**。具体权限策略请参见**权限策略**。
5. 单击**完成**。完成授权后，权限立即生效。