

ALIBABA CLOUD

阿里云

Databricks 数据洞察  
Notebook

文档版本：20200928

 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 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 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 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 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。
<b>粗体</b>	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。
<code>Courier</code> 字体	命令或代码。	执行 <code>cd /d C:/window</code> 命令，进入Windows系统文件夹。
<i>斜体</i>	表示参数、变量。	<code>bae log list --instanceid</code> <i>Instance_ID</i>
[ ] 或者 [a b]	表示可选项，至多选择一个。	<code>ipconfig [-all -t]</code>
{ } 或者 {a b}	表示必选项，至多选择一个。	<code>switch {active stand}</code>

# 目录

1.Notebook概述	05
2.管理Notebook	06
3.使用Notebook	08


# 1.Notebook概述

DataInsight Notebook是基于Web的交互式数据分析Notebook，提供了作业编辑、数据分析、数据可视化等功能。全面兼容Apache Zeppelin，您可以使用Scala、Python、Spark SQL、R等语言编写Spark程序。

## 相关操作

有关Notebook的更多操作，请参见：

- [管理Notebook](#)
- [使用Notebook](#)

 **说明** 每个Databricks数据洞察集群都会部署独立DataInsight Notebook服务。用户在使用Notebook时，需要先选择一个可用的集群。

## 2. 管理 Notebook

本文介绍如何创建、打开、删除和导入 Note 等操作，帮助您管理 Notebook。

### 前提条件

已创建集群，详情请参见[创建集群](#)。


### 创建 Note

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 **Notebook**。
4. 在 Notebook 区域，选择待操作的集群。
5. 单击 **Create new note**。
6. 在 **Create new note** 对话框中，输入 **Note Name**、从 **Default Interpreter** 列表，选择 **spark**。
7. 单击 **Create**。


### 打开 Note

在 Notebook 页面，单击已创建的 Notebook 名称，进入 Notebook 详情页面。



### 删除 Note

1. 在 Notebook 页面，单击 Notebook 名称后的  图标。
2. 在 **Move this note to trash?** 对话框中，单击 **OK**。

### 重命名 Note


1. 在 Notebook 页面，单击 Notebook 名称后的  图标。
2. 在 **Rename note** 对话框中，输入新的名称。
3. 单击 **Rename**。

### 导入 Note

1. 在 Notebook 页面，单击 **Import note**。
2. 在 **Import New Note** 对话框中，输入 note 的名称，选择导入的 note。
  - 单击  图标，选择待导入的文件，单击打开。
  - 单击  图标，输入 URL。
3. 单击 **Import note**。

### 导出 Note

1. 在 Notebook 页面，单击已创建的 Notebook 名称。

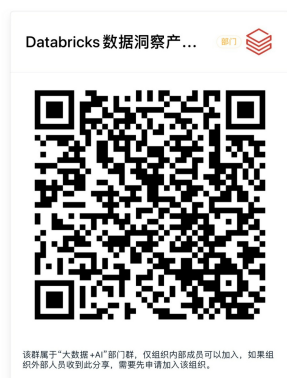
2. 单击上方的  图标。即可将Note下载到本地。

Note的备份文件支持两种格式：

- zpln格式：导入到另外的DataInsight Notebook或者Apache Zeppelin Notebook（DataInsight Notebook 100%兼容Apache Zeppelin Notebook）。
- ipynb格式：导入到Jupyter Notebook。

## 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。



# 3.使用 Notebook

Notebook是由一个或多个Note单元组成的，每个Note是一个独立的Spark任务。本文介绍如何使用 Notebook。

## 前提条件

已创建Note，详情请参见[管理 Notebook](#)。

## 开发Note

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**Notebook**。
4. 在**Notebook**区域，选择待操作的集群。
5. 在**DataInsight Notebook**页面，单击创建好的Note名。您可在单元格里编辑Spark作业。

单元格的第一行需要指定Interpreter。DataInsight Notebook目前支持以下6种Interpreter。

Interpreter	说明
<code>%spark</code>	提供了Scala环境。
<code>%spark.pyspark</code>	提供了Python环境。
<code>%spark.ipyspark</code>	提供了IPython环境。
<code>%spark.r</code>	提供了R环境，支持SparkR。
<code>%spark.sql</code>	提供了SQL环境。
<code>%spark.kotlin</code>	提供了Kotlin环境。

## 添加单元格

在DataInsight Notebook页面，将鼠标移动到任意已存在单元格的顶部或底部，单击+ Add Paragraph，即可在页面上添加新的单元格。

## 创建表

1. 单击已创建的Note名称。
2. 在DataInsight Notebook页面，在单元格中创建数据库。

```
%spark.sql

create database db_demo location 'oss://databricks-dbr/db_demo_database';
```

3. 在单元格创建表。



```
%spark.sql

use db_demo;

create table db_bank_demo(age string, job string, marital string, education string, default string,
balance string, housing string, loan string, contact string, day string, month string, duration string
, campaign string, pdays string, previous string, poutcome string, y string) ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

#### 4. 导入数据到数据库。

```
%spark.sql

use db_demo;


load data inpath 'oss://databricks-dbr/db_demo/bank/bank.csv' overwrite into table db_bank_demo;

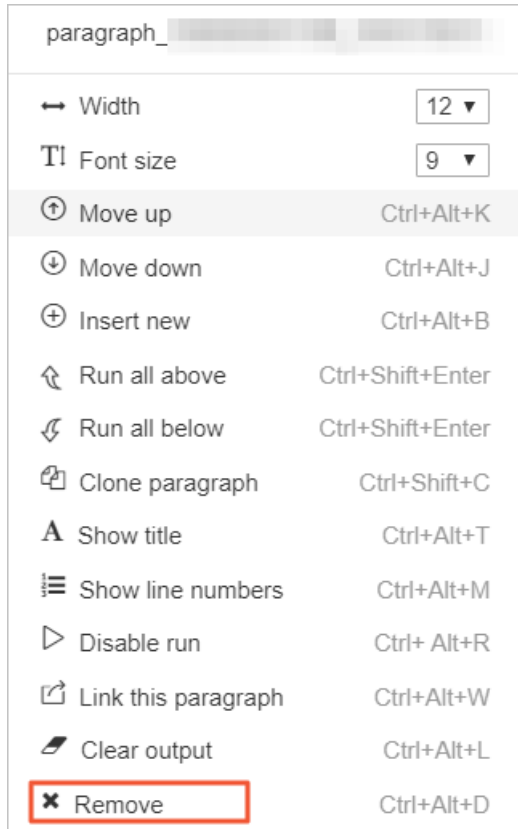
describe db_bank_demo;
```

导入成功后，查看表信息如下所示。

col_name	data_type	comment
age	string	null
job	string	null
marital	string	null
education	string	null
default	string	null
balance	string	null
housing	string	null
loan	string	null

## 删除单元格

1. 在 Datalnsight Notebook 页面，单击单元格右上角的  图标。
2. 选择 Remove。



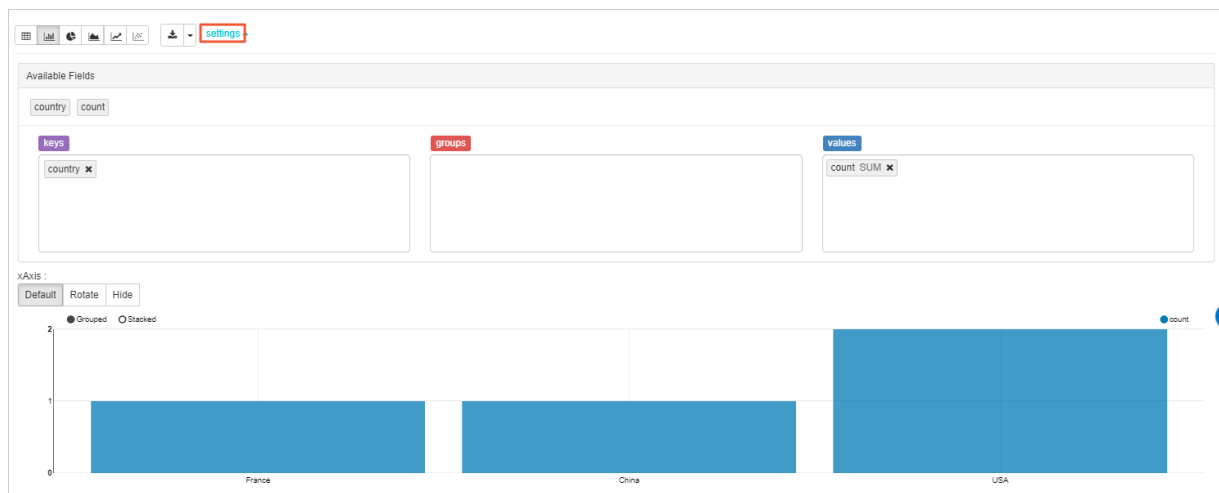
3. 在弹出框中单击OK。即可删除当前单元格。

### 运行Note

在DataInsight Notebook页面，单击单元格右上角的 图标，即可在Notebook内运行作业。

### 查看可视化运行结果

运行完Note后，在当前单元格中，可单击图形来查看运行结果。Notebook内置了多种图形来可视化Spark的DataFrame：Table、Bar Chart、Pie Chart、Area Chart、Line Chart、Scatter Chart，并且您可以单击settings对各种图形进行配置。



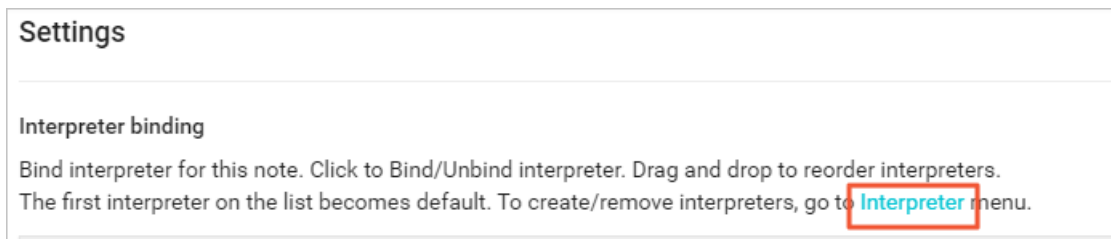
### 查看作业详情

1. 在 DataInsight Notebook 页面，单击单元格右上角的 SPARK JOB。
2. 选择待查看的作业。即可跳转至该作业的 Spark UI，查看作业执行详情。

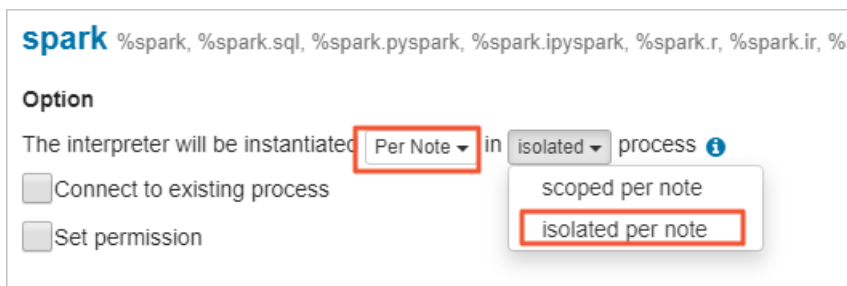
### 修改 Interpreter 模式

默认情况下 Spark Interpreter 的绑定模式是 Shared 模式，即所有的 Note 都是共享同一个 Spark App。如果是多用户场景的话，建议设置成 Isolated Per Note，这样每个 Note 都有自己独立的 Spark App，互相不会有影响。

1. 在 DataInsight Notebook 页面，单击右上角的 ⚙️ 图标。
2. 在 Settings 区域，单击 Interpreter。



3. 在 spark 区域，单击  图标，按截图设置以下参数。



4. 单击 Save。
5. 在弹出框中单击 OK。

### 配置 Interpreter

支持以下两种方式配置 Interpreter：

- 配置全局的 Interpreter。
  - i. 在 DataInsight Notebook 页面，单击右上角的 ⚙️ 图标。
  - ii. 在 Settings 区域，单击 Interpreter。
  - iii. 在 spark 区域，单击 edit，修改相关的参数。
  - iv. 单击 Save。
  - v. 在弹出框中单击 OK。

- 配置单个 Note 的 Interpreter。

通过 `%spark.conf` 来对每个 Note 的 Spark Interpreter 进行定制化，但前提是把 Interpreter 设置成 `isolated per note`。

在 DataInsight Notebook 页面的 `%spark.conf` 区域，可修改相关的参数。

```
%spark.conf
SPARK_HOME <PATH_TO_SPARK_HOME>

#set driver memory to 8g
spark.driver.memory 8g

#set executor number to be 6
spark.executor.instances 6

#set executor memory 4g
spark.executor.memory 4g
```

## 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。

