



Databricks 数据洞察 项目空间

文档版本: 20210129



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。

- 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。
- 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。
- 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。	介 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。	會学者 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。	大) 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 <i>,</i> 不是 用户必须了解的内容。	⑦ 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。
粗体	表示按键、菜单、页面名称等UI元素。	在 结果确认 页面 <i>,</i> 单击 确定 。
Courier字体	命令或代码。	执行 cd /d C:/window 命令 <i>,</i> 进入 Windows系统文件夹。
斜体	表示参数、变量。	bae log listinstanceid
[] 或者 [alb]	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 {act ive st and}

目录

1.项目管理	05
2.管理项目告警	08
3.管理作业	09
4.作业提交示例	11
5.工作流使用案例	15
6.工作流调度Zeppelin Notebook	21
7.管理项目	25
8.管理运维中心	28

1.项目管理

创建Dat abricks 数据洞察集群后,您可以在数据开发中创建项目,并在项目中进行作业的编辑和工作流的调度。新建项目之后,您可以对项目进行管理,为项目关联集群资源、添加项目成员以及设置全局变量。

前提条件

- 1. 已创建集群,详情请参见集群创建
- 2. 使用阿里云账号登录使用阿里云账号登录Databricks数据洞察控制台。使用阿里云账号登录Databricks数据 洞察控制台。

○ 注意

只有主账号才能创建项目、添加项目成员和添加集群资源,即新建项目、用户管理和集群设置只对主账 号管理员可见。

新建项目

- 1. 单击左侧导航栏的项目空间
- 2. 单击右上角的新建项目
- 3. 在新建项目对话框中, 输入**项目名称**和**项目描述**
- 4. 单击创建。

↓ 注意

1. 主账号可以查看该账号下的所有项目列表, RAM用户仅可以查看具有开发权限的项目列表。如需为 RAM用户添加项目开发权限,则需要主账号来配置,详情请参见用户管理。

项目管理

- 1. 单击左侧导航栏的项目空间
- 2. 在项目列表页面,单击待查看基本信息的项目的**项目ID**。
- 3. 单击上方的项目管理。
- 4. 在项目列表中可以管理项目基本信息、用户管理、集群设置、变量定义

项目基本信息

可以查看该项目的项目名称、创建时间、Created User和项目描述。并可以修改项目名称和描述。

用户管理

在用户管理页面,可以为RAM用户添加或删除某个项目的开发权限:

- 1. 添加用户
 - i. 单击右上角的添加用户。
 - ii. 勾选主账号下的RAM用户到该项目。
 - iii. 单击添加。

iv. 在用户管理页面, 可以查看新增的用户信息。

? 说明

被添加的RAM用户将成为该项目的成员,拥有查看、开发该项目下的作业和工作流的权限。

2. 删除用户。

在用户管理页面,单击待删除用户所在行用户信息。

集群设置

在集群设置页面,可以为项目设置集群资源,使该项目中的作业运行在关联的集群上:

- 关联集群资源。
 - i. 在集群设置页面, 单击集群所在行的修改配置。
 - ii. 从选择集群下拉列表中选择已购买的包年包月或按量付费的集群(不包括通过集群模板创建的集群)。
 - iii. 单击确定

在集群设置页面,可查看新增的集群信息。

• 取消关联集群资源。

在集群设置页面,单击集群所在行的删除,取消关联的集群资源

- 修改配置。
 - 在集群设置页面,单击集群所在行的修改配置。
 - 在修改配置对话框中,设置提交作业到该集群的队列和用户信息。

参数	描述
提交作业默认用户	设置项目使用所选集群提交作业时的默认用户,默认 值是hadoop,默认用户只能有一个。
提交作业默认队列	设置项目使用所选集群提交作业时的默认队列,默认 作业提交到default队列。
提交作业用户白名单	设置可以提交作业的用户,如果有多个用户,可以通 过英文半角逗号(,)分隔。
提交作业队列白名单	设置项目中的作业可以运行在所选集群的队列,如果 有多个队列,可以通过英文半角逗号(,)分隔。

○ 配置客户端白名单:配置可以提交作业的客户端,您可以使用Master节点或Gateway节点。通过ECS自 建的Gateway暂不支持在此处配置。

○ 单击确定。

变量定义

在变量定义页面,您可以设置项目级别的自定义变量,这些变量可以被本项目中的作业项目作为全局变量调用。

- 添加自定义变量。
 - i. 单击右上角的**添加**。
 - ii. 在添加自定义变量对话框中, 设置变量名称和变量值。

可以根据需要,选择是否为变量名的变量值加密。作业中以 \${VariableName} 的形式调用变量。例如,您添加变量名为ENV_ABC,变量值为12345,不勾选是否为密码。Shell类型作业内容示例如下。

echo \${ENV_ABC}

返回结果如下。

12345

这里环境变量的设置相当于执行了如下的脚本。

export ENV_ABC=12345

ⅲ. 单击确定。

在变量定义页面,可查看新增的变量信息。

- 删除自定义变量。
 - 在变量定义页面,单击待删除变量所在行的删除。

2.管理项目告警

本文介绍如何添加联系人和告警接收组。

添加联系人

- 1. 登录阿里云Databricks 数据洞察控制台
- 2. 在左侧导航栏单击项目空间
- 3. 单击项目空间中的告警
- 4. 在联系人区域,单击添加;
- 5. 在联系人对话框中, 输入姓名、手机号码和Email信息。
- 6. 分别单击手机号码和Email所在行的获取验证码。
- 7. 输入获取的手机验证码和邮箱验证码。
- 8. 完成上述参数配置后,单击确定。

添加告警用户组

- 1. 登录阿里云Databricks 数据洞察控制台
- 2. 在左侧导航栏单击项目空间
- 3. 单击项目空间中的告警
- 4. 在告警用户组区域,单击添加;
- 5. 在**告警用户组**对话框中,输入告警用户组的**名称**和**备注**,单击**添加**。可以添加成员信息,每个告警用 户组可以关联多个成员。
- 6. 在成员区域,从用户选择列表,选择已添加的联系人。
- 7. 完成上述参数配置后,单击确定。

添加告警钉钉群

- 1. 登录阿里云Databricks数据洞察控制台。
- 2. 在左侧导航栏单击项目空间
- 3. 单击项目空间中的告警
- 4. 在告警钉钉群区域,单击添加。
- 5. 在告警钉钉群对话框中, 输入在名称、Web Hook URL和备注信息。
- 6. 完成上述参数配置后,单击确定。

3.管理作业

在项目中,您可以通过创建作业来进行任务开发。本文为您介绍如何创建作业、设置和删除作业。

前提条件

已创建项目, 详情请参见项目管理。

创建作业

- 1. 进入项目空间页面。
 - i. 使用阿里云账号登录Databricks数据洞察控制台。
 - ii. 在Dat abricks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。
 - iii. 在左侧导航栏, 单击**项目空间**。
- 2. 在项目空间页面,单击待操作项目所在行的作业编辑。
- 3. 在作业编辑区域,在需要操作的文件夹上单击右键,选择新建作业。

☑ 作业编辑	C 🖈
请输入	Q
V 🗄 JOB	
	新建作业
	新建子文件夹

4. 在新建作业对话框中, 输入作业名称、作业描述, 选择作业类型。

⑦ 说明 创建作业时作业类型一经确定,不能修改。

- 5. 单击确定。
- 6. 在作业内容区域,填写该作业需要提供的命令行参数。例如,输入以下命令,读取Zeppelin中建立的数据表,将结果直接写入OSS。

spark.sql("USE db_demo")

val df = spark.sql("select age, count(1) as age_count from db_bank_demo where age < 50 group by age order by age")

df.write.save("oss://<mytest>/db_demo_file")

- 7. 单击右上角的保存。
- 8. 单击右上方的运行。执行作业。

设置作业

新建作业时、或者新建作业后,可对作业进行以下设置。

- 1. 单击页面右上角的作业设置。
- 2. 在作业设置页面,设置基础信息。

参数	描述
作业概要	作业描述 :可修改作业的描述。
运行资源	单击右侧的 <mark>+</mark> 图标,添加作业执行所依赖的JAR包或HDFS等资源。您需要将资源先 上传至OSS,然后在 运行资源 中直接添加即可。
配置参数	指定作业代码中所引用的变量的值。您可以在代码中引用变量,格式为 <i>钒变量名</i>]。 单击右侧的 <mark>+</mark> 图标,添加Key和Value。其中,Key为变量名,Value为变量的值。

3. 在作业设置页面,单击高级设置。

配置项	说明
模式	 提交节点::目前只支持在在Master节点提交,表示作业在分配的机器上直接运行。 预期最大运行时长:0~10800秒。
环境变量	添加作业执行的环境变量,也可以在作业脚本中直接export环境变量。
调度参数	设置作业运行调度队列、内存、虚拟核数、优先级和执行用户等信息。当未设置这 些参数时,作业会直接采用集群的默认值。

删除作业

创建作业后,不再需要当前作业时可删除。

1. 在项目空间页面,单击待操作项目所在行的删除。

2. 在**删除**对话中*,*单击**确定**。

4.作业提交示例

前提条件

- 在Dat abricks数据洞察产品中创建一个集群,详情请参见创建集群
- 已使用OSS管理控制台创建非系统目录存储空间,详情请参见创建存储空间。

△ 警告

首次使用DDI产品创建的Bucket为系统目录Bucket,不建议存放数据,您需要再创建一个Bucket来读 写数据。

? 说明

DDI访问OSS路径结构: oss://BucketName/Object

- Bucket Name为您的存储空间名称;
- Object为上传到OSS上的文件的访问路径。

例:读取在存储空间名称为databricks-demo-hangzhou文件路径为 demo/The_Sorrows_of_Young_Werther.txt的文件

// 从oss地址读取文本文档

val text = sc.textFile("oss://databricks-demo-hangzhou/demo/The_Sorrows_of_Young_Werther.txt")

步骤一:示例下载

- 示例文本下载: The_Sorrows_of_Young_Werther.txt
- 示例工程下载: spark-wordcount-examples-1.2.0.jar
- 上传OSS目录: 自定义OSS存储目录。

步骤二:创建项目空间

在左侧导航栏中,选择项目空间,单击新建项目按钮,填写项目名称、项目描述,单击创建按钮。

项目空间·<mark>作业提交示例</mark>

Databricks 数据洞察	Databricks 数据洞察 / 数据开发							
首页	项目空间							
91.87	■ 新建项目					@ C		
Notebook	项目ID/名称	損	述	创建时间	管理员		操作	
项目空间		\$5 ZB YE D				×	作业编辑 运行记录 删除	
元数据		新建坝日				~		
库		[注意] 请使	用主账号创建项目				作业编辑 运行记录 删除	
帮助		* 项目名称:	test-ddi					
		• 项目描述:	测试DDI产品。					
						_		
					0132	取消		
						-		8

二级菜单选择作业编辑,新建作业。

Databricks 数据洞察	く返回 test-ddi イ	作业开发	项目管理	這端中心	帮助
首页	◎ 作业编辑	c			
集群	请输入	Q			
Notebook	∨ ⊟ јов		新建作业		
项目空间			新建子文件夹		
元数据					
库				您可以开始以下操作:	
帮助				- 在左侧开始台膛编辑任务	
				- 工作田田重	
	<				
		~			
			-		

步骤三:编写并提交作业

--class com.aliyun.emr.example.spark.SparkWordCount oss://databricks-demo/test-sparksubmit-word-cou nt/spark-wordcount-examples-1.2.0.jar

 $oss://databricks-demo/test-sparksubmit-word-count/The_Sorrows_of_Young_Werther.txt$

oss://databricks-demo/test-sparksubmit-word-count/output/temp/011

项目空间·作业提交示例

Databricks 数据洞察	< 返回 test-wordCount ✓	作业开发 1	項目管理	运维中心																		❷ 帮助
首页	◎ 作业编辑	B word	dCount ×																		_	
集群	请输入 C	Y SPA	オ SPARK FJ-A736ED75F52989D3 作业内容: ② ▲上額 ◎ 运行 常止								保存	保存 创建快照 作										
Notebook	✓ □ JOB	1	1class com.aliyun.emr.example.spark.SparkWordCount oss://databricks-demo/test-sparksubmit-word-count/ spark-wordcount_examples_1_2_0_jar_oss://databricks-demo/test-sparksubmit-word-count/The Sorrows of Youn									nt/ Young !	a Werther.txt									
项目空间			oss://d	atabricks	s-demo/te	est-spar	arksubm	omit-wo	vord-cou	unt/out	utput/1	temp/@	01 1			 						- 1
元数据																						
d:																						
74.																						- 1
帮助																				Ť		- 1
																				+		
			实际运行	运行(仅供参考)								~	_		- 1							
	<				ubmit — class con aliyun eer example.spark.SparkWordCount oss://databricks-demo/test-sparksubmit-word-count/spark-wordcount 1.2.8.jar oss://databricks-demo/test-sparksubmit-word-count/The_Sorrows_of_Young_Werther.txt oss://databricks-demo/test-spa ord-count/output/temp/81 1							rdcount- est-spar	t-ex arks									
		日志	运行记录	审计日志	版本控制	间											+ 插入0	SS路径	∂ 去OSS	控制台上作		~ ~
		LESSO EMR_A SHELL HIVE DOB_P YARK_T SPARK HADOO CONDA FLOW SHUV SHUV	<pre>SPEN= /uss SPEN= /uss PP_HOME_DI #/bin/bash HOME=/usr./ IRS=/mnt/dis HOME=/usr./ MP_DTR=/us (DF_IDIR=/us (CONF_DIR=/us (SET_LOG_ AGENT_CON AGENT_CON =3 NATA_DIR=/u</pre>	/bin/lessp Re/usr/lib lib/hive-c iskl/flow- r/lib/emr/ kl/flow-ag sr/lib/had ofriExe-to- usr/lib/had ofFIExe-to- ITFIExe-to- sr/lib/emr	<pre>sipe.sh %s //emr/apps //emr/apps //emr/apps //emr/local- loop-curren //park-conf //emr/cem //data</pre>	al-rm/app] -rm/Local# nt/pids nt/pids t -agent/log agent-con:	pplogs/Lc alApplics d	'LocalApp	pplicatic	ion_1596	966246699	9544_2/c	contain 1596624	62466954 2_01_000	44_2_01_0							

步骤四:验证作业

作业执行完成后,查看作业运行记录。

Databricks 数据洞察	く 返回 test-ddi ~ 作业开发	项目管理	运维中心						❷ 帮助		
首页	◎ 作业编辑 C		word-count ×								
集群	请输入 Q	Y SPAR	K FJ-AFBFDB768A4F0	44F 作业内容: 🕜	▲上锁 G	9 运行选中代码	停止保存	创建快照	作业设置		
Notebook	V DJOB	1		yun.emr.example.spark.SparkWordCount oss://databricks-demo/1	test-sparksub	bmit-word-coun	nt/	100005	MEASURE		
项目空间	- a test-word-count (20204-87)	2	oss://databrick	s-demo/test-sparksubmit-word-count/The_Sorrows_of_Young_Wert			_				
		3	oss://databrick	s-demo/test-sparksubmit-word-count/output/temp/02 1							
14											
帮助								Ť			
			实际运行(仅供参考)					~ +			
			spark-submitcl	lass com.aliyun.emr.example.spark.SparkWordCount oss://databricks-demc	o/test-sparksub	bmit-word-count/	t/spark-wordcount-ex —				
			amples-1.2.0.jar oss://databricks-	-demo/test-sparksubmit-word-count/The_Sorrows_of_Young_Werther.txt				2			
	~		oss://databricks-	-demo/test-sparksubmit-word-count/output/temp/02 1	Jut/temp/02 1						
		志日	运行记录 审计日志 个	版本控制			+ 插入OSS路径	& 去OSS控制台上传 E	C 8195		
		运行实例	ID	开始时间 结束时间		ł	状态 操作				
		EJI-D500	09228356213E5	2020年8月7日 14:40:27 2020年8月7日	14:40:47	c	ок <u></u> Ш	E 400 i E 495 do 325 401			
									BA		

作业执行成功后,可在OSS相应目录查看当前作业产出数据。

项目空间·作业提交示例

Dat abricks 数据洞察

概览	上传文件 新建目录 荷片管理 授权 批量操作 > 刷新				请输入文件名前缀匹配 C
文件管理 >	文件名	文件大小	存储类型	更新时间	操作
収限管理 >	/ test-sparksubmit-word-count/ output/ temp/ 01/				
基础设置 >		OKB	标准存储	2020年8月07日 14:19	详情 更多 🗸
冗余与容错 >	part-00000	100.177KB	标准存储	2020年8月07日 14:18	详情 更多 🗸
传输管理 >					
日志管理 >					
数据处理 >					
数据统计 >					
					3

5.工作流使用案例

在Dat abricks数据开发中,您可以在项目空间的作业编辑中创建一组有依赖的作业,然后创建工作流,按照 依赖次序定义执行顺序,Dat abricks工作流支持基于有向无环(DAG)并行执行大数据作业

前提条件

- 1. 通过主账号登录阿里云 Databricks控制台。
- 2. 已创建集群,具体请参见创建集群。
- 3. 已创建好需要测试的作业,具体请参见作业提交示例。

新建项目空间

通过以下步骤,可以新建项目空间。

- 1. 在Databricks控制台页面,单机左侧导航栏的项目空间。
- 2. 单机新建项目。
- 3. 填写项目名称、项目描述、单机确定。

	[州] 👻		Q 搜索文档、控制		费用 工单	备案 企业	支持 官師	X) D_	۵.	`₩ (?)	简体	0
Databricks 数据洞察	Databricks 数据洞察 / 数据开	干发										
首页	项目空间											
集群	▣ 新建项目									告警设置	0	C
Notebook	项目ID/名称	所建项目				×	操	HF .				
项目空间		*项目名称:	ddi-test				作	业编辑 运	テ记录	删除		
元数据	doi-test	* 项目描述:	创建新的项目空间]						
<i>]库</i>	16 						作	业编辑 运行	亍记录	删除		
幣助							作	业编辑 运行	テ记录	删除		
	2: tt						作	业编辑 运行	テ记录	删除		
	EF				ėj iž	取消	作	业编辑 运往	テ记录	删除		
	da ang	tes	st-tpc-ds	2020年8月7日 18:20:33	÷ .		作	业编辑 运行	テ记录	删除	(
	nzqtest	tes	st	2020年6月10日 10:57:59	1.4		作	业编辑 运行	亍记录	删除		ă

新建工作流

通过以下步骤,可以新建工作流。

- 1. 进入项目空间。
- 2. 在工作流设计区域,在需要操作的文件夹上单击右键,选择新建工作流。
- 3. 填写工作流名称、工作流描述、执行集群。
- 4. 单机**确定**。

项目空间·工作流使用案例

⊟ (→) 阿里云 华东1	(杭州) 👻		Q 搜索文档、控制台、API、解决方案和资源	费用 工单	备案 企业	支持 官网	Þ Ú	. ∄	⑦ 简	* 📀
Databricks 数据洞察	く 返回 ddi-test 〜	作业开发 项目管理	2 运维中心							❷ 帮助
首页	計工作流设计 請輸入	新建工作流			×					
集群 Notebook		• <i>B</i> FT /	载项目: ddi-test							
项目空间		* 所属:	文件夹: FLOW/			作台。				
元数据		• 工作)	瓷名称: 工作流-01			.160.				
库		* 工作)	^{条描述:} 创建新的工作流							
释题		• 执行	「集群: 选择当前已存在集群 > Izh-test-02 > @							
				Ŕ	定 取消					

编辑工作流

- 1. 在工作流画布上,拖拽不同类型的作业节点到画布上,并进行节点编辑,选择相关作业。
- 2. 根据依赖关系,从节点底部中心可连接到另一节点上部中心。
- 3. 完成依赖关系后,再添建END组件。
- 4. 单机画布上方保存。

	(杭州) 👻	Q 提索文档、控制台、API、解决方案和资源 费用 工单 备案 企业 支持	官网 🖸 🛕 👾 🕐 简体 🔮
Databricks 数据洞察	く返回 ddi-test ✓ 作业开发 项目管	管理 运维中心	⊘ 帮助
首页	□□ III 工作流设计 C B sp.	park-s-01 × 評工作流-01 × 廖 spark-02 × 廖 spark-01 × 廖 spark-s-02 ×	
集群	🖻 _{请输入} Q 🖍		动布局 💿 运行 保存 创建快照
	11 V 🗄 FLOW		
Notebook	¥I 工作流-01 (2020年9月	节点组件 START	工作流
项目空间		控制节点	配置
元数据		START	
库		END	
帮助		作业节点	
		Spark	
	<pre>«</pre>	SparkStreaming spark-s-02	
			stwar
		END	
	运行记	记录 审计日志 版本控制	^

在编辑工作流时,可以单机画布上方**上锁**来为未工作流添加编辑锁,此时只有您可以编辑,其他项目人员无 法编辑和运行该工作流。只有解锁后,项目中其他人员才可以编辑该作业。

配置工作流调度

您可以打开工作流调度开关并配置工作流调度参数,调度系统会按照参数定时运行相关工作流,并将作业下 发到指定集群上执行。以下介绍配置工作流的基本属性、调度属性和告警设置。

完成工作流设计后,单机画布右上角配置按钮,进行相关配置。

1. 配置基本属性。

配置项		说明	
执行集群		工作流中作业	业默认的运行方法
 こう阿里云 学东1 Databricks 数据洞察 第4 	(統州) - Q 形象文档、短期台 く 返回 ddi-test > 作业开发 項目管理 道線中心 III 工作流设计 C III 工作流-01 × P 頃給入 Q へ 予 古 西 Q	API、解决方案和资源 工 Q X H	作流调度配置 × 基本属性 调度属性 告警设置 工作流印: ************************************
集群 Notebook 元武派 库 帮助	・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・		Infraster: ● 中市市市市 ● 中市市市 ● 中市市市市 ● 中市市市市 ● 中市市市市 ● 中市市市 ● 中市市市 ● 中市市市 ● 中市市市 ● 中市市市 ● 中市市市 ● 中市市 ● 中市市 ● 中市市 ● 中市 ● 中市 ● 中市 ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

2. 配置调度属性。

配置项	说明
调度状态	启动或停止工作流调度。开启后,工作流编辑画布上 方会出现调度中的状态提示。
时间属性调度	设置工作流调度的开始时间和调度周期,在此时间范 围内,系统会根据您设置的周期执行工作流。
依赖属性调度	选择当前工作流的前续工作流。当前续工作流执行完成后,当前工作流才会被调度执行。 i.选择所选项目。 ii.从所选项目中,选择依赖工作流。

项目空间·工作流使用案例

	Ⅰ (杭州)	Q 授索文档、控制台、API、解决方案和资源	工作流调度配置	>
Databricks 数据洞察	く返回 ddi-test ❤ 作业开发	项目管理 运维中心	基本属性 调度属性 告警设置	
首页	詳 工作流设计 C	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□	* 调度状态: ④ 启动 🦳 停止	
集群 Notebook			时间属性调度	
项目空间		控制节点	* 开始时间: 2020-09-23 10:33:42	
元数据库		END	结束时间: 请选择日期和时间 ====================================	
帮助		作业节点 Spark	小时 请选择 V ,分钟 00 V 运行一次	
		SparkStreaming	* CRON表达式: 0 0 **?	
			依赖属性调度	
			所 鳳项目: 请选择项目	
			依赖工作流: 请输入依赖的工作流名转	
			所属项目 工作流名称 操作	
			没有数据	
1998 - 1998 - 1998 - 1998 - 1998 - 1998 - 1998 - 1998 - 1998 - 1998 - 1998 - 1998 - 1998 - 1998 - 1998 - 1998 -		运行记录 审计日志 版本控制		

3. 配置告警设置。

配置项	说明
执行失败	设置工作流执行失败时,是否通知到用户告警组或钉 钉告警组。
节点失败	设置工作流节点失败时,是否通知到用户告警组或钉 钉告警组。
执行成功	设置工作流执行成功时,是否通知到用户告警组或钉 钉告警组。
启动超时	设置如果工作流中有节点在下发到集群后30分钟内还 没有启动时,是否通知到用户告警组或钉钉告警组。
节点执行超时	设置如果节点执行时长超过作业配置里的预期最大运 行时长时,是否通知到用户告警组或钉钉告警组。

	(杭州) 👻	Q 报素文档、控制台、API、解决方案和资源	工作流调度配置	×
Databricks 数据洞察	く 返回 ddi-test 〜 作业开发	项目管理 运维中心	基本属性 调度属性 告警设置	
首页	III III 工作流设计 C III IIII IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	コ 工作流-01 ×	执行失败: 🗋 通知到用户告警组: 请选择 💛	
集群			通知到钉钉群告警组: 请选择 ∨	
Notebook	¥# 工作流-01 (2020年9月	节点组件	节点失败: 🗾 通知到用户告警组: Izh 🗸	
项目空间		START	□ 通知到钉钉群告警组: 告繫Ding ∨	
库		END	执行成功: 通知到用户告警组: 请选择	
帮助		作业节点	□ 想知封封封件百富祖。 申[237章 *	
	·	SparkStreaming	□ 通知到打打群告警组: 请选择 ✓	
	>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>		节点执行超时: 🗌 通知到用户告警组: 请选择 🗸	
			□ 通知到钉钉群告警组: 请选择 >>	
	-	运行记录 审计日志 版本控制		

执行工作流

您也可以指定工作流的业务时间,此时工作流作业中的时间相关变量将使用指定的业务时间进行计算,一般 用于重跑某个时间段的工作流实例,可以设置单次重跑或批量重跑。如果您的作业中没有任何时间相关变 量,可以选择在当前时间立即执行,即可运行工作流。

- 1. 单机项目空间。
- 2. 在**项目列表**页面,进入对应的项目空间。
- 3. 选择工作流设计。
- 4. 单机画布上方的运行按钮。
- 5. 配置相关运行参数。
 - i. **立即运行**:立即运行一个工作流,可以将指定时间作为本工作流的业务时间,时间相关的变量将使 用该时间进行计算。
 - ii. 设置时间调度运行:设置开始时间和调度周期。打开跳过成功节点开关时,如果某个业务时间 对应的工作流实例是成功的,将会跳过该业务时间的实例,继续运行其他业务时间失败的工作流实 例。

立即运行一批工作流,设置工作流业务时间的开始时间和调度周期,运行时系统会将指定调度规则 的触发时间作为本工作流的业务时间,时间相关的变量将使用该时间进行计算。一次最多支持100 个触发时间点。

6. 单机确定。

查看工作流运行记录

运行工作流后,可通过以下步骤查看工作流运行记录。

- 1. 在工作流页面,单机画布下方运行记录标签,可以查看当前工作流的运行记录、审计日志、版本控制。
- 2. 单击工作流实例所在行的详情,跳转至运维中心。

查看工作流实例的详细情况,也可以暂停、恢复、停止和重跑工作流实例,详情请参见运维中心。

- 详情: 查看工作流实例的详细信息, 包括工作流节点实例的详细信息和运行状态。
- 停止工作流:终止正在运行的工作流实例,所有正在运行的作业节点立即停止。

- 暂停工作流:暂停正在运行的工作流实例,正在运行的作业节点会继续执行,但后续的作业节点不再执行。
- 恢复工作流:恢复已被暂停的工作流实例。
- 重跑工作流:重新运行已经结束的工作流实例。单击重跑后,可以选择只重试失败节点,也可以从头重跑 所有节点。

工作流可执行操作

在工作流设计区域,您可以在工作流名称上单击右键,执行如下操作:

- 1. 克隆工作流: 在同一文件夹下克隆出相同图形的工作流。
- 2. 重命名工作流:重新命名工作流的名称。
- 3. 删除工作流: 删除工作流。当工作流在运行状态时无法删除。

6.工作流调度Zeppelin Notebook

在Databricks数据开发中,您可以在项目空间的作业编辑中创建一组有依赖的**Zeppelin**作业,然后创建工作流,按照依赖次序定义执行顺序,Databricks工作流支持基于有向无环(DAG)并行执行大数据作业

前提条件

- 1. 通过主账号登录阿里云 Dat abricks控制台。
- 2. 已创建集群,具体请参见创建集群。
- 3. 已创建好需要使用的Notebook, 具体请参见使用Notebook。

新建项目空间

通过以下步骤,可以新建项目空间。

- 1. 在Databricks控制台页面,单击左侧导航栏的项目空间
- 2. 单击新建项目。
- 3. 填写项目名称、项目描述、单击确定。

创建zeppelin作业

- 1. 进入项目在二级菜单中新建作业。
- 2. 填写作业名称、作业描述,作业类型选择zeppline、单击确定
- 3. 编辑调度Notebook、单击保存

使用作业调度Notebook有2种方式

方式一: 使用OSS路径调度Notebook

oss://ddi-cn-hangzhou-123456789/DDI_C-464E4772BC68CBB/jfs_root/zeppelin/notebooks/case_2FS1V4 ZZB.zpln

⑦ 说明

您创建并编辑的Notebook存储在对应集群所在的OSS文件中,您可以在作业中编写Notebook所在的OSS路径来调度作业

Dat abricks 数据洞察

Databricks 数据洞察	<	返回	xuzhifeng 🗸	作业开发	项目管理	运维中/	Û.									0 帮
首页	н	₽ f	■业编辑	c	រា Flow调	腹测试 ×	⊮ 调度	则试 ×								
40.7Y	P	1	业编辑	Q	Ø ZEPPE	ELIN FJ-1EB1	IFFB00167	311D 作业内容: 🕻	9			保存	◎ 运行	作业设置	合 上锁	创建快照
300.01	21	\sim	🗂 ЈОВ		1 0	oss://ddi	-cn-han	gzhou-	2,	/DDI_C-4	/jfs_root/	zeppelin/note	books/case_	2FS1V4ZZB.zp	pln	
Notebook			🥭 调度测试	2020年11月												
项目空间			🥔 调度测试2	(2020年11月												
元数据			🥔 Note调度测	试— (2020												
di:																
-																
帮助																
	<															
				~												
															+	
															+	G
															_	L.
																88
														~	~	
					日志 注	运行记录	审计日志	版本控制					+ 插.	(OSS路径 @ #	去OSS控制台上 情	d" ^ v

方式二:使用Notebook ID调度Notebook

☰ (-) 阿里云 ≄北5	(呼和浩 ▼		Q 搜索文档、控制台、API、解决方案和资源	费用 工单	备案 企业	支持 官网 🖸	ů . ∺ () (⢠📀
Databricks 数据洞察	く返回 xuzhifeng项目測试 V f	乍业开发 项目管理	运维中心					❷ 帮助
首页	◎ 『 作业编辑 C	₿ case1 ×						
集群	□ 请输入 Q	ZEPPELIN FJ-99	E48763D9416508 作业内容: 🕜		保存	 ○ 运行 作业设 	置 合上锁	创建快照
Notebook	** ~ 🗁 JOB	1 2FTE1D39	9N					
项目空间	Case2 (2020年11月15)	1						
元数据								
库								
帮助								
	<							
	~							
							_	
							Ť	
							+	
							-	
		2575122	NI .				~ ×	
		2FTEID35 日志 运行记录	m 审计日志 版本控制			+ 插入OSS路径	⊘ 去OSS控制台上传 r	
			MAY 11 July 194			· More considering		
您可以在作业编	扁辑区填写Noteb	ookID来调	周度Notebook。					

⑦ 说明	
您可以在Zeppelin里 Notebook URL 链接尾部中找到NotebooK ID。如下图所	示
🗧 🔶 C 🚺 本安全 2d62a8a6586.cn-huhehaote.databricks.aliyuncs.com:8443/gateway/cluster-topo/zeppelin/?spm=a2c8w.12369599.notebook_list.2.75b735	5c4OjhoAu#/notebook/2FS6S2E91 🛧 🕻 :
	ini (ili 💼 dati 🏚 Manada Andare
DataInsight Notebook Notebook - Job	Q Search O dwm -
	😇 🏟 🔒 default 🗸
%spark	SPARK JOB FINISHED D ※ 圓 戀
sc.range(1,10).sum()	
res1: Double = 45.0	
Took 30 sec. Last updated by dwm at November 19 2020, 5/29:37 PM.	

△ 警告

使用Notebook ID调度Notebook只能使用**当前Notebook所在集群**运行作业,否则会无法成功运行。工作流调度作业时,请注意并设置好运行集群。

4. 设置作业参数(Knox)

作业调度Notebook,需要配置**执行集群**knox账号相关的参数来连接**作业**和Notebook,具体操作步骤 如下:

- i. 单击作业设置
- ii. 左侧抽屉中单击高级设置
- iii. 在环境变量中单击添加并填写参数
 - a. 参数一: Z_USE_KNOX true
 - b. 参数二: Z_USERNAME Knox-username
 - c. 参数三: Z_PASSWORD Knox-password

	〕 阿里云 "华东1 (杭州) ▼									作业设置	作业设置							
Databricks 数据洞察	<	iio	xuzhifeng 🗸	作业开发	项目管	理 运维中心				基础设置	高级设置							
首页		P	作业编辑	c	◎调度	测试2 × ◎ ¥	同度测试 ×											
集群		î	非输入	Q	🥏 ZEP	PELIN FJ-1EB1FF	B00167311D 作业((内容: 🕜		模式								
Notebook		~	➡ JOB	(2020年11月 ⁻	1	oss://ddi-c	n-hangzhou-1	122288651453	3982/DDI_C-464D477	2	提交节点:	在Master节点提交 🗸						
項目空间			● 调度测试2	(2020年11月							ф 1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.	S模式下,提交客户端将在M	faster节点运行。					
元数编			Note調催潮	武一(2020						75991827	C1011012C;	0 10						
库										环境变量				+				
程助																		
											參数1:	Z_USE_KNOX	true	-				
											參数2:	Z_USERNAME		-				
											參数3:	Z_PASSWORD	and the second	-				
				~														
										调度参数								
											调度队列:							
											内存 (MB):							
											虚拟核数:							
											优先级:	0-100						
										执行	i用户: 🕜							
						oss://ddi-cr	n-hangzhou-1222	2886514533982,	/DDI_C-464D4772BC681BE	31								
					日志	运行记录 审计	计日志 版本控制	RJ										

↓ 注意

参数Z_USERNAME和Z_PASSWORD对应的是你需要**执行集群**Knox账户和密码,如果输入 错误作业无法成功运行

使用工作流调度作业

- 1. 二级菜单新建工作流。
- 2. 填写工作流名称、工作流描述、执行集群。
- 3. 单击**确定**

∃ (-) 阿里云 桦东1	(杭州) 👻		Q 搜索文档、控制台、API、解决方案和资源	费用 工单	备案 企业	支持 官网	>_	Ō. Ä	0	简体	0
Databricks 数据洞察	く返回 ddi-test 〜	作业开发项目管理	运维中心								❷ 帮助
首页	詳工作流设计 请输入	新建工作流			×						
栗群 Notebook	11 ~ 日 FLOW 11 工作流-	* 所属项目:									
项目空间		* 所属文件夹:				作台:					
元数据		* 工作流名称:	工作流-01								
库		* 工作流描述:	创建新的工作流								
747 B()		* 执行集辭:	选择当蓟已存在集群 💙 Izh-test-02 🏏 🖉								
				59	定取消						

- 4. 编辑工作流
 - i. 在工作流画布上,拖拽不同类型的作业节点到画布上,并进行节点编辑,选择相关zeppelin作业。
 - ii. 根据依赖关系,从节点底部中心可连接到另一节点上部中心。
 - iii. 完成依赖关系后,再添建END组件。
 - iv. 单击画布上方保存。

	(杭州) 👻		Q 搜索文档、控制台、API、解决方案和资源 费用 工单	备案 企业 支持 官网	🖾 🛕 🔄 🧟 🚳 🔂
Databricks 数据洞察	く返回 xuzhifeng → 作业开	· 项目管理 运维中心			● 帮助
首页		和 Flow调度测试 ×			
集群	IIII iiii iiii Q		ବ୍ 🛠 🗉	☆ 上锁 ▲ 自动布)	局 ◎ 运行 保存 创建快照
Notebook		在 节点组件			工作流
项目空间		控制节点	START		配置
元数据		START			
库		END			
帮助		作业节点			
	<	SparkStreaming			
		Zeppelin			
			END		
					START
					END END
		运行记录 审计日志 版本控制			~ ~
p169322.png	^				显示全部 ×

5. 单击运行,可以调度Zeppelin作业

⑦ 说明			
你可以设置工作流配置,	配置调度属性。	详情请参考 <mark>工作流使用案例</mark>	

7.管理项目

新建项目之后,您可以对项目进行管理,为项目关联集群资源、添加项目成员以及设置全局变量。

前提条件

已创建项目, 详情请参见项目管理。

管理项目基本信息

- 1. 使用阿里云账号登录Databricks数据洞察控制台。
- 2. 在Dat abricks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, 单击**项目空间**。
- 4. 在项目空间页面,单击待操作项目所在行的作业编辑。
- 5. 单击上方的项目管理页签。
 - 在基本信息页面,可查看当前项目的详细信息。
 - 在基本信息页面,单击编辑,可修改当前项目的项目名称和项目描述信息,单击保存。

管理项目集群

- 1. 使用阿里云账号登录Databricks数据洞察控制台。
- 2. 在Dat abricks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,单击项目空间。
- 4. 在项目空间页面,单击待操作项目所在行的作业编辑。
- 5. 单击上方的项目管理页签。
- 6. 单击集群设置。
 - 在集群设置页面,可查看当前项目关联的集群信息。
 - 在集群设置页面,可为项目添加关联的集群。
 - a. 单击添加集群。
 - b. 在添加集群对话框中,从选择集群列表,选择待添加的集群。
 - c. 单击**确定**。
 - 在集群设置页面,可为项目添加关联的集群。
 - a. 单击修改配置。

b. 在修改配置对话框中,设置提交作业到该集群的队列和用户信息。

参数	描述
提交作业默认用户	设置项目使用所选集群提交作业时的默认用户,默 认值是hadoop,默认用户只能有一个。
提交作业默认队列	设置项目使用所选集群提交作业时的默认队列,默 认作业提交到default队列。
提交作业用户白名单	设置可以提交作业的用户,如果有多个用户,可以 通过英文半角逗号(,)分隔。
提交作业队列白名单	设置项目中的作业可以运行在所选集群的队列,如 果有多个队列,可以通过英文半角逗号(,)分 隔。

c. 配置客户端白名单: 配置可以提交作业的客户端, 您可以使用Master节点。

- d. 单击**确定**。
- 在集群设置页面,单击删除。

可以删除已添加的集群。

管理用户

- 1. 使用阿里云账号登录Databricks数据洞察控制台。
- 2. 在Dat abricks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,单击**项目空间**。
- 4. 在项目空间页面,单击待操作项目所在行的作业编辑。
- 5. 单击上方的项目管理页签。
- 6. 单击用户管理。
 - 在用户管理页面,可查看当前项目关联的集群信息。
 - 在用户管理页面,可为项目添加关联的RAM用户。
 - a. 单击添加用户。

⑦ 说明 添加用户前,请先在RAM控制台授权并添加子账号。

- b. 在添加用户对话框中, 勾选待添加的用户。
- c. 单击**确定**。
- 在用户管理页面,单击删除。

可以删除已添加的RAM用户。

定义变量

- 1. 使用阿里云账号登录Databricks数据洞察控制台。
- 2. 在Dat abricks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,单击**项目空间**。
- 4. 在项目空间页面,单击待操作项目所在行的作业编辑。
- 5. 单击上方的**项目管理**页签。
- 6. 单击变量定义
 - 在基本信息页面, 可添加自定义变量。
 - a. 单击右上角的**添加**。
 - b. 在添加自定义变量对话框中,设置变量名称和变量值。 可以根据需要,选择是否为变量名的value加密。
 - c. 单击确定。

在变量定义页面,可查看新增的变量信息。

○ 在基本信息页面,单击删除。

可以删除已添加的自定义变量。

8.管理运维中心

本文介绍通过运维中心可以查看作业概览,监控任务运行状态,管理流式作业和查看审计日志,便于对作业 的管理和运维。

前提条件

已创建项目, 详情请参见项目管理。

查看作业概览

- 1. 使用阿里云账号登录Databricks数据洞察控制台。
- 2. 在Dat abricks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,单击**项目空间**。
- 4. 在项目空间页面,单击待操作项目所在行的运行记录。
- 5. 单击上方的**概览**页签。 展示当前项目下的作业数量、作业类型分布图和最近24小时作业状态分布情况 。

管理作业运行记录

- 1. 使用阿里云账号登录Databricks数据洞察控制台。
- 2. 在Dat abricks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, 单击**项目空间**。
- 在项目空间页面,单击待操作项目所在行的运行记录。下方列表展示当前项目下所有作业的运行情况。还可以通过条件搜索具体的作业。
 - 单击详情,可查看所选作业的详细信息,包括作业实例信息、提交日志、YARN容器列表、审计日志 和数据流监控。
 - 单击**启动**或停止,可启动或是停止所选的作业。

管理流式作业

- 1. 使用阿里云账号登录Databricks数据洞察控制台。
- 2. 在Dat abricks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,单击**项目空间**。
- 4. 在项目空间页面,单击待操作项目所在行的运行记录。
- 5. 单击上方的流式作业页签。
 - 单击**详情**,可查看所选Spark Streaming作业的详细信息,包括作业实例信息、提交日志、YARN容器 列表、审计日志和数据流监控。
 - 单击编辑,可编辑所选作业,详情请参见管理作业。
 - 单击**启动**或停止,可启动或是停止所选的作业。

查看审计日志

1. 使用阿里云账号登录Databricks数据洞察控制台。

- 2. 在Dat abricks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,单击**项目空间**。
- 4. 在项目空间页面,单击待操作项目所在行的运行记录。
- 5. 单击上方的审计日志页签。可以查看对当前项目空间执行的操作详情。