



数据湖分析 快速入门

文档版本: 20211117



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。	⚠ 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。
⑦ 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。	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,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 listinstanceid
[] 或者 [alb]	表示可选项,至多选择一个。	ipconfig [-all -t]
{} 或者 {alb}	表示必选项,至多选择一个。	switch {act ive st and}

目录

1.使用流程	05
2.通过元信息发现功能查询并分析OSS数据	06
3.DLA Presto CU版本快速入门	10
4.DLA Spark快速入门	13

1.使用流程

云原生数据湖分析DLA(Data Lake Analytics)是无服务器(Serverless)化的云上交互式查询分析服务,支持通过Presto和Spark引擎分析多种数据源中的数据。快速入门旨在介绍如何开通DLA、构建数据湖、调用 Presto和Spark引擎进行数据分析与计算,帮助您掌握DLA的基本使用流程。

如果您是首次使用云原生数据湖分析DLA的用户,我们建议您先阅读以下部分:

- 产品简介-本内容概述了云原生数据湖分析DLA的产品概念、产品优势及应用场景等内容。
- 产品定价-本内容介绍了云原生数据湖分析DLA的产品定价、计费方式等信息。

云原生数据湖分析DLA入门指南(本指南)-本指南提供了有关使用云原生数据湖分析DLA的基本流程。操作 流程概览如下:



- 1. 开通云原生数据湖分析服务
- (可选) 创建虚拟集群:CU版适用于查询频率高、查询数据量较大的场景,同时也能够给您使用DLA的费用预算带来一定的确定性。推荐您使用CU版本来进行数据分析与计算。

⑦ 说明 如果您使用系统默认的扫描量版本来进行数据分析与计算,则无需创建虚拟集群,可以 跳过此步骤。CU版与扫描量版的具体差异,请参见扫描量版本与CU版本的差异。

- 3. 构建数据湖: 您可以通过以下多种方式来构建数据湖。例如:
 - 手工上传文件到OSS,然后通过元数据爬取功能创建表来构建数据湖。具体操作请参见上传文件和OSS数据源。
 - 通过其他产品投递文件到OSS(如通过操作审计控制台投递日志文件到OSS),然后通过元数据爬取 功能创建表来构建数据湖。具体操作请参见创建单账号跟踪和OSS数据源。
 - 通过一键建仓和多仓合并建仓(仅支持数据库)、实时数据湖(支持数据库和消息日志)功能来构建 数据湖。具体操作请参见一键建仓多库合并建仓和实时数据湖。
 - ・ 连接数据源: 您也可以通过DLA直接连接其他数据源或者OSS来进行数据分析与计算。具体操作请参见Presto引擎连接数据源和Spark引擎连接数据源。
- 4. 数据分析与计算: 您可以调用Serverless Presto和Spark引擎来进行数据分析与计算。具体请参见Serverless Presto和Serverless Spark。
- 5. 数据应用: 您可以通过DataWorks和DMS来调度DLA Presto和DLA Spark任务,也可以将OSS数据的查询 分析结果以BI报表形式进行展示。具体请参见快速搭建Quick BI可视化报表。

2.通过元信息发现功能查询并分析OSS数 据

本文档将以通过DLA的元信息发现功能查询并分析OSS数据为例,帮助您快速掌握DLA的基本使用流程。

前提条件

已注册阿里云账号并完成实名认证。

⑦ 说明 如果您还没有创建阿里云账号,系统会在您开通云原生数据湖分析服务时提示您注册账号。

操作步骤

- 1. 开通云原生数据湖分析服务。
- 2. 登录OSS管理控制台,上传文件到OSS,具体操作请参见上传文件。

例如将supplier_with_header.csv文件上传到OSS的指定目录*oss://alibaba-crawler/schema1/supply_c*eshi/。

- 3. 登录数据湖分析管理控制台,在左侧导航栏单击数据湖管理 > 元信息发现。
- 4. 在元信息发现页面的OSS数据源区域,单击进入向导。

元信息友现				_					
Lishosshillin (OSS数据源 ① 自动为OSS存量及增量文件创建和更新数据减元数	R		进入向导
Tablestore設施業者の EARSN #100.00 通知たたMerroro 上級方法第1回席の目前に、方向分析的に詳。 EARSN									
任务列表 历》	と列表								_
全部类型	~								Ľ
D	名称	schema名称/前缀	调度计划	调度状态		最近运行状态	最近运行耗时	创建时间	操作
510	("oss://alibaba-crawler-muyuan/schema1/") -> schema Titest	schemaltiest	手动执行	开启调度	•	• SUCCESS	08	2021-01-19 16:32:47	执行 编辑 历史 删除
508	Selected sis logistores of this region -> sis202101182 2	sls2021011822	手动执行	开应调度	•	• SUCCESS	12719	2021-01-18 20:45:35	B(行) 编辑 25克 最佳
503	["oss://alibaba-crawler-muyuan/dw2/"] -> test882726	test882726	手动执行	开启调度		• FAILED 😗 🔃	08	2021-01-18 17:59:01	执行 编辑 历史 删除
502	["oss://alibaba-crawler-muyuar/supply-ceshi/"] -> test	100111111	E DAN G	1039	-		08	2021-01-18-17-42-41	10.75 IBM F.P. 800

5. 在OSS数据源页签的数据源配置、调度配置、目标元数据配置区域,根据实际需要进行参数配置。

SLS的OSS投递	数据源 OSS数据源 Tablestore数据源		
数据源配置	 數合模式 自由模式 为"基于OSS而构建的标准数据合库的场景构建自动化元 * OSS目录位置 	信息发现,识别精度高。OSS路径数据布局要求为何 	科表/文件·或者"库/表/分区//分区/文件"具体参考
	> 路径过濾规则 (可选)		
	格式解析器	CSV	~
	〉 配置选项 (可选)		
调度配置	调度频率	手动执行	~
目标元数据	配置		
	* Schema名称	您新建的元信息发现的名称 元数据发现基于采样,不能保证所有数据都采样到。	漏掉的字段可以通过 DDL手动添加
	◇ 配置选项 (可选)		
	文件字段变更规则	只增加列 OSS目录下面文件的字段改变,如何处理表更新	~
	对象删除变更规则	忽略删除更新 OSS对象删除,如何处理表更新	\checkmark

参数配置说明如下表所示:

参数	说明		
数仓模式和自由模式	 您可以选择数仓模式或自由模式: 数仓模式:为"基于OSS而构建的标准数据仓库"的场景构建自动化元信息发现,识别精度高。OSS路径数据布局要求为"库/表/文件"或者"库/表/分区//分区/文件"。 自由模式:为"探索OSS上的数据进行分析"的场景构建自动化元信息发现。对OSS数据布局没有要求,可能会产生差异化的表。 		
	文件在OSS中的存储地址,以/结尾。系统会根据您选择的文件夹路径,自动设置OSS路径。		
OSS目录位置	⑦ 说明 系统会自动拉取与DLA同地域的OSS Bucket,您可以根据业务需要从下拉列表中选择Bucket。选择Bucket后,系统会自动列出该Bucket下所有的Object和文件;选中目标Object和文件后,系统会自动将其添加到右侧的OSS路径处。		
格式解析器	默认自动解析,即按照顺序调用所有内置解析器,也可指定特定文件类型的格式解析器,比如json、parquet、avro、orc、csv。		
调度频率	您可以根据需要定期计划运行元信息发现任务。		

参数	说明
Schema名称	设置Schema名称,即映射到DLA中的数据库名称(默认每个发现任务会新 创建一个独立的Schema)。
配置选项	高级自定义设置项,如字段分隔符、引用标识、表头模式、允许单个列字 段等。

6. 配置完成后,单击创建。

元信息发现任务创建成功后,在**任务列表**中您将能看到创建成功的任务信息。元信息发现任务将根据您 设置的**调度频率**,需要您手动执行或者定期自动调度该任务。

任务列表	历史列表
	-

						REBY
10	schema名称/前缀	任务名:	現行状态	预期运行时间	最后更新时间	操作
111010	schema1_test	["oss://albaba-crawler/schema1/"] -> schema1_test	• SUCCESS @	2021-01-19 16:45:45	2021-01-19 16:45:49	itti (#28 #28
110970	tablestore20201231	all ots instances of this region -> tablestore20201231	• SUCCESS @	2021-01-19 16:00:00	2021-01-19 16:00:41	ittin 1828 1820
110980	tablestore20201231	all ots instances of this region -> tablestore20201231	• SUCCESS @	2021-01-19 16:00:00	2021-01-19 16:01:17	itis B28 B28
110911	tablestore20201231	all ots instances of this region -> tablestore20201231	• SUCCESS	2021-01-19 15:00:00	2021-01-19 15:00:44	1915 M28 M28
110912	tablestore20201231	all ots instances of this region -> tablestore20201231	• SUCCESS	2021-01-19 15:00:00	2021-01-19 15:01:19	1715 I M28 I MID:
110852	tablestore20201231	all ots instances of this region -> tablestore20201231	• SUCCESS	2021-01-19 14:00:00	2021-01-19 14:00:42	(F15 1228 HIP):

元信息发现任务执行成功后,单击schema名称/前缀列下面的数据库名称链接(如单击alibaba), 跳转到Serverless Presto > SQL执行页面。您可以看到DLA自动发现创建成功的库、表、列信息。

SQL执行

schema1_test 🛛 🖸	SQL执行集群 请:	选择 🗸 同步执行(F8)	异步执行(F9) 格式化(F10)	主题 ~	
"双击"切换Schema	1 select * f	rom `schema1_test`.`supply_ceshi	<pre>` limit 20;</pre>		
✓ S schema1_test (current)	2				
∨ 🖩 supply_ceshi					
s_suppkey					
I∎ s_name					
s_address					
s_nationkey					
I s_phone					
s_acctbal					
s_comment					
	执行历史	执行结果 SQL监控 i	s.name	s. address	s_nationkey
	1	1	Supplier#00000001	N kD4on9OM lpw3,qf0JBoQDd7tgrzr	17
	2	2	Supplier#00000002	89eJ5ksX3ImxJQBvxObC,	5
Ţ.	3	3	Supplier#00000003	q1,G3Pj6OjluUYfUoH18BFTKP5aU9b	1
	4	4	Supplier#00000004	Bk7ah4CK8SYQTepEmvMkkgMwg	15
	5	5	Supplier#00000005	Gcdm2rJRzI5qITVzc	11
	6	6	Supplier#00000006	tQxuVm7s7CnK	14
	7	7	Supplier#00000007	s,4TicNGB4uO6PaSqNBUq	23
	8	8	Supplier#00000008	9Sq4bBH2FQEmaFOocY45sRTxo6yu	17
	9	9	Supplier#00000009	1KhUgZegwM3ua7dsYmekYBsK	10

7. 在Serverless Presto > SQL执行页面编写SQL语句,单击同步执行或者异步执行,执行SQL语句。
 例如在schema1_test下执行 select * from `schema1_test`.` supply_ceshi` limit 20; 。

SQL执行									通法手册	前数手册
schema1_test 0 C	SQL执行集群 语	an v Nobrie	8) 异步讯行(F9) 格式化(F10)	±m ×				#i#SOL@R/#J	# () #	BROMSBRITSQL
"双击"切换Schema	select *	fron `schenal_test`.`supply_ces	hi' limit 20;							
Schema1_test (current)										
🗸 🔳 supply_ceshi										
s_supplicity										
I s_name										
s_address										
s nhone										
s accibal										
s.comment										
	执行历史	RITIER SQLIDE							导出结果集	~ BR
	19-9	s_suppkey	s_name	s_address	s_nationkey	s_phone	s_acctbal	s_comment		detail
<	1	1	Supplier#000000001	N kD4on9OM lpw3,gf0JBoQDd7tgrzr	17	27-918-335-1736	5755.94	each siyly above the careful		1718
P	2	2	Supplier#000000002	89eJ5ksX3lmsJQBvxObC,	5	15-879-861-2259	4032.68	slyly bold instructions, idle d	ependen	1818
	3	3	Supplier#00000003	q1,G3PJ60JIuUYYUoH188FTKP5aU9b	1	11-383-516-1199	4192.40	bithely silent requests after	the expre	1918
	4	4	Supplier#000000004	Bk7ah4CK8SYQTepEmvMikkgWwg	15	25+843+787+7479	4641.08	riously even requests above	the exp	1218
	5	5	Supplier#000000005	Godm2rJRzi5d/TVzc	11	21-151-690-3663	-283.84	. slyty regular pinto bea		1918
	6	6	Supplier#000000006	tQxuVm7s7CnK	14	24-696-997-4969	1365.79	final accounts, regular dolph	ins use a	1918
	7	7	Supplier#00000007	s,4TicNGB4uO6PaSqNBUq	23	33-990-965-2201	6820.35	s unwind silently furiously re-	gular cou	1218
	8	8	Supplier#00000008	9Sq4bBH2FQEmaFOocY45sRTxx8yu	17	27-498-742-3860	7627.85	al pinto beans, asymptotes h	lege	19 ta

您可以在**执行结果**中,查看DLA从OSS目录*oss://alibaba-crawler/schema1/supply_ceshi/*下的*supplie r_with_header.csv*文件中自动发现的数据信息。

3.DLA Presto CU版本快速入门

本文主要教您如何快速上手阿里云云原生数据湖分析DLA Presto CU版本。

操作步骤

1. 创建虚拟集群

与Serverless Presto扫描版不同,在CU版本下执行SQL前,必须要创建一个虚拟集群。具体请参见虚拟集 群管理。

⑦ 说明 创建虚拟集群时,选项引擎选择Presto。

2. 配置数据源网络

⑦ 说明 如果需要连接您VPC内的数据源(如VPC内的RDS、AnalyticDB等),您需要配置数据源 网络,如不需要连接,忽略此步骤即可。

i. 在配置数据源网络前, 您需要授予DLA账户访问您VPC相关API的权限, 详细操作步骤请参见配置数据源网络。

ii. 在虚拟集群管理页面单击**详情**,进入集群详情页面。

┃虚拟集群管理	虚拟集群管理 新建直线集群						
() 目前该Region Serv	😝 目前读Region Serverless Spark已经感业化,"实例ID"为空的集群为公测集群,后续即将自动下线,请勿使用,如果继续使用,请创建新的"虚拟集群"						
集群名称	集群类型	实例ID	运行状态	创建时间小	擬作		
test	SPARK		日停止	2020-05-18 11:15	洋情 升配 删除		
abc	SPARK		已停止	2020-05-25 20:50	详情 升配 删除		
sads	SPARK		已停止	2020-06-29 11:15	详情 升配 删除		
tset	SPARK	d	运行中	2020-09-07 09:30	洋情 升配 删除		
< test-00	SPARK	c	运行中	2020-09-07 09:35	洋情 升配 删除		

iii. 单击新增数据源网络,选择您想要连接的数据源对应的虚拟网络、交换机Id和安全组Id。

新增数据源网络			×
虚拟网络:	请选择虚拟网络:	\checkmark	
交换机ld	请选择虚拟交换机:	\checkmark	
安全组ld	请选择安全组:	~	

■ 交换机ID可以在您VPC内数据源的基本信息页面获取,以RDS为例。

1000-20	男 伊令美丽的女	体 建油水中的		立即并分		S. ANALIN		geng	14		X10 PM		ц.	= '
10007098	重代业务等标米	仍,相互成功的统约	即愿奴据件上云伏束!.	77.801#01±										
云数据库	管理						⑦ RDS简介	数据导入	待处理事件	学习路径	登录数据库	性能大	盘	C H
基本信息	标签信息	高性能版												
现5.7单机基	础版升级到5.77	5可用版、且高可用	用版本地SSD盘通用型规	机格升级到独享型规林	各均5折优惠 🗙									
实例ID/名	称 🖌 🗄	输入实例名称或实)	例ID进行搜索	搜索	示签									
□ 实例	ID/名称		运行状态(全 部) 、	创建时间	实例类型(全 部) ▼	数据库类型 (全部) ▼	所在可用区	网络类型 (全部)	-	付费类	2	标签		
	n vc-test 🖌	1,008-40	运行中	2020-08-24 21:30:39	常规实例	MySQL 8.0	华东1(杭州) 可用区H	VSwitch: 71 (VPC	tre : vpc-	按量付款	1	1	注理 性	生能 羊

■ 安全组ID可以登录VPC控制台获取。

			Q 108236.19	H台、API、MI统方都和普通	费用	工单 音楽	企业 支持	官同	5.	۵. ۱	₹ ®	操体	0
	6.5.5 ♥				推送	- 1891							
默认	专有网络 吉				ClassicLink	未开启							
加入元金	让网律情 尚未加入云企业网				地域	19:351 (80H)							
	资源组 默认资源组				拥有者	出的账户							
路由器基本信息													
	ID vrt-bp1d7k5luinegae5	wpv)r			名称	- 10.50							
	別題时间 2020年8月24日 21:29	27			描述	- 1840							
資源管理 网段管	理 云企业网韵账号授权	2											6PI
基础云资源													
ECS采例	D	RDS定例	1										88
网络资源							_						_
路由表	1	交换机	1	NAT网关		0	安全	6				1	
		2. 4 0 1					<u> </u>						

⑦ 说明 这里安全组可能会有多个,选择一个可以访问您数据源的安全组即可。

3. 执行SQL

在CU版本下,当您创建好虚拟集群,首次开始执行SQL时,您会发现系统已经默认自动生成一个公共数据集。

	(杭州) 👻		2 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 🗔 🛕
数据湖分析	SQL执行	U 3	0.000		Γ
概览					
数据湖管理 HOT へ	public_dataset_tpch 🕲 C 同步执行	F8) 异步执行(F9) 格式化(F10)	主题 🖌		(i) 🔤 🖈
元数据管理	"双击"切换Schema 1 /*+pool	-selector=dladw*/SELECT * FROM	`nation` LIMIT 20;		
元数据爬取	public_dataset_tpch_1x_text (c > m_customer				
数据入湖	> 🆩 lineitem				
实时数据湖 New	> mation				
Serverless SQL	> morders				
sou this	> III partsupp				
SQCD/11	> III region				
SQL监控	> m supplier 执行历史	执行结果 SQL监控 🚺			令出
Serverless Spark New ^	序号	n_nationkey	n_name	n_regionkey	n_comment
作业管理	1	0	ALGERIA	0	haggle. carefully final deposits
独寡版 Spark New へ	2	1	ARGENTINA	1	al foxes promise slyly accordin
集群列表	3	2	BRAZIL	1	y alongside of the pending dep
数据工作台 ~	4	3	CANADA	1	eas hang ironic, silent package
	5	4	EGYPT	4	y above the carefully unusual t

在左边列表中选中public_dataset_xxx这个库,并且执行SQL语句,例如:

/*+cluster=dladw*/SELECT * FROM `nation` LIMIT 20;

⑦ 说明 cluster 是您之前创建的虚拟集群实例名称。

更多关于DLA SQL用法的详细操作,请参见常用SQL。

4.DLA Spark快速入门

熟悉Spark的开发者都了解SparkPi,它相当于Spark引擎的"Hello World!"。本文介绍如何在DLA控制台跑通SparkPi。

准备事项

1. 您需要在提交作业之前首先创建虚拟集群,具体操作请参考创建虚拟集群。

② 说明 创建虚拟集群时注意选择引擎类型为Spark。

2. 如果您是子账号登录,需要配置子账号提交作业的权限,具体操作请参考快速配置子账号权限。

操作步骤

- 1. 登录云原生数据湖分析管理控制台。
- 2. 在概览页面的左上角,选择虚拟集群所在地域。
- 3. 单击左侧导航栏的Serverless Spark > 作业管理。
- 4. 在作业编辑页面,单击创建作业模板,填写以下作业信息:

创建作业模板		×
文件名称		
文件类型	文件	1
父级	请选择 人名英格兰人名英格兰人名英格兰人名英格兰人名英格兰人名英格兰人名英格兰人名英格兰	*
作业类型	○ SparkJob ○ SparkSQL	
		确定 取消

5. 新创建的作业中包含了SparkPi作业的默认配置,在作业编辑页面,单击执行即可。

⑦ 说明 关于作业提交的详细说明,请参见创建和执行Spark作业。