阿里云 消息队列 Kafka

生态对接

文档版本: 20190914

为了无法计算的价值 | [] 阿里云

<u>法律声明</u>

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。
- 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
•	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。	禁止: 重置操作将丢失用户配置数据。
A	该类警示信息可能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,恢复业务所需 时间约10分钟。
Ê	用于补充说明、最佳实践、窍门等,不 是用户必须了解的内容。	道 说明: 您也可以通过按Ctrl + A选中全部文件。
>	多级菜单递进。	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。	单击 确定。
courier 字体	命令。	执行 cd /d C:/windows 命令,进 入Windows系统文件夹。
##	表示参数、变量。	bae log listinstanceid Instance_ID
[]或者[a b]	表示可选项,至多选择一个。	ipconfig [-all -t]
	表示必选项,至多选择一个。	<pre>swich {stand slave}</pre>

目录

汐	、律声明		•••••			I
迌	通用约定		•••••			I
1	消息队列 for A	Apach	ne Kafka 招	蒈配云 HE	Base 和 Spark 构建一	体化数
	据处理平台		• • • • • • • • • • • • • • •			1
2	使用 DTS 将数	据库	数据同步至	消息队列	for Apache Kafka	
3	将消息队列	for	Apache	Kafka	数据迁移至大数据计	算服务
	MaxComput	e	•••••	• • • • • • • • • • • • •	••••••	5

1 消息队列 for Apache Kafka 搭配云 HBase 和 Spark 构建一体化数据处理平台

云 HBase X-Pack 是基于 Apache HBase、Phoenix、Spark 深度扩展,融合 Solr 检索等技术,支持海量数据的一站式存储、检索与分析。融合云 Kafka + 云 HBase X-Pack 能够构建一体化的数据处理平台,支持风控、推荐、检索、画像、社交、物联网、时空、表单查询、离线数仓等场景,助力企业数据智能化。

下图是业界广泛应用的大数据中台架构,其中 HBase 和 Spark 选择云 HBase X-Pack。产品详情 请参见 X-pack Spark 分析引擎立即购买>>



- · 消息流入: Flume、Logstash 或者在线库的 Binlog 流入消息中间件 Kafka。
- · 实时计算:通过 X-Pack Spark Streaming 实时的消费 Kafka 的消息,写入到云 HBase 中对 外提供在线查询。
- · 实时存储与检索: 云 HBase 融合 Solr 以及 Phoenix SQL 层能够提供海量的实时存储,以及 在线查询检索。
- ・批处理、数仓及算法:在线存储 HBase 的数据可以自动归档到 X-Pack Spark 数仓。全量数据 沉淀到 Spark 数仓(HiveMeta),做批处理、算法分析等复杂计算,结果回流到在线库对外 提供查询。

该套方案的实践操作请参见 Spark 对接 Kafka 快速入门。同时,有云 HBase 和 Spark 的示例代 码请参见 Demo。

2 使用 DTS 将数据库数据同步至消息队列 for Apache Kafka

使用数据传输服务 DTS(Data Transmission Service)的数据同步功能,您可以将通过专线/ VPN 网关/智能网关接入的数据库的数据同步至消息队列 for Apache Kafka 集群,扩展消息处理 能力。

立即购买消息队列 for Apache Kafka>>

具体的前提条件、注意事项、操作步骤等信息,请参见#unique_5。

当您操作到#unique_5/unique_5_Connect_42_section_v5h_m5c_zgb时,请注意以下事项:

字段	说明
实例类型	选择通过专线/VPN网关/智能网关接入的自建数 据库。
对端专有网络	选择消息队列 for Apache Kafka 实例的 VPC 。
IP 地址	选择消息队列 for Apache Kafka 实例接入点 的任意一个 IP 地址即可,目前仅仅支持填写单 个 IP 地址。
端口	选择消息队列 for Apache Kafka 实例接入点 对应 IP 的对应端口。
数据库账号	非必填项。
数据库密码	非必填项。
Торіс	获取 Topic 列表后选择对应 Topic,Topic 建 议创建单个分区,以便保证全局顺序。
Kafka版本	选择消息队列 for Apache Kafka 实例对应的 开源版本,目前主要是 0.10 版本。



说明:

当前 DTS 仅支持通过默认接入点导入数据到消息队列 for Apache

Kafka, 假设消息队列 for Apache Kafka 实例的默认接入点为

"172.16.X.X1:9092,172.16.X.X2:9092,172.16.X.X3:9092",选择第一个 IP 地址和端口填 写即可,也即上表中的 IP地址填写"172.16.X.X1",端口填写"9092"。

配置结果如下图所示。

目标实例信息			
实例类型:	通过专线/VPN网关/智能网关接入的自建数据库		
实例地区:	华东1(杭州)		
* 对端专有网络:	vpc-bp17xemgs4hf4lm		
数据库类型:	Kafka \$		
* IP地址:	172.16.0		
• 端口:	9092		
数据库账号:		非必填项	
数据库密码:	م	非必填项	
* Topic:	333 \$	获取Topic列表	
	请先点击右侧按钮,获取Topic列表后选择具体的Topic		
* Kafka版本	0.10 \$		

3 将消息队列 for Apache Kafka 数据迁移至大数据计 算服务 MaxCompute

本文介绍如何使用 DataWorks 数据同步功能,将消息队列 for Apache Kafka 集群上的数据迁移 至阿里云大数据计算服务 MaxCompute,方便您对离线数据进行分析加工。

前提条件

在开始本教程前,确保您已完成以下操作:

- ・ 确保消息队列 for Apache Kafka 集群运行正常。本文以部署在华东1(杭州)地域(Region)的集群为例。
- · 开通 MaxCompute。
- 开通 DataWorks。
- · #unique_8。本文以在华东1(杭州)地域创建名为 bigdata_DOC 的项目为例。

示例如下。

	概览 项目列表	调度资源列表
Go DataWorks 数	据集成・数据开发・MaxCompute	
快速入口		
数据开发	数据集成	运维中心
项目		全部项目
bigdata_DOC 华东1	MaxCompute_DOC 华东2	PAltest 华东2
创建时间:2018-09-02 10:26:59 计算引擎:MaxCompute 服务模块数据开发数据集成数据管理 运维中心	创建时间:2018-07-19 09:12:37 计算引擎:MaxCompute 服务模块.数据开发 数据集成 数据管理 运维中心	创建时间:2018-05-23 13:32:29 计算引擎:MaxCompute PAI计算引擎 服务模块数据开发数据集成数据管理运维中心
项目配置 进入数据开发 进入数据集成	项目配置 进入数据开发 进入数据集成	项目配置 进入数据开发 进入数据集成
常用功能 分 创建项目 × 一键CDN		

背景信息

大数据计算服务 MaxCompute(原 ODPS)是一种大数据计算服务,能提供快速、完全托管免运 维的 EB 级云数据仓库解决方案。

DataWorks 是基于 MaxCompute 计算和存储,提供工作流可视化开发、调度运维托管的一站式海量数据离线加工分析平台。在数加(一站式大数据平台)中,DataWorks 控制台即为

MaxCompute 控制台。MaxCompute 和 DataWorks 一起向用户提供完善的 ETL 和数仓管理能力,以及 SQL、MR、Graph 等多种经典的分布式计算模型,能够更快速地解决用户海量数据计算问题,有效降低企业成本,保障数据安全。

本教程旨在帮助您使用 DataWorks,将消息队列 for Apache Kafka 中的数据导入至 MaxCompute,来进一步探索大数据的价值。

- 步骤一: 准备 Kafka 数据
 - 登录消息队列 for Apache Kafka 控制台创建 Topic 和 Consumer Group, 分别命名为 testkafka 和 console-consumer。具体步骤参见#unique_9。本示例中, Consumer Group console-consumer 将用于消费 Topic testkafka 中的数据。
 - 2. 向 Topic testkafka 中写入数据。由于 Kafka 用于处理流式数据,您可以持续不断地向其中 写入数据。为保证测试结果,建议您写入 10 条以上的数据。您可以直接在控制台使用发送消息 功能来写入数据,也可以使用消息队列 for Apache Kafka 的 SDK 收发消息。详情参见使用 SDK 收发消息。
 - 3. 为验证写入数据生效,您可以在控制台#unique_11,看到之前写入 Topic 中的数据。

步骤二: 创建 DataWorks 表

您需创建 DataWorks 表,以保证大数据计算服务 MaxCompute 可以顺利接收消息队列 for Apache Kafka 数据。本例中为测试便利,使用非分区表。

1. 登录 DataWorks 控制台,在工作空间区域,单击目标工作空间的进入数据开发。

2. 在左侧导航栏单击表管理, 然后单击新建图标。



- 3. 在新建表对话框,输入表名 testkafka,然后单击提交。
- 4. 在创建的表页面,单击 DDL模式。
- 5. 在 DDL模式对话框, 输入以下建表语句, 单击生成表结构。

```
CREATE TABLE `testkafka` (

`key` string,

`value` string,

`partition1` string,

`timestamp1` string,

`offset` string,

`t123` string,

`event_id` string,

`tag` string
```

);

建表语句中的每一列对应 DataWorks 数据集成 Kafka Reader 的默认列。

- ・key:表示消息的 Key。
- · value:表示消息的完整内容。
- · partition: 表示当前消息所在分区。
- ・headers:表示当前消息 headers 信息。
- · offset: 表示当前消息的偏移量。
- · timestamp: 表示当前消息的时间戳。

您还可以自主命名,详情参见配置 Kafka Reader。

6. 单击提交到生产环境。

详情请参见#unique_13。

步骤三:同步数据

- 1. #unique_14。此处创建的 ECS 实例将用以完成数据同步任务。
- 2. 登录 DataWorks 控制台,在工作空间区域,单击目标工作空间的进入数据开发。
- 3. 在左侧导航栏,选择数据开发 > 业务流程 > 数据迁移。
- 4. 右键选择数据集成 > 新建数据集成节点 > 数据同步。
- 5. 在新建节点对话框,输入节点名称(即数据同步任务名称),然后单击提交。
- 6. 在创建的节点页面,选择数据来源的数据源为 Kafka,选择数据去向的数据源为ODPS,选择 您在步骤二:创建 DataWorks 表中创建的表。完成上述配置后,请单击框中的按钮,转换为脚 本模式,如下图所示。

Di kafka2mc 🏾										
	L 🔍 🔂 🖉	3								
01 选择数据源	10 10	据来源			数据去向			收起		
	α ×	里配直数据的来源端机与入端;	可以是默认的数据源,也可以	以是您创建的目有	数据源宣有支持的数据来源					
* 数据源	Kafka 🗸	选择据源库 >	0	• 数据源	odps ~	odps_first	× (?			
				*表	testkafka		~			
止数据源不 点击转换为	交持向导模式,需要使用脚本 3脚本	\$模式配置同步任务,		分区信息	无分区信息					
				清理规则 写入前清理已有数据 (Insert Overwrite) >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>						
				字符串作为null(● 是 💿 否					

7. 配置脚本,示例如下。

```
{
"type": "job",
"steps": [
{
```

```
"stepType": "kafka",
           "parameter": {
    "server": "47.xxx.xxx.xxx:9092",
                 "kafkaConfig": {
    "group.id": "console-consumer"
                 },
"valueType": "ByteArray",
                 "column": [
"__key__",
"__value__",
"__partition__",
"__timestamp__",
"__offset__",
                       "'123'",
"event_id",
"tag.desc"
                 ],
"topic": "testkafka",
                 "keyType": "ByteArray",
"waitTime": "10",
                 "beginOffset": "0",
                 "endOffset": "3"
           },
"name": "Reader",
"read
           "category": "reader"
     },
{
           "stepType": "odps",
"parameter": {
                 "partition": "",
                 "truncate": true,
"compress": false,
                 "datasource": "odps_first",
                 "column": [
                       "key",
                       "value"
                       "partition1",
                       "timestamp1",
                       "offset",
                       "t123",
"event_id",
                       "tag"
                 ],
"emptyAsNull": false,
"tastkafka"
                 "table": "testkafka"
           },
           "name": "Writer",
           "category": "writer"
     }
],
"version": "2.0",
"order": {
     "hops": [
           {
                 "from": "Reader",
                 "to": "Writer"
           }
     ]
},
"setting": {
"arrorLin
      "errorLimit": {
"record": ""
     },
      "speed": {
```

```
"throttle": false,
"concurrent": 1
}
}
```

8. 在脚本页面,单击配置任务资源组,选择步骤1中创建的自定义资源组,然后单击运行图标。

Di kafka2	2mc	Di test	testkafka	test12	234	Sq test	Di test123	📄 数据集	眬			¢	, ≡
	•	ሰ 🛯	• 🔒 🖬	8									运维↗
1 2	2 "typ	e": "job",			▲			×			配置任务资源组	帮助文档	调度配
3 4 5	ste	ps : [{ stepTy	rpe": "kafka",						任务资源组	默认资源组		וו	Ē
6 7		"parame	ter": { rver": "localho	st:9093",				1		✔ 默认资源组 kafka2mc			版本

预期结果

完成运行后,运行日志中显示运行成功。

运行日志	
读写失败总数	: 0
2019-05-29 11:10:2	24 INFO
2019-05-29 11:10:2	24 INFO Exit code of the Shell command 0
2019-05-29 11:10:2	24 INFO Invocation of Shell command completed
2019-05-29 11:10:2	24 INFO Shell run successfully!
2019-05-29 11:10:2	24 INFO Current task status: FINISH
2019-05-29 11:10:2	24 INFO Cost time is: 114.983s
/home/admin/alisa	tasknode/taskinfo//20190529/diide/11/08/28/mjuicuxu5slfbv3xu7m8csqy/T3_0242504015.log-END-EOF
Exit with SUCCESS	
2019-05-29 11:10:2	28 [INFO] Sandbox context cleanup temp file success.
2019-05-29 11:10:2	28 [INFO] Data synchronization ended with return code: [0].
2019-05-29 11:10:2	28 INFO
2019-05-29 11:10:	28 INFO Exit code of the Shell command 0

后续步骤

您可以新建一个数据开发任务运行 SQL 语句,查看当前表中是否已存在从 Kafka 同步过的数据。 本文以select * from testkafka为例,具体步骤如下:

- 1. 在左侧导航栏,选择数据开发>业务流程。
- 2. 右键选择数据开发 > 新建数据开发节点 > ODPS SQL。
- 3. 在新建节点对话框,输入节点名称,然后单击提交。
- 4. 在创建的节点页面,输入select * from testkafka,然后单击运行图标。

6	5	select *	from	testka	afka;																
运	行E	志	绢	課[2]	×																
	4																				_
_	kar	A		lunkun	В		nontition1	C	D timostomo1		offeet	E		•100	F	auant id	G		•	H	
1	key	/	~	value 100		× 1	paruuon i o	~	1550100459609	~	onset		~	122	~			~	tag M		~
2	VN VN			123			ა ი		1559100458098		0			123							
3				234			9		1559100458028		0			123					\N		_
4	١N			567			0		1559100466891		0			123		\N			\N		
5	\N			123			7		1559050808437		0			123		\N			\N		
6	\N			567			1		1559100457401		1			123		\N			N/N		