DataHub

产品简介

产品简介

产品概述

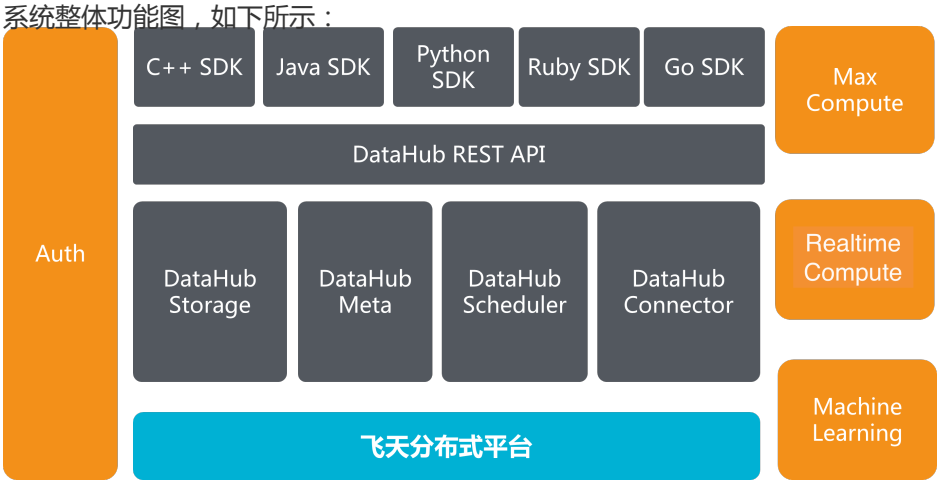
DataHub基本介绍

阿里云流数据处理平台DataHub是流式数据（Streaming Data）的处理平台，提供对流式数据的发布（Publish），订阅（Subscribe）和分发功能，让您可以轻松构建基于流式数据的分析和应用。DataHub服务可以对各种移动设备，应用软件，网站服务，传感器等产生的大量流式数据进行持续不断的采集，存储和处理。您可以编写应用程序或者使用实时计算引擎来处理写入到DataHub的流式数据，比如：实时web访问日志、应用日志、各种事件等，并产出各种实时的数据处理结果，比如：实时图表、报警信息、实时统计等。

DataHub服务基于阿里云自研的飞天平台，具有高可用，低延迟，高可扩展，高吞吐的特点。DataHub与阿里云实时计算引擎Realtime Compute无缝连接，您可以轻松使用SQL进行流数据分析。

DataHub服务也提供分发流式数据到各种云产品的功能，目前支持分发到MaxCompute（原ODPS），OSS等。

系统整体功能图，如下所示：



产品优势

高吞吐

最高支持单主题（Topic）每日T级别的数据量写入，每个分片（Shard）支持最高每日8000万Record级别的写入量。

实时性

通过DataHub ，您可以实时的收集各种方式生成的数据并进行实时的处理，对您的业务产生快速的响应。

易用性

- DataHub提供丰富的SDK包，包括C++，Java，Pyhon，Ruby，Go等语言；
- DataHub服务也提供Restful API规范，您可以用自己的方式实现访问接口。
- 除了SDK以外，DataHub还提供一些常用的客户端插件，包括：Fluentd，LogStash，Flume等，您可以使用这些客户端工具往DataHub中写入流式数据。
- DataHub同时支持强Schema的结构化数据和无类型的非结构化数据，您可以自由选择。

高可用

- 服务可用性不低于99.999%。
- 规模自动扩展，不影响对外服务。
- 数据持久性不低于99.999%。
- 数据自动多重冗余备份。

动态伸缩

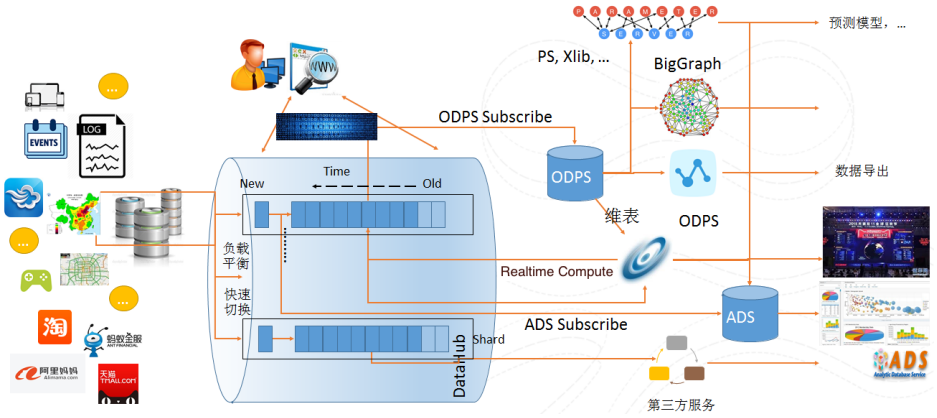
每个主题（Topic）的数据流吞吐能力可以动态扩展和减少，最高可达到每主题256000 Records/s的吞吐量。

高安全性

- 提供企业级多层次安全防护，多用户资源隔离机制。
- 提供多种鉴权和授权机制及白名单、主子账号功能。

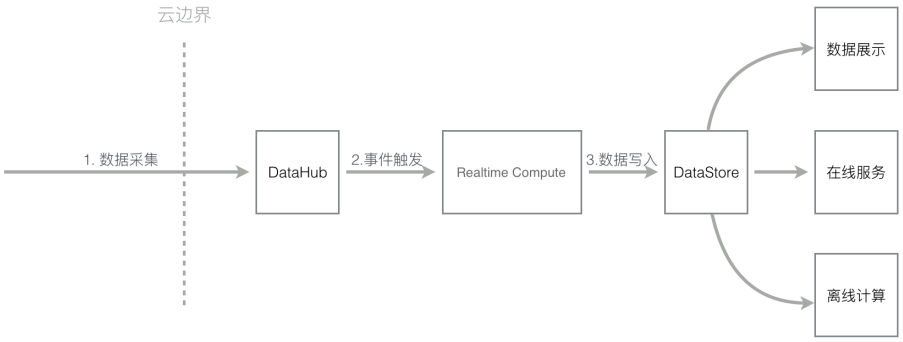
应用场景

DataHubR作为一个流式数据处理服务，结合阿里云众多云产品，可以构建一站式的数据处理服务。



实时计算Realtime Compute

Realtime Compute是阿里云提供的实时计算引擎，提供使用类SQL的语言来进行流式计算。DataHub和Realtime Compute无缝结合，可以作为Realtime Compute的数据源和输出源。



流处理应用

您可以编写应用订阅DataHub中的数据，并进行实时的加工，把加工后的结果输出。

您可以把应用计算产生的结果输出到DataHub中，并使用另外一个应用来处理上一个应用生成的流式数据，来构建数据处理流程的DAG。

流式数据归档

您的流式数据可以归档到MaxCompute（原ODPS）中，通过创建DataHub Connector，指定相关配置，即可创建将Datahub中流式数据定期归档的同步任务。

提示（新用户无需关注）：

- 目前 老版本MaxCompute DataHub已处于待下线状态，不再接入新用户。DataHub用户请参看使用文档。
- 自2016年11月21日起，新版本DataHub正式公测上线。
- 新老DataHub迁移手册。

公测约束与说明

DataHub免费公测期间，资源有限，将会不定时执行下述回收策略

- 凡15天无数据写入的Topic将有可能被系统临时关闭通道，再次使用需要重新申请资源。
- 凡15天未无数据同步的DataConnector将有可能会被系统临时暂停任务，再次使用需要重启任务。

基本概念

名词解释

名词	解释
Project	项目（Project）是DataHub数据的基本组织单元,下面包含多个Topic。值得注意的是，DataHub的项目空间与MaxCompute的项目空间是相互独立的。用户在MaxCompute中创建的项目不能复用于DataHub，需要独立创建。
Topic	Topic是 DataHub 订阅和发布的最小单位，用户可以用Topic来表示一类或者一种流数据。 更多详情请参考：Project及Topic数量限制。
Topic Lifecycle	表示一个Topic中写入数据在系统中可以保存的最长时间，以天为单位，最小值为1，最大值为7
Shard	Shard表示对一个Topic进行数据传输的并发通道，每个Shard会有对应的ID。每个Shard会有多种状态：Opening - 启动中，Active - 启动完成可服务。每个Shard启用以后会占用一定的服务端资源，建议按需申请Shard数量。
Shard Hash Key Range	每个Shard都有的属性，包括开始和结束的Key范围，写入数据的时候具有相同Key的数据会落到同一个Shard上。对一个Shard的Key范围是左闭右开。更多详情请参考：根据HashKey写入数据。
Shard Merge	Shard合并，可以把相邻的Key Range连接的Shard merge成一个Shard。更多详情请参考：Shard扩容缩容。
Shard Split	Shard分裂，可以把一个Shard分裂成Shard Key Range相连接的两个Shard
Record	用户数据和 DataHub 端交互的基本单位
RecordType	Topic的数据类型，目前支持Tuple与Blob两种类型。Tuple类型的Topic支持类似于数据库的记录的数据，每条记录包含多个列。Blob类型的

	Topic仅支持写入一块二进制数据。

数据类型介绍

- Tuple类型下只支持写入数据是有格式的数据，支持以下几种数据类型

类型	含义	值域
Bigint	8字节有符号整型。请不要使用整型的最小值 (-9223372036854775808)，这是系统保留值。	-9223372036854775807 ~ 9223372036854775807
String	字符串，只支持UTF-8编码。	单个String列最长允许1MB。
Boolean	布尔型。	可以表示为 True/False，true/false, 0/1
Double	8字节双精度浮点数。	-1.0 *10308 ~ 1.0* 10308
TimeStamp	时间戳类型	表示到**微秒**的时间戳类型

- Blob模式下支持写入一块二进制数据作为一个Record，数据将会以BASE64编码传输。

Shard状态说明

状态	说明
Opening	Topic刚创建，所有shard会处于Opening状态直至准备完成。不可读写。
Active	Shard通道打开后，状态会置为Active，此时表示Shard正常可读写。
Closing	Shard进行了Split/Merge操作，后台正在关闭该通道。该状态Shard不可读写。
Closed	Shard在Split/Merge完成后，会变为Closed态，此时Shard为只读状态。

异常描述

ErrorCode	HttpCode	含义
InvalidUriSpec	400	请求的Uri非法
InvalidParameter	400	参数错误，详细内容请看返回的ErrorMessga
Unauthorized	401	签名错误
NoPermission	403	账号权限不足

InvalidSchema	400	Schema格式错误
InvalidCursor	400	无效或过期的cursor
NoSuchProject	404	请求的Project不存在
NoSuchTopic	404	请求的Topic不存在
NoSuchShard	404	请求的ShardID不存在
ProjectAlreadyExist	400	Project已存在
TopicAlreadyExist	400	Topic已存在
InvalidShardOperation	405	非法Shard操作，如Shard已经Closed后继续写入.
LimitExceeded	400	请求参数超出限制，如Shard总数超过512个.
InternalServerError	500	未知错误或内部服务异常或系统处于升级中.

限制描述

限制描述

限制项	描述	值域范围
活跃shard数	每个topic中活跃shard数量限制	(0,10]（公测限制，流量超出10个Shard承载能力请联系管理员提升Quota）
总shard数	每个topic中总shard数量限制	(0,512]
Http BodySize	http请求中body大小限制	4MB
单个String长度	数据中单个String字段长度限制	1MB
Merge/Split频率限制	每个新产生的shard在一定时间内不允许进行Merge/Split操作	5s
QPS限制	每个Shard写入QPS限制(非Record/s，Batch写入同一Shard仅计算为1次)	1000
Throughput限制	每个Shard写入每秒吞吐限制	1MB
Project限制	每个云账号能够创建的Project上限	5
Topic限制	每个Project内能创建的Topic数	20

	量限制，如有特殊请求请联系管理员	
Topic Lifecycle限制	每个Topic中数据保存的最大时长，单位是天	[1,7]

命名规范

名词	描述	长度限制	值
Project	项目名称	[3,32]	英文字母开头，仅允许英文字母、数字及“_”，大小写不敏感。
Topic	主题名称	[1,128]	英文字母开头，仅允许英文字母、数字及“_”，大小写不敏感。

访问控制

DataHub采用阿里云RAM进行访问控制。用户对DataHub资源的访问，通过RAM进行鉴权。阿里云主账号拥有所属资源的所有权限，子用户在创建时并没有任何权限，不能访问任何资源，用户需要在RAM中对该子用户进行授权操作。关于如何创建RAM子用户与创建授权策略并进行授权可参见RAM使用文档。以下将介绍DataHub在RAM下的访问控制体系。

DataHub访问域名

对DataHub资源的访问请求，需根据资源所属服务，选择正确的域名。

DataHub域名列表

地区	Region	外网Endpoint	经典网络ECS Endpoint	VPC ECS Endpoint
华东1(杭州)	cn-hangzhou	https://dh-cn-hangzhou.aliyuncs.com	http://dh-cn-hangzhou.aliyun-inc.com	http://dh-cn-hangzhou.aliyun-inc.com
华东2(上海)	cn-shanghai	https://dh-cn-shanghai.aliyuncs.com	http://dh-cn-shanghai.aliyun-inc.com	http://dh-cn-shanghai-int-vpc.aliyuncs.co

				m
华北2(北京)	cn-beijing	https://dh-cn-beijing.aliyuncs.com	http://dh-cn-beijing.aliyun-inc.com	http://dh-cn-beijing-int-vpc.aliyuncs.com
华南1(深圳)	cn-shenzhen	https://dh-cn-shenzhen.aliyuncs.com	http://dh-cn-shenzhen.aliyun-inc.com	http://dh-cn-shenzhen-int-vpc.aliyuncs.com
亚太东南1(新加坡)	ap-southeast-1	https://dh-singapore.aliyuncs.com	http://dh-singapore.aliyun-inc.com	http://dh-singapore-int-vpc.aliyuncs.com

DataHub RAM权限控制

DataHub资源

DataHub在RAM的访问控制中的资源体系包含Project、Topic和Subscription。目前支持Project、Topic和Subscription级别的鉴权，并不支持Shard的访问控制。其中Subscription是指对某个特定Project下的Topic的一次订阅。

资源	RAM中的资源描述
Project	acs:dhs:$region:$accountid:projects/$projectName
Topic	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName
Subscription	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/subscriptions/$subId

DataHub API及对应在RAM中的授权策略

Project

API	Action	Resource
CreateProject	dhs:CreateProject	acs:dhs:$region:$accountid:projects/*
ListProject	dhs:ListProject	acs:dhs:$region:$accountid:projects/*
DeleteProject	dhs:DeleteProject	acs:dhs:$region:$accountid:projects/$projectName
GetProject	dhs:GetProject	acs:dhs:$region:$accountid:projects/$projectName

Topic

API	Action	Resource
CreateTopic	dhs:CreateTopic	acs:dhs:$region:$accountid:projects/$projectName/topics/*
ListTopic	dhs:ListTopic	acs:dhs:$region:$accountid:projects/$projectName/topics/*
DeleteTopic	dhs:DeleteTopic	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName
GetTopic	dhs:GetTopic	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName
UpdateTopic	dhs:UpdateTopic	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName

Subscription

API	Action	Resource
CreateSubscription	dhs:CreateSubscription	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/subscriptions/*
DeleteSubscription	dhs:DeleteSubscription	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/subscriptions/$subId
GetSubscription	dhs:GetSubscription	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/subscriptions/$subId
UpdateSubscription	dhs:UpdateSubscription	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/subscriptions/$subId
ListSubscription	dhs:ListSubscription	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/subscriptions/*
CommitOffset	dhs:GetSubscription	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/subscriptions/$subId
GetOffset	dhs:GetSubscription	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/subscriptions/$

		subId

Connector

API	Action	Resource
CreateConnector	dhs:CreateConnector	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/connectors/*
DeleteConnector	dhs:DeleteConnector	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/connectors/*
GetConnector	dhs:GetConnector	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/connectors/*
UpdateConnector	dhs:UpdateConnector	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/connectors/*
ListConnector	dhs:ListConnector	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName/connectors/*

Shard

API	Action	Resource
ListShard	dhs:ListShard	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName
MergeShard	dhs:MergeShard	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName
SplitShard	dhs:SplitShard	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName

PubSub

API	Action	Resource
PutRecords	dhs:PutRecords	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName
GetRecords	dhs:GetRecords	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName
GetCursor	dhs:GetRecords	acs:dhs:$region:$accountid:projects/$projectName/topics/$topicName

DataHub支持的Condition

Condition	功能	合法取值
acs:SourceIp	指定ip网段	普通ip, 支持*通配
acs:SecureTransport	是否是https协议	true/false
acs:MFAPresent	是否多设备认证	true/false
acs:CurrentTime	指定访问时间	ISO8601格式

DataHub系统授权策略

DataHub授权策略目前在RAM系统中尚无模板，需要用户自己添加策略，具体操作路径在RAM系统中：策略管理->自定义授权策略->新建授权策略。

AliyunDataHubFullAccess

```
{
"Version": "1",
"Statement": [
{
"Action": "dhs:*",
"Resource": "*",
"Effect": "Allow"
}
]
}
```

AliyunDataHubReadOnlyAccess

```
{
"Version": "1",
"Statement": [
{
"Action": ["dhs:List*", "dhs:Get*"],
"Resource": "*",
"Effect": "Allow"
}
]
}
```

WebConsole中显示

```
// 为了在WebConsole中能够显示拥有权限的project，需要在Statement中增加如下配置
// 因为WebConsole需要ListProject和GetProject，才能在页面展示project
{
"Action": ["dhs:ListProject","dhs:GetProject"] ,
```

```
"Resource": "acs:dhs:*:*:projects/*",
"Effect": "Allow"
}
```

WebConsole中创建topic

```
// 在WebConsole的project页面中显示topic需要ListTopic和GetTopic权限
// 如希望能够在WebConsole中的project：test下能够创建topic，可以使用如下配置
{
"Version": "1",
"Statement": [
{
"Action": ["dhs:ListProject", "dhs:GetProject"],
"Resource": "acs:dhs:*:*:projects/*",
"Effect": "Allow"
},
{
"Action": ["dhs:ListTopic", "dhs:GetTopic", "dhs:CreateTopic"],
"Resource": "acs:dhs:*:*:projects/test/topics/*",
"Effect": "Allow"
}
]
}
```

DataHub自定义授权策略示例

```
//只允许用户获取指定Project下topic的信息
{
"Version": "1",
"Statement": [
{
"Action": ["dhs:ListTopic", "dhs:GetTopic"],
"Resource": "acs:dhs:cn-hangzhou:12121312:projects/foo/topics/*",
"Effect": "Allow"
}
]
}


/* PubSub
* 进行发布订阅，除了需要PutRecords，GetRecords权限外
* 往往用户需要知道topic的schema和该topic的shard状态
* 所以最好同时授予用户GetTopic和ListShard权限
*/
{
"Version": "1",
"Statement": [
{
"Action": ["dhs:*Records", "dhs:GetTopic", "dhs:ListShard"],
"Resource": "acs:dhs:cn-hangzhou:12121312:projects/foo/topics/bar",
"Effect": "Allow"
}
]
```

```
}

//对所有topic进行PubSub操作
{
"Version": "1",
"Statement": [
{
"Action": ["dhs:*Records", "dhs:GetTopic", dhs:ListShard"],
"Resource": "acs:dhs:cn-hangzhou:12121312:*",
"Effect": "Allow"
}
]
}

// 新订阅功能授权Policy样例1: 给用户授权具有project foo下topic的所有订阅权限
{
"Version": "1",
"Statement": [
{
"Action": ["dhs:*Subscription"],
"Resource": "acs:dhs:cn-hangzhou:*:projects/foo/topics/*/subscriptions/*",
"Effect": "Allow"
}
]
}

// 新订阅功能授权Policy样例2: 给用户授权仅具有project foo下查询订阅的权限
{
"Version": "1",
"Statement": [
{
"Action": ["dhs:ListSubscription"],
"Resource": "acs:dhs:cn-hangzhou:*:projects/foo/topics/*/subscriptions/*",
"Effect": "Allow"
}
]
}

// 新订阅功能授权Policy样例3: 给用户授权仅具有project foo下的topic t1特定订阅'14985645198374IoCK'的提交点位权限
{
"Version": "1",
"Statement": [
{
"Action": ["dhs:GetSubscription"],
"Resource": "acs:dhs:cn-hangzhou:*:projects/foo/topics/t1/subscriptions/14985645198374IoCK",
"Effect": "Allow"
}
]
}

// 对指定Topic进行 Split/Merge shard, 包括ListShard, SplitShard, MergeShard
{
"Version": "1",
"Statement": [
{
"Action": ["dhs:*Shard"],
```

```
"Resource": "acs:dhs:cn-hangzhou:12121312:projects/foo/topics/bar",
"Effect": "Allow"
}
]
}
```