# Alibaba Cloud Elasticsearch



Document Version20200612

# 目次

1 Elasticsearchの移行1
1.1 ユーザーが作成したElasticsearchクラスターからデータを移行する1
1.1.1 外部 Elasticsearch インスタンスから Alibaba Cloud Elasticsearch インス
タンスへのデータの移行1
1.2 サードパーティのElasticsearchインスタンスからAlibaba Cloud Elasticsearchに
データを移行する9
2 データ移行10
2.1 RDS同期10
2.1.1 MySQL データの Elasticsearch との同期10
2.1.2 RDS for MySQL から ES へのリアルタイムデータ同期
2.1.3 Canal による Alibaba Cloud Elasticsearch とのデータ同期
3 データ移行
3.1 ユーザー作成の Logstash インスタンスを使用して、データを Alibaba Cloud
Elasticsearch と同期
4 ビッグデータの同期56
4.1 データの同期と移行56
4.1.1 クラウドデータのインポート56
4.1.2 DataWorks による MaxCompute と Elasticsearch 間のデータ同期58
4.2 Alibaba Cloud Realtime Compute および Alibaba Cloud Elasticsearch のベスト
5 データを収集します
5.1 Beats による可視化された O&M システムの構築81
6 Elasticsearchアプリケーション88
7 Java high-level REST Client を使用したドキュメント API の呼び出
L89
7.1 概要
7.2 Alibaba Cloud Elasticsearch インスタンスの作成と設定89
7.3 ドキュメント API の呼び出し92
8 インデックス管理103
8.1 Curator の使用103
9 ベクトル検索プラグインのベストプラクティス106
10 Alibaba Cloud Elasticsearch サイジングのベストプラクティス114

# 1 Elasticsearchの移行

### 1.1 ユーザーが作成したElasticsearchクラスターからデータを移行 する

# 1.1.1 外部 Elasticsearch インスタンスから Alibaba Cloud Elasticsearch インスタンスへのデータの移行

このトピックでは、外部 Elasticsearch インスタンスから Alibaba Cloud Elasticsearch インス タンスにデータを移行し、データのインデックスを再作成する方法について説明します。 外部 Elasticsearch インスタンスは、Elastic Compute Service (ECS) インスタンスで実行されていま す。

#### 手順

データ移行手順には、次の手順が含まれます。

- 1. インデックスの作成
- 2. データの移行

このトピックでは、よくある質問とそれに対する解決策についても説明します。 詳細については、「FAQ」をご参照ください。

#### 前提条件

このトピックの手順に従ってデータを移行する前に、次の要件が満たされていることを確認して ください。これらの要件を満たしていない場合は、別のデータ移行計画を選択してください。

- 外部 Elasticsearch インスタンスをホストする ECS インスタンスは、VPC ネットワークにデプ ロイされている必要があります。 ClassicLink を介して VPC ネットワークに接続された ECS イ ンスタンスを使用することはできません。 外部 Elasticsearch インスタンスと Alibaba Cloud Elasticsearch インスタンスが、同じ VPC ネットワークに接続されていることを確認する必要 があります。
- このトピックのスクリプトは、中間サーバーで実行できます。中間サーバーがポート 9200 を 介して両方の Elasticsearch インスタンスに接続されていることを確認します。
- Alibaba Cloud Elasticsearch インスタンスの IP アドレスを、外部 Elasticsearch インスタン スをホストする ECS インスタンスの VPC セキュリティグループに追加し、ポート 9200 を有 効にします。

- Alibaba Cloud Elasticsearch インスタンスのノードの IP アドレスを、外部 Elasticsearch インスタンスをホストする ECS インスタンスの VPC セキュリティグループに追加します。 これにより、ノードは外部 Elasticsearch インスタンスに接続できます。 Alibaba Cloud Elasticsearch インスタンスの Kibana コンソールからノードの IP アドレスをクエリできま す。
- 外部 Elasticsearch インスタンスと Alibaba Cloud Elasticsearch インスタンスは相互接続されています。
   curl -XGET http://<host>:9200 コマンドを実行して、中間サーバーで接続をテストできます。

#### インデックスの作成

外部 Elasticsearch インスタンスの既存のインデックス設定に従って、Alibaba Cloud Elasticsea rch インスタンスに新しいインデックスを作成します。 Alibaba Cloud Elasticsearch インスタン スの自動インデックスを有効にして、インデックスとマッピングを自動的に作成することもでき ます。 ただし、自動インデックスを使用しないことを推奨します。

次の例は、外部 Elasticsearch インスタンスから Alibaba Cloud Elasticsearch インスタンスに インデックスのバッチを移行するための Python スクリプトです。 デフォルトでは、新しいイン デックスのレプリカは作成されません。

#! /usr/bin/python # -\*- coding: UTF-8 -\*-# File name: indiceCreate.py import sys import base64 import time import httplib import ison ## The ECS instance that hosts the source Elasticsearch instance (IP address + Port). oldClusterHost = "old-cluster.com" # The username of the source Elasticsearch instance. The username field can be left empty. oldClusterUserName = "old-username" ## The password of the source Elasticsearch instance. The password field can be left empty. oldClusterPassword = "old-password" ## The ECS instance that hosts the destination Elasticsearch instance (IP address + Port). newClusterHost = "new-cluster.com" ## The username of the destination Elasticsearch instance. The username field can be left empty. newClusterUser = "new-username" ## The password of the destination Elasticsearch instance. The password field can be left empty. newClusterPassword = "new-password" DEFAULT REPLICAS = 0 def httpRequest(method, host, endpoint, params="", username="", password=""): conn = httplib.HTTPConnection(host) headers = {} if (username ! = "") : 'Hello {name}, your age is {age} !'.format(name = 'Tom', age = '20') base64string = base64.encodestring('{username}:{password}'.format(username = username, password = password)).replace('\n', '')

```
headers["Authorization"] = "Basic %s" % base64string;
  if "GET" == method:
    headers["Content-Type"] = "application/x-www-form-urlencoded"
    conn.request(method=method, url=endpoint, headers=headers)
  else :
    headers["Content-Type"] = "application/json"
    conn.request(method=method, url=endpoint, body=params, headers=headers)
  response = conn.getresponse()
  res = response.read()
  return res
def httpGet(host, endpoint, username="", password=""):
def httpPost(host, endpoint, host, endpoint, "", username, password)
def httpPost(host, endpoint, params, username="", password=""):
return httpRequest("POST", host, endpoint, params, username, password)
def httpPut(host, endpoint, params, username="", password=""):
return httpRequest("PUT", host, endpoint, params, username, password)
def getIndices(host, username="", password=""):
  endpoint = "/_cat/indices"
  indicesResult = httpGet(oldClusterHost, endpoint, oldClusterUserName, oldCluster
Password)
  indicesList = indicesResult.split("\n")
  indexList = []
  for indices in indicesList:
    if (indices.find("open") > 0):
       indexList.append(indices.split()[2])
  return indexList
def getSettings(index, host, username="", password=""):
    endpoint = "/" + index + "/_settings"
    indexSettings = httpGet(host, endpoint, username, password)
    print index + " The original settings: \n" + indexSettings
  settingsDict = json.loads(indexSettings)
  ## The number of shards equals those of the indexes on the source Elasticsearch
instance by default.
  number_of_shards = settingsDict[index]["settings"]["index"]["number_of_shards"]
  ## The default number of replicas is 0.
  number_of_replicas = DEFAULT_REPLICAS
  newSetting = "\"settings\": {\"number of shards\": %s, \"number of replicas\": %s}"
% (number of shards, number of replicas)
  return newSetting
def getMapping(index, host, username="", password=""):
  endpoint = "/" + index + "/_mapping"
  indexMapping = httpGet(host, endpoint, username, password)
  print index + "The original mappings: n" + indexMapping
  mappingDict = json.loads(indexMapping)
  mappings = json.dumps(mappingDict[index]["mappings"])
  newMapping = "\"mappings\" : " + mappings
  return new Mapping
def createIndexStatement(oldIndexName):
  settingStr = getSettings(oldIndexName, oldClusterHost, oldClusterUserName,
oldClusterPassword)
  mappingStr = getMapping(oldIndexName, oldClusterHost, oldClusterUserName,
oldClusterPassword)
  createstatement = "{\n" + str(settingStr) + ",\n" + str(mappingStr) + "\n}"
  return createstatement
def createIndex(oldIndexName, newIndexName=""):
  if (newIndexName == "") :
    newIndexName = oldIndexName
  createstatement = createIndexStatement(oldIndexName)
  print "New index" + newIndexName + "Index settings and mappings: \n" + createstat
ement
  endpoint = "/" + newIndexName
  createResult = httpPut(newClusterHost, endpoint, createstatement, newClusterUser,
newClusterPassword)
  print "New index" + newIndexName + "Creation result:" + createResult
```

## main indexList = getIndices(oldClusterHost, oldClusterUserName, oldClusterPassword) systemIndex = [] for index in indexList: if (index.startswith(".")): systemIndex.append(index) else : createIndex(index, index) if (len(systemIndex) > 0) : for index in systemIndex: print index + "It may be a system index that will not be recreated. You can manually recreate the index as needed."

#### データの移行

後述のいずれかの方法でデータを移行できます。 移行するデータの量に基づいて適切な方法を選 択します。

() :

- データの整合性を確保するには、データを移行する前にソース Elasticsearch インスタンスへのデータの書き込みを停止する必要があります。ただし、ソース Elasticsearch インスタン スからデータを読み取ることはできます。データが移行された後、ターゲット Elasticsearch インスタンスに切り替えて、データを読み書きできます。ソース Elasticsearch インスタン スへのデータの書き込みを停止しない場合、ターゲット Elasticsearch インスタンスのデータ はソース Elasticsearch インスタンスのデータと一致しない可能性があります。
- 後述のいずれかの方法でデータを移行する場合、IP + Port を使用してソース Elasticsearch インスタンスを接続するには、最初にターゲット Elasticsearch インスタンスの YML 設定を 変更する必要があります。 ソースインスタンスの IP アドレスを reindex ホワイトリストに追 加します。例: reindex.remote.whitelist: 1.1.1.1:9200,1.2.\*.\*:\*。
- エンドポイントを使用してソース Elasticsearch インスタンスに接続する場合、http://host: port/path 形式を使用しないでください。ドメイン名に path を含めないでください。
- 少量のデータを移行する

reindex.sh スクリプトを実行します。

```
#! /bin/bash
# file:reindex.sh
indexName="The name of the index"
newClusterUser="The username of the destination Elasticsearch instance"
Newclusterpass = "The password of the destination Elasticsearch instance"
newClusterHost="The ECS instance that hosts the destination Elasticsearch instance"
Oldclusteruser = "The username of the source Elasticsearch instance"
Oldclusterpass = "The password of the source Elasticsearch instance"
# The address of the ECS instance that hosts the source Elasticsearch instance must be
in this format: [scheme]://[host]:[port]. Example: http://10.37.1.1:9200.
Oldclusterhost = "The ECS instance that hosts the source Elasticsearch instance"
curl -u ${newClusterUser}:${newClusterPass} -XPOST "http://${newClusterHost}/
reindex? pretty" -H "Content-Type: application/json" -d'{
```

```
"source": {
    "remote": {
        "host": "'${oldClusterHost}'",
        "username": "'${oldClusterUser}'",
        "password": "'${oldClusterPass}'"
    },
    "index": "'${indexName}'",
    "query": {
        "match_all": {}
    }
    },
    "dest": {
        "index": "'${indexName}'"
    }
}'
```

・ 大量のデータを移行する (削除操作なし、ローリング更新時間ありの場合)

削除操作なしで大量のデータを移行するには、ローリング更新を使用すると、書き込み操作 の中断期間を短縮できます。 ローリング更新では、スクリプトに時系列フィールドを追加し て、更新時間を定義する必要があります。 ソース Elasticsearch インスタンスからターゲット Elasticsearch インスタンスにデータを移行した後、ソース Elasticsearch インスタンスへの データの書き込みを停止できます。 次に、ローリング更新を使用して、データ移行中に更新さ れたデータを同期します。 ローリング更新が完了したら、ターゲット Elasticsearch インスタ ンスに切り替え、読み取り操作と書き込み操作を再開します。

#! /bin/bash # file: circleReindex.sh **# CONTROLLING STARTUP:** # This is a script that uses the Reindex operation to remotely reindex the data. Requirements: #1. You have created indexes on the destination Elasticsearch instance, or the destination instance supports auto-indexing and dynamic mapping. # 2. You must configure the IP whitelist in the YML configuration of the destination Elasticsearch instance: reindex.remote.whitelist: 172.16.123. \*:9200 #3. The specified ECS instance address must be in the following format: [scheme]://[ host]:[port]. USAGE="Usage: sh circleReindex.sh <count> count: the number of times to perform the reindex operation. A negative number indicates loop execution. You can set this parameter to perform the reindex operation only once or multiple times. Example: sh circleReindex.sh 1 sh circleReindex.sh 5 sh circleReindex.sh -1" indexName="The name of the index" newClusterUser="The username of the destination Elasticsearch instance" newClusterPass="The password of the destination Elasticsearch instance" oldClusterUser="The username of the source Elasticsearch instance" oldClusterPass="The password of the source Elasticsearch instance" ## http://myescluster.com newClusterHost="The ECS instance that hosts the destination Elasticsearch instance" # You need to specify address of the ECS instance that hosts the source Elasticsearch instance in the following format: [scheme]://[host]:[port]. Example: http://10.37.1.1: 9200 oldClusterHost="The ECS instance that hosts the source Elasticsearch instance" timeField="The field that specifies the time window during which the incremental data

is migrated"

```
reindexTimes=0
lastTimestamp=0
curTimestamp=`date +%s`
hasError=false
function reIndexOP() {
  reindexTimes=$[${reindexTimes} + 1]
  curTimestamp=`date +%s`
  ret=`curl -u ${newClusterUser}:${newClusterPass} -XPOST "${newClusterHost}/
_reindex? pretty" -H "Content-Type: application/json" -d '{
     "source": {
       "remote": {
"host": "'${oldClusterHost}",
         "username": "'${oldClusterUser}'",
"password": "'${oldClusterPass}'"
       },
"index": "'${indexName}'",
       "query": {
         "range<sup>†</sup> : {
            "'${timeField}'" : {
              "gte" : '${lastTimestamp}',
              "lt" : '${curTimestamp}'
         }
      }
     dest": {
       "index": "'${indexName}'"
    }
  יינ
  lastTimestamp=${curTimestamp}
  echo "${reindexTimes} reindex operations have been performed. The last reindex
operation is completed at ${lastTimestamp} Result: ${ret}"
  if [[ ${ret} == *error* ]]; then
    hasError=true
    echo "An unknown error occurred while performing this operation. All subsequent
operations have been suspended."
  fi
function start() {
  ## A negative number indicates loop execution.
  if [[ $1 -lt 0 ]]; then
    while :
    do
       reIndexOP
    done
  elif [[ $1 -gt 0 ]]; then
    k=0
    while [[ k -lt $1 ]] && [[ ${hasError} == false ]]; do
       reIndexOP
       let ++k
    done
  fi
## main
if [ $# -lt 1 ]; then
  echo "$USAGE"
  exit 1
fi
echo "Start the reindex operation for index ${indexName}"
start $1
```

echo "You have performed \${reindexTimes} reindex operations"

・ 大量のデータを移行する (削除操作またはローリング更新時間なしの場合)

大量のデータを移行する必要があり、マッピングに更新時間フィールドが定義されていない場 合、ソースインスタンスのワークロードのスクリプトに更新時間フィールドを追加する必要が あります。フィールドを追加した後、既存のデータを移行してから、前述のデータ移行計画で 説明されているローリング更新を使用して増分データを移行できます。

```
#! /bin/bash
# file:miss.sh
indexName="The name of the index"
newClusterUser="The username of the destination Elasticsearch instance"
Newclusterpass = "The password of the destination Elasticsearch instance"
newClusterHost="The ECS instance that hosts the destination Elasticsearch instance"
Oldclusterpass = "The password of the source Elasticsearch instance"
# The address of the ECS instance that hosts the source Elasticsearch instance must be
in this format: [scheme]://[host]:[port]. Example: http://10.37.1.1:9200.
oldClusterHost="The ECS instance that hosts the source Elasticsearch instance"
timeField="updatetime"
curl -u ${newClusterUser}:${newClusterPass} -XPOST "http://${newClusterHost}/
_reindex? pretty" -H "Content-Type: application/json" -d '{
    "source": {
        "remote": {
            "host": "'${oldClusterHost}",
       "username": "'${oldClusterUser}'",
       "password": "'${oldClusterPass}'"
    },
"index": "'${indexName}'",
    "query": {
"bool": {
         "must not": {
            "exists": {
              "field": "'${timeField}"
         }
       }
    }
  },
  "dest": {
    "index": "'${indexName}'"
  }
}'
```

書き込み操作を停止せずにデータを移行する

このデータ移行計画は今後提供される予定です。

#### FAQ

問題: curl コマンドを実行すると、{"error":"Content-Type header [application/x-www-•

form-urlencoded] is not supported","status":406} が表示される。

解決策:-H "Content-Type: application/json" パラメーターを curl コマンドに追加し、もう

一度お試しください。

// Obtain all the indexes on the source instance. If you do not have the required permissions, remove the "-u user:pass" parameter. Make sure that you have replaced oldClusterHost with the name of the ECS instance that hosts the source Elasticsearch instance.

curl -u user:pass -XGET http://oldClusterHost/\_cat/indices | awk '{print \$3}' // Based on the returned indexes, obtain the setting and mapping of the index that you need to migrate for the specified user. Make sure that you have replaced indexName with the index name that you need to query.

curl -u user:pass -XGET http://oldClusterHost/indexName/ settings, mapping?pretty =true

// Create a new index on the destination Elasticsearch instance according to the settings and mapping settings that you have obtained from the preceding step. You can set the number of index replicas to 0 to accelerate the data synchronization process, and change the number of replicas to one after the migration is complete. // ewClusterHost indicates the ECS instance that hosts the destination Elasticsea rch instance, testindex indicates the name of the index that you have created, and testtype indicates the type of the index.

```
curl -u user:pass -XPUT http://<newClusterHost>/<testindex> -d '{
 "testindex" : {
"settings" : {
```

"number\_of\_shards" : "5", //Specify the number of shards for the corresponding index on the source Elasticsearch instance, for example, 5

```
"number of replicas" : "0" //Set the number of index replicas to zero
}
```

```
"mappings" : { //Set the mapping for the index on the source Elasticsearch
instance. For example, you can set the mapping as follows
```

```
"testtype" : {
    'properties" : {
      "uid" : {
         "type" : "long"
       name" : {
        "type" : "text"
       create_time" : {
        "type" : "long'
      ł
  }
}
```

}

}'

• 問題:データ移行プロセスが遅すぎる。

解決策:インデックスが大きすぎる場合、移行前にレプリカの数を0に、更新間隔を-1に設 定します。 データの移行後、レプリカと更新の設定を元の値に戻します。 これにより、同期 プロセスが高速化されます。

	<pre>// You can set the number of index replicas to 0 and disable refresh to accelerate the migration process. curl -u user:password -XPUT 'http://<host:port>/indexName/_settings' -d' {     "number_of_replicas" : 0,     "refresh_interval" : "-1" }' // After the data is migrated, set the number of index replicas to 1 and the refresh interval to 1s (default value, which means 1 second). curl -u user:password -XPUT 'http://<host:port>/indexName/_settings' -d' {     "number_of_replicas" : 1,     "refresh_interval" : "1s" }'</host:port></host:port></pre>
	<u>〕</u> 注:
-	のトピックのコンテンツの一部は 『Peindey API』から引用しています

### 1.2 サードパーティのElasticsearchインスタンスからAlibaba Cloud Elasticsearchにデータを移行する

# 2 データ移行

:

## 2.1 RDS同期

### 2.1.1 MySQL データの Elasticsearch との同期

Alibaba Cloud は、幅広いクラウドストレージとデータベースサービスを提供します。 DataWorks の Data Integration を使用して、ストレージサービスやデータベースサービスと Alibaba Cloud Elasticsearch (ES) との間でデータを同期し、データのクエリや分析をすることが できます。 Data Integration を使用すると、最小 5 分間隔でデータを同期できます。

データの同期には、インターネットデータ転送料金が発生する場合があります。

#### 概要

次の手順に従って、データベースに保存されているデータを分析・検索します。

**1.** データベースを作成します。 ApsaraDB RDS for RDS データベースを使用するか、ローカ ルサーバーにデータベースを作成することができます。 このトピックでは、ApsaraDB RDS for MySQL データベースを使用します。 2 つの MySQL テーブルを結合してから、データを Elasticsearch と同期します。 次の図に、2 つの MySQL テーブルを示します。

#### 図 2-1 : テーブル 1

	id	•	stu_id	*	c_name	▼ grade	*
1		1	901		compute		98
2		2	901		english		80
3		3	902		compute		65
4		4	902		chinese		88
5		5	903		chinese		95
6		6	904		compute		70
7		7	904		english		92
8		8	905		english		94
9		9	906		compute		90
10	1	0	906		english		85

#### 図 2-2 : テーブル 2

	id 💌	name 💌	sex	birth 💌	department 💌	address 💌
1	901	zhangda	man	1985	compute	beijing
2	902	zhanger	man	1986	chinese	beijing
3	903	zhangsan	woman	1990	chinese	hunan
4	904	lisi	man	1990	english	liaoning
5	905	wangwu	woman	1991	english	fujian
6	906	wangliu	man	1988	compute	hunan

- VPC ネットワークを介して Elasticsearch インスタンスに接続可能な Elastic Compute Service (ECS) インスタンスを購入します。 ECS インスタンスを使用して、MySQL データベー スからデータを読み取り、同期タスクを実行して Elasticsearch にデータを書き込みます。 同 期タスクは Data Integration によってディスパッチされます。
- **3.** Data Integration を有効化し、同期タスクを実行するためのリソースとして ECS インスタン スを Data Integration に追加します。
- 4. データ同期スクリプトを作成し、定期的に実行します。
- 5. Data Integration で同期されたデータを保存するための Elasticsearch インスタンスを作成します。

#### 準備

- 1. VPC ネットワークと VSwitch を作成します。
- 2. #unique\_9を行います。

## •

Elaticsearch インスタンスの [リージョン]、[VPC ネットワーク]、[VSwitch] は、手順1で指 定したものと同じでなければなりません。

**3.** Elasticsearch インスタンスと同じ VPC ネットワークに接続されている ECS インスタンスを購入し、パブリック IP アドレスか Elastic IP アドレス (EIP) を割り当てます。



- 既存の ECS インスタンスを使用することもできます。
- ECS インスタンスには、CentOS 6、CentOS 7、または AliyunOS を選択することを推奨します。
- ECS インスタンスで MaxCompute または同期タスクを実行する場合、ECS インスタンスが Python V2.6 または V2.7 を実行していることを確認します。 CentOS 5 は Python V2.4 を使用します。 他のオペレーティングシステムは、Python バージョン V2.6 以降を使用します。
- ECS インスタンスにパブリック IP アドレスが割り当てられていることを確認します。

#### リソースグループの作成とデータソースの追加

- 1. DataWorks コンソールにログインし、[Workspaces] タブをクリックします。
  - Data Integration が既に有効化されている場合、次のページが表示されます。

🕤 Dat	aWorks D	ataStudio∙Data Inte	gration · MaxComp	oute	<b>(</b> )
hortcuts					
Data Analytics		Data Integration	Maintenance Cente	r	Data Service
/orkspaces					All Workspace
asdjhasuhschj64273	China East 2	MaxCompute_DOC	Asia Pacific SE 1	test012	China East 2
Created At:Mar 15, 2018, 16:47:: Compute Engines:None Services:Data Studio Data Integ	55 ration Data Managem	Created At:Jan 21, 2019, 23: Compute Engines:MaxComp Services:Data Studio Data Int	3:23 ute egration Data Managem	Created At:Jan 02, 2018, 1 Compute Engines:MaxCom Services:Data Studio Data I	:36:40 pute PAI calculation engine ntegration Data Managem
Workspace Settings	Data Analytics	Workspace Settings	Data Analytics	Workspace Settings	Data Analytics
Data Service	Data Integration	Data Integration		Data Service	Data Integration

 Data Integration が有効化されていない場合、次のページが表示されます。 手順に従って Data Integration を有効化します。 Data Integration は有料サービスです。 ページに表示 される課金項目に基づいて料金を見積もることができます。

G	DataWorks	DataStudio · Data Integra
User Guide	eme Authentication	
Region China No Middle E Compute	orth 2 China East 1 China East 2 C ast 1 Asia Pacific SOU 1 Asia Pacific S Engines	China South 1 Hong Kong US West 1 Asia Pac
	MaxCompute Pay-As-You-Go B Allows you to develop MaxCompute SC Machine Learning Platform for Al Provides machine learning algorithms,	Buy Now       Subscription Buy Now         QL and MaxCompute MR tasks in DataWorks.         Pay-As-You-Go Buy Now         deep learning frameworks, and online prediction service
DataWorl	xs Services	
	<ul> <li>Data Integration Pay-As-You-Go</li> <li>Allows for data synchronization betwee</li> <li>Data Analytics, O&amp;M, and Administra</li> <li>Enables you to arrange workflows, sche</li> </ul>	Buy Now en a data source and a data destination. More than 20 ty tion edule recurring tasks, and query information (including p

- **2.** 購入したワークスペースの [Actions] 列の [Data Integration] をクリックします。
- [Data Integration] ページの左側のナビゲーションペインで [Resource Group] を選択し、右 上隅の [Add Resource Group] をクリックします。

- 4. リソースグループを追加します。
  - a. サーバーを追加します。

リソースグループ名を入力し、サーバー情報を指定します。 この例では、購入した ECS イ ンスタンスを追加します。 サーバー情報を次に示します。

Add Resource Group				×
Create Resource Group	Add Server	Install Agent	t Te	st Connection
* Network Type: 1 Servers	• VPC 😨			
* ECS UUID:	Enter a UUID rather the	an server name.	0	
* Server IP Address:	Enter the internal IP ad	ldress of the server.	0	
* Server CPU (Cores):				
* Server RAM (GB):				
Add Server				
		[	Previous	Next
フィールド	フィールドの記	说明		
ECS UUID	#unique_10を す。戻り値をこ	行い、dmideco このフィールドに	ode   grep UU こ入力します。	ID を実行しま

フィールド	フィールドの説明
Server IP Address/Server CPU (Cores)/Server RAM (GB)	ECS インスタンスのパブリック IP アドレス、CPU コア、メ モリサイズ。 必要な情報を取得するには、ECS コンソール にログインし、ECS インスタンス名をクリックします。 [設 定情報] フィールドに、情報が表示されます。

**b.** Security Center エージェントをインストールします。

ページの指示に従って、エージェントをインストールします。 手順 5 で、ECS インスタン スのポート 8000 を開きます。 この手順をスキップすると、デフォルト設定が使用されま す。

**c.** 接続をテストします。

5. MySQL データベースと Elasticsearch ホワイトリストを設定します。

ECS インスタンスの IP アドレスを MySQL データベースと Elasticsearch インスタンスのホワ イトリストに追加して、ECS インスタンスが MySQL データベースと Elasticsearch インスタ ンスと通信できるようにします。

左側のナビゲーションペインで [Connections] を選択し、[Add Connection] をクリックします。

7.	[MySQL] を選択します。	[Add MySQL Connection] ページで、	必要な情報を入力します。
----	-----------------	------------------------------	--------------

Add Data Source MyS	QL	×
* Data Source Type :	ApsaraDB for RDS ~	
* Data Source Name :	Enter a name.	
Description :		
* RDS Instance ID :		0
* RDS Instance :		?
Account		
* Database Name :		
* Username :		
* Password :		
Test Connectivity :	Test Connectivity	
0	The connectivity test can be passed only after the data source is added to the whitelist. Click <b>here</b> to see how to add a data source to the whitelist. Ensure that the database is available.	
	Ensure that the firewall allows the data sent from or to the database to pass by.	
	Ensure that the database domain name can be resolved. Ensure that the database has been started.	
	Previous	mplete

Connect To:この例では、ApsaraDB RDS for MySQL データベースを使用します。 [Usercreated Data Store with Public IP Addresses] か [User-created Data Store without Public IP Addresses] から選択できます。 パラメーターの詳細については、「#unique\_11」 をご参照ください。

# (!) :

ECS インスタンスがデータベースへの接続に失敗した場合、データベースのホワイトリスト を確認してください。

#### 同期タスクの作成

- 1. DataWorks コンソールにノード所有者としてログインします。
- 2. [Workspaces] タブで、[Actions] 列の [Data Analytics] をクリックします。
- 3. [Data Studio] ページで、[Create] > [Workflow] を選択します。



- [Create Workflow] ダイアログボックスで、[Workflow Name] と[Description] を指定し、[Create] をクリックします。
- **5.** 作成したワークフローを左側のワークフローリストで展開し、[Data Integration] を右クリックし、[Create Data Integration Node] > [Sync] を選択します。
- 6. [Create Node] ダイアログボックスで、[Node Name] を指定し、[Commit] をクリックしま す。
- 7. [node] タブの上部のツールバーで、[Switch to Code Editor] アイコン 7 をクリックしま

す。

8. 操作を確認して、コードエディターに切り替えます。

コードエディターの使用方法については、「#unique\_12」をご参照ください。

次のスクリプトは、2つのテーブルから学生と試験の情報を取得する例です。

```
ł
 "type": "job",
 "steps": [
      "stepType": "mysql",
      "parameter": {
         column": [
          "id"
          "name",
          "sex",
"birth",
          "department",
          "address": {
        ],
"connection": [
          ł
             "querysql":["SELECT student.id,name,sex,birth,department,address,
c name, grade FROM student JOIN score on student.id=score.stu id;"],
             "datasource": "zl_****_rdsmysql",
             "table": [
"score"
          }
        ],
"where": ""
```

```
"splitPk": ""
          "encoding": "UTF-8"
      ł,
      "name":"Reader",
      "category": "reader"
   },
{
      "stepType": "elasticsearch",
      "parameter": {

"accessId": "elastic",

"endpoint": "http://es-cn-0p*******2dpxtx.elasticsearch.aliyuncs.com:9200",

"indexType": "score",

"accessKey": "*****",

"alceapup": true
          "cleanup": true,
"discovery": false,
"column": [
             ł
                "name":"student_id",
"type":"id"
             },
             {
                "name": "sex",
"type": "text"
             },
             {
                "name": "name",
                "type": "text"
             },
            ł
                "name": "birth",
"type": "integer"
             },
             {
                "name": "quyu",
"type": "text"
             },
             {
                "name": "address",
"type": "text"
             },
             {
                "name": "cname",
                "type": "text"
             },
             {
                "name": "grades",
                "type": "integer"
             }
         j,
"index": "mysqljoin",
"batchSize": 1000,
"splitter": ","
      },
"name": "Writer",
      "category": "writer"
   }
],
"version": "2.0",
"order": {
   "hops": [
      ł
          "from": "Reader",
          "to": "Writer"
      }
```

```
    ]
    },
    "setting": {
        "jvmOption": "-Xms1024m -Xmx1024m",
        "errorLimit": {
            "record": ""
        },
        "speed": {
            "throttle": false,
            "concurrent": 1
        }
    }
}
```

スクリプトには3つのセクションがあります。

- Reader:このセクションは、#unique\_13に使用されます。querysqlを使用して、特定の条件に基づいてデータを取得する SQL 文を定義します。 querysql が設定されている場合、MySQL Reader は table、column、where、splitPk 条件を無視します。これは、querysql の優先順位は table、column、where、splitPk よりも高いためです。datasource は querysql を使用して、ユーザー名とパスワード情報を解析します。
- Writer:このセクションは、#unique\_14に使用されます。
  - endpoint: Elasticsearch インスタンスのパブリックまたはプライベートネットワーク エンドポイント。Elasticsearch インスタンスに接続するには、Elasticsearch インスタ ンスの [セキュリティ] ページでパブリックまたはプライベートネットワークのホワイト リストを設定する必要があります。
  - accessId/accessKey: Elasticsearch インスタンスのユーザー名とパスワード。デフォ ルトのユーザー名は、elasticです。
  - index: Elasticsearch インスタンスのインデックスの名前。インデックスに保存され たデータにアクセスするには、インデックス名を指定する必要があります。
  - Reader と Writer 内の列は、同じ順序で定義する必要があります。 Reader は、指定された列からデータを読み取り、データを配列に保存します。 次に、Writer は配列からデータを取り出し、定義された列に順番にデータを書き込むためです。
- setting:このセクションは、パケット損失や最大同時数などの設定に使用されます。

 スクリプトの設定後、右上隅の [Resource Group] をクリックし、作成したリソースグ ループを選択し、上部のツールバーで [Run] アイコンをクリックして MySQL データを Elasticsearch と同期します。

× Resource Group configuration ⑦										
The data integration task runs in the resource group, and the joint debugging operation with the data source is also initiated in the resource group. According to the specific scope of application of each resource group, select the appropriate resource group for your network scenario.										
• Programme: 🔵 Exclusive Resource Groups 💽 Custom DI Resource Groups 🔵 Common Resource Group										
You can use an ESC or an IDC in your VPC as a resource group to perform data integration tasks, so that you can use existing resources and ensure that resources are exclusive.										
Public Network       Accessible       Data Source         Public Network       Image: Custom Resource Group       Image: Custom Resource Group         Data Source       Custom Resource       DataWorks with source Group										
* Custom DI Resource Groups: es_test1										

#### 同期結果の確認

- 1. Elasticsearch インスタンスの Kibana コンソールにログインします。
- 2. 左側のナビゲーションペインで、[Dev Tools] を選択します。
- 3. [Console] タブで、次のコマンドを実行して同期データをクエリします。

POST /mysqljoin/\_search? pretty { "query": { "match\_all": {}} }

文字列 mysqljoin は、同期データが保存されている index の名前です。

#### FAQ

• Q:データベース接続エラーを解決するにはどうすればよいですか。

A:リソースグループ内の ECS インスタンスのパブリックまたはプライベート IP アドレス が、データベースのホワイトリストに追加されているかどうかを確認してください。 IP アド レスがホワイトリストに追加されていない場合は、追加してください。 • Q: Elasticsearch インスタンスの接続エラーを解決するにはどうすればよいですか。

A:次の手順に従って原因を特定します。

- 1. コードエディターの右上隅にある [Resource Group] をクリックし、前の手順で選択したリ ソースグループが選択されているかどうかを確認します。
  - 選択されている場合、次の手順に進みます。
  - 選択されていない場合、 [Resource Group] をクリックし、作成したリソースグループ を選択します。 次に、 [Run] アイコンをクリックしてスクリプトを実行します。
- 2. リソースグループ内の ECS インスタンスの IP アドレスが Elasticsearch インスタンスのホ ワイトリストに追加されているかどうかを確認します。
  - 追加されている場合、次の手順に進みます。
  - 追加されていない場合、ECS インスタンスの IP アドレスを Elasticsearch インスタンスのホワイトリストに追加してください。

#### (!)

ECS インスタンスのプライベート IP アドレスが使用されている場合、Elasticsearch イ ンスタンスの [セキュリティ] ページに移動し、IP アドレスを Elasticsearch システムの ホワイトリストに追加します。 ECS インスタンスのパブリックIPアドレスが使用されて いる場合、Elasticsearch インスタンスの [セキュリティ] ページに移動し、IP アドレス をパブリックネットワークホワイトリストに追加します。

 スクリプトの設定が正しいかどうかを確認します。確認する必要があるフィールドには、 endpoint (Elasticsearch インスタンスのパブリックまたはプライベートネットワークエン ドポイント)、accessId (Elasticsearch インスタンスのユーザー名。デフォルトのユーザー 名は elastic)、accessKey (Elasticsearch インスタンスのパスワード) があります。

### 2.1.2 RDS for MySQL から ES へのリアルタイムデータ同期

このセクションでは、Data Transmission Service (DTS) を使用して、RDS for MySQL インスタン スから Elasticsearch (ES) インスタンスへのリアルタイムデータ同期タスクを迅速に作成する方 法を説明します。 DTS はこの同期機能を使用して RDS for MySQL データを ES インスタンスに同 期し、データをリアルタイムでクエリします。

#### リアルタイム同期タイプ

同じ Alibaba Cloud アカウントの DTS インスタンスで、RDS for MySQL から ES。

#### SQL 操作タイプ

サポートされている主な SQL 操作タイプは次のとおりです。

- Insert
- Delete
- Update

🧾 注:

DTS では、DDL 文を使用してデータを同期することはできません。 データを同期する場合、DDL 操作は無視されます。

RDS for MySQL インスタンスで DDL を使用するテーブルが検出されると、対応するテーブルの DML 操作が失敗する可能性があります。 この問題を解決するには、以下の手順を実行します。

- **1.** 同期リストからオブジェクトを削除します。 詳細は、「」「同期オブジェクトの削除」「」を ご参照ください。
- **2.** ES インスタンスのこのテーブルに対応するインデックスを削除します。
- **3.** テーブルを同期リストに追加し直し、再初期化してください。詳細は、「」「同期オブジェ クトの追加」「」をご参照ください。

DDL を使用して列を追加したりテーブルを変更したりする場合、DDL 操作順序は次のとおりです。

- 1. ES インスタンスのマッピングと新しい列を手動で変更します。
- 2. テーブルスキーマを変更し、ソース RDS for MySQL インスタンスに新しいスキーマを追加します。
- 3. DTS でインスタンスの同期を停止します。DTS 同期インスタンスを再起動して ES で変更され たマッピング関係を再ロードします。

データ同期の設定

RDS for MySQL インスタンスから ES インスタンスにデータを同期するには、以下の手順を実行 します。 1. DTS 同期インスタンスの購入

Data Transmission Service コンソールにログインし、[データ同期] ウィンドウに移動しま す。 右上隅にある [同期タスクの作成] をクリックして同期インスタンスを購入します。 同期 インスタンスを設定できるようになります。

### 道注:

設定する前に同期インスタンスを購入する必要があります。 サブスクリプションと従量課 金の 2 つの課金方法がサポートされています。.

購入ページのパラメーター

機能

[データ同期]を選択します。

ソースインスタンス

[MySQL] を選択します。

- ソースリージョン
  - この例では、RDS for MySQL インスタンスを使用するため、RDS for MySQL インスタン スが存在するリージョンを選択する必要があります。
- ターゲットインスタンス

[Elasticsearch] を選択します。

• ターゲットリージョン

Elasticsearch インスタンスが存在するリージョンを選択します。 同期インスタンスを購入 した後、リージョンを変更することはできません。

仕様

インスタンス仕様は、同期インスタンスのパフォーマンスに対応しています。 詳細は、 「」「データ同期仕様」「」をご参照ください。

• 購入期間

- 同期インスタンスがサブスクリプションの場合、デフォルトの購入期間は1ヶ月です。

数

デフォルトの数量は1です。

📋 注:

DTS 同期インスタンスのリージョンは、選択したターゲットリージョンです。たと えば、[中国 (杭州)] リージョンの RDS for MySQL から [中国 (杭州)] リージョンの Elasticsearch への同期インスタンスの場合、DTS 同期インスタンスのリージョンは [中国 (杭州)] です。 同期インスタンスを設定するには、DTS でそのリージョンのインスタンス リストに移動し、購入したばかりの同期インスタンスを検索して、右上隅の [同期インス タンスの設定] をクリックします。

#### 2. 同期インスタンスの設定

Data Transmission	Synchronize Task List	Singapore China (H	angzhou) China (Shanghai	) China (Qingda	o) China (Beijing)	China (Shenzhen)	Hong Kong	US (Silicon Val	ley) US (Virgin	nia) UAE (Dubai)	(Synchronization Joh
Overview	target region)	Germany (Frankfurt)	Malaysia (Kuala Lumpur)	China (Hohhot)	Australia (Sydney)	India (Mumbai)	UK(London)	Japan (Tokyo)	Indonesia (Jaka	rta)	(official official of
Data Migration									OTS FAQ	C Refresh	Create Synchronization Task
Data Subscription	Synchronous Task Name	×		Search	Bank: Default	order	• Status:	All	T		
Data Synchronization	Synchronous rusk nume			Startin	Deroun	order					
Documentation	Instance ID/Task N	ame	Status	Synchroni	zation Overview	Method o	of Payment		Synchronizatio Architecture(A	on (III) 👻	Operation
	dts in findings of hangzhou-hangzhou	u-micro	Unconfigured			Рау-Аз-Ү	'ou-Go		One-Way Synchronizatio	Con	figure Synchronization Instance To subscription   Upgrade More

同期タスク名

同期インスタンスの名前に関する要件はありません。

ソースインスタンス

この例では、データソースとして RDS for MySQL を使用します。 インスタンスのタイプ、 リージョン、ID、およびデータベースのアカウントとパスワードを設定する必要があります。

Synchronous Task Name:	hangzhou-hangzhou-small	]
Source Instance Information		
Instance Type:	RDS Instance	
Instance Region:	China (Hangzhou)	
* Instance ID:	rm-bpt.st.visctonderers -	RDS instances belong to other Alicloud account
* Database account:	root	
* Database Password:	••••••• • •	
* Connection method:	${oldsymbol{\circ}}$ Non-encrypted connection ${oldsymbol{\circ}}$ SSL secure connection	

#### ターゲットインスタンス

ES インスタンスの ID、アカウント、パスワードを設定する必要があります。

Target Instance Information		
Instance Type:	RDS Instance 🔻	
Instance Region:	China (Hangzhou)	
* Instance ID:	rm-bplat2%SlideBox -	
* Database account:	elastid	
* Database Password:	<i>م</i> ه	
* Connection method:	Non-encrypted connection $\bigcirc SSL \text{ secure connection}$	

これらの設定が完了したら、[ホワイトリストを承認して次のステップに進む] をクリックし て、RDS for MySQL と ES インスタンスのホワイトリストに IPを追加します。

3. インスタンスのホワイトリストの承認

# **道**注:

ソースインスタンスが RDS for MySQL の場合、自動的に IP がホワイトリストに追加される か、セキュリティグループが追加されます。

ソースインスタンスが RDS for MySQL の場合、インスタンス IP が RDS インスタンスのホワ イトリストのセキュリティグループに追加されます。 これにより、同期タスクを作成すると き、DTS インスタンスと RDS データベース間の切断によって引き起こされるエラーを回避す ることができます。 同期タスクの安定性を確保するため、RDS インスタンスからインスタン ス IP を削除しないでください。

ホワイトリストの承認後、[次へ]をクリックして同期アカウントを作成します。

4. 同期オブジェクトの選択

同期オブジェクトとインデックスの命名規則を設定するには、次の手順を実行します。

- a. インデックスの命名規則を [テーブル名] または [データベース名\_テーブル名] から選択します。
  - [テーブル名]を選択した場合、インデックスの名前はテーブルの名前です。
  - [データベース名\_テーブル名] を選択した場合、インデックスの命名規則はデータベース
     名\_テーブル名です。たとえば、データベース名が dbtest、テーブル名が sbtest1 の場

合、テーブルが ES インスタンスと同期された後にインデックス名は dbtest\_sbtest1 に なります。

- 同じ名前の2つのテーブルが、それぞれ異なるデータベースにある場合、インデックス
   名をデータベース名\_テーブル名に設定することを推奨します。
- b. 特定のデータベース、テーブル、列を選択します。 同期オブジェクトの選択可能粒度は、 テーブルレベルの操作をサポートします。 つまり、複数のデータベースとテーブルを同期 できます。

同期オブジェクトの選択可能粒度は、テーブルレベルの操作をサポートします。 つまり、 複数のデータベースとテーブルを同期できます。

Synchronization Architecture:One-Way Synchronization		
index name: TableName •	7	
Source Database Object		Selected objects (Move the mouse to the object and click "Edit" to revise the object name or configure the filter condition) (lick here
■ Matest ■ Tables	> <	wdtest (10bjects)
All		All

**c.** デフォルトでは、すべてのテーブルの docid はプライマリキーです。 プライマリキーを持たないテーブルの場合、ソーステーブルの列に対応する docid を設定します。 右側の [選

 $\times$ 

### 択したオブジェクト]のボックスで、対応するテーブルの上にポインターを移動し、[編集] をクリックして [詳細設定] ウィンドウに移動します。

Edit table

Note: After being edited, the table or column name in the target database will be the modified name.

* Index	Name : tb1			
* Type	Name : tb1			
IsPa	rtition : 🔘 yes 🖲 r	o		
_id	value : the prima	ry key of table	•	
	Column Name	Туре	column param	column param value
	id	int(11)	index 🔻	false 🔻
	name	varchar(10	index •	false 🔻 add param
				ОК

- **d.** [詳細設定] では、インデックス名、タイプ名、パーティション列、数量、および\_id 値列を 設定できます。\_id の値がビジネスプライマリキーに設定されている場合、対応するビジ ネスプライマリキー列を選択する必要があります。
- e. 同期オブジェクトを設定したら、詳細設定に進みます。
- 5. 詳細設定

主な設定

a. 同期の初期化: [構造体の初期化] と [データの初期化] を選択することを推奨します。これ により、自動的にインデックスが作成され、データが初期化されます。 [スキーマの初期化] を選択しない場合、同期する前に ES のインデックスのマッピングを手動で定義する必要が あります。 [全データの初期化] を選択しない場合、増分 DTS データ同期の開始時刻は同期 が開始される時刻です。

- b. シャード設定:デフォルトで5つのパーティションと1つのレプリカがあります。設定の 調整後、すべてのインデックスがこの設定に従ってパーティションを定義します。
- c. 文字列インデックス:文字列を選択できるアナライザーです。デフォルトは、[Standard Analyzer] です。その他の値は、[Simple Analyzer]、[Whitespace Analyzer]、[Stop Analyzer]、[Keyword Analyzer]、[English Analyzer]、[Fingerprint Analyzer] です。すべてのインデックスの文字列フィールドは、この設定に従ってアナライザーを定義します。

1.Select the source and target in	stances of	2.Select the synchronization object	t 🔪	3.Advanced Setup	4.Pre-check	
Synchronization Initialization:	Structure Initialization [	Data Initialization				
Shard Configuration :	Please Configuration $% {\displaystyle \sum} {\displaystyle \sum}$	Please Configuration replica,				
String Index :	analyzed	Standard Analyzer				
TimeZone :	+8:00	T				
DOCID :	Default primary key, no automatic generation of ID u	primary key table, sing Elasticsearch				

- **d. Time Zone**: ES インスタンスに同期している時間フィールドが保存されている場所です。 中国のデフォルトのタイムゾーンは UTC (UTC + 8) です。
- 6. 事前チェック

同期タスクの設定が完了したら、DTS は事前チェックを実行します。 事前チェックが確認され たら、[開始] をクリックして同期タスクを開始します。

同期タスクが開始したら、同期ジョブリストに移動し、タスクのステータスが[同期初期化]か どうかを確認します。初期化にかかる時間は、ソースインスタンス内の同期オブジェクトの データ量によって異なります。初期化が完了すると、同期インスタンスのステータスは[同期 中]になります。ソースインスタンスとターゲットインスタンス間に同期リンクが確立されま す。

7. データの検証

上記のすべての手順を完了したら、ES コンソールにログインして、ES インスタンスに作成さ れたインデックスと同期データを確認します。

### 2.1.3 Canal による Alibaba Cloud Elasticsearch とのデータ同期

このトピックでは、Canal を使用して、ApsaraDB RDS for MySQL の増分データを Alibaba Cloud Elasticsearch と同期させる方法について説明します。

Canal を使用する前に、次の前提条件を満たしていることを確認してください。

() :

ApsaraDB RDS for MySQL、Alibaba Cloud Elasticsearch、Elastic Compute Service (ECS) を有効化するとき、これらのサービスに同じリージョン、ゾーン、**VPC** ネットワー ク、**VSwitch**、セキュリティ グループが指定されていることを確認します。

ApsaraDB RDS for MySQL

ApsaraDB RDS for MySQL を使用して、同期対象のソースデータと増分データを保存します。 ApsaraDB RDS for MySQL を有効化する方法の詳細は、「#unique\_18」をご参照ください。 次の図に、このトピックで使用する ApsaraDB RDS for MySQL の設定を示します。



・ canal.adapter-1.1.4.tar.gz とcanal.deployer-1.1.4.tar.gz

Canal パッケージです。 Canal は、GitHub オープンソースの extract-transform-load (ETL) ツールです。このツールを使用してデータベースログを解析し、同期対象の増分データを取得 します。 詳細は、「Canal」をご参照ください。

• Alibaba Cloud Elasticsearch

Alibaba Cloud Elasticsearch を使用して、同期対象の増分データを受信します。 Alibaba Cloud Elasticsearch を有効化する方法の詳細は、「#unique\_9」をご参照ください。 次の図 に、このトピックで使用する Elasticsearch の設定を示します。

Basic Information	
Instance ID: es-cn-	Created At: Oct 8, 2019, 10:18:24
Name: es-cn-	Status: • Active
Version: 6.7.0 with Commercial Feature	Billing Method: Pay-As-You-Go
Regions: China (Hangzhou)	Zone: cn-hangzhou-h
VPC: vpc-bp //t	VSwitch: vsw-bp9k
Internal Network Address: es-cn-	Internal Network Port: 9200
Public Network Access:	
Protocol: HTTP Edit	

• Alibaba Cloud ECS

Alibaba Cloud ECS を使用して、ApsaraDB RDS for MySQL と Alibaba Cloud Elasticsearch を接続します。 また、Canal deployer と Canal adapter を Alibaba Cloud ECS にデプロイし ます。 Alibaba Cloud ECS を有効化する方法の詳細は、「#unique\_19」をご参照ください。 次の図に、このトピックで使用する ECS の設定を示します。

	Configurations	Basic Configurations 🗹				
Selected	Billing Method : Pay-As-You-Go Quantity : 1 Units	Region : Hangzhou Zone H (6) Image : CentOS 7.6 64-bit(Security Enhancement)	Instance Type : General Purpose Type g6 / ecs.g6.large(2vCPU 8GiB) System Disk : Ultra Disk 40GiB			
		Networking 🗹				
		Network Type : VPC Network Billing Method : Pay-By-Traffic 5Mbps	VPC : dtplus_VPC/ vpc-bp wt Security Group : security-group-20190617 / sg-bp1 by5	VSwitch : hangzhouH/ vsw-bp )k/ 1		
		System Configurations 🛛 🗹				
		Logon Credentials : Password	Instance Name : `			
		Save as Launch Template O View Open API				

#### テーブル、フィールド、インデックスの作成

1. ApsaraDB RDS for MySQL にテーブルを作成し、このテーブルにフィールドを追加します。

このトピックでは、テーブル es\_test を作成します。 次の図に、テーブルに追加するフィール ドを示します。

elasticsear	- 2	
< Table	View	Progran 🔪
Enter a tak	ole name	or pa
es_tes	st	
Column (4)		
🗄 count text		
📴 id int(32)		
name text		
color text		
🗄 🗊 Index (1)		

**2.** Elasticsearch インスタンスにインデックスを作成し、マッピングを設定します。

#unique\_15を実行し、[Dev Tools] ページ の [Console] で次のコマンドを実行してインデッ クスを作成し、マッピングを設定します。

```
"mappings" : {
      " doc":{
         properties" : {
          count": {
             "type": "text"
          },
"id": {
             "type": "long"
          },
"name" : {
              "type" : "text"
           ;,
"color" : {
"type" : "text"
        }
     }
   }
}
テーブルとマッピングが正常に作成されると、次の結果が返されます。
  "acknowledged" : true,
"shards_acknowledged" : true,
"index" : "es_test"
```

#### MySQL のインストール

}

- 1. Alibaba Cloud ECS インスタンスに接続します。
- 2. MySQL ソースインストールパッケージをダウンロードします。

wget http://dev.mysql.com/get/mysql57-community-release-el7-11.noarch.rpm

3. MySQL ソースをインストールします。

yum -y install mysql57-community-release-el7-11.noarch.rpm

4. MySQL ソースが正常にインストールされたかどうかを確認します。

yum repolist enabled | grep mysql. \*

MySQL ソースが正常にインストールされると、次の結果が返されます。

```
IrootQVM01 ~]# yum repolist enabled | grep mysql.*mysql-connectors-community/x86_64MySQL Connectors Community118mysql-tools-community/x86_64MySQL Tools Community95mysql57-community/x86_64MySQL 5.7 Community Server364
```

5. MySQL サーバーをインストールします。

yum install mysql-community-server

6. MySQL サービスを起動し、MySQL サービスのステータスを確認します。

systemctl start mysqld.service

#### systemctl status mysqld.service

#### MySQL サービスが起動されていると、次の結果が返されます。



7. ApsaraDB RDS for MySQL データベースに接続します。

# ( ) :

- 次のコマンドを実行して ApsaraDB RDS for MySQL データベースに接続する前に、ECS インスタンスのプライベート IP アドレスを ApsaraDB RDS for MySQL ホワイトリストに追加する必要があります。詳細は、「#unique\_20」をご参照ください。
- Canal を使用するには、MySQL バイナリログモードを有効にする必要があります。
   ApsaraDB RDS for MySQL の場合、このモードはデフォルトで有効になっています。次の コマンドを実行すると、バイナリログモードのステータスを照会できます。

show variables like '%log bin%';

バイナリログモードが有効になっている場合、次の結果が返されます。

I Variable_name       I Value         I log_bin       I ON         I log_bin_basename       I /home/mysql/data3001/mysql/mysql/bin         I log_bin_index       I /home/mysql/data3001/mysql/master-log         I log_bin_trust_function_creators       ON         I log_bin_use_v1_row_events       I ON	mysql> show variables like 'xlog_binx';				
<pre>i log_bin i ON i log_bin_basename i /home/mysql/data3001/mysql/mysql/bin i log_bin_index i /home/mysql/data3001/mysql/master-log i log_bin_trust_function_creators i ON i log_bin_use_v1_row_events i ON</pre>		   U	¦ Variable_name		
i sql_log_bin i UN	-bin r-log-bin.index       	ors   (	<pre>i log_bin i log_bin_basename i log_bin_index i log_bin_trust_function_creators i log_bin_use_v1_row_events i sql_log_bin</pre>		

mysql -h<hostname> -P<port> -u<username> -p<password> -D<database>

変数	説明
<hostname></hostname>	ApsaraDB RDS for MySQL インスタンスのイ ントラネットアドレス。 イントラネットアド レスの情報は、インスタンスの [基本情報] で 確認できます。
変数	説明
-----------------------	--
<port></port>	ApsaraDB RDS for MySQL インスタンス の内部ポート。 デフォルトのポート番号 は、 <b>3306</b> です。 内部ポートの情報は、イン スタンスの [基本情報] ページで確認できま す。
<username></username>	ApsaraDB RDS for MySQL データベースの ユーザー名。アカウント情報は、インスタ ンスの [アカウント管理] ページで確認できま す。使用可能なアカウントがない場合、アカ ウントを作成する必要があります。 詳細は、 「#unique_21」をご参照ください。
<database></database>	ApsaraDB RDS for MySQL データベースの 名前。 データベース名は、インスタンスの [データベース管理] ページで確認できます。 使用可能なデータベースがない場合、データ ベースを作成する必要があります。 詳細は、 「#unique_21」をご参照ください。
<password></password>	ApsaraDB RDS for MySQL データベースのパ スワード。

コマンド例:

mysql -hrm-bp1u1xxxxxxx6ph.mysql.rds.aliyuncs.com -P3306 -ues -pmima -Delasticsearch

ApsaraDB RDS for MySQL に接続されていると、次の結果が返されます。

[root@UM01~]# mysql -hrm-bp h.mysql.rds.aliyuncs.com -P3306 -uj mysql: [Warning] Using a password on the command line interface can be insecure. Reading table information for completion of table and column names You can turn off this feature to get a quicker startup with -A B -Delasticsearch -p] Welcome to the MySQL monitor. Commands end with ; or \g. Your MySQL connection id is 688823 Server version: 5.7.25-log Source distribution Copyright (c) 2000, 2019, Oracle and/or its affiliates. All rights reserved. Oracle is a registered trademark of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners. Type 'help;' or 'Nh' for help. Type 'Nc' to clear the current input statement. nysql> show databases; Database information\_schema elasticsearch mysql panqbi xuyongqb i rows in set (0.00 sec) ysq1>

### JDK のインストール

1. ECS インスタンスに接続し、使用可能な JDK パッケージを照会します。

yum search java | grep -i --color JDK

 バージョンを選択し、JDK をインストールします。 このトピックでは、java-1.8.0-openjdkdevel.x86 64 を選択しています。

yum install java-1.8.0-openjdk-devel.x86\_64

- 3.環境変数を設定します。
  - a) etc フォルダーの profile ファイルを開きます。

vi /etc/profile

b) 次の環境変数をファイルに追加します。

export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.71-2.b15.el7\_2.x86\_64 export CLASSPATH=.:\$JAVA\_HOME/jre/lib/rt.jar:\$JAVA\_HOME/lib/dt.jar:\$JAVA\_HOME /lib/tools.jar export PATH=\$PATH:\$JAVA\_HOME/bin

c):wq と入力してファイルを保存し、vi モードを終了します。続いて、次のコマンドを実行 して変更を適用します。

source /etc/profile

- 4. 次のコマンドを実行して、JDK が正常にインストールされたかどうかを確認します。
  - java
  - javac
  - java -version

JDK が正常にインストールされていると、次の結果が返されます。

```
Iroot@UM01 ~]# java -version
openjdk version "1.8.0_222"
OpenJDK Runtime Environment (build 1.8.0_222-b10)
OpenJDK 64-Bit Server UM (build 25.222-b10, mixed mode)
```

### Canal サーバーのインストールと起動

1. ECS インスタンスに接続し、Canal deployer パッケージをダウンロードして解凍します。 こ

のトピックでは、Canal deployer V1.1.4 を使用しています。

wget https://github.com/alibaba/canal/releases/download/canal-1.1.4/canal. deployer-1.1.4.tar.gz

**2.** canal.deployer-1.1.4.tar.gz パッケージを解凍します。

tar -zxvf canal.deployer-1.1.4.tar.gz

**3.** conf/example/instance.properties ファイルを変更します。

vi conf/example/instance.properties



パラメーター	説明
canal.instance.master.address	<apsaradb for="" instance="" internal<br="" mysql="" rds="">endpoint&gt;:<internal port="">。必要な情報 は、ApsaraDB RDS for MySQL インスタンスの [基本 情報] ページで確認できます。例:rm-bp1u1xxxxx xxxx6ph.mysql.rds.aliyuncs.com:3306</internal></apsaradb>

パラメーター	説明
canal.instance.dbUsername	ApsaraDB RDS for MySQL データベースのユーザー 名。 アカウント情報は、 ApsaraDB RDS for MySQL インスタンスの [アカウント管理] ページで確認できま す。
canal.instance.dbPassword	ApsaraDB RDS for MySQL データベースのパスワード。

4.:wg と入力してファイルを保存し、vi モードを終了します。

### 5. Canal サーバーを起動し、Canal ログを表示します。

./bin/startup.sh cat logs/canal/canal.log

Iroot@UM01 ~ # ./bin/startup.sh cd to /root/bin for workaround relative path LOG CONFIGURATION : /root/bin/../conf/logback.xml

### Canal adapter のインストールと起動

1. ECS インスタンスに接続し、Canal adapter パッケージをダウンロードして解凍します。 こ

のトピックでは、Canal adapter V1.1.4 を使用しています。

wget https://github.com/alibaba/canal/releases/download/canal-1.1.4/canal. adapter-1.1.4.tar.gz

**2.** canal.adapter-1.1.4.tar.gz パッケージを解凍します。

tar -zxvf canal.adapter-1.1.4.tar.gz

3. conf/application.yml ファイルを変更します。

vi conf/application.yml



パラメーター	説明
canal.conf.canalServerHost	Canal deployer のエンドポ イント。デフォルトの設定 127.0.0.1:11111 を使用しま す。
canal.conf.srcDataSources.defaultDS.url	jdbc:mysql:// <apsaradb RDS for MySQL instance internal endpoint&gt;:&lt; internal port&gt;/<database name&gt;? useUnicode=true 。必要な情報は、ApsaraDB RDS for MySQL インスタ ンスの [基本情報] ページで 確認できます。例:jdbc: mysql://rm-bp1xxxxxx xxnd6ph.mysql.rds.aliyuncs .com:3306/elasticsearch? useUnicode=true</database </apsaradb 
canal.conf.srcDataSources.defaultDS.username	ApsaraDB RDS for MySQL データベースのユーザー 名。 アカウント情報 は、ApsaraDB RDS for MySQL インスタンスの [基 本情報] ページで確認できま す。
canal.conf.srcDataSources.defaultDS.password	ApsaraDB RDS for MySQL データベースのパスワード。
canal.conf.canalAdapters.groups.outerAdapters.hosts	name:es を見つけ、 <b>hosts</b> を <alibaba cloud="" elasticsea<br="">rch instance internal network address&gt;:<internal network port&gt; に置き 換えます。必要な情報 は、Elasticsearch インスタ ンスの [#unique_22] ペー ジで確認できます。例:es -cn-v64xxxxxxx3medp. elasticsearch.aliyuncs.com: 9200</internal </alibaba>
canal.conf.canalAdapters.groups.outerAdapters.mode	値を rest に設定します。

パラメーター	説明	
canal.conf.canalAdapters.groups.outerAdapters. properties.security.auth	値を <alibaba cloud<br="">Elasticsearch instance username&gt;:<password> に 設定します。例:elastic: es_password</password></alibaba>	
canal.conf.canalAdapters.groups.outerAdapters. properties.cluster.name	Alibaba Cloud Elasticsearch インスタンスの ID。 ID は、Elasticsearch インスタ ンスの [#unique_22 ] ページ で確認できます。 例:es-cn- v64xxxxxxx3medp	

- 4.:wq と入力してファイルを保存し、vi モードを終了します。
- **5.** 同じ手順で conf/es/\*.yml ファイルを変更し、ApsaraDB RDS for MySQL から Alibaba Cloud Elasticsearch にマッピングするフィールドを指定します。

dataSourceKey: defaultDS
destination: example
groupId: g1
esMapping:
_index: es_test
_type: _doc
_id: _id
#pk: id
sql: "select t.id as _id,t.id,t.count,t.name,t.color from es_test t"
commitBatch: 3000

パラメーター	説明
esMappingindex	この値には、「テーブル、フィールド、インデックスの作 成」で Elasticsearch インスタンスに作成したインデックス 名を設定します。 このトピックでは、値に <b>es_test</b> を設定し ています。
esMappingtype	この値には、「テーブル、フィールド、インデックスの作 成」で Elasticsearch インスタンスに作成したインデックス タイプを設定します。 このトピックでは、値に <b>_doc</b> を設定 しています。
esMappingid	Elasticsearch インスタンスに同期させるドキュメントの ID。 独自のドキュメント ID を指定できます。 このトピック では、値に <b>_id</b> を設定しています。

パラメーター	説明
esMapping.sql	Elasticsearch インスタンスに同期させるフィールドを取得 するための SQL 文。 このトピックでは、select t.id as _id,t. id,t.count,t.name,t.color from es_test t; 文を使用していま す。

6. Canal adapter サービスを起動し、Canal ログを表示します。

```
./bin/startup.sh
cat logs/adapter/adapter.log
```

Canal adapter サービスが起動されていると、次の結果が返されます。

2019-09-05		[Thread-2] IN	0 com.alibaba.druid.pool.DruidDataSource - {dataSource-2} inited
2019-09-05	20:16:24.928	[Thread-2] IN	0 c.a.o.canal.adapter.launcher.loader.CanalAdapterService - ## start the canal client adapters.
2019-09-05	20:16:24.929	[Thread-2] IN	0 c.a.o.canal.adapter.launcher.loader.CanalAdapterLoader - Load canal adapter: logger succeed
2019-09-05	20:16:24.929	[Thread-2] IN	0 c.a.o.canal.client.adapter.es.config.ESSyncConfigLoader - ## Start loading es mapping config
2019-09-05	20:16:24.943	[Thread-2] IN	0 c.a.o.canal.client.adapter.es.config.ESSyncConfigLoader + ## ES mapping config loaded
2019-09-05		[Thread-2] IN	0 c.a.o.canal.adapter.launcher.loader.CanalAdapterLoader - Load canal adapter: es succeed
2019-09-05		[Thread-2] IN	0 c.a.o.canal.adapter.launcher.loader.CanalAdapterLoader - Start adapter for canal instance: example succeed
2019-09-05		[Thread-2] IN	0 c.a.o.canal.adapter.launcher.loader.CanalAdapterService - ## the canal client adapters are running now
2019-09-05		[Thread-2] IN	0 c.a.o.c.a.launcher.monitor.ApplicationConfigMonitor - ## adapter application config reloaded.
2019-09-05		[Thread-8] IN	0 c.a.o.canal.adapter.launcher.loader.CanalAdapterWorker - ========>> Start to connect destination: example <====================================
2019-09-05		[Thread-8] IN	0 c.a.o.canal.adapter.launcher.loader.CanalAdapterWorker - =======>> Start to subscribe destination: example <================>
2019-09-05	20:16:25.232	[Thread-8] IN	0 c.a.o.canal.adapter.launcher.loader.CanalAdapterWorker - ==========> Subscribe destination: example succeed <===================================
2019-09-05		[pool-5-threa	-l] INFO c.a.o.canal.client.adapter.logger.LoggerAdapterExample - DML: {"data":null,"database":"mysql","destination":"example"
1,"sql":"/	rds internal	mark */ CREA	E TABLE IF NOT EXISTS mysql.ha_health_check (\n id BIGINT DEFAULT 0,\n type CHAR(1) DEFAULT '0',\n PRIMARY KEY (type)\n)\

### Canal の増分データ同期機能の検証

**1.** ApsaraDB RDS for MySQL データベースのテーブル **es\_test** に対して、データの追加、変更、 または削除を行います。

insert `elasticsearch`.`es\_test`(`count`,`id`,`name`,`color`) values('11',2,'canal\_test2','
red');

**2.** Alibaba Cloud Elasticsearch コンソールにログインし、#unique\_15を実行します。

**3.** Kibana コンソールの [**Dev Tools**] ページの [**Console**] で、同期対象の増分データを照会しま す。

GET /es\_test/\_search

増分データが正常に同期されていると、次の結果が返されます。

Console	Search Profiler	Grok Debugger			
1 GET /e	s_test/_search		<i>ب</i> :	1 • { 2 3 4 • 5 6 7 8 9 • 10 • 11 12 13 • 14 • 15 16 17 18 19 • 20 21 22 23 24 • 25 • 26 • 27 28 29 29	<pre>"took" : 1, "timed_out" : false, "_shards" : {     "total" : 5,     "successful" : 5,     "skipped" : 0,     "failed" : 0 }, "hits" : {     "total" : 1,     "max score" : 1.0,     "hits" : [     {         [</pre>

# 3 データ移行

## 3.1 ユーザー作成の Logstash インスタンスを使用して、データを Alibaba Cloud Elasticsearch と同期

Logstash は、データソースからデータを動的に取り込むことができるオープンソースのデータ 取り込みエンジンです。 Logstash では、カスタムルールを使用して、さまざまなデータソース から取り込まれたデータをフィルタリングし、そのデータをターゲットサービスに出力できま す。 このトピックでは、Elastic Compute Service (ECS) に Logstash をデプロイする方法につい て説明し、Logstash から Alibaba Cloud Elasticsearch (ES) にデータを移行する例を示します。

### 前提条件

Logstash をデプロイして使用する前に、Alibaba Cloud ECS インスタンスと Elasticsearch イン スタンスを作成し、設定する必要があります。

Alibaba Cloud Elasticsearch インスタンスの作成と設定

**1.** #unique\_9を実行します。

この例では、Elasticsearch インスタンスのバージョンは V6.7.0 です。 次の図に Elasticsea rch インスタンスの設定を示します。

Instance ID: es-cn-45914gy2	Created At: May 10, 2019, 18:01:23
Name: keepit-vpclp Edit	Status:      Active
Regions: China (Hangzhou)	Zone: cn-hangzhou-b
Version: 6.7.0	Billing Method: Pay-As-You-Go
Instance Type: Standard	Protocol: HTTP Edit
VPC: vpc-bp1 cz6d	VSwitch: vsw-bp18bzk
Internal Network Address: es-cn-4 <sup>c</sup> elasticsearch.aliyuncs.com	Internal Network Port: 9200
Public Network Access: es-cn-4: public.elasticsearch.aliyuncs.com	Public Network Port: 9200

2. Alibaba Cloud Elasticsearch コンソールにログインし、自動インデックスを有効化します。

- **3.** Elasticsearch インスタンスの Kibana コンソールにログインし、読み書き権限が付与されて いるロールを追加します。logstash-\*
  - **a.** #unique\_15を行います。
  - **b.** [Management] > [Roles] > [Create role] を選択します。

	kibana	€ Elasticsearch				
	inbana	Index Management	Roles			
Ø	Discover	Index Lifecycle Policies Apply roles to groups of users and manage permissions across the stack Rollup Jobs				
£	Visualize	Cross Cluster Replication Remote Clusters	Q Search	+ Create role		
50	Dashboard	Watcher	Role †	Reserved 🔞		
V	Timelion	7.0 Upgrade Assistant	apm_system	~		
ŵ	Canvas	Kibapa	beats_admin	~		
\$	Maps	Index Patterns	beats_system	~		
ø	Machine Learning	Saved Objects Spaces	ingest_admin	~		
	Infrastructure	Reporting	kibana_dashboard_only_user	✓		
		Advanced Settings	kibana_system	<b>~</b>		
I	Logs		kibapa usor			
	ΔΡΜ	🖡 Logstash	Kibana_usei	•		
		Pipelines	logstash_admin	×		
্ত	Uptime		logstash_role			
*	Graph	Beats	logstash_system	~		
6	Dev Tools	Central Management	machine learning admin	~		
Ч	Devitoois	(ii) Security				
æ	Monitoring	Lisers	machine_learning_user	✓		
æ	Management	Roles	monitoring_user	*		
- 25	management		remote_monitoring_agent	✓		

c. [Create role] ページで、関連するパラメーターを設定します。

Role name			
logstash_role			
A role's name cannot be changed once it has been created.			
Cluster privileges			
Manage the actions this role can perform against your :luster. Learn more		~	
Run As privileges			
Now requests to be submitted on the behalf of other users. Learn more	Add a user	~	
ndex privileges			
Control access to the data in your cluster. Learn more			
ndices	Privileges	Granted fields (optional)	
logstash-* × ⊗ ∨	read ×     create ×     delete ×     write ×       create_index ×     ⊗	*×	8

パラメーター	説明
Role name	ロールの名前。
Indices	移行するインデックスファイル。 logstash-* を入力しま す。

パラメーター	説明					
Privileges	ロールに付与する権限。 read、write、create、delete、 create_index 権限を追加します。					
Granted fields	ロールがアクセスできるフィールド。 このパラメーターは オプションです。 この例では、 <b>*</b> を入力します。					

d. [Create role] をクリックしてロールを作成します。

Alibaba Cloud ECS インスタンスの作成と設定

#unique\_19を実行します。 ECS インスタンスが Logstash インスタンスと Elasticsearch インスタンスにアクセスできることを確認してください。 要件を満たす購入済み ECS インスタンスを使用することもできます。

次の図に ECS インスタンスの設定を示します。

i-bp11k		n	k	Hangshou Zono H	Pupping	VDC	2 vCPU 8 GiB (I/O Optimized)	Pay-As-You-Go
actor_9cfdd9c0 🖍	*/	<b>• a</b> <sub>2</sub>	-	nangzhoù zone n	Okunning	VPC	ecs.g6.large 25Mbps (Peak Value)	Created at January 9, 2020, 12:04

🗎 注:

Elasticsearch インスタンスと同じゾーンかつ同じ VPC ネットワークの ECS インスタンスを 購入することを推奨します。 クラシックネットワークに接続された ECS インスタンスを購 入することもできます。 ただし、クラシックネットワークが Elasticsearch インスタンスの VPC ネットワークに接続されていることを最初に確認する必要があります。

2. ECS インスタンスに JDK をインストールします。 JDK のバージョンは、V1.8 以降でなければ なりません。

JDK をインストールする方法の詳細については、「JDK のインストール」をご参照ください。

### Logstash のインストール

**1.** Logstash V6.7.0 をダウンロードします。

公式 Elasticsearch Web サイトにアクセスし、Elasticsearch バージョンと同じバージョンの Logstash をダウンロードします。 Logstash V6.7.0 をダウンロードすることを推奨します。

wget https://artifacts.elastic.co/downloads/logstash/logstash-6.7.0.tar.gz

2. Logstash パッケージを解凍します。

tar -xzvf logstash-6.7.0.tar.gz

### Logstash での増分データの同期

1. ECS インスタンスに接続してから、Logstash ディレクトリに切り替えます。

```
cd logstash-6.7.0
```

2. test という名前の .conf ファイルを作成します。

touch test.conf

3. test.conf ファイルを設定します。以下に例を示します。

```
input {
    file {
        path => "/your/file/path/xxx"
        }
}
filter {
    output {
     elasticsearch {
        hosts => ["http://instanceId.elasticsearch.aliyuncs.com:9200"]
        user => "user-name"
        password => "logstash-password"
    }
}
```

パラメーター	説明
path	ログファイルのパス。 この例では、パスは /var/log/ meaasges です。
hosts	Elasticsearch インスタンスのエンドポイント。 instanceld を Elasticsearch インスタンスの ID に置き換えます。 Elasticsearch インスタンスの [基本情報] ページでインス タンス ID を確認できます。 例:http://es-cn-45xxxxxxxx xxxxju.elasticsearch.aliyuncs.com:9200

パラメーター	説明
user	Elasticsearch インスタンスのユーザー名。 デフォルトの ユーザー名は、elastic です。 カスタムアカウントを使用す る場合、まずアカウントのロールを作成し、そのロールに 必要な権限を付与する必要があります。 詳細については、 「ユーザーの作成」と「ロールの作成」をご参照ください。
	: ユーザー名は、二重引用符 (") のペアで囲む必要がありま す。これにより、Logstash の起動時にユーザー名に含まれ る特殊文字が原因で発生するエラーを回避できます。
password	Elasticsearch インスタンスのパスワード。 パスワードは、 二重引用符 (") のペアで囲む必要があります。 Elasticsearch インスタンスの作成後にパスワードを変更することもできま す。
	:

Logstash には、さまざまな入力、フィルター、出力のブラグインがあります。 これらのプ ラグインは、データの簡単な取り込み、変換、出力に役立ちます。 詳細については、「設定 ファイルの構造」をご参照ください。

4. logstash コマンドを実行します。

手順 3 に従って、.conf ファイルを設定し、logstash コマンドを実行します。

bin/logstash -f test.conf

コマンドが実行されると、システムは自動的に Logstash を使用してログファイルから新しい コンテンツを取り込み、Elasticsearch インスタンスに出力します。 Logstash は、ログファ イル内の変更を Elasticsearch インスタンスにインデックス付けします。 5. 結果を確認します。

同期結果を確認する前に、ロールを作成し、**logstash-\***インデックスの管理をロールに許可 する必要があります。 必要な権限は、read、write、create、delete、create\_index です。 詳細については、「ロールの作成」をご参照ください。

- a. Elasticsearch インスタンスの Kibana コンソールにログインします。
- b. 左側のナビゲーションペインで、[Dev Tools] を選択します。
- c. [Dev Tools] ページの [Console] タブで、次のコマンドを実行します。

GET /logstash-\*/\_search

コマンドの実行後、次の結果が返されます。

Console Search Profiler Grok	Debugger	
1 GFT /logstash-*/ search	ی 🖌	1 + 1
	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	<pre>"took" : 5, "timed_out" : 5, "timed_out" : 6, "stand" : { "stand" : 10, "stoped" : 0, "stoped" : 0, "failed" : 0 "hits" : { "total" : 13, "max_score" : 1.0, "hits" : {</pre>

### Logstash ノードのモニタリング

Logstash ノードをモニタリングし、モニタリングデータを収集するには、次の手順に従います。

1. Logstash ディレクトリの config フォルダーに切り替えます。

cd /logstash-6.7.0/config

2. logstash.yml ファイルを設定します。

logstash.yml ファイルを開きます。

vim logstash.yml

logstash.yml ファイルの次のパラメーターの注釈を削除し、これらのパラメーターに値を割 り当てます。

xpack.monitoring.enabled: true xpack.monitoring.elasticsearch.username: """ xpack.monitoring.elasticsearch.password: """ xpack.monitoring.elasticsearch.hosts: ["http://es-cn-45 xpack.monitoring.elasticsearch.aliyuncs.com:9200"]

パラメーター	説明
xpack.monitoring.enable	このパラメーターを true に設定します。 デフォルト値は false です。
xpack.monitoring. elasticsearch.username	Logstash をモニタリングするユーザーを作成します。 詳細 については、「ユーザーの作成」をご参照ください。
	elastic アカウントを使用することもできます。ただし、シ ステムのセキュリティリスクが生じる可能性があるため、 本番環境では elastic アカウントを使用しないことを推奨し ます。
xpack.monitoring. elasticsearch.password	Logstash のモニタリングに使用されるユーザーのパスワー ド。
xpack.monitoring. elasticsearch.hosts	Elasticsearch インスタンスのエンドポイント。 例:http:// es-cn-45xxxxxxxxxxxyu.elasticsearch.aliyuncs.com:9200 。

3. Logstash のディレクトリに戻り、Logstash サービスを起動します。

cd ../ bin/logstash -f test.conf

Logstash の起動後、次の結果が返されます。

[2019-10-24T17:34:41,007][INFO ][logstash.agent [2019-10-24T17:34:41,238][INFO ][logstash.agent

Pipelines running {:count=>2, :running pipelines=>[:".mon Successfully started Logstash API endpoint {:port=>9601} **4.** Elasticsearch インスタンスの Kibana コンソールにログインします。 左側のナビゲーション ペインで [Monitoring] を選択して、Logstash モニタリングデータを表示します。

	kibana	es-cn-45 ju										
Ø	Discover	Elasticsearch • Health is green Platinum license will expire on April 1, 2021										
佡	Visualize	Overview Nodes: 2 Indices: 24										
50	Dashboard											
₽	Timelion	Version 6.7.0	Disk Available         86.01%           33.7 GB / 39.1 GB         39.1 GB	Documents 1,677,964								
寙	Canvas	Jobs 0	JVM Heap 65.60%	Primary Shards 40								
8	Maps			Replica Shards 40								
Q9	Machine Learning											
G	Infrastructure	📕 Kibana • Health is green										
I	Logs											
G	АРМ	Overview	Instances: 1									
୍ତ	Uptime	Requests 1	Connections 0									
÷	Graph	Max. Response Time 89 ms	Memory Usage 13.46% 195.9 MB / 1.4 GB									
₿.	Dev Tools											
ŵ	Monitoring	🚛 Logstash										
٢	Management											
		Overview Nodes: 1 Pipelines: 1 A										
2	elastic	Events Received 1	Uptime 11 minutes	With Memory Queues 1								
Β	Logout	Events Emitted 1	JVM Heap 22.98%	With Persistent Queues 0								
D	Default		231.5 MB / 1,007.4 MB									

### ユーザーの作成

このセクションでは、CLI または Kibana コンソールから Logstash をモニタリングするユーザー の作成方法について説明します。

(!)

デフォルトでは、logstash\_system ユーザーは無効化されているので、logstash\_system ユー ザーを作成できません。 したがって、ogstash\_system ロールを引き受けるユーザーを作成す る必要があります。

CLI でのユーザーの作成

ECS インスタンスに接続し、次のコマンドを実行してユーザーを作成します。

curl -u elastic:es-password -XPOST http://instanceId.elasticsearch.aliyuncs.com:9200/ \_xpack/security/user/logstash\_system\_monitor -d '{"password" : "logstash-monitorpassword","roles" : ["logstash\_system"],"full\_name" : "your full name"}'

パラメーター	説明
es-password	Elasticsearch インスタンスのパスワード。 このパスワード は、Kibana コンソールへのログインにも使用されます。

パラメーター	説明
instanceld	Elasticsearch インスタンスの ID。 Elasticsearch インスタン スの [基本情報] ページでインスタンス ID を確認できます。
logstash-monitor-password	logstash_system_monitor ユーザーのパスワード。
your full name	ユーザーのフルネーム。

ユーザーが作成されると、次の結果が返されます。

Kibana コンソールでのユーザーの作成

**1.** #unique\_15を行います。

6 /]# curl -H "Content-Type: application/json"
system"]."full name" : "your full name"}'
":("created":true)."created":true)[root@6 /]#

2. [Management] > [Users] > [Create new user] を選択します。

3 -XPOST http://es-cn-45

	kibana	€ Elasticsearch					
	Discover	Index Management Index Lifecycle Policies	Users				Create new user
	Visualize	Rollup Jobs Cross Cluster Replication	Q Search				
50	Dashboard	Remote Clusters Watcher	Full Name 个	User Name	Email Address	Roles	Reserved
V	Timelion	License Management 7.0 Upgrade Assistant			No items found		
ŵ	Canvas	📕 Kibana	Rows per page: 20 🗸				
\$	Maps	Index Patterns					
۲	Machine Learning	Saved Objects Spaces					
G	Infrastructure	Reporting Advanced Settings					
ľ	Logs	Logstach					
G	APM	Pipelines					
্ত	Uptime						
÷	Graph	Beats Central Management					
铅	Dev Tools						
æ	Monitoring	Security     Users					
۲	Management	Roles					

3. [New user] ページで、ユーザー情報を入力します。

Jsername		
logstash_system_m	nonitor	
Password		
Confirm password		
•••••		
Full name		
Email address		
Roles		
logstash_system ×		⊗ ~

パラメーター	説明
Username	ユーザーの名前。 カスタム名を入力することができます。 こ のセクションでは、ユーザー <b>logstash_system_monitor</b> を 作成して、Logstash をモニタリングします。
Password	ユーザーのパスワード。

パラメーター	説明
Confirm password	確認用のパスワードをもう一度入力します。
Full name	ユーザーのフルネーム。 このパラメーターはオプションで す。
Email address	ユーザーのメールアドレス。 このパラメーターはオプション です。
Roles	ユーザーが引き受けるロール。 logstash_system ロールを 指定します。

4. [Create user] をクリックして、ユーザーを作成します。

### FAQ

 Q: Logstash を使用して Elasticsearch にデータを出力する前に、Elasticsearch インスタン スの自動インデックスを有効にする必要があるのはなぜですか。

A:データのセキュリティを確保するため、Alibaba Cloud Elasticsearch はデフォルトで自動 インデックスを無効にしています。

Logstash は Create index 操作を呼び出さずに、データを Elasticsearch に送信します。 その 後、Elasticsearch はデータに自動的にインデックスを付けます。 したがって、Logstash を使 用してデータを Elasticsearch に出力する前に、Elasticsearch インスタンスの自動インデック スを有効にする必要があります。



## (!)

自動インデックスを有効にし、操作を確定すると、Elasticsearch インスタンスは再起動され ます。操作を確定する前に、インスタンスで実行されているワークロードが悪影響を受けな いことを確認してください。 • Q:インデックスを作成する権限がないというメッセージが表示された場合、どうすればよいですか。



A:データの受信に使用される Elasticsearch アカウントのロールを確認する必要がありま す。 ロールに write、delete、create index 権限があることを確認してください。

• Q:メモリ不足のエラーを解決するにはどうすればよいですか。

Java HotSpot(TM) 64-Bit Server VM warning: INFO: os::commit\_memory(0x00000006c5330000, 986513408, 0) failed; error='Cannot allocate memory' (errno=12) # # There is insufficient memory for the Java Runtime Environment to continue. # Native memory allocation (mmap) failed to map 986513408 bytes for committing reserved memory. # An error report file with more information is saved as:

A: デフォルトでは、1 GB のメモリが Logstash に割り当てられています。 購入した ECS イ ンスタンスに Logstash 用のメモリが不足している場合、config/jvm.options を変更して Logstash に割り当てられるメモリ量を減らします。

 Q:Logstash コマンドの実行時に次のエラーが発生します。このエラーを解決するにはどう すればよいですか。

[rootHtZbpldroc0yv90nH95ae6z2 logstash-5.5.3]# bin/logstash -f task/test.com ERROR StatusLogger No log4j2 configuration file found. Using default configuration: logging only errors to the console. Sending Logstash's logs to /root/tmp/logstash-5.5.3/aw which is non configured via log4j2.properties [2017-12-01TJS:18:02.04/[ERROR]]logstash.agent ] Cannot create pipeline {reason="Expected one of #, {, } at line 12, column 22 (byte 261) after output {\n elasticsearch {\n hosts ~ [\n"hi p;//sected one of #, {, } at line 12, column 22 (byte 261) after output {\n elasticsearch {\n hosts ~ [\n"hi p;//sected one of #, {, } at line 12, column 22 (byte 261) after output {\n elasticsearch {\n hosts ~ [\n"hi p;//sected one of #, {, } at line 12, column 22 (byte 261) after output {\n elasticsearch {\n hosts ~ [\n"hi [2017-12-01TJS:18:02.465][INF0][logstash.outputs.elasticsearch] Elasticsearch pool URLs updated {:changes~{:removed~[, :added~[Intp://logstash\_system\_monitor:xxxxxxee=cn-mp90cbs/l002e]btn.elasticsearch]

A:test.confファイル内のユーザー名とパスワードに特殊文字が含まれていないことを確認す る必要があります。 ユーザー名かパスワードに特殊文字が含まれている場合、二重引用符 (") のペアで囲む必要があります。

# 4 ビッグデータの同期

### 4.1 データの同期と移行

### 4.1.1 クラウドデータのインポート

### Alibaba Cloud から ES へのデータのインポート (オフライン)

Alibaba Cloud は、豊富なクラウドストレージとデータベースプロダクトを提供しています。 これらのプロダクトのデータの分析や検索をする場合、Data Integration を使用してくださ い。Data Integration では、オフラインデータを**5**分ごとに Elasticsearch に同期させることが できます。

### サポートされているデータソース

- Alibaba Cloud データベース (MySQL、PostgreSQL、SQL Server、PPAS、MongoDB、 HBase)
- DRDS
- MaxCompute
- OSS
- Table Store
- 自社開発の HDFS、Oracle、FTP、DB2、および以前のクラウドデータベースの自社開発バージョン

**三**注:

データ同期により、パブリックネットワークのトラフィック料金が発生することがあります。

手順

オフラインデータをインポートするには、次の手順を実行します。

- VPC 内で Elasticsearch とやり取り可能な ECS インスタンスを準備します。 この ECS イン スタンスはデータソースを取得し、ES データを書き込むジョブを実行します (このジョブは Data Integration によって一元的に実行されます)。
- Data Integration サービスを有効化して、実行可能ジョブリソースとして ECS インスタンスを Data Integration サービスに登録する必要があります。
- データ同期スクリプトを設定し、定期的に実行させます。

ステップ

 Elasticsearch サービスと同じ VPC にある ECS インスタンスを購入します。 パブリック IP ア ドレスを ECS インスタンスに割り当てるか、ECS インスタンスの Elastic IP アドレスを有効に します。 コストを抑えるため、既存の ECS インスタンスを使用できます。 ECS インスタンス の購入方法は、「#unique\_19」をご参照ください。



- CentOS 6、CentOS 7、AliyunOS を推奨します。
- 追加された ECS インスタンスで MaxCompute または同期タスクを実行する必要がある 場合、ECS インスタンスの現在の Python バージョンが 2.6 または 2.7 かどうかを確認し ます (CentOS 5 の Python バージョンは 2.4 ですが、他のオペレーティングシステムの Python バージョンは 2.6 以降)。
- ECS インスタンスにパブリック IP アドレスが割り当てられていることを確認します。
- 2. Data Integration コンソールにログインして、ワークベンチを開きます。

Data Integration または DataWorks が有効化されている場合、次のように表示されます。

	Overview	Workspaces	Resources	Compute Eng	jines		
🜀 DataWorks	DataStudio •Data Inte	gration∙MaxCor	mpute	* (C	ð	0	
Shortcuts						The Data Integration Launch Support multiple development modes	3 <b>*</b> 2
Data Analytics	Data Integration	Maintenance C	Center	D	ata Service	Support more data channels	M.
Workspaces					All Workspaces		
asdjhasuhschj64273 China Er	MaxCompute_DOC	Asia Pacific SE 1	test012		China East 2		
Created At:Mar 15, 2018, 16:47:55 Compute Engines:None Services:Data Studio Data Integration Data Managemen	Created At:Jan 21, 2019, 23:1 Compute Engines:MaxCompu t DServices:Data Studio Data Into	3:23 te egration Data Management O	Created At: Compute Er Services:Da	lan 02, 2018, 15:36: gines:MaxCompute ta Studio Data Integ	40 • PAI calculation engine rration Data Management D		
Workspace Settings Data Analytics	Workspace Settings	Data Analytics	Works	ace Settings	Data Analytics		
Data Service Data Integration	Data Integration		Dat	a Service	Data Integration		
Commonly Used Features							

Data Integration または DataWorks が有効化されていない場合、次のメッセージが表示され ます。 手順に従って、Data Integration サービスを有効化します。 有料サービスなので、見 積もり価格を予算と照らし合わせてチェックしてください。  Data Integration サービスの [Project Management-Scheduling Resource Management] ページに移動し、VPC の ECS インスタンスをスケジューリングリソースとして設定します。 詳細は、「#unique\_32」をご参照ください。

Resource Groups Search	by resource group name.				Add Resource Group
Resource Group Name	Network Type	Server	Used DMU	Billing Method	Actions
Default resource group				Pay-As-You-Go	

4. Data Integration サービスでデータ同期スクリプトを設定します。 設定手順は、

「#unique\_12」をご参照ください。 Elasticsearch の設定方法は、「#unique\_14」をご参照 ください。

# 📋 注:

- 同期スクリプトの設定は3つの部分に分けられます。Readerは、アップストリームデー タソース(データ同期のクラウドプロダクト)の設定、WriterはESの設定、settingはパ ケットロス率や最大同時性などの同期の設定です。
- ES Writer の accessId と accessKey は、それぞれ Elasticsearch のユーザー名とパスワー ドです。
- 5. スクリプトを設定したら、データ同期ジョブを送信します。 ジョブの実行サイクルを設定 し、[OK] をクリックします。

🧾 注:

- 定期スケジュールを設定する場合、このポップアップウィンドウでジョブ開始時間、実行 間隔、ジョブライフサイクルなどのパラメーターを設定します。
- 設定したルールに従って、翌日の 00:00 に定期ジョブが実行されます。
- 送信後、[O&M Center-Task Scheduling] ページに移動して送信されたジョブを見つけ、デフォルトのスケジューリングリソースから、設定したスケジューリングリソースに変更します。

### リアルタイムデータのインポート

この機能は現在開発中です。今後利用可能になる予定です。

## 4.1.2 DataWorks による MaxCompute と Elasticsearch 間のデー タ同期

Alibaba Cloud は、幅広いクラウドストレージとデータベースサービスを提供します。 これらの サービスに保存されているデータの分析や検索をする場合、Data Integration を使用してデータ を Elasticsearch にレプリケートしてから、データのクエリや分析をします。 Data Integration を使用すると、最小 5 分間隔でデータをレプリケートできます。

データのレプリケートにより、パブリックネットワークトラフィックが生成され、課金が発生す る可能性があります。

#### 前提条件

オンプレミスデータの分析や検索をするには、次の手順に従います。

 #unique\_34、およびデータのインポートを行います。 Hadoop から MaxCompute ヘデータ を移行してから、データを同期します。 この例では、以下のテーブルスキームとデータを使 用します。

Column Name	Туре
create_time	STRING
category	STRING
brand	STRING
buyer_id	STRING
trans_num	BIGINT
trans_amount	DOUBLE
click_cnt	BIGINT
pt	STRING

create time	category	brand	buver id	trans num	trans amount	click ont	nt
2018-08-21 00:00:00	is in	A	null	null	null	null	1
2018-08-22 00:00:00	2.9	<b>Ø</b> #≢B	null	null	null	null	1
2018-08-22 00:00:00	2100	<b>₩</b> RC	null	null	null	null	1
	1269	A	null	null	null	null	1
2018-08-22 00:00:00	2.9	(iiikit)D	null	null	null	null	1
2018-08-23 00:00:00	2500	<b>₫</b> ₩B	null	null	null	null	1
2018-08-23 00:00:00	181	A	null	null	null	null	1
2018-08-23 00:00:00	治療	(DARE	null	null	null	null	1
2018-08-24 00:00:00	3157	₩₩G	null	null	null	null	1
2018-08-24 00:00:00	1239	- ARE	null	null	null	null	1
2018-08-24 00:00:00	州中	<b>B</b> ##A	null	null	null	null	1
2018-08-24 00:00:00	22.98	∰⊯G	null	null	null	null	1
2018-08-24 00:00:00	10.00	LC.	null	null	null	null	1

 Data Integration でレプリケートされたデータを保存するための Elasticsearch インスタンス を作成します。

- Elasticsearch と同じ VPC を共有する ECS インスタンスを購入します。 この ECS インスタン スでデータを取得し、Elasticsearch タスクを実行します (これらのタスクは、Data Integratio n によって送信されます)。
- Data Integration を有効化し、タスクを実行可能なリソースとして ECS インスタンスを Data Integration に登録します。
- データ同期スクリプトを設定し、定期的に実行します。

#### 手順

- 1. Elasticsearch インスタンスと ECS インスタンスの作成
  - **a.** #unique\_37を行います。 この例では、VPC を中国 (杭州) リージョンに作成します。 イン スタンス名は es\_test\_vpc、VSwitch 名は es\_test\_switch です。
  - **b.** Elasticsearch コンソールにログインし、Elasticsearch インスタンスを作成します。



前の手順で作成した VPC と同じリージョン、VPC、 VSwitch を選択するようにします。

	Subscription	Pay-As-You-Go					
region	Region	China (Hangzhou) China (Hong Kong) Asia Pacific SE 5 (Jakarta)	China (Beijing) US West 1 (Silicon Valley) China North 1 (Qingdao)	China (Shanghai) Asia Pacific SE 3 (Kuala Lumpur)	China (Shenzhen) Germany (Frankfurt)	Asia Pacific SOU 1 (Mumbai) Japan	Asia Pacific SE 1 (Singapore) 亚太东南 2 (澳大 利亚)
	Zone	Hangzhou Zone	F 🔹				
	Version	5.5.3 with X- Pack	6.3 with X-Pack				
	Network Typ e	VPC					
	VPC	tiel Create VPC/Subn	▼ et (Switch).Refre	sh the page after	the creation is c	omplete	
	VSwitch	Select a VSwitch	•				
	Instance Ty pe	1Core2G		for testing only 1	tie not quitable i	for the production	environment and
		is excluded from t	the SLA after-sal	es guarantee.	t is not suitable i	for the production	environment and

c. Elasticsearch インスタンスと同じ VPC にある ECS インスタンスを購入し、パブリック IP アドレスを割り当てるか、EIP を有効化します。 コストを節約するため、要件を満たす既 存の ECS インスタンスを使用することを推奨します。

この例では、杭州 (中国東部1) ゾーン F に ECS インスタンスを作成します。 [64-bit CentOS 7.4] と [パブリック IP の割り当て] を選択して、ネットワークを設定します (次 図)。

<ul> <li>How to Select a Network</li> </ul>	ann_hait_ap: If you need to create	O Private IP Addresses A     a new VPC, you can Go to Console and Create >	nalable 250,			
		VPC: ====:/msi.gov/ope-logical-solid-groundid-groundid-logical VSwitch Zone: China East 1 Zone F	VSwitch: env.jmit, anith's van lipJbecJpedathilaedit VSwitch CIDR Block: 192168.0.0/24	j		
») Network Billing Aethod	Assign public IP instance.	With this box checked, the system will automatically assign a public IP address to your instance, a	of a will be accessible from the internet. If you would like to use an existing elastic IP address (IIP), Click here to find out how to bind an IIP to you	•		
Bandwidth Pricing	Pay-By-Traffic	89-Traffic 🕜 Writh Pay-By-Traffic (traffic in GB), bandwidth usage is billed on an hourly basis. Please make sure that your default payment method is valid.				
	1M Alibaba Cloud provid	50M 100M 150M 500M 500M 500M 500M 500M 5	Mbps			



- CentOS 6、CentOS 7、または Aliyun Linux を使用することを推奨します。
- ECS インスタンスで MaxCompute タスクまたはデータ同期タスクを実行する場合、
   ECS インスタンスの Python バージョンが 2.6 か 2.7 であることを確認する必要があります。 CentOS 5 をインストールすると、Python 2.4 もインストールされます。 他のCentOS バージョンの Python は 2.6 以降です。
- ECS インスタンスにパブリック IP アドレスが割り当てられていることを確認します。

### 2. データ同期タスクの設定

- a. DataWorks コンソールにログインしてプロジェクトを作成します。 この例では、 bigdata\_DOC という名前の DataWorks プロジェクトを使用します。
  - Data Integration が既に有効化されている場合、次のページが表示されます。

		Overview	Workspaces	Resources Comput	te Engines		
🌀 Da	ataWorks Da	ataStudio•Data Inte	egration • MaxCom	pute (	6	0	
Shortcuts						The Data Integration Launch	a * 1
Data Analytics	1	Data Integration	Maintenance Cente	er	Data Service	Support monpile development modes Support more data channels	MAR A
Workspaces					All Workspaces		
asdjhasuhschj64273	China East 2	MaxCompute_DOC	Asia Pacific SE 1	test012	China East 2		
Created At:Mar 15, 2018, 16: Compute Engines:None Services:Data Studio Data In	47:55 tegration Data Manageme	Created At:Jan 21, 2019, 23: Compute Engines:MaxComp Services:Data Studio Data In	13:23 ute tegration Data Manageme	Created At: Jan 02, 2018, 1 Compute Engines:MaxCon Services:Data Studio Data	5:36:40 npute PAI calculation engine Integration Data Manageme		
Workspace Settings	Data Analytics	Workspace Settings	Data Analytics	Workspace Settings	Data Analytics		
Data Service	Data Integration	Data Integration		Data Service	Data Integration		
Commonly Used Featu	res						- 1

 Data Integration が有効化されていない場合、次のページが表示されます。Data Integration を有効化するには、次の手順を実行します。このサービスを有効化する と、料金が発生します。料金設定ルールに基づいて見積もることができます。

Real-Name Authentication	Create AccessKey	Select Region and Services	Create Workspace
Basic Information			
	* Workspace Name :		
	Display Name :		
	* Mode :	Single Environment 📝	
	Description :		
Advanced Settings			
	* Task Recurrence :	En	
	* SELECT Result Download :	En Ø	
Information of MaxCompute			
	* MaxCompute Project Name :		0
	* Identity to Access MaxCompute:	Workspace Owner 📝 🔞	
	* Resource Group:	Pay per view default resource group $\checkmark$	

**b.** DataWorks プロジェクトの下にある [Data Integration] をクリックします。

- **c.** リソースグループを作成します。
  - **A.** [Data Integration] ページの左側のナビゲーションウィンドウで、[Resource Group] を 選択し、[Add Resource Group] をクリックします。
  - B. リソースグループを追加するには、次の手順に従います。
    - **A.** リソースグループを作成します。リソースグループ名を入力します。 この例では、リ ソースグループに es\_test\_resource という名前を付けます。

	Resource Groups Search	by resource group name,		Add Resource Group
- Overview				
utasks 🖞	Resource Group Name	Add Resource Group X		Actions
Monitoring	Default resource group	- Pa	ay-As-You-Go	
🚽 Sync Resources		Create Resource Group Add Server Install Agent Test Connectivity		
A Data Source	hdfs	Resource Group Name :     Pa	ıy-As-You-Go	
Resource Group		_		
	test_dataworks_abby	VPC Pe	ıy-As-You-Go	
	test	- Pe	ay-As-You-Go	
		_		
	TestJunwen	- Pa	ay-As-You-Go	
	testabyt	VPC Pe	ay-As-You-Go	
	RDBMS	VPC Pe	ny-As-You-Go	
	es_test_resource	VPC Pe	ay-As-You-Go	

**B.** サーバーを追加します。

Add Resource Group			×
Create Resource Group	Add Server	Install Agent	Test Connectivity
* Network Type :	• VPC 🕜		
* ECS UUID :	Enter a UUID rather tha	n server name.	0
* Server IP :	Enter the internal IP ad	dress of the machine.	0
* Machine CPU (Cores) :			
* Machine RAM (GB) :			
Add Server			

Previous	Next
----------	------

ECS UUID: #unique\_10。 ECS インスタンスにログインし、dmidecode | grep
 UUID コマンドを実行して戻り値を取得します。

0017 · 200	[root@iZbp10p	Z ~]# dmidecode ¦ grep UUID 11635
[root@iZbp10 Z~]# _	[root@iZbp10	Z ~]# _

 Machine IP/Machine CPU (Cores)/Memory Size (GB): ECS インスタンスのパブ リック IP アドレス、CPU コア、メモリサイズを指定します。 ECS コンソーにログ インし、ECS インスタンス名をクリックすると、[設定情報] に、情報が表示されます。

- C. エージェントをインストールします。次の手順に従って、エージェントのインストールを完了します。 この例では、VPC を使用します。 したがって、インスタンスのポート 8000 を開く必要はありません。
- D. 接続を確認します。接続が正常に確立されると、ステータスが [Available] に変わり ます。 ステータスが [Unavailable] の場合、ECS インスタンスにログインし、tail -f /home/admin/alisatasknode/logs/heartbeat.log コマンドを実行して DataWorks

と ECS インスタンス間のハートビートメッセージがタイムアウトしたかどうかを確 認します。

- **d.** データソースを追加します。
  - A. [Data Integration] ページの左側のナビゲーションウィンドウで、[Data Source] を選択し、[Add Data Source] をクリックします。
  - B. ソースタイプとして [MaxCompute] を選択します。

Data Integrat	tion BulkTest ~								3	dtplus_docs English
≡ ✔ Overview	Data Source Data Source Type	Add Data Source					×		C Refresh	Add Data Source
<ul> <li>Tasks</li> <li>Monitoring</li> <li>Sync Resources</li> </ul>	Data Source Name	Relational Database	SQL Server	PostgroSQL PostgreSQL	ORACLE*	DM		Status	Connected At	Actions
Data Source     Resource Group     Bulk Sync	mysql_001_di_test	DRDS	POLARDB	HybridDB for MySQL	HybridDB for PostgreSQL		8	B Failed	Dec 28, 2018 17:43:45	
		Big data storage	<b>%</b> Datahub	AnalyticDB (ADS)	Lightning					
		Semi-structuredstorage	HDFS	FIP						
		NoSQL		යා		Canc	zel			

**C.** データソースに関する情報を入力します。 この例では、**odps\_es** という名前のデータ ソースを作成します (次図)。

* Connection Name :	odps_es
Description :	
* ODPS Endpoint :	http://service.odps.aliyun.com/api
Tunnel Endpoint :	
* MaxCompute Project : Name	bigdata_DOC
* AccessKey ID :	
* AccessKey Secret :	•••••

ODPS workspace name: DataWorks の [Data Analytics] ページの左上隅のアイコンの右側に、テーブルのワークスペース名が表示されます (次図)。

Detail	DataStudio bigdata_DOC	~		
III	Tables [] C	🛗 hive_doc_good_sale 🗙 🛗 ba	ank_data × D	i 314 × Di OTS
<b>(</b> )	Search by table name or description.	DDL Mode Load from Pr	oduction Environm	ent Submit to Product
*	✓ ■ Tables			
Q	✓ ➡ Others		Table Name	hive_doc_good_sale
0	🖽 bank_data	Basic Information		
G	🗰 bank_data_01			
×	demo_trade_amount	Display Name :		
□	hive_doc_good_sale			
	hive_esdoc_good_sale	Level 1 Topic :	Select an option.	
≡0	<pre>table_154812606367001ac</pre>	Description :		
fx	<pre>dutput_table_1548127871928c581</pre>			
	output_table_154822526534301a? output table 154838757651615e?			
~				
2	₩ result_table	Physical Model		
亩	🛄 system_9c676e75c4324f75b5430	Partitioning ·	Partitioned Tab	ole O Non-
	₩ t1	r di	Partitioned Table	
	test			
	test1	Table Level :	Select an option.	
	🖽 userlog1	Table Type :		
	userlog2			

 AccessKeyId /AccessKeySecrete:ユーザー名の上にポインターを移動し、[User Info] を選択します (次図)。

Deter	DataStudio	bigdata_DOC						Cross-pi	roject cloning	Operation Ce	nter 🍳	dtplus_do	cs English
Ш	Tables	C C	hive_doc_good_sale >	🧱 bank_data 🗙						data 🗙 🟯 v		H	100
3											User Info	Versions	Guide
*	🗸 🛅 Tables										*	ወ	
Q	> 🛅 Others			Table Name	hive_doc_good_s						Bugs	Logout	
©			Basic Information								А	bout DataWorks	
			Display I	lame :							Ð	¥M.	
⊞			Level 1	Topic : Select an option		Level 2 Top	ic : Select an opt	ion. v		C	- 23	£.34	2
≖			Descr	ption :							- 33	\$3193	ŭ,
fx											Ô		5
	Produc	tion 🧳											-

[Personal Account] ページでアバターの上にポインターを移動し、[accesskeys] を クリックします (次図)。

e. 同期タスクを設定します。

A. [Data Analytics] ページの左側のナビゲーションウィンドウで、[Data Analytics] アイ コンをクリックし、[Business Flow] をクリックします。

DataV	DataStudio	alatikolo,DDC V
Ξ	Data Developn 🖉 🗒 🛛	‡ C ⊕ ₪
	Enter a file or creator name	<b>™</b>
*	> Solution	
R	✓ Business Flow	
	> 🗸 works	Create Business Flow
Ë		View All Business Flows

B. ターゲットビジネスフローをクリックし、[Data Integration] を選択し、[Create Data Integration Node] > [Data Sync] を選択してから、同期ノード名を入力します。

暑 workshop 🗙					Ξ
f • • •					
<ul> <li>Data Integration</li> </ul>	Development	Blood	୯ ତ ର	ର୍ ର 🖻	Param
Di Data Sync					eters
<ul> <li>Data Development</li> </ul>					Оре
odps SQL					eration
sh Shell					n Rec
M ODPS MR					ords
VI Virtual Node					
PyODPS					Vers
SQL Component Node					ion
MP OPEN MR					

**C.** 同期ノードが正常に作成されたら、新しい同期ノードページの上部にある [Switch to Script Mode] アイコンをクリックし、[Confirm] を選択します。

<u>ل</u> ]		•	⇒								
			50	ource							Destinatio
SQL				rds_workshop_log		?					ODPS
ds_user,											ods_user
?	Tip Are you	u sure y	/ou w	ant to enter the script	t mode? You	cannot	return to th	e wizard n	node onc	e you lea	×
									Ok	Can	cel

D. [Script Mode] ページの上部にある [Apply Template] アイコンをクリックします。
 [Source Type]、[Data Source]、[Destination Type]、[Data Source] のオプションに情報を入力し、[OK] をクリックして初期スクリプトを生成します。
ę	E 🗱				
		A			
Imp	oort Template				×
	* Course Tupe :	0005			
	Source Type.	UUFS			
	* Data Source :	请选择		~	
		Add Data	Source		
	* Destination Type :	ODPS		× ?	
	10.0	1=1+177			
	* Data Source :	· 请选择	Course	<b>*</b>	
		Add Data	Source		
				Cancel	

**E.** データ同期スクリプトの設定を行います。 Elasticsearch の設定ルールの詳細は、「writer プラグインの設定」をご参照ください。

"reader": {	🛛 📝 Odps Reader 🕸
"apparenter", (	
parameter : {	
"partition": "pt=1",	
"datasource": "odps_es",	
"column": [	
"create_time",	
"category",	
"brand",	
"buyer_id",	
"trans_num",	
"trans amount",	
"click cnt"	
1.	
"table": "hive doc good sale"	
1	
j) "voitan", (	
Writer : {	
prugin": "erasticsearch",	
"parameter": {	
"accessId": "elastic",	
"endpoint": "http://es-cn-mp	.elasticsearch.aliyuncs.com:9200",
"indexType": " <mark>elasticsea</mark> rch",	
"accessKey": "",	
"cleanup": true,	
"discovery": false,	
"column": [	
{	
"name": "create time".	
"type": "string"	
3	
1)	
l "normal", "sotogony"	
Talle : Category ,	
type : string	
},	
{	
"name": "brand",	
"type": "string"	
},	
{	
"name": "buyer_id",	
"type": "string"	
},	
1	
"name": "trans num",	
"type": "long"	
The store	
د ( ۲	
"name": "trans_amount",	
"type": "double"	
},	
{	
"name": "click_cnt",	
"type": "long"	
}	
1	
"index": "es index"	

# 1 注:

同期スクリプトの設定には、Reader、Writer、Settingの3つの部分があります。
 Reader では、データを同期するソースクラウドサービスを設定します。 Write では、Elasticsearchの設定ファイルを設定します。 Setting では、パケットロスと最大同時タスクを設定します。

- Endpoint には、Elasticsearch インスタンスのプライベートまたはパブリック IP アドレスを指定します。この例では、プライベート IP アドレスを使用しま す。したがって、ホワイトリストは不要です。外部 IP アドレスを使用する場 合、Elasticsearch の [ネットワークとスナップショット] ページで、Elasticsearch へのアクセスが許可されているパブリック IP アドレスを含むホワイトリストを設 定する必要があります。ホワイトリストには、DataWorks サーバーの IP アドレ スと、使用するリソースグループを指定する必要があります。
- Elasticsearch インスタンスへのログインに使用されるユーザー名とパスワードを Elasticsearch Writer の accessId と accesskey に設定する必要があります。
- Elasticsearch インスタンスのインデックス名を index に入力します。
   Elasticsearch インスタンスのデータにアクセスするには、このインデックス名を使用する必要があります。この例では、es\_index という名前の インデックスを使用します。
- MaxCompute テーブルがパーティションテーブルの場合、partition フィールドに パーティション情報を設定する必要があります。この例のパーティション情報は、 pt=1 です。

サンプル設定コード:

```
"configuration": {
"reader": {
"plugin": "odps",
'parameter": {
"partition": "pt=1",
"datasource": "odps_es",
  "column": [
    'create_time",
   "category",
   "brand"
   "buyer_id"
   "trans_num"
   "trans_amount",
   "click_cnt"
 ],
"table": "hive_doc_good_sale"
,,
"writer": {
"plugin": "elasticsearch",
 parameter": {
"accessId": "elastic",
 "endpoint": "http://es-cn-mpXXXXXXX.elasticsearch.aliyuncs.com:9200",
"indexType": "elasticsearch",
"accessKey": "XXXXXX",
 "cleanup": true,
"discovery": false,
 "column": [
```

```
"name": "create_time",
     "type": "string"
   },
   ł
     "name": "category",
     "type": "string"
   },
   {
     "name": "brand",
      "type": "string"
   },
   ł
     "name": "buyer_id",
      "type": "string"
   ł
     "name": "trans_num",
"type": "long"
    },
   ł
     "name": "trans_amount",
"type": "double"
   },
   {
     "name": "click_cnt",
"type": "long"
   }
 ],
"index": "es_index",
"batchSize": 1000,
"splitter": ",",
},
"setting": {
"errorLimit": {
  "record": "0"
},
"speed": {
"throttle": false,
"current": 1,
  "mbps": "1",
"dmu": 1
"Type": "job ",
"version": "1.0"
}
```

 F. スクリプトが同期されたら、[Run] をクリックして、ODPS データを Elasticsearch と同 期させます。

	V.		• 🕥 (J		<b>a</b> 2	26	
> Solution		Run(F8)	How to Configu	Stop(F9)			

- 3. 結果の検証
  - a. Elasticsearch コンソールにログインし、右上隅の Kibana コンソールをクリックし、[Dev Tools] を選択します。
  - **b.** 次のコマンドを実行して、データが Elasticsearch に正常にレプリケートされたことを確認 します。

```
POST /es_index/_search? pretty
{
"query": { "match_all": {}}
```

es\_index は、データ同期中の index フィールドの値です。



データが正常に同期されると、次のページが表示されます。

c. 次のコマンドを実行して、trans\_num フィールドを基準にドキュメントをソートします。

```
POST /es_index/_search? pretty
{
"query": { "match_all": {} },
"sort": { "trans_num": { "order": "desc" } }
}
```

**d.** 次のコマンドを実行して、ドキュメント内の category フィールドと brand フィールドを 検索します。

```
POST /es_index/_search? pretty
{
  "query": { "match_all": {} },
  "_source": ["category", "brand"]
}
```

e. 次のコマンドを実行して、 category が fresh のドキュメントをクエリします。

POST /es\_index/\_search? pretty

```
"query": { "match": {"category":"fresh"} }
```

詳細は、「#unique\_39」と『Elastic help center』をご参照ください。

## よくある質問

Elasticsearch インスタンスに接続するときにエラーが発生します

- 同期スクリプトを実行する前に、前の手順で作成したリソースグループが、右側の [configuration tasks resources group] メニューで選択されているかどうかを確認してくだ さい。
  - リソースグループが選択されている場合、次の手順に進みます。
  - リソースグループが選択されていない場合、右側の [configuration tasks resources group] メニューで、作成したリソースグループを選択し、[Run] をクリックします。
- endpoint、accessId、accesskey など、同期スクリプトの設定が正しいかどうかを確認して ください。 endpoint には、Elasticsearch インスタンスのプライベートまたはパブリック IP アドレスを指定します。 パブリック IP アドレスを使用する場合、ホワイトリストを設定して ください。 accessId には、Elasticsearch インスタンスへのアクセスに使用されるユーザー名 を指定します。デフォルトは elastic です。 accesskey には、Elasticsearch インスタンスへの アクセスに使用されるパスワードを指定します。

## 4.2 Alibaba Cloud Realtime Compute および Alibaba Cloud Elasticsearch のベストプラクティス

このドキュメントでは、Alibaba Cloud Realtime Compute を使用してデータを処理し、データ を Alibaba Cloud Elasticsearch (ES) にインポートする方法について説明します。

- Alibaba Cloud Realtime Compute の有効化とプロジェクトの作成
- Alibaba Cloud Elasticsearch の有効化
- このドキュメントでは、Alibaba Cloud Log Service を例として使用しています。そのため、Alibaba Cloud Log Service の有効化、#unique\_42/unique\_42\_Connect\_42\_section\_ahq\_ggx\_ndb、および#unique\_43/unique\_43\_Connect\_42\_section\_v52\_2jx\_ndb も必要です。

Alibaba Cloud Realtime Compute は、Alibaba Cloud Flink プロダクトです。 Kafka や Elasticsearch など、複数のデータソースとデータ消費サービスをサポートしていま す。#unique\_44 Alibaba Cloud Realtime Compute および Elasticsearch を使用して、ログ取得 シナリオのビジネス要件を満たすことができます。 Kafka または Log Service のログは、単純または複雑な Flink SQL 文を使用して Realtime Compute によって処理され、検索サービスのソースデータとして Alibaba Cloud Elasticsearch にインポートされます。 Realtime Compute の強力なコンピューティング機能と Alibaba Cloud Elasticsearch の検索機能により、リアルタイムのデータ処理と検索を実現し、ビジネスをリアル タイムサービスに変換できます。 さらに、Realtime Compute は Alibaba Cloud Elasticsearch と簡単に連携できます。 次の例は、Realtime Compute を Alibaba Cloud Elasticsearch と連携 する方法を示しています。

たとえば、ログまたはデータが Log Service にインポートされ、Alibaba Cloud Elasticsearch に インポートする前にデータを処理する必要があると仮定します。 次の図は、データ消費のパイプ ラインを示しています。



#### Realtime Compute ジョブを作成する

- 1. Realtime Compute コンソール にログインし、ジョブを作成します。
- 2. Flink SQL 文を記述します。
  - a) Log Service テーブルを作成します。

```
create table sls_stream(
 a int,
 b int,
 c VARCHAR
)
WITH (
 type ='sls',
 endPoint ='<yourEndpoint>',
 accessId ='<yourAccessId>',
 accessKey ='<yourAccessKey>',
 startTime = '<yourAccessKey>',
 consumerGroup ='<yourConsumerGroupName>'
```

#### );

### WITH 変数の説明は次のとおりです。

変数	説明		
<yourendpoint></yourendpoint>	Alibaba Cloud Log Service のパブリックエンドポイント。 エンドポイントは、Log Service のプロジェクトや内部ロ グデータにアクセスするために使用される URL です。 詳細 は、「#unique_45」をご参照ください。		
	たとえば、中国 (杭州) の Log Service のエンドポイントは http://cn-hangzhou.log.aliyuncs.com です。 エンドポ イントが http:// で始まることを確認してください。		
<youraccessid></youraccessid>	Log Service へのアクセスに使用される AccessKey ID。		
<youraccesskey></youraccesskey>	Log Service へのアクセスに使用される AccessKey シーク レット。		
<yourstarttime></yourstarttime>	ログデータが消費される時間範囲の開始時間。 Realtime Compute ジョブを実行するときは、この変数で指定された 開始時間よりも早い時間を指定する必要があります。		
<yourprojectname></yourprojectname>	Log Service プロジェクトの名前。		
<yourlogstorename></yourlogstorename>	プロジェクト内の Logstore の名前。		
<yourconsumergroupnamelog service="" td="" コンシューマグループの名前。<=""></yourconsumergroupnamelog>			

WITH 変数の詳細については、「#unique\_46」をご参照ください。

## ()

:

- Realtime Compute V3.2.2 以降のバージョンでは、Elasticsearch 結果テーブルの対応 が追加されています。 Realtime Compute ジョブを作成するときに、正しいバージョ ンを選択してください。
- Elasticsearch の結果テーブルは RESTful API に基づいています。したがって、すべての Elasticsearch バージョンと互換性があります。

```
CREATE TABLE es_stream_sink(
a int,
cnt BIGINT,
PRIMARY KEY(a)
)
WITH(
type ='elasticsearch',
endPoint = 'http://<instanceid>.public.elasticsearch.aliyuncs.com:<port>',
```

b) Elasticsearch 結果テーブルを作成します。

accessId = '<yourAccessId>', accessKey = '<yourAccessSecret>', index = '<yourIndex>', typeName = '<yourTypeName>' );

#### WITH 変数の説明は次のとおりです。

変数	説明
<instances></instances>	Elasticsearch インスタンスの ID。 インスタンスの [基本 情報] ページでインスタンス ID を確認できます。
	例:es-cn-45xxxxxxxxxxk1q
<port></port>	Elasticsearch インスタンスのパブリックネットワークポート。 インスタンスの [基本情報] ページでパブリックネット ワークポートを確認できます。 デフォルトのポート番号は、9200 です。
	Flasting and 11/2 by 2 a my bhy 2 b Vibana y
<youraccessiu></youraccessiu>	Elasticsearcn ィンスタンスへのアクセスと Kibana コン ソールへのログインに使用されるユーザー名。 デフォルト のユーザー名は "elastic" です。
<youraccesskey></youraccesskey>	Elasticsearch インスタンスにアクセスし、 Kibana コン ソールにログインするために使用されるパスワード。 パス ワードは、Elasticsearch インスタンスを作成するときに指 定されます。
<yourindex></yourindex>	Elasticsearch インスタンス上のドキュメントのインデッ クス。 インデックスはデータベース名のようなもので す。 ドキュメントのインデックスが作成されていない場合 は、最初にインデックスを作成してください。 詳細は、 「#unique_47」をご参照ください。
<yourtypename></yourtypename>	インデックスのタイプ。タイプは、データベース内のテー ブルの名前のようなものです。インデックスにタイプが指 定されていない場合は、最初にタイプを指定します。詳細 は、「#unique_47」をご参照ください。

WITH 変数の詳細については、「#unique\_48」をご参照ください。

 Elasticsearch は、PRIMARY KEY フィールドに含まれるドキュメント ID に従ったドキュ メントの更新をサポートしています。PRIMARY KEY フィールドとして指定できるフィー ルドは1つだけです。PRIMARY KEY フィールドを指定すると、フィールドの値がド キュメント ID として使用されます。ドキュメント ID は、PRIMARY KEY フィールドの ないドキュメントに対してランダムに生成されます。 詳細については『Index API』を ご参照ください。

- Elasticsearch は複数の更新モードをサポートしています。 updateMode パラメーター を設定して、更新モードを指定できます。
  - updateMode=full の場合、新しいドキュメントが既存のドキュメントを上書きします。
  - updateMode=inc の場合、新しい値が、関連するフィールドの既存の値を上書きします。
- Elasticsearch のすべての更新は、データの挿入または更新を意味する UPSERT 構文に 従います。
- c) データ消費ロジックを作成し、データを同期します。

```
INSERT INTO es_stream_sink
SELECT
a,
count(*) as cnt
FROM sls_stream GROUP BY a
```

3. ジョブを送信して実行します。

ジョブを送信して実行すると、Log Service に保存されたデータが集約され、Elasticsearch に インポートされます。 Realtime Compute は他の計算操作もサポートしています。たとえば、 データビューやユーザー定義拡張機能 (UDX) を作成できます。 詳細は、「#unique\_49」をご 参照ください。

#### まとめ

Alibaba Cloud Realtime Compute と Elasticsearch を使用すると、独自のリアルタイム検索 サービスをすばやく作成できます。 Alibaba Cloud Elasticsearch にデータをインポートするた めに、より複雑なロジックが必要な場合は、Realtime Compute のカスタムシンク機能を使用し てください。 詳細は、「#unique\_50」をご参照ください。

## 5 データを収集します

## 5.1 Beats による可視化された O&M システムの構築

背景

Beats はデータシッパー専用のプラットフォームです。 Beats をインストールすると、軽量の Beats エージェントから Logstash や Elasticsearch などのターゲットオブジェクトにインスタン スのデータを送信できます。

Beats エージェントおよび軽量シッパーとして、Metricbeat はシステムやサービスからメトリッ クを収集し、そのメトリックを Elasticsearch などのターゲットオブジェクトに送信するように 設計されています。 Metricbeat は、システムやサービスの統計情報を CPU からメモリ、Redis から NGINX などに送信するための軽量な方法です。

このトピックでは、Metricbeat を使用して MacBook からメトリックを収集し、そのメトリック を Elasticsearch インスタンスに送信し、Kibana にダッシュボードを生成する方法について説明 します。



Linux または Windows システムを実行しているコンピューターからメトリックを収集し、その メトリックを Elasticserach インスタンスに送信する手順もほぼ同じです。

1. Elasticsearch インスタンスの購入と設定

Elasticsearch インスタンスがない場合は、Elasticsearch を有効化してインスタンスを作成す る必要があります (#unique\_53)。 その後、インスタンスの内部またはパブリック IP アドレス を介して、MacBook で収集したデータを Elasticsearch インスタンスに送信できるようにな ります。

\_\_\_\_\_注:

 パブリック IP アドレスを介して Elasticsearch インスタンスにアクセスする場合、[セキュ リティ]ページで [パブリックアドレス] のスイッチをオンにし、[パブリック IP アドレスの ホワイトリスト]を設定する必要があります。

- 内部 IP アドレスを介して Elasticsearch インスタンスにアクセスする場合、Elasticsearch インスタンスと同じ VPC とリージョン に Elastic Compute Service (ECS) を作成し て、Elasticsearch へのアクセスを管理する必要があります。
- a. Elasticsearch コンソールにログインし、インスタンスの名前か ID をクリックしてから、 左側のナビゲーションウィンドウで [セキュリティ] をクリックします。 [セキュリティ]
   ページで、[パブリックアドレス] のスイッチをオンにします。

Basic Information	Cluster Network Settings	
Elasticsearch Cluster		
Plug-in Settings	Elasticsearch Cluster Password: The password has been set. Reset	Kibana IP Whitelist: 110 110 Update
Cluster Monitoring	VPC IP Whitelist: CLERICHE Update	Public Address:
Logs	Public IP Address Whitelist: 🚽 127.11.03 Update	
Security		
Snapshots		
<ul> <li>Intelligent Maintenance</li> </ul>		
Cluster Overview		
Health Diagnosis		
Previous Reports		

**b.** MacBook のパブリック IP アドレスをホワイトリストに追加します。



() :

パブリックネットワークを使用する場合、パブリックネットワークのアウトバウンドネッ トワークトラフィックを制御するジャンプサーバーの IP アドレスをホワイトリストに追 加します。ジャンプサーバーの IP アドレスを取得できない場合、0.0.0.0/1,128.0.0.0 /1 をホワイトリストに追加して、特定の IP アドレスを許可します。 この設定により、 Elasticsearch インスタンスはパブリックネットワークに公開されます。 リスクを評価 し、慎重に進めてください。

**c.** 設定が完了したら、左側のナビゲーションペインで [基本情報]をクリックし、Elaticsearch インスタンスのパブリック IP アドレスをコピーします。

Instance ID:	es-cr		
Name:	forE	dit	
Elasticsearch Version:	5.5.3_with_X-Pack		
Regions:	China (Hangzhou)		
VPC Network:	vpc-l d6rwt		
VPC-connected Instance Address:	es-cn-v	/uncs.com	
Public Address:	es-cn-v	rch.aliyuncs.com	

**d.** YML 設定を変更します。 [YML 設定] ページで、[自動インデックス] を有効化します。 デ フォルトで、この機能は無効化されています。 この操作により、Elasticsearch インスタン スが再起動され、有効になるまでに時間がかかります。

YML Configurations	Modify Configuration
Create Index Automatically: Enable ③	Delete Index With Specified Name: Specify Index Name When Deleting ③
Audit Log Index: Disable 🧿	Waicher: Disable 🕥
Other Configurations: 👩	

### 2. Metricbeat のダウンロードと設定

- Metricbeat インストールパッケージ (Mac オペレーティングシステム用)
- Metricbeat インストールパッケージ (32 ビット Linux オペレーティングシステム用)
- Metricbeat インストールパッケージ (64 ビット Linux オペレーティングシステム用)
- Metricbeat インストールパッケージ (32 ビット Windows オペレーティングシステム用)
- Metricbeat インストールパッケージ (64 ビット Windows オペレーティングシステム用)
- a. Metricbeat ファイルをダウンロード、解凍し、開きます。



**b.** metricbeat.yml ファイルの **Elasticsearch output** セクションを開き、編集します。 該当 するコンテンツのコメントを解除する必要があります。



# 📋 注:

Elasticsearch のアクセス制御情報は、以下のとおりです。

- hosts: Elasticsearch インスタンスのパブリックまたは内部 IP アドレス。この例では、パブリック IP アドレスを使用しています。
- protocol: http を設定します。
- username:デフォルトのユーザー名は elastic です。
- password: Elasticsearch へのログインに使用されるパスワード。

### 3. Metricbeat の有効化

次のコマンドを実行して Metricbeat を有効化し、Metricbeat を使用して Elasticsearch イン スタンスにデータを送信します。

./metricbeat -e -c metricbeat.yml

**4.** Kibana でのダッシュボードの表示

Elasticsearch コンソールの右上にある Kibana コンソールをクリックします。 Dashboard ページに移動します (次図)。



Kibana コンソールでインデックスパターンを作成していない場合、ダッシュボードに情報 が表示されないことがあります。 この問題を解決するには、インデックスパターンを作成し て、[ダッシュボード]ページで情報をもう一度表示します。

#### a. メトリックのリスト。

K	Dashboard
Ø	Q. Search + 1-20 of 20 < >
$\odot$	Name - Description
8	Golang: Heap
ø	C Kubernetes overview
24	Metricbeat - Apache HTTPD server status
÷	Metricbeat CPU/Memory per container
o	Metricbeat Docker
•	Metricbeat Hosts Overview
	Metricbeat MongoDB
	Metricbeat MySQL
	Metricbeat filesystem per Host
	Metricbeat host overview
	Metricbeat system overview
•	Metricbeat-Rabbitmq
Ð	Metricbeat-cpu
O	Metricbeat-filesystem

#### **b.** CPU メトリック。



5 秒ごとにデータを更新してレポートを生成するようシステムをスケジュールしたり、例 外が発生したときにアラートを送信するよう webhook を設定したりすることができま す。

# 6 Elasticsearchアプリケーション

# 7 Java high-level REST Client を使用したドキュメ ント API の呼び出し

## 7.1 概要

このベストプラクティスは、Java API 6.3.x に基づいています。 Java high-level REST Client を使 用して Alibaba Cloud Elasticsearch インスタンスに接続し、Elasticsearch ドキュメント API を 呼び出す方法について説明します。 ドキュメント API を呼び出して、ローカル Java プログラム でタスクを実行できます。たとえば、ドキュメントを作成、取得、または更新できます。

このベストプラクティスは Java API 6.3.x に基づいており、Alibaba Cloud Elasticsearch V6.3 インスタンスを使用しています。 Java high-level REST Client を使用して Alibaba Cloud Elasticsearch インスタンスに接続し、Elasticsearch ドキュメント API を呼び出す方法について 説明します。 ベストプラクティスには次のトピックが含まれます。

- **1.** Alibaba Cloud Elasticsearch インスタンスの作成と設定: Alibaba Cloud Elasticsearch イン スタンスを作成し、ホワイトリストを設定して、自動インデックス作成を有効にする
- ドキュメント API の呼び出し: Alibaba Cloud Elasticsearch インスタンスに接続し、ドキュ メントを作成、取得、更新、削除する

## 7.2 Alibaba Cloud Elasticsearch インスタンスの作成と設定

このドキュメントでは、インスタンスの購入、パブリックネットワークアクセスの有効化、ホワ イトリストの設定、自動インデックス設定など、Alibaba Cloud Elasticsearch インスタンスを作 成して設定する方法について説明します。

Alibaba Cloud Elasticsearch は VPC ネットワークのみをサポートします。 Elasticsearch インス タンスを作成する前に、VPC ネットワークと VSwitch を作成する 必要があります。 **1.** Alibaba Cloud Elasticsearch コンソール にログインし、Elasticsearch インスタンスを購入 します。

次の図は、この例で購入した Elasticsearch インスタンスの情報を示しています。 Elasticsearch のバージョンを 6.3 に設定します。 他のバージョンには非互換性の問題がある 可能性があります。

Version	6.7	6.3	5.5.3



Elasticsearch インスタンスを購入するとき、パスワードを指定する必要があります。 パス ワードを記録し、安全な場所に保管してください。 このパスワードは、Java API を使用して Elasticsearch インスタンスに接続するときに使用します。

Elasticsearch インスタンスを購入すると、[インスタンス] ページに移動します。 インスタン スの [ステータス] が [有効] に代わるまで待ち、次のタスクを続けます。

Instance ID/Name	Status
es-cn-7 es-cn-7	• Active

🗎 注:

Elasticsearch がインスタンスを有効化するのに約30分かかる場合があります。

2. パブリックネットワークアクセスを有効にします。

インスタンス名をクリックします。 左側のナビゲーションウィンドウで、【セキュリティ】を選択し、【インターネットアドレス】 スイッチをクリックしてこの機能を有効にします。

**兰**注:

[インターネットアドレス] を有効にしたら、 [基本情報] ページでインスタンスのパブリック エンドポイントを確認できます。 パブリックエンドポイントを記録してください。 このエン ドポイントは、Java API を呼び出して Elasticsearch インスタンスに接続するときに使用しま す。

## **3.** ホストのパブリック IP アドレスを Elasticsearch インスタンスのパブリックネットワークホワ イトリストに追加します。

<	es-cn-o- g	Cluster Monitoring Restart Instance
Basic Information		
Cluster Configuration	Network Settings	
Plug-ins	Elasticsearch Instance Password: Password is set Reset	VPC Whitelist: Update
Cluster Monitoring	Public Network Access:	Public Network Whitelist: Update
Logs	HTTPS-	
Security		
Snapshots	Instance Interconnection: Not Configured Edit	

ホストのパブリック IP アドレスを照会するには、www.google.com にアクセスし、"What Is My IP Address" と入力して、関連するリンクをクリックします。

4. 自動インデックス作成を有効にします。



この機能を有効にするには、Elasticsearch で Elasticsearch インスタンスを再起動する必要 があります。 操作を実行する前に、ビジネスが悪影響を受けないことを確認してください。

- a) Elasticsearch インスタンスの [クラスター設定] ページに移動し、[YML 設定] の右側にある [設定の編集] をクリックします。
- b) **[YML** パラメーター設定] ページで、 [自動インデックス] を [自動インデックスを有効にす る] に設定します。
- c) [この操作は、インスタンスを再起動します。確認してから操作を行ってください。] チェッ クボックスをオンにし、 [OK] をクリックします。

YML Configuration	
Auto Indexing:	O Disable
	● Enable
	Custom true
Index Deletion:	<ul> <li>Index Names Only</li> <li>Allow Wildcard Characters</li> </ul>
Audit Log Indexing:	<ul> <li>Disable</li> <li>Enable</li> </ul>
Watcher:	<ul> <li>Disable</li> <li>Enable</li> </ul>

Elasticsearch インスタンスの再起動後、Java high-level REST Client を使用してドキュメント API の呼び出し を実行します。

## 7.3 ドキュメント API の呼び出し

このドキュメントでは、Java high-level REST Client を使用して Elasticsearch ドキュメント API を呼び出す方法について説明します。 また、Alibaba Cloud Elasticsearch クラスターに接続 し、インデックスやドキュメントを作成したり、ドキュメントを取得するためのサンプルコード も提供します。 • JDK をインストールし、Java 環境変数を設定します。

この例では、jdk1.8.0\_211 を使用します。 他の JDK バージョンの互換性は保証されません。 公式 Web サイト から JDK をダウンロードできます。

• Java プログラム開発ツールを準備します。

この例では、Eclipse を使用しています。 その他のツールも選択できます。

このドキュメントで提供するサンプルコードは、Java high-level REST Client 6.3.x に基づいています。 他のクライアントバージョンの互換性は保証されません。

1. Maven プロジェクトを作成します。

New Project		_		
Select a wizard Create a Maven Project			Ď	
Wizards:				
type filter text				
<ul> <li>&gt; General</li> <li>&gt; Gradle</li> <li>&gt; Java</li> <li>&gt; Maven</li> <li>Check out Maven Projects from</li> <li>Maven Module</li> <li>Maven Project</li> <li>&gt; Examples</li> </ul>	SCM			
? < Back Next	t > Finish		Cancel	

次の図は、Maven プロジェクトの設定を示しています。

## 図 7-1:プロジェクト名と場所を選択します。

New Maven Project		$\Box$ $\times$
New Maven project Select project name and location		M
Create a simple project (skip archetype selection)		
Use default Workspace location		
Location:	~	Browse
Add project(s) to working set		
Working set:		More
▶ Advanced		

### 図 7-2:プロジェクトを設定します。

Artifact			
Group Id:	pan		
Artifact Id:	com		
Version:	0.0.1-SNAPSHOT	~	
Packaging:	jar	~	
Name:			
Description:			
Parent Proje	ct		
Group Id:			
Artifact Id:			
Version:		~	Browse
<ul> <li>Advanced</li> </ul>			

## 2. 次の依存関係を pom.xml ファイルに追加します。

.⊖ <	<pre>xproject xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://</pre>
2	<modelversion>4.0.0</modelversion>
1	<groupid>mypro</groupid>
Ļ	<artifactid>coma</artifactid>
	<version>0.0.1-SNAPSHOT</version>
Θ	<dependencies></dependencies>
Θ	<dependency></dependency>
;	<proupid>org.elasticsearch.client</proupid>
	<pre><artifactid>elasticsearch-rest-high-level-client</artifactid></pre>
)	<pre><version>6.3.2</version></pre>
<	<pre>/project&gt;</pre>

<dependency>

<groupId>org.elasticsearch.client</groupId> <artifactId>elasticsearch-rest-high-level-client</artifactId> <version>6.3.2</version> </dependency>

3. パッケージを作成し、パッケージに必要な Java ファイルを作成します。

パッケージの設定は以下のとおりです。

🖨 New Java Pa	ackage	_	□ ×
Java Package			
Create a new J	ava package.		
Creates folders	corresponding to packages.		
Source folder:	coma/src/main/java		Browse
Name:	coma		
Create pack	age-info.java		
?		Finish	Cancel

次の図は、必要な Java ファイルを示しています。

enter or select the parent folder:
coma/src/main/java/coma
> 🔛 com
🗸 🔛 coma
🔁 .settings
Y 🦻 src
🕆 🗁 main
🗸 🗁 java
🗁 coma
🔁 resources
> 🔁 test
> 🗁 target
> 🔛 maven-demo
File name: IndexAPI use java

次の図は、Maven プロジェクトのディレクトリを示しています。



- IndexAPI\_use.java :インデックスとドキュメントを作成します。
- UpdateAPI\_use.java:ドキュメントを更新します。
- GetAPI\_use.java:ドキュメントを取得します。
- DeleteAPI\_use.java:ドキュメントを削除します。

## 🎽 注:

```
この例は参照用のみです。ビジネス要件を満たすために、他の Java ファイルを作成すること
もできます。
```

 次のコードを IndexAPI\_use.java ファイルに追加して Alibaba Cloud Elasticsearch クラス ターに接続します。次に、Java IndexRequest 操作を呼び出して、インデックスを作成しま す。

コードスニペットは次のとおりです。

```
//Connect to the Elasticsearch cluster.
final CredentialsProvider credentialsProvider = new BasicCredentialsProvider();
    credentialsProvider.setCredentials(AuthScope.ANY,
         new UsernamePasswordCredentials("elastic", "Your password"));
    RestClientBuilder builder = RestClient.builder(new HttpHost("Your Elasticsearch
instance ID.public.elasticsearch.aliyuncs.com", 9200))
         .setHttpClientConfigCallback(new RestClientBuilder.HttpClientConfigCallback
() {
           public HttpAsyncClientBuilder customizeHttpClient(HttpAsyncClientBuilder
httpClientBuilder) {
             return httpClientBuilder.setDefaultCredentialsProvider(credential
sProvider);
           }
        });
//Call the Java IndexReguest operation to create an index.
RestHighLevelClient client = new RestHighLevelClient(builder);
    try {
      IndexRequest request = new IndexRequest();
      request.index("apitest_index");
      request.type("apitest_type");
      request.id("1");
      Map<String, Object> source = new HashMap<>();
      source.put("user", "kimchy");
      source.put("post_date", new Date());
source.put("message", "trying out Elasticsearch");
      request.source(source);
      try {
        IndexResponse result = client.index(request);
         System.out.println(result);
      } catch (IOException e) {
         e.printStackTrace();
    } finally {
      client.close();
```

}

完全なサンプルコードは、 サンプルコード からダウンロードできます。 返される結果は次の

とおりです。

```
IndexResponse
[
index=apitest_index,
type=apitest_type,
id=1,
version=1,
result=created,
seqNo=0,
primaryTerm=1,
shards={"total":2,"successful":1,"failed":0}
]
```

5. 次のコードを UpdateAPI\_use.java ファイルに追加し、Java UpdateRequest 操作を呼び出し

```
てドキュメントを更新します。
```

コードスニペットは次のとおりです。

```
RestHighLevelClient client = new RestHighLevelClient(builder);
  try {
      UpdateRequest updateRequest = new UpdateRequest("apitest_index", "
apitest type", "1");
      IndexRequest indexRequest = new IndexRequest("apitest_index", "apitest_type",
"1");
      Map<String, String> source = new HashMap<>();
      source.put("user", "dingw2");
      indexRequest.source (source);
      updateRequest.doc(indexRequest);
      UpdateResponse result = client.update(updateRequest);
      System.out.println(result);
    }catch (IOException e) {
        e.printStackTrace();
    } finally {
      client.close();
    }
```

完全なサンプルコードは、 サンプルコード からダウンロードできます。 返される結果は次の

とおりです。

```
UpdateResponse
[
index=apitest_index,
type=apitest_type,
id=1,
version=2,
seqNo=1,
primaryTerm=1,
result=updated,
shards=ShardInfo{
total=2,
successful=2,
failures=[]
```

} ]

**6.** 次のコードを GetAPI\_use.java ファイルに追加し、Java GetRequest 操作を呼び出してド キュメントを取得します。

```
コードスニペットは次のとおりです。
```

```
RestHighLevelClient client = new RestHighLevelClient(builder);
try {
    GetRequest request = new GetRequest("apitest_index", "apitest_type", "1");
    GetResponse result = client.get(request);
    System.out.println(result);
}catch (IOException e) {
    e.printStackTrace();
} finally {
    client.close();
}
```

完全なサンプルコードは、 サンプルコード からダウンロードできます。 返される結果は次の

とおりです。

```
{
    "_index": "apitest_index",
    "_type": "apitest_type",
    "_id": "1",
    "_version": 2,
    "found": true,
    "_source": {
        "post_date": "2019-06-10T05:50:52.752Z",
        "message": "trying out Elasticsearch",
        "user": "dingw2"
    }
}
```

7. 次のコードを DeleteAPI\_use.java ファイルに追加し、Java DeleteRequest 操作を呼び出して ドキュメントを削除します。

コードスニペットは次のとおりです。

RestHighLevelClient client = new RestHighLevelClient(builder);

```
try {
    DeleteRequest request = new DeleteRequest("apitest_index", "apitest_type", "1
");
    DeleteResponse result = client.delete(request);
    System.out.println(result);
    } catch(Throwable e) {
        e.printStackTrace();
    } finally {
        client.close();
    }
}
```

}

完全なサンプルコードは、サンプルコードからダウンロードできます。 返される結果は次の とおりです。

```
DeleteResponse
[
index=apitest_index,
type=apitest_type,
id=1,
version=3,
result=deleted,
shards=ShardInfo{
   total=2,
   successful=2,
   failures=[]
  }
]
```

ビジネス要件を満たすために、その他のドキュメント API 操作を呼び出すことができます。 その 他のドキュメント API 操作の呼び出しの詳細については、「ドキュメント API 」をご参照くださ い。

# 8 インデックス管理

## 8.1 Curator の使用

### Elasticsearch Curator のインストール

- **1.** Elasticsearch インスタンスと同じ VPC ネットワークにある ECS インスタンスを購入します。 この例では、CentOS 7.3 64 ビットを実行する ECS インスタンスを使用します。
- 2. 以下のコマンドを実行します。
  - a. Elasticsearch Curator のインストール

pip install elasticsearch-curator



- Elasticsearch Curator 5.6.0 をインストールすることを推奨します。 このバージョン は、Elasticsearch 5.5.3 と 6.3.2 をサポートします。
- Curator と Elasticsearch のバージョンの互換性
- **b.** Curator のバージョンの表示

curator --version

返されるバージョン情報

curator, version 5.6.0

### Singleton コマンドラインインターフェイス

- curator\_cli を使用して、操作を実行します。
- Singleton コマンドラインインターフェイス



- 一度に実行できる操作は1つだけです。
- Curator を使用してすべての操作を実行できるわけではありません (Alias や Restore など)。

### Crontab によるタスクのスケジュール

crontab コマンドと curator コマンドを使用して、複数の操作を実行するようにタスクをスケ ジュールすることができます。

Curator コマンド:

```
curator [OPTIONS] ACTION_FILE
オプション:
--config PATH 設定ファイルへのパス。 デフォルト:~/.curator/curator.yml
--dry-run 何も変更しません。
--version バージョンを表示して終了します。
--help このメッセージを表示して終了します。
```

- curator コマンドを実行するとき、config.yml ファイル (公式リファレンス) を指定する必要が あります。
- curator コマンドを実行するとき、action.yml ファイル (公式リファレンス)を指定する必要 があります。

### Hot-Warm アーキテクチャの演習

Curator を使用して hot ノードから warm ノードにインデックスを移行します (公式リファレンス)。

### hot ノードから warm ノードへのインデックスの移行

1. 以下のように、config.yml ファイルをパス /usr/curator/ に作成します。

## 🙂 :

- hosts:アクセスする Elasticsearch インスタンスのアドレスに置き換えます。この例では、Elasticsearch インスタンスのプライベートアドレスを使用しています。
- http\_auth: Elasticsearch インスタンスへのログインに使用するユーザー名とパスワード に置き換えます。

```
client:
hosts:
- http://es-cn-0pp0z9p2v00031234.elasticsearch.aliyuncs.com
port: 9200
url_prefix:
use_ssl: False
certificate:
client_cert:
client_cert:
client_key:
ssl_no_validate: False
http_auth: user:password
timeout: 30
master_only: False
logging:
loglevel: INFO
```

logfile: logformat: default blacklist: ['elasticsearch', 'urllib3']

2. 以下のように、action.yml ファイルをパス /usr/curator/ に作成します。

# 📋 注:

- 以下のコンテンツでは、30分前に作成され、logstash-で始まるインデックスを hot ノードから warm ノードに移行します。
- 以下のコンテンツは、ビジネスニーズに合わせてカスタマイズできます。

actions: 1: action: allocation description: "Apply shard allocation filtering rules to the specified indices" options: key: box\_type value: warm allocation\_type: require wait for completion: true timeout override: continue if exception: false disable action: false filters: - filtertype: pattern kind: prefix value: logstash-- filtertype: age source: creation date direction: older timestring: '%Y-%m-%dT%H:%M:%S' unit: minutes unit count: 30

3. curator コマンドが正常に実行されるかどうか確認します。

curator -- config /usr/curator/config.yml /usr/curator/action.yml

コマンドが正常に実行されると、以下の情報が返されます。

2019-02-12 20:11:30,607 INFO<br/>2019-02-12 20:11:30,612 INFO<br/>allocation filtering rules to the specified indices<br/>2019-02-12 20:11:30,693 INFO<br/>require.box\_type': 'warm'}Preparing Action ID: 1, "allocation"<br/>Trying Action ID: 1, "allocation": Apply shard<br/>Updating index setting {'index.routing.allocation.2019-02-12 20:12:57,925 INFO<br/>2019-02-12 20:12:57,925 INFO<br/>2019-02-12 20:12:57,925 INFOHealth Check for all provided keys passed.<br/>Action ID: 1, "allocation" completed.<br/>Job completed.

**4.** crontab コマンドを実行して、15 分間隔で curator コマンドを実行します。

\*/15 \* \* \* \* curator --config /usr/curator/config.yml /usr/curator/action.yml

## 9 ベクトル検索プラグインのベストプラクティス

aliyun-knn プラグインは、Alibaba Cloud Elasticsearch (ES) が設計したベクトル検索エンジン です。 Alibaba DAMO Academy で設計されたベクトル検索エンジンである proxima のベクト ルデータベースを使用しています。 このプラグインは、画像検索、動画フィンガープリンティン グ、顔認識、音声認識、商品推奨などのシナリオで、ベクトル空間に基づくデータ検索要件への 対応に役立ちます。

Alibaba Cloud Elasticsearch のベクトル検索エンジンは、Pailitao、Image Search Cloud、 Youku 動画フィンガープリンティング、Qutoutiao 動画フィンガープリンティング、Taobao 商 品推奨、カスタマイズ検索、CrossMedia 検索など、Alibaba Group 内の本番環境シナリオで継 続的に使用されています。

Alibaba Cloud Elasticsearch は、ベクトル検索エンジンにアクセスするための aliyun-knn プラ グインを提供します。 これにより、エンジンの使用方法を学ぶ必要がなくなります。 aliyun-knn プラグインは、すべてのネイティブ Elasticsearch バージョンと互換性があります。 このベクト ル検索エンジンは、リアルタイム増分同期とほぼリアルタイム (NRT) 検索に加えて、マルチレ プリカ、復元、スナップショットなど、分散検索でのネイティブ Elasticsearch の他の機能もサ ポートしています。

Alibaba Cloud Elasticsearch のベクトル検索エンジンは、Hierarchical Navigable Small World (HNSW) と Linear Search のアルゴリズムをサポートします。 これらのアルゴリズムは、インメ モリストレージから少量のデータを処理するのに適しています。 次の表は、2 つのアルゴリズム のパフォーマンスを比較しています。

### 表 9-1 : HNSW と Linear Search のパフォーマンス比較

2 つのアルゴリズムのパフォーマンスは、Alibaba Cloud Elasticsearch V6.7.0 で測定していま す。 テスト環境の設定を次に示します。

- ノード設定:2つのデータノードと100 GBのクラウド SSD。 各ノードは16 コア、64 GBメ モリを搭載。
- データセット: SIFT 128 ディメンション float 型ベクトル (http://corpus-texmex.irisa.fr/)。
- ・ 総サンプル数:2000万。
- インデックス設定:デフォルト設定。
| パフォーマンス指標    | HNSW    | Linear Search |
|--------------|---------|---------------|
| リコール率トップ 10  | 98.6%   | 100%          |
| リコール率トップ 50  | 97.9%   | 100%          |
| リコール率トップ 100 | 97.4%   | 100%          |
| 遅延 (p99)     | 0.093 秒 | 0.934 秒       |
| 遅延 (p90)     | 0.018 秒 | 0.305 秒       |

# 📋 注:

p はパーセントの略です。 たとえば、遅延 (p99) は、99% のクエリのレスポンスに要する秒数 を示します。

#### インデックス計画

アルゴリズム	シナリオ	インメモリスト レージ	備考
HNSW	<ul> <li>ノードごとに 少量のデータのがす。</li> <li>低レンスをひんした。</li> <li>のと必要した。</li> <li>のと必要した。</li> <li>のとのでのでのです。</li> <li>ののでのでのでので、</li> <li>ののでので、</li> <li>ののでので、</li> <li>のので、</li> <li>のので、<!--</td--><td>はい</td><td><ul> <li>HNSW は、貪欲な検索に基づいて おり、三角不等式に従います。三 角不等式は、A から B、B から C の辺の合計が A から C の辺より も大きくなければならないことを 示しています。内積空間は三角不 等式に従いません。したがって、 HNSW アルゴリズムを適用する前 に、ユークリッド空間か球状空間 に変換する必要があります。</li> <li>Elasticsearch にデータをイン ポートした後、オフピーク時に定 期的に強制マージ API 操作を呼び 出してシャード内のセグメントを マージすることを推奨します。こ れにより、レスポンスのレイテン シを短縮できます。</li> </ul></td></li></ul>	はい	<ul> <li>HNSW は、貪欲な検索に基づいて おり、三角不等式に従います。三 角不等式は、A から B、B から C の辺の合計が A から C の辺より も大きくなければならないことを 示しています。内積空間は三角不 等式に従いません。したがって、 HNSW アルゴリズムを適用する前 に、ユークリッド空間か球状空間 に変換する必要があります。</li> <li>Elasticsearch にデータをイン ポートした後、オフピーク時に定 期的に強制マージ API 操作を呼び 出してシャード内のセグメントを マージすることを推奨します。こ れにより、レスポンスのレイテン シを短縮できます。</li> </ul>

ベストプラクティス / 9ベクトル検索プラグインのベストプラ クティス

アルゴリズム	シナリオ	インメモリスト レージ	備考
Linear Search	<ul> <li>ブルート フォース検索。</li> <li>100%のリコールを必要す。</li> <li>レイテンジンは、テンジンは、一々増加します。</li> <li>効果の比較。</li> </ul>	はい	なし

#### クラスターサイズ

アイテム	説明
データノード仕様 (必須)	本番環境の最小データノード仕様は、4 コア、16 GB でなけれ ばなりません。 2 コア、8 GB の仕様は、テスト目的でのみ使 用できます。
ノードあたりの最大データ量	各ノードに保存される最大データ量は、ノードのメモリ容量の 50% です。
入力調整	ベクトルインデックスは、CPU を大量消費するジョブです。 高 い入力レートを維持しないことを推奨します。 16 コア、64 GB メモリを備えたデータノードの場合、1 秒あたり 5,000 トラン ザクション未満のピーク入力レートを推奨します。
	Elasticsearch でクエリを処理する際、すべてのインデック スファイルがノードメモリに読み込まれます。 ノードのメモ リが不足している場合、Elasticsearch はシャードを再割り 当てします。 したがって、クエリの処理中に大量のデータを Elasticsearch にインポートしないでください。

### ベクトル検索プラグインのインストール

## () :

ベクトル検索プラグインは、Alibaba Cloud Elasticsearch V6.7 以降のバージョンでのみサポートされています。

 プラグインをインストールする前に、データノード仕様が2コア、8GBより高いことを確認 してください。2コア、8GBはテスト目的のみです。本番環境の最小仕様は4コア、16GB です。Elasticsearchインスタンスが要件を満たしていない場合、最初にデータノードを2コ ア、8GBより上にアップグレードする必要があります。詳細については、「#unique\_62」 をご参照ください。

Alibaba Cloud Elasticsearch コンソールにログインし、【インスタンス ID】> 【プラグイン 設定】 > 【ビルトインプラグインリスト】を選択します。【ビルトインプラグインリスト】から [aliyun-knn] プラグインを見つけ、インストールします。 詳細については、「#unique\_63/ unique\_63\_Connect\_42\_section\_d0y\_kyx\_fu0」をご参照ください。

Built-in Plug-ins				
D.C. ut				Colored and a second
kerresn				Enter a plug-in name.
Plug-in	Туре	Status	Description	Actions
aliyun-knn	Built-in Plug-in	<ul> <li>Not installed</li> </ul>	Aliyun Elasticsearch KNN plugin	Install

# (!)

デフォルトでは、[aliyun-knn] プラグインのステータスは[未インストール]です。

#### ベクトル検索プラグインの使用

ベクトル検索プラグインのインストール、Elasticsearch インスタンスの Kibana コンソールへの ログインの後、次の手順に従ってプラグインを使用します。

# ()

次のサンプルコードは、Elasticsearch V6.7 にのみ適用できます。 Elasticsearch V7.4 について は、公式ドキュメントをご参照ください。

1. インデックスを作成します。

} }

パラメーター	説明
index.vector.algorithm	有効値:hnsw と linear。
type	フィールドのタイプ。ベクトル型フィールドを指定するに は、値を proxima_vector に設定します。
dim	ベクトルのディメンション。 有効値:1~2048。

上記のサンプルコードでは、test という名前のインデックスが作成されます。 インデックス のタイプは \_doc です。 インデックスには 2 つのフィールド、feature と id が含まれます。 必 要に応じて、インデックスとフィールドの名前を変更できます。

2. インデックスにドキュメントを追加します。

```
POST test/_doc
{
    "feature": [1.0, 2.0],
    "id": 1
}
```

(!) .

feature フィールドの値は、float 配列でなければなりません。 配列の長さは、mapping の dim パラメーターに指定した長さと同じでなければなりません。

3. インデックス内のデータをクエリします。

```
GET test/_search
{
    "query": {
        "hnsw": {
            "feature": {
                "vector": [1.5, 2.5],
               "size": 10
            }
        }
    }
}
```

パラメーター	説明
hnsw	algorithm は、インデックスの作成時に指定したものと同じ でなければなりません。
vector	float 配列。 配列の長さは、mapping の dim パラメーター に指定した長さと同じでなければなりません。
size	返される上位のドキュメント数。

### パラメーター

### 表 9-2:アルゴリズムのパラメーター

パラメーター	説明	デフォルト
index.vector. algorithm	インデックスの作成に使用されるアルゴリズ ム。 有効値:hnsw と linear。	hnsw

#### 表 9-3 : HNSW の入力パラメーター

パラメーター	説明	デフォルト
index.vector.hnsw. builder.max_scan_n um	最悪の事態が起きた場合、グラフ作成中にス キャンされる最近傍の最大数。	100000
index.vector.hnsw. builder.neighbor_cnt	レイヤー0での各ノードの最近傍の最大数。 値を100に設定することを推奨します。 非ア クティブなインデックスを保存するために消費 されるストレージの量。 グラフの品質は、こ のパラメーターの値とともに増加します。	100
index.vector.hnsw. builder.upper_neig hbor_cnt	レイヤー 0 以外のレイヤーでの各ノードの最近 傍の最大数。 このパラメーターを neighbor_c nt の 50% に設定することを推奨します。	50
index.vector.hnsw .builder.efconstruc tion	グラフ作成中にスキャンされる最近傍の数。 スキャンされる最近傍の数が多いほど、グラフ の品質は向上します。ただし、インデックス の作成には時間がかかります。このパラメー ターを 400 に設定することを推奨します。	400
index.vector.hnsw. builder.max_level	レイヤー0を含むレイヤーの総数。1,000万 件のドキュメントがあり、scaling_factorパ ラメーターが30に設定されているとします。 30を基数として使用し、10000000の対数を 最も近い整数に切り上げます。デフォルトは5 です。	6
index.vector.hnsw .builder.scaling_fa ctor	スケーリング係数。レイヤーのデータ量は、 上位レイヤーのデータ量にスケーリング係数 を乗じたものに等しくなります。10~100 の 値を指定してください。レイヤーの数は、 scaling_factor の値とともに減少します。こ のパラメーターを 50 に設定することを推奨し ます。	50

#### 表 9-4 : HNSW の検索パラメーター

パラメーター	説明	デフォルト
ef	オンライン検索中にスキャンされる最近傍の 数。 値を大きくすると、リコール率は高くな りますが、検索は遅くなります。 有効値:100 ~1000。	100

サンプルリクエスト:

GET test/\_search
{
 "query": {
 "hnsw": {
 "feature": {
 "vector": [1.5, 2.5],
 "size": 10,
 "ef": 100
 }
 }
 }
}

#### 表 9-5:サーキットブレーカーのパラメーター

パラメーター	説明	デフォルト
indices.breaker .vector.native. indexing.limit	オフヒープメモリ使用量がこのパラメーターで 指定された値を超えると、書き込み操作は中断 されます。Elasticsearchがインデックスを作 成し、メモリをリリースすると、書き込み操作 が再開されます。このサーキットブレーカー がトリガーされた場合、システムメモリの消費 量が多いことを意味します。入力レートを調 整することを推奨します。初級ユーザーの場 合は、デフォルト設定を使用することを推奨し ます。	70%
indices.breaker. vector.native.total. limit	ベクトルインデックスの作成に使用されるオ フヒープメモリの最大割合。 実際のオフヒー プメモリ使用量が、指定された割合を超える と、Elasticsearch はシャードを再割り当てす る場合があります。 初級ユーザーの場合は、 デフォルト設定を使用することを推奨します。	80%

#### FAQ

- Q:ドキュメントのリコール率をどのように求めればよいですか。
  - A:2つのインデックスを作成し、1つはHNSW アルゴリズム、もう1つは Linear Search ア ルゴリズムを使用します。その他のインデックス設定は、どちらのインデックスも同じにしま す。 クライアントから両方のインデックスに同じベクトルデータを追加し、インデックスを更 新します。 同じクエリベクトルに対し、HNSW インデックスと Linear Search インデックス から返されるドキュメント ID を比較し、同じドキュメント ID を見つけます。

### **注**:

どちらのインデックスからも返されたドキュメント ID の数を、返されたドキュメント ID の総数で除算して、ドキュメントのリコール率を算出します。

 Q:データを Elasticsearch にインポートするときに circuitBreakingException エラーが発生 した場合、このエラーを解決するにはどうすればよいですか。

A:このエラーは、オフヒープメモリ使用量が、ndices.breaker.vector.native.indexing.limit パラメーターで指定された割合を超えたことを示します。 デフォルトの割合は 70% です。 書き込み操作は中断されます。 ほとんどの場合、Elasticsearch がインデックスを作成し、メ モリをリリースすると、自動的に書き込み操作が再開されます。 クライアントのデータイン ポートスクリプトに再試行メカニズムを追加することを推奨します。

• Q:書き込み操作が中断された後、CPU が動作し続けているのはなぜですか。

A: Elasticsearch は、更新またはフラッシュ中にベクトルインデックスを作成します。 書き 込み操作が中断された場合でも、ベクトルインデックス作成タスクは実行されている場合があ ります。 更新の最終ラウンドが完了すると、コンピューティングリソースはリリースされま す。

# 10 Alibaba Cloud Elasticsearch サイジングのベス トプラクティス

Alibaba Cloud Elasticsearch を使用する前に、必要なリソースの合計量を見積もる必要があり ます。 テスト結果とユーザーのフィードバックに基づいて、Alibaba Cloud は Elasticsearch リ ソースの量を推定し計算する一般的な方法をいくつか提供します。 これらのメソッドは参照専用 です。

#### 該当するディスクタイプ

このベストプラクティスは、[ディスクタイプ] が [SSD クラウドディスク] に設定されている Alibaba Cloud Elasticsearch インスタンスに適用できます。

	Note	Specify the disk type and capacity of the data node. The product of the storage capacity of a node and the number of nodes is the total storage of the Elasticsearch instance. Reserve space for the index, index replicas, and reserved resources. The storage configuration does not apply to any dedicated master node in the cluster.	
Storage	Disk Type	SSD An SSD supports a maximum of 2 TB data. It is used for online data analysis and searches that require high IOPS and fast data response.	
	Disk Encryption	No Yes	
	Node Storage	20 The unit is GiB. An SSD supports a maximum of 2048 GiB (2 TB). An ultra disk supports a maximum of 5120 GiB (5 TB). If the data to be stored is larger than 2048 GiB, an ultra disk can only support the following data sizes: 2560 GiB, 3072 GiB, 3584 GiB, 4096 GiB, 4608 GiB, or 5120 GiB.	

### ディスクのサイジング

Alibaba Cloud Elasticsearch インスタンスのディスク容量は、次の要因によって決まります。

- レプリカの数。各インデックスには、少なくとも1つのレプリカが必要です。
- インデックス作成のオーバーヘッド。一般的に、インデックス作成のオーバーヘッドは、 ソースデータより 10% 大きくなります。 \_all パラメーターのインデックス作成のオーバー ヘッドは含まれていません。
- オペレーティングシステムの予約領域。オペレーティングシステムは、重要なプロセス、シス テムリカバリ、およびディスクフラグメントのために、ディスク容量の5%をデフォルトで予 約します。

- Alibaba Cloud Elasticsearch のオーバーヘッド。Alibaba Cloud Elasticsearch は、セグメントのマージ、ログ、およびその他の内部的な操作のために、ディスク容量の 20% を予約します。
- セキュリティしきい値のオーバーヘッド。ディスク容量の最低 15% をセキュリティのしきい 値として予約する必要があります。

これらの要因に基づいて、必要な最小ディスク容量は次のように計算されます。最小ディスク容量= ソースデータのサイズ × 3.4

合計ディスク容量=ソースデータのサイズ×(1+レプリカの数)×(1+インデックス作成の オーバーヘッド)/(1-オペレーティングシステムの予約領域)/(1-Elasticsearchのオーバー ヘッド)/(1-セキュリティしきい値のオーバーヘッド) =ソースデータのサイズ×(1+レプリカの数)×1.7 =ソースデータのサイズ×3.4

# () :

- ・ ビジネスで必要でない限り、 \_all パラメーターを有効にしないことを推奨します。
- このパラメーターを有効にしたインデックスでは、ディスク使用率に大きなオーバーヘッド が発生します。テスト結果と実践に基づいて、推定ディスク容量の 50% を最終的なディスク 容量に追加することを推奨します。

合計ディスク容量=ソースデータのサイズ×(1 +レプリカの数)×1.7×(1+0.5) =ソースデータのサイズ×5.1

#### クラスター仕様を選択する

Alibaba Cloud Elasticsearch クラスターのパフォーマンスは、クラスター内の Elasticsearch ノードの仕様によって決まります。 Elasticsearch を使用する前に、クラスターのサイズを見積 もり、ノードを追加するか、クラスターをアップグレードすることを推奨します。 テスト結果と 実践に基づいて、次の提案をします。

- クラスターあたりの最大ノード数=各ノードの CPU コア数×5
- Elasticsearch ノードが保存できるデータの最大量は、シナリオごとに異なります。
  - データクエリの高速化と集約:ノードあたりの最大データ量=各ノードのメモリ (GB) × 10
  - ログデータのインポートとオフライン分析:ノードあたりの最大データ量=各ノードのメ モリ (GB) × 50
  - 通常のシナリオ:ノードあたりの最大データ量=各ノードのメモリ (GB) × 30

#### 表 10-1: 推奨クラスター仕様

仕様	クラスターノード の最大数	各ノードの最大メ モリ <b>(</b> クエリ <b>)</b>	各ノードの最大メ モリ <b>(</b> ログ <b>)</b>	各ノードの最大メ モリ <b>(</b> 共通)
2 コア 4 GB	10	40 GB	200 GB	100 GB
2 コア 8 GB	10	80 GB	400 GB	200 GB
4 コア 16 GB	20	160 GB	800 GB	512 GB
8 コア 32 GB	40	320 GB	1.5 TB	1 TB
16 コア 64 GB	50	640 GB	2 ТВ	2 ТВ

#### シャードサイジング

シャードの数と各シャードのサイズの両方が、Alibaba Cloud Elasticsearch クラスターの安定性 とパフォーマンスに影響します。 Elasticsearch の各インデックスは、特定の数のシャードに分 割されます。 デフォルトでは、インデックスは 5 つのシャードに分割されます。

- 小規模な Elasticsearch ノードの場合、各シャードのサイズは 30 GB 以下にすることを推奨し ます。 大規模な Elasticsearchノードの場合、各シャードのサイズは 50 GB 以下にすることを 推奨します。
- ログ分析または非常に大きなインデックスの場合、各シャードのサイズは 100 GB 以下にする ことを推奨します。
- レプリカを含むシャードの数は、ノードの数に等しいか、ノード数の等倍にしなければなりません。
- ノードのインデックスに最大5つのシャードを指定することを推奨します。



- ユーザーによって、データスキーマ、クエリの複雑さ、データサイズ、パフォーマンス、 およびデータの変更に関する要件が異なる場合があります。このドキュメントでは、 Elasticsearchのサイジングに関する参照のみを提供しています。
- 可能であれば、実際のデータとサービスのシナリオに基づいて、Elasticsearch クラスターの サイズを測定することを推奨します。
- 重いワークロードに対処する必要がある場合、Alibaba Cloud Elasticsearch のエラスティックスケーリング機能を使用して、ディスクの拡張、ノードの追加、またはノードのアップグレードを行うことができます。