Alibaba Cloud E-MapReduce

Quick Start

Issue: 20190320

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. **You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.**
2. **No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company, or individual in any form or by any means without the prior written consent of Alibaba Cloud.**
3. **The content of this document may be changed due to product version upgrades , adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.**
4. **This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults " and "as available" basis. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity , applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, special, or punitive damages, including lost profits arising from the use**

or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.

5. By law, all the content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of the Alibaba Cloud website, product programs, or content shall be used, modified , reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates . The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates).
6. Please contact Alibaba Cloud directly if you discover any errors in this document.

Generic conventions

Table -1: Style conventions

Style	Description	Example
	This warning information indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	 Danger: Resetting will result in the loss of user configuration data.
	This warning information indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	 Warning: Restarting will cause business interruption. About 10 minutes are required to restore business.
	This indicates warning information, supplementary instructions, and other content that the user must understand.	 Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	 Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	It is used for buttons, menus , page names, and other UI elements.	Click OK.
`Courier font`	It is used for commands.	Run the `cd / d C :/ windows` command to enter the Windows system folder.
Italics	It is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	It indicates that it is a optional value, and only one item can be selected.	`ipconfig` *[-all|-t]*

Style	Description	Example
{} or {a|b}	It indicates that it is a required value, and only one item can be selected.	`swich {stand | slave}`

Contents

1 Step 1 : Prerequisites

Make sure that the following prerequisites are met before you create an E-MapReduce cluster:

1. Create an Alibaba Cloud account

   To create an E-MapReduce cluster, you must have an Alibaba Cloud account, which is used to uniquely identify you in the Alibaba Cloud ecosystem. You can use this account to create E-MapReduce clusters and activate other Alibaba Cloud services, including *Object Storage Service (OSS)* and *ApsaraDB for RDS (RDS)*.

   For more information about creating an Alibaba Cloud account, see *Sign up with Alibaba Cloud*.

2. Create an AccessKey (optional)

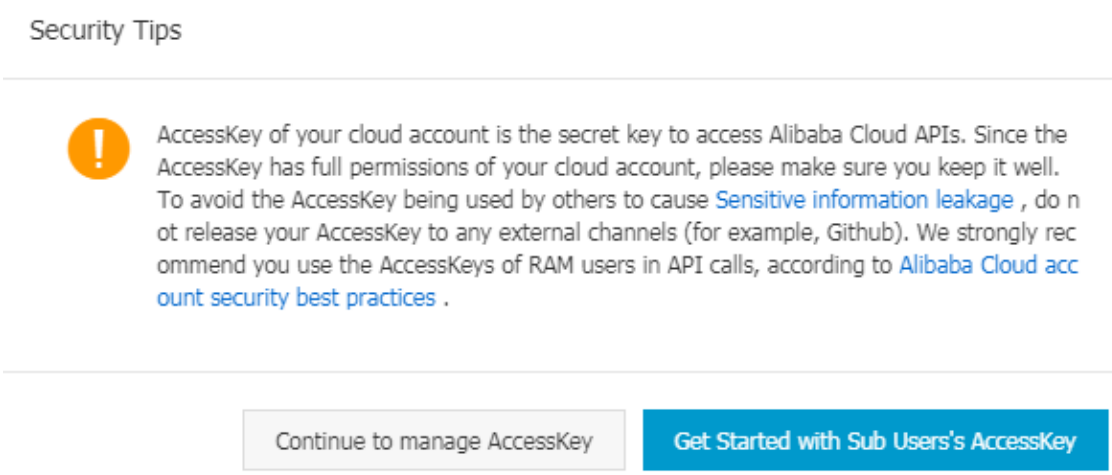
   To use E-MapReduce, you must create a minimum of one AccessKey. Follow these steps to create an AccessKey:

   a. Log on to the *Alibaba Cloud website*.

   b. Go to the Alibaba Cloud console.

   c. Hover over your profile picture and then click AccessKeys.

   > **Notice:**
   > If the following dialog box appears, click Continue to manage AccessKey .

   d. Click Create AccessKey.

   e. The AccessKey has been created.

3. Authorize default system role to the E-MapReduce service account.

   When you activate the E-MapReduce service, a default system role named AliyunEMRDefaultRole must be granted to the E-MapReduce service account. Once assigned, E-MapReduce can call the relevant services (such as ECS and OSS), create clusters, save logs, and perform other related tasks.For more information, see*Role authorization*.

4. Activate Alibaba Cloud OSS

   E-MapReduce stores the job and execution logs on Alibaba Cloud OSS. Therefore, you must activate Alibaba Cloud OSS. For more information, see *Sign up for OSS*. Create OSS buckets in the same region as the EMR cluster that you need to create. For more information, see *Create a bucket*.

5. Create high-configuration instances (optional)

   If you need to create instances with eight or more cores for a Pay-As-You-Go-based cluster, log on to your Alibaba Cloud account and submit a ticket for application. For more information, see *Support and Services*.

2 Step 2 : Create a cluster

This section describes how to create and configure an E-MapReduce (EMR) cluster.

Prerequisite

Confirm that RAM authorization has been completed. For more information, see *Role authorization*.

Go to the cluster creation page

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. Select a region where you want to create the cluster. The region cannot be changed after the cluster has been created.
3. Click Create Cluster in the upper-right corner.

Create a cluster

To create a cluster, follow these steps:

- Software configuration
- Hardware configuration
- Basic configuration

Step 1: Software configuration

Description

- EMR Version: By default, the latest version is selected.
- Cluster Type: Currently, E-MapReduce (EMR) provides the following cluster types:
  - Hadoop clusters. These clusters provide multiple ecosystem components, such as Hadoop, Hive, Spark, Spark Streaming, Flink, Storm, Presto, Impala, Oozie , and Pig. Hadoop, Hive, and Spark are semi-hosted services and are used for distributed large-scale data storage and computing. Spark Streaming, Flink, and Storm can provide stream computing. Presto and Impala are used to achieve interactive queries. For more information about the components, see Services List displayed on the cluster and service management page.
  - Kafka clusters. These clusters serve as a semi-hosted distributed message system with high throughput and high scalability. Kafka clusters provide a comprehensive service monitoring system that keeps the clusters running

stably. Kafka clusters are more professional, reliable, and secure. You do not need to deploy or maintain these clusters. These clusters are commonly used in scenarios such as log collection and monitoring data aggregation. Offline data processing and stream computing, and real-time data analysis are also supported.

- Druid clusters. These clusters provide semi-hosted and real-time interactive analysis services. Druid clusters also support querying large amounts of data in milliseconds and writing data through multiple methods. Druid clusters offer flexible and stable real-time queries together with other services such as EMR Hadoop, EMR Spark, OSS, and RDS.
- Data Science clusters. These clusters are commonly applied in big data and AI scenarios. Data Science clusters provide Hive and Spark offline big data ETL, and TensorFlow model training. You can choose the CPU plus GPU heterogeneous computing framework and the deep learning algorithm supported by NVIDIA CPUs to run computing tasks efficiently.

· Required Services: Select the default configuration. You can add, enable, or disable services on the management page later.
· High Security Mode: Indicates whether to enable the Kerberos authentication feature for the cluster. This mode is disabled by default because typically this mode is not required by clusters created for general users.
· Enable Custom Setting: You can specify a JSON file to modify software configuration. For more information about the procedure, see *Software configuration*.

Step 2: Hardware configuration

Description

· Billing Configuration
  - Billing Method: You can select Pay-As-You-Go when testing the cluster. After all tests have been passed, you can create and use a Subscription-based cluster.

- Network configuration
  - Zone: Typically, the default zone is used.
  - Network Type: VPC is selected by default. If you have not created a VPC, go to the *VPC console* to create one.
  - Zone: Geographical areas that are located in the same region. These zones are interconnected through VPCs.
  - VPC: Select a VPC that has been created in the specified region. If no VPC is available, click Create VPC/VSwitch to create one in the current zone.
  - VSwitch: Select a VSwitch for the specified VPC in the current zone. If no VSwitch is available, go to the VPC console and create one in the current zone.
  - Security Group Name: By default, no security group is available if this is the first time that you have created a cluster. You need to enter a name and then create a security group. If you have already created security groups, select a security group.
- Cluster configuration
  - High Availability: After this feature has been enabled, two master nodes will be provided to guarantee the high availability of Resource Manager and Name Node. HBase clusters support high availability by default. An HBase cluster has to use a core node as one of the two master nodes. When the high availability feature is enabled, the HBase cluster only needs one mater node to support high

availability, which is more secure and reliable. If you need to create clusters that support high availability, enable high availability during testing.

- Master node: Deploys processes, such as Resource Manager and Name node.
  - Master Instance Type: Select instance specifications based on your needs. For more information, see *Instance type families*.
  - System Disk Type: Select Ultra Disk or SSD Disk based on your needs.
  - System Disk Size: You can resize the disk based on your needs. The recommended minimum disk size is 120 GB.
  - Data Disk Type: Select Ultra Disk or SSD Disk based on your needs.
  - Data Disk Size: You can resize the disk. The recommended minimum disk size is 80 GB.
  - Master Instances: The default number of master instances is one.
- Core instance: Stores all cluster data. You can scale up the instance based on your needs.
  - Core Instance Type: Select instance specifications based on your needs. For more information, see *Instance type families*.
  - System Disk Type: Select Ultra Disk or SSD Disk based on your needs.
  - System Disk Size: You can resize the disk based on your needs. The recommended minimum disk size is 80 GB.
  - Data Disk Type: Select Ultra Disk or SSD Disk based on your needs.
  - Data Disk Size: You can resize the disk based on your needs. The recommended minimum disk size is 80 GB.
  - Core Instances: The default number of core instances is two. You can adjust the number of core instances based on your needs.
- Task Instance Group: No data is stored on task instance groups. Task instance groups are used to adjust the computing capability of clusters. This feature is disabled by default. You can enable it based on your needs.

Step 3: Basic configuration

Description

- Basic Information
  - Cluster Name: Enter the name of the cluster. It must be 1-64 characters in length and can contain Chinese characters, uppercase English letters, lowercase English letters, numbers, hyphens (-), and underscores (_).
- Running Logs
  - Running Logs: You can enable or disable this feature. This feature is enabled by default. After you enable this feature, you must specify an OSS path to save the logs. All running logs are stored in the specified OSS directory. To enable this feature, you must first activate OSS. The uploaded data is billed based on the space usage. We recommend that you enable this feature, which helps in debugging and troubleshooting.
  - Log Path: Specify an OSS path to store the log.
  - Uniform Meta Database: Hive uses a unified metadatabase, which is independent of the cluster. The meta information will not be deleted after the cluster is released. We recommend that you disable this feature if you are an E-MapReduce beginner.
- Permission Settings: Typically, the default setting is used.
- Logon Settings
  - Remote Logon: Indicates whether to open port 22 of the security group. This feature is enabled by default.
  - Key Pair:For the use of the key pair, please refer to *SSH Key Pair*
  - Password: Set the password used to log on to the master instance. The logon password must be 8 to 30 characters in length and can contain uppercase letters, lowercase letters, numbers , and special characters including exclamation marks (!), at signs (@), number signs (#), dollar signs ($), percent signs (%), ampersands (&), and asterisks (*).
- Bootstrap Operation (optional): You can set the cluster to run your custom script before it starts Hadoop. For more information, see Bootstrap action.

Configuration List

Confirm the configured items and fees in the configuration list.

Confirm the creation

After you configure the settings and make sure that all settings are valid, the Create button is highlighted. Confirm the information, and click Create to create a cluster.

Notice:

- The cluster is created immediately if the billing method is Pay-As-You-Go. You are the directed to the cluster list page. You can find an initializing cluster in the Clusters list. It may take several minutes to complete the cluster creation. After the cluster has been created, it changes its status to Idle.
- An order is generated if the billing method is Subscription. The cluster will be created after you complete the payment.

Creation failed

If the creation fails, CREATE_FAILED is displayed on the clusters list. Move the pointer over the red exclamation mark (!) to view the cause.

You do not to handle clusters that are failed to be created because no computing resources have been created for these clusters. These clusters will be automatically hidden after three days.

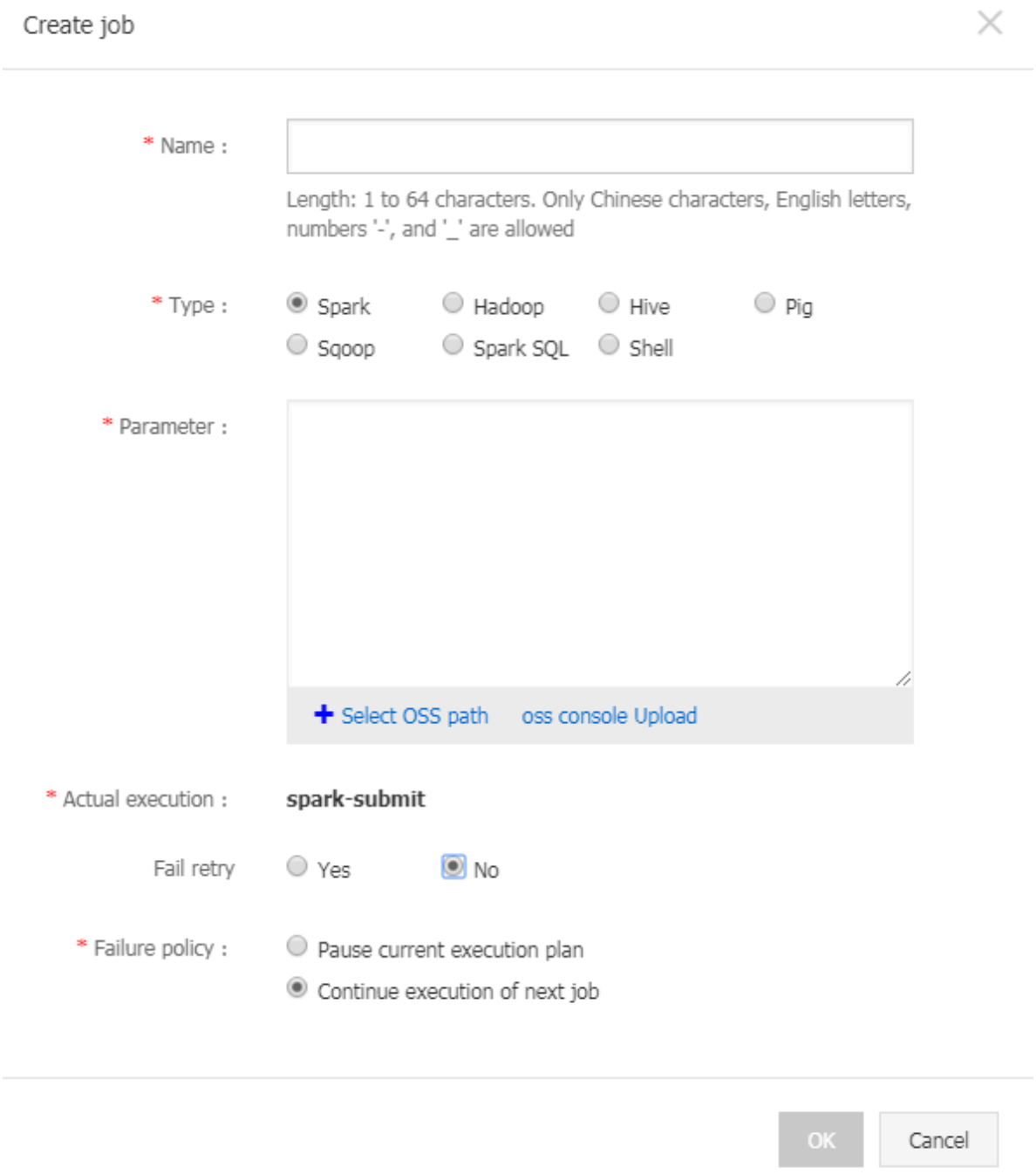
3 Create a job

This section describes how to create a job for scheduling in an early E-MapReduce version.

To run a computing task, you need to follow these steps to define a job:

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. Select a region where you want to create the job.
3. Click the Old EMR Scheduling tab to go to the jobs list page.

4. Click Create job in the upper-right corner to go to the job creation page, as shown in the following figure:



5. Enter the job name.
6. Select a job type.
7. Enter parameters of the job. The application parameters must include the following information: the JAR package run by the job, data input and output addresses of the job, and certain command line parameters. You must enter all parameters in the command line in this field. If you need to use an OSS path, click

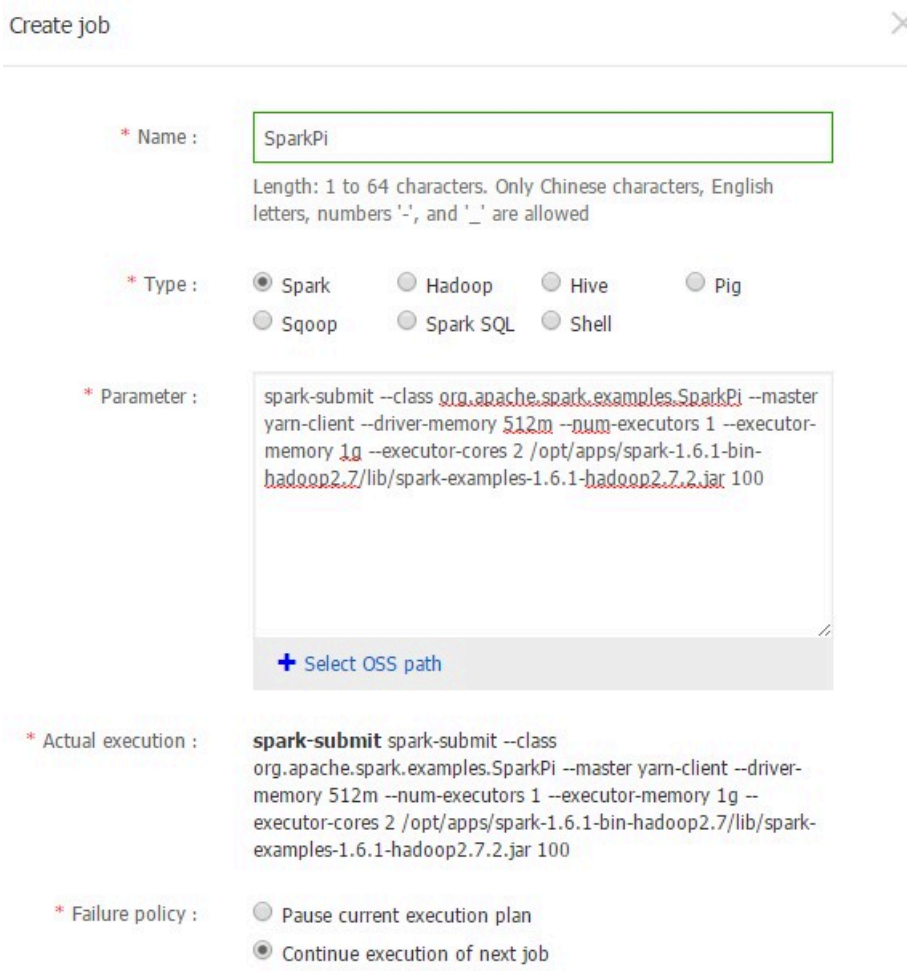
Select OSS Path to select the OSS resource path. For more information about the parameters of all job types, see the *Job* section in *User Guide*.

8. Actual execution. The job command that has been executed on ECS will be displayed in this field. You can copy the displayed command and directly run it in the command line environment on an E-MapReduce cluster.
9. Fail Retry. You can set the number of retries and their intervals. This feature is disabled by default.
10. Select a failure policy. Pause the current execution plan: Indicates that the current execution plan will be paused if this job fails and will wait for you to process. Continue execution of the next job: Indicates that errors will be ignored and the next job will be executed after this job fails.
11. Click OK to complete the creation.

Example

This is a Spark job, which sets parameters such as the input and output paths in the application parameters.

> **Notice:**
> This example is for reference only.



OSS and ossref

The oss:// prefix indicates that the data path points to an OSS path, which specifies the operation path when reading/writing the data. This is similar to hdfs://.

The ossref:// prefix also indicates that the data path points to an OSS path. However, it is used to download the corresponding code to a local disk, and then replace the path in the command line with this local path. It is easier for you to run native code. You do not need to log on to the computer to upload the code and the dependent resource packages.

In this example, the ossref://xxxxxx/xxx.jar parameter represents the JAR package of job resources. This JAR package is stored on OSS. When this path is executed in the code, the JAR package will automatically download to the cluster and be executed. The two oss://xxxx and the two values following the JAR package are processed by the main class in the JAR package as parameters.

Notice:

The ossref cannot be used to download large amounts of data. Otherwise, an error may occur.

4 Create an execution plan

This section describes how to create an execution plan in an earlier E-MapReduce version.

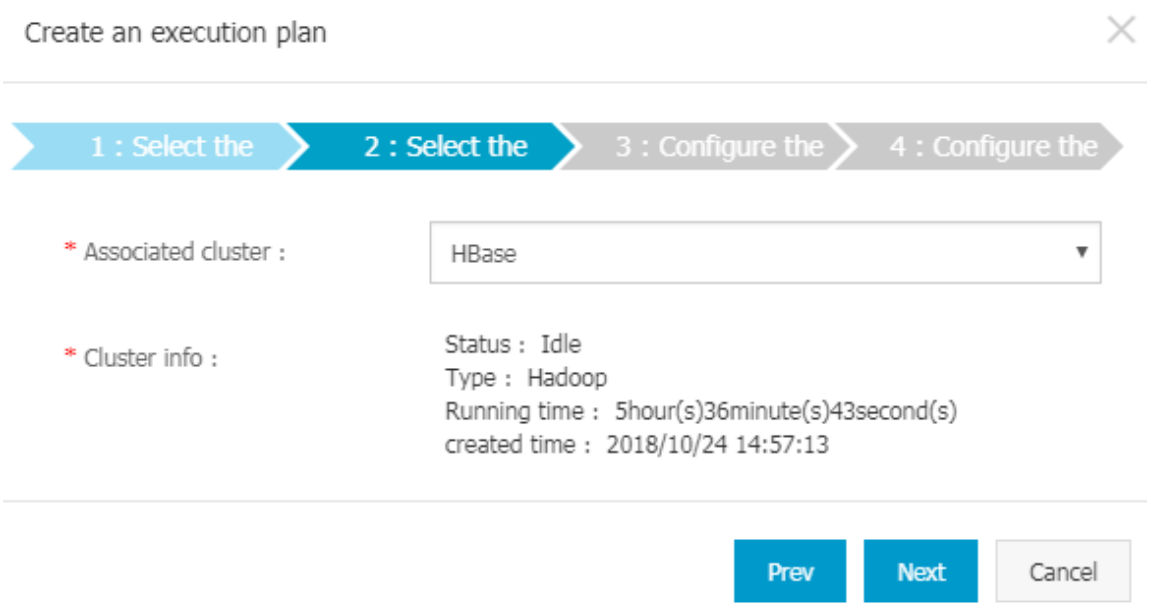
After you have created a job, you need to create an execution plan if you want to run a predefined job on a cluster. An execution plan may contain multiple jobs. You can define the execution sequence of these jobs. For example, if a scenario handles data in the following order: prepare data, process data, and clean up data, then you can define the following three jobs: prepare-data, process-data, and cleanup-data, and then you can create an execution plan that includes these three jobs.

To create an execution plan, follow these steps:

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. Select a region.
3. Click Old EMR Scheduling.
4. Click the left-side Execution Plan tab to go to the execution plan page.
5. Click Create an execution plan in the upper-right corner to go to the execution plan creation page.
6. The following two options are available on the Select the cluster page: Create as needed and Existing Clusters. Create as needed indicates that you have not created any clusters. You will run the execution plan in a temporary cluster, which will be automatically released after you have executed the plan. Existing Cluster indicates

that you have at least one running cluster, and you have to submit the execution plan to the existing cluster for execution.

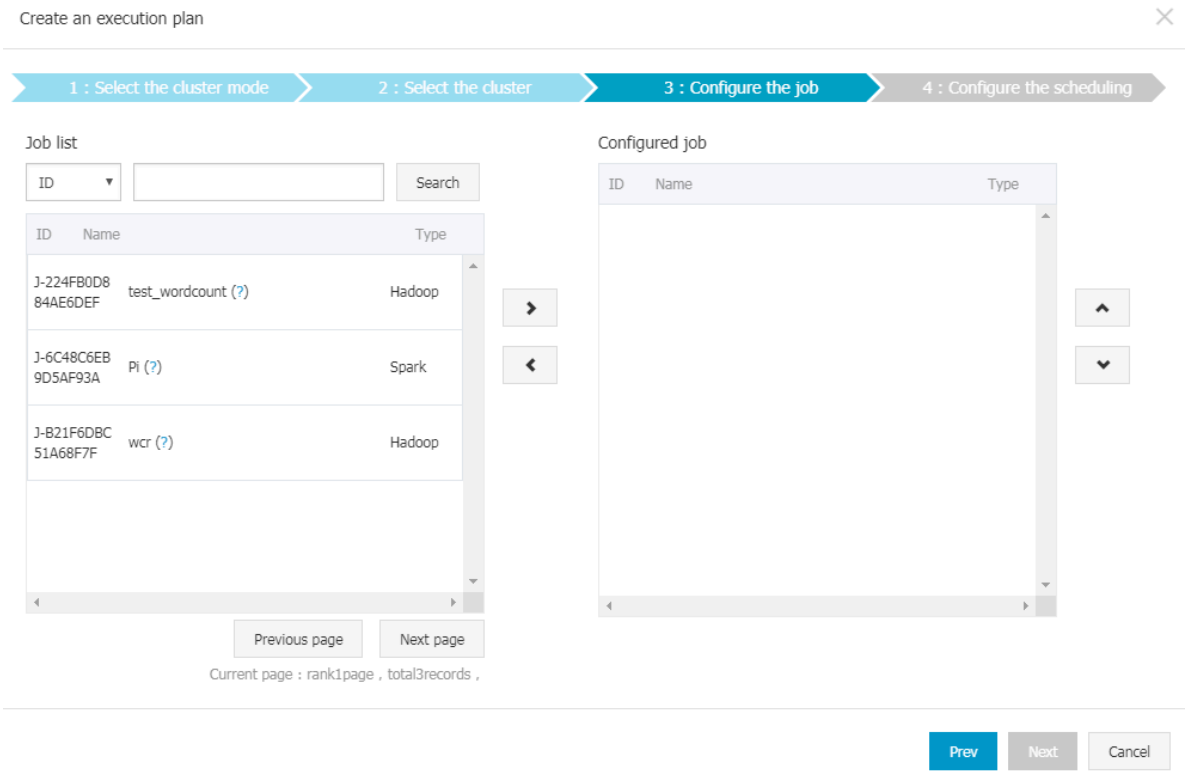
a. If you select Create as needed, you can follow the same steps as creating a cluster to create an on-demand cluster. Select configuration of the on-demand cluster, and click OK.
b. If you select Existing Cluster, you will be directed to the cluster selection page. You can select clusters that you want to associate with the execution plan.



Currently, you can only submit clusters that are in Running or Idle status.

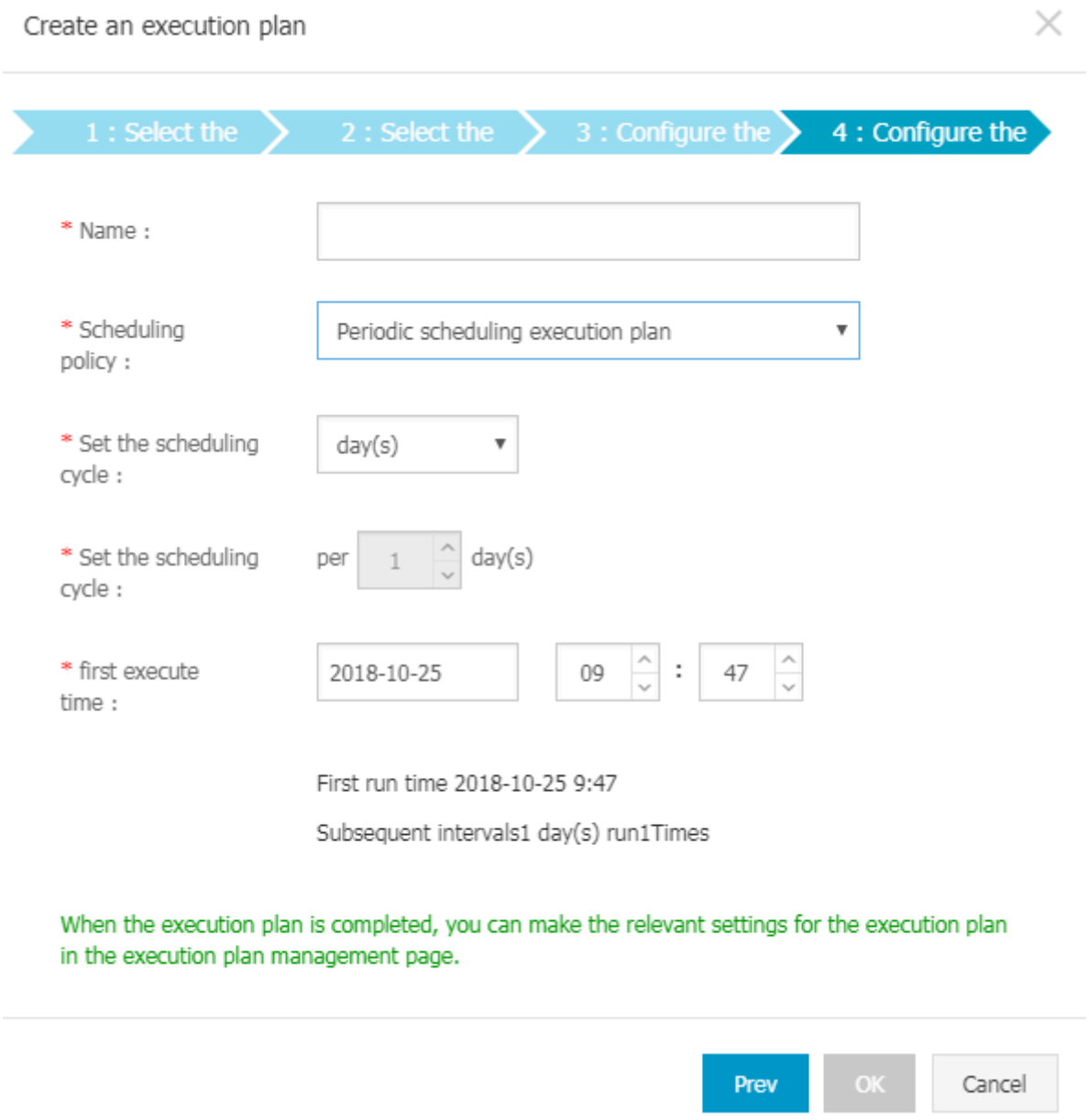
7. Click Next to go to the Configure the job page. The page displays the predefined jobs on the left-side list, and the jobs to be executed by the newly created execution plan on the right-side list. To define the execution plan, select jobs from the left-side list based on the execution sequence and add them to the right-side list. You

can click the question mark (?) to view parameters of the jobs. After you have defined the execution plan, click Next.



8. Enter an execution plan name.

9. Select a scheduling policy.
   - Manually executed indicates that the execution plan will be executed only when you click the execution plan.
   - Periodic scheduling execution plan indicates that you must set the scheduling cycle and the scheduling start time.



10. Click OK to complete the creation of the execution plan.

5 Select configuration

To use E-MapReduce (EMR), you must select appropriate Hadoop clusters. When selecting configurations for EMR, consider the use of big data in your enterprise, estimate the amount of data, the reliability of services, and your financial budget.

Big data scenarios

EMR has been applied in the following enterprise big data scenarios:

- Batch operations. This scenario requires high disk I/O throughput and high network throughput, but has low real-time capacity requirements. If you need to process large amounts of data but do not require real-time processing, you can use MapReduce, Pig, and Spark. This scenario does not require a high memory capacity. Therefore, you need to focus on the CPU, memory, and network requirements when you perform data shuffling.
- Ad hoc queries. Data scientists and data analysts use ad hoc queries to retrieve data . This scenario requires real-time queries, high disk I/O throughput, and network throughput. In this scenario you can use Impala and Presto. This scenario also has high memory requirements. You need to consider the amount of data and concurrent queries.
- Stream computing, high network throughput, and compute-intensive scenarios. For these scenarios, you can use Flink, Spark Streaming, and Storm.
- Message queues. This scenario requires high disk I/O throughput and network throughput, consumes large amounts of memory, and the storage does not depend on HDFS. Therefore, you can choose to use Kafka clusters in EMR. EMR clusters are divided into Kafka clusters and Hadoop clusters to avoid impacting Hadoop.
- Cold backup. This scenario does not require high disk I/O throughput or computing throughput and is low cost. We recommend that you use EMR d1 instances for cold backups. The storage cost of d1 instance is 0.03 USD/month/GB.

EMR instances

An EMR cluster consists of *three types of instances*: master instances, core instances, and task instances.

You can select ultra disks, SSD disks, and local disks for EMR storage. Performance of different disks is: SSD disks > local disks > ultra disks.

EMR underlying storage supports OSS (standard OSS only) and HDFS. OSS has a higher data availability than HDFS. The data availability of OSS is 99.99999999%, while the data availability of HDFS depends on the reliability of cloud disk or local disk storage.

Storage prices are as follows:

- Instance with local disks: USD 0.003/GB/month.
- OSS standard storage: 0.02 USD/GB/month.
- Ultra disk storage: 0.05 USD/GB/month.
- SSD disk storage: 0.143 USD/GB/month.

Select EMR configuration

- Select master instance configuration
  - Master instances are used to deploy the master processes of Hadoop, such as NameNode and ResourceManager.
  - We recommend that you enable high availability on production clusters. High availability is available for EMR components such as HDFS, YARN, Hive , and HBase. We recommend that you enable high availability in the "Cluster configuration" step when you create the cluster. If high availability is not enabled when you create an EMR cluster, it cannot be enabled later.
  - Master instances are used to store HDFS metadata and component log files. They are compute-intensive with low disk I/O requirements. HDFS metadata is stored in memory. The minimum recommended memory size is 16 GB based on the number of files.
- Select core instance configuration
  - Core instances are used to store data, run computing tasks and processes such as DataNode and NodeManager.
  - If the amount of data stored in HDFS (3 backups) exceeds 60 TB, we recommend that you use instances with local disks (ECS d1instances and ECS d1ne instances). The local disk capacity is calculated as follows: (the number of CPU cores/2) * 5.5 TB * the number of instances. For example, if you purchase four d1 instances with 8 cores, the local disk capacity is: 8/2 * 5.5 * 4=88 TB. HDFS requires three backups. Therefore, you need to buy at least three instances that use local disks.

We recommend that you purchase at least four instances for data reliability and disk recovery.

- If the amount of data stored in HDFS is less than 60 TB, you can use ultra disks or SSD disks.

· Select task instance configuration

- Task instances are used when the CPUs and memory of core instances have insufficient computing capability. Task instances do not store data or run DataNode. You can estimate the number of instances based on CPU and memory requirements.

EMR lifecycle

EMR supports auto-scaling, which allows you to quickly *scale up a cluster*. You can flexibly adjust the configuration of cluster nodes, or *upgrade or downgrade ECS instance configuration*.

Select a zone

We recommend that you deploy EMR and your business system in *the same zone and region*.

6 Connect to a cluster using SSH

You can connect to the master node using SSH to view detailed settings and status of the jobs in the CLI. The public IP address of the master node is shown on the cluster overview page.

Environment variables on a cluster

Environment variables have been configured on the instances. Frequently-used environment variables are as follows.

- JAVA_HOME
- HADOOP_HOME
- HADOOP_CONF_DIR
- HADOOP_LOG_DIR
- YARN_LOG_DIR
- HIVE_HOME
- HIVE_CONF_DIR
- PIG_HOME
- PIG_CONF_DIR

You can reference the variables in the script. We recommend that you do not modify the values of the variables to avoid unexpected errors in the EMR cluster.

Connect to the master node

1. Run the following command to connect to the master node using SSH. Check the public IP address of the master node in the Host section on the *#unique_16* page.

   ```
   ssh   root @ ip . of . master
   ```

2. Enter the password you have set when creating the cluster.

Connect to the master node using SSH without a password

Typically, you need to connect to the cluster for management and operations. You can quickly connect to the master node by using SSH without a password. The master node in a cluster is assigned with a public IP address by default. Procedure:

1. Connect to the master node as the root user by using a password.
2. Switch to the hadoop user or the hdfs user.

Use SSH on Linux

1. Send the private key file to your local system.

```
sz ~/. ssh / id_rsa
```

2. Go back to your local machine and connect to the master node again.

```
ssh - i private_ke y_path / id_rsa hadoop @ 120 . 26 . 221 .
130
```

If you only have one private key file, you can store it under the `~/. ssh /` directory for immediate use without specifying the key path using the `- i` option every time.

Use SSH on Windows

You have multiple methods to connect to the master node using SSH without a password on Windows. The methods are shown as follows.

- Method 1: Use PuTTY

1. Click *Download PuTTY*.
2. Download PuTTYgen through this link.
3. Start PuTTYgen and load your private key.

> Notice:
> Keep the private key safe. Generate a new key once the private key is stolen.

4. Save the private key with the default configurations. A PuTTY private key file with the .ppk extension is generated.
5. Run PuTTY and choose Session on the configuration page.
6. Enter the public IP address of the node that you want to connect to with a logon username. For example, hadoop@MasterNodeIP.
7. Select Connetion > SSH > Auth on the configuration page.
8. Select the generated PPK file.
9. Click Open to connect to the master node.

- Method 2: Use Cygwin or MinGW

  Cygwin and MinGW are easy-to-use tools for compiling Linux on Windows.

  You can refer to the operations described in the Use SSH on Linux section.

  We recommend that you use MinGW, which is the most light-weight method for using SSH on Windows. If you cannot load the official website, download Git for Windows and use the built-in Git BASH instead.

View the Web UIs of Hadoop, Spark, and Ganglia

> Note:
> Make sure that you have completed the *SSH password-less logon* procedure.

For security reasons, the ports of Web UIs for monitoring Hadoop, Spark, and Ganglia in an EMR cluster are closed. If you want to access these Web UIs, you need to create SSH tunnels and enable port forwarding. The methods are shown as follows:

> Notice:
> The following operations are based on your local machine, not on a cluster node.

- Method A: Dynamic port forwarding
  - Use a private key:

    Create an SSH tunnel to allow communication between your local machine and a dynamic port on the master node in the EMR cluster.

    ```
    ssh -i /path/id_xxx -ND 8157 hadoop@masterNode IP
    ```

  - Use a username and password:

    ```
    ssh -ND 8157 hadoop@masterNode IP
    ```

  Replace 8157 with any port number on your local machine that has not been used.

  After dynamic port forwarding is complete, you can view a Web UI using the following methods.

  - Recommended method:

    We recommend that you use the Chrome browser. You can run the following command to access a Web UI.

    ```
    chrome --proxy-server="socks5://localhost:8157" --
    host-resolver-rules="MAP * 0.0.0.0 , EXCLUDE
    localhost" --user-data-dir=/tmp/
    ```

    For Windows, an example of the *temporary file path* is *d:/tmppath*. For Linux and OSX, the format of the temporary file path is like */tmp/*.

    Chrome is installed on different paths based on the operating system. For more details, see the following table.

OS	Chrome path
Mac OS X	/Applications/Google Chrome.app/Contents/MacOS/Google Chrome
Linux	/usr/bin/google-chrome

OS	Chrome path
Windows	C:\Program Files (x86)\Google\Chrome\Application\chrome.exe

- Extension:

  ■ Use a Chrome extension to view the Web UIs.

  1. Install the Chrome extension *SwitchyOmega*.
  2. After installation is complete, click the icon of SwitchyOmega and click Options in the pop-up menu to perform configurations.
  3. Click New Profile. In the Profile name field, enter "SSH tunnel." Click PAC Profile for the type of the profile.
  4. In the PAC Script editor, enter the following code.

```
function regExpMatc h ( url , pattern ) {
  try { return new RegExp ( pattern ). test ( url
); } catch ( ex ) { return false ; }
}

function FindProxyF orURL ( url , host ) {
   // Important : replace 172 . 31 below with the
  proper prefix for your VPC subnet

    if ( shExpMatch ( url , "* localhost *")) return "
SOCKS5 localhost : 8157 ";
    if ( shExpMatch ( url , "* emr - header *")) return "
SOCKS5 localhost : 8157 ";
    if ( shExpMatch ( url , "* emr - worker *")) return "
SOCKS5 localhost : 8157 ";

    return ' DIRECT ';
}
```

  5. In the left-side navigation pane, click Apply changes to complete the configurations.
  6. Start a CLI. Choose one of the following methods and run the corresponding command.

```
// Method a : Use a private key
 ssh - i / path / id_xxx - ND 8157 hadoop @
 masterNode IP
```

```
// Method b : Use a username and password
```

```
ssh -ND 8157 hadoop@masterNode IP
```

7. Click the SwitchyOmega icon in the Chrome menu. From the drop-down list, select SSH tunnel.
8. In the address bar, enter the IP address of the node and the port number to access the Web UI. The node refers to the one that you want to connect to using the SSH command. Generally, it is a master node. Two frequently-used ports are port 8088 for YARN and port 50070 for HDFS.

By using this method, you can browse webpages and use Web UIs of clusters at the same time.

- After creating an SSH tunnel between your local machine and the master node of the EMR cluster, you need to configure a local agent for viewing the Web UIs of Hadoop, Spark, and Ganglia through browsers. Procedure:

1. Assume that you use Chrome or Firefox. Click *Download FoxyProxy Standard*.
2. After the installation is complete, restart your browser, open a text editor, and then enter the following code.

```
<?xml version="1.0" encoding="UTF-8"?>
<foxyproxy>
<proxies>
<proxy name="aliyun-emr-socks-proxy" id="
2322596116" notes="" fromSubscription="false"
enabled="true" mode="manual" selectedTabIndex
="2" lastresort="false" animatedIcons="true"
includeInCycle="true" color="#0055E5" proxyDNS="
true" noInternalIPs="false" autoconfMode="pac
" clearCacheBeforeUse="false" disableCache="false
" clearCookiesBeforeUse="false" rejectCookies="
false">
<matches>
<match enabled="true" name="120.*" pattern="http
://120.*" isRegEx="false" isBlackList="false
" isMultiLine="false" caseSensitive="false"
fromSubscription="false"></match>
</matches>
<manualconf host="localhost" port="8157"
socksversion="5" isSocks="true" username=""
password="" domain=""></manualconf>
</proxy>
</proxies>
```

```
</ foxyproxy >
```

In this example:

- Port `8157` on your local machine is used to establish an SSH tunnel to the master node of the cluster.
- Replace `120 .*` with the actual IP address of the master node.

3. Click the Foxyproxy icon in the Chrome menu. From the drop-down list, select Options.
4. Select Import/Export.
5. Select the XML file you have created. Click Open.
6. In the Import FoxyProxy Setting dialogue box, click Add.
7. Click the Foxyproxy icon. From the drop-down list, select Use Proxy aliyun-emr-socks-proxy for all URLs.
8. Enter localhost: 8088 in the address bar to view the Web UI of Hadoop.

- Method 2: Local port forwarding

Notice:
If you use this method to view a Web UI, errors will occur when you try to open the subpages.

- Use a private key:

```
ssh  - i  / path / id_rsa  - N  - L   8157 : masterNode  IP :
8088   hadoop @ masterNode  IP
```

- Use a username and password:

```
ssh  - N  - L   8157 : masterNode  IP : 8088   hadoop @
masterNode  IP
```

Parameter description:

- `path` : The path where the private key file is stored.
- `masterNode IP` : The IP address of the master node to be connected.
- `8088` : The port that is used by the Resource Manager Web UI on the master node.

7 Create E-MapReduce