ALIBABA CLOUD

Alibaba Cloud

E-MapReduce

Data Development

Document Version: 20220119

Alibaba Cloud

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 because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should 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 this document based on the "status quo", "being defective", and "existing functions" of its products and services. 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 take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, 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 contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, 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 this document 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 directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
Danger	A danger notice 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.
Warning	A warning notice 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 restart an instance.
Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	**Notice:** If the weight is set to 0, the server no longer receives new requests.
Note	A note indicates supplemental instructions, best practices, tips, and other content.	**Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.Overview

After you create an E-MapReduce (EMR) cluster, you can create a project in Data Platform. Data Platform is a workflow platform where you can develop, schedule, and monitor jobs and workflows. You can define a set of jobs that have dependencies by using a directed acyclic graph (DAG) and run the jobs in sequence based on the dependencies. You can manage jobs, schedule tasks, and monitor the status of jobs in the EMR console to manage and maintain workflows.

> **Notice** If your high-security EMR cluster is connected to an external MIT key distribution center (KDC), you cannot use the features of Data Platform.

Data Platform provides the following features:

- Project management: You can associate cluster resources with projects and add project members. For more information, see Manage projects.
- Development and editing of big data jobs: You can develop various types of jobs, such as Hive, Hive SQL, MapReduce, Spark, and Shell. For more information, see Edit jobs.
- Workflow development and scheduling: You can perform drag-and-drop operations to build a workflow. You can also configure time-based scheduling policies and dependencies among workflows. For more information, see Edit a workflow.
- Ad hoc query: Four types of ad hoc query jobs are supported: Hive SQL, Spark SQL, Spark, and Shell. For more information, see Perform ad hoc queries.
- Information viewing: You can view the running records and logs of tasks and workflows, and run failed jobs and workflows again. You can also view the operation history of project members in a project. For more information, see Scheduling center.

2.Manage projects

After you create an E-MapReduce (EMR) cluster, you can create a project on the Data Platform tab. Then, you can edit jobs and schedule workflows in the project. You can also associate a cluster with the project, add project members, and configure global variables for the project.

Prerequisites

An EMR cluster is created. For more information, see Create a cluster.

Limits

You can use only an Alibaba Cloud account to create projects, add project members, and associate clusters with projects. If you log on to the EMR console by using a RAM user, the **Create Project** button and the **Users** and **Cluster Settings** pages are unavailable.

Create a project

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.

   If you use an Alibaba Cloud account, you can view all the projects within the account. If you use a RAM user, you can view only the projects on which you have development permissions. You can use your Alibaba Cloud account to grant development permissions to a RAM user. For more information, see Manage RAM users.
2. In the upper-right corner of the **Projects** section, click **Create Project**.
3. In the **Create Project** dialog box, configure **Project Name** and **Project Description** and select an existing resource group from the **Select Resource Group** drop-down list.

   > **Note** If you do not select a resource group, the project is added to the default resource group. For more information about how to use resource groups, see Use resource groups.
4. Click **Create**.

   In the **Projects** section, you can view and manage the project you created.

View the basic information about a project

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. Go to the Projects tab.
   i. In the **Projects** section, click the ID of your project.
   ii. Click the **Projects** tab.
3. View the basic information about the project.

   On the **Basic Information** page, you can view the following information about the project: project name, creation time, description, and the user who created the project.

Configure general information

We recommend that you enable the security mode on the General Configuration page if you want to manage permissions on jobs that are run in Data Platform of the EMR console.

After you enable the security mode, you must add the EMR user account that is used to submit jobs on the Users page. For more information, see Manage user accounts. If you log on to the EMR console by using your Alibaba Cloud account and submit a job in a project for which the security mode is enabled, the job is run by the hadoop user by default. If you log on to the EMR console by using a RAM user and submit a job in a project for which the security mode is enabled, the job is run by the EMR user account that has the same name as the RAM user by default.

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. Go to the Projects tab.
   i. In the **Projects** section, click the ID of your project.
   ii. Click the **Projects** tab.
3. Configure Security Mode.
   i. In the left-side navigation pane, click **General Configuration**.
   ii. Turn on or off Security Mode based on your business requirements.

   > **Notice** After you enable the security mode, Shell and Hive jobs cannot be run in the project.

Manage RAM users

Perform the following steps to add or revoke project development permissions to or from a RAM user:

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.

iii. Click the **Data Platform** tab.

2. Go to the Projects tab.
   i. In the **Projects** section, click the ID of your project.
   ii. Click the **Projects** tab.
3. In the left-side navigation pane, click **Users**.
4. On the **Users** page, add or remove users based on your business requirements.
   - Add a RAM user.
     a. In the upper-right corner of the Users page, click **Add User**.
     b. In the **Add User** dialog box, select the RAM user that you want to add and click **Add**.

        You can view information about the added RAM user on the **Users** page.

        > **Note** The added RAM user becomes a member of the project and is granted the permissions to view and develop jobs and workflows in the project.
   - Remove a RAM user.

     On the **Users** page, find the RAM user that you want to remove and click **Delete** in the Actions column.

Configure cluster resources

Perform the following steps to configure cluster resources for a project. This way, jobs in the project can run in the cluster that is associated with the project:

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. Go to the Projects tab.
   i. In the **Projects** section, click the ID of your project.
   ii. Click the **Projects** tab.
3. In the left-side navigation pane, click **Cluster Settings**.
4. On the **Cluster Settings** page, perform the following operations:
   - Associate a cluster with the project.
     a. Click **Add Cluster** in the upper-right corner.
     b. In the **Add Cluster** dialog box, select a resource group and a cluster.

        In the Add Cluster dialog box, select a purchased subscription or pay-as-you-go cluster from the **Select Cluster** drop-down list. Clusters that are created by using a cluster template are not supported.
     c. Click **OK**.

        On the **Cluster Settings** page, you can view the information about the associated cluster.
   - Modify cluster configurations.
     a. Find the cluster whose configurations you want to modify and click **Change Configuration** in the Actions column.
     b. In the **Change Configuration** dialog box, configure the parameters that are described in the following table.

Parameter	Description
Default Job Submission User	The default user who submits jobs to the associated cluster in the project. The default value is hadoop. The default user is unique.
Default Job Submission Queue	The default queue to which jobs are submitted in the project. Default value: default.
Job Submission User Whitelist	The users who can submit jobs in the project to the associated cluster. Separate multiple users with commas (,).
Job Submission Queue Whitelist	The queues to which jobs can be submitted in the project. Separate multiple queues with commas (,).
Client whitelist	Specify the clients that can submit jobs. You can select the master node of the existing EMR cluster or a node of the gateway cluster that is associated with the EMR cluster. Self-managed gateway clusters that are deployed on ECS instances are not supported.

     c. Click **OK**.
   - Disassociate a cluster from the project.

     On the **Cluster Settings** page, find the cluster that you want to disassociate and click **Delete** in the Actions column.

Define variables

Perform the following steps to configure project-level custom variables, which can be used as global variables for jobs in a project:

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. Go to the Projects tab.
   i. In the **Projects** section, click the ID of your project.
   ii. Click the **Projects** tab.

3. In the left-side navigation pane, click **Custom Variable**.
4. On the **Custom Variable** page, you can add or remove custom variables based on your business requirements.
   - Add a custom variable.
     a. Click **Add** in the upper-right corner.
     b. In the **Add Custom Variable** dialog box, configure **Variable Name** and **Value**, and specify whether to encrypt the value of the variable. If you want to encrypt the value, turn on Set as Password.

        The variable is called in the format of `${VariableName}` in a job. For example, a variable named ENV_ABC is added, the value of the variable is 12345, and **Set as Password** is not turned on. In this example, a job that has the following content is run:

        ```
        echo ${ENV_ABC}
        ```

        The following output is returned:

        ```
        12345
        ```

        The effect of configuring the variable is equivalent to running the following script:

        ```
        export ENV_ABC=12345
        ```

     c. Click **OK**.

        On the **Custom Variable** page, you can view the information about the added variable.
   - Remove a custom variable.

     On the Custom Variable page, find the custom variable that you want to remove and click **Delete** in the Action column.

3.Edit jobs

You can create jobs to develop tasks in a project. This topic describes job-related operations.

Background information

You can perform the following operations on jobs:

- Create a job
- Configure a job
- Add annotations
- Run a job
- Operations that you can perform on jobs
- Job submission modes

Prerequisites

A project is created. For more information, see Manage projects.

Create a job

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the Projects section of the page that appears, find the project that you want to manage and click **Edit Job** in the Actions column.
3. Create a job.
   i. In the Edit Job pane on the left side of the page that appears, right-click the folder on which you want to perform operations and select **Create Job**.

   > **Note** You can also right-click the folder and select **Create Subfolder**, **Rename Folder**, or **Delete Folder** to perform the corresponding operation.

   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and then select a specific job type from the **Job Type** drop-down list.

   E-MapReduce (EMR) supports the following types of jobs in data development: Shell, Hive, Hive SQL, Spark, Spark SQL, Spark Shell, Spark Streaming, MapReduce, Sqoop, Pig, Flink, Streaming SQL, Presto SQL, and Impala SQL.

   > **Note** After the job is created, you cannot change the **type of the job**.

   iii. Click **OK**.

   After a job is created, you can configure and edit the job.

Configure a job

For more information about how to develop and configure each type of job, see Jobs. This section describes how to configure the parameters of a job on the **Basic Settings**, **Advanced Settings**, **Shared Libraries**, and **Alert Settings** tabs in the Job Settings panel.

1. In the upper-right corner of the **job** page, click **Job Settings**.
2. In the **Job Settings** panel, configure the parameters on the Basic Settings tab.

Section and parameter		Description
Job Overview	**Name**	The name of the job.
	Job Type	The type of the job.
	Retries	The number of retries that are allowed if the job fails. The value of this parameter ranges from 0 to 5.
	Actions on Failures	The action that you can perform if the job fails. Valid values: ○ **Pause**: Suspend the current workflow if the job fails. ○ **Run Next Job**: Continue to run the next job if the job fails. You can determine whether to turn on the **Use Latest Job Content and Parameters** switch based on your business requirements. ○ If you turn off this switch, a job instance is generated based on the original job content and parameters after you rerun a job that fails. ○ If you turn on this switch, a job instance is generated based on the latest job content and parameters after you rerun a job that fails.
	Description	The description of the job. If you want to modify the description of the job, you can click **Edit** on the right side of this parameter.
Resources		The resources that are required to run the job, such as JAR packages and user-defined functions (UDFs). Click the + icon on the right side to add resources. Upload the resources to Object Storage Service (OSS) first. Then, you can add them to the job.

Section and parameter	Description
Configuration Parameters	The variables that you want to reference in the job script. You can reference a variable in your job script in the format of *${Variable name}*. Click the + icon on the right side to add a variable in the key-value pair format. You can determine whether to select Password to hide the value based on your business requirements. The key indicates the name of the variable. The value indicates the value of the variable. In addition, you can configure a time variable based on the start time of scheduling. For more information, see Configure job time and date.

3. Click the **Advanced Settings** tab and configure the parameters.

Section	Parameter and description
Mode	○ **Job Submission Node**: the mode to submit the job. For more information, see Job submission modes. Valid values: ■ **Worker Node**: The job is submitted to YARN by using a launcher, and YARN allocates resources to run the job. ■ **Header/Gateway Node**: The job runs as a process on the allocated node. ○ **Estimated Maximum Duration**: the estimated maximum running duration of the job. Valid values: 0 to 10800. Unit: seconds.

Environment Variables	The environment variables that are used to run the job. You can also export environment variables from the job script.
	○ Example 1: Configure a Shell job with the code `echo ${ENV_ABC}`. If you set the ENV_ABC variable to `12345`, a value of `12345` is returned after you run the `echo` command.
	○ Example 2: Configure a Shell job with the code `java -jar abc.jar`. Content of the *abc.jar* package:
	`public static void main(String[] args) {System.out.println(System.getEnv("ENV_ABC"));}`
	If you set the ENV_ABC variable to 12345, a value of `12345` is returned after you run the job. The effect of setting the ENV_ABC variable in the Environment Variables section is equivalent to running the following script:
	`export ENV_ABC=12345`
	`java -jar abc.jar`

Scheduling Parameters	The parameters used to schedule the job, including Queue, Memory (MB), vCores, Priority, and Run By. If you do not configure these parameters, the default settings of the Hadoop cluster are used. **Note** The Memory (MB) parameter specifies the memory quota for the launcher.

4. Click the **Shared Libraries** tab.

   In the **Dependent Libraries** section, specify **Libraries**.

   Job execution depends on some library files related to data sources. EMR publishes the libraries to the repository of the scheduling center as dependency libraries. You must specify dependency libraries when you create a job. To specify a dependency library, enter its reference string, such as `sharedlibs:streamingsql:datasources-bundle:2.0.0`.

5. Click the **Alert Settings** tab and configure the alert parameters.

Parameter	Description
Execution Failed	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if the job fails.
Action on Startup Timeout	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if the job startup times out.
Job execution timed out.	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if the job execution times out.

Add annotations

You can add annotations to job scripts to configure job parameters in data development. Add an annotation in the following format:

```
!!! @<Annotation name>: <Annotation content>
```

Note Do not indent the three exclamation points (`!!!`) that start an annotation. Add one annotation in a line.

The following table describes all annotations that are supported.

Annotation name	Description	Example
rem	Adds a comment.	`!!! @rem: This is a comment.`
env	Adds an environment variable.	`!!! @env: ENV_1=ABC`

Annotation name	Description	Example
var	Adds a custom variable.	`!!! @var: var1="value1 and \"one string end with 3 spaces\"   "` `!!! @var: var2=${yyyy-MM-dd}`
resource	Adds a resource file.	`!!! @resource: oss://bucket1/dir1/file.jar`
sharedlibs	Adds dependency libraries. This annotation is valid only in Streaming SQL jobs. Separate multiple dependency libraries with commas (,).	`!!! @sharedlibs: sharedlibs:streamingsql:datasources-bundle:1.7.0,...`
scheduler.queue	Specifies the queue to which the job is submitted.	`!!! @scheduler.queue: default`
scheduler.vmem	Specifies the memory required to run the job. Unit: MiB.	`!!! @scheduler.vmem: 1024`
scheduler.vcores	Specifies the number of vCores required to run the job.	`!!! @scheduler.vcores: 1`
scheduler.priority	Specifies the priority of the job. Valid values: 1 to 100.	`!!! @scheduler.priority: 1`
scheduler.user	Specifies the user who submits the job.	`!!! @scheduler.user: root`

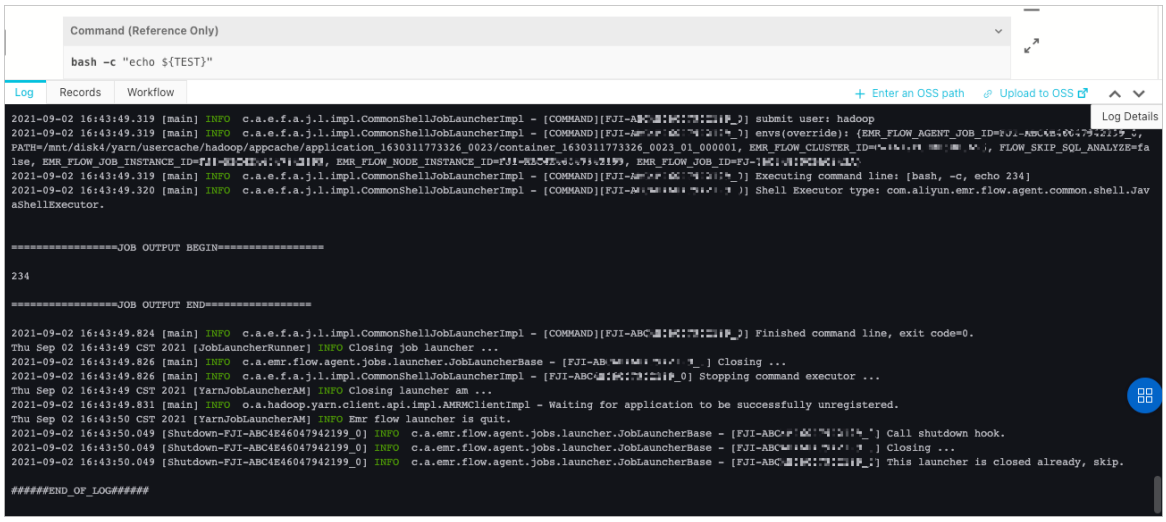
Notice

When you add annotations, take note of the following points:

- Invalid annotations are automatically skipped. For example, an unknown annotation or an annotation whose content is in an invalid format will be skipped.
- Job parameters specified in annotations take precedence over job parameters specified in the Job Settings panel. If a parameter is specified both in an annotation and in the Job Settings panel, the parameter setting specified in the annotation takes effect.

Run a job

1. Run the job that you created.
   i. On the job page, click **Run** in the upper-right corner to run the job.
   ii. In the **Run Job** dialog box, select a resource group and the cluster that you created.
   iii. Click **OK**.
2. View running details.
   i. Click the **Log** tab in the lower part of the job page to view the operational logs.



   ii. Click the **Records** tab to view the execution records of the job instance.
   iii. Click **Details** in the Action column of a job instance to go to the **Scheduling Center** tab. On this tab, you can view the details about the job instance.

Operations that you can perform on jobs

In the **Edit Job** pane, you can right-click a job and perform the operations that are described in the following table.

Operation	Description
Clone Job	Clones the configurations of a job to generate a new job in the same folder.
Rename Job	Renames a job.
Delete Job	Deletes a job. You can delete a job only if the job is not associated with a workflow or the associated workflow is not running or being scheduled.

Job submission modes

The spark-submit process, which is the launcher in a data development module, is used to submit Spark jobs. In most cases, this process occupies more than 600 MiB of memory. The Memory (MB) parameter in the Job Settings panel specifies the size of the memory allocated to the launcher.

The following table describes the modes in which jobs can be submitted in the latest version of EMR.

Job submission mode	Description
Header/Gateway Node	In this mode, the spark-submit process runs on the master node and is not monitored by YARN. The spark-submit process requests a large amount of memory. A large number of jobs consume many resources of the master node, which undermines cluster stability.
Worker Node	In this mode, the spark-submit process runs on a core node, occupies a YARN container, and is monitored by YARN. This mode reduces the resource usage on the master node.

In an EMR cluster, the memory consumed by a job instance is calculated by using the following formula:

```
Memory consumed by a job instance = Memory consumed by the launcher + Memory consumed by a job that corresponds to the job instance
```

For a Spark job, the memory consumed by a job is calculated by using the following formula:

```
Memory consumed by a job = Memory consumed by the spark-submit logical module (not the process) + Memory consumed by the driver + Memory consumed by the executor
```

The process in which the driver runs varies based on the mode in which Spark applications are launched in YARN.

Launch mode of Spark application			Process in which spark-submit and driver run
yarn-client mode	Submit a job in LOCAL mode.	The driver runs in the same process as spark-submit.	The process used to submit a job runs on the master node and is not monitored by YARN.
	Submit a job in YARN mode.		The process used to submit a job runs on a core node, occupies a YARN container, and is monitored by YARN.

Launch mode of Spark application	Process in which spark-submit and driver run
yarn-cluster mode	The driver runs in a different process from spark-submit.

4.Edit a workflow

In a data development project of E-MapReduce (EMR), you can define a group of dependent jobs, and create a workflow to allow the jobs to run in sequence based on their dependencies. An EMR workflow can be represented as a directed acyclic graph (DAG) that allows big data jobs to run in parallel. You can schedule workflows or view the status of workflows in the EMR console.

Background information

Workflow-related operations:

- Create a workflow
- Edit a workflow
- Configure workflow scheduling
- Run a workflow
- View the running details about a workflow
- Operations that you can perform on workflows

Prerequisites

- A project is created. For more information, see Manage projects.
- Jobs are edited. For more information, see Edit jobs.

Create a workflow

Perform the following steps to create a workflow:

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section of the page that appears, find your project and click **Workflows** in the Actions column.
3. Create a workflow.
   i. In the **Workflows** pane on the left side of the page that appears, right-click the folder on which you want to perform operations and select **Create Workflow**.
   ii. In the **Create Workflow** dialog box, specify **Workflow Name**, **Description**, Select Resource Group, and Target Cluster.

   Valid values of the Target Cluster parameter:

   - **Select Existing Cluster**: When the workflow is executed, the jobs run on the cluster that you selected.
   - **Create Cluster from Template**: When the workflow is executed, the jobs run on a temporary cluster that is created by using the cluster template you selected. When the workflow ends, the cluster is automatically released. For more information, see Create a cluster template.

   > **Note** Only the clusters that are associated with the project are displayed in the **Select Existing Cluster** drop-down list. Before you can select a different cluster, you must disassociate the existing clusters from the project. For more information, see Manage projects.

   iii. Click **OK**.
   After the workflow is created, you can edit and configure the workflow.

Edit a workflow

1. Drag different types of job nodes to the canvas for editing a workflow.

   After you drag a node of a specific type to the canvas, you can configure the parameters that are described in the following table in the **Edit Node** panel.

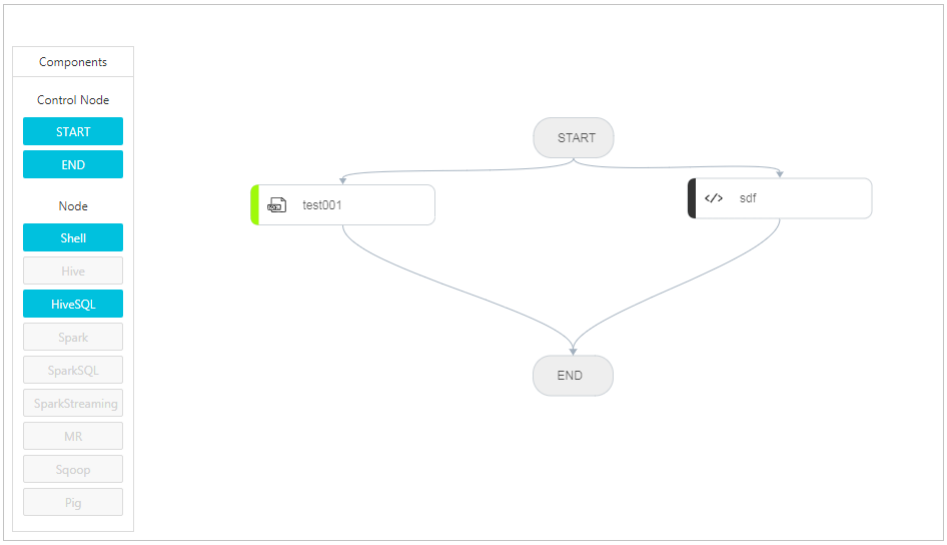
Parameter	Description
Associated Job	Select a job of the same type as the job node from the Associated Job drop-down list.
Customize Job Configuration	You can **customize job configurations** based on your business requirements. ◦ If you turn on this switch, you can change the value of the **Target Cluster** parameter. ◦ If you turn off this switch, the jobs that are associated with the job node run on the cluster that you select when you create a workflow. By default, the Customize Job Configuration switch is turned off.

2. Associate job nodes.

   On the canvas, drag a line from a job node to associate this job node with other job nodes based on the dependencies between the jobs. Arrows indicate the direction of the workflow.

3. Configure controller nodes to complete the design of the workflow.

   Drag the **END** node from the **Controller Node** section to the canvas. Then, associate the **START** node, job nodes, and **END** node to complete the design of the workflow. You can click **Auto Adjust** in the upper-right corner to adjust the layout of the job nodes in the workflow.



When you edit a workflow, you can click **Lock** in the upper-right corner to lock the workflow. This way, only you can edit or run the workflow. Other members in the project can edit the workflow only after the workflow is unlocked.

> **Note** Only the RAM user that locks the workflow and the Alibaba Cloud account can unlock the workflow.

Configure workflow scheduling

You can enable the workflow scheduling feature and configure scheduling-related parameters. Then, relevant workflows periodically run based on the parameter settings, and jobs are delivered to a specified cluster for running. Perform the following steps to configure the parameters on the Basic Attributes, Scheduling Settings, and Alert Settings tabs in the Workflow Scheduling panel:

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section of the page that appears, find your project and click **Workflows** in the Actions column.
3. On the workflow design page, click **Configure**.
4. On the **Basic Attributes** tab of the **Workflow Scheduling** panel, modify the workflow description, resource group, and the cluster used to run the jobs in the workflow based on your business requirements.
5. After the basic attributes are modified, click the **Scheduling Settings** tab and configure the parameters related to workflow scheduling.

Parameter		Description
Scheduling Status		Valid values: ○ **Start**: Start workflow scheduling. After you select Start for Scheduling Status, **Scheduling** appears in the upper-right corner of the workflow editing canvas, which indicates that the workflow is being scheduled. ○ **Stop**: Stop workflow scheduling.
Time-based Scheduling	**Start Time**	The time when workflow scheduling starts.
	End Time	The time when workflow scheduling ends. This parameter is optional.
	Recurrence	The cycle of workflow scheduling.
	CRON Expression	The CRON expression that is used to specify the cycle of workflow scheduling.
Dependency-based Scheduling	**Project**	The project to which the dependent workflow of the current workflow belongs. This parameter is optional.
	Dependent Workflow	The dependent workflow of the current workflow. The current workflow is executed only after the dependent workflow ends. This parameter is optional.

6. Click the **Alert Settings** tab and configure the alert parameters.

Parameter	Description
Execution Failed	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if the workflow fails.
Actions on Failures	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if a job node in the workflow fails to run.
Executed	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if the workflow succeeds.
Action on Startup Timeout	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if a job node in the workflow does not start within 30 minutes after it is delivered to a cluster.

Parameter	Description
Node execution timed out	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if the running duration of a job node exceeds the expected maximum running duration in the job configuration.

Run a workflow

You can specify the business time of a workflow. Time variables in jobs of the workflow are calculated by using the specified business time. The business time is used for rerunning the workflow instance in a specific period of time. You can rerun a single workflow instance or multiple workflow instances at a time. If no time variables are configured for your jobs, you can select Execute.

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section of the page that appears, find your project and click **Workflows** in the Actions column.
3. Run your workflow.
   i. On the page that appears, select a workflow and click **Run** in the upper-right corner.
   ii. In the **Run Workflow** dialog box, configure the runtime parameters.

   You can select a **running mode** based on your business requirements. The following table describes the running modes that are supported: **Execute** and **Run Periodically**.

Mode	Description
Execute	Immediately runs a workflow. You can use the **specified time** as the business time of the workflow. Time-related variables are calculated based on the business time.
Run Periodically	Runs multiple workflows at a time. The trigger time of specific scheduling rules is used as the business time of the workflows, and time-related variables are calculated based on the business time. A maximum of 100 points in time are supported at a time. If you select Run Periodically for Mode, configure the following parameters: ■ **Start Time**: the time when workflow scheduling starts. ■ **End Time**: the time when workflow scheduling ends. This parameter is optional. ■ **Recurrence**: the cycle of workflow scheduling. ■ **CRON Expression**: the CRON expression that is used to specify the cycle of workflow scheduling. ■ **Skip Successful Nodes**: specifies whether to skip a successful workflow instance. You can determine whether to turn on this switch based on your business requirements. After you turn on the Skip Successful Nodes switch, if the workflow instance that runs at a specific business time is successful, the system skips the workflow instance and continues to run the workflow instances that fail at a different business time.

   iii. Click **OK**.

View the running details about a workflow

After you run a workflow, you can perform the following steps to view the running details about the workflow:

1. Click the **Records** tab in the lower part of the **workflow design** page.

   You can view the status of a workflow instance.
2. Find your workflow instance and click **Details** in the Action column to go to the **Scheduling Center** tab.

   You can view the details about the workflow instance. You can also pause, resume, stop, or rerun the workflow instance. For more information, see Scheduling center.

Operation	Description
Details	Views the details and status of the workflow instance.
Stop Workflow	Stops all running job nodes of the workflow instance.
Pause Workflow	If you click this button, the running job nodes continue running, but the subsequent job nodes in the workflow will not start.
Resume Workflow	Resumes the workflow instance if it has been suspended.
Rerun Workflow Instance	Reruns the workflow instance if it has been terminated. After you click **Rerun Workflow Instance**, you can determine whether to rerun failed job nodes or rerun all job nodes from the START node.

Operations that you can perform on workflows

In the **Workflows** pane, you can right-click a workflow and perform the operations that are described in the following table.

Operation	Description
Clone Workflow	Clones a workflow with the same design in the same folder. **Note** The settings of the scheduling parameters for the original workflow cannot be cloned.
Rename Workflow	Renames a workflow.
Delete Workflow	Deletes a workflow. You cannot delete a running workflow.

5.Perform ad hoc queries

E-MapReduce (EMR) supports ad hoc queries, which are intended for data scientists and data analysts. You can execute SQL statements to perform ad hoc queries. When you run an ad hoc query job, relevant logs and query results appear in the lower part of the job page. This topic describes how to create, configure, run, and lock a job on the Ad Hoc Queries page in the EMR console.

Background information

You can perform the following operations on the Ad Hoc Queries page:

- Create a job
- Configure a job
- Run a job
- Lock a job

Prerequisites

A project is created. For more information, see Manage projects.

Create a job

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find the project that you created and click **Edit Job** in the Actions column.
3. Create a job for an ad hoc query.
   i. On the left side of the page, click the icon.
   ii. In the **Ad Hoc Queries** pane on the left, right-click the folder in which you want to create a job and select **Create Job**.

   > **Note** You can also right-click the folder and select **Create Subfolder**, **Rename Folder**, or **Delete Folder** to perform the required operation.

   iii. In the **Create Interactive Job** dialog box, specify **Name** and **Description**, and then select a job type from the **Job Type** drop-down list.
   EMR supports ad hoc queries based on Shell, Spark SQL, Spark Shell, and Hive SQL.

   > **Notice** After the job is created, you cannot change the type of the job.

   iv. Click **OK**.

Configure a job

For more information about how to develop and configure each type of job, see Jobs. This section describes how to configure the parameters of a job on the **Basic Settings**, **Advanced Settings**, **Shared Libraries**, and **Alert Settings** tabs in the Job Settings panel.

1. In the upper-right corner of the **job** page, click **Job Settings**.
2. In the **Job Settings** panel, configure the parameters on the Basic Settings tab.

Section	Parameter and description
Job Overview	◦ **Name**: the name of the job. ◦ **Job Type**: the type of the job. ◦ **Description**: the description of the job. You can click **Edit** on the right side of this parameter to modify the description.
Resources	The resources that are required to run the job, such as JAR packages and user-defined functions (UDFs). Click the icon on the right to add resources. Upload the resources to Object Storage Service (OSS) first. Then, you can add them to the job.
Configuration Parameters	The variables you want to reference in the job script. You can reference a variable in your job script in the format of *${Variable name}*. Click the icon on the right side to add a variable in the key-value pair format. You can select Password to hide the value based on your business requirements. The key indicates the name of the variable. The value indicates the value of the variable. In addition, you can configure a time variable based on the start time of scheduling. For more information, see Configure job time and date.

3. Click the **Advanced Settings** tab and configure the parameters.

Section	Parameter and description
Mode	◦ **Job Submission Node**: the mode to submit the job. For more information, see Job submission modes. Valid values: ▪ **Worker Node**: The job is submitted to YARN by using a launcher, and YARN allocates resources to run the job. ▪ **Header/Gateway Node**: The job runs as a process on the allocated node. ◦ **Estimated Maximum Duration**: the estimated maximum running duration of the job. Valid values: 0 to 10800. Unit: seconds.

Section	Parameter and description

Environment Variables	The environment variables that are used to run the job. You can also export environment variables from the job script.
	◦ Example 1: Configure a Shell job with the code `echo ${ENV_ABC}`. If you set the ENV_ABC variable to `12345`, a value of `12345` is returned after you run the `echo` command.
	◦ Example 2: Configure a Shell job with the code `java -jar abc.jar`. Content of the *abc.jar* package:
	`public static void main(String[] args) {System.out.println(System.getEnv("ENV_ABC"));}`
	If you set the ENV_ABC variable to 12345, a value of `12345` is returned after you run the job. The effect of setting the ENV_ABC variable in the Environment Variables section is equivalent to running the following script:
	`export ENV_ABC=12345`
	`java -jar abc.jar`

Scheduling Parameters	The parameters used to schedule the job, including Queue, Memory (MB), vCores, Priority, and Run By. If you do not configure these parameters, the default settings of the Hadoop cluster are used. **Note** The Memory (MB) parameter specifies the memory quota for the launcher.

4. Click the **Shared Libraries** tab.

   In the **Dependent Libraries** section, specify **Libraries**.

   Job execution depends on some library files related to data sources. EMR publishes the libraries to the repository of the scheduling center as dependency libraries. You must specify dependency libraries when you create a job. To specify a dependency library, enter its reference string, such as `sharedlibs:streamingsql:datasources-bundle:2.0.0`.

5. Click the **Alert Settings** tab and configure the alert parameters.

Parameter	Description
Execution Failed	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if the job fails.
Action on Startup Timeout	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if the job startup times out.
Job execution timed out.	Specifies whether to send a notification to an alert contact group or a DingTalk alert group if the job execution times out.

Run a job

1. Run the job that you created.
   i. On the **job** page, click **Run** in the upper-right corner to run the job.
   ii. In the **Run Job** dialog box, select a resource group and the cluster that you created.
   iii. Click **OK**.
2. View operational logs.
   i. After you run the job, you can view the operational logs on the **Log** tab in the lower part of the job page.
   ii. Click the **Records** tab to view the execution records of the job instance.
   iii. Click **Details** in the Action column of a job instance to go to the **Scheduling Center** tab. On this tab, you can view the details about the job instance.

Lock a job

When you edit a job, you can click **Lock** in the upper-right corner of the job page to lock the job. This way, only the account you use can edit the job. Other members in the project can edit this job only after the job is unlocked.

Note Only the RAM user that locks the job and the Alibaba Cloud account can unlock the job.

6.Scheduling center

This topic describes how to manage workflow scheduling tasks, monitor the status of tasks, and view both workflow records and audit logs in the scheduling center. These features facilitate workflow management and maintenance.

Background information

You can perform the following operations in the scheduling center:

- View project overview
- View and manage workflow records
- View audit logs

Prerequisites

A project is created. For more information, see Manage projects.

View project overview

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find the project that contains the overview information you want to view and click **Records** in the Actions column.
3. In the left-side navigation pane, click **Overview**.
   You can view the overview information of a project.

View and manage workflow records

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find the project that contains the overview information you want to view and click **Records** in the Actions column.
3. Manage workflow records.
   Perform the following operations to view and manage different types of records:
   - Workflow records
     In the left-side navigation pane, choose Workflow Records > **Workflow Records** to view the information of workflow instances. You can also click the following action items in the Actions column to manage workflow instances.

Action item	Description
Details	View the details and status of a workflow instance.
Stop	Stop a workflow instance that is running.
Pause	Pause a workflow instance that is running.
Resume	Resume a paused workflow instance.

   - Records of the jobs that are manually executed
     a. In the left-side navigation pane, choose **Workflow Records > Job Records**.
     b. On the **Job Records** page, view the running details about job instances. You can also click the following action items in the Actions column to manage job instances.

Action item	Description
Details	View the runtime parameters, content, and logs of a job instance.
Stop	Stop a job instance that is running.

   - Streaming jobs
     a. In the left-side navigation pane, choose **Workflow Records > Streaming Jobs**.
     b. On the **Streaming Jobs** page, view the details about streaming job instances. You can also click the following action items in the Actions column to manage streaming job instances.

Action item	Description
Details	View the runtime parameters, content, and logs of a streaming job.
Edit	Access the **Edit Job** pane to modify the content of a streaming job.
Start	Start a streaming job.
Stop	Stop a streaming job that is running.
History	View the history of a streaming job.

View audit logs

Perform the following steps to view the operations that a project member performed in a project:

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find the project that contains the overview information you want to view and click **Records** in the Actions column.
3. In the left-side navigation pane, click **Audit Log**.

   On the **Audit Log** page, you can view the operation history of project members.

7.Create a cluster template

Cluster templates contain saved configurations that you can use to create clusters. This topic describes how to create a cluster template.

Background information

Cluster templates are used for the system to create temporary clusters for data development workflows. If you are concerned only about the completion of the jobs in a data development workflow, you can specify a cluster template. The system creates a cluster based on the template that you specified, and then delivers the jobs to the cluster. The cluster is automatically released after the workflow is complete.

Limits

You can use cluster templates to create only Hadoop and Dataflow clusters. If you want to create a cluster of another type, submit a ticket. For more information, see submit a ticket.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, click **Cluster Template** in the upper-right corner.

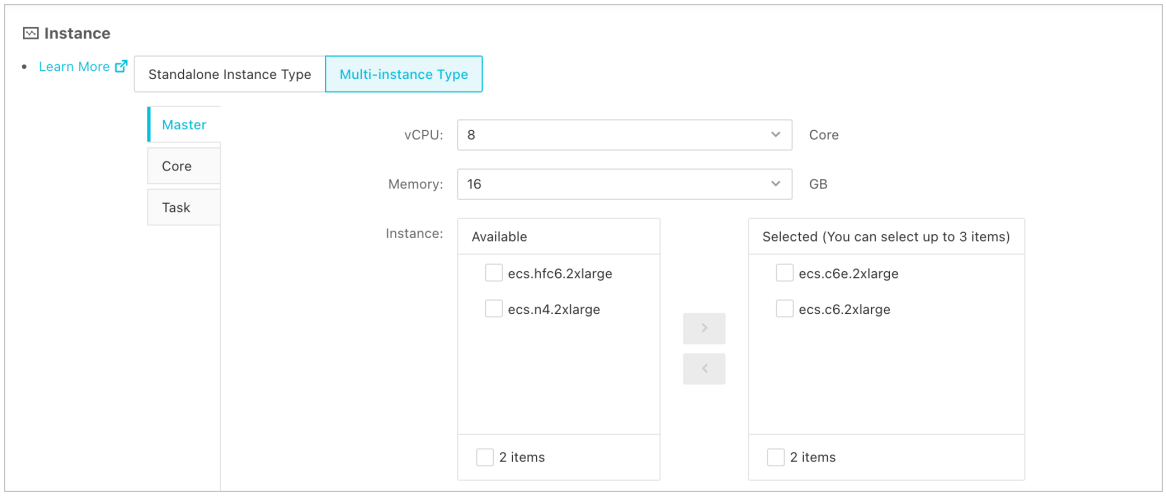
   On the cluster template list page that appears, you can modify or delete an existing cluster template based on your business requirements.

Action	Description
Edit	Click **Edit** in the Actions column that corresponds to a cluster template to modify the cluster template. After the modification is saved, the modification immediately takes effect on all workflows that are using the cluster template.
Delete	Click **Delete** in the Actions column of a cluster template to delete the cluster template. **Note** The system does not check whether the cluster template that you want to delete is being referenced by a workflow. After the cluster template is deleted, all the workflows that use the cluster template fail.

3. Create a cluster template.
   i. Click **Create Cluster Template** in the upper-right corner.
   ii. On the **Create Cluster Template** page, configure the required parameters.

   The operations that are required to create a cluster template are similar to the operations that are required to create a cluster. For more information, see Create a cluster.

   In the **Hardware Settings** step of the Create Cluster Template page, you can configure multiple instance types. This prevents cluster creation failures caused by insufficient inventory of instances of a specified type, which may affect the jobs that you want to run.



   iii. Read and select **E-MapReduce Service Terms** and click **Save**.
   You can view the cluster template that you created in the cluster template list.

8.Event codes for Cloud Monitor

In the event monitoring module of Cloud Monitor, you can subscribe to system events related to data development of E-MapReduce (EMR) to monitor the status of core components in an EMR cluster.

The following table lists event codes and the description of each event code.

Event code	Description	Event type
EMR-110401002	The workflow is completed.	FLOW
EMR-110401003	The workflow is submitted.	FLOW
EMR-110401004	The job is submitted.	FLOW
EMR-110401005	The workflow node is started.	FLOW
EMR-110401006	The status of the workflow node is checked.	FLOW
EMR-110401007	The workflow node is completed.	FLOW
EMR-110401008	The workflow node is stopped.	FLOW
EMR-110401009	The workflow node is canceled.	FLOW
EMR-110401010	The workflow is canceled.	FLOW
EMR-110401011	The workflow is restarted.	FLOW
EMR-110401012	The workflow is resumed.	FLOW
EMR-110401013	The workflow is paused.	FLOW
EMR-110401014	The workflow is stopped.	FLOW
EMR-110401015	The workflow node failed.	FLOW
EMR-110401016	The job failed.	FLOW
EMR-210401001	The workflow failed.	FLOW
EMR-210401003	The start time for the workflow node has been exceeded.	FLOW
EMR-210401004	The start time for the job has been exceeded.	FLOW

9.Job configuration

9.1. Configure job time and date

When you edit a job, you can set a time variable wildcard.

Variable wildcard format

E-MapReduce (EMR) supports the following formats of variable wildcards: *${dateexpr-1d}* and *${dateexpr-1h}*. *dateexpr* specifies the standard format of time. The following table describes the date and time formats.

> **Notice** The expression is case-sensitive.

Specifier	Description
yyyy	Indicates a 4-digit year.
MM	Indicates a 2-digit month.
dd	Indicates a 2-digit day.
HH	Indicates a 2-digit hour (24-hour clock). hh indicates a 2-digit hour (12-hour clock).
mm	Indicates a 2-digit minute.
ss	Indicates a 2-digit second.

A time variable is a combination of *yyyy* and one or more other time formats. You can also use the plus sign (+) or minus sign (-) to add or subtract a specified period of time to or from the current time. For example, *${yyyy-MM-dd}* indicates the current date.

- One year after the current date can be represented as *${yyyy+1y}* or *${yyyy-MM-dd hh:mm:ss+1y}*.
- Three months after the current date can be represented as *${yyyyMM+3m}* or *${yyyy-MM-dd hh:mm:ss+3m}*.
- Five days before the current date can be represented as *${yyyyMMdd-5d}* or *${yyyy-MM-dd hh:mm:ss-5d}*.

For example, the current time is 20160427 12:08:01.

- If *${yyyyMMdd HH:mm:ss-1d}* is configured as the variable wildcard, the time will be replaced with 20160426 12:08:01 when a job is run. One day is subtracted from the current date and the new time is accurate to seconds.
- If *${yyyyMMdd-1d}* is configured as the variable wildcard, the time will be replaced with 20160426, which indicates the day before the current date.
- If *${yyyyMMdd}* is configured as the variable wildcard, the time will be replaced with 20160427, which indicates the current date.

> **Note**
> - Only days or hours can be added or subtracted. That is, dateexpr can be followed only by +Nd, -Nd, +Nh, or -Nh. N must be an integer.
> - A time variable must start with yyyy, for example, *${yyyy-MM}*. If you want to obtain the values based on a specific period such as a month, you can use the following functions in a job:
>   - parseDate(<Parameter name>, <Time format>): You can use this function to convert a specified parameter to a date object. A parameter name indicates the variable (key) name specified in the Configuration Parameters section. A time format is the time format used by the variable name. For example, if the parameter name of the current_time variable is *${yyyyMMddHHmmss-1d}*, the time format is yyyyMMddHHmmss.
>   - formatDate(<Date object>, <Time format>): You can use this function to convert a specified date object to a time format string.
>
> Examples:
>
>   - *${formatDate(parseDate(current_time, 'yyyyMMddHHmmss'), 'HH')}* retrieves the literal hour value from the current_time variable.
>   - *${formatDate(parseDate(current_time, 'yyyyMMddHHmmss'), 'yyyy')}* retrieves the literal year value from the current_time variable.

Example

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find your project and click **Edit Job** in the Actions column.
3. Configure the job.
   i. In the **Edit Job** pane on the left, click a specific job name and click **Job Settings** in the upper-right corner.
   ii. In the **Configuration Parameters** section of the Basic Settings tab in the Job Settings panel, click the + icon to configure a variable wildcard in one of the preceding formats, as shown in the following figure.

After you complete the configurations, you can reference the key of the configured parameter in the job.

9.2. Configure a Shell job

This topic describes how to configure a Shell job.

Prerequisites

A project is created. For more information, see Manage projects.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find your project and click **Edit Job** in the Actions column.
3. Create a Shell job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **Shell** from the **Job Type** drop-down list.
   This option indicates that a Bash Shell job will be created.
   iii. Click **OK**.
4. Edit job content.
   i. Specify the command line parameters required to submit the job in the **Content** field.
   Example:

```
DD=`date`;
echo "hello world, $DD"
```

   ii. Click **Save**.

9.3. Configure a Hive job

E-MapReduce (EMR) provides a Hive environment. You can use Hive to create tables and perform operations on the tables and the data in them.

Prerequisites

- A project is created. For more information, see Manage projects.
- A Hive SQL script, for example, uservisits_aggre_hdfs.hive, is uploaded to a path in OSS, such as *oss://path/to/*.

  Content of *uservisits_aggre_hdfs.hive*:

```
USE DEFAULT;
 DROP TABLE uservisits;
 CREATE EXTERNAL TABLE IF NOT EXISTS uservisits (sourceIP STRING,destURL STRING,visitDate STRING,adRevenue DOUBLE,userAgent STRING,countryCode STRI
NG,languageCode STRING,searchWord STRING,duration INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS SEQUENCEFILE LOCATION '/HiBench/Aggr
egation/Input/uservisits';
 DROP TABLE uservisits_aggre;
 CREATE EXTERNAL TABLE IF NOT EXISTS uservisits_aggre (sourceIP STRING, sumAdRevenue DOUBLE) STORED AS SEQUENCEFILE LOCATION '/HiBench/Aggregation/
Output/uservisits_aggre';
 INSERT OVERWRITE TABLE uservisits_aggre SELECT sourceIP, SUM(adRevenue) FROM uservisits GROUP BY sourceIP;
```

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find your project and click **Edit Job** in the Actions column.
3. Create a Hive job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **Hive** from the **Job Type** drop-down list.
   This option indicates that a Hive job will be created. You can use the following command syntax to submit a Hive job:

```
hive [user provided parameters]
```

   iii. Click **OK**.
4. Edit job content.
   i. Specify the command line parameters required to submit the job in the **Content** field.
   For example, to use a Hive script uploaded to OSS, enter the following command:

```
-f ossref://path/to/uservisits_aggre_hdfs.hive
```

   Note `path` indicates the path in which `uservisits_aggre_hdfs.hive` is stored in OSS.

   Click **+ Enter an OSS path** in the lower part of the page. In the OSS File dialog box, specify File Path. The system automatically completes the path of the Hive script in OSS. File Prefix must be set to **OSSREF** to ensure that EMR can download the file.
   ii. Click **Save**.

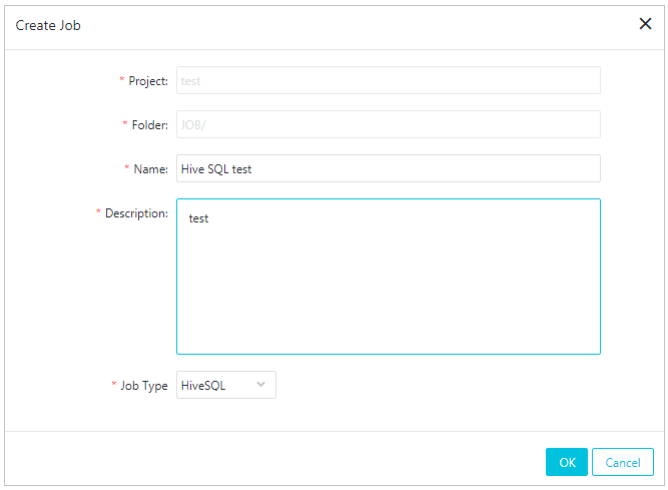
9.4. Configure a Hive SQL job

This topic describes how to configure a Hive SQL job.

Prerequisites

A project is created. For more information, see Manage projects.

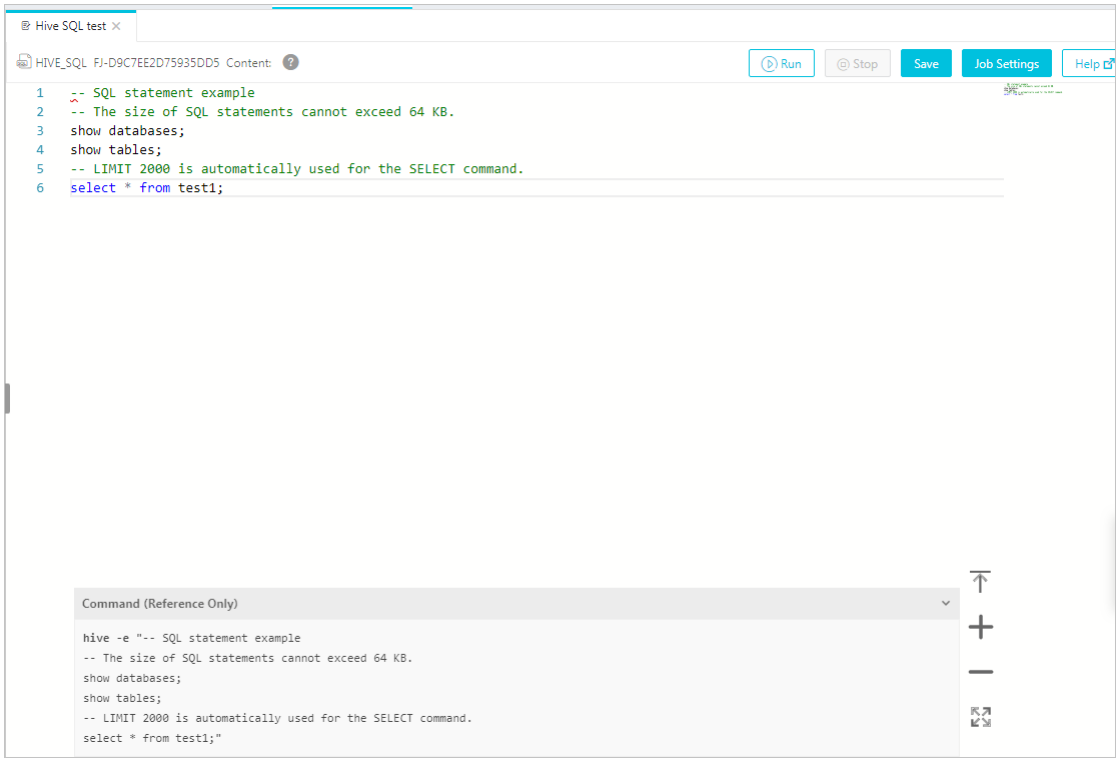
Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find your project and click **Edit Job** in the Actions column.
3. Create a Hive SQL job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **HiveSQL** from the **Job Type** drop-down list.
   This option indicates that a Hive SQL job will be created. You can use the following command syntax to submit a Hive SQL job:
   ```
   hive -e {SQL CONTENT}
   ```
   `SQL CONTENT` refers to the SQL statements that you enter in the job editor.
   

   iii. Click **OK**.
4. Edit job content.

i. Enter Hive SQL statements in the **Content** field.

```
-- SQL statement example
-- The size of SQL statements cannot exceed 64 KB.
show databases;
show tables;
-- LIMIT 2000 is automatically used for the SELECT statement.
select * from test1;
```



ii. Click **Save**.

9.5. Configure a Spark job

This topic describes how to configure a Spark job.

Prerequisites

A project is created. For more information, see Manage projects.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find your project and click **Edit Job** in the Actions column.
3. Create a Spark job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **Spark** from the **Job Type** drop-down list.
   This option indicates that a Spark job will be created. You can use the following command syntax to submit a Spark job:
   ```
   spark-submit [options] --class [MainClass] xxx.jar args
   ```
   iii. Click **OK**.
4. Edit job content.

i. Specify the command line parameters required to submit the job in the **Content** field.

Only the parameters that follow spark-submit are required.

The following examples demonstrate how to specify the parameters required to submit Spark and PySpark jobs.

- Create a Spark job.

  Create a Spark job named Wordcount. Parameter configuration example:

  - Enter the following command in the command line:

    ```
    spark-submit --master yarn-client --driver-memory 7G --executor-memory 5G --executor-cores 1 --num-executors 32 --class com.aliyun.emr.checklist.benchmark.SparkWordCount emr-checklist_2.10-0.1.0.jar oss://emr/checklist/data/wc oss://emr/checklist/data/wc-counts 32
    ```

  - Enter the following command in the **Content** field:

    ```
    --master yarn-client --driver-memory 7G --executor-memory 5G --executor-cores 1 --num-executors 32 --class com.aliyun.emr.checklist.benchmark.SparkWordCount ossref://emr/checklist/jars/emr-checklist_2.10-0.1.0.jar oss://emr/checklist/data/wc oss://emr/checklist/data/wc-counts 32
    ```

    Notice If a job is stored in OSS as a JAR package, you can reference the JAR package by using the *ossref://emr/checklist/jars/emr-checklist_2.10-0.1.0.jar* path. Click **+ Enter an OSS path** in the lower part of the page. In the OSS File dialog box, set **File Prefix** to **OSSREF** and specify **File Path**. The system automatically completes the path of the Spark script in OSS.

- Create a PySpark job.

  In addition to Scala and Java Spark jobs, you can create Python Spark jobs in EMR. Create a PySpark job named Python-Kmeans. Parameter configuration example:

  ```
  --master yarn-client --driver-memory 7g --num-executors 10 --executor-memory 5g --executor-cores 1 ossref://emr/checklist/python/kmeans.py oss://emr/checklist/data/kddb 5 32
  ```

  Notice

  - Python script resources can be referenced by using the ossref protocol.
  - The Python toolkit cannot be installed by using a PySpark job.

ii. Click **Save**.

9.6. Configure a Spark SQL job

This topic describes how to configure a Spark SQL job.

Prerequisites

A project is created. For more information, see Manage projects.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the Projects section of the page that appears, find the project that you want to manage and click **Edit Job** in the Actions column.
3. Create a Spark SQL job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and then select **SparkSQL** from the **Job Type** drop-down list.

   Note By default, a Spark SQL job is submitted in yarn-client mode.

   You can use the following command syntax to submit a Spark SQL job:

   ```
   spark-sql [options] [cli options] {SQL_CONTENT}
   ```

   The following table describes the parameters in the command syntax.

Parameter	Description
options	The setting of the SPARK_CLI_PARAMS parameter that you configure by performing the following operations: Click **Job Settings** in the upper-right corner of the job page. In the Job Settings panel, click the **Advanced Settings** tab. Click the + icon in the **Environment Variables** section and add the setting of the SPARK_CLI_PARAMS parameter, such as `SPARK_CLI_PARAMS="--executor-memory 1g --executor-cores"`.
cli options	Examples: ■ `-e <quoted-query-string>`: indicates that the SQL statements enclosed in quotation marks are executed. ■ `-f <filename>`: indicates that the SQL statements in the file are executed.
SQL_CONTENT	The SQL statements that you enter.

   iii. Click **OK**.
4. Edit job content.

i. Enter the Spark SQL statements in the **Content** field.

Example:

```
-- SQL statement example
-- The size of SQL statements cannot exceed 64 KB.
show databases;
show tables;
-- LIMIT 2000 is automatically added to the SELECT statement.
select * from test1;
```

ii. Click **Save**.

9.7. Configure a Spark Shell job

This topic describes how to configure a Spark Shell job.

Prerequisites

A project is created. For more information, see Manage projects.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the Projects section of the page that appears, find the project that you want to manage and click **Edit Job** in the Actions column.
3. Create a Spark Shell job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and then select **Spark Shell** from the **Job Type** drop-down list.
   iii. Click **OK**.
4. Edit job content.
   i. Configure the command line parameters that follow the Spark Shell command in the **Content** field.

   Example:

   ```
   val count = sc.parallelize(1 to 100).filter { _ =>
     val x = math.random
     val y = math.random
     x*x + y*y < 1
   }.count();
   println("Pi is roughly ${4.0 * count / 100}")
   ```

   ii. Click **Save**.

9.8. Configure a Spark Streaming job

This topic describes how to configure a Spark Streaming job.

Prerequisites

- A project is created. For more information, see Manage projects.
- All required resources and data to be processed are obtained.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the Projects section of the page that appears, find the project that you want to manage and click **Edit Job** in the Actions column.
3. Create a Spark Streaming job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and then select **Spark Streaming** from the **Job Type** drop-down list.
   iii. Click **OK**.
4. Edit job content.

i. Configure the command line parameters required to submit the job in the **Content** field.

You can use the following command syntax to submit a Spark Streaming job:

```
spark-submit [options] --class [MainClass] xxx.jar args
```

In the following example, a job with **Name** set to SlsStreaming is used to demonstrate the **Content** value:

```
--master yarn-client --driver-memory 7G --executor-memory 5G --executor-cores 1 --num-executors 32 --class com.aliyun.emr.checklist.benchmark
.SlsStreaming emr-checklist_2.10-0.1.0.jar <project> <logstore> <accessKey> <secretKey>
```

> **Notice**
> - If a job is stored in Object Storage Service (OSS) as a JAR package, you can reference the JAR package by using the *ossref:///xxx/.../xxx.jar* directory.
> - Click **+ Enter an OSS path** in the lower part of the page. In the OSS File dialog box, set **File Prefix** to **OSSREF** and specify **File Path**. The system completes the path of the Spark Streaming script in OSS.

ii. Click **Save**.

9.9. Configure a Hadoop MapReduce job

This topic describes how to configure a Hadoop MapReduce job.

Prerequisites

A project is created. For more information, see Manage projects.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find your project and click **Edit Job** in the Actions column.
3. Create a Hadoop MapReduce job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **MR** from the **Job Type** drop-down list.
   This option indicates that a Hadoop MapReduce job will be created. You can use the following command syntax to submit a Hadoop MapReduce job:
   ```
   hadoop jar xxx.jar [MainClass] -D xxx ....
   ```
   iii. Click **OK**.
4. Edit job content.
   i. Specify the command line parameters required to submit the job in the **Content** field.
   Start from the parameter that follows `hadoop jar`. Enter the path of the JAR package that is used to run the job. Then, specify `[MainClass]` and other command line parameters.
   For example, you want to submit a Hadoop sleep job. Instead of reading and writing data, this job submits only some mapper and reducer tasks to the Hadoop cluster, and sleeps for a period of time during the execution of each task. In Hadoop 2.6.0, this job is packaged in *hadoop-mapreduce-client-jobclient-2.6.0-tests.jar*. You can run the following command to submit the job:
   ```
   hadoop jar /path/to/hadoop-mapreduce-client-jobclient-2.6.0-tests.jar sleep -m 3 -r 3 -mt 100 -rt 100
   ```
   To configure this job in EMR, enter the following command in the **Content** field:
   ```
   /path/to/hadoop-mapreduce-client-jobclient-2.6.0-tests.jar sleep -m 3 -r 3 -mt 100 -rt 100
   ```
   > **Note** Click **+ Enter an OSS path** in the lower part of the page. In the OSS File dialog box, set **File Prefix** to **OSSREF** and specify **File Path**. The system automatically completes the path of the Hadoop MapReduce script in OSS.
   ii. Click **Save**.
   In the preceding example, the sleep job does not involve data input or output. To configure a job that reads data and provides processing results, such as a wordcount job, you must specify the data input and output paths.
   You can read data from and write data to HDFS or OSS in EMR. To read data from and write data to OSS, set the input and output paths to the paths in OSS. Sample code:
   ```
   jar ossref://emr/checklist/jars/chengtao/hadoop/hadoop-mapreduce-examples-2.6.0.jar randomtextwriter -D mapreduce.randomtextwriter.totalbytes
   =320000 oss://emr/checklist/data/chengtao/hadoop/Wordcount/Input
   ```

9.10. Configure a Sqoop job

This topic describes how to configure a Sqoop job.

Prerequisites

A project is created. For more information, see Manage projects.

Limits

E-MapReduce (EMR) V1.3.0 and later versions support Sqoop jobs. If you run a Sqoop job in an earlier version of EMR, an error is reported. For more information, see Sqoop.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find your project and click **Edit Job** in the Actions column.
3. Create a Sqoop job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **Sqoop** from the **Job Type** drop-down list.
   iii. Click **OK**.
4. Edit job content.
   i. Specify the command line parameters that follow the Sqoop command in the **Content** field.

      Example:

      ```
      sqoop [args]
      ```

   ii. Click **Save**.

9.11. Configure a Pig job

This topic describes how to configure a Pig job.

Prerequisites

- A project is created. For more information, see Manage projects.
- A Pig script is prepared. Sample code:

```
/*
 * Licensed to the Apache Software Foundation (ASF) under one
 * or more contributor license agreements.  See the NOTICE file
 * distributed with this work for additional information
 * regarding copyright ownership.  The ASF licenses this file
 * to you under the Apache License, Version 2.0 (the
 * "License"); you may not use this file except in compliance
 * with the License.  You may obtain a copy of the License at
 *
 *     http://www.apache.org/licenses/LICENSE-2.0
 *
 * Unless required by applicable law or agreed to in writing, software
 * distributed under the License is distributed on an "AS IS" BASIS,
 * WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
 * See the License for the specific language governing permissions and
 * limitations under the License.
 */
 -- Query Phrase Popularity (Hadoop cluster)
 -- This script processes a search query log file from the Excite search engine and finds search phrases that occur with particular high frequency
during certain times of the day.
 -- Register the tutorial JAR file so that the included UDFs can be called in the script.
 REGISTER oss://emr/checklist/jars/chengtao/pig/tutorial.jar;
 -- Use the PigStorage function to load the excite log file into the "raw" bag as an array of records.
 -- Input: (user,time,query)
 raw = LOAD 'oss://emr/checklist/data/chengtao/pig/excite.log.bz2' USING PigStorage('\t') AS (user, time, query);
 -- Call the NonURLDetector UDF to remove records if the query field is empty or a URL.
 clean1 = FILTER raw BY org.apache.pig.tutorial.NonURLDetector(query);
 -- Call the ToLower UDF to change the query field to lowercase.
 clean2 = FOREACH clean1 GENERATE user, time, org.apache.pig.tutorial.ToLower(query) as query;
 -- Because the log file only contains queries for a single day, we are only interested in the hour.
 -- The excite query log timestamp format is YYMMDDHHMMSS.
 -- Call the ExtractHour UDF to extract the hour (HH) from the time field.
 houred = FOREACH clean2 GENERATE user, org.apache.pig.tutorial.ExtractHour(time) as hour, query;
 -- Call the NGramGenerator UDF to compose the n-grams of the query.
 ngramed1 = FOREACH houred GENERATE user, hour, flatten(org.apache.pig.tutorial.NGramGenerator(query)) as ngram;
 -- Use the  DISTINCT command to get the unique n-grams for all records.
 ngramed2 = DISTINCT ngramed1;
 -- Use the  GROUP command to group records by n-gram and hour.
 hour_frequency1 = GROUP ngramed2 BY (ngram, hour);
 -- Use the  COUNT function to get the count (occurrences) of each n-gram.
 hour_frequency2 = FOREACH hour_frequency1 GENERATE flatten($0), COUNT($1) as count;
 -- Use the  GROUP command to group records by n-gram only.
 -- Each group now corresponds to a distinct n-gram and has the count for each hour.
 uniq_frequency1 = GROUP hour_frequency2 BY group::ngram;
 -- For each group, identify the hour in which this n-gram is used with a particularly high frequency.
 -- Call the ScoreGenerator UDF to calculate a "popularity" score for the n-gram.
 uniq_frequency2 = FOREACH uniq_frequency1 GENERATE flatten($0), flatten(org.apache.pig.tutorial.ScoreGenerator($1));
 -- Use the  FOREACH-GENERATE command to assign names to the fields.
 uniq_frequency3 = FOREACH uniq_frequency2 GENERATE $1 as hour, $0 as ngram, $2 as score, $3 as count, $4 as mean;
 -- Use the  FILTER command to move all records with a score less than or equal to 2.0.
 filtered_uniq_frequency = FILTER uniq_frequency3 BY score > 2.0;
 -- Use the  ORDER command to sort the remaining records by hour and score.
 ordered_uniq_frequency = ORDER filtered_uniq_frequency BY hour, score;
 -- Use the  PigStorage function to store the results.
 -- Output: (hour, n-gram, score, count, average_counts_among_all_hours)
 STORE ordered_uniq_frequency INTO 'oss://emr/checklist/data/chengtao/pig/script1-hadoop-results' USING PigStorage();
```

- The *script1-hadoop-oss.pig* file is saved and uploaded to a directory in OSS, such as *oss://path/to/script1-hadoop-oss.pig*.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find your project and click **Edit Job** in the Actions column.
3. Create a Pig job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **Pig** from the **Job Type** drop-down list.
   This option indicates that a Pig job will be created. You can use the following command syntax to submit a Pig job:
   ```
   pig [user provided parameters]
   ```
   iii. Click **OK**.
4. Edit job content.

i. Specify the command line parameters required to submit the job in the **Content** field.

For example, to use the Pig script uploaded to OSS, enter the following command:

```
-x mapreduce ossref://emr/checklist/jars/chengtao/pig/script1-hadoop-oss.pig
```

> **Note** Click **+ Enter an OSS path** in the lower part of the page. In the OSS File dialog box, set **File Prefix** to **OSSREF** and specify **File Path**. The system automatically completes the path of the Pig script in OSS.

ii. Click **Save**.

9.12. Configure a VVR-based Flink job

E-MapReduce (EMR) V3.27.X and earlier versions use the open source version of Flink. Versions later than EMR V3.27.X use Ververica Runtime (VVR), an enterprise-grade computing engine. VVR is fully compatible with Flink. This topic describes how to configure a VVR-based Flink job.

Background information

Enterprise-edition Flink is officially released by the founding team of Apache Flink and maintains a globally uniform brand.

VVR provides an enterprise-edition state backend whose performance is three to five times better than the performance of open source Flink. You can use the VVR engine and EMR data development feature to submit jobs in an EMR Hadoop cluster. VVR supports open source Flink 1.10 and provides business GeminiStateBackend by default, which brings the following benefits:

- Uses a new data structure to increase the random query speed and reduce frequent disk I/O operations.
- Optimizes the cache policy. If memory is sufficient, hot data is not stored in disks and cache entries do not expire after compaction.
- Uses Java to implement GeminiStateBackend, which eliminates Java Native Interface (JNI) overheads that are caused by RocksDB.
- Uses off-heap memory and implements an efficient memory allocator based on GeminiDB to eliminate the impact of garbage collection for Java Virtual Machines (JVMs).
- Supports asynchronous incremental checkpoints. This ensures that only memory indexes are copied during data synchronization. Unlike RocksDB, GeminiStateBackend avoids jitters that are caused by I/O operations.
- Supports local recovery and storage of the timer.

> **Note** If you want to use GeminiStateBackend, do not specify the type of a state backend in the code. To use GeminiStateBackend to start the Flink component, TaskManager must have 1,728 MiB of memory or more.

The basic configurations of the checkpoint and state backend in Flink also apply to GeminiStateBackend. For more information, see Configuration.

You can configure parameters based on your requirements. The following table describes some special parameters.

Parameter	Description
state.backend.gemini.memory.managed	Specifies whether to calculate the memory size of each backend based on the values of the Managed Memory and Task Slot parameters. Default value: true. Valid values: • true • false
state.backend.gemini.offheap.size	Specifies the memory of each backend when the state.backend.gemini.memory.managed parameter is set to false. Default value: 2. Unit: GiB.
state.backend.gemini.local.dir	Specifies the directory that stores local data files of GeminiDB.
state.backend.gemini.timer-service.factory	Specifies the storage location of the timer-service state. Default value: HEAP. Valid values: • HEAP • GEMINI

> **Note** For more information about parameter settings, see Manage parameters for services.

Prerequisites

- An EMR Hadoop cluster is created. For more information, see Create a cluster.
- A project is created. For more information, see Manage projects.
- Resources that are required for jobs and data files to be processed are obtained, such as JAR packages, data file names, and storage paths of the packages and files.

> **Note**
> - We recommend that you use Object Storage Service (OSS) to maintain the JAR packages that you want to run.
> - If you use a local path of a file, use the absolute path.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the Projects section of the page that appears, find the project that you want to manage and click **Edit Job** in the Actions column.
3. Create a Flink job.

i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **Flink** from the **Job Type** drop-down list.
iii. Click **OK**.

4. Edit job content.
   i. Configure the command line parameters required to submit the job in the **Content** field.
      - You can configure a Flink Datastream, table, or SQL job that is specified as a JAR package. Example:

        ```
        run -m yarn-cluster -yjm 1024 -ytm 2048 ossref://path/to/oss/of/WordCount.jar --input oss://path/to/oss/to/data --output oss://path/to/oss/to/result
        ```

      - In EMR V3.28.2 and later minor versions, you can configure a PyFlink job. Example:

        ```
        run -m yarn-cluster -yjm 1024 -ytm 2048 -py ossref://path/to/oss/of/word_count.py
        ```

        For more information about the parameters related to the PyFlink job, see Apache Flink official documentation.

   ii. Click **Save**.

> **Note** You can access the web UI of Flink based on the version of your cluster:
> - Versions earlier than EMR V3.29.0:
>
>   Use an SSH tunnel. For more information, see Create an SSH tunnel to access web UIs of open source components.
> - EMR V3.29.0 and later:
>   - Recommended. Use the EMR console. For more information, see Access the web UIs of open source components.
>   - Use an SSH tunnel. For more information, see Create an SSH tunnel to access web UIs of open source components.

9.13. Configure a Streaming SQL job

This topic describes how to configure a Streaming SQL job.

Background information

For more information about Streaming SQL, see Spark Streaming SQL.

When you configure a Streaming SQL job, you must specify dependency libraries. The following table describes the recent versions and other details about the dependency library provided by Spark Streaming SQL. We recommend that you use the latest version of the dependency library.

Dependency library	Supported version	Release date	Reference string	Description
datasources-bundle	2.0.0 (recommended)	2020/02/26	sharedlibs:streamingsql:datasources-bundle:2.0.0	Supported data sources include Kafka, LogHub, Druid, Tablestore, HBase, JDBC, DataHub, Redis, Kudu, and DTS.
	1.9.0	2019/11/20	sharedlibs:streamingsql:datasources-bundle:1.9.0	Supported data sources include Kafka, LogHub, Druid, Tablestore, HBase, JDBC, DataHub, Redis, and Kudu.
	1.8.0	2019/10/17	sharedlibs:streamingsql:datasources-bundle:1.8.0	Supported data sources include Kafka, LogHub, Druid, Tablestore, HBase, JDBC, DataHub, and Redis.
	1.7.0	2019/07/29	sharedlibs:streamingsql:datasources-bundle:1.7.0	Supported data sources include Kafka, LogHub, Druid, Tablestore, HBase, and JDBC.

For more information, see Overview.

Prerequisites

- A project is created. For more information, see Manage projects.
- Resources and data files required for a job are obtained, such as JAR packages, names of the data files, and storage paths of both the JAR packages and data files.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find your project and click **Edit Job** in the Actions column.
3. Create a Streaming SQL job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **Streaming SQL** from the **Job Type** drop-down list.
   iii. Click **OK**.
4. Edit job content.

   Specify the command line parameters required to submit the job in the **Content** field. Example:

```
--- Create a Log Service table.
CREATE TABLE IF NOT EXISTS ${slsTableName}
   USING loghub
   OPTIONS (
        sls.project = '${logProjectName}',
        sls.store = '${logStoreName}',
        access.key.id = '${accessKeyId}',
        access.key.secret = '${accessKeySecret}',
        endpoint = '${endpoint}'
   );
--- Import data to HDFS.
INSERT INTO
    ${hdfsTableName}
SELECT
    col1, col2
FROM  ${slsTableName}
WHERE ${condition}
```

Note The command used to submit a Streaming SQL job is `streaming-sql -f {sql_script}`. The SQL statements that you enter in the job editor are saved in `sql_script`.

5. Configure dependency libraries and actions on failures.
   i. Click **Job Settings** in the upper-right corner.
   ii. On the **Shared Libraries** and **Streaming Task Settings** tabs, configure dependency libraries and actions on failures.

Section	Configuration item	Description
Dependent Libraries	**Libraries**	Job execution depends on some library files related to data sources. EMR publishes the libraries to the repository of the scheduling center as dependency libraries. You must specify dependency libraries when you create a job. To specify a dependency library, enter its reference string, such as *sharedlibs:streamingsql:datasources-bundle:2.0.0*.
Actions on Failures	**Action on Current Statement Failure**	The action to perform when EMR fails to execute a statement. You can perform one of the following actions: ■ **Execute Next Statement**: Execute the next statement. ■ **Terminate Job**: Terminate the job.

   iii. Click **Save**.

9.14. Configure a Presto SQL job

If you want to use Presto SQL during data development, you can configure a Presto SQL job in the E-MapReduce (EMR) console. This topic describes how to configure a Presto SQL job in the EMR console.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the Projects section of the page that appears, find the project that you want to manage and click **Edit Job** in the Actions column.
3. Create a Presto SQL job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and then select **Presto SQL** from the **Job Type** drop-down list.
   This option indicates that a Presto SQL job will be created. You can use the following command syntax to submit a Presto SQL job:

```
presto <options> -f {SQL_SCRIPT}
```

   Note `SQL_SCRIPT` refers to the SQL statements that you entered in the job editor.

   iii. Click **OK**.
4. Edit job content.
   i. Configure the command line parameters required to submit the job in the **Content** field.
   Example:

```
SELECT * from table1;
```

   ii. Click **Save**.

Configure Presto CLI parameters

By default, Presto queries data tables under the hive catalog and default schema. You can configure Presto command-line interface (CLI) parameters to specify catalogs and schemas. You can use one of the following methods to configure Presto CLI parameters in a Presto SQL job:

- Use environment variables
  - Password: If password authentication is enabled for the Presto service, add the `PRESTO_PASSWORD` environment variable to specify a password.

- Other parameters: Configure the parameters in the `PRESTO_CLI_PARAMS` environment variable. Example: `PRESTO_CLI_PARAMS="--catalog mysql --schema db1 "`.

- Use custom variables
  - Password: Add a custom variable named `presto.password` to the job to specify a password for Presto authentication.
  - Other parameters: Add custom variables in the format of `_presto.xxx` to the job. The custom variables are added to the list of Presto CLI parameters. The corresponding option is in the format of `--xxx`.

    The following custom variables are supported:

```
## Basic parameters
* _presto.schema <schema>
* _presto.catalog <catalog>
## Control and debugging parameters
* _presto.trace-token <trace token>
* _presto.session <session>...
* _presto.source <source>
* _presto.resource-estimate <resource-estimate>...
* _presto.log-levels-file <log levels file>
## Connection parameters
* _presto.server <server>
* _presto.http-proxy <http-proxy>  * ignore-errors
* _presto.socks-proxy <socks-proxy>
## Authentication parameters
* _presto.user <user>
* _presto.password <password>
* _presto.client-info <client-info>
* _presto.client-request-timeout <client request timeout>
* _presto.client-tags <client tags>
* _presto.access-token <access token>
* _presto.truststore-password <truststore password>
* _presto.truststore-path <truststore path>
* _presto.keystore-password <keystore password>
* _presto.keystore-path <keystore path>
* _presto.extra-credential <extra-credential>...
## High-security parameters
* _presto.krb5-config-path <krb5 config path>
* _presto.krb5-credential-cache-path <krb5 credential cache path>
* _presto.krb5-disable-remote-service-hostname-canonicalization
* _presto.krb5-keytab-path <krb5 keytab path>
* _presto.krb5-principal <krb5 principal>
* _presto.krb5-remote-service-name <krb5 remote service name>
* _presto.krb5-service-principal-pattern <krb5 remote service principal pattern>
```

9.15. Configure an Impala SQL job

If you need to use Impala SQL during data development, you can configure an Impala SQL job in an E-MapReduce (EMR) cluster. This topic describes how to configure an Impala SQL job.

Prerequisites

A project is created. For more information, see Manage projects.

Procedure

1. Go to the Data Platform tab.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
2. In the **Projects** section, find your project and click **Edit Job** in the Actions column.
3. Create an Impala SQL job.
   i. In the Edit Job pane on the left, right-click the folder on which you want to perform operations and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **Impala SQL** from the **Job Type** drop-down list.

   You can use the following command syntax to submit an Impala SQL job:

```
impala-shell -f {SQL_CONTENT} [options];
```

   The following table describes the parameters in the syntax.

Parameter	Description
SQL_CONTENT	The entered SQL statement.
options	The setting of the IMPALA_CLI_PARAMS parameter that you configure by performing the following operations: Click **Job Settings** in the upper-right corner of the job page. In the Job Settings pane, click the **Advanced Settings** tab. Click the + icon in the **Environment Variables** section and add the IMPALA_CLI_PARAMS parameter. For example, set `IMAPAL_CLI_PARAMS` to "-u hive".

   iii. Click **OK**.
4. Edit job content.

i. Enter the Impala SQL statements in the **Content** field.

Example:

```
show databases;
show tables;
select * from test1;
```

ii. Click **Save**.

10.FAQ about data development

This topic provides answers to some frequently asked questions about data development.

Questions about jobs:

- What are the differences between jobs and workflows?
- What do I do if a TPS conflict occurs when multiple consumer IDs consume the same topic?
- Why does an external table created in Hive contain no data?
- Why does a Spark Streaming job stop running unexpectedly?
- Why is a Spark Streaming job still in the Running state in the EMR console after the job has been stopped?
- How do I include local shared libraries in a MapReduce job?
- How do I specify the file path of an OSS data source for a MapReduce or Spark job?
- How do I use Beeline to connect to Kerberos-authenticated clusters?
- What do I do if an out-of-memory error is reported when Spark receives Flume data?
- Why does a job run slowly?
- Why does AppMaster take a long time to start a task?
- What do I do if the timestamp field shows a delay of eight hours when I import data from ApsaraDB RDS into EMR?
- What do I do if a job stays in the SUBMITTING state for a long period of time?

Questions about logs:

- How do I view job execution records?
- How do I view logs on Object Storage Service (OSS)?
- Can I view the job logs stored in core nodes?
- How do I view the logs of a service deployed in an EMR cluster?
- How do I clear the log data of a completed job?

Questions about exception diagnosis:

- How do I fix the error "Error: Could not find or load main class"?
- What do I do if the error "Invalid authorization specification, message from server: ip not in whitelist" is reported when I connect Spark SQL to ApsaraDB RDS?
- What do I do if the following error is reported when I read data from or write data to MaxCompute tables: "java.lang.RuntimeException.Parse response failed: '<!DOCTYPE html>...'"?
- What do I do if the error "Exception in thread "main" java.sql.SQLException: No suitable driver found for jdbc:mysql:xxx" is reported?
- What do I do if the following error is reported when I run a Hive or Impala job to read Parquet data (including columns of the DECIMAL type) written by Spark SQL: "Failed with exception java.io.IOException:org.apache.parquet.io.ParquetDecodingException: Can not read value at 0 in block -1 in file hdfs://.../.../part-00000-xxx.snappy.parquet"?
- What do I do if the Thrift Server process properly runs but the "Connection refused telnet emr-header-1 10001" error is reported?
- How do I fix the Spark job error "Container killed by YARN for exceeding memory limits" or the MapReduce job error "Container is running beyond physical memory limits"?
- What do I do if the following error is reported: "Error: Java heap space"?
- What do I do if the "No space left on device" error is reported?
- What do I do if the error "ConnectTimeoutException" or "ConnectionException" is reported when I access OSS or Log Service?
- What do I do if an out-of-memory error is reported when I run a job to read a Snappy file?
- What do I do if the following error is reported: "Exception in thread main java.lang.RuntimeException: java.lang.ClassNotFoundException: Class com.aliyun.fs.oss.nat.NativeOssFileSystem not found"?
- What do I do if the error "java.lang.NoSuchMethodError:org.apache.http.conn.ssl.SSLConnetionSocketFactory.init(Ljavax/net/ssl/SSLContext;Ljavax/net/ssl/HostnameVerifier)" is reported when OSS SDK is used in Spark?
- What do I do if the following error is reported: "java.lang.IllegalArgumentException: Wrong FS: oss://xxxxx, expected: hdfs://ip:9000"?
- What do I do if the error "java.lang.IllegalArgumentException: Size exceeds Integer.MAX_VALUE" is reported when a Spark job runs?

Questions about feature usage:

- Does EMR support real-time computing?
- What do I do if the timestamp field shows a delay of eight hours when I import data from ApsaraDB RDS into EMR?
- How do I modify the spark-env configurations of the Spark service?
- How do I pass job parameters to a script?
- How do I set the authentication method of HiveServer2 to LDAP?
- How do I enable the HDFS balancer in an EMR cluster and optimize the performance of the HDFS balancer?
- How do I submit a Spark job in standalone mode?
- What do I do if the Custom Configuration button is not displayed for a service on the EMR console?

What are the differences between jobs and workflows?

- Job

  When you create a job on E-MapReduce (EMR), you configure only the JAR package, data input and output addresses, and some runtime parameters. These configurations determine how the job runs. After you complete the configurations and specify a job name, the job is created.

- Workflow

  A workflow associates a job with a cluster.

  - You can use a workflow to run a sequence of jobs.
  - You can specify an existing cluster for a workflow to run jobs. If you do not specify a cluster, a temporary cluster is automatically created for the workflow.
  - You can also schedule the execution of your workflow. After all jobs of the workflow are completed, the temporary cluster is automatically released.

- You can view the status and logs of each execution in the records of each workflow.

How do I view job execution records?

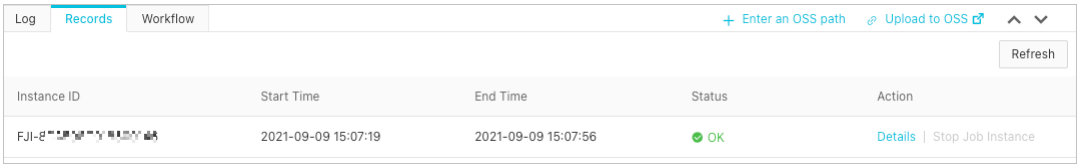
View the details of the job in the Data Platform module of the EMR console or on the YARN web UI.

- Use the Data Platform module of the EMR console

  This method is suitable for jobs that are created and submitted in the EMR console.

  i. After you submit the job, view the operational logs of the job on the **Log** tab.

  ii. Click the **Records** tab to view the execution records of the job instance.



  iii. Find the record whose details you want to view and click **Details** in the Action column. On the **Scheduling Center** page, view the information about the job instance, job submission logs, and YARN containers.

- Use the YARN web UI

  This method is suitable for jobs that are created and submitted in the EMR console or CLI.

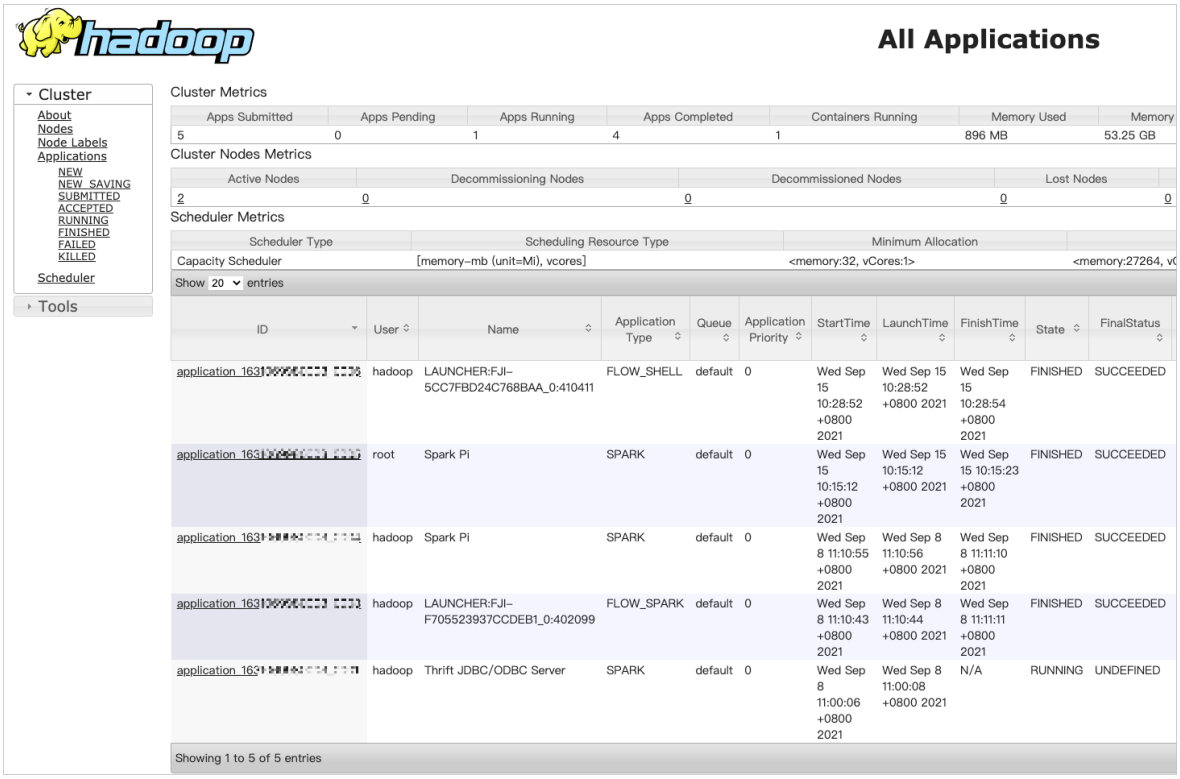
  i. Enable port 8443. For more information, see Configure security group rules.

  ii. In the left-side navigation pane of the **Clusters and Services** page for the cluster, click **Connect Strings**.

  iii. On the Public Connect Strings page, click the link for **YARN UI**.

  To access the YARN web UI by using your Knox account, you must obtain the username and password of the Knox account. For more information, see Manage user accounts.

  iv. In the Hadoop console, click the ID of the job to view the details of the job.



How do I view logs on Object Storage Service (OSS)?

1. Log on to the EMR console and click the Data Platform tab. Find the project, and click Workflows in the Actions column. In the left-side navigation pane of the page that appears, click the workflow whose logs you want to view. Click the **Records** tab in the lower part of the page.
2. On the **Records** tab, find the workflow instance that you want to view, and click **Details** in the Action column. On the **Job Instance Info** tab of the page that appears, obtain the ID of the cluster where the job is run.
3. Go to the *OSS://mybucket/emr/spark* directory and find the folder named after the cluster ID.
4. Go to the *OSS://mybucket/emr/spark/clusterID/jobs* directory, which contains folders named after job IDs. Each directory stores the operational logs of a job.

What do I do if the following error is reported when I read data from or write data to MaxCompute tables: "java.lang.RuntimeException.Parse response failed: '<!DOCTYPE html>...'"?

- Cause: The MaxCompute Tunnel endpoint may be invalid.
- Solution: Enter a valid MaxCompute Tunnel endpoint. For more information, see Endpoints.

What do I do if a TPS conflict occurs when multiple consumer IDs consume the same topic?

Cause: This topic may have been created during public preview or in other environments. This causes consumption data inconsistency among multiple consumer groups.
Solution: submit a ticket. Specify the topic and consumer IDs in the ticket.

Can I view the job logs stored in core nodes?

Yes, you can view the job logs stored in core nodes on the YARN web UI. For more information, see Use the YARN web UI.

Why does an external table created in Hive contain no data?

- Problem description: After an external table is created, the table is queried but no data is returned.

  CREATE EXTERNAL TABLE statement:

  ```
  CREATE EXTERNAL TABLE storage_log(content STRING) PARTITIONED BY (ds STRING)
      ROW FORMAT DELIMITED
      FIELDS TERMINATED BY '\t'
      STORED AS TEXTFILE
      LOCATION 'oss://log-12453****/your-logs/airtake/pro/storage';
  ```

  Command that is used to query data:

  ```
  select * from storage_log;
  ```

- Cause: Hive does not automatically associate a Partitions directory.
- Solution:
  i. Manually specify a Partitions directory.

  ```
  alter table storage_log add partition(ds=123);
  ```

  ii. Query data from the log.

  ```
  select * from storage_log;
  ```

  The following data is returned:

  ```
      OK
  abcd    123
  efgh    123
  ```

Why does a Spark Streaming job stop running unexpectedly?

- Check whether the Spark version is earlier than 1.6. If it is, update it.

  Spark versions earlier than 1.6 have a memory leak bug. This bug can cause the container to stop running unexpectedly.
- Check whether your code has been optimized in terms of memory usage.

Why is a Spark Streaming job still in the Running state in the EMR console after the job has been stopped?

- Cause: EMR cannot effectively monitor the status of Spark Streaming jobs that run in yarn-client mode.
- Solution: Change the running mode of the job to yarn-cluster.

How do I fix the error "Error: Could not find or load main class"?

Check whether the path protocol header of the JAR file for the job is `ossref`. If it is not, change it to `ossref`.

How do I include local shared libraries in a MapReduce job?

Log on to the EMR console, go to the **Configure** tab of the YARN service page, and then modify parameters on the *mapred-site* tab based on the following information:

```
<property>
    <name>mapred.child.java.opts</name>
    <value>-Xmx1024m -Djava.library.path=/usr/local/share/</value>
  </property>
  <property>
    <name>mapreduce.admin.user.env</name>
    <value>LD_LIBRARY_PATH=$HADOOP_COMMON_HOME/lib/native:/usr/local/lib</value>
  </property>
```

How do I specify the file path of an OSS data source for a MapReduce or Spark job?

You can specify the input and output data sources of a job in the `oss://[accessKeyId:accessKeySecret@]bucket[.endpoint]/object/path` format, which is similar to `hdfs://` URLs.

You can access OSS data with or without an AccessKey pair:

- (Recommended) EMR provides MetaService, which allows you to access OSS data without an AccessKey pair. You can specify a path in the *oss://bucket/object/path* format.
- (Not recommended) You can configure the AccessKey ID, AccessKey secret, and endpoint parameters on the Configuration object for a MapReduce job or the SparkConf object for a Spark job. You can also directly specify the AccessKey ID, AccessKey secret, and endpoint in a Uniform Resource Identifier (URI). For more information, see Development preparations.

What do I do if the error "Exception in thread "main" java.sql.SQLException: No suitable driver found for jdbc:mysql:xxx" is reported?

- Cause: The version of *mysql-connector-java* is not supported.

- Solution: Update *mysql-connector-java* to the latest version.

What do I do if the error "Invalid authorization specification, message from server: ip not in whitelist" is reported when I connect Spark SQL to ApsaraDB RDS?

Add the internal IP addresses of the cluster nodes into a whitelist of ApsaraDB RDS.

What do I do if the following error is reported when I run a Hive or Impala job to read Parquet data (including columns of the DECIMAL type) written by Spark SQL: "Failed with exception java.io.IOException:org.apache.parquet.io.ParquetDecodingException: Can not read value at 0 in block -1 in file hdfs://.../.../part-00000-xxx.snappy.parquet"?

Cause: The DECIMAL data type has different representations in the different Parquet conventions used in Hive and Spark SQL. Parquet data (including columns of the DECIMAL type) written by Spark SQL cannot be read properly by using Hive. Solution: To solve this issue, we recommend that you set the spark.sql.parquet.writeLegacyFormat parameter to true before you import the Parquet data written by Spark SQL to Hive or Impala. This setting makes Spark use the same convention as Hive or Impala for writing the Parquet data.

How do I use Beeline to connect to Kerberos-authenticated clusters?

- High-availability (HA) cluster (service discovery mode)

```
!connect jdbc:hive2://emr-header-1:2181,emr-header-2:2181,emr-header-3:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2;principal=hive/_HOST@EMR.${clusterId).COM
```

- HA cluster
  - Connect to emr-header-1

```
!connect jdbc:hive2://emr-header-1:10000/;principal=hive/emr-header-1@EMR.${clusterId}.COM
```

  - Connect to emr-header-2

```
!connect jdbc:hive2://emr-header-2:10000/;principal=hive/emr-header-2@EMR.${clusterId}.COM
```

- Non-HA cluster

```
!connect jdbc:hive2://emr-header-1:10000/;principal=hive/emr-header-1@EMR.${clusterId}.COM
```

What do I do if the Thrift Server process properly runs but the "Connection refused telnet emr-header-1 10001" error is reported?

Check the logs in the */mnt/disk1/log/spark* directory. This issue is caused by the Thrift Server running out of memory (OOM). You can increase memory by setting the spark.driver.memory parameter to a larger value.

How do I view the logs of a service deployed in an EMR cluster?

Log on to the master node of the cluster. View the logs of the service in the */mnt/disk1/log* directory.

How do I fix the Spark job error "Container killed by YARN for exceeding memory limits" or the MapReduce job error "Container is running beyond physical memory limits"?

- Cause: The amount of memory requested when you submit an application is too low. The JVM occupies more memory than the allocated amount during startup. As a result, the job is abnormally terminated by NodeManager. In particular, Spark jobs may consume a large amount of off-heap memory and are more likely to be abnormally terminated.
- Solution:
  - For a Spark job, log on to the EMR console, go to the **Configure** tab of the Spark service page, and then set spark.yarn.driver.memoryOverhead or spark.yarn.executor.memoryOverhead to a larger value.
  - For a MapReduce job, log on to the EMR console, go to the **Configure** tab of the YARN service page, and then set mapreduce.map.memory.mb or mapreduce.reduce.memory.mb to a larger value.

What do I do if the following error is reported: "Error: Java heap space"?

- Cause: The task has large amounts of data to process but the JVM has insufficient memory. As a result, an out-of-memory error is returned.
- Solution:
  - For a Spark job, log on to the EMR console, go to the **Configure** tab of the Spark service page, and then set spark.executor.memory or spark.driver.memory to a larger value.
  - For a MapReduce job, log on to the EMR console, go to the **Configure** tab of the YARN service page, and then set mapreduce.map.java.opts or mapreduce.reduce.java.opts to a larger value.

What do I do if the "No space left on device" error is reported?

- Cause:
  - A master or core node has insufficient storage space, which causes a job submission failure.
  - If a disk is full, an exception may occur in local Hive metadatabases such as MySQL Server, or a Hive metastore connection error may occur.
- Solution: Free up enough disk space on the master node, including the system disk and HDFS space.

What do I do if the error "ConnectTimeoutException" or "ConnectionException" is reported when I access OSS or Log Service?

- Cause: The OSS endpoint is a public endpoint, but your EMR core node does not have a public IP address. As a result, you cannot access OSS. The error is reported when you access Log Service for the same reason.
- Solution:
  - Change the OSS endpoint to an internal endpoint.
  - You can also use MetaService provided by EMR to access OSS or Log Service. If you use this method, you do not need to specify an endpoint.

For example, the `select * from tbl limit 10` command can be successfully executed, but `Hive SQL: select count(1) from tbl` fails. Change the OSS endpoint to an internal network endpoint.

```
alter table tbl set location "oss://bucket.oss-cn-hangzhou-internal.aliyuncs.com/xxx"
alter table tbl partition (pt = 'xxxx-xx-xx') set location "oss://bucket.oss-cn-hangzhou-internal.aliyuncs.com/xxx"
```

What do I do if an out-of-memory error is reported when I run a job to read a Snappy file?

- Cause: The format of standard Snappy files written by Log Service is different from that of Hadoop Snappy files. EMR processes Hadoop Snappy files, so an out-of-memory error is reported when it processes standard Snappy files.
- Solution: Log on to the EMR console, go to the **Configure** tab for the service that you use, and then configure one of the following parameters based on the job type:
  - For a Hive job, set `io.compression.codec.snappy.native` to true.
  - For a MapReduce job, set `Dio.compression.codec.snappy.native` to true.
  - For a Spark job, set `spark.hadoop.io.compression.codec.snappy.native` to true.

What do I do if the following error is reported: "Exception in thread main java.lang.RuntimeException: java.lang.ClassNotFoundException: Class com.aliyun.fs.oss.nat.NativeOssFileSystem not found"?

If you want to use Spark jobs to read or write OSS data, you must install EMR SDK. For more information, see Preparations.

What do I do if an out-of-memory error is reported when Spark receives Flume data?

Check whether the data receiving method is push-based. If it is not, change the data receiving method to push-based. For more information, see Spark Streaming and Flume integration guide.

What do I do if the following error is reported: "Caused by: java.io.IOException: Input stream cannot be reset as 5242880 bytes have been written, exceeding the available buffer size of 524288"?

The cache for network connection retries is insufficient. We recommend that you use *aliyun-java-sdk-emr* later than V1.1.0.

What do I do if the error "java.lang.NoSuchMethodError:org.apache.http.conn.ssl.SSLConnetionSocketFactory.init(Ljavax/net/ssl/SSLCor is reported when OSS SDK is used in Spark?

OSS SDK and the Spark and Hadoop running environments have version dependency conflicts. We recommend that you do not use OSS SDK in code.

What do I do if the following error is reported: "java.lang.IllegalArgumentException: Wrong FS: oss://xxxxx, expected: hdfs://ip:9000"?

The default file system of HDFS is used when you process OSS data. You must use the OSS path to initialize the file system so that it can be used to process OSS data.

```
Path outputPath = new Path(EMapReduceOSSUtil.buildOSSCompleteUri("oss://bucket/path", conf));
org.apache.hadoop.fs.FileSystem fs = org.apache.hadoop.fs.FileSystem.get(outputPath.toUri(), conf);
if (fs.exists(outputPath)) {
  fs.delete(outputPath, true);
}
```

How do I clear the log data of a completed job?

- Problem description: The HDFS space of a cluster is full. A large volume of data is stored in the */spark-history* directory.
- Solution:
  i. Go to the **Configure** tab for the Spark service. Check whether the **spark_history_fs_cleaner_enabled** parameter is specified in the **Service Configuration** section.
     - If it is, change its value to true to periodically clear the logs of completed jobs.
     - If it is not, click the **spark-defaults** tab. Click **Custom Configuration** in the upper-right corner. In the dialog box that appears, add the **spark_history_fs_cleaner_enabled** parameter and set it to true.
  ii. Select **Restart All Components** from the **Actions** drop-down list in the upper-right corner.
  iii. In the **Cluster Activities** dialog box, specify **Description** and click **OK**.
  iv. In the **Confirm** message, click **OK**.

Why does a job run slowly?

- Cause: If the size of the heap memory on the JVM where the job runs is too small, garbage collection may take a long time. As a result, the performance of the job is affected.
- Solution:
  - For a Tez job, log on to the EMR console, go to the **Configure** tab of the Tez service page, and then set hive.tez.java.opts to a larger value.
  - For a Spark job, log on to the EMR console, go to the **Configure** tab of the Spark service page, and then set spark.executor.memory or spark.driver.memory to a larger value.
  - For a MapReduce job, log on to the EMR console, go to the **Configure** tab of the YARN service page, and then set mapreduce.map.java.opts or mapreduce.reduce.java.opts to a larger value.

Why does AppMaster take a long time to start a task?

- Cause: If the number of tasks or Spark executors is large, AppMaster may take a long time to start a task. The running duration of a single task is short, but the overhead for scheduling jobs is large.
- Solution:
  - Use CombinedInputFormat to reduce the number of tasks.
  - Increase the block size (dfs.blocksize) of the data generated by former jobs.

- Set mapreduce.input.fileinputformat.split.maxsize to a larger value.
- For a Spark job, log on to the EMR console and go to the **Configure** tab of the Spark service page. Then, set spark.executor.instances to a smaller value to reduce the number of executors or set spark.default.parallelism to a smaller value to reduce the number of parallel jobs.

What do I do if the error "java.lang.IllegalArgumentException: Size exceeds Integer.MAX_VALUE" is reported when a Spark job runs?

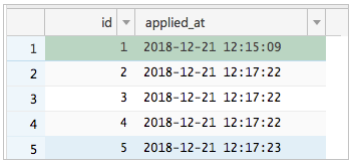
During data shuffling, the number of partitions is smaller than the value of the Integer.MAX_VALUE parameter. You can increase the number of partitions by performing the following operations: Log on to the EMR console, go to the **Configure** tab of the Spark service page, and then set spark.default.parallelism and spark.sql.shuffle.partitions to larger values. You can also perform the repartition operation before you perform data shuffling.

Does EMR support real-time computing?

EMR provides three types of real-time computing services: Spark Streaming, Storm, and Flink.

What do I do if the timestamp field shows a delay of eight hours when I import data from ApsaraDB RDS into EMR?

- Problem description:
  i. The Test_Table table in ApsaraDB RDS contains a timestamp field.



	id	applied_at
1	1	2018-12-21 12:15:09
2	2	2018-12-21 12:17:22
3	3	2018-12-21 12:17:22
4	4	2018-12-21 12:17:22
5	5	2018-12-21 12:17:23

  ii. The following command is used to import data in the Test_Table table to EMR HDFS:

```
sqoop import \
--connect jdbc:mysql://rm-2ze****341.mysql.rds.aliyuncs.com:3306/s***o_sqoopp_db \
--username s***o \
--password ****** \
--table play_evolutions \
--target-dir /user/hadoop/output \
--delete-target-dir \
--direct \
--split-by id \
--fields-terminated-by '|' \
-m 1
```

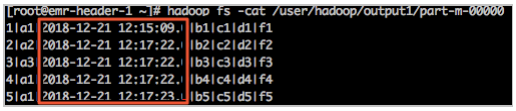
  iii. The import result is queried.

  In the query result, the timestamp field shows a delay of eight hours.

- Solution: When you import the data, delete the --direct parameter from the import command.

```
sqoop import \
--connect jdbc:mysql://rm-2ze****341.mysql.rds.aliyuncs.com:3306/s***o_sqoopp_db \
--username s***o \
--password ****** \
--table play_evolutions \
--target-dir /user/hadoop/output \
--delete-target-dir \
--split-by id \
--fields-terminated-by '|' \
-m 1
```

The query results are normal.



```
[root@emr-header-1 ~]# hadoop fs -cat /user/hadoop/output1/part-m-00000
1|a1|2018-12-21 12:15:09.|b1|c1|d1|f1
2|a2|2018-12-21 12:17:22.|b2|c2|d2|f2
3|a3|2018-12-21 12:17:22.|b3|c3|d3|f3
4|a1|2018-12-21 12:17:22.|b4|c4|d4|f4
5|a1|2018-12-21 12:17:23.|b5|c5|d5|f5
```

How do I modify the spark-env configurations of the Spark service?

Log on to the master node of the cluster and modify the configurations in the */etc/ecm/spark-conf/spark-env.sh* and */var/lib/ecm-agent/cache/ecm/service/SPARK/<Version Number>/package/templates/spark-env.sh* files.

> **Note** If you submit a task on a core node, you also need to modify the configurations on the core node.

How do I pass job parameters to a script?

You can run a script in a Hive job and use the `-hivevar` option to pass job parameters to the script.

1. Prepare a script.

   In a script, you can reference a variable in the format of `${varname}`, for example, `${rating}`. Example:

   - Script name: *hivesql.hive*
   - Path of the script in OSS: *oss://bucket_name/path/to/hivesql.hive*
   - Script content:

```
use default;
 drop table demo;
 create table demo (userid int, username string, rating int);
 insert into demo values(100,"john",3),(200,"tom",4);
 select * from demo where rating=${rating};
```

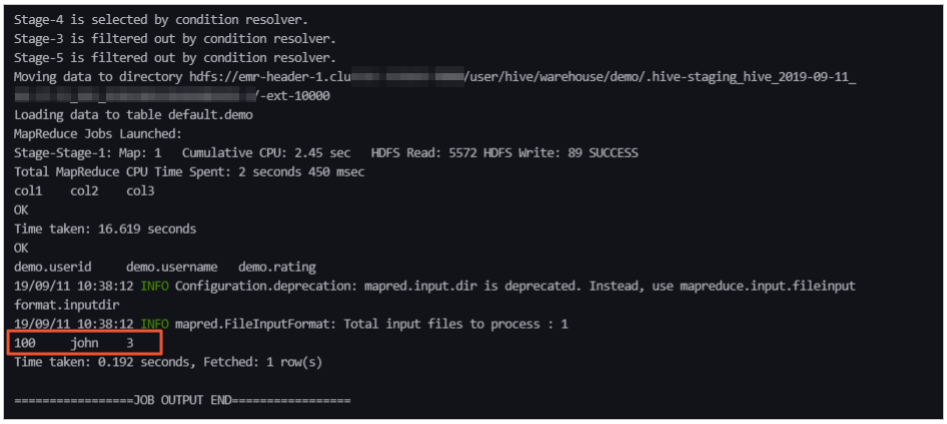
2. Go to the Data Platform page.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Data Platform** tab.
3. In the Projects section of the page that appears, find the project that you want to edit and click **Edit Job** in the Actions column.
4. Create a Hive job.
   i. In the Edit Job pane on the left, right-click the folder that you want to manage and select **Create Job**.
   ii. In the **Create Job** dialog box, specify **Name** and **Description**, and select **Hive** from the **Job Type** drop-down list.
   iii. Click **OK**.
5. Edit job content.
   i. Click Job Settings in the upper-right corner of the page. On the **Basic Settings** tab, specify **Key** and **Value** in the Configuration Parameters section. Set **Key** to a variable specified in the script, for example, **rating**.

   ii. Enter the following code in the Content field of the job to use the `-hivevar` option to pass the parameters configured in the job to the variables in the script:

```
-hivevar rating=${rating} -f ossref://bucket_name/path/to/hivesql.hive
```

6. Run the job.

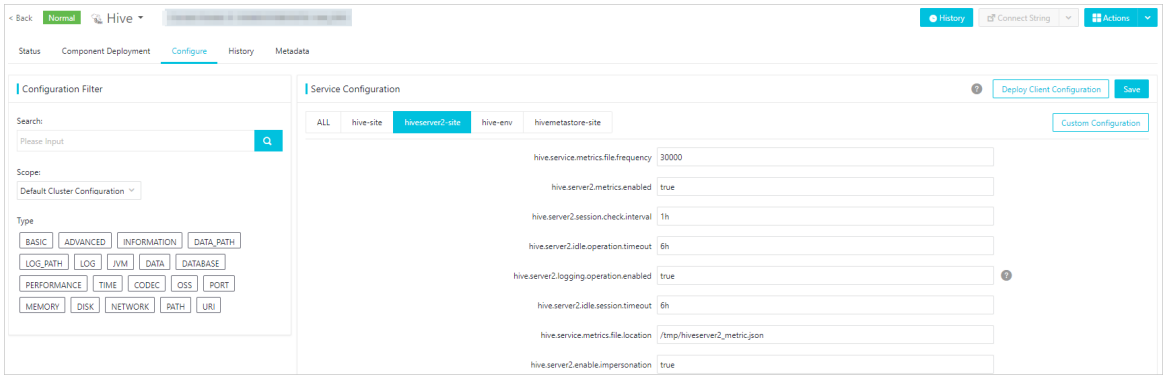
   The following figure shows the result of the job.



How do I set the authentication method of HiveServer2 to LDAP?

1. Go to the cluster details page.
   i. Log on to the Alibaba Cloud EMR console by using your Alibaba Cloud account.
   ii. In the top navigation bar, select the region where your cluster resides and select a resource group based on your business requirements.
   iii. Click the **Cluster Management** tab.
   iv. Find the cluster whose HiveServer2 authentication method you want to change and click **Details** in the **Actions** column.
2. In the left-side navigation pane, choose **Cluster Service > Hive**.
3. Set the authentication method of HiveServer2 to LDAP and restart HiveServer2.

i. Click the **Configure** tab. In the **Service Configuration** section, click the **hiveserver2-site** tab.



ii. Click **Custom Configuration** in the upper-right corner.

To set the authentication method of HiveServer2 to **LDAP**, configure the parameters described in the following table.

Parameter	Value	Description
hive.server2.authentication	LDAP	The authentication method.
hive.server2.authentication.ldap.url	Format: ldap://${emr-header-1-hostname}:10389	Replace ${emr-header-1-hostname} with your hostname. You can run the `hostname` command on the emr-header-1 node of your cluster to obtain the hostname. For more information about how to log on to the emr-header-1 node, see Log on to a cluster.
hive.server2.authentication.ldap.baseDN	**ou=people,o=emr**	None.

iii. Click **Save** in the upper-right corner.

iv. In the **Confirm Changes** dialog box, configure the parameters and click **OK**.

v. In the upper-right corner of the page, select **Restart HiveServer2** from the **Actions** drop-down list.

vi. In the **Cluster Activities** dialog box, configure the parameters and click **OK**.

vii. In the **Confirm** message, click **OK**.

4. Add an account to the LDAP service.

   In an EMR cluster, OpenLDAP is a component of the LDAP service. OpenLDAP is used to manage Knox accounts by default. HiveServer2 can reuse the Knox accounts for LDAP authentication. For more information about how to add an account, see Manage users.

   In this example, the **emr-guest** account is added.

5. Check whether you can use the new account to log on to HiveServer2.

   Use */usr/lib/hive-current/bin/beeline* to log on to HiveServer2.

```
beeline> !connect jdbc:hive2://emr-header-1:10000/
Enter username for jdbc:hive2://emr-header-1:10000/: emr-guest
Enter password for jdbc:hive2://emr-header-1:10000/: emr-guest-pwd
Transaction isolation: TRANSACTION_REPEATABLE_READ
```

   If the account or password is invalid, the following error message appears:

```
Error: Could not open client transport with JDBC Uri: jdbc:hive2://emr-header-1:10000/: Peer indicated failure: Error validating the login (state
=08S01,code=0)
```

How do I enable the HDFS balancer in an EMR cluster and optimize the performance of the HDFS balancer?

1. Log on to a node of your cluster.
2. Run the following commands to switch to the hdfs user and run the HDFS balancer:

```
su hdfs
/usr/lib/hadoop-current/sbin/start-balancer.sh -threshold 10
```

3. Check the running status of the HDFS balancer:
   - Method 1:

```
less /var/log/hadoop-hdfs/hadoop-hdfs-balancer-emr-header-xx.cluster-xxx.log
```

   - Method 2:

```
tailf /var/log/hadoop-hdfs/hadoop-hdfs-balancer-emr-header-xx.cluster-xxx.log
```

   Note If the command output includes `Successfully`, the HDFS balancer is running.

   The following table describes the HDFS balancer parameters.

Parameter	Description
Threshold	Default value: 10%. This value ensures that the disk usage on each DataNode differs from the overall usage in the cluster by no more than 10%. If the overall usage of the cluster is high, set this parameter to a smaller value. If a large number of new nodes are added to the cluster, you can set this parameter to a larger value to move data from the high-usage nodes to the low-usage nodes.
dfs.datanode.balance.max.concurrent.moves	Default value: 5. Specifies the maximum number of concurrent block moves that are allowed in a DataNode. Set this parameter based on the number of disks. We recommend that you set this parameter to `4 × Number of disks` as the upper limit for a DataNode. For example, if a DataNode has 28 disks, set this parameter to 28 on the HDFS balancer and `112` on the DataNode. Adjust the value based on the cluster load. Increase the value when the cluster load is low and decrease the value when the cluster load is high. **Note** After you set this parameter for a DataNode, restart the DataNode for the parameter setting to take effect.
dfs.balancer.dispatcherThreads	The number of dispatcher threads used by the HDFS balancer to determine the blocks that need to be moved. Before the HDFS balancer moves a specific amount of data between two DataNodes, the HDFS balancer repeatedly retrieves block lists for moving blocks until the required amount of data is scheduled. **Note** Default value: 200.
dfs.balancer.rpc.per.sec	The number of remote procedure calls (RPCs) sent by dispatcher threads per second. Default value: 20. Before the HDFS balancer moves data between two DataNodes, it uses dispatcher threads to repeatedly send the getBlocks() RPC to the NameNode. This results in a heavy load on the NameNode. To avoid this issue and balance the cluster load, we recommend that you set this parameter to limit the number of RPCs sent per second. For example, you can decrease the value of the parameter by 10 or 5 for a cluster with a high load to minimize the impact on the overall moving progress.
dfs.balancer.getBlocks.size	The total data size of the blocks moved each time. Before the HDFS balancer moves data between two DataNodes, the HDFS balancer repeatedly retrieves block lists for moving blocks until the required amount of data is scheduled. By default, the size of blocks in each block list is 2 GB. When the NameNode receives a getBlocks() RPC, the NameNode is locked. If an RPC queries a large number of blocks, the NameNode is locked for a long period of time. This slows down data writing. To avoid this issue, we recommend that you set this parameter based on the NameNode load.
dfs.balancer.moverThreads	Default value: 1000. Each block move requires a thread. This parameter limits the total number of concurrent moves.
dfs.namenode.balancer.request.standby	Default value: false. Specifies whether the HDFS balancer queries the blocks to be moved on the standby NameNode. When a NameNode receives a getBlocks() RPC, the NameNode is locked. If an RPC queries a large number of blocks, the NameNode is locked for a long period of time. This slows down data writing. If you use an HA cluster, the HDFS balancer sends RPCs only to the standby NameNode.
dfs.balancer.getBlocks.min-block-size	The minimum size of blocks to be queried by the getBlocks() RPC. After you set this parameter, the getBlocks() RPC skips blocks smaller than the minimum size. This improves the query efficiency. Default value: 10485760 (10 MB).
dfs.balancer.max-iteration-time	The maximum duration of each iteration for moving blocks between two DataNodes. Default value: 1200000. Unit: milliseconds. When the duration of an iteration exceeds the limit, the HDFS balancer automatically enters the next iteration.
dfs.balancer.block-move.timeout	Default value: 0. Unit: milliseconds. When the HDFS balancer moves blocks, an iteration may last for a long time because some block moves are still going on. You can set this parameter to avoid this issue.

The following table describes the DataNode parameters.

Parameter	Description

Parameter	Description
dfs.datanode.balance.bandwidthPerSec	Specifies the bandwidth for each DataNode to balance the workloads of the cluster. We recommend that you set the bandwidth to 100 MB/s. You can also set the dfsadmin -setBalancerBandwidth parameter to adjust the bandwidth. You do not need to restart DataNodes. For example, you can increase the bandwidth when the cluster load is low and decrease the bandwidth when the cluster load is high.
dfs.datanode.balance.max.concurrent.moves	The maximum number of concurrent threads on a DataNode used by the HDFS balancer to move blocks.

How do I submit a Spark job in standalone mode?

You can submit a Spark job only in Spark on YARN mode. The standalone mode is not supported.

What do I do if the Custom Configuration button is not displayed for a service on the EMR console?

1. Log on to the master node of the cluster. For more information, see Log on to a cluster.
2. Go to the following configuration template directory:

   ```
   cd /var/lib/ecm-agent/cache/ecm/service/HUE/4.4.0.3.1/package/templates/
   ```

   Use the `HUE` service as an example.
   - `HUE` is the name of the service directory.
   - `4.4.0.3.1` is the Hue version.
   - `hue.ini` is the configuration file.
3. Run the following command to add the required custom configuration:

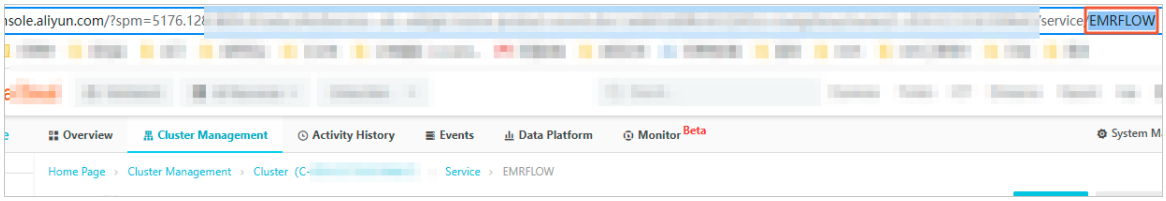
   ```
   vim hue.ini
   ```

   If the configuration item already exists, you can change the value based on the time.
4. In the EMR console, restart the service for the configuration to take effect.

What do I do if a job stays in the SUBMITTING state for a long period of time?

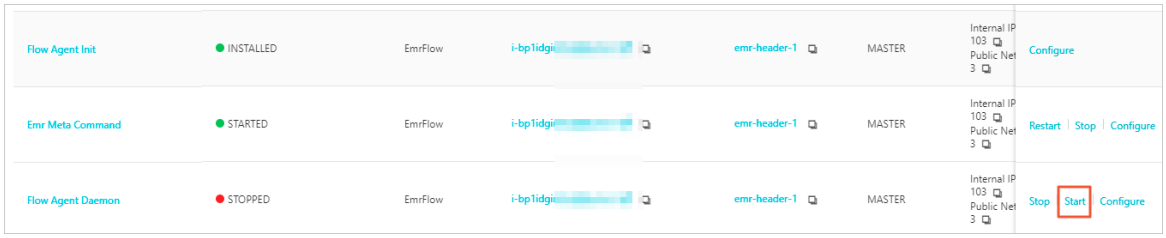
In most cases, this problem occurs because a component in the EMRFLOW service is stopped. You must start the component in the EMR console.

1. Go to the EMRFLOW page.
   i. Go to the page for a service that is deployed in your cluster, replace the service name at the end of the URL in the address bar with EMRFLOW, and then press Enter.

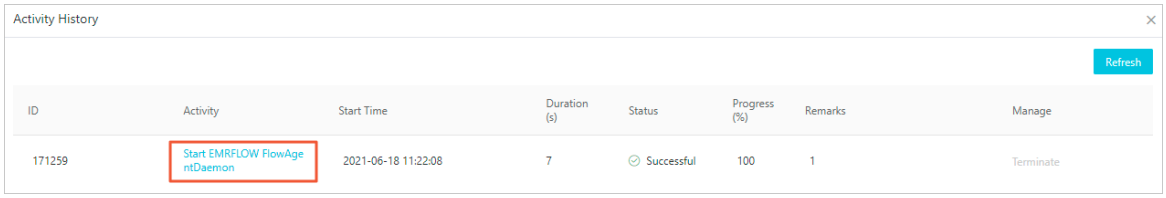
   Note In this example, you are redirected from the HDFS page to the EMRFLOW page.

   ii. Click the **Component Deployment** tab.
2. Start the component that is in the STOPPED state.
   i. On the **Component Deployment** tab, find the component that is in the **STOPPED** state and click **Start** in the Actions column.

   ii. In the **Cluster Activities** dialog box, specify **Description** and click **OK**.
   iii. In the **Confirm** message, click **OK**.
3. Check whether the component is started.
   i. Click **History** in the upper-right corner.

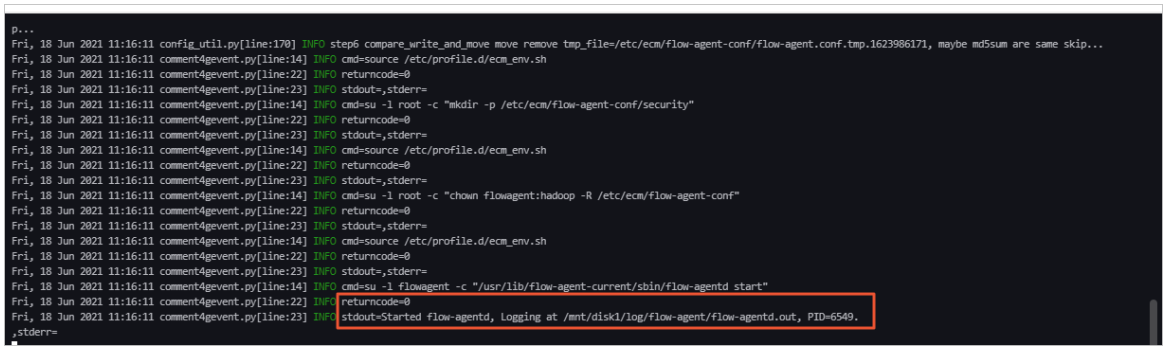
ii. In the **Activity History** dialog box, click **Start EMRFLOW FlowAgentDaemon** in the **Activity** column.



iii. Click **emr-header-1** in the **Instance Name** column.

iv. Click **START_FlowAgentDaemon_ON_emr-header-1** in the **Task Name** column.

v. View the information displayed in the **Task Log** section. If the content in the rectangle in the following figure is displayed, the component is started.



> **Note** If an error is reported after the component is started, fix the error based on the logs. If a permission error is reported, log on to the cluster in SSH mode and run the `sudo chown flowagent:hadoop /mnt/disk1/log/flow-agent/*` command to fix the error. Then, perform the preceding steps again to start the component that is in the STOPPED state.