E-MapReduce

Quick Start

Quick Start

Preparations

Before creating E-MapReduce, you need to complete the following preparations:

Apply for Alibaba Cloud account

Before applying for the E-MapReduce cluster, you need to have an Alibaba Cloud account to identify yourself to the entire Alibaba Cloud ecosystem. This account can be used not only to apply for the E-MapReduce cluster, but also to activate Alibaba Cloud services, such as [Object Storage Service, OSS] (“Object Storage Service, OSS”) and [ApsaraDB for RDS, RDS] (“ApsaraDB for RDS, RDS”).

If you don’t have an Alibaba Cloud account, apply for it by referring to [Register Cloud Account].

Create Access Key

To facilitate E-MapReduce calls and access, you need to create at least one Access Key according to the following steps:

Logon to the official Alibaba Cloud website.

Logon to the console.

Click **AccessKeys**.

Note: If the following prompt dialog box pops up, click to continue using **Accesskey**.

Click **Create Access Key**, and then click **Agree and Create** .

Input the SMS verification code, and click **OK**. Access Key is created successfully.

Activate Alibaba Cloud OSS

E-MapReduce will store your job logs and running logs in the Alibaba Cloud OSS storage space, so you need to activate Alibaba Cloud OSS by referring to the operation steps in [Open OSS] (“Open OSS”).

Create E-MapReduce

Create a cluster, job, and execution plan

Before starting the process explained in this document, confirm you have completed the required preparations.

This tutorial explains the roles of clusters, jobs, and execution plans as well as how they are used in E-MapReduce, and helps you create a Spark Pi job and run it successfully in the cluster. You will also be able to see the approximate calculation result of Pi on the console page.

1. Create a cluster.
    i. On the left side of the console, select **Cluster** and click **Create Cluster** in the top right corner.
    ii. Input the software configuration.
        a. Use the latest EMR product version.
        b. Use the default software configuration.
    iii. Input the hardware configuration.
        a. Select **Pay-As-You-Go**.
        b. If there is no security group, click **New** and enter the security group name.
        c. Select 4-core and 8G for the master node.
        d. Select 4-core and 8G for the core node (one instance).
        e. Keep others in default status.
    iv. Input the basic configuration.
        a. Enter the cluster name.
        b. Select the log path to save job logs, and confirm that the logging feature is on. In the cluster region, create an OSS bucket.
        c. Enter the password.
    v. Create a cluster.
2. Create a job.
    i. On the left side of the console, select **Job** and click **Create Job** in the top right corner.
    ii. Enter the job name.

iii. Select Spark as the job type.
iv. Enter parameters as shown.

```
--class org.apache.spark.examples.SparkPi --master yarn-client --driver-memory 512m --num-executors 1 --executor-memory 1g --executor-cores 2 /opt/apps/spark-1.6.2-bin-hadoop2.7/lib/spark-examples-1.6.2-hadoop2.7.2.jar 10
```

v. Retain other parameters in default status to create the job.

3. Create an execution plan.
   i. When a cluster is created successfully, its status on the list is shown as **Idle**.
   ii. Select **Execution Plan** on the left side of the console and click **Create Execution Plan** in the top right corner.
   iii. Select **Existing Cluster**. Select the newly created cluster and associate it with the execution plan.
   iv. Add the job created earlier to the queue.
   v. Enter the execution plan name.
   vi. Select **Manual Execution** by default.
   vii. Create an execution plan.
4. Run the execution plan.
   i. On the execution plan list page, click **Run Now**.
5. View job logs and confirm the results.
   i. Click **Management** and proceed to the management page. View the **Running Log** at the bottom of the page.
   ii. Click the right side of the running log to view the job list.
   iii. Click **stdout**, and you can see the approximate calculation of Pi: 3.14xxxx.

Create a cluster

Log on to Alibaba Cloud E-MapReduce Console Cluster Page.

Complete RAM authorization. Refer to the Role Authorization for operating steps.

Select the region in which the cluster will be created.

Click **Create a cluster** on the top right.

Cluster creation process

Note: The cluster cannot be modified after creation, except for the name. Please carefully confirm the necessary configurations during creation.

To create a cluster, you need to continue with the following steps:

Input basic information

Input the name of the cluster subject to a length limit of 1-64 characters consisting of only Chinese characters, letters, numbers, “-” and “_”.

Select the payment type. It is subscription mode by default, and you can also select pay-as-you-go mode. The subscription mode will be much cheaper than the pay-as-you-go mode, but the latter is more flexible. If you select subscription mode, you need to select a duration.

The log save function is open by default, and you need to select an OSS log directory as the location to save the job logs in the cluster. If you do not need to save job logs, you can close this function. It is strongly recommended to open the OSS log save function, which will be of great help in debugging your jobs.

Set the login password of the Master node subject to a length limit of 8-30 characters consisting of both uppercase and lowercase letters as well as numbers. After the cluster is enabled, implement SSH login using the root account with the password set.

Bootstrap Action. You can execute your custom script (optional) before Hadoop is enabled in the cluster. Refer to Bootstrap Action for detailed instructions.

Input the software configurations.

Select the product version. A lower version cluster cannot be upgraded to a higher version.

Check the required components. Note that the more components you select, the higher requirements there will be for the configuration of your computer, otherwise there may be insufficient resources to run these services.

Software configuration. You can configure software in the cluster. Refer to Software Configuration for detailed instructions.

Hardware configuration

Select the Zone where the cluster is located. Different zones have different ECS series and disks.

Select the type of network (classic network/VPC). VPC requires additional provision of the VPC and subnet (switcher). If it is not created, you can click Create VPC/Subnet (VSwitch) to go to the VPC console to create it, and then click "Refresh List" to view the newly created VPC/subnet (switch) after creation. Refer to VPC for detailed instructions of E-MapReduce VPC.

Note: The classic network is not interoperable with VPC, and the network type cannot be changed after purchase. In terms of ECS instance series, different zones have different instance series, either series I, II or both.

Select an existing security group. Generally there is no security group when you create the cluster for the first time. Turn on the "New security group" switch, and fill in name of the new security group in the "Security group name".

Select the cluster mode. At present, the default cluster is not in high availability mode, and the Master node number is 1; if the high availability mode is opened, the Master node number will be 2.

Select the Master node configuration. The Master node is configured with 8 MB public network by default, which is subject to the pay-as-you-go mode. This cost is not included in the cluster cost, so it should be subject to additional payment.

Select the Core node configuration to adjust the node number. However, the cluster should be configured with at least 2 Core nodes.

Cluster cost. The cluster cost is displayed at the right corner of the page. For pay-as-you-go clusters, hourly cost will be displayed, and for subscription clusters, the total cost will be displayed.

Confirm creation. After all valid information has been filled in, the "Create" button will be highlighted. Click **Create** to create the cluster after confirmation.

Note:

The cluster will be created immediately if it is a Pay-As-You-Go cluster. It will go back to the Cluster List page, where there is a cluster in the "Creating Cluster" status. Please be patient. It will take several minutes to create a cluster. After completion, the cluster will be switched to the **Idle** status.

For Subscription cluster, the cluster will not be created until the order is generated and paid.

Creation failed

In case of creation failure, the cluster list page shows “Cluster creation failed”. The reason for failure can be seen when the cursor is placed on the red exclamation point.

Create Job

To run a computing task, you need to define a job first according to the following steps:

Log on to Alibaba Cloud E-MapReduce Console Job Page.

Select the region in which the job will be created.

Click **Create Job** at the top right corner of the page to enter the Create Job page.

Fill in the job name.

Select the job type.

Fill in application parameters of the job. The application parameters should include full information of the jar package run by the job, data input and output addresses of the job and some command line parameters, that is, all your parameters in the command line should be filled in this field. If the OSS path is used, you can click “Select OSS Path” to select the OSS resource path. Refer to Job in the user guide for parameter configuration of all job types.

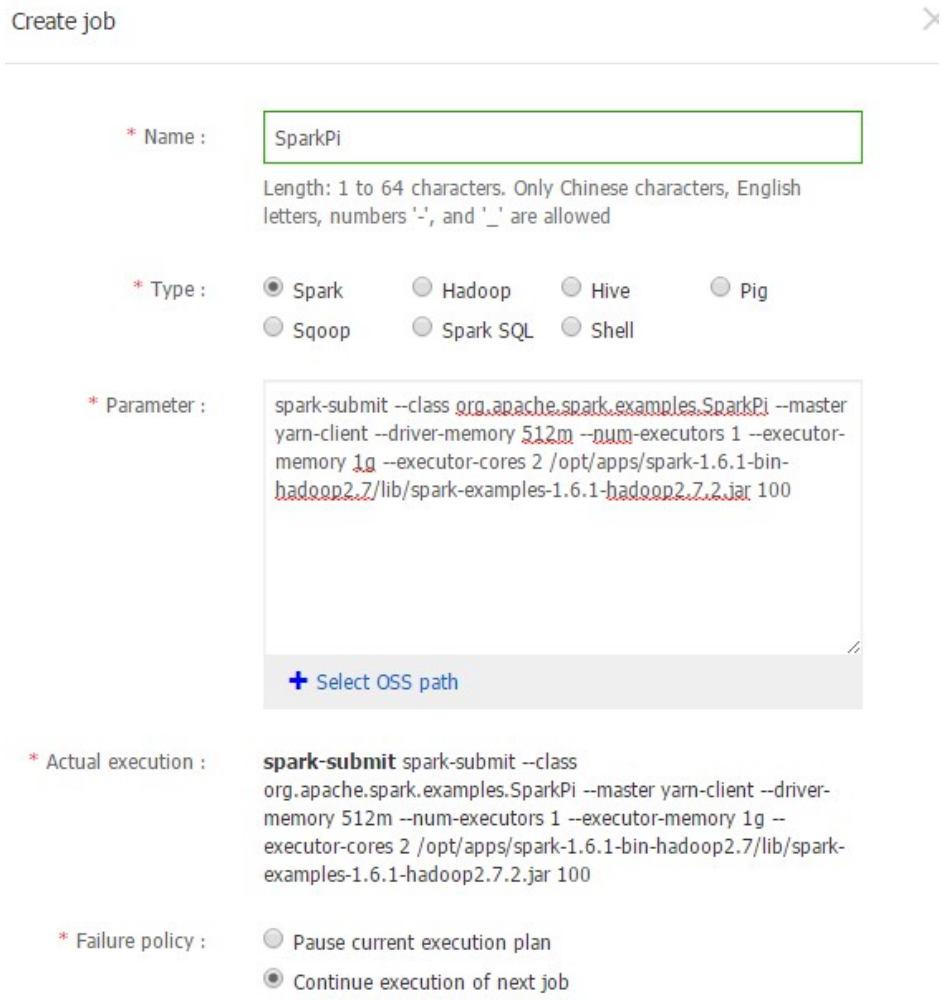
Executed command. The command actually executed for the job on ECS will displayed here. If you copy this command directly, it means the command can be run directly in the command line environment of the E-MapReduce cluster.

Select the policy for failed operations. Pausing the current execution plan will pause the entire execution plan after this job fails to wait for your handling. While continuing to execute the next job will ignore this error and continue to execute the following job after this job fails.

Click **OK** to complete the creation.

Job example

This is a Spark job, for which relevant parameters as well as input and output paths are set in the application parameters.




```
spark-submit --class org.apache.spark.examples.SparkPi --master yarn-client --driver-memory 512m --num-executors 1 --executor-memory 1g --executor-cores 2 /opt/apps/spark-1.6.1-bin-hadoop2.7/lib/spark-examples-1.6.1-hadoop2.7.2.jar 100
```

oss and ossref

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

ossref:// is also directed to an OSS path. But differently, it will be used to download the

corresponding code resource to the local disk, and then replace the path in the command line with the local path. It is used to run some native codes more easily without logging on to the computer to upload the code and the dependent resource package.

Note: The ossref cannot be used to download excessive data resources, otherwise it will lead to failure of the cluster job.

Create an execution plan

After creation of a job, if you want to run the job defined on the cluster, you need to create an execution plan. An execution plan may contain more than one job, and you can also define their order. For example, if one of your scenarios is: prepare data > process data > clean up data, you can define three jobs named "prepare-data" "process-data" and "cleanup-data" respectively, and then create an execution plan to include these three jobs.

The steps to create an execution plan are as follows:

Log on to the Execution Plan page of Alibaba Cloud E-MapReduce Console.

Select the region.

Click **Create an execution plan** at the top right corner to enter the creation page.

There are two options on the Select Cluster Mode page, "Create on Demand" and "Existing Cluster". "Create on Demand" means that you have no clusters at present, intend to run the execution plan in a temporary cluster, and have the temporary cluster automatically released after running the execution plan. While "Existing Cluster" means that you have at least one cluster running currently, and the execution plan should be submitted to and run in the existing cluster.

If "Create on Demand" is selected, you should execute the same steps as creating the cluster, select the configuration of this on-demand cluster, and then click OK.

If "Existing Cluster" is selected, enter the Select Cluster page. You can select the cluster to be associated with the execution plan.

Click **Next** to enter the Configure Job page. This page displays the job list defined previously on the left, and the job list to be run as per the newly created execution plan on the right. Select the jobs on the left to populate the right as per the execution order to

implement definition of the execution plan. You can click the question mark to view the detailed parameters. After completion, click **Next**.

Set the name of the execution plan.

Select the scheduling policy as shown in the figure below.

Periodic scheduling. Define the frequency of the periodic scheduling and the time to start the scheduling.

Manual execution. It can only be executed if you click it manually.

Click **Confirm to submit** to complete the creation of the execution plan.