Alibaba Cloud E-MapReduce

Data Development

Issue: 20190516

Legal disclaimer

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

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

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

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

Generic conventions

Table -1: Style conventions

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

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

Contents

1 Manage a workflow project

After creating an E-MapReduce cluster, you can create workflow projects so that multiple jobs can be run simultaneously or sequentially.

Create a project

1. At the top of the page, click the Data Platform tab to enter the Projects page.

   Under the master account, you can view all ofits projects and RAM useraccounts. RAM users can only view projects if they have development permissions. The granting of project development permissions must be configured in the master account. For more information about authorization, see *User management* below.
2. In the upper-right corner, click New Project. The New Project dialog box is displayed.
3. Enter the project name and description and click Create.

   > Note:
   > You can only create a project with the master account. New Projectis only visible to the administrator of the master account.

User management

After creating a new project, you can grant operational permissions for the project to RAM user accounts.

1. In the Project List page, click View Detailsin the Actions column.
2. Click the User Management tab.
3. Click Add Userto add RAM users to the project under the master account.

   The added RAM users become members of the project and are able to view and develop the jobs and workflows under the project. If you remove aRAM user from a project, click Delete in the Actions column.

   > Note:
   > You can only add project members with the master account. The User Management tab is only visible to the administrator of the master account.

Associate clusters

After creating a new project, you need to associate it with a cluster so that the workflow in the project can run on it.

1. In the Projects page, click View Detailsin the Actions column.
2. Click the Cluster Settings tab.
3. Click Add Cluster. From the drop-down menu, you can select a Subscription or Pay-As-You-Go cluster. (Clusters created by temporary jobs are not listed here.)
4. Click OK.

   To disassociate the cluster, click Delete in the Operation column.

   > Note:
   >
   > You can only associate cluster with the master account. The Cluster Settings tab is only visible to the administrator of the masteraccount.

To set both the queue and user to submit jobs to the cluster, click Modify Configuration in the Operation column. Theconfiguration items are as follows:

- Default Submit Job User: Sets the default Hadoop user who submits the job to the selected cluster in the project. The default value is hadoop. There can only be one default user.
- Default Submit Job Queue: Sets the default queue that the jobs are submitted to in the project. If you leave this blank, the job will be submitted to the default queue.
- Submit Job User Whitelist: Sets Hadoop users who can submit jobs to the selected cluster in the project. If there is more than one user, they can be separated by acomma (,).
- Submit Job Queue Whitelist: Sets the queue of the selected cluster that jobs in the project can run in. If there is more than one queue, they can be separated by a comma (,).
- Client whitelist: Configures the client that can submit jobs. You can select either the E-MapReduce master node or the E-MapReduce gateway. Gateways that you have built are not currently listed here.

2 Job editing

In the project, you can create jobs such as Shell, Hive, Spark, SparkSQL, MapReduce, Sqoop, and Pig.

Create a job

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. Click the Data Platform tab to enter the Projects page.
3. Click Design Workflow to the right of the specified project and to go to the Edit Jobs page.
4. On the left side of the Job Editing page, right-click on the folder you want to operate and select Create Job.
5. In the New Jobdialog box, enter the job name and job description and select a job type.

   The job type cannot be modified once the job has been created.
6. Click OK.

   > Note:
   >
   > You can right click on a folder and then select the corresponding option to perform New Subfolders, Rename Folder, and Delete Folder operations.

Develop a job

For more information about how to develop jobs, see the *Jobs* section of the *E-MapReduce user guide*.

> Note:
>
> When you insert an OSS UNI and select OSSREF as a File Prefix, E-MapReduce will download OSS files to your cluster and add these files to the classpath.

- Basic job settings

  In the top-right corner, click Configure Jobs, and then the Job Settings dialog box appears.

  - Failed Retries: Sets the number of retries when this job fails during the workflow running. This option will not take effect when you run the job directly on the Job Editing page.
  - Failure Policy: Sets whether to continue running the next job or suspend the current workflow when this job fails during the workflow running.
  - Add Running Resources: If you add resources such as Jar packages or UDFs that the job running depends on, you need to upload the resources to OSS first. When you select a resource, you can use this resource in a job directly.
  - Parameter Configuration: specifies the values of variables used in a job. You can use variables in your code. The format is: `${variable name}`. Click the plus (+) icon on the right side to add key-value pairs. Key is the name of a variable and value is the value of a variable. In addition, you can follow these rules to customize time variables according to the start time of a schedule.

    - yyyy represents a 4-digit year.
    - MM represents a 2-digit month.
    - dd represents a 2-digit day.
    - HH indicates that the 24-hour clock is used. hh indicates that the 12-hour clock is used.
    - mm represents a 2-digit minute.
    - ss represents a 2-digit second.

      A time variable consists of the combination of a 4-digit year and one or more other time formats. In addition, you can use plus (+) and minus (-) to add or reduce a period of time for the current time. For example, the `${yyyy-MM-dd}` variable represents the current date.

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

> **Notice:**
>
> The parameter of a time variable is required to 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>): Date converts a given parameter to a Date object. A parameter name represents the variable (key) name set in the Parameter Configuration area. A time format represents the time format used by the variable name. For example, if the parameter name 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:
>
> - To retrieve the hour literal value of the current_time variable: ${ formatDate(parseDate(current_time, 'yyyyMMddHHmmss'), 'HH')}
> - To retrieve the year literal value of the current_time variable: ${ formatDate(parseDate(current_time, 'yyyyMMddHHmmss'), 'yyyy')}

- Advanced job settings

  In the Job Settings dialog box, click the Advanced tab.

  - Mode: Submit on Worker Node and Submit on Header/Gateway Node. In the Submit on Worker Node mode, jobs are submitted to resources that are allocated by YARN as launchers. In the Submit on Header/Gateway Node mode, jobs are running on allocated nodes as processes.

- Environment Variable: The environment variables for running a job. You can also export the environment variables from a job script.
- Scheduling Parameters: YARN scheduler, CPU and memory specifications, and Hadoop users. Default values of a Hadoop cluster are applied if you do not specify the parameter values.

Execute a job

After the development and configuration of a job are complete, you can click Run in the top-right corner to run the job.

View logs

After a job runs, you can click the View Records tab to view the running logs of the job. Click View Details to go to the details page. On this page, you can view details, including the job submission log and YARN Container log.

3 Ad hoc queries

You can only select HiveSQL, SparkSQL, and Shell as the type of an ad hoc query. When you execute an ad hoc query statement, the log and query results show at the bottom of the Query page.

Create a job

When you execute a job on the Edit Jobs page and click Details, you will be directed to the Details page that shows the operation logs and run logs of this job. Ad hoc queries and jobs are used in different places. Ad hoc queries are usually used by data scientists and data analysts. In addition, you need to use SQL as a tool to implement an ad hoc query.

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. Click the Data Platform tab to enter the Projects page.
3. Click Design Workflow on the right side of the associated project to enter the Edit Jobs page.
4. In the left-side navigation pane, click the Query tab to enter the Query page.
5. In the left-side navigation pane, right-click a folder as required and select New Job.
6. In the New Jobdialog box, enter the job name and job description and select a job type.

   The job type cannot be modified once the job has been created.
7. ClickOK.

   Note:
   You can right click on a folder and then select the corresponding option to perform New Subfolders, Rename Folder, and Delete Folder operations.

Develop a job

For more information about how to develop jobs with HiveSQL, SparkSQL, and Shell types, see the *jobs* section of E-MapReduce user guide.

Note:
When you insert an OSS UNI and select OSSREF as a File Prefix, E-MapReduce will download OSS files to your cluster and add these files to the classpath.

- Basic job settings

  In the top-right corner, click Configure Jobs, and then the Job Settings dialog box appears.

  - Resource File: If you want to add resources such as jar packages or UDF that a job execution depends on, you must upload these files to OSS. When you select a resource, you can use this resource in a job directly.
  - Parameter Configuration: specifies the values of variables used in a job. You can use variables in your code. The format is: `${variable name}`. Click the plus (+) icon on the right side to add key-value pairs. Key is the name of a variable and value is the value of a variable. In addition, you can follow these rules to customize time variables according to the start time of a schedule.

    - yyyy represents a 4-digit year.
    - MM represents a 2-digit month.
    - dd represents a 2-digit day.
    - HH indicates that the 24-hour clock is used. hh indicates that the 12-hour clock is used.
    - mm represent a 2-digit minute.
    - ss represents a 2-digit second.

    A time variable consists of the combination of a 4-digit year and one or more other time formats. In addition, you can use plus (+) and minus (-) to add or reduce a period of time for the current time. For example, the `${yyyy-MM-dd}` variable represents the current date.

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

    Notice:
    The parameter of a time variable is required to 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 represents the variable (key) name set in the Parameter Configuration area. A time format represents the time format used by the variable name . For example, if the parameter name 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:

- To retrieve the hour literal value of the current_time variable: ${ formatDate(parseDate(current_time, 'yyyyMMddHHmmss'), 'HH')}
- To retrieve the year literal value of the current_time variable: ${ formatDate(parseDate(current_time, 'yyyyMMddHHmmss'), 'yyyy')}

- Advanced job settings

  In the Job Settings dialog box, click the Advanced tab.

  - Mode: Submit on Worker Node and Submit on Header/Gateway Node. In the Submit on Worker Node mode, jobs are submitted to resources that are allocated by YARN as launchers. In the Submit on Header/Gateway Node mode, jobs are running on allocated nodes as processes.
  - Scheduling Parameters: YARN scheduler, CPU and memory specifications, and Hadoop users. Default values of a Hadoop cluster are applied if you do not specify the parameter values.

Execute a job

After the development and configuration of a job are complete, you can click Run in the top-right corner to run the job.

View logs

After you execute a job, you can view run logs on theLog tab at the bottom of the Query page.

4 Manage a workflow

E-MapReduce workflows support the parallelexecution of big data jobs based on DAG. You can also suspend, stop, rerun workflows, and view their running statuses in the webUI.

Create a workflow

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. At the top of the page, click the Data Platform tab.
3. Click Design Workflownext to the target project in the Actions column. Then select the Design Workflow tab.
4. On the left side, right-click the folder you want to operate on and select New Workflow.
5. In the New Workflow dialog box, enter the workflow name and description, and select the E-MapReduce cluster where you want to run the workflow.

   You can select a Subscription or Pay-As-You-Go E-MapReduce cluster that has been created and associated with the project. Alternatively, you can create a new temporary cluster using the cluster template.
6. Click OK.

Edita workflow

You can drag different types of jobs to the workflow editing canvas and specify the order of job instances by curve. After the jobhas been dragged, drag the END component from the control node area to the canvas. Thisindicates that the entire workflow is complete.

Configure a workflow

On the right of the Workflow Design page, click Configureto configure the workflow scheduling.

- Run In: The E-MapReduce cluster where the workflow is to run can be modified.

- Scheduling Policy: After workflow scheduling has been enabled, period schedule are mandatory by default, and dependency schedule can be added.
    - Time Scheduler: Sets the start and end timesfor the workflow scheduling. The system then runs the workflow according to the schedule you set.
    - Dependency: Select the dependency workflow of the current workflow from the selected project. After the dependency workflow has been completed, the current workflow is scheduled to run. Currently, only one workflow can be selected.

Run a workflow

Once a workflow has been developed and configured, you can click Run in the top right corner to run the workflow.

View and operate workflow instances

After the workflow is running, click the View Records tab on the left to view the running status of the workflow instance. Click View Detailsnext to the workflow instance to view the running status of the job instance. You can also suspend, resume, stop, and rerun workflow instances.

- Suspend workflow instance: The job instance continues to run, but subsequent instances do not. By clickingResume Workflow, the system continues to run the subsequent jobs.
- Stop workflow instance: All running job instances stop immediately.
- Rerun workflow instance: The system runs the workflow from the start component .

5 Jobs

5.1 Job operations

You can create, clone, modify, and delete jobs.

Job creation

A new job can be created at any time. Currently, a job can only be used in the region where it is created.

Job cloning

Configurations that already exist for a job can be cloned. A cloned job can also only be used in the region where it is created.

Job modification

Before you can modify a job that needs to be added to an execution plan, you must first ensure that the execution plan is not running and that its periodic scheduling is not in progress.

Before you can modify a job that needs to be added to several execution plans, you must first ensure that none of the execution plans are running and that none of their periodic scheduling is in progress. Modifying a job may result in changes to all of the execution plans that use this job.

If you need to debug, we recommend that you perform cloning instead. After you debug, the original jobs in the execution plan are replaced.

Job deletion

As with modification, a job can only be deleted when the execution plan where the job is located is not running and its periodic scheduling is not in progress.

5.2 Time and date variables

When you are creating a job, variable wildcards are supported in the job parameters for both time and date.

Variable wildcard format

The format of the variable wildcards supported by E-MapReduce is either `${ dateexpr - 1d }` or `${ dateexpr - 1h }`. For example, assuming the current date and time is `2016 / 04 / 27 12 : 08 : 01 :`

- If `${ yyyyMMdd HH : mm : ss - 1d }` is displayed, the parameter wildcard is replaced with `20160426 12 : 08 : 01` when executed, which is the current date minus one day, and time accurate to the second.
- If `${ yyyyMMdd - 1d }` is displayed, the parameter wildcard is replaced with `20160426` when executed, which is the current date minus one day.
- If `${ yyyyMMdd }` is displayed, the parameter wildcard is replaced with `20160427` , which is the current date.

dateexpr represents the standard format of expressing time. Time is therefore formatted according to this expression and is followed by the amount of time that you want to add or deduct, which can be written as N. For example, `${ yyyyMMdd - 5d }`, `${ yyyyMMdd + 5d }`, `${ yyyyMMdd + 5h }`, or `${ yyyyMMdd - 5h }`.

Note:

E-MapReduce currently supports the addition and deduction of hours and days only.

Example

1. Click Job Settings on the top right of the Edit Jobs page.
2. Click the add icon to add new parameters on the Parameter Configuration part, and fill in the parameter according to the Variable wildcard format that mentioned above.
3. You can now use the reference of the parameter key in the job editing.

5.3 Configure a Hive job

When you apply for clusters in E-MapReduce, you are provided with a Hive environment by default. Using Hive, you can create and operate tables and data.

Procedure

1. Prepare the Hive script in advance. For example:

```
USE   DEFAULT ;
 DROP   TABLE   uservisits ;
 CREATE   EXTERNAL   TABLE   IF   NOT   EXISTS   uservisits
( sourceIP   STRING , destURL   STRING , visitDate   STRING ,
adRevenue   DOUBLE , user
 Agent   STRING , countryCod  e   STRING , languageCo  de   STRING
, searchWord   STRING , duration   INT  )  ROW   FORMAT   DELIMITED
  FIELDS   TERMI
 NATED   BY  ','  STORED   AS   SEQUENCEFI  LE   LOCATION  '/
HiBench / Aggregatio  n / Input / uservisits ';
 DROP   TABLE   uservisits  _aggre ;
 CREATE   EXTERNAL   TABLE   IF   NOT   EXISTS   uservisits  _aggre
 (  sourceIP   STRING ,  sumAdReven  ue   DOUBLE )  STORED   AS
SEQUENCEFI  LE   LO
 CATION  '/ HiBench / Aggregatio  n / Output / uservisits  _aggre
';
 INSERT   OVERWRITE   TABLE   uservisits  _aggre   SELECT
sourceIP ,  SUM ( adRevenue )  FROM   uservisits   GROUP   BY
sourceIP ;
```

2. Save this script into a script file, such as `uservisits_aggre_hdfs.hive`, and upload it to an OSS directory (for example, `oss://path/to/uservisits_aggre_hdfs.hive`).
3. Log on to the *Alibaba Cloud E-MapReduce console*.
4. At the top of the navigation bar, click Data Platform.
5. In the Actions column, click Design Workflow next to the specified project.
6. On the left of the Job Editing page, right-click the folder you want to operate and select New Job.
7. In the New Job dialog box, enter the job name and description.
8. Select the Hive job type to create a Hive job. This type of job is submitted in the background using the following method.

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

9. Click OK.

Note:

> You can also create subfolders, rename folders, and delete folders by right-clicking on them.

10. Enter the parameters in the Content field after the Hive commands. For example, if you want to use a Hive script uploaded to OSS, enter the following.

    ```
    - f   ossref :// path / to / uservisits  _aggre_hdf  s . hive
    ```

    You can also click Select OSS path to view and select from OSS. The system will automatically complete the path of the Hive script on OSS. Switch the Hive script prefix to ossref by clicking Switch resource type. This ensures that the file is correctly downloaded by E-MapReduce.

11. Click Save to complete the Hive job configuration.

5.4 Configure a Pig job

When you apply for clusters in E-MapReduce, a Pig environment is provided by default. Using Pig, you can create and operate tables and data.

Procedure

1. Prepare the Pig script in advance. For example:

~~~
``` shell
 /*
 *  Licensed   to   the   Apache   Software   Foundation   ( ASF )
 under   one
 *  or   more   contributo  r   license   agreements .   See   the
 NOTICE   file
 *  distribute  d   with   this   work   for   additional
 informatio  n
 *  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
 *  distribute  d   under   the   License   is   distribute  d   on
   an  " AS   IS "  BASIS ,
 *  WITHOUT   WARRANTIES   OR   CONDITIONS   OF   ANY   KIND ,
 either   express   or   implied .
 *  See   the   License   for   the   specific   language
 governing   permission  s   and
 *  limitation  s   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   NonURLDete  ctor   UDF   to   remove   records
if   the   query   field   is   empty   or   a   URL .
 clean1  =  FILTER   raw   BY   org . apache . pig . tutorial .
NonURLDete  ctor ( 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
YYMMDDHHMM  SS .
--  Call   the   ExtractHou  r   UDF   to   extract   the   hour  (
HH )  from   the   time   field .
 houred  =  FOREACH   clean2   GENERATE   user ,  org . apache .
pig . tutorial . ExtractHou  r ( time )  as   hour ,  query ;
--  Call   the   NGramGener  ator   UDF   to   compose   the   n -
grams   of   the   query .
 ngramed1  =  FOREACH   houred   GENERATE   user ,  hour ,  flatten
( org . apache . pig . tutorial . NGramGener  ator ( 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_frequ  ency1  =  GROUP   ngramed2   BY  ( ngram ,  hour );
--  Use   the   COUNT   function   to   get   the   count  (
occurrence  s )  of   each   n - gram .
 hour_frequ  ency2  =  FOREACH   hour_frequ  ency1   GENERATE
flatten ($ 0 ),  COUNT ($ 1 )  as   count ;
--  Use   the   GROUP   command   to   group   records   by   n -
gram   only .
--  Each   group   now   correspond  s   to   a   distinct   n -
gram   and   has   the   count   for   each   hour .
 uniq_frequ  ency1  =  GROUP   hour_frequ  ency2   BY   group ::
ngram ;
--  For   each   group ,  identify   the   hour   in   which
this   n - gram   is   used   with   a   particular  ly   high
frequency .
--  Call   the   ScoreGener  ator   UDF   to   calculate   a  "
popularity "  score   for   the   n - gram .
 uniq_frequ  ency2  =  FOREACH   uniq_frequ  ency1   GENERATE
  flatten ($ 0 ),  flatten ( org . apache . pig . tutorial .
ScoreGener  ator ($ 1 ));
--  Use   the   FOREACH - GENERATE   command   to   assign   names
  to   the   fields .
```

~~~
 uniq_frequ  ency3  =  FOREACH   uniq_frequ  ency2   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_u niq_freque  ncy  =  FILTER   uniq_frequ  ency3   BY
score  >  2 . 0 ;
--  Use   the   ORDER   command   to   sort   the   remaining
records   by   hour   and   score .
 ordered_un  iq_frequen  cy  =  ORDER   filtered_u  niq_freque  ncy
  BY   hour ,  score ;
--  Use   the   PigStorage   function   to   store   the   results
.
--  Output : ( hour , n - gram , score , count , average_co
unts_among  _all_hours )
 STORE   ordered_un  iq_frequen  cy   INTO  ' oss :// emr /
checklist / data / chengtao / pig / script1 - hadoop - results '
USING   PigStorage ();
```
~~~

2. Save this script into a script file, such as `script1 - hadoop - oss . pig` , and upload it to an OSS directory (for example, `oss :// path / to / script1 - hadoop - oss . pig`).
3. Log on to the *Alibaba Cloud E-MapReduce console*.
4. At the top of the navigation bar, click Data Platform.
5. In the Actions column, click Design Workflow next to the specified project.
6. On the left of the Job Editing page, right-click the folder you want to operate and select New Job.
7. In the New Job dialog box, enter the job name and description.
8. Select the Pig job type to create a Pig job. This type of job is submitted in the background using the following method.

```
pig  [ user   provided   parameters ]
```

9. Click OK.

> Note:
>
> You can also create subfolders, rename folders, and delete folders by right-clicking on them.

10. Enter the parameters in the Content field after the Pig commands. For example, if you want to use a Pig script uploaded to OSS, enter the following.

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

You can click Select OSS path to view and select from OSS. The system will automatically complete the path of Pig script on OSS. Switch the Pig script prefix to ossref by clicking Switch resource type. This ensures that the file is correctly downloaded by E-MapReduce.

11.Click Save to complete the Pig job configuration.

5.5 Configure a Spark job

In this tutorial, you will learn how to configure a Spark job.

Procedure

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. At the top of the navigation bar, click Data Platform.
3. In the Actions column, click Design Workflow next to the specified project.
4. On the left of the Job Editing page, right-click the folder you want to operate and select New Job.
5. In the New Job dialog box, enter the job name and description.
6. Click OK.

   Note:

   You can also create subfolders, rename folders, and delete folders by right-clicking on them.

7. Select the Spark job type to create a Spark job. This type of job is submitted in the background using the following method.

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

8. Enter the parameters in the Content field that are required to submit this job. Only the parameters after `spark - submit` can be entered. The following example shows how to enter the parameters for creating a Spark job and a PySpark job.

   - Create a Spark job

     Create a Spark WordCount job:

     - Job name: WordCount
     - Type: Select Spark
     - Parameters:

       ■ Enter the following command:

       ```
       spark - submit  -- master   yarn - client  -- driver -
       memory   7G  -- executor - memory   5G  -- executor -
       cores   1  -- num - executors   32  -- class   com . aliyun
       . emr . checklist . benchmark . SparkWordC  ount   emr
       - checklist_  2 . 10 - 0 . 1 . 0 . jar   oss :// emr /
       checklist / data / wc   oss :// emr / checklist / data / wc
       - counts   32
       ```

       ■ Enter the following in the E-MapReduce job Content field:

       ```
       -- master   yarn - client  -- driver - memory   7G  --
        executor - memory   5G  -- executor - cores   1  -- num -
        executors   32  -- class   com . aliyun . emr . checklist .
        benchmark . SparkWordC  ount   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:

       Job jar packages are saved in OSS. In the example above, the way to reference the Jar package is *ossref :// emr / checklist / jars / emr - checklist_ 2 . 10 - 0 . 1 . 0 . jar* . Click Select OSS path to view and select one from OSS. The system will automatically complete the absolute path of the Spark script on OSS. Switch the default OSS protocol to the ossref protocol.

- Create a PySpark job

  In addition to Scala and Java job types, E-MapReduce also supports Python job types in Spark. Create a Spark K-means job for the Python script:

  - Job name: Python-Kmeans
  - Type: Spark
  - Parameters:

    ```
    -- master   yarn - client  -- driver - memory   7g  -- num -
     executors   10  -- executor - memory   5g  -- executor - cores
       1  -- jars   ossref :// emr / checklist / jars / emr - core
     - 0 . 1 . 0 . jar   ossref :// emr / checklist / python /
     wordcount . py   oss :// emr / checklist / data / kddb   5
     32
    ```

  - References of Python script resources are supported, and the ossref protocol is used.
  - For PySpark, the online Python installation kit is not supported.

9. Click Save to complete the Spark job configuration.

5.6 Configure a Spark SQL

In this tutorial, you will learn how to configure a Spark SQL job.

> Note:
> By default, the mode of Spark SQL used for submitting a job is YARN.

Procedure

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. At the top of the navigation bar, click Data Platform.
3. In the Actions column, click Design Workflow next to the specified project.
4. On the left of the Job Editing page, right-click the folder you want to operate and select New Job.
5. In the New Job dialog box, enter the job name and description.
6. Click OK.

   > Note:

> You can also create subfolders, rename folders, and delete folders by right-clicking on them.

7. Select the Spark SQL job type to create a Spark SQL job. This type of job is submitted in the background using the following method.

```
spark - sql  [ options ] [ cli   option ]
```

8. Enter the parameters in the Content field after the Spark SQL commands.

   - -e option

     -e options can be written to the running SQL by inputting them into the Content field of the job. For example:

     ```
     - e  " show   databases ;"
     ```

   - -f option

     -f options can be used to specify a Spark SQL script file. Uploading well-prepared Spark SQL script files to OSS can provide greater flexibility. We recommend that you use this operation mode. For example:

     ```
     - f   ossref :// your - bucket / your - spark - sql - script . sql
     ```

9. Click Save to complete Spark SQL job configuration.

5.7 Configure a Shell job

In this tutorial, you will learn how to configure a Shell job.

> Notice:
> By default, Shell scripts are currently run by Hadoop. If you need to use the root user, the `sudo` command can be used. Use Shell script jobs with caution.

Procedure

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. At the top of the navigation bar, click Data Platform.
3. In the Projects area, select a target project ID to go to the Project Management tab page.
4. In the left-side navigation bar, click Edit Jobs next to the specified project.
5. On the left of the Edit Jobs tab page, right-click the folder you want to operate and select New Job.

6. In the New Job dialog box, enter the job name and description.
7. Select the Shell job type to create a Bash Shell job.
8. Click OK.

   Note:

   You can also create subfolders, rename folders, and delete folders by right-clicking on them.

9. Enter the parameters in the Content field after the Shell commands.

   - -c option

     -c options can be used to set Shell scripts to run by inputting them into the Content field of the job. For example:

     ```
     - c  " echo  1 ; sleep  2 ; echo  2 ; sleep  4 ; echo  3
      ; sleep  8 ; echo  4 ; sleep  16 ; echo  5 ; sleep  32
      ; echo  6 ; sleep  64 ; echo  8 ; sleep  128 ; echo
      finished "
     ```

   - -f option

     -f options can be used to run Shell script files. By uploading a Shell script file to OSS, Shell scripts on OSS can be defined in the job parameters, making it more flexible than the -c option. For example:

     ```
     - f  ossref :// mxbucket / sample / sample - shell - job . sh
     ```

10. Click Save to complete Shell job configurations.

5.8 Configure a Sqoop job

In this tutorial, you will learn how to configure a Sqoop job.

Note:

Only E-MapReduce products with version V1.3.0 or later support the Sqoop job type. Running a Sqoop job on lower versions will fail and errlog will report "Not supported" errors. For more information on parameters, see *Sqoop*.

Procedure

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. At the top of the navigation bar, click Data Platform.
3. In the Actions column, click Design Workflow next to the specified project.

4. On the left of the Job Editing page, right-click the folder you want to operate and select New Job.
5. In the New Job dialog box, enter the job name and description.
6. Select the Sqoop job type to create a Sqoop job. This type of job is submitted in the background using the following method.

   ```
   sqoop  [ args ]
   ```

7. Click OK.

   > **Note:**
   > You can also create subfolders, rename folders, and delete folders by right-clicking on them.

8. Enter the parameters in the Content field after the Sqoop commands.
9. Click Save to complete Sqoop job configuration.

6 Old EMR Scheduling (Soon will be unavailable)

6.1 Notebooks

6.1.1 Introduction

Notebooks allow you to compile and run Spark, Spark SQL, and Hive SQL tasks directly on the E-MapReduce console. You can then view the running results in the notebook. Notebooks are ideal for processing debugging tasks that require a shorter runtime and whose results need to be viewed directly. For tasks that have a longer runtime and require regular execution, the job and execution plan function must be used. This section describes how to create and run a notebook demo task.

Create a demo task

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. At the top of the navigation bar, click Old EMR Scheduling.
3. In the navigation bar on the left, click Notebook.
4. Click New notebook demo.

5. A confirmation box is displayed, indicating the required cluster environment. Click OK to create a demo task. Three examples of interactive tasks are created.

Run a Spark demo task

1. Click EMR-Spark-Demo to display the example of a Spark notebook. Before running the notebook, you need to associate the task to a created cluster. Select a created cluster in the list of available clusters. Note that the associated cluster must be E-MapReduce 2.3 or later and have no less than three nodes, each with at least 4 cores and 8 GB of memory.

2. After a cluster is associated, click Run. When the associated cluster executes the Spark or Spark SQL notebook for the first time, it takes about one minute to build the Spark context and running environment. It does not need to be built in subsequent executions. The running result is displayed under the Run button.

Run a SparkSQL demo task

1. Click EMR-Spark-Demo to display the SparkSQL notebook example. Before running the notebook, you need to associate it to a created cluster. In the upper-right corner, select a created cluster from the list of available clusters.
2. The SparkSQL demo contains several demo sections that can be run individually or together by clicking Run All. After running, you can see the returned data results of each section.

   Note:

   If the section for creating a table is run multiple times, an error is reported indicating that the table already exists.

Run a Hive demo task

1. Click EMR-Hive-Demo to display the Hive notebook example. Before running the notebook, you need to associate it to a created cluster. In the upper-right corner, select a created cluster from the list of available clusters.
2. The Hive demo task contains several demo sections that can be run individually or together by clicking Run All. After running, you can see the returned data results of each section.

   Note:

   - When the associated cluster executes the Hive notebook for the first time, it takes a few seconds to build the Hive client running environment. It does not need to be built in subsequent executions.
   - If the section for creating a table is run multiple times, an error is reported indicating that the table already exists.

Cancel the association with clusters

After a notebook is run in a cluster, the cluster creates a process for caching some context running environments in order to ensure a quick response upon re-execution. If you do not need to execute other notebooks, and you want to release the cluster resources occupied by caching, you can disassociate all interactive tasks that have been run from the associated clusters. In this way, you can release the memory resources occupied on the original associated clusters.

6.1.2 Operations

This section details how to perform a number of notebook operations, including how to create a new notebook task on the E-MapReduce console.

Create a new notebook task

> Note:
>
> The cluster on which an interactive task is run must be E-MapReduce 2.3 or later and have no less than three nodes, each with at least 4 cores and 8 GB of memory.

1. Log on to the *Alibaba Cloud E-MapReduce console*.
2. At the top of the navigation bar, click Old EMR Scheduling.
3. In the navigation bar on the left, click Notebook.
4. Click New notebook or File > -New notebook.
5. Enter a name and select the default type. Associating a cluster is optional. Click OK to create a notebook.

   Three types of notebook task are supported. Spark can be used to write Scala code, Spark SQL can be used to write SQL statements supported by Spark, and Hive can be used to write SQL statements supported by Hive.
6. An associated cluster must be E-MapReduce 2.3 or later and have no less than three nodes, each with at least 4 cores and 8 GB of memory. You can also associate the cluster before running the task.

   Up to 20 interactive tasks can be created in one account.

Enter and save a section

A paragraph is the smallest unit for running a notebook. Multiple paragraphs can be entered into a notebook. Each paragraph starts with either `% spark` , `% sql` , or `% hive` , indicating whether it is a Scala code paragraph, Spark SQL paragraph, or Hive SQL paragraph. The type prefix is separated by a blank space or by line feed and actual content. If the type prefix is not specified, the default type of the interactive task is used as the run type of this paragraph.

The following example shows how to create a temporary Spark table:

Paste the following code into the section and a red * symbol is displayed, indicating that this notebook has been changed. Click Save Paragraph or run to save the modifications to the paragraph. Click + under the paragraph to create a new paragraph. Up to 30 paragraphs can be created in one notebook.

```
% spark
 import   org . apache . commons . io . IOUtils
 import   java . net . URL
 import   java . nio . charset . Charset
//  load   bank   data
 val   bankText  =  sc . paralleliz  e (
     IOUtils . toString (
         new   URL (" http :// emr - sample - projects . oss - cn -
 hangzhou . aliyuncs . com / bank . csv "),
         Charset . forName (" utf8 ")). split ("\ n "))
 case   class   Bank ( age :  Integer ,  job :  String ,  marital :
 String ,  education :  String ,  balance :  Integer )
 val   bank  =  bankText . map ( s  =>  s . split (";")). filter ( s
 =>  s ( 0 ) ! = "\" age \""). map (
     s  =>  Bank ( s ( 0 ). toInt ,
             s ( 1 ). replaceAll ("\"", ""),
             s ( 2 ). replaceAll ("\"", ""),
             s ( 3 ). replaceAll ("\"", ""),
             s ( 5 ). replaceAll ("\"", ""). toInt
         )
). toDF ()
 bank . registerTe  mpTable (" bank ")
```

Run a paragraph

Before running a notebook, you must first associate it to a created cluster. If a created notebook is not associated with a cluster, Not Attached is displayed in the upper-right corner of the page. Click it to select a cluster from the list of available clusters. Note that the associated cluster must be E-MapReduce 2.3 or later and have no less than three nodes, each with at least 4 cores and 8 GB of memory.

Click Run to save the current paragraph and run the content. If this is the last paragraph, a new paragraph is created automatically.

PENDING indicates that the paragraph has not run yet, RUNNING indicates that the paragraph is running, FINISHED indicates that the running has finished, and ERROR indicates that an error has occurred. The running result is displayed beneath the Run button. During running, you can click Cancel beneath the Run button to cancel running. ABORT is displayed after running has been canceled.

The paragraph can be run multiple times, but only the result of the last running is retained. You cannot modify the content of a paragraph while it is running. It can only be modified after the running has finished.

Run all

For a notebook, you can click Run All on the menu bar to run all paragraphs. The paragraphs are then submitted sequentially for running. Different types have independent execution queues. If a notebook contains multiple paragraph types, the order for executing them on the cluster is decided based on type after they have been submitted sequentially. Spark and Spark SQL support one-by-one execution. Hive supports concurrent execution, with the maximum number of concurrently executed interactive paragraphs on the same cluster is 10. Note that all concurrently executed paragraphs are restricted by cluster resources. If the cluster size is small and many paragraphs need to be executed concurrently, the paragraphs still need to queue in YARN.

Cancel the association with clusters

After a notebook is run in a cluster, the cluster creates a process for caching some context running environments to ensure a quick response upon re-execution. If you do not need to run other notebooks, and you want to release the cluster resources occupied by caching, you can disassociate all notebooks that have been run from the associated clusters. In this way, you can release the memory resources occupied on the original associated clusters.

Other operations

- Paragraph operations

  - Hide and display the results

    Hide the paragraph results and only display the entered content of the paragraph.
  - Delete a paragraph

    Delete the current paragraph. Paragraphs that are running can also be deleted.
- File menu

  - New notebook

    Create a notebook and switch to the created notebook interface.
  - Create Paragraph

    Add a new paragraph to the end of a notebook. A notebook can have up to 30 paragraphs.
  - Save all paragraphs

    Save all modified paragraphs.
  - Delete notebook

    Delete the current notebook. If a cluster has been associated, it will be disassociated.
- View

  Only display codes or display codes and results.

6.1.3 Examples

6.1.3.1 Query bank employee information

1. Create a temporary table

```
% spark
 import   org . apache . commons . io . IOUtils
 import   java . net . URL
 import   java . nio . charset . Charset
//  Zeppelin   creates   and   injects   sc  ( SparkConte  xt )  and
 sqlContext  ( HiveContex  t   or   SqlContext )
//  So   you   don ' t   need   create   them   manually
//  load   bank   data
```

```
 val   bankText  =  sc . paralleliz  e (
     IOUtils . toString (
         new   URL (" http :// emr - sample - projects . oss - cn -
 hangzhou . aliyuncs . com / bank . csv "),
         Charset . forName (" utf8 ")). split ("\ n "))
 case   class   Bank ( age :  Integer ,  job :  String ,  marital :
 String ,  education :  String ,  balance :  Integer )
 val   bank  =  bankText . map ( s  =>  s . split (";")). filter ( s
 =>  s ( 0 ) ! = "\" age \""). map (
     s  =>  Bank ( s ( 0 ). toInt ,
             s ( 1 ). replaceAll ("\"", ""),
             s ( 2 ). replaceAll ("\"", ""),
             s ( 3 ). replaceAll ("\"", ""),
             s ( 5 ). replaceAll ("\"", ""). toInt
         )
). toDF ()
 bank . registerTe  mpTable (" bank ")
```

2. Query the table structure

```
% sql
 desc   bank
```

3. Query the number of employees in each age group under 30

```
% sql   select   age ,  count ( 1 )  value   from   bank   where   age
  <  30   group   by   age   order   by   age
```

4. Query the information of employees younger than or equal to 20

```
% sql   select  *  from   bank   where   age  <=  20
```

6.1.3.2 Video playback data

Preparations

In this example, you need to download data from OSS and upload it to your OSS bucket. This data includes:

- *User table sample data*
- *Video table sample data*
- *Video playback table sample data*

Upload this sample data respectively to the specified UserInfo, Videoinfo, and Playvideo on your OSS bucket. For example, upload the data to the Demo or UserInfo directory under Bucket Example.

In the following table, replace the SQL [bucketname] with your bucket name, replace [region] with your OSS region name, and replace [bucketpath] with your specified OSS path prefix, such as Demo.

1. Create a user table

```
% hive
 CREATE   EXTERNAL   TABLE   user_info ( id   int , sex   int , age
 int ,  marital_st  atus   int )  ROW   FORMAT   DELIMITED   FIELDS
 TERMINATED   BY  ','  LOCATION  ' oss ://[ bucketname ]. oss - cn -[
 region ]- internal . aliyuncs . com /[ bucketpath ]/ userinfo '
```

2. Create a video table

```
% hive
 CREATE   EXTERNAL   TABLE   video_info ( id   int , title   string
 , type   string )  ROW   FORMAT   DELIMITED   FIELDS   TERMINATED
   BY  ','  LOCATION  ' oss ://[ bucketname ]. oss - cn -[ region ]-
 internal . aliyuncs . com /[ bucketpath ]/ videoinfo '
```

3. Create a video playback table

```
% hive
 CREATE   EXTERNAL   TABLE   play_video ( user_id   int , video_id
   int ,  play_time   bigint )  ROW   FORMAT   DELIMITED   FIELDS
 TERMINATED   BY  ','  LOCATION  ' oss ://[ bucketname ]. oss - cn -[
 region ]- internal . aliyuncs . com /[ bucketpath ]/ playvideo '
```

4. Count the user tables

```
% sql   select   count (*)  from   user_info
```

5. Count the video tables

```
% sql   select   count (*)  from   video_info
```

6. Count the video playback tables

```
% sql   select   count (*)  from   play_video
```

7. Count the video playbacks for each video type

```
% sql   select   video . type ,  count ( video . type )  as   count
 from   play_video   play   join   video_info   video   on  ( play .
 video_id  =  video . id )  group   by   video . type   order   by
 count   desc
```

8. Display the video information for the top 10 video playbacks

```
% sql   select   video . id ,  video . title ,  video . type ,
 video_coun  t . count   from  ( select   video_id ,  count ( video_id
 )  as   count   from   play_video   group   by   video_id   order
 by   count   desc   limit   10 )  video_coun  t   join   video_info
 video   on  ( video_coun  t . video_id  =  video . id )  order   by
 count   desc
```

9. Display the age of the viewers watching the video with the most video playbacks

```
% sql   select   age  ,  count (*)  as   count    from  ( select
 distinct ( user_id )  from   play_video    where   video_id  = 49
  )  play   join   user_info   userinfo   on  ( play . user_id  =
 userinfo . id )  group   by   userinfo . age
```

10. Display the gender, age, and marital status of the viewers watching the video with the most video playbacks

```
% sql   select   if ( sex = 0 ,' Female ',' Male ')  as   title ,
 count (*)  as   count , ' Gender '  as   type   from  ( select
 distinct ( user_id )  from   play_video    where   video_id  = 49
  )  play   join   user_info   userinfo   on  ( play . user_id  =
 userinfo . id )  group   by   userinfo . sex
 union   all
 select   case   when   userinfo . age < 15   then  ' Less   than
 15 '  when   age < 25   then  ' 15 - 25 '  when   age < 35   then
  ' 25 - 35 '  else  ' More   than   35 '  end    ,  count (*)  as
   count , ' Age   Group '  as   type   from  ( select   distinct (
 user_id )  from   play_video    where   video_id  = 49 )  play   join
   user_info   userinfo   on  ( play . user_id  =  userinfo . id )
 group   by   case   when   userinfo . age < 15   then  ' Less   than
   15 '  when   age < 25   then  ' 15 - 25 '  when   age < 35   then
 ' 25 - 35 '  else  ' More   than   35 '  end
 union   all
 select   if ( marital_st  atus = 0 ,' Unmarried ',' Married ')  as
   title , count (*)  as   count , ' Marital   Status '  as   type
 from  ( select   distinct ( user_id )  from   play_video    where
 video_id  = 49  )  play   join   user_info   userinfo   on  ( play .
 user_id  =  userinfo . id )  group   by   marital_st  atus
```

6.2 Execution plans

6.2.1 Create an execution plan

An execution plan is a set of jobs that can be executed either at one time or periodically by means of scheduling. It can be executed on an existing E-MapReduce cluster and can also create a temporary cluster to execute the jobs dynamically. Its biggest advantage is that it only uses the resources it needs during execution.

Procedure

To create an execution plan, follow these steps:

1. Log on to the *Alibaba Cloud E-MapReduce console* .
2. Select a region.
3. In the upper-right corner, click Old MER Scheduling to go to the Jobs page.

4. In the navigation panel on the left, click Execution plan.
5. In the upper-right corner, click Create an execution plan.
6. In the Create an execution plan page, select between Create as needed and Existing clusters.

    a. Create as needed: Create a new cluster to run jobs.

    - Execution plan for one-time scheduling: Clusters with corresponding configurations are created when the execution starts and are then released upon completion of the operation. For more information about creation parameters, see *Step 3 : Create a cluster*.
    - Execution plan for periodic scheduling: A new cluster is created based on the scheduling settings you define and is then released upon completion of the operation.

    b. Existing clusters: Use an existing cluster that complies with the following requirement:

    - Execution plans can only be added to clusters that are Running or Idle.

    Select Existing clusters and then enter the Select Cluster page. Here, you can select a cluster to associate with the execution plan.
7. Click Next to enter the job configuration page. All user jobs are listed in the table on the left. You can select jobs for execution from this table. By clicking the right-facing button, the checked jobs are added to the job queue. Jobs in the queue are then submitted to the cluster for execution in order. The same job can be added and executed several times. If you have not created any jobs, see *Jobs*.
8. Click Next to enter the scheduling mode configuration page. The configuration items are as follows:

    a. Name: Must be between 1-64 characters and may only consist of Chinese characters, English letters, numbers, hyphens (-), and underscores (_).

    b. Scheduling policy

    - Manual execution: The execution plan is not executed automatically after it is created. Instead, it must be executed manually. Once the execution is in progress, it cannot be executed again.
    - Periodic scheduling: If you select this function, it is enabled immediately after the execution plan is created. The execution then begins from the configured scheduling time. Periodic scheduling can be disabled in the list

page. If a scheduling execution starts, but its last execution is not completed, the scheduling is ignored.

c. Set the scheduling cycle: There are two scheduling periods: days and hours. The day cycle is one day by default and cannot be changed. However, you can set a specific time interval for hours. The range must be 1-23.
d. First execution time: The effective start-time of the scheduling. From this point onwards, periodic scheduling is conducted according to the intervals specified.

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

Other information

- Example of periodic scheduling

  These configurations indicate that the scheduling started on 11/01/2018 at 17:35 with an interval of one day. This means that the second scheduling was conducted on 11/02/2018, at 17:35.
- Sequence of jobs

  Jobs in the execution plan are executed from first to last according to the sequence that you defined in the job list.
- Sequence of multiple execution plans

  When multiple execution plans are submitted to the same cluster, each one submits jobs from its own job sequence. This means that jobs run parallel with each other.
- Example of early job debugging

  During the debugging of a job, it may take some time to create and start a cluster on demand. We recommend that you create a cluster manually first, select Associate the cluster in the execution plan to run jobs, and then set the scheduling mode to Execute immediately. During debugging, you can view the results by clicking Run now on the execution plan list page. Once the debugging is finished, modify the execution plan, modify the way you associate an existing cluster to create a new cluster on demand, and then modify the scheduling mode to periodic scheduling as required. Jobs are then executed automatically on demand.

6.2.2 Manage an execution plan

You can view, manage, and modify your execution plans as follows.

1. Log on to the *Alibaba Cloud E-MapReduce console* .
2. Select a region.
3. In the upper-right corner, click Old MER Scheduling to go to the Jobs page.
4. In the navigation panel on the left, click Execution plan.
5. Click Manage next to a plan to go to the execution plan detail page. Here, you can perform the following operations:

   - View details of the execution plan

     You can view the basic information of the execution plan, such as its name, associated clusters, job configurations, scheduling mode and status, and alarm information.
   - Modify the execution plan

     > Notice:
     >
     > Jobs can only be modified if they are not currently running or being scheduled. For an execution plan to be executed immediately, it can only be modified when it is not currently running. If the execution plan is scheduled periodically, wait

for the completion of its current operation and verify whether it is in periodical scheduling. If it is, click Stop scheduling before modifying it.

Each separate module can be modified independently. Click the pen icon to modify information.

- Configure alarm notifications

  There are three types of alarm notifications:

  - Booting timeout: If the periodical scheduling has not been conducted correctly at the specified time and is not executed within 10 minutes of timeout, an alarm is sent.
  - Failed execution: If any job in the execution plan fails, an alarm is sent.
  - Successful execution: If all jobs in the execution plan are executed successfully, a notification is sent.
- Run and view results

  If the execution plan can be run, in Basic Information, there will be a Run now button to the right of Scheduling status. If you click this button, a schedule will be executed.

  At the bottom of the page, there are running records displaying the execution plan instances executed each time, making it easy to view the corresponding job list and logs.

6.2.3 Execution plan list

An execution plan list displays basic information about all of your execution plans, as shown in the following figure.

- ID/Name: The ID and name of the execution plan.
- Last run cluster: The last cluster to execute this execution plan. This can either be a cluster created on demand or an existing associated cluster. If a cluster is created automatically on demand, (Automatically created) is displayed beneath it, indicating that the cluster was created on demand by E-MapReduce and will be released automatically after running.

- Last run: The running status of the last execution plan.
  - Start time: The time at which the last execution plan started.
  - Running time: The duration for which the last plan ran.
  - Running status: The running status of the last execution plan.
- Scheduling status: This indicates whether scheduling is in progress or has been stopped. Only periodic jobs have a scheduling status.
- Operation
  - Manage: View and modify execution plans.
  - Run now: A job can only be run manually when it is neither running nor being scheduled. Click Run now to run the execution plan immediately.
  - More
    - Start/Stop scheduling: If the scheduling is stopped, Enable scheduling is displayed, which you can click to start the scheduling. If Stop scheduling is displayed during scheduling, you can click it to stop the scheduling. This button is only available for periodic execution plans.
    - Running log: Click to enter the job log viewing page.
    - Delete: Deletes an execution plan. A running execution plan or one in the process of scheduling cannot be deleted.

6.2.4 View job results and logs

In this tutorial, you will learn how to view job results and logs.

View execution records

1. Log on to the *Alibaba Cloud E-MapReduce console* .
2. Select a region.
3. In the upper-right corner, click Old MER Scheduling to go to the Jobs page.
4. In the navigation panel on the left, click Execution plan.

5. To the right of the execution plan, click More > Running log.

   - Execution order ID: The sequence of execution for the execution record, which indicates its position in the execution queue. For example, 1 stands for the first position.
   - Running status: The running status of each execution record.
   - Start time The time at which the execution plan starts.
   - Running time: The total running time until the page is viewed.
   - Execute cluster: The cluster run by the execution plan can either be created on demand or it can be an existing associated cluster. Click to view the cluster details page.
   - Operation

     View job list: Click to enter the job list page.

View job records

On the Job list page, you can view the job list in the execution records of a single execution plan as well as the details of each job, as shown in the following figure.

- Job execution order ID: After a job is executed, a corresponding ID is created, which is different from the job ID. The job execution ID is the unique identifier for viewing logs on OSS.
- Name: The name of the job.
- Status: The running status of the job.
- Type: The type of job.
- Start time: The time at which the job starts. This is converted into local time.
- Running time: The total running time of the job, in seconds.

- Operation
  - Stop job: You can stop a job if it is in the process of submission or running. If a job is in submission, stopping it will cancel execution. If the job is running, it will be killed.
  - stdout: Records all output content from the standard output (Channel 1) of the master process. If log saving is not enabled for the cluster where jobs are run, this function cannot be executed.
  - stderr: Records all output content from the diagnostic output (Channel 2) of the master process. If log saving is not enabled for the cluster where jobs are run, this function cannot be executed.
  - Workers log: Views the logs of all job worker nodes. If log saving is not enabled for the cluster where jobs are run, this function cannot be executed.

View job worker logs

- Cloud server instance IP: The ECS instance ID of a running job and the corresponding intranet IP address.
- Container ID: The container ID that YARN runs.
- Type: Different log types. stdout and stderr come from different outputs.
- Operation

  View the log: Click different types to view the corresponding logs.

6.2.5 Parallel execution of multiple execution plans

To maximize the use of a cluster's available computing resources, multiple execution plans can be associated to the same cluster and executed in parallel.

The main points are summarized as follows:

- Jobs in the same execution plan are executed in sequence. By default, preceding jobs are executed before new jobs can be submitted and executed.
- If you have enough cluster resources, you can create multiple different execution plans and associate them to the same cluster to run and execute jobs in parallel. Clusters support a maximum of 20 execution plans by default.
- The management and control system currently supports the submission to YARN of multiple execution plans associated to the same cluster. However, if the cluster itself has insufficient resources, it may take some time for jobs in the YARN queue to wait for scheduling.

For more information on how to create execution plans and associate them to a cluster, see *#unique_32*.