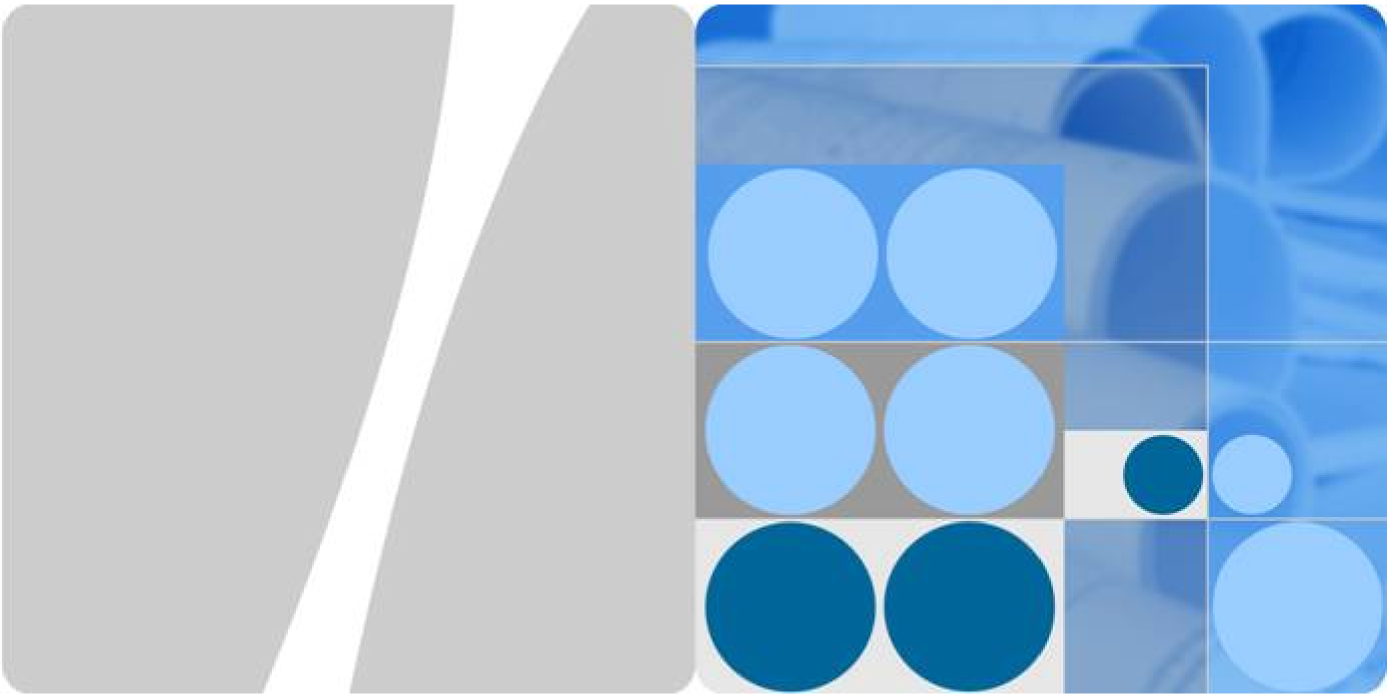


Data Lake Factory

Service Overview

Issue	01
Date	2020-12-08

HUAWEI TECHNOLOGIES CO., LTD.

Copyright © Huawei Technologies Co., Ltd. 2022. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Trademarks and Permissions

HUAWEI and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

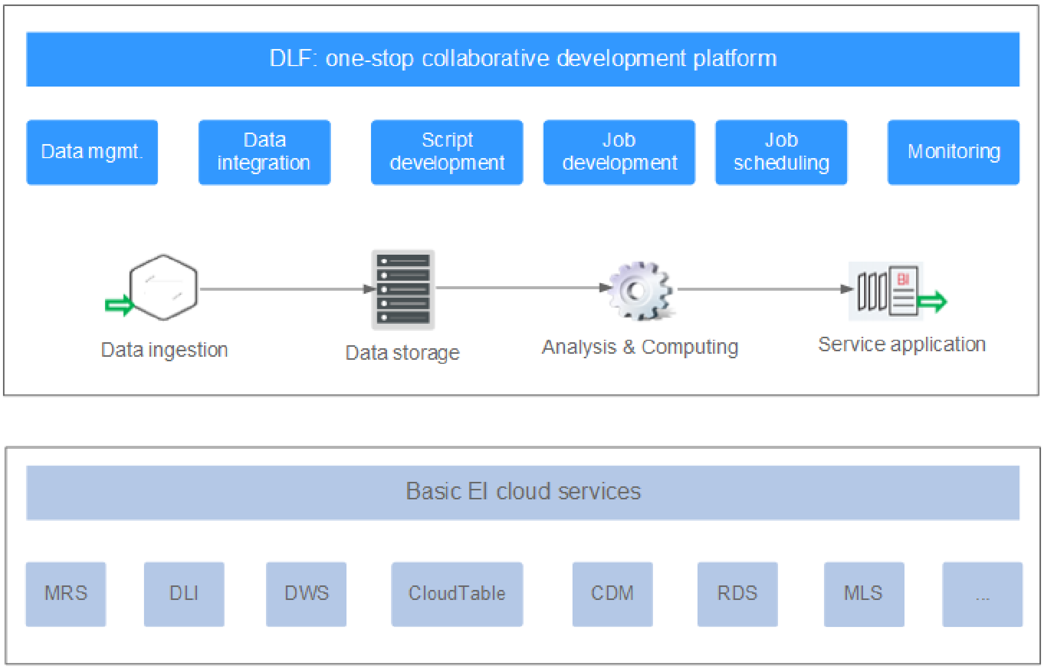
Contents

1 Overview

Data Lake Factory (DataArts Factory) is a big data platform designed specifically for the HUAWEI CLOUD. It manages diverse big data services and provides a one-stop big data development environment and fully-managed big data scheduling capabilities. Thanks to DLF, big data is more accessible than ever before, helping you effortlessly build big data processing centers.

DataArts Factory enables a variety of operations such as data management, data integration, script development, job scheduling, and monitoring, facilitating the data analysis and processing procedure.

Figure 1-1 DLF process



For details about data management, data integration, script development, job development, job scheduling, and monitoring, see **Functions**.

Copyright © Huawei Technologies Co., Ltd.

2 Product Advantages

One-Stop Data Warehouse Building

DLF supports one-stop building of cloud data warehouses, where you can complete data integration, script development, job development, job scheduling, job monitoring, and data management without the need for multiple tools.

Data Lake Development

DLF manages a variety of Big Data services such as DWS and DLI and allows data to be orchestrated and scheduled in different types of data services.

Diverse Data Types

DLF allows online collaborative development, supports online editing of SQL and Shell scripts and real-time script query, and enables job development for types of data processing nodes such as Data Migration, SQL, MR, Shell, Machine Learning, and Spark.

Powerful Job Scheduling Capabilities

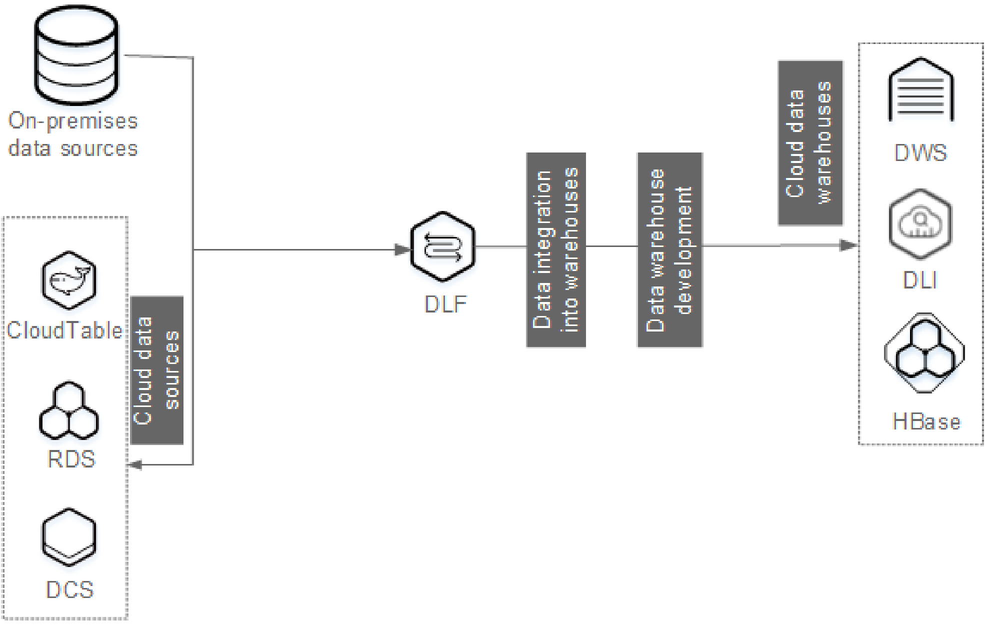
DLF provides you with diverse scheduling policies and powerful scheduling capabilities, supporting manual, periodic, and event-driven scheduling.

3 Scenarios

Quick Building of Cloud Data Warehouses

DataArts Factory can migrate offline data to the HUAWEI CLOUD and integrate the data into the HUAWEI CLOUD big data services for data development in DataArts Factory.

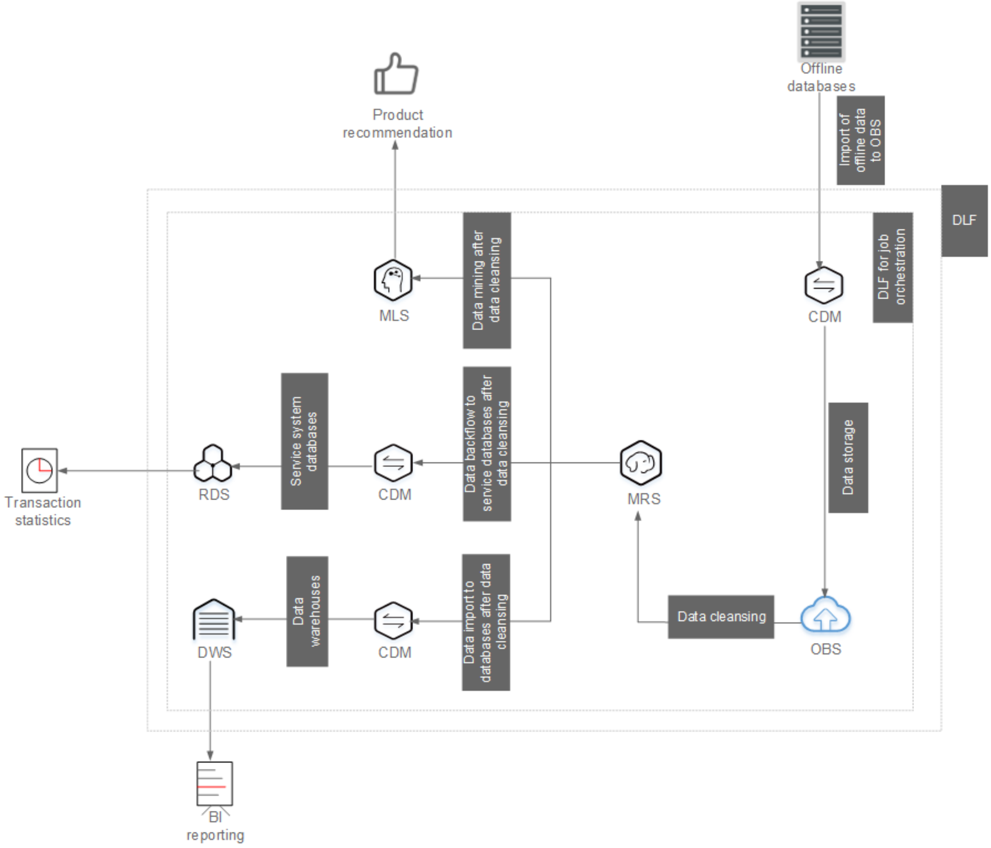
Figure 3-1 Scenario example



Automated Data Analysis Service Flow

DataArts Factory automates the E2E procedure from data import, data cleaning, machine learning, data backhaul, to report generation.

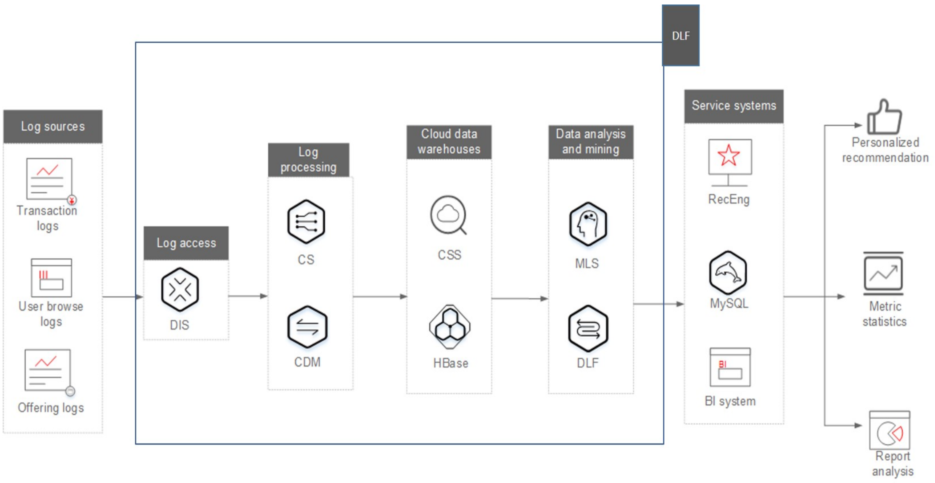
Figure 3-2 Scenario example


Product recommendation
Offline databases
Import of offline data to OBS
DLF
DLF for job orchestration
Data mining after data cleansing
MLS
CDM
Service system databases
Data backflow to service databases after data cleansing
Data storage
Transaction statistics
RDS
CDM
MRS
Data warehouses
Data import to databases after data cleansing
Data cleansing
DWS
CDM
OBS
BI reporting

Easy Analysis and Mining of Massive Amounts of Logs

After ingesting logs into Object Storage Service (OBS) or Cloud Search Service through Data Ingestion Service (DIS), DLF helps you analyze and mine massive amounts of logs by compiling data development scripts and data mining scripts.

Figure 3-3 Scenario example



4 Functions

Data Management

- Manages multiple data warehouses, such as DWS, MRS Hive, and DLI.
- Manages data tables using the visual interface or data definition language (DDL).

Data Integration

Works with Cloud Data Migration (CDM) to enable reliable and effective data transmission between 20+ disparate data sources and effortlessly integrate data sources into data warehouses.

Script Development

- Provides an online script editor that allows more than one operator to collaboratively develop and debug SQL and Shell scripts online.
- Allows usage of variables and functions.

Job Development

- Provides a graphical designer that allows you to quickly build a data processing workflow by drag-and-drop.
- Presets multiple task types such as data integration, MR, Spark, machining learning, SQL, and shell and completes data analysis and processing by dependency between tasks.
- Supports job import and export.

Resource Management

Supports unified management of file, jar, and archive resources used during script and job development.

Job Scheduling

Supports **Run once**, **Run periodically**, and **Event-driven**. If **Run periodically** is selected, you can run a job by **Minute**, **Hour**, **Day**, **Week**, or **Month**.

Monitoring

- Supports basic management of a job, including run, pause, restore, and terminate.
- Allows you to view the operation details of each job and each node in the job.
- Supports diverse alert methods so that the related personnel can be notified when a job or task error occurs.

5 Permissions Management

If you need to assign different permissions to employees in your enterprise to access your DLF resources, IAM is a good choice for fine-grained permissions management. IAM provides identity authentication, permissions management, and access control, helping you secure access to your HUAWEI CLOUD resources.

With IAM, you can use your HUAWEI CLOUD account to create IAM users for your employees, and assign permissions to the users to control their access to specific resource types. For example, some software developers in your enterprise need to use DLF resources but must not delete them or perform any high-risk operations. To achieve this result, you can create IAM users for the software developers and grant them only the permissions required for using DLF resources.

If your HUAWEI CLOUD account does not need individual IAM users for permissions management, you may skip over this chapter.

IAM can be used free of charge. You pay only for the resources in your account. For more information about IAM, see **IAM Service Overview**.

DLF Permissions

By default, new IAM users do not have permissions assigned. You need to add them to one or more groups, and attach permissions policies or roles to these groups. Users inherit permissions from the groups to which they are added and can perform specified operations on cloud services based on the permissions.

DLF is a project-level service deployed in specific physical regions. Therefore, DLF permissions are assigned to users in specific regions (such as **CN North-Beijing1**) and only take effect for these regions. If you want the permissions to take effect for all regions, you need to assign the permissions to users in each region. When accessing DLF, the users need to switch to a region where they have been authorized to use cloud services.

DLF system permissions lists all the system-defined roles and permissions supported by DLF.

You can grant users permissions by using roles and policies.

- Role: IAM initially provides a coarse-grained authorization mechanism to define permissions based on users' job responsibilities. This mechanism provides only a limited number of service-level roles for authorization. When using roles to grant permissions, you need to also assign other roles on which the permissions depend to take effect. However, roles are not an ideal choice for fine-grained authorization and secure access control.

- Policies: A type of fine-grained authorization mechanism that defines permissions required to perform operations on specific cloud resources under certain conditions. This mechanism allows for more flexible policy-based authorization, meeting requirements for secure access control. For example, a user group is not allowed to delete jobs but can only perform basic operations on jobs, such as creating jobs and querying the job list. For the API actions supported by DLF, see **Permissions Policies and Supported Actions**.

Table 5-1 DLF system permissions

Role/Policy Name	Description	Type
DLF FullAccess	All permissions for DLF	System-defined policy
DLF Development	Developer permissions for DLF. Users granted these permissions can use DLF to develop scripts and orchestrate jobs, but cannot add, delete, or modify workspaces.	System-defined policy
DLFOperationAnd-MaintenanceAccess	O&M permissions for DLF. Users granted these permissions can perform O&M operations on DLF scripts and jobs, but cannot add, delete, or modify resources.	System-defined policy
DLF ReadOnlyAccess	Read-only permissions for DLF. Users granted these permissions can only view DLF resources.	System-defined policy
DLF Administrator	Administrator permissions for DLF.	System-defined role

Common operations supported by each policy or role lists the common operations supported by each policy or role of DLF. Select the policies or roles as required.

Table 5-2 Common operations supported by each policy or role

Operation	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
Querying workspaces	√	√	√	√	√

Operation	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
Creating workspaces	√	x	x	x	√
Updating workspaces	√	x	x	x	√
Deleting workflows	√	x	x	x	√
Querying environment variables.	√	√	√	√	√
Updating environment variables	√	√	x	x	√
Importing environment variables	√	√	x	x	√
Exporting environment variables	√	√	x	x	√
Querying tables	√	√	√	√	√
Creating tables	√	√	x	x	√
Updating tables	√	√	x	x	√
Deleting tables	√	√	x	x	√
Querying databases	√	√	√	√	√
Creating databases	√	√	x	x	√

Operation	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
Updating databases	√	√	x	x	√
Deleting databases	√	√	x	x	√
Querying modes.	√	√	√	√	√
Creating modes	√	√	x	x	√
Updating modes	√	√	x	x	√
Deleting modes	√	√	x	x	√
Querying directories	√	√	√	√	√
Creating directories	√	√	x	x	√
Updating directories	√	√	x	x	√
Deleting directories	√	√	x	x	√
Querying solutions	√	√	√	√	√
Creating solutions	√	√	x	x	√
Updating solutions	√	√	x	x	√
Deleting solutions	√	√	x	x	√
Importing solutions	√	√	√	x	√

Operation	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
Exporting solutions	√	√	√	x	√
Starting solutions	√	√	√	x	√
Stopping solutions	√	√	√	x	√
Querying scripts	√	√	√	√	√
Creating scripts	√	√	x	x	√
Updating scripts	√	√	x	x	√
Deleting scripts	√	√	x	x	√
Checking script syntax	√	√	√	x	√
Executing scripts	√	√	√	x	√
Canceling script execution	√	√	√	x	√
Importing scripts	√	√	√	x	√
Exporting scripts or script execution results	√	√	√	x	√
Querying jobs	√	√	√	√	√
Creating jobs	√	√	x	x	√
Updating jobs	√	√	x	x	√

Operation	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
Deleting jobs	√	√	x	x	√
Renaming jobs	√	√	x	x	√
Importing jobs	√	√	√	x	√
Exporting jobs	√	√	√	x	√
Verifying validity of job definitions	√	√	√	x	√
Testing jobs	√	√	√	x	√
Starting jobs	√	√	√	x	√
Stopping jobs	√	√	√	x	√
Suspending jobs	√	√	√	x	√
Resuming jobs	√	√	√	x	√
Querying job instances	√	√	√	√	√
Rerunning job instances	√	√	√	x	√
Stopping job instances	√	√	√	x	√

Operation	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
Forcibly making jobs instances succeed	√	√	√	x	√
Continuing with job instances	√	√	√	x	√
Enabling job nodes	√	√	√	x	√
Disabling job nodes	√	√	√	x	√
Rerunning job nodes	√	√	√	x	√
Skipping job node execution	√	√	√	x	√
Suspending job nodes	√	√	√	x	√
Resuming job node execution	√	√	√	x	√
Forcibly making jobs nodes succeed	√	√	√	x	√
Querying data connections	√	√	√	√	√
Creating data connections	√	√	x	x	√

Operation	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
Updating data connections	√	√	x	x	√
Deleting data connections	√	√	x	x	√
Testing connectivity of data connections	√	√	x	x	√
Importing data connections	√	√	√	x	√
Exporting data connections	√	√	√	x	√
Querying resources	√	√	√	√	√
Creating resources	√	√	x	x	√
Updating resources	√	√	x	x	√
Deleting resources	√	√	x	x	√
Uploading resources	√	√	x	x	√
Exporting resources	√	√	√	x	√
Importing resources	√	√	√	x	√

Operation	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
Querying backup information	√	√	√	√	√
Starting backup	√	√	√	x	√
Stopping backup	√	√	√	x	√
Querying notifications	√	√	√	√	√
Creating notifications	√	√	x	x	√
Updating notifications	√	√	x	x	√
Deleting notifications	√	√	x	x	√
Querying PatchData	√	√	√	√	√
Creating PatchData tasks	√	√	x	x	√
Stopping PatchData tasks	√	√	√	x	√

6 Related Services

MRS

The big data type of nodes (such as SparkSQL) in DataArts Factory runs in Map Reduce Service (MRS).

OBS

DataArts Factory supports data import from Object Storage Service (OBS), and also can use OBS to store data, results, log files, and user programs.

RDS

DataArts Factory allows data to be stored to Relational Database Service (RDS) and completes RDS data processing.

DEW

Data Encryption Workshop (DEW) is used to encrypt and decrypt user passwords and keys of data connections in DataArts Factory.

DWS

DataArts Factory allows data to be stored to Data Warehouse Service (DWS) and completes DWS data processing.

CDM

DataArts Factory relies on Cloud Data Migration (CDM) to complete migration data processing.

MLS

DataArts Factory relies on Machine Learning Service (MLS) to complete machine learning data processing

DLI

DataArts Factory relies on Data Lake Insight (DLI) to complete data insight processing.

Cloud Search Service

DataArts Factory relies on Cloud Search Service to complete cloud search data processing

SMN

DataArts Factory relies on Simple Message Notification (SMN) to achieve notification management by sending job information to the user.

CS

DataArts Factory relies on Cloud Stream Service (CS) to analyze streaming Big Data in real time.

CloudTable

DataArts Factory allows data to be stored in CloudTable Service (CloudTable).

DIS

DataArts Factory relies on Data Ingestion Service (DIS) to complete dump related processing.

IAM

The Identity and Access Management (IAM) provides the authentication function for DataArts Factory.

7 Constraints

Before using DataArts Factory, be aware of the following use constraints:

- The supported browser types and versions are listed as follows:
  - Google Chrome: 54.0 or later

8 Regions and AZs

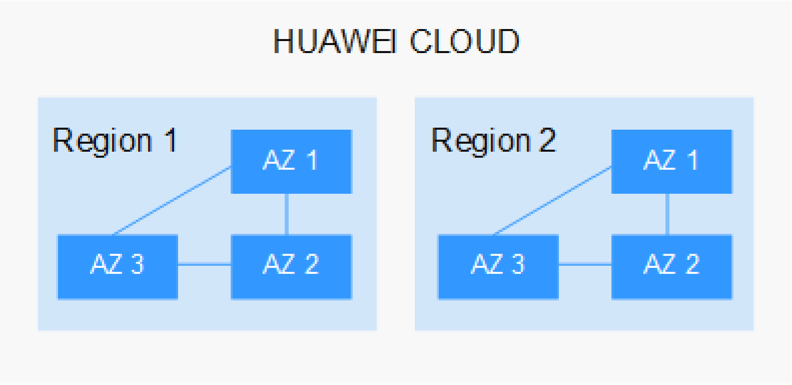
What Are Regions and AZs?

A region and availability zone (AZ) identify the location of a data center. You can create resources in a specific region and AZ.

- Regions are classified based on geographical location and network latency. Public services, such as Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Object Storage Service (OBS), Virtual Private Cloud (VPC), Elastic IP (EIP), and Image Management Service (IMS), are shared within the same region. Regions are classified as universal regions and dedicated regions. A universal region provides universal cloud services for common tenants. A dedicated region provides services of the same type or only provides services for specific tenants.
- An AZ contains one or more physical data centers. Each AZ has independent cooling, fire extinguishing, moisture-proofing, and electricity facilities. Within an AZ, computing, network, storage, and other resources are logically divided into multiple clusters. AZs within a region are interconnected using high-speed optical fibers to allow you to build cross-AZ high-availability systems.

Figure 8-1 shows the relationship between regions and AZs.

Figure 8-1 Regions and AZs



HUAWEI CLOUD provides services in many regions around the world. You can select a region and AZ as needed. For more information, see **Global Products and Services**.

How to Select a Region?

When selecting a region, consider the following factors:

- Location

  You are advised to select a region close to you or your target users. This reduces network latency and improves access speed. However, Chinese mainland regions provide the same infrastructure, BGP network quality, as well as resource operations and configurations. Therefore, if you or your target users are in the Chinese mainland, you do not need to consider the network latency differences when selecting a region.

  The countries and regions outside the Chinese mainland, such as Bangkok and Hong Kong, provide services for users outside the Chinese mainland. If you or your target users are in the Chinese mainland, these regions are not recommended due to high access latency.

  - If you or your target users are in the Asia Pacific area (excluding the Chinese mainland), select the **AP-Bangkok**, or **AP-Singapore** region.
  - If you or your target users are in Africa, select the **AF-Johannesburg** region.
  - If you or your target users are in Europe, select the **EU-Paris** region.

- Resource price

  Resource prices may vary in different regions. For details, see **Product Pricing Details**.

Regions and Endpoints

When using an API to use resources, you must specify its region and endpoint. For details about HUAWEI CLOUD regions and endpoints, see **Regions and Endpoints**.

Copyright © Huawei Technologies Co., Ltd.

9 Quota Description

Each user can create a maximum of 1,000 jobs by default.

NOTE

For details about the quotas of other resource services involved, see the quotas of other services.

Copyright © Huawei Technologies Co., Ltd.