数据湖探索

产品介绍

文档版本	01
发布日期	2024-12-04

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：
https://www.huawei.com/cn/psirt/vul-response-process
如企业客户须获取漏洞信息，请参见如下网址：
https://securitybulletin.huawei.com/enterprise/cn/security-advisory

目 录

1 图解数据湖探索

2 数据湖探索简介

什么是数据湖探索

数据湖探索（Data Lake Insight，简称DLI）是完全兼容**Apache Spark**、**Apache Flink**、HetuEngine生态，提供一站式的流处理、批处理、交互式分析的Serverless融合处理分析服务。用户不需要管理任何服务器，即开即用。

DLI支持标准SQL/Spark SQL/Flink SQL，支持多种接入方式，并兼容主流数据格式。数据无需复杂的抽取、转换、加载，使用SQL或程序就可以对云上**CloudTable**、**RDS**、**DWS**、**CSS**、**OBS**、**ECS**自建数据库以及线下数据库的异构数据进行探索。

功能介绍

DLI用户可以通过可视化界面、Restful API、JDBC、Beeline等多种接入方式对云上CloudTable、RDS和DWS等异构数据源进行查询分析，数据格式兼容CSV、JSON、Parquet和ORC主流数据格式。

- 三大基本功能
  - SQL作业支持SQL查询功能：可为用户提供标准的SQL语句。具体内容请参考**《数据湖探索SQL语法参考》**。
  - Flink作业支持Flink SQL在线分析功能：支持Window、Join等聚合函数、地理函数、CEP函数等，用SQL表达业务逻辑，简便快捷实现业务。具体内容请参考**《数据湖探索SQL语法参考》**。
  - Spark作业提供全托管式Spark计算特性：用户可通过交互式会话(session)和批处理(batch)方式提交计算任务，在全托管Spark队列上进行数据分析。具体内容请参考**《数据湖探索API参考》**。
- 多数据源分析：
  - Spark跨源连接：可通过DLI访问CloudTable，DWS，RDS和CSS等数据源。具体内容请参考**《数据湖探索用户指南》**。
  - Flink跨源支持与多种云服务连通，形成丰富的流生态圈。数据湖探索的流生态分为云服务生态和开源生态：
    - 云服务生态：数据湖探索在Flink SQL中支持与其他服务的连通。用户可以直接使用SQL从这些服务中读写数据。如DIS、OBS、CloudTable、MRS、RDS、SMN、DCS等。

  - 开源生态：通过增强型跨源连接建立与其他VPC的网络连接后，用户可以在数据湖探索的租户授权的队列中访问所有Flink和Spark支持的数据源与输出源，如Kafka、Hbase、ElasticSearch等。

  具体内容请参见**《数据湖探索开发指南》**。

- 存算分离

  用户将数据存储到OBS后，DLI可以直接和OBS对接进行数据分析。存算分离的架构下，使得存储资源和计算资源可以分开申请和计费，降低了成本并提高了资源利用率。

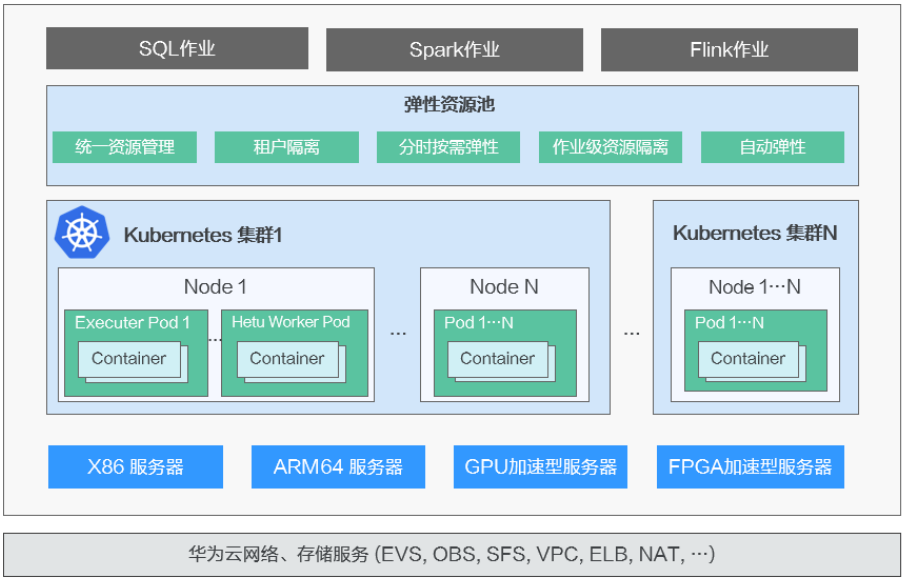
  存算分离场景下，DLI支持OBS在创建桶时数据冗余策略选择单AZ或者多AZ存储，两种存储策略区别如下：

  - 选择多AZ存储，数据将冗余存储至多个AZ中，可靠性更高。选择多AZ存储的桶，数据将存储在同一区域的多个不同AZ。当某个AZ不可用时，仍然能够从其他AZ正常访问数据，适用于对可靠性要求较高的数据存储场景。建议优选使用多AZ存储的策略。
  - 选择单AZ存储，数据仅存储在单个AZ中，但相比多AZ更加便宜。收费详情请参见**OBS产品价格详情**。

- 弹性资源池

  弹性资源池后端采用CCE集群的架构，支持异构，对资源进行统一的管理和调度。详细内容可以参考用户指南的**弹性资源池**。

  图 2-1 弹性资源池架构图



  弹性资源池的优势主要体现在以下几个方面：

  - **统一资源管理**
    - 统一管理内部多集群和调度作业，规模可以到百万核级别。
    - 多AZ部署，支持跨AZ高可用。
  - **租户资源隔离**

    不同队列之间资源隔离，减少队列之间的相互影响。

- **分时按需弹性**
  - 分钟级别扩缩容，从容应对流量洪峰和资源诉求。
  - 支持分时设置队列优先级和配额，提高资源利用率。
- **作业级资源隔离（暂未实现，后续版本支持）**

  支持独立Spark实例运行SQL作业，减少作业间相互影响。
- **自动弹性（暂未实现，后续版本支持）**

  基于队列负载和优先级实时自动更新队列配额。

弹性资源池解决方案主要解决了以下问题和挑战。

维度	原有队列，无弹性资源池时	弹性资源池
扩容时长	手工扩容时间长，扩容时长在分钟级别	不需要手工干预，秒级动态扩容。
资源利用率	不同队列之间资源不能共享。 例如：队列1当前还剩余10CU资源，队列2当前负载高需要扩容时，队列2不能使用队列1中的资源，只能单独对队列1进行扩容。	添加到同一个弹性资源池的多个队列，CU资源可以共享，达到资源的合理利用。
	配置跨源时，必须为每个队列分配不重合的网段，占用大量VPC网段。	多队列通过弹性资源池统一进行网段划分，减少跨源配置的复杂度。
资源调配	多个队列同时扩容时不能设置优先级，在资源不够时，会导致部分队列扩容申请失败。	您可以根据当前业务波峰和波谷时间段，设置各队列在弹性资源池中的优先级，保证资源的合理调配。

- BI工具

  对接永洪BI：与永洪BI对接实现数据分析。具体内容请参考**《数据湖探索开发指南》**。

DLI 核心引擎：Spark+Flink+HetuEngine

- Spark是用于大规模数据处理的统一分析引擎，聚焦于查询计算分析。DLI在开源Spark基础上进行了大量的性能优化与服务化改造，不仅兼容Apache Spark生态和接口，性能较开源提升了2.5倍，在小时级即可实现EB级数据查询分析。
- Flink是一款分布式的计算引擎，可以用来做批处理，即处理静态的数据集、历史的数据集；也可以用来做流处理，即实时地处理一些实时数据流，实时地产生数据的结果。DLI在开源Flink基础上进行了特性增强和安全增强，提供了数据处理所必须的Stream SQL特性。
- HetuEngine是提供交互式查询分析能力的开源分布式SQL查询引擎，具备高性能、低延迟的查询处理能力，支持在大规模数据存储中进行数据查询和分析。

DLI 服务架构：Serverless

DLI是无服务器化的大数据查询分析服务，其优势在于：

- 按量计费：真正的按使用量（扫描量/CU时）计费，不运行作业时0费用。
- 自动扩缩容：根据业务负载，对计算资源进行预估和自动扩缩容。

如何访问 DLI

云服务平台提供了Web化的服务管理平台，既可以通过管理控制台和基于HTTPS请求的API（Application programming interface）管理方式来访问DLI，又可以通过JDBC客户端连接DLI服务端。

- 管理控制台方式

  提交SQL作业、Spark作业或Flink作业，均可以使用管理控制台方式访问DLI服务。

  登录管理控制台，从主页选择“EI企业智能”>“EI大数据”>“数据湖探索”。
- API方式

  如果用户需要将云平台上的DLI服务集成到第三方系统，用于二次开发，可以使用API方式访问DLI服务。

  具体操作请参见**《数据湖探索API参考》**。
- JDBC

  DLI支持使用JDBC连接服务端进行数据查询操作。具体内容请参考**《数据湖探索开发指南》**。
- Spark-submit

  DLI支持通过Spark-submit提交作业。具体内容请参考**《数据湖探索开发指南》**。
- 数据治理中心DataArts Studio

  数据治理中心DataArts Studio具有数据全生命周期管理、智能数据管理能力的一站式治理运营平台，支持行业知识库智能化建设，支持大数据存储、大数据计算分析引擎等数据底座，帮助企业快速构建从数据接入到数据分析的端到端智能数据系统，消除数据孤岛，统一数据标准，加快数据变现，实现数字化转型。

  在DataArts Studio管理中心控制台创建数据连接即可访问DLI，进行数据分析。关于DataArts Studio的操作指导请参考《**数据治理中心产品文档**》。

3 产品优势

纯 SQL 操作

DLI提供标准SQL接口，用户仅需使用SQL便可实现海量数据查询分析。SQL语法全兼容标准ANSI SQL 2003。

存算分离

DLI解耦计算和存储负载，存算分离架构，存储资源和计算资源按需灵活配置，提高了资源利用率，降低了成本。

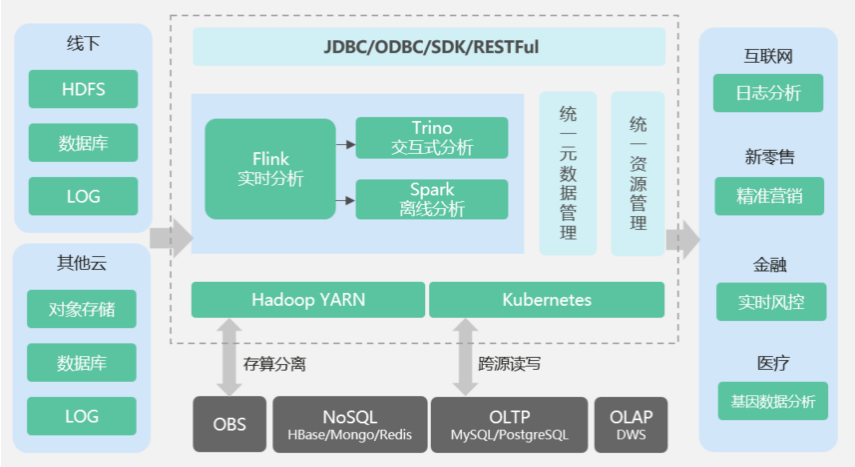
企业级多租户

支持计算资源按租户隔离，数据权限控制到队列、作业，帮助企业实现部门间的数据共享和权限管理。

Serverless DLI

DLI完全兼容Apache Spark、Apache Flink生态和接口，是集实时分析、离线分析、交互式分析为一体的Serverless大数据计算分析服务。线下应用可无缝平滑迁移上云，减少迁移工作量。采用批流融合高扩展性框架，为TB~EB级数据提供了更实时高效的多样性算力，可支撑更丰富的大数据处理需求。产品内核及架构深度优化，综合性能是传统MapReduce模型的百倍以上，SLA保障99.95%可用性。

图 3-1 DLI Serverless 架构



与传统自建Hadoop集群相比，Serverless架构的DLI还具有以下优势：

表 3-1 Serverless DLI 与传统自建 Hadoop 集群对比的优势

优势	维度	数据湖探索 DLI	自建Hadoop系统
低成本	资金成本	按照实际扫描数据量或者CU时收费，可变成本，成本可节约50%。	长期占用资源，资源浪费严重，成本高。
	弹性扩缩容能力	基于容器化Kubernetes，具有极致的弹性伸缩能力。	无。
免运维	运维成本	即开即用，Serverless架构。	需要较强的技术能力进行搭建、配置、运维。
	高可用	具有跨AZ容灾能力。	无
高易用	学习成本	学习成本低，包含10年、上千个项目经验固化的调优参数。同时提供可视化智能调优界面。	学习成本高，需要了解上百个调优参数。
	支持数据源	• 云上：OBS、RDS、DWS、CSS、MongoDB、Redis。 • 云下：自建数据库、MongoDB、Redis。	• 云上：OBS。 • 云下：HDFS。
	生态兼容	DLV、永洪BI、帆软。	大数据生态工具。
	自定义镜像	支持，满足业务多样性。	无。
	工作流调度	DataArts Studio-DLF调度。	自建大数据生态的调度工具，如Airflow。
	企业级多租户	基于表的权限管理，可以精细化到列权限。	基于文件的权限管理。
高性能	性能	基于软硬件一体化的深度垂直优化。	大数据开源版本性能。

跨源分析

支持多种数据格式，云上多种数据源、ECS自建数据库以及线下数据库，数据无需搬迁，即可实现对云上多个数据源进行分析，构建企业的统一视图，帮助企业快速完成业务创新和数据价值探索。

4 应用场景

DLI服务适用于海量日志分析、异构数据源联邦分析、大数据ETL处理。

海量日志分析

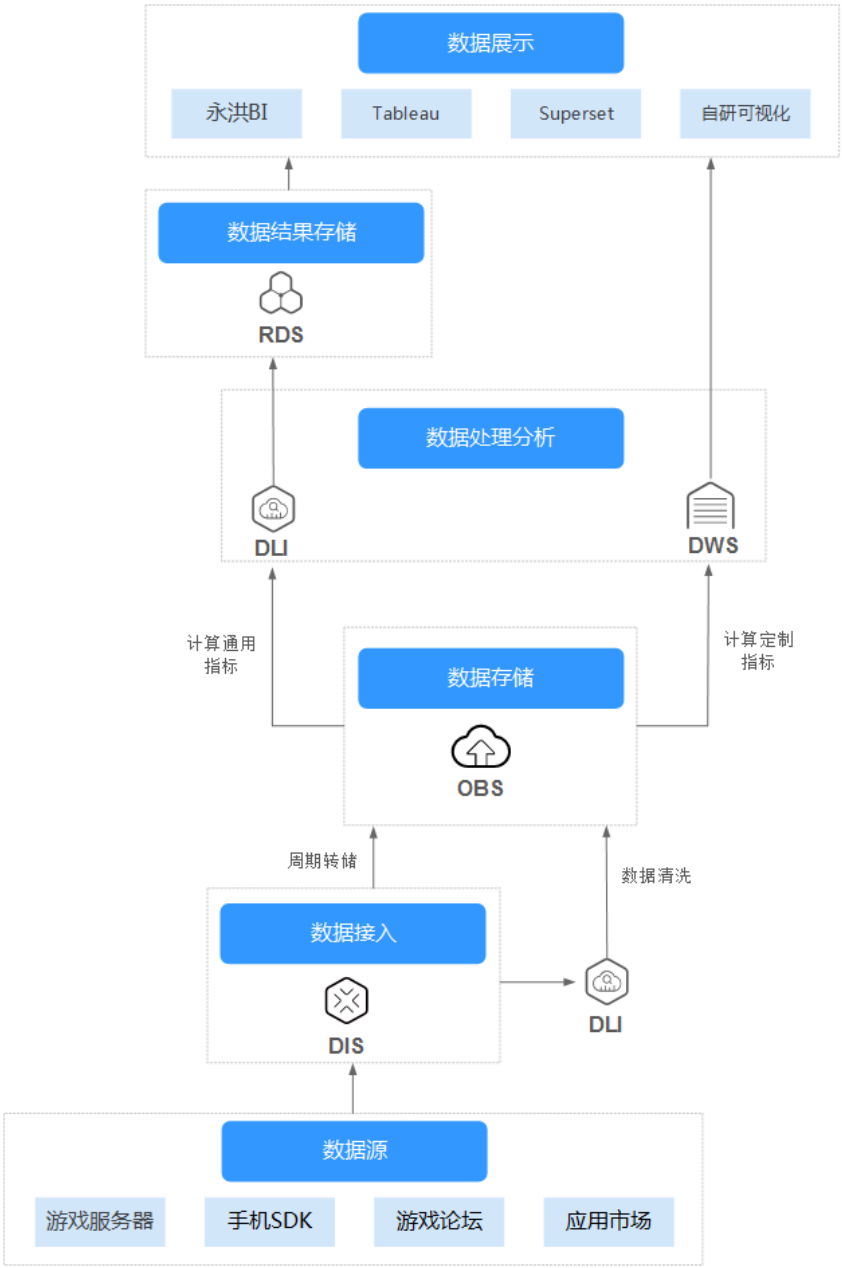
- 游戏运营数据分析

  游戏公司不同部门日常通过游戏数据分析平台，分析每日新增日志获取所需指标，通过数据来辅助决策。例如：运营部门通过平台获取新增玩家、活跃玩家、留存率、流失率、付费率等，了解游戏当前状态及后续响应活动措施；投放部门通过平台获取新增玩家、活跃玩家的渠道来源，来决定下一周期重点投放哪些平台。

- 优势
  - 高效的Spark编程模型：使用DLI直接从DIS中获取数据，进行数据清理等预处理操作。只需编写处理逻辑，无需关心多线程模型。
  - 简单易用：直接使用标准SQL编写指标分析逻辑，无需关注背后复杂的分布式计算平台。
  - 按需计费：日志分析按时效性要求按周期进行调度，每次调度之间存在大量空闲期。DLI按需计费只在使用期间收费，成本较独占队列降低50%以上。
- 建议搭配以下服务使用

  OBS，DIS，DWS，RDS

图 4-1 游戏运营数据分析



异构数据源联邦分析

- 车企数字化服务转型

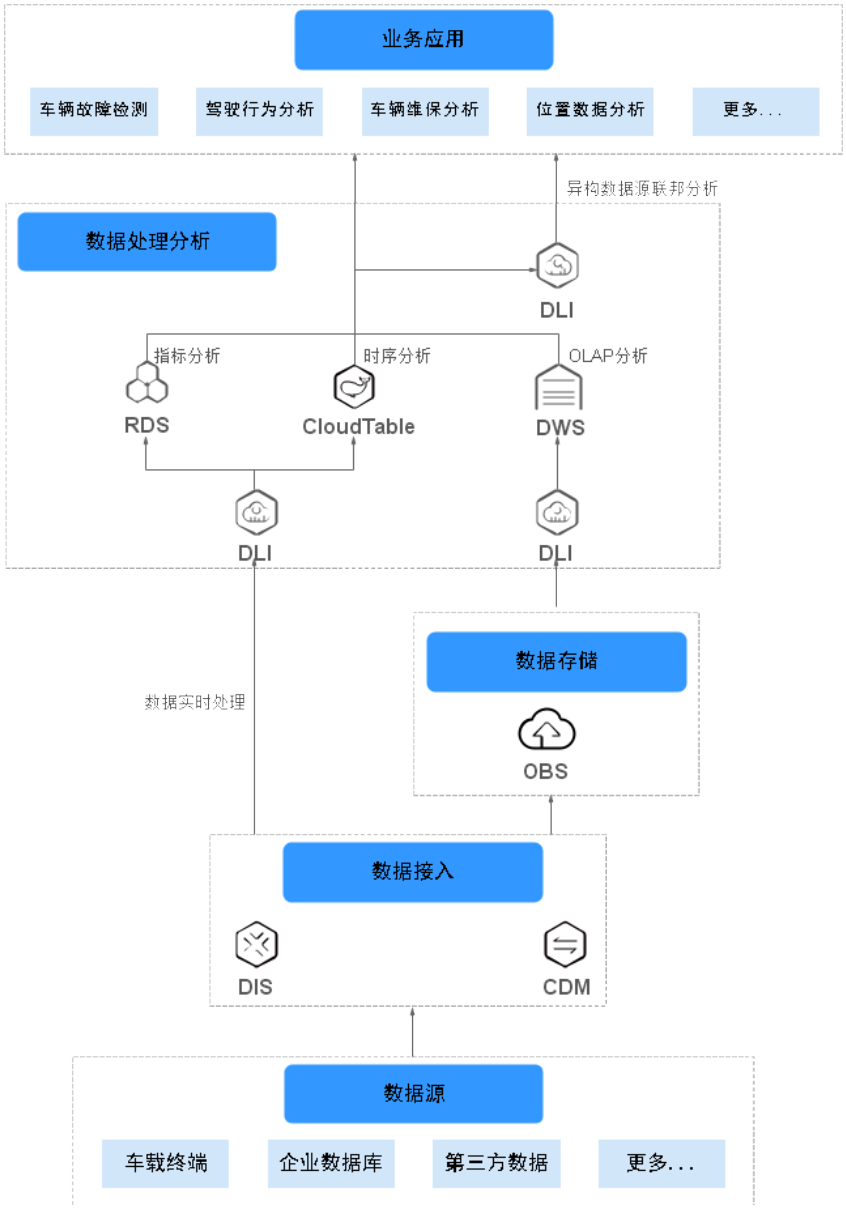
  面临市场新的竞争压力及出行服务不断变革，车企通过构建车联云平台和车机OS，将互联网应用与用车场景打通，完成车企数字化服务转型，从而为车主提供更好的智联出行体验，增加车企竞争力，促进销量增长。例如：通过对车辆日常指标数据（电池、发动机，轮胎胎压、安全气囊等健康状态）的采集和分析，及时将维保建议回馈给车主。

- 优势
  - 多源数据分析免搬迁：关系型数据库RDS中存放车辆和车主基本信息，表格存储CloudTable中存放实时的车辆位置和健康状态信息，数据仓库DWS中存放周期性统计的指标。通过DLI无需数据搬迁，对多数据源进行联邦分析。

  - 数据分级存储：车企需要保留全量历史数据支撑审计类等业务，低频进行访问。温冷数据存放在低成本的对象存储服务OBS上，高频访问的热数据存放在数据引擎（CloudTable和DWS）中，降低整体存储成本。
  - 告警快速敏捷触发服务器弹性伸缩：对CPU、内存、硬盘空间和带宽无特殊要求。
- 建议搭配以下服务使用

  DIS、CDM、OBS、DWS、RDS、CloudTable

图 4-2 车企数字化服务转型



大数据 ETL 处理

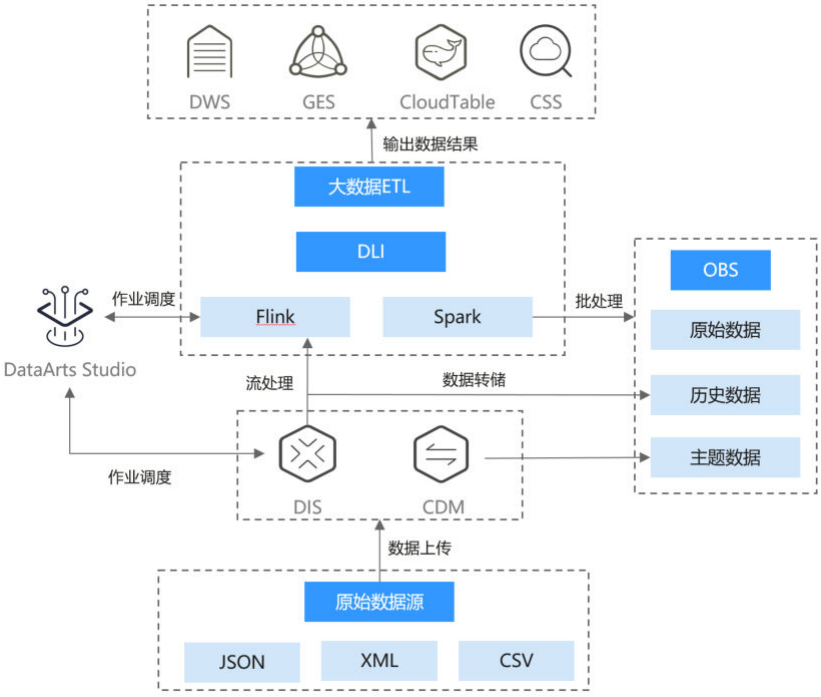
- 运营商大数据分析

  运营商数据体量在PB~EB级，其数据种类多，有结构化的基站信息数据，非结构化的消息通信数据，同时对数据的时效性有很高的要求，DLI服务提供批处理、流处理等多模引擎，打破数据孤岛进行统一的数据分析。

- 优势
  - 大数据ETL：具备TB~EB级运营商数据治理能力，能快速将海量运营商数据做ETL处理，为分布式批处理计算提供分布式数据集。
  - 高吞吐低时延：采用Apache Flink的Dataflow模型，高性能计算资源，从用户自建的Kafka、MRS-Kafka、DMS-Kafka消费数据，单CU每秒吞吐1千~2万条消息。
  - 细粒度权限管理：P公司内部有N个子部门，子部门之间需要对数据进行共享和隔离。DLI支持计算资源按租户隔离，保障作业SLA；支持数据权限控制到表/列，帮助企业实现部门间数据共享和权限管理。
- 建议搭配以下服务使用

  OBS、DIS、DataArts Studio

图 4-3 运营商大数据分析



地理大数据分析

- 地理大数据分析

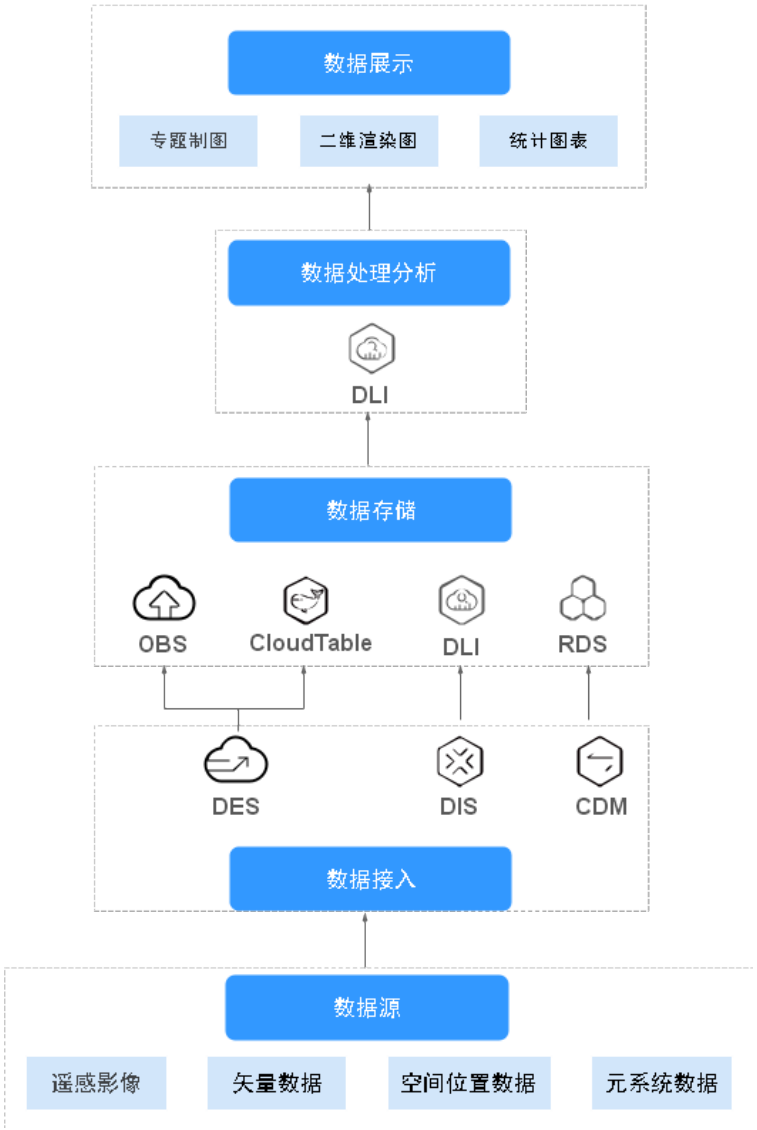
  地理大数据具有大数据的相关特征，数据体量巨大，例如，全球卫星遥感影像数据量达到PB级。数据种类多，有结构化的遥感影像栅格数据、矢量数据，非结构化的空间位置数据、三维建模数据；在大体量的地理大数据中，通过高效的挖掘工具或者挖掘方法实现价值提炼，是用户非常关注的话题。
- 优势
  - 提供地理专业算子：支持全栈Spark能力，具备丰富的Spark空间数据分析算法算子，全面支持结构化的遥感影像数据、非结构化的三维建模、激光点云等巨量数据的离线批处理，支持带有位置属性的动态流数据实时计算处理。
  - CEP SQL：提供地理位置分析函数对地理空间数据进行实时分析，用户仅需编写SQL便可实现例如偏航检测，电子围栏等地理分析场景。

- 大数据治理能力：能快速将海量遥感影像数据接入上云，快速完成影像数据切片处理，为分布式批处理计算提供弹性分布式数据集。

- 建议搭配以下服务使用

  DIS、CDM、DES、OBS、RDS、CloudTable

图 4-4 地理大数据分析



5 约束与限制

作业相关约束限制

- DLI支持的作业类型：Spark SQL、SparkJar、Flink SQL、Flink Jar
- DLI支持的Spark版本：Spark 3.3.1、Spark 3.1.1（EOM）、Spark 2.4.5（EOM）、Spark 2.3（EOS）
- DLI支持的Flink版本：Flink Jar 1.15、Flink 1.12（EOM）、Flink 1.10（EOS）、Flink 1.7（EOS）
- SQL作业支持Spark和HetuEngine两种引擎。
  - Spark：显示执行引擎为“Spark”的作业。
  - HetuEngine：显示执行引擎为“HetuEngine”的作业。
- DLI配置SparkUI只展示最新的100条作业信息。
- 控制台界面查询结果最多显示1000条作业结果数据，如果需要查看更多或者全量数据，则可以通过该功能将数据导出到OBS获取。
- 导出作业运行日志需要具有OBS桶的权限，请提前在“全局配置 > 工程配置”页面配置DLI作业桶。
- default队列下运行的作业或者该作业为同步作业时不支持归档日志操作。
- 仅Spark作业支持使用自定义镜像。了解**自定义镜像**。
- 当前弹性资源池最大的计算资源 32000CUs。
- 弹性资源池中可创建队列的最小CU：
  - 通用队列：4CUs
  - SQL队列：Spark SQL队列：8CUs；HetuEngine SQL队列：16CUs

DLI 套餐包使用约束限制

- 套餐包购买后区域固定无法更换，购买的套餐只能在绑定的区域使用，不能在非绑定区域使用。
- 套餐包购买后不支持退订。
- 计费时优先使用套餐中的资源，套餐中资源使用完后，超出部分按需付费。
- 套餐包不支持抵扣已使用的资源。
- 套餐包到期后，按需资源不会自动关闭，将会以按需付费的方式继续使用。
- 存储套餐的额度每个小时会重置。其他类型套餐包额度按月重置。

队列使用约束限制

- DLI服务预置了名为“default”的队列供用户体验，资源的大小按需分配。运行作业时按照用户每个作业的数据扫描量（单位为“GB”）收取计算费用。
- 队列类型：
  - SQL类型队列：SQL队列支持提交Spark SQL作业。
  - 通用队列：支持Spark程序、Flink SQL、Flink Jar作业。

  不支持队列类型切换，如需使用其他队列类型，请重新购买新的队列。
- 不支持切换队列的计费模式。
- 队列不支持切换区域。
- 16CUs队列不支持扩容和缩容。
- 64CUs队列不支持缩容。
- 创建队列时，仅支持包年包月队列和按需专属队列选择跨AZ双活，且跨AZ的队列价格为单AZ模式下的2倍。
- 新创建的队列需要运行作业后才可进行扩缩容。
- DLI队列不支持访问公网。

更多队列使用约束限制请参考**队列使用约束限制**。

弹性资源池使用约束限制

- 不支持切换弹性资源池的计费模式。
- 弹性资源池不支持切换区域。
- 按需计费的弹性资源池默认勾选专属资源模式，自创建起按自然小时收费。
- Flink 1.10及其以上版本的作业支持在弹性资源池运行。
- 弹性资源池网段设置后不支持更改。
- 弹性资源池关联队列：
  - 仅支持关联按需计费模式的队列（包括专属队列）。
  - 队列和弹性资源池状态正常，资源未被冻结。
- 当前仅支持包年包月计费模式的弹性资源池进行规格变更。
- 仅支持查看30天以内的弹性资源池扩缩容历史。
- 弹性资源池不支持访问公网。
- 弹性资源池CU设置、弹性资源池中添加/删除队列、修改弹性资源池中队列的扩缩容策略、系统自动触发弹性资源池扩缩容时都会引起弹性资源池CU的变化，部分情况下系统无法保证按计划扩容/缩容至目标CUs：
  - 弹性资源池扩容时，可能会由于物理资源不足导致弹性资源池无法扩容到设定的目标大小。
  - 弹性资源池缩容时，系统不保证将队列资源完全缩容到设定的目标大小。

    在执行缩容任务时，系统会先检查资源使用情况，判断是否存在缩容空间，如果现有资源无法按照最小缩容步长执行缩容任务，则弹性资源池可能缩容不成功，或缩容一部分规格的情况。

    因资源规格不同可能有不同的缩容步长，通常是16CUs、32CUs、48CUs、64CUs等。

    示例：弹性资源池规格为192CUs，资源池中的队列执行作业占用了68CUs，计划缩容至64CUs。

执行缩容任务时，系统判断剩余124CUs，按64CUs的缩容步长执行缩容任务，剩余60CUs资源无法继续缩容，因此弹性资源池执行缩容任务后规格为128CUs。

更多弹性资源池约束限制请参考**弹性资源池使用约束限制**。

DLI 存储资源使用约束限制

DLI提供了存储资源的能力，用于存储数据库和DLI表，DLI存储按存储数据量计费。

资源相关约束限制

- **数据库**
  - “default”为内置数据库，不能创建名为“default”的数据库。
  - DLI支持创建的数据库的最大数量为50个。
- **数据表**
  - DLI支持创建的表的最大数量为5000个。
  - DLI支持创建表类型：
    - Managed：数据存储位置为DLI的表。
    - External：数据存储位置为OBS的表。
    - View：视图，视图只能通过SQL语句创建。
    - 跨源表：表类型同样为External。
  - 创建DLI表时不支持指定存储路径。
- **数据导入**
  - 仅支持将OBS上的数据导入DLI或OBS中。
  - 支持将OBS中CSV，Parquet，ORC，JSON和Avro格式的数据导入到在DLI中创建的表。
  - 将CSV格式数据导入分区表，需在数据源中将分区列放在最后一列。
  - 导入数据的编码格式仅支持UTF-8。
- **数据导出**
  - 只支持将DLI表（表类型为“Managed”）中的数据导出到OBS桶中，且导出的路径必须指定到文件夹级别。
  - 导出文件格式为json格式，且文本格式仅支持UTF-8。
  - 支持跨账号导出数据，即B账户对A账户授权后，A账户拥有B账户OBS桶的元数据信息和权限信息的读取权限，以及路径的读写权限，则A账户可将数据导出至B账户的OBS路径中。
- **程序包**
  - 程序包支持删除，但不支持删除程序包组。
  - 支持上传的程序包类型：
    - JAR：用户jar文件。
    - PyFile：用户Python文件。
    - File：用户文件。

- ModelFile：用户AI模型文件。

更多资源相关约束限制请参考**数据管理**。

增强型跨源连接约束限制

- 在同一队列中，如果同时使用了经典型跨源连接和增强型跨源连接，则经典型跨源连接优先于增强型跨源连接。推荐使用增强型跨源连接。
- DLI提供的default队列不支持创建跨源连接。
- Flink作业访问DIS，OBS和SMN数据源，无需创建跨源连接，可以直接访问。
- 增强型跨源仅支持包年包月队列、按需计费模式下的专属队列。
- 增强型跨源连接需要使用VPC、子网、路由、对等连接功能，因此需要获得VPC（虚拟私有云）的VPC Administrator权限。

  可在**服务授权**中进行设置。
- 使用DLI增强型跨源时，弹性资源池/队列的网段与数据源网段不能重合。
- 访问跨源表需要使用已经创建跨源连接的队列。
- 跨源表不支持Preview预览功能。
- 检测跨源连接的连通性时对IP约束限制如下：
  - IP必须为合法的IP地址，用“.”分隔的4个十进制数，范围是0-255。
  - 测试时IP地址后可选择添加端口，用":"隔开，端口最大限制5位，端口范围：0~65535。

    例如192.168.xx.xx或者192.168.xx.xx:8181。
- 检测跨源连接的连通性时对域名约束限制如下：
  - 域名的限制长度为1到255的字符串，并且组成必须是字母、数字、下划线或者短横线。
  - 域名的顶级域名至少包含两个及以上的字母，例如.com，.net，.cn等。
  - 测试时域名后可选择添加端口，用":"隔开，端口最大限制为5位，端口范围：0~65535。

    例如example.com:8080。

更多增强型跨源连接约束限制请参考**增强型跨源连接概述**。

跨源认证使用约束限制

- 仅Spark SQL、和Flink OpenSource SQL 1.12版本的作业支持使用跨源认证。
- 仅在2023年5月1日后创建的队列，支持Flink作业使用跨源认证。
- DLI支持四种类型的跨源认证，不同的数据源按需选择相应的认证类型。
  - CSS类型跨源认证：适用于“6.5.4”及以上版本的CSS集群且集群已开启安全模式。
  - Kerberos类型的跨源认证：适用于开启Kerberos认证的MRS安全集群。
  - Kafka_SSL类型的跨源认证：适用于开启SSL的Kafka。
  - Password类型的跨源认证：适用于DWS、RDS、DDS、DCS数据源。

更多跨源认证约束限制请参考**跨源认证简介**。

SQL 语法相关约束限制

- SQL语法限制
  - 不支持在创建DLI表时指定存储路径。
- SQL语句大小限制
  - 须小于500000字符。
  - 须小于1MB。

其他约束限制

- DLI配额相关约束限制请参考**配额管理**操作指导。
- 建议使用支持的浏览器登录DLI服务。
  - Google Chrome : 43.0及更高版本
  - Mozilla FireFox : 38.0及更高版本
  - Internet Explorer : 9.0及更高版本

  更多浏览器的兼容性列表请参考**支持的浏览器有哪些?**

6 产品规格

弹性资源池产品规格

弹性资源池为DLI作业运行提供所需的计算资源（CPU和内存）。弹性资源池的单位为CU，1CU包含1CPU和4GB内存。

您可以在弹性资源池中创建多个队列， 队列之间的计算资源支持共享。 通过合理设置队列的计算资源池分配策略，提高计算资源利用率。

DLI提供以下规格的计算资源，如表6-1所示。

表 6-1 弹性资源池规格

类型	规格	约束限制	适用场景
基础版	16-64CUs规格	• 不支持高可靠与高可用。 • 不支持设置队列属性和作业优先级。 • 不支持对接Notebook实例。 其他弹性资源池使用相关约束限制请参考弹性资源池使用约束限制。	适用于对资源消耗不高、对资源高可靠性和高可用性要求不高的测试场景。
标准版	64CUs及以上规格	弹性资源池使用相关约束限制请参考弹性资源池使用约束限制。	具备强大的计算能力、高可用性、及灵活的资源管理能力，适用于大规模计算任务场景和有长期资源规划需求的业务场景。

7 安全

7.1 责任共担

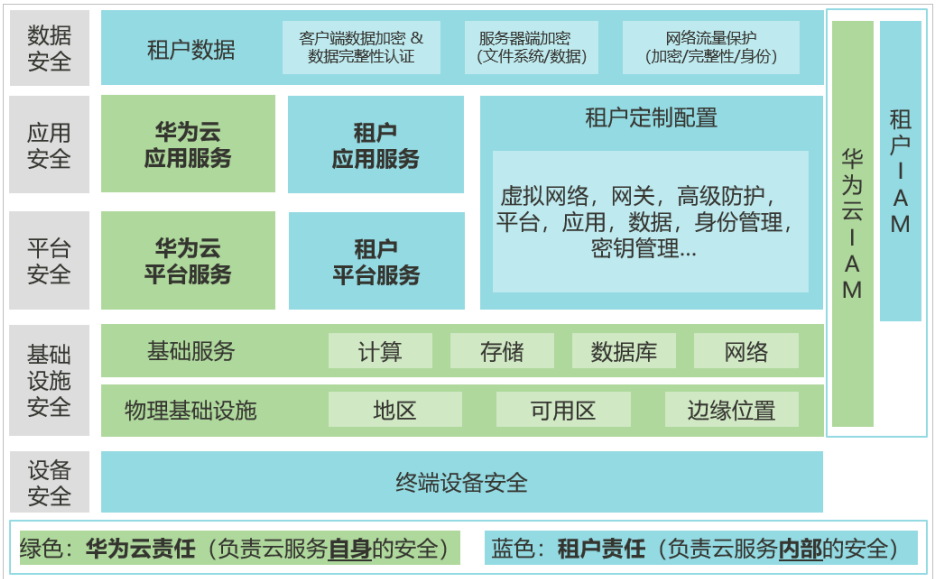
华为云秉承“将公司对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

安全性是华为云与您的共同责任，如图7-1所示。

- **华为云**：负责云服务**自身**的安全，提供安全的云。华为云的安全责任在于保障其所提供的IaaS、PaaS和SaaS类云服务自身的安全，涵盖华为云数据中心的物理环境设施和运行其上的基础服务、平台服务、应用服务等。这不仅包括华为云基础设施和各项云服务技术的安全功能和性能本身，也包括运维运营安全，以及更广义的安全合规遵从。
- **租户**：负责云服务**内部**的安全，安全地使用云。华为云租户的安全责任在于对使用的IaaS、PaaS和SaaS类云服务内部的安全以及对租户定制配置进行安全有效的管理，包括但不限于虚拟网络、虚拟主机和访客虚拟机的操作系统，虚拟防火墙、API网关和高级安全服务，各项云服务，租户数据，以及身份账号和密钥管理等方面的安全配置。

《华为云安全白皮书》详细介绍华为云安全性的构建思路与措施，包括云安全战略、责任共担模型、合规与隐私、安全组织与人员、基础设施安全、租户服务与租户安全、工程安全、运维运营安全、生态安全。

图 7-1 华为云安全责任共担模型



7.2 资产识别与管理

DLI 可以通过标签实现资源的标识与管理。

使用场景

通常您的业务系统可能使用了华为云的多种云服务，您可以为这些云服务下不同的资源实例分别设置标签，各服务的计费详单会体现这些资源实例设置的标签。如果您的业务系统是由多个不同的应用构成，为同一种应用拥有的资源实例设置统一的标签将很容易帮助您对不同的应用进行使用量分析和成本核算。

对DLI来说，标签用于标识购买的队列和创建数据库，对购买的DLI队列和数据库进行分类。为队列或数据库添加标签时，该队列或数据库上所有请求产生的计费话单里都会带上这些标签，您可以针对话单报表做分类筛选，进行更详细的成本分析。

例如：某个队列作用于A部门，我们可以用该部门名称作为标签，设置到被使用的集群上。在分析话单时，就可以通过标签分析该部门的开发使用成本。

DLI以键值对的形式描述标签。一个队列默认20个标签。每个标签有且只有一对键值。键和值可以任意顺序出现在标签中。同一个集群标签的键不能重复，但是值可以重复，并且可以为空。

使用方式

DLI支持通过控制台方式创建队列和数据库标签，详情请参见**队列标签管理**。

7.3 身份认证与访问控制

身份认证

用户访问DLI的方式主要有两种，包括DLI Console界面、DLI Open API等，其本质都是通过DLI提供的REST API接口进行请求。

DLI的接口均需要通过认证鉴权才能访问，控制台发送的请求与调用API接口的请求均支持Token认证鉴权。

访问控制

您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制华为云资源的访问。

关于IAM的详细介绍，请参见**IAM产品介绍**。

权限根据授权精细程度分为角色和策略。

- 角色：IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于华为云各服务之间存在业务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。
- 策略：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。例如：不允许某用户组删除集群，仅允许DLI基本操作（如创建、查询作业等）。

  DLI支持的授权项请参见**权限管理概述**。

如表3-1所示，包括了DLI的所有系统权限。

系统角色/策略名称	描述	类别	授权方式
DLI FullAccess	数据湖探索所有权限。	系统策略	具体的授权方式请参考**创建IAM用户并授权使用DLI**以及《**如何创建子用户**》和《**如何修改用户策略**》。
DLI ReadOnlyAccess	数据湖探索只读权限。	系统策略	
Tenant Administrator	租户管理员。 • 操作权限：具有所有云服务的管理和使用权限。创建后，可通过ACL赋权给其他子用户使用。 • 作用范围：项目级服务。	系统角色	
DLI Service Administrator	DLI服务管理员。 • 操作权限：具有数据湖探索服务队列、数据的管理和使用权限。创建后，可通过ACL赋权给其他子用户使用。 • 作用范围：项目级服务。	系统角色	

7.4 数据保护技术

数据存储安全

为了确保您的个人敏感数据（例如用户名、密码、手机号码等）不被未经过认证、授权的实体或者个人获取，DLI对用户数据的存储和传输进行加密保护，以防止个人数据泄露，保证您的个人数据安全。

数据销毁机制

用户删除DLI队列后，存储在集群上的用户个人敏感数据会随之删除。

用户在控制台上删除填写的手机号、邮箱，并关闭消息通知功能后，数据库中会同步删除用户的手机号、邮箱信息。

数据传输安全

用户个人敏感数据将通过TLS 1.2、TLS1.3进行传输中加密，所有华为云DLI服务的API调用都支持 HTTPS 来对传输中的数据进行加密。

Spark 作业传输通信加密

Spark作业支持通过配置表7-1中的参数开启通信加密。

说明

请确保已上传密钥和证书到指定的OBS路径下，并在作业配置中的其他依赖文件中引入。

表 7-1 Spark 作业传输开启通信加密配置项

参数	说明	配置示例
spark.network.crypto.enabled	该参数用于启用或禁用数据在节点之间传输时的加密。当设置为true时，Spark会加密Executor和Driver之间以及Executor之间的所有通信。这是确保数据传输安全的重要配置。	true
spark.network.sasl.serverAlwaysEncrypt	该参数用于配置服务器端是否使用加密来与客户端通信。当设置为true时，服务器将要求所有客户端使用加密连接，这可以提高通信的安全性。	true
spark.authenticate	该参数用于配置是否对Spark应用程序的组件进行身份验证。启用身份验证可以防止未授权的访问。这个参数可以设置为true来启用身份验证。	true

Flink 作业传输通信加密

在Flink作业可以通过配置表7-2中的参数来开启SSL传输。

说明

- 打开Task Manager之间data传输通道的SSL，会对性能会有较大影响，建议结合安全和性能综合考虑是否开启。
- 证书文件还需要在作业配置页面的“其他依赖文件”中完成配置。
- OBS路径/opt/flink/usrlib/userData/为默认存储依赖文件路径。
- 请确保已上传密钥和证书到指定的OBS路径下，并在作业配置中的其他依赖文件中引入。

表 7-2 Flink 作业传输通信加密配置参数

参数	说明	是否必须	配置示例
security.ssl.enabled	打开SSL总开关。	是	true
akka.ssl.enabled	打开akka SSL开关。	否	true
blob.service.ssl.enabled	打开blob通道SSL开关。	否	true
taskmanager.data.ssl.enable	打开taskmanager之间通信的SSL开关。	否	true
security.ssl.algorithms	设置SSL加密的算法。	否	TLS_DHE_RSA_WITH_AES_128_GCM_SHA256,TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256,TLS_DHE_RSA_WITH_AES_256_GCM_SHA384,TLS_ECDHE_RSA_WITH_AES_256_GCM_SHA384
security.ssl.keystore	keystore的存放路径，“flink.keystore”表示用户通过generate_keystore.sh*工具生成的keystore文件名称。	是	/opt/flink/usrlib/userData/flink.keystore
security.ssl.keystore-password	keystore的password，-表示需要用户输入自定义设置的密码值。	是	-

参数	说明	是否必须	配置示例
security.ssl.key-password	ssl key的password，-表示需要用户输入自定义设置的密码值。	是	-
security.ssl.truststore	truststore存放路径，“flink.truststore”表示用户通过generate_keystore.sh*工具生成的truststore文件名称。	是	/opt/flink/usrlib/userData/flink.truststore
security.ssl.truststore-password	truststore的password，-表示需要用户输入自定义设置的密码值。	是	-
security.ssl.rest.enabled	REST API接口是否启用SSL/TLS加密。	是	false
security.ssl.verify-hostname	用于控制在建立SSL/TLS连接时是否验证对端的主机名（hostname）与证书中的信息是否匹配。	否	false
security.ssl.protocol	指定SSL/TLS连接时所使用的协议版本	否	TLSv1.2、TLSv1.3
security.ssl.encrypt.enabled	Flink集群内部以及与其他组件之间通信时是否启用数据加密	否	false

开启Flink作业传输通信加密配置示例：

```
security.ssl.enabled: true
security.ssl.encrypt.enabled: false
security.ssl.key-password: ***
security.ssl.keystore-password: Admin12!
security.ssl.keystore:  /opt/flink/usrlib/userData/*.keystore
security.ssl.protocol: TLSv1.2
security.ssl.rest.enabled: false
security.ssl.truststore-password: ***
security.ssl.truststore:  /opt/flink/usrlib/userData/*.truststore
security.ssl.verify-hostname: false
```

7.5 审计与日志

- **DLI对接云审计服务**

云审计服务（Cloud Trace Service，CTS），是华为云安全解决方案中专业的日志审计服务，提供对各种云资源操作记录的收集、存储和查询功能，可用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。

CTS可记录的DLI操作列表详见**云审计服务支持的DLI操作列表说明**。用户开通云审计服务并创建和配置追踪器后，CTS开始记录操作事件用于审计。关于如何开通云审计服务以及如何查看追踪事件，请参考《**云审计服务快速入门**》中的相关章节。

CTS支持**配置关键操作通知**。用户可将与IAM相关的高危敏感操作，作为关键操作加入到CTS的实时监控列表中进行监控跟踪。当用户使用DLI服务时，如果触发了监控列表中的关键操作，那么CTS会在记录操作日志的同时，向相关订阅者实时发送通知。

- **DLI的作业日志**

  在创建DLI作业时，可以在作业编辑页面，通过保存作业日志功能，将作业运行时的日志信息保存到OBS。

  查询作业日志信息，参考**查看DLI SQL日志**。

  作业日志为日常的服务运维提供了重要保障，包括跟踪资源使用情况、检测作业运行安全性、追踪资源消耗、检测错误等。

7.6 服务韧性

DLI通过流量限制、跨AZ容灾、备份恢复等技术方案，保障数据的持久性和可靠性。

- 流量限制：DLI通过设置流量控制机制，防止服务过载并保持服务的稳定性。
- 跨AZ容灾：DLI云服务采用跨可用区容灾部署，减少单点故障的风险，提高系统的可用性和弹性。
- 备份恢复：DLI自动化的备份策略和恢复计划，确保在发生故障时可以迅速恢复服务和数据。

7.7 监控安全风险

云监控服务为用户的云上资源提供了立体化监控平台。通过云监控您可以全面了解云上的资源使用情况、业务的运行状况，并及时收到异常告警做出反应，保证业务顺畅运行。

DLI服务提供基于云监控服务CES的资源监控能力

DLI已对接云监控服务，提供基于云监控服务的资源监控能力，帮助用户监控账号下的DLI队列，执行自动实时监控、告警和通知操作。用户可以实时掌握队列中的运行作业网络流入速率、网络流出速率、CPU使用率、内存使用率、磁盘利用率、失败作业率、等待作业数等信息。还可以通过云监控服务提供的管理控制台或API接口来检索数据湖探索服务产生的监控指标和告警信息。

关于DLI支持的监控指标请参见**数据湖探索监控指标说明及查看指导**。

7.8 故障恢复

- **系统级故障恢复**

  DLI系统采用存算分离的架构，计算集群基于K8s资源调度和故障切换机制，在系统故障时，支持自动故障恢复。

- **作业级故障恢复**

  Flink、Spark作业支持配置自动重启恢复机制，在开启自动重启功能后，当作业出现异常时将自动重启恢复作业。

7.9 更新管理

更新漏洞

DLI云服务通过华为云安全公告密切跟踪漏洞，如Apache Log4j2 远程代码执行漏洞（CVE-2021-44228）、Fastjson存在反序列化漏洞（CNVD-2022-40233）等。

一旦发现服务模块涉及漏洞影响，会迅速通过官方解决方案升级现网更新漏洞。

更新配置

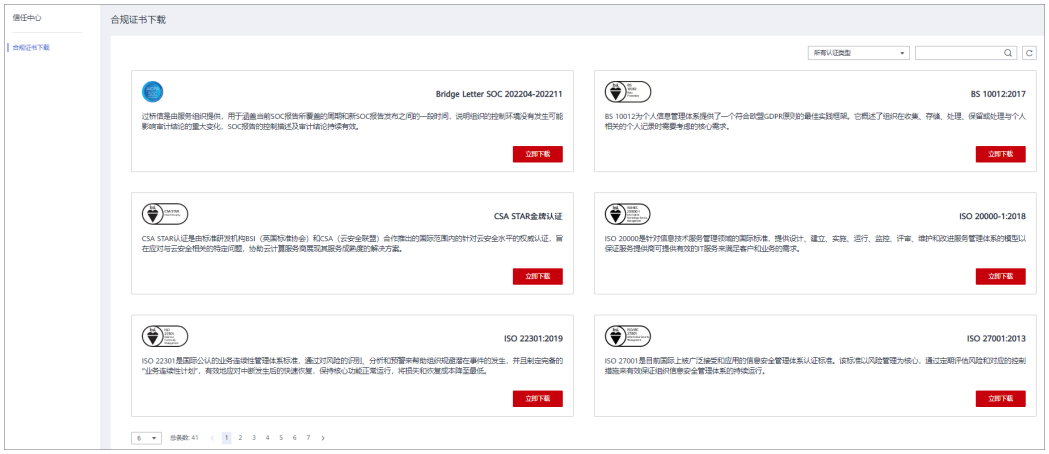
DLI云服务通过版本更新升级更新配置，确保服务的安全性和稳定性。

7.10 认证证书

合规证书

华为云服务及平台通过了多项国内外权威机构（ISO/SOC/PCI等）的安全合规认证，用户可自行**申请下载**合规资质证书。

图 7-2 合规证书下载



资源中心

华为云还提供以下资源来帮助用户满足合规性要求，具体请查看**资源中心**。

图 7-3 资源中心

销售许可证&软件著作权证书

另外，华为云还提供了以下销售许可证及软件著作权证书，供用户下载和参考。具体请查看**合规资质证书**。

图 7-4 销售许可证&软件著作权证书

8 权限管理

在华为云上购买DLI资源后，如果您需要给企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全地控制华为云资源的访问。

通过IAM，您可以在账号中给员工创建IAM用户，并使用策略来控制他们对华为云资源的访问范围。例如您的员工中有负责软件开发的人员，您希望他们拥有DLI的使用权限，但是不希望他们拥有删除DLI等高危操作的权限，那么您可以使用IAM为开发人员创建用户，通过授予仅能使用DLI，但是不允许删除DLI的权限策略，控制他们对DLI资源的使用范围。

如果账号已经能满足您的需求，不需要创建独立的IAM用户进行权限管理，您可以跳过本章节，不影响您使用DLI服务的其他功能。

IAM是华为云提供权限管理的基础服务，无需付费即可使用，您只需要为您账号中的资源进行付费。关于IAM的详细介绍，请参见**《IAM产品介绍》**。

DLI 权限

默认情况下，管理员创建的IAM用户没有任何权限，您需要将其加入用户组，并给用户组授予策略或角色，才能使得该用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

DLI部署时通过物理区域划分，为项目级服务。授权时，“作用范围”需要选择“区域级项目”，然后在指定区域对应的项目中设置相关权限，并且该权限仅对此项目生效；如果在“所有项目”中设置权限，则该权限在所有区域项目中都生效。访问DLI时，需要先切换至授权区域。

权限类别：根据授权精程度分为角色和策略。

- 角色：IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于华为云各服务之间存在业务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。
- 策略：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。例如：针对DLI服务，管理员能够控制IAM用户仅能对某一类云服务器资源进行指定的管理操作。DLI支持的API授权项请参见《**权限策略和授权项**》。

表 8-1 DLI 系统权限

系统角色/策略名称	描述	类别	依赖关系
DLI FullAccess	数据湖探索所有权限。	系统策略	该角色有依赖，需要在同项目中勾选依赖的角色： ● 创建跨源连接：VPC ReadOnlyAccess ● 创建包年/包月资源：BSS Administrator ● 创建标签：TMS FullAccess、EPS EPS FullAccess ● 使用OBS存储：OBS OperateAccess ● 创建委托：Security Administrator
DLI ReadOnlyAccess	数据湖探索只读权限。 只读权限可控制部分开放的、未鉴权的DLI资源和操作。例如创建全局变量、创建程序包以及程序包组、default队列提交作业、default数据库下建表、创建跨源连接、删除跨源连接等操作。	系统策略	无
Tenant Administrator	租户管理员。 ● 操作权限：具有数据湖探索服务资源的所有执行权限。创建后，可通过ACL赋权给其他子用户使用。 ● 作用范围：项目级服务。	系统角色	无
DLI Service Administrator	数据湖探索管理员。 ● 操作权限：具有数据湖探索服务资源的所有执行权限。创建后，可通过ACL赋权给其他子用户使用。 ● 作用范围：项目级服务。	系统角色	无

表8-2列出了DLI SQL常用操作与系统权限的授权关系，您可以参照该表选择合适的系统策略。

更多SQL语法赋权请参考《数据湖探索SQL语法参考》>《数据控制》>《权限列表》章节。

表 8-2 DLI 常用操作与系统权限的关系

资源	操作	说明	DLI FullAccess	DLI ReadOnlyAccess	Tenant Administrator	DLI Service Administrator
队列	DROP_QUEUE	删除队列	√	×	√	√
	SUBMIT_JOB	提交作业	√	×	√	√
	CANCEL_JOB	终止作业	√	×	√	√
	RESTART	重启队列	√	×	√	√
	GRANT_PRIVILEGE	队列的赋权	√	×	√	√
	REVOKE_PRIVILEGE	队列权限的回收	√	×	√	√
	SHOW_PRIVILEGES	查看其他用户具备的队列权限	√	×	√	√
数据库	DROP_DATABASE	删除数据库	√	×	√	√
	CREATE_TABLE	创建表	√	×	√	√
	CREATE_VIEW	创建视图	√	×	√	√
	EXPLAIN	将SQL语句解释为执行计划	√	×	√	√
	CREATE_ROLE	创建角色	√	×	√	√
	DROP_ROLE	删除角色	√	×	√	√
	SHOW_ROLES	显示角色	√	×	√	√
	GRANT_ROLE	绑定角色	√	×	√	√

资源	操作	说明	DLI FullAccess	DLI ReadOnlyAccess	Tenant Administrator	DLI Service Administrator
	REVOKE_ROLE	解除角色绑定	√	×	√	√
	SHOW_USERS	显示所有角色和用户的绑定关系	√	×	√	√
	GRANT_PRIVILEGE	数据库的赋权	√	×	√	√
	REVOKE_PRIVILEGE	数据库权限的回收	√	×	√	√
	SHOW_PRIVILEGES	查看其他用户具备的数据库权限	√	×	√	√
	DISPLAY_ALL_TABLES	显示数据库中的表	√	√	√	√
	DISPLAY_DATABASE	显示数据库	√	√	√	√
	CREATE_FUNCTION	创建函数	√	×	√	√
	DROP_FUNCTION	删除函数	√	×	√	√
	SHOW_FUNCTIONS	显示所有函数	√	×	√	√
	DESCRIBE_FUNCTION	显示函数详情	√	×	√	√
表	DROP_TABLE	删除表	√	×	√	√
	SELECT	查询表	√	×	√	√
	INSERT_INTO_TABLE	插入	√	×	√	√
	ALTER_TABLE_ADD_COLUMNS	添加列	√	×	√	√
	INSERT_OVERWRITE_TABLE	重写	√	×	√	√

资源	操作	说明	DLI FullAccess	DLI ReadOnlyAccess	Tenant Administrator	DLI Service Administrator
	ALTER_TABLE_RENAME	重命名表	√	×	√	√
	ALTER_TABLE_ADD_PARTITION	在分区表中添加分区	√	×	√	√
	ALTER_TABLE_RENAME_PARTITION	重命名表分区	√	×	√	√
	ALTER_TABLE_DROP_PARTITION	删除分区表的分区	√	×	√	√
	SHOW_PARTITIONS	显示所有分区	√	×	√	√
	ALTER_TABLE_RECOVER_PARTITION	恢复表分区	√	×	√	√
	ALTER_TABLE_SET_LOCATION	设置分区路径	√	×	√	√
	GRANT_PRIVILEGE	表的赋权	√	×	√	√
	REVOKE_PRIVILEGE	表权限的回收	√	×	√	√
	SHOW_PRIVILEGES	查看其他用户具备的表权限	√	×	√	√
	DISPLAY_TABLE	显示表	√	√	√	√
	DESCRIBE_TABLE	显示表信息	√	×	√	√
弹性资源池	DROP	删除弹性资源池	√	×	√	√
	RESOURCE_MANAGEMENT	弹性资源池资源管理	√	×	√	√
	SCALE	扩缩容弹性资源池	√	×	√	√

资源	操作	说明	DLI FullAccess	DLI ReadOnlyAccess	Tenant Administrator	DLI Service Administrator
	UPDATE	更新弹性资源池	√	×	√	√
	CREATE	创建弹性资源池	√	×	√	√
	SHOW_PRIVILEGES	查看其他用户具备的弹性资源池权限	√	×	√	√
	GRANT_PRIVILEGE	赋予指定用户弹性资源池权限	√	×	√	√
	REVOKE_PRIVILEGE	移除指定用户弹性资源池权限	√	×	√	√
增强型跨源连接	BIND_QUEUE	增强型跨源连接绑定队列 仅用于跨项目授权。	×	×	×	×

如果系统策略不满足授权要求，管理员可以创建自定义策略，并通过给用户组授予自定义策略来进行精细的访问控制，自定义策略是对系统策略的扩展和补充。详细操作请参考**创建自定义策略**。

相关链接

- **《IAM产品介绍》**
- **《创建用户组、用户并授予DLI权限》**
- **《策略语法》**
- **《如何修改用户策略》**
- **《队列赋权》**（API赋权）
- **《数据赋权》**（API赋权）
- **《设置队列权限》**（Console赋权）
- **《数据库权限管理》**（Console赋权）
- **《表权限管理》**（Console赋权）

9 配额管理

什么是配额?

为防止资源滥用，平台限定了各服务资源的配额，对用户的资源数量和容量做了限制。

如果当前资源配额限制无法满足使用需要，您可以申请扩大配额。

怎样查看我的配额

1. 登录管理控制台。
2. 单击管理控制台左上角的，选择区域和项目。
3. 在页面右上角，选择“资源 > 我的配额”。
   系统进入“服务配额”页面。

   图 9-1 我的配额

4. 您可以在“服务配额”页面，查看各项资源的总配额及使用情况。
   如果当前配额不能满足业务要求，请参考后续操作，申请扩大配额。

如何申请扩大配额?

1. 登录管理控制台。
2. 在页面右上角，选择“资源 > 我的配额”。
   系统进入“服务配额”页面。

图 9-2 我的配额

3. 单击“申请扩大配额”。
4. 在“新建工单”页面，根据您的需求，填写相关参数。
   其中，“问题描述”项请填写需要调整的内容和申请原因。
5. 填写完毕后，勾选协议并单击“提交”。

10 与其他云服务的关系

与对象存储服务（OBS）的关系

对象存储服务（Object Storage Service）作为DLI的数据来源及数据存储，与DLI配合一起使用，关系有如下四种。

- 数据来源：使用DLI服务提供API，将OBS对应路径的数据导入到DLI。
  具体API请参考《**导入数据**》。
- 存储数据：DLI中支持创建OBS表，该类型表在DLI服务中只有元数据，实际数据在该表对应的OBS路径中。
  创建OBS表的SQL语法请参考《**使用DataSource语法创建OBS表**》和《**使用Hive语法创建OBS表**》。
- 备份数据：使用DLI提供导出API，将DLI的数据导出到OBS中备份。
  具体API请参考《**导出数据**》。
- 存储查询结果：DLI提供API供用户将日常作业的查询结果数据保存到OBS。
  具体API请参考《**导出查询结果**》。

与统一身份认证服务（IAM）的关系

统一身份认证服务（Identity and Access Management）为DLI提供了华为云统一入口鉴权功能。

具体操作请参考《**创建用户并授权使用DLI**》和《**DLI自定义策略**》。

与云审计服务（CTS）的关系

云审计服务（Cloud Trace Service）为DLI提供对应用户的操作审计。

云审计服务支持的DLI操作请参考《**云审计服务支持的DLI操作列表说明**》。

与云监控服务（CES）的关系

云监控（Cloud Eye）为DLI提供监控数据，监控作业中的多项指标，从而集中高效地呈现状态信息。

具体指标请参考《**数据湖探索监控指标说明**》。

与消息通知服务（SMN）的关系

消息通知服务（Simple Message Notification）可以在DLI发生作业运行异常时给用户发送通知。

具体操作请参考《**创建消息通知主题**》。

与表格存储服务（CloudTable）的关系

表格存储服务（CloudTable Service）作为DLI的数据来源及数据存储，与DLI配合一起使用，关系有如下两种。

- 数据来源：DLI服务提供DataFrame和SQL方式从CloudTable中导入数据到DLI。
- 存储查询结果：DLI使用标准SQL的Insert语法将日常作业的查询结果数据存放到CloudTable表中。

通过DLI跨源连接访问CloudTable数据请参考《**跨源分析开发方式参考**》。

与关系型数据库服务（RDS）的关系

关系型数据库（Relational Database Service）作为DLI的数据来源及数据存储，与DLI配合一起使用，关系有如下两种。

- 数据来源：DLI服务提供DataFrame和SQL方式从RDS中导入数据到DLI。
- 存储查询结果：DLI使用标准SQL的Insert语法将日常作业的查询结果数据存放到RDS表中。

通过DLI跨源连接访问RDS数据请参考《**跨源分析开发方式参考**》。

与数据仓库服务（DWS）的关系

数据仓库服务（Data Warehouse Service）作为DLI的数据来源及数据存储，与DLI配合一起使用，关系有如下两种。

- 数据来源：DLI服务提供DataFrame和SQL方式从DWS中导入数据到DLI。
- 存储查询结果：DLI使用标准SQL的Insert语法将日常作业的查询结果数据存放到DWS表中。

通过DLI跨源连接访问DWS数据请参考《**跨源分析开发方式参考**》。

与云搜索服务（CSS）的关系

云搜索服务（Cloud Search Service）作为DLI的数据来源及数据存储，与DLI配合一起使用，关系有如下两种。

- 数据来源：DLI服务提供DataFrame和SQL方式从CSS中导入数据到DLI。
- 存储查询结果：DLI使用标准SQL的Insert语法将日常作业的查询结果数据存放到CSS表中。

通过DLI跨源连接访问DWS数据请参考《**跨源分析开发方式参考**》。

与分布式缓存服务（DCS）的关系

分布式缓存服务（Distributed Cache Service）作为DLI的数据来源及数据存储，与DLI配合一起使用，关系有如下两种。

- 数据来源：DLI服务提供DataFrame和SQL方式从DCS中导入数据到DLI。
- 存储查询结果：DLI使用标准SQL的Insert语法将日常作业的查询结果数据存放到DCS表中。

通过DLI跨源连接访问DWS数据请参考《**跨源分析开发方式参考**》。

与文档数据库服务（DDS）的关系

文档数据库服务（Document Database Service）作为DLI的数据来源及数据存储，与DLI配合一起使用，关系有如下两种。

- 数据来源：DLI服务提供DataFrame和SQL方式从DDS中导入数据到DLI。
- 存储查询结果：DLI使用标准SQL的Insert语法将日常作业的查询结果数据存放到DDS表中。

通过DLI跨源连接访问DWS数据请参考《**跨源分析开发方式参考**》。

与 MapReduce 服务（MRS）的关系

MapReduce服务（MapReduce Service）作为DLI的数据来源及数据存储，与DLI配合一起使用，关系有如下两种。

- 数据来源：DLI服务提供DataFrame和SQL方式从MRS中导入数据到DLI。
- 存储查询结果：DLI使用标准SQL的Insert语法将日常作业的查询结果数据存放到MRS表中。

通过DLI跨源连接访问DWS数据请参考《**跨源分析开发方式参考**》。

与数据治理中心（DataArts Studio）的关系

在数据治理中心DataArts Studio中，数据开发是一个一站式的大数据协同开发平台，提供全托管的大数据调度能力。它可管理多种大数据服务，极大降低用户使用大数据的门槛，帮助用户快速构建大数据处理中心。

- 通过数据治理中心的DLI SQL节点传递SQL语句到DLI中执行，请参考《**DLI SQL**》。
- 通过数据治理中心的DLI Flink Job节点执行一个预先定义的DLI作业，请参考《**DLI Flink Job**》。
- 通过数据治理中心的DLI Spark节点执行一个预先定义的Spark作业，请参考《**DLI Spark**》。

11 基本概念

弹性资源池

专属的计算资源，不同弹性资源上的计算资源完全隔离，弹性资源池内的不同队列资源可以共享，并可以根据队列资源负载配置策略进行分时弹性扩缩容，满足不同的业务需求。

DLI 存储资源

DLI存储资源是DLI服务内部存储的资源，用于存储数据库和DLI表，是向DLI导入数据的必备条件，体现用户数据存储在DLI中的数据量。

弹性资源池的实际 CUs、已使用 CUs、CU 范围、规格（包周期 CU）

- **实际CUs**：弹性资源池当前分配的可用CUs。
- **已使用CUs：**已经被作业或任务占用的CU资源。这些资源可能正在执行计算任务，暂时不可用。

  说明

  HetuEngine已使用CUs和实际CU一致。
- **CU范围**：CU设置主要是为了控制弹性资源池扩缩容的最大最小CU范围，避免无限制的资源扩容风险。
  - 弹性资源池中所有队列的最小CU数之和需要小于等于弹性资源池的最小CU数。
  - 弹性资源池中任意一个队列的最大CU必须小于等于弹性资源池的最大CU。
  - 弹性资源池至少可以满足弹性资源池中所有队列按最小CU运行，尽量满足队列按最大CU运行。
- **规格（包周期CU）**：购买弹性资源池时选择的CU范围的最小值即弹性资源池规格。规格是包周期弹性资源池特有的。规格部分以包周期的计费，规格之外的部分按需计费。

数据库

数据库即按照数据结构来组织、存储和管理数据的仓库。DLI服务管理权限的基础单元是数据库，赋权以数据库为单位。

在DLI中，表和数据库是定义底层数据的元数据容器。表中的元数据让DLI知道数据所在的位置，并指定了数据的结构，例如列名称、数据类型和表名称。数据库是表的逻辑分组。

OBS 表、DLI 表、CloudTable 表

不同表类型表示不同的数据存储位置。

- OBS表表示数据存储在OBS服务的桶中。
- DLI表表示数据存储在本服务内部的表中。
- CloudTable表表示数据储存在CloudTable服务的表中。

可通过DLI创建表，与其他服务的数据进行关联，以此来实现不同数据源的联合查询分析。

元数据

元数据（Metadata）是用来定义数据类型的数据。主要是描述数据自身信息，包含源、大小、格式或其它数据特征。数据库字段中，元数据用于诠释数据仓库的内容。

SQL 作业

在SQL作业编辑器执行的SQL语句、导入数据和导出数据等操作，在系统中对应的执行实体，称之为SQL作业。

SQL作业适用于使用标准SQL语句进行查询的场景。通常用于结构化数据的查询和分析。

Flink 作业

Flink作业专为实时数据流处理设计，适用于低时延、需要快速响应的场景。适用于实时监控、在线分析等场景。

- Flink OpenSource作业：提交作业时可以使用DLI提供的标准的连接器（connectors）和丰富的API，快速与其他数据系统的集成。
- Flink Jar作业：允许用户提交编译为Jar包的Flink作业，提供了更大的灵活性和自定义能力。适合需要自定义函数、UDF（用户定义函数）或特定库集成的复杂数据处理场景。可以利用Flink的生态系统，实现高级流处理逻辑和状态管理。

Spark 作业

Spark作业是指用户通过可视化界面和RESTful API提交的作业，支持提交Spark Core/DataSet/MLlib/GraphX等Spark全栈作业。

CU

CU是DLI计算资源的单位。 1CU= 1Core 4GMem。不同规格的计算资源对应的计算能力不一样，规格越高计算能力越好。

常量与变量

环境变量中，常量与变量的区别如下：

- 常量在程序运行过程中，所表示的值是无法被改变的。
- 变量是“可读、可写”，而常量是“只读”的。变量是在程序运行过程中，内部存储的值，随时可以被改变的一段内存地址。比如：int a = 123，这里的a就是一个整型变量。

表生命周期

DLI表数据的生命周期管理功能（dli.lifecycle.days），指表（分区）数据从最后一次更新的时间算起，在经过指定的时间后没有变动，则此表（分区）DLI自动回收。这个指定的时间就是生命周期。生命周期管理功能方便您释放存储空间，简化回收数据的流程。同时提供数据备份与恢复功能，避免因误操作丢失数据。

