

目录

目录	1
产品概述	2
名词解释	2
产品优势	2
统一管理	2
安全管控	2
高效开发	2
规模扩展	2
兼容开放	2
生态完善	2
产品功能	2
多用户协同开发	2
多源异构数据集成	2
全链路的数据管理	3
海量数据批量处理	3
流式数据实时处理	3
多数据源统一查询	3
多维数据分析与可视化	3
大规模作业智能调度	3

产品概述

大数据云平台（DataCloud）是金山云在智能数字化转型时代推出的新一代大数据智能产品。在主流Hadoop、Spark框架之上构建端到端的大数据智能开发管理闭环，以高度统一的标准提供数据汇聚、数据开发、智能调度、数据管理、数据分析与服务等一站式大数据Web化处理环境，帮助您轻松挖掘数据价值，助力企业数字化运营。

名词解释

- **CU** 大数据平台计算单元。CU描述了作业的最小运行能力，作业可跑在多个CU上运行。
- **DCU** 容器计算单元。DCU描述了作业的最小运行能力，作业可跑在多个DCU上运行。
- **资源组** 购买的CU/DCU资源可按组进行管理，划分多个资源组，每个资源组按业务量任意分配不同的CU/DCU数。如购买了30个CU，可划分2个资源组，分别分配10CU和20CU。
- **项目** 项目是大数据云平台最基本的业务组织单元，主账号可以根据不同的项目需求创建多个项目，每个项目包含不同的大数据云组件，如项目1包含数据采集和流计算组件，项目2包含数据集成和批计算组件。项目管理员可以将不同的人加入到不同的项目中，实现多人协作开发，项目制管理。
- **任务** 数据处理逻辑的封装，如数据同步任务完成数据从源库到目的库的数据迁移，批计算任务完成一批数据的ETL处理。用户可新建多个不同的任务，完成相应的处理逻辑开发。任务以工作流（DAG图）的形式进行组织，一个工作流可包含一个节点或多个节点，每个节点可以完成一项数据处理操作。
- **实例** 任务运行后会产生一个实例，如调度系统配置了某个任务每小时运行一次，那每天会产生24个该任务的实例。

产品优势

统一管理

统一接入和管理多存储引擎、计算引擎、分析引擎等，并提供统一的访问入口和编程接口。

安全管控

提供数据加密、脱敏等多种安全措施以及完善的多租户权限管理，满足合规性要求。

高效开发

大数据加工全流程可视化界面操作，拖拽式工作流编排，智能的监报告警体系。

规模扩展

具备高度的可扩展性，支持高达万台规模，破除Hadoop、MPP等关键系统的规模瓶颈。

兼容开放

兼容开源Hadoop标准，支持SQL2003标准的运算引擎SparkSQL。

生态完善

为金融、政务、工业等多行业提供专业的一体化的大数据解决方案，赋能企业对外输出行业云能力。

产品功能

多用户协同开发

多租户机制，项目制管理，支持不同角色用户协同开发，支持严格的权限管控和资源隔离，提升大数据项目整体效率和安全性。

多源异构数据集成

- 支持多种异构数据源的批量和实时数据接入
- 支持复杂网络环境（有公网/无公网）下的数据接入
- ETL过程支持脱敏、加密、标准化等企业特性

全链路的数据管理

- 支持查看企业所有数据资产分布情况
- 支持元数据统一管理，库表统一管理
- 支持数据血缘、数据质量、数据生命周期管理
- 支持企业内部数据权限申请与审批

海量数据批量处理

- 支持拖拽式构建任务工作流
- 支持SQL、MR、shell、数据同步等多种处理算子
- 支持异构集群任务调度，支持容器化运行环境
- 支持隔离的测试和生产环境

流式数据实时处理

- 支持拖拽式工作流编辑，支持任务上下游依赖关系编排
- 支持多种流处理算子
- 支持在线编写业务逻辑SQL
- 支持流数据运维，实时展现数据指标

多数据源统一查询

- 统一SQL查询Hadoop、GreenPlum、ES、关系库等异构数据系统；
- 提供标准的分析接口；
- 适用于企业存在多种数据存储系统需要统一提供标准化分析接口的场景。

多维数据分析与可视化

- 支持海量数据毫秒级查询
- 支持丰富的可视化展现形式
- 支持灵活构建分析主题仪表盘
- 支持大屏设计与实时动态刷新

大规模作业智能调度

- 支持每天百万级任务调度；
- 单个工作流支持千个以上任务
- 分钟级作业流定时调度，秒级节点间调度延迟
- 支持异构集群作业调度，支持非Yarn类工作流容器化调度