云数据仓库套件 Sparkling

产品简介

产品文档

腾讯云

版权所有：腾讯云计算（北京）有限责任公司

【版权声明】

©2013-2019 腾讯云版权所有

本文档著作权归腾讯云单独所有，未经腾讯云事先书面许可，任何主体不得以任何形式复制、修改、抄袭、传播全部或部分本文档内容。

【商标声明】

腾讯云

及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。

【服务声明】

本文档意在向客户介绍腾讯云全部或部分产品、服务的当时的整体概况，部分产品、服务的内容可能有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或模式的承诺或保证。

文档目录

产品简介
- 产品概述
- 产品优势
- 应用场景
- 版本历史
- 多租户策略

产品简介
产品概述

最近更新时间：2020-03-16 18:22:59

简介

云数据仓库套件 Sparkling（Tencent Sparkling Data Warehouse Suite）基于业界领先的 Apache Spark 框架，为用户提供一套全托管、简单易用、高性能的 PB 级云端数据仓库解决方案，实现一站式大数据开发与数据科学计算。企业可以借助 Sparkling 跨数据源联合分析特性，轻松分析位于云上的数据，从而帮助企业专注于数据价值的挖掘和探索。

云数据仓库套件 Sparkling 支持弹性扩缩容，用户可以根据业务发展需要调整集群资源大小；支持多种异构数据源的导入及作业定时调度，用户可以实现对多路数据源的汇聚分析；提供了统一的、图形化的交互方式，帮助用户更加便捷、高效地进行数据开发与数据分析操作。

产品功能

集群管控

Sparkling 集群是云数据仓库套件 Sparkling 为用户提供服务的载体。一个 Sparkling 集群由 Master 节点和工作节点组成。工作节点包括核心节点和弹性计算节点，其中核心节点提供数据存储能力和计算能力，弹性计算节点提供计算能力。Sparkling 集群的大小，决定了云数据仓库套件 Sparkling 所能提供的存储能力和计算能力的上限。

Sparkling 支持创建高可用的集群，来保证服务的高可用性。在高可用模式下，Master 节点服务在机架级硬件故障情况下仍然可用。

- 云数据仓库套件 Sparkling 提供了非常简洁易用的 集群管理 功能，方便用户指定符合业务需求的 Sparkling 集群规模。同时，随着用户业务的发展，存储和计算需求的增加，用户也可以很方便的 扩容和缩容 集群规模。随着节点规模的增加，整个集群的存储容量和计算性能也得到线性的提升。
- 当用户不再需要使用某个集群时，可以选择 销毁集群。被销毁的集群无法恢复，同时集群中存储的数据也会在一段时间之后无法再访问。
- Sparkling 的集群管理除了支持集群的创建、销毁、扩缩容，还可以查看集群当前运行状态、资源信息、所在地域、节点型号等信息。

数据集成

提供各类异构数据源的接入集成。用户可以通过 Data Studio 控制台将传统关系型数据库、对象存储 COS、Kafka 流式数据经过抽取、转换和装载，接入到云数据仓库套件 Sparkling 的存储中。

版权所有：腾讯云计算（北京）有限责任公司

Sparkling 提供丰富的功能供用户定制数据接入过程。用户可以：

- **行裁剪**：设置过滤条件，对导入数据进行行裁剪。
- **导入部分列**：删除导入数据的某些列，只将部分列导入数据仓库。
- **列顺序调整**：对导入列的顺序进行调整。
- **设置分区**：在某些列上设置分区，提高后续数据查询的效率。
- **多种格式存储**：支持多种数据文件存储格式。
- **管理数据源**：支持保存和管理数据源，方便之后设置新的数据导入任务以及数据溯源。
- **导入已有数据表**：支持将数据导入到数仓上一个已经存在的数据表中。
- **设置映射关系**：用户可以设置源数据表与目标数据表之间的映射关系，使数据导入更加灵活。
- **预览数据**：数据导入过程中用户可以预览导入数据。
- **定时导入**：用户可以进行单次导入，也可以设置定时任务进行周期性导入。
- **全量/增量导入**：支持数据的全量导入和增量导入。
- **增量导入自定义语法**：支持增量导入条件中提供一些自定义语法使增量导入条件更加灵活。

数据开发

Sparkling 为用户提供基于 Notebook 的线上交互式环境。用户可以通过在 Notebook 中执行代码，对数据仓库中的数据进行分析和加工。

用户可以在 Sparkling Notebook 中执行 SQL、Python 和 Spark 代码。Sparkling SQL 是一种结构化的查询语言，语法与 MySQL/Oracle/Hive SQL 类似，兼容业界 SQL 标准 ANSI SQL 2003。熟悉传统数据库或 Hive 的用户可以很容易上手。除了标准的 SQL 操作，Sparkling SQL 还嵌入了多种高级函数，这些函数包含了常见的数学运算、统计分析、时间日期等方面的操作。

用户在 Sparkling Notebook 里还可以运行 spark 和 pyspark 程序，方便用户开发更灵活的数据分析程序。

Sparkling Notebook 提供了数据可视化工具。通过拖拽组件的方式，用户可以在 Notebook 中通过多种方式（如饼状图、散点图等）对数据进行可视化。通过结合交互式编程和数据可视化，用户可以方便的分析和调试数据。用户还可以将数据分析结果进行报表展示，以及将分析结果导出下载到本地。

Sparkling 还提供一些辅助功能提高用户效率，例如用户可以按照项目组织自己 Notebook，可以在一个 SQL IDE 中查看数据表。

任务管理

对于持续更新的数据，Sparkling 支持用户将数据导入和 Notebook 设置为按一定周期定时执行。周期设置的范围从小时到月。Sparkling 提供可靠的周期调度，支持任务的回填调度。

除了基本的数据导入和 Notebook 定时任务，用户可以将数据导入和 Notebook 进行组合，组成一个 DAG 工作流任务，Sparkling 将对整体 DAG 工作流任务进行按依赖关系进行调度，这在复杂的数据分析流水线和数据科学等场景下尤为有用。

版权所有：腾讯云计算（北京）有限责任公司

通过 Sparkling 统一的任务管理界面可以查看和管理数据导入和 Notebook 定时运行任务。用户可以进行如查看任务状态、历史信息、临时触发任务或终止任务等操作。

弹性伸缩

云数据仓库套件 Sparkling 提供强大的弹性扩缩容能力。计算存储分离，集群工作节点包括核心节点和弹性计算节点。用户通过 Data Studio 控制台或云 API，实现手动和自动对大规模节点的快速线性横向的扩容，以及纵向计算和存储能力的变配。弹性计算节点同时支持自动化弹性缩容，以适配业务的发展。

数据管理

提供元数据管理模块，支持技术元数据、管理元数据和业务元数据的注册、导入、存储、检索、导出、发布等管理功能，同时为用户提供数据地图、数据字典、数据血缘追踪和影响分析、元数据版本管理、元数据统计分析、数据质量报表等多种数据管理能力。

项目管理

提供项目管理模块，支持客户按照企业内部产品线、团队和项目方式来创建项目空间并进行项目人员和笔记簿管理。

版权所有：腾讯云计算（北京）有限责任公司

产品优势

最近更新时间：2020-03-16 18:23:20

云数据仓库套件 Sparkling 主要有以下特点与显著优势：

一站式创建

用户只需要在腾讯云终端界面选择产品的参数指标即可完成对云数据仓库套件 Sparkling 服务的创建。具体创建流程对用户完全屏蔽，由后台完成对 CVM、TencentDB、CLB 等资源的申请，及对云数据仓库套件 Sparkling 集群的搭建、服务的拉起、监控等创建工作。

统一的交互方式

云数据仓库套件 Sparkling 提供统一的交互方式，用户可以使用数据开发页面进行交互式的数据处理，同时云数据仓库套件 Sparkling 也为用户提供了 JDBC/ODBC 接口，用户可以程序化的方式与数仓进行交互。

专业的集群管控

独享模式为用户提供集群管理和监控模块，支持集群创建、自动扩缩容、集群配置、启停、资源智能监控报警等功能。Sparkling 集群是全托管集群，用户无需过多关注集群底层架构，减少运维压力。

丰富的异构数据集成

提供各类异构数据源的接入集成。用户可以通过 Data Studio 控制台将传统关系型数据库 RDBMS、对象存储 COS、Kafka 消息队列中的数据经过抽取、转换和装载，接入到云数据仓库套件 Sparkling 的存储中，同时也提供了丰富的抽取条件和抽取任务调度，以满足用户不同的数据导入需求。

齐全的数据管理

提供元数据管理模块，支持技术元数据、管理元数据和业务元数据的注册、导入、存储、检索、导出、发布等管理功能，同时为用户提供数据地图、数据字典、数据血缘追踪和影响分析、元数据版本管理、元数据统计分析、数据质量报表等多种数据管理能力。

企业级作业调度

云数据仓库套件 Sparkling 提供了企业级的作业调度引擎，可以帮助用户以复杂的依赖关系构建完整的作业流程。同时提供任务编排调度管理模块，支持时间驱动与事件驱动的 DAG 任务编排和调度。同时提供完备的任务监控，方便用户运维数据 ETL 和数据加工分析作业。

完备的项目管理与账号服务

提供项目管理模块，支持客户按照企业内部产品线、团队和项目方式来创建项目空间并进行项目人员和笔记簿管理。有完备的账号服务相关方案，确保主账号和子账号之间的资源共享和权限管理。

高性能、高可用及高可扩展性

版权所有：腾讯云计算（北京）有限责任公司

云数据仓库套件 Sparkling 依托腾讯云提供的 IaaS 服务以及自身组件的能力，提供了高性能、高可用性以及高可扩展性的数仓产品。

数据安全保障

通过用户专用网络，权限管理、多租户隔离等方式，保证用户数据的安全。

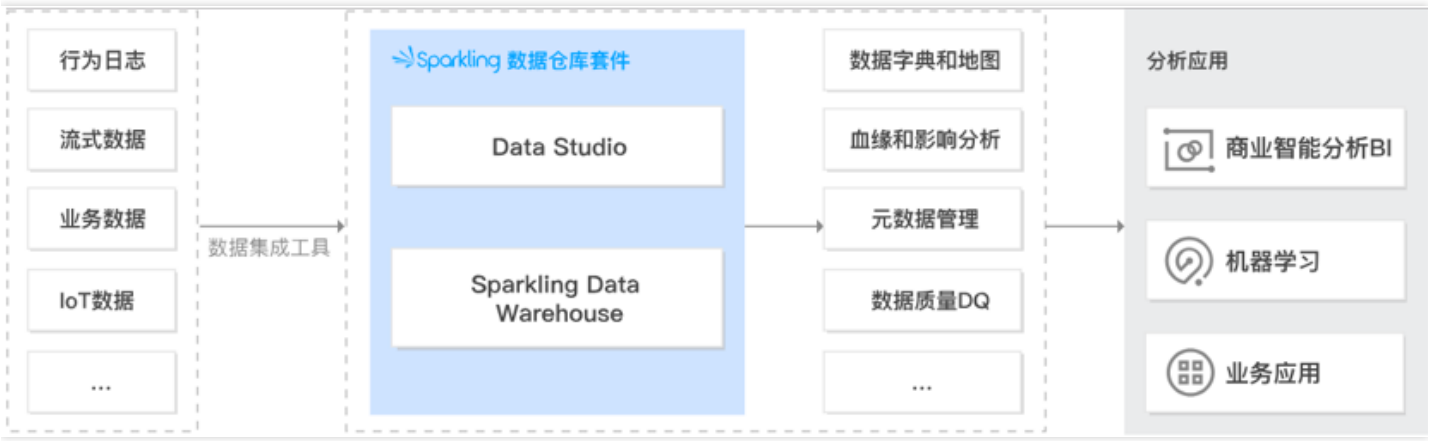
应用场景

最近更新时间：2020-03-16 18:24:05

企业全域数据资产管理

在游戏、金融、零售、工业等多个行业，迫切需要对用户行为、人员、采购、销售、资产、供应链等管理和业务数据进行汇总管理和分析，构建完备的企业全域数据视图，以便掌握公司总体经营情况，进行快速精准决策。

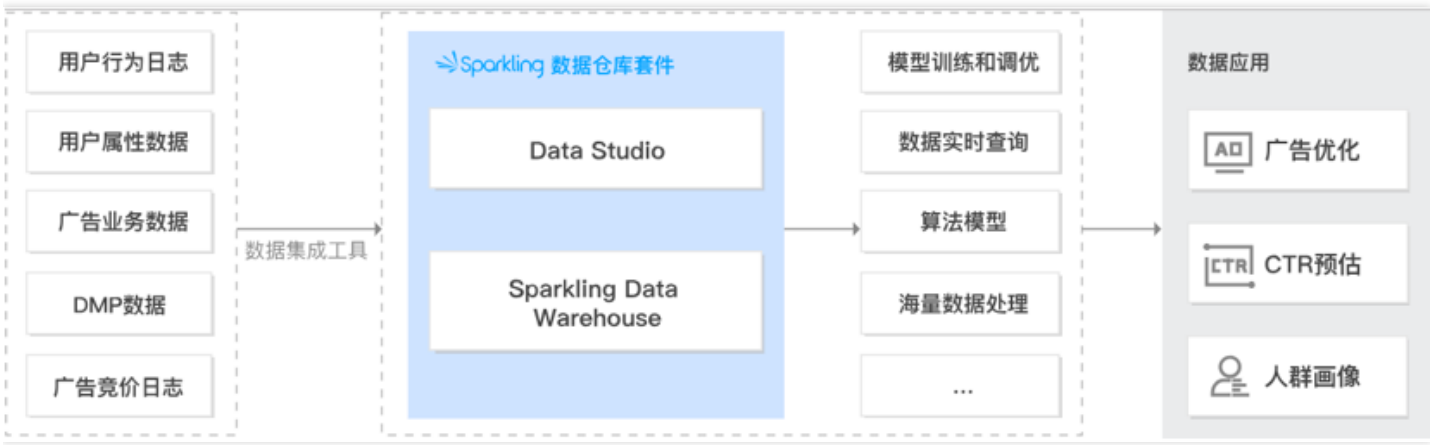
云数据仓库套件 Sparkling 通过数据集成工具将分散在传统数据库、业务系统、服务器上的各类数据源汇总，构建公司级别数据地图门户，利用高性能的异构数据源聚合分析能力，进行数据管理和业务价值提炼。



海量日志分析和精准营销

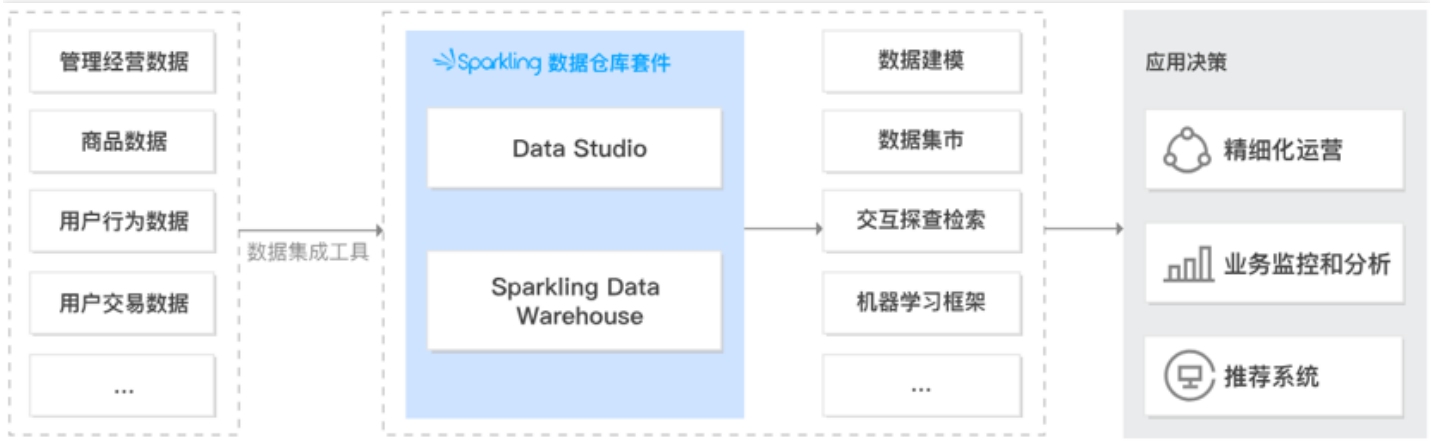
通过日志标准化和规范化机制，用户可以便捷的分析 PB 级用户行为和系统日志等结构化或半结构化数据，构建基于 cookie 的消费者人群画像，对用户进行个性化推荐，有效提升商品投放精准度。云数据仓库套件 Sparkling 支持实时数据接入、COS 深度集成。

版权所有：腾讯云计算（北京）有限责任公司



数据科学决策

通过提供低门槛的机器学习框架、交互式的协作编程环境和数据实时查询分析能力，为数据科学家提供数据建模的有力工具，为企业管理者提供精细化运营的能力，赋能企业，提升企业业务洞察能力。



版本历史

最近更新时间：2019-09-19 16:32:54

2019.08 正式版 V2.0 发布

新增功能

- 增加项目空间管理功能：
  a. 支持账号管理（包括主账号、子账号、协作者账号）。
  b. 支持账号权限管理（包括所有权限、项目管理员权限、集群管控者权限、数据管控者权限、数据开发师/分析师权限）。

优化

- Kafka 数据接入：
  a. 新增支持接入 JSON、CSV 格式。
  b. 支持数据源高级配置。
  c. 支持对 Schema 标注字段手动排序。
  d. AVRO 数据支持“表格模式”和“Schema 脚本”两种 Schema 标注形式。
  e. CSV 数据 Schema 标注支持“表格模式”，支持用户手动键入字段序列号。
- 新增定时导入 COS 数据功能，包括增量追加和整表全量导入/覆盖。
- 优化集群监控功能：新增集群级别的 CPU、momory、Disk IO & Network IO 的监控指标。

2019.07.02 公测版 V1.5 发布

新增功能

- 新增数据源管理功能。
- 增加 TDSQL 数据导入功能。
- 更新 COS 数据导入功能：支持文件夹及多类型文件导入。

2019.05.28 公测版 V1.4 发布

新增功能

- 新增集群监控功能。
- 新增 Kafka 数据导入功能模块。
- 新增服务授权模块。

优化

- 任务管理页面重构：新增任务搜索、查看任务创建时间、删除任务等模块。
- COS 数据导入：支持多文件格式数据导入，支持自定义分隔符。

版权所有：腾讯云计算（北京）有限责任公司

2019.03.21 公测版 V1.3 发布

优化

- 数据导入页面重构：新增保存数据源、导入数据预览等页面。
- 后台集群创建流程优化。

2019.01.18 内测版 V1.2 发布

新增功能

- 新增集群高可用功能。
- 新增时间戳增量导入、设置分区、任务定时调度功能。
- 新增任务管理功能。

2018.11.23 内测版上线

核心功能

- 集群管理：集群创建、扩缩容、销毁、信息查询等。
- 数据开发：工作区 Notebook 管理及 SQL IDE 等。
- 数据集成：RDBMS 及 COS 数据导入。

版权所有：腾讯云计算（北京）有限责任公司

多租户策略

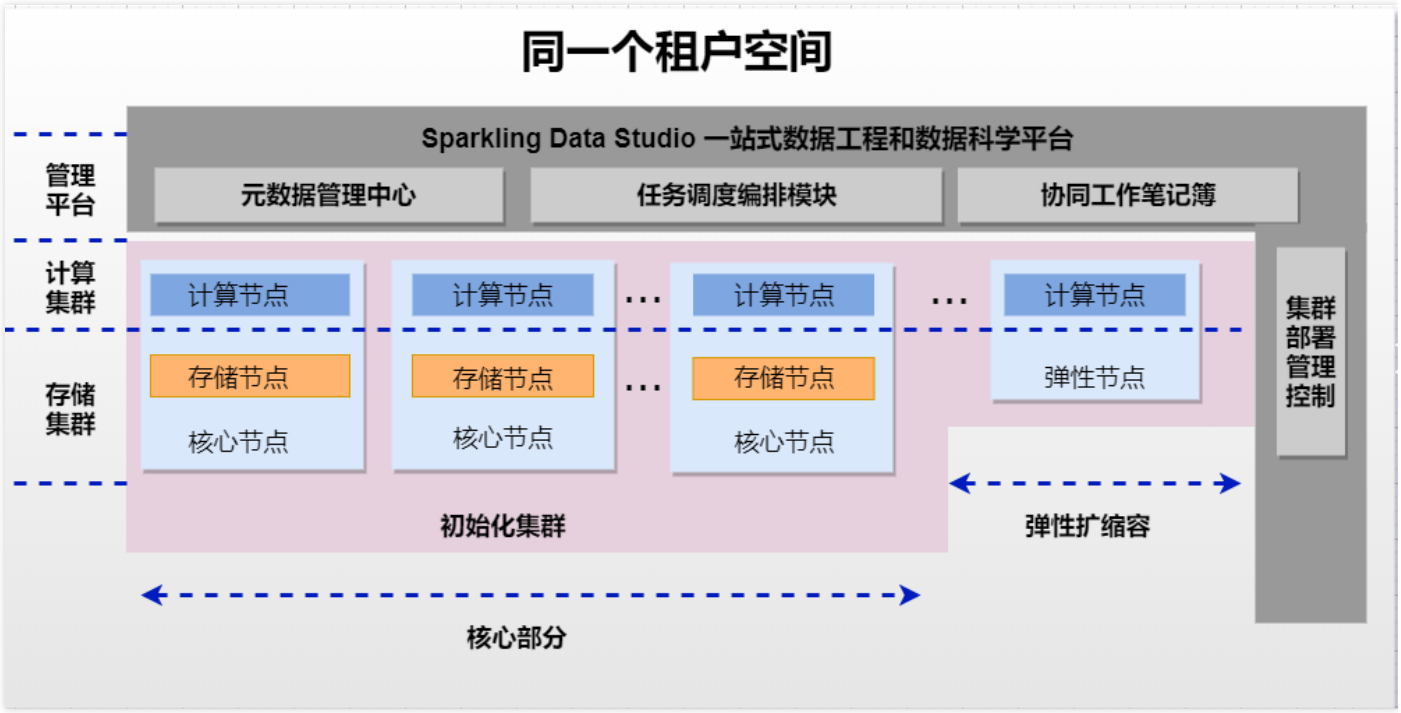
最近更新时间：2020-03-16 18:24:20

Sparkling 采用多租户策略进行权限及角色管理，进而帮助用户管理项目、计算、存储及数据等资源，确保用户之间的资源、故障、安全的隔离性。本节将为您介绍 Sparkling 多租户策略，及其中涉及的账号登录体系、认证机制和用户角色。

特点

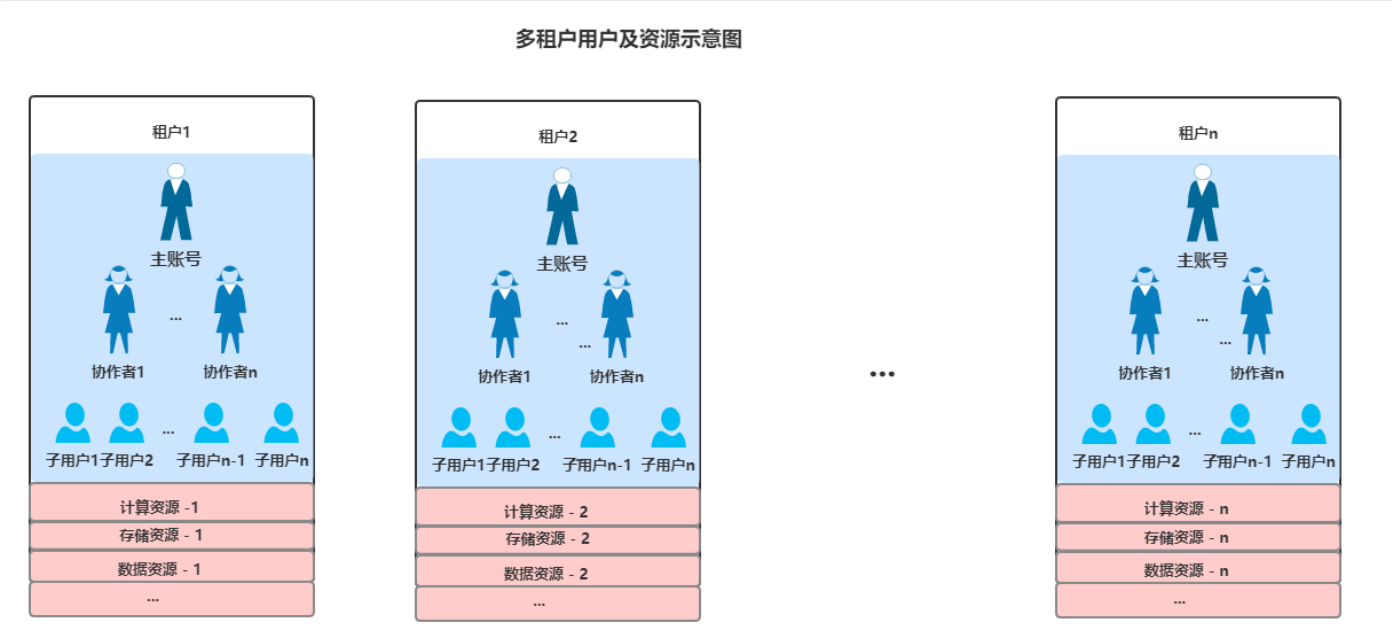
租户是使用 Sparkling 的一个用户或一组用户（包括主账号用户、子用户和协作者用户）。租户申请到的资源只能被租户内的用户使用，与其他租户之间隔离。

- **同租户共享：**
  同一租户空间下的用户共享计算/存储资源以及统一的 Sparkling 管理平台。同一租户空间架构图如下：



- **多租户隔离：**
  使用 Sparkling 服务的多租户之间资源隔离且计费独立。租户是 Sparkling 资源申请及计费的基本单位，租户独享数据源、数据、Notebook 笔记簿、任务等对象实例，独立管理所有的数据、权限、用户及角色。Sparkling

集群及资源只允许租户内的用户使用，各租户之间互相隔离。



策略

Sparkling 多租户策略贯穿整个用户使用流程，包括账号登录、身份认证、角色与权限管控。

账号登录

Sparkling 账号登录体系与腾讯云保持一致。您注册的腾讯云账号身份为主账号，并可以通过 用户管理 为主账号创建子账号进行协作。Sparkling 支持主账号和子账号（包括 子用户 和 协作者）账号登录使用 Sparkling。

- **主账号**
  - 以主账号身份进行集群开通的创建者，自动拥有 Sparkling 内部的所有权限，且不能被其他的项目管理者进行撤权和删除操作。
  - 主账号身份登录，可以拉取子用户和协作者账号，向当前项目空间进行用户添加和授权操作。
- **子用户**
  - 由主账号创建，完全归属于创建该子用户的主账号。
  - 子用户身份登录，在被主账号授予【项目管理员】SparklingProjectGovernance 权限策略或者 AdministrorAccess 权限策略后，可以拉取其主账号名下的子用户账号信息，进行添加和授权操作。
- **协作者账号**

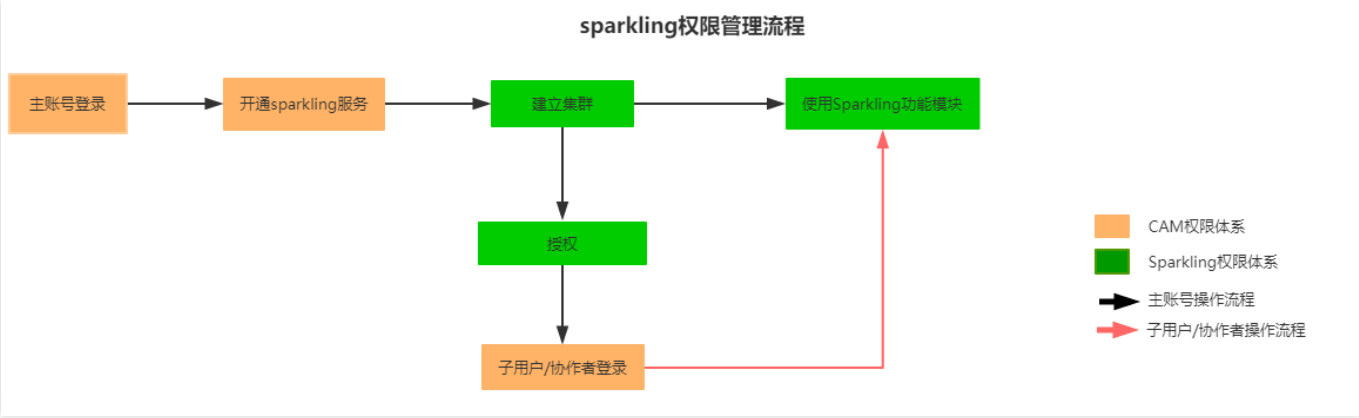
- 协作者账号本身拥有主账号身份，可以被添加作为当前主账号的协作者。被添加的协作者账号为当前主账号的子账号之一，可切换回主账号身份。
- 协作者账号无法将自己的子账号加入到主账号的项目空间中。
- 协作者账号不可作为 Sparkling 项目管理者。

主、子账号管理操作请参见 CAM 用户指南。

身份认证

Sparkling 将腾讯云 CAM 权限管理体系和 Sparkling 内部权限管理这两套机制统一管理，帮助管理员轻松管控 Sparkling 用户权限。

- **CAM 权限管理体系**
  CAM 权限管控体系主要用于：**用户授权使用 Sparkling 服务；获取主账号所有的子账号信息。**
  CAM 权限管控体系包括公有云租户/用户管理集群管控模块、数据管控模块、数据开发模块、任务调度模块、机器学习模块等，涵盖 SparklingFullAccess、SparklingClusterGovernance、SparklingProjectGovernance、SparklingDataGovernance、SparklingDataDevelop、SparklingDataAnalytics 六类权限。使用 Sparkling 服务前，用户必须提前在 CAM 账号体系中创建主账号，详情请参见 CAM 权限定义。
  - **Spakrling 权限管理系统**
    Sparkling 内部账户权限管控系统主要是管理提供集群内部各类对象实例的权限管控服务，同时可与企业 LDAP/AD 系统深度集成，用于用户使用 Sparkling 内部的数据源、数据、Notebook 笔记簿、任务等对象实例。
    Spakrling 权限管理流程如下图所示：



用户角色及权限管控

角色是权限的载体，其拥有对 Sparkling 操作和资源的权限。权限附加到角色而不附加到具体的用户，不同的角色拥有的对 Sparkling 进行操作和访问的权限不同。Sparkling 中支持的用户角色包括项目管理员、集群管控者、数据管控者、数据开发者、数据分析师。

说明：
所有表格中1表示角色拥有该权限点，0表示没有。

- **角色与权限：**

集群管控

权限点	项目管理员	集群管控者	数据管控者	数据开发者	数据分析师
查看集群	1	1	1	1	1
创建集群	1	1	0	0	0
集群监控	1	1	0	0	0
集群续费	1	1	0	0	0
集群扩容	1	1	0	0	0
集群缩容	1	1	0	0	0
集群销毁	1	1	0	0	0
SparkUI/log 权限	1	1	1	1	1

工作区

权限点	项目管理员	集群管控者	数据管控者	数据开发者	数据分析师
新建文件夹	1	0	0	1	1
新建 Notebook	1	0	0	1	1
搜索 Notebook	1	0	0	1	1
搜索文件夹	1	0	0	1	1
删除 Notebook	1	0	0	1	1
重命名 Notebook	1	0	0	1	1
定时调度	1	0	0	1	1
运行所有命令	1	0	0	1	1
运行单行命令	1	0	0	1	1

版权所有：腾讯云计算（北京）有限责任公司

查看命令结果	1	0	0	1	1
可视化查询结果	1	0	0	1	1
下载查询结果	1	0	0	1	1
添加代码段	1	0	0	1	1
删除命令	1	0	0	1	1
SQLIDE 访问	1	0	0	1	1
SQLIDE 查看数据目录	1	0	0	1	1
编辑 SQL	1	0	0	1	1

数据集成

权限点	项目管理员	集群管控者	数据管控者	数据开发者	数据分析师
数据源配置	1	0	1	0	0
数据预览	1	0	1	0	0
目标配置	1	0	1	0	0
抽取任务配置	1	0	1	0	0
数据源预览	1	0	1	0	0

任务调度与监控

权限点	项目管理员	集群管控者	数据管控者	数据开发者	数据分析师
开启/关闭任务	1	0	0	1	1
立即执行/删除任务	1	0	0	1	1
查看任务创建时间	1	0	0	1	1
查看创建人	1	0	0	1	1
查看最近执行时间	1	0	0	1	1

项目空间管理

版权所有：腾讯云计算（北京）有限责任公司

权限点	项目管理员	集群管控者	数据管控者	数据开发者	数据分析师
添加成员	1	0	0	0	0
编辑成员权限	1	0	0	0	0
删除成员	1	0	0	0	0
搜索查看成员	1	0	0	0	0

- **角色与账号**

创建 Sparkling 项目的主账号是整个集群和 Sparkling 资源的拥有者（包括数据源、数据、Notebook、任务等），只有主账号及拥有主账号授权许可的子账号或协作者可以访问 Sparkling 中的资源。Sparkling 中每种角色及可作为其角色载体的账号类型如下：

账号类型	项目管理员	集群管控者	数据管控者	数据开发者	数据分析师
主账号	1	1	1	1	1
协作者账号	0	1	1	1	1
子账号	1	1	1	1	1

版权所有：腾讯云计算（北京）有限责任公司