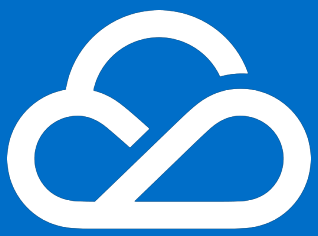
腾讯云大数据处理套件

产品简介

产品文档



腾讯云

【版权声明】

【商标声明】

【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

产品概述

什么是 TBDS

腾讯大数据处理套件 TBDS（Tencent Big Data Suit）是在腾讯多年海量数据处理经验之上，结合开源 Hadoop 生态和自研组件服务，对外提供可靠、安全、易用的大数据处理平台。用户可以按需部署大数据处理服务以实现企业的大数据处理需求，例如：数据提取、处理、分析、报表展示、客户画像、机器学习等大数据应用，以提高企业在大数据背景下的核心竞争力。

我们的理念

1. 屏蔽系统规划、安装及部署细节，降低使用成本
通过控制台规划集群，安装和部署大数据组件；
通过控制台管理系统配置，启停和上下线大数据服务；
尽可能降低用户上机操作的几率；
基于解决方案的一键式部署；

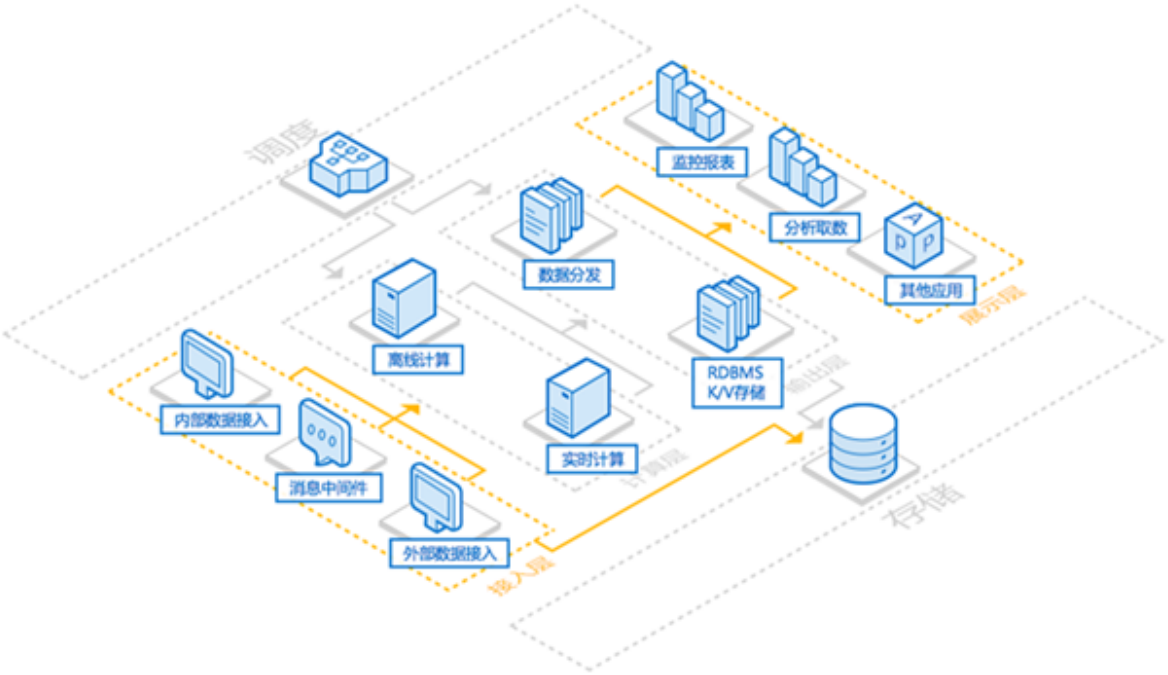
2. 系统可用性
借鉴腾讯相关产品在大数据领域的先进经验，在用户端快速复制腾讯相关产品的高可用大数据系统，做到开箱即用；

3. 系统可扩展
系统提供接口方便后续引入新的大数据服务；

4. 系统可维护性
系统提供丰富的日志帮助用户定位问题；
系统升级不影响现有业务；

我们的架构

一条完整的数据处理流水线通常由“接入-存储-计算-输出-展示”多环节衔接而成。大数据技术经过阶段性地发展，各环节都涌现出一批相互借鉴、相互补充的基础系统。大数据套件将常见的基础系统（包含社区版系统、社区改造版系统以及腾讯自研系统）集成封装，形成统一的大数据平台。数据开发人员可以从大数据平台自由选择不同的基础系统来构建数据流水线，以满足不同场景的数据处理需求。



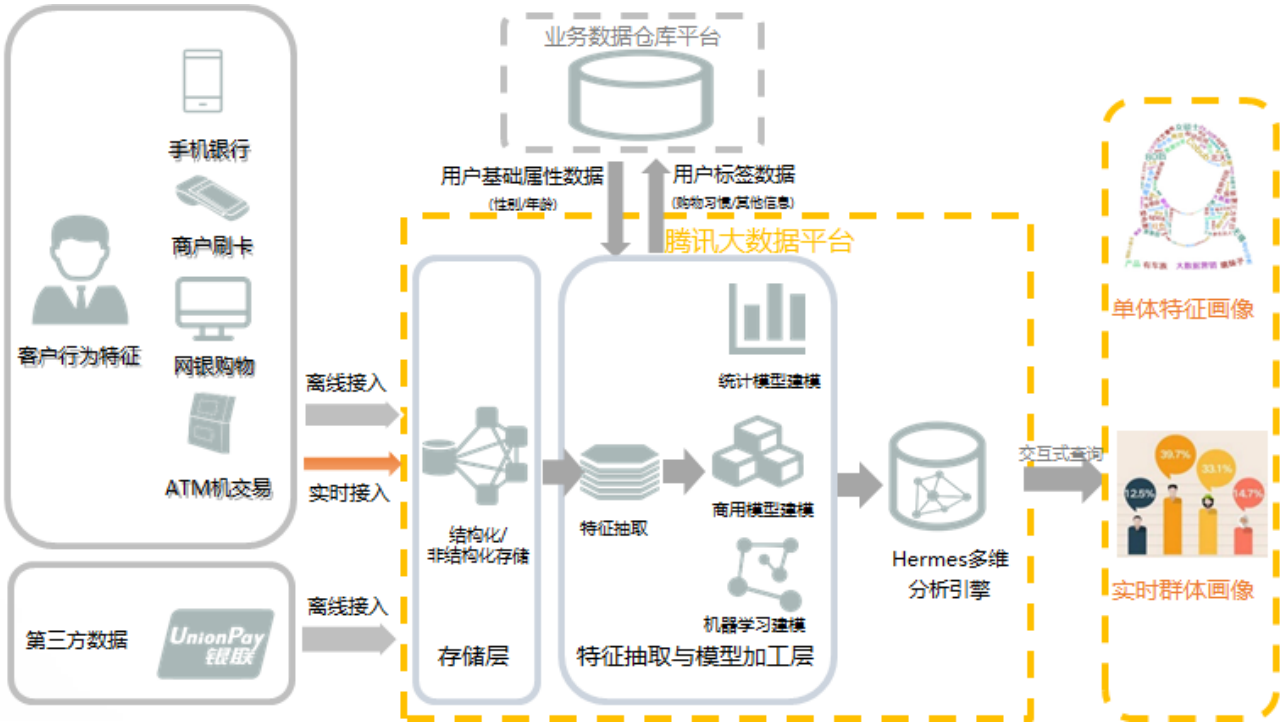
社区方案

社区版系统各自关注于大数据生态某个环节，没有端到端闭环解决方案，同时在使用便利和稳定性上往往无法达到企业级标准。TBDS 经过腾讯内部多业务系统海量数据打磨，通过对社区版加固，自研，提供友好的 IDE 开发环境，从开发，部署，运维全方位满足线上系统高可用要求。

社区方案VS腾讯方案

	社区方案	腾讯方案
方案完整性	无，流程无法打通	完整的闭环解决方案
使用成本	高，需专业人士支持	系统自动化和界面化 降低学习和使用成本
稳定性	一般，没有考虑处理异常场景	线上生产系统验证 可靠性和稳定性有保障
支持力度	较弱，反应慢	专业团队提供支持

实践与案例



功能介绍

基本架构

腾讯大数据处理套件基本构架如下：

功能描述

腾讯大数据处理套件 TBDS（Tencent Big Data Suit）功能如下：

全链路数据开发

易用、安全、稳定、高性能的全链路大数据开发引擎。提供拖拽式的可视化数据开发 IDE，为用户的大数据集成、存储、计算环节提供完整而稳定的企业级解决方案。用户能借助于大数据套件获取到强大的大数据开发能力，聚焦于进行企业的业务创新。包括但不仅限于以下功能：

- 多渠道数据集成

  实时数据接入：支持 Flume、Tube、Kafka

  数据接入，包含结构化、半结构化、非结构化的异构数据毫秒级实时接入。

  离线数据导入：支持 Mysql、Postgre、Oracle

  等主流关系数据库高效导入，支持文本类日志数据离线导入。

  压缩及传输安全：支持高强度数据压缩及加密传输，历经每天 400T、2 万亿条数据接入考验。
- 企业级大数据存储解决方案

  多类型存储支持：支持块存储、分布式文件、对象存储、SQL、NoSQL 从 GB 到 PB

  量级的存储解决方案，满足企业客户复杂存储应用场景。

  可靠：存储系统高可靠容灾设计,可靠性可达

  99.996%，用户按需选择数据热备数量，支持冷热数据分治，支持数据冷备策略自定义。

  可扩展：高可扩展设计，存储系统可动态随企业数据量增加从 G 到 P

  级的动态扩容，支持系统不停机动态扩容。

  集群数据平衡成本最小化。
- 离线批处理计算

  灵活：支持 MapReduce、Hive、Pig

  等批处理计算作业，可支撑企业数仓建设中的数据清洗、转换、汇集、主题提取等数据处理需求；支持 Spark 分布式内存计算框架，在内存中对数据集进行快速的多次迭代，以支持复杂的数据挖掘算法和图计算算法。

  丰富的调度策略：支持丰富的作业调度策略，包含分钟、小时、天、月级周期或非周期的任务执行策略。
- 实时流处理计算

  灵活：支持 TStorm(用 Java 语言重写的 Storm 流处理引擎)、Storm

  流式任务作业引擎，覆盖实时要求极高的流式作业场景；支持基于 Spark 上的 Spark Streaming，满足毫秒级的实时计算场景需求，如实时推荐、用户行为分析等。
- 可视化工作流开发 IDE

  数据开发 IDE：拖拽式的工作流开发 IDE，简单 Web

  式拖拽操作来完成整个大数据工作流的任务开发。

  丰富的任务类型：内置丰富的处理器，囊括离线数据导入导出、在线实时数据接入、MR

  程序、TStorm 程序、Java 程序、Shell 脚本、Email、Hive 脚本、机器学习等多种任务的配置集成。
- 文本检索及检索分析

  文本检索：基于 Lucene 的文本搜索服务器 ElasticSearch 向用户提供友好的 RESTful

  接口的分布式多用户能力的全文搜索引擎，支持 TB 级别的全文检索应用。

  检索分析：万级纬度、千亿数据规模下向用户提供毫秒级高性能检索分析服务，满足用户的检索分析场

景需求。

强大的数据分析与探索挖掘引擎

一站式数据分析、探索、挖掘平台。包含基于纬度建模的多维分析、交互式探索分析、机器学习、深度学习、可视化敏捷报表门户等功能，向用户提供强大的数据分析与数据挖掘能力，助力用户大数据的价值发现。包括但不仅限于以下功能：

- 多维分析引擎
  SQL 兼容：基于 Apache Kylin 开源分布式分析引擎，为用户提供基于 Hbase 存储的数据 Cube 预建模及百亿行规模的 SQL 数据分析能力，满足企业级用户面向部门的数据集市建设需求。
- 交互式数据探索
  任意纬度组合、秒级分析：采用列存储技术、万维标签查询处理技术为用户提供实时的多维交互式 SQL 查询、统计、分析系统，支撑万级维度、千亿级规模下的秒级数据统计分析需求，支持数据离线导入及在线数据实时接入。
- 分布式数据库
  支持 SQL 标准：支持 SQL 2003 核心扩展的分布式关系数据库，完全兼容 PostgreSQL 的 SQL 语法，支持主键、触发器、约束、函数、存储过程、跨节点 join 等绝大部分的 SQL 特性，同时满足百 T 级数据规模的 OLTP 和 OLAP 应用场景。
  高性能、可扩展：单机可达 20000TPS，支持服务器在线扩容，扩容后性能表现接近线性扩展。
  内核级分库分表：内核级支持数据库分库分表，分库分表逻辑对业务完全透明化，简化业务的数据访问逻辑。
  内核级冷热数据分治：内核级支持冷热数据分治，业务无需感知底层存储介质的差异，对外提供一个统一的数据库视图，可有效降低服务器硬件成本。
  高可靠：可选多份数据热备，保障系统高可用，故障秒级切换。
- 敏捷报表门户
  易用：可视化数据源配置，可视化自助创建报表门户，无任何 IT
  基础也可配置专业级运营报表门户，轻松把握业务脉搏，助力企业决策。
  多渠道触达：勾选报表门户中已有指标源即可实现数据内容的可视化配置推送，支持邮件、微信渠道报表定向推送，助力决策者实时调整商业决策。
  丰富的可视化模板：内置十余种图表模板，表格、曲线图、柱状图、饼图、雷达图等主流图表模板一应囊括。
- 机器学习
  丰富：面向数据科学家的专业机器学习算法开发平台。集成 Spark、Python、R、XGBoost 等 4

种机器学习框架，支持图计算和深度学习，内置分类、回归、聚类、关联规则等 60 余种丰富算法。
拖拽式开发：可视化的 Web 拖拽式任务流开发，能够让算法工程师和数据科学家从数据和模型的角度以最自然的流方式来思考，充分激活企业大数据活力。
团队协作与知识共享：支持机器学习任务的团队协作开发，提高企业数据探索发现效率，有效助力团队的知识沉淀与共享

开箱即用的数据治理工具

开箱即用的数据治理工具，面向企业数据治理需求，提供完善的数据元信息管理功能。支持细到字段级别的数据权限管控，包含库表数据字典、数据血缘跟踪与溯源、热点数据分析等特色功能，以帮助企业客户提高海量数据资产的管理效率。包括但不尽限于以下功能：

- 数据权限管控
  细粒度权限管控：提供对关系数据库、分布式数据库、HDFS 文件、Hive 库和 Hbase 的文件、库、表、字段级的数据权限控制能力，满足用户在不同场景下对不同粒度数据的安全管控能力。
  安全：支持基于项目、用户、角色纬度的数据权限验证和授权，及时阻断用户的数据非法访问请求，保障企业数据资产安全。
  数据访问审计：实时记录敏感数据访问行为以支持定期的安全审计，支持自定义敏感数据访问预警策略，严控内部数据安全风险。
- 数据字典
  海量数据的元信息管理：可视化 Web 式元信息管理工具，满足用户对海量数据的元信息检索、标注、数据口径标准化等诉求。让用户能高效的对企业数据资产进行管理、索引、查找，有效提高企业数据资产管理效率。
- 血缘分析、直系分析和重要性分析
  元数据分析及追溯：包含血缘分析、直系分析、重要性分析等数据治理工具。
  用户可通过元数据分析直观了解到数据的来源、数据之间的关系、数据与任务的计算关系、数据流向、数据被引用次数等重要信息，便于用户直观的把握数据资产状况。
- 自助提数
  成本优化：降低提数门槛，无 SQL 基础的业务专家也可以自助提数，大大释放 IT
  技术部门压力，有效降低企业人力成本。
  效率：业务人员也可自助提数，不再需要提交需求到技术团队，减少沟通环节，提数周期从周降低到分钟级别，可大大提高企业的商业决策效率。
  安全管控：完备的数据权限管控机制始终贯穿自助提数的整个环节，在降低成本的同时更降低数据安全风险。

一站式运维管理平台

一站式的可视化运维管理平台，包含一键式集群部署、增量部署、丰富的可视化运维工具、完善的面向多租户的计算资源管控体系和完善的用户权限管理体系为客户提供企业级的大数据平台运维管理能力支撑。包括但不仅限于以下功能：

- 便捷部署

  一键部署：企业级大数据平台一键式部署，用户搭建部署稳健大数据平台的耗时从数周降低到数小时。

  增量部署：存储、计算节点适应企业数据从 GB 到 PB 的数据规模增长一键式线性扩容。

  30 余组件一键式增量部署，用户可根据企业的快书发展实时调整大数据架构。
- 仪表盘式运维

  可视化运维：集群运维仪表盘的实时呈现，支持仪表盘自定义配置，服务健康度跟踪告警，为用户提供集群运行状态实时感知能力。

  监控告警：支持自定义短信、邮件渠道的服务异常告警。

  20 余纬度的指标监控覆盖所有 30 余组件的运行状态，让运维人员对集群服务监控指标了如指掌。

  备份：支持 Hdfs 文件、文件夹的自定义备份策略，支持 Hbase 表的自定义备份策略。
- 资源管控

  资源隔离：完整多租户方案面向企业提供部门级的计算、存储资源分配与隔离，有效防止租户间任务的相互影响。

  动态调整：支持资源的动态调整，结合完善的资源指标监控系统可为用户极大程度的提升系统吞吐量，降低总体IT硬件成本。

  灵活：支持以项目，角色为主体的数据资源、计算资源申请与使用，系统管理员分配资源更便捷直观。
- 项目管理

  可视化：项目任务的可视化运维，包含项目内的实时、离线、机器学习等任务的运行状态指标。

  基于角色的项目管理：内置项目管理员、运维工程师、开发工程师三种项目角色，满足大多数部门级大数据处理场景。

  支持基于项目的角色自定义，企业客户可根据企业特点打造专属的大数据项目管理模型。
- 用户及租户管理

  灵活：支持基于用户、用户组、项目的用户管理体系，满足不同发展阶段企业客户的需求。

  安全：单点登录，统一访问策略体系有效支撑管理员对用户的功能访问进行权限管控，杜绝内部风险。

产品优势

为什么选择腾讯云大数据处理套件？

全链路大数据探索平台

企业可依托平台安全、便捷的进行数百 PB
级别的大数据的集成、处理、存储、分析、展现、机器学习等数据开发任务。

安全性

- 国际认证的系统安全加固技术保障系统级数据安全。
- 自定义算法的数据加密，确保数据在传输、存储过程中的安全管控。
- 全平台单点登录，统一策略管控中心，支持基于角色的列级数据管控体系保障数据访问安全。
- 健全的访问审计及预警模型，助力安全事件的事后追踪和企业的定期安全审计。

稳定性

经过多年的海量数据处理，腾讯沉淀了大量的数据分析挖掘技术和运维、数据开发、运维、数据治理等工具，能有效助力客户的大数据应用开发和大数据平台的平稳运营。

易用性

- 数据接入、处理、存储、分析、展现、机器学习的拖拽式全链路大数据开发。
- 企业级大数据平台一键式部署。
- 开箱即用的数据治理工具集。

开放性

技术源于开源社区，知识迁移平滑，运维管理简单，无需投入大量的人力物力替换原有体系。

低成本

- 冷热数据区分及差异化高强度压缩技术有效降低至 72% 的存储成本。
- 调度算法优化，高计算、高 IO 的高效分时混合技术可让内存、CPU、网络资源利用率同时达到 90%，有效降低服务器硬件成本。
- 低门槛的数据分析与挖掘平台，业务专家也可进行数据的分析挖掘，有效降低企业人力成本。

可用性

- 数据节点分布式部署，可选多份备份。
- 所有系统控制节点主从热备，故障秒级切换，腾讯 95% 业务考验，可用性 99.999%。

可运维

- 超大规模服务支撑，单集群可支撑近万节点。
- 涵盖服务器运维、组件运维、任务运维、诊断等功能的一站式运维平台。
  组件热插拔设计、秒级部署到端。
- 监控指标覆盖所有 30 余组件，运行异常实时感知。
- 无缝对接自有监控告警系统的实时邮件、短信告警。

高性能

- 高性能数据接入引擎，内部业务日接入五万亿条数据。
- 性能全面超越社区方案，数据处理能力提升 30% 左右。
- 支持上万维度、千亿规模数据的秒级交互式多维分析。

服务与支持

- 专家级架构咨询及技术咨询服务。
- 7x24 小时服务支持，一对一指导。
- 支持电话咨询、QQ 远程协助。

应用场景

腾讯大数据处理套件 TBDS（Tencent Big Data Suit）适用于企业从 GB 到 TB、PB 级的大数据处理场景，包括但不限于以下场景：

数据仓库建设

- 大数据处理套件完整覆盖数据抽取、转换、加载、建模、分析、报表呈现、数据治理等数仓建设环节，用户可借助 TBDS 大数据套件在公有云、私有云、非云化环境快速建设 TB 到 PB 级的企业数据仓库和数据集市，搭建专属的大数据应用。
- 通过大数据处理套件，用户可显著降低基于企业数据仓库的数据应用开发周期，降低开发成本，还可大大降低数据仓库、数据处理、数据应用的运维成本。

实时流式数据处理

- 用户可基于腾讯大数据套件快速开发本行业在实时流式场景下的大数据处理、分析的应用程序，以实现对企业实时业务的风险监控与告警，以占据大数据时代的优势地位。
- 流式数据处理可用于金融行业的风险管控、物联网的海量传感器数据处理、工业生产线的实时故障预警、病人特征数据实时分析、实时交通流量分析、互联网实时流量分析等应用场景。

离线数据处理

- 腾讯大数据套件基于 Hadoop 体系的 MapReduce、HIVE、PIG、SPARK 技术向企业用户提供的强大的数据离线批处理能力，用户可以便捷的使用腾讯大数据套件对企业数据进行抽取、转换、加载等离线数据处理加工。
- 通过离线数据处理引擎，用户可迅速的对企业所积累的数据进行 ETL 处理，快速发掘海量历史数据的商业价值和社会价值。

数据分析与探索挖掘

- 通过腾讯大数据处理套件所提供的强大数据分析与探索挖掘能力，用户可快速对企业在 PB 级规模下的大数据进行可视化的数据分析探索，在纷繁复杂的商业数据中快速获取数据洞察力，占领商业先机。
- 用户还可通过腾讯大数据处理套件所提供的强大机器学习能力对企业数据进行深度挖掘，进一步发掘海量数据中蕴藏的无限价值。

相关术语

工作流相关术语

工作流：腾讯自研的任务调度系统，具有毫秒级任务下发，高可靠的特性，同时支持插件式扩展任务类型
数据分析：提供简单的 SQL 查询功能，可以将 SQL 语句转换为 MapReduce/Spark 任务运行，进行在线 Scala、Python、SQL 脚本调试。
机器学习：
让数据科学家和算法工程师提供更棒体验的机器学习平台，通过拖拽式任务流设计，灵活多变的运行模式，丰富的内置机器学习算法，支持多种机器学习框架，
并提供可视化效果，同时还有强大的团队协作和分享能力，支持多种场景下的多实例调度，
让用户享受机器学习的乐趣。

平台管理相关术语

项目：
项目是大数据平台所有资源管理的基础，所有应用系统上线，都必须运行在分配好的项目之内，项目可以认为是一个大数据开发项目，也可以按照组织部门逻辑划分项目。项目包含唯一的资源队列，包括
CPU、内存、存储空间等，不同资源队列之间分配优先级；
用户：大数据平台的使用用户，隶属于项目，拥有开发、运维、管理员等数种项目成员权限身份。
角色：
分配给用户在大数据平台的权限划分，包括系统管理员（超级用户）、项目管理员、项目开发、项目运维。
资源：包括计算资源和存储资源，计算资源是 yarn 资源可以调度分配 CPU 和内存资源；存储资源是 HDFS 的存储空间。
资源池：是 yarn 分配指定资源队列，提供计算任务时需要的资源。

数据管理相关术语

可管理库表：数据库表的责任人具有对库表的可管理权限。
可读写库表：
数据库表所属项目的成员对库表具有可读写权限，其它项目成员可通过申请库表权限功能开通可读写权限。
无归属库表：系统扫描到的数据库表（一般是用户在登录机器命令行创建），无责任人和项目归属关系。
数据血缘：数据产生的链路或者路径，比如通过数据 A 数据 B 产生了数据 C，那么 C 的父血缘就是 A 和 B，反之亦然。在大数据套件中描述数据“父子”关系，以思维导图形式展现了数据变化影响和数据生产溯源，清

晰刻画表与表之间、任务与任务之间的关系。

直系关系：父级到子级数据表的任务处理流程关系。

血缘回溯：分析原始数据到目标数据表的计算路径。

数据字典：

对数据的数据项、数据结构、数据存储等进行定义和描述，其目的是对数据流程图中的各个元素做出详细的说明，使用数据字典为简单的建模项目。简而言之，数据字典是描述数据的信息集合，是对系统中使用的所有数据元素的定义的集合。

数据展现相关术语

数据源：是报表展示的数据来源，一般指数据库表；

维度：

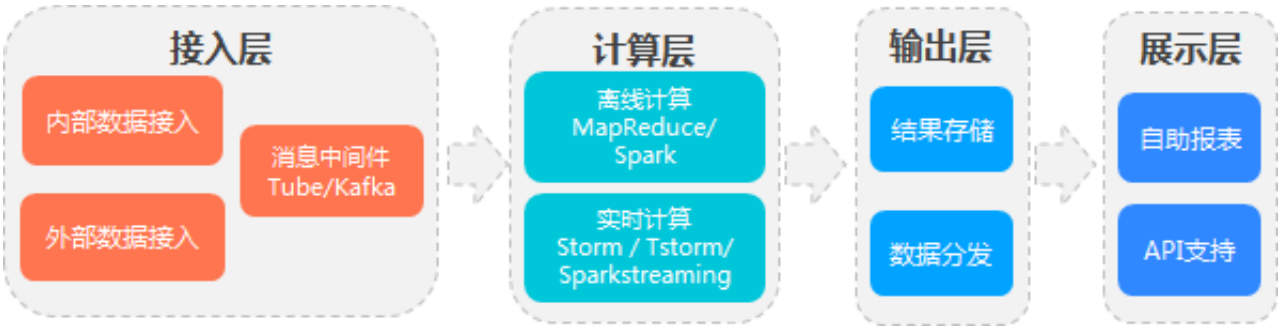
是报表展示的可指定不同值的对象的描述性属性或特征，是报表展示分析的角度。例如，网页浏览统计的浏览器类型维度，该维度包含 IE、Chrome、Firefox 等浏览器，地理分析中的城市维度，该维度包含北京、上海、广州等地点。用户可以使用维度来整理、细分和分析数据，也可以通过添加和删除维度来查看数据的不同统计特征。

指标：

是报表展示的聚合统计对象，通常是指数据字段，可以按总数或比值衡量的具体维度元素。例如，维度“城市”可以关联指标“人口”，其值为具体城市的居民总数。

数据处理流程

通常情况下数据处理主要包括数据接入、工作流、数据分析、数据展现与数据管理。流程图如下：



各模块说明如下：

数据接入

支持 Hippo 的数据实时接入，可以满足具有高可靠高可用应用场景的业务需求；支持多种主流关系数据库高效导入，支持文本类日志数据离线导入；支持高强度数据压缩及加密传输，历经每天峰值数据接入考验。

工作流

腾讯自研的任务调度系统，具有毫秒级任务下发，高可靠的特性，支持多种离线任务类型和调度策略，满足复杂数据挖掘要求；可视化工作流开发IDE，拖拽式任务开发；丰富的任务模板，满足快速便捷流计算任务创建。

数据分析

通过 web 界面与集群之间的操作来进行交互分析处理数据。提供数据分析与探索挖掘功能，助力用户发现大数据价值；开源分布式分析引擎，满足企业级用户数据集市建设需求；支撑万级维度、千亿级规模下的秒级数据统计分析，支持数据离线导入及在线实时接入。

数据展现

利用可视化工具，有效展现数据分析结果。拥有专业报表工具，丰富图表控件，无需代码即可上手；提供多样、高效实时的全流程数据平台体系；自定义多终端推送，线上线下无缝掌握数据脉搏。

数据管理

腾讯大数据平台提供细而全数据权限验证和授权，便于用户直观的把握数据资产状况清晰了解全局信息。支持多渠道全方位进行统一维护；支持血缘、直系、重要性分析，便于用户直观的把握数据资产状况；清晰了解全局信息，支持多渠道全方位进行统一维护