JDT 京东科技

大数据产品白皮书

京东科技大数据团队出品

目录

前言

大数据时代的超大数据体量和数据的多样性，已经超越了传统数据库的管理能力，大数据技术是新一代的技术与架构，它将帮助人们存储管理好大数据并从大体量、高复杂度的数据中提取价值。

从国内的大数据技术和行业应用发展来看，大数据技术的基础架构技术已日趋成熟，大数据领域由技术创新驱动转向应用创新驱动的趋势开始显现，但更多的传统企业在如何建设大数据平台，如何利用大数据来驱动企业业务发展上仍然缺乏经验，制约了大数据技术的大规模产业应用。

京东科技大数据平台建设了完整的技术体系，包括离线计算、实时计算和机器学习平台，可以满足多种复杂应用场景的计算任务，元数据管理、数据质量管理、任务调度、数据开发工具、流程中心等构成了全面的数据运营工具，数据可视化平台产品提供了便利的数据分析功能，敏感数据保护，数据权限控制等策略，最大程度的保护了数据资产的安全。

京东科技期望以数字、技术、科技驱动价值新增长，随着国家大数据战略的逐步落实，我们希望能输出京东科技的大数据技术能力，建设开放的技术平台，借助技术和数据为企业、政府和社会创造更多价值，同时，我们也欢迎更多的合作伙伴一起，在大数据领域一起深入探索，为我国大数据产业的发展贡献力量。

京东科技大数据产品体系

京东科技大数据平台源于京东丰富场景的最佳实践，提供便捷、稳定、安全的一站式大数据服务，可实现数据资产的集成、存储、开发、管理、服务、应用全数据流程场景，帮助用户专注数据价值的探索与挖掘。

在数据内容构建层面，基于多年大数据实践经验，京东科技沉淀出了自有的数据建设方法论，可以支持全域数据内容体系的快速构建，全局控制、科学管理，从而高效的使用数据资产。

在数据资产管理层面，基于数据资产管理的战略规划、数据治理的制度规范、工作流程以及数据治理的实践，建立起数据的标准化、规范化的管理方案，同时孵化数据治理相关工具，对数据进行深度治理，保障数据的可靠性、可用性，准确辅助决策，获得更精确的应用效果。

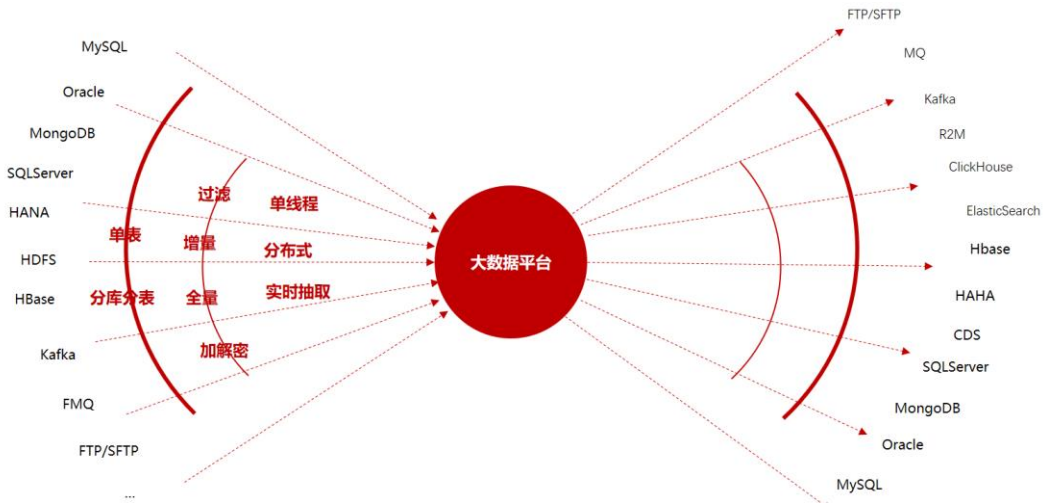
1.1 数据集成

数据采集是大数据的基石，企业内部各业务系统经过业务生产活动产生了大量的数据，但同时也形成了一个个独立的数据孤岛，通过数据集成工具，将

企业内部的全域数据进行汇集，在大数据平台侧形成企业统一的数据中心，经过数据的加工、处理将数据之间的各种信息重新组合，实现数据上的深度融合，从而更好的构建企业内部的数据资产，为上层数据应用提供的数据支撑。为了将企业内部的结构化和非结构化数据进行采集，京东科技搭建了一套标准化的数据采集工具-数据管道，可以实现各类异构数据库、数据来源的数据汇集到大数据平台。

数据管道提供了一套完整的数据集成解决方案，为后续的查询、分发、计算和分析提供数据基础。数据管道提供丰富多样、简单易用的数据采集功能，可满足离线计算、集成分发等多种需求，并进行全程状态监控。

基于数据入仓和出仓场景，数据管道支持数据接入和数据推送两种不同的数据同步方式。其中数据接入可以完成各类数据库的数据接入大数据平台，支持的数据源类型包括：MySQL、SQL Server、Oracle、MongoDB、HBase、ElasticSearch、离线文件等；数据推送则可以将在数仓中加工好的数据推送至指定的目标数据库内，从而帮助应用系统更好的基于本地数据进行数据应用，目前支持推送的目标包括 MySQL、Oralce、ElasticSearch、HAHA、Hbase、ClickHouse 等。



数据集成工具支持单表或者分库分表等不同来源场景的数据抽取，适配各类复杂的数据存储情况。

在抽取策略上，支持增量抽取和全量抽取数据，也通过条件过滤抽取的数据，已保证抽取的数据满足每日更新所需的数据要求。

在抽取方式上，支持通过批量方式和实时方式抽取，同时也可以启动分布式抽取，进行批量数据抽取，保证数据抽取的能力，尽早完成数据抽取

1.2 数据存储

基于 HDFS 京东科技完成对大数据分布式存储平台的自研升级，采用分布式存储技术，满足大数据高效可靠的存储需求，提供较高的持久性、较高的吞吐量和较低的延迟速度，具备高可用和高可靠的特点,容易扩展，并支持水平扩展至百 PB 级存储容量，同时拥有较高的硬件故障容忍能力，提供全面的安全性和多样化的权限功能.

分布式存储采用将元数据集群与数据集群分离并可实现独立扩展，用户既可以通过扩展元数据集群获得更多文件管理的能力，又可通过扩展数据存储集群获得更大的聚合带宽与存储容量，灵活、无缝、平滑的扩展方式可以为用户高效的计算环境提供坚实的数据保障。

- **数据高可靠和平台高可用**

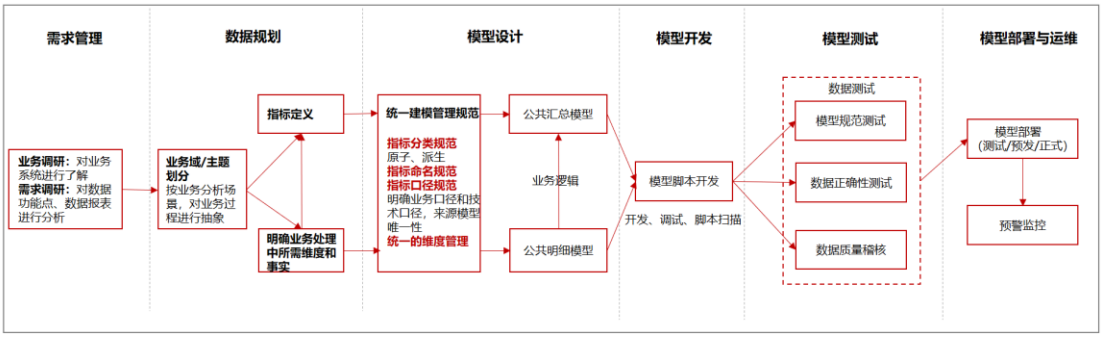
平台可用性是最重要的指标之一，需要保证在机器发生故障时，系统可用性不受影响。我们使用多副本策略，副本分布在不同的机器，机器故障引起某些副本失效时，其它副本仍然能提供服务。

- **集群水平扩展能力**

存储平台中的元数据服务器和存储节点是拥有横向水平扩展能力，存储节点扩展是存储节点数量的扩展，存储节点扩展带来容量上的增长，扩展过程中无需中断存储系统上应用的运行，扩展的容量即插即用。而且随着存储服务器数量的增多，整套存储平台的流量也会线性的增长,成为核心存储集群之一。

1.3 离线数据开发

数据建模、数据开发、调度服务，作为离线开发的组件，为离线数据设计、开发、调试、运维提供的全流程的工具服务。



数据开发流程

数据建模

通过建模工具提供的拖拽，所见即所得的交互方式，可快速完成数仓的整体架构设计，包括构建数仓的整体分层、数仓的集市划分，主题域与主题的设计，从而指导模型体系的建设。基于业务诉求、数仓模型体系的规划，进行逻辑模型、物理模型的设计，快速完成整个模型体系的建设。

数据开发

数据开发平台提供在线可视化交互 IDE 工具，支持多人协同开发、在线调试、版本管理以及脚本的调度。

产品优势：

产品特色：

调度服务

京东科技自主研发用于管理和运行大数据离线作业的分布式调度系统，用于控制作业的运行顺序、时机和资源。

产品特点：

- 作业管理：支持各类作业（集成、计算、质量、推送、同步等）的创建和管理；
- 调度引擎：支持日、周、月、不定期、小时级调度，包括对作业进行实例化、调度、执行作业
- 补数中心：对历史作业进行批量、链路补数补充执行；
- 监控预警：对作业运营情况进行监控，延迟启动、延迟完成、任务报

错、数据为空等等异常情况进行监控预警；

- 对外接口：提供丰富的api供三方进行集成
- 运维运营：支持对作业的空间、节点、标签进行管理

1.4 实时数据开发

RIC 平台是一个提供高性能低延迟处理实时流式大数据的计算服务平台计算能力上可以支持 PB 级毫秒级别的处理延迟，并且可以通过简单的 SQL 开发就可实现复杂的实时场景。能够支持多种数据格式接入或落地。

京东科技选择 Flink 作为实时计算平台的技术框架，是在技术、底层框架和业务场景三者权衡选择后的落地，Storm 作为第一代增量计算的高速事件处理框架，它的毫秒级延迟满足对延迟要求较高的场景。而 SparkStreaming 的有状态计算，恰好一次的传递，对延迟要求不高以及图形操作、机器学习、SQL 支持等特性恰恰是对 Storm 的完美补充。Flink 的出现可以说是以上框架优势的集大成者，Flink 支持 Storm 式的增量迭代，毫秒级延迟，吞吐量比 Storm 高，并且具有对迭代自动优化的功能，支持 Exactly Once 、状态管理和窗口统计，在迭代式数据处理上，更是比 Spark 更突出。

在技术层面，实时计算平台不是对开源技术的照搬，我们汲取社区开源框架的优势，取长补短，对开源技术进行二次开发；产品层面，通过支持 SQL 引擎，提升用户体验，提高使用效率，降低用户的学习成本，让用户通过 SQL 方式体验实时技术所带来的魅力。

产品特点：

- 平台化：提供完善的流计算开发平台，时刻保证作业的高效和稳定，让平台的使用者更专注业务实现和场景的应用。
- SQL 化：提供 SQL 语义的流式数据分析能力，通过简单的配置即可获取 JMQ 实时数据来完成复杂业务加工，大幅降低流数据分析门槛，高效轻松的完成实时业务开发。
- 智能化：针对运行中常见的问题，平台会根据相关策略进行自动化的配置，提供作业优化方案，帮助用户更好地管理自己的项目和作业。
- 稳定可靠：平台为计算作业提供了全方位的监控与告警服务，遇到问题第一时间通知到用户，让用户在任何时间任何地点都可以了解到作业运行情况。

1.5 数据服务

数据服务平台为数据转服务提供了便捷工具，可基于数仓数据、业务应用表数据快速配置化生成接口，同时也提供数据接口注册、管理、监控、治理、网关服务等涵盖数据服务全生命周期的能力，实现数据的快速应用。

数据服务平台提供高性能、安全可靠的数据获取服务，有效促进了企业内部的数据共享和服务共享，同时为数据服务提供了全生命周期的管理能力，提升数据服务的管理效能。

产品特性：

- 数据源管理：基于用户数据获取需求进行数仓模型表、业务应用表等接口数据源管理，支持 MySQL、分库分表、Hbase 等各类异构数据源

- 基于数仓模型表、业务应用表快速配置、测试、上线 API，帮助快速获取相应数据源中的数据，支持将已有 API 服务快速注册至平台中，利用平台能力进行全面的管理

- 接口服务管理：针对在平台上配置或注册的 API 进行全面管理，包含限流、降级、熔断、授权、鉴权、分流、编排等，保障接口服务的高效和稳定

- 服务及数据监控报警：实时监控接口服务、字段运行状态及关键指标等，支持自定义监控大屏进行实时展示或离线统计，同时支持短信、邮件等方式对异常情况进行报警，快速发现并处理问题

- API 市场：平台管理的服务可公开至 API 市场，供所有 API 使用者进行 API 的查看和选用，有效实现数据和 API 的共享

- 租户及应用管理：按照租户级别进行服务的开发和管理，实现租户间的资源隔离，按照应用级别进行服务的授权鉴权和使用

1.6 画像标签

画像平台为企业立体刻画用户画像提供了一站式服务，通过构建全面、精准、多维的用户画像体系，快速圈选目标人群，全方位的人群探查分析，与投放渠道链接，进行精准触达，从而助力发掘运营高转化率，提高业务增长，让营销更加精准、高效。

产品服务对象

服务	用户	场景	支持内容	服务方式
画像标签池	（各业务线）产品、运营、分析师、算法、建模人员	建模分析 行为分析 营销效果分析等	（标签、画像）数据开发+数据存储+数据分析计算	1、数据开发 2、数据处理和分析
画像系统	（各业务线）产品、运营、分析师、算法、建模人员	画像分析 精准投放 用户识别 广告投放 智能客服 风险洞察 内部审计 业务决策	标签市场 创建人群 人群画像 人群应用 效果评估 智能标签推荐 数据监控	1、标签/人群/分析运营操作页面（画像系统） 2、效果数据评估 3、人群优化和推荐
接口服务	业务系统场景：投放系统、应用系统、用户管理系统、分析系统等	精准投放 广告营销 用户识别 风险管控	标签取值 人群创建 人群命中 人群画像 人群下载 人群上传 人群数量预估	1、整套数据接口服务 2、接口级数据监控

产品核心能力：

- 标签市场：提供企业全量标签的展示平台，通过标签市场用户可查看标签的
- 群体圈选：支持圈选、文件、库表、二次筛选等多种群体创建方式，快速完成用户聚类，锁定目标群体，分析更精准
- 画像分析：全面立体洞察目标群体，勾勒细分群体的特征偏好，为荆溪壶运营提供全方位的画像支撑
- 应用触达：具备成熟度完备的开放接口，实现系统快速对接，充分发挥群体数据价值，辅助策略分析、风险洞察、营销决策等应用场景的闭环的支持，帮助企业实现千人前面、精准营销、精细化运营、风险识别业务应用。

1.7 数据可视化

作为京东科技大数据可视化自助服务平台，Agile BI 提供了多种数据可视化图形，支持 MySQL、Oracle、Presto 多种数据源类型接入，拥有丰富的图表样式、多样化的展现渠道，为数据提供了一个解读的平台。帮助业务快速看到数据，识别数据，追踪数据，为用户进行数据分析、数据挖掘、数据可视化提供重要的工具和支持。

用户无需任何 SQL 基础，仅需简单灵活的拖拽即可实现各类主题报表，即可快速完成报表开发工作，为每个人提供解读数据的可能。借助订阅周期性邮件任务，让晨报自动发送，让用户从容安排自己的时间。最终用户也可通过移动设备随时随地查看数据，让数据驱动更好的走进每个人的工作中。

功能特性：

- 丰富的数据源：支持关系数据库、文件、Hive、接口等各类数据源
- 秒级响应：页面秒级展现，支持数据缓存，支持实时看板
- 多终端发布：支持 PC、APP、内嵌等多种方式进行多终端数据展示
- 灵活的多维分析：支持进行数据的多维分析、数据探查
- 丰富酷炫的图片：支持看板、报告、大屏等多种展示形式，同时提供了丰富多样的图表类型

1.8 机器学习平台

机器学习平台 KuAI 是面向开发者的一站式 AI 平台，基于海量的数据及强劲的计算资源，搭载 Sklearn、XGBoost、TensorFlow、Pytorch 等机器学习/深度

学习框架，提供从模型开发、训练到部署、监控的一站式服务，帮助用户快速构建、部署模型，并实现 AI 工作流全生命周期管理。

KuAI 平台基于云原生架构，实现内核级的服务器虚拟化、秒级启停，基于 K8S 对 Docker 进行编排调度，实现对服务器资源的灵活管理调度和动态扩/缩容，有效提高服务稳定性和资源利用率。

产品特性：

1.9 图数据分析

在没有图计算平台时，业务会面临着各种各样的问题，例如在进行一些案件跟踪时跨多系统多行为链多支付链的数据多度关联难以实现，海量数据下传统的图计算引擎计算性能不足无法满足计算性能要求，各类个性化的图可视化需求无法获得满足，由此继续一个简单、便捷、高效的工具，能够将图数据

库、图数据、图分析的场景进行简单化，为挖掘数据关系背后的潜在价值提供高效的帮助。

京东科技自研的图计算平台，通过构建数据库、数据分析引擎、数据分析一体化的工具，便于进行数据关联和深层次分析。自研的高性能图数据库，能够支持万亿级带属性的节点和边的数据管理，并支持高并发场景下满足毫秒级低延时的查询诉求。

通过对关系数据的组织、可视化及深度分析和探索，帮助业务人员从复杂关联数据网络中洞察有效信息并挖掘出更深层次的价值，目前广泛应用于风控、营销、宏观分析、数据资产管理等业务场景。

产品特征：

- 强大的可视化能力：提供图数据建模、图数据处理、图数据查询和图智能分析的各类功能，全流程可视化操作
- 自研的 JoyGraph 系图分析引擎：覆盖主流图分析算法和图神经网络算法，深度优化的自研单机图分析引擎 JoyGraph、图分析引擎 ReJoyGraph 和图神经网络引擎 DeepJoyGraph，性能百倍于主流开源图分析引擎
- 低门槛：可支持业务、数据、算法、策略分析师等各类人群快速、简单的获取图相关数据能力，支持业务进行图数据探索分析

1.10 数据治理

京东科技自主研究了数据健康评价体系，通过对数据资产的各类角度进行健康状况探查，快速分析出企业数据资产当前的健康问题，同时通过提供可视化工具，为数据生产者和数据管理者实现线上化治理。

通过生产者视角的治理工具，可帮助数据生产者快速自查自有资产的健康状况，并提供治理工具快速实现问题的自治。

通过管理者视角的治理工具，可帮助管理者快速分析全域资产健康状况，各个问题的治理进展情况，以及当前遗留的问题情况，通过制定专项治理计划，强制用户配合进行相应的数据治理，以此完成数据的高效治理。

数据标准

当企业内部数据严格进行数据的规范化管理，企业内部的数据共享、数据应用、数据问题追溯等问题上能够更加高效和便捷。而数据标准的贯彻执行除了靠运营管理外，离不开工具的支持，通过标准化工具进行标准的管理、审核、发布、应用、快速制定数据标准、应用数据标准。

数据标准管理实施过程主要分为六个步骤：标准体系规划、标准编写制定、标准评审发布、标准落地执行、标准运营维护、标准评估监控，数据标准是数据质量管理的重要抓手，通过建立企业级的数据标准，达成业务部门和 IT 部门对数据特征的一致意见，形成信息在各个业务模块和系统之间的统一规范，解决数据不一致、数据不完整、信息孤岛、无法有效整合等情况引起的数据质量问题。

一个企业在数据标准方面核心建设的内容主要包括：

1) 数据标准管理体系的框架搭建。包括组织人员、制度、流程、技术四个方面。

2) 数据标准管理主要流程的设计：包括数据标准编制流程，数据标准执行流程和数据标准修订流程；

3) 数据标准体系设计和编制规划；遵循行业标准，监管部门（银监会等）数据标准化要求，并借鉴行业优秀经验，按照数据业务主题，包括客户主题、财务主题、风险管理、运营管理主题以及相关公共代码的数据标准编制。

4) 数据标准的落地实施：通过对源系统进行数据标准的满足度评估来推动数据标准的落地执行。

数据质量

数据质量是有效分析和利用大数据的前提,是大数据产生跨区域、跨行业、跨部门价值的保障。引发数据质量问题的原因总结起来如下：

数据创建产生：

如数值为空、数据内容和描述不符、数据精度不足、数据默认值使用不当和数据录入的校验规则不当等；

数据获取产生

如数据结构错误、数据获取不完整、数据采集点不正确、取数时点不正确等等；

数据传递产生

如接口数据传递延时、接口数据漏传、网络传输不可靠丢包、数据传递不及时等；

数据加工产生

如数据清洗和加工逻辑不正确、算法错误导致数据多算、漏算等；

围绕上述数据质量产生的原因，结合京东信息系统建设的特点，京东大数据定义数据质量好坏可以从以下几个指标来描述：

及时性：数据平台是否满足业务应用对数据的时间要求；

完整性：数据平台是否包含了业务应用所需要的所有数据；每一份数据的记录是否完整无缺；

准确性：获取的每一份数据是否存在异常或者错误信息；数据平台在数据的获取、传递、加工过程中是否能保证数据的准确；

可用性：多维度、多渠道获取的数据是否能够易于理解并使用；

京东科技大数据自研了数据质量监控平台，实现了对数据质量的管理，主要功能如下：

- 数据准确性监控

针对每张数据表，可根据需要定制规则，对数据记录行数和数据值进行统计，如订单金额等，可与历史的数据值进行比较，通过自定义的阈值进行告警，比对的周期可以定义为日、周等不同周期，实现数据准确性的预警。

- 数据及时性监控

将数据加工任务按照服务的业务线归类，根据服务 SLA 设置完成时间的阈值，用户可以随时观察任务运行的状态、时间等，也可以通过设置阈值指标，将告警值通过邮件、短信等方式及时反馈给运维人员。可根据应用、业务、任务三层监控对象进行告警配置。

- 数据完整性监控

针对单行数据记录，可定义空值检测、枚举值范围检测等，可根据定义的规则进行告警。

- 数据质量事件

当发生数据质量异常告警后，将生成一个数据质量事件，该事件由数据运维人员发起，及时查明异常原因并记录在知识库中，该事件的完成必须经过上级负责人的审批。

- 数据质量分析报告

用户可根据需要查询数据表的质量运行报告，包括原因、状态、处理结果等，生成的质量分析报告对于数据质量的改进提供参考。

除了在技术手段上进行数据质量管理外，在管理制度上我们也进行了规范，如上游数据库表变更通知、数据质量问题必须当日进行记录及解决等。

数据安全

数据安全对于各家企业都非常重要，京东科技大数据平台围绕数据的全生命周期提供了细致全面的数据安全管控措施，主要围绕以下几个方面：

- 敏感数据加解密

将涉及用户隐私的数据在入仓时就进行应用级的解决，然后在通过数仓特有的加密逻辑实现数仓级别的加密，实现数据在数据仓库内部的加密存储，存储采用了国家认证的密钥算法，在数据模型上提供可逆列和非可逆列，其中数据可逆列方便数据在被应用层使用时可先在数仓解密再按应用加密的方式保障数据的安全，数据不可逆列则主要用户数据的匹配，保障数据在关联时不会存在安全问题，同时在用户使用上采取严格的审批机制，必须经过多级领导、数据安全部门、大数据部门同时签批才可使用敏感数据。

■ IP 黑白名单机制

设立白名单，则在白名单中的用户会优先通过，不会被元数据服务器拒绝，白名单以外的用户都不能访问到元数据服务器. 白名单可以提高用户工作效率，并保持系统以最佳性能运作，规避威胁大数据平台的非法用户的违反行为，并提醒有关团队未经授权程序正在访问集群，让安全人员立即采取行动。

■ 最小化授权策略

大数据平台权限系统提供了一个和 HDFS 原生权限相匹配适应的授权模型管理访问策略，可以将用户的授权细化到文件级别，保证最小化的授权策略，所有授权均需通过相关数据集市相关的业务负责人审批。

■ 数据导出统计

针对所有从平台导出的数据以及集市之间交换的数据进行统计，制作分析报告，防范风险。

■ 全面的日志审计

所有的大数据平台的日志都将保留并定期收集进行审计。

成本治理

成本治理主要是站在资源管理的角度审视企业数据的监控状况，通过对存储、计算的分析，治理无效的存储、提升资源的利用率，降低或减少无价值的使用，帮助企业节省存储和计算资源。

在存储成本治理层面，京东科技提供了僵冷数据、EC、数据删除、生命周期管理、存储策略管理等多样化的方式实现对存储成本的精细化管理，通过主动治理僵冷数据，设置合理的生命周期，将有效改善企业内部存储数据的范围，将有价值的数据保留下来。

在计算成本治理层面，则通过僵尸任务治理，无效任务，孤点任务，降低计算资源申请，去除链路成环等各项技术，减少无效计算自研的浪费，同时通过合理优化资源队列，保障重点任务的执行资源，为企业核心模型时效提供相应的保障机制。