



天冕数据中台系统 产品技术白皮书

产品版本：V4

文档版本：20211008

目录

1. 产品概述.....	4
1.1 产品介绍	4
1.2 产品系统架构	4
1.3 产品逻辑架构	5
2. 产品核心功能.....	7
2.1 数据总线	7
2.2 流式计算	7
2.3 数据工坊	8
2.4 分析引擎	8
2.5 数据可视化	9
2.6 数据监控	10
2.7 数据安全	10
2.8 机器学习	11
3. 产品特性.....	11
3.1 一站式可视化交互	11
3.2 数据治理	12
3.3 性能和可靠性	12
3.4 全域数据整合建模	13
3.5 智能算法组件集成	13
3.6 灵活参数配置化	14
3.7 开放性和扩展性	14
3.8 便捷性	15
4. 大数据应用.....	15
4.1 实时数据共享	15
4.2 数据监控	15
4.3 数据可视化	16
4.4 模型应用	16
5. 大数据平台优劣分析.....	17
6. 系统运行环境.....	19
6.1 硬件环境	19
6.2 系统及软件环境	20
7. 性能与可靠性.....	20

7.1	稳定性	20
7.2	响应时间	20
7.3	并发访问	21

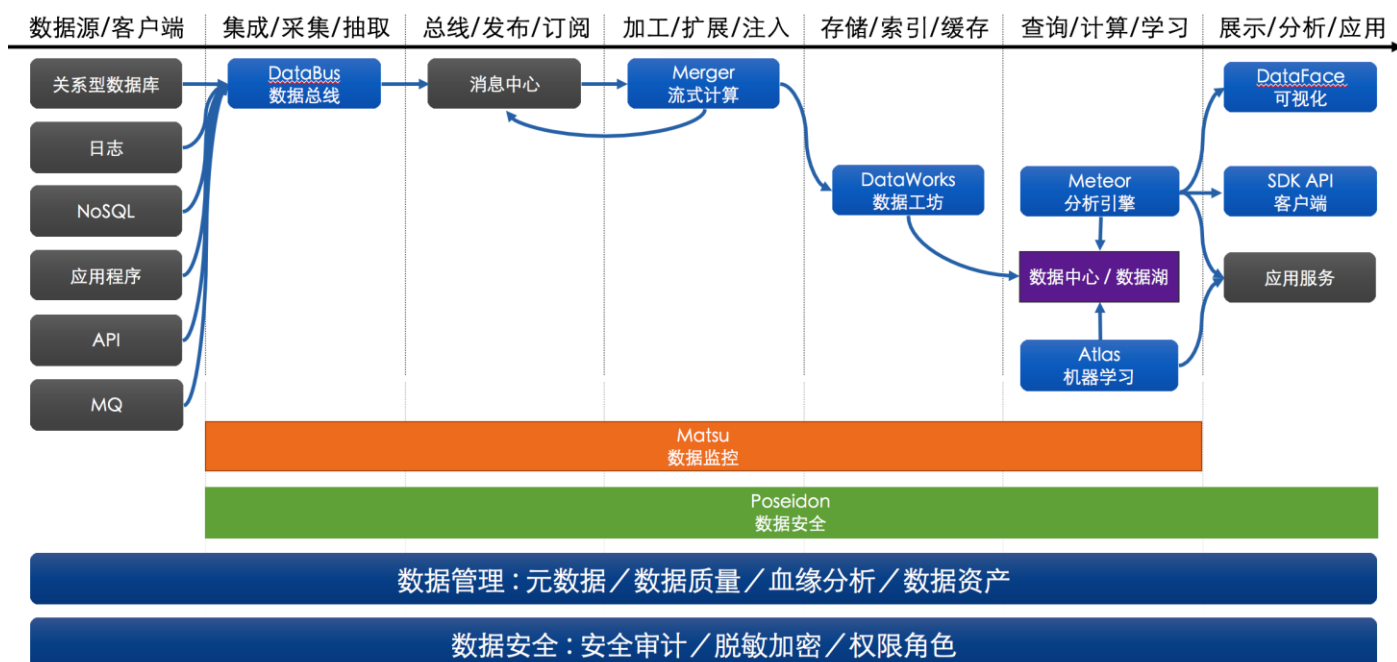
1. 产品概述

1.1 产品介绍

天冕数据中台服务（简称天冕数据中台）是天冕根据自身多年的实践经验推出的一款实时处理海量数据的一站式大数据平台。产品面向具有海量数据处理需求的企业，快速搭建数据的实时采集、存储、集成、挖掘、分析、可视化、机器学习、流计算的全流程数据处理平台，平台提供一系列工具和可视化交互将数据的管理自动化、流程化、规范化、智能化，用来支撑更轻量、灵活、低门槛、快速迭代的大数据应用，从而大幅度提升企业数据资产的质量和管理效率，帮助企业洞察数据中的关键信息，智能、快速的交付大数据应用。

1.2 产品系统架构

天冕数据中台在系统架构上采用组件化、服务化的设计理念，平台建立在开源的 Hadoop 大数据服务技术生态上，支持 TB / PB 级别数据的存储和计算，具有强大的数据整合、处理、分析、计算、挖掘能力。

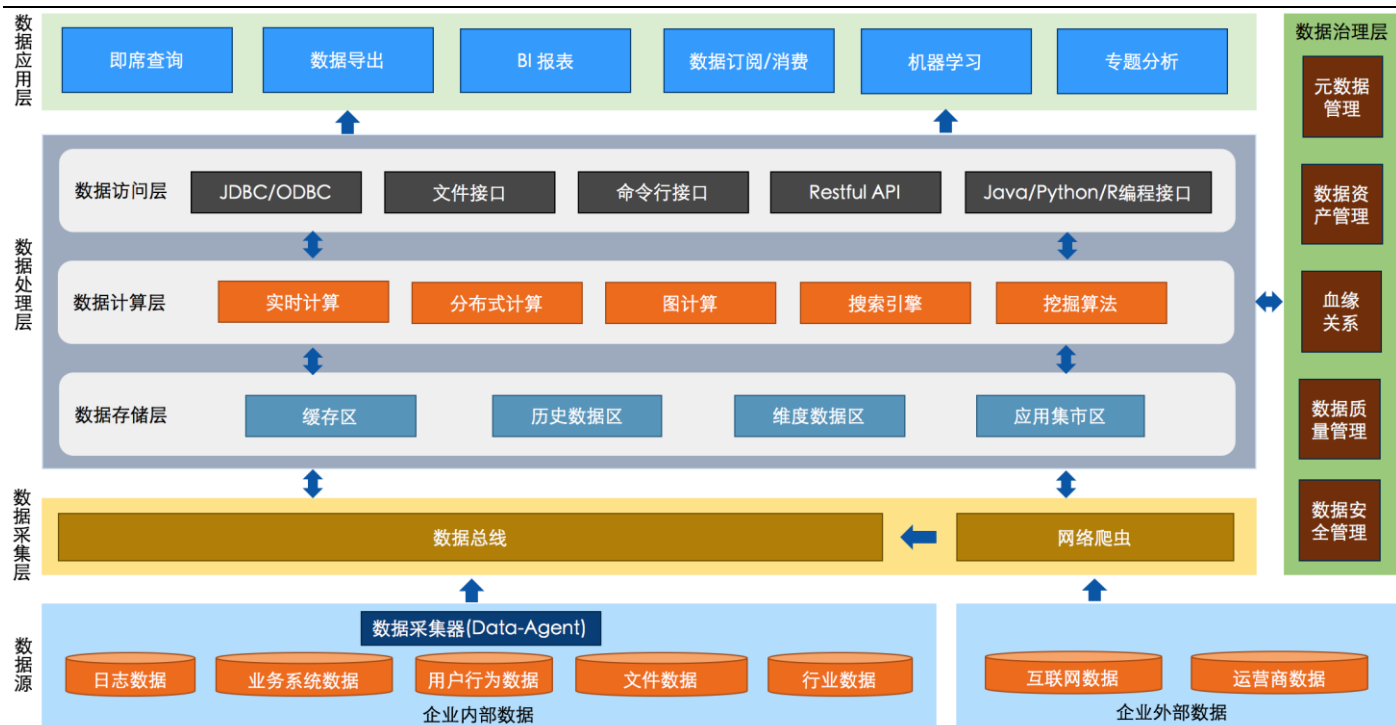


整个平台从系统架构上分为数据采集、数据处理、数据存储、数据分析、数据监控、数据安全、机器学习等组件，并将这些组件采用平台化的方式组织起来，提供可视化配置化的方式开箱即用，极大的提升了大数据应用效率。

- 数据采集：基于异步数据源的无侵入式的实时采集，支持丰富的数据源
- 流式计算：将采集的数据进行清洗、加工、扩展、合并、标签化等，并完成数据的质量管理，为全域数据汇聚与融合、数据治理与共享提供支持
- 数据工坊：将清洗完成的数据选择指定的数据库或存储平台存入数据中心，并进行智能化的索引和缓存的构建，整个数据流的处理无需人工干预
- 分析引擎：提供可视化方式进行数据实时多维分析
- 机器学习：基于机器学习沉淀出八十多种模型算法，可快速支持绝大多数场景下的数据建模和挖掘
- 数据可视化：提供数据展示的可视化界面
- 数据监控：自定义数据流转各个阶段的指标并预警
- 数据安全：覆盖全流程的安全审计、加密/解密、威胁拦截、单点登录等管理

1.3 产品逻辑架构

天冕数据中台在服务化的基础上采用了分层构建的理念。通过不同的层级来管理服务化组件，并实现层级之间良好的扩展性和开放性。



整个平台从逻辑架构上分为数据采集层、数据治理层、数据处理层、数据应用层，每一个层级之间独立且通过服务化组织到一起。

- 数据采集层：主要是通过源数据库上安装数据收集器（Data-Agent）将源数据的数据变更（新增、修改、删除）实时同步给数据总线，并通过数据总线来管理数据收集和分发的策略，安装数据收集器（Data-Agent）对源数据库无影响
- 数据治理层：主要是完成数据的标准化的工作，通过元数据管理、数据资产管理、血缘分析、数据质量管理、数据安全治理，确保大数据平台的数据一致、安全、有效。数据运维贯穿整个数据治理体系的流程中，实现平台化的管理思路
- 数据处理层：数据处理包含数据存储层、数据计算层和数据访问层
 1. 数据存储层：主要提供数据的存储的区域划分、冷热数据管理、索引管理、缓存管理等
 2. 数据计算层：主要提供数据查询优化、数据计算、数据合并、数据过滤等
 3. 数据访问层：主要提供访问数据的接口和数据展现的接口
- 数据应用层：提供数据的消费和应用的能力，同时可进一步对数据进行加工和挖掘

2. 产品核心功能

2.1 数据总线

数据总线（简称 DataBus）是数据中台的数据治理模块，主要包括数据的收集、集成、抽取、分发、数据清洗/加工/扩展、元数据管理、数据质量、数据资产管理等功能。

在数据源服务器上安装数据采集的 Data-Agent，在 DataBus 中通过可视化配置完对关系型数据库（MySQL、PostgreSQL、Oracle）、NoSQL（HBase、ElasticSearch、MongoDB）、日志文件、视频、图片、消息系统等数据的标准化管理工作，并建立实时的数据流管道。

功能特性

- 一键部署数据采集 Data-Agent，无侵入的方式接入多种数据源
- 海量数据实时传输与多路消息订阅分发
- 自动感知数据源 Schema 变更，并预警
- 数据实时脱敏、加密、标准化
- 可视化处理复杂的数据汇集与数据清洗策略
- 可视化元数据血缘分析、数据质量管理、数据资产统计

2.2 流式计算

流式计算（简称 Merger）是数据中台的数据流管理模块，提供可视化的操作将数据流的处理简单化、规范化、流程化、自动化。

功能特性

- 可视化配置和 SQL 开发流式应用
- 处理增删改三种类型消息流

- 流上跨表数据混算
- 基于日志多数据质量管理
- 数据血缘分析、数据标签化、数据分组
- 基于事件时间戳策略流式处理
- 数据的补齐、ID Mapping

2.3 数据工坊

数据工坊（简称 DataWorks）是数据中台的数据存储管理模块，提供将一份数据实时流式写入多种不同的存储或数据库，并根据选择的数据库类型智能的创建索引、缓存、元数据等。通过收集业务方使用习惯来智能化的存储加速和查询加速。

功能特性

- 多目标数据系统异步等入库
- 流上跨存储系统写入
- 在线的 DDL
- 智能小文件处理及文件合并
- 智能动态分区存储
- 智能动态索引
- 数据流式存储实时监控预警

2.4 分析引擎

分析引擎（简称 Meteor）是数据中台的数据分析和挖掘模块，面向数据仓库工程师/数据分析师/开发人员等，提供数据虚拟化分析解决方案，提供可视化、SQL、API 与 Meteor 交互，即可透明屏蔽异构数据系统异构交互方式，轻松实现跨异构数据系统透明混算，并智能化优化查询

策略，可较大幅度提升数据分析性能和分析效率，支持 TB / PB 级别数据亚秒级到秒级的响应能力。

功能特性

- 虚拟化能力，多源异构系统无缝混算（支持 MySQL、PostgreSQL、Kudu、HBase、Redis、HDFS、Hive 等）
- 统一扩展 SQL 语法查询操作
- 支持四种调用方式（REST 服务、JDBC 连接、ODBC 连接、SDK 客户端）
- 智能的查询优化策略（查询替换、复杂算子下推、整体快速下推、数据立方体临时缓存、抽样快速查询）
- 智能优化数据分析策略大幅度提升数据多维分析性能

2.5 数据可视化

数据可视化（简称 DataFace）是数据中台的数据可视化模块，数据可视化模块提供对大数据高效分析与展现的轻量级工具，支持 PC、移动端、大屏的三屏合一。通过对数据进行即席的分析与查询，实现数据的实时高效可视化展现。内置的智能查询加速引擎可以对海量数据进行实时在线分析，而无需提前进行大量的数据预处理，极大提高了分析效率。通过拖拽式操作和丰富的可视化图表控件，轻松自如地完成数据透视分析、自助取数、业务数据探查、报表制作和数据门户搭建等工作。极大降低了数据的获取成本和使用门槛。

功能特性

- 在几分钟之内创建数据仪表板和可视化效果
- 图形化编辑界面
- 支持丰富的图表组件
- 支持 PC、移动端、大屏的三屏合一

- 自定义数据概览、领导舱
- 数据上下钻取、透明钻取、旋转、分页层钻、跨维钻取等交互式分析

2.6 数据监控

数据监控（简称 Matsu）是数据中台的数据预警模块，基于在线的自定义监控流程和告警策略提供全流程的任意数据实时监控告警解决方案。

功能特性

- 基于数据流的实时的监控
- 自定义告警策略及告警阈值
- 支持数据质量监控
- 支持任意数据指标监控预警
- 支持扩展性的告警通知方式

2.7 数据安全

数据安全（简称 Poseidon）是大数据平台的全链路的安全防护模块，提供全流程的应用安全和数据安全的控制机制，除了权限分配与管理外，通过记录资源操作将审计事件保存到数据中心，利用保存的操作记录，实现安全分析、资源变更追溯以及合规性审计，为数据平台的安全提供行之有效的认证、授权、访问和审计的能力。

功能特性

- 支持统一身份认证和单点登录
- 支持应用安全和数据安全，可控制库、表、字段级别的细粒度权限
- 支持基于角色的权限定义
- 支持可视化的水印

- 支持数据安全审计、加密/解密、威胁拦截
- 支持全局的数据安全访问“堡垒机”
- 符合国家三级等保安全要求

2.8 机器学习

数据学习（简称 Atlas）是大数据平台的数据深度挖掘和分析模块，提供基于 Spark MLlib、TensorFlow、PyTorch 的数据模型开箱即用的数据挖掘方案。

功能特性

- 基于算法库实现数据的动态感知，支持常见的八十多种算法（线性支持向量机、逻辑回归、K 近邻随机、森林朴素贝叶斯、K 均值聚类、线性回归、GBDT 回归、协同过滤、多分类评估、二分类评估、回归模型评估、聚类模型评估预测、PS 线性回归等）
- 支持成熟的用户画像、特征工程、关系网络、风险评估模型

3. 产品特性

3.1 一站式可视化交互

天冕数据中台提供一站式的可视化的处理实时数据收集、计算、挖掘、分析、可视化；可视化交互界面结合关系网络、地图分析、信息立方、时序行为等多个维度的在线交互式探索可以快捷的完成数据的分析和挖掘，方便用户从网络、时间、空间三个维度视角来洞察数据。

核心指标

- 一站式完成数据收集、计算、挖掘、分析的实时处理
- 基于可视化的界面所见即所得的在线完成数据清洗、挖掘、分析
- 支持丰富的数据可视化，支持包括线图、饼图、柱图、累积、表格、地图等丰富数据展现形

式

- PC、移动端、大屏三屏合一的灵活自定义实时数据看板

3.2 数据治理

数据治理是数据管理的基石，天冕数据中台数据治理模块覆盖全方位的数据治理能力，包含数据标准管理、元数据管理、数据质量管理、数据资产管理等各模块协同运作，确保大数据平台的数据一致、规范、有效。

核心指标

- 实现数据标准化管理的系统化、可视化、简单化，并衔接元数据管理、数据质量管理和数据资产管理形成数据治理的闭环
- 提供血缘/影响分析功能，便于对数据进行分析判断、问题定位
- 提供统一的元数据管理工具，以支持大数据平台海量元数据的获取、存储、管理及应用
- 提供数据质量监控可视化，可以在界面上完成监控对象选择、规则配置、告警及结果查看、问题处理等数据质量全流程监控

3.3 性能和可靠性

天冕数据中台支持千亿级数据 join，百万级并发 job，作业 I/O 可达 TB 级/天，具备跨集群（机房）数据共享能力，支持数千节点的集群数，扩容不受限制；提供功能强大易用的 SQL、MR 引擎，兼容 SQL92 语法；系统采用多重备份、读写请求鉴权、应用沙箱、系统沙箱等多层次数据存储和访问安全机制保护用户的数据不丢失、不泄露、不被窃取；基于 MPP 架构并融合了分布式检索技术的分布式实时计算技术。

核心指标

- 拥有快速处理迁移海量数据的能力，使得数据分析中使用的数据可以不再是抽样的，而是业

务系统中产生的全量数据，使得数据分析的结果具有最大的代表性

- 采用分布式计算技术，拥有强大的实时计算能力，通常可以在数百毫秒内完成百亿级的数据计算，使得使用者可以根据自己的想法在海量数据中自由的进行探索
- 支撑较高并发查询量，并且通过动态的多副本数据存储技术来保证较高的系统可用性
- 在线业务 TB 到 PB 级别的数据量实时处理，亚秒到秒级响应，离线数据挖掘分钟级响应；灵活根据数据量和并发要求配置计算资源
- 任意数据访问和用户操作可事前监控、事中拦截、事后审计

3.4 全域数据整合建模

天冕数据中台以逻辑模型全流程自动化的建立替代耗时耗力的传统数据仓库物理模型，为分布在多个计算存储资源上的明细数据建立统一的数据中心。用户在完成业务数据梳理后，通过天冕数据中台管理后台轻松配置定义业务逻辑模型，以及应用场景的参数即可以完成业务分析功能的建模和定义。逻辑模型支持根据数据源的变动或业务变动，进行灵活的修改和实时生效。

以目前天冕数据中台实际落地的项目来看，在充分理解业务数据的基础上，通常在一天就能完成系统部署应用和上线使用。

核心指标

- 全域的数据可视化收集和实时处理
- 快速建立统一的数据中心，提供数据实时的交互共享能力
- 在线完成功能的建模和定义

3.5 智能算法组件集成

天冕数据中台机器学习引擎为关系型数据的挖掘提供了多种针对业务优化的数据模型和算法模型，平台结合经典算法，融入机器学习实现智能分析，如用户画像、特征工程、关系网络、风

险评估等，帮助用户快速实现对海量数据的复杂挖掘。

核心指标

- 开箱即用的机器学习引擎
- 支持成熟的用户画像、特征工程、关系网络、风险评估模型

3.6 灵活参数配置化

天冕数据中台支持包括数据源、业务模型，图表样式、用户角色权限、数据流、系统参数等参数的配置化管理。用户可以根据项目的实际情况在管理控制台灵活自定义配置，满足具体项目的实际业务需求。

核心指标

- 基于参数化的数据采集、存储、集成、挖掘、分析
- 灵活的应对数据需求变化

3.7 开放性和扩展性

天冕数据中台基于开源的 Hadoop 构建（Kafka、Kudu、HBase、Impala、Spark、Zookeeper、Hive），支持主流的大数据生态，每一个服务组件（DataBus 数据总线、Merger 流式作业平台、Meteor 计算服务平台、DataWorks 数据工坊、DataFace 数据可视化、Matsu 数据监控、Poseidon 数据安全、Atlas 机器学习）都可以根据需求自定义，支持标准 SQL(SQL92)、协议、技术框架，无单点故障，支持多数据中心（多机房）、故障自动恢复。

核心指标

- 平台基于开源的 Hadoop 技术生态
- 组件服务化，支持灵活的扩展
- 支持标准的 SQL（SQL92）、协议、技术框架

3.8 便捷性

天冕数据中台将大数据处理中涉及的复杂的流程、数据治理、数据安全、数据分析、数据挖掘采用可视化的方式来管理，大幅度提高了大数据的应用效率。

- 可视化界面操作的方式处理数据实时采集、存储、集成、挖掘、分析，非常快捷的应对数据需求
- 所见即所得的方式完成数据治理和标准化工作，大幅度提升数据质量和管理效率
- 提供成熟的用户画像、特征工程、关系网络、风险评估四类算法模型开箱即用，极大提高模型演练的效率

4.大数据应用

4.1 实时数据共享

天冕数据中台中通过部署数据总线的插件和定义数据采集流程将各大业务系统的数据实时采集到大数据平台中，实现数据的中心化管理，非常方便的完成了各大业务系统数据共享和交换；

- 数据共享：全域的数据实时整合到数据中心后，基于数据的中心化管理可实现全域数据的实时共享
- 数据交换：基于消息订阅的方式订阅数据流中的消息，实时监听数据的变化（新增、修改、删除），实现数据的实时变更通知

4.2 数据监控

天冕数据中台中的数据流转是基于消息流完成的，通过订阅消息流任意一个阶段的数据；数据监控模块可以自定义数据的监控指标和告警策略，完成对数据质量和数据指标的监控。

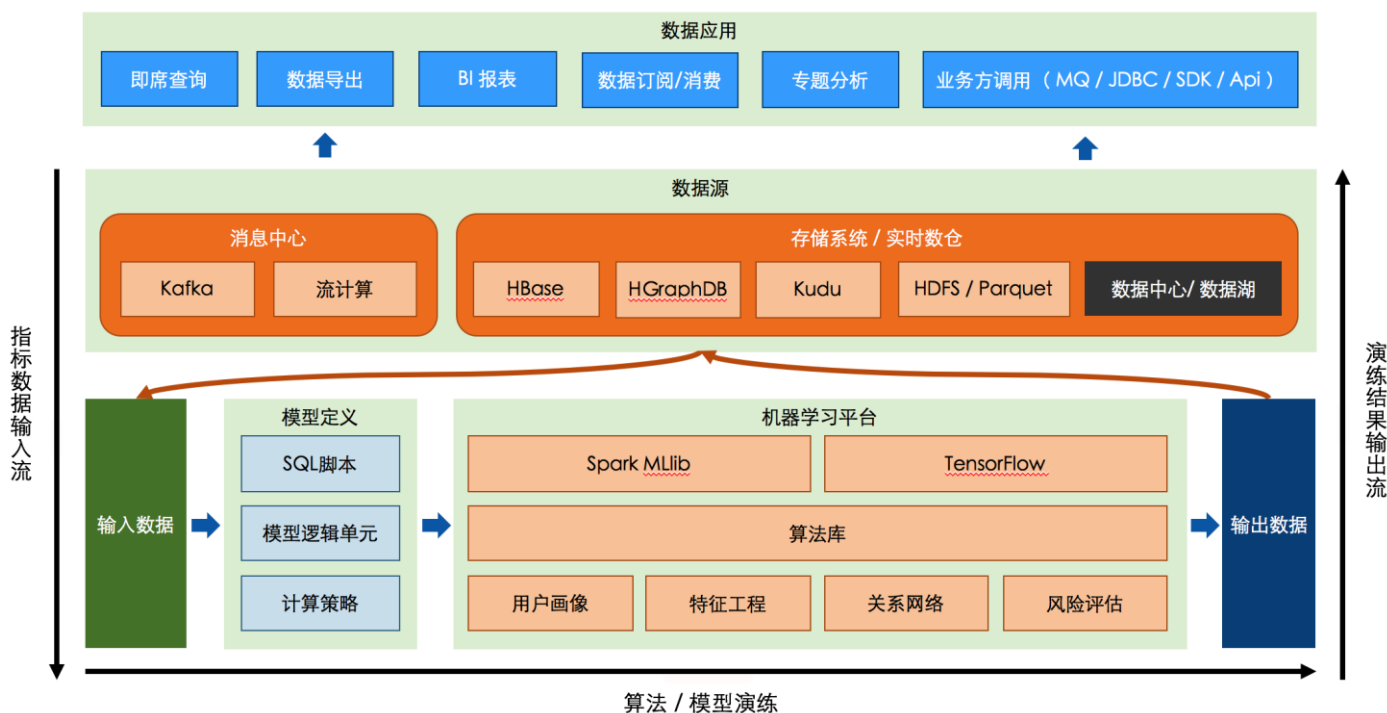
4.3 数据可视化

天冕数据中台中的数据可视化帮助非专业的工程师通过图形化的界面轻松搭建专业水准的可视化应用，满足会议展览、业务监控、风险预警、地理信息分析、BI 报表等多种业务的展示需求。

4.4 模型应用

天冕数据中台内置的机器学习模块整合了 Spark MLlib 和 TensorFlow 的五十多种经典算法，输出了四类成熟度较高的模型包括用户画像、特征工程、关系网络、风险评估，可以开箱即用的方式应用到大数据的挖掘中。

模型的应用需要配置数据源和计算数据逻辑，通过实时增量和全量的模型演练，将演练的模型结果存入数据中心，业务方通过 SDK、SQL 或者消息订阅的方式调用模型结果。



- 模型演练的方式：

1. 指定数据源中的数据并配置演练模型计算 SQL 的流程，生成模型演练结果并入库
2. 指定数据源通过封装的模型 SDK 选择算法库，生成模型演练结果并入库

3. 基于数据中心的数据自定义机器学习的演练模型，生成模型演练结果并入库
- 模型结果的使用：
 1. 业务方通过调用 SDK 访问数据中心存储的模型演练结果
 2. 业务方通过 SQL 的 JDBC 方式调用数据中心存储的模型演练结果数据
 3. 业务方通过订阅模型演练过程中的消息，通过消息总线可以实时通知业务方每一步骤的演练结果

5. 大数据平台优劣分析

天冕数据中台与云厂商在系统生态、功能实现、实施风险、扩展性、合规性等方面侧重点各有不同，特做如下对比：

指标	大数据平台优劣势对比	
	天翼大数据平台	云厂商大数据平台
系统生态	优势： 1. 基于开源的 CDH Hadoop 生态支持标准的技术框架、SQL、协议，在技术开放性、扩展性、掌控力上更有优势，人才储备有优势； 2. 天翼大数据平台各组件已整合完善，无需大量的调整即可投入使用； 劣势： 1. 在超大规模数据集、万级别的服务器节点的应用规模上云厂商更有优势；	优势： 1. 在超大规模的数据集上有成熟的使用经验； 劣势： 1. 大数据基础服务建立在私有封闭的平台上，开放性和生态不完善；外部企业使用上人力资源储备匮乏及功能居多限制； 2. 云厂商平台提供都是比较技术化的服务，需要业务方根据需求改造或重新整合后方可投入使用；
功能实现	优势： 1. 无侵入式的强大的各种数据快速整合能力； 2. 数据的采集、存储、集成、挖掘、分析全部实时化； 3. 完善的实时全域数据监控的能力； 4. 强大的在线数据清洗/加工/扩展的能力； 5. 完善的全域数据的治理、血缘关系、元数据管理、数据资产管理的能力； 6. 全流程的安全认证、授权、访问以及数据多维度的事前监控、事中拦截、事后审计； 7. 全域数据的实时数仓和数据中心； 8. 全域数据的实时订阅与分发；	劣势： 1. 数据整合能力弱，全域数据支持的复杂度高； 2. 数据处理不实时或离线计算的居多，大多采用传统的 ETL 模式； 3. 缺乏或不支持实时数据监控； 4. 缺乏在线数据清洗/加工/扩展的能力； 5. 缺乏全域数据的治理、血缘关系、元数据管理、数据资产管理的能力或支持的复杂度高； 6. 安全管理未覆盖整个大数据体系的全生命周期； 7. 缺乏或不支持实时数据数仓； 8. 缺乏或不支持全域数据的实时订阅与分发；
实施风险	优势： 1. 无侵入现有系统和架构，快速实现数据智能流转；	劣势： 1. 不同程度需要改变现有系统、业务、部署方式等，存在落地周期长和对现有架构和业务的入侵性；

	2. 架构简单，依赖的技术组件少，技术体系更容易掌控； 3. 数据治理能快速落地； 4. 数据收集、计算、挖掘、分析可见即所得；	2. 依赖的技术组件多，整体架构复杂，不易掌控；
扩展性	优势： 1. 可支持灵活的需求定制；	劣势： 1. 基本不支持需求定制或支持周期不可控；
合规性	优势： 1. 系统技术和软件完全遵守开源授权协议，不存在法律合规性风险； 2. 数据安全和备份更灵活，更方便支持合规性对数据的安全、容灾的要求；	劣势： 1. 云厂商在技术平台上未完全遵守开源授权协议，存在法律合规性争议（如 MongoDB (SSPL)、Neo4j、Redis (Commons Clause) 等； 2. 部分平台自建的封闭的备份机制和备份策略，缺乏灵活性

6. 系统运行环境

6.1 硬件环境

阿里云、腾讯云 ECS 服务器 或 X86 通用型服务器

推荐配置：

- CPU: Intel(R) Xeon(R) CPU
- 内存: 32G 及以上
- 硬盘: 1T 及以上
- 网络: 20M/S 及以上
- 服务器至少为 3 台，具体服务器数量根据数据量的规模和并发用户数而定

6.2 系统及软件环境

操作系统：CentOS 6.4 及以上

运行时依赖：

- JAVA 环境 JDK 1.8 及以上版本
- NODE 环境 Node.js 4.4.2 及以上版本
- CDH 5.0 及以上版

7.性能与可靠性

天冕数据中台稳定和性能经过包括内外部的多个重要级项目的考验和验证；同时作为一个大数据处理平台，仍具有较高的用户并发访问支持。

7.1 稳定性

由于业务工作是全天候行为，所以平台的稳定运行是能够产生持续效益的有力保障。平台支持 7×24 无故障运行，对一般错误用容错机制保证其它操作不受影响，出现较大问题时，可快速重启恢复。支持 7×24 小时调用访问，有效工作时间 $\geq 99.9\%$ 。

7.2 响应时间

天冕数据中台将被广泛使用在日常业务的过程中，因此要求平台响应时间要在最终用户可以接受的范围之内。以实际案例为例，实时的数据分析响应时间在毫秒级到秒级，机器学习的算法计算响应时间在秒级到分钟级，实际响应时间需要根据数据量和集群数而定。

7.3 并发访问

以实际案例为例，天冕数据中台使用者数量较多。因此，设计时充分考虑并发访问量，尤其是高峰期的并发访问量，以保障业务的顺利运行。并发用户数峰值预计在 5000 人左右，平台支持在线人数 ≥ 100000 人，实际并发数根据服务器集群的规模而定。