**大数据管理平台  
建设方案**

**XXX公司**

目录

[1.大数据管理平台(数据中台) 1](#_Toc9171)

[1.1.系统概述 1](#_Toc7926)

[1.1.1需求描述 1](#_Toc25863)

[1.1.2.需求分析 2](#_Toc3544)

[1.1.3.重难点分析 3](#_Toc32267)

[1.1.4.重难点解决措施 3](#_Toc6766)

[1.2.系统架构设计 4](#_Toc14207)

[1.2.1.系统架构图 4](#_Toc7694)

[1.2.2.关键技术 7](#_Toc13263)

#### 1.大数据管理平台(数据中台)

##### 1.1.系统概述

###### 1.1.1需求描述

建设大数据管理中台，按照统一的数据规范和标准体系，构建统一数据采集-治理-共享标准、统一技术开发体系、统一接口API，实现数据采集、平台治理，业务应用三层解耦，并按照统一标准格式提供高效的数据支撑服务，保证上层业务应用系统能够更加专注于业务流与数据流的梳理和重构，从而释放数据压力，节约系统开发周期，提升系统建设效率，满足企业、政府数据精细治理、分类组织、精准服务、安全可控的数据资源体系管理目标。

###### 1.1.2.需求分析

在业务系统较复杂的情况下，大数据管理平台中的主流的解决方案为技术中台解决方案，主要目的为打破信息孤岛。

在数据中台建设中，我们主要有以下目标：

1. 去重,避免重复建设，规避重复造轮子/烟囱式架构节约公司成本；
2. 复用,系统快速复用，标准化、组件化、松耦合，支持前台业务快速迭代、灵活创新；
3. 做强,能力不断沉淀，服务可被不断滋养，系统可统一把控；

建设数据中体采用如下方法：

1. 明确中台建设目标及领域；
2. 梳理价值链并萃取功能点；
3. 设计中台并梳理前中后关系;

在决策支持方案种，我们采用的数据存储于数据仓库中，有质量管控，决策支持，采用的思路为：有什么样的数据，我们分析什么样的数据；在数据中台方案中我们采用数据驱动的方式，以数据价值为驱动，采集更多的数据，是干什么样的事，采集什么样的数据。数据运营优化中，采用规范数据，规范数据应用，优化中台，形成数据应用闭环（数据分析闭环）。

解决数据孤岛，让数据发挥价值，回馈企业。采用以数据为驱动，按照需要什么样的数据，采集什么样的方式。数据的来源以多系统的真实数据和设备采集的实时数据为基础。

在大数据系统（数据中台）中，我们需要对各个子系统静态数据与采集实时流数据进行标准化。在数据治理的过程中，我们采用对不同数据源的接入，提供对元数据的整体管理能力，提供目录和标签的管理，同时针对数据本身提供权限访问，数据探查和质量报告等提高数据治理的能力和水平。

###### 1.1.3.重难点分析

从布点、收集、存储、刷新、识辨、关联、挖掘、决策、行动、再到反馈，这样一个闭环，才能让数据驱动业务。在数据真实性的前提下，数据驱动业务，才能以数据为驱动，为企业提供经营决策，创造效益。构建大数据分析系统（数据中台），存在一下问题：

1. 数据采集的正确性确定难，从一端流向另一端，不可避免会产生转化率的问题，如何将非真实的数据与异常的数据过滤掉，收集真实的数据，实现完整的数据闭环，才能让数据挖掘、数据分析变成商业价值。数据真实性的收集，异常数据的过滤，在大数据平台建设过程中，是建设的难点之一。
2. 技术选型困难，不同业务方有不同的数据需求，技术选型时依据这些客观需求及主观偏好，会选择不同的计算框架和数据组件，搭建一个合适特定行业的业务架构与技术架构，让企业转型变简单，也成为建立大数据分析系统的难点；
3. 数据需求多样，业务部门需求多样化，包括报表计算、可视化看板、数据探索、数据服务、结果推送、数据采集及迁移、标签体系、用户触达、数据应用；
4. 数据管理复杂，数据的可解释性、可管理性要求越来越高，各种新存储架构的加入，使得元数据管理和数据流程标准化更加复杂；
5. 数据权限管理，数据赋能的体系中权限控制是很关键的功能，需要实现各种级别的数据权限，组织架构、角色、权限策略自动化，以及对新的计算架构的权限管理。
6. 数据成本高量化难，数据成本包括集群成本、运维成本、人力成本、时间成本等，持续系统地计算这些成本需要在系统架构中加入相应的统计接口，而现有的大多数中台并没有将这些接口考虑在内。
7. 数据标准建立和协调困难，数据标准化建设主要包括，数据建设规范中的数据安全规范、数据存储规范、数据建模规范、数据接入规范；数据消费规范中的数据销毁规范、数据调用规范、数据权限规范；

###### 1.1.4.重难点解决措施

在大数据的难点分析中，主要有技术性的困难与组织性的针对以上问题，我们提出了一下措施：

1. 对齐业务价值，不在业务场景还没有明确、优先级还不清晰、价值度量体系尚未建立起来的时候，就建立大而全的数据平台，并且把所有的数据都存起来，多考虑投入产出比。大而全的数据平台往往会面临尴尬的局面，一堆功能看上去很有用，应该都能用上，但是缺乏应用场景，真的有了场景，发现也不能开箱即用，还需要众多的定制化。
2. 从小数据、小场景做起，数据中台是面向场景而非面向技术的，这种与客户的业务、企业的结构和信息化发展阶段有着紧密的相关性的业务基础架构，是很难买一个大而全的产品来一劳永逸解决的。利用低代码平台内置的组件及场景，构建3D可视化界面，实现高效、高质量软件开发，同时结合数据中台的开发工具，为低代码平台所搭建的场景应用提供数据。
3. 在技术选型上，建立在Hadoop集群上，数据集成、数据准备、数据挖掘（全量运算）、数据分析、数据可视化（数据输出），形成一个循环的过程，连接数据就可以实时得到分析结构。

以数据中台结合低代码平台的思路，可让所有人（特别是管理者和业务人员）都有能力开发在线的协作和管理工具，从小数据、小场景做起，使组织的效率和协作能力获得根本性提升。当低代码平台兼具低代码数据平台和低代码开发平台能力时，可以使企业数字化管理得到高效进展。

##### 1.2.系统架构设计

###### 1.2.1.系统架构图

1. 数据中台功能架构

数据中台整体功能架构如图5.3.3-1所示



图5.3.3-1 数据中台功能架构

数据工具平台层，是数据中台的载体，包含了大数据处理的能力技术，如数据的采集，存储，计算、数据安全等一体的大数据平台；还包含建设数据中的一系列研发工具，如离线或实时数据研发工具、数据联通工具、标签设计工具、算法平台工具、数据服务工具及自助分析工具。



图5.3.3-2 工具平台层架构（数据中台）

数据资产层，数据资产层是数据中的核心层，可划分为主题域模型区、标签模型区、算法模型区。主题域模型是面向业务分析的，将业务过程或维度进行抽象的集合，业务过程可以概括为一个个不可以分隔的行为事实；标签模型，一般会涉及企业经营过程中的实体对象，这些实体对象一般来说都会穿插到各个业务流中；算法模型更加贴近业务场景，在设计算法模型的时候，要反复推演算法模型的使用场景，包括模型的冷启动等问题。



图5.3.3-3 数据资产层架构（数据中台）

数据应用层，数据中台的使命就是为业务服务的，几乎所有的企业在建设数据中台的同时，都已经规划好了数据应用。数据应用可按照使用场景来进行划分，分为分析与决策应用，标签应用，智能应用。

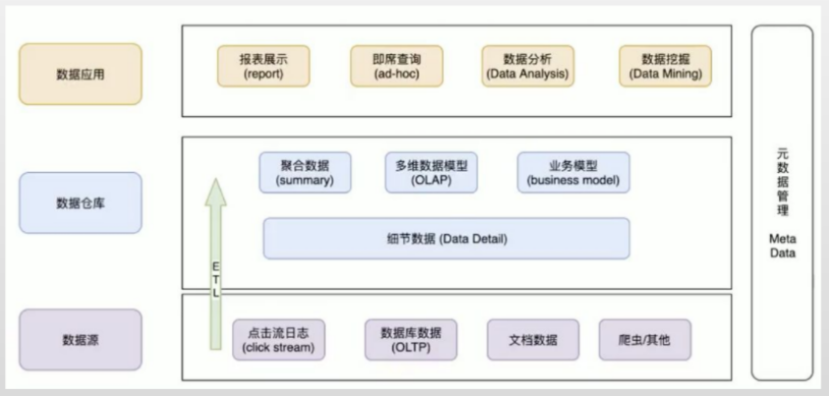


图5.3.3-4 数据应用层架构（数据中台）

1. 数据中台技术架构如图5.3.3-5所示：

****

图5.3.3-5 数据中台技术架构

* 数据来源层，提供数据仓库的数据源，该技术架构中支持结构化数据、非机构化数据、文件日志；
* 数据传输层，结构化数据同步工具采用Sqoop，非结构化数据同步工具采用DataX，日志文件采用flume工具进行数据采集；
* 数据存储层，采用hdfs(分布式文件系统)作为数据存储底层，在该系统上提供HBase数据库，用于存储时序型数据为业务层提供数据服务；该存储层支持对kafaka中的数据进行二次加工并存储；
* 资源管理层，主要提供对硬件资源节点的管理，节点中资源中CPU，内存的管理；
* 数据计算层，在该层中提供离线计算与实时计算两种计算方式，并对计算结果进行分层建模处理。才该技术方案中采用DWD（数据明细层），DWS(数据服务层/数据主题层)，ADS（数据应用层）；
* 业务层，主要用于展示数据计算结果。

###### 1.2.2.关键技术

1. 数据仓库分层技术



图5.3.3-6 数据仓库分层结构

* ODS 层：原始数据层，存放原始数据，直接加载原始日志、数据，数据保持原貌不做处理；
* DWD 层：对ODS层数据进行清洗（去除空值，脏数据，超过极限范围的数据）、维度退化脱敏等；
* DWS 层：以DWD为基础，按天进行轻度汇总；
* DWT 层：以DWS为基础，按主题进行汇总；

1. 数据标准化设计技术
2. ODS原始数据层

Ods层命名空间namespace=ods；

Ods层表命名方式tablename=源系统名+表名；

该层数据首先从kafka接入，然后通过一定的解析方式，最后逐条存入相应的表中。

1. DWD明细数据层

Dwd层命名空间namespace=dwd；

Dwd层表命名方式tablename=源系统名+表名；

Dwd层与Ods层是一对一的关系，对Ods层数据进行清洗（去除空值，脏数据，超过极限范围的数据）、维度退化脱敏等；

1. DWS服务数据层

Dws层命名空间namespace=dws；

Dws层表命名方式tablename=自定义表名；

Dws层与Dwd层是一对多关系，以Dwd层为基础，按天进行轻度汇总；

1. DWT数据主题层

Dwt层命名空间namespace=dwt；

Dwt层表命名方式tablename=自定义表名；

Dwt层与Dws层是一对多关系，以Dws层为基础，按主题进行汇总。

数据仓库分层技术；



图5.3.3-7 系统架构图

* 清晰数据结构：每一个数据分层都有它的作用域和职责，在使用表的时候能更方便地定位和理解。
* 减少重复开发：规范数据分层，开发一些通用的中间层数据，能够减少极大的重复计算。
* 统一数据口径：通过数据分层，提供统一的数据出口，统一对外输出的数据口径。
* 复杂问题简单化：将复杂的任务分解成多个步骤来完成，每一层只处理单一的步骤，比较简单和容易理解。当数据出现问题之后，不用修复所有的数据，只需要从有问题的步骤开始修复。

屏蔽原始数据的异常：不必改一次业务就需要重新接入数据。

1. 标签类目体系设计技术

在数据中台中，标签的所处的位置如图5.3.3-2所示,



图5.3.3-8 标签位置

属性标签，就是对业务实体各种属性的真实刻画。

统计标签，就是对业务实体从某个维度的度量进行的汇总。

算法标签，就是通过某些算法推理得到的特性。算法标签相对比较复杂，但非常有用。它既可以设计得简单易行，如企业的行业地位、交易成功率、客户开拓能力、客户忠诚度、企业成长度等。

在标签体系目录的设计设计方案种，遵循以下原则：

1. 标签规范，考虑原始数据是否可以用于加工成标签；考虑标签是否能体现业务价值或者能创新业务场景；
2. 安全等级，构建1～4等级的安全定级（L1～L4），第一等级的标签（L1）是公开标签，可对外公开，是最为开放的数据标签，安全等级最低；第二等级的标签（L2）是内部标签，是在企业/机构内部跨部门可直接流通、申请、使用的数据标签，安全等级较低；第三等级的标签（L3）是保密标签，企业内部跨部门使用需要申请授权，批准后才能使用的标签，安全等级较高；第四等级的标签（L4）是机密标签，是企业/机构内部少数人才可以使用的标签，且不可传播，安全等级最高。
3. 标签需要与底层物理表映射，才能在生成数据服务时进行真实数据流动。