数据交换平台概要设计

（v1.0）

大数据组 . 张峻 . 2020-06-01

目 录

平台建设目标-1

- 东师理想大数据中心 数据交换平台（以下简称“数据交换平台”），是一套面向云计算、大数据、人工智能与物联网等典型应用场景，结合教育行业规范与数据标准，支持基础数据规范管理、业务数据互联互通、用户行为与日志数据实时汇集的后台支撑系统。

- 数据交换平台支持数据的纵向（多个中心、分布式部署）和横向（集中式部署）的交换与汇集；
- 支持平台基础数据、系统业务数据、系统日志数据和用户行为数据的交换与汇集；
- 支持结构化数据、半/非结构化数据和一般文件的交换与汇集，并支持使用ETL方式集成遗留系统数据（遗留系统需要开放数据库）。

- 数据交换平台能够与东师理想基础数据管理系统、大数据分析与呈现系统实现无缝对接，高效的实现端到端大数据应用，为用户实施大数据项目节约成本、缩短周期、降低风险。

平台建设目标-2

- 应用数据交换平台，能够有效帮助教育局与学校用户解决以下问题：

- 一、数据不规范、不一致：完全参照《JY/T 1001 - 1007 2012》系列教育行业数据标准，实现数据校验和数据过滤，保证交换与汇集数据的规范性、一致性和有效性，提高数据质量。
- 二、系统信息孤岛难题：提供跨平台、多语言支持的通用数据交换接口，高效、快速整合各个接入系统，实现数据交换与数据汇集，打通信息孤岛，实现数据的互联互通、有效治理。
- 三、“大数据”汇集难题：提供半结构化/非结构化数据、用户行为数据与系统日志数据的汇集接口，实现对大数据的高并发、高性能汇集、处理和存储，为大数据分析和AI智能决策提供数据支撑。
- 四、遗留系统整合难题：提供成熟的ETL整合框架，支持全部主流数据库的点对点数据集成（遗留系统需要开放数据库），保证用户数据资产的充分利用、全面治理。

平台建设原则-1

- 规范性：

一、支持元数据数据项规范检查，参照《JY/T 1004-2012 普通中小学校管理信息》、《JY/T 1005-2012 中职学校管理信息》数据标准，实现对结构化/半结构化数据的数据项检测；

二、支持元数据数据字典规范检查，参照《JY/T 1001–2012 教育管理基础代码》数据标准，实现对结构化/半结构化数据的数据字典检测；

三、支持数据类型检查，包括数据必填检测和常规数据类型、数据格式检测；

四、支持数据所属机构检查，根据数据交换组织机构目录，对交换和汇集的数据进行所属机构检测；

五、支持按需自定义数据规范检查，基于JSON Mapping进行数据格式检测。

平台建设原则-2

- 可伸缩：

支持可伸、可缩的技术架构，支持逐步扩展的部署架构；

一、单机部署，对于简单应用场景和部署条件，支持收缩为使用MySQL单一数据持久化模式；

二、标准部署，支持MySQL、ElasticSearch、Kafka多种数据持久化模式，可以按照用户实际需求和业务数据特征灵活调整模式（结构化数据一般使用MySQL，半/非结构化数据使用ElasticSearch，行为数据、日志数据等一般使用Kafka）；

三、集群部署，可以按照用户实际业务量灵活扩展、提升数据交换与汇集处理容量，全部模块支持横向扩展（Web服务和应用服务模块使用负载均衡，数据持久化模块均原生支持集群部署）。

平台建设原则-3

- 可复用：

支持与其他厂商的基础数据管理系统集成，不绑定东师理想基础数据管理系统。支持与大数据中心其他模块的无缝集成，包括数据仓库、统计分析、AI智能分析、大数据呈现等。

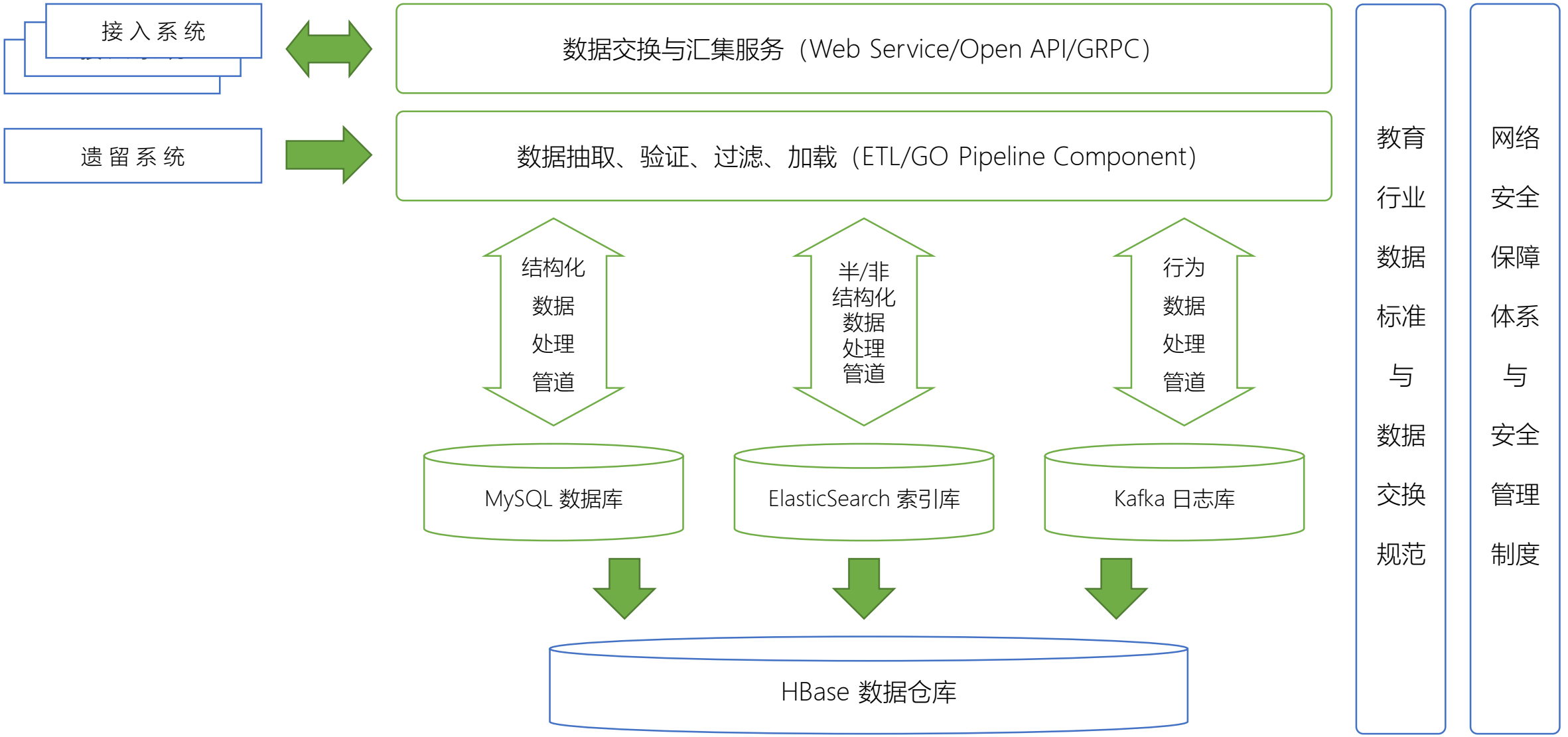
- 高可用：

支持负载均衡、集群部署，可以有效消除各个模块的单点故障，提高服务质量，缩短故障恢复时间，实现数据交换平台的高可用性。

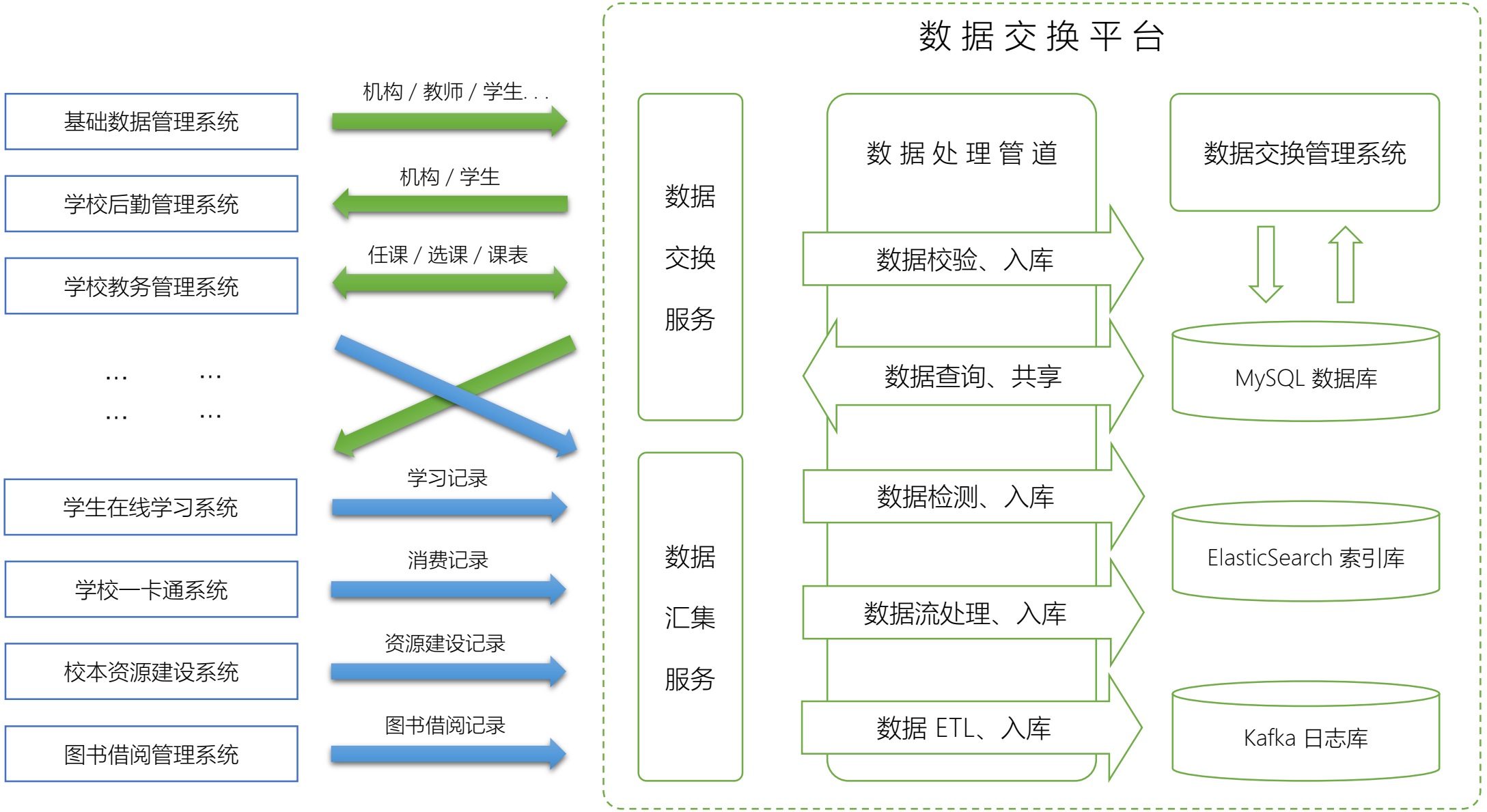
- 高性能：

采用Go语言搭建Web服务和应用服务平台，性能远高于一般的技术平台（Java、.Net、PHP等）；采用ElasticSearch和Kafka数据持久化方案，提供高性能数据处理和数据查询。

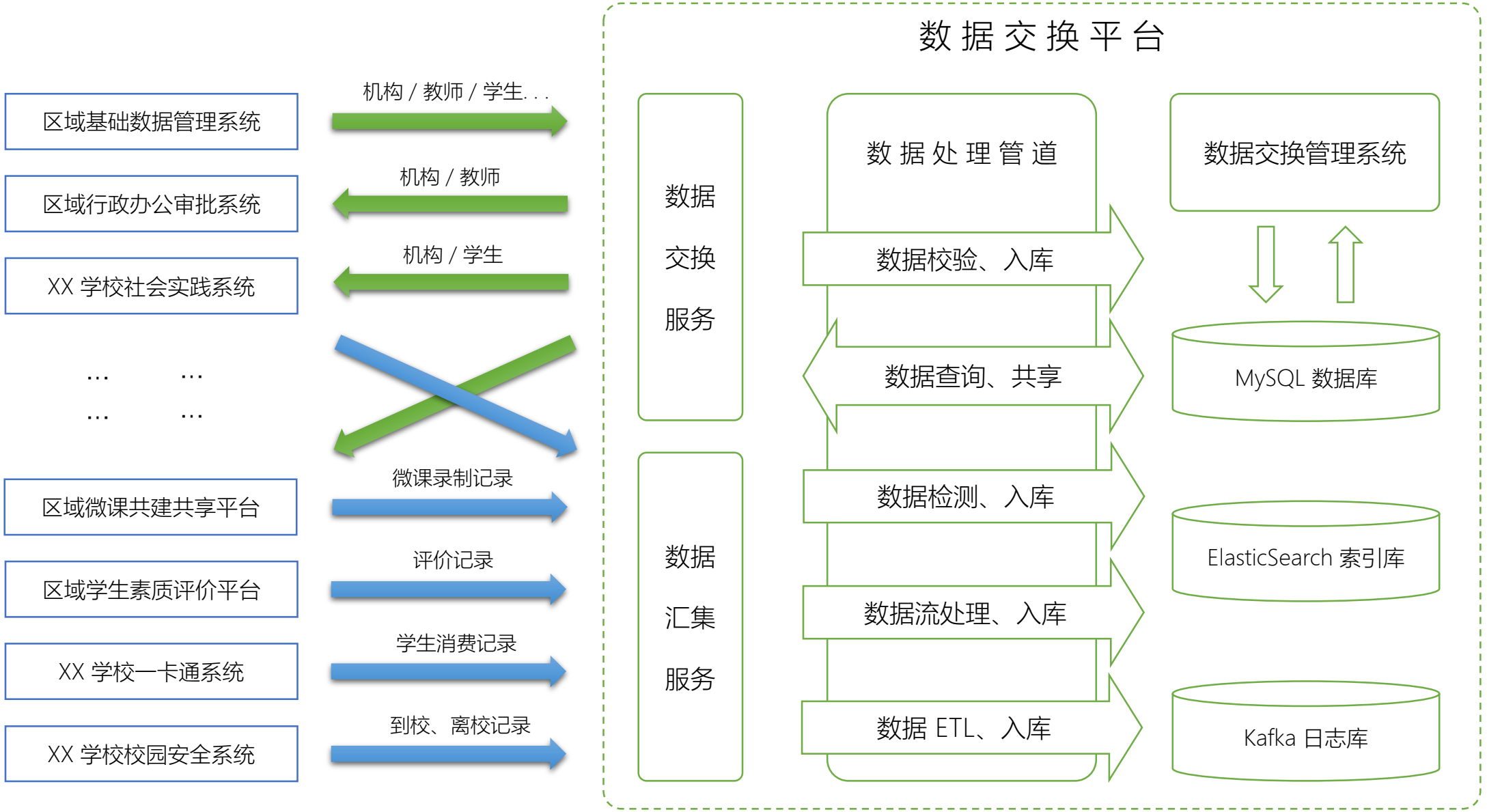
平台核心流程


接入系统
遗留系统
数据交换与汇集服务（Web Service/Open API/GRPC）
数据抽取、验证、过滤、加载（ETL/GO Pipeline Component）
结构化数据处理管道
半/非结构化数据处理管道
行为数据处理管道
MySQL 数据库
ElasticSearch 索引库
Kafka 日志库
HBase 数据仓库
教育行业数据标准与数据交换规范
网络安全保障体系与安全管理制度

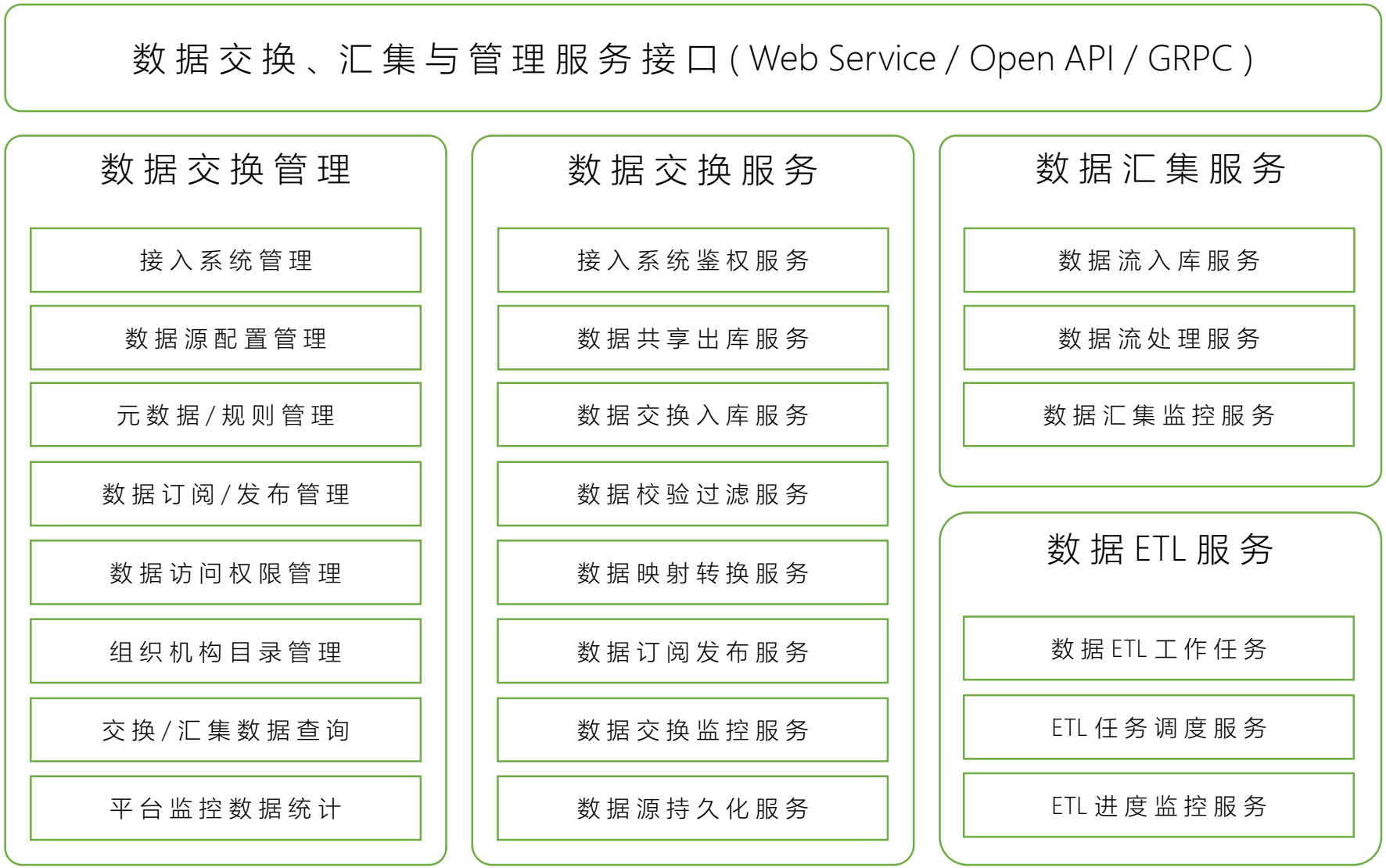
横向交换流程示例


数 据 交 换 平 台
基础数据管理系统
学校后勤管理系统
学校教务管理系统
… …
… …
学生在线学习系统
学校一卡通系统
校本资源建设系统
图书借阅管理系统
机构 / 教师 / 学生. . .
机构 / 学生
任课 / 选课 / 课表
学习记录
消费记录
资源建设记录
图书借阅记录
数据交换服务
数据汇集服务
数 据 处 理 管 道
数据校验、入库
数据查询、共享
数据检测、入库
数据流处理、入库
数据 ETL、入库
数据交换管理系统
MySQL 数据库
ElasticSearch 索引库
Kafka 日志库

纵向交换流程示例


数 据 交 换 平 台
区域基础数据管理系统
区域行政办公审批系统
XX 学校社会实践系统
… …
… …
区域微课共建共享平台
区域学生素质评价平台
XX 学校一卡通系统
XX 学校校园安全系统
机构 / 教师 / 学生. . .
机构 / 教师
机构 / 学生
微课录制记录
评价记录
学生消费记录
到校、离校记录
数据交换服务
数据汇集服务
数 据 处 理 管 道
数据校验、入库
数据查询、共享
数据检测、入库
数据流处理、入库
数据 ETL、入库
数据交换管理系统
MySQL 数据库
ElasticSearch 索引库
Kafka 日志库

平台核心功能-1


数据交换、汇集与管理服务接口（Web Service / Open API / GRPC）
数据交换管理
接入系统管理
数据源配置管理
元数据/规则管理
数据订阅/发布管理
数据访问权限管理
组织机构目录管理
交换/汇集数据查询
平台监控数据统计
数据交换服务
接入系统鉴权服务
数据共享出库服务
数据交换入库服务
数据校验过滤服务
数据映射转换服务
数据订阅发布服务
数据交换监控服务
数据源持久化服务
数据汇集服务
数据流入库服务
数据流处理服务
数据汇集监控服务
数据ETL服务
数据ETL工作任务
ETL任务调度服务
ETL进度监控服务

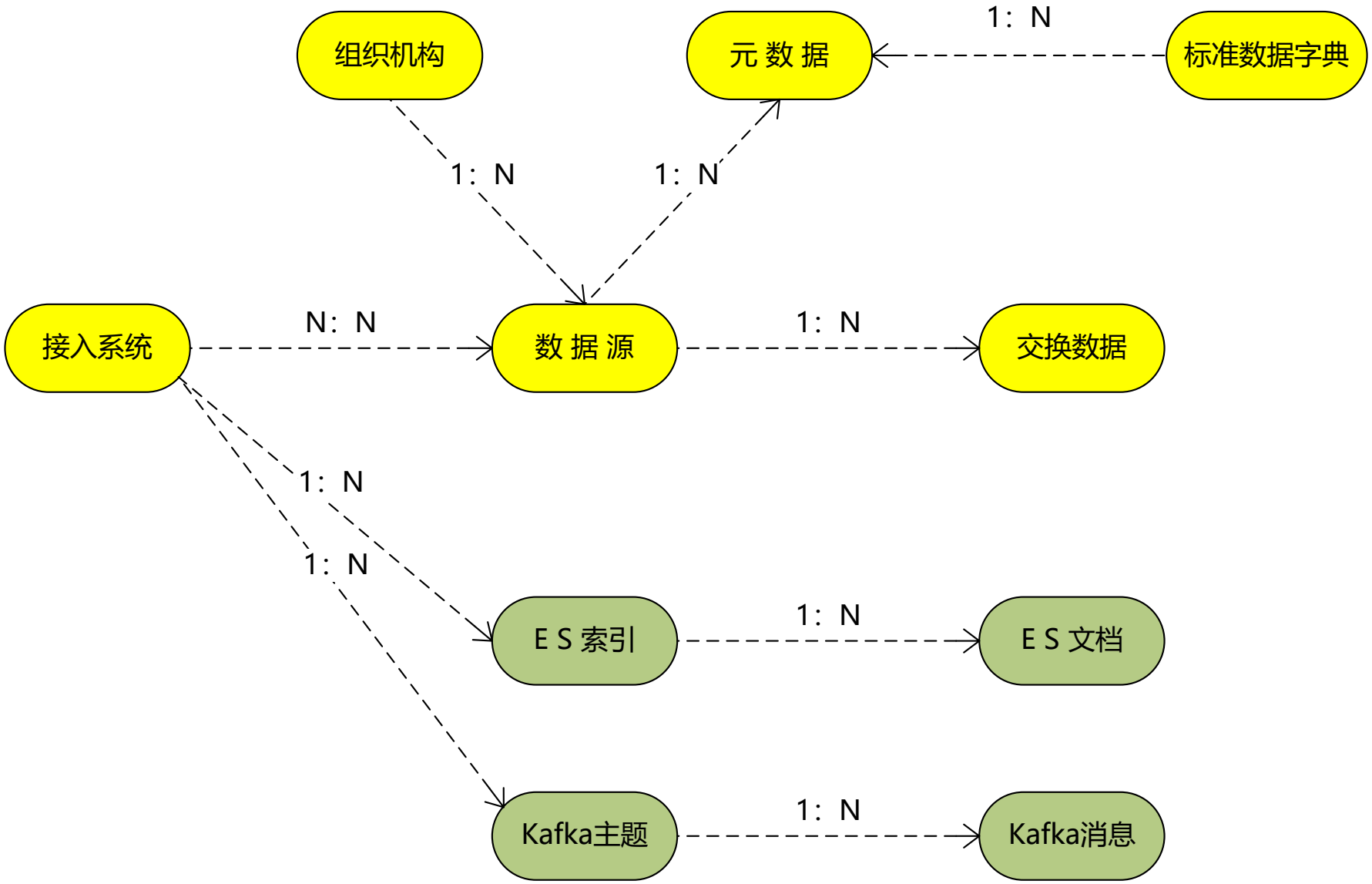
平台核心功能-2

- A、数据交换管理系统【平台管理员角色】
  - ✓ A1、接入系统管理：支持第三方接入系统的信息配置，【包括：系统名称、系统类型【基础数据、管理系统、资源系统、教学工具】、厂商名称、接入系统ID、接入系统Key，是否启用数据汇集，支持增删改查】；
  - ✓ A2、数据源管理：支持第三方接入系统提供的数据源信息配置，【包括：数据源名称、数据源编码【6位字符】、数据源描述、数据源提供系统【外键】、数据提供类型【可增加PUT、可修改POST、可删除DELETE】、数据提供级别【不限制、机构、机构及以下】、数据存储方式【MySQL数据库、ES索引库、Kafka日志库】，支持增删改查】，支持配置数据源消费系统【关系表、包括：数据消费类型【只读、可增加、可修改、可删除】、数据消费级别【不限制、机构、机构及以下】，支持增删改查】；
  - ✓ A3、元数据管理：支持数据源的元数据【即：数据的数据】信息配置【每个字段一条数据描述、可选填】，支持字段名称校验【参照国标、不区分大小写】、支持字段字典校验【参照国标】、支持字典必填校验、支持字段类型校验【整数、浮点、字符【带长度】、布尔、日期、时间、日期+时间】、支持内容格式校验【身份证、邮件、电话、手机】。

平台核心功能-3

- ✓ A4、订阅发布管理：支持接入系统对数据源数据变更的订阅【数据提供者产生数据变更，发布消息到数据交换平台，平台转发消息给数据订阅者】；
- ✓ A5、数据权限管理：控制数据访问范围、数据操作类型，合并到A2；
- ✓ A6、机构目录管理：支持数据交换相关系统所属机构的目录管理，【横向交换全部指定到根目录，纵向交换需要区分上下级机构【机构树】】；
- ✓ A7、数据查询：支持对数据交换平台各类数据源数据的查询【支持简单条件、排序、分页查询】；
- ✓ A8、数据统计：支持对交换数据、汇集数据的整体统计【交换：有进有出，汇集：只进不出】，支持按时间、按数据类型、按数据源、按接入系统的统计【只统计数据量】。

数据模型概述


组织机构
元 数 据
1：N
标准数据字典
1：N
1：N
接入系统
N：N
数 据 源
1：N
交换数据
1：N
1：N
E S 索引
1：N
E S 文档
Kafka主题
1：N
Kafka消息

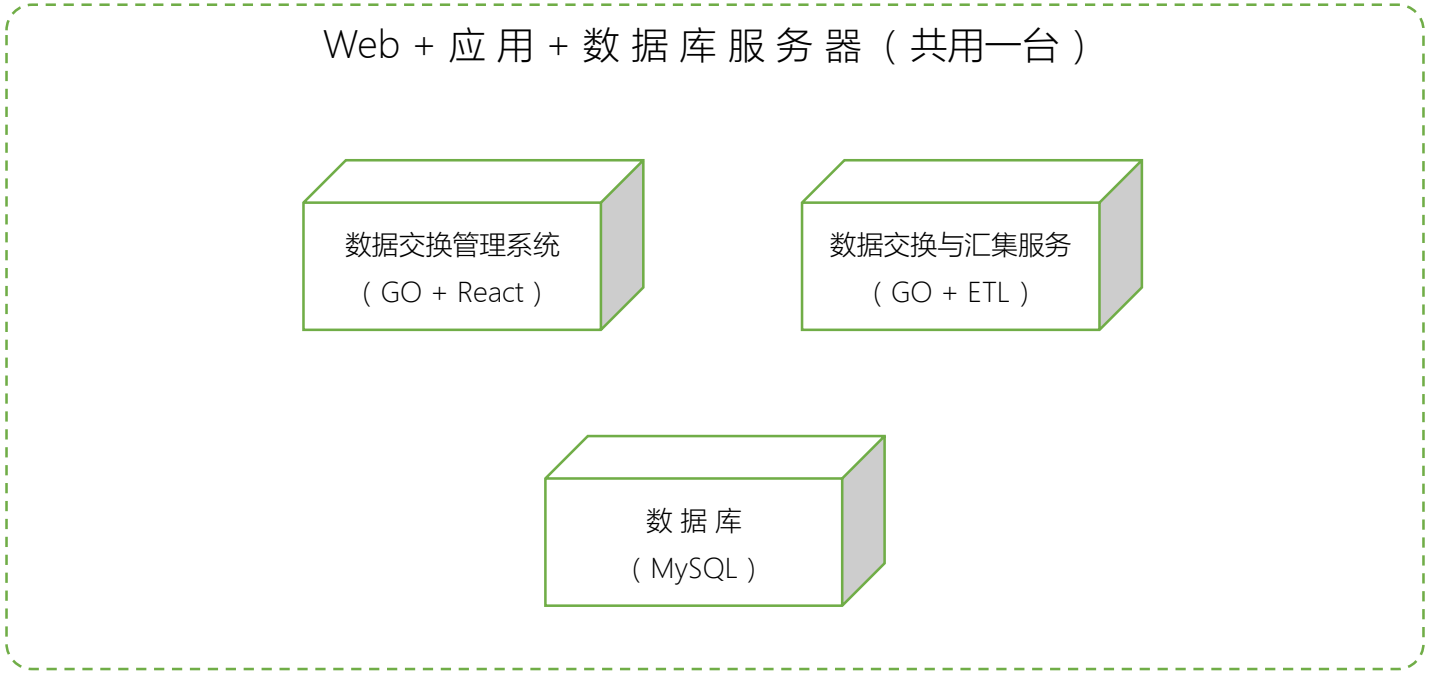
数据接口概述

【说明】暂不支持订阅、发布相关接口

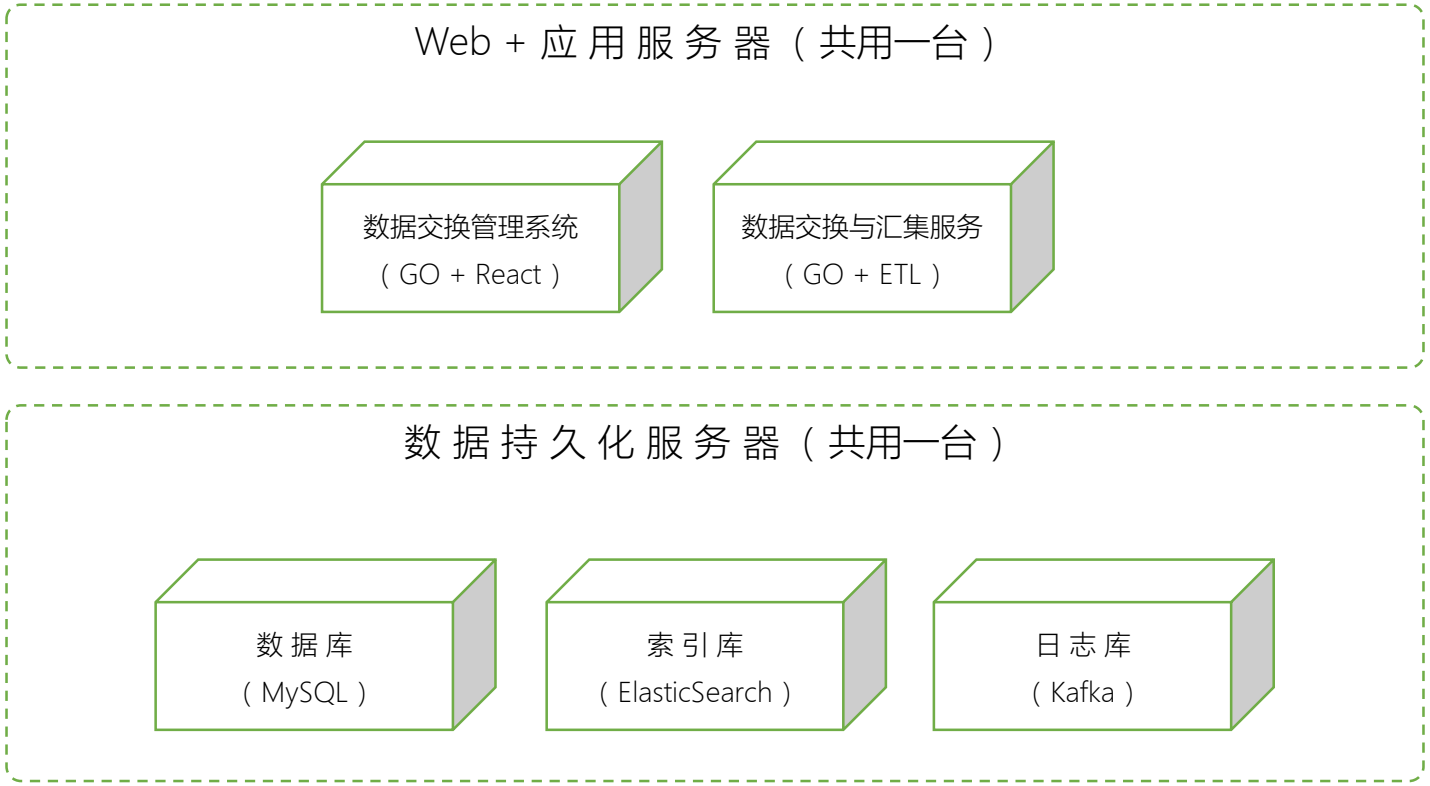
数据交换与汇集服务（Web Service/Open API/GRPC）

系统鉴权接口	数据增加接口	数据汇集接口	数据修改接口	数据删除接口	数据查询接口
SystemAuth	DataexAdd	DataexCollect	DataexUpdate	DataexDelete	DataexQuery
【参数】 system_id system_token auth_time 【返回】 auth_token	【参数】 system_id auth_token data_source org_id data_ids datas 【说明】 批处理数小于100 【返回】 succes_ids fail_results message	【参数】 system_id auth_token data_source org_id data_ids datas 【说明】 批处理数小于100 【返回】 message	【参数】 system_id auth_token data_source org_id data_ids datas 【说明】 批处理数小于100 【返回】 succes_ids fail_results message	【参数】 system_id auth_token data_source org_id data_ids 【说明】 批处理数小于100 【返回】 succes_ids fail_results message	【参数】 system_id auth_token data_source query_time query_page 【说明】 每页100条数据 【返回】 succes_ids datas message

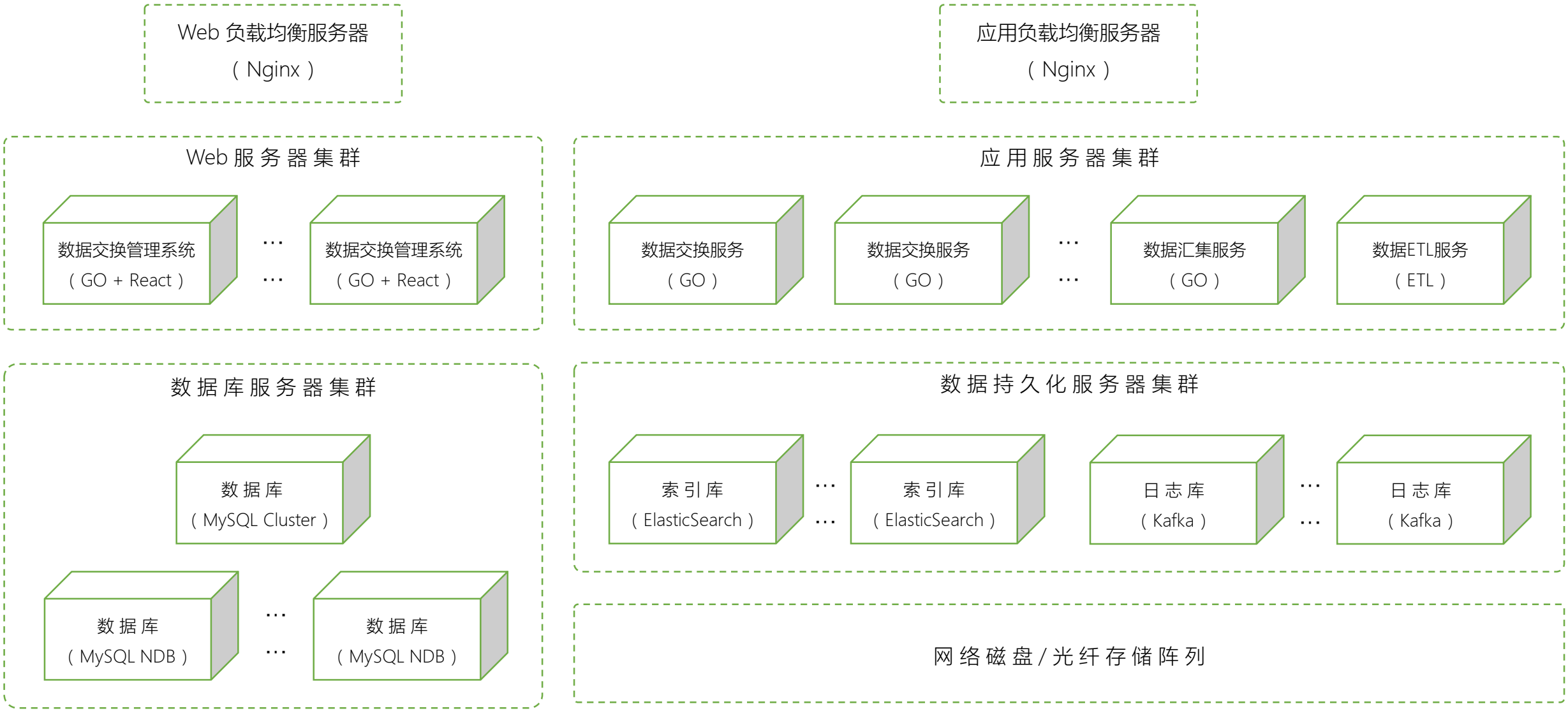
单机部署架构


Web + 应 用 + 数 据 库 服 务 器 （ 共用一台 ）
数据交换管理系统
（ GO + React ）
数据交换与汇集服务
（ GO + ETL ）
数 据 库
（ MySQL ）

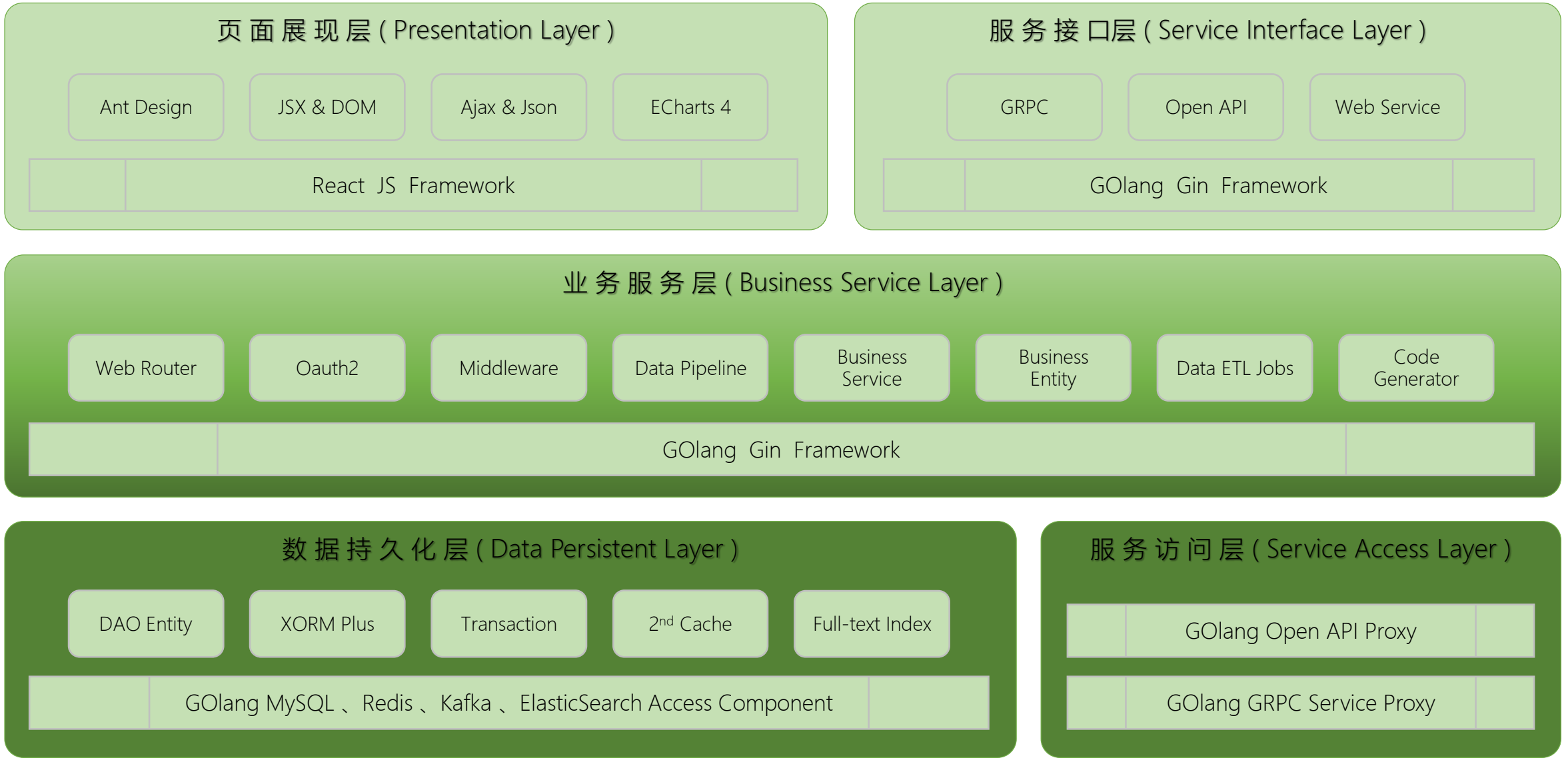
双机部署架构


Web + 应用服务器（共用一台）
数据交换管理系统
（GO + React）
数据交换与汇集服务
（GO + ETL）
数据持久化服务器（共用一台）
数据库
（MySQL）
索引库
（ElasticSearch）
日志库
（Kafka）

集群部署架构

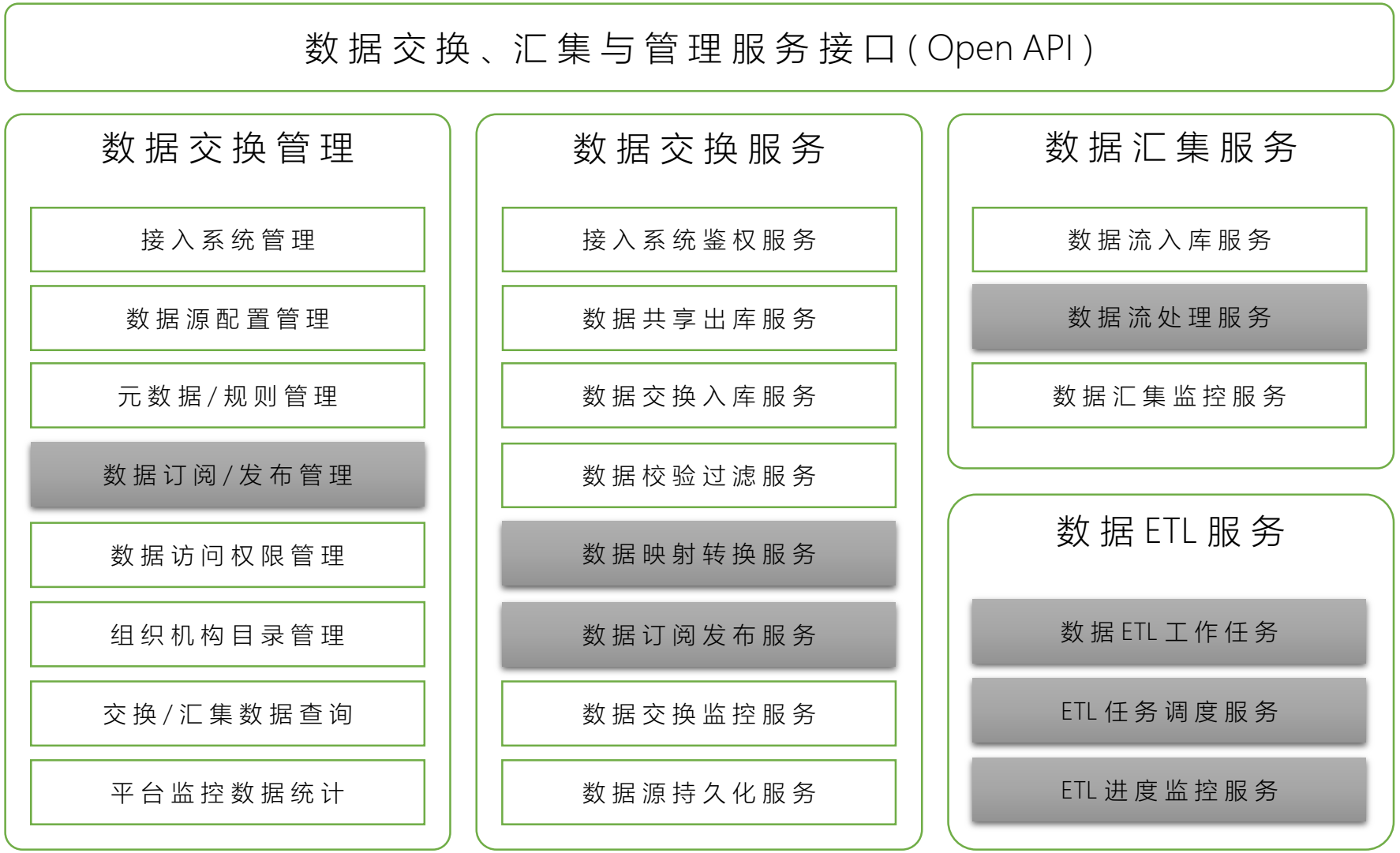

Web 负载均衡服务器
(Nginx)
应用负载均衡服务器
(Nginx)
Web 服 务 器 集 群
数据交换管理系统
(GO + React)
数据交换管理系统
(GO + React)
应 用 服 务 器 集 群
数据交换服务
(GO)
数据交换服务
(GO)
数据汇集服务
(GO)
数据ETL服务
(ETL)
数 据 库 服 务 器 集 群
数 据 库
(MySQL Cluster)
数 据 库
(MySQL NDB)
数 据 库
(MySQL NDB)
数 据 持 久 化 服 务 器 集 群
索 引 库
(ElasticSearch)
索 引 库
(ElasticSearch)
日 志 库
(Kafka)
日 志 库
(Kafka)
网 络 磁 盘 / 光 纤 存 储 阵 列

平台技术架构


页面展现层 (Presentation Layer)
Ant Design
JSX & DOM
Ajax & Json
ECharts 4
React JS Framework
服务接口层 (Service Interface Layer)
GRPC
Open API
Web Service
GOlang Gin Framework
业务服务层 (Business Service Layer)
Web Router
Oauth2
Middleware
Data Pipeline
Business Service
Business Entity
Data ETL Jobs
Code Generator
GOlang Gin Framework
数据持久化层 (Data Persistent Layer)
DAO Entity
XORM Plus
Transaction
2nd Cache
Full-text Index
GOlang MySQL 、Redis 、Kafka 、ElasticSearch Access Component
服务访问层 (Service Access Layer)
GOlang Open API Proxy
GOlang GRPC Service Proxy

一期开发计划-1

- 一期开发范围：



一期开发计划-2

- 一期开发人员：

待定！

- 一期时间节点（一期大致估计 2.5 - 3个月开发周期、不含集成测试）：

第 1 周：完成系统概要设计（张峻）【6月7日】

第 2 周：完成系统核心技术验证（张峻）、概要设计优化和评审（黄海、张峻）【6月14日】

第 3 – 6 周：完成数据交换与汇集服务后端核心功能开发、部分完成数据交换管理后端服务开发（张峻+1人待定）

第 7 -10/12 周：完成数据交换管理前端+后端开发、完成数据交换与汇集服务单元自测（张峻+3人待定）