# 政务大数据共享管理平台设计说明书

# 引言

1. 目的

随着我国现代信息技术的蓬勃发展，信息化建设模式发生根本性转变,一场以云计算、大数据、物联网、移动应用等技术为核心的“新 IT”浪潮风起云涌，信息化应用进入一个“新常态”。*\**(某政府部门)为积极应对“互联网+”和大数据时代的机遇和挑战，适应全省经济社会发展与改革要求，大数据平台应运而生。

大数据平台整合省社会经济发展资源，打造集数据采集、数据处理、监测管理、预测预警、应急指挥、可视化平台于一体的大数据平台，以信息化提升数据化管理与服务能力，及时准确掌握社会经济发展情况，做到“用数据说话、用数据管理、用数据决策、用数据创新”，牢牢把握社会经济发展主动权和话语权。

1. 背景

近年来，全球新一代信息技术发展突飞猛进，移动互联网、大数据、云计算、物联网、人工智能等信息技术的突破和融合发展促进了数字经济的快速发展，在去IOE的大背景下，中国基础软件技术发展迅猛，国内相关大数据技术，云计算技术，人工智能技术已完全赶超欧美等发达国家，基础软件国产化的发展也如火如荼，国内高科技公司的相关产品完全可以替代甚至赶超原有外国相关产品，IBM小型机，EMC存储，Oracle数据库独霸一方的时代已一去不复返。近两年，中美贸易战如火如荼，国家网络安全和信息化建设加快了信息化自主安全可控建设的步伐，国产信息化产品替代成为国家网络安全建设的根本。

依据国务院关于印发《“十三五”国家信息化规划》的通知，明确发展目标：到2020年，“数字中国”建设取得显著成效，信息化发展水平大幅跃升，信息技术和经济社会发展深度融合，数字鸿沟明显缩小，数字红利充分释放；信息化全面支撑党和国家事业发展，促进经济社会均衡、包容和可持续发展，为国家治理和治理能力现代化提供坚实支撑。

在国务院印发的《促进大数据发展行动纲要》中明确提出：将大数据作为提升政府治理能力的重要手段，通过高效采集、有效整合、深化应用政府数据和社会数据，提升政府决策和风险防范水平，提高社会治理的精准性和有效性，增强乡村社会治理能力；助力简政放权，支持从事前审批向事中事后监管转变，推动商事制度改革；促进政府监管和社会监督有机结合，有效调动社会力量参与社会治理的积极性。推动宏观调控决策支持、风险预警和执行监督大数据应用。统筹利用政府和社会数据资源，探索建立国家宏观调控决策支持、风险预警和执行监督大数据应用体系。

基于大数据和人工智能技术，整合现有信息化系统，最大发挥数据价值，探索政务大数据共享管理平台，服务众多业务处室部门，为政府领导提供更多决策依据。

# 总体设计

## 设计原则

本平台在设计和建设中，严格遵循以下原则：

* **稳定性**

制定并实施平台高可用性方案、运行管理监控制度、运行维护制度、故障处理预案等，保证平台在多用户、多节点等复杂环境下的可靠性、稳定性。

* **安全性**

充分考虑数据的安全，保障数据全生命周期的安全性，避免用户受到异常攻击或敏感数据窃取；对用户接入层、平台服务层、数据资源层、平台支撑层、基础设施层进行全面一体的安全防护设计，保障平台安全；主动评估业务系统的安全状况及提供弥补措施，并提供各种操作行为的可回溯能力。

* **可扩展性**

充分考虑客户业务长期发展的需求，平台能够快速新响应新的用户，新的业务的新增要求；支持在统一系统架构中服务器、存储、I/O设备等的可扩展性。

* **先进性**

云平台的建设采用业界主流的云计算理念，广泛采用容器化、分布式存储、分布式计算等先进技术与模式，并于具体业务相结合，确保先进技术与模式应用的有效与适用。云平台所包含的各种工具、技术和组件的更新换代应在满足业务需求和运行安全的基础上，与开源社区最新或次新技术保持同步。

* **灵活性**

建立一个灵活的硬件基础架构，满足资源的随时随地按需分配，硬件基础架构通常由虚拟的服务器池、共享的存储系统、网络和硬件管理软件组成。

* **易用性**

有良好的人机接口和灵活多样的展现方式，提供友好的用户操作界面，支持资源的自动化部署，能够以云产品为单位安装并管理大数据应用组件，自动建立大数据服务间的依赖。

* **易维护性**

平台具备全节点安装部署、日常监控和运行维护一体化运维管理功能，界面简单易懂，逻辑清晰，易用性强。能够对于平台故障提供及时地预警报警，保证云计算平台的稳定运行。

* **开放性**

依照相关的国际标准和国家标准，引入业界开放的、成熟的标准，提供标准结构及接口从而保证系统成为符合标准又不失灵活性的开放平台，为未来的系统的灵活开放奠定基础。

* **自主可控性**

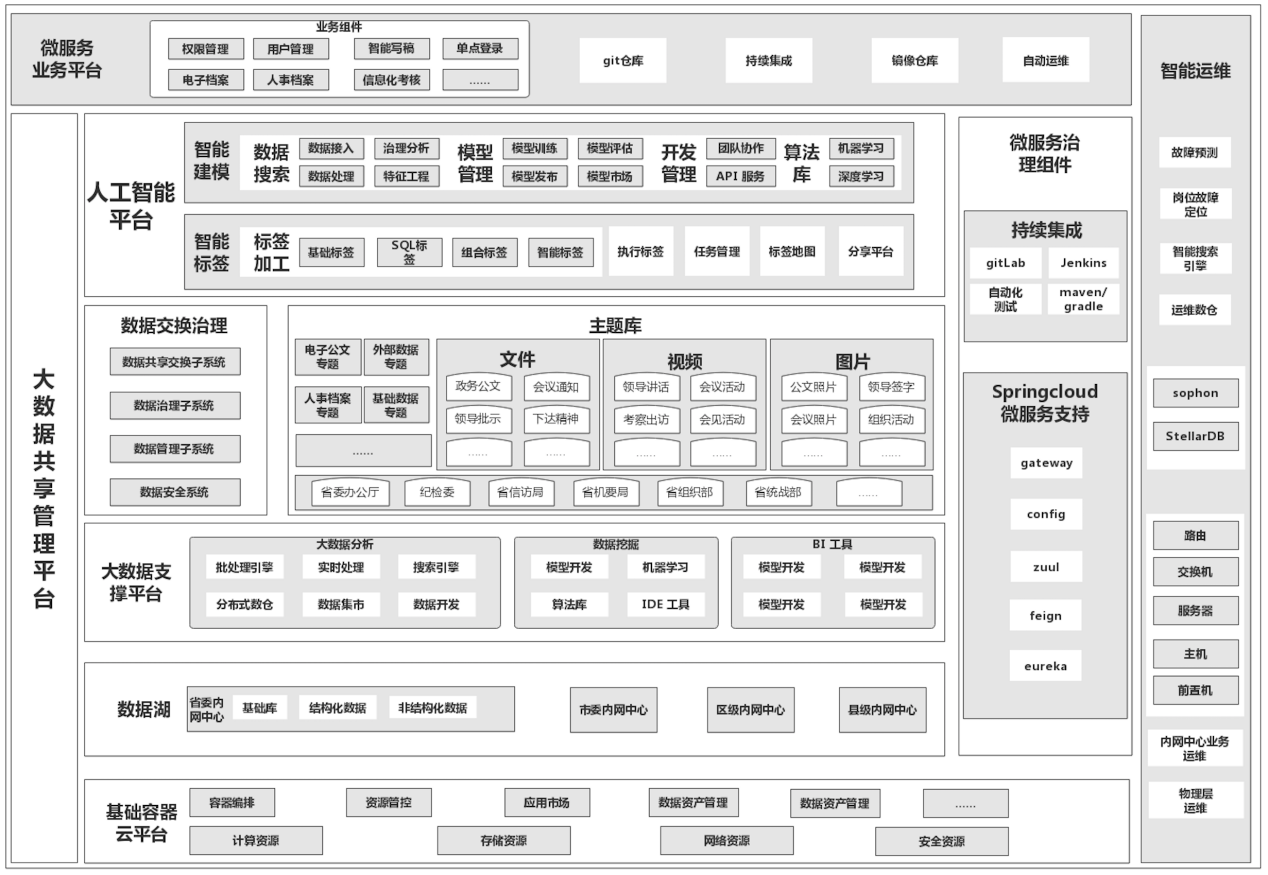
平台为国产自主软件，云计算技术完全自主可控。

## 总体架构

大数据共享管理平台，采用云产品构建私有云平台，提供计算资源、网络资源、存储资源和安全资源等服务。在 PaaS层之上，建立大数据湖，抽取数据到数据湖，打通数据壁垒，破除信息孤岛。然后按照业务主题，建立公文主题库、人事档案主题库、基础数据主题库。汇总数据之后，基于大数据技术，对数据进行智能分析，为部门领导的决策提供支持，高自由度展示各个维度的数据。

在大数据人工智能平台，在大数据基础之上，可以做智能预测，智能推荐，图形分析，为各个部门提供数据参考。

顶层业务层采用微服务治理架构，提高平台建立之后服务的开发效率，提高数据流通性，对外提供 API 接口，便于二次开发。本次业务开发先进行电子档案的业务、人事档案、信息化统计这三个业务的开发。采用微服务架构，便于扩展。



总体架构示意图

## 业务架构

大数据共享管理平台业务首先采用 ETL ，将各个部门的数据抽取到数据湖平台上。抽取的时候采用前置物理机作为跳板，确保数据的安全性；其次，使用数据治理平台，抽取元数据，做数据治理分析，确保数据的完整性、及时性、有效性；然后，进行数据清洗，存放入数据集市，对业务层提供数据支持。业务层使用 Springcloud 开发，对数据进行分析和展示，处理业务流程，对外提供 restful 形式的API 业务接口。采用前后分离的开发方式，前端静态页面部署到 nginx，直接调取后端的业务 API 实现业务流程。整个流程采用数据驱动的方式进行开发。

采用这样设计的优势是：

* **数据抽取方便快捷**

在每个部门的信息化项目上设置前置机，部署 ETL 服务，按照一定测策略抽取数据。ETL 支持集群部署，当业务数据量大的时候，可以水平扩展。同时支持各种常规数据源，包括 MySQL，Oracle，DB2，MSSQL-server 和非关系型数据库，也支持自定义的数据源。

* **安全性保证**

前置机的安装部署安装按照国家网络安全等级要求，对抽取数据做单向流向规定，设置安全防火墙。对用户权限设置详细的规划，保证用户在可以满足业务的情况下，分配最小的权限。

* **数据质量保证**

对抽取的到数据，进行数据治理。从源头到终端再到源头，形成一个闭环，统计收集所产生的数据，制定数据质量标准，分析数据血缘影响，建立元数据管理系统。通过成熟的，由上到下的元数据管理体系，保证数据的及时性、准确性、完整性等，从而为大数据提供准确分析的基础。

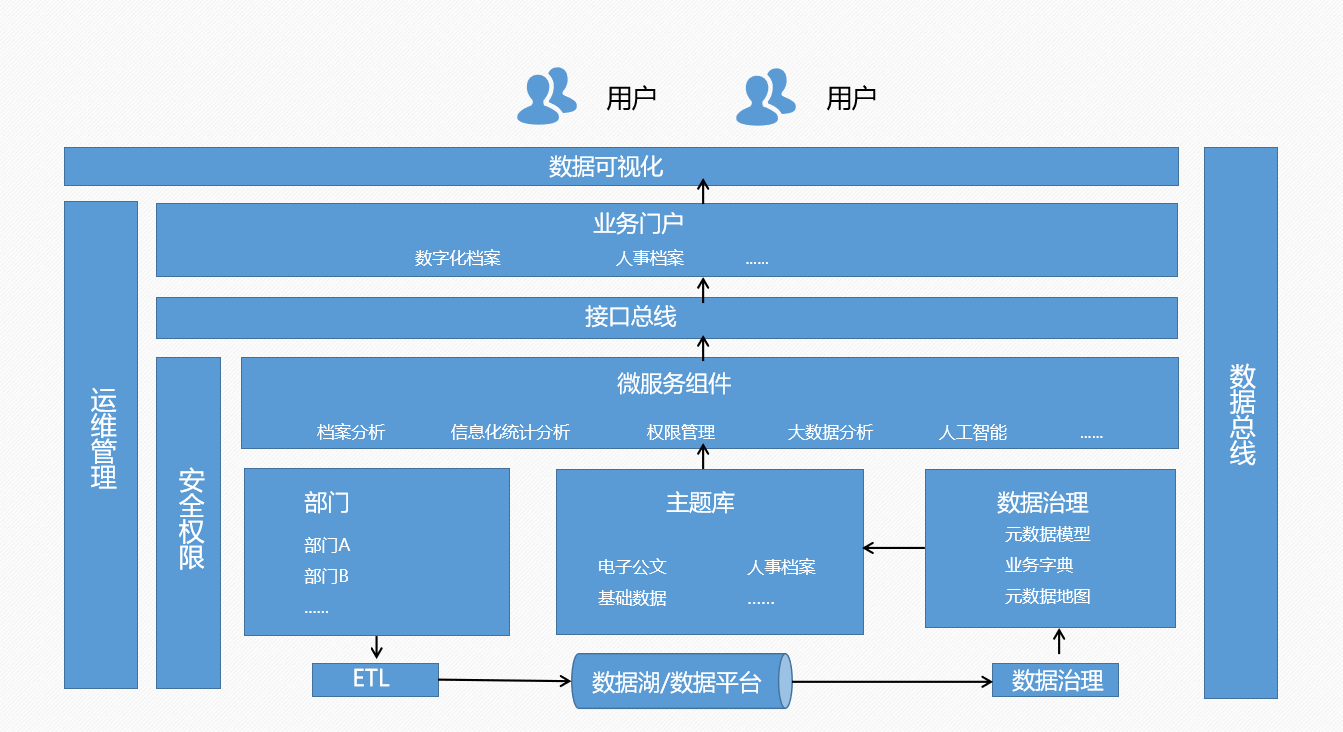
* **快速迭代**

业务流程开发采用敏捷开发，结合业界成熟的微服务技术，从需求的产生到落地，进行敏捷快速的迭代开发，保证政务业务快速的落地，提高SW部门的工作的执行效率。当需求产生的时候，收集到需求开发库，有产品经理进行需求评估，设置需求优先级，细化需求，有研发部门设计详细开发计划，进入开发流程，同时测试部门进行同步测试案例设计。需求开发完毕进行系统测试，通过灰度发布进行上线。系统产生 bug，则再次进入需求池，通过以上过程，重复迭代开发。

微服务目前有成熟的持续集成技术，对于线上项目的升级有良好的支持，可以无缝进行热机升级，用户可以正常使用系统，甚至无法感知系统已经升级。

* **职能解耦**

按照微服务架构的设计理念，同一个项目将拆分为多个系统级微服务模块。不同的业务小组负责不同的业务开发，互相不影响。例如业务小组 A负责 A 项目的开发，业务小组B 负责 B 项目的开发，开发完成的项目，部署到服务总线进行按照部署。A 小组和 B 小组可以是不同的公司或者开发部门，互相独立，开发各自的业务，充分的解耦业务。由于同时都部署到总线，A 项目和 B 项目也可以做到轻松的数据打通服务，内部设置权限进行安全管理，按照 rest 接口就能轻易完成数据交互。



业务流程示意图

## 4.架构优势

大数据和云计算技术已进入快速发展的第二个十年，随着人工智能技术的兴起和爆发，将这三个新兴技术进行结合，打造新一代智能大数据云平台，打通并整合企业应用、数据资产和AI模型，构建新一代的智能数据业务基础架构，完成数据时代背景下企业基础架构的升级和改造。

基于云计算技术构建的服务器虚拟化资源池，可为大数据平台系统的各类应用、分布式计算和存储服务组件提供多租户隔离的容器资源调配管理、应用打包部署及SLA管理、作业调度管理以及统一运维监控管理，该技术架构具有众多优势：

* **便捷快速部署**

用户可以通过Web UI、REST API或者命令行一键瞬间安装和部署大数据集群，能自动根据服务的依赖性安装所需的其他服务组件。在虚拟技术之前，部署硬件资源满足新的应用需求需要几天时间，通过虚拟化技术把这个时间降到了分钟级别，而目前基于Docker的云平台把时间降到了秒级别。Docker作为装载进程的容器，不必重新启动操作系统，几秒内能关闭，可以在数据中心创建或销毁，没有额外消耗。

* **高效率资源利用率**

典型的数据中心硬件资源利用率是25%，原因是传统的虚拟化技术，需要额定占用云平台资源且不释放，导致云平台资源利用率很低，容器化平台通过更积极的资源分配，以低成本方式对新的实例实现更合理的资源分配，同时支持动态分配，灵活调度系统资源，从而大大提高平台的资源利用效率。

* **完整的资源隔离**

该平台通过优化Kubernetes资源管理框架实现了基于Docker容器对CPU、内存、硬盘和网络更好的隔离。平台中Docker容器的隔离目前是由Linux内核提供的六项隔离，包括主机名与域名的隔离，信号量、消息队列和共享内存的隔离，进程编号的隔离，网络设备、网络栈、端口的隔离，挂载点（文件系统）的隔离，用户和用户组的隔离。这些隔离保证了不同容器的运行环境是基本不受影响的，比如挂载点的隔离，就保证了一个容器中的进程不能随意访问另外一个容器中的文件。

该平台相比于传统的Apache Yarn管理框架和开源Kubernetes的资源管理框架而言，在资源粒度方面可以管控磁盘和网络，而传统资源调度框架只能管理到CPU和内存；在隔离性方面，容器技术有天然的优势；在依赖性和通用性方面，不依赖于Hadoop组件以及技术，这意味着可以实现所有上层应用的云化开发、测试、升级以及管理调度。

资源隔离对比表

| **产品** | **资源粒度** | **隔离程度** | **依赖性** | **通用性** |
| --- | --- | --- | --- | --- |
| YARN | CPU/MEM | 进程级别、不精确 | 依赖某个HDFS | 支持少量计算引擎 |
| Kubernetes | CPU/MEM | Container | 不依赖Hadoop | 支持通用Linux负载 |
| 大数据共享管理平台 | CPU/MEM，DISK，NETWORK | Container+Quota+VLAN | 不依赖Hadoop | 支持大数据及通用应用 |

* **灵活资源调度**

平台支持对大数据组件自动扩容或者缩容，同时也允许其他应用服务和大数据服务共享集群，从而提高资源的使用率。平台创新的支持抢占式资源调度模型，能在保障实时业务的同时，提高集群空闲时的资源占用，让批量作业和实时业务在互不干扰的情况下分时共享计算资源。通过支持动态扩容和收缩集群，从何实现了资源的动态分配与调度，整个过程热插拔业务、服务无需重启。

* **自动修复**

平台的Replicator模块负责检测集群规模，当集群中服务发生问题时可以另起一个服务实例实现集群的自我修复功能。举例而言，某个Hyperbase Region Server由于硬件原因服务停止，平台能够实时感知，并在管理的资源范围内另起一个Region Server的Docker容器接替因故停止的容器，动态的保证了服务集群的整体稳定性。

* **应用隔离**

该平台的大数据支撑平台TDH以及应用服务还引入微服务的架构，显著降低用户部署环境对TDH以及应用服务稳定性的影响，提高了部署的可用性，并且能让用户在不停业务以及服务的前提下，享受到快捷的TDH以及应用服务更新版本滚动升级。

* **大数据产品一键部署**

用户不需要在关注大数据集群该如何部署，云平台提供按需一键式配置集群大数据平台。能够一键式部署云产品：数据仓库，数据集市，实时计算，搜索引擎，NewSQL数据库，数据挖掘平台，人工智能建模平台等。

* **多租户微服务创新架构**

1. 多租户

动态调配资源，实现资源的共享的优势，并可以通过设置，享受资源隔离的性能优势。

1. 微服务

通过服务实现模块化设计，让微服务成为软件开发中的砖瓦，实现服务即产品的技术理念。

1. 应用优势

资源碎片化，实现最大限度的共享（多租户模式的资源是动态分配，并可以设置最低资源以及抢占模式）和最大限度的安全隔离（单个微服务或者多租户宕机后，其他微服务和多租户不受影响）。

由于平台具有全面的数据、完善的功能、可扩展的架构、高性能的计算能力，业务应用就会逐步叠加，而这些业务应用都应该各自独立，互不影响，而微服务是这种方式的最佳架构。

* **成熟完整的大数据平台产品线**

支持涵盖数据生命周期的全图形化的大数据服务，包含ETL、数仓、报表、检索、数据挖掘和数据库等大数据产品；提供云化的数据库与数据服务。

* **全流程DevOps开发工具链**

提供全套的DevOps开发工具，结合星环科技多年项目实战经验，实现DevOps落地，实现应用快速迭代上线。提供丰富的微服务治理功能包括服务注册、服务发现、服务监控、熔断、容错、调用链管理等，有效解决微服务的分布式特点带来的管理复杂性；支持存量应用的平滑迁移。

# 功能模块

## 大数据支撑平台

大数据支撑平台是整个大数据共享管理平台提供分析计算和存储的引擎和工具，是大数据共享管理平台的底层功能支撑，为数据仓库的构建提供高性能、高稳定、高可用的数据库软件，并通过离线处理、流处理、全文检索、数据开发、资源管理、安全控制等组件为数据交换与治理系统的开发提供支撑。通过数据采集、数据交换、数据存储、数据治理到数据共享与服务，为大数据共享管理平台建立起整个数据共享交换体系。

#### 1.1设计原则

大数据支撑平台在考虑本期项目建设需求的基础上，同时考虑满足后续系统平滑扩展升级要求，平台设计遵循如下原则：

* **开放性**

1. 平台开放性

平台保持开放性，并在可靠性、安全性、管理性、高效性方面进行增强。

1. 接口开放性

平台提供标准JDBC（包括JDBC type 4 driver）、ODBC驱动（至少兼容linux、windows（64位））。

* **兼容性**

1. 兼容通用硬件

平台支持运行在X86架构的通用PC服务器上；支持运行在基于ARM国产芯片架构的Linux服务器上。

1. 兼容通用操作系统

平台兼容国产操作系统（红旗、麒麟等），兼容主流的Linux操作系统（SuSe、RedHat、CentOS等）；

1. 兼容标准接口

平台组件兼容对应开源组件开发接口；兼容标准JDBC、ODBC接口；兼容标准SQL与存储过程语法。

* **扩展性**

平台跟随业务需求的扩展，支持基础数据模型、应用分析模型和前端应用的扩展性；支持统一系统架构中的计算、存储、网络等设备的动态扩容；支持通过平台可视化管理界面进行集群和相应角色的扩展。

* **安全性**

平台内部安全和外部安全同步建设，应符合大数据建设安全标准与电子政务内网保密安全的相关要求，按照国家标准、行业标准、安全规范等实现大数据服务平台的安全建设。

#### 1.2 大数据分析

大数据支撑平台核心功能是提供平台数据开发和业务开发所需要的大数据分析组件。平台提供分析型数据仓库，用于构建数据湖和主题库，对结构化数据的批量分析与存储；提供分布式的NoSQL数据库，用于对半、非结构化以及视频影像等大对象的存储和实时在线高并发检索查询；支持实时流式处理；能够支持数据检索，提供搜索引擎；提供一系列完整的数据开发工具，一方面通过统一的工具对以上各组件进行操作，另一方面也为数据治理、数据共享交换、数据资源目录、平台安全运维等子系统的开发提供对应的组件。

* **分析型数据仓库**

在本期项目中，数据湖和主题库中结构化区域的构建主要基于分析型数据仓库。分析型数据仓库提供高可靠的分布式存储，以及高性能的数据分布式计算与分析。考虑到数据湖和主题库在本期项目中的构建特点，具体对于分析型数据仓库应该具备的功能描述参见【数据仓库开发】小节。

* **分布式NoSQL数据库**

分布式NoSQL数据库是用于海量半、非结构化或视频影像数据存储和数据实时在线数据处理的非关系型数据库，能够满足OLTP、OLAP、批处理和搜索等高并发在线业务需求。在本期项目中，分布式NoSQL数据库可用于构建视频、图像、日志等数据在数据湖和主题库中的存储区域，以及与业务相关的对高并发在线查询要求较高的结构化数据存储区域。

1. 功能设计：支持标准SQL语言和存储过程，能够对数据进行高效批处理；分布式NoSQL数据库可支持多种索引，如局部索引、全局索引、高维索引、全文索引、智能索引等；分布式NoSQL数据库通过k-v存储结构以及辅助索引进行快速、精确的并发查询，满足OLAP低延时数据分析需求和离线批处理；分布式NoSQL数据库可支持多种数据类型存储和检索能力，包括结构化数据、半结构化、非结构化数据，以及文档、图片、音者视频等大对象；分布式NoSQL数据库具有自动容错能力。通过对索引分片，并对每个分片创建多个子任务。一个子任务失败后，集群会在其它节点上重建失败的子任务；支持事务日志，确保索引操作无丢失，即使整个集群因故重启或者任务没有索引到磁盘，重启后索引模块仍然可以根据日志记录，快速恢复索引任务。
2. 指标要求

为满足数据湖和主题库对半、非结构化数据的存储以及并发查询分析的基本需求，NoSQL数据库满足以下性能指标要求：按关键字检索单表记录延时小于200ms，单个节点并发度超过1000；按关键字检索多表关联记录延时小于1s，单个节点并发度超过1000。

* **实时计算**

实时计算组件如Spark Streaming、Storm、Flink等，主要用于海量高并发、低时延要求的实时数据的计算和分析场景，可进行基于微批或事件驱动两种方式对实时数据进行处理。在区大数据局应用场景中，将主要用于IoT、实时影像、视频分析等以实时流方式接入数据湖的数据类型。

功能设计：

1. 语言支持

实时计算组件支持Java、Scala等语言的流应用开发，提供类SQL接口，支持以标准SQL、存储过程和UDF的方式实现流业务开发。

1. 输入输出

实时计算组件支持文件系统、Kafka、Flume、TCP Socket等多种数据源以及自定义数据源作为输入源；支持HDFS、Kafka、NoSQL数据库等为输出，并支持多种输出格式，如text、orc、hbase或自定义自定义存储等。

1. 处理模式

实时计算组件支持基于微批和事件驱动的数据处理模式，支持复杂事件处理CEP。基于微批模型，将一批任务对应一个分布式弹性数据集，从而提高吞吐量；基于事件驱动模型，逐条读取数据流中的数据，从而降低时延。

1. 流式挖掘

实时计算组件支持用户在流上运行数据挖掘模型，支持将训练成熟的算法模型用于实时数据流中，实现数据价值挖掘、信息告警等准实时流式挖掘任务；支持主流流式挖掘模型，如Top-K排序操作、静态表关联、多数据流关联、实时机器学习等。

1. 时间窗口

实时计算组件支持时间窗口统计，如滑动窗口、跳动窗口、无限窗口和会话窗口等，满足对一定的时间窗口区间做多表关联、聚合或者统计。

1. 流任务管理

实时计算组件提供RESTful接口查询流任务的实时指标，支持持久化流任务的历史指标；提供可视化管理界面对流任务进行生命周期管理，包括流任务的创建、运行、监控、停止和删除等。

* **数据检索**

数据检索组件主要通过文本分词提供大规模数据的模糊查询、全文检索等。在本期应用场景中，数据检索组件可用于为数据湖和主题库中存储的海量数据建立索引，并作为各部门对底层数据检索查询的搜索引擎。

* **大数据开发工具**

大数据支持平台除了提供上面提到的基础大数据分析组件外，还需要提供一系列完整的大数据开发工具，一方面对基础的大数据分析组件进行统一的操作和开发，另一方面能够满足对数据共享交换子系统、数据治理子系统、数据资源目录、数据安全子系统等的开发和建设。

为了满足共享交换体系各子系统的建设，大数据开发工具需涵盖但不限于涵盖数据开发工具、数据迁移工具、任务调度工具、元数据管理工具、平台BI分析工具、Cube设计工具、日志管理工具等。

1. 数据开发

大数据支撑平台需提供数据开发工具以针对大数据分析组件进行统一的操作和管理。通过数据开发工具能够可视化地利用SQL等语言操作分析型数据仓库、NoSQL数据库、实时计算组件和数据检索引擎。数据开发工具主要用于SQL、UDF和存储过程等辅助开发与调试，能够支持数据预览、关键字高亮、语法检查、格式化辅助、内嵌命令行等功能。

1. 数据迁移

数据迁移工具是利用大数据支撑平台构建和开发数据共享交换子系统的核心工具。大数据支撑平台需提供数据迁移工具用于数据在RDBMS和Hadoop平台之间的迁移（即数据交换、数据共享过程），或实现Hadoop平台内部数据的流转（即数据治理与集成过程）。数据迁移工具需能够通过可视化界面来设计和创建ETL任务，支持多种格式的数据源，包括CSV、JDBC、XML、JSON、关系数据库以及跨集群导入；支持多种常用的数据转换操作，如连接、去重、聚合、关联、清洗、卸数等；支持多种数据导出方式，包括text、orc、关系数据库、kafka以及跨集群导出；支持多种迁移方式，包括全量、增量、准实时等。

1. 任务调度

任务调度工具是通过可视化界面进行工作流的设计、调试、调度与分析的工具，也是配合数据共享交换子系统、数据治理子系统等进行工作流程构建与开发的工具。任务调度工具需要支持图形化交互式配置；支持Shell、SQL、JDBC、HTTP等任务类型；支持自定义Java任务；支持多周期调度；支持多种分析能力，如依赖关系、执行历史、甘特图等。

1. 元数据管理

元数据管理是通过可视化界面对集群中存储的库表、对象等的元数据进行查看、管理的工具，是数据治理工作的主要目的之一，也是帮助构建数据资产目录的重要工具。大数据支撑平台提供的元数据管理工具需要能够支持底层存储库表与对象的元数据在web界面自动地、可视化地展示，并支持通过SQL或者可视化地方式查询元数据（表、字段、存储过程）的全文检索与完整的活动监控；支持监控所有数据和程序的更改历史；支持进行数据血缘分析和影响分析。

1. 平台BI分析

与能够支持上层应用级展示的可视化BI分析工具相区别，大数据支撑平台提供的BI分析工具应该是一种轻量级的、能对平台内部数据指标或展示效果要求不高的业务报表进行交互式分析的工具。它的特点是能够支持轻量、灵活、快速的部署；支持提供数多种报表样式进行多维度分析和；支持自助式分析时序数据的展现；支持报表导入和导出；支持团队协作和共享。

1. Cube设计

大数据支持平台需提供Cube设计工具服务于主题库中需要进行高效维度查询的业务库。Cube设计是通过可视化界面进行Cube设计，主要满足于M-OLAP场景数据集市的构建，基于已有的数据表建立符合业务特点的优化存储数据表，使得基于源数据表的查询能够落在优化存储的数据表上，进而加速查询性能，提高查询并发度。支持所建Cube实例化于HDFS；支持增量构建和降维优化；能支持雪花模型和星形模型；能支持多种格式的数据源，包括HDFS、RDBMS；支持对对Cube进行生命周期管理，包括Cube建立、更新、删除的监控管理；能支持CUBE模型导入和导出。

1. 日志管理

平台日志管理工具对平台日志进行收集和存储，并通过可视化界面进行日志的查询和分析，能满足不同场景的日志分析业务。日志管理工具也是数据安全子系统的组成部分。大数据支撑平台需提供统一的日志管理工具对平台的日志进行统一存储、查询和管理。应支持日志监控告警；支持全链路高吞吐、高安全和高可用。

#### 1.3 数据挖掘

数据挖掘组件是针对数据科学家提供的能够用编程语言针对海量结构化、半结构化和非结构化数据进行统计学习分析或机器学习算法模型开发的平台，一般通过IDE工具进行组件的交互式操作。数据挖掘组件是基于分布式计算场景开发的，需要能够支持R、Spark等框架，并提供相应且常用的分布式统计算法、机器学习算法以及关联关系网络等图分析算法，提供R语言、Python语言和API等开发接口。

* **可视化挖掘**

1. 数据接入：数据挖掘组件需支持多源、多方式的数据接入，包括但不限于本地文件、数据库、HDFS、Glusterfs等，支持主流数据库驱动。
2. 数据探索：数据挖掘组件通过提供可视化IDE工具进行交互式数据探索，并提供数据视图对关系型数据库、本地文本和大数据平台内多种数据源进行集成和展示，包括元数据、主数据、数据连接等。
3. 数据预处理

* 模块功能：能够在平台上构造包括数据清洗，数据变换和数据规约在内的数据预处理流水线。
* 数据分析：提供对数据特征的统计工具，能够在数据集合进行分布分析，对比分析，统计量分析和相关分析，为数据建模人员提供基本的特征描述。
* 数据清洗：对原始数据的无效异常数据的过滤，缺失数据的补齐。并将预定义清洗模式统一应用在数据云平台的全量数据上。
* 数据变换：提供数据属性转换、新属性生成在内的处理能力。为机器学习任务和算法的需要提供有效的样本输入。例如可以通过定义的数据标记模板为监督学习提供标记过的数据样本。
* 数据规约：提供基本数据属性的归一化工具，为机器学习任务和算法的需要提供高质量的样本输入。

1. 特征工程

支持常见特征工程流程，如特征变换、特征重要性评估、特征选择、特征生成等，包括且不限于归一化、标准化、离散化、one-hot编码等。

1. 模型训练

* 模型构建：可以通过图形化方式或自编码构建模型与算法。
* 资源容器化的调度与封装：针对训练任务，对现有物理资源进行容器的动态构建与释放，提供模型训练的物理设施。
* 模型训练：支持多任务排队；支持多任务并发；能够实时保存训练日志及模型。
* 训练监控：快速构建模型训练任务，实时启动、监控和停止各个训练任务。同时可以根据任务的状态对资源进行动态调度。
* 模型评估：支持主流模型评估方法。包括但不限于KS，Lift，AUC，R2，MSE等。

1. 模型发布

* 一键发布API：基于模型的训练结果快速构建应用服务，并通过容器化的方式进行发布与运行。
* 模型应用的构建：基于模型训练结果，通过图形界面构建相应的应用服务，并通过容器化的方式打包与运行。
* 模型导出：支持模型以通用的格式一键导出。

1. 模型管理

组件支持模型管理，包括内置行业模板、自定义模型模板，例如担保链分析、用户画像、风险图谱、营销推荐、用户流失预警、商圈聚类分析、客户精分、实施推荐、垃圾短信检测、实时人流密度估计等，能够直接切合公共安全、公共管理、经济发展等应用方向，用户可以基于组件快速地基于自身业务系统构建解决方案。

1. 模型运行

* 运行管理：对所有模型的应用服务统一管理， 构造机器学习服务的集中运维平台，能够查看模型被调用次数；自带负载均衡器，能够对流量带宽进行管理；能够实现弹性伸缩以及手动扩容缩容；
* 运行监控：运行期间提供实时监控页面，自定义告警阈值等。

1. 模型协作

平台支持模型协作，多人协作模型开发模式。

* **分布式算法**

为了能让数据科学家基于分布式平台进行数据挖掘分析和算法模型开发，数据挖掘组件能提供基本的统计学习和机器学习的分布式算法。其中，机器学习算法库包括逻辑回归、朴素贝叶斯、支持向量机、随机森林、聚类、线性回归、推荐算法等；统计算法库包括均值、方差、中位数、直方图、箱线图等；支持在平台上搭建多种分析型应用，例如用户行为分析、精准营销、标签分类等。

1. 统计学算法

提供常见统计算法的分布式化，统计算法运行在分布式计算引擎上。现有的分布式统计算方法如下：Min，Max,Mean,Variance,Percentile Approx,Minmaxnormalization,Znormalization,Median,Percentile,Boxplot,Cardinality,Correlation,Histogram,BInning,Pie,Boxplotapprox,Boxplots.

1. 机器学习算法

集成多种常见机器学习算法的分布式化。现有的分布式机器学习算法如下：LogisticRegression逻辑回归，NaïveBayes朴素贝叶斯，SVM支持向量机，K-Means聚类算法，LinearRegression线性回归，CollaborativeFiltering协同过滤，Apriori频繁项集，AssociateRule关联挖掘，DecisionTree决策树，ANN人工神经网络算法，RandomForest随机森林，Gradient-BoostedTrees梯度提升树，保序回归，Apriori算法，协同滤波（CF），FP-growth，PageRank，TF-IDF

Word2Vec，奇异值分解(SVD)，主成分分析(PCA)，Histogram，BoxPlot，Binning，Pie，流处理统计算法，流处理统计中位数算法，流处理统计终值算法，流处理统计分位数算法，流处理数据统计算法，流处理数据数据清洗，流处理数据直方图，流处理数据箱型图算法，流处理数据直方图算法，流处理数据密度算法，流处理数据kmean算法，流处理数据线性回归算法，

1. 自定义算法

当组件自带的库函数、库算法无法满足开发需求时，开发人员可通过平台自主开发分布式算法并加载使用，如R语言算法、Python语言算法。

## 人工智能服务

大数据共享管理平台提供企业级AI应用平台。提供数据清洗、数据分析挖掘、机器学习、深度学习、模型管理，API部署，工作流调度等功能，助力企业AI时代业务创新与变革。

可以完成包括数据预览、数据预处理、特征工程、建模、模型评估以及模型部署等整个数据分析流程。对于大多数企业用户，要通过自己的力量从无到有打造人工智能平台的代价是巨大的，需要有技术实力强大的大数据团队和AI 团队作为基础。打通了大数据平台和人工智能平台，业务分析师和数据分析师可以通过自动建模以及内置的行业模板轻松构建对应AI 模型，从而提升业务价值。

* **完整的数据挖掘流程**

对于一款实用的数据挖掘工具来说是必要的。机器学习产品不仅能够胜任完整的数据挖掘流程，在流程的每个步骤都能提供足够丰富和方便的算子，供用户灵活实用。

* **丰富的机器学习算法**

包含分布式算法支持、流式机器学习支持、自定义算子支持与标准模型导入导出。

* **强大的企业级特性**

包括完整的多租户能力支持，对于计算资源、数据资源等到细粒度隔离与共享；支持模型协作功能；集成了LDAP、Kerberos等常见的权限认证手段；支持工作流与定时调度；结合容器技术对于大型集群进行高效的管理和调度。

* **完善的深度学习支持**

包括网络结构图形化拖拽、深度学习框架整合、分布式GPU优化、深广结合与经典网络结构支持。

## 可视化监控

大数据支撑平台提供统一的可视化运维监控界面用于对整个大数据支撑平台组件和工具的配置、管理和运维。通过可视化运维监控界面，用户只需通过几个手动步骤，就能在服务器或云平台上部署大数据支撑平台；可以对集群的资源进行可视化监控、监测、度量和告警；可以浏览各服务的状态，并且在告警出现时采取恰当的措施以处理应对；可以对软硬件角色进行管理、升级和迁移等。

## 应用设计

### 4.1档案数字化系统

* 建立数字化档案系统，将纸质文档扫描入库，同时把扫描件、电子文件等非结构化数据智能转化为结构化数据入库存储
* 文件上传，建立适合客户现有和未来档案的多种模板，智能化实现文档的自动归类、自动属性识别（含文档名称、发文部门、日期、关键词等内容），文件检索整理，详情分析等多方位结合的数据整合平台
* 所有档案包括领导批示，均进行结构化存储，实现聚合统计和查询
* 查询结果，实现档案的数字化结果和原件同时展示
* 档案可视化查询

1. 按照人物、机构、时间、事件、档案类型等内容实现档案的统计分析展示
2. 系统默认一种展示页面，根据用户选择或设置主题、内容等实时形成图表统计结果，也可以查询对应的档案细节
3. 用户可以任意选择档案时间、范围、主题等形成对应档案归类分析报告，实时掌握档案库情况。
4. 可以按照标题、关键词、类型、时间等条件进行档案的搜索

* **接口开发**

1. 档案数字化录入
2. 档案搜索展示
3. 档案上传
4. 人物工作关联大数据分析展示

### 4.2智能人事档案系统

人事档案负责部门有大量纸质档案，由于纸质档案信息的统计十分低效和耗费精力，所以对人事档案进行信息化建设。同时，人事档案对于安全性有较高的要求，需要详细设计业务权限和部署的权限。

第一步就是录入档案信息，采用图片识别的人工智能算法，对图片信息进行读取，结构化保存到数据库，同时对所有入库数据进行加密。

第二步，数据治理，确保录入数据的质量。

第三步，聚合分析，为用户提供高效的查询结果。

**接口开发**

1. 聚合查询接口开发
2. 档案上传接口开发
3. 档案详情接口开发
4. 数据导入接口开发

### 4.3信息化应用监测与分析系统

各个信息化系统的使用率目前统计信息空白，需要抽取各个系统的日志，对该系统的使用情况进行分析：

* + 统计各个部门各个信息化系统的使用情况，进行分类展示
  + 按照部门统计，按照个人统计，按照系统统计。
  + 统计各个部门的使用的比例，查看哪些部门使用率高；
  + 统计所有人的使用率，查看哪些职员活跃使用；
  + 统计各个系统的使用率，查看哪些系统使用率高；
  + 每个部门的系统使用 top10 / last10
  + 部门活跃 top5 / last5
  + 统计人员在线时长，做排名展示
  + 协助管理部门分析信息化系统使用情况，加快推进电子化办公

**接口开发**

1. 抽取日志接口开发
2. 按照部门统计接口
3. 按照个人统计接口
4. 按照系统统计接口
5. 各种聚合统计
6. 各种排名统计