# 数博目标

## 建设目标

数博数据平台总体架构图如下：



数博数据平台的总体目标为：

一、通过研发和集成多种大数据技术，建立从采集融合、存储组织、查询分析和可视化展示等各种数据处理环节的大数据处理各分析机制，支持大数据应用的公共技术的形成和技术解决方案的支撑。

二、通过将数据以关系型、索引型、多维数据集和图的方式进行建模和组织，充分发挥各种数据组织类型的优势，使应用分析和设计者在应用分析和设计时可以根据用户的具体场景选择适合的技术来进行数据的处理和分析功能的设计，从而快速、低成本构建出适应用户场景并具有一定扩展性的应用解决系统。

## 主要组成

数据采集

索引数据组织

图数据组织

## 关键技术

### 支持多种数据来源、数据格式的采集和融合

### 基于分布式索引的数据检索和聚合技术

### 基于知识图谱的关联数据分析技术

## 领域智能检索目标

领域智能检索是数博大数据分析平台的子系统之一，主要解决在领域环境下，以知识图谱为主线建立数据融合、数据存储组织和数据检索的机制，本次建设开发如下工具套件：

**一、本体定义工具**

本体定义工具用来完成进行领域本体（即领域概念模型）的定义工作。依据领域检索和应用的需求确定需求，建立本领域的实体以及实体间的关系。在本系统中，使用属性图进行领域本体的表达。

**二、数据融合工具**

要实现领域智能检索的目标，除了定义和开发领域本体之外，最重要的工作是基于本体从多种异构的数据源中抽取数据，形成领域知识库，以知识库为基础向用户提供数据检索的功能。数据融合工具实现从各种不同类型的数据源中以多种方式获取不同格式的数据，并依据本体定义进行数据的清洗加工、转换、实体的识别、对齐和消歧等操作，使数据符合检索和分析的需要。

**三、数据检索工具**

数据检索工具可以实现自然语言检索、关键字检索和过滤条件检索等多种检索方式，并可以将多种方式结合使用。工具内置机器学习模型会自动识别出输入的自然语言中描述的检索条件，对数据进行检索和过滤。检索结果按相关度排序展现。在线数据检索工具也是一套定制工作，可以根据领域中信息的特点定义不同实体、关系和属性的各种展现方式，并通过内置的链接协议建立这些展现的关联，从而达到数据关联浏览和分析的作用。

**四、图浏览工具**

对以关系为主的业务，系统提供基于图的分析工具，可以通过节点和关系的方式展现检索的结构，基于检索结果还可以进一步关联检索更多的内容，也可以了解当前节点或节点集合的详细信息。

# 本体定义工具

## 需求分析

领域知识图谱基于领域内的概念体系进行构建。为了可以定义完整的、符合检索和分析需要的领域概念体系，应该提供概念体系定义工具，支持领域知识图谱概念体系的构建和演化。

概念体系定义工具应该支持知识图谱中的实体、关系和属性的定义，属性可以支持多种数据格式，以满足结构化和非结构化数据存储的需要。并可以将定义的领域概念以关系图方式进行展现，提供以图为中心的友好的展现和操作方式。

概念体系定义工具应该支持领域知识的定义，包括业务校验逻辑和业务推理逻辑等领域知识的定义，并可以在数据质量管理、业务分析场景下使用。

概念体系定义工具应该支持领域概念体系的演化，根据领域检索和分析的需要，可以建立、推理或演化出新的实体、关系和属性，并自动进行数据结构的同步。

## 主要特性

|  |  |
| --- | --- |
| 主要特性 | 详细描述 |
| 领域模型定义 | 支持将领域模型定义为本体、关系和属性的方式 |
| 领域模型演化 | 当业务分析需要领域模型进行变更时，可以对本体、关系和属性进行添加、删除和变更操作。  添加：新添加的本体、关系或属性需重新从数据源同步数据  删除：删除后的本体、关系或属性，期实例数据一并删除  变更：实体、关系或属性发生变更时，支持通过模型定义操作，在模型发生变化时，对应的实例数据一并发生变化 |
| 数据关联 | 统计关联：解决统计数字与具体信息的关系  导出信息关联：解决原始信息与导出信息的关系 |
| 索引定义 | 关键字索引：指定需要索引的实体和关系上的属性，索引后，可以快速查找  全文索引：指定需要建立全文索引的实体、关系及其属性，索引后，可以进行全文检索  索引变更：在过些发生变化时，支持基于知识图谱数据重建索引 |
|  |  |

## 设计方案

## 扩展性设计

# 数据融合工具

数据融合工具主要面向基于工具的实施人员和最终用户，应用系统开发人员可以基于工具提供的扩展机制进行扩展功能开发。

## 需求分析

当前，数据源包括缉私案件生产库和深加工数据库，这两部分又分别由多个库组成，每个库中存储不同的案件信息。库中的数据主要为结构化数据，包括时间、名称、文本、金额、日期等格式，为了能够快捷方便地构建案件知识图谱，应该提供数据融合工具将当前各种数据源中的多种数据整合形成案件知识图谱的功能。

融合工具应该支持各种常见的数据库如 Oracle、DB2、MySQL 等，还支持各种平面文件和Hadoop、HBase、MongoDB 等 NoSQL 数据源。工具应该具备数据同步作业的可视化定义、同步作业的运行监控功能，支持从数据源到目标图数据库中的数据格式转换、信息拆分、歧义消除和冗余处理等功能，由于同一案件的信息可能会在多个数据库中同时出现，融合工具还应该能够在作业定义时，根据案件的数据源、信息的新鲜程度预定义的规则自动合并相同的案件或案件的部分信息的功能。

刑事和行政案件也涉及大量的附件、照片、报告等非结构化信息，这些信息以文件的方式存储在磁盘上，融合工具应该能自动提取这些文件中的文本、表格、图片等信息，与案件信息进行关联，同时还可以将文本和表格的数据转化为结构化数据，支持知识图谱的构建和进一步的关联分析的需要。

融合工具能够进行全量和增量的数据同步，在数据源首次接入时，进行全量的数据同步，之后建立周期性的同步机制，在数据源中的数据发生变化时，可以识别增量的变化信息，自动同步到目标知识库中。

融合工具应该能够支持大数据量同步时的高性能和高并发的要求，能够进行分布式集群的部署和数据同步，提供高效的运行服务。

## 主要特性

|  |  |
| --- | --- |
| 主要特性 | 详细描述 |
| 关系型数据库支持 | 支持从Mysql、SQL Server、Oracle等主流关系型数据库中获取数据  支持从人大金仓、达梦等国产关系型数据库中获取数据 |
| 数据文件支持 | 支持从Excel、CSV、XML等数据文件中读取数据 |
| Web数据源支持 | 支持从Web Service、REST API和普通Web页面中读取数据 |
| 数据可视化映射 | 支持将数据源中的数据以可视化的方式建立与本体的对应关系，可以支持数据源同步为本体、关系或属性 |
| 可视化作业定制 | 支持图形化的作业定制机制，通过拖拽完成数据融合过程的定义  支持通过JS脚本和Java程序实现复杂逻辑的转换控制 |
| *内置文本数据* | *内置中文自然语言处理程序，实现文本数据的格式化处理*  *内置中文全名实体识别程序，实现文本数据中识别实体、关系和属性* |
| 多种数据同步方式 | 1、实时同步方式：通过定制工具实时运行同步作业完成数据同步  2、定时同步方式：通过Web控制台设定同步周期，定期同步作业  3、并行同步：可以设置多线程并行数据同步方式，提升同步效率 |
| 完整的监控机制 | 支持数据同步过程的日志监控，可以设定日志的详细程度  支持Web在线监控作业的运行情况 |
| 插件化的扩展机制 | 通过Java插件化开发机制，支持对数据源、转换等逻辑的控制 |
| 支持数据源的管理 | 可以管理需要融合的数据源，并对数据源中的数据架构进行浏览和查询；  支持建立数据源中的数据视图，以方便数据融合时直接使用。 |

## 设计方案

## 扩展性设计

# 质量管理工具

## 需求分析

质量管理工具主要实现对案件知识图谱的完整性、基准性、权威性、主权性、鲜活性、安全性以及稳定性的监控。基于质量管理工具监控和管理图谱和概念体系数据的一致性、完整性和准确性，并可根据预置的规则进行数据和结构的更新、补全等操作。

质量管理工具支持案件知识图谱的演化，可以支持敏捷的知识图谱开发过程，包括但不限于：1）演化出新的实体；2）演化新的关系；3）演化出新的属性；4）变更当前实体或关系的名称；5）变更属性的名称、类型等；6）实体、关系或属性废弃。在进行知识图谱演化时，可以自动同步概念体系和实例数据的一致性。

基于质量管理工具，可以进行知识图谱一致性检测，包括结构与实体的一致性、实体之间的一致性等。当结构与实体不一致性时，可以根据需要进行结构或实体的更新。当实体与实体之间不一致时，可以根据检测规则确定数据的错误点，并进行数据更新。

质量管理工具可以监控知识图谱中结构和实例数据的变更，生成监控日志，并根据监控信息统计和分析知识图谱的变更情况。

## 主要特征

## 设计方案

## 扩展性设计

# 语义检索工具

## 需求分析

语义检索引擎可以通过用户输入自然语言，理解用户查询意图，根据查询意图找到符合结果，并按照符合度排序。具体功能包括以下几个方面。

1、查询输入识别功能

实现从无模式的“自然语言输入”到“查询图”的转换。“自然语言”是检索用户任意输入的，查询图识别方法通过自然语言处理获得相应的词列表，并基于领域词汇表、行为特征表，对词意进行理解（例如，有些词是代表概念、实体的，有些词是代表关系的、有些词是代表实例数据值的等等），并按照查询图模型构造查询图。

2、节点匹配功能

实现从“查询图”到“匹配节点”集合的转换。查询图中包含节点类型、节点属性及属性值、节点标签、节点之间关系类型、关系的属性及属性值、关系标签等信息。节点匹配方法根据给定的信息，实现复杂逻辑匹配，在领域知识图谱中找到与查询图中节点（边）近似匹配的节点（边）集合。对于查询图中每个节点（边）在领域知识图谱中会找到零到多个对应节点，每个对应关系都拥有一个0到1的相似度值。

3、图结构匹配功能

实现基于“查询图”和“匹配节点”集合到“匹配子图”集合的转换。图结构匹配在领域知识图谱中找到包含“匹配节点”集合节点的若干联通子图，再将子图与查询图进行匹配，得到相似度，然后按照子图相似度进行排序，找到领域知识图谱中与查询图匹配度高的子图。

4、实体摘要功能

实现“匹配子图”集合到“结果摘要”集合的转换。基于“匹配子图”可以得到符合查询意图的实体集合。实体拥有众多属性、关联着大量其它实体，需要对这些信息进行裁剪、计算，以简洁、符合用户意图的方式展现出来。

## 主要特性

|  |  |
| --- | --- |
| 主要特性 | 详细描述 |
| 支持实体各种展现方式的定制 | 支持通过数据检索定制工具实现本体定义中的各类实体的列表视图、详细信息视图等的展现方式的定义 |
| 支持多种检索方式 | 关键字检索：通过输入关键字，匹配相应的实体、关系或属性  自然语言检索：通过将输入的自然语言解析成数据检索条件，查询过滤知识库中的业务数据  条件检索：通过类似电商的查询方式，简单点选后获取满足条件的数据信息 |
| 返回结果智能排序 | 当检索的结果是一类实体的集合时，根据实体的匹配度自动排序结果，返回用户最关心的数据 |
| 丰富的可视化组件 | 信息页签：实体信息以页签的方式展示，可以根据内容的类型和文字数量自动以最优布局展示  统计图形：支持多种统计图形如饼图、折线图、雷达图等方式展现统计和分析数据  时间线：支持以时间线的方式展现实体间的时间相关度  地图：支持在地图上展示带有地理位置信息的实体和关系数据 |
| 强大的全文检索功能 | 通过输入关键字，在建立全文检索的属性中进行匹配，按匹配度返回最优结果  在全文检索的结果中高亮显示关键字信息  只需通过简单配置即可实现全文索引属性的修改 |
| 高效的检索效率 | 近实时的检索能力，千万级的数据可秒级响应  可支持在线平滑扩展系统部署结构，通过分布式提升系统并行处理能力 |
|  |  |

## 设计方案

## 扩展性设计

# 图探索工具

## 需求分析

为了能够对存储的知识数据进行宏观层面的探索以及分析，平台应集成图关联探索引擎，支持用户对数据以图数据结构的形式进行关联搜索，主要包括：

* 支持用户以统计图表的形式浏览查询图数据结构的基本信息；
* 支持用户通过常规图数据结构算法对图数据结构进行基本的分析；
* 支持用户通过常规关联分析算法对图数据结构进行分析，包括但不限于聚类、社区发现等算法；
* 支持用户通过自定义方式（脚本、程序等）对图数据结构进行分析；
* 以上功能应达到如下具体要求：
* 关于图数据结构基本信息的指标展示达到20项以上；
* 关于图数据结构的基本分析算法达到10种以上；
* 关于图数据结构的关联分析算法达到5种以上；
* 支持的自定义分析语言/方法达到5种以上；

## 主要特性

|  |  |
| --- | --- |
| 主要特性 | 详细描述 |
|  |  |
|  |  |

## 设计方案

## 扩展性设计

# 交互式分析工具

## 需求分析

为了能够对存储的知识数据进行细节层面的探索和分析，平台应集成交互分析工具，支持用户对数据以前端交互的形式进行探索，主要包括：

* 支持用户以近似平面图的形式直观的查看其请求的图数据结构；
* 支持用户对图中结点进行拓展，了解该结点相关联数据的细节信息；
* 支持用户以基于结点/连边属性的方式对数据进行筛选；
* 支持用户以基于时间轴的方式对数据进行筛选；
* 以上功能应达到如下具体要求：
* 平面图绘制时应以不存在边交叉为主要原则，在无法达成时，交叉的边数应不超过总边数的0.1倍；
* 平面图绘制应支持超过1000个元素（点、边）以上的同时展示；
* 筛选的条件应包含结点/连边90%以上的相关属性；

# 非功能需求

## 安全需求

1、用户认证方式

2、访问控制方式

3、日志审计

## 性能需求

1、实体节点数量2亿，检索响应时间1秒以内；

2、批量数据加载50条/秒；增量数据更新，150条/秒；

3、索引更新延迟时间小于5秒；

## 管理需求

1、用户管理

支持组织机构管理，包括用户组增加、删除、修改功能；用户增加、删除、修改。

2、权限管理

支持角色管理，包括角色增加、删除、修改功能；角色授权功能；用户角色分配功能。

3、语义检索模型管理

支持词汇表抽取、调整功能；支持匹配模型设置、训练功能。

## 接口需求