## **总体设计目标**

根据养老保险全国统筹工作总体部署，依托人力资源和社会保障服务网络及养老保险全国统筹云平台，构建与养老保险全国统筹相适应的大数据管理应用平台。平台以养老保险全国统筹业务管控等数据资源为基础，整合人社其他相关数据资源，打造集数据抽取、数据处理、数据管理、监测分析、预测分析、模型构建、辅助决策、数据可视化、应用支撑等功能及服务于一体的大数据管理应用平台，以信息化提升养老保险全国统筹数据资源管理与服务能力，支撑人社业务工作开展，打造用数据说话、用数据管理、用数据决策、用数据创新、用数据服务的养老保险全国统筹大数据管理应用模式，更好地实现信息化对养老保险全国统筹及人社相关业务的支撑作用，更好地实现信息化为民、便民服务。

## **系统技术路线规划**

### 基于J2EE体系的B/S/S的三层架构

J2EE平台提供了一个基于组件的方法，进行设计、开发、装配及部署企业应用程序，并提供了多层的分布式应用模型、组件重用、一致化的安全模型以及灵活的事务控制。J2EE平台提供了一个多层的分布式应用模型，将整个应用划分为数据层，业务层和表示层，三层应用体系结构有助于提升应用的伸缩性和扩展性，各层通过接口进行交互，实现动态部署和扩展。

本项目将采用基于J2EE体系的B/S/S三层架构，优先基于人力资源和社会保障应用系统技术支撑框架进行开发，支持与人社应用云技术框架及其支撑平台对接。采取参数化、组件化设计的思想，以保证系统功能的可扩展性与灵活性。

### **采用大数据分布式计算框架**

采用Map/Reduce技术实现分布式计算，分布式计算是通过将一个大的任务划分成多个部分，分别交给多个计算节点进行处理，综合得到最终结果的计算技术，是进行数据计算、数据分析和数据挖掘的有效工具。

分布式计算通过调度批量任务操作静态数据，可用于大规模数据集（大于1TB）的并行运算。可以采用分布式计算架构进行全文搜索、分布式查询、比对、日志分析统计、大规模索引、海量数据排序、词频统计和历史数据挖掘分析等数据研判业务操作。

### **采用分布式列存储的HBase架构**

HBase是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统，利用HBase技术可在廉价PC Server上搭建起大规模结构化存储集群。

hbase介于nosql和RDBMS之间，仅能通过主键(row key)和主键的range来检索数据，仅支持单行事务(可通过hive支持来实现多表join等复杂操作)。主要用来存储非结构化和半结构化的松散数据。Hbase目标主要依靠横向扩展，通过不断增加廉价的商用服务器，来增加计算和存储能力。

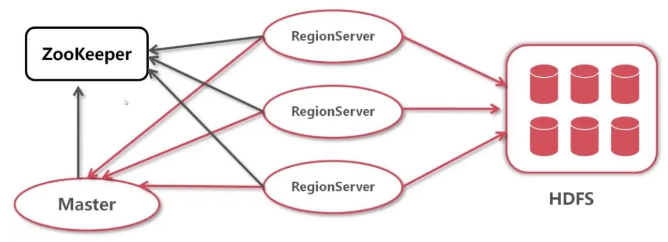
HBase拥有以下的特点：

1 大：一个表可以有上亿行，上百万列

2 面向列:面向列(族)的存储和权限控制，列(族)独立检索。

3 稀疏:对于为空(null)的列，并不占用存储空间，因此，表可以设计的非常稀疏。

HBase的架构如下：

说明：

Zookeeper是分布式的协调。RegionServer将自己的信息写到ZooKeeper中。

HDFS是Hbase运行的底层存储文件系统。

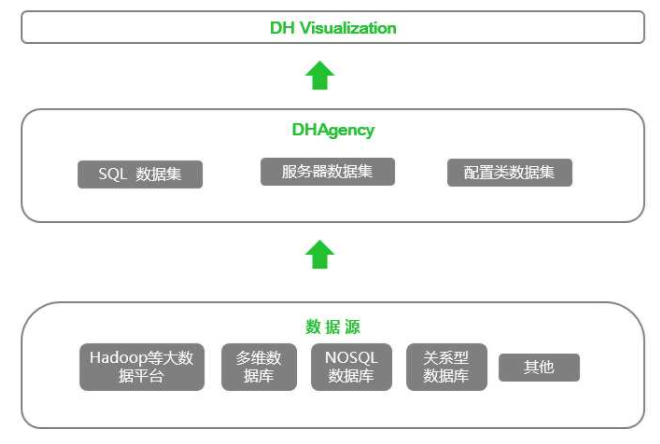
RegionServer，作为数据节点，用来存储数据。

Master RegionServer需实时的向Master报告信息。Master监控全局的RegionServer运行情况，可控制RegionServer的故障转移和Region的切分。

### **基于Hadoop的技术扩展和封装**

围绕Hadoop衍生出相关的大数据技术，应对传统关系型数据库较难处理的数据和场景，例如针对非结构化数据的存储和计算等，充分利用Hadoop开源的优势，伴随相关技术的不断进步，其应用场景也将逐步扩大，目前最为典型的应用场景就是通过扩展和封装Hadoop来实现对互联网大数据存储、分析的支撑。对于非结构、半结构化数据处理、复杂的ETL流程、复杂的数据挖掘和计算模型，Hadoop平台更擅长。

### **采用可视化分析引擎**



可视化引擎支持连接企业已有的大数据计算平台，如 Hadoop，Kylin，Greenplum，Elasticsearch 等。与此同时，借助这些大数据计算平台的性能，也解决了 TB 到 PB 级超大数据量多维分析的难题。

DHAgency 是 DataHunter 推出的大数据可视化利器，旨在帮助企业用户更好地处理超大数据量的分析，满足企业对数据源实时性的需求。

通过 DHAgency 可视化引擎，可以直接对接现有的数据源，无论关系型数据库（Mysql，Oracle，Postgresql 等），还是 Hadoop 相关大数据平台、NOSQL 数据库、MPP 架构的解决方案，都可自助完成分析并展现，最终做出更敏捷的、更及时的分析和决策。

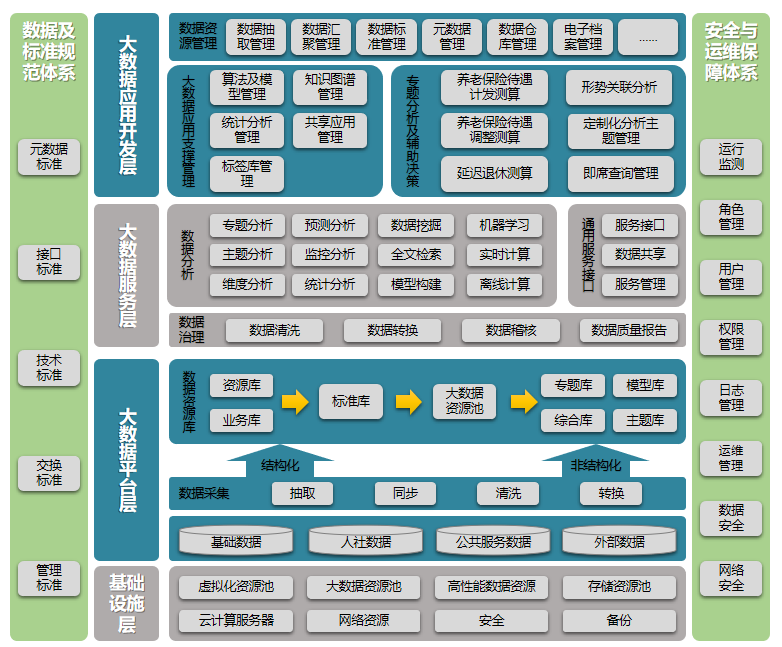
### **支持国内外主流中间件、数据库、操作系统、及浏览器**

为了保障本项目的信息系统从硬件到软件的自主研发、生产、升级、维护的全程可控，同时可对信息和信息系统实施安全监控管理，防止非法利用信息和信息系统，实现信息安全。本项目的建设所使用的中间件产品使用开放的体系标准，兼容国内外中间件的主流产品，不依赖某些特定的服务器、操作系统、数据库或者中间件。后台数据库兼容国内外数据库的主流产品，支持国产大数据云平台相关数据库存储产品，支持主流国产数据库产品，支持MPP架构数据库及HADOOP架构数据库。服务器端支持的操作系统包括Windows2008（或更高版本）、UNIX和Linux等；客户端支持 WindowsXP、Windows7及更高版本，或国产桌面操作系统；浏览器支持IE、火狐、Chrome等主流浏览器及相应版本，并支持国产浏览器。

## **总体架构设计**

### 总体架构

养老保险全国统筹信息化建设项目大数据管理应用平台遵循信息资源服务平台建设要求来设计总体架构图，总体架构图基于标准大数据平台的四层架构设计理念，主要包括基础设施层、大数据平台层、大数据服务层和大数据应用开发层四部分，以及数据标准规范体系和安全与运维保障体系两套支撑体系，具体如下图所示：



1、基础设施层

本项目大数据管理应用平台所需的基础环境，采用养老全国统筹已采购云平台等大数据、虚拟化技术、网络资源或网络安全，实现数据动态的资源分配与管理。

2、大数据平台层

大数据平台层从资源库和业务库获取数据，通过智能的数据采集工具ETL，将养老全国统筹相关的基础数据、人社数据、公共服务数据和外部数据进行采集，经过数据抽取、同步、清洗与转换，形成标准的数据资源库，并落地大数据资源池中，最终形成本项目所需的各种专题库、综合库、模型库和主题库。同时可对各数据库数据资源进行台账式管理，协助管理者实时掌控资源的构成、分布以及动态变化。

3、大数据服务层

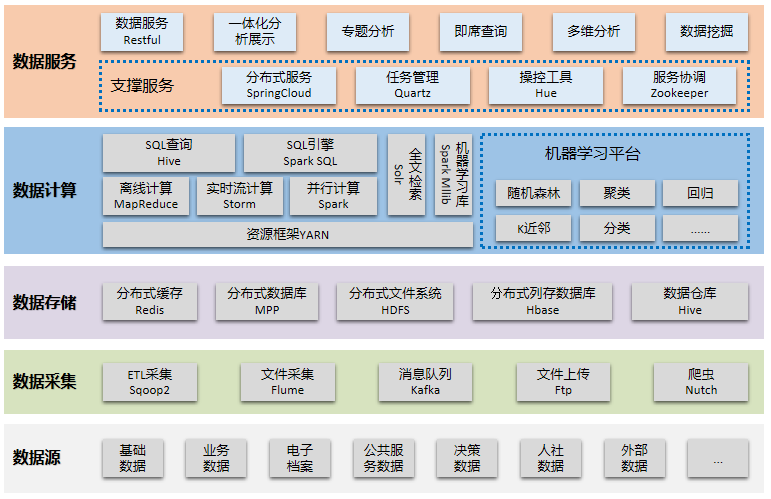
在大数据服务层，本项目将构建数据治理服务、数据分析服务和通用服务接口。数据治理提供数据清洗、数据转换、数据稽核以及数据质量报告等功能，依托数据及标准规范体系，为数据治理中台的建设提供支撑；数据分析提供一系列的数据分析、模型构建以及机器学习能力，为养老全国统筹大数据管理应用平台提供业务预测、风险防控、数据分析、模型构建和BI报表等服务能力；通用接口服务提供服务接口、数据共享以及服务管理能力，将汇聚管理的人社业务系统数据统一管理、分析及利用，支撑全国统筹业务工作，同时通过大数据管理应用平台向人社业务协同平台、公共服务平台等进行数据或服务输出。

4、大数据应用开发层

在大数据应用开发层，将构建数据资源管理、大数据应用支撑管理和专题分析及辅助决策三大应用。数据资源管理包括数据源及数据抽取管理、数据切片及分层汇聚管理、数据标准管理、数据质量管理、元数据管理、数据过程管理、数据仓库管理、人社电子档案袋管理、数据安全管理，为养老保险全国统筹等相关业务提供数据支撑；大数据应用支撑管理包括算法及模型管理、统计分析管理、标签库管理、知识图谱管理、共享应用管理，对养老保险全国统筹业务管控等相关系统提供服务支撑；专题分析及辅助决策包括养老保险待遇计发测算、养老保险待遇调整测算、延迟退休测算、社会保险与就业、失业形势关联分析、定制化分析主题管理、即席查询管理，支持养老保险全国统筹业务预测、风险防控等。

### 系统技术架构

基于本次养老保险全国统筹总体的科学规划、统筹资源的大前提下进行本项目大数据管理应用平台的合理性设计，以养老保险全国统筹业务管控等数据资源为基础，构建大数据应用中心平台。本项目技术架构主要围绕数据采集、数据存储、数据计算及数据服务几个方面进行技术选型与设计。总体技术架构如下：



在数据源层，主要包括养老全国统筹数据、人社其他业务数据、人社业务决策数据、外部其他部委数据，统一构成本项目大数据平台的数据基础。

在数据采集层，提供了ETL采集Sqoop2、文件采集Flume、消息队列采集Kafka、文件上传采集Ftp、爬虫抓取Nutch等技术手段，提供丰富多样的采集工具，满足养老全国统筹的业务应用需要。

在数据存储层，大数据存储采用Hbase分布式列式数据库、Hive数据仓库工具、MPP分布式数据库、Redis分布式缓存数据库、HDFS分布式文件系统相结合的方式进行设计，满足养老全国统筹的业务应用场景的存储需求。

在数据计算层，是基于数据存储的基础，在资源框架YARN的统一管理下，通过离线计算MapReduce、实时计算Storm、并行计算Spark等计算技术，提供即席查询能力，如SQL实时查询Hive、Spark SQL实现，精确查询采用Hbase库实现，批处理采用MapReduce实现，全文检索采用ES/Solr实现。在数据计算层提供机器学习平台，使用机器学习库Spark Mllib，提供随机森林、分类算法、聚类算法、回归算法等常用的算法。

数据服务提供支撑服务，采用分布式服务SpringCloud、任务管理Quartz、操控工具Hue以及服务协调Zookeeper等工具，支撑大数据服务能力。通过数据封装和服务总线以接口访问的方式提供服务，访问方式支持restful、Dubbo等多种访问方式；应用多种数据展示技术，实现一体化分析展示、专题分析、即席查询、多维分析、数据挖掘等功能的展示。

#### 数据流架构



人社部全国养老保险全国统筹大数据管理应用平台的数据从采集到最终提供服务，历经接入、采集、清洗、转换、存储、融合、服务应用等多个阶段：

数据接入和采集：养老保险全国统筹业务经办、风险管控数据资源、养老保险全国统筹业务管控系统、基础信息库以及与养老保险全国统筹相关的其他人社业务系统的数据资源，通过不同的方式接入和交换到大数据管理应用平台，平台具有丰富的适配和采集能力，支持实时数据采集、批量数据采集、日志数据采集和应用数据采集等多种方式，通过分布式消息队列将数据传输到数据中心。

数据清洗、转换和存储：大数据管理应用平台提供大数据中心作为平台数据存储中心，存储从采集到服务过程中各阶段的数据，包括汇集数据、基础数据、融合数据、统计分析数据等，随着人社数据的不断采集，大数据中心数据会持续扩充。并对数据按照数据标准管理和数据质量管理进行数据的清洗、转换，保障数据的标准可用。

数据融合：采用元数据管理、ETL、BETL（大数据ETL）、数据融合、数据质量管理、数据标准管理等多类组件构建，是平台数据处理的核心。

数据服务应用：经过数据融合处理后的数据，可以通过多种途径对外共享提供服务，通过平台开放的数据文件接口、数据库接口以及API接口等多种方式为人社部业务系统上层应用提供数据支撑。通过数据共享交换平台提供数据资源的内外部共享服务。

# 应用功能设计

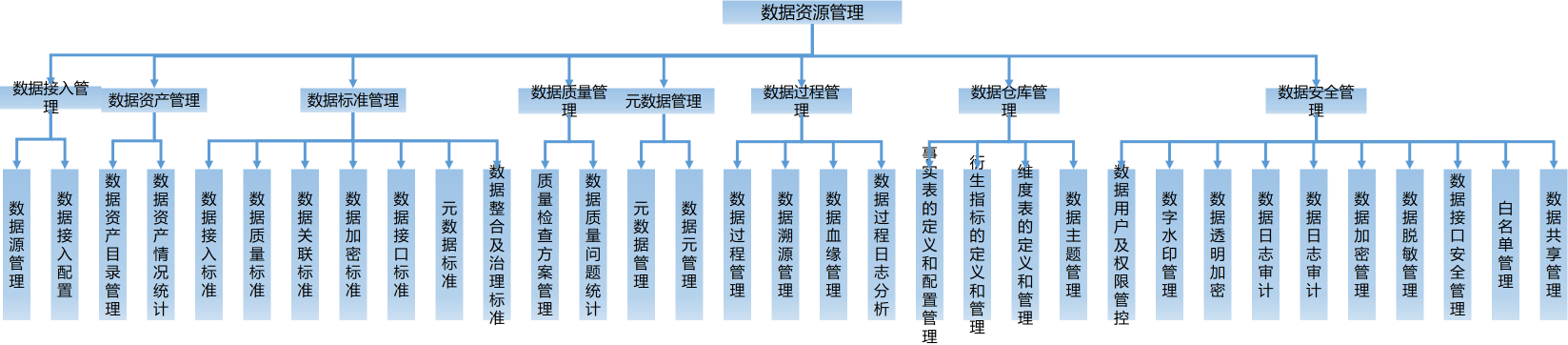
遵循养老保险全国统筹信息化建设总体规划和部署原则，依托人力资源和社会保障服务网及养老保险全国统筹云平台，构建与养老保险全国统筹相适应的大数据管理应用平台。本项目建设，我们将做好全面、完整的设计，以满足养老保险全国统筹大数据管理应用平台项目业务要求和技术需求。平台整体建设将以养老保险全国统筹业务管控等数据资源为基础，整合人社其他相关数据资源，实现对各类业务信息的归集和整合，打造集数据抽取、数据处理、监测分析、模型构建、辅助决策、数据可视化、应用支撑等功能与服务于一体的大数据管理体系，全面提升养老保险全国统筹数据资源管理和服务能力，更强有力地支撑养老保险统筹业务办理，为参保人提供更便捷、更人性化的服务。

下面将从应用功能设计、重点业务功能设计、应用系统衔接设计三个方面进行详细说明。

## **应用功能设计方案**

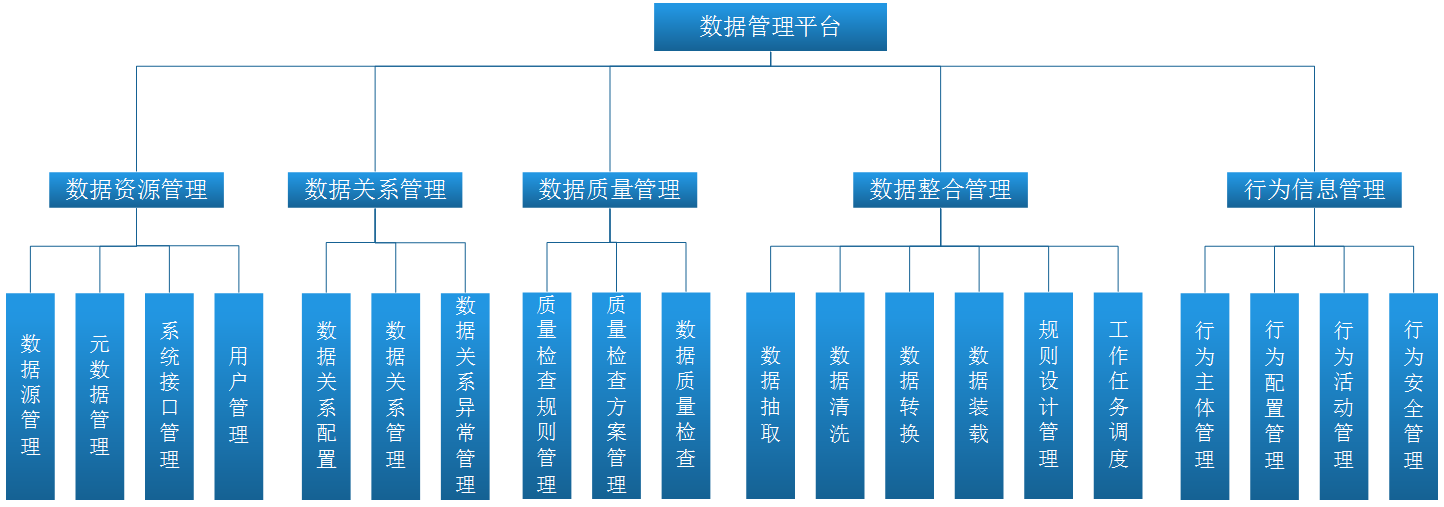
### 数据资源管理应用功能设计

数据资源管理平台实现对养老保险全国统筹业务经办、风险管控数据资源及与养老保险全国统筹相关的其他人社数据资源的统一管理，包括数据接入管理、数据资产管理、数据标准管理、数据质量管理、元数据管理、数据过程管理、数据仓库管理、数据安全管理等，并能够支持各项数据管理工作的流程、日志、结果等信息的可视化展示与可视化配置，支持对数据管理工作进行统一调度、监控、问题预警及问题排查等，总体功能结构示意图如下：



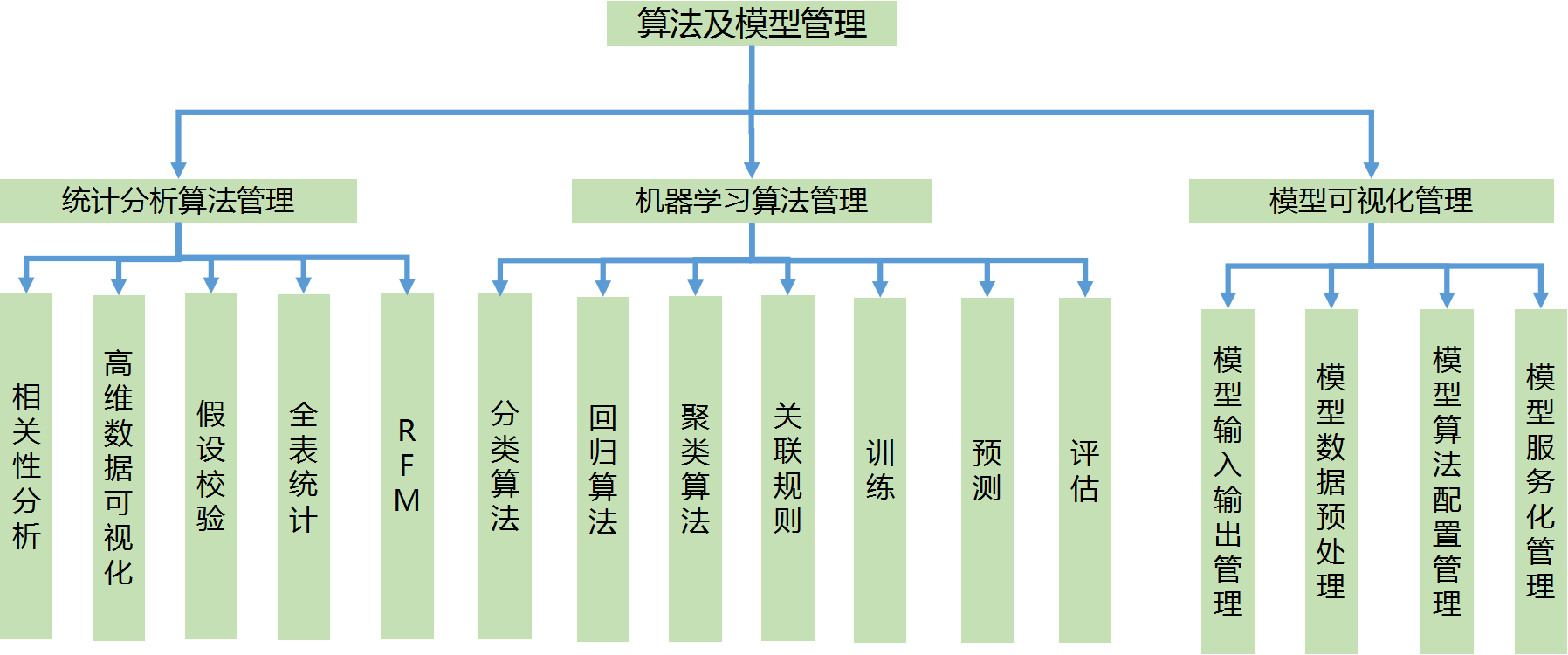
### 大数据应用支撑管理应用功能设计

大数据应用支撑管理功能包括：算法及模型管理、报表及BI展示管理、标签库管理、知识图谱管理、共享应用管理，总体功能结构示意图如下：



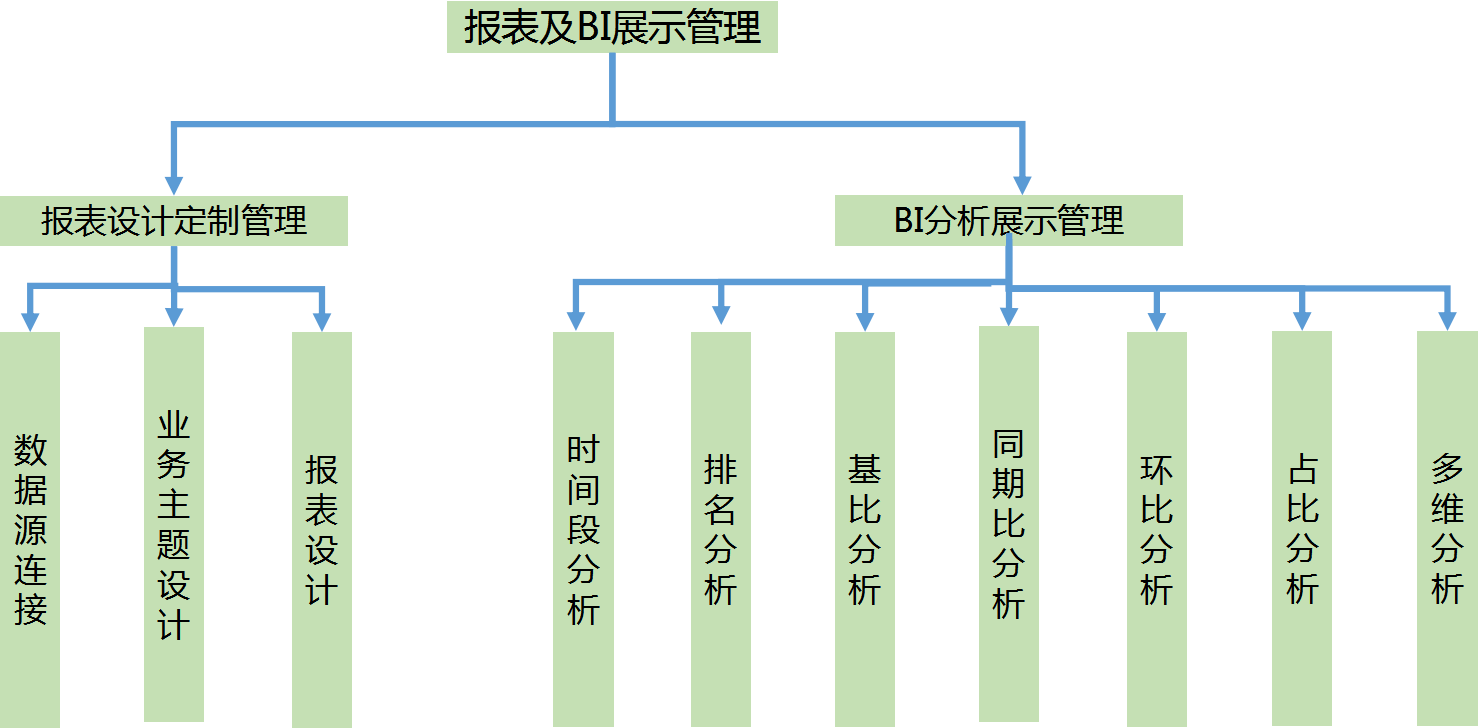
#### 算法及模型管理

算法及模型管理是统一对大数据平台使用的基本统计分析算法、机器学习算法（如分类、聚类、逻辑回归、神经网络等）进行管理和配置，支持算法的更新、修改、新增及删除等。能够将相关算法与业务数据相结合，构建模型，并对建模过程进行可视化配置和统一管理，支持联邦学习，支持对模型进行统一管理和展示。具体功能模块如下图所示：



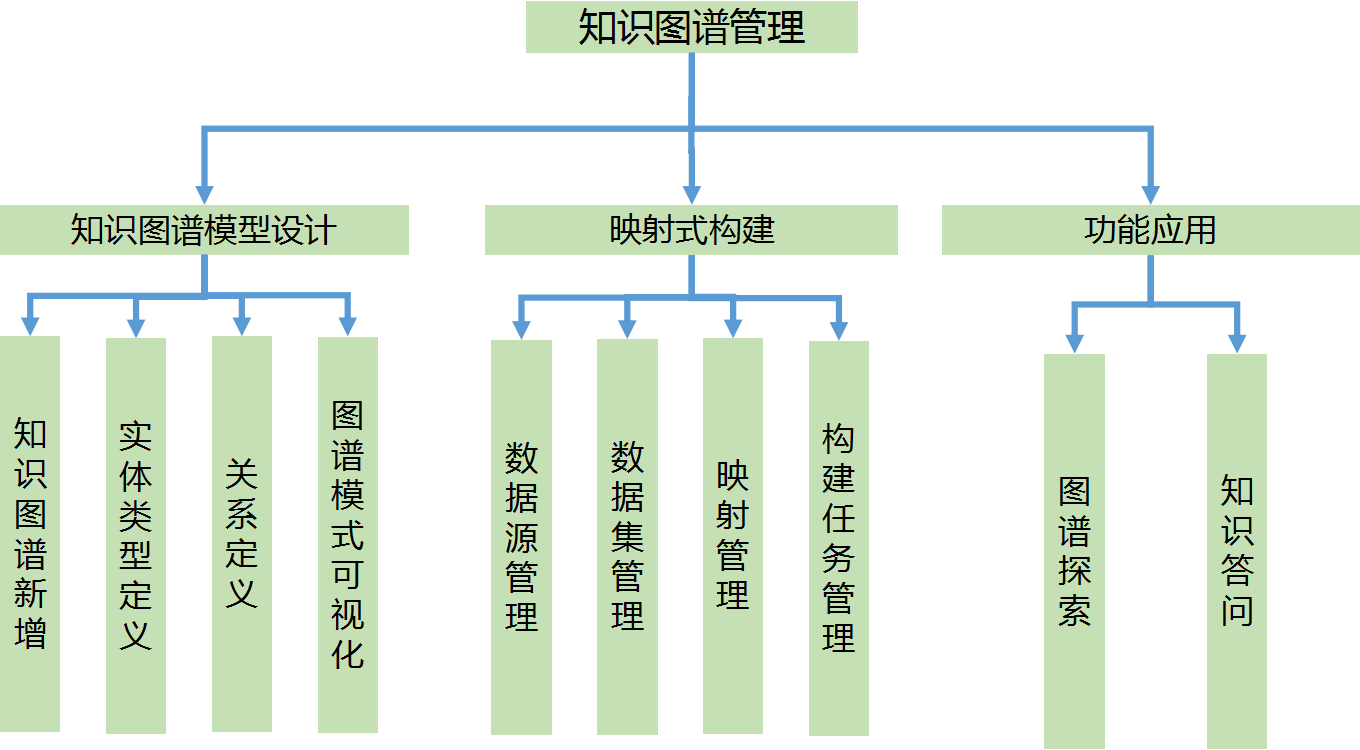
#### 报表及BI展示管理

统计分析管理是支持对养老保险全国统筹及相关业务数据，进行各类数据统计报表、动态图文报告的规划、制定及配置。支持按日、月度、季度、半年度、年度等时间维度定制，支持统计报表及动态图文报告数据的同比、环比及对比分析，支持按时间等维度进行趋势化分析和多维分析，支持在线编辑模板，支持数据统计报表及动态图文报告的设计、导入、统计、汇总及多形式多类型分析展现，支持基于大屏的制作和展现，支持灵活定制及扩展，支持在统一的应用安全管控范围内集成到相关系统中实现共享应用。支持数据统计报表及动态图文报告的自动定时刷新等。具体功能模块如下图所示：



#### 知识图谱管理

知识图谱管理是支持知识库的可视化配置管理，支持知识库新增、修改、删除等管理，支持知识的定义、抽取、关联、映射等，支持知识图谱的可视化构建过程，支持知识图谱的查询、预览、编辑、导出、检索等。具体功能模块如下图所示：



### 专题分析及辅助决策应用功能设计

#### 养老保险待遇计发测算

养老保险待遇计发测算是支持基于各地上报的待遇项目和计发办法，结合养老保险全国统筹数据，对待遇水平进行预测、核查及计算评估，支持养老保险全国统筹待遇风险防控。具体功能模块如下图所示：



#### 人社档案袋管理

人社档案袋管理是构建基于持卡库、单位库、职工养老保险为主，其他社会保险及相关人社业务数据辅助的人社个人电子档案袋、单位电子档案袋，整合个人及单位信息，形成统一的单位及个人业务组合模式，并为相关系统提供支撑和服务等。

#### 养老保险待遇调整测算

养老保险待遇调整测算是基于现有业务规则、生命表及相关信息，构建养老保险待遇调整测算模型，支持待遇调整测算，为业务决策提供模型支持。

#### 延迟退休测算

延迟退休测算是对职工养老保险延迟退休产生的社会、基金、参保人员、待遇人员等的影响进行深入分析，构建延迟退休测算模型，形成可供决策使用的支撑能力。具体功能模块如下图所示：



#### 社会保险与就业、失业形势关联分析

社会保险与就业、失业形势关联分析是基于养老保险全国统筹数据，结合人社就业、失业等其他业务数据及国家宏观经济数据，研究分析社会保险与就业、失业的关联关系，研究分析新业态从业人员，灵活就业人员、农民工等重点群体参保、就业、失业等情况，研究分析劳动力结构变化情况，对参保、就业、失业等的影响，形成社会保险与就业、失业关系模型，为业务提供支撑和服务。具体功能模块如下图所示：



#### 定制化分析主题管理

定制化分析主题管理是支持养老保险全国统筹及人社其他相关业务，开展定制化的主题分析，建立分析及模型框架，支持领导视图、监测预警、预测分析等，支持多维数据分析，支持汇总、下钻、上卷等，支持对分析主题及对象的标签化查询、管理等，支持分析结果按权限导出。具体功能模块如下图所示：



##### 领导视图

1. 功能设计原型如下：

