# 创智数据挖掘平台软件架构描述

### 1、目的

文档使用软件架构图描述了创智数据挖掘平台的软件架构，并在后面描述了各模块的主要功能。

### 2、软件架构

创智数据挖掘平台采用MVC设计模式，架构图如下：



### 3、模块功能描述

平台作为一个提供公共挖掘算法服务的平台，与消费服务的客户端之间仅通过消息进行通讯。

平台通过北向接口与客户端进行通讯，北向接口支持用户数据上传、数据挖掘、挖掘结果查询和用户管理四类接口。

消息队列用于保存北向接口收到的数据挖掘、挖掘结果查询消息，北向接口收到这两类消息后会先将消息保存到消息队列里，并返回客户端消息已经受理。队列里的消息会由流程控制模块来读取。

数据汇聚模块用于接收从客户端上传到服端的数据，并通过数据预处理模块将数据转换成Key-Value格式后 保存到数据仓库里的HDFS中。数据汇聚同时还提供用户删除数据挖掘结果的服务。

数据预处理模块主要有两个功能：1、用于将客户端上传的数据转换后保存到数据仓库中存储；2、为数据挖掘模块从数据仓库中准备适用于算法的数据，并将这些临时数据保存在HDFS中。

数据挖掘模块实现数据挖掘功能，并将挖掘结果保存到数据集市中以便数据查询模块查询。数据挖掘算法过程分为两步：1、启动挖掘算法后调用数据预处理模块准备算法运行所需的数据，并将准备好的数据保存到HDFS中；2、调用数据挖掘算法进行运算，预算结果保存在数据集市中用于查询。

数据查询模块实现在数据集市里的挖掘结果中查询需要的数据。

预处理算法库用于向数据汇聚模块和数据预处理模块提供数据预处理算法。

挖掘算法库用于向数据挖掘模块提供数据挖掘算法，Mahout算法库集成在这个模块中，以后扩展新的数据挖掘算法库时，比如Spark，都集成在这个模块里面。

查询算法库用于向数据查询模块提供数据查询算法。

资源调度模块主要用于监控平台运行环境的资源使用情况以及为新任务分配资源。模块分成资源监控和资源分配两个部分。资源监控统一监控整个平台的资源使用情况，资源分配分成四个子模块，分别为数据汇聚、数据预处理、数据挖掘和数据查询模块分配运算所需的资源。

流程控制模块用于控制平台数据处理流程。对于数据汇聚，流程控制模块会先到资源监控模块中查询是否有 足够的资源，如果有足够的资源则分配资源并启动数据汇聚模块接收客户端上传的数据，如果资源不足则拒绝接收数据。对于数据挖掘和数据查询，流程控制模块从消息队列里取到消息后查询资源监控模块，如果有足够的资源则分配资源，然后调用数据挖掘/查询模块进行处理。处理完成后将结果按回调地址返回给客户端。对于用户管理，流程控制直接调用用户/角色管理模块进行处理。

用户/角色管理模块用于用户注册、鉴权等操作，并将鉴权结果返回客户端。鉴权时考虑使用License中心进行鉴权。

控制台是在服务端提供的一个用户界面，便于平台维护人员进行平台资源设置、用户管理等维护操作。

南向接口用于提供平台逻辑层到数据层的接口。

数据层用于保存数据。关系型数据库主要用于保存用户信息。经过预处理（ELT过程）后的数据长期保存到数据仓库中用于挖掘计算，这部分数据物理上保存在HDFS或 HBase中；挖掘时从数据仓库中提取出来用于计算的数据保存到HDFS中（不在数据仓库中）；挖掘结果存放到数据集市中以便查询，这部分数据物理上保存在HDFS、 HBase或关系型数据库中，数据位置保存到元数据中，以区分不同的运算结果。

运行支撑平台为数据挖掘平台提供运行环境，比如云平台。