(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106250987 A
(43)申请公布日 2016.12.21

(21)申请号 201610587879.3

(22)申请日 2016.07.22

(71)申请人 无锡华云数据技术服务有限公司
地址 214000 江苏省无锡市滨湖区科教软件园6号

(72)发明人 许广彬 郑军 张银滨 强亮 周曙刚 段石石

(74)专利代理机构 北京商专永信知识产权代理事务所(普通合伙) 11400
代理人 高之波 储振

(51)Int.Cl.
G06N 99/00(2010.01)

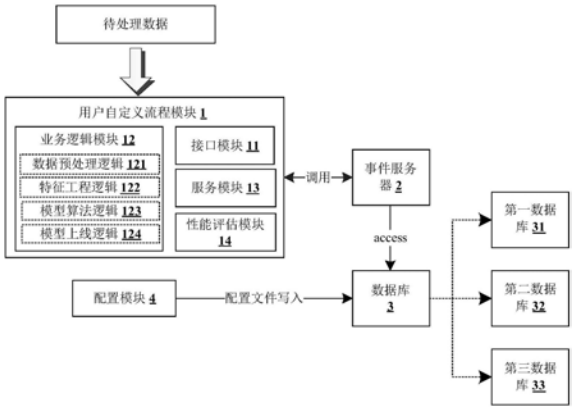
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种机器学习方法、装置及大数据平台

(57)摘要

发明公开了一种机器学习装置，基于机器学习装置的一种机器学习方法，以及运用上述机器学习装置及其机器学习方法的一种大数据平台，该机器学习装置包括：用户自定义流程模块、配置模块、数据库；事件服务器；用户自定义流程模块包含一个逻辑，该逻辑能够接收用户发起的请求所包含的可执行文件，并被事件服务器所调用；数据库通过配置模块所写入的配置文件，将前端开发应用与所述可执行文件进行绑定。通过用户自定义模板来完成业务逻辑的构件，实现了对各种应用场景的适应性与通用性，实现了对标准化大数据开发过程中涉及到的数据挖掘、机器学习的高效运行，简化了标准化大数据的开发流程，提高了标准化大数据的开发部署效率。



CN 106250987 A

1.一种机器学习装置,其特征在于,其包括:

用户自定义流程模块、配置模块、数据库;以及

事件服务器;其中,

用户自定义流程模块包含一个逻辑,该逻辑能够接收用户发起的请求所包含的可执行文件,并被事件服务器所调用;

数据库通过配置模块所写入的配置文件,将前端开发应用与所述可执行文件进行绑定。

2.根据权利要求1所述的机器学习装置,其特征在于,所述用户自定义流程模块包括接口模块、业务逻辑模块、服务模块及性能评估模块。

3.根据权利要求2所述的机器学习装置,其特征在于,所述业务逻辑模块包含对可执行文件执行逻辑操作的至少一种规则,所述规则包括机器学习算法规则、文本数据处理规则、图形用户界面处理规则。

4.根据权利要求3所述的机器学习装置,其特征在于,所述性能评估模块根据该业务逻辑模块中所包含的规则对用户发起的请求所包含的可执行文件获取机器学习算法模型,并根据用户场景在指定模型参数内进行超参数调整优化,以得到机器学习算法模型参数。

5.根据权利要求1所述的机器学习装置,其特征在于,所述业务逻辑模块包含数据预处理逻辑、特征工程逻辑、模型算法逻辑及模型上线逻辑;其中,所述模型上线逻辑包括RESTfull API及数据库存储索引。

6.根据权利要求5所述的机器学习装置,其特征在于,还包括加密模块,其通过访问关键字绑定RESTfull API,以将配置文件与所述可执行文件进行绑定。

7.根据权利要求6所述的机器学习装置,其特征在于,所述访问关键字包括Access key或者Secret key。

8.根据权利要求1所述的机器学习装置,其特征在于,所述数据库通过创建不同的数据表,并按服务请求时间类型将数据库分割成第一个数据库、第二数据库及第三数据库;其中,

所述第一数据库,用于存储元数据;

所述第二数据库,用于存储事件类型、配置参数、模型训练参数;

所述第三数据库,用于存储完成训练的模型。

9.根据权利要求1所述的机器学习装置,其特征在于,所述数据库支持Hbase交互模式、Elasticsearch交互模式或者Mysql交互模式。

10.根据权利要求1所述的机器学习装置,其特征在于,所述事件服务器包括导入引擎、处理引擎、模型训练引擎及服务提供引擎。

11.根据权利要求1至10中任意一项所述的机器学习装置,其特征在于,所述可执行文件包括可执行程序、计算机组件、系统插件、可视化界面应用或者计算机可执行文档。

12.一种机器学习方法,其特征在于,包括以下步骤:

S1、由用户自定义流程模块接收用户发起的请求所包含的可执行文件;

S2、将可执行文件调用至事件服务器;

S3、根据用户的环境变量构建配置文件;

S4、在数据库中根据配置文件的内容,将前端开发应用与所述可执行文件进行绑定。

13.一种大数据平台，其特征在于，包括如权利要求1至10中任意一项所述的机器学习装置以及至少一个平台引擎，所述平台引擎包括spark引擎、tensorflow引擎或者mxnet引擎。

一种机器学习方法、装置及大数据平台

技术领域

[0001] 本发明涉及大数据技术领域，尤其涉及一种机器学习方法、机器学习装置，以及基于该机器学习装置的一种大数据平台。

背景技术

[0002] Spark是Databricks开源的大数据计算处理引擎，于2010年成为Apache顶级项目，其核心计算是弹性分布式数据集(RDD)，提供了比Hadoop更加丰富的MapReduce模型，能够快速在内存中对数据集迭代计算，支持复杂的机器学习算法与图论算法。

[0003] 现有技术中的机器学习方法如下所示。

[0004] 首先，执行步骤1)原始数据的收集：数据生产方会生成多种类型的数据，如log文件、图像数据、文本数据等等，数据质量会随着用户的不适当行为或者系统的一些问题产生很多噪声数据，很难避免错误数据的产生；另外一些多媒体数据如文本、图像还需要一些特别的工具来进行数据载入。

[0005] 然后，执行步骤2)数据预处理：从步骤1)中收集到的数据中含有很多脏数据、无效数据、以及一些多媒体数据，必须经过一系列的措施来进行处理，常采用Hive、MR、以及Spark的preprocess模块来对数据进行去脏处理、缺失值填充等等；针对多媒体数据的处理，会有第三方的工具包用来转换为计算机能够处理的数据，如OpenCV，Word2vec。

[0006] 然后，执行步骤3)特征工程：特征工程主要包括对预处理数据的再次处理、数据的格式化，采样、数据的转换以及特征的设计与选择，常使用MR、Spark features模块以及一些专业的第三方工具来对数据进行处理，处理后输出特征数据，用来做模型的训练。

[0007] 然后，执行步骤4)模型训练：模型训练主要是将实际业务问题通过数据方法来进行建模，常用Spark MLlib、Mahout、以及第三方包如sklearn等等来对数据进行业务建模，训练好的模型进行持久化保存。

[0008] 然后，执行步骤5)模型上线：模型上线主要是讲训练好的模型接入线上数据，来为用户提供常用的分类、回归、推荐等机器学习服务。

[0009] 最后，架构依赖：整个基于spark的机器学习技术方案依赖与HDFS、S3等大数据文件系统，以及包括hadoop、spark在内的各种大数据处理和部署工具。

[0010] 由此可见，现有技术中的基于spark大数据的机器学习方法中的整个架构支撑涉及面广，业务难度大。因此，现有技术中基于spark大数据的机器学习算法模型的技术方案，十分复杂，设计的技术层面覆盖分布式计算、架构部署、模型计算、数据开发等方方面面，花费很大的人力物力才能完成。每一个流程都需要与文件系统频繁地进行互访操作，大大降低整个系统的性能，从而导致建模、预测及应用的可靠性降低；更重要的是会导致编程的灵活性、易维护性、代码或者组件的重用性受到较大影响，因此导致用户体验较差。

发明内容

[0011] 本发明的目的在于公开一种机器学习装置，基于该机器学习装置的一种机器学习

方法，以及运用上述机器学习装置及其机器学习方法的一种大数据平台，用以实现对标准化大数据开发过程中涉及到的数据挖掘、机器学习的高效运行，简化标准化大数据的开发流程，提高标准化大数据的开发部署效率，并提供简洁统一的接口。

[0012] 为实现上述第一个发明目的，本发明提供了一种机器学习装置，其包括：

[0013] 用户自定义流程模块、配置模块、数据库；以及

[0014] 事件服务器；其中，

[0015] 用户自定义流程模块包含一个逻辑，该逻辑能够接收用户发起的请求所包含的可执行文件，并被事件服务器所调用；

[0016] 数据库通过配置模块所写入的配置文件，将前端开发应用与所述可执行文件进行绑定。

[0017] 作为本发明的进一步改进，所述用户自定义流程模块包括接口模块、业务逻辑模块、服务模块及性能评估模块。

[0018] 作为本发明的进一步改进，所述业务逻辑模块包含对可执行文件执行逻辑操作的至少一种规则，所述规则包括机器学习算法规则、文本数据处理规则、图形用户界面处理规则。

[0019] 作为本发明的进一步改进，所述性能评估模块根据该业务逻辑模块中所包含的规则对用户发起的请求所包含的可执行文件获取机器学习算法模型，并根据用户场景在指定模型参数内进行超参数调整优化，以得到机器学习算法模型参数。

[0020] 作为本发明的进一步改进，所述业务逻辑模块包含数据预处理逻辑、特征工程逻辑、模型算法逻辑及模型上线逻辑；其中，所述模型上线逻辑包括RESTfull API及数据库存储索引。

[0021] 作为本发明的进一步改进，该机器学习装置还包括加密模块，其通过访问关键字绑定RESTfull API，以将配置文件与所述可执行文件进行绑定。

[0022] 作为本发明的进一步改进，所述访问关键字包括Access key或者Secret key。

[0023] 作为本发明的进一步改进，所述数据库通过创建不同的数据表，并按服务请求时间类型将数据库分割成第一个数据库、第二数据库及第三数据库；其中，

[0024] 所述第一数据库，用于存储元数据；

[0025] 所述第二数据库，用于存储事件类型、配置参数、模型训练参数；

[0026] 所述第三数据库，用于存储完成训练的模型。

[0027] 作为本发明的进一步改进，所述数据库支持Hbase交互模式、Elasticsearch交互模式或者Mysql交互模式。

[0028] 作为本发明的进一步改进，事件服务器包括导入引擎、处理引擎、模型训练引擎及服务提供引擎。

[0029] 作为本发明的进一步改进，所述可执行文件包括可执行程序、计算机组件、系统插件、可视化界面应用或者计算机可执行文档。

[0030] 为实现上述第二个发明目的，本发明还提供了一种机器学习方法，包括以下步骤：

[0031] S1、由用户自定义流程模块接收用户发起的请求所包含的可执行文件；

[0032] S2、将可执行文件调用至事件服务器；

[0033] S3、根据用户的环境变量构建配置文件；

[0034] S4、在数据库中根据配置文件的内容，将前端开发应用与所述可执行文件进行绑定。

[0035] 为实现上述第三个发明目的，本发明还提供了一种大数据平台，包括上述任意一项机器学习装置以及至少一个平台引擎，所述平台引擎包括spark引擎、tensorflow引擎或者mxnet引擎。

[0036] 与现有技术相比，本发明的有益效果是：通过用户自定义模板来完成业务逻辑的构件，实现了对各种应用场景的适应性与通用性，实现了对标准化大数据开发过程中涉及到的数据挖掘、机器学习的高效运行，简化了标准化大数据的开发流程，提高了标准化大数据的开发部署效率，并能够提供简洁统一的接口，从而使得算法开发、应用开发与构架开发能够实现模块化操作，极大的提高了大数据平台的部署效率及对数据进行挖掘的效率。

附图说明

[0037] 图1为本发明一种机器学习装置的结构图；

[0038] 图2为图1中的机器学习装置中的事件服务器的结构图；

[0039] 图3为本发明一种机器学习装置在一种变形例中的结构图；

[0040] 图4为本发明一种机器学习方法的流程图。

具体实施方式

[0041] 下面结合附图所示的各实施方式对本发明进行详细说明，但应当说明的是，这些实施方式并非对本发明的限制，本领域普通技术人员根据这些实施方式所作的功能、方法、或者结构上的等效变换或替代，均属于本发明的保护范围之内。

[0042] 实施例一：

[0043] 请参图1与图2所示的本发明一种机器学习装置的一种具体实施方式。

[0044] 在本实施方式中，一种机器学习装置，其包括：用户自定义流程模块1、配置模块4、数据库3；以及事件服务器2。用户自定义流程模块1包含一个逻辑，该逻辑能够接收用户发起的请求所包含的可执行文件，并被事件服务器所2调用。数据库3通过配置模块4所写入的配置文件，将前端开发应用与所述可执行文件进行绑定。具体的，该可执行文件包括可执行程序、计算机组件、系统插件、可视化界面应用或者计算机可执行文档。

[0045] 该用户自定义流程模块1包括接口模块11、业务逻辑模块12、服务模块13及性能评估模块14。具体的，参图2所示，在本实施方式中，事件服务器2包括导入引擎21、处理引擎22、模型训练引擎23及服务提供引擎24。导入引擎21负责配置数据源参数、对待处理数据进行读操作/写操作等基本处理，并支持与数据库3之间进行数据交互。处理引擎22负责对待处理数据执行文本数据处理。模型训练引擎23，负责对实际业务问题通过数据方法来进行建模，常用用Spark MLlib、Mahout、以及第三方包如sklearn等等来对数据进行业务建模，训练好的模型进行持久化保存，并保存至第二数据库32中。具体的，模型训练引擎23支持在线模型训练与离线模型训练，从而提高了用户在大数据部署时的便捷性。服务提供引擎24，其接受服务模块13中的模型，并直接通过网络向用户提供在线服务操作。

[0046] 该业务逻辑模块12包含对可执行文件执行逻辑操作的至少一种规则，所述规则包括机器学习算法规则、文本数据处理规则、图形用户界面处理规则。在本实施方式中，业务

逻辑模块12通过加入上述四种逻辑，使得整个用户自定义流程模块1具备了集中化处理业务的逻辑。

[0047] 性能评估模块14根据该业务逻辑模块中所包含的规则对用户发起的请求所包含的可执行文件获取机器学习算法模型，并根据用户场景在指定模型参数内进行超参数调整优化，以得到机器学习算法模型参数。性能评估模块14可让用户按需或者按自己设定的规则输入规则，并用于实现模型的线上部署。

[0048] 业务逻辑模块12包含数据预处理逻辑121、特征工程逻辑122、模型算法逻辑123及模型上线逻辑124；其中，所述模型上线逻辑124包括RESTfull API及数据库存储索引。RESTfull API，一种软件架构接口，其提供了一组设计原则和约束条件。它主要用于客户端和服务器交互类的软件。

[0049] 数据库3通过创建不同的数据表，并按服务请求时间类型将数据库3分割成第一个数据库31、第二数据库32及第三数据库33，其中，第一数据库31，用于存储元数据；第二数据库32，用于存储事件类型、配置参数、模型训练参数；第三数据库33，用于存储完成训练的模型。

[0050] 通过RESTfull API可实现自用户自定义流程模块1所生成的可执行文件，同时也可接收用户或者管理员所发出的数据查询请求，并可在第二数据库32中保存事件服务器2所输出的模型或者服务或者应用。

[0051] 用户或者管理员在构建大数据平台时，可将app、系统插件、程序、图形用户界面(GUI)、文本数据等一切可被计算机读取的数据被用户自定义流程模块1所捕获，并形成用户自定义模板。该用户自定义模板可与配置模块4所导入的配置文件，在数据库3中与前端开发的web应用、app或者服务实现封装，并存储于第三数据库33中，从而为后续服务或者应用提供一体化的大数据服务。前端开发的web应用、app或者服务可通过JAVA、Python、PHP或者Ruby等语言编写而成。

[0052] 具体的，在本实施方式中，该数据库3支持Hbase交互模式、Elasticsearch交互模式或者Mysql交互模式，并优选为Elasticsearch交互模式。Elasticsearch是一个基于Lucene的搜索服务器。它提供了一个分布式多用户能力的全文搜索引擎，基于RESTful web接口。Elasticsearch是用Java开发的，并作为Apache许可条款下的开放源码发布，并可实现分布式全文检索。

[0053] 在本实施方式中，通过用户自定义模板1来完成业务逻辑的构件，实现了对各种应用场景的适应性与通用性，实现了对标准化大数据开发过程中涉及到的数据挖掘、机器学习的高效运行，简化了标准化大数据的开发流程，提高了标准化大数据的开发部署效率，并能够提供简洁统一的接口，从而使得算法开发、应用开发与构架开发能够实现模块化操作，极大的提高了大数据平台的部署效率及对数据进行挖掘的效率。

[0054] 同时，也能够实现后端业务逻辑的适应开发、设计及设定，并通过向事件服务器2提交新建app请求，确定app ID、app NAME、Access Key等相关信息。确定完毕这些信息之后，可直接在数据库3中进行模板的构建与部署，以完成模板与app的绑定；然后通过事件服务器2依次对构建与部署完毕的模板进行编译、训练等操作，从而生成模板。生成后的模板可与前端开发的各种应用、程序、插件等计算机可执行文件通过Access Key进行绑定，实现了业务与架构的分离。

[0055] 在业务部署与开发过程中，业务与架构分离，算法工程师只需关注算法逻辑工作、负责算法逻辑模板的开发工作；应用开发工程师只需参与app、web开发工作，提供数据接入与数据呈现逻辑；架构开发工程师只需关注架构细节，所有业务均有事件触发，事件由业务人员自定义，所有模型相关工作均由配置模板4向数据库3中所写入的配置文件所控制，整个大数据平台分工明确、部署简单，在spark的底层计算与大数据生态工具的帮助下，能大大地减少传统机器学习装置中的数据存储冗余、性能低下、开发流程复杂的难题。

[0056] 实施例二：

[0057] 结合参照图3所示，本实施例与实施例一的主要区别在于，在本实施方式中，该机器学习装置还包括加密模5块，其通过访问关键字绑定RESTfull API，以将配置文件与所述可执行文件进行绑定。优选的，该访问关键字为Access key，也可为Secret key。前端应用通过绑定Access Key绑定RESTful API服务与模型时间服务器6进行交互，完成数据的查询服务。

[0058] 本实施例与实施例一相同的技术方案请参实施例一所述，在此不再赘述。

[0059] 实施例三：

[0060] 参图4所示，本实施例揭示了一种机器学习方法，包括以下步骤：

[0061] S1、由用户自定义流程模块接收用户发起的请求所包含的可执行文件；

[0062] S2、将可执行文件调用至事件服务器；

[0063] S3、根据用户的环境变量构建配置文件；

[0064] S4、在数据库中根据配置文件的内容，将前端开发应用与所述可执行文件进行绑定。

[0065] 实施例四：

[0066] 本实施例公开了一种大数据平台，其包括一个或者多个机器学习装置以及至少一个平台引擎，所述平台引擎包括spark引擎、tensorflow引擎或者mxnet引擎，具体的，该平台引擎根据业务需求选择不同计算引擎，例如在基于图像和视频等多媒体数据服务的场景下，我们提供tensorflow或mxnet引擎来做平台的计算框架，在基于结构化数据的业务场景下，采用spark作为平台计算引擎。

[0067] 本实施例中的机器学习装置配合参照本说明书实施例一和/或实施例二所述。

[0068] Spark是一个开源的数据处理平台，由一组功能强大、高级别的库组成，目前这些库主要包括Spark SQL、Spark Streaming、MLlib、GraphX，支持包括Scala、Java、Python、R在内的API调用，能够与Hadoop生态系统和数据源进行高效集成。

[0069] Spark主要包括结构化数据查询与分析引擎(SparkSQL)、分布式机器学习库(MLlib)、并行图计算框架(GraphX)、流计算框架(Spark Streaming)、第三方子项目(例如BlinkDB、Tachyon、Mesos等)。MLlib是Spark中负责机器学习的组件，常用模块包括Classification、Regression、Clustering、Collaborativefiltering、Frequentpatternmining以及常用的数据预处理和特征工程模块。MLlib是Spark下提供机器学习算法的模块，内置多种机器学习算法。

[0070] Mxnet和tensorflow是深度学习计算工具，常用来构建基于多媒体数据的深度学习框架，其优点在于无需人工设计特征，给定业务需求，构建多层神经网络，通过海量的迭代计算，来挖掘用户感兴趣的多媒体数据需求。常见的业务场景包括安防、关口异常检测、

目标识别等等。

[0071] 上文所列出的一系列的详细说明仅仅是针对本发明的可行性实施方式的具体说明，它们并非用以限制本发明的保护范围，凡未脱离本发明技艺精神所作的等效实施方式或变更均应包含在本发明的保护范围之内。

[0072] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化囊括在本发明内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。

[0073] 此外，应当理解，虽然本说明书按照实施方式加以描述，但并非每个实施方式仅包含一个独立的技术方案，说明书的这种叙述方式仅仅是为清楚起见，本领域技术人员应当将说明书作为一个整体，各实施例中的技术方案也可以经适当组合，形成本领域技术人员可以理解的其他实施方式。

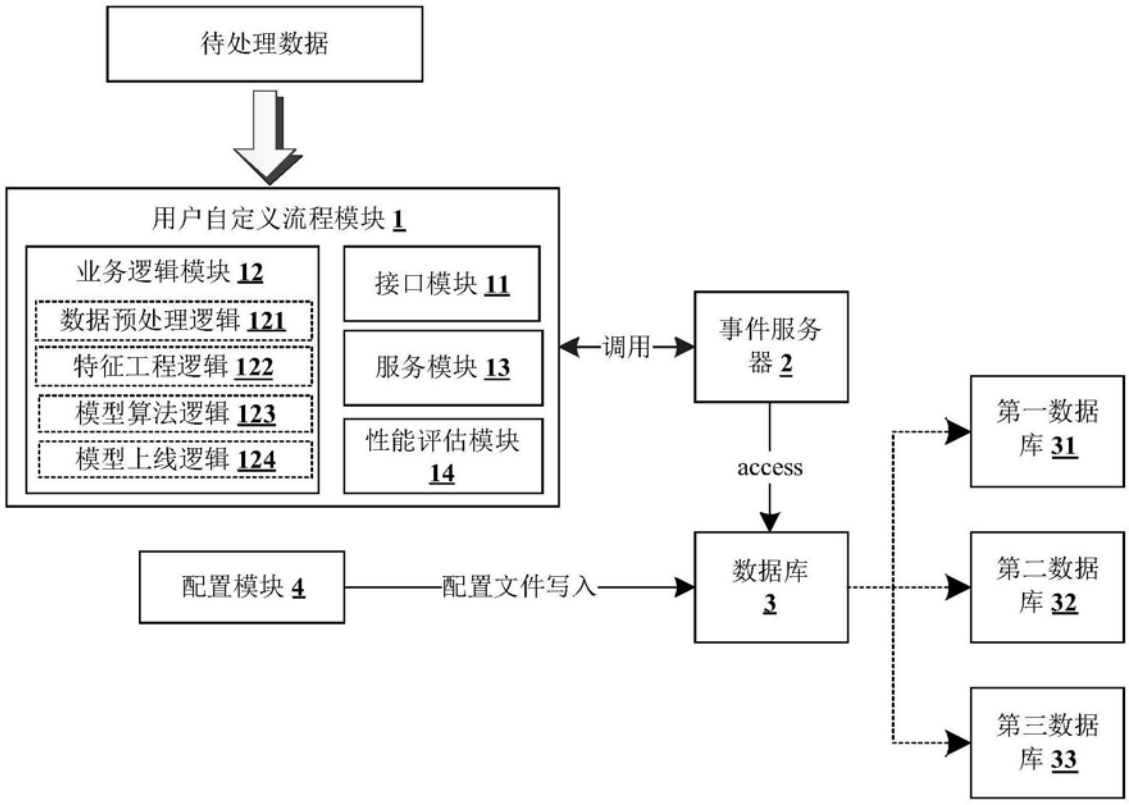

待处理数据
用户自定义流程模块 1
业务逻辑模块 12
数据预处理逻辑 121
特征工程逻辑 122
模型算法逻辑 123
模型上线逻辑 124
接口模块 11
服务模块 13
性能评估模块 14
调用
事件服务器 2
access
数据库 3
配置模块 4
配置文件写入
第一数据库 31
第二数据库 32
第三数据库 33

图1

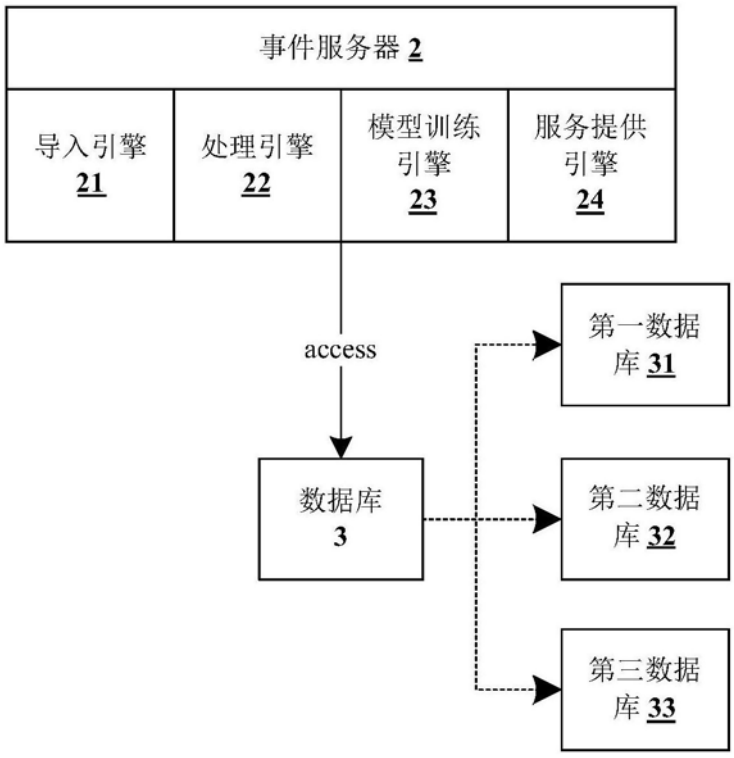

事件服务器 2
导入引擎 21
处理引擎 22
模型训练引擎 23
服务提供引擎 24
access
数据库 3
第一数据库 31
第二数据库 32
第三数据库 33

图2

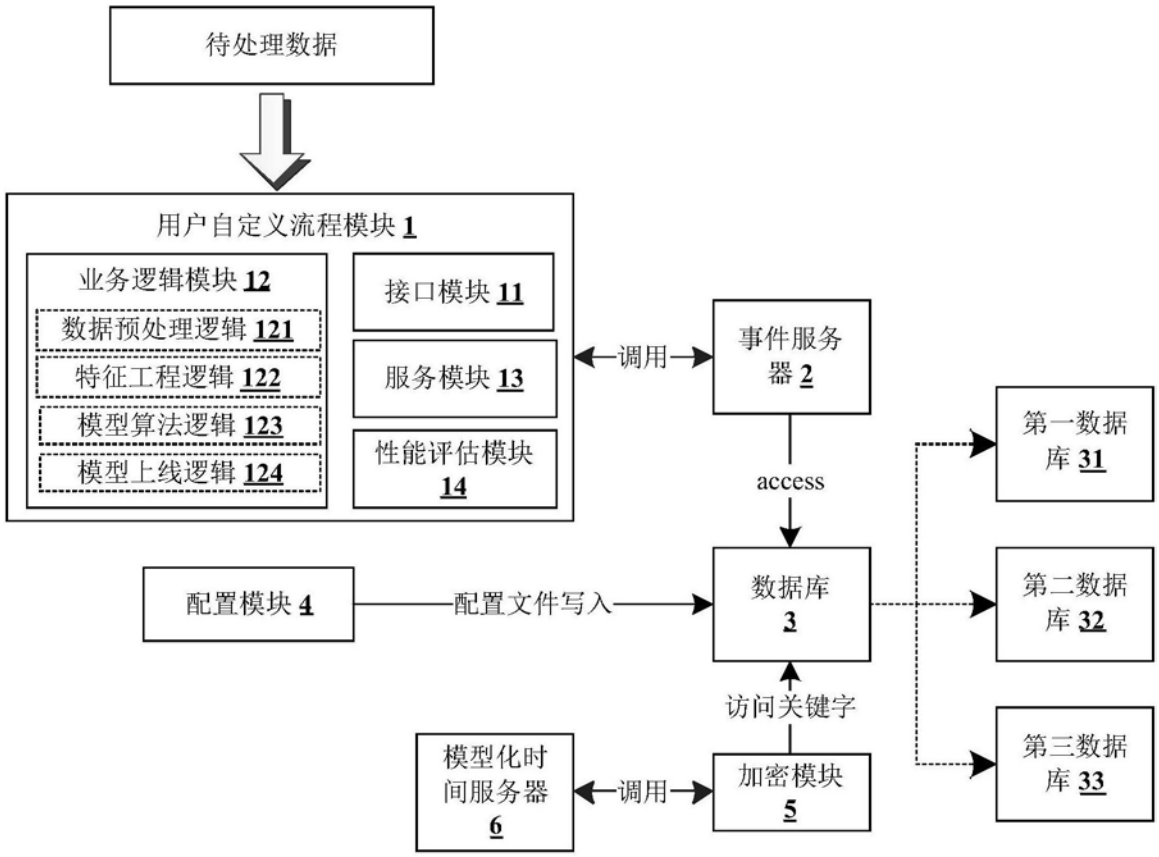

待处理数据
用户自定义流程模块 1
业务逻辑模块 12
数据预处理逻辑 121
特征工程逻辑 122
模型算法逻辑 123
模型上线逻辑 124
接口模块 11
服务模块 13
性能评估模块 14
调用
事件服务器 2
access
第一数据库 31
第二数据库 32
第三数据库 33
配置模块 4
配置文件写入
数据库 3
访问关键字
模型化时间服务器 6
调用
加密模块 5

图3

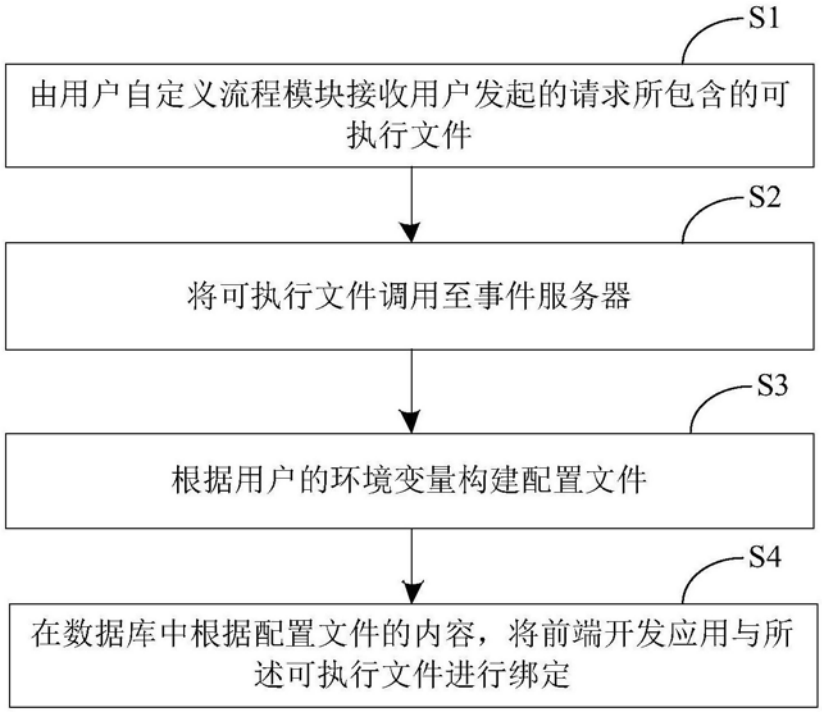

S1
由用户自定义流程模块接收用户发起的请求所包含的可执行文件
S2
将可执行文件调用至事件服务器
S3
根据用户的环境变量构建配置文件
S4
在数据库中根据配置文件的内容，将前端开发应用与所述可执行文件进行绑定

图4