(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 109063842 A
(43)申请公布日 2018.12.21

(21)申请号 201810738131.8

(22)申请日 2018.07.06

(71)申请人 无锡雪浪数制科技有限公司
地址 214000 江苏省无锡市滨湖区金融一街1号昌兴国际金融大厦6楼

(72)发明人 王峰

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 孟金喆

(51)Int.Cl.
G06N 99/00(2010.01)
G06F 9/50(2006.01)
G06F 17/30(2006.01)

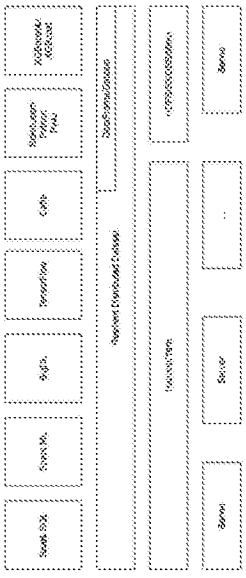
权利要求书1页 说明书3页 附图2页

(54)发明名称

一种兼容多种算法框架的机器学习平台

(57)摘要

本发明公开一种兼容多种算法框架的机器学习平台，该平台通过搭建一个集群兼容所有主流的机器学习框架，统一进行资源调度和用户隔离。主流的机器学习框架搭建在Hadoop+Spark的集群之上，并使用RDD作为数据存储。本发明优点如下：一、利用yarn进行统一的资源管理，因此继承所有yarn的优点；二、使用Spark作为统一底层计算框架，Spark RDD作为统一数据存储，因此继承所有Spark优点；三、调度多种计算框架，支持所有框架的原生优点，包括同步/异步训练，模型/数据并行计算，在线预测等；四、完成了异构计算框架的整合；五、底层支持CPU和GPU。



CN 109063842 A

1.一种兼容多种算法框架的机器学习平台，其特征在于，该平台通过搭建一个集群兼容所有主流的机器学习框架，统一进行资源调度和用户隔离。

2.根据权利要求1所述的兼容多种算法框架的机器学习平台，其特征在于，所述主流的机器学习框架搭建在Hadoop+Spark的集群之上，并使用RDD作为数据存储。

3.根据权利要求2所述的兼容多种算法框架的机器学习平台，其特征在于，所述通过搭建一个集群兼容所有主流的机器学习框架，统一进行资源调度和用户隔离，具体包括：Spark原生支持Spark ML；对于深度并行学习框架TensorFlow、Caffe以及BigDL，在启动每个Spark Executor之后：

启动TensorFlow，Caffe，BigDLworker，并且为数据和事件添加侦听；通过各个框架对应的feed机制，将Spark RDD数据从包括但不限于HDFS在内的存储介质进行读取并发送到TensorFlow，Caffe，BigDL；各个框架进行分布式并行计算，并自动进行模型的跨节点同步；结算结束后，关闭TensorFlow，Caffe，BigDLworker。

4.根据权利要求1至3之一所述的兼容多种算法框架的机器学习平台，其特征在于，对于机器学习框架ScikitLearn，并不支持分布式计算，但是同样可以利用yarn的调度机制，启动单个的Spark Executor；在启动Spark Executor后，将RDD数据转换为ScikitLearn所需数据类型进行单机计算，计算完成后将模型或者数据表写回包括但不限于HDFS在内的存储介质。

一种兼容多种算法框架的机器学习平台

技术领域

[0001] 本发明涉及一种机器学习平台，尤其涉及一种兼容多种算法框架的机器学习平台。

背景技术

[0002] 机器学习已经发展了几十年，在不同时期，为了解决不同的场景的问题，出现了多种机器学习框架，如传统的机器学习框架ScikitLearn，分布式并行计算框架Spark ML，以及深度学习框架TensorFlow，Caffe，Intel BigDL等。在实际的场景中，往往需要同时使用这些异构的计算框架解决具体的问题，来达到最优的效果。如果为每一种计算框架都搭建一个单独的集群，这样做存在较大的硬件资源浪费，开发和运维成本也非常高，并且不利于数据共享。

发明内容

[0003] 本发明的目的在于通过一种兼容多种算法框架的机器学习平台，来解决以上背景技术部分提到的问题。

[0004] 为达此目的，本发明采用以下技术方案：

[0005] 一种兼容多种算法框架的机器学习平台，该平台通过搭建一个集群兼容所有主流的机器学习框架，统一进行资源调度和用户隔离。

[0006] 特别地，所述主流的机器学习框架搭建在Hadoop+Spark的集群之上，并使用RDD作为数据存储。

[0007] 特别地，所述通过搭建一个集群兼容所有主流的机器学习框架，统一进行资源调度和用户隔离，具体包括：Spark原生支持Spark ML；对于深度并行学习框架TensorFlow、Caffe以及BigDL，在启动每个Spark Executor之后：

[0008] 启动TensorFlow，Caffe，BigDLworker，并且为数据和事件添加侦听；通过各个框架对应的feed机制，将Spark RDD数据从包括但不限于HDFS在内的存储介质进行读取并发送到TensorFlow，Caffe，BigDL；各个框架进行分布式并行计算，并自动进行模型的跨节点同步；结算结束后，关闭TensorFlow，Caffe，BigDLworker。

[0009] 特别地，对于机器学习框架ScikitLearn，并不支持分布式计算，但是同样可以利用yarn的调度机制，启动单个的Spark Executor；在启动Spark Executor后，将RDD数据转换为ScikitLearn所需数据类型进行单机计算，计算完成后将模型或者数据表写回包括但不限于HDFS在内的存储介质。

[0010] 本发明提出的兼容多种算法框架的机器学习平台优点如下：一、利用yarn进行统一的资源管理，因此继承所有yarn的优点；二、使用Spark作为统一底层计算框架，Spark RDD作为统一数据存储，因此继承所有Spark优点；三、调度多种计算框架，支持所有框架的原生优点，包括同步/异步训练，模型/数据并行计算，在线预测等；四、完成了异构计算框架的整合；五、底层支持CPU和GPU。

附图说明

[0011] 图1为本发明实施例提供的兼容多种算法框架的机器学习平台示意图；

[0012] 图2为本发明实施例提供的兼容多种算法框架的机器学习平台的工作原理示意图。

具体实施方式

[0013] 下面结合附图和实施例对本发明作进一步说明。可以理解的是，此处所描述的具体实施例仅仅用于解释本发明，而非对本发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非全部内容。除非另有定义，本文所使用的所有的技术和科学术语与属于本发明的技术领域的技术人员通常理解的含义相同。本文中在本发明的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本发明。本文所使用的术语“及/或”包括一个或多个相关的所列项目的任意的和所有的组合。

[0014] 本实施例中兼容多种算法框架的机器学习平台通过搭建一个集群兼容所有主流的机器学习框架，统一进行资源调度和用户隔离。如图1所示，在本实施例中所述主流的机器学习框架搭建在Hadoop+Spark的集群之上，并使用RDD作为数据存储。需要说明的是，图中各英文名词均为计算机领域的惯用技术术语，在计算机领域的释义唯一，对本领域的普通技术人员来说无任何异议，因此，在此就不再赘述。

[0015] 具体的，如图2所示，通过搭建一个集群兼容所有主流的机器学习框架，统一进行资源调度和用户隔离，具体包括：Spark原生支持Spark ML；对于深度并行学习框架TensorFlow、Caffe以及BigDL，在启动每个Spark Executor之后：启动TensorFlow，Caffe，BigDLworker，并且为数据和事件添加侦听；通过各个框架对应的feed机制，将Spark RDD数据从包括但不限于HDFS在内的存储介质进行读取并发送到TensorFlow，Caffe，BigDL；各个框架进行分布式并行计算，并自动进行模型的跨节点同步；结算结束后，关闭TensorFlow，Caffe，BigDLworker。其中，图中各英文名词均为计算机领域的惯用技术术语，在计算机领域的释义唯一，对本领域的普通技术人员来说无任何异议，因此，在此就不再赘述。

[0016] 值得一提的是，对于机器学习框架ScikitLearn，并不支持分布式计算，但是同样可以利用yarn的调度机制，启动单个的Spark Executor；在启动SparkExecutor后，将RDD数据转换为ScikitLearn所需数据类型进行单机计算，计算完成后将模型或者数据表写回包括但不限于HDFS在内的存储介质。

[0017] 本发明提出的技术方案优点如下：一、利用yarn进行统一的资源管理，因此继承所有yarn的优点；二、使用Spark作为统一底层计算框架，Spark RDD作为统一数据存储，因此继承所有Spark优点；三、调度多种计算框架，支持所有框架的原生优点，包括同步/异步训练，模型/数据并行计算，在线预测等；四、完成了异构计算框架的整合；五、底层支持CPU和GPU。

[0018] 本领域普通技术人员可以理解实现上述实施例中的全部部分是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，所述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory，ROM)或随机存储记忆体(Random Access Memory，RAM)

等。

[0019] 注意，上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解，本发明不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此，虽然通过以上实施例对本发明进行了较为详细的说明，但是本发明不仅仅限于以上实施例，在不脱离本发明构思的情况下，还可以包括更多其他等效实施例，而本发明的范围由所附的权利要求范围决定。

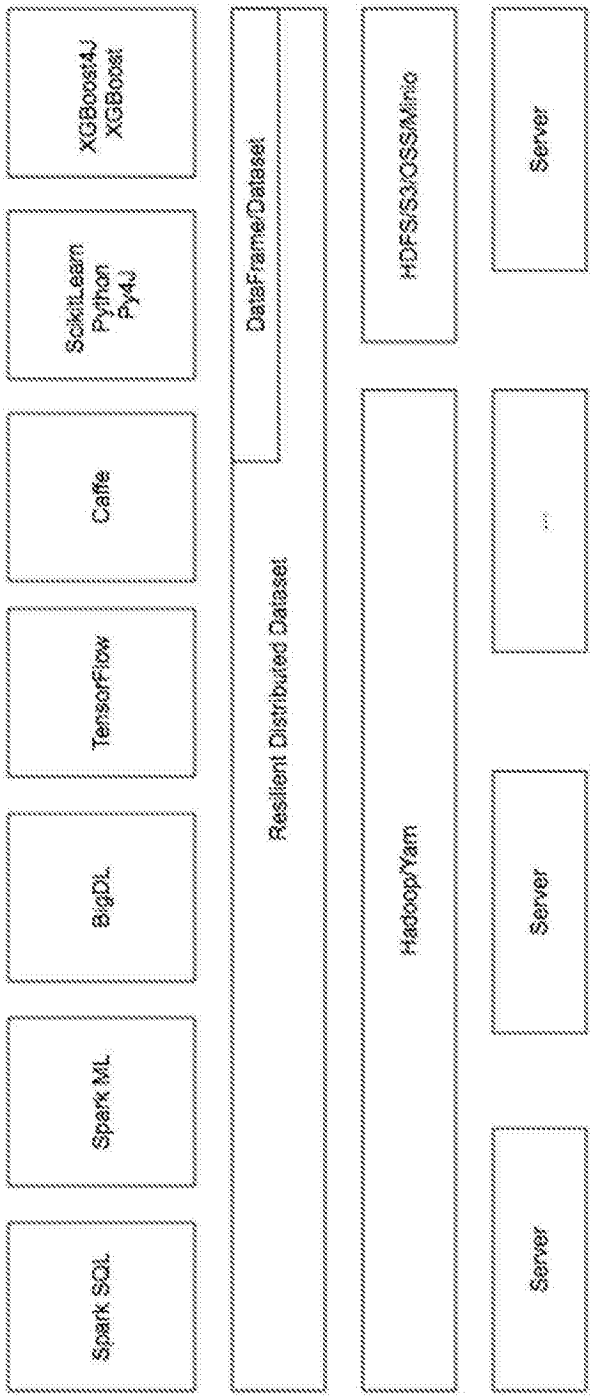

XGBoost4J
XGBoost
SciKitLearn
Python
Py4J
Caffe
TensorFlow
BigDL
Spark ML
Spark SQL
DataFrame/Dataset
Resilient Distributed Dataset
HDFS/S3/OSS/Minio
Hadoop Yarn
Server
...
Server
Server

图1

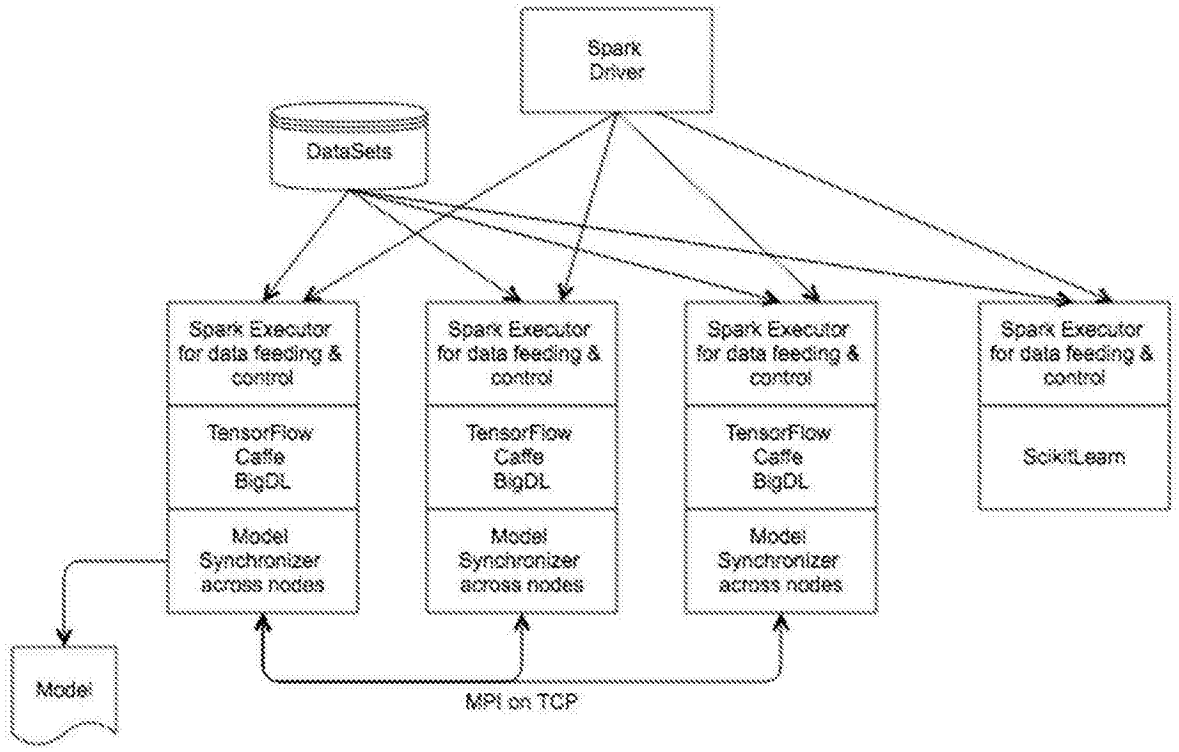

Spark Driver
DataSets
Spark Executor for data feeding & control
TensorFlow Caffe BigDL
Model Synchronizer across nodes
Spark Executor for data feeding & control
TensorFlow Caffe BigDL
Model Synchronizer across nodes
Spark Executor for data feeding & control
TensorFlow Caffe BigDL
Model Synchronizer across nodes
Spark Executor for data feeding & control
SciKitLearn
Model
MPI on TCP

图2