# MapReduce软件产品

## 术语定义

**序列化** 序列化是将结构化的对象状态转换为可持久性或传输的字节流格式的过程，与序列化相对的是反序列化，它将字节流转换为一系列结构化的对象。这两个过程相结合，可实现轻松自如地存储和传输数据。

## 需求分析

随着现代社会数据量的不断增长，工程项目领域需要强大的计算能力和对海量数据的分析能力。这种情况下，对计算和数据处理的要求不仅是能够完成工作获得结果，而且在计算速度、实时性方面都有严格的要求。传统的单处理器和串行程序在这方面已经显得捉襟见肘，在大规模数据处理和计算中已经显得无能为力，在这种情况下并行机和并行编程应运而生。

在并行机的体系结构方面，并行机一般需要多台计算机或者多个处理器协同工作，每个处理器或者计算机完成一部分任务；在并行算法方面，先要把整个计算任务划分成几个阶段或者几部分，它们之间可以并发地进行计算，然后把它们分布到各个处理器或者机器上。通常并行计算的过程需要进行数据交换，一般采用网络通信或者共享存储的方法。当所有计算都完成以后，再由某个机器完成对各部分结果的合并工作，从而得出最终结果。

为方便合理高效的利用并行计算环境，必须要有一种并行程序设计模式与体系结构相配合。在并行编程中，OpenMP主要应用于共享存储器的计算机上，而MPI(Message Passing Interface)作为基于消息传递的程序设计模式，已经得到了广泛的支持和应用，但如果让普通的串行程序员来学习OpenMP 或MPI来开发并行应用程序无疑大大增加它们的负担，而且并行程序运行过程中存在的数据和任务的分配，进程或线程间的消息通信，容错处理等各种细节都会使程序员不能集中精力解决并行任务的表述。

由于需要处理的数据量很大，MapReduce模型必须将计算分布在成百上千台的机器上执行。同时，MapReduce模型应以简化并行编程为目标，使程序员不再关心底层实现细节，而只关注于如何表述并行任务。故MapReduce模型需要解决如下问题：

（1）提供简化的计算模型，尽可能地为用户屏蔽MapReduce模型底层细节

（2）如何把一个需要非常巨大的计算能力才能解决的问题分成许多小的部分问题

（3）将这些小部分计算分配给许多计算机进行处理，这需要实现任务调度策略

（4）把这些计算结果综合起来得到最终的结果

（5）由于执行任务的进程可能失效，因此必须具备相应的容错机制重新调度该任务。

（6）由于需要在节点间传递数据以及报告节点运行状态，MapReduce模型必须能自动处理各计算节点间的互相通信。

## 功能定义

针对需求分析，MapReduce产品需实现如下功能：

（1）简化复杂的数据处理计算过程：为简化复杂的数据处理过程，MapReduce模型将数据处理过程分为两个阶段，即map阶段和reduce阶段，每个阶段都将一系列key/value对作为输入和输出，其中的键和值的类型为MapReduce软件产品的用户指定。同时，用户指定map函数和reduce函数。

根据用户指定的map函数，接受一个输入key/value对，然后产生一系列临时中间key/value对。根据用户指定的reduce函数，接受一个中间key和相关的一个value集。它合并这些value，形成一个比较小的value集合。通常，每次reduce调用只输出1个value。

（2）分割输入数据：根据用户定义的map任务数量，生成若干个map任务，每一个map任务负责将一部分输入数据转换成一系列中间key/value对；

（3）在大规模集群上执行调度：执行调度即任务分配问题，为避免网络流量激增或某些任务服务器负载过重导致效率下降，需要将合适作业的合适任务分配到合适的服务器上，涉及两个步骤，先选择作业，然后在此作业中选择任务；

（4）合并中间结果：当所有map任务完成后，执行reduce任务，它负责到执行过map任务的任务服务器上收集中间结果，并计算出最终结果；

（5）具备三种故障处理功能：（a）任务失败时，由于用户自定义的代码可能有bug，错误本身无法修复，记录错误发生点到日志，并将错误报告发送给作业主节点；（b）任务节点故障处理，当任务节点故障时，将所有原本分配给这个任务节点的任务安排给其它的任务节点；（c）作业主节点故障处理，一旦作业主节点故障，立即启动作业主节点的备用机器，该机器将作为集群中作业主节点。

## 系统架构

MapReduce软件产品采用master/slave架构，即由一个任务服务器JobTracker和一定数目的任务服务器TaskTracker组成。其中，MapReduce集群由成百上千台机器组成，一个MapReduce集群系统结构如图 1所示。

JobTracker负责接收用户提交的作业（Job），并把一个个任务（Task）分派到执行节点上，监控任务的执行、重新执行失效的任务，它是调度各个作业分配任务的核心。为了简化master间同步任务的高度复杂性，MapReduce软件产品的JobTracker作为单节点存在。

TaskTracker是任务（Task）的执行节点，具体负责执行用户定义的操作，每一个作业被拆分成很多的任务，包括map任务和reduce任务等。任务是具体执行的基本单元，它们都需要分配到合适的任务服务器上去执行，任务服务器一边执行一边以发送心跳消息的方式向JobTracker汇报各个任务的状态，以此来帮助作业服务器了解作业执行的整体情况，分配新的任务等。每个TaskTracker直接和JobTracker通信，为了保证任务间的独立性，便于软件产品的容错处理，它们互相之间不允许通信。

作业服务器为每个map和reduce任务存储它们的完成状态（idle，in-progress或者completed），并且通过保存一定的数据结构识别出不同的节点。同时，map任务生成的中间文件的位置通过JobTracker传递到reduce任务，对每个完成的map任务，JobTracker存储着由map任务产生的中间文件的大小和位置，这些信息接下来由JobTracker增量推送到处于in-progress状态的reduce任务节点上。



图 1 MapReduce软件产品结构图

## 产品组成

MapReduce产品主要由一个JobTracker（主结点）和若干TaskTracker（子结点）组成，如图 2所示。JobTracker主要负责任务划分和调度，包括任务状态监控和容错。TaskTracker主要负责任务的执行，状态上报等。应用程序通过JobClient提交作业ID和作业程序，其输入数据由TaskTracker从MapReduce挂接的分布式文件系统中读取，作业被提交到JobTracker后，根据作业的配置，JobTracker将作业划分为多个任务，并分配给TaskTracker，TaskTracker定时向JobTracker上报作业执行状态。

MapReduce软件产品采用RPC协议接口来实现JobTracker和TaskTracker的交流，实现者作为RPC服务器，调用者经由RPC的代理进行调用，如此完成JobTracker和TaskTracker的通信。除此，MapReduce软件产品还提供一个客户端接口，用于平台用户向MapReduce软件产品提交作业。用户提交作业后，只需要监视和等待作业的完成即可，其余工作全部由MapReduce软件产品后台执行。



图 2 MapReduce软件产品组成图

## 接口关系

### 外部接口

表 ‑1 MapReduce系统外部接口表

| 序号 | 调用方 | 服务方 | 接口名称 | 用途 | 调用方式 | 参数类型 | 参数意义 |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 用户 | MapReduce | JobClient.getAllJobs() | 获取集群中已提交的全部作业 | JobStatus[]  JobClient.getAllJobs() | JobStatus[] | JobStatus数组中每一成员对应已提交的每一作业的状态 |
|  | 用户 | MapReduce | JobClient.getMapTaskReports | 根据MapMap作业ID，获得map任务状态报告 | TaskReport[]  JobClient.getMapTaskReports(JobID jobId) | JobID | MapReduce作业ID |
| TaskReport[] | TaskReport数组，数组中每一成员代表指示的map任务的状态 |
|  | 用户 | MapReduce | JobClient.getReduceTaskReports | 根据MapReduce作业ID，获得reduce任务状态报告 | TaskReport[]  JobClient.getReduceTaskReports(JobID jobId) | JobID | MapReduce作业ID |
| TaskReport[] | TaskReport数组，数组中每一成员代表指示的reduce任务的状态 |
|  | 用户 | MapReduce | JobClient.submitJob | 提交作业给MapReduce系统 | RunningJob  JobClient.submitJob(String jobFile) | String | 标识MapReduce作业文件名 |
| RunningJob | 返回用于标识MapReduce作业的对象 |
|  | 用户 | MapReduce | InputFormat | 依据MapReduce作业输入源数据的大小，将输入数据分割成一系列InputSplit，其中，每一InputSplit将交给一个map任务处理。 | InputSplit[]  InputFormat.getSplits(JobConf job, int numSplits) | JobConf | 用于描述MapReduce作业所必备信息 |
| int | 描述分割源数据而得到的一系列InputSplit的个数 |
| InputSplit[] | 返回MapReduce作业输入源数据分割成的一系列InputSplit |
|  | 用户 | MapReduce | RecordReader | 读取InputSplit所表示的字节流中的下一个key/value键值对 | RecordReader.next(K key,V value) | K | 用于生成供map/reduce任务处理的键 |
| V | 用于生成供map/reduce任务处理的值 |
|  | 用户 | MapReduce | Mapper<KeyIn, ValueIn, KeyOut, ValueOut> | 用户提供的Map函数，TaskTracker负责根据mapper函数执行map任务 | Mapper<KeyIn, ValueIn, KeyOut, ValueOut>.map() | KEYIN | Map方法的输入键值对中的键 |
| KEYOUT | Map 方法的输出键值对中的键 |
| VALUEIN | Map方法的输入键值对中的值 |
| VALUEOUT | Map 方法的输出键值对中的值 |
|  | 用户 | MapReduce | Reducer<KeyIn, ValueIn, KeyOut, ValueOut> | 用户提供的Reduce函数，TaskTracker负责根据reduce函数执行reduce任务 | Reducer<KeyIn, ValueIn, KeyOut, ValueOut>.reduce() | KEYIN | Reduce方法的输入键值对中的键 |
| KEYOUT | Reduce方法的输出键值对中的键 |
| VALUEIN | Reduce方法的输入键值对中的值 |
| VALUEOUT | Reduce方法的输出键值对中的值 |
|  | 用户 | MapReduce | Partitioner<K2,V2> | 用于分割map任务输出的中间key-value键值对，分割后的个数等于reduce任务数目。 | Partitioner<K2,V2>.getPartition(key,value,numReduceTasks) | K2 | map任务输出的中间key-value键值对中的每一key |
| V2 | map任务输出的中间key-value键值对中的每一value |
| int | 用户期望的reduce任务数目 |
|  | 用户 | MapReduce | OutputCollector<K2,V2> | 用于收集map任务输出的中间键值对或reduce任务输出的结果键值队 | OutputCollector<K,V>.collect(K key,V value) | K | 指定的map任务输出的中间键或reduce任务输出的结果键 |
| V | 指定的map任务输出的中间值或reduce任务输出的结果值 |
|  | 用户 | MapReduce | OutputFormat.getRecordWriter | 获取OutputFormat类中实现的RecordWriter，其中RecordWriter用于将MapReduce作业运行结果输出到文件系统上。 | OutputFormat.getRecordWriter(FileSystem fs, JobConf job, String name, Progressable progress) | FileSystem | 当前作业所挂载的文件系统，通常情况下默认为hdfs |
| JobConf | 描述当前MapReduce作业的对象 |
| String | 输出文件名称，具有全局唯一性 |
| Progressable | 设置用于报告RecordWriter写入文件进展状态的对象 |
|  | 用户 | MapReduce | RecordWriter<K,V> | 向挂载的文件系统中，以字节流的形式，写入一对key/value键值 | RecordWriter<K,V>.write(K key,V value) | K | 待写入的key |
| V | 待写入的value |

表 ‑1 MapReduce系统外部接口表

| 序号 | 调用方 | 服务方 | 接口名称 | 用途 | 调用方式 | 参数类型 | 参数意义 |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | JobTracker | TaskScheduler | JobTracker.startTracker | 当JobTracker启动时，用于设置MapReduce作业调度服务器taskScheduler的jobTracker，集群只有一个JobTracker节点 | taskScheduler.setTaskTrackerManager(JobTracker jobTracker) | JobTracker | 控制全部MapReduce作业的类 |
|  | JobTracker | TaskScheduler | assignTasks | 对给定的MapReduce集群上指定节点，从任务调度队列中找出合适该节点运行的任务。 | List<Task> assignTasks(TaskTracker taskTracker) | List<Task> | 分配给指定taskTracker的Task任务列表，Task任务包括map、reduce、JobSetup、JobCleanup，TaskCleanUp |
| TaskTracker | 用于控制MapReduce集群上指定节点上各子任务的类 |
|  | JobTracker | TaskScheduler | JobTracker.getJobsFromQueue | 返回全局MapReduce作业队列中的所有作业 | JobStatus[] getJobsFromQueue(String queue) | JobStatus[] | 作业状态组成的数组 |
| String | 指定的MapReduce作业队列名称，目前系统只默认支持一个作业队列，因此队列名称暂无意义 |
|  | JobTracker | TaskScheduler | JobClient.submitJob | 提交作业给MapReduce系统 | RunningJob  JobClient.submitJob(String jobFile) | String | 标识MapReduce作业文件名 |
| RunningJob | 返回用于标识MapReduce作业的对象 |
|  | JobTracker | TaskScheduler | InputFormat | 依据MapReduce作业输入源数据的大小，将输入数据分割成一系列InputSplit，其中，每一InputSplit将交给一个map任务处理。 | InputSplit[]  InputFormat.getSplits(JobConf job, int numSplits) | JobConf | 用于描述MapReduce作业所必备信息 |
| int | 描述分割源数据而得到的一系列InputSplit的个数 |
| InputSplit[] | 返回MapReduce作业输入源数据分割成的一系列InputSplit |
|  | JobTracker | TaskScheduler | RecordReader | 读取InputSplit所表示的字节流中的下一个key/value键值对 | RecordReader.next(K key,V value) | K | 用于生成供map/reduce任务处理的键 |
| V | 用于生成供map/reduce任务处理的值 |
|  | JobTracker | TaskScheduler | Mapper<KeyIn, ValueIn, KeyOut, ValueOut> | 用户提供的Map函数，TaskTracker负责根据mapper函数执行map任务 | Mapper<KeyIn, ValueIn, KeyOut, ValueOut>.map() | KEYIN | Map方法的输入键值对中的键 |
| KEYOUT | Map 方法的输出键值对中的键 |
| VALUEIN | Map方法的输入键值对中的值 |
| VALUEOUT | Map 方法的输出键值对中的值 |
|  | JobTracker | TaskScheduler | Reducer<KeyIn, ValueIn, KeyOut, ValueOut> | 用户提供的Reduce函数，TaskTracker负责根据reduce函数执行reduce任务 | Reducer<KeyIn, ValueIn, KeyOut, ValueOut>.reduce() | KEYIN | Reduce方法的输入键值对中的键 |
| KEYOUT | Reduce方法的输出键值对中的键 |
| VALUEIN | Reduce方法的输入键值对中的值 |
| VALUEOUT | Reduce方法的输出键值对中的值 |
|  | JobTracker | MapReduce | Partitioner<K2,V2> | 用于分割map任务输出的中间key-value键值对，分割后的个数等于reduce任务数目。 | Partitioner<K2,V2>.getPartition(key,value,numReduceTasks) | K2 | map任务输出的中间key-value键值对中的每一key |
| V2 | map任务输出的中间key-value键值对中的每一value |
| int | 用户期望的reduce任务数目 |
|  | 用户 | MapReduce | OutputCollector<K2,V2> | 用于收集map任务输出的中间键值对或reduce任务输出的结果键值队 | OutputCollector<K,V>.collect(K key,V value) | K | 指定的map任务输出的中间键或reduce任务输出的结果键 |
| V | 指定的map任务输出的中间值或reduce任务输出的结果值 |
|  | 用户 | MapReduce | OutputFormat.getRecordWriter | 获取OutputFormat类中实现的RecordWriter，其中RecordWriter用于将MapReduce作业运行结果输出到文件系统上。 | OutputFormat.getRecordWriter(FileSystem fs, JobConf job, String name, Progressable progress) | FileSystem | 当前作业所挂载的文件系统，通常情况下默认为hdfs |
| JobConf | 描述当前MapReduce作业的对象 |
| String | 输出文件名称，具有全局唯一性 |
| Progressable | 设置用于报告RecordWriter写入文件进展状态的对象 |
|  | 用户 | MapReduce | RecordWriter<K,V> | 向挂载的文件系统中，以字节流的形式，写入一对key/value键值 | RecordWriter<K,V>.write(K key,V value) | K | 待写入的key |
| V | 待写入的value |

### 内部接口

## 业务流程

系统级业务流程，可以从子系统间的合作关系出发，通过子系统间彼此协同，为客户端(产品外部接口调用者)提供的主要功能服务的分析，抽取若干核心的业务流程。

## 研究内容

研究内容部分，是在1.1-1.6节的产品总体设计的基础上，对组成产品的每个子系统进行方案设计。一般而言，可以认为，1.4节中产品分解出的每个子系统，将在1.7节中展开描述。

### ××子系统

#### 软件架构

#### 系统组成

列出子系统核心的模块，及其组成关系，可以画组成结构图，对每个模块的功能进行描述。

#### 接口关系

##### 外部接口

##### 内部接口

#### 业务流程

## 关键技术

关键技术的描述，要特性几个要素，**必要性(或技术研究所要解决的具体问题)、研究内容(研究×××模型、方法、理论、技术等)、解决什么问题、达到什么效果**。具体参考描述方法，参见软件产品技术方案撰写方法。

## 关键技术解决途径

技术途径或技术解决途径或技术方案，与关键技术和研究内容的关系是，用于说明如何开展研究内容的研究工作，以实现所述关键技术，也就是说，用什么方法、步骤、算法、流程、架构等来开展研究工作，解决问题。

对于一个项目而言，多个关键技术之间必定存在相互关系，在描述技术途径时，可以用图的方式，对各个关键技术间的关系进行表述。同理，对于软件产品的关键技术也是这样。

技术途径的要素，一般可包括：针对某一关键技术或研究内容的细粒度的软件架构图、应用模式图、模型图(数据模型、规范标准模型等)、关键算法描述、方法步骤或流程。

技术途径的描述方式，多种多样，根据解决的技术类型而定，目前，还抽象不出可复用的模式。

## 主要创新点

在关键技术、研究内容、和技术途径的基础上，突出牛逼之处，例如：

1. 提出了××××
2. 达到了××××
3. 满足了××××
4. 填补了××××
5. 实现了××××效果
6. 具备×××××特征
7. 优化了××××
8. 提高了××××
9. 减低了××××
10. 达到××××目的/目标/效果

## 技术指标

## 部署方式

对于软件产品而言，软件实体(子系统、模块之类)是研制成果的载体，一个完整的软件服务或系统，如何部署到硬件平台上，包括网络拓扑关系。

以分布式归档服务为例，需要部署归档管理服务和归档实例服务，且必须描述归档服务系统与存储系统(Hdfs、盘阵、关系数据库)间的拓扑关系。