通过典型数据分析应用场景、算法与系统架构，结合6大案例，全面、深入讲解Spark大数据分析的各种技术和方法

大数据技术丛书

Spark Big Data Analytics in Action

Spark大数据分析实战

高彦杰 倪亚宇◎著

机械工业出版社
China Machine Press

大数据技术丛书

Spark Big Data Analytics in Action

Spark大数据分析实战

高彦杰 倪亚宇◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Spark 大数据分析实战 / 高彦杰，倪亚宇著 . —北京：机械工业出版社，2015.12
（大数据技术丛书）

ISBN 978-7-111-52307-9

I. S…　II. ①高…　②倪…　III. 数据处理软件　IV. TP274

中国版本图书馆 CIP 数据核字（2015）第 297614 号

Spark 大数据分析实战

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：高婧雅　　责任校对：董纪丽
印　　刷：中国电影出版社印刷厂　　版　　次：2016 年 1 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：14
书　　号：ISBN 978-7-111-52307-9　　定　　价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

Preface 前　言

为什么要写这本书

Spark 大数据技术还在如火如荼地发展，Spark 中国峰会的召开，各地 meetup 的火爆举行，开源软件 Spark 也因此水涨船高，很多公司已经将 Spark 大范围落地并且应用。Spark 使用者的需求已经从最初的部署安装、运行实例，到现在越来越需要通过 Spark 构建丰富的数据分析应用。写一本 Spark 实用案例类的技术书籍，是一个持续了很久的想法。由于工作较为紧张，最初只是将参与或学习过的 Spark 相关案例进行总结，但是随着时间的推移，最终还是打算将其中通用的算法、系统架构以及应用场景抽象出来，并进行适当简化，也算是一种总结和分享。

Spark 发源于美国加州大学伯克利分校 AMPLab 的大数据分析平台，它立足于内存计算，从多迭代批量处理出发，兼顾数据仓库、流处理和图计算等多种计算范式，是大数据系统领域的全栈计算平台。Spark 当下已成为 Apache 基金会的顶级开源项目，拥有着庞大的社区支持，生态系统日益完善，技术也逐渐走向成熟。

现在越来越多的同行已经了解 Spark，并且开始使用 Spark，但是国内缺少一本 Spark 的实战案例类的书籍，很多 Spark 初学者和开发人员只能参考网络上零散的博客或文档，学习效率较慢。本书也正是为了解决上述问题而着意编写。

本书希望带给读者一个系统化的视角，秉承大道至简的主导思想，介绍 Spark 的基本原理，如何在 Spark 上构建复杂数据分析算法，以及 Spark 如何与其他开源系统进行结合构建数据分析应用，让读者开启 Spark 技术应用之旅。

本书特色

Spark 作为一款基于内存的分布式计算框架，具有简洁的接口，可以快速构建上层

数据分析算法，同时具有很好的兼容性，能够结合其他开源数据分析系统构建数据分析应用或者产品。

为了适合读者阅读和掌握知识结构，本书从 Spark 基本概念和机制介绍入手，结合笔者实践经验讲解如何在 Spark 之上构建机器学习算法，并最后结合不同的应用场景构建数据分析应用。

读者对象

本书中一些实操和应用章节，比较适数据分析和开发人员，可以作为工作手边书；机器学习和算法方面的章节，比较适合机器学习和算法工程师，可以分享经验，拓展解决问题的思路。

- Spark 初学者
- Spark 应用开发人员
- Spark 机器学习爱好者
- 开源软件爱好者
- 其他对大数据技术感兴趣的人员

如何阅读本书

本书分为 11 章内容。

第 1 章 从 Spark 概念出发，介绍 Spark 的来龙去脉，阐述 Spark 机制与如何进行 Spark 编程。

第 2 章 详细介绍 Spark 的开发环境配置。

第 3 章 详细介绍 Spark 生态系统重要组件 Spark SQL、Spark Streaming、GraphX、MLlib 的实现机制，为后续使用奠定基础。

第 4 章 详细介绍如何通过 Flume、Kafka、Spark Streaming、HDFS、Flask 等开源工具构建实时与离线数据分析流水线。

第 5 章 从实际出发，详细介绍如何在 Azure 云平台，通过 Node.js、Azure Queue、Azure Table、Spark Streaming、MLlib 等组件对用户行为数据进行分析与推荐。

第 6 章 详细介绍如何通过 Twitter API、Spark SQL、Spark Streaming、Cassandra、D3 等组件对 Twitter 进行情感分析与统计分析。

第 7 章 详细介绍如何通过 Scrapy、Kafka、MongoDB、Spark、Spark Streaming、

Elastic Search 等组件对新闻进行抓取、分析、热点新闻聚类等挖掘工作。

第 8 章 详细介绍了协同过滤概念和模型，讲解了如何在 Spark 中实现基于 Item-based、User-based 和 Model-based 协同过滤算法的推荐系统。

第 9 章 详细介绍了社交网络分析的基本概念和经典算法，以及如何利用 Spark 实现这些经典算法，用于真实网络的分析。

第 10 章 详细介绍了主题分析模型（LDA），讲解如何在 Spark 中实现 LDA 算法，并且对真实的新闻数据进行分析。

第 11 章 详细介绍了搜索引擎的基本原理，以及其中用到的核心搜索排序相关算法——PageRank 和 Ranking SVM，并讲解了如何在 Spark 中实现 PageRank 和 Ranking SVM 算法，以及如何对真实的 Web 数据进行分析。

如果你有一定的经验，能够理解 Spark 的相关基础知识和使用技巧，那么可以直接阅读第 4 ~ 11 章。然而，如果你是一名初学者，请一定从第 1 章的基础知识开始学起。

勘误和支持

由于笔者的水平有限，加之编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。如果你有更多的宝贵意见，我们会尽量为读者提供最满意的解答。你也可以通过微博 @ 高彦杰 gyj，博客：http://blog.csdn.net/gaoyanjie55，或者邮箱 gaoyanjie55@163.com 联系到高彦杰。你也可以通过邮箱 niyayu@foxmail.com 联系到倪亚宇。

期待能够得到大家的真挚反馈，在技术之路上互勉共进。

致谢

感谢微软亚洲研究院的 Thomas 先生和 Ying Yan，在每一次迷茫时给予我鼓励与支持。

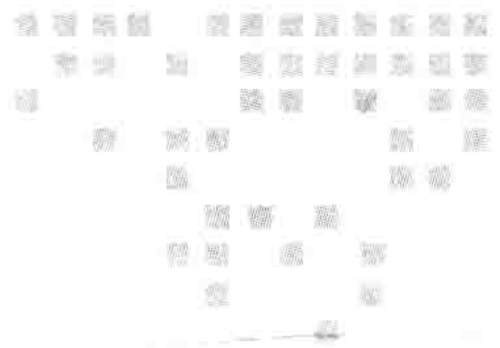
感谢机械工业出版社华章公司的杨福川和高婧雅，在近半年的时间里始终支持我们的写作，你们的鼓励和帮助引导我顺利完成全部书稿。

特别致谢

谨以此书献给我最亲爱的爱人，家人，同事，以及众多热爱大数据技术的朋友们！

高彦杰

目　录 *Contents*



第1章 Chapter 1

Spark 简介

本章主要介绍 Spark 框架的概念、生态系统、架构及 RDD 等，并围绕 Spark 的 BDAS 项目及其子项目进行了简要介绍。目前，Spark 生态系统已经发展成为一个包含多个子项目的集合，其中包含 SparkSQL、Spark Streaming、GraphX、MLlib 等子项目，本章只进行简要介绍，后续章节会有详细阐述。

1.1 初识 Spark

Spark 是基于内存计算的大数据并行计算框架，因为它基于内存计算，所以提高了在大数据环境下数据处理的实时性，同时保证了高容错性和高可伸缩性，允许用户将 Spark 部署在大量廉价硬件之上，形成集群。

1. Spark 执行的特点

Hadoop 中包含计算框架 MapReduce 和分布式文件系统 HDFS。

Spark 是 MapReduce 的替代方案，而且兼容 HDFS、Hive 等分布式存储层，融入 Hadoop 的生态系统，并弥补 MapReduce 的不足。

（1）中间结果输出

Spark 将执行工作流抽象为通用的有向无环图执行计划（DAG），可以将多 Stage 的任务串联或者并行执行，而无需将 Stage 的中间结果输出到 HDFS 中，类似的引擎包括 Flink、Dryad、Tez。

（2）数据格式和内存布局

Spark 抽象出分布式内存存储结构弹性分布式数据集 RDD，可以理解为利用分布式的数组来进行数据的存储。RDD 能支持粗粒度写操作，但对于读取操作，它可以精确到每条记录。Spark 的特性是能够控制数据在不同节点上的分区，用户可以自定义分区策略。

（3）执行策略

Spark 执行过程中不同 Stage 之间需要进行 Shuffle。Shuffle 是连接有依赖的 Stage 的桥梁，上游 Stage 输出到下游 Stage 中必须经过 Shuffle 这个环节，通过 Shuffle 将相同的分组数据拆分后聚合到同一个节点再处理。Spark Shuffle 支持基于 Hash 或基于排序的分布式聚合机制。

（4）任务调度的开销

Spark 采用了事件驱动的类库 AKKA 来启动任务，通过线程池的复用线程来避免系统启动和切换开销。

2. Spark 的优势

Spark 的一站式解决方案有很多的优势，分别如下所述。

（1）打造全栈多计算范式的高效数据流水线

支持复杂查询与数据分析任务。在简单的“Map”及“Reduce”操作之外，Spark 还支持 SQL 查询、流式计算、机器学习和图算法。同时，用户可以在同一个工作流中无缝搭配这些计算范式。

（2）轻量级快速处理

Spark 代码量较小，这得益于 Scala 语言的简洁和丰富表达力，以及 Spark 通过 External DataSource API 充分利用和集成 Hadoop 等其他第三方组件的能力。同时 Spark 基于内存计算，可通过中间结果缓存在内存来减少磁盘 I/O 以达到性能的提升。

（3）易于使用，支持多语言

Spark 支持通过 Scala、Java 和 Python 编写程序，这允许开发者在自己熟悉的语言环境下进行工作。它自带了 80 多个算子，同时允许在 Shell 中进行交互式计算。用户可以利用 Spark 像书写单机程序一样书写分布式程序，轻松利用 Spark 搭建大数据内存计算平台并充分利用内存计算，实现海量数据的实时处理。

（4）与 External Data Source 多数据源支持

Spark 可以独立运行，除了可以运行在当下的 Yarn 集群管理之外，它还可以读取已有的任何 Hadoop 数据。它可以运行多种数据源，比如 Parquet、Hive、HBase、HDFS 等。这个特性让用户可以轻易迁移已有的持久化层数据。

（5）社区活跃度高

Spark 起源于 2009 年，当下已有超过 600 多位工程师贡献过代码。开源系统的发

展不应只看一时之快，更重要的是一个活跃的社区和强大的生态系统的支持。

同时也应该看到 Spark 并不是完美的，RDD 模型适合的是粗粒度的全局数据并行计算；不适合细粒度的、需要异步更新的计算。对于一些计算需求，如果要针对特定工作负载达到最优性能，还需要使用一些其他的大数据系统。例如，图计算领域的 GraphLab 在特定计算负载性能上优于 GraphX，流计算中的 Storm 在实时性要求很高的场合要更胜 Spark Streaming 一筹。

1.2　Spark 生态系统 BDAS

目前，Spark 已经发展成为包含众多子项目的大数据计算平台。BDAS 是伯克利大学提出的基于 Spark 的数据分析栈（BDAS）。其核心框架是 Spark，同时涵盖支持结构化数据 SQL 查询与分析的查询引擎 Spark SQL，提供机器学习功能的系统 MLBase 及底层的分布式机器学习库 MLlib，并行图计算框架 GraphX，流计算框架 Spark Streaming，近似查询引擎 BlinkDB，内存分布式文件系统 Tachyon，资源管理框架 Mesos 等子项目。这些子项目在 Spark 上层提供了更高层、更丰富的计算范式。

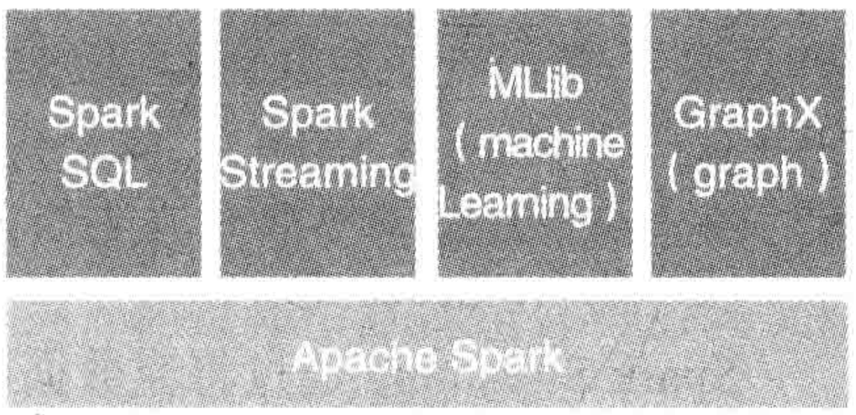


图 1-1　伯克利数据分析栈（BDAS）主要项目结构图

图 1-1 展现了 BDAS 的主要项目结构图。

下面对 BDAS 的各个子项目进行更详细的介绍。

（1）Spark

Spark 是整个 BDAS 的核心组件，是一个大数据分布式编程框架，不仅实现了 MapReduce 的算子 map 函数和 reduce 函数及计算模型，还提供了更为丰富的算子，例如 filter、join、groupByKey 等。Spark 将分布式数据抽象为 RDD（弹性分布式数据集），并实现了应用任务调度、RPC、序列化和压缩，并为运行在其上层的组件提供 API。其底层采用 Scala 这种函数式语言书写而成，并且所提供的 API 深度借鉴函数式的编程思想，提供与 Scala 类似的编程接口。

图 1-2 所示即为 Spark 的处理流程（主要对象为 RDD）。

Spark 将数据在分布式环境下分区，然后将作业转化为有向无环图（DAG），并分阶段进行 DAG 的调度和任务的分布式并行处理。

（2）Spark SQL

Spark SQL 提供在大数据上的 SQL 查询功能，类似于 Shark 在整个生态系统的角色，它们可以统称为 SQL on Spark。之前，由于 Shark 的查询编译和优化器依赖 Hive，使得 Shark 不得不维护一套 Hive 分支。而 Spark SQL 使用 Catalyst 作为查询解析和优化器，并在底层使用 Spark 作为执行引擎实现 SQL 的算子。用户可以在 Spark 上直接书写 SQL，

相当于为 Spark 扩充了一套 SQL 算子，这无疑更加丰富了 Spark 的算子和功能。同时 Spark SQL 不断兼容不同的持久化存储（如 HDFS、Hive 等），为其发展奠定广阔的空间。

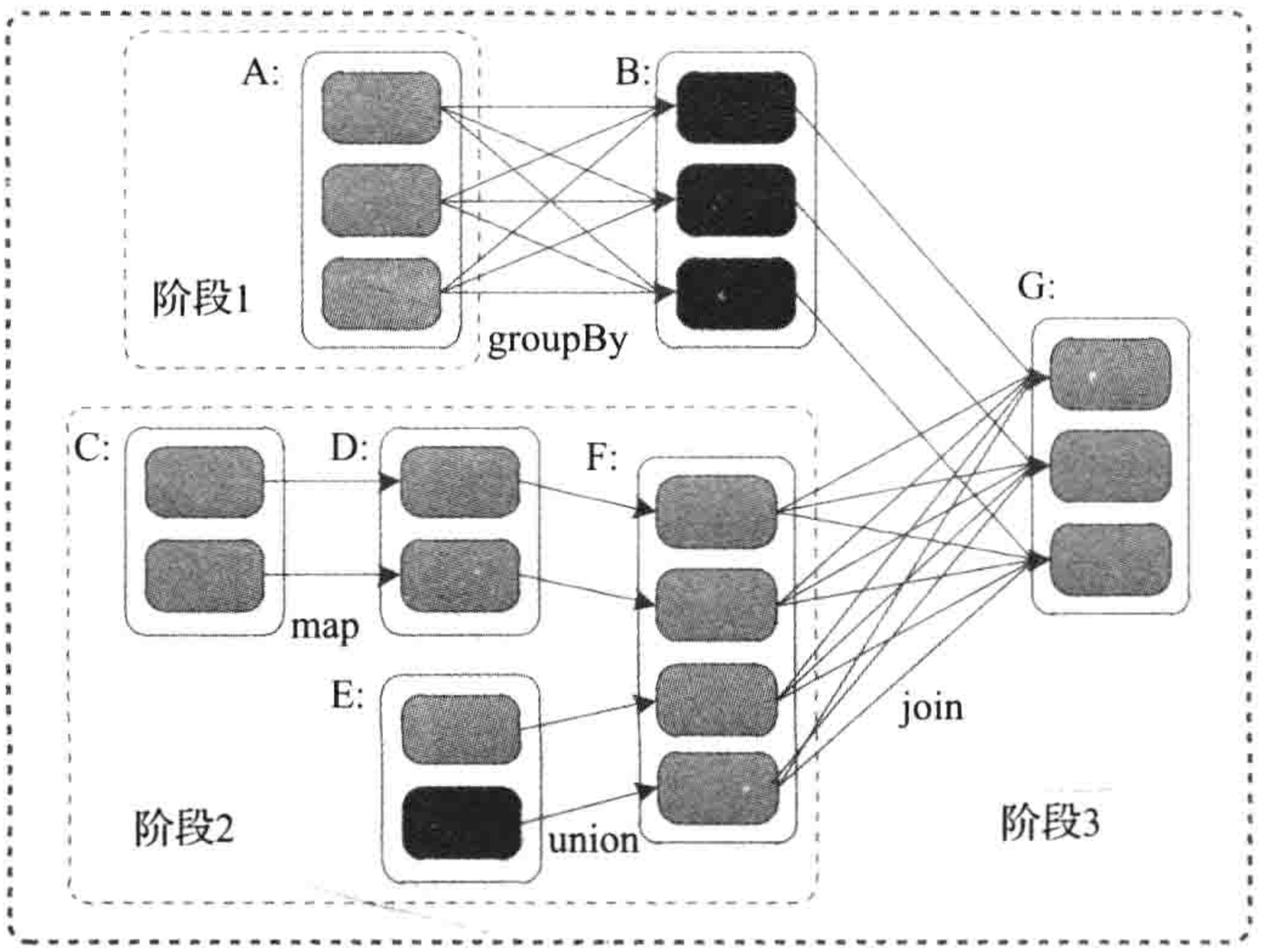


图 1-2　Spark 的任务处理流程图

（3）Spark Streaming

Spark Streaming 通过将流数据按指定时间片累积为 RDD，然后将每个 RDD 进行批处理，进而实现大规模的流数据处理。其吞吐量能够超越现有主流流处理框架 Storm，并提供丰富的 API 用于流数据计算。

（4）GraphX

GraphX 基于 BSP 模型，在 Spark 之上封装类似 Pregel 的接口，进行大规模同步全局的图计算，尤其是当用户进行多轮迭代的时候，基于 Spark 内存计算的优势尤为明显。

（5）MLlib

MLlib 是 Spark 之上的分布式机器学习算法库，同时包括相关的测试和数据生成器。MLlib 支持常见的机器学习问题，例如分类、回归、聚类以及协同过滤，同时也包括一个底层的梯度下降优化基础算法。

1.3　Spark 架构与运行逻辑

1. Spark 的架构

- Driver：运行 Application 的 main() 函数并且创建 SparkContext。
- Client：用户提交作业的客户端。

- Worker：集群中任何可以运行 Application 代码的节点，运行一个或多个 Executor 进程。
- Executor ：运行在 Worker 的 Task 执行器，Executor 启动线程池运行 Task，并且负责将数据存在内存或者磁盘上。每个 Application 都会申请各自的 Executor 来处理任务。
- SparkContext：整个应用的上下文，控制应用的生命周期。
- RDD：Spark 的基本计算单元，一组 RDD 形成执行的有向无环图 RDD Graph。
- DAG Scheduler：根据 Job 构建基于 Stage 的 DAG 工作流，并提交 Stage 给 TaskScheduler。
- TaskScheduler：将 Task 分发给 Executor 执行。
- SparkEnv：线程级别的上下文，存储运行时的重要组件的引用。

2. 运行逻辑

（1）Spark 作业提交流程

如图 1-3 所示，Client 提交应用，Master 找到一个 Worker 启动 Driver，Driver 向 Master 或者资源管理器申请资源，之后将应用转化为 RDD 有向无环图，再由 DAGScheduler 将 RDD 有向无环图转化为 Stage 的有向无环图提交给 TaskScheduler，由 TaskScheduler 提交任务给 Executor 进行执行。任务执行的过程中其他组件再协同工作确保整个应用顺利执行。

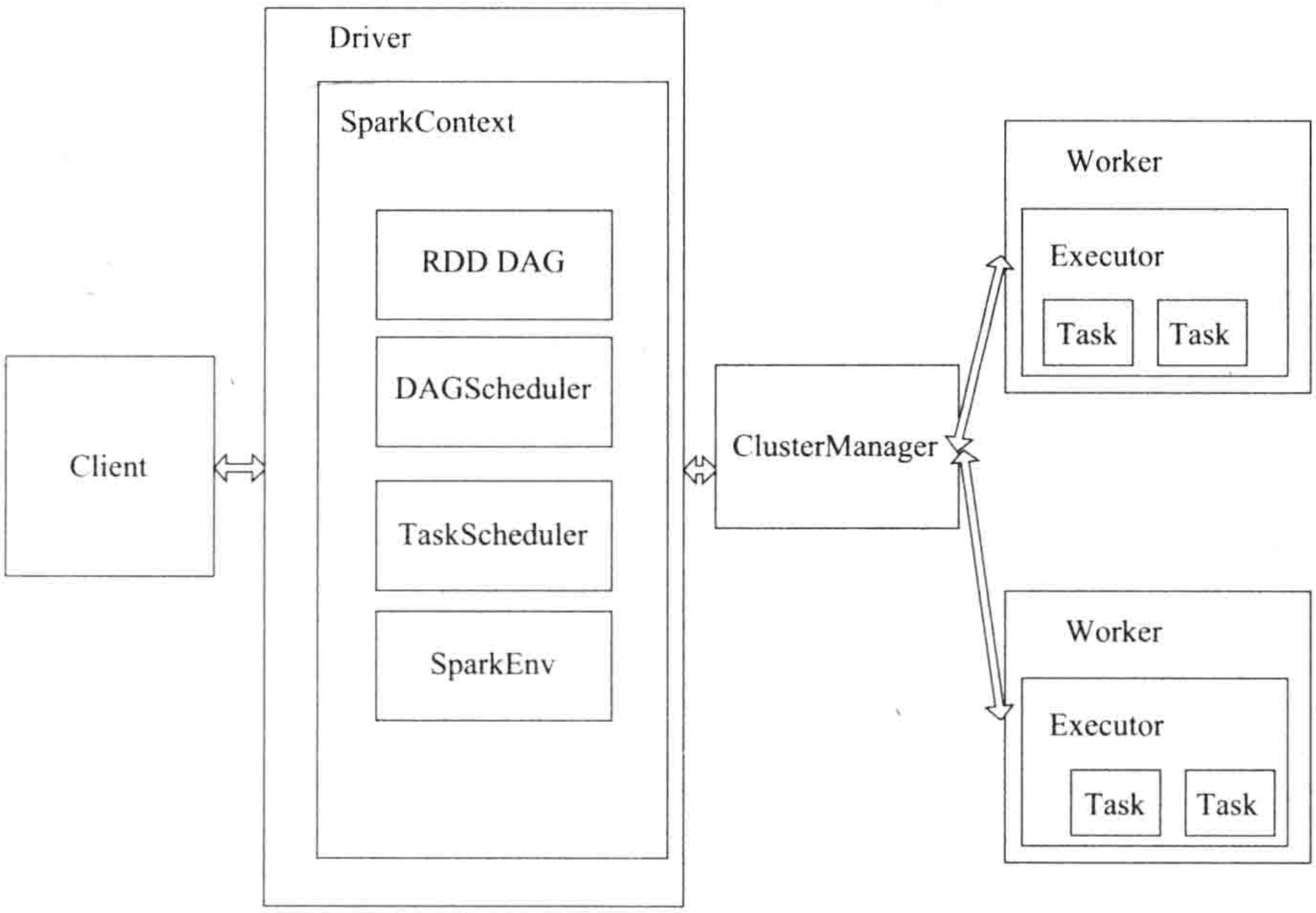


图 1-3　Spark 架构

（2）Spark 作业运行逻辑

如图 1-4 所示，在 Spark 应用中，整个执行流程在逻辑上运算之间会形成有向无环图。Action 算子触发之后会将所有累积的算子形成一个有向无环图，然后由调度器调度该图上的任务进行运算。Spark 的调度方式与 MapReduce 有所不同。Spark 根据 RDD 之间不同的依赖关系切分形成不同的阶段（Stage），一个阶段包含一系列函数进行流水线执行。图中的 A、B、C、D、E、F，分别代表不同的 RDD，RDD 内的一个方框代表一个数据块。数据从 HDFS 输入 Spark，形成 RDD A 和 RDD C，RDD C 上执行 map 操作，转换为 RDD D，RDD B 和 RDD E 进行 join 操作转换为 F，而在 B 到 F 的过程中又会进行 Shuffle。最后 RDD F 通过函数 saveAsSequenceFile 输出保存到 HDFS 中。

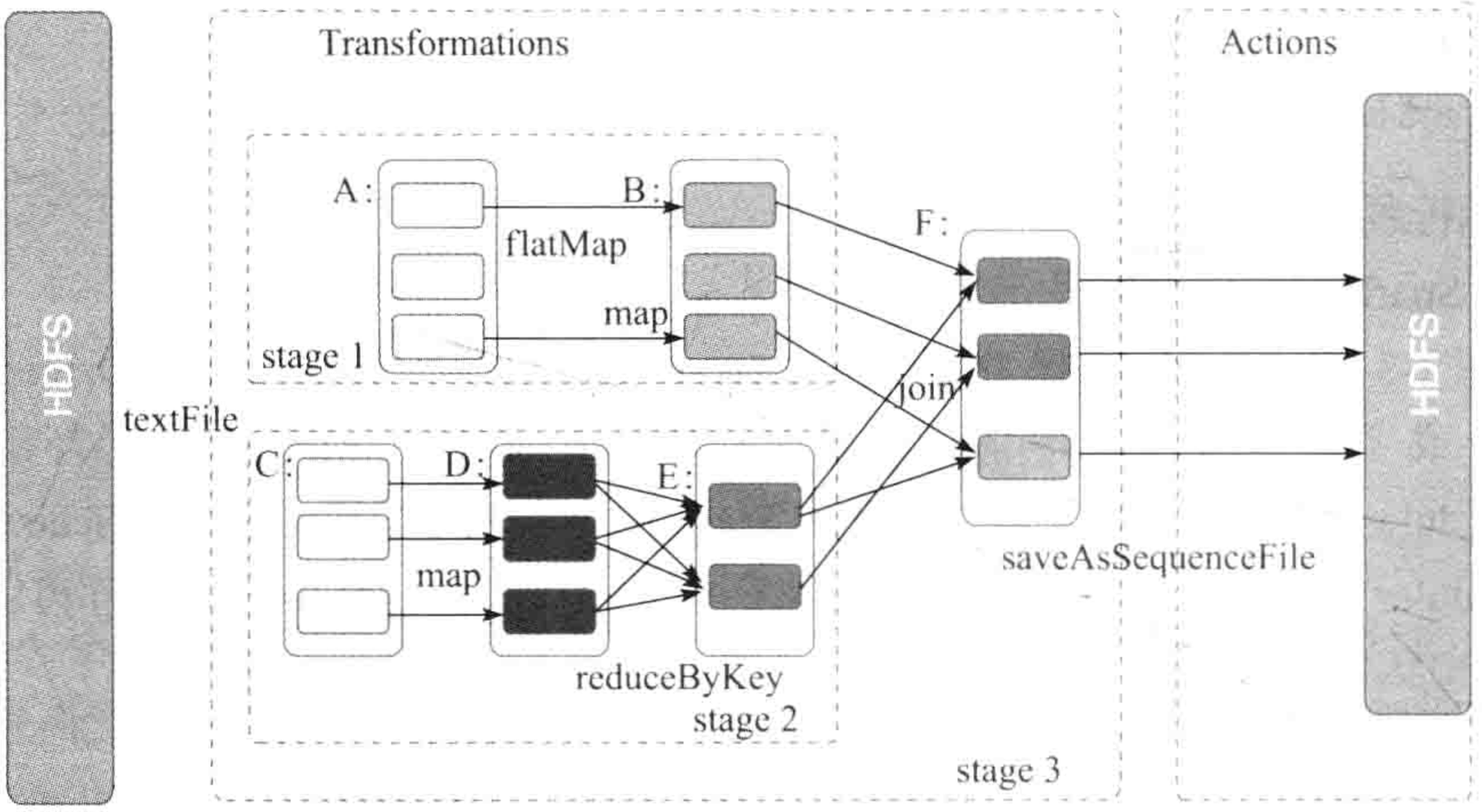


图 1-4　Spark 执行有向无环图

1.4　弹性分布式数据集

本节将介绍弹性分布式数据集 RDD。Spark 是一个分布式计算框架，而 RDD 是其对分布式内存数据的抽象，可以认为 RDD 就是 Spark 分布式算法的数据结构，而 RDD 之上的操作是 Spark 分布式算法的核心原语，由数据结构和原语设计上层算法。Spark 最终会将算法（RDD 上的一连串操作）翻译为 DAG 形式的工作流进行调度，并进行分布式任务的分发。

1.4.1　RDD 简介

在集群背后，有一个非常重要的分布式数据架构，即弹性分布式数据集（Resilient Distributed Dataset，RDD）。它在集群中的多台机器上进行了数据分区，逻辑上可以认