O'REILLY®

弹性MapReduce编程

Programming Elastic MapReduce

Kevin J. Schmidt, Christopher Phillips 著
陈新 唐晓 译

中国电力出版社

弹性MapReduce编程

Kevin J. Schmidt & Christopher Phillips 著

陈新 唐晓 译

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo O'REILLY®

O'Reilly Media, Inc.授权中国电力出版社出版

中国电力出版社

图书在版编目（CIP）数据

弹性MapReduce编程/（美）施密特（Schmidt,K.），（美）菲利普斯(Phillips,C.)著；陈新，唐晓译. —北京：中国电力出版社，2015.8

书名原文：Programming Elastic MapReduce

ISBN 978-7-5123-7944-2

I. ①弹… II. ①施… ②菲… ③陈… ④唐… III. ①软件工具－程序设计 IV. ①TP311.56

中国版本图书馆CIP数据核字（2015）第140896号

北京市版权局著作权合同登记

图字：01-2015-3236号

封面设计/ Randy Comer，张健

出版发行/ 中国电力出版社（http://www.cepp.sgcc.com.cn）

地　　址/ 北京市东城区北京站西街19号（邮政编码100005）

经　　销/ 全国新华书店

印　　刷/ 北京丰源印刷厂

开　　本/ 787毫米×980毫米　16开本　9.75印张　185千字

版　　次/ 2015年8月第一版　2015年8月第一次印刷

印　　数/ 0001－3000册

定　　价/ 28.00元（册）

O'Reilly Media, Inc.介绍

O'Reilly Media通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O'Reilly一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了Make杂志，从而成为DIY革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项O'Reilly的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar博客有口皆碑。”

——Wired

“O'Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference是聚集关键思想领袖的绝对典范。”

——CRN

“一本O'Reilly的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

目录

前言 1

第1章 亚马逊弹性MapReduce介绍 9

1.1 本书中使用的亚马逊Web服务 10

1.2 亚马逊弹性MapReduce 12

1.3 亚马逊EMR及Hadoop生态系统 14

1.4 亚马逊弹性MapReduce安装与传统的Hadoop安装 15

1.5 应用程序构建块 17

第2章 AWS的数据采集及数据分析 20

2.1 日志分析应用 21

2.2 日志消息数据集分析 21

2.3 理解MapReduce 22

2.4 收集阶段 24

2.5 模拟系统日志数据 25

2.6 开发一个MapReduce应用程序 32

2.7 自定义JAR MapReduce工作 32

2.8 运行一个亚马逊EMR集群 36

2.9 查看结果 38

2.10 调试Job Flow 38

2.11 应用程序的实际使用 47

第3章 数据过滤设计模型及调度工作 48
3.1 扩展应用程序示例 49
3.2 理解Web服务器日志 49
3.3 使用数据过滤发现Web日志中的错误 52
3.4 在数据集中构建汇总统计 58
3.5 Job Flow调度 62
3.6 AWS数据管道调度 65
3.7 实际使用 75
第4章 亚马逊EMR上使用Hive和Pig进行数据分析 77
4.1 亚马逊Job Flow技术 78
4.2 Pig是什么？ 79
4.3 在亚马逊EMR上使用Pig 79
4.4 Hive是什么？ 90
4.5 在亚马逊EMR上使用Hive 91
4.6 我们应用程序中的Hive和Pig 98
第5章 使用EMR进行机器学习 99
5.1 机器学习快速导览 99
5.2 Python和EMR 101
5.3 接下来干什么呢？ 110
第6章 规划AWS项目并管理开销 111
6.1 开发项目开销模型 111
6.2 优化AWS资源来减少项目开销 117
6.3 亚马逊为预估项目开销提供的工具 127
附录A 亚马逊Web服务资源和工具 129
附录B 云计算、亚马逊Web服务及其影响 133
附录C 安装和设置 141

前言

很多组织拥有存储在多个地方的珍贵的信息宝藏。为了揭开这些信息中隐藏的秘密，将其用于竞争的市场环境中，不少组织已经开始研究Hadoop和“大数据”，将其作为获取竞争优势的关键。然而许多组织缺乏知识型资源和数据中心空间来为它们的数据分析项目提供大规模Hadoop解决方案。

亚马逊弹性MapReduce（Elastic MapReduce，EMR）是亚马逊的Hadoop解决方案，运行在亚马逊的数据中心。亚马逊的解决方案使得组织只需要关注需要解决的数据分析问题，而无需规划数据中心建设和维护巨大的机器集群。亚马逊的按使用付费模式是另外一个优势，使得组织在开始项目之前无需前期投入，并可以随着项目增长立即进行扩展。我们希望本书可以激发你研究亚马逊Web 服务（Amazon Web Services，AWS）和亚马逊EMR的动力，并使用本书来帮助你启动接下来的重大项目，使用强大的亚马逊云来解决用户巨大的数据分析问题。

本书关注需要构建使用AWS和EMR应用程序的核心亚马逊技术。本书选择一个分析日志文件的应用程序来作为贯穿全书的实例分析，用于展示强大的EMR日志分析对于面临大量数据分析问题的组织来说是一个很好的研究案例。计算机日志文件包含了来自不同来源的大量各式各样的数据，以及可以用来挖掘并获取有价值的情报。更为重要的是，日志文件贯穿计算机系统，它们无所不在，并且能够提供准备好的可用数据集，用户可以使用这些数据，并开始解决数据分析问题。

下面是本书的一个大纲：

- 关于第三方软件的样本配置。

- 关于AWS的逐步的配置。
- 示例代码。
- 最佳实践。
- 需要注意的地方。

本书的目的不是为了提供所有代码、配置等，而是为了使用户可以立即开始在AWS上运行应用程序。本书将提供指导来帮助读者了解如何将一个系统或应用程序放到云环境中，并描述在AWS上构建自己项目时可能会面临的核心问题。

如果你有为传统数据中心开发应用程序或者管理应用程序开发的经验，那么将可以最大限度地利用本书，但是现在希望学习如何将应用程序和数据迁移到云环境中。可能你比较习惯使用开发工具集和审阅代码示例、体系结构图和配置示例，这样可以理解本书中涉及到的基本概念。在许多示例中，本书将使用UNIX下的命令行和命令行工具，所以对使用基本UNIX命令行工具的人员来说，操作命令行是十分熟悉的。本书中的示例也可以用在Windows系统中，但是可能需要使用类似Cygwin（*http://www.cygwin.com/*）这样的第三方工具来运行。

本书给你带来的挑战是需要采用新的方法来看待在传统数据中心之外的应用程序，但愿这将让你意识到可以集中关注于希望解决的问题，而不是关心许多私有数据中心建设的管理性问题。

什么是AWS？

亚马逊Web服务（Amazon Web Services, AWS，*http://aws.amazon.com/cn/what-is-aws/*）是亚马逊在2006年启动的一个计算平台的名称。AWS为企业和第三方开发者提供了一整套服务来构建解决方案，该解决方案可以使用位于亚马逊全球数据中心的计算和软件资源。亚马逊弹性MapReduce（Amazon Elastic MapReduce，EMR，*http://aws.amazon.com/cn/elasticmapreduce/*）是AWS许多服务中的一种。开发者和公司只需要使用按使用付费模式来为所使用的AWS中的资源付费。该模式改变了许多企业看待新项目和新提议的途径。新的提议可以立即启动，并根据顾客基础增长在AWS上进行扩展，而没有购买新的服务器和基础设施等传统的预先开支。通过使用AWS，公司现在可以关注创新和构建重大的解决方案。它们可以更少地关注如何建设和维护数据中心和物理的基础设施，以便集中关注开发解决方案。

云服务及其影响

贯穿本书，我们将讨论许多AWS和云服务的优势。尽管这些服务在很多地方确实为组织提供了惊人的价值，但并不是对所有项目都是最好的选择。在AWS中运行应用程序会伴随与使用VMware等其他虚拟技术一样的影响和效果。这些影响可能会影响到应用程序的性能和安全性，并且运行在云中的应用程序可能会和其他用户的应用程序运行在同一台物理服务器上。对于大多数应用程序来说，云计算带来的优势远远大于它带来的影响。在附录B中，本书讨论了许基于云的应用程序的因素和影响。建议在启用你的应用程序之前认真阅读附录B中的条目，确认该项目非常适合AWS和云计算。

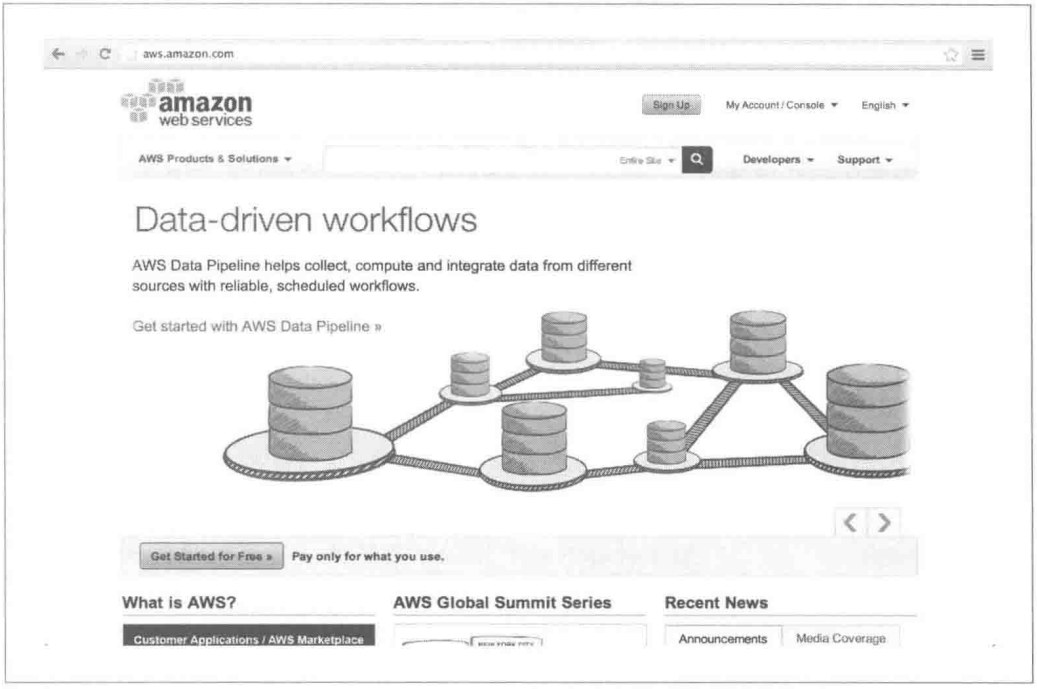
本书中有什么？

本书组织结构为：第1章介绍云计算的基本知识，并帮助读者理解亚马逊Web服务（AWS）和亚马逊弹性MapReduce（EMR）。第2章开始介绍亚马逊工具，将会使用这些工具来审计日志文件，并在亚马逊EMR中执行第一个工作流。在第3章，开始着手处理企业使用多种MapReduce设计模式的亚马逊EMR可以分析的类型，并检查可以从日志数据中获取的结果。在第5章，本书介绍机器学习技术，并研究如何实现这些技术，并在应用程序中利用这些技术来构建可以提出或给出推荐问题解决方案的智能系统。最后，在第6章，我们评审了AWS和EMR应用程序的预估项目开销，以及如何为项目进行开销分析。

注册AWS

首先需要注册AWS。如果你已经是一个AWS用户，那么可以略过这部分，因为你已经可以访问本书中需要使用的AWS服务。如果你是一名新用户，那么请从现在开始做。

请访问AWS网站（*http://aws.amazon.com/cn/what-is-aws/*），注册AWS，如图P-1所示。

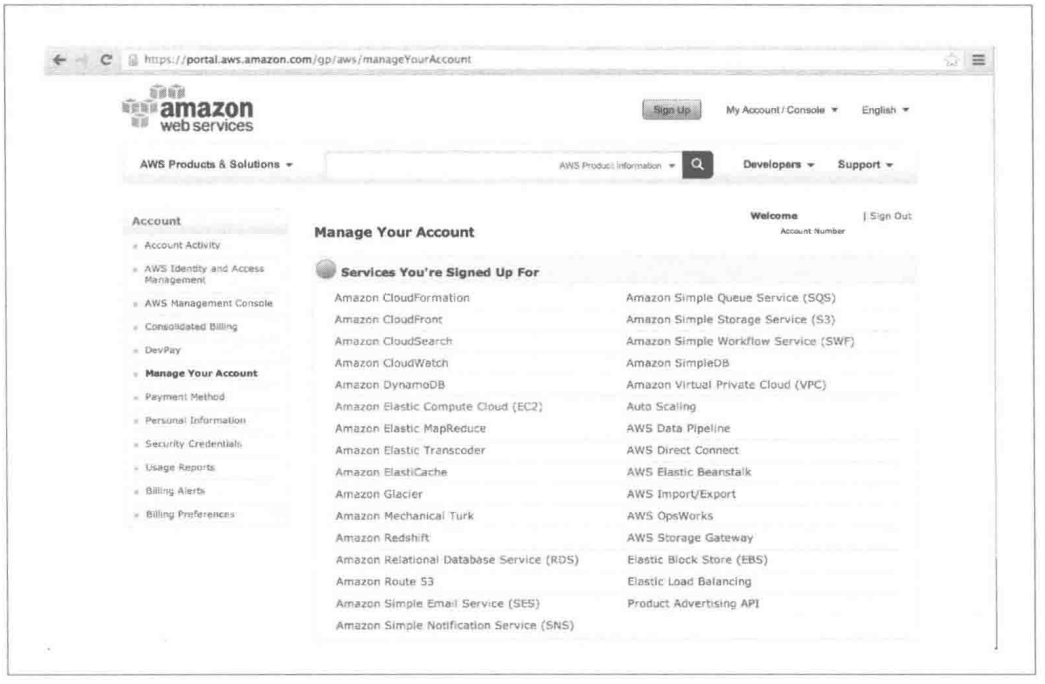


图P-1：AWS主页

在这里需要提供手机号码来确认创建了一个有效的账户，并且还需要提供信用卡号码来使得亚马逊可以按照AWS服务的使用来收费。本书将在第6章讨论如何在AWS上估计、评审和建立账单警告。

在创建完AWS账户后，请进入“我的账户”页面来查看可以访问哪些服务。图P-2显示了当前用户可用的服务，有可能你看到的界面会有一些不同。

注意： 请牢记，AWS的使用是有相关费用的，本书中许多示例和练习会导致你的账户付费。对于一个新的AWS账户，有一个免费套餐（*http://aws.amazon.com/cn/free/*）。为了最小化学习亚马逊EMR的开销，请务必仔细阅读免费套餐的限制，在运行完练习后请关闭实例，并学习第6章提到的如何预估开销。



图P-2：注册后可用的AWS服务

本书中代码示例

贯穿全书有大量的代码样本和示例。许多示例是使用Java编程语言或Hadoop Java库构建的。为了最大限度地利用本书，需要搭建一个系统来进行Java开发，并使用Hadoop Java JAR来构建亚马逊EMR可以使用和执行的应用程序。为了准备开发和构建接下来的应用程序，请查阅附录C来搭建自己的开发环境。这不是必须的，但是这将帮助你最大限度地利用本书中提供的资源。

本书使用的排版约定

本书使用如下排版约定：

斜体(italic)

用来表示新术语、URL、email地址、文件名、文件扩展名等。

等宽字体（`constant width`）

用来表示程序列表，同时在段落中引用的程序元素（例如变量、函数名、数据库、数据类型、环境变量、声明和关键字等）也用该格式表示。

等宽黑体（**constant width bold**）

用于表示需要用户逐字符键入的命令或其他文本。

等宽斜体（*constant width italic*）

用于表示应该以用户提供的值或根据上下文决定的值加以替换的文本。

注意：表示一个技巧、建议或者一般性注释。

警告：表示一个警告或者一个提醒性注释。

如何使用示例代码

本书可以帮助你完成你的工作。一般来说，可以在自己的程序和文档中使用本书中的代码。除非将重新编译代码中重要的部分，否则你是不需要联系我们来获得授权的。举例来说，使用本书中几段程序来编程是不需要获得我们的授权的，但是销售或发布O'Reilly出版书籍中配套光盘中代码是需要授权的。通过引用本书中示例来回答问题是不需要授权的，而将本书中重要部分的示例代码整合到你产品的文档中是需要授权的。

我们感谢你在使用我们代码的时候给出引用说明，但这不是硬性规定。一个引用说明通常包括了标题、作者、出版社和ISBN。举例来说，本书的引用说明“*Programming Elastic MapReduce* by Kevin J. Schmidt and Christopher Phillips (O'Reilly). Copyright 2014 Kevin Schmidt and Christopher Phillips, 978-1-449-36362-8.”

如果不确定所使用的示例代码是否超出了上面给定的权限，可以随时通过电子邮件联系我们。我们的电子邮件地址是*permissions@oreilly.com*。

Safari® Book Online

Safari® Book Online 是一个按需定制的数字化图书馆，它提供来自全球技术和商业上最顶尖的专家的书籍和视频。

技术专业人员、软件开发者、网页设计者和商业创新专业人员都可以使用Safari Book Online来作为研究问题、解决、学习和认证培训的主要资源。

Safari Book Online为组织者、政府机构和个人提供了各种价格范围的资源。订阅者可以通过一个统一的可搜索数据库访问成千上万的书籍、培训视频和预先出版的手

稿，提供资源的出版社包括O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology等。如需更多关于Safari Book Online的信息，请访问我们的网站。

如何联系我们

请将关于本书的意见和问题发送给出版社：

美国：

O'Reilly Media，Inc.
1005 Gravenstein Highway North
Sebastopol，CA 95472

中国：

北京市西城区西直门南大街2号成铭大厦C座807室（100035）
奥莱利技术咨询（北京）有限公司

我们为本书提供了网页，该网页上面列出了勘误表、示例和任何其他附加的信息。你可以访问如下网页获得：

http://oreil.ly/Prog-Elastic-MapReduce

要询问技术问题或对本书提出建议，请发送电子邮件至：

bookquestions@oreilly.com

要获得更多关于我们的书籍、会议、资源中心和O'Reilly网络的信息，请参见我们的网站：

http://www.oreilly.com

http://www.oreilly.com.cn

我们的Facebook地址：*http://facebook.com/oreilly*

我们的Twitter地址：*http://twitter.com/oreillymedia*

也可以在YouTube查看我们的信息，地址是：*http://www.youtube.com/oreillymedia*

致谢

我的妻子Michelle给了我巨大的勇气来完成本书。当然我在Dell公司的同事也是非常值得感谢的，他们支持我来完成本项目。接下来需要感谢的是提供给我许多有价值资料的合作者：Rob Scudiere、Wayne Haber和 Marco Arguedas。最后，技术评审提供了关于如何使得本书更好的出色的指导，他们是Jennifer Davis、Michael Ducy、Kirk Kimmel、Ari Hershowitz、Chris Corriere、Matthew Gast和Russell Jurney。

——Kevin

我想要感谢我漂亮的妻子Inna和我可爱的孩子Jacqueline和Josephine。他们的善良、幽默和爱在撰写本书和整个生命旅程中给我极大的鼓舞和支持。我还要感谢技术评审，感谢他们富有洞察力的反馈，这极大地改进了本书中许多学习的例子。特别要感谢Matthew Gast，他提供了贯穿全书所有章节的反馈，他深刻理解了技术带来的商业和技术价值，并且这些例子是非常宝贵的。非常感谢Wayne Haber、Rob Scudiere、Jim Birmingham，以及我在Dell公司的同事，感谢他们给予的贯穿全书编写过程的有价值的努力和定期评审。最后要感谢我的合作者Kevin Schmidt和编辑Courtney Nash，感谢给我这个机会来完成这本重要的书籍，以及感谢他们在编写过程中艰苦的工作和努力。

——Chris

第1章

亚马逊弹性MapReduce介绍

在很多领域，编程最难的部分不是如何解决问题，而是决定要解决什么问题。

——保罗·格雷厄姆

伟大的黑客

2012年8月6日，好奇号火星探测器（Mars rover Curiosity）成功登陆了距地球数百万英里外的火星。这项任务带来了大量基于工程技术专业知识的信息数据。令人兴奋的是这项任务背后的信息技术以及使用美国国家航空航天局的喷气推进实验室（NASA’s Jet Propulsion Laboratory，JPL）提供的AWS服务（*http://aws.amazon.com/solutions/case-studies/nasa-jpl-curiosity*）。在着陆之前不久，美国国家航空航天局能够提供大量的AWS（*http://aws.amazon.com/solutions/case-studies/nasa-jpl-curiosity*）基础设施来对25 Gb/s（千兆比特每秒）的吞吐率进行支持，从而有能力为它的众多粉丝和科学家提供关于好奇号及有关它登陆的最新信息。如今，美国国家航空航天局继续利用AWS来进行数据分析并且支持科学家们能够对该任务背后的科学数据进行快速访问。

为什么本书要把亚马逊弹性MapReduce作为一个重要内容呢？因为曾经这种类型的资源只提供给政府或者大型的企业使用。而如今，任何一一上拥有笔记本电脑和信用卡的用户就可以使用这种对海量数据进行分析以及支持瞬间大流量的能力。过去要建立一个大数据中心，计算机硬件以及网络需要花费几个月的时间，而现在一些短期的项目可以通过AWS瞬间完成。

如今，商家们需要通过了解它们的客户来确定发展趋势，以保持领先的竞争力。在金融和企业安全方面，商家们被大量的信息所淹没。IT部门被要求在最经济的预算下去弄清楚不断增长的数据量中所包含的信息，从而帮助企业在这场游戏中保持领先的地

位。Hadoop 以及MapReduce 框架可以提供强有力的帮助。然而，传统的数据中心需要大量的时间与开销来建造和维护支持其运行所需的庞大的IT基础设施。

EMR是一个亚马逊数据中心的云端托管的解决方案，它提供了计算能力以及随需应变的基础设施，从而来解决如找准问题发展趋势以及理解海量的数据的真正含义这样一些复杂的问题。

通过本书，我们将探索亚马逊EMR以及如何利用它去解决所遇到的数据分析问题。在众多例子中，我们将关注许多企业同样面临的共同问题：跨越不同的系统分析计算机日志信息。很多企业被要求遵循已有的法规，比如医疗保险可携性和责任法案（HIPAA），支付卡行业数据安全标准（PCI DSS），即使不能每天，也要定期地对日志信息进行分析和回顾。这些大企业的日志信息会非常容易地发展成为TB级或PB级的数据。我们将利用EMR为这些应用建立若干构建模块，来装载计算机的日志信息以及对它们的趋势进行分析。我们将介绍如何利用亚马逊EMR服务去完成这些分析，并且讨论做这些事所需的费用问题。

1.1 本书中使用的亚马逊Web服务

最早的AWS是一种通过虚拟计算机来提供远程托管的基础设施，我们称它为EC2。近年来AWS的发展非常迅速，如今，即使不是全部，AWS也为许多应用程序提供了非常多的构件块。在本书中，我们主要关注于亚马逊提供的各种关键服务。

亚马逊弹性MapReduce（Amazon Elastic MapReduce，EMR）

如果一本书中的EMR没有使用亚马逊关键的AWS服务，那么这本书是不完整的。总而言之，在本书中我们将进行更为详细的探究，亚马逊EMR是Hadoop框架下的云端的主力，我们可以通过一个可配置、可扩展的计算能力来分析大量的数据。亚马逊EMR大量使用亚马逊S3来存储分析结果以及对主机的数据集进行并行处理，并且通过利用亚马逊的EC2的可扩展计算资源来运行我们开发的作业流，从而完成分析任务。这里有一个额外收取30%费用的EMR EC2实例。可以通过浏览亚马逊EMR Web网页（*http://aws.amazon.com/elasticmapreduce*）来阅读亚马逊EMR的概述。作为本书的重点，本书在很多示例中中反复使用了亚马逊EMR。

亚马逊简单存储服务（Amazon Simple Storage Service，S3）

亚马逊S3是AWS的持久存储器。它提供了简单的网络服务接口，用户通过它可随时在Web上的任何位置存储以及检索任何大小的数据。但是，也会有一些限制，数据必须存储在S3中已经命名的存储桶中，且每个单独文件的大小不能超

过5TB。存储在S3上的数据具有高持久性，它们存放在多个设施中，且在一个设施中又有多个设备。通过本书，我们将利用S3存储器去存储许多亚马逊EMR脚本、数据资源，以及我们的分析结果。

与大多数AWS服务一样，S3提供了标准的基于REST和SOAP的Web服务API，通过使用这些接口可以交互操作存储在S3上的文件。它为每一个开发者提供了一个高扩展性、高可靠性、高安全性，且快速、廉价的存储平台，这也是亚马逊通过运行自己的全球网站所用到的。该服务的目的是获得最大化的规模效益，并将这些好处分享给开发人员。可以通过浏览亚马逊S3的Web网页（*http://aws.amazon.com/s3*）来阅读亚马逊的S3的概述。亚马逊S3的永久存储器将被用来存放那些EMR Job Flow生成的数据集及计算结果集。亚马逊EMR建立的应用程序将需要用到一些S3服务来存储数据。

亚马逊弹性计算云（Amazon Elastic Compute Cloud，EC2）

亚马逊EC2使得在任意一个AWS区域中按需运行多个虚拟机实例成为了可能。该服务的迷人之处就在于可以按自己的需求启动不限个数的实例，而无需再去购买或租赁类似于传统主机服务一样的物理硬件。这就意味着在亚马逊EMR环境下可以将我们的Hadoop集群大小扩展到任何我们需要的尺寸，而不需要考虑新的硬件采购以及进行容量规划。单个EC2实体有各种各样的尺寸及规格，可以用来满足不同类型的应用程序的需要。它有针对高CPU负载、高内存、高速I/O及更多的实例存在。在本书中，我们将使用本地EC2实例来调度亚马逊EMR上的Job Flow，运行日常的管理工作以及与我们应用程序构件块相关的数据操作工作。当然，我们将会利用亚马逊EMR EC2去做繁重的数据处理与分析工作。

可以通过浏览亚马逊EC2的Web网页（*http://aws.amazon.com/ec2*）来阅读亚马逊的EC2的概述。在本书中，亚马逊EC2实例将被作为亚马逊EMR集群的一部分使用。我们同样也会将C2实例用在行政职能、模拟实时路况及数据集上。在你自己构建的应用中，可以在自己的主机上运行管理和生活数据，这些单独的EC2实例在通过亚马逊EMR构建一个应用程序时并不是必需的服务。

亚马逊云存档存储服务（Amazon Glacier）

Amazon Glacier是AWS的一个新产品。Glacier与S3很相似，因为它能以一种安全、持久的存储方式存储一个任意给定大小的数据。使用Glacier的主要目的是用于存放长期的数据，因为它在存储和检索数据时具有高延迟性。亚马逊想要实现在Glacier中进行数据检索请求可能需要花费几小时的时间。正因为这些原因，我们在亚马逊Glacier中存储那些不是经常使用的数据。亚马逊Glacier的优势在于节省大量开销。就在撰写本书的同时，美国东部地区的存储开销是每月每十亿字节0.01美元，相比于S3的存储成本每月每十亿字节0.076美元到0.095美