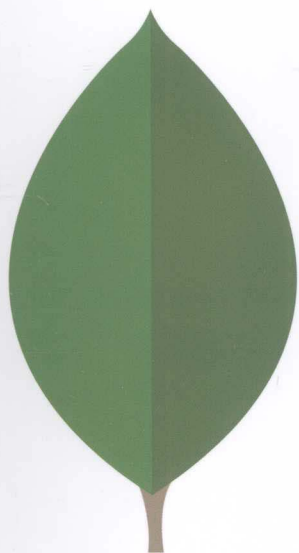


开源、高性能，针对当前最热门的 NoSQL 数据库  
注重理论实践，深入剖析 MongoDB 存储原理与应用  
提炼经典案例，深刻把握大数据应用开发技巧

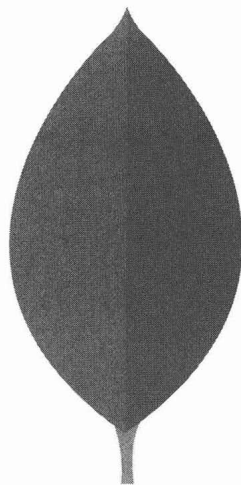


# 大数据存储

# MongoDB

## 实战指南

郭远威 著



大数据存储  
**MongoDB**



实战指南

郭远威 著

人民邮电出版社

北京

## 图书在版编目 (C I P) 数据

大数据存储 : MongoDB实战指南 / 郭远威著. -- 北京 : 人民邮电出版社, 2015. 2  
ISBN 978-7-115-37655-8

I. ①大… II. ①郭… III. ①关系数据库系统—指南  
IV. ①TP311.138-62

中国版本图书馆CIP数据核字(2014)第299457号

## 内 容 提 要

MongoDB 是一种面向文档的分布式数据库。时至今日, MongoDB 以其灵活的数据存储方式逐渐成为 IT 行业非常流行的一种非关系型数据库 (NoSQL)。

本书从学习与实践者的视角出发, 本着通俗精简、注重实践、突出精髓的原则, 精准剖析了 MongoDB 的诸多概念和要点。全书共分 4 个部分, 分别从基础知识、深入理解 MongoDB、监控与管理 MongoDB 和应用实践几个维度详细地介绍了 MongoDB 的特点及应用实例。

本书适合有海量数据存储需求的人员、数据库管理开发人员、数据挖掘与分析人员以及各类基于数据库的应用开发人员阅读。读者将从书中获得诸多实用的知识和开发技巧。

- 
- ◆ 著 郭远威  
责任编辑 陈冀康  
责任印制 张佳莹 彭志环
  - ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号  
邮编 100164 电子邮件 315@ptpress.com.cn  
网址 <http://www.ptpress.com.cn>  
三河市海波印务有限公司印刷
  - ◆ 开本: 800×1000 1/16  
印张: 11.75  
字数: 220 千字 2015 年 2 月第 1 版  
印数: 1-3 500 册 2015 年 2 月河北第 1 次印刷
- 

定价: 39.00 元

读者服务热线: (010)81055410 印装质量热线: (010)81055316  
反盗版热线: (010)81055315

# 前言

多年来，我一直在和数据库存储技术打交道，深知数据存储技术在整个 IT 系统中起着至关重要的作用，尤其是随着云计算时代的到来，所有企业都面临着海量的数据信息，如何处理这些数据成为当前研究的热点。在过去二十几年中，数据的存储是关系数据库的天下，它以高效、稳定、支持事务的优势几乎统治了整个行业的存储业务；但是随着互联网的发展，许多新兴产业如社交网络、微博、数据挖掘等业务快速增长，数据规模变得越来越庞大，高效存储、检索、分析这些海量的数据，关系数据库变得不再适用。前几年我们还可以看到网络上关于关系数据库与 NoSQL 数据库谁优谁劣的激烈讨论，如今 NoSQL 几乎占据了各大数据库论坛讨论的大部分版块。一些行业领头公司也逐渐将业务迁移到非关系数据库上，NoSQL 类型的数据库也变得越来越成熟。当然，在未来一段时间里关系数据库如 Oracle、DB2、SQL Server 等仍会在事务性要求比较高的行业（如银行、电信等）发挥它的作用。

另一方面，在信息技术领域，计算与存储一直是密不可分的，当前我们身处云计算的浪潮中，因此对应的各种云存储技术也呼之欲出。本书将介绍的 NoSQL 数据库 MongoDB 正是众多分布式海量数据存储技术中最出色的一种。MongoDB 是一种面向文档的分布式数据库，可扩展，表结构自由，支持丰富的查询语句与数据类型，旨在为未来的大数据应用提供高性能的云存储解决方案。当然 MongoDB 并不是万能的，随着了解的深入，我们也会发现它的缺点，这也是本书的宗旨，尽量让读者明白它的长处与短处，对于特定的业务选择最合适的数据存储方案。最后我们希望本书介绍的 MongoDB 知识能为您在未来的项目中处理海量数据时提供帮助。

## 本书内容

本书尽量从一个学习与实践者的角度，本着力求精简、突出精髓的原则，剖析了 MongoDB 在生产环境中使用需要知道的所有内容，全书分 4 部分，共 13 章，每章的内容简单介绍如下。

第 1 章 本章主要从什么是 MongoDB 以及几个核心进程两方面概述了 MongoDB，使读者整体对 MongoDB 的体系结构有个认识。

第 2 章 本章主要介绍了 MongoDB 的查询语言系统，包含各种查询选择器以及查询选项，这是对任何一个数据库都有的内容。

第 3 章 本章主要介绍了 MongoDB 的索引与查询优化。

第 4 章 本章主要介绍了 MongoDB 的增、删、改语句。

第 5 章 本章主要从底层存储视图与写操作流程剖析了 MongoDB 的 Journaling 日志功能。

第 6 章 本章主要介绍了 MongoDB 的聚集分析框架与 MapReduce 的编程模型。

第 7 章 本章主要介绍了复制集的功能与工作机制，包含数据同步、故障转移、写关注等，这些是 MongoDB 的核心。

第 8 章 本章主要介绍了分片集群，包含部署架构、分片、读写分离、片键选择等内容，这是 MongoDB 不同于传统关系数据库地方，也是实现海量数据分布式存储的关键。

第 9 章 本章主要介绍了分布式文件系统的 GridFS 文件，实现二进制数据的存储。

第 10 章 本章主要介绍了对 MongoDB 的管理与监控，包括数据的导入导出、备份恢复以及运行状态的监控。

第 11 章 本章主要介绍权限控制，实现不同数据库对不同角色用户的权限分配。

第 12 章 本章主要从应用开发角度，介绍了 MongoDB 的 PHP 驱动接口。

第 13 章 本章主要介绍了一个完整的电商平台，数据库使用的是 MongoDB 并对前面所有章节的知识进行总结，内容包含电商平台数据库表的设计、核心代码的编写、前台界面的原型图设计等，还介绍了开发 Web 应用程序常用的 PHP 框架 Codeigniter 和前端开发框架 Bootstrap 等。

## 本书特色

- 注重实践，本书为多年一线数据库存储，部署开发经验的总结。
- 注重效率，本书用最精简的篇幅直接阐明问题的本质，节省宝贵的阅读时间。
- 注重基础，本书用计算机领域相关的基础理论知识来解释某些难于理解的概念。
- 案例丰富，本书使用完整的例子与代码注释，使读者可以直接上手操作。
- 把握未来，大数据势不可挡，本书介绍的 MongoDB 特性与此息息相关。

## 读者对象

- 有海量数据存储需求的人员。
- 数据库管理与开发人员。
- 数据挖掘与分析人员。
- 各类基于数据库的应用程序开发人员。

谨以此书献给热爱技术、热爱 MongoDB 的朋友们!

# 目录

## 第一部分 基础知识

第 1 章 大数据与云计算 .....	3	2.4 小结 .....	22
1.1 什么是大数据 .....	3	第 3 章 索引与查询优化 .....	23
1.2 什么是云计算 .....	4	3.1 索引 .....	23
1.3 大数据与云计算 .....	4	3.1.1 单字段索引 .....	24
1.4 什么是 MongoDB .....	5	3.1.2 复合索引 .....	27
1.5 大数据与 MongoDB .....	6	3.1.3 数组的多键索引 .....	28
1.6 MongoDB 特点 .....	7	3.1.4 索引管理 .....	29
1.7 安装 MongoDB .....	9	3.2 查询优化 .....	30
1.8 几个重要的进程介绍 .....	9	3.3 小结 .....	30
1.8.1 mongod 进程 .....	10	第 4 章 增改删操作 .....	31
1.8.2 mongo 进程 .....	10	4.1 插入语句 .....	31
1.8.3 其他进程 .....	10	4.2 修改语句 .....	32
1.9 适合哪些业务 .....	11	4.3 删除语句 .....	33
1.10 小结 .....	14	4.4 锁机制 .....	33
第 2 章 查询语言系统 .....	15	4.5 小结 .....	34
2.1 查询选择器 .....	16		
2.2 查询投射 .....	18		
2.3 数组操作 .....	19		

## 第二部分 深入理解 MongoDB

第 5 章 Journaling 日志功能 .....	37	6.2 MapReduce 模式聚集 .....	44
5.1 两个重要的存储视图 .....	37	6.3 简单聚集函数 .....	47
5.2 Journaling 工作原理 .....	38	6.4 小结 .....	49
5.3 小结 .....	41	第 7 章 复制集 .....	50
第 6 章 聚集分析 .....	42	7.1 复制集概述 .....	50
6.1 管道模式进行聚集 .....	42		

7.2 复制集工作机制 .....	55	8.2.1 使集合分片 .....	74
7.2.1 数据同步 .....	55	8.2.2 集群平衡器 .....	79
7.2.2 故障转移 .....	58	8.2.3 集群的写与读 .....	80
7.2.3 写关注 .....	63	8.2.4 片键选择策略 .....	86
7.2.4 读参考 .....	65	8.3 小结 .....	87
7.3 小结 .....	66		
<b>第 8 章 分片集群 .....</b>	<b>67</b>	<b>第 9 章 分布式文件存储系统 .....</b>	<b>88</b>
8.1 分片部署架构 .....	67	9.1 小文件存储 .....	88
8.2 分片工作机制 .....	73	9.2 GridFS 文件存储 .....	90
		9.3 小结 .....	94

### 第三部分 监控与管理 MongoDB

<b>第 10 章 管理与监控 .....</b>	<b>97</b>	10.3.3 Web 控制台监控 .....	112
10.1 数据的导入导出 .....	97	10.4 小结 .....	115
10.2 备份与恢复 .....	100		
10.2.1 单节点 dump 备份与 恢复 .....	101	<b>第 11 章 权限控制 .....</b>	<b>116</b>
10.2.2 集群 dump 备份恢复 策略 .....	102	11.1 权限控制 API .....	116
10.3 监控 .....	103	11.1.1 针对所有数据库的 角色 .....	116
10.3.1 数据库角度监控 命令 .....	103	11.1.2 针对单个数据库的 角色 .....	118
10.3.2 操作系统角度监控 命令 .....	107	11.2 复制集与集群的权限控制 .....	120
		11.3 小结 .....	120

### 第四部分 应用实践

<b>第 12 章 PHP 驱动接口 .....</b>	<b>123</b>	12.4 几个重要的类、方法与 参数 .....	133
12.1 开发环境安装 .....	123	12.5 复制集上的操作 .....	136
12.2 驱动介绍 .....	126	12.6 分片集群上的操作 .....	139
12.3 单实例上的增删改查 .....	132		

12.7	分布式小文件存取操作	140	13.4	CodeIgniter 框架	162
12.8	分布式大文件存取操作	143	13.4.1	基本介绍	162
12.9	小结	145	13.4.2	下载与安装	163
<b>第 13 章</b>	<b>案例：高度定制化的</b>		13.4.3	执行原理	164
	<b>电商平台</b>	146	13.4.4	代码示范	166
13.1	功能需求	146	13.5	Bootstrap 框架	169
13.2	数据库表设计	147	13.6	前台界面原型图	170
13.3	编写 MongoDB_driver 类	152	<b>附录 常见问题</b>		174



# 第一部分 基础知识

这一部分主要介绍 MongoDB 方面的基础知识，熟悉关系数据库的读者能够快速认识到 MongoDB 是什么以及与其他数据库的区别，这一部分的基础知识很重要，贯穿整本书，建议多实践和测试。

第 1 章 本章介绍了大数据、云计算的基本概念以及云存储与 MongoDB 的关系，还介绍了 MongoDB 是什么、它的特点以及如何在各种平台上部署 MongoDB 等，最后介绍了 MongoDB 部署启动后一些关键的进程。

第 2 章 本章介绍了各种查询操作，这是数据库上最常用的一个操作。MongoDB 的查询与关系数据库的语法区别很大，但它们很多设计思想是相同的，查询选择器相当于关系数据库中经常用到的 where 语句，查询选项相当于过滤出需要返回的字段。最后介绍了一种特殊对象的查询操作，这在关系数据库中是没有的。

第 3 章 本章介绍了查询用到的索引以及利用索引对查询的优化，这个思想和关系数据库也是一致的，利用索引来提高查询效率。

第 4 章 本章介绍了对 MongoDB 插入、删除、修改操作，至此一系列完整的增删改查的操作都介绍完了，对于一般的应用程序开发都能支持了。



# 第 1 章

## 大数据与云计算

### 1.1 什么是大数据

对于各种规模大小的组织机构而言，由于数据爆炸式的增长，传统的数据处理技术变得越来越难适应，需要有变革的技术来存储、分析这些大数据。谁能够掌握这些存储、分析技术，谁就有可能成为未来市场的主导者。财富 500 强公司在这个方面已走在前列，他们认识到大数据不仅仅是一门技术，而且是未来商业的发展趋势，并且已经开始从创新的大数据业务中受益。例如，企业能够分析用户的 Web 点击习惯，总结出用户喜好，进而有针对性地开展促销；政府部门能够利用大数据预测疾病的传播趋势，进而提前进行干预。

具体来说，大数据技术涉及到数据的创造、存储、获取和分析，数据的主要特点有以下几个。

**数据量大。**一个典型的 PC 机在 2000 年前后其存储空间可能有 10GB，今天 Facebook 一天增加的数据量就将近有 500TB；一架波音 737 的飞机围绕美国飞行一周将会产生 240TB 的数据；移动互联网的发展，智能手机的普及，人们每时每刻都在产生数以百万计的数据。

**数据变化快。**高速的股票交易市场，产生的数据以微秒计算；基础设施系统、实施系统每秒都产生大量的变化的日志，每秒都处理大量的并发。

**数据多样性。**大数据的类型不仅仅是简单的数字、日期和字符串，它可能包含地理数据、3D 数据、音视频以及无结构的文档，而且这么多类型的数据可能需要保存在一起。

大数据技术的战略意义不仅在于掌握庞大的数据信息，而且也在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键在于提高对数据的“存储和加工能力”，通过“加工”实现数据的“增值”。大数据技术

能够利用修改过的硬件取代原来高消耗和昂贵的老系统。由于许多大数据技术是开源的，它们实施起来更快且更便宜，例如，将它的数据存储技术迁移到 MongoDB 上来。

## 1.2 什么是云计算

云计算的定义有多种说法，对于到底什么是云计算，我们至少可以找到 100 种解释。目前广为接受的是美国国家标准与技术研究院定义：云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池（资源包括网络、服务器、存储、应用软件、服务），这些资源能够被快速提供，只需投入很少的管理工作，或服务供应商进行很少的交互，本质上就是虚拟化技术的延伸，以服务的形式提供客户。按照服务的形式，目前主要有如下 3 种形式的云计算。

### 1. IaaS：基础设施即服务

IaaS (Infrastructure-as-a-Service)：基础设施即服务。消费者通过 Internet 可以从完善的计算机基础设施获得服务，例如硬件服务器租用。

### 2. SaaS：软件即服务

SaaS (Software-as-a-Service)：软件即服务。它是一种通过 Internet 提供软件的模式，用户无需购买软件，而是向提供商租用基于 Web 的软件，来管理企业经营活动。例如：阳光云服务器。

### 3. PaaS：平台即服务

PaaS (Platform-as-a-Service)：平台即服务。PaaS 实际上是指将软件研发的平台作为一种服务，以 SaaS 的模式提交给用户。因此，PaaS 也是 SaaS 模式的一种应用。但是 PaaS 的出现可以加快 SaaS 的发展，尤其是加快 SaaS 应用的开发速度，例如软件的个性化定制开发。

## 1.3 大数据与云计算

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式计算架构。它的特色在于对海量数据的

挖掘，但它必须依托云计算的分布式处理，也就说大数据就像做饭用的一堆原材料，云计算就像做饭用的工具。云计算解决了大数据的运算工具问题，而对大数据的存储我们需要相应的云存储工具。云存储是在云计算概念上延伸和发展出来的一个新的概念，是指通过集群应用或分布式文件系统等功能，将网络中大量的存储设备通过应用软件集合起来协同工作，共同对外提供数据存储和业务访问功能的一个系统。所以云存储是一个以数据存储和管理为核心的云计算系统，本书介绍的 MongoDB 就可以当作一个云存储系统使用。

## 1.4 什么是 MongoDB

MongoDB 是一个可扩展、开源、表结构自由、用 C++ 语言编写且面向文档的数据库，旨在为 Web 应用程序提供高性能、高可用性且易扩展的数据存储解决方案。

MongoDB 是一个介于关系数据库和非关系数据库之间的产品，是非关系数据库当中功能最丰富、最像关系数据库的 NoSQL 数据库；它支持的查询语言非常强大，其语法有点类似于面向对象的查询语言，可以实现类似关系数据库单表查询的绝大部分功能，而且还支持对数据建立索引。

MongoDB 不是在实验室里面凭空想象出来的产品，它是 10gen 公司的工程师根据实际的需求而设计的，主要基于以下几点考虑，需要一种新的数据库技术来满足数据存储层的水平扩展，而且要容易开发，能够存储海量的数据；一种非关系的结构是使数据库能支持水平扩展的最好方案；文档数据模型（BSON）容易编码和管理，将内部相关的数据放在一起能够提高数据库的操作性能。

MongoDB 服务端可运行在 Linux、Windows 或 OS X 平台，支持 32 位和 64 位应用，默认监听端口为 27017。MongoDB 的内存管理依赖于操作系统的自动内存管理机制，而且通过 Map 对数据文件进行内存映射，因此推荐 MongoDB 运行在 64 位平台上，否则在 32 位模式受虚拟内存地址大小的限制，而且运行时支持的最大文件尺寸也只能为 2GB。当然对于测试和开发环境我们可以在 32 位模式下进行，生产环境上最好是部署在 64 位上。

MongoDB 发展迅速，无疑是当前 NoSQL 领域的人气王，就算与传统的关系数据库比较也不甘落后，数据库知识网站 DB-Engines 根据搜索结果对 223 个数据库系统进行流行度排名，2014 年 7 月的数据库流行度排行榜前 12 名如图 1-1 所示。

我们可以看到前三甲依然是 Oracle、MySQL 和微软的 SQL Server，值得关注的是，第

五名 MongoDB 与第四名 PostgreSQL 之间的积分差距已不足 1 分。前四名由于历史原因都是关系数据库，许多大型的垄断行业仍然在使用这些关系数据库。

Rank	Last Month	DBMS	Database Model	Score	Changes
1.	1.	Oracle	Relational DBMS	1485.12	-15.80
2.	2.	MySQL	Relational DBMS	1295.78	-13.77
3.	3.	Microsoft SQL Server	Relational DBMS	1246.60	+22.81
4.	4.	PostgreSQL	Relational DBMS	239.46	-0.53
5.	5.	MongoDB	Document store	238.78	+7.33
6.	6.	DB2	Relational DBMS	202.01	+3.98
7.	7.	Microsoft Access	Relational DBMS	144.62	+2.26
8.	8.	SQLite	Relational DBMS	91.16	+1.97
9.	↑	10. Sybase ASE	Relational DBMS	83.72	+3.03
10.	↓	9. Cassandra	Wide column store	81.58	-0.26

图 1-1 2014 年 7 月 DB-Engines 上的数据库排行榜

MongoDB 只通过 6 年时间就将公司市值发展到 12 亿美元，其成果相当于著名开源公司 Red Hat 20 年的发展。MongoDB 的成功之路，一大部分归功于 Web 开发者。作为一个面向文档数据库，在许多场景下它都优于 RDBMS，同时还可以获得非常高的读写性能。此外，动态、灵活的模式更可以让用户在商用服务器上轻松地进行横向扩展。

## 1.5 大数据与 MongoDB

大数据意味着新的机会，企业能够创造新的商业价值。MongoDB 这样的数据库可以支撑很多大数据系统，它不仅可以作为实时的可操作的大数据存储系统，也能在离线大数据分析系统中使用。利用 MongoDB 作为大数据的云存储系统，企业能够在全世界范围内存储更多的数据，吸引更多的用户，挖掘更多用户的喜好，创造更多的价值。

选择正确的大数据存储技术，对使用者的应用和目标是非常重要的。MongoDB 公司提供的产品和服务能让使用者担更少的风险、花更少的精力提供更好的生产系统产品。事实上，MongoDB 天生就是为云计算而生的，其原生的可扩展架构，通过启用分片和水平扩展，能提供云存储所需的技术；此外，它的自动管理被称为“副本集”的冗余服务器，以保持数据的可用性和完整性。MongoDB 目前已经成为多家领先的云计算供应商，其中包括亚马逊网络服务、微软和 SoftLayer 等。

MongoDB 还支持 Google 提出的 MapReduce 并行编程模式，为大数据的分析提供了强有力的保障。MongoDB 同时提供了与 Hadoop 的接口，与其他第三方数据分析工具完美结合。

## 1.6 MongoDB 特点

它的存储模型与关系数据库的比较如表 1-1 所示。

表 1-1 MongoDB 存储模型与 MySQL 的对比

关系数据库 ( MySQL )	MongoDB
database	database
table	collection
row	document/object

关系数据库中最基本的单元是行，而 MongoDB 中最基本存储单元是 document，典型结构如下所示。

```
{
  "_id" : ObjectId("51e0c391820fdb628ad4635a"),
  "author" : { "name" : "Jordan","email" : "Jordan@123.com" },
  "postcontent" : "jordan is the god of basketball",
  "comments" : [
    {"user" : "xiaoming", "text" : "great player"},
    { "user" : "xiaoliang", "text" : "nice action" }
  ]
}
```

它用与 JSON 格式类似的键值对来存储（在 MongoDB 中叫 BSON 对象），其中值的数据类型有常见的字符串、数字、日期，还可以是 BSON 对象、数组以及数组的元素，也可以是 BSON 对象，通过这种嵌套的方式，使 MongoDB 的数据类型变得相当丰富。

MongoDB 与传统关系数据库还有一个重大区别就是：可扩展的表结构。也就是说 collection（表）中的 document（一行记录）所拥有的字段（列）是可以变化的，下面文档对象 document（一行记录）比上面列出的文档对象 document（一行记录）多一个 time 字段，但它们可以共存在同一个 collection（表）中。

```
{
  "_id" : ObjectId("51e0c391820fdb628ad4635a"),
  "author" : { "name" : "Jordan", "email" : "Jordan@123.com" },
  "postcontent" : "jordan is the god of basketball",
  "comments" : [
    { "user" : "xiaoming", "text" : "great player" },
    { "user" : "xiaoliang", "text" : "nice action" }
  ],
  "time" : "2013-07-13"
}
```

MongoDB 查询语句不是按照 SQL 的标准来开发的，它围绕 JSON 这种特殊格式的文档型存储模型开发了一套自己的查询体系，这就是现在非常流行的 NoSQL 体系。关系数据库中常用的 SQL 语句在 MongoDB 中都有对应的解决方案。当然也有例外，MongoDB 不支持 JOIN 语句。我们知道传统关系数据库中 JOIN 操作可能会产生笛卡尔积的虚拟表，消耗较多系统资源，而 MongoDB 的文档对象集合 `collection` 可以是任何结构，我们可以通过设计较好的数据模型尽量避开这样的操作需求。如果真的需要从多个 `collection`（表）中检索数据，那我们可以通过多次查询得到。

在关系数据库中经常用到的 `group by` 等分组聚集函数，在 MongoDB 中也有，而且 MongoDB 提供了更加强大的 MapReduce 方案（GOOGLE 提出的并行编程），为海量数据的统计、分析提供了便利。

MongoDB 支持日志功能 `Journaling`，对数据库的增、删、改操作会记录在日志文件中。MongoDB 每 100ms 将内存中的数据刷到磁盘上，如果意外停机，在数据库重新启动时，MongoDB 能通过 `Journaling` 日志功能恢复。

MongoDB 支持复制集（`Replset`），一个复制集在生产环境中最少需要 3 台独立的机器（测试的时候为了方便可能都部署在一台机器上），一台作主节点（`primary`），一台作次节点（`secondary`），一台作仲裁节点（只负责选出主节点），备份、自动故障转移，这些特性都是复制集支持的。

MongoDB 支持自动分片 `Sharding`，分片的功能实现海量数据的分布式存储，分片通常与复制集配合起来使用，实现读写分离、负载均衡，当然如何选择片键是实现分片功能的关键。如何实现读写分离我们后面会详细分析。

总之，MongoDB 最吸引人的地方应该就是自由的表结构、MapReduce、分片、复制集，通过这些功能实现海量数据的存储、高效地读写以及数据的分析。



## 1.7 安装 MongoDB

MongoDB 官方已经提供了 Linux、Windows、Mac OS X 以及 Solaris 4 种平台的二进制分发包，最新的稳定版本是 2.6.3，下载地址是：<http://www.mongodb.org/downloads>，如图 1-2 所示。



图 1-2 各平台二进制分发包

下载完成后，解压，我们就能直接运行里面的二进制文件，这里所讨论的安装 MongoDB，一般指的是运行 MongoDB 服务器端的进程 `mongod`。

解压后，在 `bin` 目录下，我们可以看到一个名为 `mongod.exe` 的可执行程序，这个就是服务器端进程对应的程序。因为 MongoDB 启动时需要指定数据文件所在的目录，所以先要建立一个保存数据文件的目录，如 `D:\mongodb-win32-i386-2.6.3\test_single_instance\data`；启动时也可以指定一个日志文件，如 `D:\mongodb-win32-i386-2.6.3\test_single_instance\logs\123.log`，我们通过以下命令就可以启动。

```
> mongod --config E:\MongoDB-win32-i386-2.6.3\test_single_instance\123.conf
```

上述步骤在 Linux 平台上也是一样的，只不过要注意目录和文件的读写权限。

还有一种安装方式就是直接通过各 Linux 分发版本对应的包管理器，如 RedHat、Debian、Ubuntu 等都有自己的包管理器，通过包管理器安装时，系统会自动创建数据目录和日志文件，找到这些目录和文件所在的位置，后续分析问题可能会经常要读取日志文件。

## 1.8 几个重要的进程介绍

通过官网下载的二进制包中有几个重要的可执行文件，这些可执行文件运行后都会对