异步图书
www.epubit.com

Prometheus 技术秘笈

百里燊 编著

- 畅销图书作者再出新作，从原理到应用清晰透彻地阐述 Prometheus 技术难点
- 层层深入、循序渐进，配合实际应用场景对 Prometheus 技术进行系统介绍
- 入门到实战，一本书阐述 Prometheus 生态中核心组件的工作原理以及核心实现

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

人民邮电出版社
北京

图书在版编目（CIP）数据

Prometheus技术秘笈 / 百里桑编著. -- 北京 : 人民邮电出版社, 2019.12
ISBN 978-7-115-52156-9

Ⅰ. ①P… Ⅱ. ①百… Ⅲ. ①计算机监控系统 Ⅳ. ①TP277.2

中国版本图书馆CIP数据核字(2019)第220248号

内 容 提 要

Prometheus 是一款当前迅速崛起的新兴监控系统。本书主要以 Prometheus 2.5.0 版本为基础进行介绍。全书分为 11 章，从 Prometheus 的基础入手，系统地介绍了 Prometheus 配置、Prometheus TSDB、scrape 模块、storage 模块、HTTP API 接口、PromQL 语句、Rule 配置、Discovery、AlertManager 以及 Client 等内容，读者阅读本书后，将会全面了解并掌握 Prometheus 的原理与应用，并在实际场景中进行实践。

本书适合监控运维人员 、Prometheus 二次开发人员 、Golang 工程师以及时序数据库开发人员阅读。

◆ 编　　著　百里桑
责任编辑　陈聪聪
责任印制　焦志炜

◆ 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
大厂聚鑫印刷有限责任公司印刷

◆ 开本：800×1000　1/16
印张：23.5
字数：423 千字　　2019 年 12 月第 1 版
印数：1－2 400 册　　2019 年 12 月河北第 1 次印刷

定价：89.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

前言

无论是在互联网公司还是在传统IT公司，监控系统都占有非常重要的地位。运维人员通过监控数据以及告警通知可以实时了解系统的运行状态，开发人员可以通过监控数据快速定位系统的性能瓶颈。

目前在市场上除Zabbix等老牌的监控系统之外，新兴监控系统也逐步崛起，例如本书将要介绍的Prometheus、小米公司开源的Open-Falcon等。另外，很多时序数据库（如InfluxDB和OpenTSDB等）也被作为自研监控系统的存储层。相信读者即使在生产实践中没有接触过时序数据库，也一定对其有所耳闻。例如，在2016年，百度云在其物联网平台上发布了国内首个多租户的分布式时序数据库产品TSDB；阿里云在“2017云栖大会•上海峰会”上发布了面向物联网场景的高性能时间序列数据库HiTSDB等。时序数据库作为物联网中的基础设施之一，得到了各个互联网巨头企业的重视，其热门程度可见一斑。

与其说Prometheus是一个监控系统，不如说Prometheus提供了一个完备的监控生态。本书将会深入介绍Prometheus生态中的核心组件，例如Prometheus TSDB、Prometheus Server、AlertManager和Client等，并且详细剖析这些核心组件的工作原理以及核心实现。

如何阅读本书

由于篇幅限制，本书并没有详细介绍Go语言的基础知识，但为了便于理解Prometheus的实现细节，需要读者对Go语言的基本语法有一定的了解。

本书共分为11章，主要从源码角度深入剖析Prometheus各个组件的核心原理和代码实现。各章之间的内容相对独立，对Prometheus有一定了解的读者可以有目标地选择合适的章节开始阅读，当然也可以从第1章开始向后逐章阅读。本书主要以Prometheus的2.5.0版本为基础进行介绍。

第1章首先介绍了Prometheus在时序数据库以及监控领域所处的地位，然后详细介绍了InfluxDB、Graphite、OpenTSDB和Open-Falcon的架构特点及其优缺点。接下来对Prometheus生态的核心架构和关键组件的功能进行概述，对其各个核心组件的功能进行了简要说明。最后还介绍了Prometheus的安装流程、源码环境的搭建流程以及Grafana接入Prometheus的操作步骤。

第2章介绍prometheus.yml配置文件中核心的配置项，其中结合示例介绍了各核心配置项的含义和功能。

第3章介绍了Prometheus TSDB，它是Prometheus本地时序存储的实现，也是Prometheus的核心模块。这里首先阐述Facebook Gorilla论文的核心思想，该思想是Prometheus TSDB实现的基础。之后介绍了Prometheus TSDB在磁盘上的目录与文件的组织方式、含义和功能。随后详细分析了Prometheus TSDB时序存储的核心实现，其中介绍了Chunk接口实现、Meta元数据结构以及读写时序数据用到的ChunkWriter和ChunkReader的实现等内容。

接下来介绍的是Prometheus TSDB中的index文件，深入剖析了index文件中各个部分内容的读写流程。之后对Prometheus TSDB使用的WAL日志文件的物理结构和逻辑结构进行了深入分析，同时详细分析了Checkpoint机制的相关内容。随后介绍了Prometheus TSDB如何通过tombstones文件实现“标记删除”功能。

在第3章中还对Prometheus TSDB的压缩计划生成以及具体压缩操作的执行逻辑和实现进行了全方位的剖析。最后，深入介绍了Prometheus TSDB中内存Head窗口涉及的基础组件以及内存Head窗口中数据的存储方式、读写等内容。

第4章介绍了Prometheus中的scrape模块，主要涉及scrape模块如何根据prometheus.yml文件中的配置信息周期性地从客户端、exporter或PushGateway抓取时序数据，以及Relabel操作的具体实现。

第5章介绍了Prometheus Server中的storage模块，该模块的核心功能是对本地存储和远程存储进行封装和适配。

第6章介绍了Prometheus Server中V1版本的HTTP API接口，该版本主要提供了执行PromQL语句、查询时序元数据、根据Label Name查询Label Value、查询target和查询Rule的功能。另外，HTTP API接口还提供了一些Admin管理的功能。

第7章详细介绍了PromQL语句的执行流程，其中涉及PromQL的解析、抽象语法树中每个节点的执行流程等内容。

第8章介绍了Prometheus Server中与Rule相关的模块。首先介绍了Recording Rule配置以及Alerting Rule配置在内存中的抽象，以及Prometheus如何管理这些Rule配置；然后介绍了Recording Rule以及Alerting Rule的执行流程；最后分析了notifier模块的实现，它的核心逻辑是将Alerting Rule产生告警的时序信息发送到AlertManager集群。

第9章介绍了Prometheus Server中discovery模块的核心接口和实现。discovery模块负责接入多种服务发现组件，让Prometheus Server能够动态发现target信息以及AlertManager信息。

第10章介绍了AlertManager。首先介绍了AlertManager中核心模块的功能以及整个AlertManager的核心架构；随后的章节深入分析了AlertManager中每个核心模块的工作原理和具体实现。

第11章介绍了Prometheus Client（Golang版本）的核心原理以及相关实现。首先介绍了Prometheus Client中的4种基本数据类型的特点和使用场景；然后以Gauge为例，深入分析了Prometheus Client记录监控的思想以及涉及的核心组件；接下来以Node Exporter为例介绍了Exporter大致实现原理。

如果读者在阅读本书的过程中，发现任何不妥之处，请将您宝贵的意见和建议发送到邮箱shen_baili @163.com，也欢迎读者朋友通过此邮箱与我进行交流。

关于作者

百里燊，硕士研究生毕业，小时候想成为闯荡江湖的侠客，结果着迷于代码，最终成为辛勤工作的程序员。目前关注各种开源时序数据库，期待与大家共同进步。联系邮箱：shen_baili@163.com。

致谢

感谢人民邮电出版社的陈聪聪老师，是您的辛勤工作让本书的出版成为可能。同时还要感谢许多我不知道名字的幕后工作人员为本书付出的努力。

感谢三十在技术上提供的帮助。

感谢三白和陈默同学对我的鼓励和支持。

感谢冯玉玉同学和李成伟同学，你们是我生活中的灯塔。

感谢我的母亲，谢谢您的付出和牺牲！

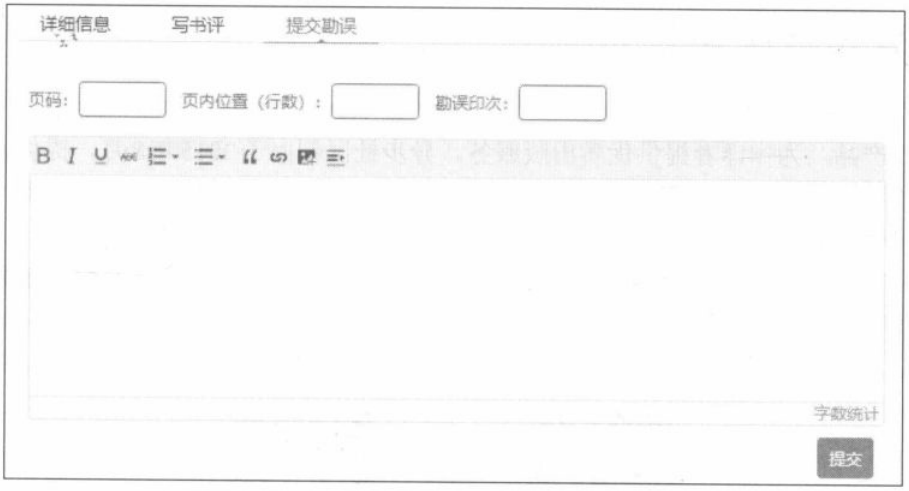
资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，点击“提交勘误”，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的100积分。积分可用于在异步社区兑换优惠券、样书或奖品。



扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问www.epubit.com/selfpublish/submission即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下IT专业图书社区，致力于出版精品IT技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于2015年8月，提供大量精品IT技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网https://www.epubit.com。

“异步图书”是由异步社区编辑团队策划出版的精品IT专业图书的品牌，依托于人民邮电出版社近30年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

目录

第1章
Prometheus基础入门

Prometheus是一款时下比较先进的时序数据库，也被认为是下一代的监控系统。在时序数据库2019年3月的排名中，Prometheus排名已经超越老牌时序数据库OpenTSDB，跃居第5名的位置，如图1-1所示。

☐ include secondary database models　　30 systems in ranking, March 2019

Rank Mar 2019	Rank Feb 2019	Rank Mar 2018	DBMS	Database Model	Score Mar 2019	Score Feb 2019	Score Mar 2018
1.	1.	1.	InfluxDB	Time Series	16.17	+0.41	+5.53
2.	2.	2.	Kdb+	Time Series, Multi-model	5.60	+0.19	+2.50
3.	3.	↑4.	Graphite	Time Series	3.07	+0.12	+1.00
4.	4.	↓3.	RRDtool	Time Series	2.75	+0.05	-0.33
5.	5.	↑6.	Prometheus	Time Series	2.72	+0.21	+1.67
6.	6.	↓5.	OpenTSDB	Time Series	2.28	+0.04	+0.26
7.	7.	7.	Druid	Multi-model	1.57	+0.08	+0.58
8.	8.	↑16.	TimescaleDB	Time Series	0.91	+0.03	+0.82
9.	9.	↓8.	KairosDB	Time Series	0.66	+0.14	+0.21
10.	↑11.	↑13.	FaunaDB	Multi-model	0.52	+0.16	+0.41

图1-1

1.1　时序数据库对比

下面将从数据存储和监控系统两个层面介绍一下常见的时序数据库和监控系统，帮助读者迅速了解它们的特性，方便读者根据自己的使用场景进行选型。

1.1.1　InfluxDB简介

InfluxDB是使用Golang语言编写的一款时序数据，目前较新的版本为2.0 Alpha版本，稳定版是1.7.5版本。InfluxDB在时序数据库方面的市场占有率较大，其热度之所以如此之

高，与以下特点有着直接关系。

1. 读写性能

InfluxDB在时序数据写入、数据压缩以及实时查询等方面的表现都非常出众。InfluxDB官方网站将InfluxDB与Cassandra、Elasticsearch、MongoDB、OpenTSDB进行了性能以及磁盘占用量的比较，结果表明InfluxDB在时序数据读写性能方面，较市面上其他的数据库产品有较大的优势。

2. 支持多种接口

InfluxDB提供了多种通用接口，例如HTTP API和GRPC等；也支持多种时序数据库协议，例如Graphite、Collectd和OpenTSDB等。这就方便了时序数据的写入以及InfluxDB与其他时序产品之间的数据迁移。

3. 支持类SQL的查询语句

用户可以通过书写InfluxQL语句来查询InfluxDB中的时序数据。InfluxQL是一种类SQL的查询语言，这就降低了InfluxDB的使用门槛；同时InfluxQL也支持多种函数和表达式，方便用户实现一些高级功能。

4. 数据压缩

对于近期的时序数据，InfluxDB会保存其原始数据；对于较久的时序数据，InfluxDB会进行Downsampling处理，对数据进行聚合处理，聚合之后的时序数据精度会降低，但数据量会减少，这样就可以降低磁盘占用量，这也算是InfluxDB在数据精度和磁盘使用量之间的折中设计。另外，InfluxDB可以开启定期清理过期数据的功能，进一步释放磁盘空间。

单从时序数据的存储方面来看，InfluxDB已经非常先进，Prometheus TSDB在某些方面的设计与InfluxDB非常类似。从一个监控系统的角度来看，InfluxDB之前的相关生态比较匮乏，但是近几年InfluxData以InfluxDB为中心，打造了很多配套组件，形成了一个完整的生态系统，也被称为“TCIK Stack”，如图1-2所示。

这里简单介绍一下InfluxDB相关组件的功能。

- Telegraf：Agent组件。Telegraf用于收集各个系统产生的时序数据以及事件信息，并将其push到InfluxDB进行持久化。
- Kapacitor：流处理引擎。Kapacitor可以从InfluxDB或是Telegraf获取时序数据或时