<http://www.uml.org.cn/sjjm/2016032210.asp?artid=17785>

**编者按**作者刘斌是OneAPM后端研发工程师，拥有10多年编程经验，参与过大型金融、通信以及Android手机操作系的开发，熟悉Linux及后台开发技术。曾参与翻译过《第一本Docker书》、《GitHub入门与实践》、《Web应用安全权威指南》、《WEB+DB PRESS》、《Software Design》等书籍，也是Docker入门与实践课程主讲人。本文所阐述的「时间序列数据库」，系笔者所负责产品 Cloud Insight对性能指标进行聚合、分组、过滤过程中的梳理和总结。

通过上一章《时序列数据库武斗大会之什么是TSDB》的介绍，相信大家已经知道了什么是时序列数据库，以及对它能干什么，具有什么特点。

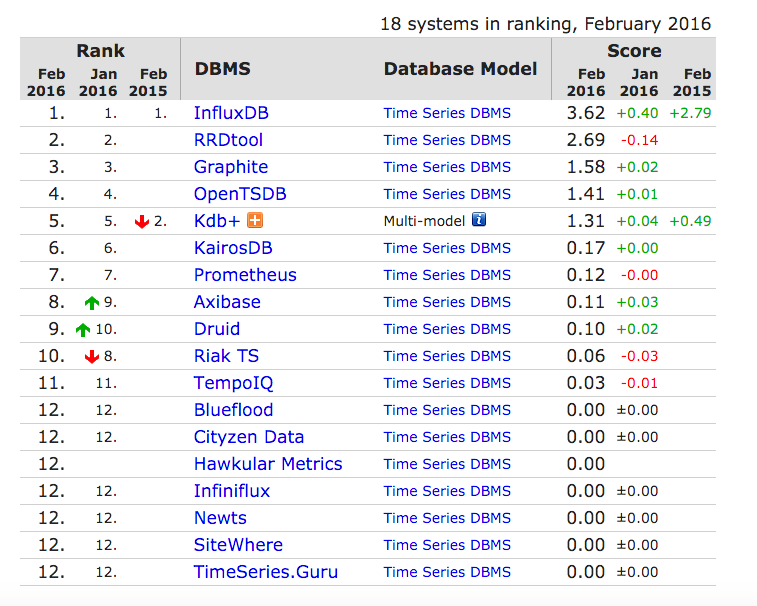
那么在这一篇文章中，我们将介绍一下目前都有哪些 TSDB，以及它们各自的特点，并基于个人观点，给出一定的（喜好）评判。

由于个人能力所限，有些地方调查可能不到位，再加上一定的个人主观因素，跟其他人的结论可能不一样，不过这应该也正常。没有调查过就没有发言权，只有真正的深度用户的发言，才具有说服务力，你权当这里就是我抛砖了。

虽然也有人用 ElasticSearch 或者 MongoDB 来存储时序列数据，作为更适合分类为 NOSQL 的这两个数据库软件，我们这里就不对它们做介绍了。

**DB-Engines 中时序列数据库排名**

我们先来看一下DB-Engines中关于时序列数据库的排名，这是当前（2016年2月的）排名情况：



下面，我们就按照这个排名的顺序，简单介绍一下这些时序列数据库中的一些。下面要介绍的 TSDB 以开源的为主，如果是商业或者 SaaS 服务，也简单介绍一下其特点，让大家能对其他领域的事物也有所了解。

这里有一个例外，就是 Pinot 并不在这个排名里，但是我也把它列在了这里。

**1. InfluxDB**

InfluxDB 由 Golang 语言编写，也是由 Golang 编写的软件中比较著名的一个，在很多 Golang 的沙龙或者文章中可能都会把 InfluxDB 当标杆来介绍，这也间接帮助 InfluxDB 提高了知名度。



InfluxDB的主要特点包括下面这些：

* schemaless(无结构)，可以是任意数量的列
* 可扩展（集群）
* 方便、强大的查询语言
* Native HTTP API
* 集成了数据采集、存储、可视化功能
* 实时数据 Downsampling
* 高效存储，使用高压缩比算法，支持retention polices

InfluxDB 是 TSDB 中为数不多的进行了用户和角色方面实现的，提供了 Cluster Admin、Database Admin 和 Database User 三种角色。

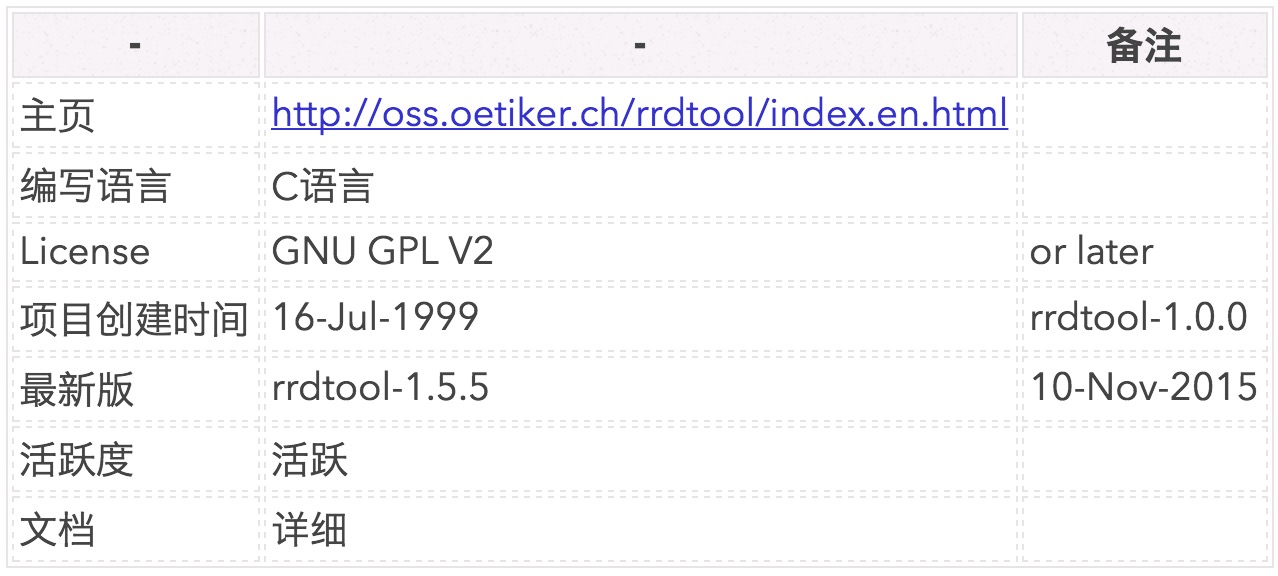
InfluxDB 的数据采集系统也支持多种协议和插件： - 行文本 - UDP - Graphite - CollectD - OpenTSDB

不过 InfluxDB 每次变动都较大，尤其是在存储和集群方面，追求平平安过日子，不想瞎折腾的可以考虑下。

注意：由于InfluxDB开发太活跃了，很可能你在网上搜到的资料都是老的，会害到你，所以你需要以官方文档为主。

**一句话总结：欣欣向荣、值得一试。**

**2. RRDtool**



RRDtool 全称为 Round Robin Database Tool，也就是用于操作 RRD 的工具，简单明了的软件名。

什么是 RRD 呢？简单来说它就是一个循环使用的固定大小的数据库文件（其实也不太像典型的数据库）。

大体来说，RRDtool 提供的主要工具如下：

* 创建RRD（rrdtool create）
* 更新RRD（rrdtool update）
* 画图（rrdtool graph）

这其中，画图功能是最复杂也是最强大的，甚至支持下面这些图形，这是其他 TSDB 中少见的：

* 指标比较，对两个指标值进行计算，描画出满足条件的区域
* 移动平均线
* 和历史数据进行对比
* 基于最小二乘法的线性预测
* 曲线预测
* 总之，它的画图功能太丰富了。

**一句话总结：老牌经典、艺多不压身。**

**3. Graphite**



Graphite 由 Orbitz, LLC 的 Chris Davis 创立于 2006 年，它主要有两个功能：

* 存储数值型时序列数据
* 根据请求对数据进行可视化（画图）

相应的，它的特点为：

* 分布式时序列数据存储，容易扩展
* 功能强大的画图Web API，提供了大量的函数和输出方式
* Graphite本身不带数据采集功能，但是你可以选择很多第三方插件，比如适用于\* collectd、Ganglia或Sensu的插件等。同时，Graphite也支持Plaintext、Pickle和AMQP这些数据输入方式。

Graphite主要由三个模块组成：

* whisper：创建、更新RRD文件
* carbon：以守护进程的形式运行，接收数据写入请求
  + carbon-cache：数据存储
  + carbon-relay：分区和复制，位于carbon-cache之前，类似carbon-cache的负载均衡
  + carbon-aggregator：数据集计，用于减轻carbon-cache的负载
* graphite-web：用于读取、展示数据的Web应用

whisper 使用了类似 RRDtool 的 RRD 文件格式，它也不像 C/S 结构的软件一样，没有服务进程，只是作为 Python library 使用，提供对数据的 create/update/fetch 操作。

如果你对它的性能比较在意，这里有一份老的数据可供参考。

Google、Etsy、GitHub、豆瓣、Instagram、Evernote 和 Uber 等很多知名公司都是 Graphite 的用户。有此背景，其可信度又加一层，而且网上的资料也相当的多，值得评估一下。

**一句话总结：群众基础好、可以参考。**

**4. OpenTSDB**



OpenTSDB 是一个分布式、可伸缩的时间序列数据库。它支持豪秒级数据采集所有 metrics，支持永久存储（不需要 downsampling），和 InfluxDB 类似，它也是无模式，以 tag 来实现维度的概念。

比如，这就是它的一个metric例子：

|  |
| --- |
| mysql.bytes\_received 1287333217 66666666 schema=foo host=db1 |

OpenTSDB 的节点称为 TSD（Time Series Daemon (TSD)），它没有主、从之分，消除了单点隐患，非常容易扩展。它主要以HBase作为存储系统，现在也增加了对 Cassandra 和 Bigtable（非云端）。

OpenTSDB 以数据存储和查询为主，附带了一个简单地图形界面（依赖Gnuplot），共开发、调试使用。

**一句话总结：好用，Cloud Insight也在用这项技术来实现对性能指标进行聚合、分组、过滤。**

**5.KDB+**



所有 TSDB 中，估计就数这个最酷了，我说的是域名，只有两个字母，猥琐地想一下，域名就值很多钱 ：-）。

kdb+是一个面向列的时序列数据库，以及专门为其设计的查询语言q（和他们的域名一样简短）。Kdb+ 混合使用了流、内存和实时分析，速度很快，支持分析 10 亿级别的记录以及快速访问TB级别的历史数据。

不过这是一个商业产品，但是也提供了免费版本（貌似还限制在32位）。

**6.KairosDB**



KairosDB 是一个 OpenTSDB 的 fork，不过是基于 Cassandra 存储的。由于 Cassandra 的行比 HBase 宽，所以 KairosDB 的 Cassandra 的默认行大小为 3 星期，而 OpenTSDB 的 HBase 则为 1 小时。

KairosDB 支持通过 Telnet、Rest、Graphite 等协议写入数据，你也可以通过编写插件自己实现数据写入。

KairosDB 也提供了基于 Web API 的查询接口，支持数据聚合、持过滤和分组等功能。

同时 KairosDB 提供了一个供开发用的 Web UI，图形绘制引擎使用了 Flot。

和 OpenTSDB 类似，KairosDB 也提供了插件机制，你可以使用插件完成如下工作：

* 添加数据点（data point）监听器
* 添加新的数据存储服务
* 添加新的协议处理程序
* 添加自定义系统监视服务

**7.Druid**



Druid 是一个快速、近实时的海量数据 OLAP 系统，并且是开源的。Druid 诞生于 Metamarkets，后来一些核心人员创立了 IMPLY 公司，进行 Druid 相关的产品开发。

Druid 会按时间来进行分区（segment），并且是面向列存储的。它的主要特性如下：

* 支持嵌套数据的列式存储
* 层级查询
* 二级索引
* 实时数据摄取
* 分布式容错架构

根据去年底 druid.io 的白皮书，现在生产环境下最大的集群规模如下：

* >3M EVENTS / SECOND SUSTAINED (200B+ EVENTS/DAY)
* 10 – 100K EVENTS / SECOND / CORE
* >500TB OF SEGMENTS (>50 TRILLION RAW EVENTS)
* >5000 CORES (>400 NODES, >100TB RAM)
* QUERY LATENCY (500MS AVERAGE)
* 90% < 1S 95% < 2S 99% < 10S
* 3+ trillion events/month
* 3M+ events/sec through Druid’s real-time ingestion
* 100+ PB of raw data
* 50+ trillion events

Druid 企业用户比较多，比如 Netflix、Paypal 等。具体可以参考http://druid.io/druid-powered.html。

Druid 架构比较复杂，因此对部署和运维也有一定的负担，比如需要的机器多、机器配置要高（尤其是内存）。

**一句话总结：好用，我们在用。**

**8.Prometheus**



Prometheus 是一个开源的服务监控系统和时序列数据库，由社交音乐平台 SoundCloud 在2012年开发，最近也变得很流行，最新版本为 0.17.0rc2。

Prometheus 从各种输入源采集 metric，进行计算后显示结果，或者根据指定条件出发报警。

和其他监控系统相比，Prometheus 的特点包括：

* 多维数据模型（时序列数据由metric名和一组key/value组成）
* 灵活的查询语言
* 不依赖分布式存储，单台服务器即可工作
* 通过基于HTTP的pull方式采集是序列数据
* 可以通过中间网关进行时序列数据推送
* 多种可视化和仪表盘支持

由于 Prometheus 采用了类似 OpenTSDB 和 InfluxDB 的 key/value 维度机制，所以如果你对任一种 TSDB 有了解的话，学习起来会简单些。

**一句话总结：貌似比较火，何不试一试？**

**9.Pinot**



Pinot 是一个开源的实时、分布式 OLAP 数据存储方案。它来自 Linkedin，虽然 Linkedin 最近估价表现很差，但是他们创建的各种软件、中间件实在太多了。这一点我们做软件的都应该向 Linkedin 表示感谢。

Pinot 就像是一个 Druid 的 copy，不过两者的灵感都来源于SenseiDB（Sensei 在日语里为老师的意思，写成汉字为“先生”）。

Pinot 也像 Druid 一样，能加载 offline 数据（Hadoop 文件）和实时数据（Kafka）。Pinot 从设计上就面向水平扩展。

Pinot 主要特点：

* 面向列
* 插拔式索引引擎：排序索引、位图索引和反向索引
* 根据查询语句和segment信息对查询/执行计划进行优化
* 从 Kafka 实时数据摄取（ingestion）
* 从 Hadoop 进行批量摄取
* 类似 SQL 的查询语言，支持聚合、过滤、分组、排序和唯一处理。
* 支持多值字段
* 水平扩展和容错

Pinot 的特点和 Druid 很像，两者可互为参考。

**一句话总结：背靠大树好乘凉。**

**小结**

这里我们为大家介绍了几种常见 TSDB，如不出意外，你可能会在这里选择某一种来使用。

尽管如此，我们还是会为大家介绍更多一些的项目，让大家能更多的了解一些不同的 TSDB 及其特点，也能帮助读者深入了解 TSDB 的各种场景，开阔思路。在下一篇文章中，我们将会为各位再介绍几种时序列数据库。