Terark 产品和技术白皮书

领先的存储引擎和数据库解决方案提供商

郭宽宽(CEO)

+86-18513000640

royguo@terark.com

TERARK

目录

公司介绍

奇简软件（北京）有限公司（Terark），创立于 2015 年，是一家专注于数据库核心存储引擎及数据库整体解决方案研发的科技公司。

公司在中国北京、美国硅谷设有研发公司，同时也是美国著名孵化器 Y Combinator 的成员企业（该孵化器曾孵化 Airbnb, Dropbox 等独角兽）。

Terark 的使命是为企业打造安全、易用、低成本以及高效的数据库基础设施，帮助企业构建强大的底层数据支撑。

www.terark.com

公司介绍

TERARK

Y Combinator

YC W17 届孵化企业

Y Combinator 是硅谷最成功的孵化器，孵化了 Dropbox、Airbnb 等知名独角兽

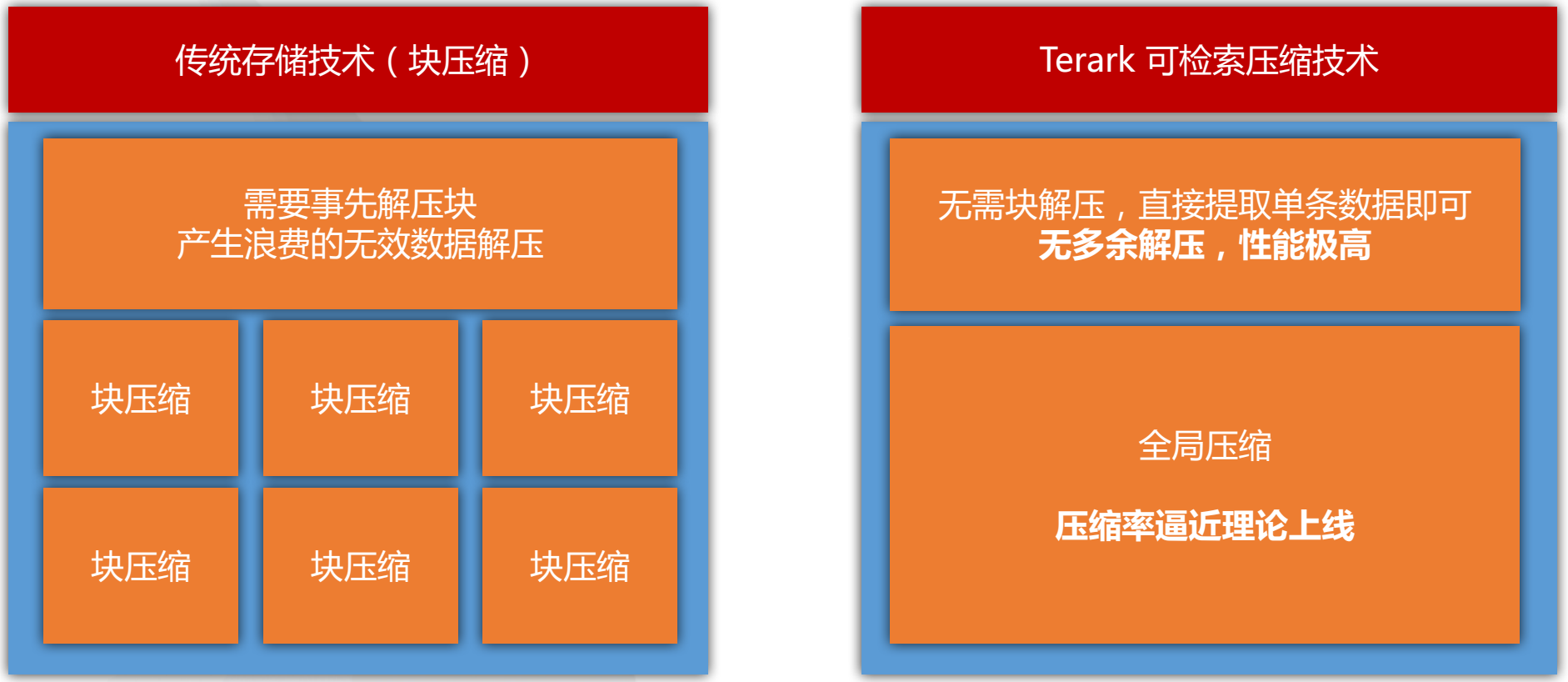
公司介绍 – 服务内容

- 数据库存储引擎商业授权
- 数据库整体解决方案建设
- 数据库性能和成本优化
- 数据库底层算法优化和咨询

标杆客户

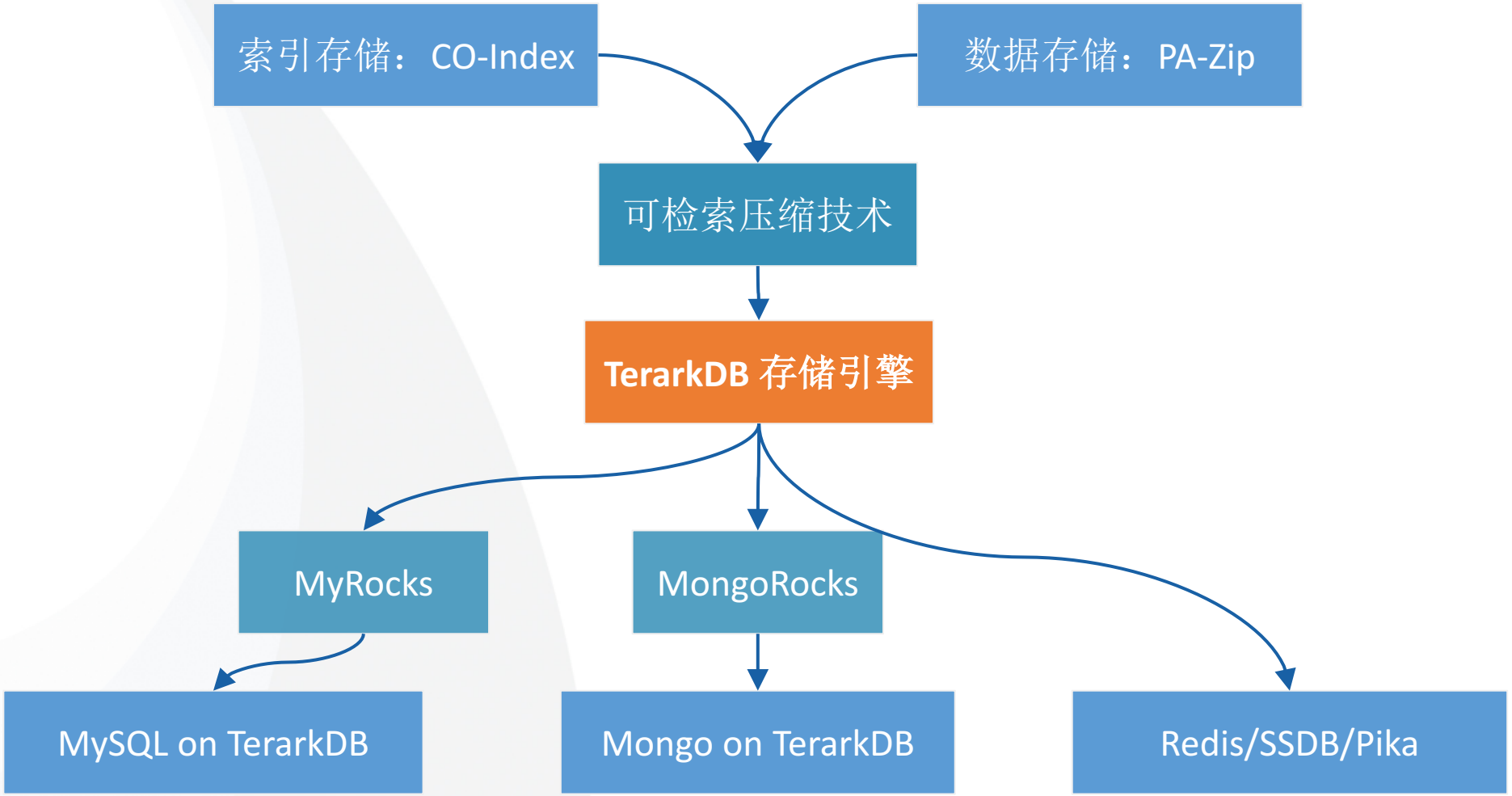
Terark 为阿里云数据库提供了底层的核心存储引擎，做为国内唯一一个商业存储引擎公司，我们的技术实力受到广泛认可。

核心技术 – 可检索压缩技术



Terark 全球首创的“可检索压缩算法”可以无需块解压即可提取单条数据，压缩率和性能双重提高，即节约用户存储成本，又提升了应用响应效率。

核心技术 – 技术栈


索引存储：CO-Index
数据存储：PA-Zip
可检索压缩技术
TerarkDB 存储引擎
MyRocks
MongoRocks
MySQL on TerarkDB
Mongo on TerarkDB
Redis/SSDB/Pika

产品体系

TerarkDB 存储引擎

- 超高的数据压缩率(3倍以上)
- 超高的随机访问性能(10倍以上)
- 成熟稳定的算法

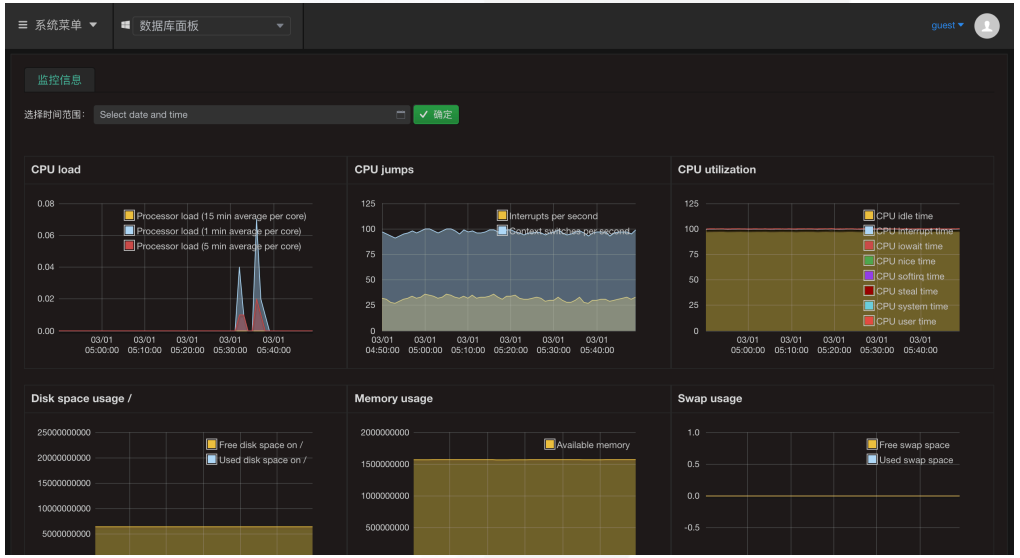
MySQL / MongoDB / Redis 数据库

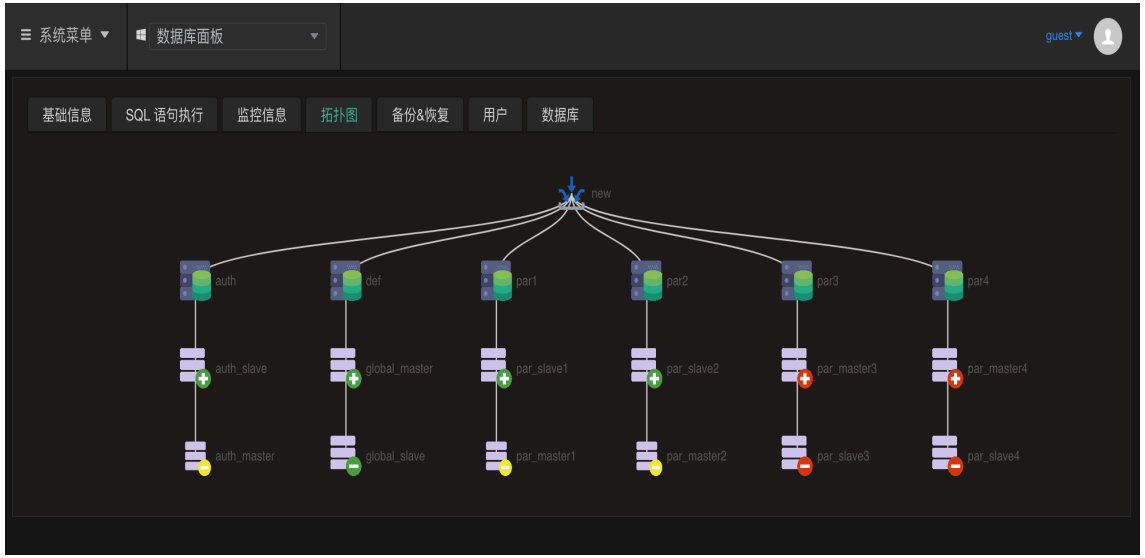
- 基于 TerarkDB 存储引擎，替换 MySQL/MongoDB 引底层擎
- 压缩率3倍以上，随机访问性能3～5倍
- 经过阿里云的实际测试和使用验证

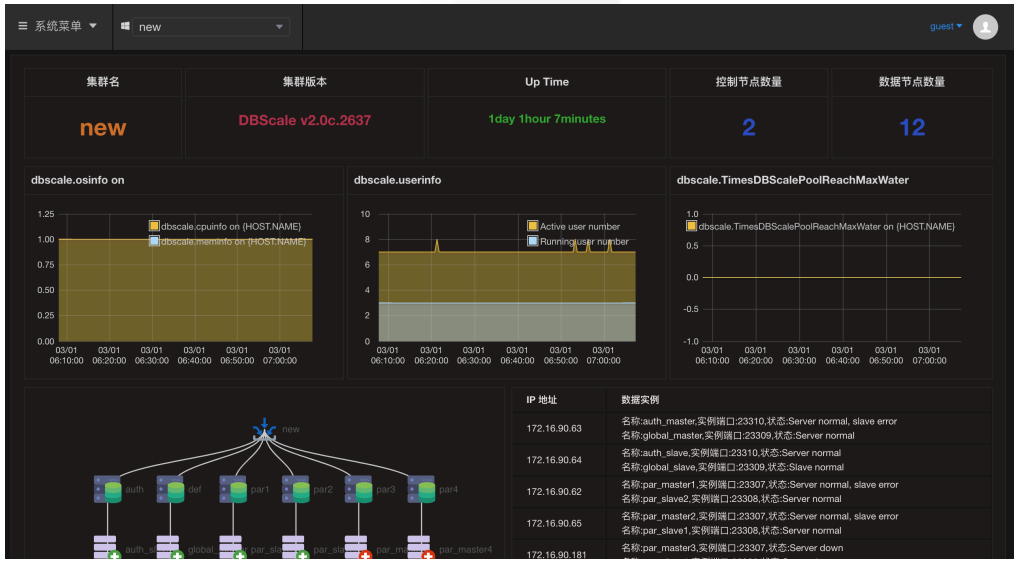
一站式数据库管理平台

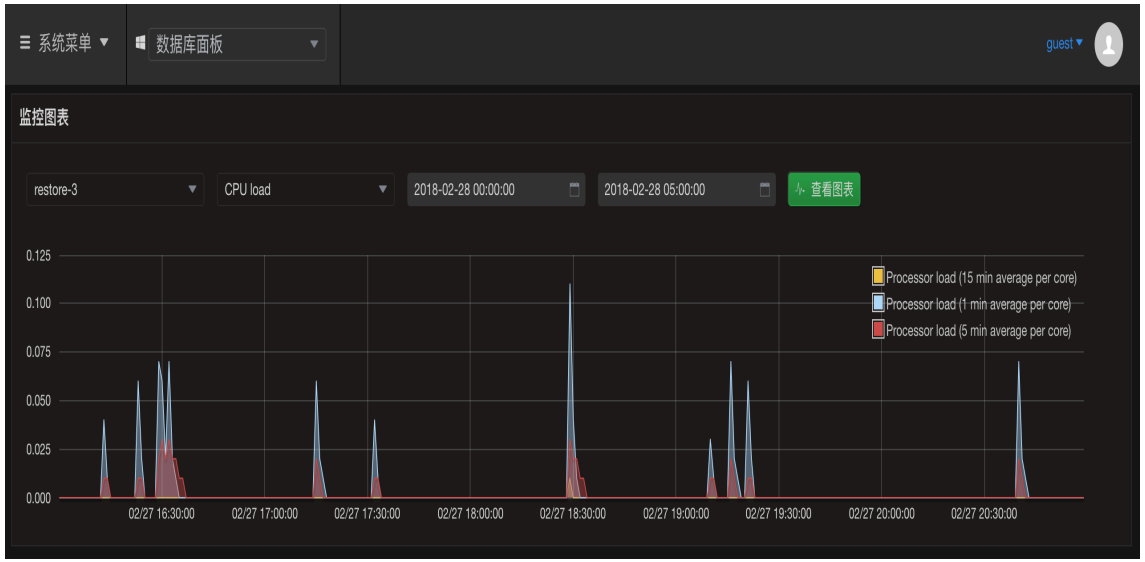
- 基于 B/S 开发模式的数据库集群管理平台

管理平台

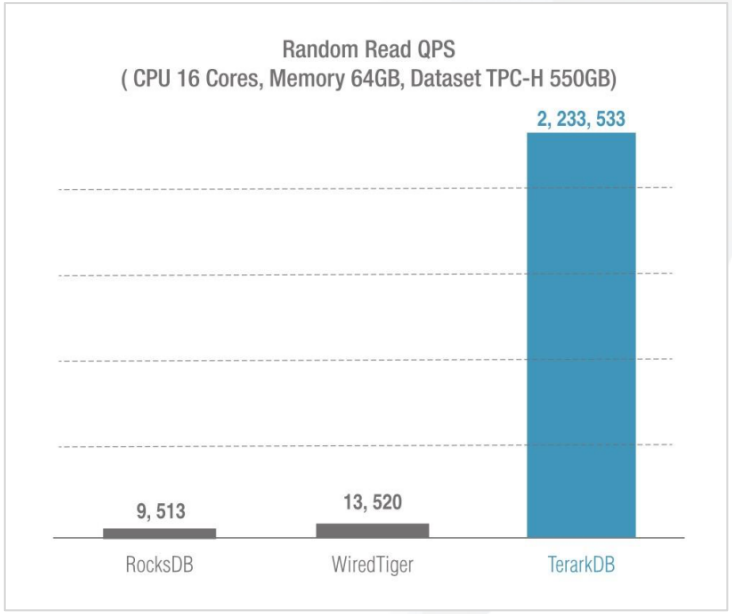

系统菜单
数据库面板
guest
监控信息
选择时间范围：
Select date and time
确定
CPU load
Processor load (15 min average per core)
Processor load (1 min average per core)
Processor load (5 min average per core)
CPU jumps
Interrupts per second
Context switches per second
CPU utilization
CPU idle time
CPU interrupt time
CPU iowait time
CPU nice time
CPU softirq time
CPU steal time
CPU system time
CPU user time
Disk space usage /
Free disk space on /
Used disk space on /
Memory usage
Available memory
Swap usage
Free swap space
Used swap space


系统菜单
数据库面板
guest
基础信息
SQL 语句执行
监控信息
拓扑图
备份&恢复
用户
数据库
new
auth
def
par1
par2
par3
par4
auth_slave
global_master
par_slave1
par_slave2
par_master3
par_master4
auth_master
global_slave
par_master1
par_master2
par_slave3
par_slave4

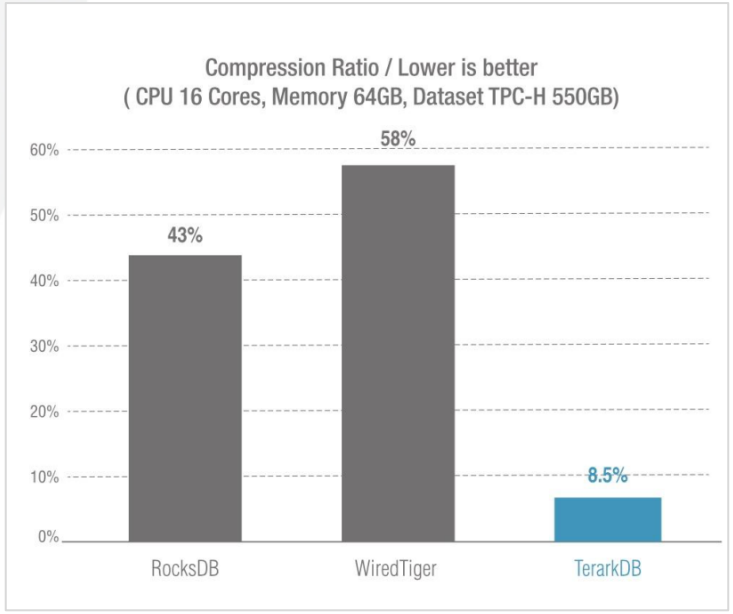

系统菜单
new
guest
集群名
集群版本
Up Time
控制节点数量
数据节点数量
new
DBScale v2.0c.2637
1day 1hour 7minutes
2
12
dbscale.osinfo on
dbscale.cpuinfo on {HOST.NAME}
dbscale.meminfo on {HOST.NAME}
dbscale.userinfo
Active user number
Running user number
dbscale.TimesDBScalePoolReachMaxWater
dbscale.TimesDBScalePoolReachMaxWater on {HOST.NAME}
IP 地址
数据实例
172.16.90.63
名称:auth_master,实例端口:23310,状态:Server normal, slave error
名称:global_master,实例端口:23309,状态:Server normal
172.16.90.64
名称:auth_slave,实例端口:23310,状态:Server normal
名称:global_slave,实例端口:23309,状态:Slave normal
172.16.90.62
名称:par_master1,实例端口:23307,状态:Server normal, slave error
名称:par_slave2,实例端口:23308,状态:Server normal
172.16.90.65
名称:par_master2,实例端口:23307,状态:Server normal, slave error
名称:par_slave1,实例端口:23308,状态:Server normal
172.16.90.181
名称:par_master3,实例端口:23307,状态:Server down


系统菜单
数据库面板
guest
监控图表
restore-3
CPU load
2018-02-28 00:00:00
2018-02-28 05:00:00
查看图表
Processor load (15 min average per core)
Processor load (1 min average per core)
Processor load (5 min average per core)
02/27 16:30:00
02/27 17:00:00
02/27 17:30:00
02/27 18:00:00
02/27 18:30:00
02/27 19:00:00
02/27 19:30:00
02/27 20:00:00
02/27 20:30:00

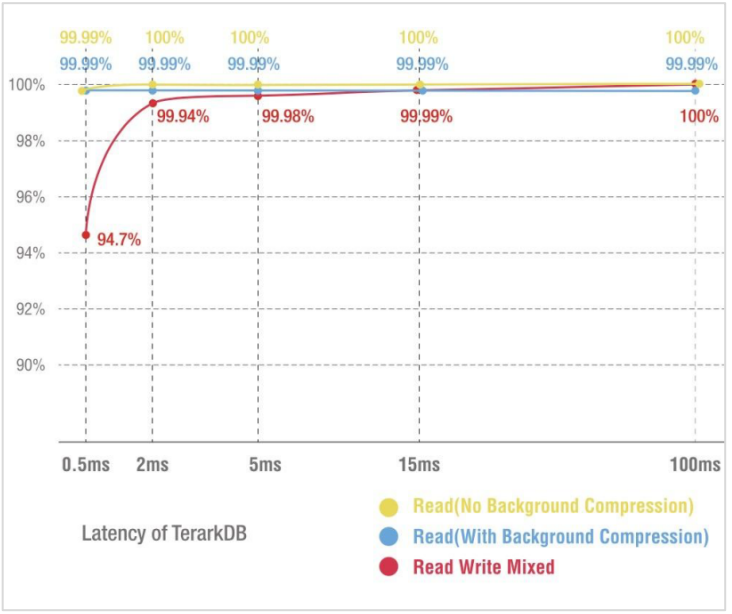
产品体系 - TerarkDB 性能指标对比



随机读性能



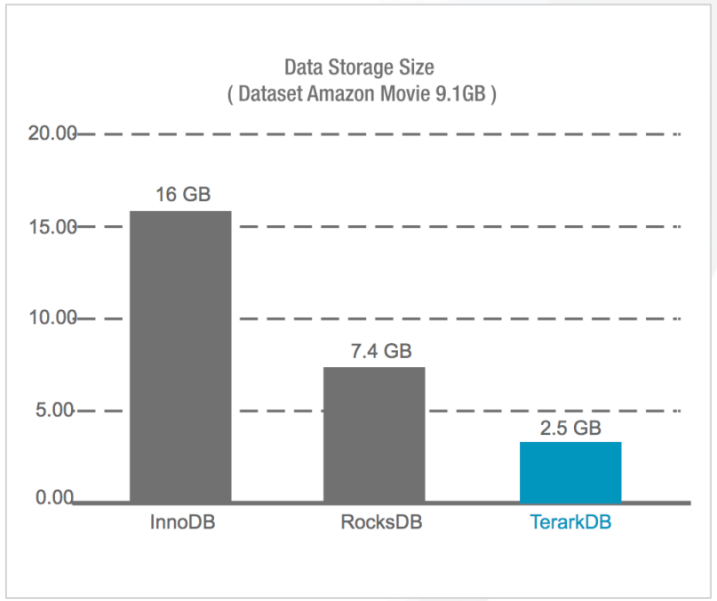
压缩率



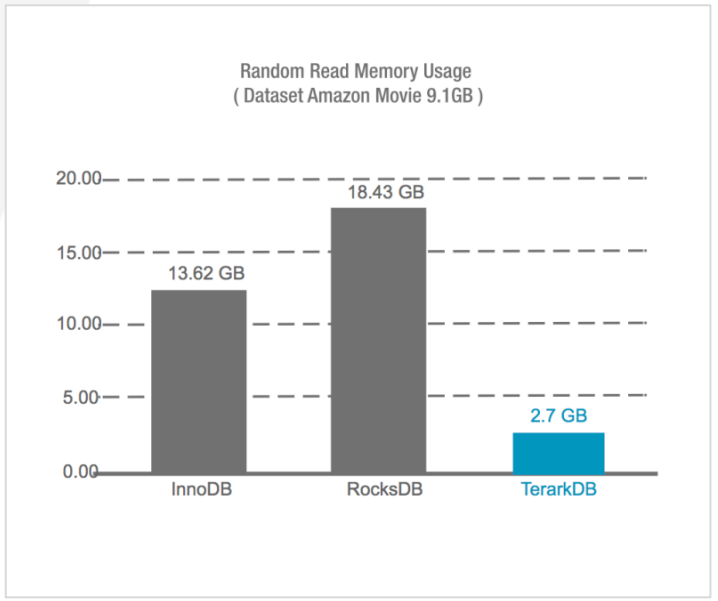
访问延迟

注：在不同的数据集和硬件环境下，测试结果可能不同

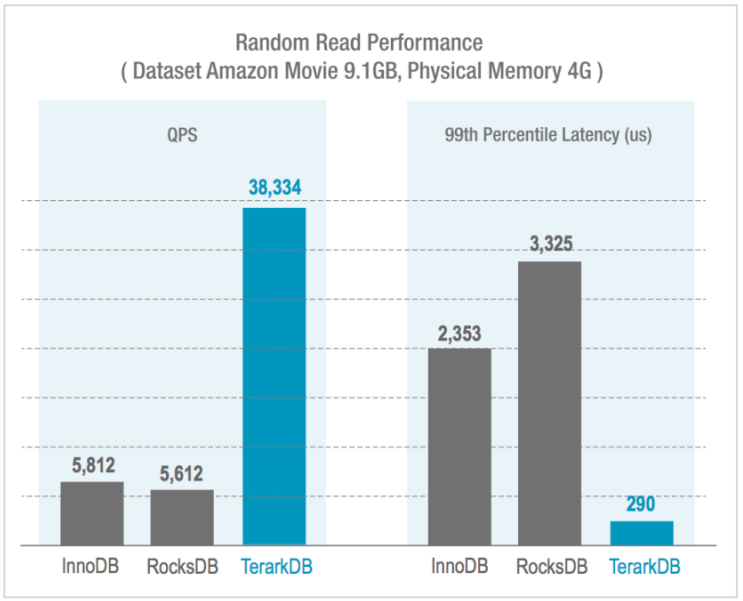
产品体系 - MySQL 性能指标对比



压缩率对比



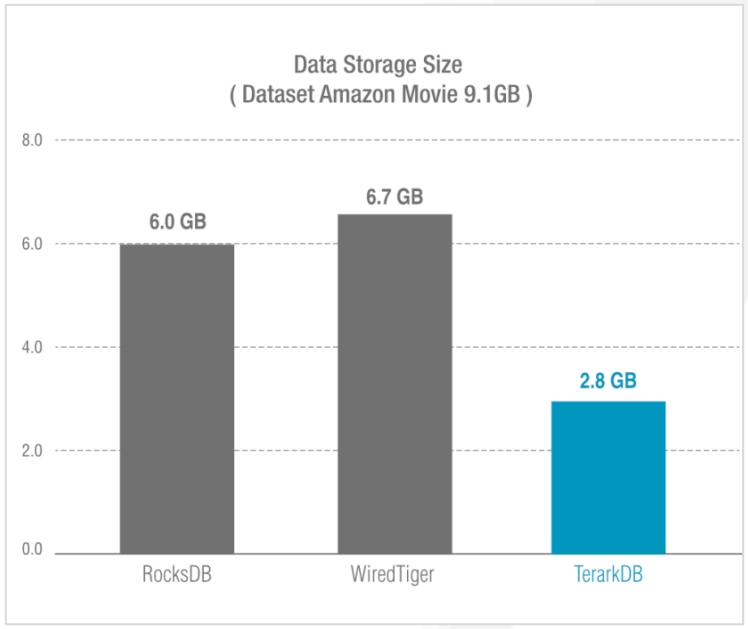
内存使用对比



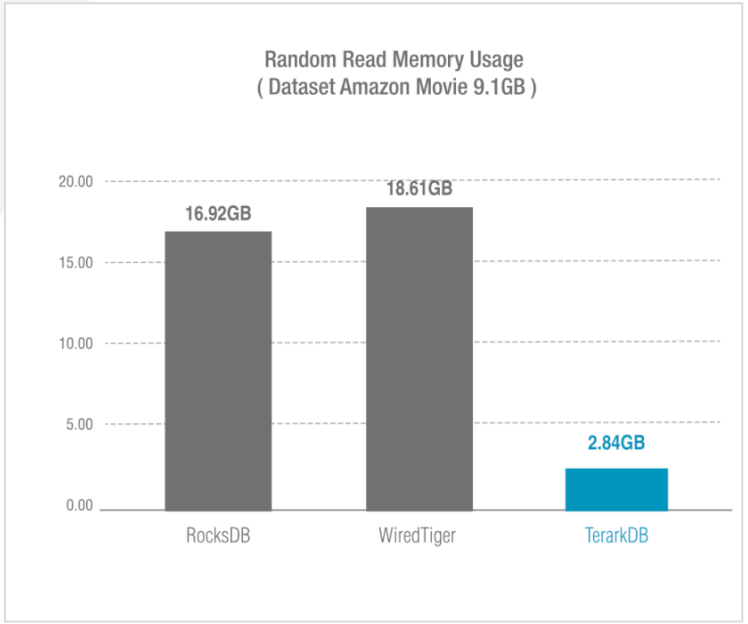
随机读性能

注：在不同的数据集和硬件环境下，测试结果可能不同

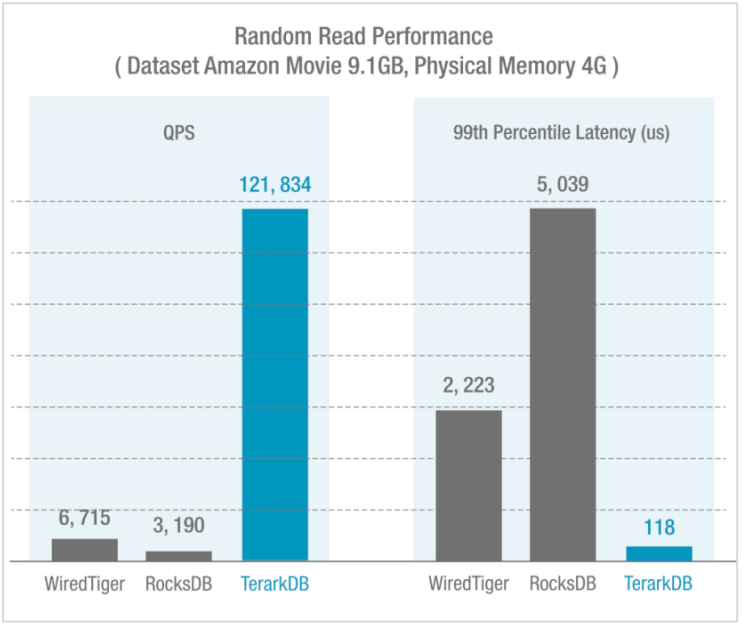
产品体系 - MongoDB 性能指标对比



压缩率对比



内存占用对比



随机读性能

注：在不同的数据集和硬件环境下，测试结果可能不同

性能测试 – 压缩率

数据源	InnoDB 无压缩	InnoDB 压缩	TokuDB 压缩	TerarkDB 压缩
百度知道	1.82 TB	970 GB	661 GB	472 GB
百度网盘	4.65 TB	3 TB	1.8 TB	1.23 TB
百度贴吧	1.4 TB	1.05 TB	710 GB	0.99 TB

- 该测试由百度专业DBA团队进行，InnoDB 经过高度优化
- TokuDB 是一款追求极端压缩率，牺牲性能的引擎，线上基本被抛弃
- TerarkDB 在压缩率的基础上，性能优势更大

数据源	记录尺寸	原始数据	RocksDB压缩	TerarkDB 压缩
TPCH-LineItem	128 bytes	40 GB	19 GB	8.1 GB
TPCH-LineItem	512 bytes	104 GB	40 GB	12 GB

- 该测试由阿里巴巴验收报告提供

性能测试 – 压缩率

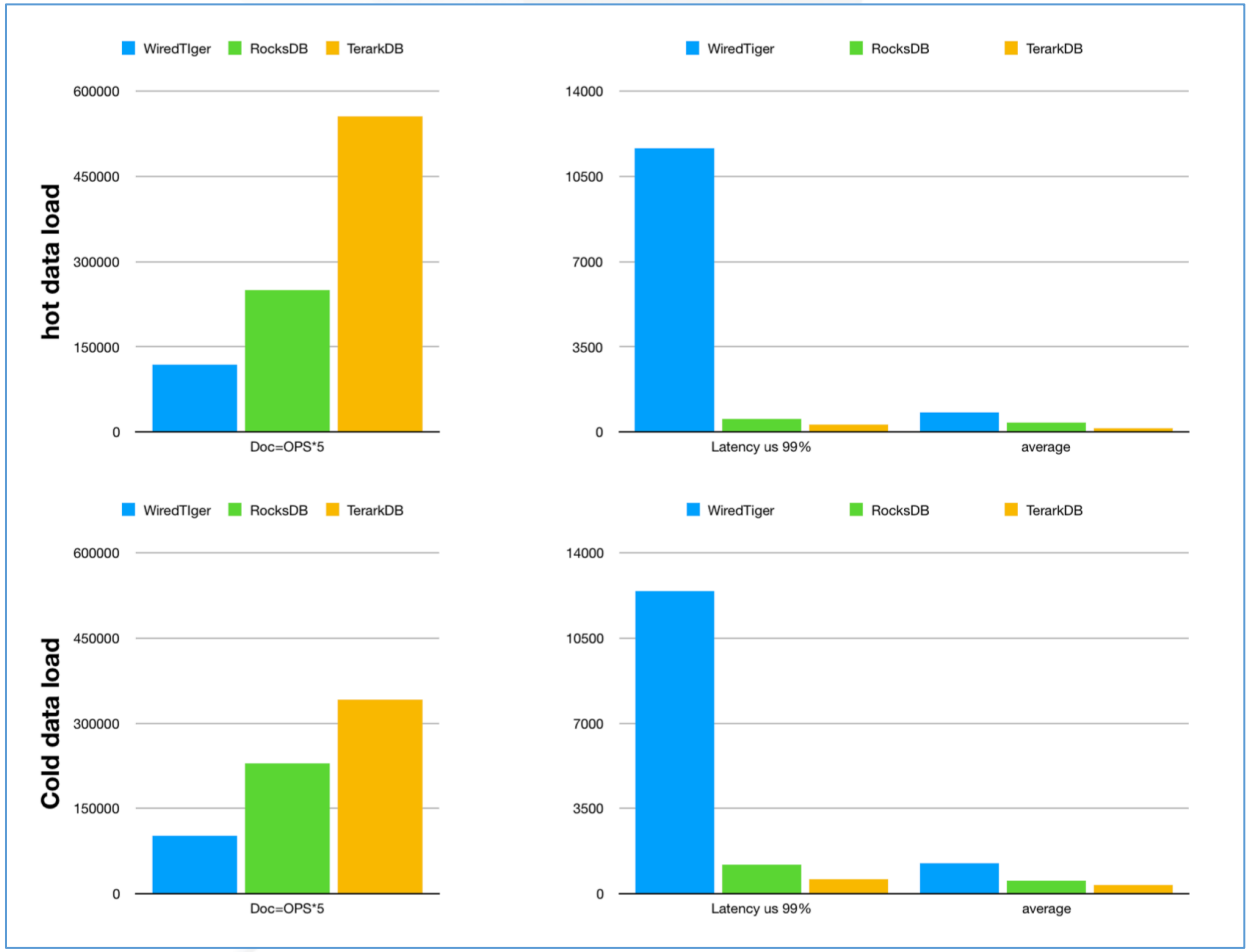
数据源	原始数据	WiredTiger	RocksDB	TerarkDB
阿里某业务	3.8 TB	2.25 TB	1.85 TB	1.2 TB

- 该测试由阿里巴巴验收报告提供
- 该测试场景包含大量的数字、随机字符串等难以压缩的内容

数据源	原始数据	WiredTiger 压缩	TerarkDB 压缩
维基百科	102 GB	62 GB	23 GB

- 该测试由 Terark 自测
- 该测试场景是我们的优势场景，有大量的文本冗余数据

性能测试 – 读性能(MongoDB)



- 该测试由阿里巴巴验收报告提供
- 对比 WiredTiger 有 3 ~ 5 倍的性能优势
- 该测试是在 MongoDB 作为数据库的情况下进行的

性能测试 –MySQL 综合测试

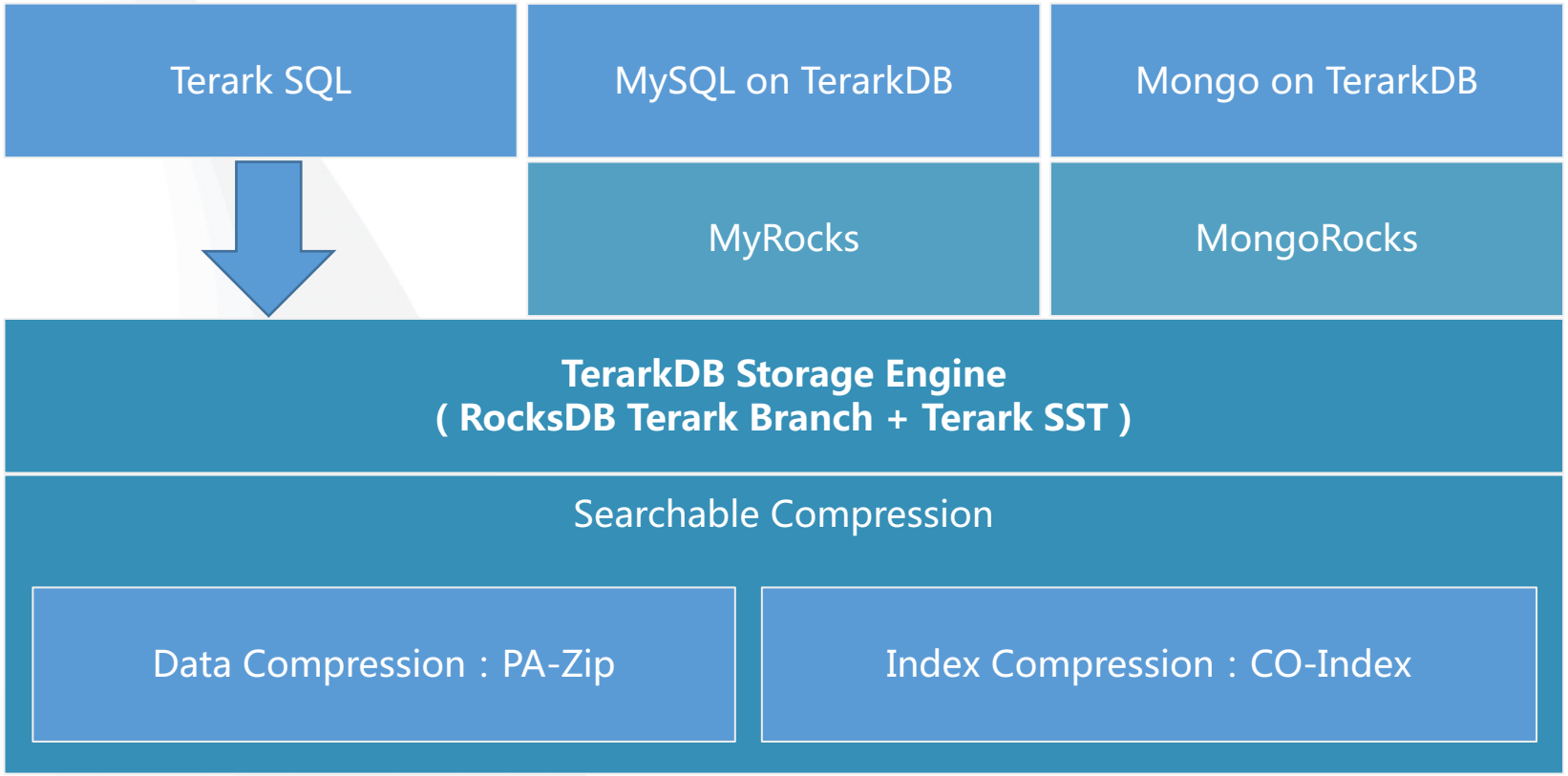
	InnoDB QPS	TerarkDB QPS
不限内存	47,000	97,000
96 GB	28,000	97,000
32 GB	19,000	95,000
16 GB	14,500	47,000
8 GB	14,000	28,000

	原始数据	InnoDB	TerarkDB
写入后尺寸	120GB	210GB	35GB

	InnoDB	TerarkDB
批量写入耗时	21 小时	3.5 小时

- TPCH Lineitem 数据，共 120 GB（100个表，每个表 9 个索引）
- 该测试是在 MySQL 作为数据库的情况下进行的
- 该测试由 Terark 和百度合作进行

附录1 - 技术介绍 – 架构图


Terark SQL
MySQL on TerarkDB
Mongo on TerarkDB
MyRocks
MongoRocks
TerarkDB Storage Engine
(RocksDB Terark Branch + Terark SST)
Searchable Compression
Data Compression：PA-Zip
Index Compression：CO-Index

附录1 - 技术介绍 – CO-Index

- CO-Index 表示 "Compressed Ordered Index"
- CO-Index 是 TerarkDB 的索引结构
- CO-Index 对不同的数据类型，有多种不同实现
- 其中 NST(Nested Succinct Trie) 用于普通的字符串
  - Trie 结构采用 Succinct Data Structure 实现
  - Trie 本身使用的是 Patricia Trie
    - 通过嵌套多重 Patricia Trie 增强压缩能力
  - 实现这样的结构，需要兼顾低内存和高性能，实现难度极大
  - 我们通过优秀的工程能力，完美实现了低内存和高性能

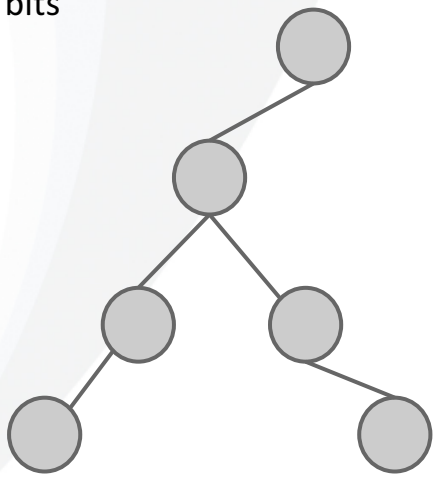
附录1 - 技术介绍 – 索引对比

	Hash	B+Tree	CO-Index
Compression	None	Good	✔✔✔ Excellent
Searching	✔✔ Very Fast	Good	✔ Fast
Exact Searching	✔ Support	✔ Support	✔ Support
Range Searching	Not Support	✔ Support	✔ Support
Prefix Searching	Not Support	✔ Support	✔ Support
Regex Searching	Not Support	Not Support	✔ Support
Reverse Searching(id to key)	Not Support(can be work-around)	Not Support	✔ Support

附录1 - 技术介绍 – Succinct Tree

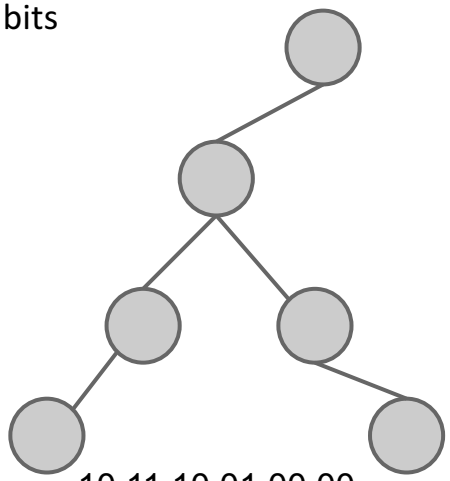
- Succinct Tree is a kind of data structure that implement Trees via bitmap. In general, it only uses about 1/30 of memory to represent a tree, compared to traditional pointer-based structures.
- Succinct data structures have relatively bad performance (compared to pointer-based technology)
- Terark uses its own implementation, which has a much better performance than open source versions

Each node uses two bits
Pre-Order
DFUDS



Needs **findopen, findclose, enclose**, which are much slower than rank/select, rarely used

Each node uses two bits
Level-Order
LOUDS

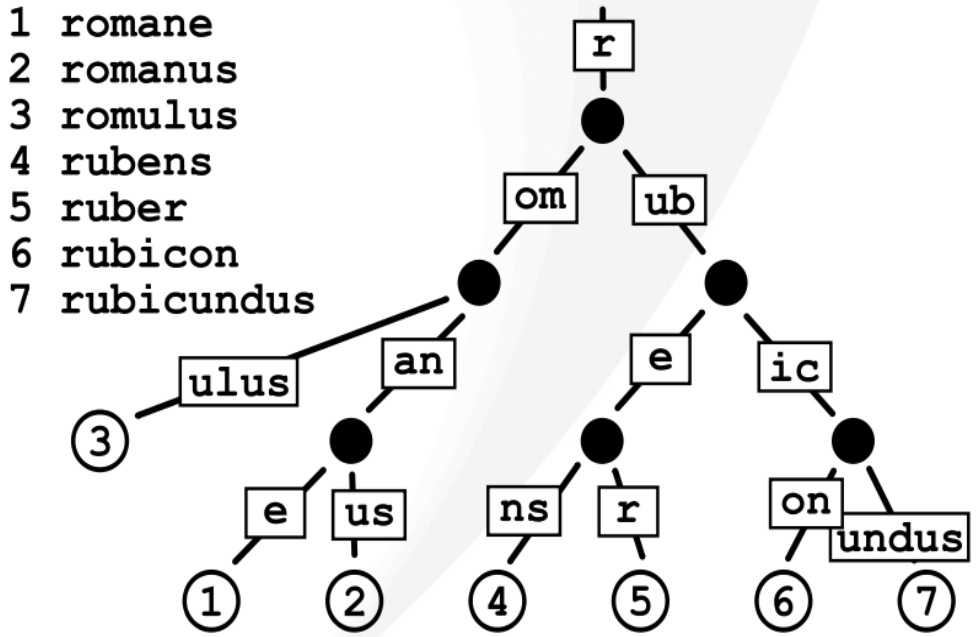


Simple and fast, small:

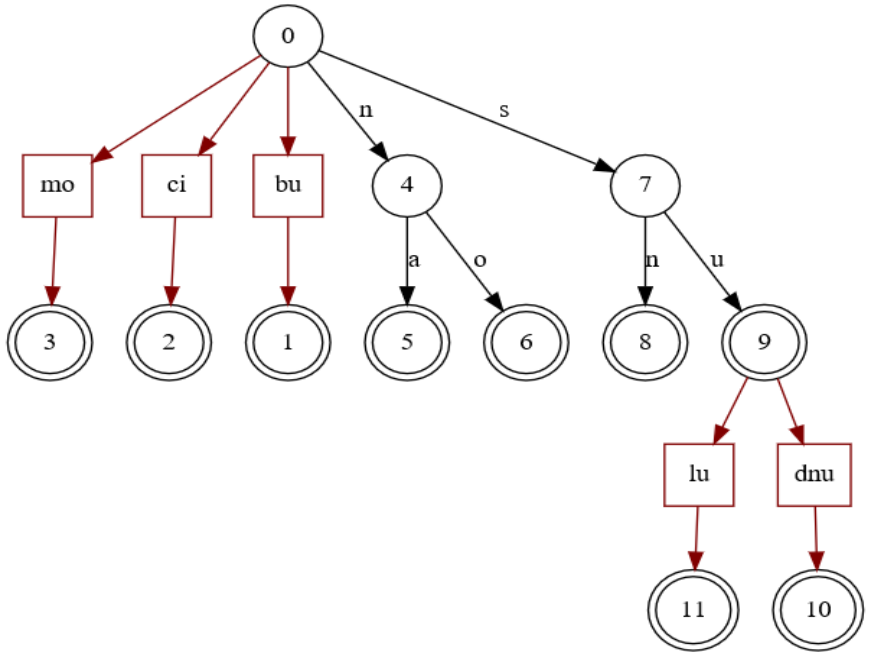
Parent(c) = rank0(select1(c))
Child(p, i) = select0(p) – p + i

附录1 - 技术介绍 – Nested Patricia Trie

Patricia Trie: Path Compression Trie
Path Compression: Compress all single-child nodes into one node



Nested: Store the compressed path into a new Trie.
Notes: Trie supports “reverse search”（except Succinct，Double Array Trie also support reverse search, but without compression)



附录1 - 技术介绍 – PA-Zip

- PA-Zip (Point Accessible Zip) 用于 TerarkDB 的值存储
- 相对于传统压缩技术
  - PA-Zip 可以将数百万到数十亿条记录压缩到一个大对象中
    - 压缩率逼近理论上限（传统块压缩分块压缩）
  - 压缩后的大对象允许针对单条记录随机访问（基于 Record ID）
    - 传统块压缩需要现解压块，在到块里面查找对应的记录
    - 传统块压缩解压后的块在会耗费不必要的内存（无效解压）

附录1 - 技术介绍 – 值压缩对比

	Block-based: leveldb, rocksdb, wiredtiger…	Short data: CO-Index	Medium data: PA-Zip
Compression ratio	OK	✔✔✔ Excellent	✔✔✔ Excellent
Random Read	Slow	✔ Fast	✔✔✔✔ Fast
Sequential Read	✔✔ Fast	Slow	✔✔✔ Fast
Double Cache Problem	YES	NO	NO
Compression Speed	✔ Fast(snappy…)	Slow	Slow(comparable to gzip)

附录2 – 部分资质证书

北京市新技术新产品（服务）证书

单位名称：奇简软件（北京）有限公司

产品（服务）名称：奇简数据库引擎软件 Terark

产品型号：TerarkDB V1.0；TerarkDB V0.13；TerarkX；SiriusDB V1.0

证书编号：XCP2017DZ0984

发证日期：2017 年 12 月

有 效 期：3 年

批准机关：

市科委　市发展改革委　市经济信息化委　市住房城乡建设委　市质监局　中关村管委会

附录2 – 部分资质证书

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第1576558号

软件名称：奇简TerarkDB数据库引擎软件 V1.0

著作权人：奇简软件（北京）有限公司

开发完成日期：2015年11月25日

首次发表日期：2015年11月25日

权利取得方式：原始取得

权利范围：全部权利

登记号：2016SR397942

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局 计算机软件著作权登记专用章

2016年12月27日

No. 01385622

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第1576560号

软件名称：奇简SiriusDB数据库引擎软件 V1.0

著作权人：奇简软件（北京）有限公司

开发完成日期：2016年08月25日

首次发表日期：2016年08月25日

权利取得方式：原始取得

权利范围：全部权利

登记号：2016SR397944

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局 计算机软件著作权登记专用章

2016年12月27日

No. 01385623

附录2 – 部分资质证书

荣誉证书

Terark 荣获 2016 年全国移动互联创新大赛 社会 组

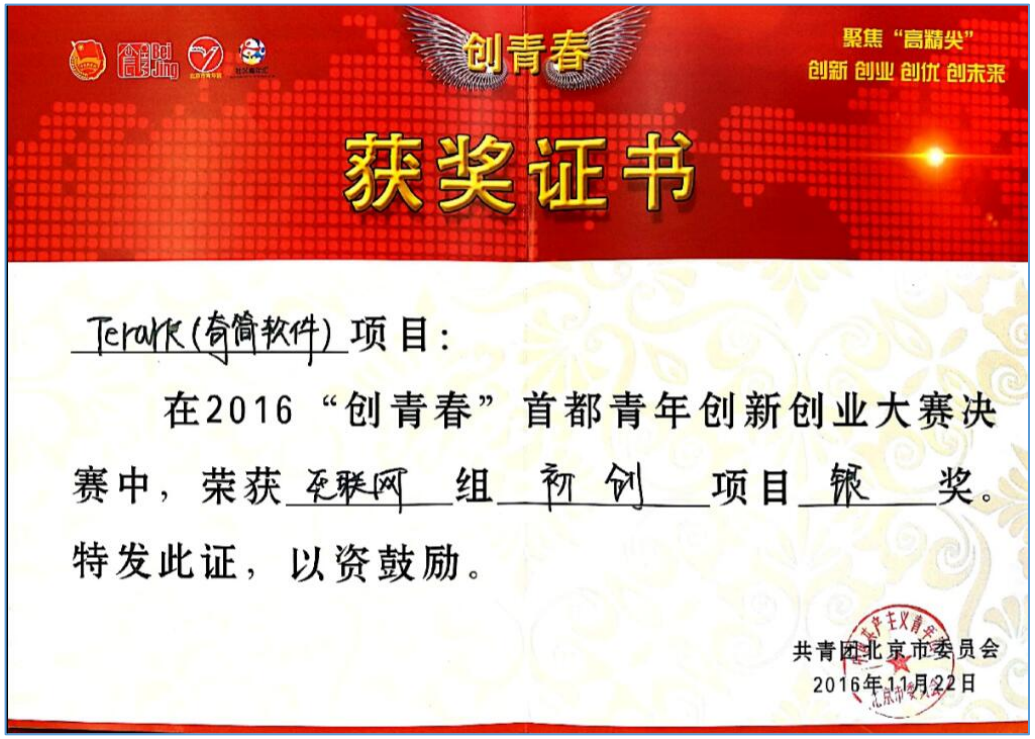
一 等奖。特颁此证，以资鼓励。

获奖单位：奇简软件（北京）有限公司

获 奖 者：付新元、雷鹏、郭宽宽

中国通信学会

二零一六年十月



创青春

聚焦“高精尖”

创新 创业 创优 创未来

获奖证书

Terark(奇简软件) 项目：

在2016“创青春”首都青年创新创业大赛决赛中，荣获 互联网 组 初创 项目 铜 奖。

特发此证，以资鼓励。

共青团北京市委员会

2016年11月22日

附录2 – 部分资质证书

荣誉证书

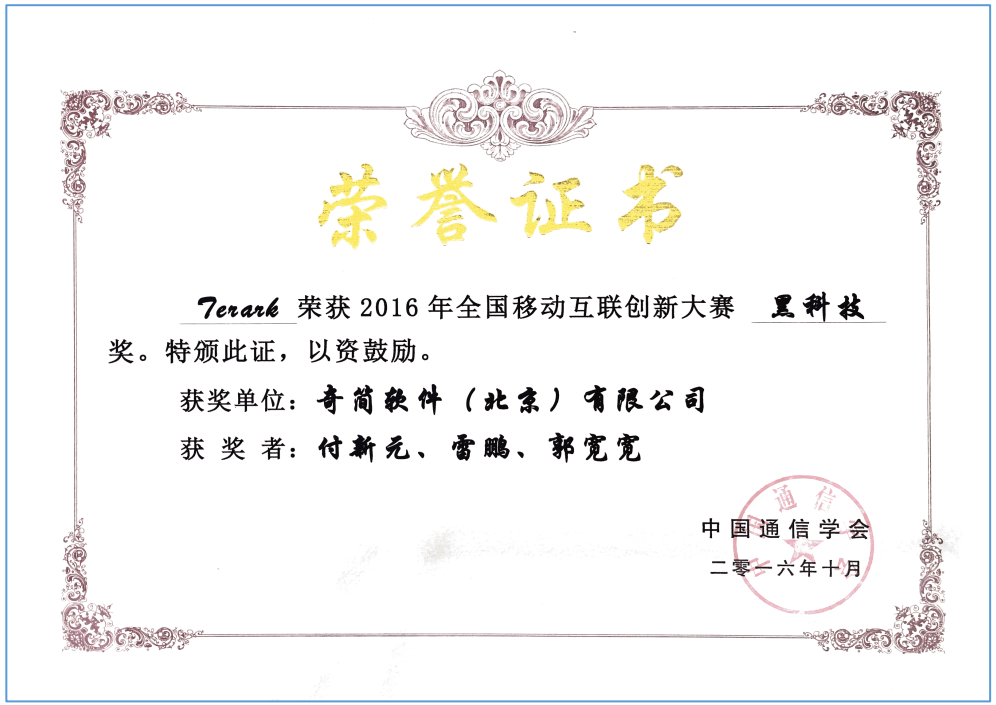
Terark：

在2017年北京市优秀创业项目遴选中被认定为优秀创业项目。

特发此证，以资鼓励。

北京市人力资源和社会保障局

2017年7月



荣誉证书

Terark 荣获2016年全国移动互联创新大赛 黑科技 奖。特颁此证，以资鼓励。

获奖单位：奇简软件（北京）有限公司

获 奖 者：付新元、雷鹏、郭宽宽

中国通信学会

二零一六年十月

附录2 – 部分资质证书

065395 8001CN01

中华人民共和国国家知识产权局

100005

北京市东城区建国门北大街8号华润大厦20层
北京市君合律师事务所 宋海宁,唐宇

发文日：

2016年03月03日

申请号或专利号：201610118629.5　　发文序号：2016030301121410

专利申请受理通知书

根据专利法第28条及其实施细则第38条、第39条的规定，申请人提出的专利申请已由国家知识产权局受理。现将确定的申请号、申请日、申请人和发明创造名称通知如下：

申请号：201610118629.5
申请日：2016年03月02日
申请人：奇简软件（北京）有限公司
发明创造名称：进行数据压缩的方法、装置、系统和计算机程序产品

经核实，国家知识产权局确认收到文件如下：
发明专利请求书 每份页数:4页 文件份数:1份
权利要求书 每份页数:8页 文件份数:1份 权利要求项数： 44项
说明书 每份页数:12页 文件份数:1份
说明书附图 每份页数:10页 文件份数:1份
说明书摘要 每份页数:1页 文件份数:1份
摘要附图 每份页数:1页 文件份数:1份
向外国申请专利保密审查请求书 每份页数:1页 文件份数:1份

提示：

1. 申请人收到专利申请受理通知书之后，认为其记载的内容与申请人所提交的相应内容不一致时，可以向国家知识产权局请求更正。

2. 申请人收到专利申请受理通知书之后，再向国家知识产权局办理各种手续时，均应当准确、清晰地写明申请号。

审 查 员：陈晨(电子申请)　　审查部门：专利局初审及流程管理部-15

200101 2010.2　纸件申请，回函请寄：100088 北京市海淀区前门桥西土城路6号 国家知识产权局受理处收
电子申请，应当通过电子专利申请系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

065395 8002CN01

中华人民共和国国家知识产权局

100005

北京市东城区建国门北大街8号华润大厦20层
北京市君合律师事务所 宋海宁,毛健

发文日：

2016年03月03日

申请号或专利号：201610119044.5　　发文序号：2016030301152810

专利申请受理通知书

根据专利法第28条及其实施细则第38条、第39条的规定，申请人提出的专利申请已由国家知识产权局受理。现将确定的申请号、申请日、申请人和发明创造名称通知如下：

申请号：201610119044.5
申请日：2016年03月03日
申请人：奇简软件（北京）有限公司
发明创造名称：为输入数据搜索匹配候选项的方法以及数据库创建方法

经核实，国家知识产权局确认收到文件如下：
发明专利请求书 每份页数:4页 文件份数:1份
权利要求书 每份页数:5页 文件份数:1份 权利要求项数： 26项
说明书 每份页数:17页 文件份数:1份
说明书附图 每份页数:5页 文件份数:1份
说明书摘要 每份页数:1页 文件份数:1份
摘要附图 每份页数:1页 文件份数:1份
向外国申请专利保密审查请求书 每份页数:1页 文件份数:1份

提示：

1. 申请人收到专利申请受理通知书之后，认为其记载的内容与申请人所提交的相应内容不一致时，可以向国家知识产权局请求更正。

2. 申请人收到专利申请受理通知书之后，再向国家知识产权局办理各种手续时，均应当准确、清晰地写明申请号。

审 查 员：陈晨(电子申请)　　审查部门：专利局初审及流程管理部-15

200101 2010.2　纸件申请，回函请寄：100088 北京市海淀区前门桥西土城路6号 国家知识产权局受理处收
电子申请，应当通过电子专利申请系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

期待和您的合作!

TERARK