CHINA DATA ANALYST SUMMIT

CDA 数据分析师 www.cda.cn

数据库的前世今生

演讲人：韩锋

跨界互联 数聚未来

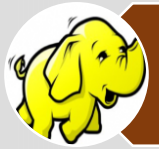
第四届中国数据分析师行业峰会

CHINA DATA ANALYST SUMMIT

北京 中国大饭店 2017.07

- 技术的发烧友
- 虔诚的布道者
- 曾经的码农
- 如今的数据从业者
- 有时动动笔的人
- 玩过Oracle、MySQL、Informix、GreenPlum、PostgreSQL、SQLServer、FoxPro、Redis、MongoDB、Java、PowerBuilder、Visual Basic、JavaScript、Python、Shell、PowerDesigner、Rose、LoadRunner、Jmeter…

数据库发展现状



大数据与数据库

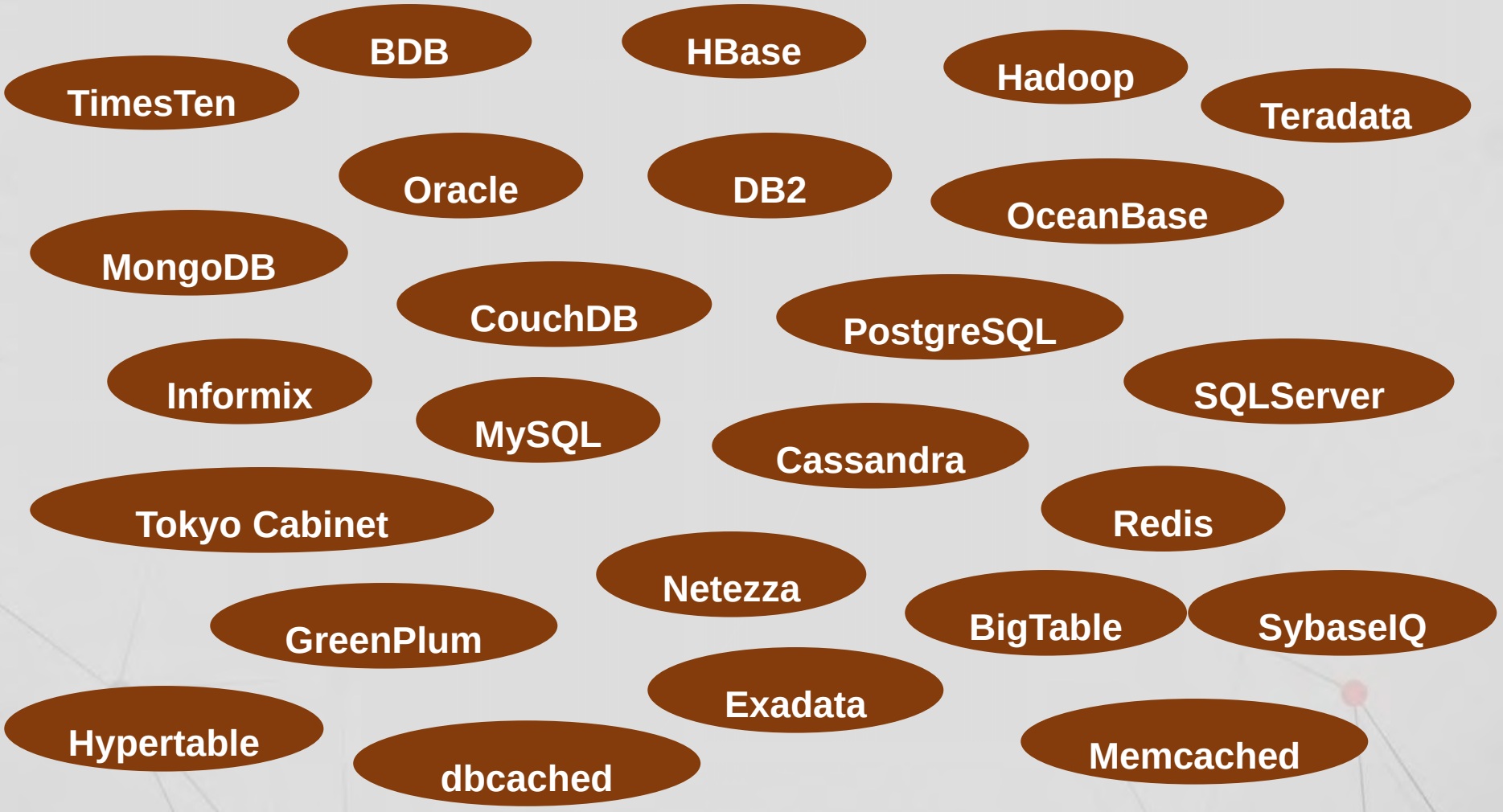
云与数据库

硬件与数据库

虚拟化与数据库

数据库管理的变化

数据库发展现状


BDB
HBase
Hadoop
TimesTen
Teradata
Oracle
DB2
OceanBase
MongoDB
CouchDB
PostgreSQL
Informix
SQLServer
MySQL
Cassandra
Tokyo Cabinet
Redis
Netezza
BigTable
SybaseIQ
GreenPlum
Exadata
Hypertable
Memcached
dbcached

关系型数据库

- Oracle
- DB2
- SQL Server
- Sybase IQ
- PostgreSQL
- MySQL
- OceanBase
- GreenPlum
- BDB
- TimesTen
- …

非关系数据库

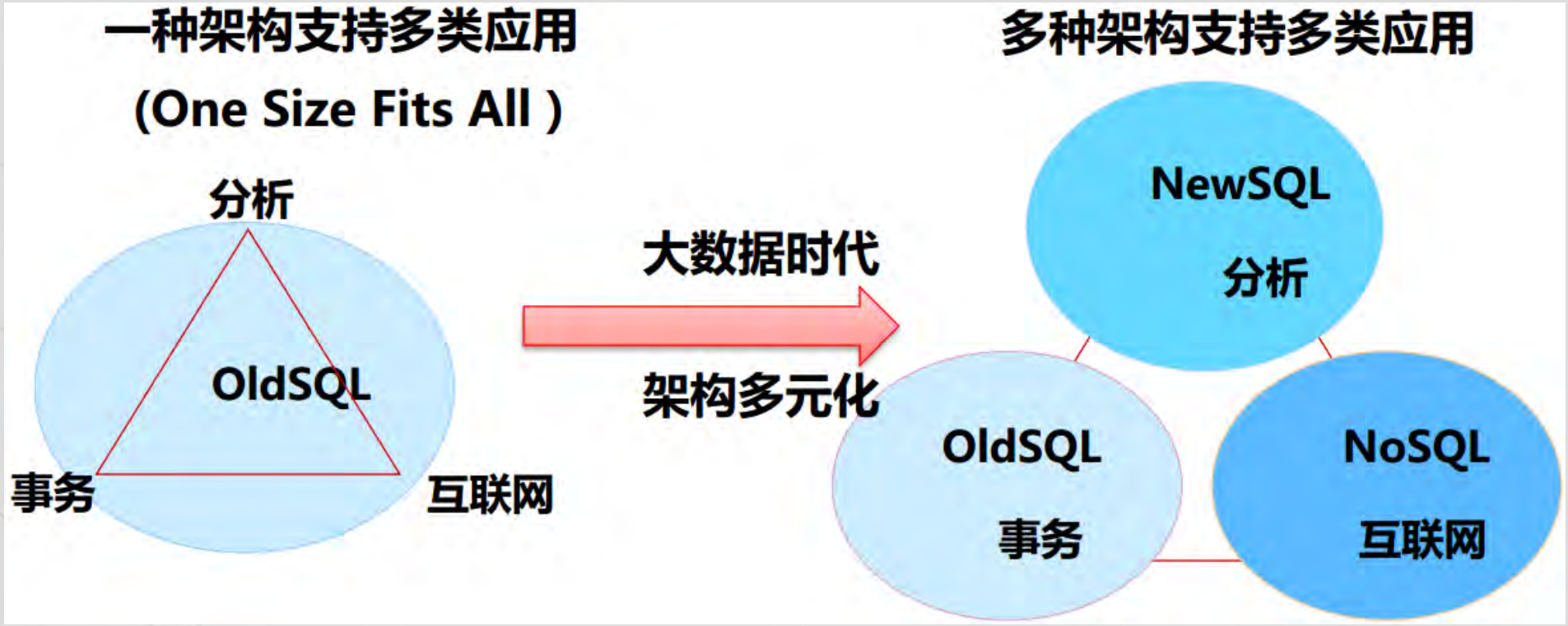
- Hbase
- MongoDB
- Redis
- CouchDB
- BigTable
- Tokyo cabinet
- Dynamo
- Voldemort
- …

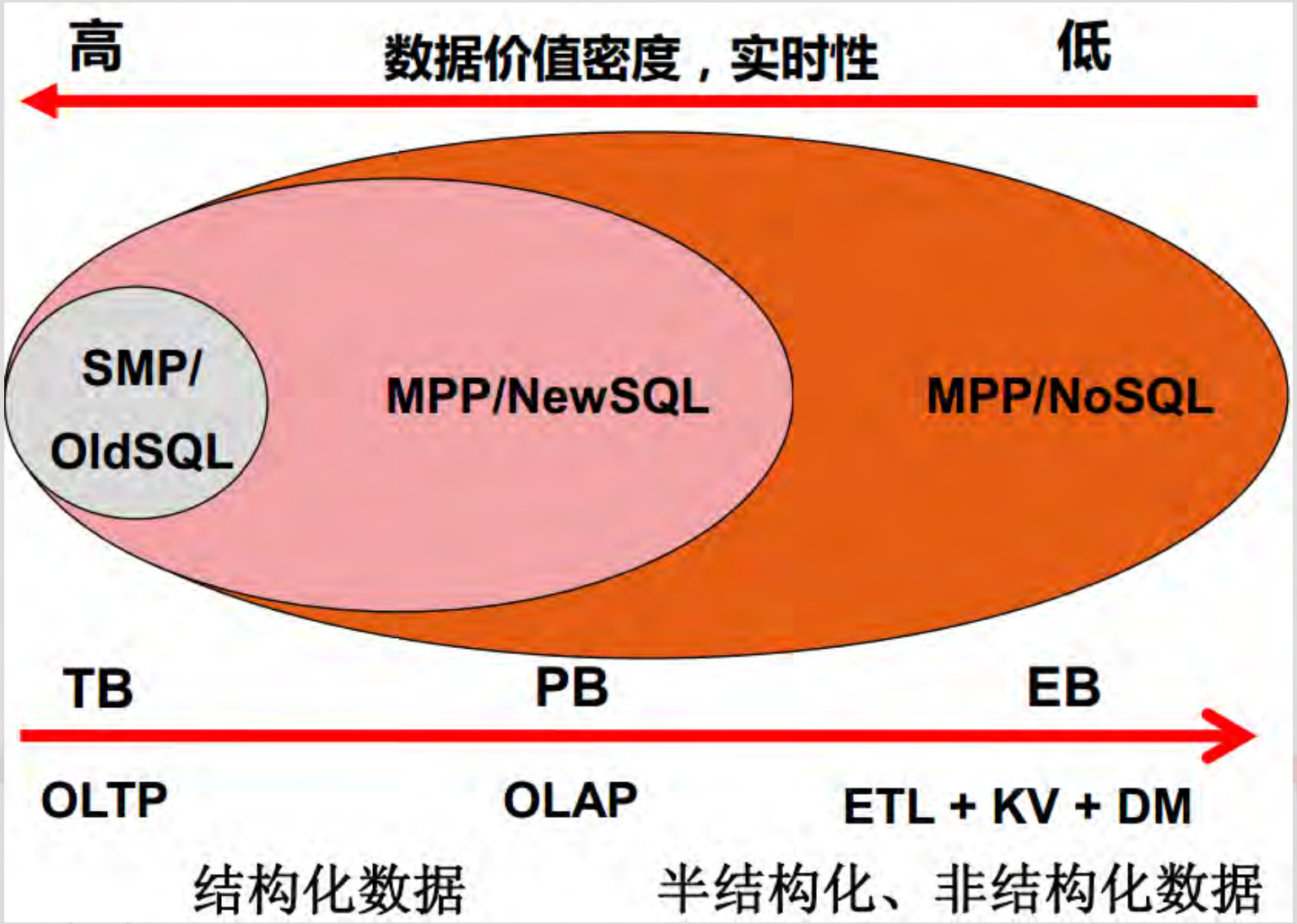
商业产品

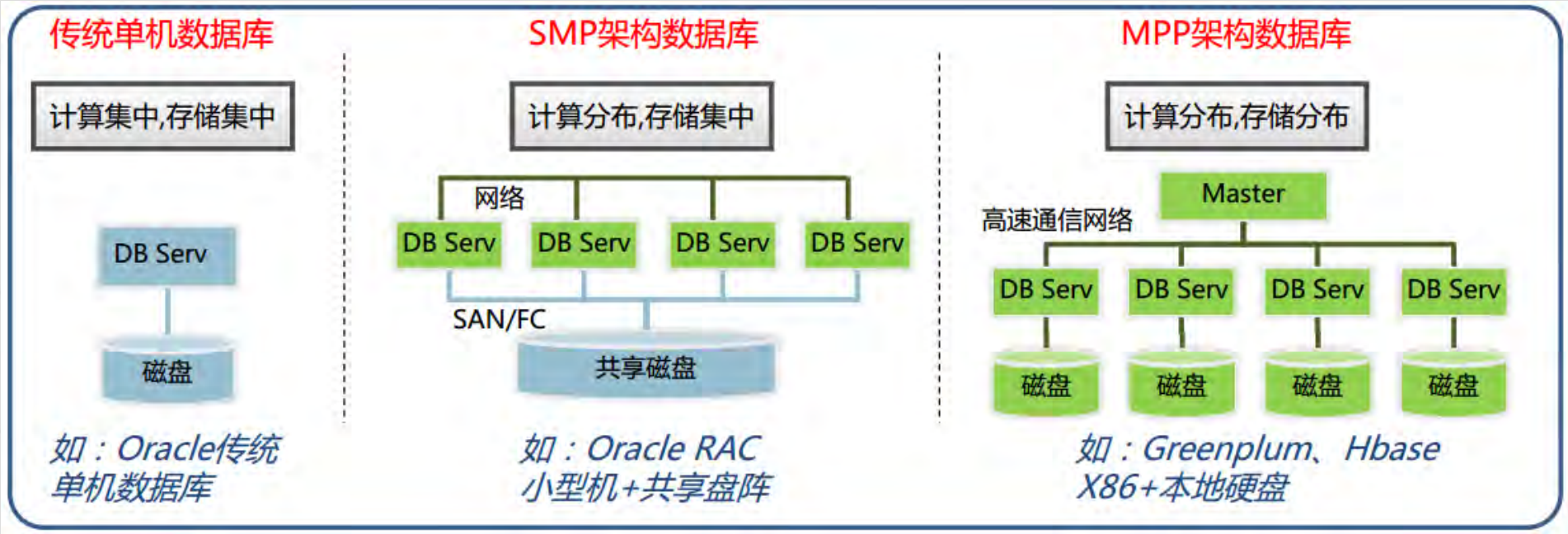
- Exadata
- Netezza
- Teradata
- …

非数据库

- Memcached
- Dbcached
- Hadoop
- …


一种架构支持多类应用
(One Size Fits All)
分析
OldSQL
事务
互联网
大数据时代
架构多元化
多种架构支持多类应用
NewSQL
分析
OldSQL
事务
NoSQL
互联网

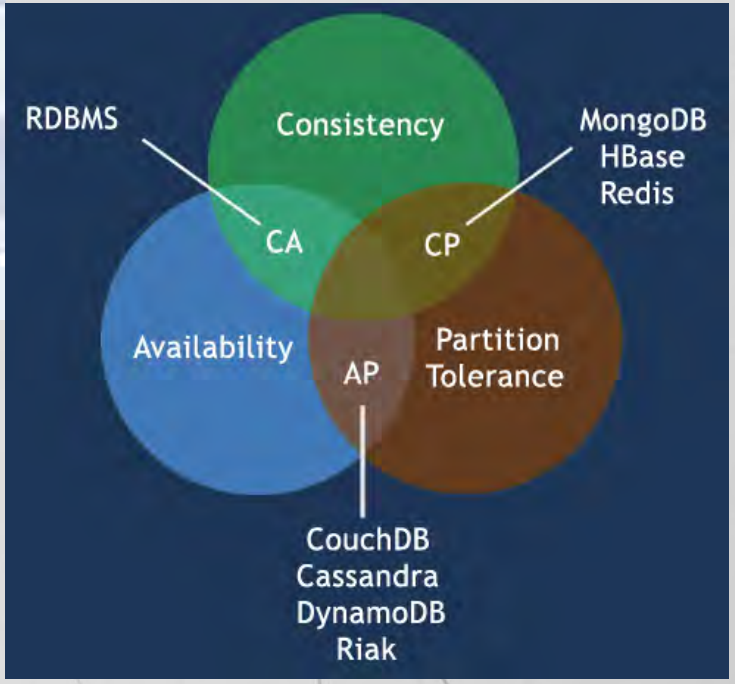

高
数据价值密度，实时性
低
SMP/
OldSQL
MPP/NewSQL
MPP/NoSQL
TB
PB
EB
OLTP
OLAP
ETL + KV + DM
结构化数据
半结构化、非结构化数据



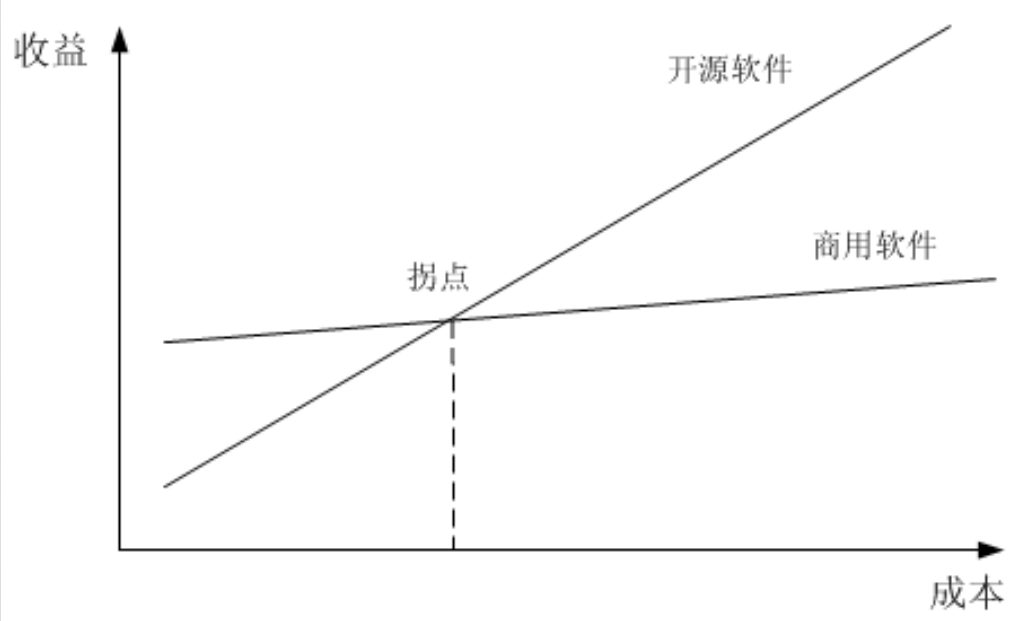
SMP:对称多处理器		
两台以上服务器，通过总线，共享磁盘数据。	扩展能力有限，只能通过提升节点能力达到扩容。	磁盘访问往往成为性能瓶颈。

MPP:大规模并行处理		
每个节点有独立计算、存储能力。	扩展能力强，通过增加服务器数量扩展处理能力。	多软件要求较高，需要协调调度个节点运行。

类型	主要产品	简介
KV存储	Redis Memcached	使用key快速查到其value，Memcached支持string类型的value，Redis除string类型外还支持set、hash、sort set、list等类型
文档存储	MongoDB CouchDB	使用JSON或类JSON的BSON数据结构，存储内容为文档型，能实现部分关系数据库的功能
列存储	HBase Cassandra	按照列进行数据存储，便于存储结构化和半结构化数据，方便做数据压缩和针对某一列和某几列的数据查询
图存储	Neo4J FlockDB	图形关系的存储，能够很好弥补关系数据库在图形存储的不足
对象存储	Db4o Versant	通过类似面向对象语言的方式操作数据库，通过对象的方式存取数据
XML数据库	Berkeley DB XML BaseX	高效存储XML数据，支持XML的内部查询语法，如XQuery、XPath




1
去“IOE”提出
2008年，前微软亚洲研究院常务副院长王坚加盟阿里巴巴成为集团首席架构师，即现在的首席技术官。2009年提出“去IOE”战略。
2
去“IOE”实践
2010年不再购买小型机，2012年不购买EMC的设备，2011年7月份完成商品库的首个去IOE实践。
3
去“IOE”实现
2013年5月17日，阿里最后一台IBM小机在支付宝下线。7月10日，淘宝重中之重的广告系统Oracle数据库下线。
计划
尝试
行动
外界宣传及媒体介入
“去IOE”
4
去“IOE”运动
2013年8月《商业价值》杂志发表王坚采访文章《阿里巴巴为什么会“去IOE”》，8月底中国软件开发者大会上，周宏方分享了《阿里“去IOE”战略》，引起业界广泛关注。
宣传
王坚
阿里巴巴集团首席技术官



人才积累

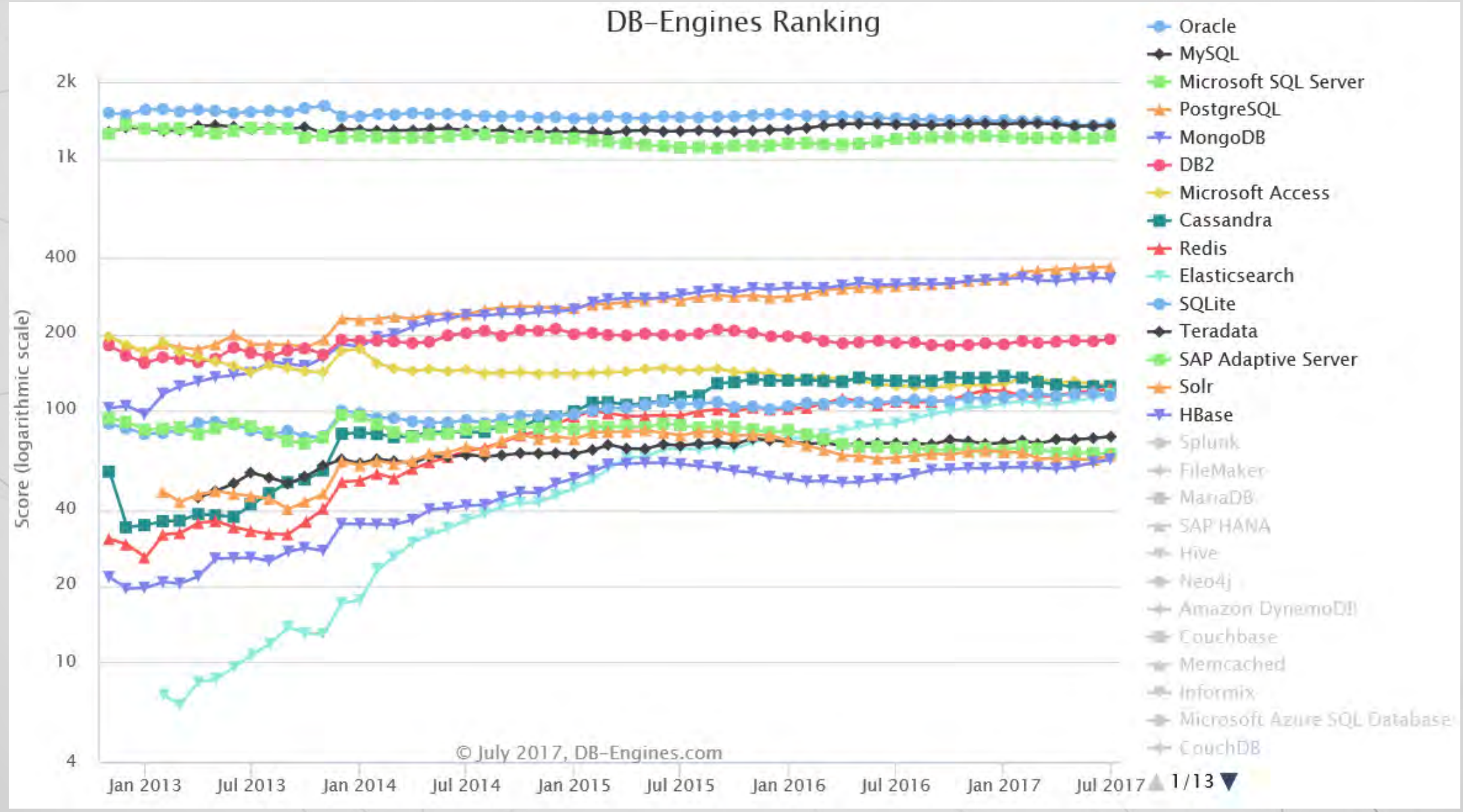
开源软件

硬件革命

摩尔定律

开源

商业


DB-Engines Ranking
Score (logarithmic scale)
2k
1k
400
200
100
40
20
10
4
Jan 2013
Jul 2013
Jan 2014
Jul 2014
Jan 2015
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
© July 2017, DB-Engines.com
Oracle
MySQL
Microsoft SQL Server
PostgreSQL
MongoDB
DB2
Microsoft Access
Cassandra
Redis
Elasticsearch
SQLite
Teradata
SAP Adaptive Server
Solr
HBase
Splunk
FileMaker
MariaDB
SAP HANA
Hive
Neo4j
Amazon DynamoDB
Couchbase
Memcached
Informix
Microsoft Azure SQL Database
CouchDB
1/13

- 弱化关系数据库的部分特性(例如跨表JOIN、事务等)，针对自身场景开发“自定义DB”。
- 充分利用开源数据库的功能，通过引入中间层达到高可用、水平扩展强的能力。
- 集群数据库层出不穷，可有的选择很多。不开发中间层，也可以达到很好的效果。
- 在合适的场景，大胆使用NoSQL，但要处理选择场景。
- 利用高速发展硬件技术，提高数据库整体能力。
- 传统数据库仍有应用场景，要发掘其潜力，压榨资源。

大数据与数据库

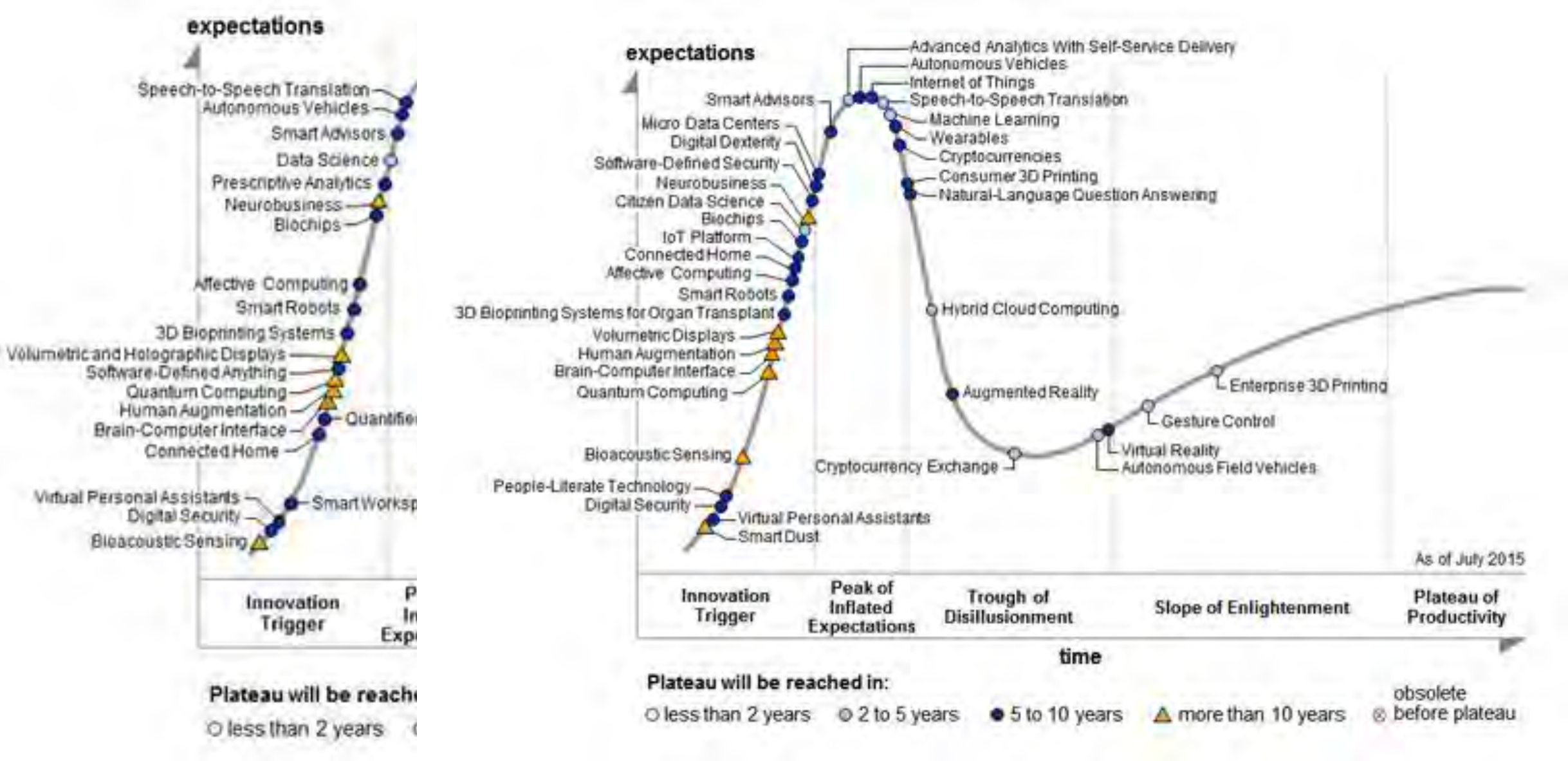
大数据（巨量资料（IT行业术语））

编辑

大数据(big data)，是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。

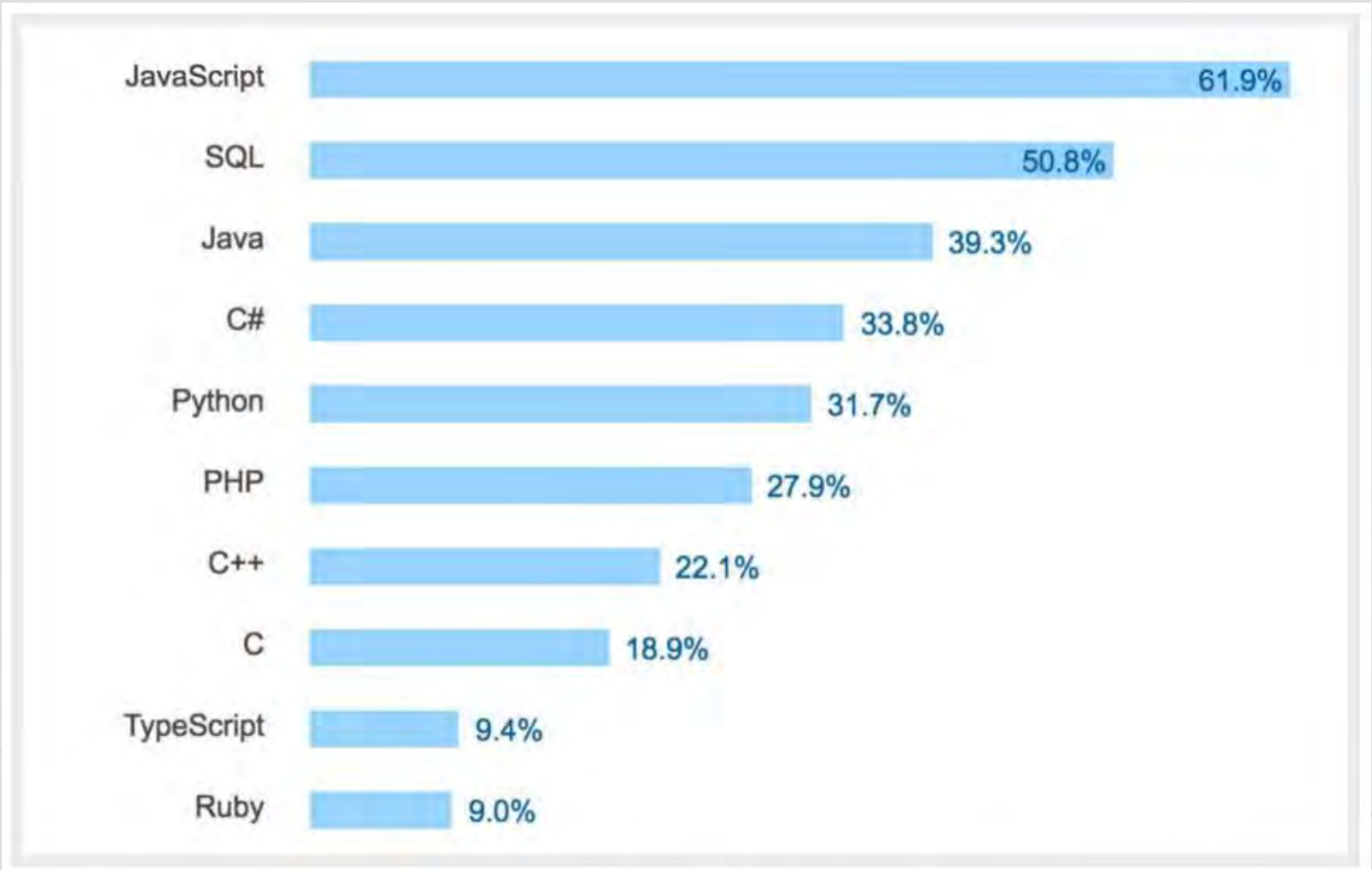
在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》[1]中大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据进行分析处理。大数据的4V特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值）。[2]

Volumn	• 数据规模爆炸式增长（TB->PB->EB）
Variety	• 结构化、半结构化和非结构化数据
Value	• 价值密度低
Velocity	• 海量数据快速获得信息


expectations
Advanced Analytics With Self-Service Delivery
Autonomous Vehicles
Internet of Things
Speech-to-Speech Translation
Machine Learning
Wearables
Cryptocurrencies
Consumer 3D Printing
Natural-Language Question Answering
Smart Advisors
Micro Data Centers
Digital Dexterity
Software-Defined Security
Neurobusiness
Citizen Data Science
Biochips
IoT Platform
Connected Home
Affective Computing
Smart Robots
3D Bioprinting Systems for Organ Transplant
Volumetric Displays
Human Augmentation
Brain-Computer Interface
Quantum Computing
Hybrid Cloud Computing
Augmented Reality
Enterprise 3D Printing
Gesture Control
Virtual Reality
Autonomous Field Vehicles
Cryptocurrency Exchange
Bioacoustic Sensing
People-Literate Technology
Digital Security
Virtual Personal Assistants
Smart Dust
As of July 2015
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau

大数据与传统数据库的没有本质区别...

其核心都是“数据”载体，承担存储与计算的能力。



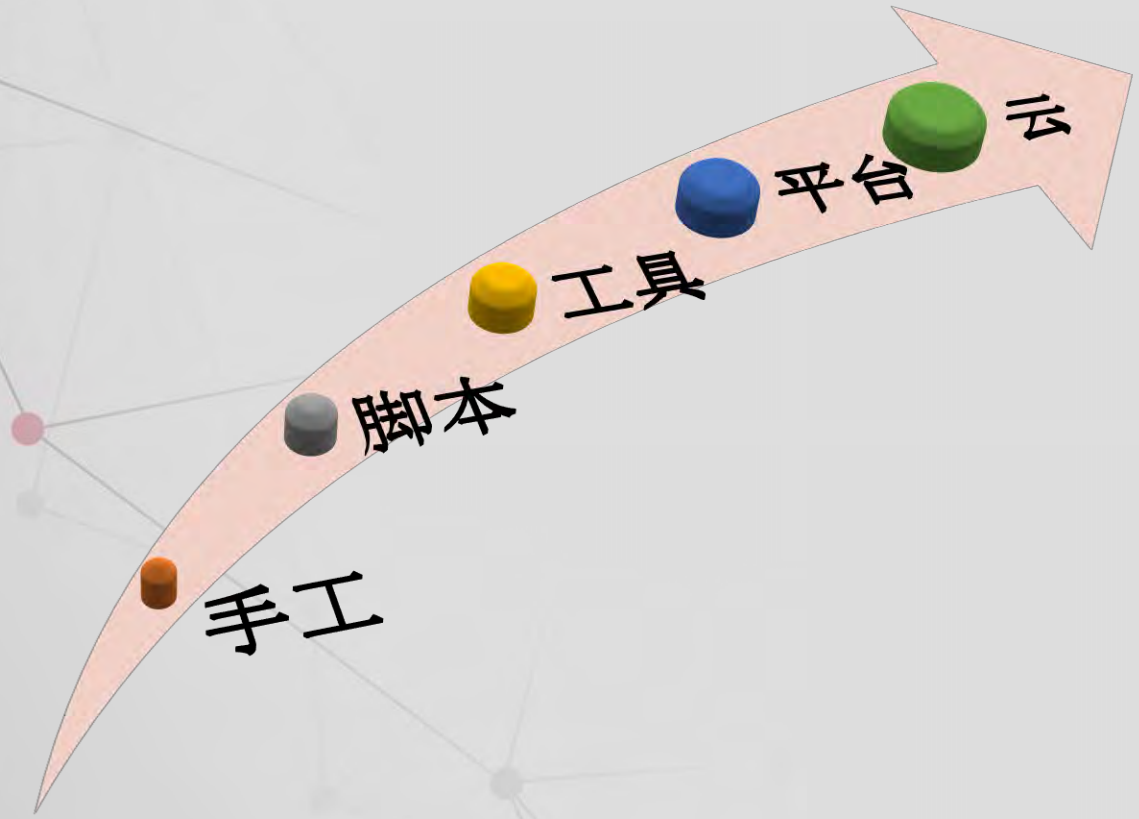
开发者最常用的语言排行, Stack Overflow 2017

云与数据库

数据库上云，是未来的发展趋势！

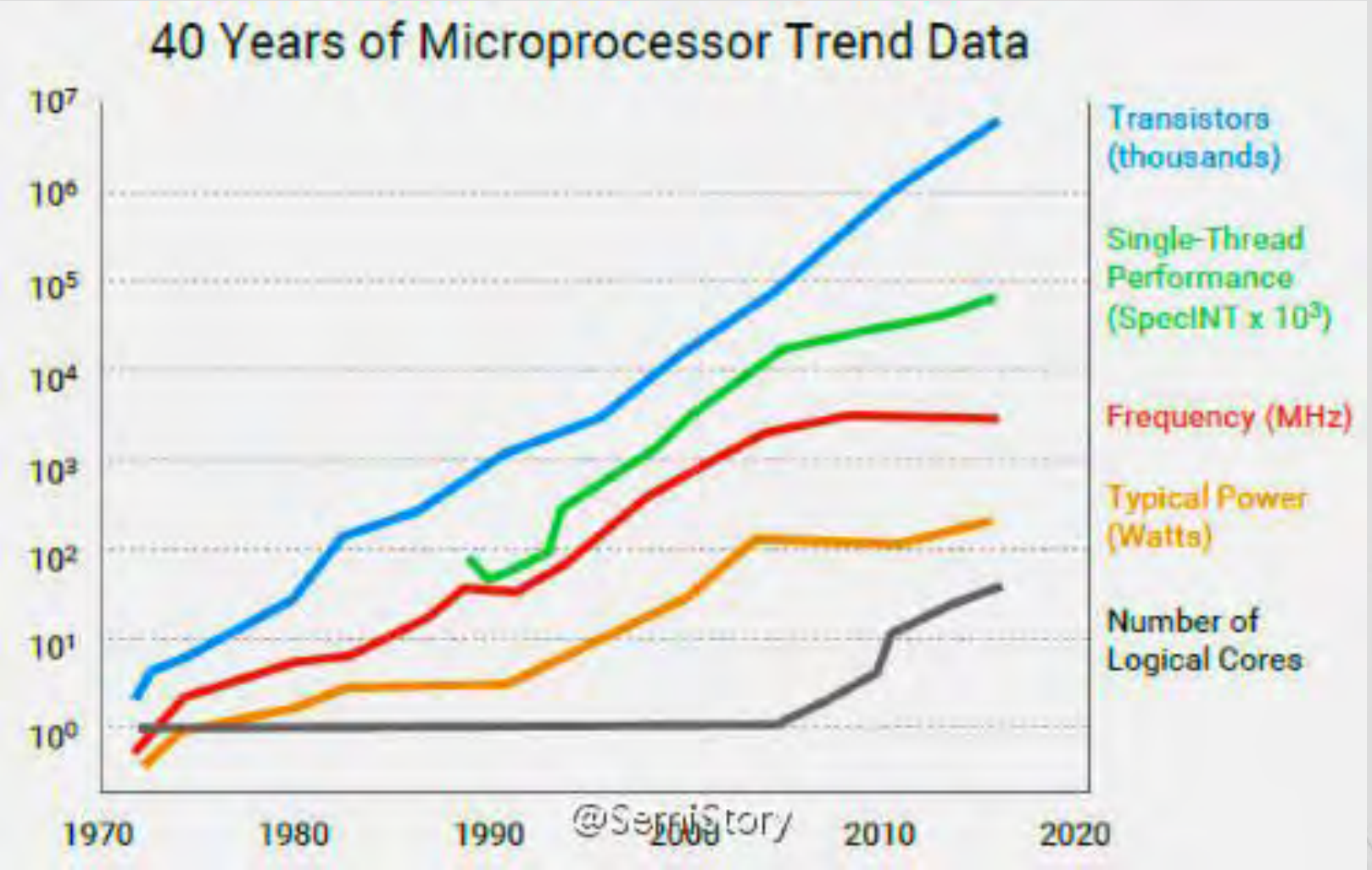
网易 amazon web services 腾讯云 阿里云 aliyun.com 百度云

系统部署	日常监控	资源分配	资源回收	主备切换	增减备库
实例迁移	数据迁移	备份恢复	异常诊断	负载均衡	故障切换
性能分析	变更发布	安全控制	结构审核	语句审核	元数据



- 文档/标准化
- 脚本/工具化
- 自动/平台化
- 智能/云化

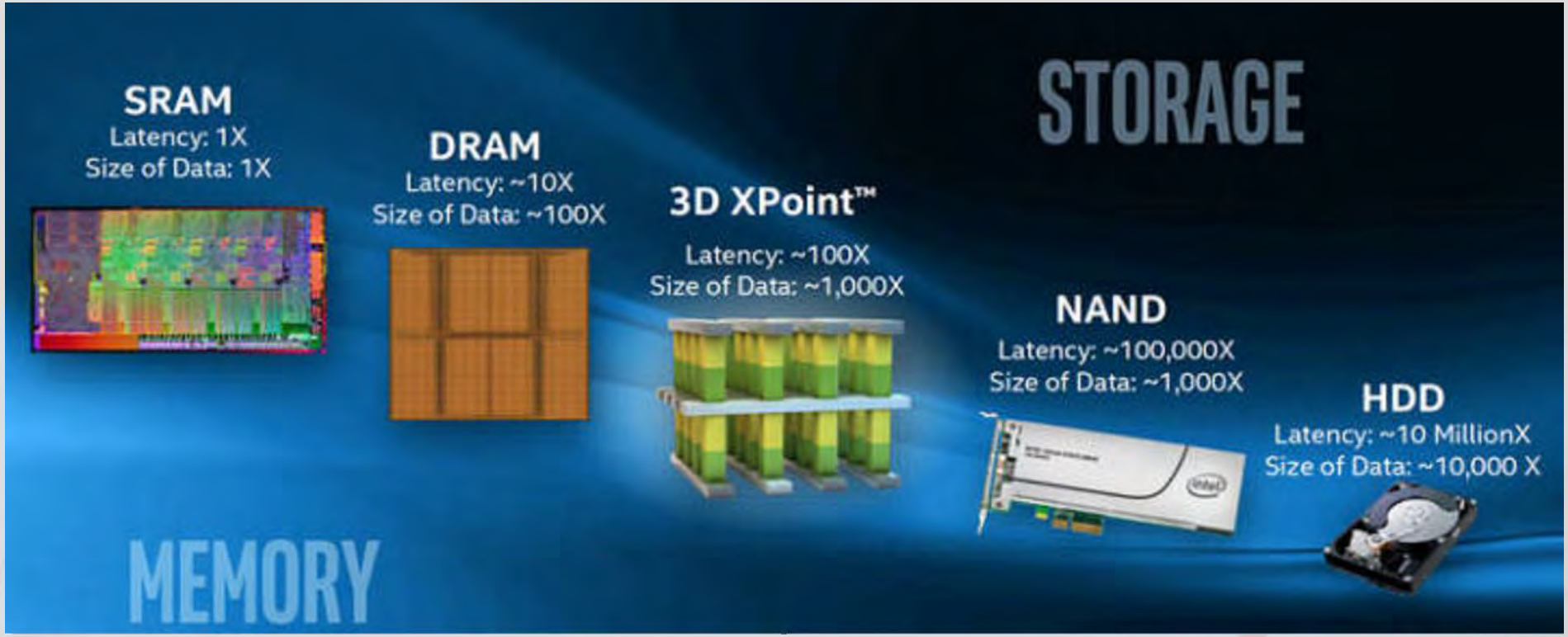
硬件与数据库


40 Years of Microprocessor Trend Data
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
1970
1980
1990
2000
2010
2020
Transistors (thousands)
Single-Thread Performance (SpecINT x 10^3)
Frequency (MHz)
Typical Power (Watts)
Number of Logical Cores
@SerialStory

体系结构	吞吐量 (int ops)	延迟	功耗	灵活性
CPU	~1 T	N/A	~100W	很高
GPU	~10 T	~1 ms	~300W	高
FPGA (Stratix V)	~1 T	~1 us	~30W	高
FPGA (Stratix 10)	~10 T	~1 us	~30W	高
ASIC	~10 T	~1 us	~30W	低

计算密集型任务，CPU、GPU、FPGA、ASIC 的数量级比较（以 16 位整数乘法为例，数字仅为数量级的估计）

- NVRAM
- 3D XPoint
- 3D NAND
- NVMe



随着GigE、10GbE、InfiniBand技术的飞速发展，低延迟、高带宽的服务品质给数据库乃至整个IT系统带来了很多变化。常见的应用领域有：

- 加速分布式数据库，例如Oracle RAC。
- 加速大数据处理，例如提升Hadoop MapReduce处理。
- 存储架构的变革，从Scale-Up向Scale-Out演变。
- 容灾方案，主备策略...

硬件技术的飞速发展，促进了数据库软件技术不断发展，为新一代数据铺垫了基础（例如分布式）。

传统数据库对硬件结束的发展需要加快适应过程，这一次硬件在推动软件革命。

新兴数据库的不断涌现，可更好地利用硬件资源，也为系统架构提供了更多的选择。

IO不在是瓶颈?

磁盘IO模型已落伍?

聚簇因子不再重要了？

NoSQL好像适应的更好?

分布式数据库的春天来了？

数据库优化还重要吗？

虚拟化与数据库

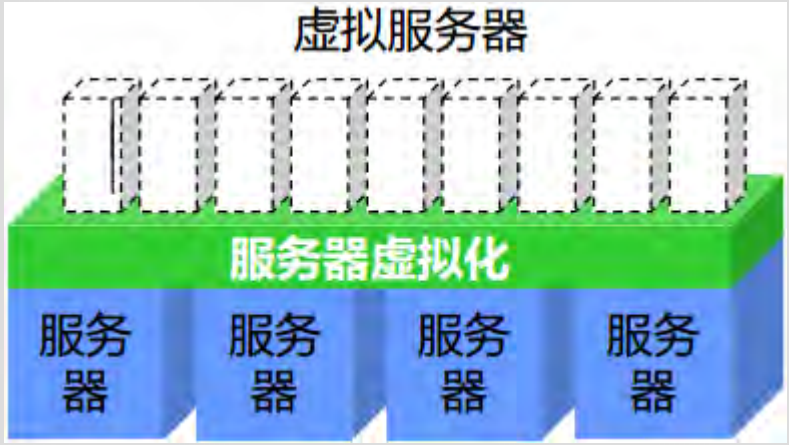
虚拟化

锁定

本词条由"科普中国"百科科学词条编写与应用工作项目审核。

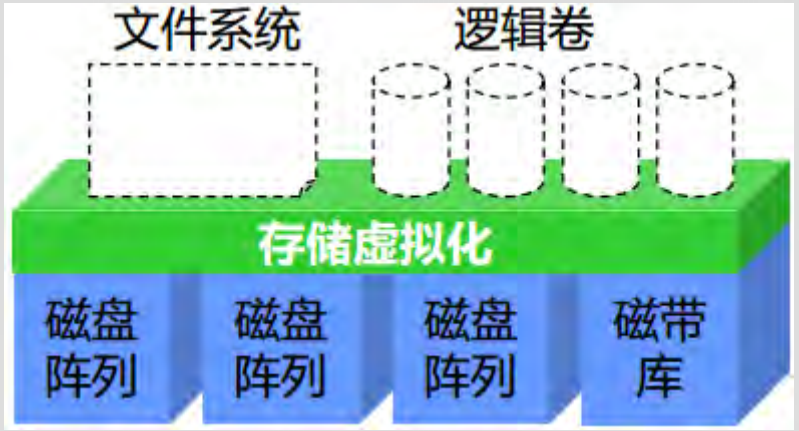
虚拟化，是指通过虚拟化技术将一台计算机虚拟为多台逻辑计算机。在一台计算机上同时运行多个逻辑计算机，每个逻辑计算机可运行不同的操作系统，并且应用程序都可以在相互独立的空间内运行而互不影响，从而显著提高计算机的工作效率。

服务器虚拟化

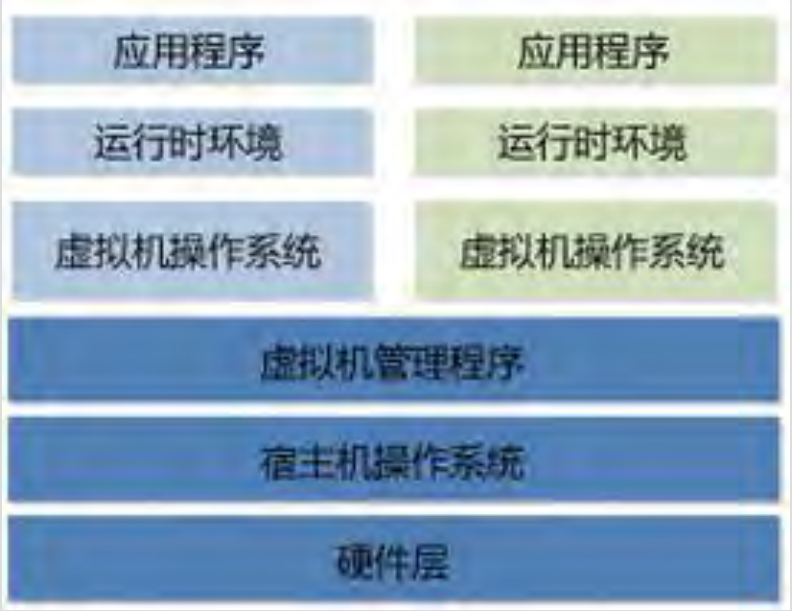


- 整合主机资源
- 提高主机利用率

存储虚拟化



- 整合存储资源
- 和数据库技术结合
  - 数据库高可用
  - 数据库本地保护



虚拟化

容器化

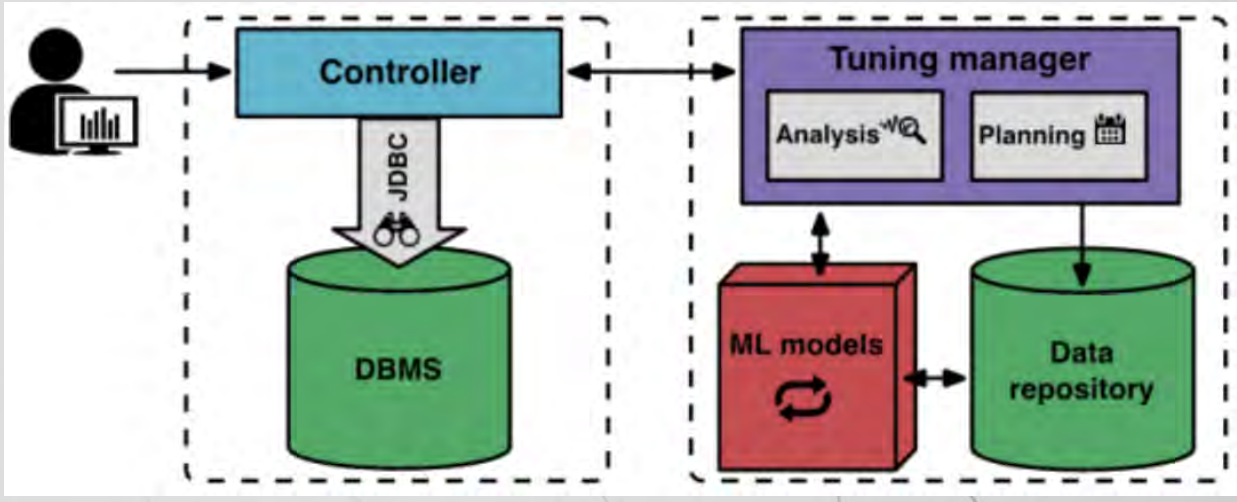
容器化在数据库领域目前应用不多，常见的是在MySQL的单机多实例混跑环境中，使用容器中的资源隔离技术—cgroup，限制单实例可使用的CPU、MEM 、IO资源。

数据库管理变迁

厉害｜DBA泪奔了！亚马逊用机器学习自动调优数据库管理系统！！

导读：最近亚马逊和卡内基梅隆大学一起开发了一套名叫“OtterTune”的机器学习自动化调整DBMS的系统，并公布起设计论文和开源项目，重点解决DBMS长期存在的一些问题：1.对管理人员专业性要求高；2.管理成本高；3.无法实现配置资源最优化等一系列问题。

注意！这无疑是让那些经验丰富待遇丰厚的DBA人员直接失业呀！




Baloo


性能
优化
运维
管理
架构设计

开发 运维 架构 分析 治理 管理 业务

- 数据库技术发展很快，作为DBA不要害怕变革，要勇于拥抱变化，紧跟时代脚步。
- 在纷繁复杂的技术中，不要盲从。各种技术万变不离其中，学好一种，可以融会贯通。
- 结合公司情况、自身情况，不追求技术的”高、精、尖”。脚踏实地，做好现有的工作，一样可以发挥很大作用。
- 深入公司业务，只有这样才能发挥技术的最大价值。

CHINA DATA ANALYST SUMMIT

CDA 数据分析师 www.cda.cn

THANKS

跨界互联 数聚未来

第四届中国数据分析师行业峰会

CHINA DATA ANALYST SUMMIT