



Elasticsearch在vivo互联网的最佳实践

杨振涛

yangzhentao@vivo.com



SPEAKER

杨振涛

vivo互联网搜索团队负责人

2014年加入vivo互联网，带领搜索引擎技术团队。

- 负责搜索引擎及NLP相关的技术与产品开发。
- 负责互联网研发团队的软件过程改进，持续提升交付效率和质量。

之前曾先后任职于 华大基因、TCL，并参与过B2C、IM等创业项目。

- 专注于数据的存储、检索和可视化，及软件过程改进。
- 开源社区支持者，发起Elasticsearch、Redis及Jenkins（国内首次）技术沙龙。
- 技术翻译爱好者，InfoQ中文社区编辑，TED Translator。



CONTENTS

01

vivo的搜索需求

02

不只是Lucene : Elasticsearch从入门到普及

03

案例分析：典型问题

干货

04

关于数据的存储、索引、检索与可视化

05

也谈AI: LTR 与NLP应用现状及未来方向

务虚

vivo的搜索需求

- 垂直搜索/综合搜索
- Android本地搜索
- 企业级搜索
- 基于搜索的实时数据分析

1

vivo的搜索需求

- 互联网产品中的垂直搜索
 - 应用商店
 - 游戏中心
 - 主题/壁纸/铃声/字体
 - 商城
 - 社区
 - 官网
 - 音乐
 - 阅读



vivo的搜索需求

- 综合搜索
 - 浏览器入口
 - 系统全局搜索入口
 - 其他入口



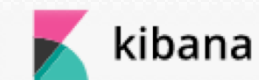
vivo的搜索需求

- 企业级搜索
 - 各种企业级业务系统产生的各种数据，
需要一个统一的搜索入口
 - 知识管理
 - 客服QA
 - 系统日志搜索



vivo的搜索需求

- 基于搜索的实时数据分析
- 轻量灵活的BI



搜索团队建设

Since 2014 ~

- 技术团队：工程、算法（Rank、NLP）、数据、商业策略
- 产品、运营、品质评测

不只是Lucene：Elasticsearch从入门到普及

- 搜索引擎的技术选型
- Java技术栈
- Elasticsearch

2

搜索引擎的技术选型

开源方案 **vs** 自研

C/C++ **vs** Java /Python /Go...

sphinx

Xapian

Java技术栈

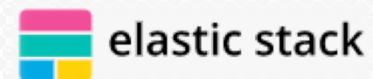
- Lucene
- Solr/SolrCloud
- Elasticsearch

Zoie

Nutch

sensei

bobo



Elasticsearch应用场景划分

索引/请求	小请求	大请求
小索引	Lucene ? Solr ? Elasticsearch ?	Elasticsearch
大索引	Elasticsearch	Elasticsearch

Elasticsearch入门

- 基本数据类型
- 数据写入
- 常用查询
- 常用聚合
- 集群拓扑与容量规划
- 最佳实践
- 使用指导

默认不使用别名，
不算入门ES！



<https://www.elastic.co/guide/cn/index.html>

Elasticsearch普及

除了搜索产品外，诸多项目中都使用了Elasticsearch：

- 推荐系统
- 广告系统
- DMP
- PUSH
- 云诊断
- 智能客服
- 智能问答
- 企业内部应用

存储三剑客

MySQL

Redis

Elasticsearch



案例分析

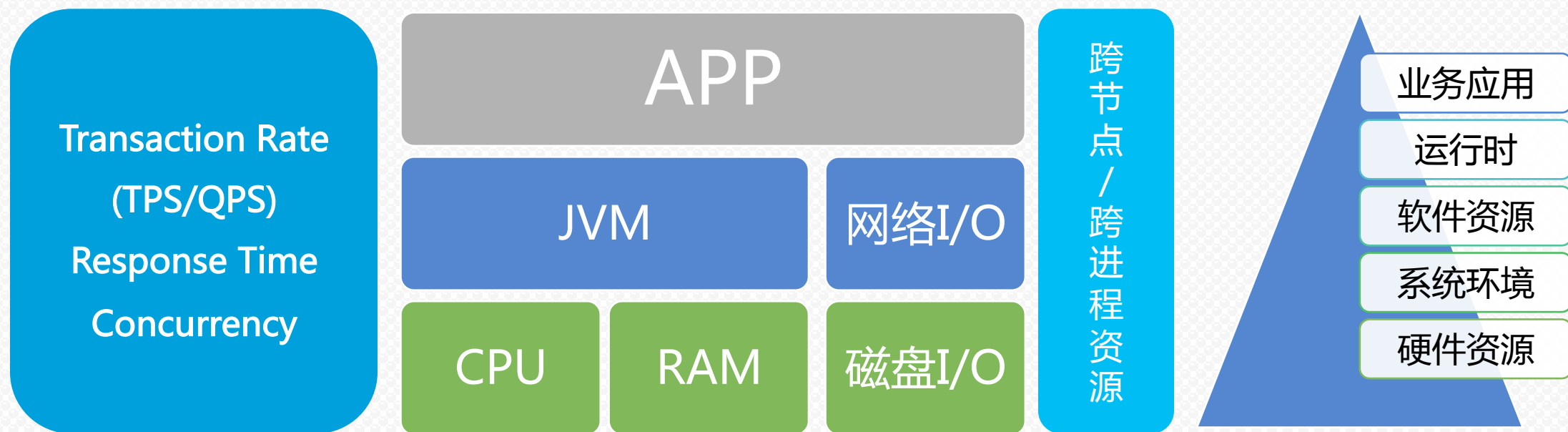
一个优化案例

3

基础功能

- 爬虫与采集
- 索引数据
 - 写入速度
 - 批量写入
- 检索
 - 查询
 - 聚合
 - 排序
 - Scroll ID与翻页
- 可视化
 - Kibana

性能分析模型



优化案例：问题描述与优化目标

- 某业务的ES集群，初始默认3个节点，5 shards 0 replicas，版本 2.x, JDK7, CentOS 6
- 节点配置 物理机 64G 32核 SATA盘
- 一个索引库，业务上线后持续增长，目前存储1.6亿+的用户标签数据，大小1T
- 采用scroll方式每次根据业务条件检索数万到数千万用户ID

当前QPS
5k/s

优化目标
4w~5w/s

分析优化过程

- 基准数据
 - 抽样一个典型的业务查询，Scroll + filter 测试QPS 5k/s
- 排除缓存
 - ES缓存和操作系统缓存，同一个查询第二次可以达到2w+/s
- 主观判断优化点
 - 单分片数据量过大，可以尝试调大分片数，减少单分片数据量
 - 集群拓扑结构不合理，可以尝试新增节点，分配不同角色

分析优化过程

- 建立观测指标
 - CPU负载
 - 内存消耗
 - 本地磁盘I/O，网络I/O
 - JVM CPU消耗，堆与GC，栈与线程，类对象
 - ES的各个线程池与队列负荷

分析优化过程

- 疑似瓶颈：磁盘I/O，且从bigdesk看到大部分耗时在fetch阶段
- 针对措施：
 - 每个节点有6块硬盘，共3个节点
 - 先设置索引库分片为18个，暂不设置副本
 - 配置 path.data，保证一块硬盘一个分片
- 基本效果：
 - 检索速度由5k/s提升到2.8w/s，5倍+

```
29 # ----- Paths -----
30 #
31 # Path to directory where to store the data (separate multiple locations by comma):
32 #
33 path.data: /data/sd2/es,/data/sd3/es,/data/sd4/es,/data/sd5/es,/data/sd6/es,/data/sd7/es
34 #
35 # Path to log files:
36 #
37 path.logs: /path/to/logs
38 #
```

分析优化过程

- 从代码层面梳理业务逻辑，寻找优化点
 - 检查mapping，默认将所有字段写入 _all 字段，实际业务并不需要，设置 _all 为 false；存储空间占用从 1T 下降到 892G，QPS 从 2.8w/s 提升到 3w+/s，约 10%
 - 根据业务条件查询标签数据时，只需要知道用户 ID，而无需其他字段，所以不返回 _source；QPS 从 3w/s 提升到 3.4w/s，约 11%

分析优化过程

- 再次检查磁盘负载，使用率较高

Device:	rrqm/s	wrqm/s	r/s	w/s	rMB/s	wMB/s	avgrq-sz	avgqu-sz	await	r_await	w_await	svctm	%util
sdm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdj	3.50	0.00	1411.50	0.00	35.69	0.00	51.79	3.18	2.29	2.29	0.00	0.62	87.45
sdc	2.00	0.00	1380.00	0.00	32.89	0.00	48.81	2.93	2.12	2.12	0.00	0.56	76.90
sde	3.00	0.00	1429.00	0.00	35.85	0.00	51.38	3.88	2.72	2.72	0.00	0.65	93.15
sdd	2.00	0.00	1353.00	0.00	35.65	0.00	53.96	2.06	1.52	1.52	0.00	0.54	73.65
sdf	2.00	0.00	1440.00	0.00	38.15	0.00	54.26	3.11	2.35	2.35	0.00	0.56	81.05
sdg	1.00	0.00	1426.00	0.00	35.74	0.00	51.32	2.56	1.79	1.79	0.00	0.53	75.85
sdh	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdi	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdj	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdh	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdb	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sda	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

最终结果

- 在review代码时发现的其他优化点：
 - TransportClient 连接方式改为 NodeClient方式，可减少一次网络连接开销，QPS提升至4.02w/s，约18%
 - 排序优化：默认scroll排序优化为使用 _doc排序，依赖于Lucene内部id排序，QPS提升至4.81w/s，约20%

继续优化的空间

- Elasticsearch 版本升级到5.x或6.x ，同步升级JDK到8或9，升级Lucene 6
 - 可以使用 Sliced Scroll 增加并发；可以使用 profile API；可以使用 shrink API 收缩shard数
 - 可以使用 reindex 命令；原生的RestClient ；Ingest Node & Pipeline ；String -> text & keyword
 - 无缝滚动升级，提供迁移助手，支持序列ID，支持排序索引
- 资源允许时升级为SSD盘
- 期间的无效尝试
 - 调大es的search线程数和队列数，无显著差别
 - 对于and条件检索，分别对比TermQuery或BoolQuery，无显著差别
 - 每次检索返回数据量500,1000,2000,5000，速度无显著差别
 - 添加副本数，无显著差别（本案例中每个主分片独占一块磁盘，所以只能提升可靠性，无法提高读取性能，而且会降低写入速度）

性能问题参考 Qbox官博系列文章

How to Maximize Elasticsearch Indexing Performance

- <https://qbox.io/blog/maximize-guide-elasticsearch-indexing-performance-part-1>
- <https://qbox.io/blog/maximize-guide-elasticsearch-indexing-performance-part-2>
- <https://qbox.io/blog/maximize-guide-elasticsearch-indexing-performance-part-3>

The Authoritative Guide to Elasticsearch Performance Tuning

- <https://qbox.io/blog/authoritative-guide-elasticsearch-performance-tuning-part-1>
- <https://qbox.io/blog/elasticsearch-performance-tuning-part-2-zen>
- <https://qbox.io/blog/maximize-guide-elasticsearch-indexing-performance-part-3>

The Authoritative Guide to Elasticsearch Search Tuning

- <https://qbox.io/blog/elasticsearch-search-tuning-5-0-ultimate-guide>
- <https://qbox.io/blog/elasticsearch-search-tuning-part-2>
- <https://qbox.io/blog/authoritative-elasticsearch-search-tuning-part-3>

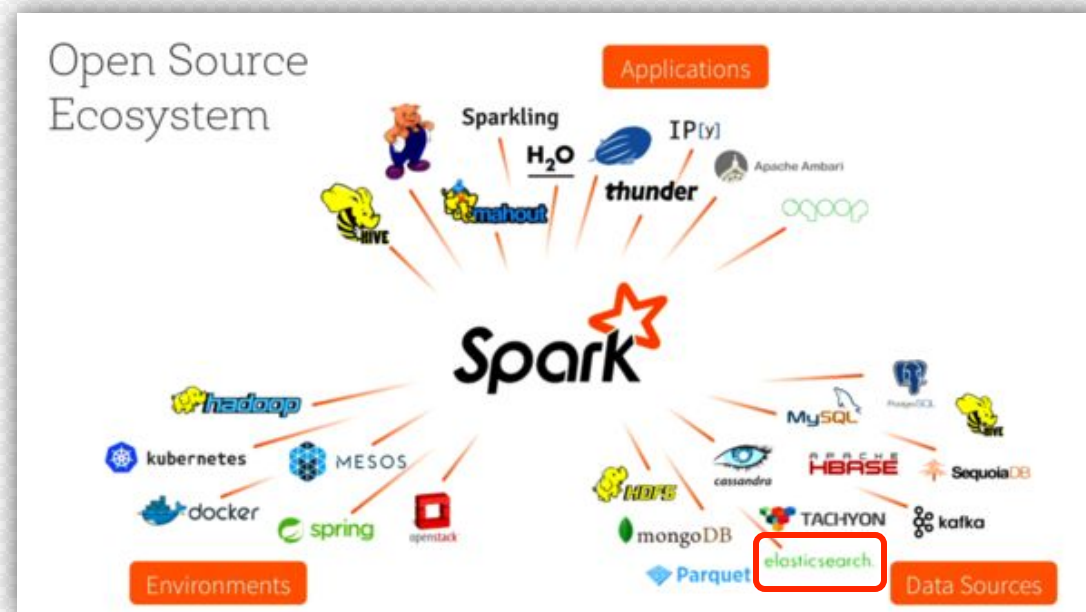
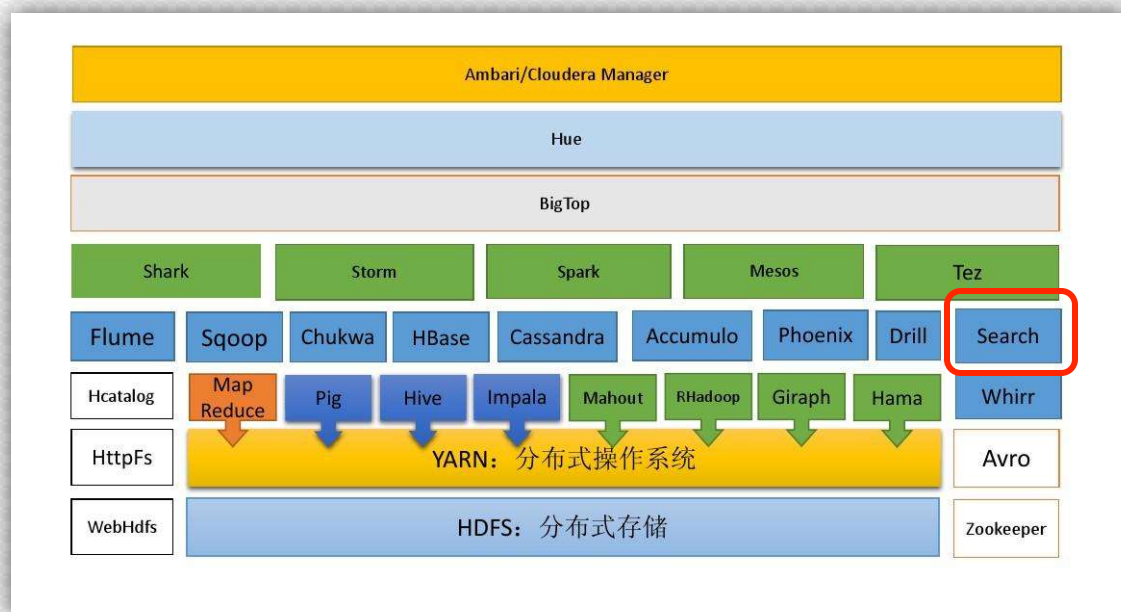


关于数据的存储、索引、检索与可视化

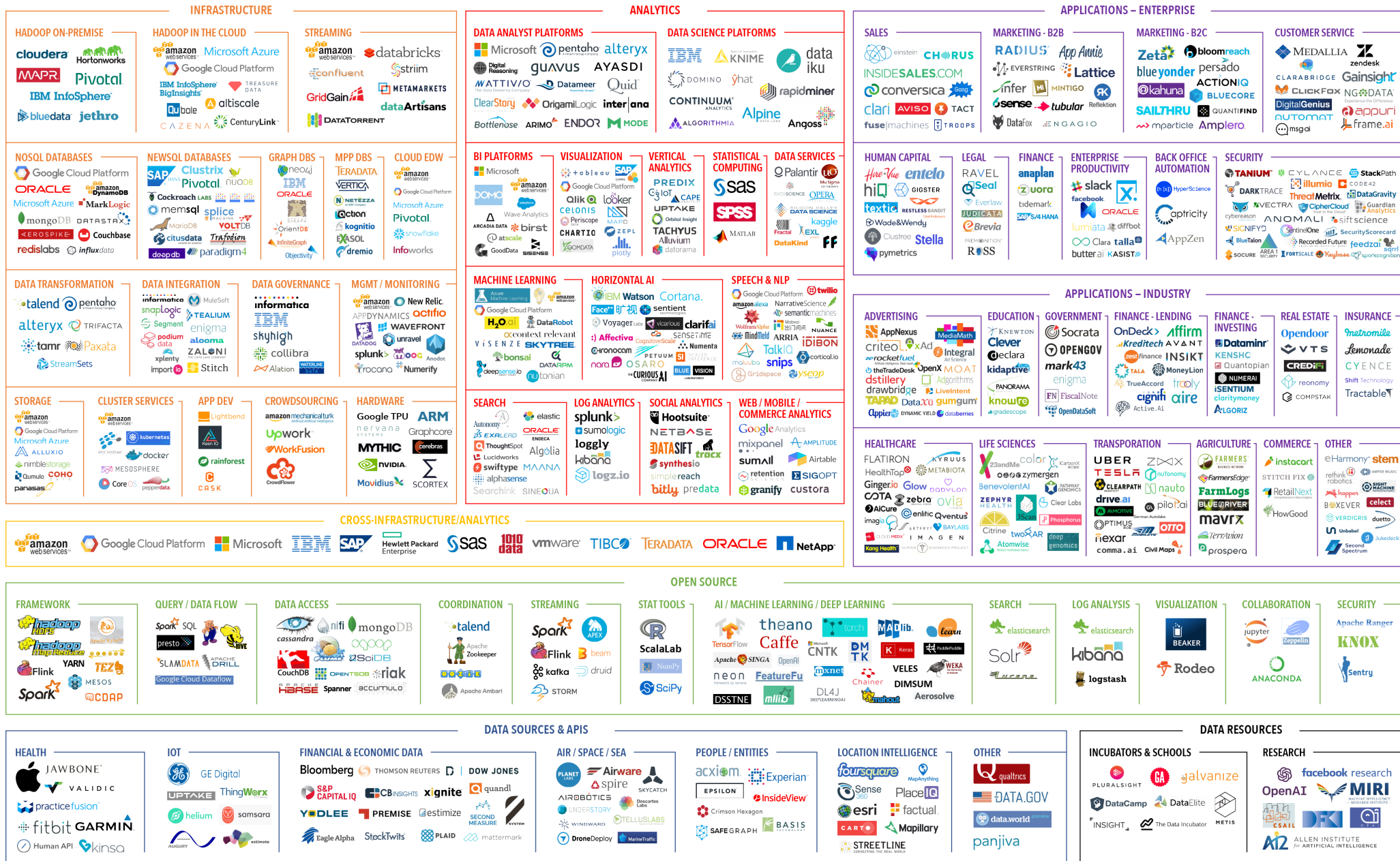
- 存储
 - 索引
 - 检索
 - 可视化
-
- 与大数据、人工智能的关系？

4

大数据Hadoop生态圈



BIG DATA LANDSCAPE 2017



Flink	Open Source	Streaming
Kafka	Open Source	Streaming
Spark	Open Source	Streaming
Storm	Open Source	Streaming
NumPy	Open Source	Stat Tools
R	Open Source	Stat Tools
ScalaLab	Open Source	Stat Tools
SciPy	Open Source	Stat Tools
Knox	Open Source	Security
Ranger	Open Source	Security
Sentry	Open Source	Security
ElasticSearch	Open Source	Search
Lucene	Open Source	Search
Solr	Open Source	Search
Drill	Open Source	Query/Data Flow
Google Cloud DataFlow	Open Source	Query/Data Flow
HIVE	Open Source	Query/Data Flow
PIG	Open Source	Query/Data Flow
Presto	Open Source	Query/Data Flow
SlamData	Open Source	Query/Data Flow
Spark SQL	Open Source	Query/Data Flow
Elasticsearch	Open Source	Log Analysis
Kibana	Open Source	Log Analysis
Logstash	Open Source	Log Analysis
CDAP	Open Source	Framework
Flink	Open Source	Framework
HDFS	Open Source	Framework

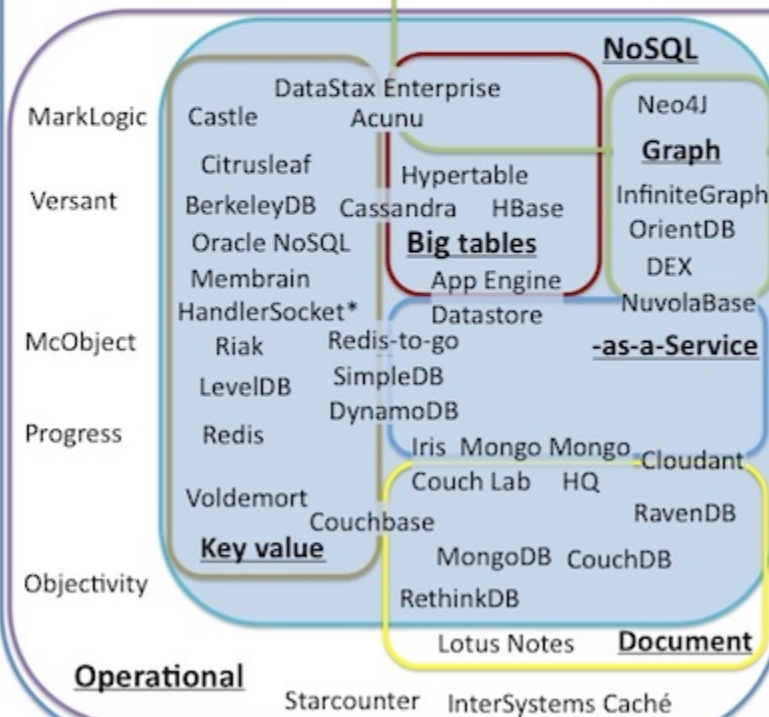
存储

- RDBMS
 - MySQL、PostgreSQL
- NoSQL
 - Key-Value : Riak、Redis、Memcached、Amazon DynamoDB
 - Document Oriented : MongoDB、CouchDB、RavenDB
 - Wide Column Store : Cassandra、HBase
 - Graph Oriented : Neo4J、Infinite Graph、OrientDB
- NewSQL
 - Clustrix、GenieDB、ScalArc、Schooner、VoltDB、RethinkDB、ScaleDB、Akiban、CodeFutures、ScaleBase、Translattice、NimbusDB、Drizzle

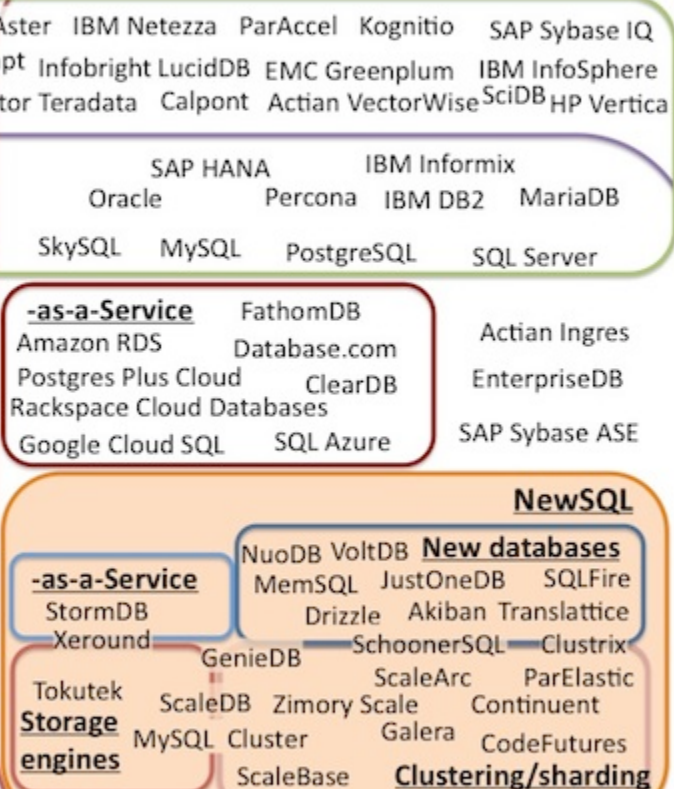
The evolving database landscape

451 Research

Non-relational



Relational



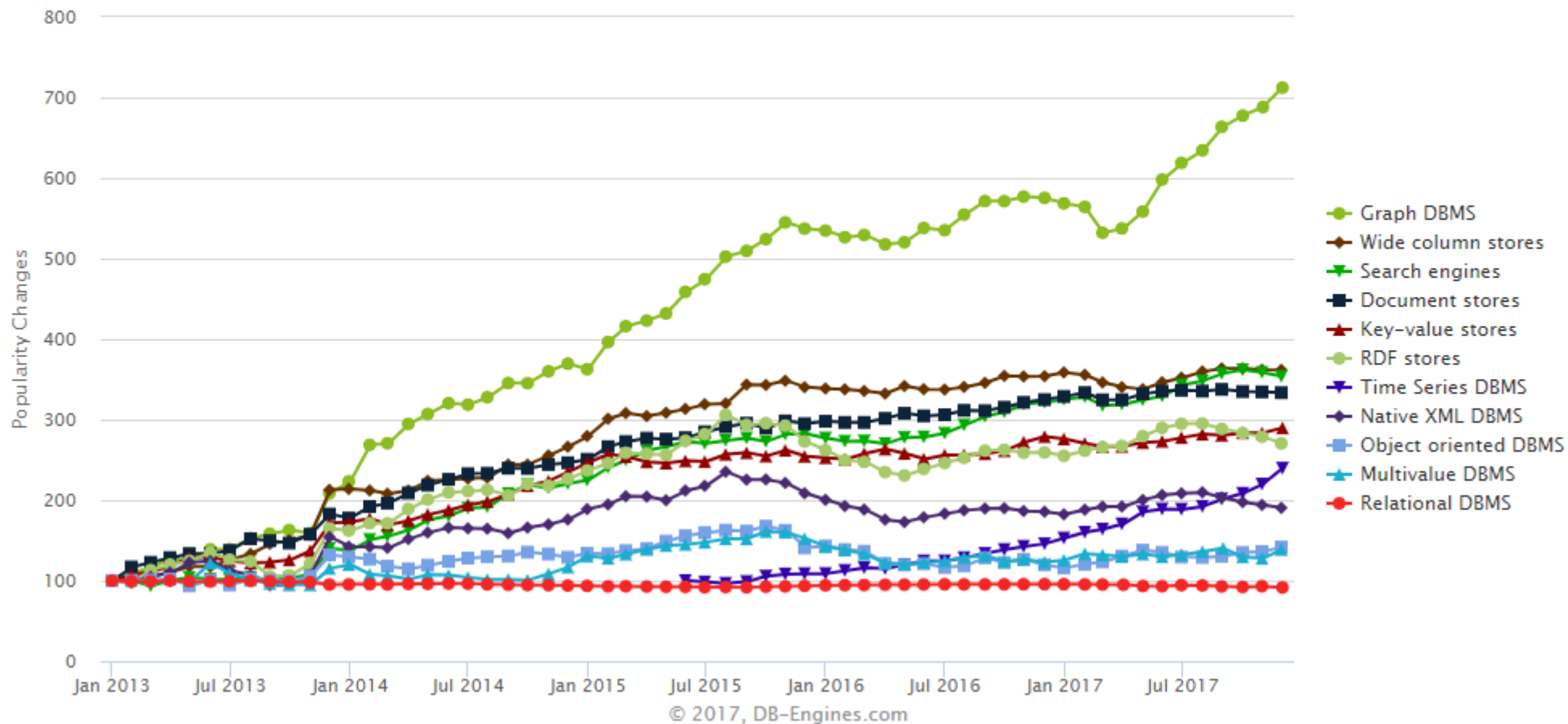
© 2012 by The 451 Group. All rights reserved

Rank			DBMS	Database Model	Score		
Dec 2017	Nov 2017	Dec 2016			Dec 2017	Nov 2017	Dec 2016
1.	1.	1.	Oracle +	Relational DBMS	1341.54	-18.51	-62.86
2.	2.	2.	MySQL +	Relational DBMS	1318.07	-3.96	-56.34
3.	3.	3.	Microsoft SQL Server +	Relational DBMS	1172.48	-42.59	-54.17
4.	4.	4.	PostgreSQL +	Relational DBMS	385.43	+5.51	+55.41
5.	5.	5.	MongoDB +	Document store	330.77	+0.29	+2.09
6.	6.	6.	DB2 +	Relational DBMS	189.58	-4.48	+5.24
7.	7.	8.	Microsoft Access	Relational DBMS	125.88	-7.43	+1.18
8.	9.	9.	Redis +	Key-value store	123.24	+2.05	+3.34
9.	8.	7.	Cassandra +	Wide column store	123.21	-1.00	-11.07
10.	10.	11.	Elasticsearch +	Search engine	119.78	+0.37	+16.51
11.	11.	10.	SQLite +	Relational DBMS	115.19	+2.44	+4.36
12.	12.	12.	Teradata	Relational DBMS	74.74	-3.49	+1.37
13.	13.	14.	Solr	Search engine	66.30	-2.86	-2.70
14.	14.	13.	SAP Adaptive Server	Relational DBMS	65.68	-1.35	-4.74
15.	15.	16.	Splunk	Search engine	63.79	-1.08	+8.87
16.	16.	15.	HBase	Wide column store	63.41	-0.15	+4.79
17.	18.	20.	MariaDB +	Relational DBMS	56.73	+1.44	+12.64
18.	17.	17.	FileMaker	Relational DBMS	55.20	-3.64	+1.08
19.	19.	19.	Hive +	Relational DBMS	54.67	+1.42	+5.27
20.	20.	18.	SAP HANA +	Relational DBMS	46.49	-2.69	-5.28
21.	21.	21.	Neo4j +	Graph DBMS	38.73	+0.28	+1.90
22.	22.	22.	Amazon DynamoDB +	Document store	36.71	-0.41	+6.70
23.	23.	23.	Couchbase +	Document store	31.31	-1.01	+1.58
24.	24.	24.	Memcached	Key-value store	28.43	+0.45	-0.34
25.	25.	25.	Informix	Relational DBMS	27.45	-0.26	+0.38
26.	26.	28.	Microsoft Azure SQL Database	Relational DBMS	21.85	-0.26	+1.02
27.	27.	27.	Vertica +	Relational DBMS	20.81	-0.84	-0.28
28.	28.	26.	CouchDB	Document store	20.66	+0.16	-1.69
29.	30.	30.	Firebird	Relational DBMS	17.25	-0.02	+1.57
30.	29.	29.	Netezza	Relational DBMS	17.24	-1.29	-2.19
31.	31.	34.	Impala	Relational DBMS	14.46	+0.61	+4.77
32.	32.	59.	Microsoft Azure Cosmos DB +	Multi-model	14.13	+1.10	+10.71



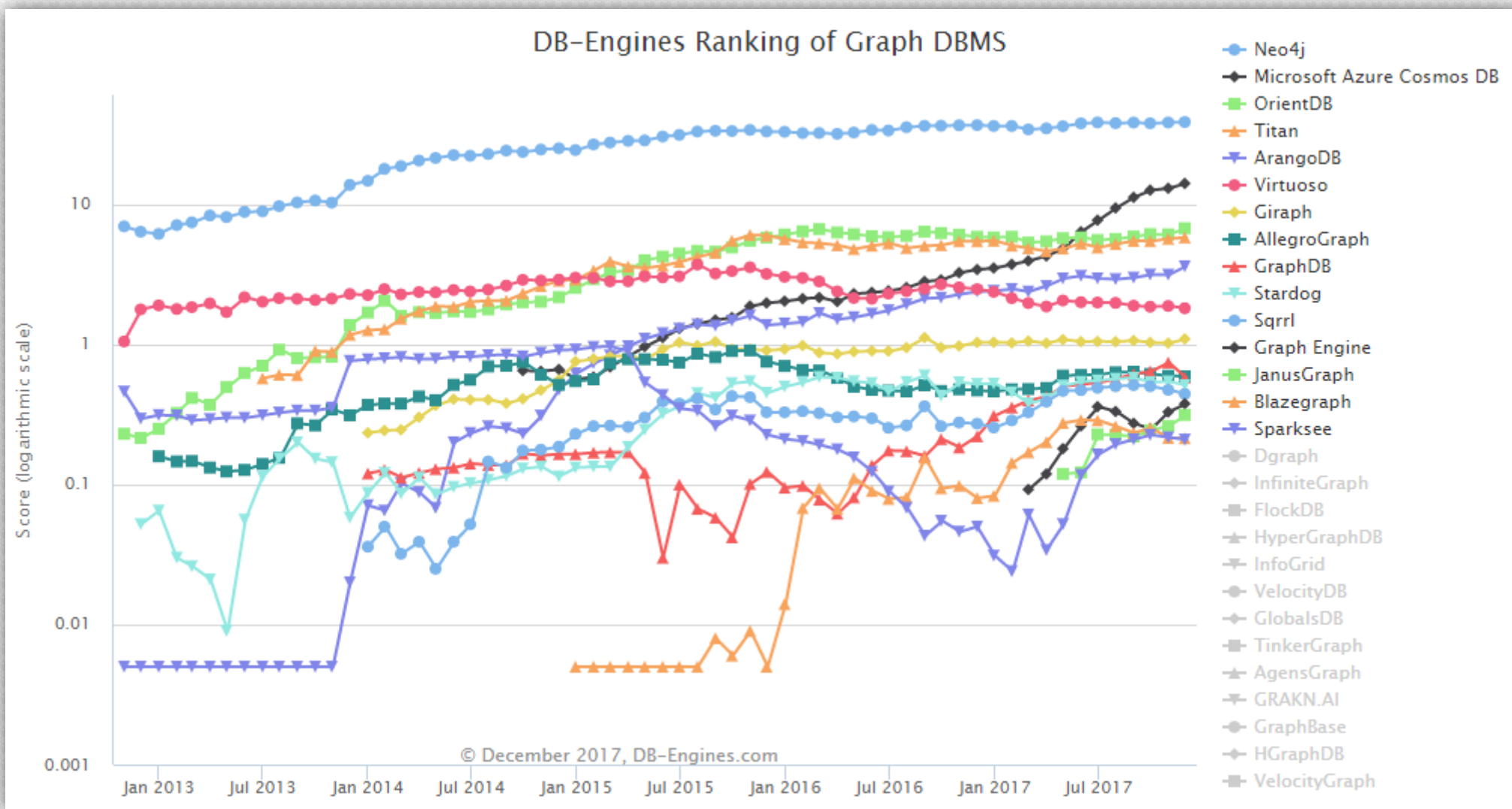
GraphDB

Complete trend, starting with January 2013



https://db-engines.com/en/ranking_categories

Neo4j

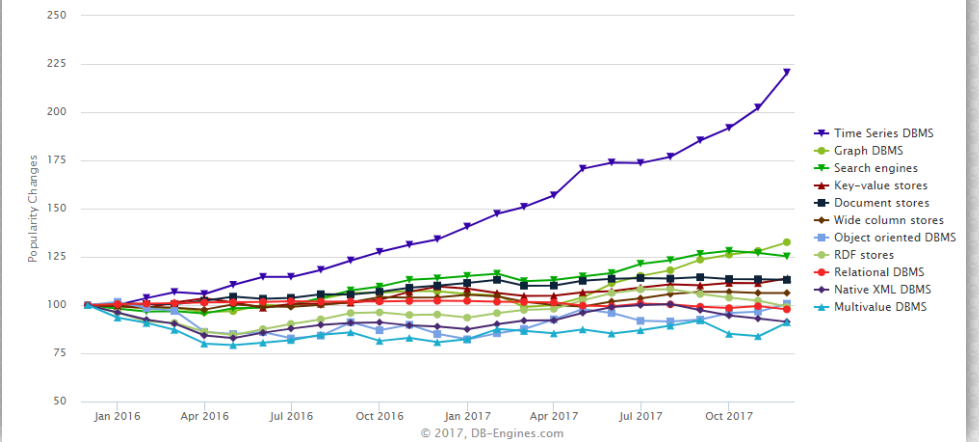


TSDB

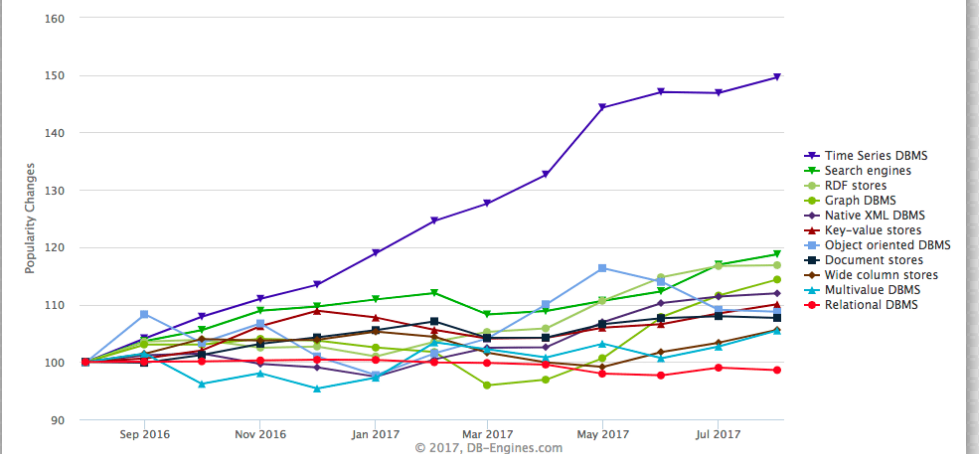
23 systems in ranking, December 2017

Rank			DBMS	Database Model	Score		
Dec 2017	Nov 2017	Dec 2016			Dec 2017	Nov 2017	Dec 2016
1.	1.	1.	InfluxDB	Time Series DBMS	9.87	+0.53	+4.00
2.	4.	5.	Kdb+	Multi-model	3.73	+1.88	+2.52
3.	2.	2.	RRDtool	Time Series DBMS	3.17	-0.03	+0.73
4.	3.	3.	Graphite	Time Series DBMS	2.65	-0.21	+0.75
5.	5.	4.	OpenTSDB	Time Series DBMS	1.99	+0.28	+0.55
6.	6.	6.	Druid	Time Series DBMS	0.96	-0.02	+0.29
7.	7.	7.	Prometheus	Time Series DBMS	0.83	+0.02	+0.48
8.	8.	8.	KairosDB	Time Series DBMS	0.44	-0.02	+0.18
9.	9.	9.	eXtremeDB	Multi-model	0.29	-0.01	+0.10
10.	10.	10.	Riak TS	Time Series DBMS	0.22	+0.01	+0.10
11.	12.	16.	Blueflood	Time Series DBMS	0.16	-0.01	+0.13
12.	11.	19.	Hawkular Metrics	Time Series DBMS	0.15	-0.01	+0.15
13.			TimescaleDB	Time Series DBMS	0.05		
14.	13.	18.	Axibase	Time Series DBMS	0.04	-0.05	+0.04
15.	16.	13.	Warp 10	Time Series DBMS	0.04	-0.01	-0.01
16.	14.	17.	Machbase	Time Series DBMS	0.03	-0.04	+0.02
17.	15.	15.	TempoIQ	Time Series DBMS	0.02	-0.04	-0.00
18.	17.	12.	Heroic	Time Series DBMS	0.00	-0.02	-0.06
18.	18.		IRONdb	Time Series DBMS	0.00	±0.00	
18.	18.	14.	Newts	Time Series DBMS	0.00	±0.00	-0.03
18.	18.		SiriDB	Time Series DBMS	0.00	±0.00	
18.	18.	19.	SiteWhere	Time Series DBMS	0.00	±0.00	±0.00
18.	18.	11.	Yanza	Time Series DBMS	0.00	±0.00	-0.06

Trend of the last 24 months

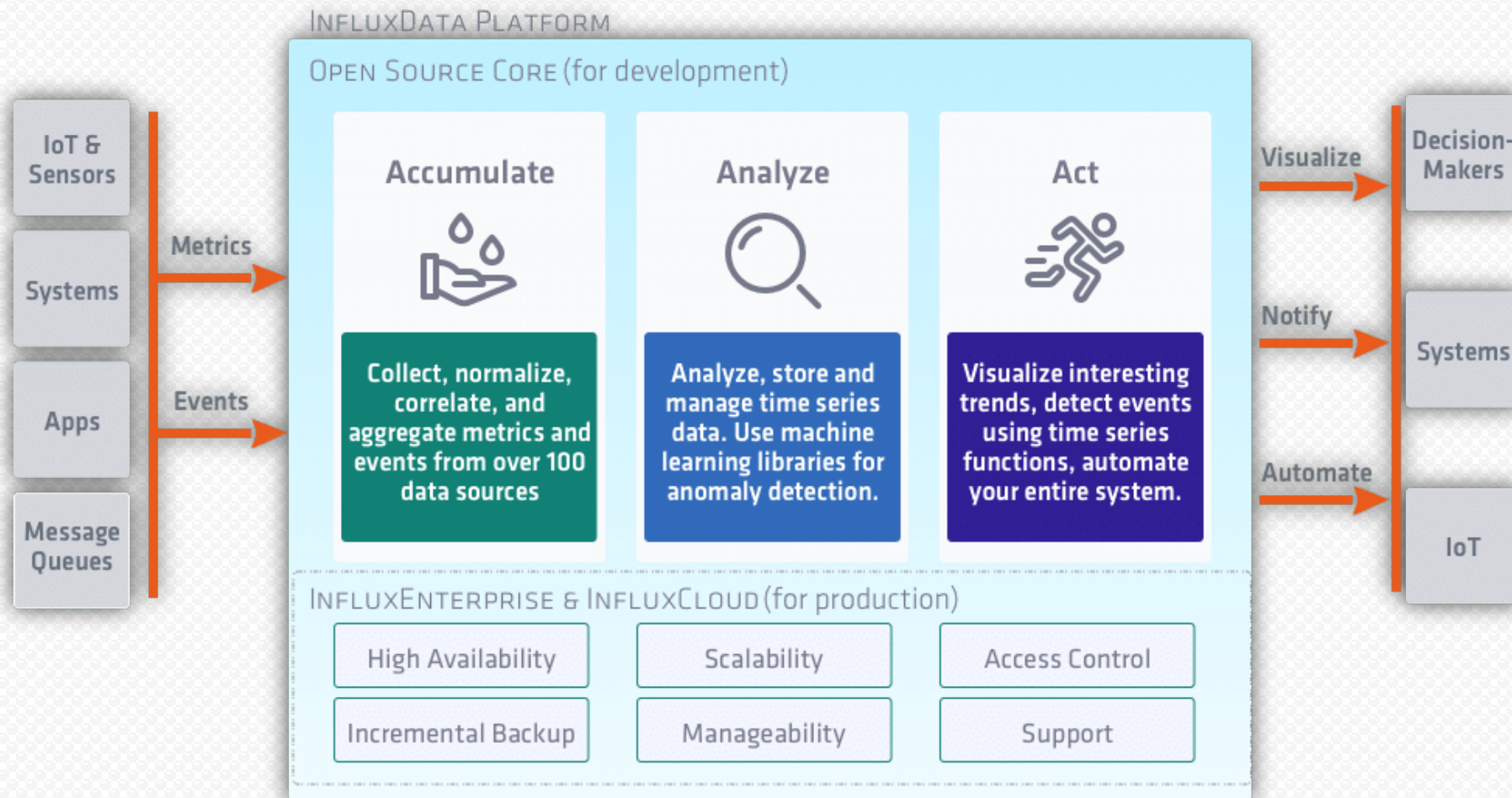


Trend of the last 12 months



InfluxDB

TSDB 时序数据库代表



典型应用

- 智能搜索
- 知识图谱
- 日志分析



索引（一个查找问题）

- B-tree索引（B+tree /平衡树）
- 哈希映射索引 (Hash Bucket)
- BitMap索引
- 机器学习索引
 - 可学习索引框架
Learning Index Framework, LIF
 - 递归模型索引
Recursive-Model Indexes, RMI
 - 基于标准误差的搜索策略

倒排索引：词典+倒排表

数据结构	优缺点
排序列表	实现简单，但性能差
哈希表	性能高，内存消耗大
跳跃表	占用内存小，且可调，但对模糊查询支持不好
B树	磁盘索引，更新方便，但检索速度慢，数据库应用较多
字典树	查询效率只跟字符串长度有关，但只适合英文词典
双数组字典树	可做中文词典，内存占用小，分词工具应用较多
Finite State Transducers (FST)	共享前缀，内存消耗小，但要求输入有序，更新不易

<http://blog.csdn.net/njpjsoftdev/article/details/540>

Jeff Dean出品：用机器学习索引替代B-Trees，3倍性能提升，10-100倍空间缩小
<http://mp.weixin.qq.com/s/MpuUdZi8CWcu0b-ij-bHjA>

检索

- 查询理解
 - 分析
 - 重写
 - 意图识别
- 召回
 - 相关性 TF-IDF BM25
 - 重要性/质量度
- 排序/重排序
- 效果评估

如何做好搜索？
每天坚持分析Bad Case！

可视化

- Kibana – ELK
- Grafana (JavaScript语言)
 - + collectd (C语言) ; 可选替代 Metrics (Java语言) <http://metrics.dropwizard.io>
 - + InfluxDB (GO语言)
 - collectd/Metrics + InfluxDB + Grafana = CIG/MIG ?
- ECharts/D3.js – 画个图 ?

可视化 Circos

<http://circos.ca> 以圆形风格见长的开源数据可视化工具，Perl语言编写，主要用于基因组领域，正在更多地应用到其他领域的关系型数据可视化。



也谈AI：LTR与NLP应用现状及未来

- LTR
 - 机器学习排序
- NLP
 - 搜索引擎与NLP
 - DL与NLP
 - 知识表示、知识库、知识图谱

5

AI、ML、DL、IR、NLP

人工智能概述：七大目标，三大阶段，三大类型

NLP、人工智能、机器学习、深度学习、神经网络之间的区别与联系

NLP与文本挖掘（或文本分析）之间的不同

大数据中的NLP

什么是聊天机器人

NLP中为什么需要深度学习

日志分析与日志挖掘中的NLP

深度自然语言处理

NLP的其他典型应用领域：自动摘要 情感分析 信息提取 文本分类

一篇入门好文：

人工智能与自然语言处理概述：AI三大阶段、NLP关键应用领域 <http://36kr.com/p/5074076.html>

英文原文 <https://www.xenonstack.com/blog/overview-of-artificial-intelligence-and-role-of-natural-language-processing-in-big-data>

Goals of Artificial Intelligence

-  Reasoning
-  Automated Learning & Scheduling
-  Machine Learning
-  Natural Language Processing
-  Computer Vision
-  Robotics
-  General Intelligence

• 三大阶段

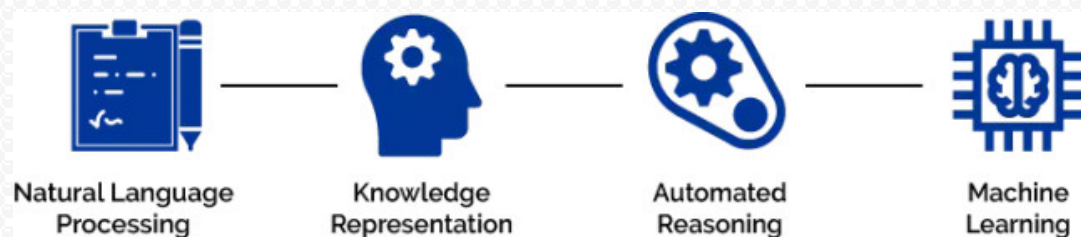
机器学习；机器智能；机器意识

• 三大类型

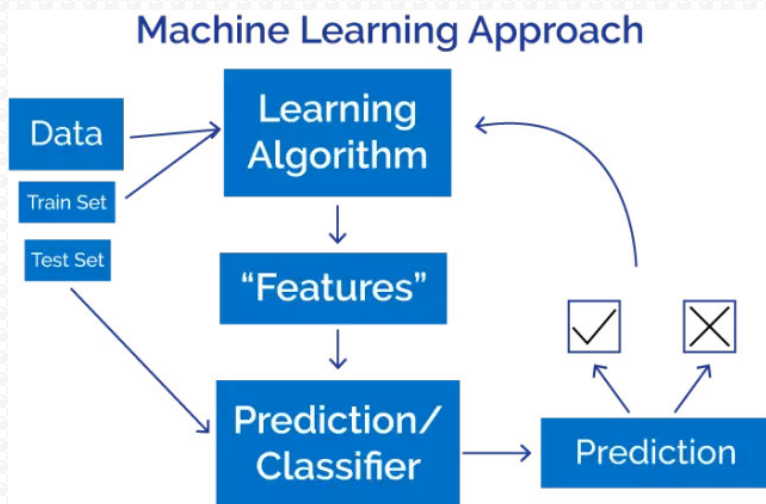
ANI（狭义人工智能）；AGI（通用人工智能）；ASI（强人工智能）

AI、ML、DL、IR、NLP

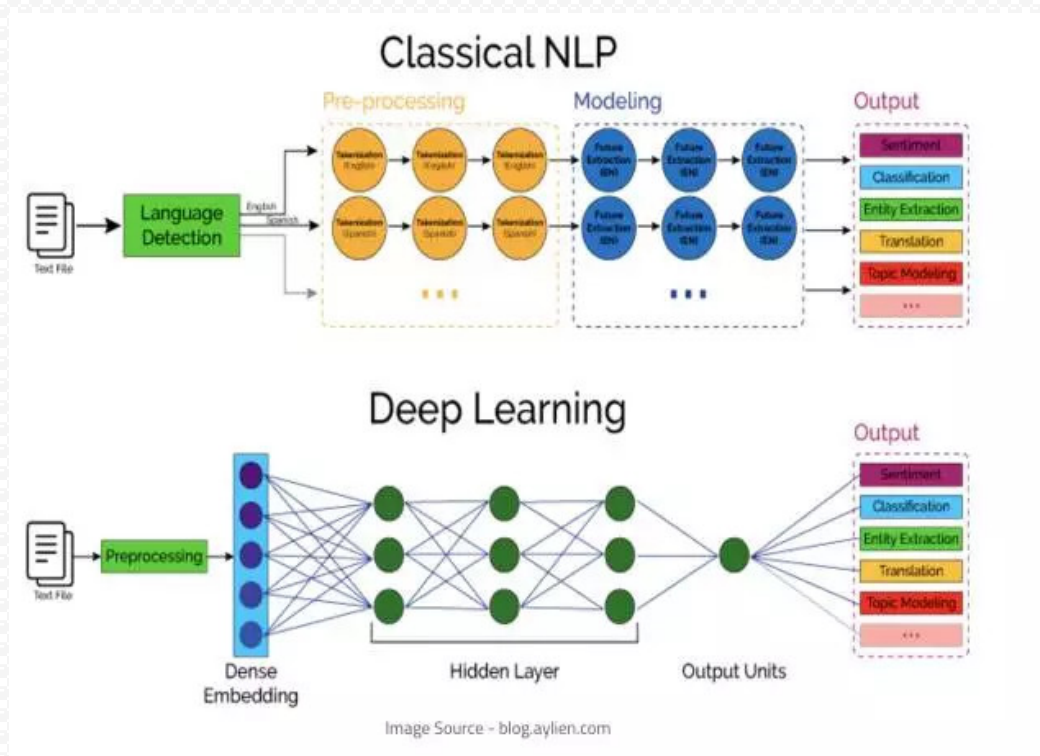
什么使得系统智能化？



典型机器学习过程



传统NLP vs 深度学习



NLP

- 自然语言处理
 - 分析
 - 理解
 - 生成
- 任务类型
 - 分类
 - 匹配
 - 翻译
 - 结构化预测
 - 序贯决策过程

Task	Description	Model	Applications
Classification	assign a label to a string	$s \rightarrow c$ s : string, c : label	text classification, sentiment analysis
Matching	matching two strings	$s, t \rightarrow R^+$ s : string, t : string R^+ : non - negative real values	search, question answering, single turn dialogue (retrieval based)
Translation	transform one string to another	$s \rightarrow t$ s : string, t : string	machine translation, automatic speech recognition, single turn dialogue (generation based)
structured prediction	map a string to a structure	$s \rightarrow [s]$ s : string, $[s]$: structure	named entity recognition, word segmentation, part-of-speech tagging, dependency parsing, semantic parsing
sequential decision process	take actions in states in dynamically changing environment	$\pi : s \rightarrow a$ π : policy, s : state, a : action	multi-turn dialogue

典型应用：Learning to Rank

- es-ltr By **OpenSource Connections**

<https://github.com/o19s/elasticsearch-learning-to-rank>

- 在Elasticsearch中应用机器学习排序LTR

<http://www.infoq.com/cn/articles/we-are-bringing-learning-to-rank-to-elasticsearch>

- 认识机器学习排序LTR

<http://www.infoq.com/cn/articles/machine-learning-sequencing-ltr>

- 机器学习排序LTR入门——线性模型

<http://www.infoq.com/cn/news/2017/10/Machine-learn-LTR-linear-model>

典型应用：Elastic X-Pack Machine Learning



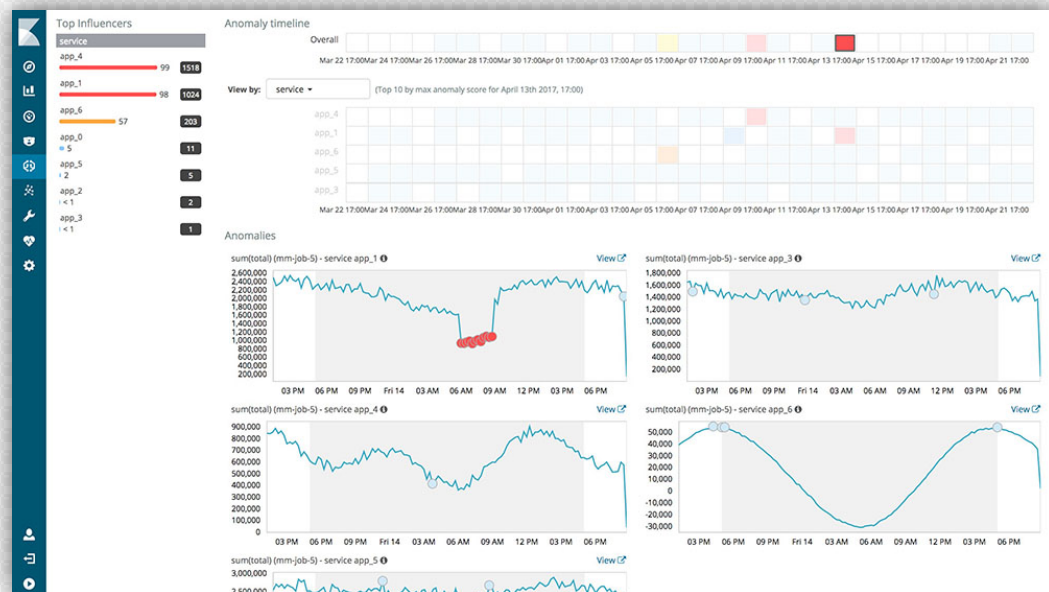
IT 运维: 发现应用请求数量异常降低, 然后深入探究导致异常的问题服务器。



安全性分析: 发现异常网络活动或用户行为, 趁攻击者制造破坏之前准确地定位攻击者。



业务分析: 如果您的电子商务网站中废弃购物车异常增多, 则会给您发出通知。

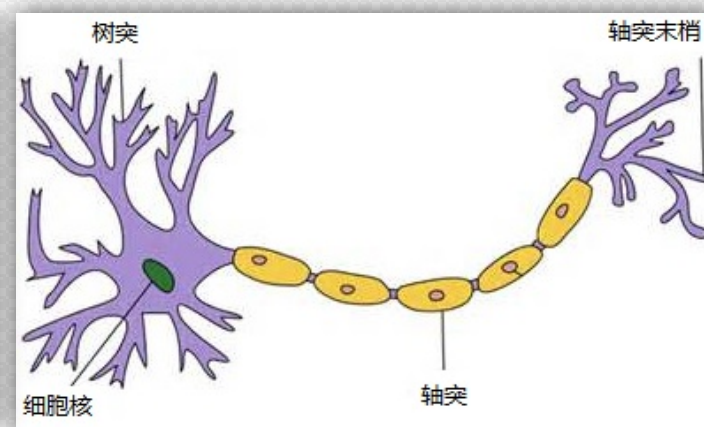
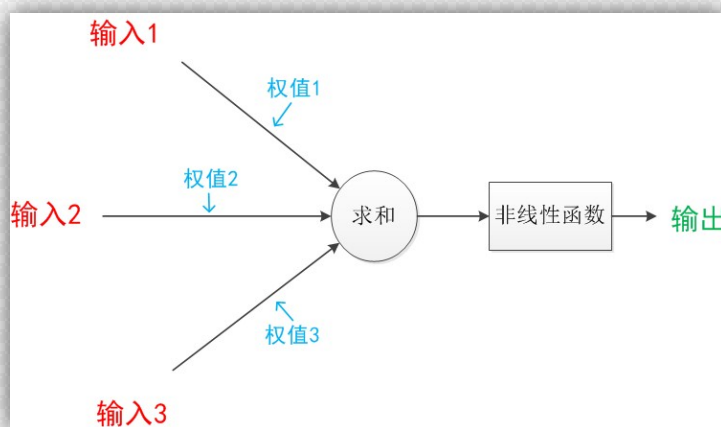


DL在NLP领域的应用

- RNN – LSTM – seq2seq
Encoder-Decoder , Attention Mechanism

- 应用领域

- 机器翻译
- 智能对话与问答
- 自动编码与分类器训练



Chatbot

- 意图分类 Intent Classification
- 实体提取 Entity Extraction
- Word Embedding : word2vec



知识图谱

- 智能搜索/语义搜索
- 查询理解
- 知识问答
- 反欺诈，不一致性验证
- 异常分析

27 systems in ranking, December 2017

Rank			DBMS	Database Model	Score		
Dec 2017	Nov 2017	Dec 2016			Dec 2017	Nov 2017	Dec 2016
1.	1.	1.	Neo4j	Graph DBMS	38.73	+0.28	+1.90
2.	2.	4.	Microsoft Azure Cosmos DB	Multi-model	14.13	+1.10	+10.71
3.	3.	2.	OrientDB	Multi-model	6.75	+0.65	+0.88
4.	4.	3.	Titan	Graph DBMS	5.81	+0.13	+0.35
5.	5.	6.	ArangoDB	Multi-model	3.61	+0.47	+1.22
6.	6.	5.	Virtuoso	Multi-model	1.81	-0.07	-0.67
7.	7.	7.	Giraph	Graph DBMS	1.09	+0.07	+0.06
8.	9.	9.	AllegroGraph	Multi-model	0.59	-0.00	+0.12
9.	8.	11.	GraphDB	Multi-model	0.58	-0.17	+0.36
10.	10.	8.	Stardog	Multi-model	0.51	-0.03	-0.01

<https://db-engines.com/en/ranking/graph+dbms>

知识图谱

- Google : 2012年收购MetaWeb , 推出Knowledge Graph、语音搜索、Google Now
- 百度搜索 : 2013年上线 百度知心 (<http://tupu.baidu.com>)
- 搜狗搜索 : 2012年11月上线 知立方
- Facebook : 2013年 Graph Search
- 微软Bing : 2013年 Satori

《Introducing the Knowledge Graph: Things, Not Strings》

“The world is not made of strings, but is made of things”

~ 5亿+实体 35亿条事实 (实体-属性-值 , 实体-关系-实体)

Amit Singhal



AT&T – Google - Uber

搜索的未来

多模式、多媒介、个性化、基于位置、社交关系、跨语言、上下文



Marissa Mayer

人机交互方式的改变



搜狗 王小川

个人观点

一切并非偶然

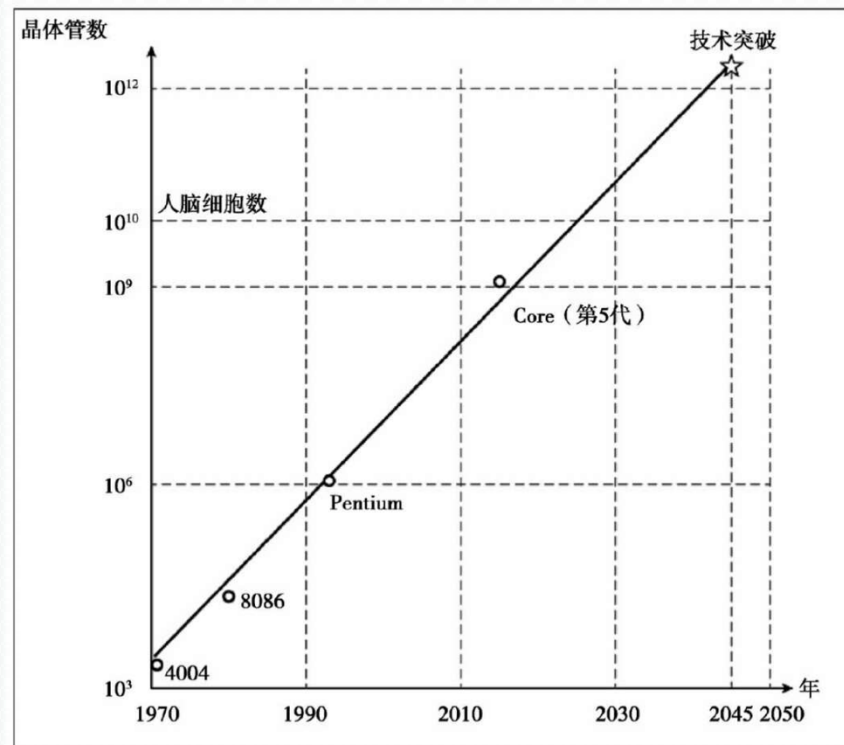
2007年，我在做基因序列数据的搜索

2017年，我在做移动互联网数据的搜索

神经网络就是算法界的“仿生学”应用！

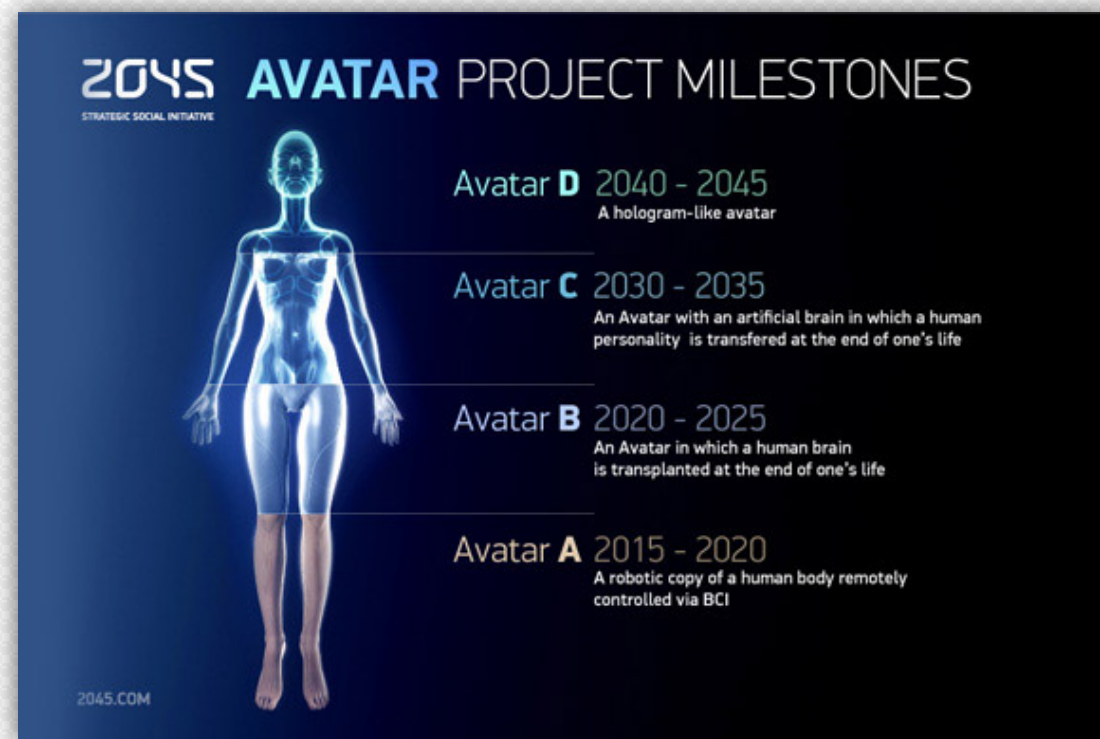
NLP领域有不少的开源项目来自计算生物学家的贡献

DNA计算



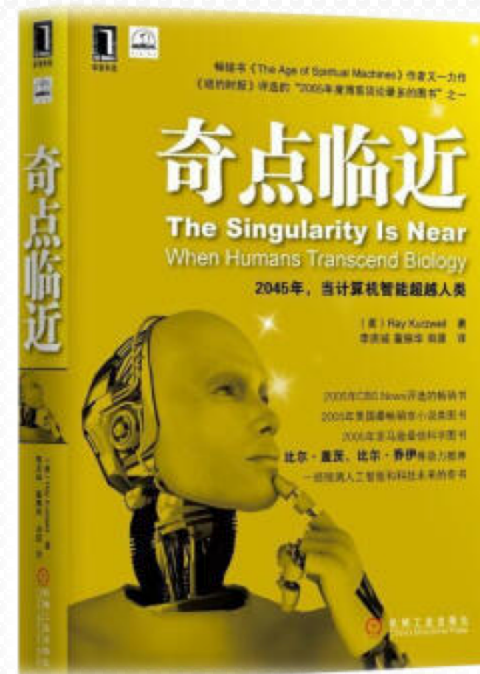
2045

<http://2045.com>



推荐书籍

- 《数学之美》第二版 吴军 著
- 《The Singularity is Near: When Humans Transcend Biology》
Ray Kurzweil 著 中文版：《奇点临近》 李庆诚 董振华 田源 译
- 《The Evolution of Everything: How New Idea Emerge》
Matt Ridley 著 中文版：《自上而下》 阎佳 译





Q & A

