Michael Zilberstein

Lost in the wealth of database products? We will help you to make sense of them

Michael Zilberstein

Data Architect, DBA

Wealth of databases

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

345 systems in ranking, April 2019

Rank Apr 2019	Rank Mar 2019	Rank Apr 2018	DBMS	Database Model	Score Apr 2019	Score Mar 2019	Score Apr 2018
1.	1.	1.	Oracle	Relational, Multi-model	1279.94	+0.80	-9.85
2.	2.	2.	MySQL	Relational, Multi-model	1215.14	+16.89	-11.26
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	1059.96	+12.11	-35.55
4.	4.	4.	PostgreSQL	Relational, Multi-model	478.72	+8.91	+83.25
5.	5.	5.	MongoDB	Document	401.98	+0.64	+60.57
6.	6.	6.	IBM Db2	Relational, Multi-model	176.05	-1.15	-12.89
7.	↑ 8.	↑ 9.	Redis	Key-value, Multi-model	146.38	+0.25	+16.27
8.	↑ 9.	8.	Elasticsearch	Search engine, Multi-model	146.00	+3.21	+14.64
9.	↓ 7.	↓ 7.	Microsoft Access	Relational	144.65	-1.55	+12.43
10.	10.	↑ 11.	SQLite	Relational	124.21	-0.66	+8.23
11.	11.	↓ 10.	Cassandra	Wide column	123.61	+0.81	+4.52
12.	12.	↑ 14.	MariaDB	Relational, Multi-model	85.23	+0.92	+20.67

Most important

- No more “one size fits all”
- Don’t drive technology to an edge. One step further and…

Considerations for choosing "the right one"

Workload type

- OLTP
- Data Warehouse
- Mixed
- Particular use-case
  - distributed cache
  - time-series data
  - graph data
  - …

Sizing

- Fixed size
- Scale Up is enough
- Scale Out required

Ecosystem & Maturity

- Ecosystem
  - Supported Languages
  - Connectors to Kafka/Spark/S3 etc
  - Monitoring options

- Maturity
  - Deployment base
  - Developing company

Location

- On Premise
- Cloud only (aka Vendor Locking)
- Mixed (can run on both)

License type

- Freeware
- Freeware with limitations (GPL V1-3, Apache V1,2)
- Free limited community edition vs Full-featured Enterprise
- Appliance
- SaaS / PaaS
- “There ain't no such thing as a free lunch”

Pricing model

- Per core
- Per server
- Per TB of loaded data
- Per GB of cache
- Per hardware type per hour (cloud)

Dev and Migration cost

- Organization DNA
- Cost = TCO (Total Cost of Ownership)
  - License and Hardware cost
  - Development
  - Migration
  - Monitoring
  - Hiring new developers vs educating company employees

Combination of technologies

- Read about common architectures
- Read about Kafka and Spark

General-purpose databases (aka “Mixed Workload”)

Most popular/recommended

- Microsoft SQL Server
- Oracle
- PostgreSQL
- MySQL

Let’s drill down

- MySQL – free (commercial forks exist; most recommended: Percona)
- PostgreSQL – free (commercial forks exist)
- MSSQL – not free
- Oracle – very-very not free

Let’s drill down

- MySQL
  - doesn’t support parallelism in a single session.
  - supports only Nested Loops for JOIN

- MSSQL and Oracle
  - huge ecosystem
  - proven enterprise-level support

Let’s drill down

- PostgreSQL
  - open source
  - ecosystem grows very fast
  - has many extensions: TimescaleDB for timeseries data, PipelineDB for timeseries and stream etc.
  - different from others in Updates treatment – beware and do VACUUM!
  - other strange behaviors; for example: CTE materialization

OLTP

Most popular products

- Relational: MSSQL, MySQL, Oracle, PostgreSQL
- NoSQL: Hbase, Cassandra, MongoDB, Couchbase
- NewSQL (relational, scale out): MemSQL, NuoDB, Google Spanner

*** Check your requirements before making a choice

Reporting

Most popular/recommended columnar dbs

- Vertica
- Redshift
- MemSQL
- Google BigQuery
- Azure Datawarehouse

Caveats

- All columnar databases have same inherent limitation: CONCURRENCY
- Bad at OLTP-type operations

Special Case: Distributed Cache

Typical use-cases

- Dashboards
- Real-time analytics

… anything that requires submillisecond latency

Most popular/recommended

- Redis
- Memcached
- Couchbase

Special Case: Extreme Write Scale

When you need 1M Inserts / second…

Typical use-cases:

- IoT (sensors data, telemetry)
- Dev-Ops monitoring
- User-activity (website serving events, ad views etc.)

Most popular/recommended

- Cassandra
  - Apache - Free
  - DataStax - Enterprise
- ScyllaDB
- HBase

Special Case: Document Store

Most popular technologies

- MongoDB
- Couchbase

Features

- Both have Community and Enterprise editions
- Secondary indexes supported
- Couchbase for Mobile
- Scale!

Special Cases: Graph

Nodes, edges, properties…

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

☐ include secondary database models — 31 systems in ranking, April 2019

Rank Apr 2019	Rank Mar 2019	Rank Apr 2018	DBMS	Database Model	Score Apr 2019	Score Mar 2019	Score Apr 2018
1.	1.	1.	Neo4j	Graph	49.49	+0.91	+8.59
2.	2.	2.	Microsoft Azure Cosmos DB	Multi-model	26.28	+1.45	+9.09
3.	3.	3.	OrientDB	Multi-model	6.19	+0.06	+0.55
4.	4.	4.	ArangoDB	Multi-model	4.29	+0.03	+0.49
5.	5.	5.	Virtuoso	Multi-model	3.31	+0.12	+1.51
6.	↑ 8.	↑ 7.	Amazon Neptune	Multi-model	1.39	+0.36	+0.70
7.	↓ 6.	↑ 13.	JanusGraph	Graph	1.38	+0.06	+1.09
8.	↓ 7.	↓ 6.	Giraph	Graph	1.20	+0.16	+0.16

https://en.wikipedia.org/wiki/Graph_database

In computing, a graph database (GDB[1]) is a database that uses graph structures for semantic queries with **nodes**, **edges** and **properties** to represent and store data. A key concept of the system is the graph (or edge or relationship), which directly relates data items in the store a collection of nodes of data and edges representing the relationships between the nodes.

Typical use-cases

- Social network
- Network topology

*** Beware of scaling limitation: scales for reads, not for writes

Most popular technologies

- Neo4J

Special Cases: Timeseries

Nodes, edges, properties…

https://db-engines.com/en/ranking/time+series+dbms

☐ include secondary database models — 30 systems in ranking, April 2019

Rank Apr 2019	Rank Mar 2019	Rank Apr 2018	DBMS	Database Model	Score Apr 2019	Score Mar 2019	Score Apr 2018
1.	1.	1.	InfluxDB	Time Series	17.22	+1.04	+6.46
2.	2.	2.	Kdb+	Time Series, Multi-model	5.85	+0.25	+2.77
3.	3.	↑ 4.	Graphite	Time Series	3.12	+0.05	+0.93
4.	↑ 5.	↑ 7.	Prometheus	Time Series	2.91	+0.20	+1.86
5.	↓ 4.	↓ 3.	RRDtool	Time Series	2.70	-0.05	-0.05
6.	6.	↓ 5.	OpenTSDB	Time Series	2.37	+0.09	+0.67
7.	7.	↓ 6.	Druid	Multi-model	1.65	+0.07	+0.59
8.	8.	↑ 19.	TimescaleDB	Time Series, Multi-model	0.95	+0.04	+0.92
9.	9.	↓ 8.	KairosDB	Time Series	0.64	-0.02	+0.20
10.	↑ 11.	↓ 9.	eXtremeDB	Multi-model	0.40	+0.00	+0.08
11.	↓ 10.	11.	FaunaDB	Multi-model	0.37	-0.15	+0.26
12.	↑ 13.		Amazon Timestream	Time Series	0.33	+0.06	

https://en.wikipedia.org/wiki/Time_series_database

A time series database (TSDB) is a software system that is optimized for handling time series data, arrays of numbers indexed by time (a datetime or a datetime range).

Typical use-cases

- IoT
- Stock market
- Personalization

Most popular/recommended

- InfluxDB
- TimescaleDB
- Couchbase*
- Cassandra*

Products worth checking

TimescaleDB

- Time-series database
- Powered by PostgreSQL
- Community and Enterprise edition available

MemSQL

- Handles OLTP and Reporting workloads simultaneously
- Scale Out
- Tables are in-memory (for OLTP) or columnar (for Reports)
- Pricing: per GB of cache

ScyllaDB

- “Cassandra on steroids”
- Developed in Israel

ClickHouse

- Columnar database opensourced by Yandex
- Backbone of Yandex Metrica – web analytics engine second only to Google Analytics
- Has first production deployment in Israel (Apps Flyer)
- Developers actively communicate and answer questions in Telegram channel

No more “one size fits all”. Choose your database wisely!

Q & A

Please fill our surveys

- Session survey:

https://www.sqlsaturday.com/823/Sessions/SessionEvaluation.aspx

- Event survey:

https://www.sqlsaturday.com/823/EventEval.aspx

Our generous sponsors

Global SQLSaturday Partner

Microsoft Azure

PASS

Gold Sponsor

EXPERDA

Quest

elastic

BARRY KATZ
Communicating Success

Couchbase

actifio

Silver Sponsor

DATASITE

VALINOR
YOUR DATA. OUR EXPERTS.

Bronze Sponsor

inspire
TECHNOLOGIES