Gen™

Introduction to MongoDB®

CTO Cloud Meetup @ Gen

2024-05-30

Petr Zemek

norton Avast LifeLock by norton Avira AVG Reputation Defender by norton CCleaner

Outline

All the commands, queries, etc. I will use in the live demos are available here

- What is MongoDB?
- Why am I having this talk?
- Getting to know MongoDB: basics, selected features, infrastructure, caveats
- Why/when (not) to consider using MongoDB?
- MongoDB Atlas: Managed MongoDB in the cloud
- Summary and discussion

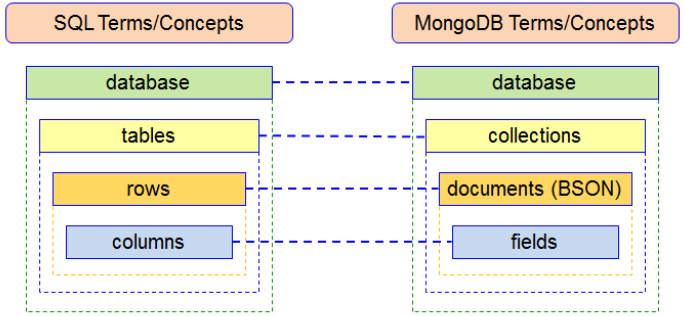
What is MongoDB?

- Source-available, cross-platform, **document-oriented NoSQL database**
- One of the most **popular databases** (Stackoverflow survey 2023, DB engines ranking 2024)
- Developed by **MongoDB Inc.**, initially released in 2009
- **Available drivers** for major programming languages (18+)
- Three hosting options
  - **Free-to-use, self-managed** version via **MongoDB Community**
  - **Paid, self-managed** version via **MongoDB Enterprise**
  - **Paid, database-as-a-service** solution via **MongoDB Atlas**
- Data stored in the **BSON** format (Binary JSON), compressed, usually **denormalized**
- **Custom query language** called MQL (but it is more an API)
- The default MongoDB client **mongosh** is based on the JavaScript driver
- Part of the **MEAN** stack (MongoDB, Express.js, Angular, and Node.js)

Why am I having this talk?

- MongoDB has been **historically** used by our team
- **Currently** used by several teams at the company
- I have **experience** with using and managing MongoDB clusters (1/3/5 nodes, up to 12 TB of data)
- To **share knowledge** – you never know when the knowledge might be useful ;-)

Basics



(ref)

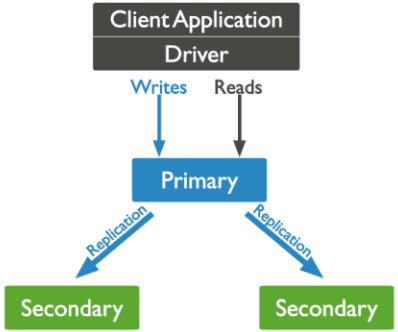
- **BSON**, differences from JSON, MongoDB's extended JSON format
- **Schema-on-read** ("schema-less") vs **schema-on-write**
- **Databases**, **collections**, **documents** vs SQL terms/concepts
- **Creating**, **updating**, **deleting**, and **querying** documents *(live demo)*
- The **_id** field and the **ObjectID** type *(live demo)*
- Document **indexing**, query **analyzer** *(live demo)*
- **Atomicity** of single-document updates *(live demo)*
- Multi-document **transactions** (ACID)
- **Read isolation** (read concern), **write acknowledgements** (write concern)

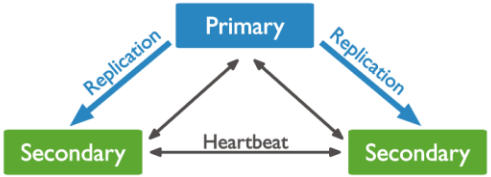
Selected features

- **Aggregation** operations *(live demo)*
- Special **index types** (e.g. TTL or partial indexes)
- **Capped collections**
- **Joining** of multiple documents via **$lookup**
- **Bulk write** operations
- **Server-side JavaScript** execution
- **Security** (auth, RBAC, in-flight and at-rest encryption, auditing)
- Database/query **profiler**

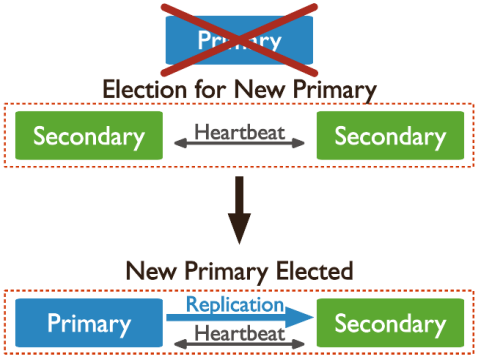
Infrastructure (part 1/2)

- Three deployment options
  - **Single server** (dev/test)
  - **Replica set** with a primary and secondaries (HA)
  - **Sharded cluster**
- Asynchronous **replication**, **heartbeats**, operation log (**oplog**)
- **Automatic failover**, election, voting



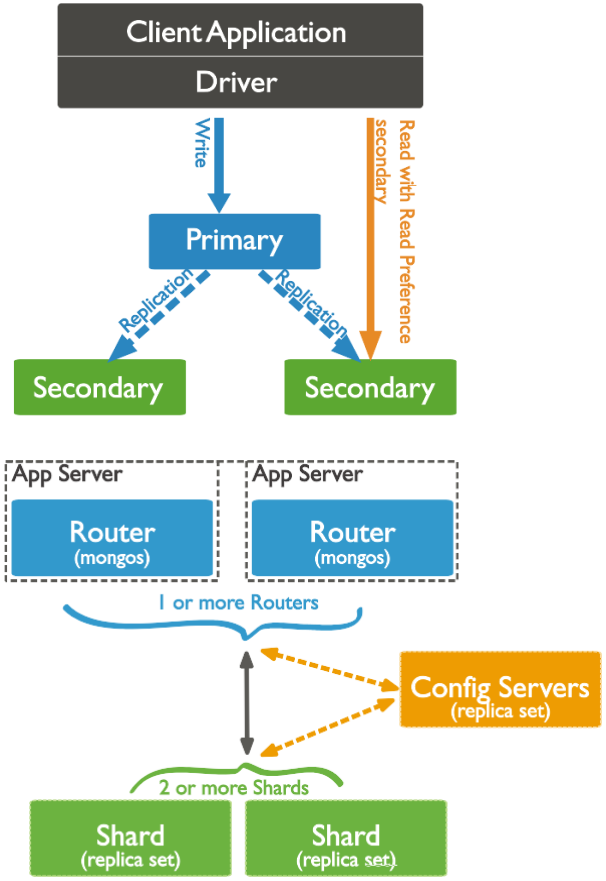


(ref)



Infrastructure (part 2/2)

- **Scaling reads** via secondaries vs **scaling reads+writes** via sharding
- MongoDB supports **mirrored reads** for pre-warming secondaries
- **Write acknowledgements** (write concern)
- **Read isolation** (read concern)



(ref)

Caveats

- Max. **16 MB** per document (hardcoded)
- Data deletion **does not reclaim space back to the OS** (but is usable by MongoDB)
- **Many reported (and fixed) bugs** throughout the history (ref)
- Some people **hate** MongoDB (e.g. non-SQL data model, issues with older versions, lack of understanding)
- Some **controversies** around MongoDB / MongoDB Inc. (e.g. relicensing in 2018)

Battle stories

- Cluster crashed due to shared SAN storage between nodes
  - **Take away**: Anything can fail, ensure that nodes are as independent as possible (duh!)
- Cluster crashed due to a failed transition from one node to a three-node replica set
  - **Take away**: A bootstrapping secondary is still counted as an active member
- Cluster got overloaded when forgetting an index for a frequent query
  - **Take away**: Monitor your cluster, ensure that you have indexes in place that support your queries
- Replication was unable to keep up when the primary got overloaded due to large writes
  - **Take away**: Monitor your cluster, ensure that it is able to deal with the load
- Short oplog window due to no-ops (MongoDB tickets: SERVER-45442, SERVER-45653)
  - **Take away**: Monitor your cluster, examine the oplog via avast/mongodb-oplog-stats (created by me in 2020)

However, overall, my experience with MongoDB is positive ;-)

Why/when to consider using MongoDB?

- When your data fit into the **document model** better than into the relational (SQL) model
- When you can benefit from a **loose schema** (schema-on-read)
- When your single-primary/replica(s) database is unable to handle the workload (**built-in horizontal scaling**)
- When you need to **self-manage** a database (my viewpoint: it is easier than e.g. PostgreSQL)
- When **prototyping** (MongoDB is easy to set up and use)
- When you need an **enterprise platform** and/or **cross-cloud support** (MongoDB Atlas)

Why/when NOT to consider using MongoDB?

- **SQL** databases/concepts/skills are **more common** (design, support, tooling, maintainability, ...)
- When the **relational (SQL) data model** is more suitable for your data
- When you **do not need** "infinite" **scaling**
- When you cannot reasonably fit your data into **16 MB documents**
- When you **do not trust** MongoDB or MongoDB Inc.
- MongoDB is **not fully open-source** (only source-available via the SPPL license)
- **PostgreSQL** is more extensible, configurable, battle-tested
- **My rule of thumb**: If you are unsure, just use PostgreSQL

MongoDB Atlas

- A **fully managed** MongoDB in the **cloud** (GCP/AWS/Azure)
- Provides some **additional features** (a suite of integrated data services)
- How to **get** it?
- How to **connect** to it?
- An overview of the **web management console** *(live demo)* - clusters, monitoring, alerting
- **Billing** (support plan, database cluster, data transfer, PSC, backup)

Available documentation

- (Online) Official MongoDB documentation
- (Online) Official MongoDB Atlas documentation
- (Book) MongoDB: The Definitive Guide (3rd Edition, 2019) - MongoDB 4.2
- (Book) Mastering MongoDB (4th edition, 2024) - MongoDB 7.0

Summary and discussion

- MongoDB is a source-available, **document-oriented NoSQL database**
- **Both free as well as paid hosting options**, including a fully managed solution in the cloud via **MongoDB Atlas**
- Data stored in the **BSON** format (Binary JSON), compressed, usually **denormalized**
- **Schema-on-read** ("schema-less") vs **schema-on-write** via schema validation
- **Custom query language** (but it is more a set of APIs than a language)
- MongoDB structures data into **documents** stored inside **collections** located in **databases**
- MongoDB supports **transactions**, **indexing**, **query analysis**, **joining** of documents, and many other features
- Three deployment options: **standalone**, **replica set** (HA), **sharding**
- Asynchronous **replication**, **automatic failover**, **oplog**, write acknowledgements, read isolation
- **Scaling reads** via reading from secondaries vs **scaling writes+reads** via sharding
- There are some **caveats**, e.g. max. 16 MB per document, and some **hate/controversies**
- MongoDB is just a *tool* - there are both cases **when to use it** and **when not to use it**