Week 11: NoSQL and Cloud Databases

LSE MY472: Data for Data Scientists
https://lse-my472.github.io/

Autumn Term 2024

Ryan Hübert

Slides last updated: 30 September 2024

Outline

- Cloud solutions for databases
- SQL vs. noSQL
- Coding
  - Online database example with SQL: BigQuery
  - NoSQL example: MongoDB

Cloud solutions

Why remote solutions?

- Last week we learned about relational databases
- Worked with SQL to manipulate data stored within tables
- In our applications, the data were **local**
- At scale, we invariably want to store data **remotely**
- Trade-offs, as always!

Some exemplary services

Database Type	AWS	GCP	Azure
Managed RDS	Amazon RDS	Cloud SQL	Azure SQL
Data Warehousing	Redshift	BigQuery	Snowflake
NoSQL (simple key-value)	DynamoDB	BigTable	Azure Tables
NoSQL (document)	DocumentDB	MongoDB on GC	Cosmos DB

Google Cloud Platform: BigQuery

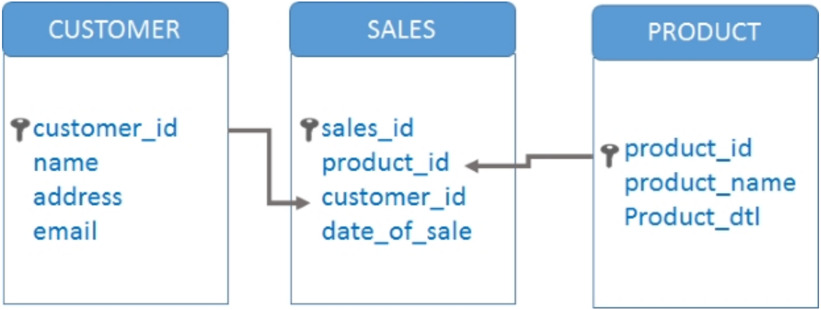
- To create and query online databases, we will look at Google BigQuery's sandbox version as an example
- Database warehouse with other features, used by many financial and commercial companies
- Queried via SQL syntax (API access allows integration with R or Python)
- Scalable to very large databases
- Good documentation
- Many similar databases exist from other providers

SQL vs noSQL

SQL

- Relational databases have a strict structure

A simple e-commerce example:

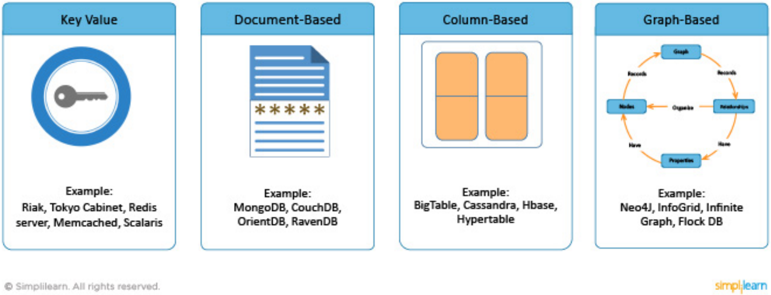


noSQL

- Originally referring to “non SQL”, “non relational” or “not only SQL”
- Provides a mechanism for storage and retrieval of data which is modeled in means other than the tabular relations used in relational databases
- No strict structure/schema
- noSQL databases are good for data with
  - High **velocity** – Lots of data coming in very quickly
  - High **variety** – Data can be structured, semi-structured, and unstructured
  - High **volume** – Total size of data
  - High **complexity** – Stored in many locations

noSQL types

Some examples from recent years:



Key Value

Example:
Riak, Tokyo Cabinet, Redis server, Memcached, Scalaris

Document-Based

Example:
MongoDB, CouchDB, OrientDB, RavenDB

Column-Based

Example:
BigTable, Cassandra, Hbase, Hypertable

Graph-Based

Example:
Neo4J, InfoGrid, Infinite Graph, Flock DB

© Simplilearn. All rights reserved.

From: Simplelern

noSQL: Pros and Cons

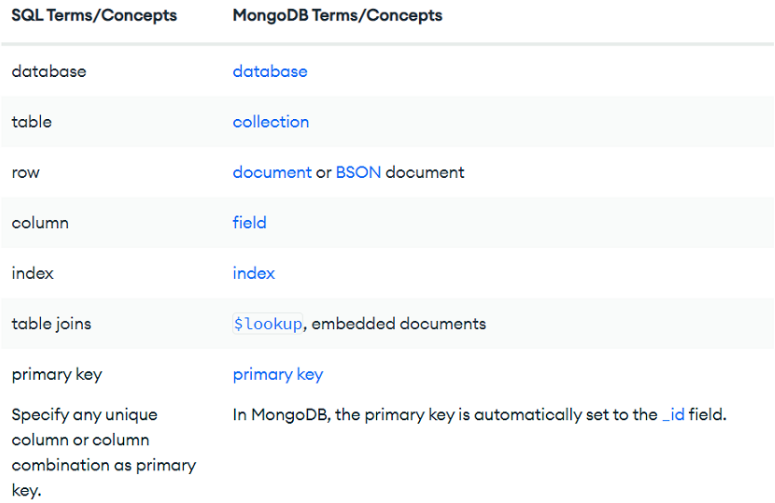
PROS	CONS
Massive scalability	Limited query capabilities
High availability	Not standardized
Schema flexibility	Not matured
Sparse and semistructured data	Developer heavy

MongoDB

- **Document-based** database
- Mapping of concepts
- Each document is constructed as a **BSON** (Binary JSON)
- Not UTF-8 string encoded document
- Like JSON, but binary - machine readable only (very lightweight)
- Can store more data types: Dates, separate kinds of numerics (int, float, etc.)

Reference:
https://docs.mongodb.com/manual/reference/sql-comparison/

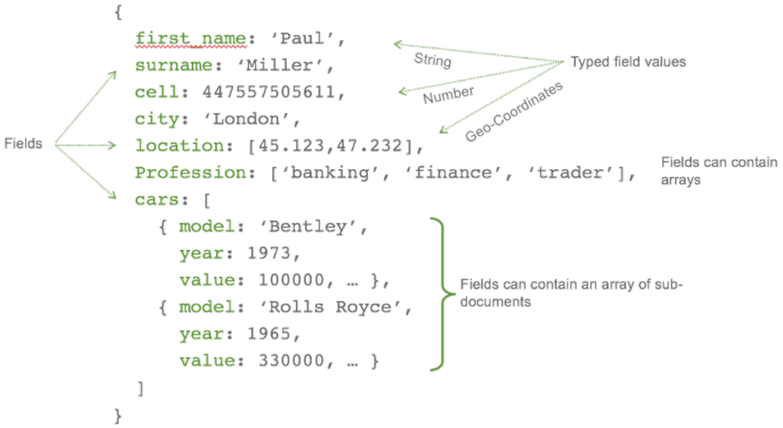
MongoDB vs. SQL



SQL Terms/Concepts	MongoDB Terms/Concepts
database	database
table	collection
row	document or BSON document
column	field
index	index
table joins	`$lookup`, embedded documents
primary key	primary key
Specify any unique column or column combination as primary key.	In MongoDB, the primary key is automatically set to the _id field.

MongoDB documents

A document looks like this:



```
{
  first_name: ‘Paul’,
  surname: ‘Miller’,
  cell: 447557505611,
  city: ‘London’,
  location: [45.123,47.232],
  Profession: [‘banking’, ‘finance’, ‘trader’],
  cars: [
    { model: ‘Bentley’,
      year: 1973,
      value: 100000, … },
    { model: ‘Rolls Royce’,
      year: 1965,
      value: 330000, … }
  ]
}
```

Typed field values
String
Number
Geo-Coordinates
Fields

Fields can contain arrays

Fields can contain an array of sub-documents

From: datawow.io

MongoDB in R (optional)

- We will look at MongoDB as an example of a popular noSQL database this week
- We thereby try to replicate basic queries from last week using MongoDB via R with the package `mongolite`
- For a simple selection of documents (i.e. rows in SQL), we will use its `find()` method
- For a bit more sophisticated queries, we will use the `aggregate()` method
- Search queries are in JSON like notation
- Detailed documentation of MongoDB commands and operators
- Resource 1 (pdf) and resource 2 (website) for the R package `mongolite`

Coding

Coding

- ➜ 01-bigquery-create-own-database.Rmd
- ➜ 02-bigquery-examples.Rmd
- ➜ 03-mongodb-demo.Rmd