DATA ENGINEER

Learning Progress Review 6th Week

We are Gold D. S.

- Rinaldy Widyantoro
- Rian Pauzi
- Hanief Fatchudin

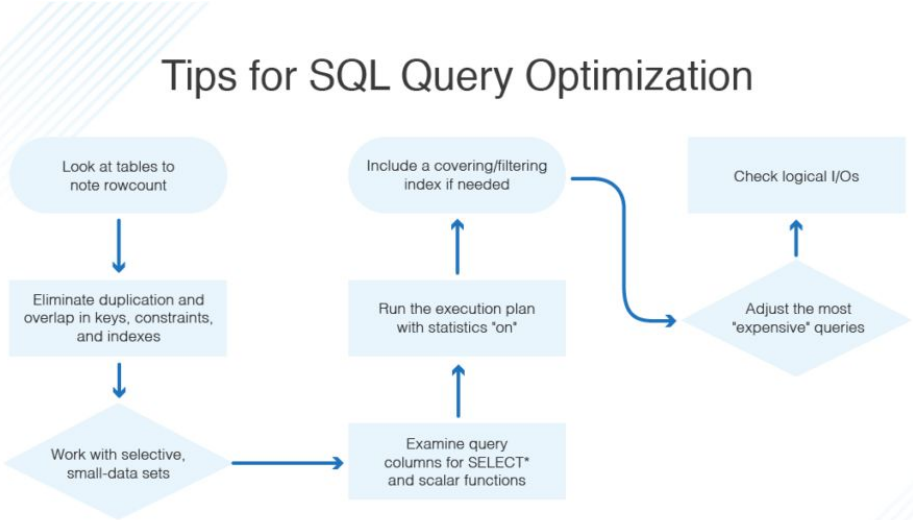
ADVANCED SQL: IMPROVING SQL PERFORMANCE

HOW TO MAKE READABLE QUERY?

1. QUERY WRITING SUGGESTION
    a. Consistency
    b. Uppercase for SQL Syntax
    c. Make new line for every (SELECT, FROM, WHERE)
    d. Add indentation for everty (Sub Query, ON, AN/OR)
    e. Add semicolon to end the statement
    f. Make an alias for long query
    g. Adding comment if necessary

HOW TO DECIDE WHICH SYNTAX TO BE IMPROVED?

1. Syntax Order
2. Processing Order
3. Logical Processing Order



HOW TO FILTER DATA

1. WHERE
    a. We using **WHERE** we want to filter in and/or filter out the result **FROM** statement
    b. WHERE statement will be checked on every row
    c. WHERE statement will be processed first before SELECT statement
    d. There will be no difference on the small set of data but will be matters on huge data
2. HAVING
    a. HAVING statement working similarly like WHERE statement except HAVING will be FILTER OUT data that already being grouped.
    b. Only can be used to filter Numeric type of Data

HOW TO EXTRACT DATA EFFICIENTLY?

1. Process after SELECT statement
    a. Need a huge resource to run after this (expensive)
    b. Only use with caution
2. USING SELECT
    a. Only pick the column that you really need
    b. Forget it if not necessary
3. USING LIMIT
    a. Limit will be limiting the result of the statement
    b. Will be executed on the very last
4. USING ORDER BY
    a. Will ordering the result (can be ascending or descending)
    b. Dont use it if not necessary because take lots of resource

HOW TO MANAGE DUPLICATION?

Duplications happened due to bad design of data or wrong query statement to get the data

To avoid that you can use:

1. DISTINCT
    a. Can be used to remove duplicate or make unique result
    b. Use with caution due to large resource can be taken
2. UNION OR UNION ALL
    a. Merging two set different tables with the same column name
    b. UNION will merge data and will remove duplicate data eep all
    c. UNION ALL will merge data as it is

HOW TO SPEED UP SEARCHING?

Using index: index is a structure that can be used to speed up the query result. it works is like a postman sending mail to the receiver, the postman already know the address, so they doesn't need to knock every doors one by one. Just knock the recipient doors

Type of INDEX:

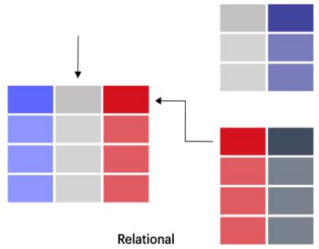
1. CLUSTERED
    a. Like a dictionary, every rows already ordered by the value
2. NON CLUSTERED
    a. Like a glossary, have a layered page that mapped into unordered row
    b. Process insert and update can be faster than clustered

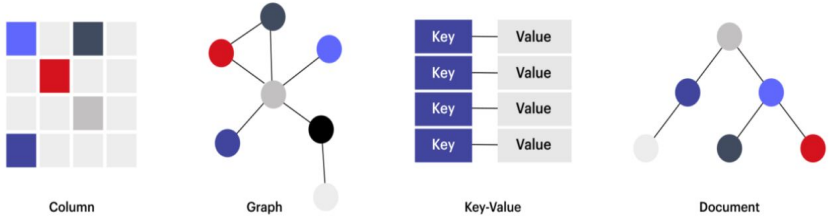
No SQL (MongoDB)

WHAT IS MongoDB ?

MongoDB is an open source NoSQL database management program. NoSQL is used as an alternative to traditional relational databases. NoSQL databases are quite useful for working with large sets of distributed data. MongoDB is a tool that can manage document-oriented information, store or retrieve information.

SQL DATABASE VS NoSQL DATABASE





Relational Data Model

Pros

- Easy to use and setup
- Universal, compatible with many tools
- Good at high-performance workloads
- Good at structure data

Cons

- Time consuming to understand and design the structure of the database
- Can be difficult to scale

Document Data Model

Pros

- No Investment to design model
- Rapid development cycles
- In general faster than SQL
- Runs well on the cloud

Cons

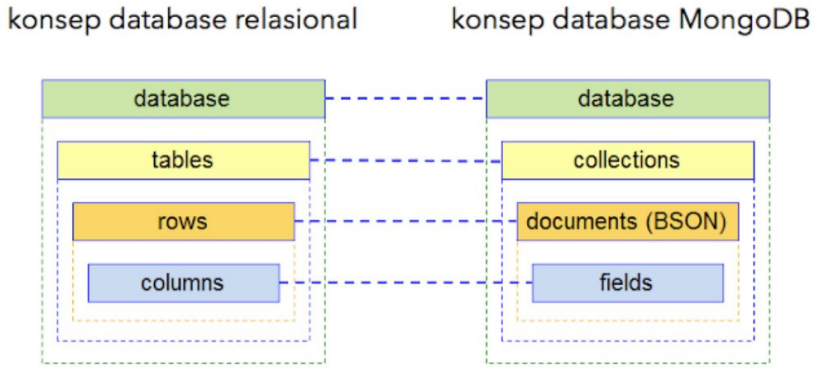
- Unsuited for interconnected data
- Technology still maturing
- Can have slower respone time

NoSQL MongoDB

- MongoDB stores data in flexible, JSON-like documents, meaning fields can vary from document to document and data structure can be changed over time
- The document model maps to the objects in your application code, making data easy to work with
- Ad hoc queries, indexing, and real time aggregation provide powerful ways to access and analyze your data
- MongoDB is a distributed database at its core, so high availability, horizontal scaling, and geographic distribution are built in and easy to use

SQL VS MongoDB

Database: is the element at the top level. In a relational database, a database generally consists of tables and views. In MongoDB, a database is a physical container that contains a structure called a collection. Each database has its own set of files in the storage media filesystem. Generally, a single MongoDB server consists of a number of databases.



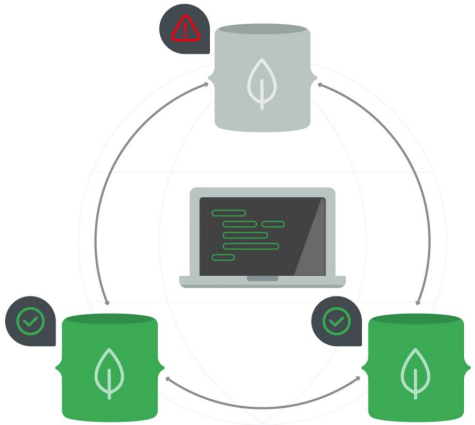
Collection: is a group of MongoDB documents. Collections can be matched with tables in an RDBMS. Multiple collections can exist in the same database but must have different names.

Documents: is the smallest unit of data in MongoDB. Basically composed of a group of key-value pairs. Unlike records in an RDBMS, documents have a dynamic schema, meaning that documents that are in a collection do not have to have the same set of fields.

MongoDB Atlas

MongoDB Atlas is a fully-managed cloud database developed by the same people that build MongoDB. Atlas handles all the complexity of deploying, managing, and healing your deployments on the cloud service provider of your choice (AWS, Azure, and GCP).

A **replica set** in MongoDB is a group of mongod processes that maintain the same data set connected from multiple machines. Replica sets provide redundancy and high availability, and are the basis for all production deployments. replica sets can ensure that if something happens to one of the machines, the data will remain intact.



Export and Import MongoDB

Export is the process of outputting data or copying data from a file stored in MongoDB collection to our computer, usually in the form of a .json file.

Import is the process of entering data or copying data from files stored on our computer into a MongoDB collection, usually in the form of a .json file.

Export data .JSON

```
mongoexport --uri="<Atlas Cluster URI>"
            --collection=<collection name>
            --out=<filename>.json
```

Import data .JSON

```
mongoimport --uri="<Atlas Cluster URI>"
            --drop=<filename>.json
```

Collection Method

NAME	DESCRIPTION
db.collection.deleteOne()	Deletes a single document in a collection.
db.collection.deleteMany()	Deletes multiple documents in a collection.
db.collection.drop()	Removes the specified collection from the database.
db.collection.find()	Performs a query on a collection or a view and returns a cursor object.
db.collection.findOne()	Performs a query and returns a single document.
db.collection.updateOne()	Modifies a single document in a collection.
db.collection.updateMany()	Modifies multiple documents in a collection.

ETL FROM SCRATCH WITH PYTHON (SCRAPING)

EXTRACT

- WE CAN CATEGORIZE ‘WEB SCRAPING’ AS ‘E’ (EXTRACT) PROCESS IN ETL
- TECHNIC IN WEB SCRAPING:
  - COPY-PASTE
  - PATTERN MATCHING
  - HTML PARSING
  - ETC.
- IN WEB SCRAPING, WE CAN USE ‘PANDAS’ FOR SPECIFIC READ_TABLE:
  - PANDAS.READ_HTML(“SOME_WEBSITES_URL”)

TRANSFORM

- WE CAN CATEGORIZE ‘CLEANING DATA’ AS ‘T’ (TRANSFORM) PROCESS IN ETL
- CLEANING DATA TO TRANSFORMATION DATA FROM ONE FORM TO ANOTHER
- FOR EXAMPLE TRANSFORM VALUE COLUMN FROM STRING TO NUMERIC AND ADD COLUMN YEAR IN TABLE:

Before:

No.	Nama	Kekayaan bersih (USD)	Usia	Kebangsaan	Sumber kekayaan
NaN	Jeff Bezos	$112.0 miliar	54	Amerika Serikat	Amazon.com
NaN	Bill Gates	$90.0 miliar	62	Amerika Serikat	Microsoft
NaN	Warren Buffett	$84.0 miliar	87	Amerika Serikat	Berkshire Hathaway

After Transform:

nomor_urut	tahun	nama	kekayaan_bersih_usd_juta	usia	kebangsaan	sumber_kekayaan
1	2018	Jeff Bezos	112000.0	54	Amerika Serikat	Amazon.com
2	2018	Bill Gates	90000.0	62	Amerika Serikat	Microsoft
3	2018	Warren Buffett	84000.0	87	Amerika Serikat	Berkshire Hathaway

LOAD

- WE CAN CATEGORIZE ‘STORING DATA TO DATABASE AS ‘L’ (LOAD) PROCESS IN ETL
- PROCESS LOAD NOT ONLY STORING TO DATABASE, BUT ALSO TO FILE LIKE CSV, PARQUET, ORC, ETC.
- METHOD TO STORING DATABASE USUALLY USING 2 TOOLS:
  - PANDAS
  - SQLALCHEMYS