Data Engineering Day 17

The credit for this course goes to Coursera. Click More

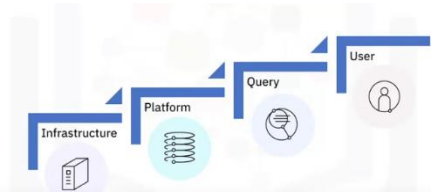
Another link : Azure data Engineer

Monitoring and Optimizations of Databases

 # Overview of Database monitoring.

- The core reason for monitoring the database is to maintain its health so that it performs its tasks very smoothly without any obstacles.

 # different monitoring levels

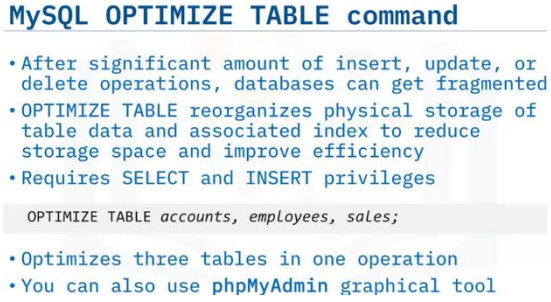


1 Infrastructure: the infrastructures such as operating systems, servers, Networks, and hardware should be well maintained for regular performance.
2 Platform: managing databases like MySQL, Db2, or any other related databases and writing optimized codes for robustness performances.
3 Query: solving latency, handling errors, can be a best practice for maintaining the health of a query.
4 User: misleading of monitoring, can be one of the causes of low performance of database. Therefore, a Database Admin should monitor it time and again to smooth its performance.

 # Optimizing Database.

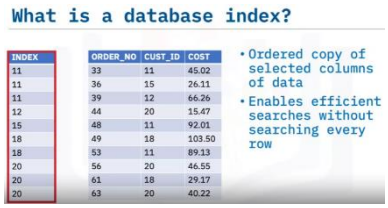
1. Improved Performance: Optimizing databases increases the speed and efficiency of data retrieval and processing, leading to faster response times which is good for users when they access it.
2. Resource Efficiency: It helps in using hardware and software resources like CPU, memory, and storage more effectively, preventing unnecessary expenditure.
3. Scalability: Optimization ensures that a database can handle increasing volumes of data and numbers of users without degradation in performance.
4. Cost Reduction: Efficient databases can defer or eliminate the need for costly hardware upgrades or additional storage.
5. Data Integrity: Ensures consistent and reliable data across different users and applications, even under heavy loads.
6. Lower Latency: Optimization reduces delays in data processing and access, critical for real-time applications.
7. Enhanced Concurrency: Allows more users to access and modify the database simultaneously without significant performance drops.
8. Maintenance and Upkeep: Regular optimization prevents gradual performance degradation and maintains overall system health.

The figure on the right shows the reasons and the basic commands used for optimizing the MySQL databases.



 # Optimizing Database.

- Indexing helps to get or retrieve the data faster in the database saving the user time.
- primary is the special type of indexing and is always unique, non-nullable and is one per table which gives a unique identification.

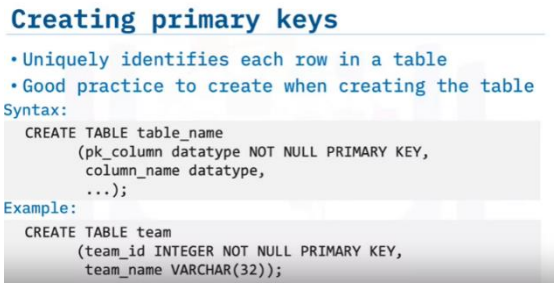


The figure on the right-hand side and the figure below try to explain columns.

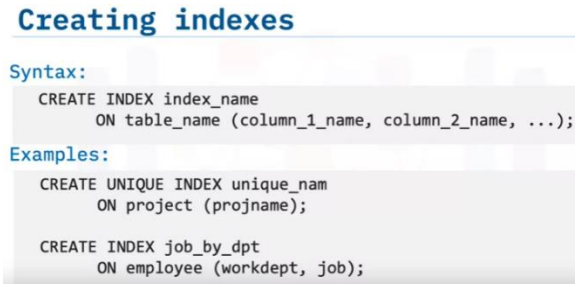
Primary key with multiple columns

Syntax:

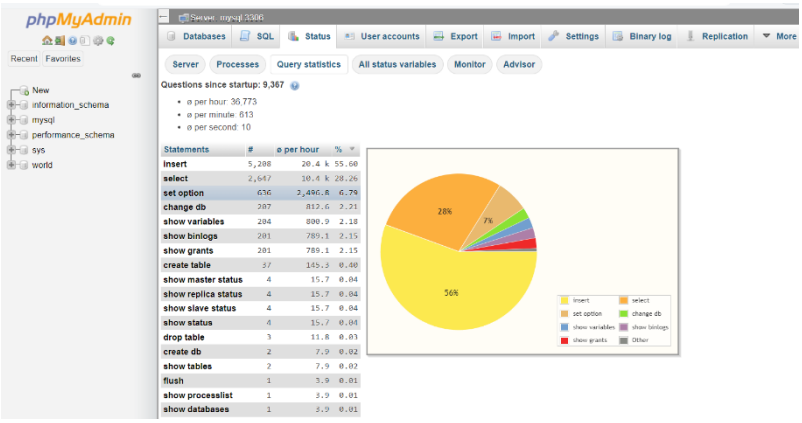
```
CREATE TABLE table_name
      (column_1_name datatype NOT NULL,
       column_2_name datatype NOT NULL,
       ...
       PRIMARY KEY(column_1_name, column_2_name));
```



creating primary keys in single and multiple



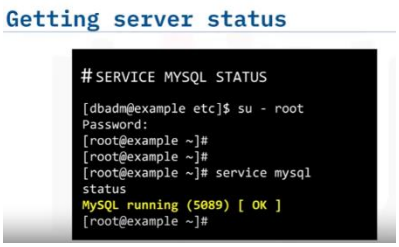
The figure on the right-hand side below tries to explain creating indexes.



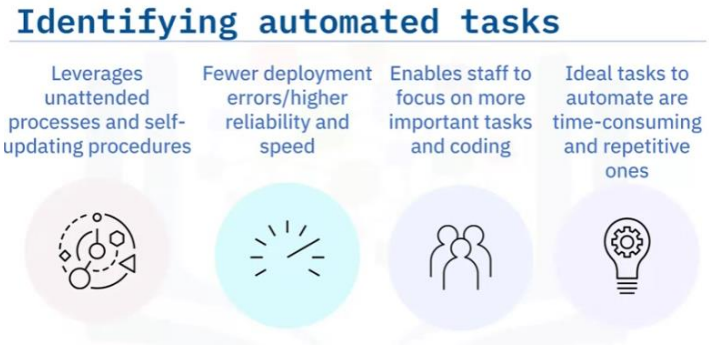
The figure below shows the status of the query in phpMyAdmin's panel.

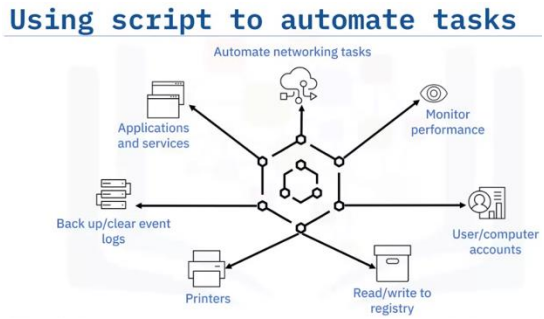
 # Overview of Database monitoring.

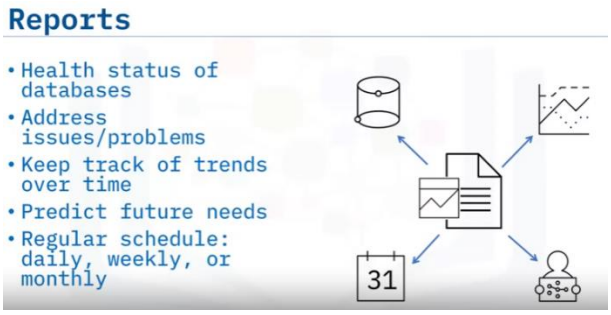
- Trouble shooting is another name for identifying and solving the problems.



 # Automations of Database tasks:







#Creating a database, backup the database using an automated script, and finally truncate and restore it back.

Question: You will create a shell script that does the following:

- Writes the database to a SQL file created with a timestamp, using the `MySQL dump` command.
- Zips the SQL file into a zip file using the `gzip` command.
- Removes the SQL file using rm command.
- Deletes the backup after 30 days.

Answer:

Step 1: create the file named sqlbackup.sh.

Step 2: open the file and write the following commands.

```
#!/bin/sh
#  the above line  tells the interpreter this code needs to run as a shell
script.

# Set the database name to a variable.

DATABASE = 'sakila'

# This will be printed on to the screen. In the case cron job, it will be
printed to the logs.
echo = "Pulling Database: This may take a few minutes"

#set the folder where the database backup will be stored
backupfolder = /home/theia/backups

# Number of the days to store the backup
keep_day = 30

sqlfile = $backupfolder/all-database-$(date +%d-%m-%Y_%H-%M-%S).sql
zipfile = $backupfolder/all-database-$(date +%d-%m-%Y_%H-%M-%S).gz


# create a backup

if mysqldump $DATABSE > $sqlfile ; then
   echo " Sql dump created"
   # Compress backup
```

```
   if gzip -c $sqlfile > $zipfile; then
      echo " The backup was successfully compressed"
    else
      echo "Error compressing backup.Backup was not created"
      exit
    fi
    rm $sqlfile
else
  echo "pg_dump return non-zero Code No backup was created.”
  exit
fi

# Delete old backups
find $backupfolder -mtime +keep_day -delete
```

Step 3: save your script and add following command in terminal. ***sudo chmod u+x+rkup.sh***

Step 4: create a new directory to save the backup file and name directory as backups.

Step 5: Now we need to create the crontab file in which we will write the code for performing the tasks automatically. To create the file run this command in terminal: ***crontab -e.***

Step 6: paste this: ****/2 * * * * /home/project/sqlbackup.sh > /home/project/backup.log***

Step 7: hit this command in the terminal: ***sudo service cron start.***

Step 8: hit this command ***ls -l /home/theia/backups.***

Step 9: hit this command ***sudo service cron stop.***

You can see that the process works smoothly and if any error occurs, do check your file directory to sort out the errors.

Reading: Improving Performance of Slow Queries in MySQL

Estimated time needed: 20 minutes

In this reading, you'll learn how to improve the performance of slow queries in MySQL.

Objectives

After completing this reading, you will be able to:

1. Describe common reasons for slow queries in MySQL
2. Identify the reason for your query's performance with the `EXPLAIN` statement
3. Improve your query's performance with indexes and other best practices

Software Used

In this reading, you will see usage of MySQL. MySQL is a Relational Database Management System (RDBMS) designed to efficiently store, manipulate, and retrieve data.

Common Causes of Slow Queries

Sometimes when you run a query, you might notice that the output appears much slower than you expect it to, taking a few extra seconds, minutes or even hours to load. Why might that be happening?

There are many reasons for a slow query, but a few common ones include:

1. The size of the database, which is composed of the number of tables and the size of each table. The larger the table, the longer a query will take, particularly if you're performing scans of the entire table each time.

2. Unoptimized queries can lead to slower performance. For example, if you haven't properly indexed your database, the results of your queries will load much slower.

Each time you run a query, you'll see output similar to the following:

```
+--------+-----------+----------------+-----------------+--------+-----------+
300024 rows in set (0.34 sec)
```

As can be seen, the output includes the number of rows outputted and how long it took to execute, given in the format of `0.00` seconds.

One built-in tool that can be used to determine why your query might be taking a longer time to run is the `EXPLAIN` statement.

EXPLAIN Your Query's Performance

The `EXPLAIN` statement provides information about how MySQL executes your statement—that is, how MySQL plans on running your query. With `EXPLAIN`, you can check if your query is pulling more information than it needs to, resulting in a slower performance due to handling large amounts of data.

This statement works with `SELECT`, `DELETE`, `INSERT`, `REPLACE` and `UPDATE`. When run, it outputs a table that looks like the following:

```
mysql> EXPLAIN SELECT * FROM employees;
+----+-------------+-----------+------------+------+---------------+---
| id | select_type | table     | partitions | type | possible_keys | ke
+----+-------------+-----------+------------+------+---------------+---
|  1 | SIMPLE      | employees | NULL       | ALL  | NULL          | NU
+----+-------------+-----------+------------+------+---------------+---
1 row in set, 1 warning (0.00 sec)
```

As shown in the outputted table, with `SELECT`, the `EXPLAIN` statement tells you what type of select you performed, the table that select is being performed on, the number of rows examined, and any additional information.

In this case, the `EXPLAIN` statement showed us that the query performed a simple select (rather than, for example, a subquery or union select) and that 298,980 rows were examined (out of a total of about 300,024 rows).

The number of rows examined can be helpful when it comes to determining why a query is slow. For example, if you notice that your output is only 13 rows, but the query is examining about 300,000 rows—almost the entire table!—then that could be a reason for your query's slow performance.

In the earlier example, loading about 300,000 rows took less than a second to process, so that may not be a big concern with this database. However, that may not be the case with larger databases that can have up to a million rows in them.

One method of making these queries faster is by adding indexes to your table.

Indexing a Column

Think of indexes like bookmarks. Indexes point to specific rows, helping the query determine which rows match its conditions and quickly retrieves those results. With this process, the query avoids searching through the entire table and improves the performance of your query, particularly when you're using SELECT and WHERE clauses.

There are many types of indexes that you can add to your databases, with popular ones being regular indexes, primary indexes, unique indexes, full-text indexes and prefix indexes.

Type of Index	Description
Regular Index	An index where values do not have to be unique and can be NULL.
Primary Index	Primary indexes are automatically created for primary keys. All column values are unique and NULL values are not allowed.
Unique Index	An index where all column values are unique. Unlike the primary index, unique indexes can contain a NULL value.
Full-Text Index	An index used for searching through large amounts of text and can only be created for **char**, **varchar** and/or **text** datatype columns.
Prefix Index	An index that uses only the first N characters of a text value, which can improve performance as only those characters would need to be searched.

Now, you might be wondering: if indexes are so great, why don't we add them to each column?

Generally, it's best practice to avoid adding indexes to all your columns, only adding them to the ones that it may be helpful for, such as a column that is frequently accessed. While indexing can improve the performance of some queries, it can also slow down your inserts, updates and deletes because each index will need to be updated every time. Therefore, it's important to find the balance between the number of indexes and the speed of your queries.

In addition, indexes are less helpful for querying small tables or large tables where almost all the rows need to be examined. In the case where most rows need to be examined, it would be faster to read all those rows rather than using an index. As such, adding an index is dependent on your needs.

Be SELECTive With Columns

When possible, avoid selecting all columns from your table. With larger datasets, selecting all columns and displaying them can take much longer than selecting the one or two columns that you need.

For example, with a dataset of about 300,000 employee entries, the following query takes about 0.31 seconds to load:

```
1. 1
```

```
1. SELECT * FROM employee;
```

Copied!

```
| 499998 | 1956-09-05 | Patricia       | Breugel          | M      | 1993-10-13 |
| 499999 | 1958-05-01 | Sachin         | Tsukuda          | M      | 1997-11-30 |
+--------+------------+----------------+------------------+--------+------------+
300024 rows in set (0.31 sec)
```

But if we only wanted to see the employee numbers and their hire dates (2 out of the 6 columns) we could easily do so with this query that takes 0.12 seconds to load:

```
1. 1
```

```
1. SELECT employee_number, hire_date FROM employee;
```

Copied!

```
| 499998 | 1993-10-13 |
| 499999 | 1997-11-30 |
+--------+------------+
300024 rows in set (0.12 sec)
```

Notice how the execution time of the query is much faster compared to the when we selected them all. This method can be helpful when dealing with large datasets that you only need select specific columns from.

Avoid Leading Wildcards

Leading wildcards, which are wildcards (`"%abc"`) that find values that end with specific characters, result in full table scans, even with indexes in place.

If your query uses a leading wildcard and performs poorly, consider using a full-text index instead. This will improve the speed of your query while avoiding the need to search through every row.

Use the UNION ALL Clause

When using the `OR` operator with `LIKE` statements, a `UNION ALL` clause can improve the speed of your query, especially if the columns on both sides of the operator are indexed.

This improvement is due to the `OR` operator sometimes scanning the entire table and overlooking indexes, whereas the `UNION ALL` operator will apply them to the separate `SELECT` statements.

Next Steps

Congratulations! Now that you have a better understanding of why your query may be performing slow and how you can improve that performance, let's take a look at how we can do that with MySQL in the Skills Network Labs environment.