horizontal line

Student Questions

**SDA Data Engineering Bootcamp (Feb 2024)**

## Week 1 - SQL

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
| When to use having clause, and why exactly it conflict with where ? and why it works sometimes while where doesnt work ?   * (Answer) WHERE is a filter used at the table level. When you select from a table, you use WHERE to filter out the data you don't need so the query can run faster and read less data. * HAVING is a filter that's applied with the GROUP BY operation. So it filters on the output of a group by <https://www.sqlshack.com/the-difference-between-where-and-having-clause-in-sql/> check this one out. | 1 | SQL | Resolved |
| Can you kindly explain how to construct subqueries? Like how in details i should expect to use something inside something inside something to create a multiple subqueies   * (Answer) Nested query is very useful but general rule of thumb is that if your query requires many levels of nesting (query inside query insider query...) then you need to consider splitting up the queries. |  | SQL | Resolved |
| i have some misunderstanding in sql workbench that every time i open for new code i should create the same database(like superstore db) again and again doesnt know how can i save it ,and how combine between two database and use them in the same code,sorry if its common but really want to be familiar with basics.   * (Answer) SQL workbench and MySQL workbench are different tools. MySQL Workbench is what we use for teaching I believe. * I should keep the data and your scripts I believe. If it keeps disappearing, you'd better reach out the teaching assistant for support. |  | SQL | Resolved |
| How can I effectively determine which SQL function is best suited for different scenarios?   * (Answer) Most SQL function achieves specific goals. If you're dealing with a text column then apply string functions. If you're dealing timestamp, then use the time functions. The experience comes with practice. Chatgpt would be helpful in this case :-) | 1 | SQL | Resolved |
| What is the most efficient way to handle those problems in databases, should it be taking a deep look at the schema and break our problem down? I know it’s a bit of stupid question sorry :P   * (Answer) Read the error logs. The error message usually tells you what went wrong.   + 1. Read the error message first.   + 2. Google the error message if it doesn't help you right away * Most of the time you will find similar questions asked by other people online. | 1 | SQL | Resolved |
| When to use right-left join, depends on what? Would you explain it with an example   * (Answer) I don't see people use RIGHT JOIN a lot in practice. If you switch the table A and B, RIGTH can often be expressed with a LEFT JOIN. * LEFT JOIN Is more commonly used for sure. | 1 | SQL | Resolved |
| What is the best way to analyze the problems if it has more things to do like subqueries, window functions , etc..., to get the results ? because sometimes I do brainstorming, and that can be overwhelming.   * (Answer) It becomes easier when you have done more SQL challenges. Generally speaking, I’d suggest   + Read the question carefully.   + First, try to check if the data should be filtered (is the question about sales in a specific region, department, product line? )   + Then, check if the question requires joining different tables in order to get the columns you need   + Then, check if you can use those tables directly for joins or is there any further filtering/aggregation required (this raises the need for doing subqueries)     - If you’re dealing with many table joins and subqueries, then it’s opportunity for CTE (with…) or you can create one temp table for each before joining them   + After that, check if you need to do aggregation either on one of the tables     - Also check if aggregation is also required on the joined output (on top of the subqueries)   + Lastly, I’d check if we need to return things like ranks, top N items, running totals, etc. ⇒ these bring up the need for window functions |  | SQL | Resolved |
| 1. How to understand the question, and breakdown to steps to make it easier to solve ?    1. (Answer) Please read the previous question. It’s related 2. While we using Group By , how to know which columns that we need to use with ? because sometimes I use some column but it didn't work or it was wrong usage    1. (Answer) The column you usually need for grouping should be categorical (strings) such as city, country, product name, ID, department, etc.    2. When you read the business questions, look for things like “Sales by region”, “monthly sales”, etc. that usually means you need to aggregate the data by region or month. 3. What is the meaning of use subquery as filter or table ? What is the difference?    1. (Answer) Can you elaborate this question? 4. How can I know which function or way to solve the problem ?    1. (Answer) First check what type of column you’re dealing. If you want to transform a date column, then use the date function. If it’s a string, then use string function.    2. The documentation of MySQL is always helpful. 5. As Mr.Shaohua said if we have many subqueries inside each other we need to splitting them up, But how we can do that while we actually have difficulties with nested subqueries??    1. (Answer) Whatever that’s nested inside a query can be split up. For example, you can create a table using the subquery and then replace the subquery with the table name. Or you can use the CTE with statement. It may not help you remove the nesting but for sure it helps you with the following:       1. Code is more readable       2. It’s easier to test each part of your query so debugging is way easier.    2. You do very nested subqueries when you get more experienced? It doesn’t necessarily give you performance gain. 6. How much time should I spend trying to solve one problem ? After that if I can't solve it I need to see the solution. Because I think I spend a lot in one question    1. (Answer) Don’t be too hard on yourself. At the beginning of the journey, feel free to check out the solution. I’d say       1. If you couldn’t understand the question very well, ask the TA.       2. If you understand the question but just don’t know how to get started, then check the solution       3. If you think you have a chance of solving it, then spend more time on it and try to avoid looking at solutions. You will improve fast when you do this. But don’t get stuck on something longer than 15-20 minutes. | 1 | SQL | Pending |
| * How can I determine the order of queries, specifically which query precedes the one mentioned in the sup query ?   + (Answer) Good questions. Here’s the sequence in general:     - SELECT FROM     - WHERE     - GROUP BY     - HAVING     - ORDER BY * Why we use primary key and foreign key , how there is working? We don't uses in exercises   + (Answer) Yes. It’s useful for application developers and database developers. Foreign key allows you to achieve “referential integrity” so you don’t delete a customer from the customer table while leaving all the transactionals related to this customers as garbage data in other tables. It creates the “references” between different tables through primary and foreign keys so the integrity can be enforced.   + It’s important for applications.   + But for data engineering, especially data warehouse, people don’t normally do that. It’s not necessary. That’s why we didn’t emphasize on it. * I have difficulty solving nested queries. Do you have any advice or reference that can help me?   + (Answer) At the beginning, whenever you can try to use multiple queries to solve the problem.     - Write one query, save the output to a table     - Write a second query and use the table generated in the previous query.   + Once you get more comfortable with SQL, you will notice yourself leaning towards subqueries. * Will the level of projects and exams in the upcoming stages of the bootcamp be similar to the level of practical exercises in the labs? I spend a lot of time solving some of the labs, even for a single query. Will I face difficulties in the upcoming stages of the bootcamp? Will I be able to succeed?   + (Answer) I think your struggle at the beginning is mainly due to the lack of experience with SQL. Afterall, you only learned it for 4 days. A data analysis or engineer write SQL on a daily basis and they also read someone else’s SQL code when they join a company.   + Don’t be too hard on yourself. We will practice SQL again later in the analytics engineering part of the program.   + The journey will be bumpy but the reward is big. |  | SQL | Resolved |
| So in a previous question I was asking the wrong question, but I can clarify the question I asked: how can I fully understand the business question and how to solve the problem by breaking it down to pieces, because when I try to read the problem or business question for the first time I get confused and may take time to realize I have to use this and that. I also have other question regarding the accountability for a data engineer to focus on the execution time for the query and how important for me in the current understanding to fully get my brain around performance.   * (Answer) Taking the time to understand business questions and breaking them down to pieces is part of the learning process. You will get more experienced as you work on more questions like that. Understanding the business itself is also important. For example, you need know what is asked, why it’s needed, etc. Of course, that sounds more like data analytics than data engineering. Data engineers usually work with BI, DS, and Business teams so a lot of requirements can be broken down for them. There will be technical documentation and data transformation discussions. * (Answer) Regarding performance, it’s more advanced topics. We will discuss performance here and there in different parts of the bootcamp. We will discuss performance of keys, spark job tunings, etc. Don’t worry about it for now. It will start to make more sense once you have learned more tools. In the first 2 weeks, we’re just practicing the fundamental programming skills to get prepared for the later content. |  | SQL | Resolved |
| In the first exercise when we load the product table from the csv file , why did we need to use the extra line ( character set 'latin1' ), is it because of the product table data or what? And what does this line means or do ?   * (Answer) I believe it’s due to the data itself. You don’t always need to use latin` It depends on the underlying data you’re dealing with. If you dont’ use latin1, you may get some warning or error message. |  | SQL | Resolved |

## Week 2 - Python

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
| In pandas1 lap , question 3  1- .replace(' ',0) # will replace ' ' with 0 ?  2- .fillna(0) # will fill NAN with 0 ? | 2 | Python | Pending |
| In pandas1 lap , question 6  how can i do this list comprehension with regular for loop and if else ?  data['loyalty'] = ['Bronze' if tenure < 12  else 'Silver' if (12<= tenure <=24)  else 'Gold'  for tenure in data['tenure']]  (Answer) If if the below works.  loyalty = []  for tenure in data['tenure']:  if tenure < 12:  loyalty.append('Bronze')  elif 12<= tenure <=24:  loyalty.append('Silver')  else:  loyalty.append('Gold')    data['loyalty] = pd.Series(loyalty) | 2 | Python | Resolved |
|  |  |  | Pending |

## Week 3 - Web Scraping + Linux

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

## Week 4 - Web Scraping + AWS

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

## Week 5

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

## Week 6

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

## Week 7

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

## Week 8

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

# 

# 

## Week 9

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

# 

## Week 10

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

## Week 11

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

## Week 12

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

# 

## Week 13

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

# 

## Week 14

| **Question** | **Week** | **Type** | **Status** |
| --- | --- | --- | --- |
|  | 1 | SQL | Pending |
|  |  |  | Pending |
|  |  |  | Pending |

# 

# 

# 

# 

# 

# 

# 