Task 2 & 3

2.2.1) Querying top 10 posts by score

In [1]:

```
from pyhive import hive
from tabulate import tabulate
import pandas as pd


host_name = "localhost"
port = 10000
user = "hdoop"
password = "a"
database="default"

def hiveconnection(host_name, port, user,password, database):
    conn = hive.Connection(host=host_name, port=port, username=user, password=password,
                           database=database, auth='CUSTOM')
    return conn


conn = hiveconnection(host_name, port, user,password, database)
cur = conn.cursor()

## Usage example from https://github.com/dropbox/PyHive
```

In [11]:

```
cur.execute('select ID, Title, Score from stackexchange_view order by score desc limit 10')
result = cur.fetchall()
print(tabulate(result, tablefmt='orgtbl'))
```

```
| 11227809 | Why is processing a sorted array faster than processing an unsorted array? | 25933 |
|   927358 | How do I undo the most recent local commits in Git?                        | 23348 |
|  2003505 | How do I delete a Git branch locally and remotely?                         | 18514 |
|   292357 | What is the difference between 'git pull' and 'git fetch'?                 | 12834 |
|   231767 | What does the "yield" keyword do?                                          | 11551 |
|   477816 | What is the correct JSON content type?                                     | 10921 |
|   348170 | How do I undo 'git add' before commit?                                     | 10079 |
|  5767325 | How can I remove a specific item from an array?                            |  9931 |
|  6591213 | How do I rename a local Git branch?                                        |  9792 |
|  1642028 | What is the "-->" operator in C/C++?                                       |  9560 |
```

2.2.2) The top 10 users by post score

In [16]:

```
### Join data as usernames data was extracted and added post data pulling

cur.execute("""
    select
           OwnerUserId,
           DisplayName,
           sum(score) as score
    from stackexchange_view
    group by OwnerUserId,DisplayName
    order by score desc
    LIMIT 10

""")
result = cur.fetchall()
print(tabulate(result, tablefmt='orgtbl'))
```

```
|  87234 | GManNickG       | 37672 |
|   4883 | readonly        | 28817 |
|   9951 | e-satis         | 26878 |
|   6068 | pupeno          | 25944 |
|  89904 | Hamza Yerlikaya | 24024 |
|  51816 | Joan Venge      | 23763 |
|  49153 | Ali             | 20203 |
| 179736 | TIMEX           | 19603 |
|  95592 | Matthew Rankin  | 19479 |
|  63051 | flybywire       | 19362 |
```

2.2.3) The number of distinct users, who used the word “cloud” in one of their posts

In [18]:

```
cur.execute("""
SELECT
     COUNT(DISTINCT owneruserid) as user_count
FROM stackexchange_view
WHERE title LIKE '% cloud %' or Body LIKE '% cloud %'
""")
result = cur.fetchall()
display(result)
```

```
[(248,)]
```