Dist Data Engineering

Assignment 02

Muhammad Usman Malik	19i-0402
Shaheer Asif	19i-2152
Musa Ishaq	19i-0731

Instructor: Dr.Adnan Tariq

https://github.com/GhostByteX/Dist-Data-Eng-Assignment-02

Database Design

- Created an AWS account
- Created a table using the Dynamo DB service
- Connected the Python script using AWS access tokens
- Inserted the data into the table
- Username is the primary index as it is unique
- Secondary index is the location of the user

Muhammad Usman Malik 19i-0402 Shaheer Asif 19i-2152 Musa Ishaq 19i-0731

Dr.Adnan Tariq

Insertion

```python
dynamodb = session.resource('dynamodb', region_name = region_name)
table = dynamodb.Table(table_name)

# Open the CSV file and read the data into a dictionary
with open('Bitcoin_tweets.csv', 'r', encoding='utf-8') as file:
    reader = csv.DictReader(file)
    for row in reader:
        # Convert the date value to a datetime object
        date_str = row['date']
        date_obj = datetime.strptime(date_str, '%Y-%m-%d %H:%M:%S')

        # Convert the user_created value to a datetime object
        created_str = row['user_created']
        created_obj = datetime.strptime(created_str, '%Y-%m-%d %H:%M:%S')

        # Convert the user_followers value to a Decimal type
        followers = Decimal(str(row['user_followers']))

        # Create a new item to add to the table
        item = {
            'username': row['user_name'],
            'location': row['user_location'],
            'description': row['user_description'],
```

```python
# Create a new item to add to the table
item = {
    'username': row['user_name'],
    'location': row['user_location'],
    'description': row['user_description'],
    'created': created_obj.isoformat(),
    'followers': str(followers),
    'friends': row['user_friends'],
    'favorites': row['user_favourites'],
    'verified': row['user_verified'].lower() == 'true',
    'date': date_obj.isoformat(),
    'text': row['text'],
    'hashtags': row['hashtags'],
    'source': row['source'],
    'is_retweet': row['is_retweet'].lower() == 'true'
}

response = table.put_item(Item=item)
```


Muhammad Usman Malik 19i-0402 Shaheer Asif 19i-2152 Musa Ishaq 19i-0731

Dr.Adnan Tariq

Queries

```python
# Get all tweets of a user
def get_tweets_by_user(username):
    response = table.query(
        KeyConditionExpression='username = :u',
        ExpressionAttributeValues={
            ':u': username
        }
    )
    return response['Items']


def get_tweets_by_location(my_location):
    response = table.query(
        IndexName='location-index',
        KeyConditionExpression='location = :l',
        ExpressionAttributeValues={
            ':l': my_location
        }
    )
    return response['Items']
```

```python
def get_top_users(k):
    response = table.scan(
        ProjectionExpression='username, followers',
        Limit=k,
        ReturnConsumedCapacity='TOTAL'
    )
    users = sorted(response['Items'], key=lambda u: u['followers'], reverse=Tr
    return users[:k]


def get_tweets_by_top_users(k):
    users = get_top_users(k)
    tweets = []
    for user in users:
        username = user['username']
        response = table.query(
            KeyConditionExpression='username = :u',
            ExpressionAttributeValues={
                ':u': username
            }
        )
        tweets.extend(response['Items'])
    return tweets

# Get top k tweets with the most matching tags
def get_top_tweets_by_tags(k, tags):
    response = table.scan(
        FilterExpression='contains(hashtags, :tags)',
        ExpressionAttributeValues={
            ':tags': tags
        },
        ProjectionExpression='hashtags, favorites',
        Limit=k,
        ReturnConsumedCapacity='TOTAL'
```

```python
# Delete all posts of user with followers less than a threshold
def delete_tweets_by_followers(threshold):
    response = table.scan(
        ProjectionExpression='username, followers',
        FilterExpression='followers < :f',
        ExpressionAttributeValues={
            ':f': threshold
        },
        ReturnConsumedCapacity='TOTAL'
    )
    for item in response['Items']:
        username = item['username']
        table.delete_item(
            Key={
                'username': username,
                'date': item['date']
            }
        )


# Test the functions
if __name__ == '__main__':
    print(get_tweets_by_user('Mike Chambers at Northey Point'))
    print(get_tweets_by_location('UK'))
    print(get_top_users(3))
    print(get_tweets_by_top_users(3))
    print(get_top_tweets_by_tags(3,['blockchain','bitcoin']))
    delete_tweets_by_followers(2000)
```


Muhammad Usman Malik 19i-0402 Shaheer Asif 19i-2152 Musa Ishaq 19i-0731

Dr.Adnan Tariq

Results

```
GET TWEETS BY USER:

 [{'hashtags': "['crypto', 'Bitcoin', 'BTC']", 'is_retweet': False, 'location': 'London, Essex and Suffolk', 'followers': '1335.0
 'created': '2015-02-11T12:44:29', 'friends': '797', 'source': 'Twitter for iPhone', 'text': 'MasterCard to integrate digital cur
ncy payments “directly on our network” \n\n#crypto #Bitcoin #BTC \u2066@MastercardUK\u2069… https://t.co/DQTb1JlFM3', 'verified':
alse, 'date': '2021-02-10T23:19:38', 'favorites': '2394', 'username': 'Mike Chambers at Northey Point', 'description': 'Unpacking
he UK’s retail payments landscape with my newsletter - subscribe at: https://t.co/LE1EJIArzG'}]
PS E:\FAST\Semester 8\Dist Data\Assignments\Dist-Data-Eng-Assignment-02>
```

```
GET TOP k USERS:

 [{'followers': '55350.0', 'username': 'Shira Rubinoff'}, {'followers': '385.0', 'username': 'GenZ Investor'}, {'followers': '17405
.0', 'username': 'U and i R 1'}]
```

```
GET TWEETS by TOP k USERS:

 [{'hashtags': "['blockchain', 'security', 'BTC', 'blockchaintechnology', 'CyberSecurity', 'Infographic', 'bitcoin']", 'is_retweet
: False, 'location': 'New York, NY', 'followers': '55350.0', 'created': '2011-10-04T15:14:07', 'friends': '30024', 'source': 'Twit
er for Android', 'text': 'The power of #blockchain \nExplained...\n#security #BTC #blockchaintechnology #CyberSecurity #Infographi
 #bitcoin… https://t.co/dYNWOALGqv', 'verified': False, 'date': '2021-02-09T15:44:19', 'favorites': '38212', 'username': 'Shira Ru
inoff', 'description': '#Cybersecurity & #Blockchain #Advisor, #ThoughtLeader #KeynoteSpeaker & #Author\nTop ranked #Influencer gl
bally in cybersecurity \n🔴 YouTube/ShiraRubinoffTV'}, {'hashtags': '', 'is_retweet': False, 'location': '', 'followers': '385.0',
'created': '2021-01-09T23:43:31', 'friends': '283', 'source': 'Twitter Web App', 'text': "Had to reupload because of some issues.\
If you haven't watched this yet I highly suggest it!… https://t.co/tVWAUAdlfV", 'verified': False, 'date': '2021-02-09T22:48:29',
favorites': '535', 'username': 'GenZ Investor', 'description': 'Here for the memes\nONLY 💎🖐🤚\n#Bitcoin & Stonks'}, {'hashtags':
"['bitcoin', 'like', 'cryptocurrency', 'news', 'btc', 'ethereum', 'eth']", 'is_retweet': False, 'location': '', 'followers': '1740
.0', 'created': '2013-09-09T10:52:17', 'friends': '148', 'source': 'Bitly', 'text': 'Enter Your Email To Claim FREE 3600 Coins ( $
6 Value )  #bitcoin #like #cryptocurrency #news #btc #ethereum #eth… https://t.co/7FGawL19oS', 'verified': False, 'date': '2021-02
10T19:41:25', 'favorites': '7', 'username': 'U and i R 1', 'description': ''}]
```


Muhammad Usman Malik 19i-0402 Shaheer Asif 19i-2152 Musa Ishaq 19i-0731

Dr.Adnan Tariq

Results

GET TWEETS BY SAME USERS IN A LOCATION:


```
 [{'hashtags': "['btc', 'Bitcoin', 'btcusd']", 'date': '2021-02-10
T19:41:21', 'userlocation': 'England, United Kingdom', 'followers'
: '8.0', 'favorites': '96', 'created': '2021-02-08T19:12:31', 'fri
ends': '53', 'text': '“Bitcoin is going to $100k” 🤣🤣\nIf this cl
oses like this on the daily, see you sub $20k.\n\n#btc #Bitcoin #b
tcusd… https://t.co/3cxApX1g0S', 'source': 'Twitter for iPhone', '
username': 'Charlie Scrutton', 'description': '21 | Fx & Crypto Jo
urney | ‘Just Swinging It’ |'}, {'date': '2021-02-10T22:18:53', 'u
serlocation': 'England, United Kingdom', 'followers': '70.0', 'fav
orites': '265', 'created': '2020-06-13T14:38:29', 'friends': '1356
', 'text': "BTC payment still not done\nIt's 21:18 pm now\nI know
it will take at least another hour.. maybe 3 hours+.. I can't t… h
ttps://t.co/qmmsZGibqM", 'source': 'Twitter for Android', 'usernam
e': 'Omar Arrawi', 'description': "BCH is the best cash so far. Tw
eet at me!\nInterested in art, technology, finance, blockchain and
 movies.\n(I'm now team BCH)"}, {'hashtags': "['help', 'bitcoin',
'BTC', 'Support']", 'date': '2021-02-10T22:55:42', 'userlocation':
 'England, United Kingdom', 'followers': '143.0', 'favorites': '16
3', 'created': '2016-04-10T13:35:49', 'friends': '453', 'text': '@
CoinbaseSupport - please will you reply to my DMs + 4 emails?  Tha
nks. #help #bitcoin #BTC #Support', 'source': 'Twitter Web App', '
username': 'Artceptual', 'description': "Gaming with a bit of art
thrown in!  I'm open to suggestions so feel free to drop it on me!
"}]
```


Muhammad Usman Malik 19i-0402 Shaheer Asif 19i-2152 Musa Ishaq 19i-0731

Dr.Adnan Tariq

THANK YOU

Muhammad Usman Malik 19i-0402 Shaheer Asif 19i-2152 Musa Ishaq 19i-0731

Dr.Adnan Tariq