Good morning every one

I am srinivas studying in Aditya engineering college and working as an intern at ADP

Initially we will start with Internship details followed by area of study post which we will discuss the main domain elasticsearch where I worked on then we will dive deep into requirements that I was given on elasticsearch, how I worked on the requirement and how I contributed my part to the project. Then we will throw some light on how my work was utilized in the project then we will walk through other requirements that I was given and finally we will discuss my learning outcomes then the floor will be open for questions.

Coming to the project

The project I was assigned to is N8 output tax compare and the project is to compare the output files from mainframe system and N8 cloud, the domains I was explored in the project is elasticsearch, cloud computing, python modules like pandas & boto3.

In the area of study, to implement elasticsearch in local and storing 2D data into it.

AWS cloud service like lambda, S3, SQS, SES. Python packages like panda and boto3. Cloud databases like DynamoDB and Aurora DB, understanding how the unit testing and pytest works and exposure to the software development tools like JIRA (story management) and GIT (code versioning)

In elasticsearch

* Learnt the concept of Elastic search
* Created mapping structure to insert the data
* Inserted data according to the requirement.
* Querying the data according to the requirement.

Data given to me to analyse

* An Excel sheet data will be given and the fields are

agency\_name

form\_name

ids\_with\_errors

ids\_without\_errors

* There may be any number of forms with a single agency and a single form contains many ids(with and without error)

The requirement is explained to me as a student, semester and passed and failed subjects

Need to get the student name who passed all the subjects (whole course)

The task given is to insert the data into elastic search in such a way that requirements are satisfied by querying on the data.

* Need to get agencies having all of it’s ids passed under all the forms.
* Need to get forms having all of it’s ids passed
* Get all the failed ids irrespective of the form and agency

Data should be inserted in a structured manner that is called mapping

* Selected a suitable mapping with nested structures.
* Level 1 with “agency\_name” and array of “forms”.
* Level 2 with “form\_name” and array of “ids” in array of forms.
* Level 3 with “id” and “no\_error” in array of ids.

Data insertion

* The data I will be having with columns

agency\_name

form\_name

ids\_with\_errors

ids\_without\_errors

* Firstly checking for the agency\_name in the index.
* If agency\_name is present in the index it will be retrieved and modified and inserted in the total document again using same document id.
* If the agency\_name is not present then the new document is created.
* If the agency\_name is present and the form\_name is not present in array of forms then it is added to the array and the document is inserted with the same document id
* Similarly all the data will be inserted in the index

Querying

* The query is to retrieve agency\_name in which there are no id\_with\_errors
* I used nested query inside a bool query
* First checked that there is no id\_with\_errors in ids and then checked for no forms with failed checks in ids and finally for no form\_name failure in forms
* Similarly other two queries are also done

My contribution to the project is data insertion, querying and output storing.

* As I wrote the code for data insertion and data querying the code helped the project in identifying

- agencies with no error ids

- forms with no error ids and

- ids with errors

* Code was implemented using AWS OpenSearch.
* Python boto3 module is used in connecting with the OpenSearch dashboard.

Another requirement I was given is to reduce the memory consumption while loading data into data frames

* When data is loaded into data frames those consume more memory usage for loading the data.
* The data is in deque files when those are loaded into data frames then all the column fields are in same data type this consume more memory.
* My task is to reduce the memory while loading the data into data frames.
* Successfully completed my part to reduce the data consumption while loading into data frames.

Finally the learning outcomes of the technologies that I have learnt are

Elasticsearch – Inserting ,querying and visualisation

AWS – serverless services

Databases – understanding cloud based data bases

Others – docker, JIRA, GIT, agile process and etc..