MILESTONE-1

TASK COMPLETION STATUS

***Pre-Requisites:***

1. Install MongoDB: <https://docs.mongodb.com/manual/tutorial/install-mongodb-on-windows/>
2. Connect to local host on Compass: mongodb://localhost:27017/admin?readPreference=primary&appname=MongoDB%20Compass&ssl=false
3. Basic Knowledge of query processing using cmd.
4. Knowledge of Agile Methodology.

***Task1: Data Pre-processing***

***Subtask1: Identify the type of data that we wish to use for this project and research to verify if large volume of data is available.***

Databases Includes: 4

Link to the Database and Information:

1. <https://data.edd.ca.gov/Industry-Information-/Current-Employment-Statistics-Mono-County/ru45-rtii>

Rows: 7062 Columns: 9

1. <https://data.edd.ca.gov/Employment-Projections/Short-Term-Occupational-Employment-Projections/guh4-bakw>

Rows: 781 Columns: 18

1. <http://eforexcel.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/>

Rows: 1000000 Columns: 14

1. <http://eforexcel.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/>

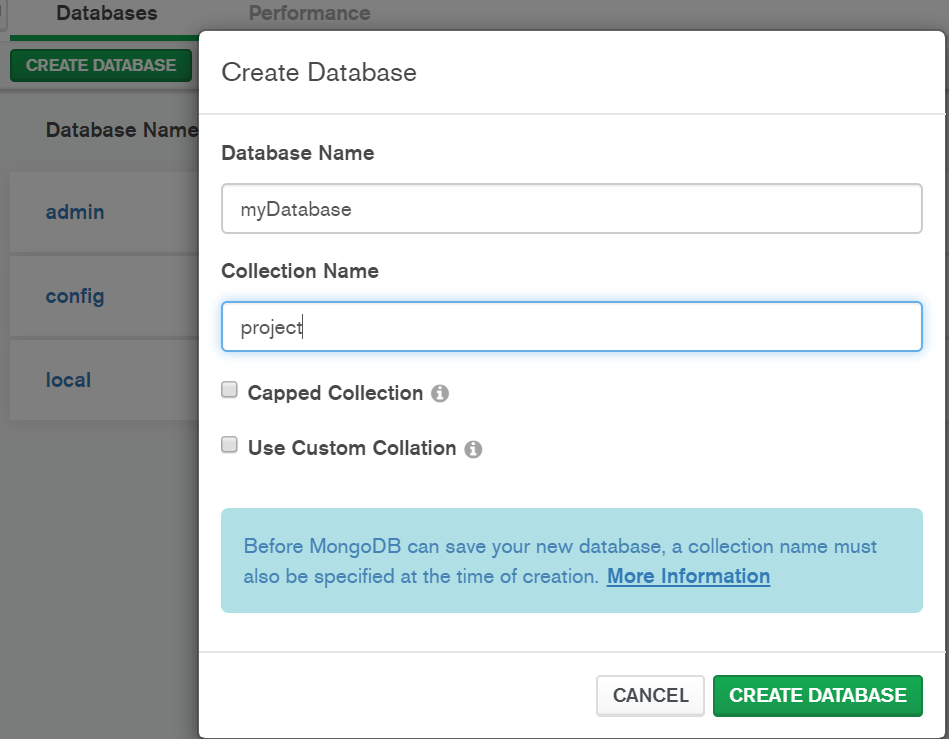
Rows: 2000000 Columns: 14

***Subtask2: Identify a smaller sub set of chosen data which would be used to populate the Database created***

Since it is a different variety of datasets involved, we did not consider a subset of the same data set .

***Subtask3: Create a database which would be populated with data from previous steps.***

Creating a Database in Compass:



Importing csv file to MongoDB using cmd:

mongoimport --db myDatabase --collection movies --type csv --headerline --file Desktop\MAC\Sem-2\ADT\Project\ amazontest.csv

***Subtask4: Identify the type of query that is common on the data and finalize it for execution.***

***Subtask5: Identify the various query performance metrics used widely in the industry.***

Evaluating Query Performance: <https://docs.mongodb.com/manual/tutorial/analyze-query-plan/>

<https://youtu.be/0WkJKa_Nv_o>

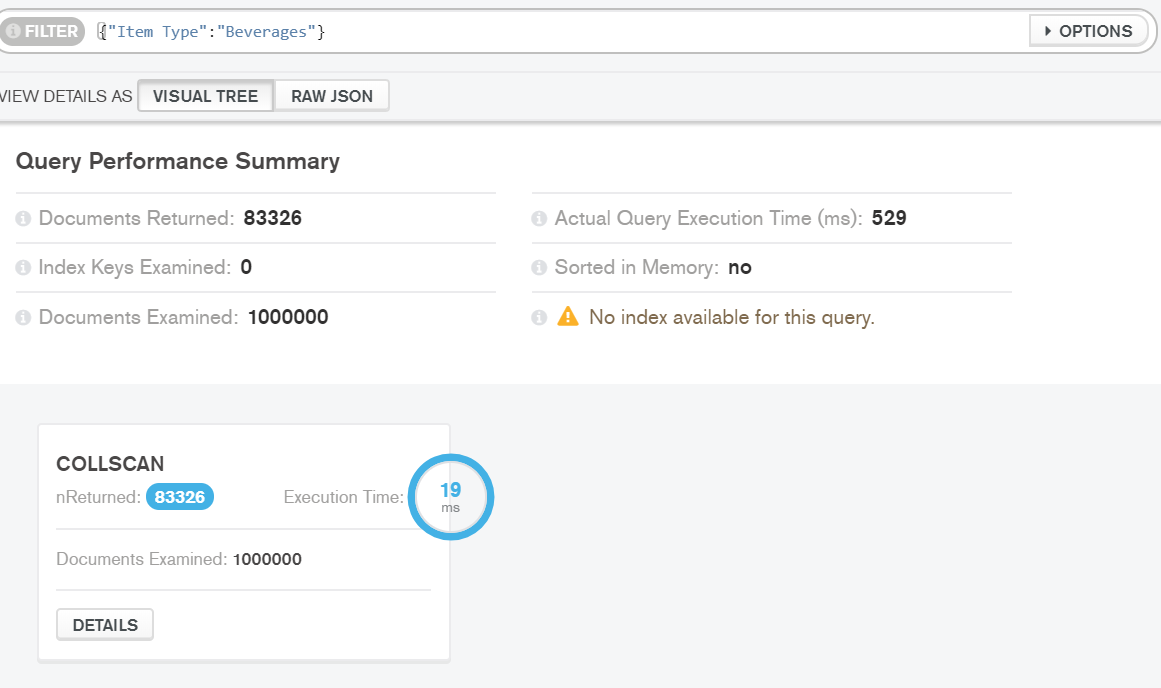
***Subtask6: Run the query from above step and document performance stats***

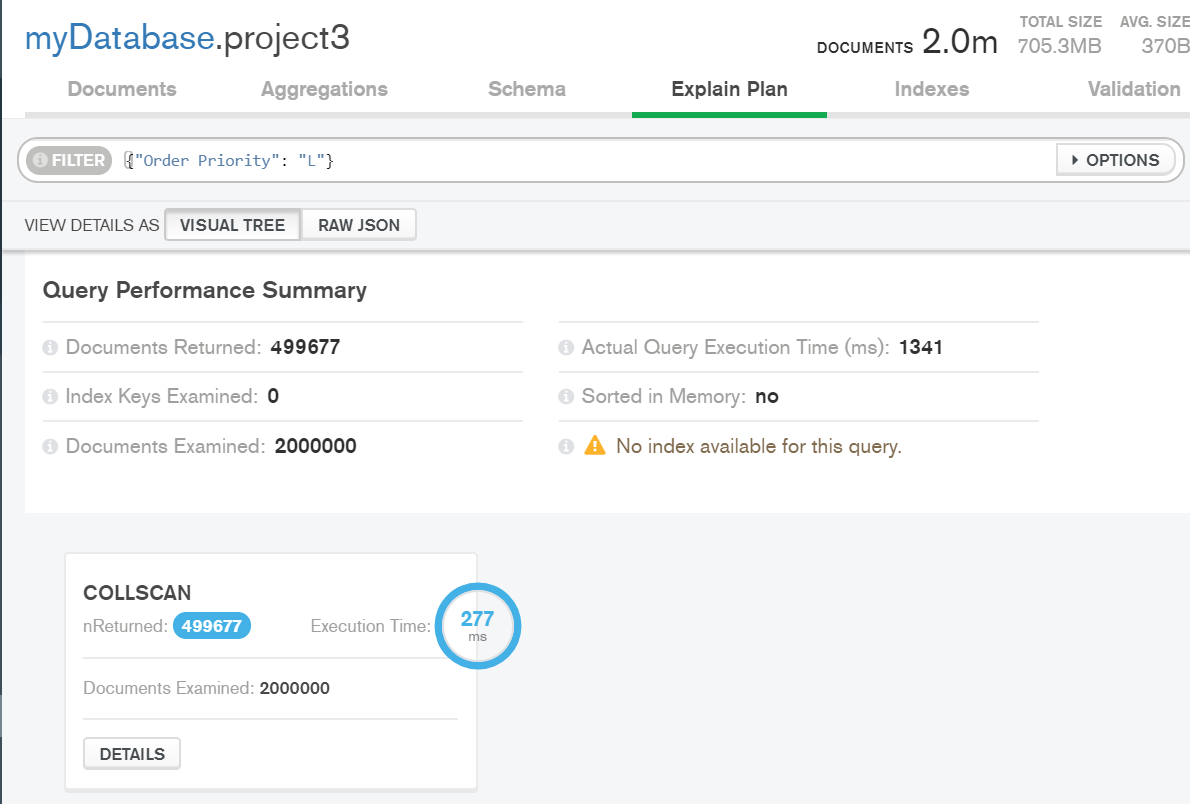
Run the find() query on all 3 datasets and note down their execution Stats.(Before Indexing)

1. db.project.find( { “Current Employment”: { $gt: “500” } } )

******

1. db.project1.find({"SOC Level":"4"}).limit(500)
2. db.project2.find({“Item Type”: “Beverages”})

******

******

***TASK2: INDEXING***

***Subtask1: Create an index.The index stores the value of a specific field or set of fields, ordered by the value of the field***

Indexing : <https://docs.mongodb.com/manual/indexes/>

Set Indexes: db.project.createIndex({Current Employment:1})

***Subtask2: Indexing will attempt to re-index any items that have previously failed to be indexed***

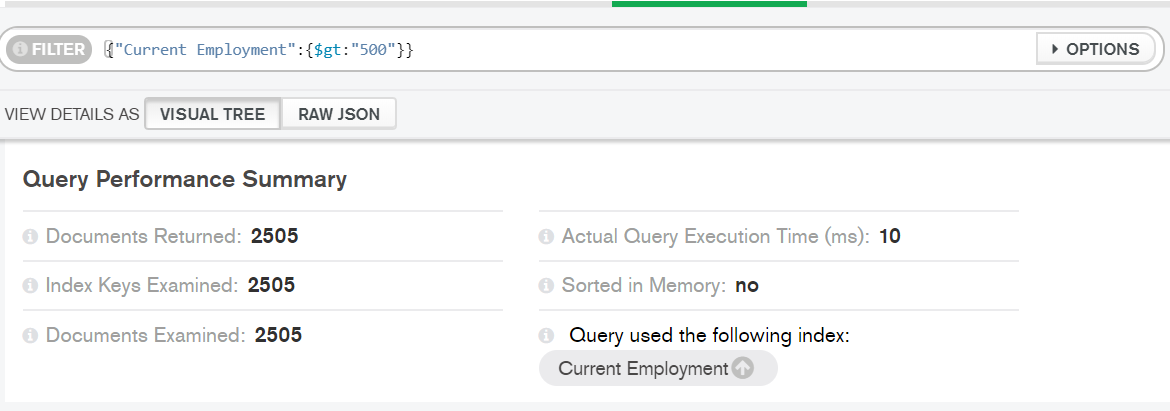
***Subtask3: Different checks will then be performed depending on the Verification level***

***Subtask4: Once the upgrade has completed successfully, the old index volume will be replaced with the new volume.***

***Subtask5: Rebuild index volumes. Once the rebuild has completed successfully, the old index volume will be replaced with the new volume.***

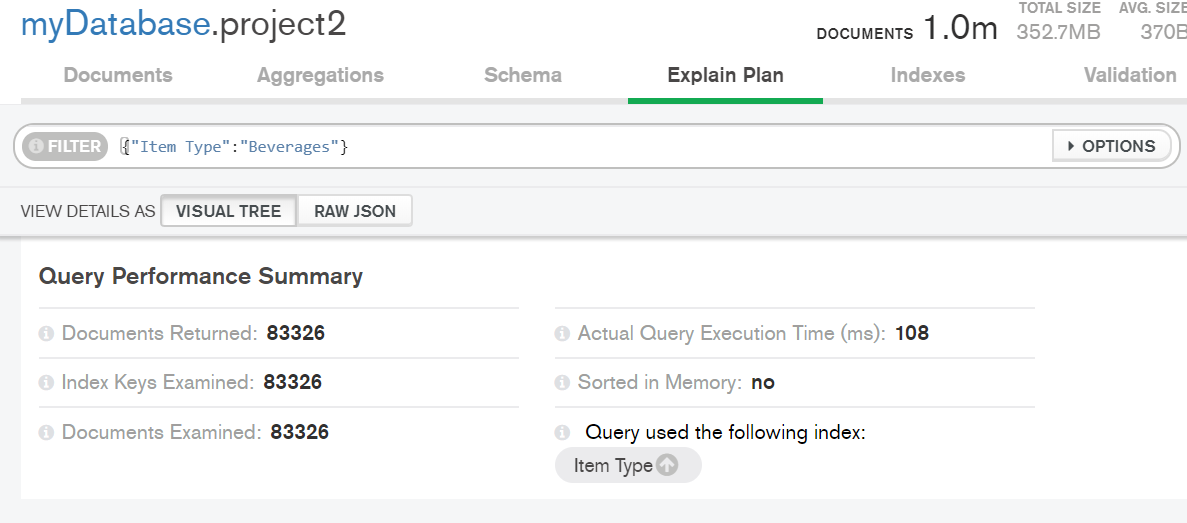
Results after Indexing:

Collection: project



Collection: project1:

Collection: project2:



Collection: project3

