**Activity 12**

This activity will illustrate how query efficiency can be improved with the use of indexes. To begin this activity, create the *sample* database based on the following SQL schema:

|  |
| --- |
| CREATE DATABASE sample;  \c sample  \timing  DROP TABLE Sample;  CREATE TABLE Sample (  id INT PRIMARY KEY,  rnd INT NOT NULL  ); |

\timing will enable timing for the queries.

SQL template is available here:

Use the COPY command below to populate the Sample table with increasingly large files. Between the loads you will perform a query on the non-prime attribute “rnd” two times: before and after an index on that attribute has been created. You will record how long it takes for Postgres to answer those queries and whether using an index affects those query response times.

COPY command:

|  |
| --- |
| -- replace ? by 1, 2, 3, or 4 depending which file you want to \COPY Sample (id, rnd) FROM '<YOUR PATH>/sample?.csv' DELIMITER ',' CSV HEADER; |

INDEX commands:

|  |
| --- |
| CREATE INDEX rnd ON Sample(rnd);  -- make sure you drop the index before timing a new sample  DROP INDEX rnd; |

Files to use:

* [sample1.csv](https://msudenver-my.sharepoint.com/:x:/g/personal/tmota_msudenver_edu/EafzeeFDBjpPuiR1NtarYAEBOJdvNNdgEyWngyJNd7-5Gw?e=EcJaVV) (10KB to a final Sample size of 1K rows);
* [sample2.csv](https://msudenver-my.sharepoint.com/:x:/g/personal/tmota_msudenver_edu/ESrBMmIIiu9Dig9ePizZk1cBUYXZOkAyMUuySUH2Dch4zA?e=BxuZpv) (1.2MB to a final Sample size of 100K rows);
* [sample3.csv](https://msudenver-my.sharepoint.com/:x:/g/personal/tmota_msudenver_edu/ET_LlfUpLc5CoeAROBNigd4BLIUZL07_uev_Jafnbk31SA?e=mKlDSZ) (11.6MB to a final Sample size of 1M rows);
* [sample4.csv](https://msudenver-my.sharepoint.com/:x:/g/personal/tmota_msudenver_edu/EaAnO04PBWpPnccN3AMA22kBHsxlUhqeF5XEubA4W-FfzA?e=BapJrZ) (125MB to a final Sample size of 10M rows).
* [sample.zip](https://msudenver-my.sharepoint.com/:u:/g/personal/tmota_msudenver_edu/EQB0RZ6iYsdKrPR_zE130OIBRWNOsb3obomCfWH9fWBobg?e=mDuKQn) (ALL 4 files)

Queries to perform:

* after loading sample1.csv file:
  + SELECT \* FROM Sample WHERE rnd = 500;
* after loading sample2.csv file:
  + SELECT \* FROM Sample WHERE rnd = 50000;
* after loading sample3.csv file:
  + SELECT \* FROM Sample WHERE rnd = 500000;
* after loading sample4.csv file:
  + SELECT \* FROM Sample WHERE rnd = 5000000;

Record your times on the following table:

|  |  |  |
| --- | --- | --- |
| **Sample** | **Before/After Index** | **Query Time (s)** |
| sample1.csv - 1K rows | Before | **3.281** |
| sample1.csv - 1K rows | After | **1.759** |
| sample2.csv - 100K rows | Before | **9.400** |
| sample2.csv - 100K rows | After | **4.653** |
| sample3.csv - 1M rows | Before | **94.298** |
| sample3.csv - 1M rows | After | **2.211** |
| sample4.csv - 10M rows | Before | **442.871** |
| sample4.csv - 10M rows | After | **1.658** |