Activity 2 – Data on Electricity Use

Please note: You may do this activity alone or in groups of two.

The course website contains materials needed for this question, including three tab delimited data files:

name	size (bytes)	size (records)	No. Attributes
device.txt	18452	554	3
facility.txt	21851	337	6
worklog.txt	7693369	100,000	10

Below you will find an ER model. Examine the ER model and before starting please be sure that you understand how the three tables work together.

Using this data conduct an investigation into the use of indexes for improving performance and write up a report.

1. **Getting starting**. To get started, create three tables, one for each of the above text files. Create the primary and foreign keys and load the tables with the data. Please submit your SQL script for creating the database in your report (described below).

   Please note: You will find some data integrity problems, which is very common with real data sets. When you encounter a problem (e.g., a violation in a referential constraint, that is, a foreign key).

2. **Query analysis**. A common query is to determine how much electricity is used by a device for a given date range. For example, the following query will determine how many readings are between two dates and the sum of those readings for one device (143):

   ```
   select count(e), sum(e) from work_log
   where did=143 and timelogged
   between '2007-10-01 08:30:00-07' and '2007-10-01 09:30:00-07';
   ```

   Before proceeding make sure that you understand this query. Once you understand it, please follow these steps:

   a) **Describe the query**. Describe exactly what the query does, how the query works, and what operations must be performed in order to compute a correct result. What is the correct result?

   b) **Baseline performance measure.** Determine how long it takes to complete this query. Here is a simple SQL script that can be used to generate an estimate (but use whatever approach you like):

   ```
   select date_trunc('milliseconds',now()) as "start";
   /* Your SQL Query Goes Here */
   select date_trunc('milliseconds',now()) as "end";
   ```

   c) **Query analysis**. Carefully analyze the relation WORK_LOG and the query. In addition, use the SQL EXPLAIN command to examine how precisely the query is "planned." Report your analysis of the relations, the query, and the results of the SQL EXPLAIN command and draw conclusions on the performance of this query. How might the query be sped up?

d) **Tuning**. Using indexes and/or other methods seek to improve the performance of the above query.

e) **Experimental performance measure**. Repeat the performance measure and query analysis to determine if your approach for improving the performance of the query worked.

3. **Additional queries**. Repeat step #2 for two additional queries of your own invention. One of your queries must join data from two or more tables.

   The goal of these experiments is to show how indexes and/or other physical database design methods can speed up or slow down database access and update operations.

   Write up your experiments and findings neatly, providing sufficient information so that we could replicate your experiments. Wherever appropriate give explanations for the performance "speed up" or "slow down."

4. **Data partitioning**. Outline a partitioning strategy for improving the performance of this system. Document some reasonable assumptions about: 1) Data insertion; 2) Data update; and 3) Major types of queries.

Deliverable and grading

Prepare a project report of the above components and submit it to the dropbox.

1. Neatness and attention to detail; Absence of spelling and grammatical errors; Clear organization and concise writing; A rigorous and careful analysis.

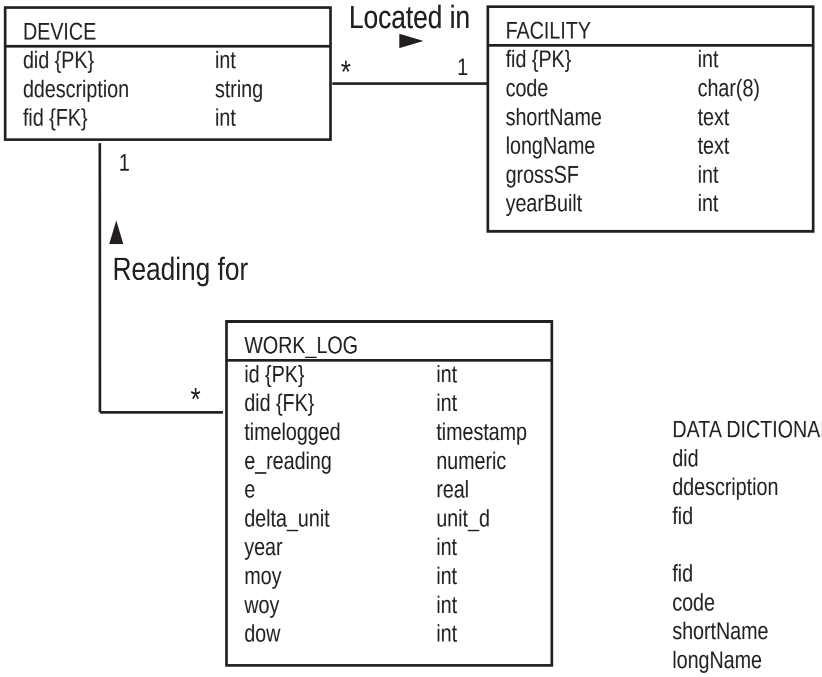
2. For each query:
   a) A clear description of the query and interpretation of the results of the EXPLAIN command;
   b) A well-described approach, with rationale, for how the query can be sped up;
   c) A clear presentation of empirical data showing how the indexes or other methods improved the performance of the query;
   d) A strong discussion of the empirical data and its interpretation.

3. A brief introduction and conclusion of this activity.

4. A brief reflective statement on what you have learned.

A3 ER MODEL

This is a simplified schema for storing electricity data. The table WORK_LOG contains time series data of electricity readings. Conceptually, each device (an electricity meter) sends a new reading into the system every 15 minutes. These readings are stored in the the table WORK_LOG. Devices are installed in facilities, which have a code, name and other basic information. (Note: The actual data model for this system is more complex.)



DATA DICTIONARY

did	device id -- the unique identifier for each meter
ddescription	a short description of the device
fid	see below
fid	facility id -- each device is located in a facility
code	A short code name for a facility (ie., a buidling)
shortName	A short name for the facility
longName	A long name for the facility
grossSF	The size of the facility in gross square feet
yearBuilt	The year that the facility was built
id	unique id for each record in the time series
did	see above
timelogged	time that the reading was posted
e_reading	the electricity reading (kWh) from the meter -- it is the running total
e	this is the electricity used in an interval (kWh)
delt_unit	a code for different interval, either A, B, C, D (A - 15min; B-1h; C-4hr; D-24hr) (In WORK_LOG this value is always A)
year	year of the reading
moy	month of year (1-12)
woy	week of year (1-53) of reading, with week #1 having January 4 in it (ISO-6801)
dow	‘day of week’ of reading (0-Sunday ... 6-Saturday)