**Database Design Exercise**

**Problem Definition:**

You are given a comma-separated file containing 6 columns. You have to design the database schema(s), process to insert the data and how to retrieve the data (retrieving is not too difficult once the schema is designed efficiently.)

**Input File Format:**

HOUSEHOLD\_ID, BRAND, VERTICAL, AD\_SPOT\_ID, AD\_DATE, VIEW\_DURATION

HOUSEHOLD\_ID: Identifier of a household

BRAND: Name of the brand for which ad was aired

VERTICAL: Classification of a brand

AD\_SPOT\_ID: ID of the ad being aired

AD\_DATE: Date on which the ad was aired

VIEW\_DURATION: Time in seconds for which the household watched the ad.

**Size of input file:**

Input file contains around a billion (or more) entries.

**Inputs/Outputs/Constraints:**

1. Required inputs:

a. VERTICAL (only one will be present at a time)

2. Optional Inputs:

a. List of Brand(s)

b. Date Range

c. List of household\_id(s)

3. Outputs:

a. Number of Ads viewed

b. Number of Distinct Ads

c. Total duration of the Ads viewed (distributed by households\_id)

4. Constraints:

a. Input file always contains a week worth of data.

b. Brand and Vertical have a one-to-one relationship

**Further information:**

Relationship between Vertical/Brand: An Example of vertical is Automotive, and Brands could be Toyota, Honda, BMW etc. Similarly, another example of vertical could be Fast Foods and Brands could be McDonald’s, KFC, etc.

In short Vertical could be thought of a collection of Brand(s).

**Sample File:**

HOUSEHOLD\_ID, BRAND, VERTICAL, AD\_SPOT\_ID, AD\_DATE, VIEW\_DURATION

1, Toyota, Automotive, 1, 2016-01-01, 10

2, Toyota, Automotive, 1, 2016-01-01, 9

1, KFC, Fast Foods, 2, 2016-01-02, 15

**Possible Starts:**

Start with a base (un-optimized) schema for the table. For example, have a single table containing all the columns. Then check the output requirements and try to figure out what are the bottlenecks and work through optimizations step-by-step.