See:

<http://stackoverflow.com/questions/8862845/how-to-model-airport-flight-data-in-a-graph-database-like-neo4j>

Trucking, Inc. is a trucking broker that arranges truckers to carry loads of goods. Trucking, Inc. connects truckers to loads that need to be delivered. Trucking, Inc. makes its money by getting a percentage of the cost of the load.

For example, a factory wants to ship furniture. They ask Trucking, Inc. to find the cheapest trucker. Trucker A says they will ship the furniture from the factory in New York to Chicago for $200. Trucking, Inc. then charges the factory $220.

Up to know, Trucking, Inc. only has the zipcode of the factory and the zipcode of the delivery point.



Trucking, Inc. has historically calculated zipcode start to zipcode end profitability.

And because there was not a lot of data of zip code to zip code travel a profitability could not be profitably established.

We believe that we can determine profitability and pricing decisions by analyzing the data at the route level.

Our goal is to model this structure in neo4J



That way we can look at the frequency of travel on a route:



If we had this data:

|  |  |  |
| --- | --- | --- |
| **Starting Zip** | **Ending Zip** | **Profitability** |
| Zipcode A | Zipcode D | $23.34 |
| Zipcode A | Zipcode F | $54.78 |
| Zipcode B | Zipcode E | $12.98 |
| Zipcode C | Zipcode A | $87.53 |
| Zipcode E | Zipcode B | $52.39 |
| Zipcode F | Zipcode D | $83.23 |

Initially, we are loading data only from the state of Pennsylvannia.

We will have three different data sets that we need to integrate.

First, there is the zipcode to zipcode data which is in the spreadsheets. There are two sheets: one for Total Load and the other for Partial Load. A total load means the entire truck is filled with a single product. Partial Load means that only a partial truck was filled. If it is a total load than it means the truck only had one starting point and one ending point. If it is a Partial Load than truck may take an indirect route and need to stop at multiple locations to pick up or drop off Partial Loads.

We will need to differentiate the two types of loads in the data.

These spreadsheets should be loaded using the technique described in the URL you were provided. If you have an alternative method we can discuss.

Second, there is the zipcodes of the highways. We will identify each highway in the state and determine the zipcodes through which the highway runs.

Data will have this kind of structure:

|  |  |  |
| --- | --- | --- |
| Highway Number | Sequence | Zipcode |
| 475 | 1 | 53776 |
| 475 | 2 | 53785 |
| 475 | 3 | 53748 |
| 475 | 4 | 53715 |
| 475 | 5 | 53796 |
| 475 | 6 | 53703 |
| 475 | 7 | 53793 |

We will then be able to see if a pick-up zip code matches a highway zipcode and be able to see which route a trucker likely took. We will not be certain of this fact.

Third, there is data on the closest highway to a starting zip code. If all routes started on a highway found in data set #2 then everything would be very simple. However, in many cases a load is picked at a location in a specific zipcode and then the trucker drives to the nearest highway and drives to the destination. We will have the distance of the centroid of the zipcode to the closest highway.

The data may have a structure like this:

|  |  |  |  |
| --- | --- | --- | --- |
| **StartingZipcode** | **Distance** | **Highway #** | **HighwayZipcode** |
| 53743 | 9 | 475 | 53776 |
| 53744 | 3 | 65 | 53936 |
| 53719 | 19 | 85 | 53790 |
| 53719 | 10 | 475 | 53785 |
| 53793 | 0 | 475 | 53793 |

Our goal is to model this:



Since I know the profitability and margin data on the route from zipcode 53743 and 53965 I should be able to then determine the following:

For trucks passing through zipcode 53785 (presumedly on highway 475) what is the average margin across all trucks that traversed this route.

And, for the highway 475 what is the average margin across all of the zipcodes (53776, 53785, 53748, and 53715)

The tasks of this demo is to:

1. Deploy a neo4J instance
2. Deploy the data
3. Deploy a R-neo4J interface
4. Make the R interface available to the statistics company
5. Construct graph query ourselves to see if we can answer the question differently than the statistics company using R.
6. Create a simple UI that would allow you to enter a starting zipcode and an ending zip code and provide margin and revenue