WASHINGTON STATE UNIVERSITY
VANCOUVER

Project 1

For
CS 453 Cloud Data Management
Washington State University
Vancouver, Washington

Tyler Bounds	**11258134**	**tyler.bounds@wsu.edu**
Stephen Brown	**11466405**	**stephen.p.brown@wsu.edu**
Chris Hight	**11483028**	**chris.hight@wsu.edu**

Date: 10/7/2016

Part 1: Data Modeling

Key/Value Model

The key value model uses a key to search for a specific section in a table and returns the rest of the table relating to the key that was initially searched.

- CSV = Freeway_detectors
  - Key = detectorID
  - Values = highwayID, milePost, locationText, detectorClass, laneNumber, stationID

- CSV = Freeway_loopdata
  - Key = detectorid
  - Values = starttime, volume, speed, occupancy, status, dqflags

- CSV = Freeway_stations
  - Key = stationid
  - Values = highwayid, milepost, locationtext, upstream, downstream, stationclass, numberlanes, latlon, length

- CSV = Highways
  - Key = highwayID
  - Values = shortDirection, direction, highwayName

Document Model

For all data files, we would use a JSON file to organize the data. Depending on the database used the document setup may have an object ID or a key.

Highways.csv

In the case of Highway.csv, every object's key would be highwayid and values would be shortdirection, direction, and highwayname. Some databases like MongoDB may use a hashvalue as well which was added into the example. Here is an example of what it would look like:

```
[
{
 “_id” : ObjectId(“hashkey”),
  "highwayid": 3,
  "shortdirection": "N",
  "direction": "NORTH",
  "highwayname": "I-205"
},
{
 “_id” : ObjectId(“hashkey”),
  "highwayid": 4,
  "shortdirection": "S",
  "direction": "SOUTH",
  "highwayname": "I-205"
}
]
```

Freeway_stations.csv

In the case of Freeway_stations.csv, every object's key would be stationid and values would be highwayid, milepost, locationtext, upstream, downstream, stationclass, numberlanes, latlon, length.

Freeway_loopdata.csv

In the case of Highway.csv, every object's key would be detectorid and values would be starttime, volume, speed, occupancy, status, dqflags.

Freeway_detectors.csv
In the case of Highway.csv, every object's key would be highwayID and values would be shortDirection, direction, highwayName.

Column Data Model

For a basic data model, the database is a collection of key/value pairs, and the key consists of 3 parts: a row key, a column key, and a timestamp. This is a flexible schema which means the set of columns isn't fixed, and may differ row-to-row. The column key consists of two parts: a column family and a qualifier.

- CSV = Freeway_detectors
    - Time = timestamp
    - RowID = ID number
    - ColumnIDs = detectorid, highwayid, milePost, locationText, detectorClass, laneNumber, stationID

- CSV = Freeway_loopdata
    - Time = timestamp
    - RowID = ID number
    - ColumnIDs = detectorid, starttime, volume, speed, occupancy, status, dqflags

- CSV = Freeway_stations
    - Time = timestamp
    - RowID = ID number
    - ColumnIDs = stationid, highwayid, milepost, locationtext, upstream, downstream, stationclass, numberlanes, latlon, length

- CSV = Highways
    - Time = timestamp
    - RowID = ID number
    - ColumnIDs = highwayid, shortDirection, direction, highwayName

Problem 2

Describe in words or pseudo-code how you could answer the queries below:

- **Count high speeds: Find the number of speeds > 100 in the data set.**
    - **Key/Value**
        - Using freeway_loopdata, use the key "speed" to find all values > 100, and then find number of results for the count.
    - **Document**
        - Same process as key/value model.
    - **Column**
        - Same process as key/value model.
        - Count all rows where the speed values are > 100 and where columnID = "speed".

- **Volume: Find the total volume for the station Foster NB for Sept 21, 2011.**
    - **Key/Value**
        - Using freeway_detectors, find all detectorids where locationtext == "Foster NB".
        - Using the result and freeway_loopdata, find and sum all volumes associated with the previously gathered detectorids and where the starttime is 9/21/2011.
    - **Document**
        - Same process as key/value model.
    - **Column**
        - Using freeway_detectors, get all detectorids where columnID = "locationtext" and locationtext == "Foster NB"
        - Using the result and freeway_loopdata, find and sum all volumes associated with the previously gathered detectorids where the starttime is 9/21/2011

- **Single-Day Station Travel Times: Find travel time for station Foster NB for 5-minute intervals for Sept 22, 2011. Report travel time in seconds.**
    - **Key/Value**
        - Using Freeway_stations, find stationID where locationtext == FosterNB
        - Using Freeway_detectors, find detectorIDs where stationID == previous result
        - Using freeway_loopdata, find all results matching detectorID found in previous result.
        - Search the data for the first five minutes based on starttime
        - Find average speed in result delivered.
        - Using length and average speed queried calculate (length/avg speed) * 3600
        - Repeat for 5 minute intervals until 24 hours is completed
    - **Document**
        - Same process as key/value model.
    - **Column**
        - Using Freeway_stations, find stationID values where columnID == locationtext and locationtext == FosterNB
        - Using Freeway_detectors, find detectorID value where columnID == stationID and stationID == previous result
        - Using freeway_loopdata, find all results matching detectorID found in previous result.
        - Search the data for the first five minutes based on starttime
        - Find average speed in result delivered.
        - Using length and average speed queried calculate (length/avg speed) * 3600
        - Repeat for 5 minute intervals until 24 hours is completed

- **Peak Period Travel Times: Find the average travel time for 7-9AM and 4-6PM on September 22, 2011 for station Foster NB. Report travel time in seconds.**

- **Key/Value**
  - Using freeway_stations, find length for when locationtext == Foster NB
  - Using freeway_detectors, find detectorid for when locationtext == Foster NB
  - Using freeway_loopdata and the result from the previous query, find speed for when detectorid == the previously found detectorid and when date == 9/22/11.
  - Query for data based on starttime (time >= 7:00 and <= 9:00 and >= 16:00 and <= 18:00)
  - Find average speed in results returned by previous query for each time interval.
  - Calculate using (length/avg speed) * 3600
- **Document**
  - Same process as key/value model.
- **Column**
  - Using freeway_stations, find length for when columnID == locationtext and locationtext == Foster NB
  - Using freeway_detectors, find detectorid values for when columnID == locationtext and locationtext == Foster NB
  - Using freeway_loopdata and the result from the previous query, find speed for when columnID == detectorid and detectorid == the previously found detectorid and when columnid == date and date == 9/22/11.
  - Query for data based on starttime (time >= 7:00 and <= 9:00 and >= 16:00 and <= 18:00)
  - Find average speed in results returned by previous query for each time interval.
  - Calculate using (length/avg speed) * 3600

- **Peak Period Travel Times: Find the average travel time for 7-9AM and 4-6PM on September 22, 2011 for the I-205 NB freeway. Report travel time in minutes.**
  - **Key/Value**
    - Using highways, find highwayid where direction == NORTH
    - Using freeway_stations, find detectorID where highwayid == previous result
    - Using freeway_loopdata, find all data based on previous result and on starttime (date is 9/22/11, time >= 7:00 and <= 9:00 and >= 16:00 and <= 18:00)
    - Find average speed and for each length in query delivered
    - Get the average travel time by using sum((length)/avg(speed)) * 60
  - **Document**
    - Same process as key/value model.
  - **Column**
    - Using highways, find highwayid where columnid == “direction” and direction == NORTH
    - Using freeway_stations, find detectorID where columnid == “highwayid” and highwayid == previous result

        - Using freeway_loopdata, find all data based on previous result and when columnid = “starttime” and starttime == 9/22/11 between the time intervals >= 7:00 and <= 9:00 and >= 16:00 and <= 18:00)
        - Find average speed and for each length in query delivered
        - Get the average travel time by using sum((length)/avg(speed)) * 60

- **Route Finding: Find a route from Johnson Creek to Columbia Blvd on I-205 NB using the upstream and downstream fields.**
    - **Key/Value**
        - Using Freeway_stations find locationtext == Johnson Cr NB and downstream value.
        - Concatenate locationtext to a string variable.
        - If locationtext != Columbia Blvd 205 NB, concatenate locationtext, and using freeway_stations again based on previous results downstream result.
    - **Document**
        - Same process as key/value model.
    - **Column**
        - Using Freeway_stations and columnID == locationtext find locationtext == Johnson Cr NB and value where columnID == downstream.
        - Concatenate locationtext to a string variable.
        - If locationtext != Columbia Blvd 205 NB, concatenate locationtext, and using freeway_stations again based on previous results downstream result.
-