Introduction

1. Application of Big Data - Hadoop
2. Hadoop File Systems – HDFS
3. Data Extraction, loading and transformation – ETL
4. Reporting and Analytics with Hadoop and Excel

Objective

By using the provided HDP tools and techniques, the students will process the data to create their own analytics reports and create a visual presentation that will benefit decision-makers.

This virtual lab-based course will use truck fleet data to refine and analyse trucking movement in order to meet the organizational goal of better understanding risk. The use case involves geographic data, vehicles, average mileage, gas consumption, events, risk factors, and other supporting information.

Each truck has been equipped to with devices to log location and event data. These events are streamed back to a datacentre where students will process the data and revise truck movements to increase safety.

Business Objective

Accidents caused by large trucks remain one of the leading causes of injuries and deaths in the United States. The objective is identifying dangerous commercial truck drivers nationwide

Upon completing this lab, the student will be able to answer common business questions related to the trucking business, such as:

Which truck has the highest risk factor based on geographic location and time?

Exercise 1

In this exercise, the student will create the set of geographic data needed for the remaining exercises.

Objective

Using geographic and census data from public resources (such as “census.gov”), the student will download geographic information including latitude, longitude, city and state. Then, the

data can be enriched by truck fleet information such as total mileage, average mileage, and gas consumption. The truck fleet information can be created by the student or obtained by using the sample data provided in this lab.

Outlines

1) Search the Internet to obtain geographic information including latitude, longitude, city and state from public internet resources.
2) Research and find truck fleet information from public internet resources related to the geographic information you have identified in Step 1. The truck fleet information includes: truck id, mileage, average mileage, total mileage, and gas consumption.
3) Download the data, examine it, and store it in a structure table or Excel sheet.
4) Enrich the geographic data obtained from step 1 with the truck fleet information obtained from step 2.
5) Store the final set of documents into your own folder.

Sample Data (Public free samples)

File name	Source	attachment
Geographic information	https://www.uscitieslist.org (Free sample) https://www.cpsc.gov https://www.data.gov https://www.cms.gov	us-cities-sample.xls
Truck fleet information		trucks.csv trucks_mg.csv
Geographic information with fleet information		GEOLOCATION.xlsx

Exercise 2

Loading Data into Hadoop File System (HDFS)

Objectives

In this exercise, you will load the geographic data you created in your local machine into HDFS, which is in Cloudera VM. You can perform tasks like create directories, navigate file systems, and upload files to HDFS. In addition, you'll perform a few other HIVE related tables loading tasks.

Outlines

1. Transfer your input files into Cloudera VM by utilizing the system copy command.

   "SCP". Example: **$ scp D:\truck.csv training@192.168.37.130:/home/training/test_input**

2. Create a Hive Table for geolocation tab in Geolocation.xls file

```
CREATE TABLE geolocation_stage
(
truckid string,
driverid string,
event string,
latitude DOUBLE,
longitude DOUBLE,
city string,
state string,
velocity BIGINT,
event_ind BIGINT,
idling_ind BIGINT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```

3. Load the imported file into HIVE by utilizing "LOAD DATA INPATH" (see pg#31-chapter 5)
4. You may create additional tables to load the tab "riskfactor" and the remaining tabs if needed.

Exercise 3

Integrating HDFS with Tableau via JDBC drivers for Impala and Hive

Objectives

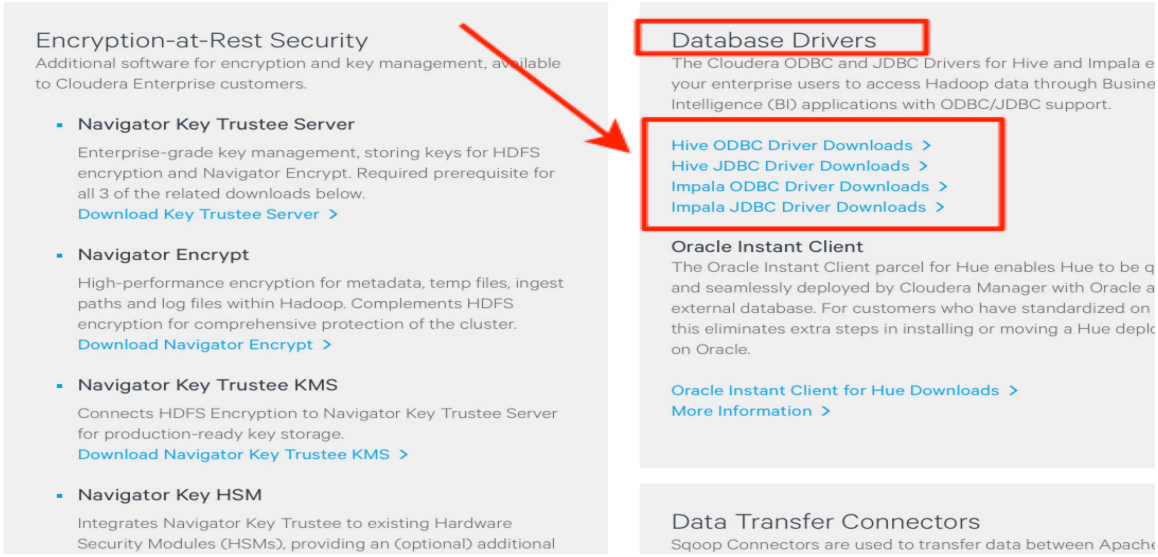
In this exercise, you will be integrating HDFS instance including the tables with an external analytical tool. Then you can perform some analytics or calculations, creating visualizations or dashboard if needed.

Outlines

1) Identify the correct driver
2) Install the JDBC driver
3) Load the tables into the analytical tools via ETL transactions
4) Create the Charts needed.

Steps

1) Go to the url https://www.cloudera.com/downloads.html



Step 2

Install ODBC and JDBC drivers for Hive and Impala.

Step3

If you are using Tableau as an analytical tool, follow the below steps to connect to HDFS.

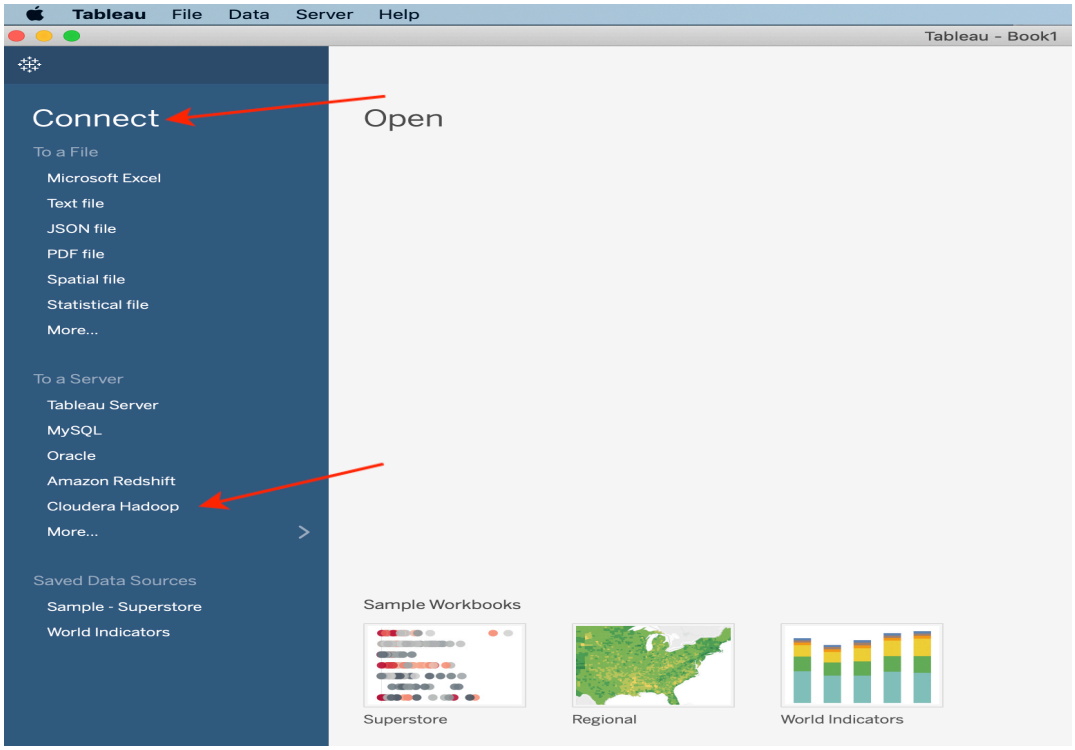


Tableau File Data Server Help
Tableau – Book1
Connect
To a File
Microsoft Excel
Text file
JSON file
PDF file
Spatial file
Statistical file
More...
To a Server
Tableau Server
MySQL
Oracle
Amazon Redshift
Cloudera Hadoop
More...
Saved Data Sources
Sample - Superstore
World Indicators
Open
Sample Workbooks
Superstore
Regional
World Indicators

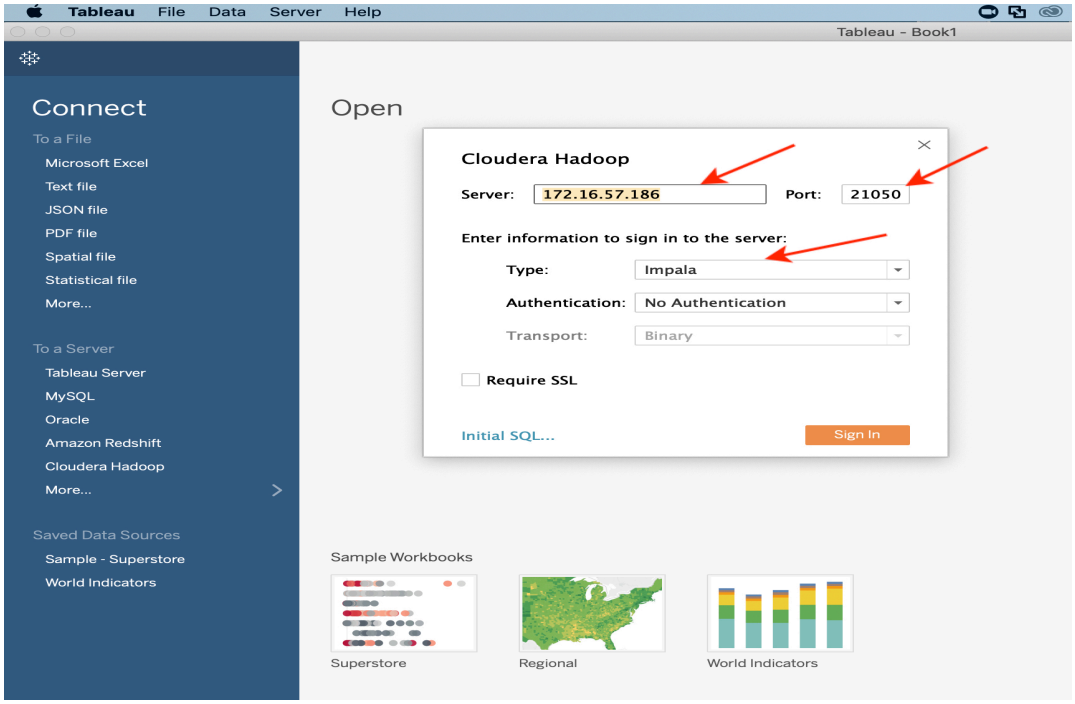
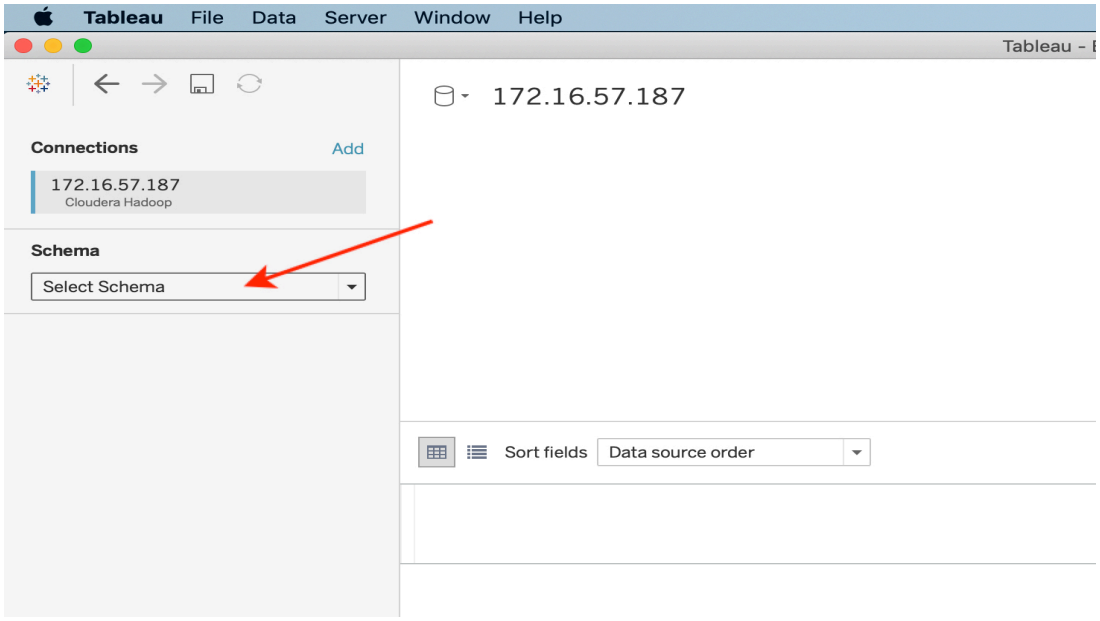
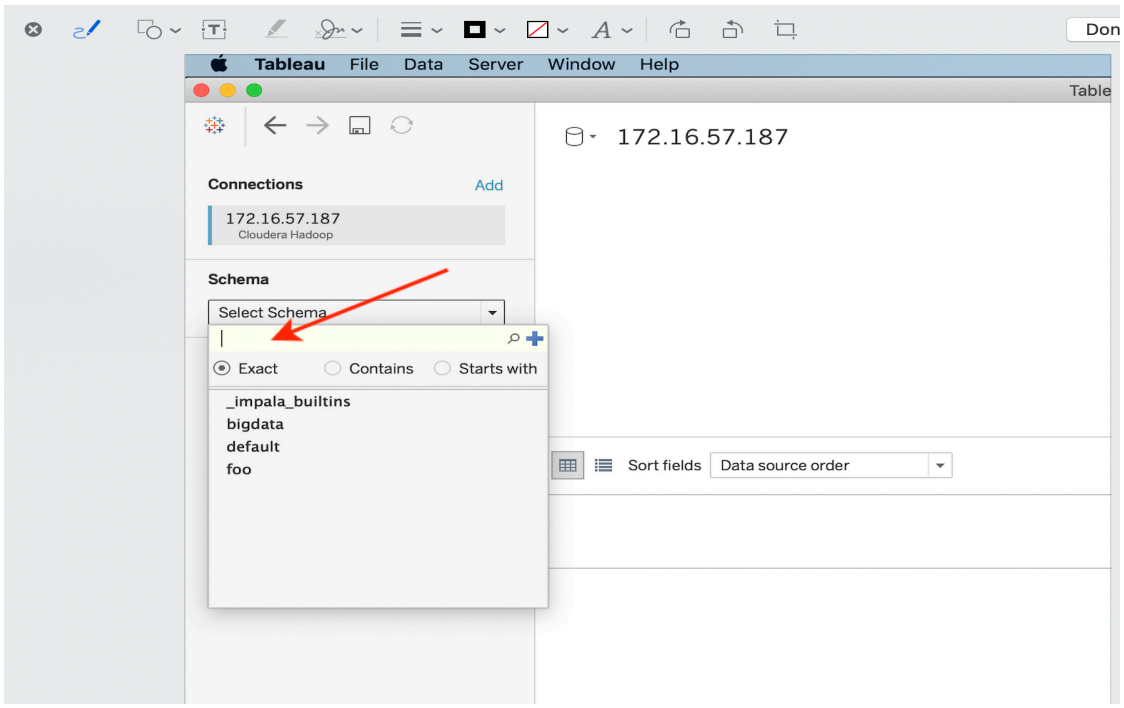


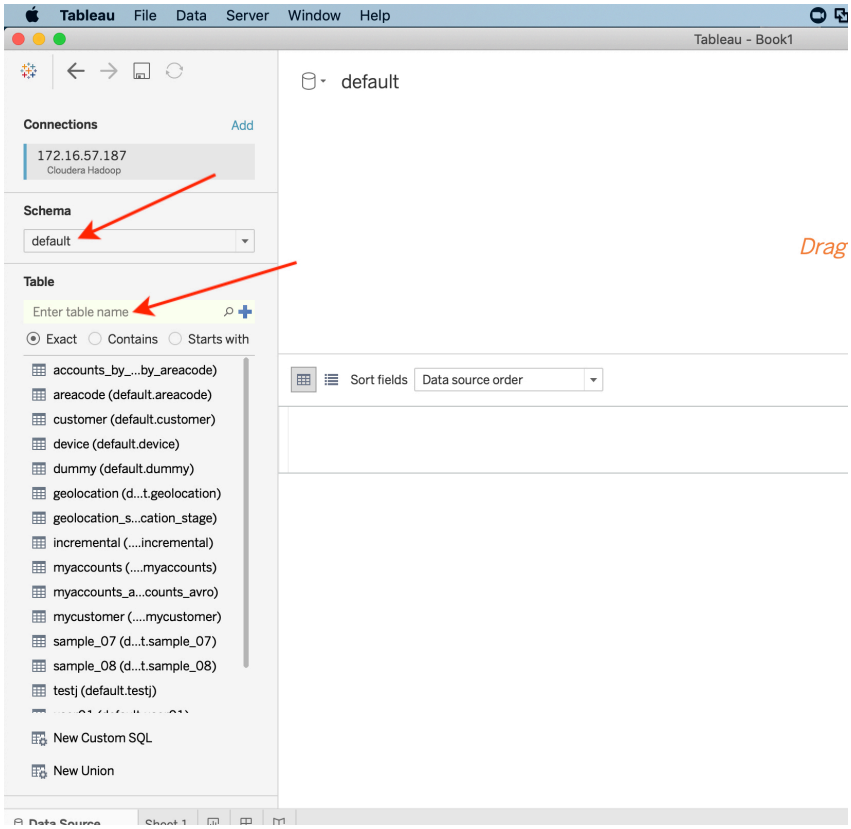
Tableau File Data Server Help
Tableau – Book1
Connect
To a File
Microsoft Excel
Text file
JSON file
PDF file
Spatial file
Statistical file
More...
To a Server
Tableau Server
MySQL
Oracle
Amazon Redshift
Cloudera Hadoop
More...
Saved Data Sources
Sample - Superstore
World Indicators
Open
Cloudera Hadoop
Server: 172.16.57.186
Port: 21050
Enter information to sign in to the server:
Type: Impala
Authentication: No Authentication
Transport: Binary
Require SSL
Initial SQL...
Sign In
Sample Workbooks
Superstore
Regional
World Indicators



Place the cursor at the Schema Bar and hit Enter.



Please the cursor at the Table bar and hit Enter. Then you will start getting the Hive/Impala tables.



Exercise 4

For this project you will take on the role of a fleet manager for Az National Trucking (ANT), a fictional national trucking corporation headquartered in California. In your role, you are in charge of ensuring that all the drivers in the fleet are in compliance with the rules and regulations of the corporation and do not present an insurance risk due to factors including speeding, unsafe following, lane departure, and other unsafe driving practices. The company details and context scenario are described below.

The Company

ANT is a 10 year old organization with 400 employees, most of whom are long-distance truck drivers. ANT's primary business is to provide long distance trucking services of all general and non-specialized cargo types (excluding any HAZMAT cargo) in the western United States ANT has licensure and approval for transport in 14 states. General organizational details are provided below.

Organizational Structure: ANT is a privately held organization.

Overall organizational details:

- 400 Employees
    - 40 - Staff, 15 - Mechanics, 20 - Administration, 15 - Management
    - 310 Drivers
        - 300 -W2 employees
        - 10 - 1099 contractors
- States and Countries approved for trucking licensure
    - States: AZ, CA, CO, NV, UT, WA, OR, ND, SD, KS, MO, NE, NM, WY
    - Countries and Provinces:
        - Canada: BC, Alberta, Saskatchewan
        - Mexico: Baja only
- Hubs
    - HQ= City of Industry, California
    - Secondary hub locations
        - Salt Lake, UT
        - Seattle, WA
        - Fargo, ND
        - Kansas City, MO
- Trucks
    - 300- ANT Owned
    - 10- Individually contracted

Compliance guidelines

The organization's compliance and risk management abides by the FMCSA regulations for the trucking industry at no less than the minimum identified in the applicable sections, and by the Department of Transportation regulations, and any applicable State specific laws and regulations. However, the organization reserves the right in the future to apply more stringent rules surrounding certain risk factors as may be warranted as determined by telematics and/or business needs, and allowable by law.

Risk thresholds

The thresholds for risk mitigation are determined by the requirements of the governing bodies for our industry combined with reasonable business needs and observed telematics data.

The following is a partial list of the organization's current risk factors:

- Miles driven
- Speed
- Individual State highway law compliance
- Number of trailers towed
- Freight weight (tonnage)
- Driver logs
- Cargo limitations – any general and non-specialized cargo is transportable by ANT. Exception: HAZMAT(e.g. BioHazard, radio active, explosives, weapons, etc.) which is specifically excluded as transportable by ANT.
- Off hours, as per section 395.3 of the FMCSA regulations
- Drug test results as per section 382 of the FMCSA regulations
- Driving record per section 391.27 of the FMSCA regulations (e.g. DUIs, tickets, reported accidents)
- Recorded telematics sensor violations, e.g. unsafe following, speeding, lane departure.
- Road closures, rerouting
- Vehicle maintenance and inspection
- Driver qualifications per section 391 of the FMCSA regulations

The Truck Driver Risk Factor Cases:

In this project you will be using the Risk factor threshold "Mileage / Speed" to determine the risk factor for each driver so that if it is exceeded or equal to 7.0 (on scale of 10) an alert or a system notification would be triggered and sent to the ANT corporate management and the insurance. The Risk Factor for a driver can be calculated by dividing the "totalmiles" / "riskfacor" (see tab "riskfactor" in the Geolocation excel file provided*).

The Task

Your task as a Fleet Manager, is to highlight the drivers who has the (mileage/riskfactor) greater than or equal to 7.0.

As an example, the below two cases have been generated out of Microsoft Excel Power Map component (Add-on). You can process the input files provided and generate the same report output in Tableau.

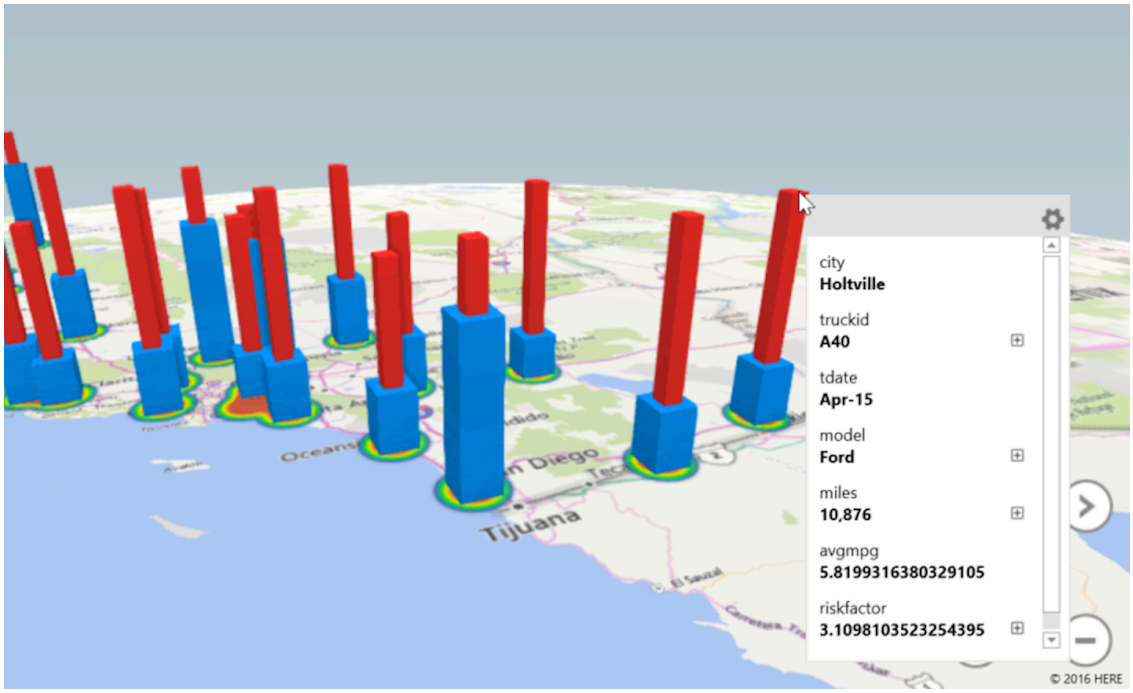
Example: From the below map visualization, the following two stories can be concluded:

The Truck story # 1

Truck id # A88 – (Hino model) drove a total of 9,653 miles in the city of San Diego in the month of April of 2015 with an average of 5.4 mpg and a risk factor of 7.51 (on a scale of 1-10).

The Truck Story # 2

Truck id # A40 – (Ford model) drove a total of 10,876 miles in the city of Holtville in the month of April, 2015 with an average of 5.8 mpg and a risk factor of 3.1 (on a scale of 1-10).



The product (Your deliverable)

1) You will need to discuss and analyse the data provided with your team members and be able to create the charts that highlight the riskfactor for each driver.
2) Create a power point for your analysis to showcase the riskfactors
3) The power point should have only three slides including:
    a) Problem statements and Objectives
    b) The analysis' workflow diagram. You may need to create a process flow diagram to show the components and the Echo systems you have used during the analysis from end to end. (Keep it simple)
    c) Dashboard or report chart that shows the Riskfactor per driver and be able to highlight those who has exceeded the threshold.

Note:

The above two stories (Truck story #1 and #2) are just an examples. They have been created in Excel with Power-Map add-on and you don’t have to create the same. It is just for you to visualize one option of the solution.