# **Open Source Summit 2016 – Industrial Automotive Scenario**

**Overview**

The purpose of this demonstration is to show how a data pipeline can be built using Bluemix to discover insights hidden in big data. Your demonstration must highlight how Bluemix tools and technologies can provide a solution from raw source data to a consumable format (reports, visualizations, APIs, etc.). Watson Analytics can also be leveraged for additional exploration, predictions and dashboard visualization.

The demonstration can be as simple as a report or something more sophisticated such as creation of APIs. It should be as interactive as possible and have a user interface that a non-technical person can understand and appreciate.

**Sales Scenario**

Ford Motor Company puts safety first! Initially, they would like to understand vehicular fatalities that occurred in 2014. If a reasonable solution is demonstrated, they will extend the analysis at a later time to include data that is available since 1975 from the NHTSA (National Highway Transportation Safety Administration).

Ford has asked you to provide an architecture, pipeline, analysis and output demonstration that provides insights into this data. They are interested in comparisons with other vehicle makes, years, vehicle weights, etc. They are also interested in understanding if other factors play a significant role, such as road classification, speeding, impaired driving, driver height and weight, weather, light conditions, etc.

The overall main objective is to help Ford:

* Focus less on gathering, cleansing and organizing the data
* Provide a platform where multiple tools and approaches are available to analyze the data for 2014, but also capable of extending into the past
* Produce an executive summary and consumption of insights to a broad audience.

**Pre-Requisites**

1. You should already have an account for Bluemix, Watson Analytics and Data Science Experience.
2. Limited weather data is provided in these NHTSA reports, so you might consider enhancing this with historical weather data provided by The Weather Company. If so, an API key will need to be requested and may take a few days. This request can be made at <https://goo.gl/oPVzPa>
3. You may also wish to have a Compose trial account available for MongoDB, PostgreSQL and/or Elasticsearch.

**Demo Expectations**

The following must be accomplished in the demonstration:

1. Ingest NHTSA Fatality Analysis Reporting System data for the year 2014
2. Prepare data and persist to a target of your choosing but some suggestions include:
   1. Cloudant
   2. DashDB
   3. Object Storage
   4. Compose MongoDB
   5. Compose Elasticsearch
   6. BigInsights on Cloud
3. Create visualizations for the results. Watson Analytics can also be leveraged for additional exploration, predictions and dashboard visualization.
4. Stretch Goals:
   1. Enrich the limited weather data provided in this data with weather data provided by The Weather Company or provide weather data for those incidents that have weather reported as ‘Not Reported’ or ‘Unknown’
   2. Use Spark machine learning or RStudio to provide predictions
   3. Incorporate data-in-motion streaming

**Data Details**

The data to be used will be the NHTSA Fatality Analysis Reporting System and provides raw data from 1975 in .dbf format. This data is located at <http://www.nhtsa.gov/FARS>

For 2014, you will find a FARS2014.zip files that contains 19 different files with much data. Complete detailed documentation can be found in the 560 page document, ‘FARS Analytical User’s Manual’ located at  
<ftp://ftp.nhtsa.dot.gov/fars/FARS-DOC/USERGUIDE-2014.pdf>

It is recommended that you only concentrate on 2 of the 19 files, the ‘vehicle.dbf’ and the ‘accident.dbf’.

Detailed documentation on the ‘vehicle.dbf’ can be found starting on page 13 of the ‘FARS Analytical User’s Manual’. This starts with a list of the columns and provides additional details for each column later in the document. For example, the vehicle make codes and names are provided starting on page 91 for 1991-Later. Other data in this file includes:

1. State of occurrence
2. State case number
3. Day, month, hour and minute of the crash
4. Vehicle make, year, weight
5. Road classification
6. Speed traveled and speed limit
7. Drunk driving related
8. Driver height and weight
9. And many others

Detailed documentation on the ‘accident.dbf’ can be found starting on page 11 of the ‘FARS Analytical User’s Manual’. This starts with a list of the columns and provides additional details for each column later in the document. For example, the weather codes and descriptions are provided starting on page 59. Other data in this file include:

1. State of occurrence
2. State case number
3. County and city of crash in GLC code
4. Latitude and longitude of crash site
5. Light conditions
6. Weather conditions
7. And many others

This data contains a wealth of information but we suggest that you only work with the tables and columns aforementioned.

**Helpful Links**

There are many artifacts that may help you in this scenario located here: <http://github.com/dxkikuchi/auto>

There are snippets of Spark code that may help you located here: <http://github.com/dxkikuchi/SparkSnips>