SAND2016-3592C

NLIT Summit 2016

National Laboratories Information Technology

May 1-4, 2016 • Albuquerque Convention Center • Albuquerque, NM

Tableau Visualizations Leveraging Predictive Analytics on Virtualized Structured and Unstructured Data

Presenters: Ron Hooks
Jeremy Myers

Technical Contributor: Brian Anderson

Sandia National Laboratories

Data Sciences

Sandia National Laboratories

Data Sciences Department 9534

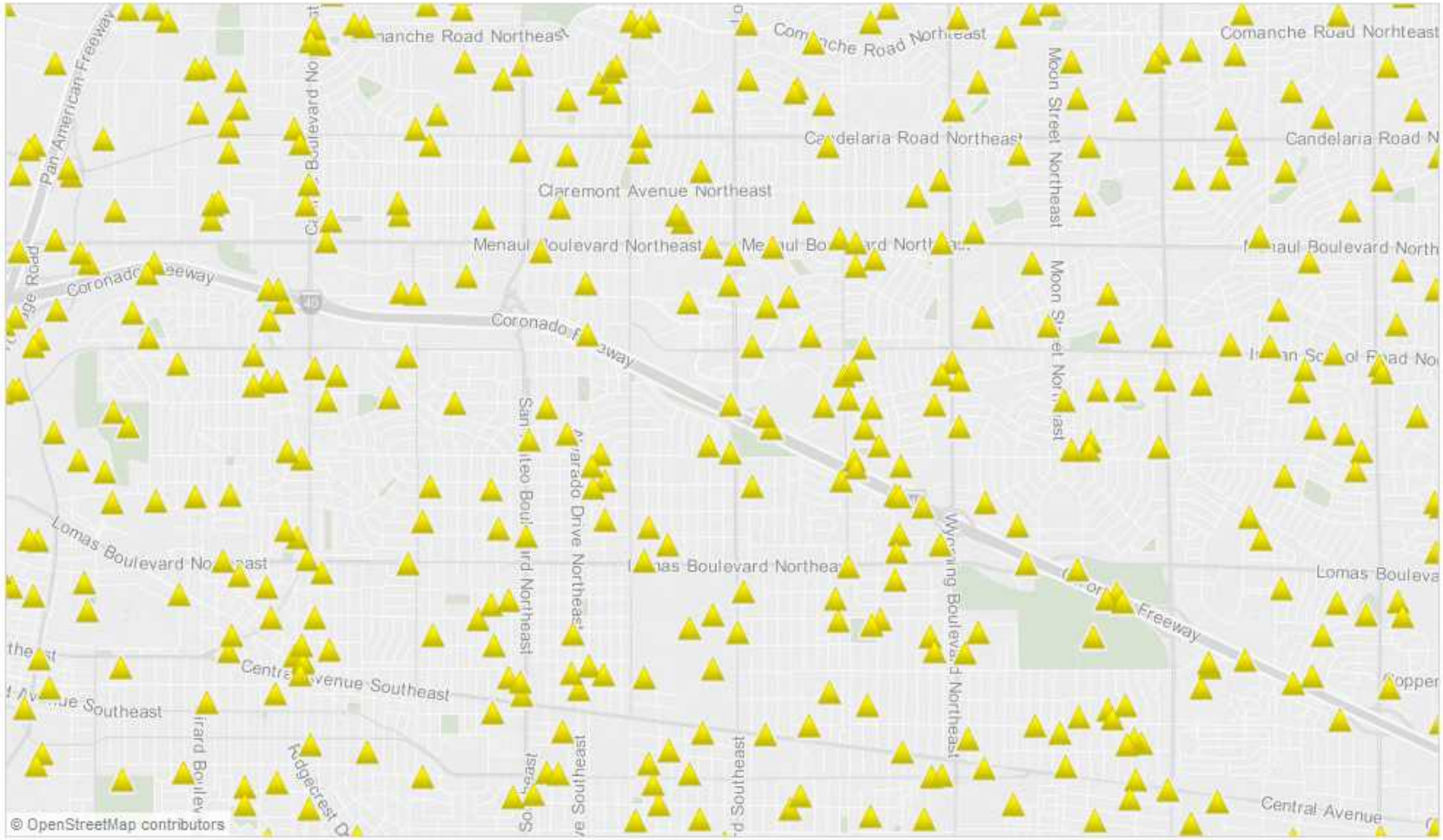
- Data Visualizations (Tableau, SSRS, D3, Custom)
- Data Virtualization (Cisco Data Virtualization Platform)
- Data Transformations (ODI, Pentaho, SSIS)
- Application Development (Wildfly, JavaScript, Spring, Angular JS)
- Predictive Analytics (R)
- Machine Learning (Weka)

Show Interactive Tableau Visualization At this Time

Albuquerque, NM
79°F
Humidity: 12%
Wind: 11 mph

Data Sciences

Based on current weather, there are **2** accidents predicted for today.



Data Sources

Weather Data from Weather Underground

https://www.wunderground.com/history/airport/KABQ

- Weather Data for 2011-2012 in CSV virtualized with Cisco

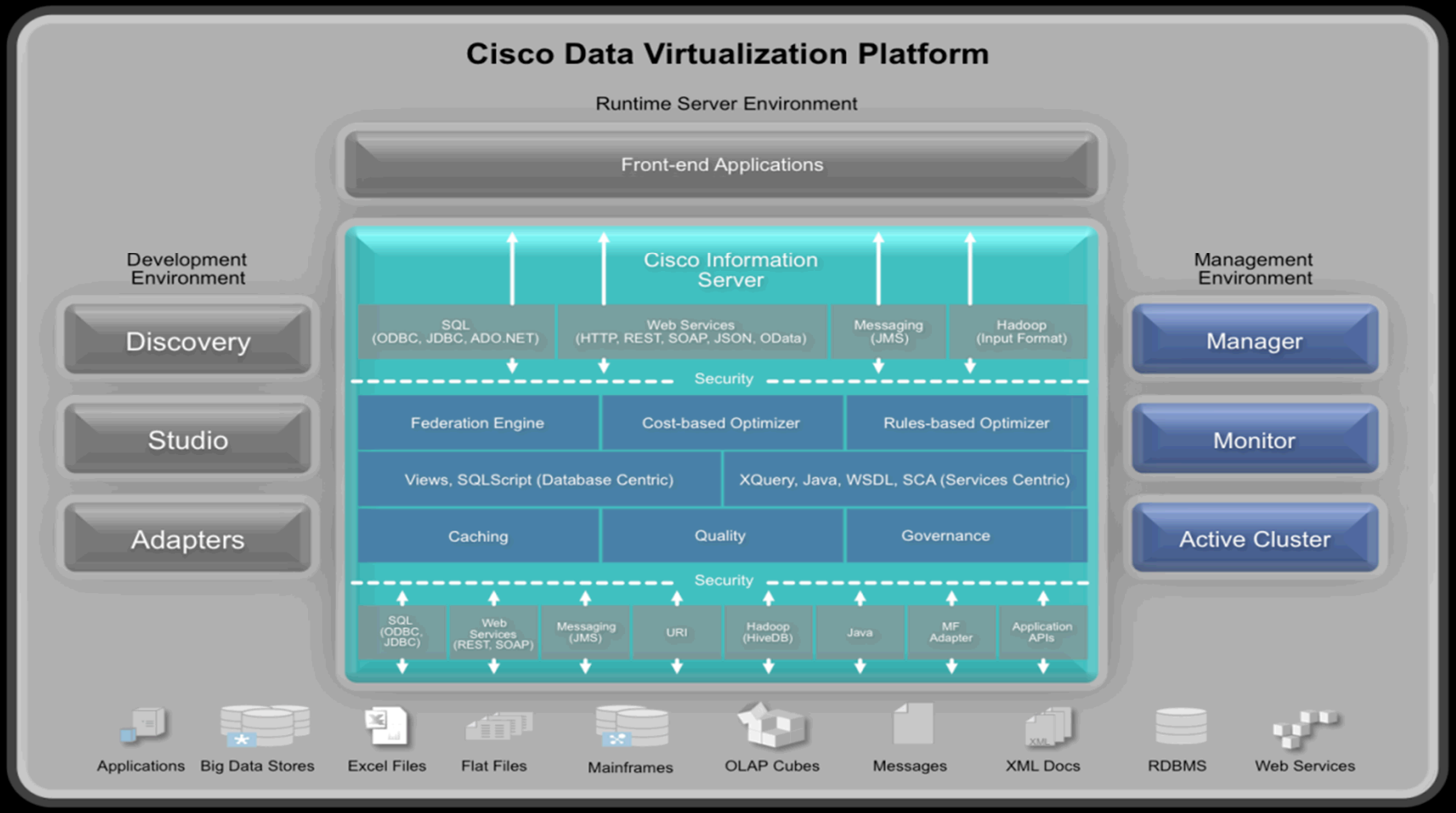
Injury Information from Department of Labor

http://developer.dol.gov/

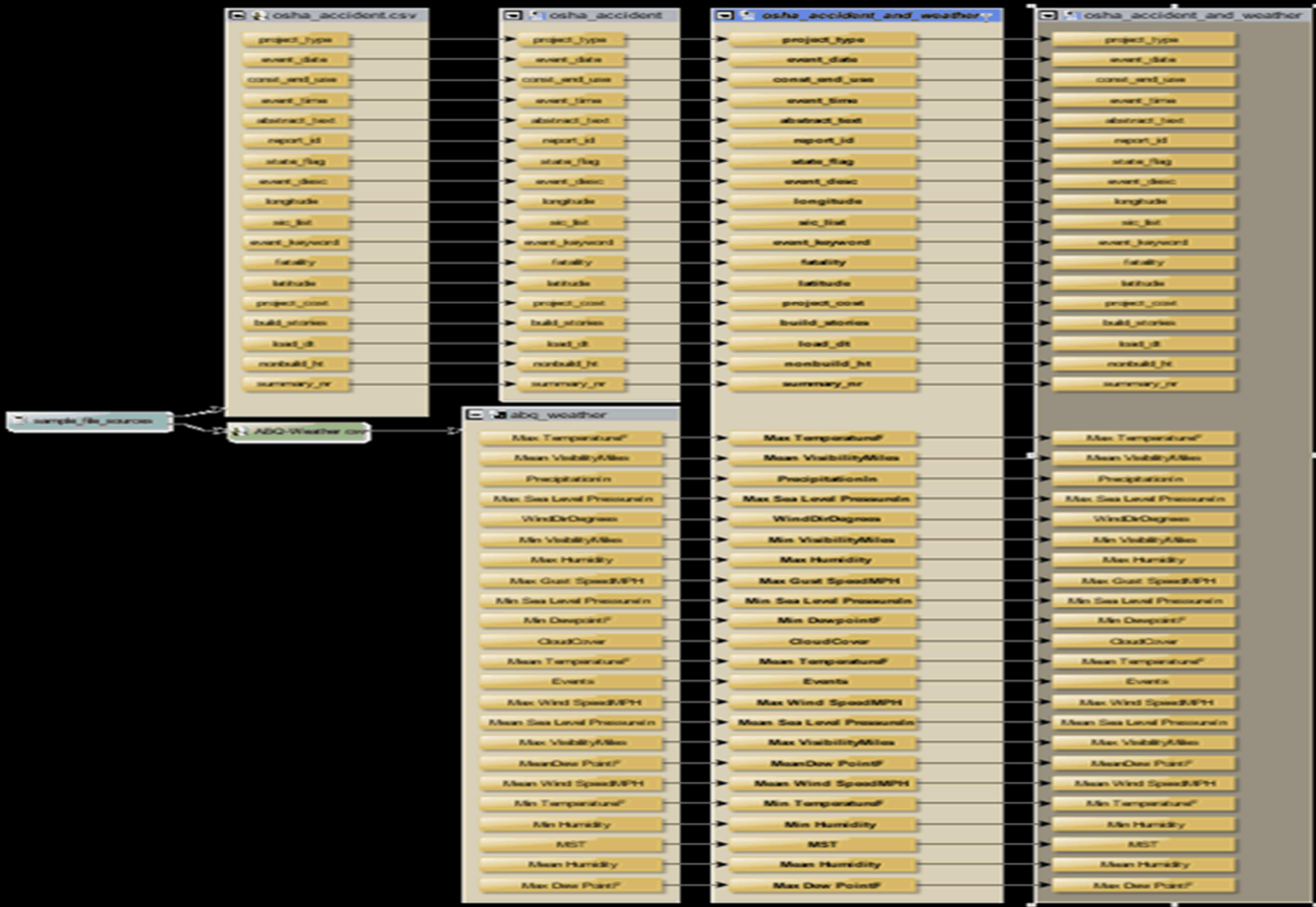
- Could not access Albuquerque specific data so created data based off National accident information for 2011-2012
- Latitude and Longitude were added programmatically
- Accident Information obtained in CSV virtualized with Cisco
- Accident Abstract Information for 2011-2012 in CSV loaded into MongoDB and virtualized with Cisco

Why Use Data Virtualization?

- For any application usage where data needs to quickly and easily be served up as REST, SOAP, or as a database procedure or view
- As a virtual data warehouse (data warehouse of data warehouses) combining disparate data sources able to be joined and transformed via an SQL paradigm
- To support a variety of analytic tools that consume data differently (IE Tableau favors data objects while D3 only supports web services)
- Tableau provides a built-in virtualization mechanism of sorts. However, it is limited. Disparate data sources must be blended not joined. Blending limits many functions, does not provide every join type.


Cisco Data Virtualization Platform
Runtime Server Environment
Front-end Applications
Development Environment
Discovery
Studio
Adapters
Cisco Information Server
SQL (ODBC, JDBC, ADO.NET)
Web Services (HTTP, REST, SOAP, JSON, OData)
Messaging (JMS)
Hadoop (Input Format)
Security
Federation Engine
Cost-based Optimizer
Rules-based Optimizer
Views, SQLScript (Database Centric)
XQuery, Java, WSDL, SCA (Services Centric)
Caching
Quality
Governance
Security
SQL (ODBC, JDBC)
Web Services (REST, SOAP)
Messaging (JMS)
URI
Hadoop (HiveDB)
Java
MF Adapter
Application APIs
Management Environment
Manager
Monitor
Active Cluster
Applications
Big Data Stores
Excel Files
Flat Files
Mainframes
OLAP Cubes
Messages
XML Docs
RDBMS
Web Services

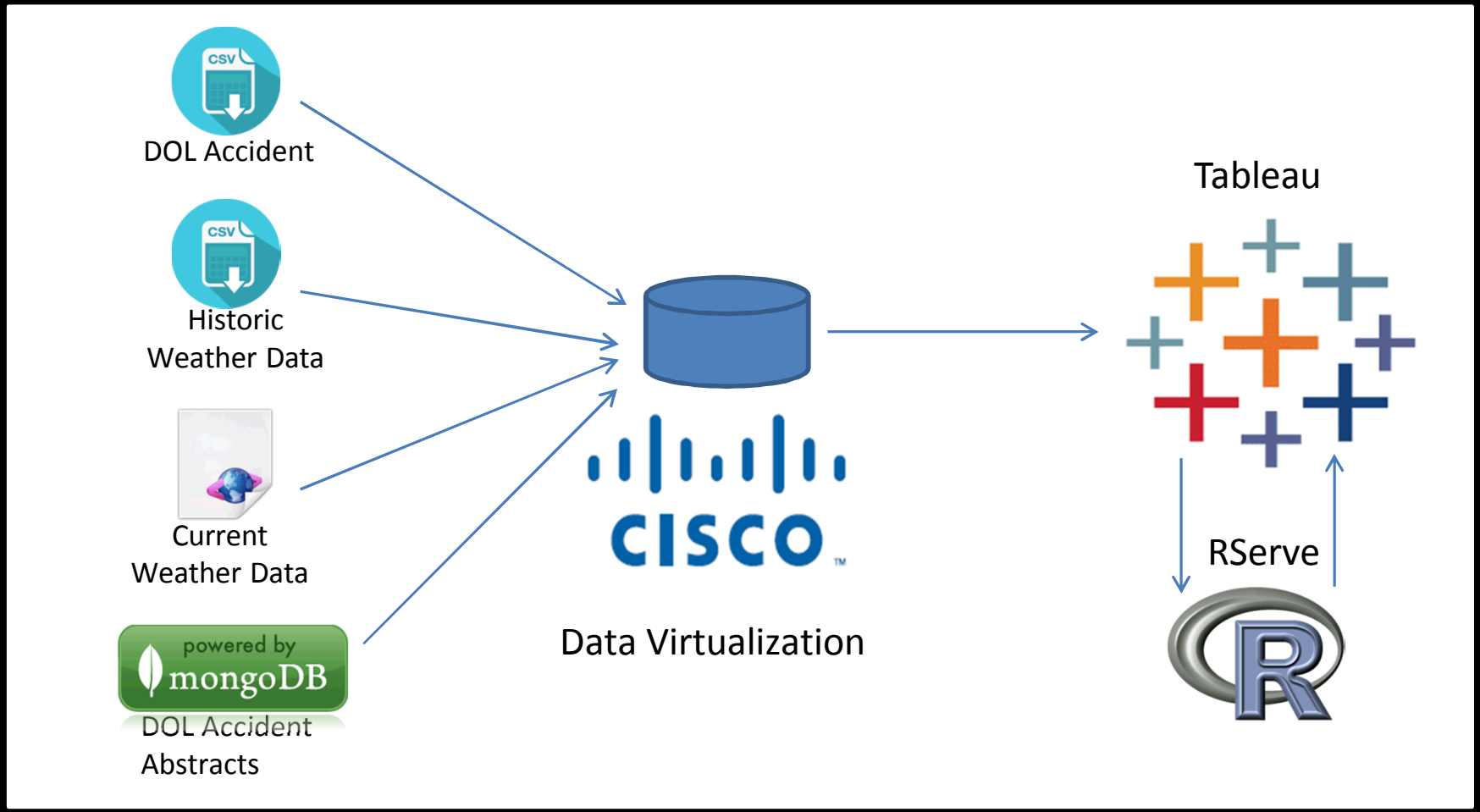
Cisco Data Virtualization Lineage:



osha_accident_and_weather:

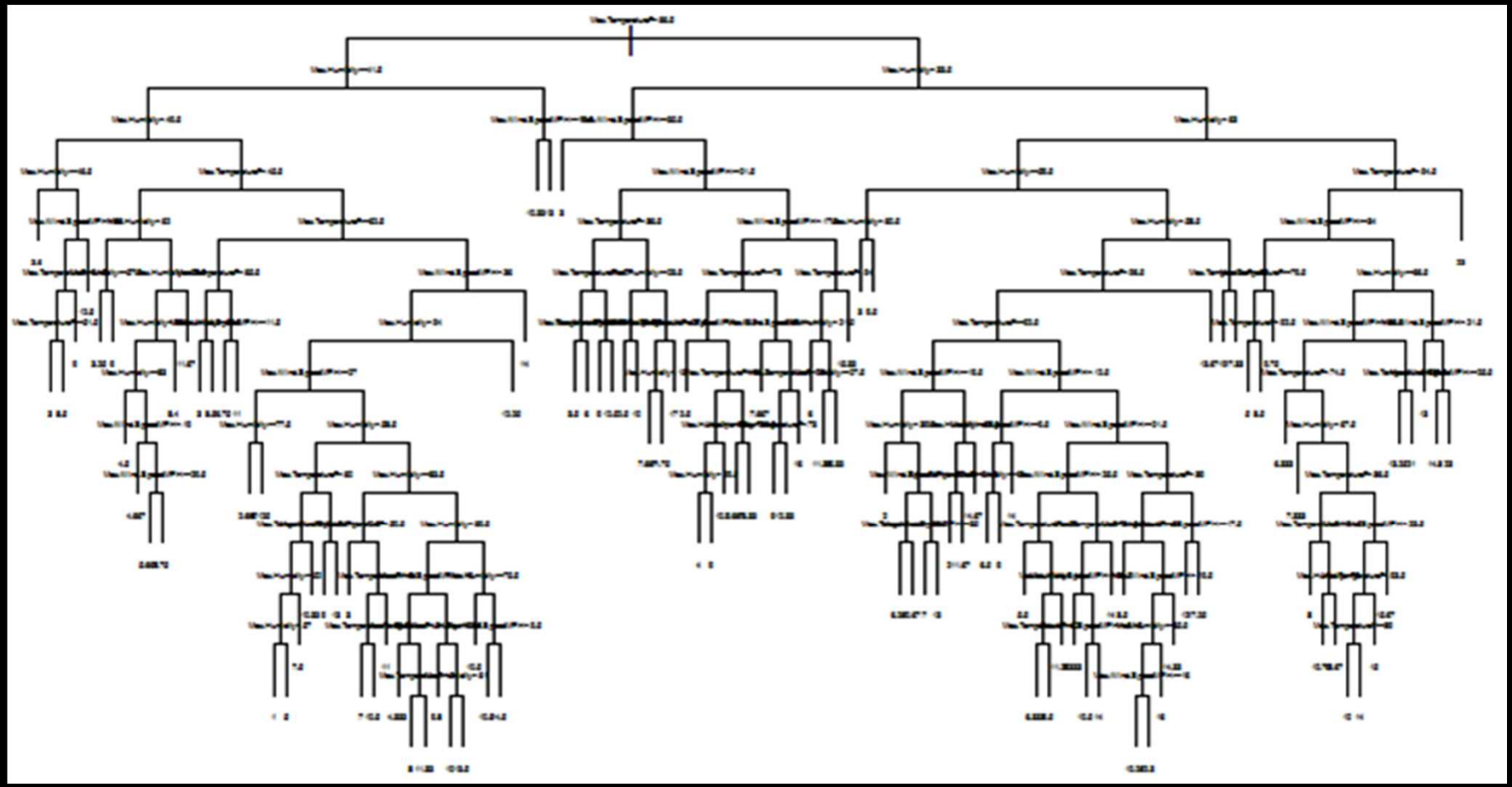
```
SELECT
*
FROM
/shared/NLIT/Formatting/abq_weather
ABQ_Weather_rest
        JOIN
/shared/NLIT/Formatting/osha_accident
osha_accident_csv
        ON
        ABQ_Weather_rest.MST =
osha_accident_csv.event_date
```

Integration of Data by Cisco Data Virtualization:

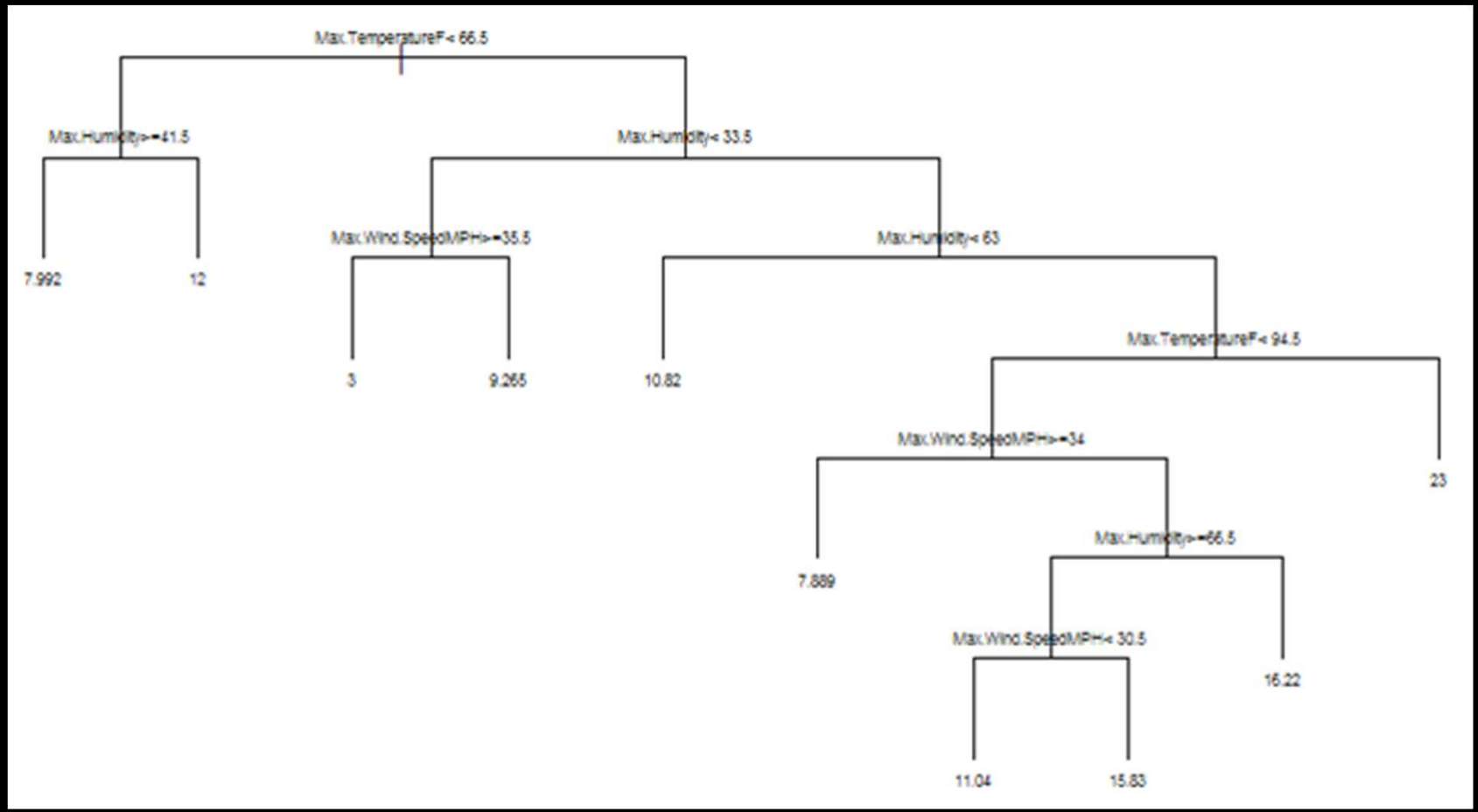


- Used regression tree (rpart) to output the number of accident using the inputs of temperature, wind speed and humidity to build the model.
  - Selected rpart for recursive partitioning and probability matrix for accidents.
- Using generic predict function for rpart class.

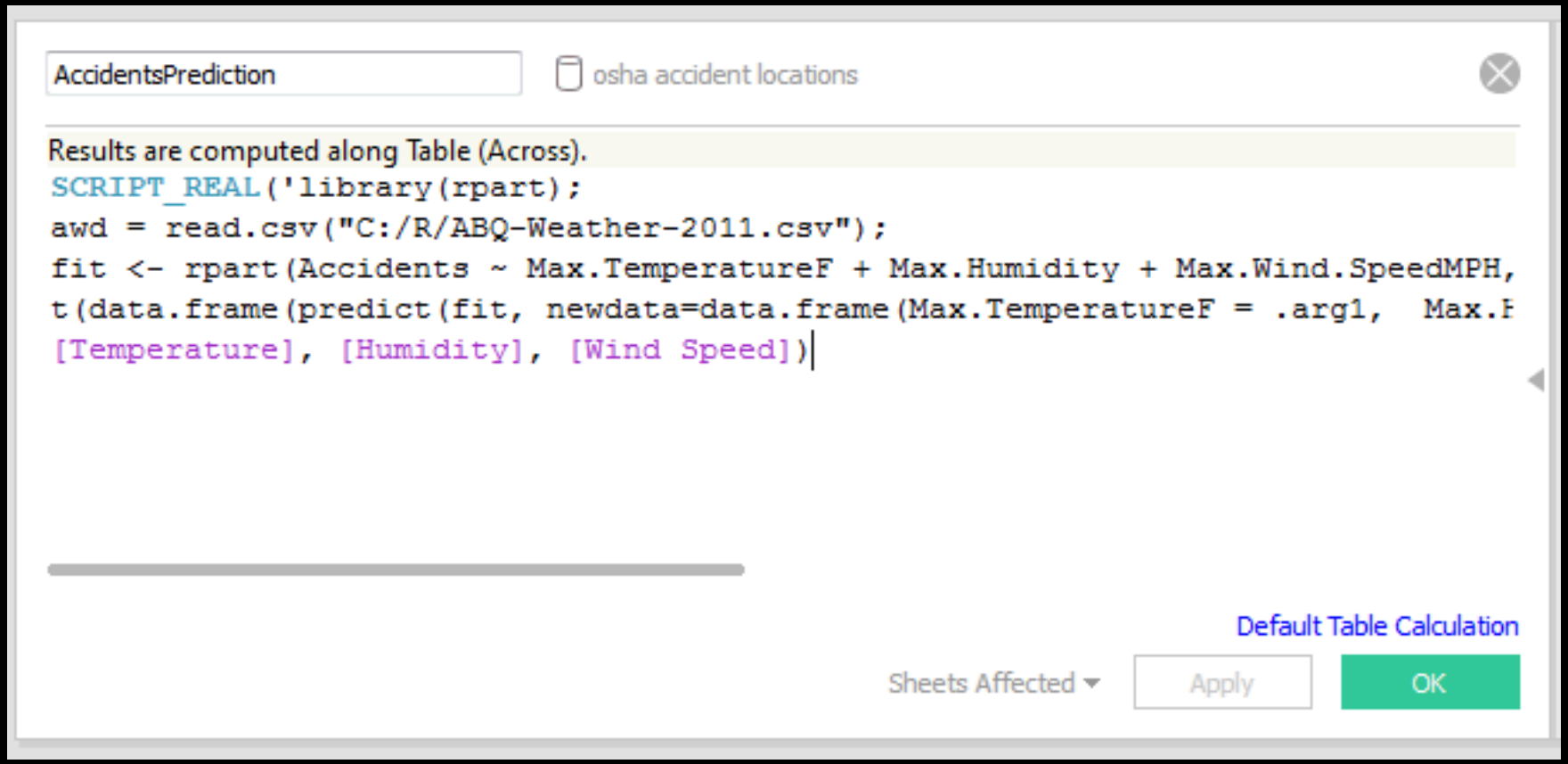
Unpruned Regression Tree:



Pruned Regression Tree:



Integrating R and Tableau



R Script in Tableau

```
SCRIPT_REAL('
    library(rpart);

    awd = read.csv("C:/R/ABQ-Weather.csv");

    fit <- rpart(Accidents ~ Max.TemperatureF + Max.Humidity + Max.Wind.SpeedMPH,
    data=awd));

    t(data.frame(predict(fit, newdata=data.frame(Max.TemperatureF = .arg1,  Max.Humidity =
    .arg2, Max.Wind.SpeedMPH = .arg3))))[1,];',

    [Temperature], [Humidity], [Wind Speed]
)
```

Conclusion

- Virtual Database Environments such as Cisco or Denodo greatly aid in joining disparate data sources (structured, unstructured, Cubes, Big Data etc.) into the SQL paradigm for analytic consumption
- Powerful Predictive Analytics can be calculated and visualized by leveraging the integration of R and Tableau

NLIT Summit 2016

National Laboratories Information Technology

May 1-4, 2016 • Albuquerque Convention Center • Albuquerque, NM