Big Data Driven Transportation Analytics

The TMW Systems architecture for concentrating and unifying operational data to deliver actionable insight at both the operational and strategic levels for individual customers and the transportation industry.
Big Data Driven Transportation Analytics

Authors
Timothy Leonard – EVP Operations and Technology
Piyush Kumar – Operations Research Specialist
Steve Pembridge – Data Architect
Donnie Wheat – Big Data Architect

Contents
3 Executive Summary
4 What is Data-Driven Decision Making?
5 Guiding Principles for Building a D^3M Enterprise
6 Technological Challenges Facing the Transportation Industry
6 Alignment between Business Requirements and Information Systems and Services
8 Technology Infrastructure to Meet the Guiding Principles
8 Hadoop Framework
11 Enterprise Architecture
14 Big Data ELT for Data Warehousing in the Cloud
15 Applying Data Science to the TMW Transportation Data Cloud
16 Principles in Building Data Science Capabilities
16 Statistical Analysis of Transportation Data
18 Data Quality Case Study: Market Rate Index
22 Predictive Analysis
23 Prescriptive Analytics

© 2017 TMW Systems, Inc. | A Trimble Company. All Rights Reserved.
Executive Summary

As with companies in nearly every industry, transportation enterprises are challenged to make good decisions quickly, providing competitive advantage through innovation and quality-of-service. Making these decisions requires intuition, information and agility; i.e., the ability to move in the right direction, right now.

The cornerstone of this triad is information. Intuition is important for deciding among alternatives, but intuition without information is guesswork. Agility is a key aspect of success in the modern economy, but agility without information is blind churning. It is information that binds intuition and agility, making possible those good, quick decisions that drive you in the right direction.

In many respects this challenge is a more difficult for the transportation industry than it is for many others. While there are certainly major players in transportation, MCMIS (Motor Carrier Management Information Systems) data suggest that 91% of all trucking companies operate fleets of 6 or fewer trucks, and 97% operate 20 or fewer. As a result of this extreme fragmentation, getting solid, timely information to serve as the basis for business decisions is very difficult for many companies.

<table>
<thead>
<tr>
<th>Carrier Size In Trucks</th>
<th>Industry %</th>
<th>Industry Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 6</td>
<td>91.3</td>
<td>91.3</td>
</tr>
<tr>
<td>7 - 20</td>
<td>6.4</td>
<td>97.7</td>
</tr>
<tr>
<td>21 - 50</td>
<td>1.5</td>
<td>99.2</td>
</tr>
<tr>
<td>51 - 100</td>
<td>0.5</td>
<td>99.7</td>
</tr>
<tr>
<td>101+</td>
<td>0.3</td>
<td>100.0</td>
</tr>
</tbody>
</table>

The TMW Systems Vertical Data Domain is designed to span this fragmented reality. TMW’s goal is to provide a foundation of flexible, integrated information systems and services necessary to:

- Deliver actionable data to the business, across the entire organizational hierarchy, enabling quick, nimble and well-informed (data-driven) decisions right at the point of impact;
- Evolve the systems and services to maintain alignment with rapidly changing business circumstances and industry requirements; and
- Deliver industry level analytics capabilities to measure and benchmark transportation customers against industry trends, and provide the data necessary to make decisions on corporate direction in a constantly changing industry.
What is Data-Driven Decision Making?

Data-driven decision making (D³M) involves listening to what the Six Sigma folks call the “voice of the process”, or paying attention to the actual data that emerges from your day-to-day execution of business processes. It means bringing the objective, measureable outcomes of your performance into the middle of your decision making so it can inform, guide and shape your decisions.

For many of us this involves a significant change in business culture. Traditionally, we have made decisions based primarily on our experience and the sharpened intuition that experience provides.

Figure 1: Typical Decision Making Process

Shifting to a D³M approach involves putting data ahead of intuition and experience. The latter are still crucial, but we put them to work in the context of an objective set of measures and insights that focus our attention and our efforts.
Changing the culture of a business is difficult, requiring sustained effort. Moreover, leadership must “walk the talk” to be successful. Some of what you can expect from a transition to a D3M approach includes:

- A definable process for approaching and making business decisions;
- The ability to trace the lineage of a decision back to the data that engendered it;
- Increased accountability among your staff (the data is what it is; there is no place to hide);
- Increased emphasis on the quality, completeness and timeliness of your data; and
- Support for a consistent and sustainable “Evaluate – Decide – Do – Measure – Evaluate” cycle of operational and strategic innovation.

**Guiding Principles for Building a D³M Enterprise**

Several principles guiding the development of a platform to enable the technology transition needed to support D³M.

- Data is a highly valuable strategic asset for the business
  - The foundation for innovation and competitive advantage
  - Comes in all shapes, sizes and types, requiring consolidation from across the enterprise and organization by subject area to be leveraged
  - Must be governed, managed and secured
Data should be decoupled from individual users
- Owned, governed, managed and secured by the enterprise
- Served to users and other consumers via defined data services

Binding Data to Business Processes, Rules, Terminology and Models should be done later rather than earlier
- Collect and consolidate all data—structured, semi-structured and unstructured—in a common, informal format with no or few presuppositions
- Model data that supports analytics in consistent and stable use cases
- Relate (late binding) data on demand for maximum agility in meeting business needs and integration requirements

Technological Challenges Facing the Transportation Industry

The technological challenges facing the Transportation industry, particularly the carrier segment are often a result of the industry’s tremendous diversity

- Complex silos of data for Transportation Management, Asset Maintenance, Mobile Communications, and Regulatory Compliance.
- Demand for visibility to freight under management and the ability to understand, interpret, and have a unified view of their company, customers, drivers and freight.
- High cost of software and hardware solutions to integrate and analyze data from multiple data sources.
- Visibility into industry trends and direction driven by data.

Alignment between Business Requirements and Information Systems and Services

The core mission of the TMW Business Intelligence approach is to meet the challenges facing the Transportation Industry today and in the future. The company’s Business Intelligence and Value Engineering teams utilize surveys, data-driven benchmarks, and TMW customer engagements to identify the challenges and the data required to measure, monitor and analyze these challenges in order to provide the tools needed to overcome them. TMW has worked with the industry to identify the 5 key challenges facing the industry, and how they affect a company’s key metric, Operating Ratio.
**Business Objectives**

![Diagram of Transportation Challenges](image)

1. **Resource Velocity** – Profitable transportation companies are those that focus on Dollars per Truck per Week. They focus on having drivers and tractors available to move freight in the face of driver shortages and complex regulatory requirements.

2. **Driver Churn Mitigation** – With an ever increasing demand for qualified drivers, transportation companies need the Business Intelligence tools to see which drivers are getting miles, getting paid, and being successful to retain qualified drivers.

3. **Maintenance Expense Mitigation** – Efficient maintenance and repairs enable companies to keep tractors on the road and enhance resource utilization while being able to monitor excess costs and repairs.

4. **Fuel Expense Mitigation** – Helping to reduce the second largest expense that faces a carrier under all fuel market conditions/pricing levels.

5. **Freight Network Management** – An effective freight network will allow carriers to keep trucks moving with freight on board, to and from profitable geographies.

The Business Intelligence team worked side by side with the Value Engineering team to develop a data model for measuring the challenges facing carriers. The Transportation Data Model identifies measurable business process, which may cross multiple data sources, to provide decision makers with the data to measure and improve processes. TMW is able to bridge the gap between data sources, and merge data from different systems to provide a single view of real-world, physical entities such as tractors and trailers. By moving the Business Intelligence application to the cloud, TMW has enabled its customers to leverage high computing resource utilization to reduce costs. This approach also has empowered TMW’s Business Intelligence and Value Engineering teams to more quickly deliver new content and solve emerging business problems.
Technology Infrastructure to Meet the Guiding Principles

TMW Reveal has proven to be a valuable tool in aiding customer decision-making and analytics. However, with the original on-premise solution came additional hardware and database requirements and related resources. By moving to a SaaS model, TMW reduced implementation time and changed the conversation from implementation and IT infrastructure to how to use the tools and analyze data and business problems. The SaaS delivery model offers several important advantages:

1. **Minimize TMW Customer Hardware and Software Cost**
   Users of TMW Reveal are not required to have local database servers and minimize ETL machines. The new SaaS application requires only a small extract client to be installed and running.

2. **Minimize Implementation Time**
   The Implementation team has reduced the complexity of the implementation process, and now moves most implementation activities to the SaaS environment with new tools to monitor, manage and configure.

3. **Simplify Content Delivery**
   By centralizing content and content delivery, new Business Intelligence data is made available to all SaaS customers seamlessly through upgrades managed by TMW.

4. **Support New Varieties of Data**
   By moving to a Hadoop infrastructure, TMW now has the tools to consume and analyze data beyond that of traditional database systems. This will enable new statistical and machine learning processes to solve new problems and identify new data trends and patterns.

**Hadoop Framework**

TMW has designed and implemented an architecture built on industry tools and best practices to ingest, secure, clean and analyze data for individual TMW customer needs and industry data-driven benchmarks. At the heart of the system is the Apache Hadoop framework for distributed computing. This framework allows TMW to horizontally scale data processing for scalable growth. TMW is able to leverage multiple software applications within Hadoop for rapid development of content and business analytics.
HDFS – Hadoop Distributed File System

HDFS is a Java based file system that provides scalable and reliable data storage with the ability to span large clusters of servers. The storage provides availability and redundancy by distributing data across multiple machines for storage and parallel processing. TMW can collect and store customer data in its SaaS environment to reduce cost of hardware and processing, while providing the redundancy and power of cloud computing.

YARN and MapReduce2

YARN (Yet Another Resource Negotiator) is the central resource manager that manages and allocates all system resources including CPU and memory used across the Hadoop cluster. Combined with MapReduce2, a key and value pair parallel computing processing framework, TMW is able to distribute processing of data sets across a large computing cluster to analyze and benchmark the transportation industry.
**Hive**
Hive data warehouse software allows for reading, writing and managing data within the Hadoop environment, via the familiar SQL interface. While the SQL may be familiar, Hive provides schema on read capabilities to allow TMW to work with diverse sets of transportation data. This in turn, allows TMW customers using different versions of software to leverage the SaaS Business Intelligence offering. This familiar SQL interface enables the data integration and data science teams to perform the complex transformations required for the TMW Enterprise Data Warehouse, while providing new capabilities in benchmarking and data classification as well as access to data from advanced statistical analysis tools such as R.

**HBase**
The HBase NoSQL, distributed, big data store enables real-time read/write access to big data. HBase empowers TMW to look beyond the relational database world, and enables real-time performance in a data store that is able to elastically scale across a distributed cluster.

**TEZ**
TEZ is the application framework, which from a developer’s perspective, works like MapReduce, but improves performance of Hive code. This framework allows TMW to enable high-performance batch data processing.

**R**
The R statistical computing platform is used by the TMW Data Science Team to move beyond rules based benchmarking to statistical analysis in the Market Intelligence benchmark. R is able to statistically validate the tribal transportation knowledge gained by TMW through years of Market Rate Indexes, while allowing data scientists to take a statistical approach to data analysis.

**ELT on Hive**
For deployment of the SaaS Business Intelligence solution, TMW has taken an ELT (Extract, Load, and Transform) approach to data processing. The extract application that exists in the TMW customer IT infrastructure is a lightweight application to minimize the impact of collecting data and moving it to the secure, SaaS environment. The data is the loaded into the Hadoop cluster for processing, where all cleaning, classification and data transformations happen. In a centralized environment, new content and bug fixes occur seamlessly so clients receive updates in one location, applied to all TMW SaaS customers during regular updates.

**Ambari**
The Ambari management console allows a simple interface for the Hadoop admins to manage, maintain, and upgrade large numbers of servers with the click of a button. The interface also provides valuable monitoring software so engineers and admins can see applications running in real time and analyze performance of applications and machines.
**Hue**
The Hue application has allowed the TMW team to move to the cloud alongside our customers for application development. This interface provides the Data Integration team a simple interface to the complex Hadoop framework.

**Knox**
Apache Knox, combined with LDAP and Kerberos, enables a secure gateway to control Authentication, Authorization and Auditing in the Hadoop environment. This controlled access enables only authorized TMW personnel to access the Hadoop cluster.

**Ranger**
Ranger is a framework to enable, monitor and manage comprehensive data security across the Hadoop platform. The vision with Ranger is to provide comprehensive security across the Apache Hadoop ecosystem. With the advent of Apache YARN, the Hadoop platform can now support a true data lake architecture. Enterprises can potentially run multiple workloads, in a multi-tenant environment. Data security within Hadoop needs to evolve to support multiple use cases for data access, while providing a framework for central administration of security policies and monitoring of user access.

**Enterprise Architecture**
The TMW SaaS Business Intelligence architecture was built to provide TMW customers with secure and scalable access for decision makers without the upfront costs of the high-throughput servers required to run business intelligence applications. TMW has partnered with the Microsoft Azure and Hortonworks teams to ensure that it is taking the right approach in moving to the cloud and working with a Big Data platform. These resources and technical experts were able to help TMW’s Business Intelligence specialists in designing an architecture that is scalable, capable, and secure. The teams looked at industry best practices and applications to develop a secure infrastructure with Apache Knox, Apache Ranger, Kerberos, and the secure Azure environment to provide an enterprise architecture that would meet the needs of TMW customers.

With the requirement to provide a multi-tenant capable architecture to efficiently utilize computing resources for multiple TMW customers, the product was moved to an ELT data processing model. The ELT model moves the CPU cycles and memory requirements of business intelligence from the TMW customer premise to the cloud. This started with a complete re-envisioning of how TMW is able to efficiently extract data from a remote site and securely transfer that data to the hosting environment. Data access must be efficient to minimize impact. Transfer must be minimized to reduce bandwidth cost and secured to keep data private. The client application was designed with new capabilities to apply advance change data capture (CDC) techniques to find and deliver the smallest amount of data possible for processing. The loading of this data was accomplished with the tools available to Hadoop to minimize development time and ensure successful, repeatable data ingestion.
The transform logic, which is core to the TMW Enterprise Data Warehouse and crucial to the complex nature of enterprise data, was re-designed and implemented from the ground up. The Data Model, with years of transportation knowledge contributed by TMW and customers, remained the foundation of the Data Warehouse. However, Hadoop now empowered TMW data integrators to look at compute power and processing from a 21st Century perspective. The initial design started with a multi-tenant data model perspective. The Apache Ranger application, combined with Ranger and Knox, allowed the data design team to secure and segment data, while auditing and logging access to the data. From there data processing must be modularized and streamlined to process the variety of data. The current release contains modeling of data for .Suite (TMWSuite) and AMS (asset maintenance). The data model supported systems that had either or both of these products. The data integration team created a configurable and customizable pipe lining architecture to support wide varieties of source data. With this pipeline, the team is able to add new content (dimensions and facts) and seamlessly add the content to the process by configuring the new code in the existing process.

Figure 5: TMW ELT Architecture

- Minimal Client Side Install
- Light Weight CDC Client for minimal Database Impact
- Files compressed for minimum transfer impact.
- SSL Connection for Secure data transfer.
▪ Client Side – TMW Customers install a light weight extract client locally, and are able to access the Enterprise Data Warehouse via TMWBI.com.
▪ Benchmark data is also available on TMWBI.com.
▪ Loading data is only available via SSL (TLS) into a landing zone provided for each TMW customer. That data is then loaded into a segregated Hadoop Landing Zone.
▪ Data cleansing is applied to weed through the rough edges of data and provide an Enterprise Data Warehouse built on years of industry data analysis.
▪ The TMW transformation converts cleaned and normalized data into the TMW Enterprise Data Warehouse, using customizable parameters and TMW’s years of industry analysis and expertise.
▪ TMW Market Intelligence is able to anonymize and aggregate with our data share partners to produce an ever-growing set of transportation industry benchmarks for use by the industry.
▪ Once complete, the Enterprise Data Warehouse is loaded from the Hadoop System, and made available on TMWBI.com.

**Metadata Purpose – this is needed for control of data flow for keeping track of data points from end to end – data linage is important part of the Process Architecture**

The metadata tables that support the Data lake ETL process provide an augmentation to the metadata provided by ETL tool repositories and form the basis for enabling the following functionality within the ETL Architecture:

▪ **Stores the tables and fields being sourced** to provide a consolidated list that can be used as the cross-reference for user definitions or other metadata-related functionality.
▪ **Summarizes content analysis characteristics about the data being sourced** (e.g. number of rows per table, max/min/counts for selected fields, number of nulls for a selected field, domain distributions and observations)
▪ **Stores rules used to validate the source data** (works in conjunction with the content analysis statistics. These rules along with analysis against the contents forms the foundation for enabling the Data Stewardship function.)
▪ **Tracks the last extract date for each of the ETL processes to aid in delta processing of source records.**
Big Data ELT for Data Warehousing in the Cloud

TMW is able to leverage the capabilities and scale of Hadoop to enable Big Data ELT (Extract, Load and Transform) processing of multiple TMW customers into the TMW Transportation Data Model. TMW customers are provided a proprietary CDC (Change Data Capture) collection engine that is able to minimally impact production databases, and deliver data to the cloud for analytics.

Once data is moved to the cloud, TMW is able to run complex data translation processes and present the transportation data back to business decision makers. The TMW Data Warehouse is a transportation data model that allows users to slice and dice their business data with a dimensional data model. The data model is driven by business requirements to support the success and failure analysis of data driven decision making. Through dimensional data, the model is able to maintain historical master data and embrace a simplified data model to understand data from multiple systems. At the core of the model are fact tables that break down business processes into measurable facts, which are attached to dimensional master data. Dimensional data de-normalizes the properties of those facts, which the model allows for analyzing the business process.

Figure 6: Star Schema Data Model for Trip Activity
The initial Extract has the largest impact since it is required to collect. This requires the CDC client to capture the entirety of the database for analysis in the cloud. With the goal of minimizing equipment requirements and cost, the ELT application was able to reduce what could have been multiple hours of on-premise ETL to just a few minutes of load on a utility machine.

<table>
<thead>
<tr>
<th>Extract Type</th>
<th>Avg Size MB</th>
<th>Avg Table Count</th>
<th>Avg Record Count</th>
<th>Avg Extract Time Seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>421</td>
<td>59</td>
<td>12,303,054</td>
<td>671</td>
</tr>
<tr>
<td>Ongoing</td>
<td>9</td>
<td>59</td>
<td>494,343</td>
<td>8</td>
</tr>
</tbody>
</table>

Leveraging the Hadoop environment, the TMW ELT is able to use horizontally scalable machines to process large volumes of data into the TMW Data Model in a multi-tenant, parallel processing ELT application.

- **Average Dimensional Records Processed (Per Customer Daily):** 680,249
- **Average Fact Records Processed (Per Customer Daily):** 6,201,376
- **Total Records (Per Customer Daily):** 6,881,624

**Applying Data Science to the TMW Transportation Data Cloud**

Given the size and scale of the collective TMW customer base, as well as the capabilities of the TMW Transportation Cloud to collect and aggregate the data, TMW has engaged its data science team in implementing advance data science methodology to better understand the transportation industry. Currently, the data being captured is prone to a variety of issues due to incorrect/incomplete recording of events with human involvement. This necessitates an analysis of how to organize and clean the existing data before it can be used in any type of machine learning.

Trucking analysts within various companies have set their own parameters based on the data available to them which might not reflect the overall trucking environment. After data cleaning, unsupervised learning from the data helps validate/change the industry’s widespread assumptions. As the next step, supervised learning allows us to move away from descriptive analytics to prescriptive analytics, where instead of the user manually exploring the data and drawing conclusions which affect their businesses, statistical tools within the big data framework lead to real-time recommendations based on daily changes within the user’s operating conditions as well as the industry as a whole.
Principles in Building Data Science Capabilities

There are several principles guiding the development of a platform to enable cutting edge data science capabilities:

- Create easily readable, reusable scripts for data exploration using R.
- Use unsupervised learning techniques such as clustering analysis for validating various groupings of data such as mileage bands in conjunction with experience from industry experts.
- Use supervised learning to create predictions which would be valuable to TMW customers.
- Create a platform to calculate and deliver the predictions on data in near real time.
- Integrate both optimization and predictive analytics platforms to create a recommendation engine.

Statistical Analysis of Transportation Data

Analyzing data with respect to the trucking industry affords the following advantages for TMW customers:

1. Understanding the distribution of the industry is valuable in determining which direction in which to expand/contract a business. For example, the contour plot below describes the distribution of rate per mile vs. service miles charged on a load. As we can observe, the density of loads serviced is highest near the rate of $2 per mile and 500 service miles. Depending on which part of the plot constitutes the transportation company’s business, it may choose to move towards the denser part of the plot if the goal is to expand by soliciting more loads. This type of analysis is termed “descriptive analytics.”

![Contour Plot](image-url)
2. The next step is “predictive analytics.” The plot below is the output of a time series forecast using R and D3JS. It shows a series of forecasts and prediction intervals (80% to 90%) based on historical data only. It uses 6 years of historical data and returns a 52-week (1 year) period forecast series for the latest point of the time series. It also returns the historical forecast time series for the preceding 2 years, making it easier for the user to assess forecast accuracy.

![Forecast vs Actual Data](image)

**Forecast vs Actual Data**

<table>
<thead>
<tr>
<th>Date</th>
<th>Forecast Mean</th>
<th>Actual</th>
<th>Delta</th>
<th>Delta %</th>
<th>95% High</th>
<th>80% High</th>
<th>80% Low</th>
<th>95% Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 Jul 2016</td>
<td>544</td>
<td>512</td>
<td>32</td>
<td>6.27</td>
<td>619</td>
<td>593</td>
<td>497</td>
<td>473</td>
</tr>
<tr>
<td>7 Aug 2016</td>
<td>537</td>
<td>612</td>
<td>586</td>
<td>9.58</td>
<td>491</td>
<td>467</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 Aug</td>
<td>537</td>
<td>617</td>
<td>589</td>
<td>9.58</td>
<td>487</td>
<td>461</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Quality Case Study: Market Rate Index

The market rate index (MRI) data contains details pertaining to any loads serviced in North America, including but not limited to arrival and departure dates, pickup and delivery addresses, type of load, breakdown of revenue, etc.

The steps taken in the process of analyzing MRI data are as follows:

Data Organization

The data is first classified into four categories: strings, numerics, factors, and time. These categories are universal in nature and, thus, the first step towards analyzing the data. We add derivative columns, which illuminate details of interactions within the data. For example, the drop-off and pick-up dates in conjunction with loaded miles are used to determine load velocity, which is of interest when calculating the relative velocity of the lanes.

Data Exploration

Next we inspect the data to determine the characteristics based on the category it belongs to.

1. **String:** Identify any readily problematic strings that may be a result of input error or missing strings.
2. **Numeric:** Various plots such as box plots, normality plots, and histograms are used to identify the distribution of the data. This allows us to determine outliers which can then be used to normalize the data for further use.
For example, in the violin/box plot above, we can observe the distribution of loaded miles grouped by different equipment type. The width of the plot implies the density of the points. For Flat Bed (FB), the maximum number of orders are approximately 400-500 miles in length whereas for Refrigerated (RF), it is steadily decreasing from 250 miles onwards until 600 miles. All the plots have a long up tail and a short tail towards 0. It indicates that the rate/mile follows approximately the normality assumption (that is the population follows the normal distribution) except FB and VN have thinner tails as compared to RF.

![Total Revenue Distribution](image)

In the Frequency histogram above, total revenue per load can be shown to also follow the normality assumption, but it is skewed left which implies that the results from any prediction analysis involving total revenue as a factor become less reliable as the revenue increases. Unsupervised learning such as k-means clustering is used to isolate orders that generally have similar revenue, allowing us to further segment the data to obtain finer-grained analysis.
3. **Factors:** Expert knowledge from the industry analysts is used to isolate correct factors. It also helps to identify the frequency of incorrect factors present in the data. For example, the equipment column might contain only three types of equipment, namely Van (VN), Flat Bed (FB), or Refrigerated (RF). But the raw data also had other entries, such as CO, TX, INC., etc., which had to be cleaned.

![Rate/Mile Time Series](image)

4. **Time:** The dates in the data allow us to create a timeline of behavior of the numeric variables. This can help us identify seasonality and predict any future numeric variable. For example, in the rate/mile time series above, we can observe the variations in rate/mile grouped by equipment over time. This shows that the Flat Bed prices are always higher than Van or Refrigerated prices. The blue line at the top represents the retail fuel prices in North America and shows a downward trend, whereas the rate/mile plots do not show any appreciable correlation. We can tentatively conclude that rate/mile is insulated from changes in fuel prices and, hence, is accounted for in fuel surcharges. In terms of seasonality, we can observe that Flat Bed prices tend to rise near the start of the year and then dip toward the end of the year.
Data Cleaning

At this step, we determine parameters to identify the various errors in the data. In this step, each error is defined as follows:

1. **Outlier**: For any numeric variable, outliers are determined based on the normality assumptions of the variable.
2. **Undesired strings**: Some strings in the data that occur frequently can be the result of errors.
3. **Desired factors**: There are a limited number of factor variables that are of concern to the industry and any other entry in their place is incorrect. Hence, orders that do not have the desired factors are marked as incorrect.
4. **Deduction based on equation**: For example, the total revenue in MRI data is a combination of line haul, fuel surcharge and accessorial charges. This equation can be expressed in R to validate which orders do not meet this criteria, and are marked as such.

After applying the criteria outlined above to the available MRI data only 55% of the total data was free of errors. Distribution of the number of errors is shown below. We can observe that more than 75% of the orders only have up to 2 errors per order.
The types of errors along with their respective percentages are in the table given below.

<table>
<thead>
<tr>
<th>Description</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero Errors</td>
<td>55.93</td>
</tr>
<tr>
<td>Outlier Linehaul Cost, Outliers Rate per Mile</td>
<td>6.09</td>
</tr>
<tr>
<td>Outlier Unload Time</td>
<td>4.01</td>
</tr>
<tr>
<td>Outlier Load Velocity MPH</td>
<td>3.36</td>
</tr>
<tr>
<td>Outlier Load Time</td>
<td>3.13</td>
</tr>
<tr>
<td>Outlier Total Revenue, Outlier Linehaul Cost, Outlier Rate per Mile</td>
<td>2.30</td>
</tr>
<tr>
<td>NA Router Miles, Outlier Router Miles</td>
<td>2.05</td>
</tr>
<tr>
<td>Outlier Router Miles, Inconsistent Mileage</td>
<td>1.94</td>
</tr>
<tr>
<td>Outlier Rate per Mile</td>
<td>1.72</td>
</tr>
<tr>
<td>Inconsistent Mileage</td>
<td>1.47</td>
</tr>
<tr>
<td>Outlier Linehaul Cost, Outliers Router Miles, Outlier Rate per Mile, Inconsistent Mileage</td>
<td>0.84</td>
</tr>
<tr>
<td>Outlier Total Revenue, Outlier Linehaul Cost, Outlier Rate per Mile, Inconsistent Mileage</td>
<td>0.69</td>
</tr>
<tr>
<td>Outlier Total Revenue, Outlier Linehaul Cost, Outlier Router Miles, Outlier Rate per Mile, Inconsistent Mileage</td>
<td>0.63</td>
</tr>
<tr>
<td>NA Mileage Band, Load Velocity MPH, Outlier Total Revenue, Outlier Linehaul Cost, Outlier Loaded Miles, Outlier Router Miles, Outlier Rate per Mile, Inconsistent Mileage</td>
<td>0.51</td>
</tr>
<tr>
<td>Outlier PTA Empty Miles</td>
<td>0.48</td>
</tr>
</tbody>
</table>

**Predictive Analysis**

Predictive analytics is the perceived “black art” of data science even though it is only a combination of math and algorithms. The application of predictive analysis can be seen from word prediction while typing on a mobile device to product predictions while shopping online. There is a balance to be found between performance, scale, and the trade-off between variance and bias. The steps toward building a prediction model are:

1. **Question**: The first step is to ask the question that would be of value to TMW customers. It allows us to determine what we would like to predict and which variables would be significant in its prediction based on industry/domain experience.

2. **Model building**: Based on the dependent variable (i.e., the prediction) and the independent variables which can explain the behavior of the dependent variable, we choose which technique to use for building the model. There are a variety of techniques available based on the combination of independent variables. Linear regression is the simplest and fastest technique. This builds a prediction model for a numerical dependent variable. Because of its simplicity, it is not applicable for predicting complex nonlinear behavior.
3. **Model validation**: The model is built using a training data set that is separate from the test data set. The model is then tested for validity by running predictions on the test data set and then comparing the training and test error rate. If they differ by a wide margin, the model must be reworked.

4. **Model update and real time prediction based on dynamic query**: The model would be continuously updated with the new data so that the model is sensitive to the changes in the business environment of the TMW customers. These customers also should be able to ask for a prediction based on any number of factors that would constitute the model. For example, given the model as Total_Revenue ~ Pickup_region + DropOff_region, users might only choose to specify the Pickup_region, and the result would be the prediction of Total_Revenue for every O-D pair with Origin being the Pickup_region specified and the destination would be all the DropOff_regions in the data.

**Prescriptive Analytics**

The cutting-edge for application of data science is the step beyond predictive analytics which is prescriptive analytics. This refers to integrating predictive analytics with simulation and optimization algorithms to provide recommendations based on a range of future scenario’s. Prescriptive analytics moves from simply predicting the future to advising on optimal decision making given near real time changes in the environment.

**Ready to make better data-driven decisions?**

**TMW Systems provides a foundation of flexible, integrated information systems and services necessary to:**

- Deliver actionable data to your business across the entire organizational hierarchy.
- Evolve your systems and services to maintain alignment with rapidly changing business circumstances and industry requirements.
- Deliver industry level analytics capabilities for you to measure and create benchmarks with industry trends.

**WIN THE BIG PICTURE WITH TMW SYSTEMS.**

TMW is a leading transportation software provider to commercial and private fleets, brokerage and 3PL organizations. Founded in 1983, TMW has focused on providing enterprise software to the transportation industry, including asset-based and non-asset-based operations as well as heavy-duty vehicle service centers. With offices in Cleveland, Dallas, Indianapolis, Nashville, Oklahoma City, Raleigh, and Vancouver, the company serves over 2,000 customers, including many of the largest, most sophisticated and complex transportation service companies in North America. TMW is a Trimble Company (NASDAQ: TRMB) and part of the International Transportation and Logistics Division.