How to Measure, Monitor and Improve Data Quality in Transportation Visibility

Many organizations are trying to use insights from their transportation visibility platforms to make supply chain decisions. Supply chain technology leaders using advanced analytics need to understand and ensure the quality of the data in order to make decisions based on accurate information.

Overview

Key Findings

- Many companies want to leverage the data and insights from their real-time transportation visibility platform (RTTVP) to assist with upstream and downstream decisions. However, the quality of the data and predictions must be vetted in order to ensure accuracy in those decisions.

- Users often do not take data quality into account as a key requirement when selecting a vendor for their transportation visibility solution.

- Most organizations do not measure the quality of their transportation visibility data on an ongoing basis, so they do not know how or where to improve it. Those that do normally address data quality issues reactively, so they have no effective way to prevent problems.

Recommendations

Supply chain technology leaders looking to ensure data quality within transportation visibility should:

- Investigate what data quality means and how to measure it. Measure the data quality of your real-time transportation visibility platform using data quality dimensions by focusing on those most important to your data elements and use cases.
Strategic Planning Assumptions

By 2023, 50% of supply chains will be using advanced analytics in their organization to advance insights around transportation spend, fleets, sustainability, transportation capacity and freight rates.

By 2024, 60% of the data used for the development of artificial intelligence (AI) and analytics solutions will be synthetically generated.

By 2024, use of synthetic data and transfer learning will halve the volume of real data needed for machine learning (ML).

Introduction

Data quality has been recognized as a key component of any data and analytics strategy for over 20 years. The emphasis on quality comes in and out of fashion with each new technology trend. With the start of each new technology trend, data quality is thought to become redundant or automated, only to find that the initiative replacing data quality cannot succeed without it. Figure 1 provides a brief history of data quality in relation to new technology trends.
When insights are used to manage supply chains, leaders depend on trusted, high-quality data. However, data quality is often treated as an inconvenient afterthought. Organizations should create preventative approaches to fundamentally consider and improve data quality at the onset of transportation visibility initiatives.

This becomes even more important as companies start moving toward advanced analytics, where the data is used to make predictions and prescriptions. If the quality of the data is low, then the predictions and prescriptions based on that data will not be accurate, undermining trust in the system and therefore producing little to no value. So in order to create worthwhile actions, and in the future automate those actions based on data insights, it is critical that the data quality is as high as it can be.
When it comes to visibility, and specifically transportation visibility, data compliance and data quality remain the two main challenges users face when implementing these solutions (see the challenges regarding data in Smart Insights for the Real-Time Transportation Visibility Platforms Market). To ensure companies are prepared to be successful on their visibility journey, this guidance provides an introduction to the core concepts of data quality and how they affect the achievement of value in transportation visibility initiatives.

This document looks at data quality and its impact on transportation visibility in four main sections:

- What is data quality and what are the dimensions of data quality?
- How does data quality apply to transportation visibility?
- How does data quality impact the transportation visibility vendor selection?
- How do companies continue to improve their data quality?

**Analysis**

**Investigate What Data Quality Means**

Most companies understand the value of data, but they don't always understand the difference between “just any old data” and “clean, good data.” So we need to start with explaining why data quality matters and what data quality consists of.

**Definition of Data Quality**

The term “data quality” relates to the processes and technologies for identifying, understanding and correcting flaws in data that support effective data and analytics governance across operational business processes and decision making (see Magic Quadrant for Data Quality Solutions).

**The Dimensions of Data Quality**

As shown in Figure 2, there are many dimensions of data quality, each of which has metrics to be reviewed and measured. You should focus your efforts on the metrics that are important to the data elements required in your use cases. These metrics can be used to determine your data quality levels and can help indicate where data improvement is needed.
All these dimensions can be applied to table level or attribute, structure data or unstructured, and can be executed in real time or batch mode. Not all data is equal or used in the same way, so not all data requires 100% data quality. Different use cases may tolerate different levels of data quality even for the same dataset. You need to use appropriate dimensions for different data objects and for different purposes. It's common to use multiple dimensions to measure one data object. Add them as appropriate. Choose only the metrics that are important to support your use cases.

**Explore the Use Cases of Data Quality in the Different Stages of Transportation and Transportation Visibility Platform Usage**

As it relates to transportation visibility platforms, all of the data quality elements apply, but some of the key dimensions to consider for data quality specific to visibility are:
The first step to ensuring quality of data is to assess and ensure the dataset is complete. This relates to the carrier onboarding and integration on the network (Stage 1 in Figure 3). This is a critical first stage to ensuring complete data is reliable in the feed of data from the carrier network. This is why the carrier network is one of the key influencing factors in selecting a RTTVP vendor. The carrier network can be looked at from three different angles:

- The shipper's individual carrier network at a point in time.
- The collective carrier network on the platform.
- The annual changing carrier base as shippers execute new carrier contracts.
Once the connections to the carrier network are made, it is critical that the RTTVP ensures the connection is maintained, and that the data is fed timely and consistently from the carriers to the platform. Stage 2 focuses on the ability to process the data: The foundation of these transportation visibility platforms is the ability to not just collect the data, but to process the data. During this foundational process, it will consider accessibility, consistency, timeliness, accuracy, completeness, validity and relevance of the data to ensure data quality.

Once the connections are made and monitored, the next most critical piece is the accuracy. A Gartner study revealed data accuracy as the No. 1 issue when it comes to data quality. ¹
The data feed to the platform is used to create predictive insights in Stage 3. Predictive analytics allow companies to get forward-looking insights into their shipments, which can be shared with their customers and partners. This information can be used to share updates before the events occur, but they also allow the shipper to take corrective actions. An example of predictive analytics is the predictive estimated time of arrival (ETA). If the data being leveraged to create the ETA isn’t complete, accurate, timely and consistent, the output of the ETA will also be inaccurate. Many organizations use the ETAs to affect downstream teams and decisions, and rely on the accuracy of the prediction. In Stage 4, the system can, through the use of machine learning, prescribe over time these suggested actions to the shipper, typically in the form of scenarios that show an impact on cost, inventory and customer service. The final step, Stage 5, would be to use the automation of workflows leveraging advanced algorithms and ML techniques. In this case, we are simulating the operator’s intelligence through automation that allows the user to handle more volume, faster, more efficiently, with higher quality of outcome, and creates a better work environment for that user.

Much like the visibility journey Gartner sees companies going through in order to leverage more of the platform and insights realizing more value and cost over time and maturation, there is also a maturity journey to data quality in this market. For more information, see Smart Insights for the Real-Time Transportation Visibility Platforms Market. Both the RTTVP vendors and customers using the data have to ensure the data collection and quality are in place before true predictive analytics and automation can take place (see Figure 3). Data quality and cleansing remains the biggest challenge within the transportation visibility platforms.

Select the Right Vendor That Provides High-Quality Data

Selecting the right partner for your transportation visibility needs is critical. Asking questions about data quality should be a part of the vendor selection process and RFP. Vetting vendors for their commitment and processes on data quality is a critical step in ensuring the vendor can meet your organization’s data quality expectations and requirements. As vendors have different approaches and processes around data quality, we urge users to take a deeper dive when selecting a partner for their visibility journey.

In a recent study that Gartner did with several real-time transportation visibility vendors, we noticed that:

- The data is not measured in a consistent way by the vendors.
Some vendors automate the process, whereas others also have a data quality team staffed to ensure the quality of master data and transaction data, investigating abnormal event data.

Vendors do not all have the same commitment to the accuracy and support of data cleansing.

Some vendors provide good insights on how they improve and guarantee data quality, whereas other vendors do not provide the insights into the inner workings of the data.

Vendors do not calculate ETAs in the same manner, making it hard to compare and benchmark the ETA quality.

Some vendors do not provide any metrics around ETA accuracy, therefore, they cannot give commitment to customers or track for improvement.

Very few vendors offer a SLA to data quality or ETA accuracy.

To demonstrate the variance in the way vendors approach data quality within their platforms, we’ve included a few examples of vendors’ approach and practices for users to consider:

Here is one way vendors look for accuracy of the data: verification events across multiple sources. Where possible, the vendor retrieves the same, or similar, events from different sources. The vendor applies an ML-enabled algorithm for detecting the quality of incoming data, attempts to repair data and flags data it cannot apply or repair. For example, the vendor might capture and compare data from three providers — ocean carriers, an automatic identification system (AIS) and marine terminals — to identify when a vessel departs or has a port of call at a marine terminal. This significantly reduces the chances of missing data while processing the event as early as possible. An additional benefit is that each source typically has a different set of associated attributes, which allows the vendor to further enrich master data and transaction data.²
Another vendor uses supervised ML algorithms to cleanse masses of input data before dynamically updating the digital representation of the supply chain. A typical example is cleansing of shipping partner entity names, which may all refer to the same or related entities. These are automatically reconciled by machine learning algorithms and mapped to a common entity, or parent, as per the business use case. In addition, as part of an automated process, data is enriched with missing information through multiattribute validation, such as filling in ZIP Codes based on city and state locations. The use of supervised AI significantly increases the rate of autoclassified matches, cutting manual effort by roughly 98%. The result of our cleansing, normalization and enrichment is decision-grade and machine-ready data.²

Data quality through collaboration. One of the vendors created a shippers council, which holds regular meetings about carrier data quality for shippers on the platform. In these meetings, metrics and quantitative measures of data quality are shared, including timeliness and completeness from key carriers and nonvessel operating common carriers (NVOCCs) contributing to the platform. Shippers are provided with regularly updated rankings for all carriers about their data performance so that the shippers can engage with their key carriers to request improvements in timeliness or completeness. Shippers can also consider changes in their carrier or NVOCC mix to shift to service providers who are ranked more highly on data quality. This is designed to improve the overall visibility value that enterprise shippers derive from the aggregate data they see on the platform.²

Using a canonical data model to improve data accuracy. One vendor who has direct electronic data interchange (EDI) connections with the top ocean carriers noticed these messages are frequently incorrect, inaccurate or latent. The vendor runs hundreds of millions of these raw transactional messages through their canonical data model, which automatically enriches, deduplicates and normalizes data across all datasets into a single representative model. Through this process, they discard roughly 1 in 5 EDI messages to provide their customers with a clean baseline set of data.²

Due to the variation in process and commitment across vendors, it’s an important attribute to vet during a RFP process for users. Leveraging Gartner’s RFP Toolkit (see Toolkit: RFP for Real-Time Transportation Visibility Platforms) is a good start for most organizations, as many questions related to the vendor’s approach to data quality and accuracy are included:

- Describe how the vendor ensures clean data in its solution.
Improve Data Quality Through the Use of Real and Synthetic Data

Data quality is not a one-and-done exercise. The data needs continuous attention to make sure it reaches “high fidelity,” meaning, just as in audio, you get the highest quality of the data. Vendors need to evolve their data journey by continuing to collect more data, clean and enrich the data, as well as add synthetic data to the real data. Synthetic data is a class of data that is artificially generated.

Vendors in the RTTVP market are collecting data from a variety of sources to visualize shipments in transit and predict things like ETA. The data has to be collected from a variety of sources and normalized together in order to leverage the information to make predictive insights. The attention and monitoring of the various data sources is critical to ensuring accuracy.

Some example of data sources being collected are:

- Carrier-fed data — location, status
- Port data
- Terminal data
- Address data
- Shipment or order data
- Shipper network data — network performance information
- Weather data
- Traffic data
Vendors need to continue to focus on collecting data from all these data sources to get the best possible data quality.

Besides improving the real data being used by the vendors through the integration with the sources of the data, these solutions also must bring in synthetic data. This is a practice not used widely today, but should be given weighted consideration for the future to drive more value. Synthetic data uses AI and ML to enrich the information about transportation visibility and the network that can be inferred or simulated, but might not otherwise be collected from real sources. As more AI models are used for advanced analytics, the importance and the value of synthetic data will grow in the next decade (see Figure 4).

**Figure 4. Synthetic Data**

Currently, most organizations believe that wide data (previously referred to as big data; see Top Trends in Data and Analytics for 2021: From Big to Small and Wide Data) is the way to drive AI value. Consider all of your organization’s investments in this space: data warehouses, data lakes, data pipelines and data quality programs. What they all have in common is an assumption: Your data has an untapped value significantly beyond its current use. But this is often false, as evidenced by the poor return on investment we typically see for these initiatives. Real data is not the new oil, nor is it the invaluable asset many believe it to be. Why not? Well, for a start, the data has several problems:
Real data is incomplete: AI needs both large and diverse datasets, but real data is often incomplete, excluding infrequent scenarios that are critical for AI performance.

Real data is expensive: It is hard to collect, integrate, store and maintain.

Real data is biased: Even if data perfectly reflects reality, it can encode biases present in the real world that we would like to remove.

Real data is restricted: Regulation is increasingly limiting data use for AI.

Synthetic data will become even more important as companies move across the stages of transportation visibility (see Figure 3). In the prescriptive or automation stage, users will need to leverage synthetic data to enrich the AI models to improve the insights that will drive decisions.

Synthetic data will transform the use of AI over time, providing vast opportunities. Early investors in this technology will be uniquely positioned to drive value from AI, tackling business problems that wouldn’t be solvable otherwise and opening up use cases for which they have limited data now. With synthetic data, they should be able to reliably test their complex AI models to ensure they will work once deployed. The future of AI is connected to synthetic data and it belongs to those who embrace it early.

Evidence

1 2020 Gartner Digital Supply Chain Decisions Survey: 268 respondents provided input on the key data quality issues. The question asked was: Thinking of your supply chain data, what are the most typical quality issues?

2 This research incorporates vendor-provided process details, use-case examples and information gathered from our interviews and surveys with real-time transportation visibility vendors or international transportation platform vendors. They included Blume Global, Descartes, E2open, FourKites, Infor, Logit One, project44, Shippeo, Transporeon and Trimble.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Magic Quadrant for Real-Time Transportation Visibility Platforms

Smart Insights for the Real-Time Transportation Visibility Platforms Market