DATA CLEANSING FOR BETTER DECISION MAKING

If you have ever performed supply chain analysis or design projects, you probably know the feeling - A project with straightforward objectives grinds to a halt once you start digging into the data and discover problems: missing information, duplication of records, time period issues and inconsistent naming conventions are just a few examples. These common data challenges present two potential problems. First, significant time can be spent in cleansing and preparing the data to make it ready for analysis. A survey conducted by Harvard Business Review reported that data scientists spend 80% of their time preparing and discovering data[1]. The second potential problem is that significant assumptions or exclusions may have to be used to address these underlying issues, that may compromise the purpose of the analysis. The phrase "garbage in, garbage out" certainly holds true for supply chain analysis and design projects. If the data feeding analytical models (and ultimately decisions) is not representative of the supply chain under evaluation, then deriving meaningful decisions can be both arduous and dangerous.

If the data feeding analytical models (and ultimately decisions) is not representative of the supply chain under evaluation, then deriving meaningful decisions can be both arduous and dangerous.

To help avoid this snag, and the potential waste of time and resources, there are some sound practices that can be put in place to address the data preparation process, as well as the development of project assumptions which is often necessary. In this article, we will share a few of the processes and best practices that we deploy to tackle common problems. You may experience the same or related issues in your own supply chain analysis and design projects. We will focus on the two key phases of supply chain analysis projects: Data Diagnostics and Data Preparation. Figure 1 provides a typical data analysis project lifecycle.

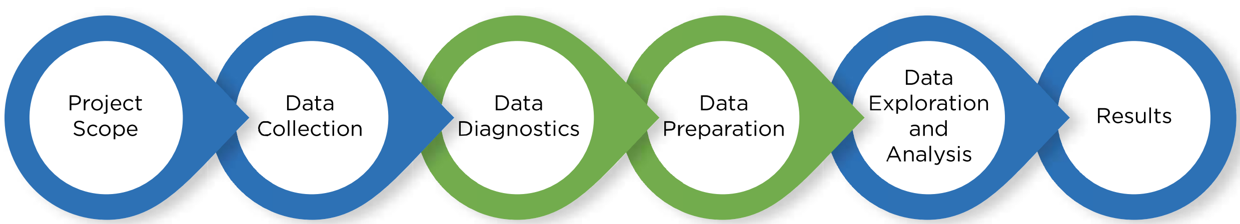

Figure 1 – Data Analysis Project Lifecycle

[1]https://hbr.org/2017/05/whats-your-data-strategy

DATA DIAGNOSTIC

Every good project begins with a clear scope. Business specifications are translated to the data requirements that guide the data collection, both in terms of what data is produced, and through the lens we look at it. To begin to understand the data and any gaps, a thorough diagnostic evaluation against the scope should take place. Conducting a thorough data diagnostic enables you to familiarize yourself with the data structures, content and understand potential gaps that may affect the project scope and expectations. For example, if you had expected to see a group of major suppliers from Vietnam, but the data is indicating China as the largest supply source, then it is a good point to pause and clarify. This could mean incomplete data or perhaps a misunderstanding of the current supply chain. These validation checks allow you to question, and ultimately, align the data to the stated scope and objectives with key stakeholders.

The first step is to capture each data source and key elements in a consistent way to help organize and validate that all data elements are present. In Table 1 below, we have provided an example of a data components summary.

Data Set	1	2	3	4	5	
Workbook	Q3 2019 Shipments	Q3 2019 BO Shipments	Model_Numbers_With_ Weight_or_Dimension_ Values 3-09-2019	Global Tariff - 4.1.2019	Worksheet in Global Tariff - 4.1 - AMS Transcon	
Worksheet	Q3 2019 Shipments	Q3 2019 BO Shipments	Model_Numbers_With_ Weight_or_Di	8 Sheets	5 Sheets	
Contents	Shipments	Shipments	Models and Weights	Rates	Rates	
Columns	39	43	11	–	–	**93**
Rows	135,998	118,838	26,465	–	–	**281,301**
Cells	5,303,922	5,110,034	291,115	–	–	**10,705,071**
Shipments	17,827	15,138	–	–	–	**32,965**
Sales Orders	17,416	11,632	–	–	–	**29,048**
Qty	943,001	727,321	–	–	–	**1,670,322**
Weight	3,382,098	–	–	–	–	**3,382,098**
Wgt Units	lb	–	–	–	–	
Start Date	7/1/2019	7/1/2019	–	–	–	
End Date	9/30/2019	9/30/2019	–	–	–	
Received	6/24/2019	6/24/2019	7/6/2019	6/24/2019	6/24/2019	
Used	Yes	Yes	Yes	Yes	Yes	

Table 1 – Data Components Summary

In this example, we knew we had to merge data sources one, two, and three together to build our “core file” because there were unique elements contained in each. The last two data sets were simply rating tariffs so no further actions were needed for these. Developing this “core file” containing all critical items will make your modeling work easier by streamlining your actions against a single data source, instead of trying to connect, query and validate across multiple source tables. As illustrated in Table 2, we had a common reference field (Sales Order Number) that enabled us to capture unique elements from the three data files and create a single file. We used the transportation detail fields from Data Set 1, added in the freight terms and product family detail from Data Set 2 and then added the part information from Data Set 3. However, it may not always be possible to merge all the files together for comprehensive analysis. At this point, another pause and validation should occur before moving forward.

1	1	2	2	3
Q3 2019 Shipments	**Q3 2019 Shipments**	**Q3 2019 BO Shipments**	**Q3 2019 BO Shipments**	**Model_Numbers_With_Weight_or_Dimension_Values 3-09-2016**
Q3 2019 Shipments	**Q3 2019 Shipments**	**Q3 2019 BO Shipments**	**Q3 2019 BO Shipments**	**Model_Numbers_With_Weight_or_Di**
39,071	**39,071**	**35,861**	**35,861**	**26,465**
sales_order	Parent_ID	**Sales Order Number**	Forwarder Confirmation Date	Part Type
BillTo_ID	**Item**	**Freight Terms**	Rev Rec Type	**Number**
BillTo_Name	Item_Type	Reporting Ship to GEO	Book Date	Description
BillTo_Address_Line1	Product_Line	Reporting Ship to Country	Ship To Address 1	Lifecycle Phase
BillTo_Address_Line2	Slot	SO Channel Code	Ship To Address 2	Part Data.User Item Type(*)
BillTo_City	Port	Bill To Name	Ship To Address 3	Part Data.Item Status(*)
BillTo_State	Ordered_Quantity	Ship To Name	Ship To Zip Code	**Part Data.Packaged Weight (in lbs):**
BillTo_ZipCode	**Shipped_Quantity**	Ship Date	Fob	**Part Data.Dimensions UOM**
BillTo_Country	Material_Cost	Partial Allowed	Shipment Key	Part Data.Dimensions Width
ShipTo_ID	**Ship_Date**	Product	Shipment Priority Description	Part Data.Dimensions Length
ShipTo_Name	Month	Supplier	Business Unit	Part Data.Dimensions Height
ShipTo_Address_Line1	Inv_Org_Code	Order Type	**Product Family**	
ShipTo_Address_Line2	**CM**	Ship To City	Due Date	
ShipTo_Address_Line3	PO_Number	Ship To State	Promise Date	
ShipTo_City	**Tracking_Number**	Ship Via	Require Date	
ShipTo_State	Ship_Via	Tracking Number	CM Commit Date	
ShipTo_ZipCode	**Weight**	Quote Line	Payee Cm	
ShipTo_Country	Pieces	Order Line	Shipment Net $	
Quote_Line	**Region**	Ship To Country	Shipment Cost $	
Line_Number		Parent Customer Name	Total Shipped Products	
		End User Name	Revenue	
		Customer Name		

Table 2 – Joining Files

COLOR AND CATEGORY

Nodes
Dates
Parties
References
Measures
Values

Columns used for analysis are in **bolded text**.

Columns also used to join tables are in **red text**.

Tables 1 and 2 are joined on sales_order = Sales Order Number.

Tables 1 and 3 are joined on Item = Number.

After merging the files together, there will be exclusions. These could be duplicate entries, missing weights, mode or service level inconsistencies, time frames, date formats (common amongst global data sets) or records that might be out of scope (e.g.. if the project was related to ocean consolidation, then perhaps we exclude all the air shipments).

Documenting all of these items in a structured manner provides a clear path for understanding the magnitude of the exclusions. Once exclusions are captured, it is likely that assumptions are necessary to fill in significant gaps. It is an important step to summarize and display the exclusions and assumptions so they are transparent to key stakeholders. Unsound methodologies and unrepresentative data erodes not only the confidence for results but also trust from key stakeholders for both the current project and potentially future projects that you are summoned to lead.

In Table 3, we have provided an example of how to share the exclusions and key assumptions so their impact is visible.

Data Set	**1**	**1**	**1,2**	**1,3**	
Workbook	Q3 2019 Shipments	Q3 2019 Shipments	Q3 2019 BO Shipments	Q3 2019 Shipments	
Worksheet	Q3 2019 Shipments	Q3 2019 Shipments	Q3 2019 BO Shipments	Q3 2019 Shipments	
Column	Region	CM	Sales Order	Dims	
Value	Out of Scope (region)	Misc	Not Found	0	
Total Records	94,755	7,915	19,222	1,947	**TOTAL 123,839**
% Total Records Affected	5%	33%	14.10%	1.40%	
Count of Remaining Records	90,018	2,677	16,531	1,920	**REMAINING 111,146**

Assumptions

- Air shipments from Shanghai to Los Angeles with a blank service level were assumed to be standard service, 5% of shipments
- Where the supplier is blank in Shanghai we assumed everything was from a single supplier, 3% of shipments

Table 3 – Data Exclusions and assumptions

Unsound methodologies and unrepresentative data erodes not only the confidence for results but also trust from key stakeholders for both the current project and potentially future projects that you are summoned to lead.

DATA PREPARATION: COMMON DATA ISSUES AND SOLUTIONS

With almost every supply chain analysis, we see common data issues that need to be addressed to create a more accurate, representative and meaningful analytical product. This step of data preparation can be described as cleansing and harmonizing our core data records. Table 3 summarizes possible solutions to many of these common problems that you are likely to face.

<table>
<tr><th>PROBLEM</th><th>SOLUTION</th></tr>
<tr><td>Geo Coding: Identifying the actual locations for physical links in your supply chain (suppliers, customers, warehouse locations) is critical for model accuracy as distance is usually a proxy for transportation cost. However, in transactional data stores it is unlikely that they contain latitude and longitude locations as a native process and there are various formats containing physical locations and addresses.</td><td>Using a variety of tools from web-based services to stored geographic database tables it is possible for us to geocode many different types of geographic information – ranging from fully detailed address information down to single, misspelled city names. The less detailed the information is though the less confidence we would have in the results and it is possible that the geographic info is so incomplete we can't find a location and must be excluded, revalidated or geo located to another close by location.</td></tr>
<tr><td>Data Inconsistencies: Extracts from structured data sources often contain various naming conventions for the same thing (multiple version of the truth). The culprit is usually free form text or too many drop down boxes in internal systems. An example might be selection of service level, with an aggregated view of records as follows:
<table>
<tr><th>Service Level</th><th>Count of Records</th></tr>
<tr><td>Two Day</td><td>5670</td></tr>
<tr><td>TwoDay</td><td>4967</td></tr>
<tr><td>2 Day</td><td>3506</td></tr>
<tr><td>2nd Day</td><td>1656</td></tr>
<tr><td>2Day</td><td>1123</td></tr>
<tr><td>GND-2 DAY</td><td>576</td></tr>
</table>
Similar issues are seen for supplier and customer names, Sku/products, sale channels, divisions and many other tables for key supply chain data elements.</td><td>To perform the cleansing we will typically bring the data into a database where we can store and manipulate more easily. In this example, we might make larger updates to the transactions by looking for specific text such as "2" or "Two" with the service level column. These can then be easily converted to our desired naming convention.

The Levenshtein distance algorithm (also known as fuzzy lookup) can be used by identifying how close two strings are to one another and how many character changes would be needed to make the strings match. A simple example: ABC is a closer match to ABD (1 character substitution) than ACB (2 character substitutions). This allows you to group a large set of unique spellings like customer names into a smaller, more concise list.</td></tr>
<tr><td>Transportation Mode Inconsistencies: Accurate representation of modes is important to validate model spend levels and identify network opportunities. However, there are often incorrect classifications, or lack of detail, in the transactional data set containing transportation information. An example would be LCL vs. FCL for ocean movements or LTL & FTL for domestic transportation. We often see weight data listed as "LTL" that is well over what would typically be an LTL shipment, and the record probably should have been flagged as FTL.</td><td>When this occurs, we tend to use general rules to either prescribe what the network should look like (if there is no data) or identify opportunity (if the data shows a different profile that typical rules). We follow standard transportation "pivot points" where it is generally less expensive to bump up to the next service, e.g. in the US, the pivot point from parcel to LTL is 150 lbs. and from LTL to FTL is 10,000 lbs. While the actual pivot point will depend on the mode, service and tariff, these general rules provide a great starting point to look at the network.</td></tr>
<tr><td>Dates: When there are multiple files extracted from different geographies' databases, we often see different formats: Month/day/year vs. day/month year. When these files are merged, it can be problematic to decipher what date is accurate, e.g. 12/6/2019 = June 12 or December 6th?</td><td>This is also usually addressed in a SQL database using reference tables to convert the date from one format to another based on a logical rule (for example origin country or source data). Doing this across a large data file by hand can be very tedious, but using database software in this manner greatly speeds it up.</td></tr>
</table>

Table 3 – Common Data Problems

AUTOMATING CLEANSING: LOADING DATA

For our examples above, we described a manual approach to data cleansing. However, another method is available. The process is known as an acronym: ETL (extract, transform, load). ETL uses code to clean and harmonize the raw data input sources. It's most common application is found in a digital twin modeling environment. In this environment, routine data refreshes feed a standard set of models. However, the setup and development of an ETL process requires significant development time as each assumption and cleansing rule is coded into the ETL process. When your analysis is a custom, onetime event the manual process we described above is the preferred approach.

To learn more about this type of modeling environment and Expeditors digital twin service, follow this link to Expeditors' digital twin - [The Living Model](#).

CONCLUSION

Data analytics continues to gain in importance and more investments are being made to expand and enable data-driven decision making across organizations. As with any enabler, fundamentals and foundations need to be in place to ensure that analysis is handled effectively and sound decisions are being made.

The challenges with data cleanliness don't appear to be diminishing anytime soon. One can argue that it will increase more with the rapid growth of larger and more complex data stores brought about by digitization.

While there are many sophisticated tools and techniques that can be deployed in this space, they may not always be the best fit or most practical. In many situations, the process is faster and more accurate to do it manually by following the project's scope and focusing efforts on the items that are most impactful to the results . As you work through your own data diagnostic and preparatory phases, we hope that some of these learnings and practices can be leveraged to help produce meaningful and sound decisions to propel both you and your company forward.