Case study for tabular data

This case study is a simplified abstraction of an actual project undertaken by Central Moment's analysts. And all data have been disguised to preserve confidentiality. Still, however, the main principles are preserved. The example highlights how Central Moment assists clients in combining multiple sources of information—especially when that information comes in large data tables. Indeed, a lot of the information that company decision-makers use lies in the form of tables. Data on things like sales volumes by product feature, product returns by customer type, or total market size by geography, often come in tabular form. For example, here below is a table summarizing quarterly Smartphone shipments by vendor.

Table 1: Data source #1 reporting smartphone shipments by vendor

Vendor	**Quarterly shipments**
Samsung	58,942
Apple	44,469
Other	209,589
Grand total	**313,000**

Data reported in thousands
Source: IDC Quarterly Mobile Phone Tracker, July 28, 2021

Consider the client in this simplified example to be one of the vendors listed in Table 1—a large, world-wide smartphone manufacturer. The client in this case had multiple data sources like that in Table 1. Each data source described different sides or aspects of the smartphone market reflecting that data source's particular strengths and weaknesses In addition, each data source reported a completely different estimate of total smartphone shipments. For example, while one source described quarterly smartphone shipments by vendor as in Table #1 above, another source described smartphone shipments by screen size. And another source described quarterly smartphone shipments by price category. This presented multiple problems. The client was forced to switch among different data sources to understand different parts of the smartphone market and the client was forced to judge between sources when multiple sources described the same part of the market. In addition, each source contained different estimates for total smartphone shipments. So, whenever the client sought to learn about one part of the smartphone market, that source invariably contained smartphone estimates that were at odds with those from other data sources. This client, however, couldn't plan its marketing strategy around partial views and conflicting shipment estimates. What the client wished for was one single, complete, and accurate view of the smartphone market—one that reflected the strengths of all its data sources, and a view that reduced the inevitable error among sources. The client needed Central Moment, Inc. to evaluate the data sources and use its analytic expertise to distill all available information into one single source of market truth.

This is one of Central Moment's key strengths. We can take every bit of a client's available information and integrate or combine it into one single source of truth. At the same time, this single source of truth tends to reduce error as it "borrows strength" from the different inputs and errors tend to cancel-out. This single source of truth is also more complete than any one individual source. It contains all dimensions of interest contained in the various ingredient data sources.

The data source in Table 1 (IDC's Mobile Phone Tracker) was quite good at reporting smartphone vendor market share, e.g., quarterly smartphone shipments by Apple, Samsung, and others. By contrast, the second data source (shown in Table 2 below) had earned the client's trust for accurately reporting smartphone shipments by screen size. Note that the vendor shipment totals in Table 2 below differ from those in Table 1 above due to differences in the data collection methods of the two sources (and different reporting errors).

Table 2: Data source #2 reporting smartphone shipments by vendor and screen size

Screen size	Apple	Samsung	Other
Less than 5 inches	-	8,697	30,299
5 - 5.4 inches	12,946	17,191	61,897
5.5 - 6 inches	19,491	25,942	92,581
Greater than 6 inches	12,898	17,351	20,522
Total	**45,335**	**69,181**	**205,299**

Grand total = 319,815
Data reported in thousands

The client's third data source, shown in Table 3 below, was the only data source that reported smartphone shipments by major geography which was another value aspect of the market for this client. Again, the three subtotals by vendor, and also the grand total shipment estimates, all differ between Source #1, Source #2, and Source #3 due to methodological differences and reporting errors. However, all data sources are attempting to measure the same fundamental thing—worldwide quarterly smartphone shipments. They each, however, emphasize and report on different dimensions or aspects of it. They each reflect a different expertise.

Table 3: Data source #3 reporting smartphone shipments by vendor and geography

Geography	Apple	Other	Samsung
Asia	17,889	24,090	86,111
USA	10,436	14,211	50,482
Europe	8,785	11,763	41,740
ROW	6,430	8,879	31,433
Total	**43,540**	**58,943**	**209,766**

Grand total = 312,249
Data reported in thousands

Finally, data source #4 shown in Table 4 reported on retail smartphone sales over the same time period. And while this source didn't measure smartphone shipments per se like the other sources, it was close. And the source provided an essential view of smartphone consumption by both screen size (like in Source #3) and by price band (which no other source included). Table 4 reveals that relatively few large phones ever sell for less than $100; and relatively few small phones sell for $600 or more.

Table 4: Data source #4 reporting smartphone shipments by screen size and price

Size	Less than $100	$100 - $199	$200 - $599	$600 or more
Less than 5 inches	5,103	18,037	17,233	515

5 - 5.4 inches	10,249	36,615	34,257	2,927
5.5 - 6 inches	15,419	54,041	51,063	30,149
Greater than 6 inches	61	1,190	11,344	6,818

Grand total = 295,021

Data reported in thousands

The last data source we consider in this case study are the smartphone shipment estimates generated from within the client's own internal company reporting system. That is, the client in this case study was actually one of the smartphone shipment vendors reported by some of the earlier data sources. Now, client information like this isn't essential to a data integration project. But usually, clients consider their own shipment data to be more accurate than anything reported by external data providers. And clients often prefer seeing their own data included in the market model provided that the client's actual shipment values won't get leaked to the wrong audience. Again, it's not essential, but in this case the client provided estimates of its own smartphone shipments arranged by all of dimensions of interest. The client's internal estimates contained its own shipment estimates arranged by geography, screen size, and price. The client's estimates did not include any information regarding the other smartphone vendors, i.e., its competitors.

Table 5 compares all of the data sources used in this analysis in terms of their market coverage. And for all data sources that include an estimate of total worldwide quarterly shipments, the last column in Table 5 summarizes those estimates. Although each data source contains a different worldwide estimate, all estimates are in the neighborhood of ~300 million units.

Table 5: The relative smartphone market coverage of each data source

Data source	Vendor (3 levels)	Geography (4 levels)	Screen size (4 levels)	Price band (4 levels)	# Data rows in the source	Total quarterly shipment estimate (in thousands)
Source 1	✓				3	313,000
Source 2	✓		✓		11	319,815
Source 3	✓	✓			12	312,249
Source 4			✓	✓	16	295,021
Client actuals	(Just one level)	✓	✓	✓	24	NA

One can see in the above comparison that no one single data source contains a complete view of the smartphone market. Each data source describes just one or two aspects of it. In addition, each of the worldwide data sources contains a different estimate of total smartphone shipments. This is typical of the projects and data inputs that Central Moment works with. Again, what the client wants from this effort is a complete view of smartphone shipments—one that leverages the information contained in all the data sources in Table 5 and that can estimate smartphone shipments for all major dimensions of the market in one single picture.

And that's exactly what Central Moment achieved. In this project, Central Moment integrated all five data sources listed in Table 5 above. That is, we combined the four external data sources (Sources #1 -

#4) along with the client actuals. A key step in the process involved researching the data sources to learn the relative confidence we should attribute to each one. (Client confidence was highest in the source reflecting their own shipment actuals). This confidence then became an essential ingredient to a Bayesian statistical model we created to integrate all of the different data sources. The end result was data table with complete dimensionality of the smartphone market. That is, the resulting view could contain up to 192 data points (since 3 vendors * 4 geographies * 4 screen sizes * 4 price bands = 192). However, because Apple doesn't sell any smartphones with a screen size less than 5 inches (see Table 2), the actual deliverable contained just 152 data points or table cells.

Table 6a shown below contains a partial view of the end result or final deliverable. As one can see, this view contains smartphone shipment estimates for all dimensions of interest. That is, it contains shipment estimates for each of the four geographies, crossed by all three vendors, crossed by all four screen sizes, crossed by all four price bands—a row for every possible geography*vendor*price*size combination. The deliverable now contains more rows than any one individual data source. This view enabled the client to see all aspects of the smartphone market without conflicting estimates.

Table 6a: The final integrated result—a complete view of smartphone quarterly shipments

Record #	Geography	Vendor	Price	Screen size	Shipment volume
1	Asia	Apple	$201 - $600	5 - 5.4 inches	4,580
2	Asia	Apple	$201 - $600	5.5 - 6 inches	6,779
3	Asia	Apple	$201 - $600	Greater than 6 inches	1,514
4	Asia	Apple	Greater than $600	5 - 5.4 inches	2,669
5	Asia	Apple	Greater than $600	5.5 - 6 inches	3,997
6	Asia	Apple	Greater than $600	Greater than 6 inches	933
7	Asia	Samsung	Less than $100	Less than 5 inches	370
8	Asia	Samsung	Less than $100	5 - 5.4 inches	719
9	Asia	Samsung	Less than $100	5.5 - 6 inches	1,070
10	Asia	Samsung	Less than $100	Greater than 6 inches	21
…					
151	ROW	Other	Greater than $600	5.5 - 6 inches	2,441
152	ROW	Other	Greater than $600	Greater than 6 inches	636

Data reported in thousands

Table 6b: The final integrated result--summarized by vendor

Apple	Samsung	Other	Grand total
50,201	**60,509**	**205,267**	**315,977**

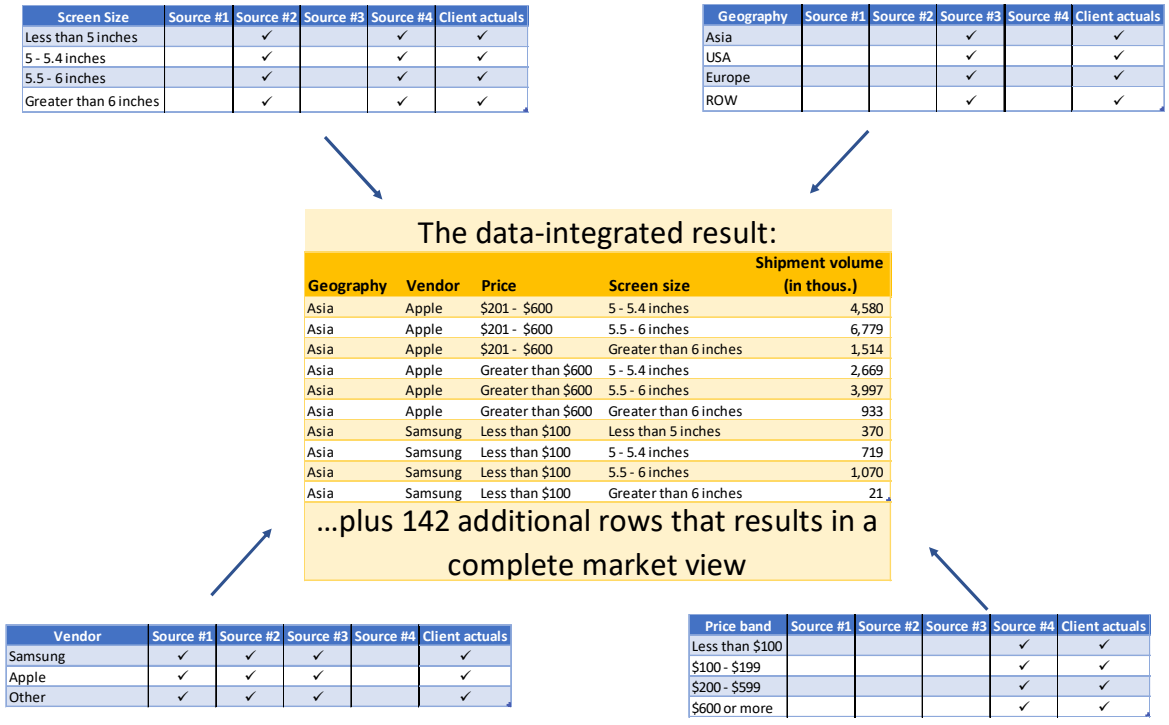
Data reported in thousands

Table 6b shows the finished shipment estimates summarized by vendor and also the grand total quarterly shipment estimate. Note that the final estimates for shipments by vendor in Table 6b reflect something of a compromise between the vendor shipments described in Tables 1, 2, and 3. Similarly, the total market estimate in the far-right cell of Table 6b lies in between the source data estimates

shown on the right-hand side of Table 5. Indeed, one can look at the integrated result as something of a weighted average of the input data sources—where the data source weights reflect the relative confidence given to each source. The final shipment estimates in Table 6b are almost always more accurate than the shipment estimates in any one single data source. Errors in the individual data sources tend to cancel out once these data sources are combined. The final estimate tends to "borrow strength" from all of the separate data sources.

Figure 1 below provides a picture of the data integration process. Again, the client wanted to understand smartphone shipments along four dimensions in combination--smartphone shipments by screen size, geography, vendor, and price band. The client had five data sources addressing these areas, but none of the sources contained complete information on all four dimensions. At best, each data source included information on just two parts or dimensions of the smartphone market. The end-result, however, was a complete view of the smartphone market, one that described smartphone shipments along all four dimensions of interest. The finished result leveraged the information contained in each of the individual sources. The final estimate filled-in the incompleteness found in the separate data sources.

Figure 1: A visual image of the data integration process



This kind of data integration—assisted by Central Moment's extensive demographic and firmographic database—is our core competency. It permits our clients to leverage all available sources of market information. It resolves ambiguities and results in one complete source of market truth. And while the above example is a greatly simplified version using just five data sources, Central Moment has produced

similar results from dozens of data sources containing thousands or even millions of data points. In fact, our methodology is far faster and more efficient than standard Bayesian techniques when it comes to integrating large data sets.

We're happy to talk about our methodologies and how we might bring them together to help solve the conflicting data sources affecting your organization.