DATA IN HARMONY

Confessions of a Data Salesperson

What You Need to Know BEFORE Shopping for Data.

Tom Myers
20+ Year Data Veteran

PURPOSE OF THIS GUIDE

Acquiring the data you need can be a very time consuming and frustrating endeavor. Whether you are just scouting for new data sets, or have an immediate need for specific data, navigating the data landscape can be daunting. There are plenty of data providers selling similar data sets, but far fewer ones providing data solutions that address your unique needs.

This guide is based on my experiences over a 20+ year career in the data industry. I have been both a consumer of data and provider of data. So I think I offer a unique perspective on the data procurement process.

This guide is meant to help you navigate the data acquisition process so you can get the highest quality data for an affordable price under friendly terms and conditions.

I also include INSIDER TIPS that reveal little secrets that data providers don't disclose and probably don't want you to know.

CLEARLY DEFINE WHAT YOU NEED

The first step in successfully procuring data is to clearly define your data requirements. This may sound obvious, but very often data acquisition projects launch without considering

Data Shopping Checklist

- ✓ Cleary define what you need
- ✓ Make a list of potential providers
- ✓ Define your sample data request
- ✓ Schedule a call with each provider
- ✓ What is included in the data
- ✓ What are common uses of the data
- ✓ What is the coverage of the data
- ✓ What is the source of the raw data
- ✓ How is the data collected
- ✓ What data cleansing is done
- ✓ When do files become available
- ✓ How is the data delivered
- ✓ What file formats are available
- ✓ What are the storage req's
- ✓ What IT & expertise is required
- ✓ How to get support
- ✓ How can the data be used
- ✓ What if you cancel

such important aspects as:

- What exactly is the problem or pain point you are trying to address? What question(s) do you want the data to help you answer?
- Which pieces of information are "must haves" in the data set (e.g. identifiers to map to existing data)?
- When must you receive data updates in order to meet internal deadlines?
- Who within your organization needs to be able to view and use the data?
- Who will be conducting the technical evaluation of the new data? Does it require a senior, more experienced data scientist or can a more junior analyst effectively evaluate the data?
- Will you ever need to share the data with anyone outside of your organization (e.g. customers, regulators, etc.)?
- What is your budget for this data? What is the maximum amount of money you are willing/able to spend for this data?

It is important to document all of your requirements BEFORE starting off to look for data. This will save time, and ensure you get the right data solution for your use case.

MAKE A LIST OF POTENTIAL PROVIDERS

It is rare for there to be only one source for any particular data set. Most likely there are multiple sources and solutions available to you. Often a good first place to look for the data you need is from your existing data providers. Even if they don't have the data you need, they may know of some sources.

In some industries, there are long established legacy data providers. Many of these firms are great businesses with which to work. However, some exploit their size and position atop the food chain to bully their clients on pricing and contractual terms. They are also not too motivated to innovate or customize their solutions. Therefore, I recommend you cast a wide net and look at a variety of providers – from big/established firms to smaller companies and startups.

Hopefully you'll be able to create a list of several potential providers. Not only will this help determine a fair price for the data, you'll also be able to ask more questions and gleam more details on the sources of the raw data, potential challenges of working with the data that you had not yet considered, and more.

DEFINE YOUR SAMPLE DATA REQUEST

When shopping for data it is critical to receive data samples BEFORE you make any final decisions. Almost all data providers are used to providing data samples and will likely have a predefined sample available.

Clearly define exactly what data you need.

However, you should have your OWN requirements for any data sample you receive. When defining your sample data requirements, there are several factors to consider, including:

- Which data points do you wish to see in the sample?

- Ask for data from the geographic areas or categories of most interest to you.

- If you have a file format preference, make sure your sample data is delivered in that format.

By requesting the same sample data from each potential provider, it will make it easier to compare "apples-to-apples".

Note that a data sample is useful to determine whether the provider has the data you need. That is, the data contains the critical information you require and can be delivered in a format with which you can easily work.

To truly evaluate a data provider and its solutions, you may need to request a "data trial". During a trial, you want to be able to access a larger amount of data that goes further back in history. You also want to connect to production servers (not a demo environment) to download data updates for a given amount of time. This can help you gauge the reliability and capabilities of the provider's data delivery systems.

If your use case requires customization of the provider's solution, or a deeper dive over a longer period of time, ask for a Proof of Concept (POC). POCs usually require some small payment up front, but are the best way to determine if a data solution and provider are best for you. Sometimes data providers refund all or a portion of the POC fee when you subscribe for a full data license.

INSIDER TIP: If a provider refuses to give you a custom data sample that meets your requirements, or says you have to pay for a trial, these should be red flags. At the same time, don't expect a data provider to give you free access to the complete data set with all of its history, or a free trial term greater than three months (30 days is most common). Data providers often operate under redistribution contracts that prohibit them giving away large swaths of their data without having to pay fees back to the sources of the raw data.

SCHEDULE A CALL WITH EACH PROVIDER

It may be tempting to only communicate with potential data providers via email. However, you'd be taking a huge risk and missing out on the opportunity to learn a great deal about the data you seek and the various potential providers. Schedule a call with each provider you are considering. This is your opportunity to pick their brains, not just about their data, but also other providers.

You can also learn how other users are maximizing their investment in the data. What hurdles and challenges did they face that you could avoid?

Data providers are humans, too. Humans tend to share a great deal more over the phone than they do in email.

INSIDER TIP: Most data providers and their salespeople are very professional and do a great job. However, a small number can put their interests ahead of their clients. Often these salespeople have quotas and earn a considerable amount of their compensation in the form of commissions and/or bonuses. This may benefit the data provider's business model, but it doesn't always benefit you. The salesperson assigned to you might be under tremendous pressure to meet his/her short-term quota. So unless you work for a huge company, or request enough data to clearly help the salesperson meet that immediate quota, you may not receive the attentive, prompt service that you deserve. Obviously, this should disqualify the data provider from your search.

For example, it is very common in the capital markets industry to speak to a more junior employee whose job is to screen and qualify YOU before you are permitted to speak with a more knowledgeable salesperson or technical data person. This screening process often includes questions about your assets under management (AUM) because this is an indication of how large your company is, how much data you may consume, and (most importantly to the data provider) how much they can charge you for the data. If they have not heard of your firm, or your AUM is not great enough, your emails and phone calls will likely go unanswered. That's horrible, but sadly how many large, legacy data providers operate. Needless to say if this happens to you, you should scratch the data provider off your list.

WHAT IS INCLUDED IN THE DATA

Often data providers offer different versions of their data. This could be a pricing strategy to entice clients with a lower priced option to start using their data. Or they have noticed that different use cases require different data points. When you request your data sample

Have conference calls with each data provider to uncover more details about their data and competitors.

and/or trial, be sure to ask for ALL available fields. This ensures you can determine exactly what data points are available.

Also, you should ask for a file specifications document or “data dictionary”. This will tell you the fields that are available and the format of each field. This information is critical.

For example, perhaps the data contains a time stamp. You need the time stamps to be to the nano second, but the provider only has millisecond time stamps.

Also, learning how each provider defines each field can avoid confusion down the road. This is not just important on how a number is calculated, but how other non-numeric data fields are defined.

COMMON USES OF THE DATA

Understanding the key attributes, strengths, and weaknesses of a data set is crucial to determining if it can help you. Especially with multiple use cases and/or end-users, often data is acquired and only after weeks, months or even years do the users realize the data is not what they need.

One of the simplest ways to determine if a data set is right for you is to ask the data provider how their existing clients use the data. If their existing clients are trying to answer similar questions to yours, then the data may be helpful to you.

However, if none of the existing users of the data have the same problem or use case as you, further investigation is likely necessary.

COVERAGE OF THE DATA

Data coverage is a crucial consideration with many data sets. This may refer to geographic coverage, or some other means of categorizing data.

For example, real estate data about a particular metropolitan area may include only properties that are listed with one of the Multiple Listing Services (MLSs) in the area, omitting properties listed with other area MLSs.

Coverage can also refer to how much historical data a provider offers. Perhaps there was an important event in the past which is critical to your analysis. You'll want to make sure the provider's history includes such an event.

INSIDER TIP: Data providers are almost always expanding their coverage. However, for business reasons, they prefer to have a client asking for the additional data before they dedicate the resources, time, and money to implement the expansion. So if a provider appears to be lacking in coverage, ask them about it. There is a good chance they have or can get the data you need.

SOURCE OF THE RAW DATA

Knowing where the raw data that underpins a provider's data product comes from is important for several reasons. First, where the raw data comes from can influence the quality of the final data product. The accuracy, completeness, and freshness of the data you receive can be greatly determined by the source(s) of the raw data.

Also, you can get an idea of the reliability of the data service going forward. Does the source of the raw data appear steady and likely to be around in the future? Or is its longevity suspect, raising questions about whether you'll still be able to receive data into the future?

Equally important is the compliance of the data. That is, does the data provider from whom you'd be receiving your data have the legal right to possess and/or redistribute the raw data? Given the exponential explosion of new data sets in the world, and changing privacy laws, this question has become more important than ever.

Ask about the sources of the raw data.

For example, let's say you sign up to receive an exciting new data set. The data contains all the key pieces of information you need. The history is deep and robust. Updates arrive on time every day. And you were able to get a price below your budget! But then you get a notice from your provider that they can no longer provide the data to you. Then you get a letter from a law firm you've never heard of saying you have their client's data without his permission. You can see where this is going – a very uncomfortable conversation with your management team, legal counsel, etc. So be sure to ask a lot of questions about the sources of the data you'll be receiving.

HOW IS THE DATA COLLECTED

Much like the sources of the raw data, how the provider is collecting their raw data can be directly related to data quality and compliance. For example, stock quotes are broadcast in real-time from exchanges and trading venues — imagine a huge fire hose spewing tons of data. If your data provider recorded such a real-time feed over the internet in order to build its historical data product, there is a very good chance they have dropped packets at some point causing omissions and errors in their data.

Another consideration is web scraping. This is the automated process of collecting data from websites. Bots scour the internet collecting interesting data from websites, usually without the owners of the websites even knowing. Inherently there is nothing illegal or nefarious about web scraping. However, if the web scraping bots collect any copyrighted material for which their owner does not have the express right to use and/or redistribute, then a huge legal and compliance issue can arise.

So again… you should ask a lot of questions about how the raw data is being collected from the provider's source(s).

WHAT CLEANSING IS DONE TO THE DATA

Data quality is the #1 concern of most data users. So having data that is accurate, complete, and timely is extremely important. However, in the quest for the highest quality data possible, too often data is over scrubbed. Outliers are removed. Apparent duplicate records are deleted. The result can be a data set that has lost its reality.

An example may help…

Let's say you are developing an algorithm to trade the stock market. To confirm the accuracy of your algorithm you obtain a deep history of stock prices and back-test your strategy. Once you've fine-tuned your algorithm to where it is consistently producing good results, you begin using your algorithm in the live market.

However, the results you see in your live trading are dramatically worse than you saw in your research. Data over scrubbing may very likely be the culprit. By removing entire records or data values that looked suspicious, the data provider has made the historical data look nothing like the real, live data. Sometimes in the real world trades are reported late, and quotes are manually entered with decimal points transposed.

Because humans are involved, no data is 100% perfect.

Remember that almost all data sets are reporting on human behavior, and all data feeds are programmed by humans at some level. Humans make mistakes or ignore rules and processes from time to time.

The result… no data set is perfect!

That can be hard to accept, but the sooner you understand this, the better for your data analysis (and sanity).

So why not ask for your data to be 100% raw? That would eliminate the risk of over scrubbing. Unfortunately, raw data is often unusable. So instead of doing data analysis, you will waste hours, weeks, or even months making the data ready to use. It is better to let your data provider — who knows the data better than anyone — normalize raw data and perform checks to make the data useable.

INSIDER TIP: The best way to avoid over scrubbing a data set is to better understand the cleaning process. What are the cleaners looking for exactly? Also important is to NOT remove any suspect data, but instead flag it and provide details of why the data should be questioned.

WHEN DO FILES BECOME AVAILABLE

Assuming you wish to keep your data set current going forward, you'll want to receive periodic data updates. The timing of when these updates become available for download can be critical to your internal workflows. Such data updates often are fed into multiple reports, analytics, etc. that take time to process. So be sure to confirm when you can expect data updates to be ready for download, and that you'll have enough time to process the data internally.

INSIDER TIP: Some data providers can make their data available earlier if you are willing to receive an uncleansed version. That is, the data provider can skip its internal cleansing process and make a rawer version of the data available to you earlier in the day. Then you can download the second, more pristine version of the data, later in the day.

DELIVERY METHODS OF DATA

Most data providers offer several methods of receiving their data. These can include:

- Bulk file download
- API delivery
- GUI display

Which delivery method is right for you depends on your use case, internal technology stack, and the end-users who will be interacting with the data. If you wish to do broad research across large portions of the data set, a bulk file download may be the best solution for you because it would allow you to download the largest amount of data at one time.

If your use case involves displaying the data in an application to your end-users, an API may be the right solution for you.

Finally, if your use case involves end-users who are not very technology savvy, a GUI display solution may be more appropriate.

If you're downloading data, there are several technologies that data providers offer (e.g. ftp, REST API, python, java, XML, etc.). So be sure to ask what is possible.

AVAILABLE FILE FORMATS

The file format that data providers use to deliver their data is usually determined by the nature of the data.

Structured data is highly organized and fits nicely into rows and columns. So it is easy to search and analyze using readily available computing tools. Examples of structured data are dates, names, identifiers (e.g. product

SKUs), transaction information, and so forth. This data lends itself to being delivered in a set file format. Often structured data is delivered in a delimited file format (e.g. .CSV, .TXT).

Unstructured data is much more common and does not fit nicely into any one format. So such data is often stored in its native formats. A challenge of unstructured data is that traditional methods and tools can’t be used to analyze and process it. Examples include such information as text files, PDF documents, social media posts, comments, images, audio/video files, and emails.

Be sure to confirm exactly what file formats are available from the data provider.

STORAGE REQUIREMENTS

Another question to ask potential providers is the file storage requirements for their data. You’ll want to understand how large the historical files are, as well as the current and predicted size of the daily update files going forward.

Some data sets can get very large. Many estimate that by 2025 there will be 175 zettabytes of data in the world. That is 175 trillion gigabytes! Obviously at some point it becomes unrealistic for every customer of a data provider to maintain their own copy of the data set.

Rather than storing all of the data in your infrastructure, a Data-as-a-Service (DaaS) solution might make more sense for you. With DaaS, the data is stored and maintained in the cloud by the data provider. You are given access to the data, often with scalable computing resources and analysis tools so you can ask your questions of the data and download the answers. Many DaaS solutions also allow you to upload other data to the cloud to add to your analysis.

DATA ANALYSIS TOOLS & EXPERTISE

Structured data sets that are provided in delimited file formats lend themselves to be analyzed with databases or other commonly used software (e.g. Excel®). However, not all types of data are so easily analyzed.

For example, satellite and aerial data sets are

Data-as-a-Service (DaaS) solutions store and maintain data in the cloud with scalable computing resources and analysis tools.

often intended for use with specialized remote sensing software with specific applications (e.g. spectral analysis). So you'll need such software and expertise to effectively analyze such data.

The data provider should be able to tell you what software you'll need to best interact with their data. They can also shed light on how to extract "nuggets" of knowledge from their data.

INSIDER TIP: It is common practice to leave the evaluation of new data sets to less experienced data analysts. No one wants to ask senior data scientists to stop the important work they are doing to evaluate some new data that you may or may not end up using.

However, depending on the type of data, it may require a very skillful data scientist to unlock the potential of the data for your organization. That is, the answers to the important questions you hope to ask the data may not be obvious to a more junior data analyst. So you need to nail down who will be evaluating any new data sets, and don't hesitate to ask the data provider to help you uncover the value of their data.

HOW TO GET SUPPORT

As was covered earlier, no data set is 100% perfect. You will find issues with almost every data set. So you need to know who to contact if you have a question, and how soon your issue can be resolved.

Also, how does the data provider handle additions and corrections to their data sets? Do they notify all clients when they find and correct an omission or error in their data? Do they provide a file with the correction, or do they force you to download the entire data set?

INSIDER TIP: Most data providers provide support during normal business hours via phone or email. If your use case requires more support than that, be sure to raise this with the data provider as soon as possible. They may be able to make alternative arrangements for you (e.g. having a support person on-call should an issue arise outside of normal business hours).

HOW CAN THE DATA BE USED

The old saying, "The devil is in the details" is very true when it comes to data. Not only can issues crop up within the data itself, or the delivery of the data, but also on the contractual side.

To avoid any potential snags at the end of the procurement process, I recommend you ask for a copy of the data agreement up front. Few people enjoy reading legal contracts, but you absolutely should read the entire document.

INSIDER TIP: I am NOT an attorney, and I *strongly* recommend you have an attorney review any data agreement before you sign it.

However, here are some specific things to look for in data agreements:

- Does the data provider confirm it has the right to redistribute the data in question?
- Who may view or use the data? Can users be in different locations?
- Are there any restrictions on how the data may be used? Is your use case permitted?

- What about any derived works you create from the data? Can you share or even sell such derived works?
- Are you required to secure data in any special ways (e.g. issuing usernames and passwords to allow an end-user access to the data)?
- Are there any audit provisions in the agreement that allow the data provider to confirm you are abiding with the contract terms (these are very common in many industries and usually non-negotiable)?

WHAT IF YOU CANCEL

There may come a time in the future when you no longer need the data and wish to cancel your subscription. We have seen an uptick in the number of data providers who have "Purge Clauses" in their data agreements. These say that if you cancel your subscription for any reason, you must delete and/or return ALL of the data you have already received – even though you've paid for the data.

That may sound outrageous. However, such clauses have become pretty common, especially in the financial data market.

The reasoning given for such clauses is you do NOT own the data, but rather are licensing it. The sources of the raw data own the data. So if you cancel your license, you should no longer have access to the data.

To add insult to injury, there is often a buyout portion to Purge Clauses whereby you can pay a cash amount to retain the right to keep the data you've already received and paid for. If that sounds like paying a ransom, well… no need to say more. Try to avoid purge clauses.

About Data In Harmony (DIH)

Data In Harmony (DIH) provides financial data for multiple asset classes and markets around the world. We help firms get better results with complete and accurate data that is still as raw as possible. Our data experts ensure our clients get the data they need, reliably delivered under user-friendly license terms, and for a price that fits their budget. We hope this guide has been helpful. If you have any questions, please to contact us:

support@datainharmony.com USA +1 512 333 2686 UK +44 20 3287 1280

Copyright © DIH Solutions, LLC. All rights reserved. v2022.08