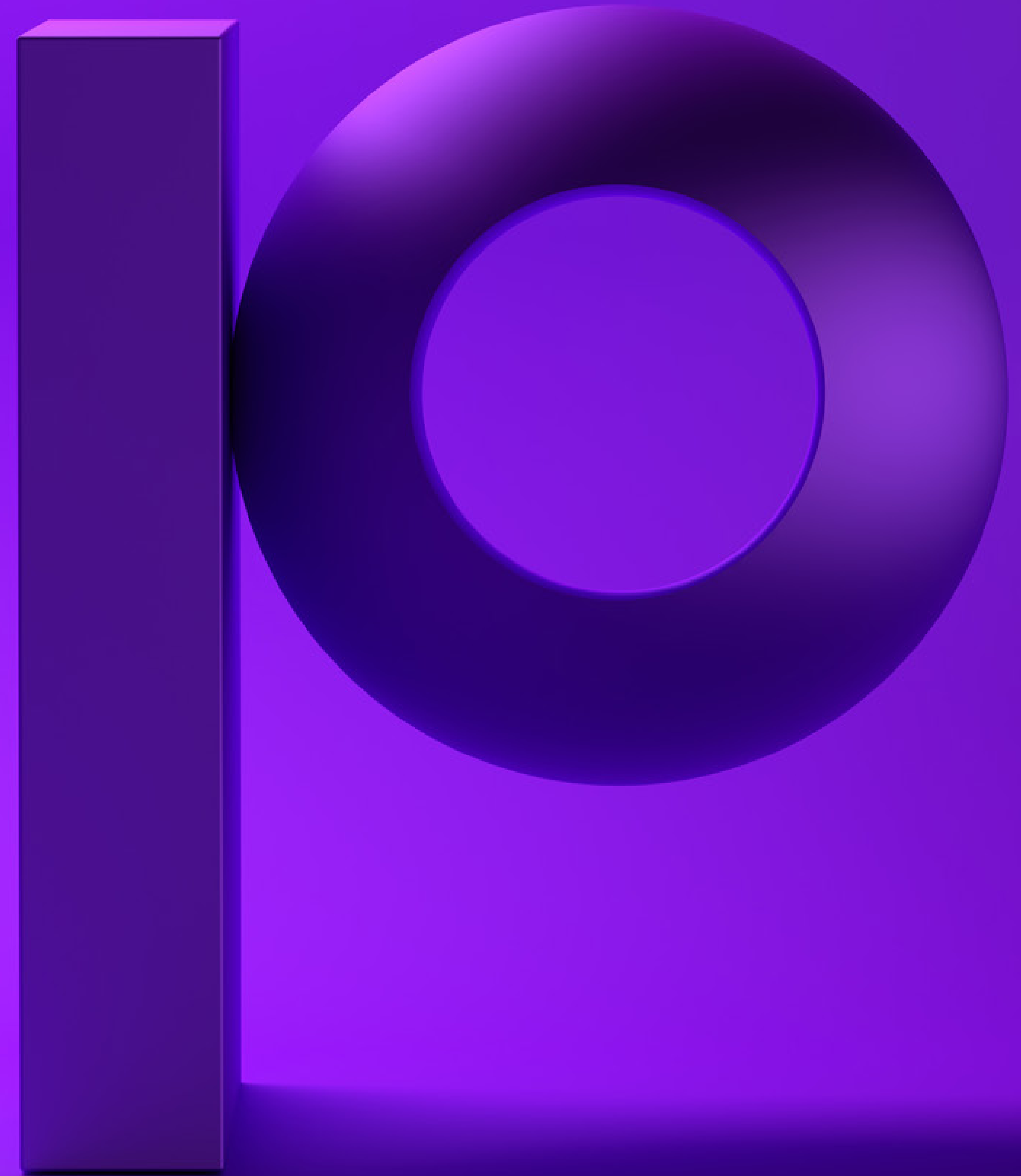precisely

Six Steps to Sourcing Enrichment Data for Your Business

Introduction

Data is the driving force behind business today. On their own, companies generate vast quantities of data — called first-party data — about their customers and their operations. But businesses also depend on external third-party data to learn more about their customers so they can create rich omnichannel marketing campaigns and help sellers take the next best action. Or, for example, they use external data to help evaluate business risk, ensure compliance, make informed planning decisions, or select the right location for stores, restaurants, and infrastructure.

The uses of enrichment data are virtually unlimited. The challenge is finding the right data and data sources for meeting your business goals. Whether you are using third-party data for business intelligence dashboards, problem-solving, and analytics, or AI/ML applications, your results will depend on the quality of the data you use. In this eBook, we cover the six steps to successfully sourcing third-party enrichment data for your business:

1. Understand the use case for the data
2. Determine what data you will need
3. Identify potential data sources
4. Evaluate your top candidates
5. Understand terms and conditions
6. Consider data delivery methods

STEP 01
Understand the use case for the data

While this requirement may seem self-evident, in many organizations the person acquiring the data is several steps away from the person or team that will actually use it. The buying decision may be made by purchasing, for example, or the IT organization. The farther away the decision-maker is from the project owners, the more important this step is for requirements gathering.

The intended use could be a proof-of-concept project, for example, where fast availability and low cost are more important than long-term availability and the age of the data. Volume can be the key factor in data that will be used to train machine learning (ML) models. If the use case is for business-critical decision support, however, data quality and frequent updates are likely to be the top requirements.

It is also important to understand the end-use of the data. Will it be used for internal analysis, for example, or will you be publishing the results? Will the data be available to the public, directly or indirectly? Are you going to combine it with your own data and resell it as a data monetization strategy?

All these factors will come into play in making data acquisition decisions.

Key takeaway:

The best practice is to gather requirements from the development team, the intended end-users, and the data professionals in your organization to ensure you get data that is fit for use.

STEP 02
Determine what data you need

The next step is to determine what data points you need for your use case. For example, your use case may be to build a consumer web application to help people identify the right neighborhood for their next home purchase. You would start with location data, in particular neighborhood boundaries, because they are much more precise than postal codes. You'd need street data to understand accessibility as well as proximity to major highways. Geospatial data enables you to display data on maps. And you'd want to know what schools, businesses, parks, entertainment, malls, government buildings, and the like — called points of interest — are close to each neighborhood.

You'd need historical property purchase and economic data to calculate average home prices and living costs for the neighborhood.

And you would need demographic and consumer segmentation and demographic data such as income, homeownership, age range, interests, and the like to create a profile of each neighborhood.

The takeaway for this step is that it is important at this early stage to brainstorm and dig deeper into what your end goals are and what you need to know to get there. It's also important to understand whether this is structured data, unstructured, or a combination. Any gaps that you leave in determining data requirements will be much harder to address once software and databases have been designed.

Key takeaway:
Brainstorm and dig deeper into what your end goals are and what you need to know to meet those goals.

National or global scope?

At this step, you also will want to assess geographical requirements. Once you step outside of any one country for data sourcing, you will find the ensuing steps more complex. Availability of the same or similar datasets across other countries is likely to vary widely. If the data is available, it is likely to be structured differently. And it will most likely have different terms and conditions that may or may not fit your use case.

27

The average number of third-party data sources an enterprise currently has integrated with its data architecture.[1]

[1] Data Integrity Trends 2021, Corinium Research, 2021

STEP 03
Identify potential data sources

This step may be your most difficult because there are thousands of different sources for third-party data. For starters, it helps to understand the different categories of data providers, with the caveat that there can be a significant crossover between categories. These include:

- **Primary data sources.** These organizations perform the research or gather the data and make it available for external use, including academics and universities, government agencies, NGOs, scientific organizations, industry associations, research firms, and think tanks. They can also be organizations that source data from the internet using a process called "web scraping" that uses a software tool to extract data from a web page. Primary source data is also known as "raw data."

- **Public use and open-source data.** Public use or open-source data is typically the data that is available at no charge, and in many cases, it comes from the primary data sources. Open-source data certainly has a place in data sourcing for commercial use, but "free" doesn't mean "no cost." Incorrect use of collected data, missing data, or inaccurate statistical inferences can be problematic. And the use of data from open-source projects may include requirements to "give back" any improvements you make to the data.

- **Data aggregators.** A data aggregator is any organization that collects data from multiple sources, adds some value by processing it, combines it into a new dataset, and repackages the result in a usable form.

- **Data providers and single-source data marketplaces.** Data providers such as Precisely practice data aggregation on a large scale to produce marketable data products. These data products are designed for commercial use, with value-add characteristics such as standardized formatting for easy integration, quality assurance, data governance, reliable metadata, and product documentation. For convenience and a good customer experience, data products are offered through a marketplace environment, the [Precisely Data Experience.](#)

- **Data marketplaces.** Data marketplaces such as the [Snowflake Data Marketplace](#) and [AWS Data Exchange](#) bring market-relevant products and services of multiple data providers together in a single, searchable environment. There are also data marketplaces targeted for specific industries or purposes, such as segmented consumer audience data.

How do you find all these data sources?

Visiting a data marketplace or doing a Google search are great places to start. You can search by type of provider, by providers for your industry, or data category (think "weather" or "agriculture"). And you can take a deeper dive into understanding the various types of data sources to help in your decision process.

Key takeaway:

Criteria for evaluating data providers and data marketplaces include:

- **Comprehensiveness of data products** — Can the provider or the marketplace meet all your data needs from a single source?

- **Sample datasets** — Are data samples available for download and evaluation?

- **Online software tools** — Can you explore datasets online, for example using a mapping application?

- **Ease of use** — How precise is the search function, and is it easy to navigate the site?

STEP 04
Narrow your list of candidates and assess them

Whether dealing with open-source data or a commercial data provider, you will want to evaluate each candidate carefully. This can be a time-consuming process — particularly when you consider that most use cases will require multiple datasets that may come from more than one provider.

Precisely works with more than 130 data suppliers. Here are the requirements we expect them to meet because we know these factors are essential to our customers:

- **Data quality.** Data is correct (correct values, clean geometry, and no invalid duplicates), complete (acceptable coverage, feature completeness, and no missing values), and current (the data is fresh and will be updated appropriately).

- **Data structure.** The product structure, including download location, folder structure, files and file names, and field names) is consistent with previous downloads.

- **Structure changes.** The supplier provides at least 90-days advance notice of any planned changes.

- **Documentation and metadata.** The supplier product documentation includes field layouts, definitions, product metrics (file size, record counts), release notes, age of the data, and sources (primary, secondary, or tertiary).

- **Effective issue resolution.** With large datasets, issues are not uncommon. Is there an established method of communicating problems? How quickly will the supplier provide and implement solutions?

- **Product timing.** The data supplier has a published, consistent schedule for updates and a product roadmap for content improvement.

As a rule of thumb, the closer you are to raw data, the more critical this assessment process is. With raw or open-source data, you want to make sure the dataset is stable, well maintained, and meets quality standards. For commercial data products, including those sourced from a marketplace, you should still perform due diligence to ensure that the data provider requires criteria like this of its suppliers.

Key takeaway:

The closer your sources are to raw or open-source data, the more critical it is to determine that the dataset is stable, well maintained, and meets quality standards.

> "Along with most organizations, we are prioritizing being able to trust our data more and focusing on key strategic pillars like data quality."
> — Gladwin Mendez, Data Officer, Fisher Funds[2]

[2] Data Integrity Trends 2021, Corinium Research, 2021

STEP 05
Understand terms, conditions, and compliance risks

This step is a crucial part of the assessment process, and we are calling it out as a separate step because of its importance. The terms and conditions associated with any third-party data will determine its fitness for your use case. This includes open-source data. Legal terms and conditions specify what use of the data is allowed and under what circumstances.

Key takeaway:

Terms and conditions specify what you can and cannot do with third-party data under what circumstances. Make sure your use case is allowable before you invest in data.

> **"Companies are enriching their own data with data from third-party sources. But most say doing this consistently at scale is challenging."** [3]

[3] Data Integrity Trends 2021, Corinium Research, 2021

Risk factors

The risk of not doing the due diligence at this step is that you find out, after the fact, that you cannot do what you intended with the data you've acquired. Even worse is finding out after you've done it, and you find yourself the liable party in a lawsuit. Misusing primary data in a way that violates the primary data source's intellectual property is another risk. For that reason, larger organizations, in particular, will want their legal departments to participate in this assessment.

Compliance concerns

Compliance should be a consideration when using third-party data, particularly when customer or consumer data is involved. You want to know if a dataset will include any personally identifiable information (PII) and, if so, who will be responsible for anonymizing it. Also, discover if the data was collected in a way that adheres to industry standards like the Open Financial Exchange (OFX) standard or governmental regulations such as the General Data Production Regulation (GDPR) in the EU or the California Consumer Privacy Act. Data sourced by screen scraping, without the consent of either individuals or the content provider, is a particular risk factor when it comes to compliance.

STEP 06
Consider data delivery methods

As a final step in sourcing data, consider the many ways that it can be delivered and determine what makes the most sense operationally for your organization and your use case.

Providers may use an FTP site, a cloud storage site, or a web page to make data available for download. Data marketplaces like Snowflake Data Marketplace and AWS Data Exchange are designed to simplify data delivery and allow you to access multiple data providers, including Precisely, through one interface.

Another consideration is the file format the data is delivered in, such as a CSV file, an ASCII file, an extended tab format, a database file (Microsoft Access, MySQL), or a delimited text file. File format is likely to be determined more by the type of data than the data provider.

More mature data consumers should consider finding a provider that offers an application programming interface (API) or software developer's toolkit (SDK) that can streamline and automate data acquisition.

Key takeaway:
Confirm that the data you intend to purchase will be delivered by a method and in a format that you can easily use.

Summary

The world is awash in data, and the critical path to finding, evaluating, and acquiring it is the internet. Options range from simply downloading and using free open-source or public-use data files — often a "buyer beware" exercise — to purchasing data from one or more established commercial providers. Data marketplaces are emerging to simplify the search process by gathering multiple commercial data providers and products under a single digital umbrella.

Working with Precisely can help save you time and money. Your data professionals spend about 80 percent of their time finding, prepping, and managing data.[4] That means only 20 percent of their time is spent applying data to your business operations—that includes running models to understand risk and opportunities, improving efficiencies, improving customer experience and retention, and all of the other use cases that you hired them for.

Our data products are easy to find on the Precisely Data Experience. Moreover, we have already assessed the suppliers of data that goes into our data products. We hold our suppliers to high standards for data quality, freshness, and documentation as well as governance, provenance, terms and conditions, and compliance. That means you can focus on meeting your business goals with data, data is ready for your professionals to use with minimal prepping.

There are many data providers in the marketplace. Choose Precisely and make more confident business decisions based on data you can trust.

> **"If people don't trust the insights, they're not going to act on them, especially when the insights conflict with their so-called gut reaction."**
> — Dan Power, MD of Data Governance, Global Markets, State Street[5]

[4] The 80/20 Data Science Dilemma, IDG/InfoWorld, 2021.

[5] Data Integrity Trends 2021, Corinium Research, 2021

precisely

About Precisely

Precisely is the global leader in data integrity, providing accuracy and consistency in data for 12,000 customers in more than 100 countries, including 97 of the Fortune 100. Precisely's data integration, data quality, data governance, location intelligence, and data enrichment products power better business decisions to create better outcomes. Learn more at www.precisely.com.

www.precisely.com