URBAN LOGIQ

Data Evaluation Checklist and Guide for New Data Acquisition

From UrbanLogiq ➡

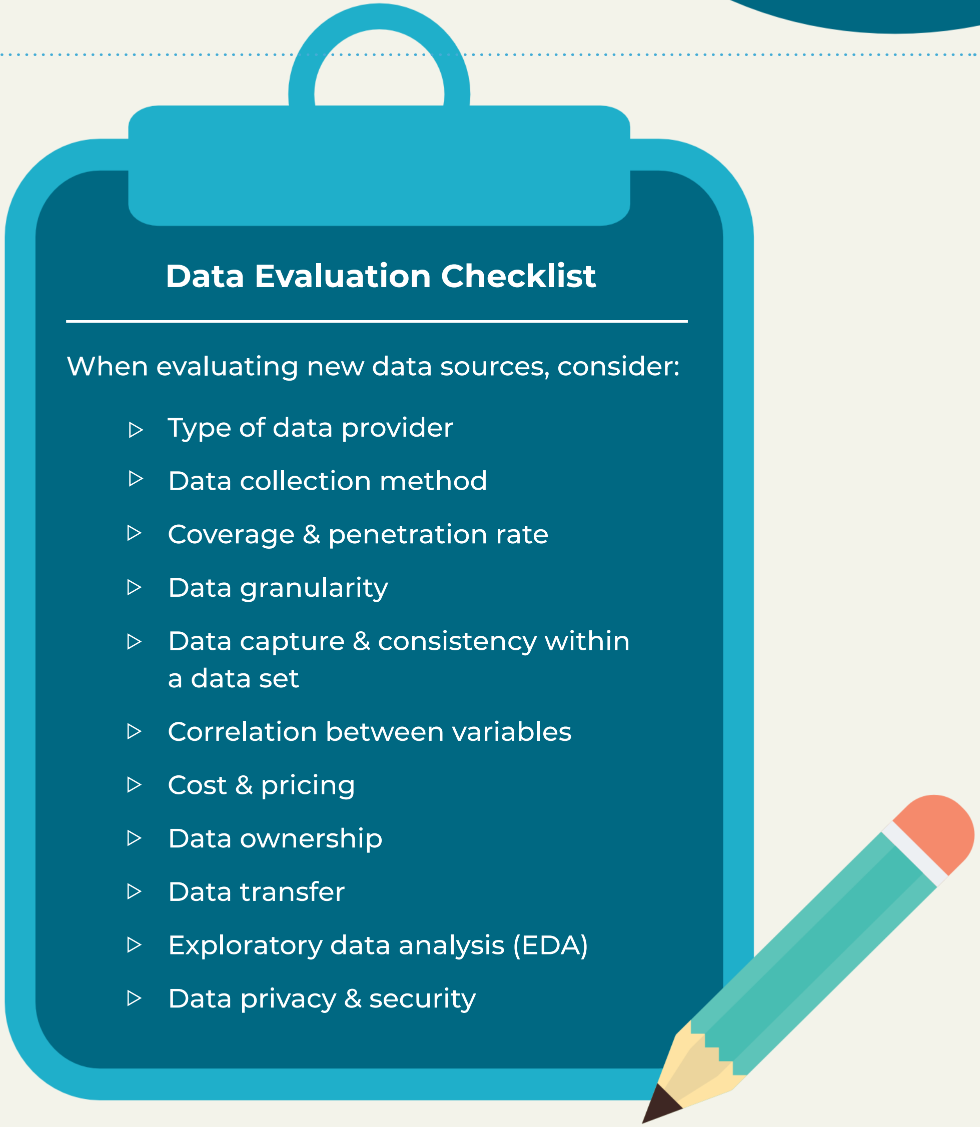
Contents

Introduction

From our blog post, "How to Write a Data Acquisition Strategy: An In-Depth Guide"

Within the public sector, there is an ever-increasing need for data to support decision making. Despite this fact, many public officials are unfamiliar with the processes involved in evaluating, purchasing, and storing new data assets in formats that can be easily accessed and applied to use cases in the future. The sheer quantity of data sources on the market, combined with the number of evaluative factors to consider can make the process feel quite daunting.

We designed this checklist to help guide government officials through the data evaluation process safely and efficiently. Whether your agency is just starting to experiment with emerging data sources, or simply looking to improve the overall management of existing data assets, this guide is sure to help.



Data Evaluation Checklist

When evaluating new data sources, consider:

- Type of data provider
- Data collection method
- Coverage & penetration rate
- Data granularity
- Data capture & consistency within a data set
- Correlation between variables
- Cost & pricing
- Data ownership
- Data transfer
- Exploratory data analysis (EDA)
- Data privacy & security

✓ Types of data providers

Tips for evaluating the different types of data providers on the market today.

There are several different **types of data providers** on the market today that generally fall into two categories:

Those that generate their own data
This includes telecom providers, vehicle manufacturers, sensor manufacturers and others who generate and own their own data. These companies sell their data to consumers in a raw or aggregated form.

Those that purchase external sources for resale
Otherwise known as the Data Marketplace, this includes companies that purchase data from multiple sources to amalgamate and resell. This data is generally sold to consumers in an aggregated form or in a package of preprocessed insights (as opposed to raw).

Clarifying who owns supplier relations at your organization will help the evaluation process stay organized.

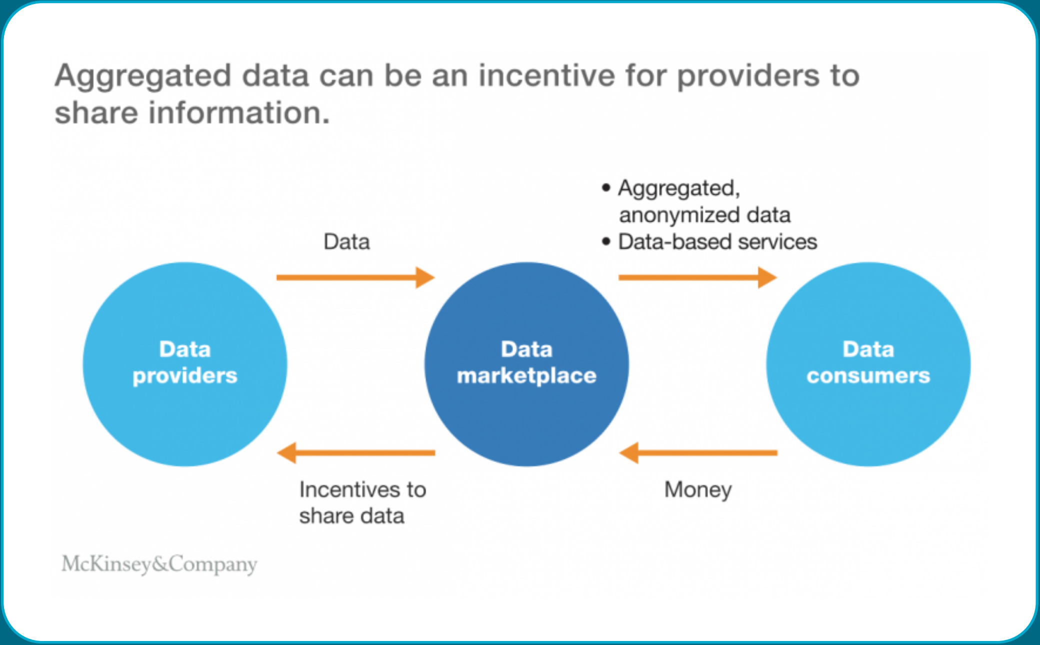
Also, thinking about data providers like partners can help you to build a strong foundation of trust and transparency from the start.

Note! *Data from Marketplace providers can be more expensive due to the effort involved in sourcing and storing the data.*

Most Marketplace providers will set their prices higher for any additional processing work performed on the data. In these cases, you must trust that the provider has accurately processed the data and extrapolated any insights or statistics.

Keep in mind that data providers may be reluctant to discuss how they arrived at the insights they provide, as their methods may be proprietary.

Consumers willing to pay money for aggregated data provides incentives for data providers to share information with consumers



Aggregated data can be an incentive for providers to share information.

McKinsey&Company

Source: McKinsey & Company

Aggregated data can be an incentive for providers to share information

Note! Data from providers exists on a spectrum going from very raw data (for instance, one row=one device ping) to highly aggregated (for instance one row= one day of traffic counts between two regions).

Consider whether you want to purchase your data pre-processed, or whether you would rather undertake this work yourself.

Data collection methods

An overview of the predominant types of data collection methods.

Data collection methods will vary depending on the type of data and the data provider. Some examples of common methods include:

- **Manual data collection** (interviews, surveys, questionnaires, etc.)
- **Sensor technology** (including Bluetooth, IoT, GPS, and other physical hardware such as induction loops or pneumatic tubes)
- **Location-based services (LBS) data**
- Data scraping

Understanding different data collection methods will better prepare you to ask clarifying questions throughout the evaluation process.

✓ Data coverage and penetration rate

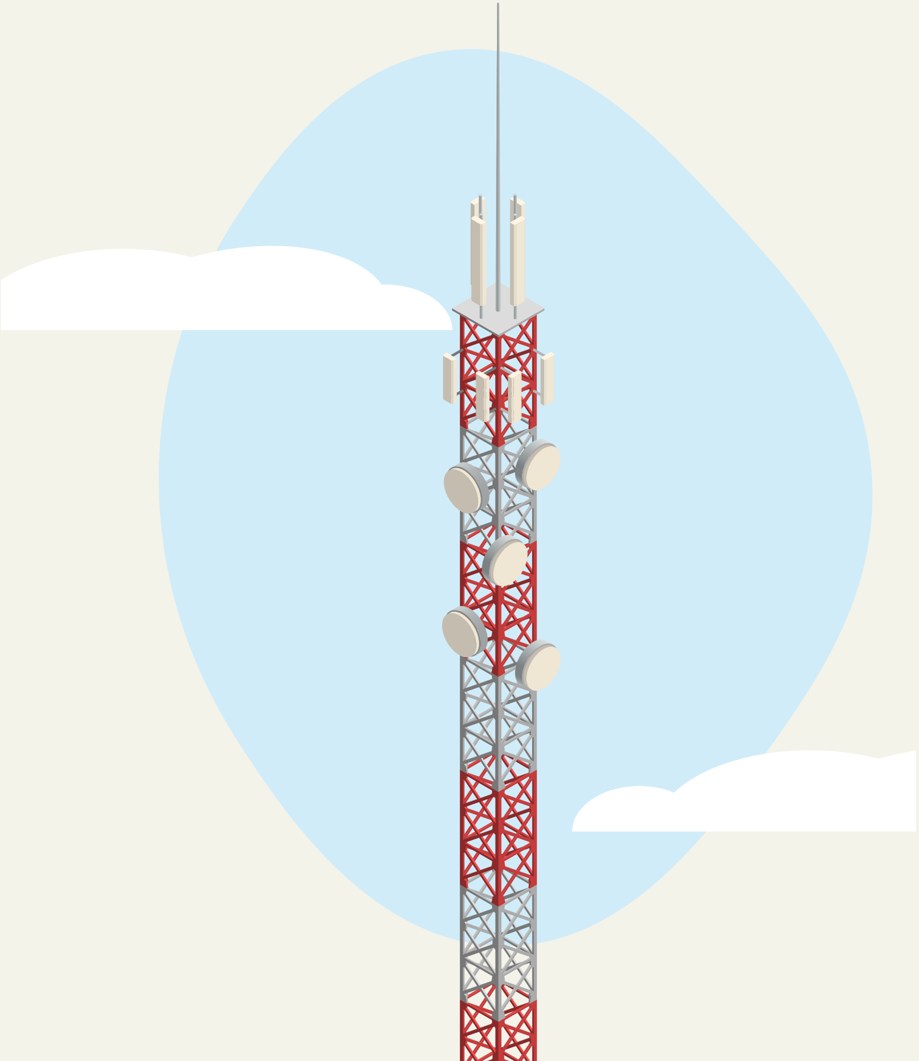
A brief overview of data coverage and penetration rate, including important questions to ask a data provider.

Data coverage refers to the total scope of the information collected from a specific sample group across time (how far back) and space (across what geographic region).

Many factors impact data coverage: from population density to the market share of different data suppliers, to the maturity of certain technology.

When considering coverage, keep in mind that very low statistics (for example data collected in rural and remote areas) can make it too easy to identify individuals. As a result, those who perform the data analysis may exclude this data from the set altogether rather than perform complex anonymization.

This can ultimately result in the underrepresentation of certain groups or regions in the data.



URBAN LOGIQ

✓ Data coverage and penetration rate cont.

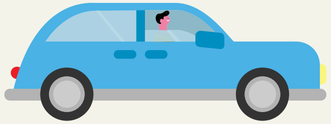
A brief overview of data coverage and penetration rate, including important questions to ask a data provider, continued.

Here are a few questions to ask a data provider:

- What data collection methods do you use?
- Is your data specific to a certain county or geographic region?
- Are there certain regions in which your data coverage is stronger or weaker?
- When is the data collected (i.e every time a user opens an application), and how far back does it date?
- What percentage of the population does your data represent?
- Is there any personally identifiable information (PII) in this data set?

Note! *When considering data coverage, keep in mind that very low statistics can make it too easy to identify individuals.*

As a result, those who perform the data analysis may exclude this data from the set altogether rather than perform complex anonymization which could lead to underrepresentation.



URBAN LOGIQ

“The most common question about emerging data sets is whether the data represents sufficient penetration rate or sample size to produce reliable insights.

The question in itself is a really good one and raises very important considerations that come with using third-party data effectively. Knowing the scope of your use case you intend to use the emerging data for is very important. Different use cases will have different error tolerances, which means certain penetration rates might influence results in your use case.

Julien Refour

Chief Data Scientist, UrbanLogiq

✓ Data granularity

What you need to know about the granularity of a data set.

Data granularity refers to the degree of detail contained in a data set, and how well-defined the data fields are.

Some projects or analyses will require more granular data or data with specific attributes. Note that with a greater granularity of data, the risk of exposure to personally identifiable information increases, and therefore the extent of de-identification and anonymisation needed could increase as well.

✓ Data capture and consistency within a data set

Quick facts to help you understand data capture and consistency within a data set.

Data capture refers to the cadence of data collection for any given source. For example, census data is only collected every couple of years, whereas vital statistics such as marriages, births, or deaths are recorded continuously.

When evaluating a data set, inquire as to whether the data is collected periodically, sequentially, continuously, etc. This information, combined with latency (the time between when data is created by a source and the time it is available for end-users for analysis) will inform the recency and potential accuracy of your analysis.

It is also valuable to know whether the data is static (unchanging once recorded) or dynamic (subject to change after being recorded).

Consistency within a data set refers to the consistency in data collection or transmittal. Consider ping rates: vehicles send out a point, or ping, every few seconds. Some vehicles may ping every 3-5 seconds, but others every 2-4 seconds. Perhaps the manufacturer collects the data from these vehicles every month except for November.

These types of inconsistencies will affect your analysis, and it's important to be aware of them.

✓ Correlation between data variables

Why it's important to understand any correlation between variables.

Correlation between variables refers to the relationships within a given data set. Changes in one variable are consistently associated with changes in another variable.

Consider building footprints, for example. The number of floors in a given building may correlate with the height of that building.

Determining correlation helps illuminate how the data is constructed, and whether any variables should be omitted or transformed during analysis.

> *"Determining correlation helps illuminate how the data is constructed, and whether any variables should be omitted or transformed during analysis."*

Cost and pricing

Learn the three primary influencers of data cost.

In our eBook, "The Definitive Guide to Mobility Data Sources for Government" we explain that the three primary influencers of data **cost and pricing** are volume, velocity, and variety. Here is a brief description of each:

Volume: Determined by the geographic region you require data for, how frequently you need it, and for what time period.

Velocity: Dictated by how fast the data will be consumed i.e in a real-time stream or by historical dumps.

Variety: Determined by the number of different attributes you require to be delivered in the data set. Certain vendors may offer deals where the data is transferred in bulk at a discounted rate.

Download our free eBook to learn more on this topic.

URBANLOGIQ PRESENTS:
The Definitive Guide to Mobility Data Sources for Government

The Definitive Guide to Mobility Data Source for Government

Download Now

✓ Data ownership

Data ownership will vary depending on the type of transaction; here's what you need to know.

Data ownership clauses in contracts are varied and nuanced and generally fall into the different buckets below:

- **Complete ownership** of the data and any derived insights (rare)
- **Partial ownership** of the data for a defined time period and any derived insights
- **Partial ownership** of the data for a defined time period but shared ownership of any derived insights
- **Shared ownership** with neither able to do anything with the insights without the other's permission

For most data purchases, the purchaser cannot resell the data or send it to other parties outside of their organization.

The raw data will also generally need to be deleted after a period specified in the contract.

Clarifying the terms of ownership and how you are allowed to use derived insights before purchase will eliminate headaches in the long term.

✓ Data transfer

Key things to know about data transfer before you make a purchase.

Data transfer refers to how you will actually receive raw data from the selected vendor, including the method of transfer (such as FTP or API), as well as the file format you receive it in.

Note that some file formats will be easier to transform into useful information than others, so consider the data science and data engineering resources you have available at your organization.

Certain data formats will help decrease file size, saving you storage room. Non-standard formats can make it difficult to read and use the data you receive and may require more work than other file formats.

The best solution depends on the needs of your organization and how you plan to integrate the data with your existing databases.

For instance, if you are purchasing data that is similar to data you have purchased in the past, ensuring that the formats are compatible will eliminate the need for manual manipulation and reformatting.

URBAN LOGIQ

Exploratory data analysis (EDA)

A brief overview of what an exploratory data analysis entails.

A critical step in the data evaluation process is to receive sample data on which to perform an **EDA**.

The EDA is performed on data to determine the types of information present, whether there are any missing values, and whether there are any obvious outliers or correlations. The process often employs statistical techniques to uncover patterns, relationships and trends in data that might otherwise be overlooked.

Your EDA may lead to additional questions, hypotheses or other areas of interest which are worth further investigation. Taking your time to be thorough during this step will help you down the road when you need to reuse this data for additional purposes.

An exploratory analysis also provides the opportunity to validate the data and ensure it accurately reflects your community. Validation using ground source truth data, such as highly accurate traditional sources (census data, traffic counts), will give you a much better idea of how complete the data set is and help identify any limitations.

Here are a few specific things to look out for at this step:

- **What's present in the data set?** i.e which columns, units
- **Missing data**
- **Coverage** (in time and space) as well as coverage gaps
- **Data bias**
- **Correlations** (note the discussion on 'correlation between variables' above)
- **Privacy concerns** i.e whether the data contains PII or partial PII

Data privacy & security

Key data privacy and security considerations for your strategy.

Baking **data privacy** considerations into your data acquisition strategy is vital to mitigating risk and fostering a culture of critical thinking. Below are a few actionable tips:

- **Form a Data Governance Board** to plan out policies for handling privacy issues (or take lessons from others)
- **Write clear instructions** surrounding access control of all data
- **Write a clear policy** for handling PII

With regards to **data security**, requiring all IT vendors to provide proof of third-party certification of their security and privacy procedures is a good place to start. For example, UrbanLogiq's extensive work with emerging data sets has led us to seek certification in the world's leading standards in cybersecurity and data privacy compliance. As a result, our team is able to help agencies navigate the application of big data responsibly, transparently, and ethically.

Below are a few actionable tips for public officials considering data security in their data acquisition strategies:

- **Know your ISO 2700 series and NIST 800 series**
- **Join an ISACS group** to share threat information and best practices
- **Form alliances with neighboring municipalities** to cooperatively procure cybersecurity services
- **Schedule regular security conversations** with your procurement partners (both existing and new)
- **Bake security terms into contracts** with IT providers, i.e. specify that you must be notified of any security breaches
- **Proactively communicate** your data management privacy policies to data providers

URBAN LOGIQ

About UrbanLogiq

UrbanLogiq is a software platform that integrates and visualizes data to provide fast insights for government. We work with agencies ranging from municipalities under 100,000 in population, to regional governments, states and provinces, and many agencies in-between. As a data agnostic platform, we integrate diverse data sets to produce insights and metrics configured to department-specific needs.

By centralizing data from multiple sources and combining them into one easy to use geospatial interface, UrbanLogiq enables public officials to make faster, more affordable and accurate decisions. We provide streamlined data processing and integration, meaning agencies do not have to spend time data wrangling and cleaning; instead, they can quickly and easily access all the data they need without any of the headache.

Through our work with governments across North America, we've been recognized on the inaugural list of GovTech Magazine's Best International Technology Companies, named a favorite company by TechCrunch, and one of the top five smart city companies to watch by TechNYC.

Stay connected

f facebook.com/urbanlogiq/
twitter.com/urbanlogiq
in linkedin.com/company/urbanlogiq

Additional resources:

- ➡ How to Write a Data Acquisition Strategy: An In-Depth Guide
- ➡ The Definitive Guide to Mobility Data Sources for Government
- ➡ Data Engineering 101: A Guide for Government Officials
- ➡ The Role of Emerging Data in Traffic Data Collection Programs

URBAN LOGIQ

Building better communities with data