CENTER FOR DEMOCRACY & TECHNOLOGY

What is a Database?

A Technical Guide to the Personal and Institutional Considerations of Government Data Consolidation

July 2026

Authored by
Hannah Quay-de la Vallee, *Senior Technologist*

Limited and targeted data sharing between agencies has long been understood as a way to reduce administrative burdens (on both administrators and individuals interacting with government); improve service coordination; and assist in preventing fraud, waste, and abuse. However, the Trump Administration has made unprecedented moves to go beyond such targeted sharing and consolidate the various collections of data held by both federal, and state and local agencies despite decades of bipartisan consensus against the federal government amassing personal data on Americans. This large-scale data consolidation poses risks to personal privacy, government surveillance, and cybersecurity vulnerabilities.

These developments raise fundamental questions about how to know when the risks of data consolidation outweigh the benefits. Understanding how specific risks relate to technical design choices is important to making this assessment. Therefore, this explainer provides background on government data consolidation, and more details about the risks and benefits of consolidations as well as the conditions that can enable or discourage consolidation. Using these core concepts as its foundation, the explainer lays out the technical elements of consolidating data and how they can contribute to or ameliorate specific risks to Americans.

Background on Government Data Consolidation

At its core, data consolidation is mingling and combining data from multiple different sources to allow it to be analyzed or viewed as a whole. The federal government's most recent push for consolidation and repurposing has intensified public debate about privacy

and data security, and led to over 20 lawsuits against various federal agencies alleging that the government has violated long-standing privacy laws designed to protect personal information. It has also led to allegations that the federal government is building a "master database," a centralized collection of an enormous amount of data about Americans and others. In turn, the government has insisted that no such master database exists.

This ultimately raises a question of what constitutes a "master database." There are different methods of data consolidation, namely warehousing, where all data is stored in a central location, and federation, where data is stored in disparate places but users across a range of organizations are able to access it. The former is very clearly a "master database," but the latter provides much of the same functionality. Trying to delineate the two is largely a distinction without a difference for many of the concerns (e.g. privacy and security risks, discussed in more detail below) raised with respect to government data consolidation in the current context.

Why Consolidate Personal Data, and Risks to Consider

Governments of all levels might choose to consolidate personal data for a number of reasons. Some of these reasons seem quite banal, while others are far more controversial, and still others verge on totalitarian:

- **Improve access to and quality of government services** by assessing the performance of programs, shortening review timelines, and reducing application burden (as when programs share data about applicants, allowing them to access multiple benefits programs without completing full applications for each).
- **Fraud detection and management** is another common reason for agencies to share data, such as ensuring that an applicant provides consistent information across programs or that people found to have committed benefits fraud against one agency are removed from other programs (though the specific approach and implementation of fraud management can vary widely, with commensurate variation in risks).
- **Immigration enforcement** is a controversial purpose for data consolidation, because it is unrelated to the purpose for which the information was shared and can have chilling effects on benefits adoption, service access, and compliance (e.g., payment of taxes). Data can assist in immigration enforcement when it is used for purposes such as identifying immigrants and locating immigrants.
- **Surveillance and intimidation** is enabled by this sort of consolidated information and gives the Executive Branch the power to suppress dissenting voices and chill speech and protest. Such expansions of executive power are rarely clawed back, which provides a laid table for not just this administration, but the unpredictable actions of future administrations as well.

Data consolidation also carries numerous risks, impacting both the people whose data is implicated as well as the governments responsible for consolidating that data:

- **Running afoul of laws and regulations** such as the Privacy Act of 1974 and the E-Government Act of 2002 that protect the data held by the federal government. Data sharing actions that do not comport with these frameworks leads to litigation costs for both the government as well as those trying to ensure the government is complying with the law.
- **Risks of misuse** increase as the data is used for reasons further and further away from the purpose associated with the original collection, which increases the likelihood that those who provided the original data would be surprised or concerned with the new use.
- **Security risks** grow because the more ways there are to access a given piece of data, the higher the risk that that data will be breached, either through human error or a security flaw. Additionally, a consolidated store of data may be an enticement for hackers, leading to a higher number of hacking attempts.

Technical Considerations for Data Consolidation

A number of technical considerations during implementation will impact the capabilities, and thus the consequences, of a combined data system, including how the data is stored and maintained, and what it is used for. In particular, the technical considerations that inform levels of risk are:

- Warehousing vs. federation;
- Data quality, definitions, and matching; and
- Use of artificial intelligence (AI).

Each of these technical considerations affects different policy elements of a sharing program, and different decisions will allow agencies to achieve different specific policy outcomes.

Warehousing vs Federation

There are two main architectures when sharing data:

- **Warehousing**, where all shared data is stored in a centralized location that every party has access to; and
- **Federation**, where each organization maintains its own data and provisions access to the other organizations.

In one-way sharing, the analogous warehouse approach would be the sharing agency providing a copy of the data to the receiving agency, while the analogous federation approach would be the sharing agency providing access to their own version of the data to the receiving agency.

In the federal government's rush to consolidate data, it is not always clear which approach is being used, but pre-existing data sharing frameworks are a mix. For instance, the National Center for Education Statistics' (NCES) Common Core of Data is a warehouse – schools and districts share their data with NCES, which maintains its own database. On the other hand, Treasury's Do Not Pay database appears to be at least in part a federation model, accessing other datasets such as the Social Security Administration's Numident system.

Considerations for Warehousing vs. Federation

Warehousing and federation have different outcomes for who manages the data, how up-to-date the data is when accessed by various parties, and how easy it is to stop sharing data, all of which have important policy impacts. The following considerations tease out the implications of whether data is linked through warehousing or federation:

- **Determining who controls the data**
  In warehouse approaches, generally one party is ultimately responsible for maintaining and managing the centralized store and coordinating with the sharing parties. This can make access more straightforward, as access to the warehouse can provide the ability to view data from a range of sources. In a federated model, each party maintains control of its own data and provides receiving parties with some way of accessing that data. Often there still needs to be an "owner" of the sharing project to ensure effective responsibilities and roles for building the sharing infrastructure, but controlling the project is different from controlling the data itself. Federation can also complicate access, as each pair of sharing and receiving parties need their own relationships.

- **Keeping data up-to-date**
  Federation eliminates the need to keep data up to date, as receiving parties are accessing data from the original source, and will generally see updates as they happen. Warehouse approaches require infrastructure in place to ensure that the data is kept up to date, as the sharing agency's data changes over time. This means the sharing agency needs a way to share updates.

- **Process for ending sharing**
  The centralized nature of warehouses means that it can be challenging for sharing parties to "claw back" any data they have already shared, should they choose to stop sharing data. Ultimately, stopping sharing would mean deleting data from the shared warehouse, and that would rely on whoever controls the warehouse agreeing to delete the data. This can be ameliorated by providing sharing parties with some level of control over the data they have contributed (e.g., through a contract or agreement), but there would still need to be an ultimate "owner" of the database who would have the power to adjust these permissions. Federated models generally provide sharing parties

with more control over stopping sharing, as they can generally revoke any access they have granted to view their data (this does not preclude receiving parties from having copied the data, which then presents similar challenges as a warehouse model in that it requires the receiver to delete the data themselves).

- **Cybersecurity impacts**
  Centralized warehouses can add an additional cybersecurity burden, as they add an extra location where the data lives and there are more access points to this central point (since a breach in one agency could allow a hacker to exploit that agency's access to the centralized store to reach all the agencies' data). Typically this burden must be managed by all agencies feeding into the warehouse. The potential benefits to a hacker are also greater when accessing a warehouse as they can obtain more information at once, which creates a greater incentive to hack. In a federation context, each agency is generally responsible for the data that is in their care, whether that is their data or data they have received from another agency. Because this means that the sharing agency loses some amount of control over how their data is protected, data sharing agreements can lay out expectations for how the receiving agency will protect the data.

Data Quality, Definitions, and Matching

Another technical element that can impact the potential utility and risk level of consolidated data is the quality of the data itself, and what infrastructure is in place to ensure that data from different databases can be effectively compared. Many government databases are not particularly "clean," meaning the data they contain may be incorrect, incomplete, or in a non-standard format.

Irregularity and mismatches can cause numerous issues in data sharing programs. Irregularities or "workarounds" in data (such as using a specific date like May 20, 1875 to mean "birthdate unknown" when the design of the database will not allow a black or non-number entry in the birthdate field) often rely on institutional knowledge to function appropriately. Of course, this can cause problems outside the context of data consolidation, but it is particularly relevant when data is shared across agencies, as it moves outside its original context, away from that institutional knowledge or controls. This lack of institutional knowledge was the source of erroneous claims by the Department of Government Efficiency (DOGE) that there were numerous people receiving Social Security payments who were over 150 years old, and thus obviously fraudulent recipients.

Mismatches can be similarly damaging. Mismatches are the conflating of two records that are actually about different individuals, or the failure to connect two records that are actually about the same person. The latter generally does not produce the benefits that data sharing was meant to provide, while the former can result in dire consequences, such as people not receiving benefits to which they are entitled, or false arrests and imprisonment.

Additionally, the definitions that underlie elements of databases are not always as straightforward as they may seem. Consequently, data typically has to be standardized, and definitions have to be clarified. This can be as simple as one database that has a "name" field interacting with another that has a "first name" and a "last name" field. Trying to use data with these disparate schemas without clarifying can introduce numerous errors.

Considerations for Data Quality, Definitions, and Matching

There are a few considerations to keep in mind when assessing how the quality and structure of the data being shared will impact the outcomes of the sharing.

- **Matching framework**
  Matching frameworks are designed to ensure that the agencies sharing information have a shared understanding of definitions for data elements and a clearly defined mechanism for determining when records are considered a match and when they are not. Considering these elements ahead of time can help to avoid errors.

- **Protocol for investigating and correcting matching and data quality errors**
  Matching protocols and standard definitions can help to minimize errors, but they will not eliminate them entirely. Consequently, there should be a process in place for correcting errors in the data or matching framework. This must include a way to receive information about errors as well as steps to correct the error in a given case and, if applicable, avoid the error in the future.

The Impact of AI

AI has complicated data consolidation because it can provide solutions to some of the concerns identified above (for instance, AI may be able to match records across databases even if they do not have standard definitions or data schemas), but it also introduces new risks, such as hallucinations, where AI produces false content or inferences, or unpredictable mismatches that are difficult to guard against. Additionally, it complicates the questions of "what is a database" – if AI can access data across databases and present them seamlessly to a user, much of the delineations that separate individual databases are functionally dissolved. Consequently, AI's role in data consolidation will have to be carefully tailored and monitored.

Considerations for AI in Data Sharing

How AI is used and governed will determine how it impacts a data sharing program. Some uses will introduce fewer risks, and different governance frameworks will determine how those risks are managed.

- **Role of AI**
  If AI is used with consolidated data to, for example, analyze overall trends in applicants to public benefits across a variety of programs, the impacts of errors

would be diffuse, affecting overall policy decisions. If, however, it is used to identify individuals for immigration enforcement, any errors or shortcomings can result in enormously harmful actions impacting individuals.

- **AI governance structures**
  While AI introduces risk, governance can help to ameliore those risks. In the context of data consolidation, this governance will often have to be a cross-agency endeavor. For instance, one of the important aspects of AI governance is ensuring that the data is appropriate to the tasks. Agencies who are the experts on their own data will need to support any AI uses by receiving agencies who are less familiar with the data.

Conclusion

As the government continues to push for data consolidation, the question of what constitutes a database is becoming increasingly important. Efforts to reassure the American public that a "master database" does not exist (here and here) are based on technicalities, but as this explainer illustrates, data consolidation, regardless of whether it is accomplished through a single database or federate model, raise significant privacy and civil liberties concerns.

As more agencies contribute to consolidated datasets, the potential for errors expands. This will likely lead to government action based on these erroneous data. This has the potential to cause enormous harm to people caught up in those government actions, which could include inappropriate benefits stoppages, false arrests, and the overall degradation of government services. More data sharing also creates more opportunities for data to be used for purposes divorced from the original purpose for the collection of the data. This may not comport with the expectations and consent of those who provided the data, violating their privacy and degrading trust in government agencies. The potential for these wide-ranging and damaging impacts require careful attention to and governance of government data sharing programs.

Ultimately, there are legitimate reasons that agencies would wish to share data in service of furthering their missions. Because data consolidation can have significant impacts to the privacy and well-being of people whose data is being shared, however, data consolidation must be rigorously assessed to ensure that appropriate protentions are in place, the data sharing is in service of the agencies' missions, and there are frameworks and processes in place to ensure the data is not repurposed in ways that harm the people the agencies are meant to serve.

Find more work from CDT's Equity in Civic Technology team at *cdt.org*.

CENTER FOR DEMOCRACY & TECHNOLOGY

The Center for Democracy & Technology (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet.