SLDS Issue Brief

Centralized vs. Federated: State Approaches to P-20W+ Data Systems

States' needs for information about the progress of students, schools, districts, and programs continue to expand from the boundaries of K12 education to encompass the broader spectrum of P-20W+[1] information. To meet this need, data must be brought together from multiple sources across agencies and organizations. States have approached the challenge to maintain and provide secure access to data linked across organizations through two predominate strategies, each with its own challenges and advantages. The centralized P-20W+ data repository is a single, integrated data repository that contains, maintains, and provides secure access to data from all participating agencies and organizations. The federated P-20W+ model is an alternative in which data from participating organizations are temporarily linked to create a report or to generate a dataset.

This document is intended to help state agencies through the process of determining whether a centralized or federated model (or a hybrid[2] approach) will best suit their environment and stakeholder needs. This issue brief will address key questions that should be considered early in the development of a P-20W+ system and describes how federated and centralized models bring together data from agencies across a state's P-20W+ environment and make those data useful for and accessible to education stakeholders.

Key Questions to Consider

A clear understanding of your state's unique environment will inform decisions about your system's development and will improve the likelihood that it will meet your end users' information needs. No matter if you choose to develop a centralized or federated system, or to include some aspects of both approaches, all agencies must address certain fundamental questions and issues. For example, each approach includes the need for P-20W+ data governance as a solid foundation because clear roles, responsibilities, and ownership are critical to the success of a P-20W+ system.

The following issues, many of which apply well beyond the system design conversation, should be considered early in P-20W+ system planning:

1. **State policy/legislation.** What are your state policies regarding data consolidation and exchange? For example, does any legislation limit your state's ability to maintain linked data across agencies? Does any legislation mandate the development of a certain type of system?
2. **Stakeholder information needs.** What P-20W+ longitudinal data do your stakeholders need to inform education policy and evaluate programs? Do you need a system solely to respond to data requests from researchers and produce standard reports on a scheduled basis, or one that can support a broader and less structured array of users and uses?

[1] P-20W+ refers to data from preschool (early childhood), K12, and postsecondary through post-graduate education, along with workforce and other outcomes data (e.g., public assistance and corrections data). The specific agencies and other organizations that participate in the P-20W+ initiative vary across states.

[2] In one hybrid approach, a linkage is established via identifiers (e.g., Social Security number, name, date of birth, and student identifiers), while the data to be shared with researchers or other data recipients (e.g., enrollment, attainment, and assessment data) are kept separate and pulled into the dataset only when needed for a specific purpose.

This product of the Institute of Education Sciences (IES) SLDS Grant Program was developed with the help of knowledgeable staff from state education agencies and partner organizations. The information presented does not necessarily represent the opinions of the IES SLDS Grant Program.

For more information about the IES SLDS Grant Program, additional SLDS publications, or support with system development or use, please visit http://nces.ed.gov/programs/SLDS.

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

3. **Governance.** Will a single agency own the system or will ownership be shared among contributing agencies? Do agencies and organizations in your state adhere to a common data standard? Would all participating agencies abide by a common set of rules, or would the agencies require their own rules that would need to be mapped? Can statewide data cleansing processes be implemented to ensure high quality and consistency? Do you have a process for reliably matching records across systems and for reconciling discrepancies that are identified?
4. **Startup funding.** What funding is available for the development and implementation of a P-20W+ system?
5. **Sustainability and responsibility.** How will resources be acquired and allocated for ongoing support and maintenance? Will your existing resources be sufficient to support the system over time or will additional staff and funding be needed? If you are currently using grant funding to develop your statewide longitudinal data system (SLDS), how will your state sustain the SLDS after that funding is exhausted? What agency or agencies will be assigned or assume responsibility for maintaining the system over the long term?
6. **Staffing capacity.** What are the staffing resources available from participating agencies and how would the work be allocated across organizations/agencies for a federated system? How would staffing needs differ for a centralized data repository?
7. **Timeline.** What is your timeline for implementation? Do you intend to have all integration work completed at the same time, or will you look for quick wins and schedule incorporation of data and data links to the separate datasets over time?
8. **Scalability.** How scalable must your system be? Should you develop a system that would accommodate other data sources after the system has been developed?
9. **Data sharing culture.** What are your partner agencies' stances toward data sharing and ownership? Are they open to sharing datasets with researchers and other outside entities?
10. **Privacy protection.** How will federal, state, and local laws affect interagency data sharing in your state? What are the participating agencies' responsibilities around governance and the protection of combined data sets in either a federated or centralized scenario? Are your data truly de-identified,[3] or will the data be subject to requirements of the Family Educational Rights and Privacy Act (FERPA) or other laws (e.g., the need for memoranda of understanding or contracts for multi-agency data sharing)? How will this play out over time? Is there a limit to the number of years that data can be shared or used?

Centralized and Federated P-20W+ Models: What Are They and How Do They Compare?

Centralized and federated P-20W+ SLDSs have several key structural differences regarding whether and how data are integrated and stored. But these system types also share basic characteristics in terms of data sources and the ultimate presentation of data to users.

In a **centralized data system**, all participating source systems copy their data to a single, centrally located data repository where they are organized, integrated, and stored using a common data standard. As depicted in figure 1 (next page), data in a P-20W+ centralized SLDS are periodically matched, integrated, and loaded into a central repository. Users query the system and can access the data that they have been authorized to view and use.

In a P-20W+ **federated data system**, individual source systems maintain control over their own data, but agree to share some or all of this information to other participating systems upon request.

System users submit queries via a shared intermediary interface that then searches the independent source systems. As depicted in figure 2 (next page), data are queried from source systems and records are matched to fulfill a data requestor's information needs. The linked data are not stored by the system, but rather are cached and delivered, then removed.

[3] De-identification of data refers to the process of removing or obscuring any personally identifiable information from student records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them. While it may not be possible to remove the disclosure risk completely, de-identification is considered successful when there is no reasonable basis to believe that the remaining information in the records can be used to identify an individual. De-identified data may be shared without the consent required by FERPA (34 CFR §99.30) with any party for any purpose, including parents, general public, and researchers (34 CFR §99.31(b)(1)).

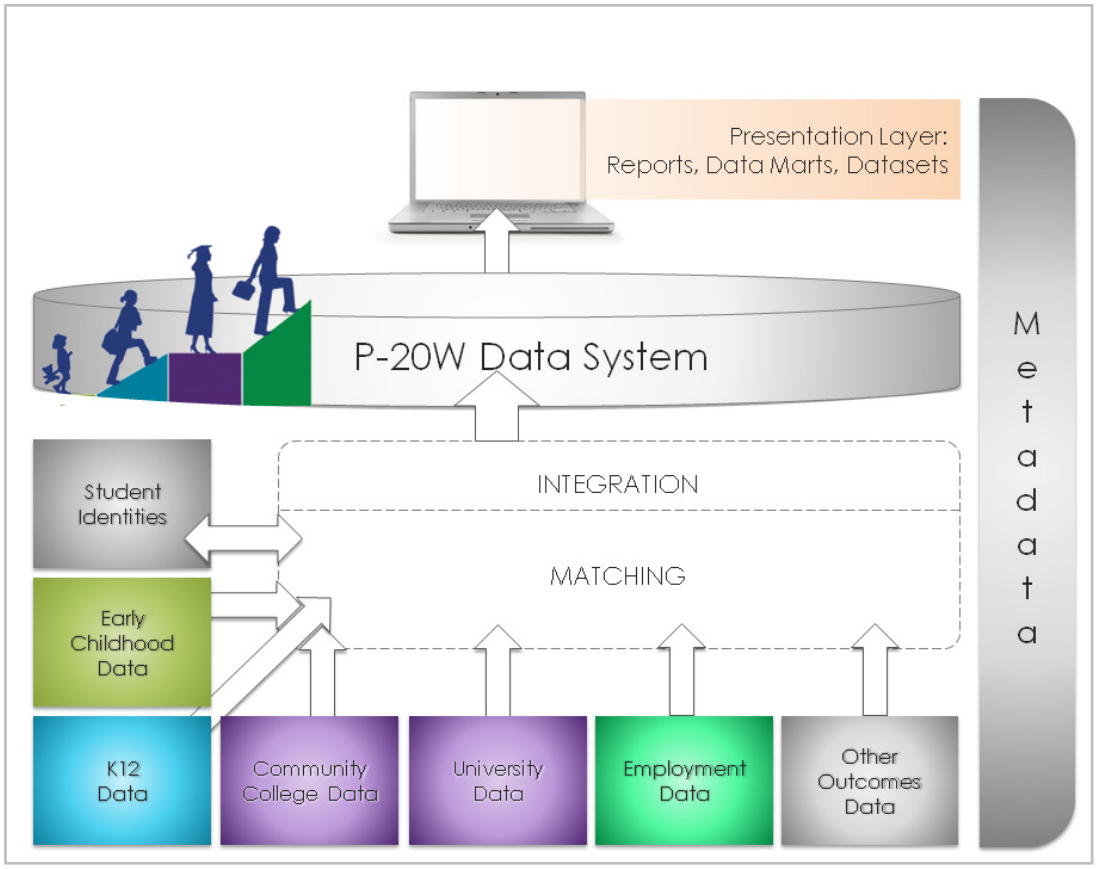

Figure 1. Basic structure of centralized data system

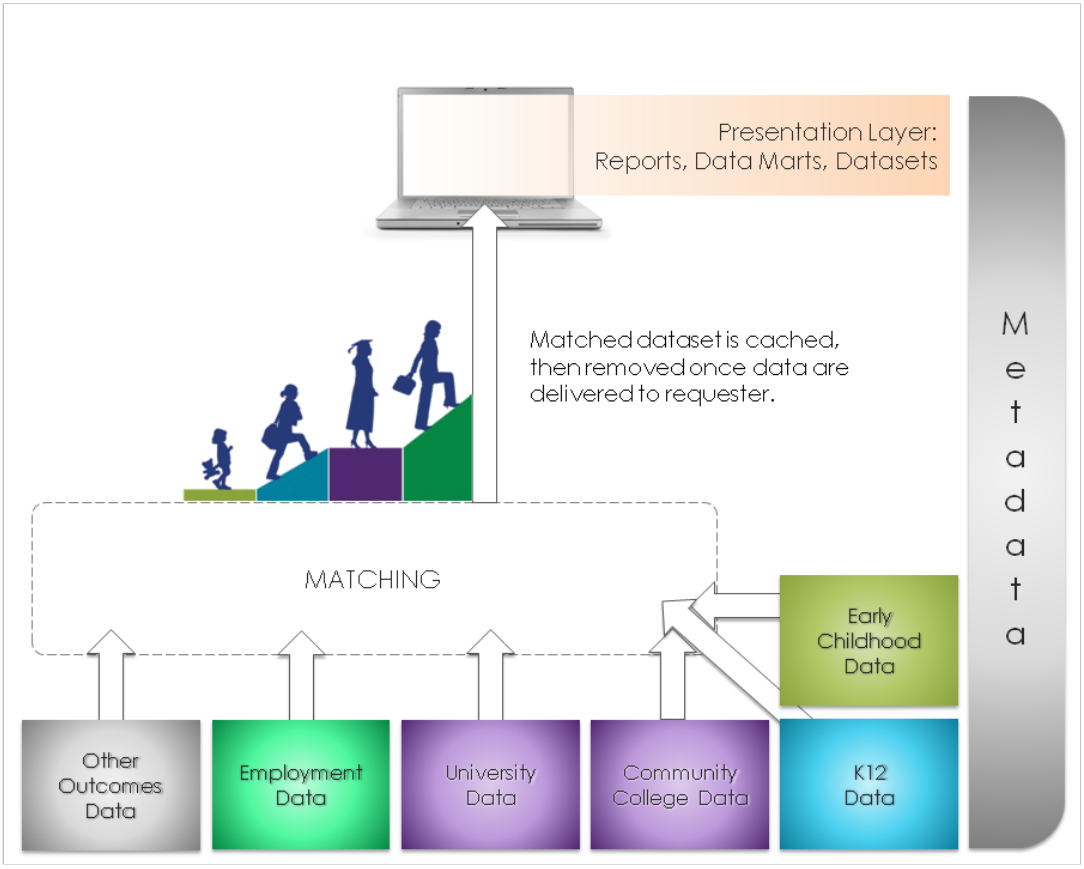

Figure 2. Basic structure of federated data system

Comparison of Centralized and Federated System Characteristics

Component	Centralized	Federated
Data ownership	Source agencies own the data and share data stewardship[4] with the centralized data warehouse entity. Data stewardship responsibility should be spelled out in memoranda of understanding (MOU).	Source agencies own and have stewardship over their data. There is no shared data stewardship.
Staff resources	Staff resources are required of each source system to oversee and maintain required data access. In addition, support will be needed for the extract, transform, and load (ETL) processes to reflect changes in source data systems and data element modifications. Staff will also be needed to support the centralized database system.	Staff resources are required of each source system to oversee and maintain required data access. In addition, support will be needed for the extract, transform, and load (ETL) processes to reflect changes in source data systems and data element modifications. Staff resources are required from each participating agency to review and approve data requests.
Technical requirements	Each source system will need to allow access for pulling the data or have the technical capability to provide data to the centralized data system. The system infrastructure will require ETL tools as well as capabilities to match the data, store the results, and deliver matched datasets to stakeholders via tools such as portals or business intelligence solutions.	Each source system will need to allow access for pulling the data. It will also need hardware and bandwidth to process external queries with ETL tools, match the data, and return the resulting dataset to stakeholders via tools such as portals or business intelligence solutions.
System performance	Data extraction is generally fast since all data matches have occurred in the transformation and load steps. Data are matched once and used many times. Scheduled extracts can occur on source systems during off- peak hours to minimize impact on sources. Centralized data system architecture can be designed specifically for this purpose, thus increasing response times.	Data delivery is subject to longer delays due to load and scheduling on source systems. Agency-specific performance issues and other priorities can affect delivery times for requested datasets.
Privacy/security	Primary responsibility is with the centralized data system entity as the data steward, but policies are dictated by source system agencies via MOUs. Security is handled through access rules for users. The centralized repository may make it easier to preserve data integrity. Although records are typically de-identified during loading, the stakes may be higher in the event of a breach because all data are stored in a single location.	Primary responsibility is with the source system agencies. Secure processes are needed for handling data queries. Data are diffused, allowing for tailored protection based on sensitivity of each source system's data and reducing the amount of data that could be accessed through a breach.
Data updates/corrections	ETL processes take place either when data are changed (if the data system is required to have near real-time data) or at specific intervals to capture changes, corrections, or updates.	Data reside within each agency. Each agency is responsible for communicating and possibly updating the data extractraction processes to reflect changes, corrections, or updates.
Data availability	Data must be finalized in the source system and then integrated into the centralized repository before they are available. Access to data is determined by source agency via MOU.	Based on when data are available in the source and made available for extract. Access to data is determined by source agency.
Data quality	Consistent data cleansing processes and data quality checks apply to all data as agreed by the source systems. Data may be more reliable because they are validated as part of load process from each source system.	Quality depends on processes implemented and supported at each agency.

Table 1. Comparison of centralized and federated data systems by key characteristics

[4] Data stewardship is a comprehensive approach to data management to ensure quality, integrity, accessibility, and security of the data.

Component	Centralized	Federated
Implementation	The implementation period is longer due to the need to build the centralized database or warehouse. Significant time is also needed to determine requirements and establish processes for ETL and data provision.	Dependent on processes implemented and supported at each agecy.
Scalability	Adding data sources potentially requires supplementing or expanding centralized data system architecture. Adding data sources requires writing ETL processes and implementing matching/integration rules.	Adding data sources does not require the addition of any hardware or other resources. Adding data sources requires writing ETL processes and implementing matching/integration rules.
Production of standard reports	Standard reports can be automated to save time and cost.	Report production requires one or more agencies to accept this as a responsibility.
Sustainability	Possible approaches include a state appropriation to the centralized data system entity for the development and ongoing support and maintenance of the centralized system. This approach would have no fiscal impact on the participating agencies. Another approach would be for each participating agency to pay a proportional part of the funds needed to support the centralized system in a cost-recovery model. This approach could discourage some agencies from participating.	Possible approaches include asking each agency to contribute toward supporting data system processes. This approach may discourage some agencies from participating. Alternatively, state appropriations could be made to each participating agency for data system support based on a funding formula.
Usability	Longitudinal data are all in one place, facilitating and streamlining data mining.	Multiple years of data must be queried from partner agencies, which requires assurance of comparability. If additional years of data are needed for a given cohort, the entire dataset may need to be rebuilt.

Table 1. Comparison of centralized and federated data systems, by key characteristics, continued

At a Glance: Key Pros and Cons to Consider

	Centralized	Federated
Pros	• Better performance for pulling data • More streamlined for data mining • Easier to account for data integrity/security • (Single) central data policy • Easier to ensure data quality • Quicker data results • Avoids issues of disparate and non-compatible technologies	• Shorter time for repository • Mitigates turf battles and trust issues • Diffuses data and allows for tailored protection of data based on sensitivity • More scalable
Cons	• Higher costs for infrastructure development and training • Data only as current as most recent load • Higher risk in event of breach due to amount of data contained in single repository • More difficult to distribute costs across participating agencies, if needed	• Requires development and maintenance of multiple data sharing policies • Requires data to be pulled and linked every time a dataset is generated • Investment and support of intermediary interface by each of the participating agencies • Limited P-20W+ data integration

Table 2. Major pros and cons of centralized and federated data systems