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

Historically, efforts to create a P-20W[1] data repository resulted in the development and use of a single, centralized data system that contains, maintains, and provides secure access to data from all participating agencies. In recent years, however, an alternative model has emerged in some states for reporting P-20W data—a federated model in which data from participating agencies are temporarily linked to create a report or generate a dataset. This approach, while relatively new and untested in the education field, has typically been adopted to align with states' data sharing cultures or to deal with issues such as state legislative prohibition of permanently establishing a linkage between certain data.

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. We begin with some key questions that should be considered early on. Next, a matrix presents a side-by-side comparison of these two approaches to bringing together data from agencies across a state's P-20W environment and making those data useful for and accessible to education stakeholders.

Key Questions to Consider Up Front

A clear understanding of your state's unique environment will inform decisions about your system's development and, ultimately, improve the likelihood that it will meet your end users' information needs. Regardless of whether you choose to develop a centralized or federated system, there are certain fundamental questions and issues that all agencies will need to address. For example, neither approach will allow you to avoid the need for P-20W data governance as a solid foundation of clear roles, responsibilities, and ownership are critical to any P-20W system's success.

The following issues, many of which apply well beyond the centralized/federated conversation, should be considered early on in any P-20W effort:

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 do your stakeholders need in terms of education policy and program evaluation concerning P-20W longitudinal data? Do you need a system solely to respond to data requests from researchers or one that can support a broader array of users and uses? For instance, will the system need to support the generation of standard reports on a regular basis?
3. **Governance**: Will a single agency own the system or will ownership be shared among contributing agencies? Does your state adhere to a common data standard? Can/would all participating agencies abide by the same 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?

---

[1] P-20W refers to data from prekindergarten (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 from state to state.

[2] In one promising hybrid approach, a linkage is established via identifiers (for example, Social Security number, name, date of birth, and student identifiers), while the data to be shared with researchers or other data recipients (for example, enrollment, attainment, and assessment data) are kept separate.

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

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 SLDS, how will your state sustain the SLDS after that funding is exhausted? What agency(ies) will be assigned or assume responsibility for maintaining the system over the long term?
6. **Staffing capacity**: Do participating agencies have the staffing resources to meet the ongoing needs of a federated system (e.g., quick turnarounds to fulfill ad hoc data requests)? Or, would dedicated, separate resources in support of a centralized system be more in line with agencies' ability to participate?
7. **Timeline**: What is your timeline for implementation?
8. **Scalability**: How scalable does your system need to be? Should you develop a system that will be able to 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?
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)?

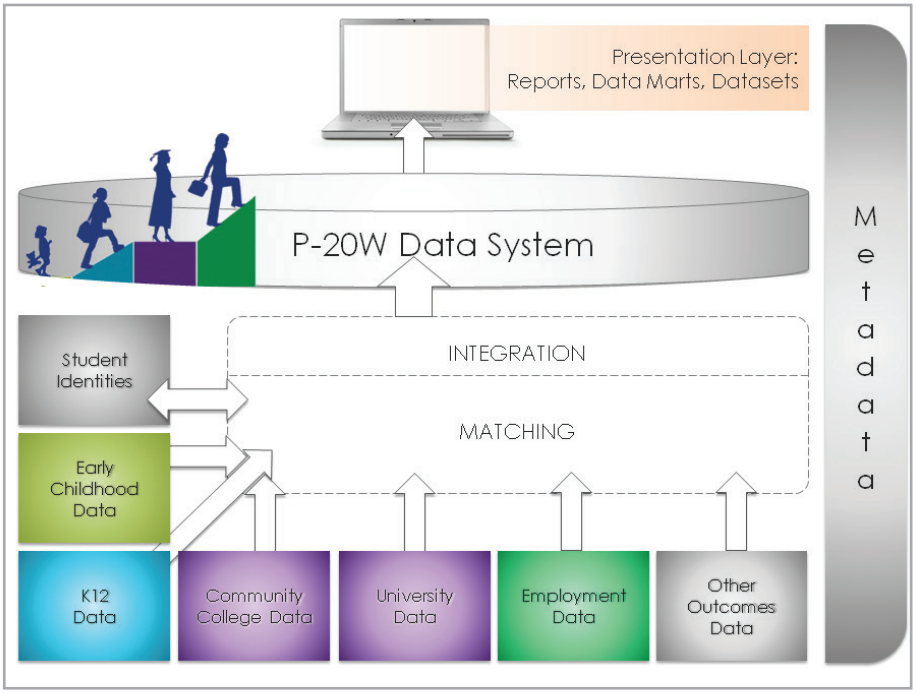
[3] De-identification of data refers to the process of removing or obscuring any personally identifiable information from student records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them. While it may not be possible to remove the disclosure 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)).

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 (for example, in how (or if) 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.

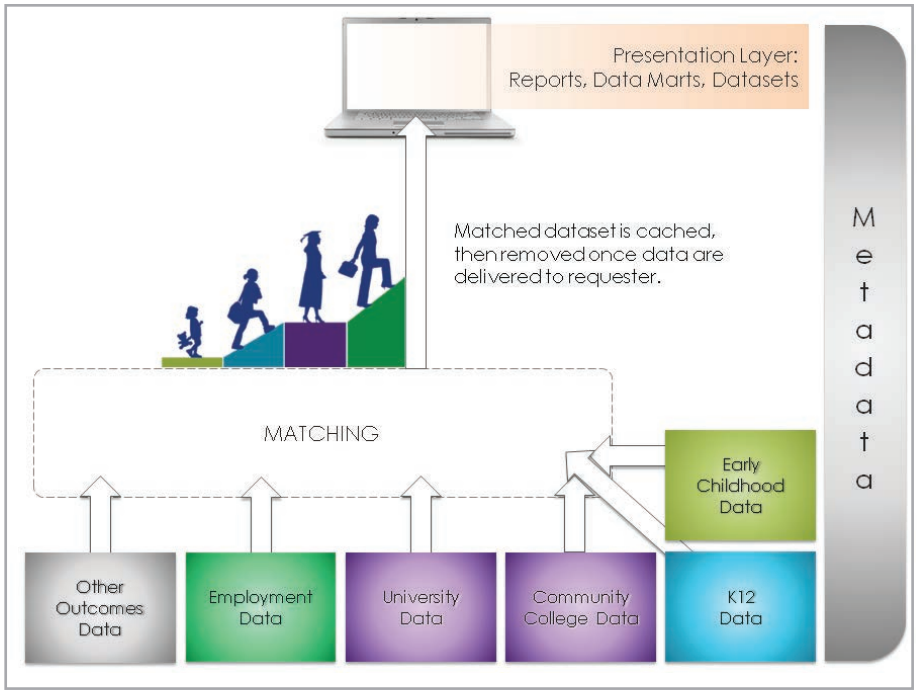
In a **centralized data system**, all participating source systems copy their data to a single, centrally-located data repository where they are organized, integrated, and stored using a common data standard. As depicted in Figure 1, data in a P-20W centralized SLDS are periodically matched, integrated,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

Figure 1. Basic structure of centralized data system



In a **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. In a P-20W federated system, as depicted in Figure 2, 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 removed once cached and delivered. The individual sources of data maintain control of their data, storing and securing them, and providing them to the system only upon request.

Figure 2. Basic structure of federated data system



Comparison of Centralized and Federated System Characteristics

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

	Centralized	Federated
Data ownership	Data ownership is with the source agency with shared data stewardship with the centralized data warehouse agency/entity. Responsibility for this data stewardship should be spelled out in memoranda of understanding (MOU).	Data ownership is with the source agency with no need for shared data stewardship.
Staff resources	Staff resources are required of each source system to oversee and maintain required data access. In addition, support will need to be given to 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 data base system.	Staff resources are required of each source system to oversee and maintain required data access. In addition, support will need to be given to 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 be willing to allow access or provide the data to be included in the centralized data system. An infrastructure to support the centralized system along with ETL tools, conduct matching processes and storing the results. There will also be a need to deliver the matched resulting dataset (e.g., via portal or business intelligence (BI) solution).	Each source system will need the required hardware and network bandwidth to facilitate and process external queries (ETL tools), conduct matching processes and returning the resulting dataset. There will also be a need to deliver the matched resulting dataset, i.e. portal or business intelligence (BI) solution.
System Performance	Data extraction is generally fast since all data matches have occurred in the transformation and load steps. Match once, use 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. Established technology and procedures; proven technology.	Subject to longer delays in data delivery due to load on source systems, etc. Agency specific performance issues can affect the performance of the entire system. Also the possibility of limited or narrow windows of processing time due to other/competing priorities. Relatively new technology; accounts for less than 10 percent of all data warehouse projects; not a proven technology.
Privacy/ Security	Primary responsibility is with the centralized data system agency/entity as the data steward, but is dictated by source system agencies via memoranda of understanding. Security is handled through access rules for users. May make it easier to account for data integrity. Stakes may be higher in event of a breach since all data are stored in one location (though typically records are deidentified as part of load process).	Primary responsibility is with the source system agencies. Secure process needed for handling of 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	Establish process for ETL either when data are changed (if required to have near real-time data in centralized data system) or at a specific periodicity to capture changes, corrections, or updates.	Data reside within each agency. Each agency is responsible for communicating and possibly updating the data extract processes to reflect changes, corrections or updates.
Data availability	Based on when data are available in the source and made available for extract. 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.

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

	Centralized	Federated
Data quality	Process for data cleansing apply to all data as agreed upon by the source system agencies; consistency of data cleansing processes and data quality checks. May provide more reliable data since the compiled data from various systems are validated as part of load process.	Dependent on processes implemented at each agency.
Implementation	Longer implementation period due to the need to build the centralized data system database/warehouse. But equal time is also needed to determine requirements and processes for ETL and data provision.	Generally requires less time; although equal time is needed to determine requirements and processes for ETL and data provision.
Scalability	Potentially supplementing or expanding centralized data system architecture to accommodate additional agency source system data. Writing ETL processes and matching/integration rules.	The addition of any required hardware and other resources (as mentioned above) required for data queries/matches across the system. Writing ETL processes and matching/integration rules.
Production of standard reports	Can be an automated process; less expensive and timelier to accomplish.	Dependent on an agency accepting this as a responsibility.
Sustainability	Possible approaches include a state appropriation to the centralized data system agency/entity for the development and ongoing support and maintenance of the centralized system. This would have no fiscal impact on the participating agencies. Another approach would be for each participating agency to pay for a proportional part of the needed funds for the support of the centralized system, in a cost recovery model. This could be a deterrent for agencies to participate.	Possible approaches are for each participating agency to make their contribution for the corporate support of the processes needed for the federated system. This may be a deterrent for agencies to participate. Another approach would be specific appropriation that is allocated to each participating agency, based on a funding formula.
Usability	Longitudinal data all in one place. Facilitative of 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, entire data set will need to be rebuilt.

At a Glance: Key Pros and Cons to Consider

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

	Centralized	Federated
Pros	X Proven technology X Better performance X Better for data mining X Easier to account for data integrity/ security X Central data policy X Easier to ensure data quality X Quicker data results	X Shorter development time X Mitigates turf battles/get around trust issues X Diffuses data and allows for tailored protection of data based on sensitivity X More easily 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	– Requires development and maintenance of multiple data sharing policies – Data linked every time a dataset is generated. – Unproven technology (for example, response time not yet tested) – Investment and support of intermediary interface by each of the participating agencies – Limited P-20W data integration

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