Modern Data System Design and Key Decisions

Contents

* *note*: P20W refers to state longitudinal data systems that include information from preschool, K-12, postsecondary, and workforce data sources

P20W Trade-Offs and Considerations | Overview & Approach

This document describes key decisions and options that need to be considered in designing a P20W data pipeline technology infrastructure.

Analysis Approach

- The organizing structure for these considerations is the P20W Capability Model, which outlines the organizational and technical capabilities needed in a modern P20W data pipeline. See the next section for orientation to the Capability Model.
- Each Capability in the model is associated with at least one key decision, and a few Capabilities have more than one
- In many cases, these decisions have governance and business process implications in addition to technology choice implications

P20W Capability Model

I. Overview
II. Capability Descriptions

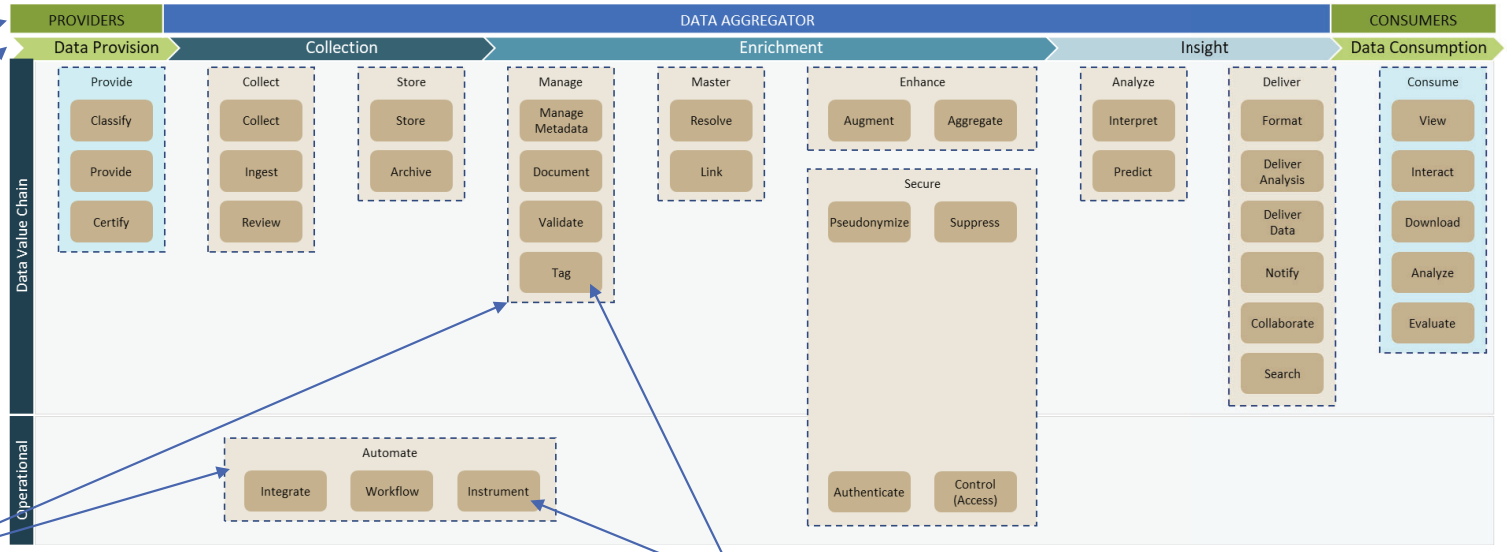
P20W Capability Model | Overview

The P20W Capability Model outlines organizational and technical capabilities needed in a modern P20W data pipeline. General left-to-right progression is assumed, although data may move through the pipeline capabilities iteratively at various stages, depending on the unique business processes established by the Data Aggregator.

Actors: There are three entities/actors assumed in the model: Data Providers, the Data Aggregator (e.g., P20W Data Agency), and Data Consumers (e.g., researchers, legislators, public constituents)

Data Value Chain: There are three stages of the Data Aggregator's value chain: Data Collection, Data Enrichment, and Data Insight.

Business Capabilities: The dotted boxes represent logical groupings of core business capabilities, which include people, process and technology that enable the work of the Data Aggregator and the data pipeline



Technical Capabilities: The boxes within the dotted boxes represent technical capabilities, technology functions that are core to supporting the activities of the data value chain.

P20W Capability Descriptions

Business Capability	Technical Capability	Typical Design Elements Required
Collect	**Collect** Receive data in secure manner from authorized sources	• Secure data transfer methods - API, SFTP, User Interface for manual HTTPS upload • Authentication method for identity verification • Access control by organization and role • Data encryption methods applied to transfer and storage
	Ingest Parse and store data in well-defined structure	• Methods to transform data and metadata received in variety of formats to well-defined structure • Integrate with Store, Archive and Manage Metadata solutions • Fully automated
	Review Review summary of submitted data	• Methods to review summaries of the data • Integration with Authenticate, Control and Automate capabilities to support authorized access
Store	**Store** Securely store data in a flexible structure with appropriate performance level	• Methods to enable access to element-level structured data • Structured but flexible data store to accommodate growth of data scope • Mechanism to enable access to the data only by authorized individuals and services • Mechanism to identify and capture links between data elements • Data encryption methods for data at rest and while in transit into and out of storage
	Archive Retain data as received for long term	• Methods to support retention of data with appropriate performance and cost over long term • Linkage with metadata including provenance

P20W Capability Descriptions, cont'd

Business Capability	Technical Capability	Typical Design Elements Required
Manage	**Manage Metadata** Collect, store and allow retrieval of metadata	• Methods to collect, generate, store and retrieve metadata relating to submitted data sets, which populates the data catalog in support of Search and Request capabilities; metadata attribute examples include provenance, usage and sensitivity classification, descriptive text, links to applicable data dictionary elements and code lists, and summary statistics • Mechanism to consistently apply metadata definitions (e.g., enterprise-wide data dictionary and code lists)
	Document Publish agreed data formats and structures for the collection of data	• Data encoding schemas, data dictionary and code lists • Methods to document and publish data formats, structures and supporting files for data providers to contribute data • Support consumption in both human-readable and machine-processable form • Online publication for easy discoverability and interactive (hyperlinked) usage
	Validate Automated validation of collected data against pre-defined constraints	• Mechanism to define and maintain pre-defined business rules for data quality • Automated validation of data quality against pre-defined constraints (e.g., valid code lists, data types, ranges, business rules) • Sharing of validation results with data contributor at a granular level, sufficient to support correction and resubmission • Methods to track data quality over time and apply additional support as needed • Alignment with data dictionary to support documentation and application of element-level constraints
	Tag Manage and apply data sensitivity and allowable use classification metadata	• Methods to collect and store data sensitivity and allowable use classification tagging of the source data, along with other critical metadata elements • Mechanisms to inform enforcement of classification and usage tags in downstream data consumption components (e.g., Data Request, Access Control and Data Suppression)
Master	**Resolve** Resolve and maintain unique master IDs for each mastered entity	• Methods to resolve a set of identifying information to one or more "golden" IDs with a computed confidence level • Mechanism to create new golden IDs • Mechanism to merge and split golden IDs when presented with new identifying information
	Link Link different data sets through predefined or discovered identitifiers	• Methods to link different data sets through predefined identifiers • Discovery of new links between data sets

P20W Capability Descriptions, cont'd

Business Capability	Technical Capability	Typical Design Elements Required
Enhance	**Aggregate** Data is summarized at various levels of aggregation, and sliced across various dimensions	• Methods to automatically summarize and store data at predefined aggregation level • Methods to summarize and store data for ad hoc analysis • Integration with Data Suppression mechanism to prevent inadvertent disclosure
	Augment Enhance data with standardized calculated values	• Mechanisms to enhance data with calculated values using various formulas and rules that standardize or derive important variables from source data
Secure	**Pseudonymize** Form of deidentification in which Personally Identifiable Information (PII) is replaced with deterministically generated or assigned tokens	• Method to replace PII and other sensitive information with tokens, including generated IDs, that are consistent across data sets for the same represented individual • Integration with classification tagging of elements to identify those that require pseudonymization • Mechanism to prevent "reverse engineering" tokens to reidentify individuals • Mechanism to re-identify records for authorized users (if required)
	Suppress Aggregated data is evaluated for small cells; primary and complementary data suppression is applied to limit disclosure risk	• Method to identify and suppress small cells (below a configurable threshold) within data aggregations (primary suppression) • Mechanism to apply complementary suppression across elements of a data aggregation • Application of differential privacy to produce randomized, yet statistically equivalent, data sets for modeling purposes • Mechanism to estimate disclosure risk and flag high risk (above defined thresholds) for review and approval or mitigation
	Authenticate Internal and external user identity is authenticated	• Mechanism for the Organization to manage identities and authentication protocol for their internal and external users • Methods to support federated identity management with partner organizations • Mechanism to integrate identity with role-based access control
	Control (Access) Role based user access control to data and systems	• Mechanism to incorporate granular role-based access control for internal and external personas across all system components • Mechanism to monitor access patterns, including logging and audit capabilities

P20W Capability Descriptions, cont'd

Business Capability	Technical Capability	Typical Design Elements Required
Analyze	**Interpret** Statistical models are developed and applied to derive meaning from historical unitary data (Descriptive Analytics)	• Methods to aggregate, slice and compare data across dimensions/variables of interest • Methods to identify and isolate/adjust data anomalies, errors and outliers • Methods to produce and test statistical models to uncover historical trends • Ability to automate the production of predefined analyses as new data becomes available, as well as to perform ad hoc analysis
	Predict Historical data is used for predictions using statistical and machine leaning techniques, e.g., to predict future academic outcomes (Predictive Analytics)	• Prediction mechanism using simple statistical models; e.g., regression models • Mechanism to enable predictions using machine learning models
Deliver	**Format** Statistical summaries are formatted as tables and graphs with added commentary	• Methods to develop public dashboards and tables with quantitative data • Mechanism to include and track qualitative commentaries with the dashboards
	Deliver Analysis Results of analysis are delivered via appropriate channels	• Methods to deliver analysis to the public anonymously (not requiring login), such as through interactive visualizations and summary data download • Method to deliver sensitive analysis results securely to an organization • Methods to deliver sensitive analysis results securely to individuals
	Deliver Data Aggregated or individual-level data sets (de-identified or identifiable, per data request) are generated and transferred to the requester securely	• Methods to support download of set of aggregated public data • For approved data request, methods to generate and transfer set of data from provider cloud repositories per data use and sensitivity tagging • Mechanism for 3rd party data requester to access the secure data location, apply research tool and perform analysis • Mechanism to monitor and control secure access of 3rd party requesters

P20W Capability Descriptions, cont'd

Business Capability	Technical Capability	Typical Design Elements Required
Deliver	**Notify** Individuals register subscriptions and are notified when specific public dashboards or data sets are refreshed, or when stored searches result in a new match in the data catalog	• Methods for public and private individuals to create an account and subscribe to one or more dashboards, data sets, or catalog query results • Methods to deliver notifications through users' choice of channels, based upon registered subscriptions
	Collaborate Users may connect with each other based on expression and matching of mutual interests and have the ability to share analysis content and commentary in an online community forum	• Methods for the users to express their research interests • Mechanism for connecting users who have common interests • Mechanisms for users to share comments and content, including analysis code and links to relevant data • Integration with the Notify capability, to inform users of new content or users that may be of interest
	Search Users can search for content on the platform	• Mechanism to index all content and expose appropriate descriptive metadata upon request • Data catalog functionality, which captures and allows query of key metadata identified and managed according to the Manage Metadata capability
Automate	**Integrate** A data integration platform to orchestrate automated transformation and flow of data between components	• Methods to move the data through any necessary extract, transform and load activities • Mechanism to interface between components via standard methods, which may include API calls, file transfers, and database queries, among others • Mechanism to support error handling
	Workflow Business process workflows (i.e., orchestrated and repeatable patterns of activity) that support the data pipeline are designed, managed and automated	• Method to define and configure workflow solution to manage portions of the data pipeline and ancillary processes • Mechanism to manage human activity tasks, such as to submit and respond to requests for intervention or approval; should include submission of constrained data forms, documents, performance of electronic signatures, notification of task assignment and status changes, and user interfaces for performing all such tasks • Mechanism to support error handling and recovery • Monitoring of process status, with support for escalation and exception recovery operations
	Instrument Measure system performance and user engagement with data and reports; establish feedback mechanism for improvement	• Mechanism to collect and analyze system events to support operations such as capacity and scale adjustment, identification and correction of failures, etc. • Mechanism to collect and process user engagement measures for reports and data • Methods to allow reporting, tracking and resolution of issues experienced by users • Method to proactively solicit feedback from users (e.g., context-sensitive survey prompts)

P20W System Design | Key Decisions

P20W Design | Foundations

Assumptions: Two key assumptions underlie the P20W Capability Model and the technology decisions as a whole:

1. The model assumes that the P20W data agency will be implementing the architecture on a public cloud (e.g., Azure, AWS, Google Cloud)
2. The P20W data agency will use high-level analytics and services provided by the cloud vendor using cloud-native components rather than building custom versions of these tools or integrating with additional third-party solutions

Key Strategic Question: There is a foundational question to be addressed that will guide the rest of the technology choices made:

Should the P20W data agency use a commercial data transformation & master data management solution ('MDM solution") in its core architecture OR use cloud-based tools to manage those functions?

Options	Commercial, Third-Party Solution	Public Cloud Vendor Services
Description	Specialized, off-the-shelf data transformation and master data management solution product	Data transformation, integration and master data management services offered by the public cloud providers.
Considerations & Implications	• Requires the P20W data agency to design the solution architecture around the functionality of selected solution, and integrate third-party vendors' products into the data processing "pipeline" • Potentially faster development with a comprehensive solution suite that meets requirements of multiple P20W Capabilities (*Collect, Manage, Master* Capabilities) • A specialized solution vendor has in-house skills to maintain and update the product over time • Potentially higher total cost of ownership vs. usage-based model of a cloud service, unless solution cost is driven down through high-volume enterprise license agreement	• Using services of the selected public cloud vendor would provide a natural synergy with the end-to-end data pipeline toolchain • Serverless pricing model would provide a cost-effective solution when processing demand varies significantly throughout the day/week/year with considerable idle time • Aligning the selected services with current and anticipated staff skillset would help accelerate solution development
Solution Example(s)	Informatica Power Center, Talend Data Fabric, Dell Boomi	Azure Data Factory, AWS Glue, GCP Data Fusion

P20W Capabilities | Key Decisions List

There are many key technology decisions that influence the technical design of a P20W Data Pipeline. The decisions on the list below are organized by P20W Capability, and detailed considerations for each decision are elaborated on the slides that follow.

#	Business Capability	Technical Capabilities	Decision
1	Collect	Collect	**Level of prescription of data structure**: To what extent should the P20W data agency prescribe the structure for data submission from contributors?
2			**Control on data provided:** How much direct control will data contributors retain on the data they provide?
3			**Data transfer mechanisms:** What are the methods by which data will be securely transferred to the P20W system?
4		Ingest	**Data transformation:** What technology should the P20W data agency use to perform data transformation (e.g., to a standardized format)?
5		Review	**Data quality:** How will data quality be ensured?
6	Store	Store	**Data storage structure:** What cloud storage structure and services should be selected for various purposes across the P20W data pipeline?
7		Archive	**Data Backup:** What are the options for implementing data durability requirements through backup?
8			**Data Archival:** To optimize the storage cost, while providing the necessary durability to ensure availability, what are the considerations for migrating data between storage tiers, including archival storage class?
9	Manage	Document, Manage Metadata	**Data structure documentation:** What tool(s) should be used for documenting, maintaining and exposing data structures and code lists used by both human and machine for processing and understanding the structure across the P20W data system? (This is sometimes referred to as a data dictionary.)
10		Validate	**Data Quality:** What solution should be used to measure and report data quality issues?
11		Tag	**Data Tagging:** In what format, and using what tool, will the agency collect sensitivity and usage metadata from contributors?

P20W Capabilities | Key Decisions List, cont'd

There are many key technology decisions that influence the technical design of a P20W Data Pipeline. The decisions on the list below are organized by P20W Capability, and detailed considerations for each decision are elaborated on the slides that follow.

#	Business Capability	Technical Capabilities	Decision
12	Master	Resolve, Link	**Data Mastering:** What tools should the Agency use for Master data and Identity management?
13		Link	**Data Linking:** What are the implications of providing linking as a service to data providers, potentially for non-P20W data?
14	Secure	Control	**Data Use Enforcement:** How will data sensitivity and usage classification be enforced?
15		Pseudonymize/Suppress	**Data Suppression:** How will data sensitivity metadata be applied in the pseudonymization and/or suppression of data elements?
16	Enhance	Aggregate, Augment	**Data Enhancement:** What tools should the P20W data agency use to enhance, summarize and augment the data?
17	Analyze	Interpret, Predict	**Data Analysis:** To what extent will the P20W data agency be adding information value through statistical analysis or other analytical techniques?
18	Deliver	Format, Deliver Analysis	**Data Visualization:** What tools will the P20W data agency use for data visualization, e.g., dashboards?
19		Deliver Data	**Data Delivery Methods:** How will the P20W data agency provide sensitive, linked data to requesters?
20			**Data Delivery Methods:** What technology should the P20W data agency use for a Secure Data Enclave?
21		Collaborate	**Collaboration Opportunities:** What are the P20W data agency's required features and solution options for an online research collaboration platform?
22		Search	**Search:** What technology will be used to implement the Data Catalog, and how will it be integrated with the data request/approval/delivery process?

P20W Capabilities | Key Decisions List, cont'd

There are many key technology decisions that influence the technical design of a P20W Data Pipeline. The decisions on the list below are organized by P20W Capability, and detailed considerations for each decision are elaborated on the slides that follow.

#	Business Capability	Technical Capabilities	Decision
23	Automate	Integrate	**Integration:** Are there parts of the solution that require general or specific integration capabilities, beyond what is built into existing solution components that are designed to be interconnected?
24		Workflow	**Workflow:** What tool should the P20W data agency use for managing workflow use cases, e.g., the data request/approval/provisioning process, data submission process, data quality review process, data format/dictionary update process?
25		Instrument	**Instrumentation:** What tools should the P20W data agency use for instrumenting and monitoring the operation of the pipeline and engagement with its data products?

P20W Capability: Collect (1 of 3)

Technical Capability Overview

Collect: Receive data in a secure manner from authorized sources

Key Design Question

Level of prescription of data structure: *To what extent should the P20W data agency prescribe the structure of data submissions from contributors?*

Options	High Level of Prescription	Medium Level of Prescription	Low Level of Prescription
Description	Requires a specific schema for the data, a prescribed encoding format, and agreed-upon standardized code list values, (e.g., M = Male, F = Female, N=Nonbinary, U=Unspecified)	Requires a specific schema for the data, in addition to a prescribed encoding format, but not values	Requires a prescribed encoding format for data submission (e.g., Text/Fixed, XML, CSV, JSON) but allows for multiple schemas and code lists
Considerations & Implications	• Transfers the burden of data standardization to data contributors, potentially reducing data contributors' willingness or ability to provide the data • May positively impact the quality of the data • Increases potential for interoperability of the data outside of the P20W data system • Submission of standardized demographic data (which is notoriously variable or inaccurate) separate from measures data could increase overall data quality	• Reduced burden of data standardization for P20W data agency	• Requires more effort by the P20W data agency to standardize the data across data contributors • Increased ambiguity of quality of the submitted data, if measured
Solution Example(s)	Develop and provide schemas, code lists, semantic definitions (data dictionary) and validation rules to all data providers, with commonly used elements standardized across domains	Develop and provide schemas and definitions to data providers, who supply mappings for their code lists; P20W agency must perform mapping	Data providers supply their own schemas, code lists and definitions; P20W agency must rationalize them into the format used for the linked data set

P20W Capability: Collect (2 of 3)

Technical Capability Overview

Collect: Receive data in a secure manner from authorized sources

Key Design Question

Control on data provided: *How much direct control will data contributors retain over the data they provide?*

Options	High Level of Control	Low Level of Control
Description	Data contributors retain complete control over their data. The P20W agency uses the data per narrow guidelines agreed to with the providers. Data, including aggregations, is deleted if requested.	The P20W agency's data use agreements with the contributors allow for broad usage flexibility. Need for data deletion is not anticipated.
Considerations & Implications	• Mechanisms to enable contributor control of data will need to be established; for example, establishing contributor-controlled data repositories, usage-specific data tagging, and data deletion capability • More willingness of data contributors to share data if they have more control over its use • Slower response time to requests for new, additional reporting and insights from the linked data	• Lower willingness of data contributors to share data under broad use agreement • Faster response time for new data linking and reporting increases the P20W agency's ability to create value for stakeholders
Solution Example(s)	N/A	**N/A**

P20W Capability: Collect (3 of 3)

Technical Capability Overview

Collect: Receive data in a secure manner from authorized sources

Key Design Question

Data transfer mechanisms: *What are the methods by which data will be securely transferred to the P20W system?*

Options	Low Maturity	High Maturity
Description	Data transferred over SFTP with a low degree of automation and mostly manual mechanism for data quality feedback	Data transferred through Web Portal (manual upload) or API (automated integration) with near real-time data quality feedback
Considerations & Implications	• Low effort required to get the SFTP data transfer mechanism established • Manual mechanism for data quality feedback may reduce the overall data quality and processing time • Normally applicable for low frequency of data submission activity from a small number of contributors	• Portal / API development required • Improves data quality due to near real-time feedback • More frequent data submission activity requires higher degree of automation for data transfer and processing
Solution Example(s)	Tibco MFT, AWS Transfer	Customized data submission and management portal and API

P20W Capability: Ingest

Technical Capability Overview

Ingest: Parse and store data in a well-defined structure

Key Design Question

Data transformation: *What technology should the P20W data agency use to perform data transformation (e.g., to a standardized format)?*

Options	Data Transformation Solution Providers	Public Cloud Vendor Services
Description	Specialized, off-the shelf data transformation and master data management solution product	Data transformation services offered by the public cloud providers
Considerations & Implications	• Potentially faster development with a comprehensive solution suite that meets requirements of multiple P20W Capabilities (*Collect, Manage, Master* Capabilities) • A specialized solution vendor has in-house skills to maintain and update the product over time • Potentially higher total cost of ownership vs. usage-based model, unless solution cost is driven down through high-volume enterprise license agreement	• Using data transformation services of the selected public cloud vendor would provide a natural synergy from end to end across the data pipeline tool chain • Serverless pricing model would provide a cost-effective solution when processing demand varies, with significant periods of idle time • Aligning the selected services with current and anticipated skill set would help accelerate solution development
Solution Example(s)	Informatica Power Center, Talend Data Fabric, Dell Boomi	Azure Data Factory, AWS Glue, GCP Data Fusion

P20W Capability: Review

Technical Capability Overview

Review: Review summary of submitted data

Key Design Question

Data quality: *How will data quality be ensured?*

Options	Data Contributor Responsibility	Joint Responsibility	Data Aggregator Responsibility
Description	Data contributors have primary responsibility for providing accurate data, per agreed-upon data definitions	Contributors and the aggregator share responsibility for data quality, with both parties cooperating to perform validation, feedback and correction activities	The aggregator has primary responsibility for data quality, with contributors providing data as-is from their systems
Considerations & Implications	• Managing data quality at the source could be more efficient, as contributors understand their own data best • Reliance on contributors to manage data quality may expose the P20W data agency to the risk of ingesting low quality data, depending on the level of diligence of the contributor	• Contributors manage the specific nuances of their system, and how their submitted data conforms to agreed-upon standards, while the P20W data agency performs automated validation and provides feedback to ensure adherence with the agreed definitions • This approach balances the cost, data quality risk and speed of making data available for insight generation	• Managing the data quality by the P20W data agency could be expensive, due to the need for building out quality mechanisms unique to each contributor • Reliance on the P20W data agency for data quality may cause long cycle time from data submission to being used for insight
Solution Example(s)	N/A	N/A	N/A

P20W Capability: Store

Technical Capability Overview

Store: Securely store data in a flexible structure with appropriate retrieval methods and performance

Key Design Question

Data storage structure: *What cloud storage structure and services should be selected for various purposes across the P20W data pipeline?*

Use Case	Raw Data	Web Applications	Data Warehouse	Reporting Data Mart
Characteristics	• Support file and object data storage (bulk, multi-record, not transactional) upon which structure can be applied • Support hierarchical name space for partitioning • Provide use case specific performance	• Support structured data • Provide low latency to support user interactivity for write and read • Provide auto-scaling capability to match capacity with load	• Provide support for large, typically longitudinal, linked sets of data • Support structured data while accommodating schema drift as those structures change over time • Can tolerate some latency	• Support structured data • Provide low latency to support interactive read
Solution Example(s)	• GCP Cloud storage • Amazon S3 • Azure Data Lake storage	• SQL family – MSSQL, MySQL, PostgreSQL • Document DBs – Cosmos, AWS Document DB, GCP Firestore	• Synapse Virtual SQL • Snowflake Virtual SQL • Synapse Dedicated SQL DW • AWS Redshift • GCP Big Query	• Virtual SQL • SQL Family • Data Mart functionality built into Visualization Tool

P20W Capability: Archive (1 of 2)

Technical Capability Overview

Archive: Retain data as received for as long as needed to accommodate data retention and business continuity requirements

Key Design Question

Data Backup: *What are the options for implementing data durability requirements through backup?*

Options	Solutions Offered by Data Protection Vendors	Data Protection Services of Public Cloud Vendors
Description	Solutions offered by pure play data protection vendors provide robust features for managing data backup	Emerging backup services from public cloud vendors provide native cloud solutions for data protection
Considerations & Implications	• Data backup aligned with overall Data Protection strategy across all storage components could optimize cost and overall durability of the data	• Backup capabilities offered by the Public cloud vendor may be sufficient for the needs of the cloud-based components of the P20W data system, but strategies and options may be different for each data storage type • Portfolio of backup services of Public cloud vendor may not cover on-prem workloads as robustly as solutions from pure-play data protection vendors
Solution Example(s)	Veritas NetBackup, Dell EMC Data Protection Solutions, Commvault Backup and Recovery	Azure Backup, AWS Backup

P20W Capability: Archive (2 of 2)

Technical Capability Overview

Archive: Retain data as received for as long as needed to accommodate data retention and business continuity requirements

Key Design Question

Data Archival: *To optimize the storage cost, while providing the necessary durability to ensure availability, what are the considerations for migrating data between storage tiers, including archival storage class?*

Overview	
Description	Critical data stores across the P20W data pipeline must be protected against loss by selecting adequate levels of durability (storage redundancy) while balancing against retrieval performance requirements
Considerations & Implications	• What is the appropriate default storage tier for each data store within the P20W data system environment? (Choices are typically defined by cloud vendors as high-performance, hot, cool, cold, archive) • What is the required level of redundancy (online storage replication) for each data store? (Choices typically include single-zone multi-copy redundancy, multi-zone redundance, multi-region redundancy) • What classes of data could be moved to lower performance, less expensive storage tiers, including archival storage? Potential candidates may include backup images, raw data files provided by contributors, upstream transitory copies of the data used in the data transformation process • What timeframe of data could be moved to the archival storage? • What timeframe is acceptable to access the data from archive, and what are the use cases for doing so? How frequently might this be needed? • What mechanism should be used to move the data into archive? (Choices typically include manual or automated processes, potentially time-based)
Solution Example(s)	N/A

P20W Capability: Manage & Document Metadata

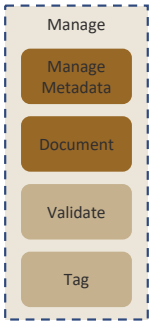
Technical Capability Overview

- **Manage Metadata:** Collect, store and allow retrieval of metadata for use during processing
- **Document**: Publish metadata describing agreed data formats and structures for the collection and publication of data

Key Design Question

Data structure documentation: *What tool(s) should be used for documenting, maintaining and exposing data structures and code lists used by both human and machine for processing and understanding the structure across the P20W data system? (This is sometimes referred to as a data dictionary.)*

Option	
Requirements	This tool needs to support the following use cases: • P20W data agency and data contributors can collaboratively define data structures for submission covering schema, code lists, and validation rules such as data types and allowed value ranges • All data structures, code lists, and validation rules should be version-controlled, so the correct metadata can be applied to each data set in a time-dependent manner • Data sets are described through a catalog containing semantic definition, taxonomy, and descriptive attributes of the data e.g., counts, range • Data is discoverable by internal and eternal users who can search the catalog • Data contributors and consumers can interact with the data dictionary through an online portal • The data dictionary can be applied by automated processing components of the pipeline (e.g., parsing, validation, linking) using the same metadata as as published and available to users
Considerations & Implications	There may be a lack of available solutions for this capability, requiring P20W data agencies to build custom tools, potentially leveraging or expanding upon related solutions to serve all the use cases defined above.
Solution Example(s)	N/A



24

P20W Capability: Validate

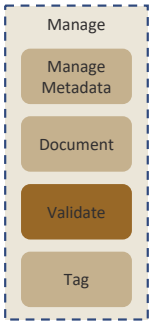
Technical Capability Overview

Validate: Automated validation of collected data against pre-defined constraints

Key Design Question

Data Quality: *What solution should be used to measure and report data quality issues?*

Options	Packaged Data Quality Solutions	Custom Developed Solution
Description	Packaged solution offered by a Data Quality Management software vendor	Custom-developed solution
Considerations & Implications	• Use of familiar, traditional data quality solution could accelerate development • Capabilities of packaged data quality solution may not be fit for purpose for P20W data quality • Packaged data quality solution must provide a platform to extend with customizations that meet any unique data quality needs • Cost and value tradeoffs of the packaged solutions should be considered	• Customized data quality solution using cloud-native services could provide a more fit-for-purpose solution, compared with a less flexible packaged solution • Time required to design and develop the solution should be compared against alternative solutions • Cost of ongoing maintenance and enhancements should be considered for total cost of ownership
Solution Example(s)	Informatica Data Quality, IBM InfoSphere Server for Data Quality, SAS Data Quality	N/A



P20W Capability: Tag

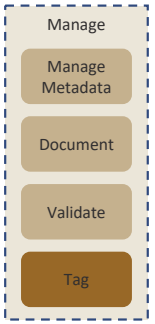
Technical Capability Overview

Tag: Manage and apply data sensitivity and allowable use classification metadata

Key Design Question

Data Tagging: *In what format and using what tool will the P20W data agency collect sensitivity classification and usage metadata from contributors?*

Option	
Requirements	This tool/mechanism needs to support the following use cases: • Data sensitivity and usage classification should be defined for submitted data at both data set and element levels • Data classification and usage allowance may vary by contributor; the system must accommodate alternative or conflicting contributor-specific tags • It should be assumed that data sensitivity and usage classification would not vary by records within submitted data, though withdrawal of consent on an individual basis may need to be accommodated through this or other means • Changes to the classification for the same type of data or element over time would need to be supported, including whether or not changes would apply retroactively • The mechanism for documenting the data structure for submission could be leveraged to collaboratively document sensitivity and usage metadata, collected at the same time as the data format (dictionary) metadata
Considerations & Implications	The classification and sensitivity classification may be the same across contributors for some commonly defined data elements and may be different for other data elements
Solution Example(s)	N/A



P20W Capability: Master (Resolve & Link)

Technical Capability Overview

Resolve: Resolve and maintain unique master IDs for each mastered entity
Link: Link different data sets through predefined or discovered identifiers

Key Design Question

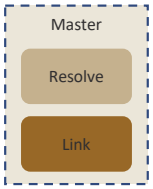
Data Mastering: *What tools should be used for master data management and identity resolution (linking)?*

Options	Packaged Data Quality Solutions	Custom Developed Solution
Description	Packaged solutions offered by Data Management software vendors	Custom-developed solution
Considerations & Implications	• Use of a commonly used, traditional master data management solution could accelerate development of this capability • Capabilities of a packaged MDM solution may be more or less than what is needed for Resolving and Linking P20W data • A packaged data quality solution could provide a platform to extend with customizations to meet unique P20W needs • Cost and value tradeoffs of the packaged solutions should be considered	• Customized mastering solution built upon cloud-native components could provide a more fit-for-purpose solution • Time required to design and develop the solution should be compared against alternative solutions • Cost of ongoing maintenance and enhancements should be considered for total cost of ownership
Solution Example(s)	IBM InfoSphere MDM, Informatica MDM, Tibco EBX MDM	N/A

P20W Capability: Link

Technical Capability Overview

Master: Link different data sets through predefined or discovered identifiers



Key Design Question

Data Linking: *What are the implications of providing linking as a service to data providers, potentially for non-P20W data?*

Option	
Description	As linking is a core capability of the P20W data agency, providing links between the P20W dataset and additional, adjacent data sets could be a high-demand service for the P20W data agency.
Considerations & Implications	• Providing linking as a service could increase interoperability of the data across data partners to create more value from the data • The linking service could be automated through a data pipeline specifically designed for this purpose, with a secure API • Use cases need to be defined; for example, bulk vs. individual linking requests, what categories of entities can be linked, and what data needs to be sent to enable the most accurate matches • In an alternative model, the agency could also provide common linking keys to data providers, to enable them to link with their own data, rather than sending the data to be linked
Solution Example(s)	N/A

P20W Capability: Control

Technical Capability Overview

Control: Role-based user access control to data and systems

Key Design Question

Data Use Enforcement: *How will data sensitivity and usage classification tags be enforced at different points across the system?*

Option	
Description	Design a mechanism by which data sensitivity and usage classification tags specified by the data contributors are applied.
Considerations & Implications	1. The P20W data set generation process would need to apply usage and sensitivity classification requirements, as recorded in the Metadata store 2. Similarly, Data Catalog browsing and query results, and the data request/provisioning process will also need to apply classification requirements 3. Public query builder may also need to refer to the Metadata store to determine data suppression needs, potentially with different results by role (e.g., fewer restrictions for authorized users vs. anonymous public queries) 4. A service interface that meets all use cases will need to be designed
Solution Example(s)	N/A

P20W Capability: Secure (Pseudonymize, Suppress)

Technical Capability Overview

Pseudonymize: Form of deidentification in which Personally Identifiable Information (PII) is replaced with deterministically generated or assigned tokens
Suppress: Aggregated data is evaluated for small cells; primary and complementary data suppression is applied to limit disclosure risk

Key Design Question

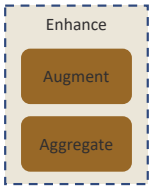
Data Suppression: *How will data sensitivity metadata be applied in the pseudonymization and/or suppression of data elements?*

Option	
Description	Data disclosure protection will need to be enabled through deidentification of data to protect PII. Additional data suppression techniques will need to be applied to protect against data disclosure through potential reidentification.
Considerations & Implications	1. No out-of-the-box commercial software solution has been identified yet to address all of the use cases of pseudonymization and suppression 2. Pseudonymization could be applied as a by-product of the Master –Resolve & Link capability, if the MDM service also generates pseudonymized identifiers to replace PII data 3. Additional approaches are likely needed, such as providing primary and complementary suppression to protect against small cell sizes that can lead to identity disclosure, but recent research suggests these measures may not be sufficient. An emerging practice which applies differential privacy through noise injection to provide directionally accurate but less precise information could be considered.
Solution Example(s)	N/A

P20W Capability: Enhance (Augment, Aggregate)

Technical Capability Overview

- **Aggregate**: Data is summarized at various levels of aggregation, and sliced across various dimensions
- **Augment**: Enhance data with standardized calculated values



Key Design Question

Data Enhancement: *What tools should the P20W data agency use to enhance, summarize and augment the data?*

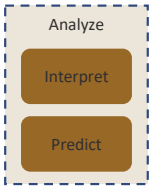
Solution options for the Enhance capability are most likely identical to the transformation tools supporting the Ingest capability.

Options	Data Transformation Solution Providers	Public Cloud Vendor Services
Description	Data transformation solutions provided by independent software vendor with data integration products being the core part of their business	Data transformation services offered by the public cloud providers.
Considerations & Implications	• Faster progress is possible with traditional ETL solutions in the ecosystem, with corresponding in-house or integrator skills • Potentially higher total cost of ownership, unless solution cost is driven down through high volume enterprise license agreements	• Using data transformation services of the selected public cloud vendor would provide a natural synergy across the data pipeline tool chain • Serverless pricing model would provide a cost-effective solution when processing demands vary significantly, with a majority of time spent idle • Aligning the selected services with current skill set would help accelerate progress
Solution Example(s)	Informatica Power Center, Talend Data Fabric, Dell Boomi	Azure Data Factory, AWS Glue, GCP Data Fusion

P20W Capability: Analyze (Interpret, Predict)

Technical Capability Overview

- **Interpret**: Statistical models are developed and applied to derive meaning from historical data (Descriptive Analytics)
- **Predict**: Historical data is used for predictions using statistical and machine leaning techniques, e.g. to predict future academic outcomes (Predictive Analytics)



Key Design Question – Analyze

Data Analysis: *To what extent will the P20W data agency be adding information value through statistical analysis or other analytical techniques?*

Interpretive and predictive analytics add value by finding answers to critical questions using longitudinal linked data. Key considerations that influence a P20W data agency's analytics offering include:

1. To what extent the P20W data agency is chartered to provide research and analytics
2. Technical and business process capacity to provide analysis
3. Clear process and buy-in from data providers on planned methodologies and uses of research and analytics

Tool Set	Interpret	Predict
Considerations	• Ability to generate simple statistical measures for the data – e.g., distribution, median, standard deviation • Capability to conduct complex statistical analysis – clustering, pattern tracking, regression analysis • Ability to slice and filter the data across combinations of dimensions not provided through public reporting dashboards and queries	• Chosen tool would need the ability to create machine learning and AI models • Ability to run on public cloud platforms, ideally using compute resources that are on-demand or optimized to analysis workload • Support MLOPs (Machine Learning Operations) for end-to-end ML model lifecycle management
Solution Example(s)	R, Python, Spark, Jupyter, SAS, Power BI, Tableau, QlikSense, Azure Synapse Analytics, Google Looker or Data Studio	SAS, IBM SPSS Statistics, Dataiku, MathWorks MATLAB, Azure or AWS Machine Learning, Google Vertex AI

P20W Capability: Deliver (Format, Deliver Analysis)

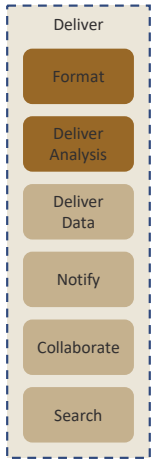
Technical Capability Overview

- **Format**: Statistical summaries are formatted as tables and graphs with added commentary
- **Deliver Analysis:** Results of analysis are delivered via appropriate channels

Key Design Question – Analyze

Data Visualization: *What tools will the P20W data agency use for data visualization, e.g., dashboards?*

Option	
Description	Many good data visualization solutions are available that may meet the requirements of dashboard delivery to P20W data consumers
Considerations & Implications	1. Use of a cloud-native data visualization solution tightly integrated with the other infrastructure components could accelerate development of this capability 2. Solution must have the ability to operate on the selected public cloud platform, and integrate with the necessary data stores 3. Solution must have the ability to embed the visualizations into web pages for publication to users 4. Software licensing structure must provide a model for consumption of analysis by a large number of public (anonymous, unauthenticated, unlicensed) users 5. Solution may have the ability to run on serverless/auto-scaling compute services to dynamically match capacity with demand
Solution Example(s)	Microsoft Power BI, Tableau, Qlik Sense, Google Looker, Amazon QuickSight



P20W Capability: Deliver (Deliver Data, 1 of 2)

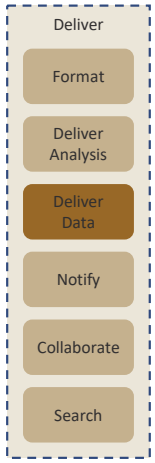
Technical Capability Overview

Deliver Data: Aggregated or individual-level data sets (de-identified or identifiable, per data request) are generated and transferred securely to the requester

Key Design Question – Deliver Data

Data Delivery Methods: *How will the P20W data agency provide sensitive, linked data to requesters?*

Options	Pre-Built Restricted Dashboards	Query Builder	Execute Requesters' Analytical Models	Deliver Data via Secure Enclave	Deliver Data to Trusted Partner Repository
Description	Deliver analysis and data through pre-built access-controlled dashboards	Deliver data through query builder with ability to apply data classification and access control restrictions on-the-fly	Accept and execute analytical models from requesters, returning results without exposing sensitive data	Deliver data to an access-controlled secure enclave which provides hands-on analytical tools	Transfer sensitive requested data to secure cloud repository of trusted partners upon request and approval
Considerations & Implications	• Suitable for regular access to commonly requested, sensitive analysis and data • Request, provisioning, access and revocation process would need to be defined	• Query builder design would need to support dynamic data suppression based on data classification • An access-controlled version could provide greater access to sensitive data, compared with public version	• Capability to build and test analytical models compatible with the P20W analytics platform and data sets must be created • Model execution would need to incorporate data classification and usage requirements	• Agency would need to consider choices of outsourcing secure enclave as a service versus designing and developing its own custom solution	• Request, provisioning and access process would need to be defined
Solution Example(s)	Same as Deliver Analysis capability	Custom solution	Custom solution built on cloud-native components	Outsourced: Coleridge ADRF; In-House: RIPL or custom	Refer to Data Contributor Repository and Data Provisioning components



P20W Capability: Deliver (Deliver Data, 2 of 2)

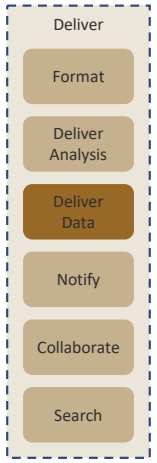
Technical Capability Overview

Deliver Data: Aggregated or individual-level data sets (de-identified or identifiable, per data request) are generated and transferred securely to the requester

Key Design Question – Deliver Data

Data Delivery Methods: *What technology should the P20W data agency use for a Secure Data Enclave?*

Options	Outsource to an ecosystem partner	Leverage technology of ecosystem partners	Assemble components to build own platform
Description	Ecosystem partners provide secure data enclave platforms and administration as a service.	Ecosystem partners like RIPL have developed reference implementations of public cloud components to help automate the creation of a secure data enclave.	Create and operate secure data enclave built on public cloud platform
Considerations & Implications	• Fully managed service for secure data enclave could be used to eliminate development and operations of in-house solution • Request, provisioning and access process and integration with the managed service would need to be defined	• An available reference implementation could be used to accelerate the design and development of a secure enclave capability • Request, provisioning and access process interactions with the enclave component would need to be defined	• Secure data enclave should be designed and built using components from the selected public cloud provider • Request, provisioning and access process would need to be defined
Solution Example(s)	Coleridge ADRF	RIPL Research Data Lakes CloudFormation Templates	Custom solution



P20W Capability: Deliver (Collaborate)

Technical Capability Overview

Collaborate: Users may connect with each other based on expression and matching of mutual interests, and have the ability to share analysis content and commentary in an online community forum

Key Design Question – Collaborate

Collaboration Opportunities: *What are the P20W data agency's required features and solution options for an online research collaboration platform?*

Option	
Description	Enabling researchers and others to collaborate on the P20W data platform is an optional service the P20W data agency may decide to provide, to amplify the value and insight derived from the data it makes available.
Considerations & Implications	1. The P20W data agency must determine the scope of collaboration it will facilitate between researchers; this scope may include collaboration on, for example, data sets, analytical models, analysis tools and techniques, and research results 2. Collaboration likely includes managing the identity of the researchers in the platform, which may require a B2C identity management solution or federation with external identity providers (e.g., Google, Microsoft, Facebook) 3. Based on the scope of collaboration features to be supported, the P20W data agency must select a collaboration platform that meets the requirements 4. What, if any, integration points would a collaboration platform have with, e.g., the P20W Dashboards, Query Builder and Data Catalog?
Solution Example(s)	• Purpose-built, managed data collaboration platform, e.g., Coleridge ADRF • General research collaboration platforms, e.g., Open Science Framework • Broad online community platforms, e.g., Hivebrite

P20W Capability: Deliver (Search)

Technical Capability Overview

Search: Users can search, filter and browse rich metadata in order to find the most relevant content available on the platform, including data sets

Key Design Question – Search

Search: What technology will be used to implement the Data Catalog, and how will it be integrated with the data request/approval/delivery process?

Considerations & Implications

1. The agency will need to design for and incorporate the synergy between Data Catalog and Data Dictionary; the two should be integrated in order to link reference material with data set metadata and summaries; a well integrated solution that includes both Data Dictionary and Data Catalog capabilities may be preferred, if one can be found
2. The Data Catalog search result interface should provide an easy means by which to access data that is found (if public or preauthorized), or to request desired data; this means that public data sources should be included in the data catalog, and it should be integrated with the data request/approval/provisioning process for data that requires this level of control
3. Data Catalog should have robust search capability to discover available data, including full-text search, faceting/filtering, and other common and useful discovery features
4. Data Sets should be summarized in codebook form (metadata that contains statistical summarization of each variable or measure contained in the data set), and researchers should be able to use this rich descriptive metadata to discover relevant data sets based on a selection of desired characteristics of the data

P20W Capability: Automate (Integrate)

Technical Capability Overview

Integrate: A data integration platform to orchestrate automated transformation and flow of data between components

Key Design Question – Integrate

Integration: Are there parts of the solution that require general or specific integration capabilities, beyond what is built into existing solution components that are designed to be interconnected?

Considerations & Implications

1. Will the agency need to public cloud-based data and services with on-premises systems?
2. Will the agency need to capture and process streaming data?
3. Will the agency need to integrate third-party SaaS/PaaS solutions with
4. Will the agency build any complex, distributed applications or services that require an Enterprise Service Bus (ESB), microservice architecture, event-driven architecture, or other middleware technology?

If the answer to any of the above questions is “yes”, then what are the integration requirements, and can they be met by other components in the solution architecture? Do any additional integration solutions (middleware) need to be added to the design and procured?

P20W Capability: Automate (Workflow)

Technical Capability Overview

Business process workflows (i.e., orchestrated and repeatable patterns of activity) that support the data pipeline are designed, managed and automated

Key Design Question – Workflow

Workflow: What tool should the P20W data agency use for managing workflow use cases, e.g., the data request/approval/provisioning process, data submission process, data quality review process, data format/dictionary update process?

Considerations & Implications

1. Consider solution candidate choices, which may include cloud-native workflow tools (e.g., Azure Power Automate, AWS Simple Workflow Service) or best-of-breed workflow management tools (e.g., Appian, Pega Platform, Decisions, etc.)
2. The tool should be able to support a number of key capabilities, including a "low-code" development model, role-based work queues, notifications, exception and escalation branches, automated decision logic, event-based triggers, and the ability to manage long-running processes over the required durations
3. The tool should include customizable user interfaces components for form data submission, to indicate decisions and acknowledgments, electronic signatures, etc.
4. The tool must be able to interface easily with other system components through common integration technologies, e.g., REST API calls, SQL queries, identity and access integration
5. The tool must allow business process owners to manage the processes and user/role assignments
6. A solution for managing the status, legal agreements and other attributes of relationships with external partner organizations and individuals is needed, either as part of the Workflow solution, or a separate partner management capability integrated with the appropriate business process workflows

P20W Capability: Automate (Instrument)

Technical Capability Overview

Measure system performance and user engagement with data and reports; establish feedback mechanism for improvement

Key Design Question – Instrument

Instrumentation: What tools should the P20W data agency use for instrumenting and monitoring the operation of the pipeline and engagement with its data products?

Considerations & Implications

1. The agency must consider the functionality required and choose between cloud-native instrumentation and analytics tools (e.g., Azure Monitor, AWS Cloud Watch) or a best-of-breed monitoring solution (e.g., Dynatrace, NewRelic)
2. A pure-play log analytics platform might be used as an alternative, if service logs are collected through built-in functionality; examples include Azure Monitor, AWS CloudWatch, Splunk
3. The instrumentations tools would need to cover the spectrum of application, cloud infrastructure and website performance and usage monitoring, and would provide both operational and business intelligence levels of analytics
4. Additional capabilities or requirements may include monitoring the performance of website transactions, code-level performance profiling, monitoring of cloud data pipeline performance, storage capacity and performance, error and exception monitoring, etc.
5. A potentially separate solution may be required to provide security monitoring; however, operational security and risk management is out of scope of this document and must be addressed as part of a complete information security program