EXHIBIT F

Feasibility and Cost Study on the Development of a Statewide Data Warehouse Program

Findings and Recommendations Report

Prepared by the Arkansas Data and Transparency Panel

Table of Contents

Executive Summary:

The purpose of this study is to evaluate the feasibility and costs of the development of a statewide data warehouse program, its ability to address identified gaps and objectives, and to make recommendations for an optimal implementation approach based on existing public sector programs and industry best practices.

In Act 1282 of 2015, the general assembly identified needs for increased data access, use, and management and called for an Open Data and Transparency Task force to study these findings and make recommendations. An Arkansas Legislative Audit special report published in 2015 identified the potential benefits of a centralized data warehousing approach in addressing some of the findings identified. In January 2017, the Open Data and Transparency Task force recommended a cost and feasibility study for a statewide data warehouse program, which was codified by Act 912. A gap analysis with executive agency leadership reinforced data integration, data sharing, and data access as top needs, with a focus on a citizen master record, and subsequent 2019 legislation emphasized the need for real-time data access for continuous program alignment with evolving needs.

The identified needs and recommendations were synthesized into six program objectives aimed at establishing and supporting the accessibility, sharing, integration, quality, timeliness, and analysis of state data assets. The objectives identified are:

- **Enable openness, transparency, and pervasive, self-service data access and delivery**
- **Share data to enhance its value while enforcing privacy and security**
- **Support data-driven decision making and analytic maturity through development and support of analytic skills and shared services**
- **Integrate data for improved cross-agency analysis and reduced duplication of data and efforts**
- **Enable real-time assessment, support, alignment, and automation of decisions, programs, and resources**
- **Manage enterprise data as a state asset.**

As part of this study, research was conducted of existing data warehouse, analytics, and information management programs in multiple US states, 1 Australian state, and a Canadian federal agency. Industry analyst interactions and a study of published research were leveraged to assess current best practices from both public and private sector. The study found many successful examples using a variety of different approaches that resulted in cost savings, operational efficiencies, increased program performance, and effective policy insights.

Feasibility of addressing the identified objectives was found to be high with key recommended approaches including:

- **Pervasive Business Intelligence, web, mobile, and Application Programming Interfaces for data access**
- **Compliant, secure data sharing, via a layered approach of authorized role and organization-based access to identified data and broader access to deidentified aggregate data**
- **Analytic tools and training for increased value from existing data and more data-driven decision making**
- **Integration of cross-agency data via a Data Hub approach leveraging centralized Master Data Management**
- **Real-time data access supported as needed via data federation**
- **Self-service, agency-level Data Quality Management, Master Data Management, and Stewardship at the source**
- **Comprehensive Data Governance for standards, security, privacy, compliance, and change management**

The annual cost to meet Arkansas' identified needs is estimated at $4M. The cost of other statewide data programs is over $8-9M annually with larger efforts having up to 30 support staff. This is a significant investment, but other states report returns from statewide data warehousing programs ranging from $25M per year to $1M per day. Other programs use a combination of funding strategies including federal and private grants, general revenue appropriations, and usage-based chargeback.

The key recommended next steps include:

- **Develop a program charter to formalize the scope of the program**
- **Formalize a multi-department agreement to allow departments for compliant, secure and efficient data sharing**
- **Determination of initial and sustaining program funding approaches**
- **Enablement of a data hub (value-driven broker for cross-agency data sharing and analytics) to act as an agent of individual agencies in integrating and providing secure, compliant access to and analysis of inter-agency data**

Background

Act 1282 of 2015 outlined findings related to state data and formed the Open Data and Transparency Task Force (ODTF) to determine the best practices for maintaining and delivering state data. The task force met throughout calendar year 2016 and published their findings and recommendations in January 2017. Act 912 of 2017 included many of the task force's recommendations. Two key provisions of this act were the creation of the position of Chief Data Officer (CDO) and the completion of a feasibility and cost study on the development of a statewide data warehouse program. The purpose of this study is to **evaluate the feasibility and cost** of implementing a statewide data warehouse program and to **recommend best practices and approaches** for addressing the identified state data needs.

Methodology

The approach taken to conduct this study included:

- Analysis of identified Arkansas data needs and opportunities as outlined by the general assembly, Open Data and Transparency Task Force, and executive agency leadership
- Synthesis of related identified needs into program objectives
- Study of current data management and analytic best practices from research and advisory firm Gartner and data management professional organizations via analyst inquiries and analysis of published research.
- Study of other public sector entities[1] who have implemented centralized data warehouse programs or other approaches to the delivery and support of integrated data management and analytics
    - Analysis of published literature and presentations
    - Interviews with public sector chief data officers and analytic program leadership
    - Analysis of budget data from state transparency portals
- Application of successful best practice approaches to program objectives and feasibility assessment
- Cost and benefit estimation of implementing the identified approaches
- Recommendation of next steps towards program implementation

Figure 1 - Study Methodology

[1]Alabama, Connecticut, Florida, Indiana, Iowa, Kentucky, Massachusetts, Michigan, Minnesota, Mississippi, New Jersey, New South Wales Australia, North Carolina, Tennessee, Texas, Treasury Board of Canada

Identified Data Needs

Act 1282 of 2015

The General Assembly identified the following data-related findings in section 1.b.1 of Act 1282 of 2015 (To create the Open Data and Transparency Task Force to determine the best practices for the state to achieve the most efficient system for maintaining and delivering the state's public records and data.):

(A) State agencies contain great amounts of valuable information and reports on all aspects of life for the citizens of this state, including without limitation health, business, public safety, labor, and transportation data;
(B) The tremendous amount of data maintained by state agencies can result in **duplication of efforts, data, records, and parts of data and records** that may result in the maintenance of **inconsistent data and records concerning the same citizen**;
(C) The **lack of a quick and efficient delivery system** to respond to legislative and executive branch inquiries is harmful to the policy-making process and ultimately costs taxpayers money;
(D) Progressive states have evolved to **become data-driven governments that use data as a strategic asset** to improve the delivery of services to the state's citizens and to **become more efficient stewards of citizens' data**;
(E) **Ensuring the quality and consistency of public data** is essential to maintaining the data's value and utility;
(F) New information technology has fundamentally changed the way people search for and expect to find information and can aggregate large quantities of data to allow the state to **provide better information to citizens** with increasing efficiency and thoroughness; and
(G) The state should:
    (i) Evaluate ways to appropriately, efficiently, and securely **share data between and within state agencies** to **allow for quicker, more impactful cross-agency analysis** to allow policymakers to make quicker, more informed decisions; and
    (ii) Use the innovations in information technology to **enhance public access to public data** to make the state more transparent and to promote public trust while **eliminating waste, fraud, and abuse in the execution and delivery of government services**.

Arkansas Legislative Audit Special Report on Centralized Data Warehousing

A special report published by Arkansas Legislative Audit (ALA) in November 2015 on "Potential Benefits of a Centralized Data Warehouse for the State of Arkansas" studied other states, counties, and municipalities that are leveraging centralized data warehouses to reduce tax and program fraud, conserve state dollars, enhance public safety efforts, and improve health care. The identified benefits of centralized data management and access include:

- **Appropriate and authorized access** to large data sets for reporting and analytics.
- **Improved quality and accuracy of data**.
- **Sharing of data** among state and local entities.
- Greater efficiency through **reduction of duplicate efforts**.

The study made the following recommendations for steps towards implementation of a statewide data warehouse:

- Regarding a statewide centralized data warehouse, ALA staff recommend that the General Assembly consider:
    - **Authorizing a feasibility study** identifying the IT requirements and costs associated with centralized data warehousing.
    - Creating legislation **authorizing a Chief Data Officer** to lead the State's research into and potential development and implementation of a centralized data warehouse project.
- Should the feasibility study conclude that a centralized data warehouse would be beneficial to the State, ALA staff recommend that, during the development and implementation process:
    - Access to the centralized data warehouse is controlled
    - Secure transmission and storage of data are ensured.
    - Current facilities and other resources available are used.

Open Data and Transparency Task Force and Act 912 of 2017

Per Act 1282 of 2015, the Open Data and Transparency Task Force (ODTF) met throughout calendar year 2016 to study best practices related to the identified findings. Based on recommendations from the ALA special report and other interviews, testimony, and analysis, the ODTF determined that:

"A feasibility and cost study should be performed to **determine the specific requirements needed for a statewide data warehouse solution**. At a minimum, the hardware, software, physical location, staffing and communication networking requirements should be considered and evaluated. A cost/benefit analysis should also be performed."

"The results of the feasibility and cost study should be provided to the General Assembly so that the objectives, magnitude, and scope of the data warehouse program may be revised as appropriate. At this time the General Assembly may take action on appropriating any necessary funds for the remainder of the program."

Along with the creation of the positions of Chief Data Officer and Chief Privacy Officer and the formation of the Data and Transparency Panel (DTP), this requirement was codified as A.C.A. § 25-4-127.c.1 by Act 912 of 2017.

CDO Gap Analysis of Agency Data Needs

Upon formation in September 2017, the office of the Chief Data Officer (CDO) conducted a gap analysis of data and data management capabilities needing improvement. Interviews were held with the director and executive staff of each executive agency resulting in an inventory of identified gaps/needs. Analysis of the identified gaps identified the following core data needs:

Data Need	Percentage of Responses
Data Sharing	48.9%
Data Integration	26.7%
Data Access	6.7%
Data Quality	6.7%
Data Standardization	6.7%
Data Governance	4.4%

Figure 2 - Gap Analysis Identified Data Needs

The CDO team recommended undertaking three major initiatives to address the gaps identified in the analysis and to increase the overall data management capability and maturity across all agencies. These initiatives include:

- Creation of a Comprehensive **Multi-Agency Data Sharing Agreement**
- Implementation of a **Master Data Management** (MDM) System and **Master Citizen Record**
    - A central MDM hub ensures consistent identification
    - A central MDM hub affords more accurate identification
    - A central MDM Hub saves time and effort in data integration
- Adoption of a **Data Governance Model** and **Data Quality Management Standards**
    - Creating a data catalog
    - Creating a data asset inventory
    - Establish data governance framework
    - Establish data quality management program

Act 936 of 2019

During the 2019 legislative session, Act 936 amended the law concerning the duties of the data and transparency panel to add additional member agencies and to add the following duties:

(7) Develop a unified longitudinal system that **links existing siloed agency information** for education and workforce outcomes to **continuously conduct a business systems assessment** to:

(A) Help the leaders of this state and service providers **develop an improved understanding of individual outcomes** resulting from education and workforce pipelines in Arkansas;

(B) **Identify opportunities for improvement by using real-time information**; and

(C) **Continuously align programs and resources** to the evolving economy of this state.

Act 943 of 2019

Act 943 of 2019 (To create the data sharing and data-driven decision-making task force) reiterated the findings listed in Act 1282 of 2015, noted that the ODTF began to address these problems through creation of the CDO, CPO, and DTP, and found that:

(6) The state should continue those efforts by **evaluating ways to implement a shared services model for statewide data sharing** in order to drive innovation and facilitate efficiency across state agencies, improve the delivery of services, and to better serve the citizens of this state.

To this end, the act created the **Data-Sharing and Data-Driven Decision-Making Task Force.** The task force shall meet between July 1, 2019 and December 31, 2019 and file a written report of its activities, findings, and recommendations.

The task force shall:

(B) Recommend specific solutions and legislation necessary to **create a statewide data sharing system** for maintaining and sharing public data that is owned, controlled, collected, or maintained by a state agency; and

(H) Recommend funding mechanisms to support the use of statewide data sharing, including without limitation **data analytics, machine learning**, and **innovative technologies to link data between agencies, to support data driven decision making** for all state agencies.

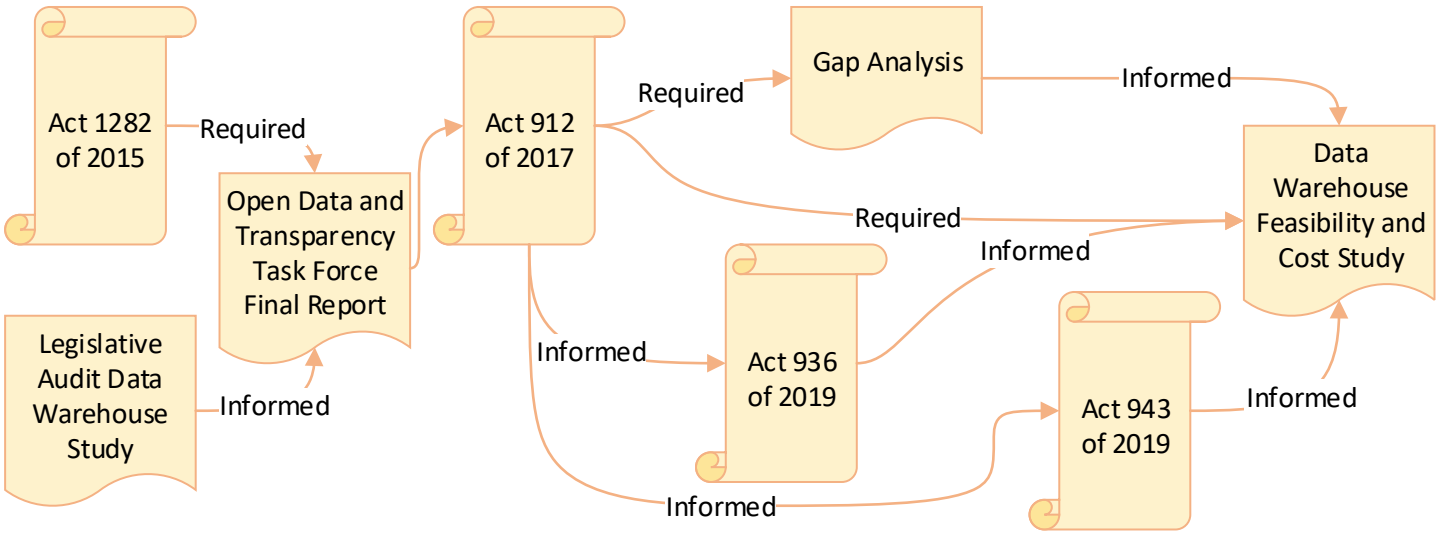


Figure 3 - Legislation and Studies Requiring or Informing the Data Warehouse Study

Program Objectives

The identified needs were synthesized into six program objectives aimed at establishing and supporting the accessibility, sharing, integration, quality, timeliness, and analysis of state data assets.

Identified Requirement / Need	Category	Program Objective
Improve data access and delivery (Act 1282 of 2015, Gap Analysis)	**Accessibility**	**Enable openness, transparency, and pervasive, self-service data access and delivery**
Enhance accessibility and quality of public data (Act 1282 of 2015)		
Improve data sharing between and within state agencies (Act 1282 of 2015)(Gap Analysis)	**Sharing**	**Share data to enhance its value while enforcing privacy and security**
The state should continue those efforts by evaluating ways to implement a shared services model for statewide data sharing (Act 943 of 2019)		
Create a statewide data sharing system for maintaining and sharing public data that is owned, controlled, collected, or maintained by a state agency (Act 943 of 2019)		
Become more data driven (Act 1282 of 2015)	**Analysis**	**Support data-driven decision making and analytic maturity through development and support of analytic skills and shared services**
Data analytics, machine learning, and innovative technologies to link data between agencies, to support data driven decision making (Act 943 of 2019)		
Improved capacity for cross-agency analysis (Act 1282 of 2015)	**Integration**	**Integrate data for improved cross-agency analysis and reduced duplication of data and efforts**
Reduce duplication of efforts and data (Act 1282 of 2015)		
Inconsistent data and records concerning the same citizen (Act 1282 of 2015)		
Eliminate waste, fraud, and abuse in the execution and delivery of government services (Act 1282 of 2015)		
Links existing siloed agency information (Act 936 of 2019)		
Implementation of a Master Data Management System and Master Citizen Record (Gap Analysis)		
Continuously conduct a business systems assessment (Act 936 of 2019)	**Timeliness**	**Enable real-time assessment, support, alignment, and automation of decisions, programs, and resources**
Identify opportunities for improvement by using real-time information (Act 936 of 2019)		
Continuously align programs and resources (Act 936 of 2019)		
Improve data stewardship (Act 1282 of 2015, Gap Analysis)	**Quality**	**Manage enterprise data as a state asset**
Ensure data quality (Act 1282 of 2015)		
Adoption of a Data Governance Model and Data Quality Management Standards (Gap Analysis)		

These objectives align with the top three business expectations identified by respondents in the Gartner 2017 Third Annual Chief Data Officer survey which included "**enhance data quality, reliability and access**", "**enhance analytical decision making**" and "**create internal and/or operational efficiencies**."

Data Warehouse Program Approaches

Act 912 of 2017 called for a feasibility and cost study on the development of a **statewide data warehouse program**.

In defining a data warehouse program, it is important to note that:

A) The **definition of and approaches** used to deliver the intended benefits of data warehousing **have evolved over time** and **continue to mature** rapidly in response to evolving needs such as new data types, new processing requirements, data science, and faster development cycles as well as the advent of new enabling technologies.

B) As noted in the Open Data and Transparency Task Force Final Report, **data warehousing alone can only partially address the findings** identified in Act 1282. In order to fully address all findings, **additional complementary solution components** such as business intelligence, data quality management, and master data management **are required**. These are common components of mature data warehousing or enterprise information management programs and were **considered as part of the scope of a statewide data warehouse program**.

The following is a brief summary of the common styles of data warehousing currently in use.

Traditional Data Warehouse

A traditional repository-style data warehouse, or enterprise data warehouse (EDW), is a **collection of data** in which **disparate data sources** can be brought together in an **integrated, time-variant** information management strategy.

This style of data warehouse generally:

- Houses **well-known and structured data** in a **persistent repository** of data that is copied from source systems
- Supports high performance access for **well-known, predefined and repeatable analytics needs**
- Contains data that is **highly modeled** during the development process
- Transforms data during **batch extract, transform, and load (ETL) processes**, typically nightly
- Supports **retention of historical data** that may not be retained by the source systems
- May support smaller, subject-oriented datamarts for specific analytic needs

There are numerous successful examples of traditional repository-style data warehouses that support program or agency-level needs currently deployed across Arkansas state government, some since the 1990's.

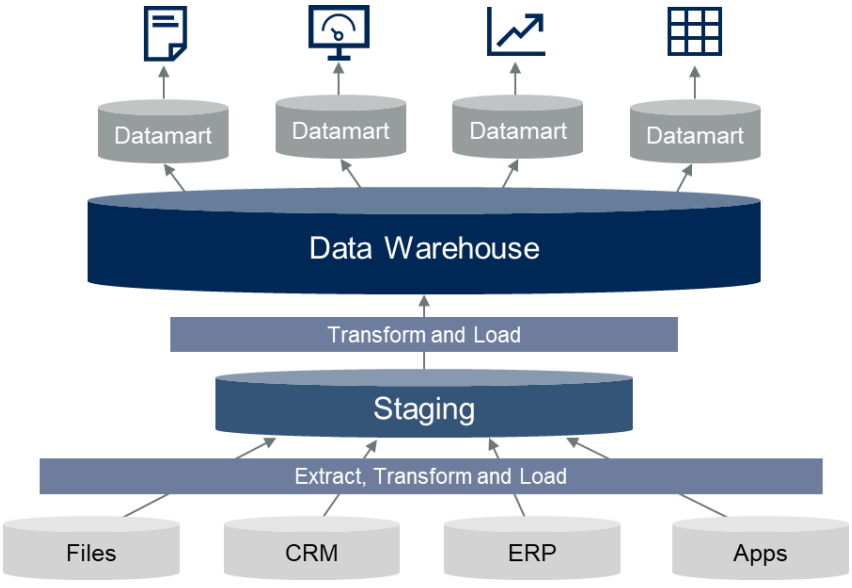


Figure 4 - Traditional Repository-Style Data Warehouse (Source: Gartner)

Data Lake

The concept of the Data Lake emerged in the mid-2000's to **supplement the Traditional Data Warehouse** in order to **support new data types, new processing requirements**, and the **need for development agility**.

A data lake is a **collection of generally raw data and events** that have been captured and ingested for use with **limited structure, transformation** to a specific schema, **or quality assurance**. Data lakes usually support **exploratory analysis**, **rapid prototyping, and data science activities**, often with data of **unknown quality and utility**. The data lake generally supports **minimal controls for data governance** and is typically intended more for use by data analytics professionals than by business users.

The data lake differs from the data warehouse primarily in that the data isn't modeled prior to use. This has the benefit of reducing the time to transform and ingest the data, which increases agility and reduces time wasted on data that is found to be of limited value or that was required for one-time use. Because the data isn't pre-modeled for specific known uses, it has more flexibility. However, this requires the data to be modeled by the user upon retrieval, requiring more time and technical capability.

Data lakes have gotten a poor reputation as "data swamps" because of the lack of quality controls, but this stems from a frequent misconception of the intended purpose of the data lake. The data lake's goal is to exploit new, unknown data sources. The Arkansas Department of Information Systems (DIS) leverages the data lake as a means for rapid ingestion and use of new data in conjunction with the stable, governed data warehouse environment. If the value, utility, and quality of data are established, they are promoted into the more governed environment where they are more intentionally modeled, placed under stewardship, and made available for widespread use.

Data lakes are sometimes misunderstood as a replacement for data warehouses, but **modern data and analytics initiatives require the capabilities of both types**. The data warehouse is optimized for performance-oriented, scalable, repeatable delivery on consistent data. The data lake is optimized for rapid access, free-form discovery, and flexibility. These two optimization goals are at odds, so these environments are complementary, not interchangeable. The evolving form of the logical data warehouse encompasses both concepts.

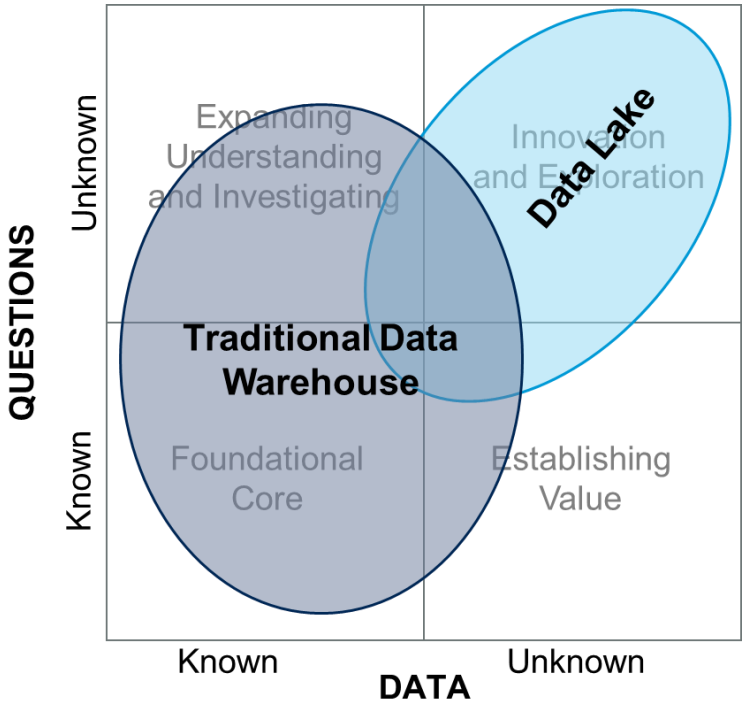


Figure 5 - Data Lake Relationship to Data Warehouse (Source: Gartner)

Logical Data Warehouse

In the early 2010's, the Logical Data Warehouse (LDW) brought the data warehouse and data lake together into a **logically consolidated view across all types of data**. It also introduced **Data Federation as the primary interface** for data access and analysis over multiple data sources.

Data federation is where the data stored in multiple sources (of the same or different type) are made accessible to data consumers by using **on-demand data integration**, rather than executing data movement and physically storing integrated data. Data federation acts as a **real time virtual repository** that hides the underlying complexity of the data landscape from consumers so that changes can be made in a nondisruptive way behind the scenes of the data federation tier.

The logical data warehouse **prioritizes data connection over data collection** but **still includes the data warehouse** as a system of compromise for meeting performance or historical retention needs, data lakes for agility, and possibly other forms of data sources and stores such as event processing or streaming data.

The Arkansas Department of Information Systems adopted the Logical Data Warehouse architecture in 2013 and has experienced many benefits including faster development cycles (and lower accompanying development costs), support for real-time reporting, lower storage costs, and reduced impact from changes to the underlying data sources. Some critical components required to enable maturation to logical data warehousing included the establishment of agency-level master data management and the accompanying stewardship and governance over naming and coding of data, the inclusion of a data federation layer in the agency's business intelligence platform, and support for business intelligence semantic layers that can span multiple distributed data sources of different types. These topics are covered in more depth later in this study. While the agency uses a federate-first strategy to decision support, traditional data warehouse and data lake approaches are still key components as needed for performance, historical retention, and agility.

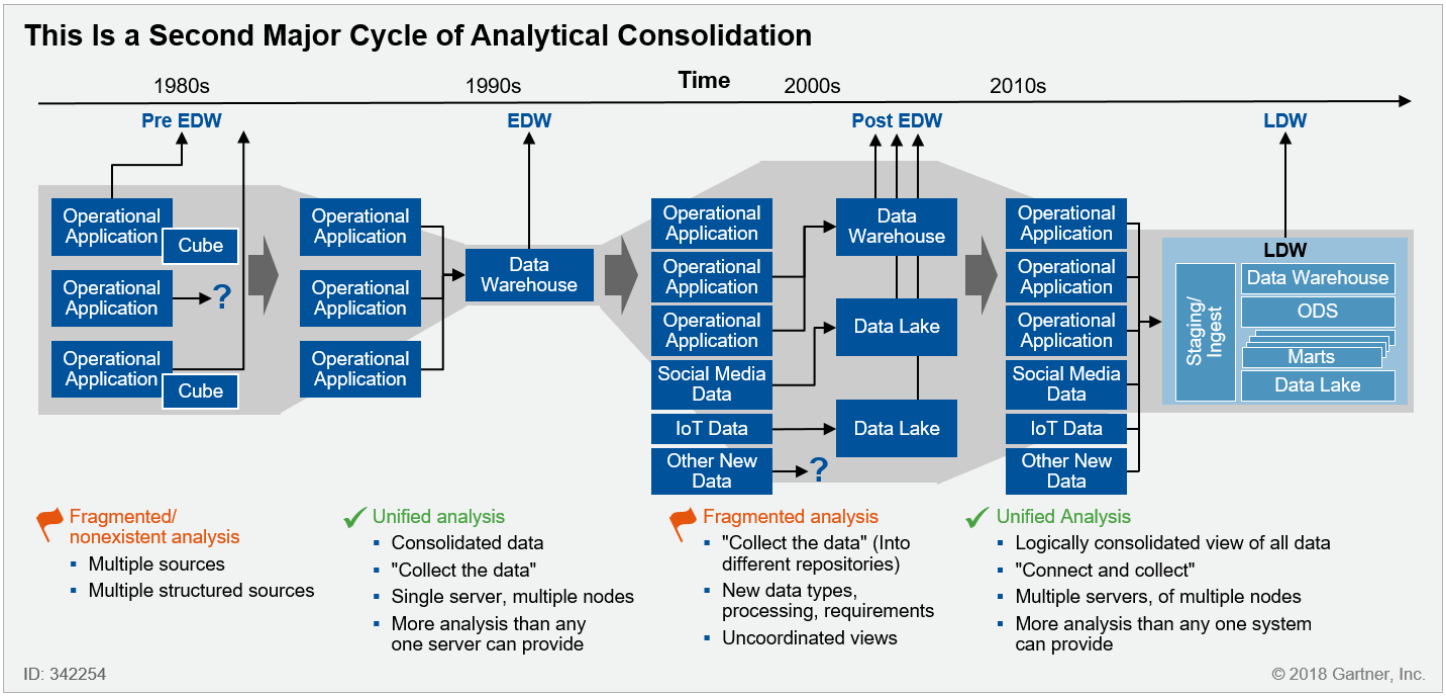


Figure 6 - Evolution of Logical Data Warehouse

Data Hub

The most recent evolution of approaches for serving data sharing and access needs is the Data Hub, an architectural pattern for enabling the **seamless flow and governance** of data across a range of business systems**, including applications, data warehouses and data lakes**.

Producers and consumers of data connect with each other via the data hub, with governance controls and standard models applied to enable effective data sharing. Predicated on the concept of common models, data hubs are mainly focused on driving consistent semantics (consistent naming and meaning) but can support a range of use cases (both operational and analytical) and processing strategies.

Data hubs are about sharing of data with effective governance. By having endpoints (systems, processes and organizations) that connect to the hub as either providers or consumers of data, the hub becomes a point of mediation and a place to apply governance controls. The data hub enables better scalability and manageability of data flows, because it provides transparency and independence between data producers and consumers.

Many organizations have identified a need to share data faster, but they also want to maintain governance guardrails. This is where a hub-centric approach comes in. In this context, the data hub acts as a consistent arbiter between producers and consumers of data, providing consistent semantics and governance policy across operational use cases.

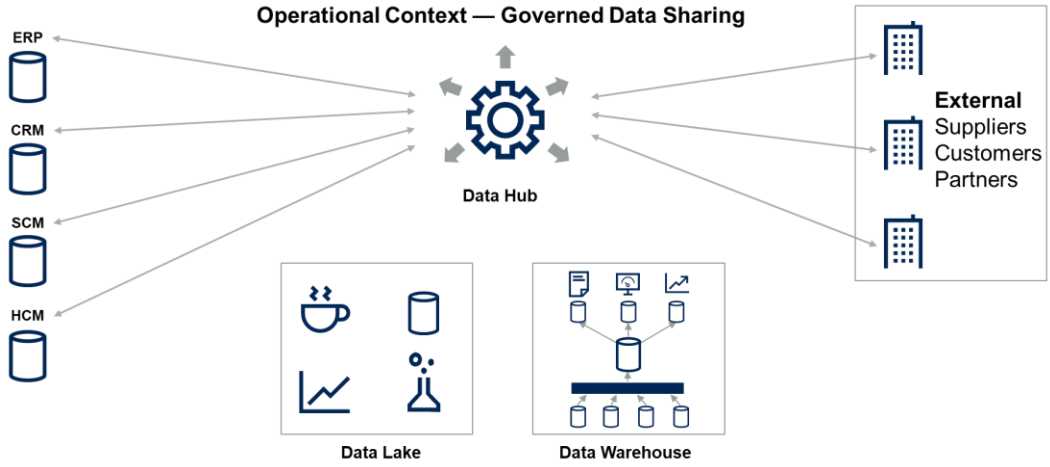


Figure 7 - Data Hub (Source: Gartner)

Recommended Approach

In order to support all identified program objectives, it is recommended to use the current "data hub" approach, which includes the use of more traditional approaches where applicable.

- Use the data hub's capabilities for enabling seamless sharing of all types of data.
- Use the data hub's capabilities around consistent semantics to feed both data warehouses and the data lake.
- Use a data warehouse as needed to collect and organize data for known questions and data.
- Use a data lake for unknown, less-structured data, unclear questions and discovery-oriented analysis.

Including a combination of interconnected data warehouse, data lake and data hub capabilities will enable:

- A wider range of use cases (both analytical and operational)
- A greater reach in connecting data assets across the state
- A more flexible approach to governing and provisioning data

Key Case Studies

State of Indiana

One of the most well publicized and successful statewide data warehousing, sharing, and analytics programs is Indiana's Management Performance Hub (MPH).

History

The MPH is an advanced data analytics management system for the State of Indiana government that performs business analytics and in-depth case studies on specific problems. MPH also houses a public transparency site enabling citizens to access a variety of real-time information about state government.

The MPH initiative began with the kick-off of an infant mortality study in August 2013 and was followed with an executive order by Governor Mike Pence in March 2014 requiring interagency data sharing with the MPH. The MPH website was launched in June 2014, and a physical MPH center opened in September 2014. In 2017, House bill 1470 codified and funded MPH as a state agency.

Approach

Over 5 trillion rows of data are received from agencies via flat file and loaded into a 3TB in-memory data platform. Access to detailed data is highly controlled in the most secure system in the state. This system is used for data integration and for data science tasks across the full breadth of state data. Analysts who work directly with the privacy data have all undergone background checks and work in a secure room. There are also strict agreements in place on how the data can be used in the form of MOUs. A separate instance houses anonymized data for broader agency and public consumption. Much of the anonymized data is available for public access on the Indiana Data Hub.

The implementation of in memory database technologies has greatly improved query speeds (1000x faster than SQL for simple queries, 5000x faster than SQL for complex queries) and data compression (90+% compression).

Results

MPH has conducted studies to guide policy and increase transparency in areas including recidivism, income tax fraud, highway crash prediction, Medicaid optimization, education and workforce, and the opioid epidemic.

MPH automatically calculates real time performance measures for all executive agencies.

Now that data is accessible in one location and can be pulled quickly, analysts can spend more time examining data.

Cost

MPH currently has 20 full time employees and an annual budget of $9M consisting of $6M in general funds, $1.7M in database management funds, and $1.3M from the Department of Insurance fund.

Lessons Learned

Don't try to include all state data at once. Take a more phased approach. Start with use cases that make a difference.

Treasury Board of Canada

The Treasury Board of Canada Secretariat has had great results by starting with widespread access to data.

History

The Treasury Board of Canada Secretariat (TBS) is a Federal agency that manages 250K employees across 86 departments and an annual budget of $250B CAD. The agency's strategic outcome is that government is well managed and accountable, and that resources are allocated to achieve results. Faced with 100's of silos, little data sharing, and inconsistent reporting, TBS developed and began implementing a phased, multi-year data sharing and access plan.

Approach

TBS started by conducting analysis and research on leading public sector examples and showing "Art of the Possible" examples to executive stakeholders. They formed a strategy in 2013 to consolidate data silos, encourage and enable data sharing, and establish common data elements. In early 2014, TBS visited Washington, DC to visit with the Recovery Accountability and Transparency Board for a deep dive into the Recovery.Gov implementation to learn from a successful example.

Software was purchased in March 2014 and an enterprise data warehouse was developed, starting with mature, standardized data sets. Data was provisioned from multiple sources and formats and included direct connection to source data for operational reports. Each business unit has data asset owners with stewardship over their own data in the data warehouse. In Fall 2014, the Central Online Reporting System (CORS) was launched to early adopter clients at first and then rolled out to 250K users starting with ad-hoc and predefined Business Intelligence reporting.

New capabilities have been rolled out yearly including visual data discovery, dashboards, planning and budgeting, and in-memory analytics. Their focus is on master data management and providing tools, training, and hosting. Their goal is not to provide reporting but to provide departments the ability to provide their own reporting.

Results

CORS has had an excellent adoption rate and resulted in gains in operating efficiencies. Stakeholders who used to wait 2 weeks for someone else to provide data are now accessing data on their own instantly during meetings. As the analytic maturity of all users evolved, they began using more advanced tools and the users are now demanding predictive and prescriptive analytics and driving the business case for growing the program. Agencies are even contributing staff to the development and admin team.

Cost

The cost of this program was not available. It started with centralized funding and is moving to a chargeback model for recovery of consumption-based expenditures as adoption and use grows. The starting team size was 12.

Lessons Learned

- Find "real world" public sector examples
- Strategic "business-focused" IT model
- Leverage a multi-year project plan
- Communicate project status updates and outcomes clearly and often
- They trained staff proactively before project
- User groups
- Super end users from business embedded in technology team during development sprints
- Core active users drive future direction

State of Michigan

Michigan's statewide data warehouse is the longest running and has the highest reported return on investment.

History

Michigan's statewide data warehouse started out in 1993 as the first Medicaid data warehouse in the nation and had an early focus on Temporary Assistance for Needy Families (TANF) and Medicaid fraud. It has now grown to span 21 agencies and more than 10K users from across Michigan state government. It is operated by the Data Center Operations group within the Department of Technology, Management, and Budget (DTMB).

Approach

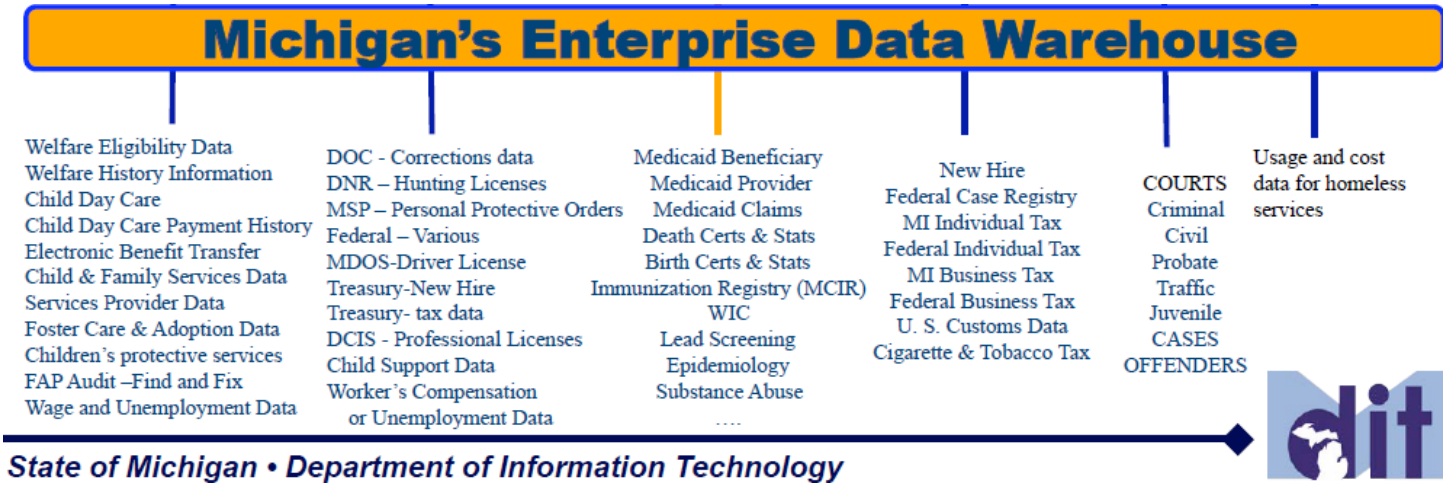
Michigan uses a traditional repository-style data warehouse. A unique client identifier (UCI) has been in place for many years to assist with the unique identification of individuals. The UCI is now being replaced by the Master Person Index (MPI) from the state's Master Data Management (MDM) platform, which was established to improve the quality of data held by each of the departmental systems.

Results

- The return on investment (ROI) for the program is **one million dollars per day**.
- Provides one place for multiple agencies and organizations to get information about large segments of Michigan's citizens and the government efforts to serve them.
- Used to reduce health care fraud, support health care analysis and outreach, and analyze mental health care
- Statewide Homeless Assistance Data online Warehouse (SHADoW) provides de-identified longitudinal database for studying and improving homeless assistance
- Used for rate setting for hospitals and managed and long-term care providers
- Used to determine parent location for child support
- Supports tax audits and validate tax filings
- Used to manage partner performance
- Supports fraud detection
- Used for supporting policy improvements
- Used for epidemiology studies

Cost

The annual operating cost is $8.1M and is recovered via chargeback to agencies served.



Recommended Approach and Feasibility

The methodology used for identifying and assessing the feasibility of optimal approaches for addressing the program objectives included:

- Evaluating each needs-based program objective to identify successful approaches in use by other states, successful approaches in use at a smaller scale in Arkansas, and industry best practice.
- Recommending one or more approaches to address each program objective.
- Evaluating the feasibility of each recommended approach through assessment of successful Arkansas implementations, prototyping with representative tools and data, and studying successful applications in other states.
- Mapping approaches into a cohesive framework to mitigate known barriers to compliant sharing, access to, and analysis of timely, integrated, high quality data between data providers and consumers in Arkansas.

Accessibility

Objective: *Enable openness, transparency, and pervasive, self-service data access and delivery*

According to the Data Management Body of Knowledge (DMBOK), all data has associated costs and risks, but **data has value only when it is actually used** or can be useful in the future. To achieve the maximum value and utility from state data assets and any data warehousing initiative, enablement of widespread access to the data should be a key component. This should include the ability to easily find, understand, search, download, analyze, and retrieve data through self-service user interfaces in both human-readable and machine-readable formats.

Recommended Approach: *Business Intelligence*

All data warehousing approaches put data into technologies that require special programming dialects to access. The earliest decision support systems required programming staff to handle any data requests and no self-service access to data by business users. A key enabler to self-service data access emerged in 1991 with the invention of the "semantic layer", which translates technical data structures into user-friendly business terms and allows for easy, drag-and drop query and analysis without technical knowledge of the data stores or query languages. The semantic layer provides the ease of use and flexibility to answer any question on the fly. Because the data relationships, calculations, and business rules are centrally defined, it also ensures consistency between analysis across users and time.

Self-service access to and analysis of data through the semantic layer are provided through Business Intelligence (BI) tools and platforms. BI platforms are used to query and analyze data retrieved via semantic layers, format the results, make the analysis repeatable and parameter-driven for flexibility, and schedule and distribute the results to the intended audience in a wide variety of formats.

Many users do not even need direct access to BI tools to receive benefits. A common use of BI platforms is to schedule distribution of data and reporting to a wider audience via email, file server, web server, etc. This combination of automation and distribution is used widely to gain operational efficiencies in state business processes.

Publication of data from BI tools can also include public audiences. During the hurricane Katrina relief effort, family members who were put on different buses in New Orleans were evacuated to relief areas in different cities or states. BI tools were used to facilitate overnight development of searchable public-facing reporting of evacuee locations to help reunite families. Public-facing transparency dashboards for the American Reinvestment and Recovery Act (ARRA) including interactive maps of recovery project spending and benefits were also easily provided via BI tools.

As noted in the case study, the Treasury Board of Canada has seen tremendous value from widespread BI access.

While data warehousing has little utility without business intelligence, business intelligence can provide early value without requiring a data warehouse. Arkansas Community Corrections (ACC) leveraged a BI semantic layer over the electronic Offender Management Information System (eOMIS) without a specialized data warehouse schema starting in 2009 and have realized many benefits to operational efficiency and decision support. They have since developed and provided thousands of reports and automatically distribute many reports to the field daily. ACC has recently built an eOMIS data warehouse for increased performance, historical retention, and incorporation of external data sources.

The feasibility of this approach is considered to be high because thousands of Arkansas state and school district employees currently leverage BI tools to access and analyze data from hundreds of state data sources, it is an extremely mature and established technology that has become a basic commodity asset for most organizations. Many Arkansas organizations are already using one or more business intelligence tools, and the prevailing industry trend is to use multiple business intelligence tools as needed for specific business needs over standardizing on a single tool. To maximize access to the largest audience of consumers while minimizing disruption, it is recommended to provide access to a state shared business intelligence platform but also support data access to other state business intelligence tools through secure, compliant database connectivity.

It is recommended that Business Intelligence tools be a key component of the statewide data warehouse strategy to ensure easy, self-service access to and detailed, flexible analysis of state data for all consumers.

Recommended Approach: *Web Portals and Mobile Access*

Not all users or use cases require detailed analysis of data. There is also value and utility in publishing raw structured data sets for consumption. Other uses cases may be better served by purpose-built web or mobile device interfaces to provide specific information via a guided experience and customized interfaces. This is a common starting point for government transparency initiatives.

Arkansas already has financial data sets available for exploration or download via the transparency.arkansas.gov website provided by the Arkansas Department of Finance and Administration. The myschoolinfo.arkansas.gov website provided by the Arkansas Department of Education supports online search, download, comparison, and analysis of public school data. Arkansas also has several successful web applications for use by citizens including the Gov2Go app developed by the Information Network of Arkansas. Given the prevalence and maturity of both open and closed web and mobile access portals, the feasibility of this approach is considered high.

In order to support new and existing web and mobile access to Arkansas data assets, it is recommended that support for web and mobile application access to the statewide data warehouse data be a key consideration.

Recommended Approach: *Application Programming Interfaces*

While Business Intelligence tools and web/mobile access can provide people with easy, self-service access to data, it is also becoming increasingly important to also provide applications with access to data via machine-readable formats to support automated processing. This need is most commonly met through the provisioning, documentation, and support of Application Programming Interfaces (APIs) or reasonably structured, non-proprietary data sets, without licensing restrictions, and easily downloadable in bulk. When applied to open data, these same interfaces could provide benefits to both authorized systems and the general public.

Application Programming Interfaces are widely used for internal information exchange in Arkansas government and are a mature and widespread technology. The Arkansas GIS Office makes Arkansas location data and services available via open APIs, and the City of Little Rock's Open Data portal supports API access for all published data sets.

To serve the broadest range of purposes and users, it is recommended to include support for access to data via Application Programming Interfaces as a component of the statewide data warehouse program.

Sharing

Objective: *Share data to enhance its value while enforcing privacy and security*

Increased data sharing was identified as the top need by the CDO Gap Analysis of Agency Data Needs, and the first recommendation from that study is the **creation of a comprehensive multi-agency data sharing agreement**. This would address issues with the lack of consistency between agreements and the lack of agility in creating or expanding agreements. The Public Safety Interagency Data Exchange Agreement signed in November 2017 was a vital first step to enable a culture of data sharing within the state, but **standardized data sharing agreements will need to be in place across all agencies** in order for a statewide data warehouse program to be effective.

The State of Texas has addressed this challenge via a **statewide data sharing compact** with standard terms in place of ad hoc peer to peer data sharing agreements. The State of Indiana **required all agencies to share data** by Governor's executive order (14-06 of 2014), which was later codified by Act 1470 of 2017. Interagency data sharing is facilitated by the Indiana Management Performance Hub acting as a centralized broker, which is discussed in further detail below.

From an operational perspective, the primary concerns that can impede or prevent more widespread data sharing are related to the need to ensure data security, privacy, and compliance, which necessarily complement and overlap each other. Compliance is ensuring that an organization meets its obligations under applicable state or federal laws, regulations, contractual obligations, and institutional policies. Privacy is related to ensuring appropriate collection, sharing, and use of personal data and personally identifiable information. Both privacy and compliance are typically the responsibility of the Chief Privacy Officer (CPO). Privacy and compliance risks are often barriers to sharing personally identifiable, regulated, or sensitive information.

Security is important for both compliance and privacy. Security is the primary responsibility of the Chief Information Security Officer (CISO) and involves protecting data from impermissible access.

The key challenge lies in establishing methods for sharing personal, sensitive, or regulated data in a manner that complies with all applicable policies but still retains value and utility.

Recommended Approach: *Authentication and Authorization*

Two means for providing secure, compliant access to data include the use of Authentication (verifying a user's identity) and Authorization (granting permissible access based on organization and role) are very common and mature.

The most basic form of authorization is restricting access to entire data sources, reports, dashboards, semantic layers, or collections of content based on a user's organization and role. However, most modern BI and data platforms have robust functionality allowing for more nuanced controls that support restriction to specific rows, columns, or other elements of the same data set. This supports compliant access to the greatest number of users via the smallest number of standardized delivery mechanisms to develop and maintain. Row-level security has been successfully used since 1999 to provide human resource managers at all agencies appropriate access to their agency's sensitive historical human resource data in the Arkansas Human Resources Management System (AHRMS) data warehouse via standard reports used by all agencies while the stewards at the Office of Personnel Management (OPM) have unrestricted access.

Authenticated use also allows for complete auditing of user activity, which is required by some regulations and can provide additional benefits for assessing and driving adoption and use of tools and data.

Recent Business Intelligence development efforts by the Department of Finance and Administration (DFA), such as the Performance Goals and Compensation calibration dashboards used by all agencies, are accessed via the statewide employee self-service portal (EASE) using the same authentication mechanism. Because the user account is tied to the employee record in the Arkansas Administrative Statewide Information System (AASIS), access is revoked automatically when an employee's status, role, or organization changes. This pre-existing authentication source could be effectively

leveraged by the statewide data warehouse for a large percentage of the potential user population. Additional authentication sources would likely be necessary for non-AASIS users or for data access requiring more stringent multi-factor authentication.

Unauthenticated access is an option for public access to data determined to be open and sharable without requirement for authenticated, authorized access.

The feasibility of this approach is considered to be high because it has widespread successful use across many Arkansas agencies and systems.

It is recommended that the statewide data warehouse include support for providing both authenticated, authorized access to restricted or closed data and unauthenticated access to open data.

Recommended Approach: *Compliance Views* **(***Deidentification, Aggregation, Masking)*

While authorization and authentication are established mechanisms for ensuring permissible organization and role-based access, their primary purpose is to restrict the sharing of sensitive data publicly or between organizations. However, there are closed data sets that are beneficial for public, research, or inter-agency use that are unavailable due to security, privacy, or compliance. The most common privacy and compliance barrier is related to the sharing of personally identifiable information.

Much of the value of the closed data than cannot be shared in its raw form is not lost when personal identities are omitted. There are many valuable use cases that require only aggregate statistics and de-identified data.

One option for mitigating sharing barriers is to provide multiple views to the same data for different audiences and purposes:

- **Personal and Identifiable Individual-Level Data View** – Authenticated and authorized access to personal and identifiable individual-level data can be limited to users with an authorized role and purpose.
- **De-Identified Individual-Level Data View** - Some data can be more broadly shared at the individual record level if personal identifiers are sufficiently obscured or removed so that the record cannot be attributed back to the person to which it belongs. These views would require authenticated and authorized access.
- **De-Identified Aggregate Data View** - Some regulations prohibit the disclosure of even deidentified data at the individual record level because data could potentially be re-identified in the case of very small populations. In these cases, data can only be shared as aggregate counts. Some regulations require that aggregate data not be shared below a minimum statistical cell size or n-size in order to avoid re-identification risk. This requirement can be met through the use of suppression (excluding data), perturbation (adding noise), or blurring (reducing the precision of data) as data is accessed through the view depending on the data set being accessed and the allowed minimum cell size. These views could be openly available if sufficiently aggregated.

A similar approach could be applied to other data such as location that might only be shareable if sufficiently obscured to ensure privacy, compliance, and security.

It is not known if this exact approach is currently in use by Arkansas stakeholders, but similar approaches are currently in use to deliver data to users from different tables, views, or data sources in the same report based on user profile. Anonymization and aggregation views have been tested successfully on a representative data platform and are generally considered to be a feasible approach that requires additional testing and validation.

It is recommended that the statewide data warehouse include support for various compliance views in order to increase the compliant sharing of sensitive data with additional authorized or public users.

Analysis

Objective: *Support data-driven decision making and analytic maturity through development and support of analytic skills and shared services*

While not currently available to every consumer or for every relevant data source, access to historical Arkansas data via business intelligence outputs such as reports, dashboards, scorecards, and ad hoc query and analysis are currently available to over 10K state and school employees. These tools and techniques are categorized as Descriptive Analytics because they describe what happened in the past, or in the case of real time access, what is currently happening.

The Accessibility recommendations in this study focus on increasing the availability of access to data consumers, but there are also important considerations for increasing the maturity of analytic capabilities available to existing and future data consumers. Increasing awareness of, skills in, and technical capabilities for more mature types of analytics can enable more efficient and effective decisions and processes by reducing time-to-decision, recommending optimal actions and interventions, predicting future events or performance (which can support early intervention), and automating some decisions or processes where appropriate (with human oversight).

Three broad categories of business analytics or advanced analytics beyond traditional business intelligence include:

- Diagnostic Analytics – Used for diagnosing root causes and drivers of events based on relationship and patterns in the data. While descriptive analytics tools can support some of the same capabilities for known questions, tools and techniques for diagnostic analytics tend to offer more support for automated detection of unknown relationships, anomalies, or patterns without always knowing the question in advance.
- Predictive Analytics – Leverages predictive modeling and machine learning techniques to predict likely future outcomes based on patterns and relationships in current and historical data.
- Prescriptive Analytics – Extends predictive analytics by recommending the decisions and actions that are most likely to lead to the desired outcome based on current and historical data.

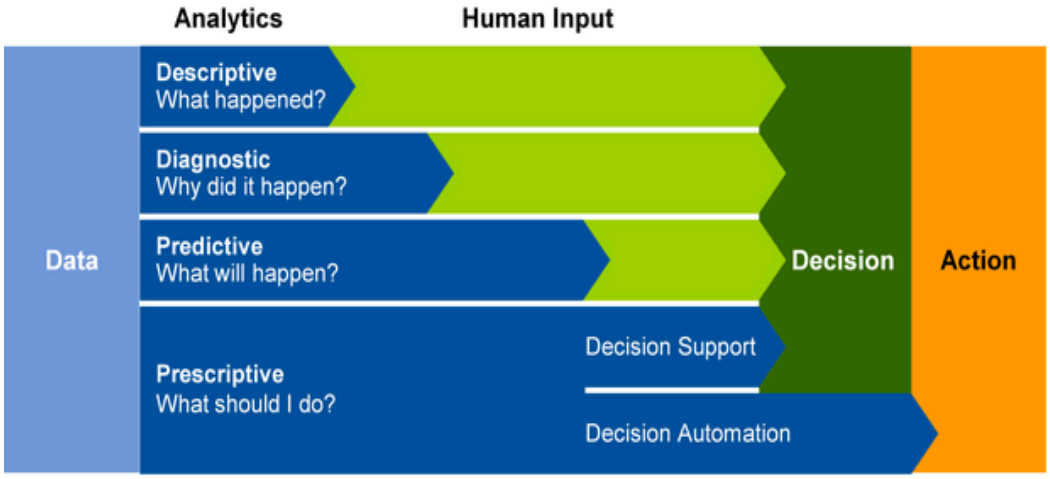


Figure 8 - Analytic Maturity (Source: Gartner)

While these are the three most commonly known and addressed categories of more advanced analytics, it is a broad and rapidly evolving field with new tools and approaches constantly emerging. Some other prevalent types of analytics include spatial (location-oriented), time series (patterns and forecasts over time), graph (focuses on relationships), text (sentiment and other derived meaning from unstructured text), and process mining (mapping and optimizing flow of events). This is not an exhaustive list and merely seeks to illustrate the need for versatile analytics capabilities to address evolving needs and approaches.

Recommended Approach: *Analytics Tools (Predictive/Prescriptive, Machine Learning, Spatial, Data Science Automation)*

As noted above, analytics technologies, approaches, and requirements evolve constantly and rapidly. An exhaustive recommendation for analytics capabilities is beyond the scope of this study and would become quickly dated. It is instead recommended to include in the data warehouse approach certain key, established capabilities that have either broad adoption and versatility or known applicability to Arkansas use cases. These key, foundational analytics capabilities include:

- **Predictive and Prescriptive Analytics** – The capability to predict outcomes and prescribe effective actions are broadly applicable to strategic and operational needs and can improve the efficiency and effectiveness of many decisions, processes, and services.
- **Machine Learning** – There are a core set of mature, foundational machine learning algorithms that can be applied and combined to perform a wide variety of analytic tasks. These should include Decision Trees, Support Vector Machines, Nearest Neighbor, Linear Regression, Logistic Regression, Neural Networks, K-Means, Association Rules, Random Forest, and Naïve Bayes.
- **Spatial Analytics** – Location is one of the few common elements and intersection points across many data sets and enables incorporation of valuable location-based external data such as that from the Census Bureau and the Bureau of Labor and Statistics.
- **Data Science Automation** - Up to 80% of the data science process can be spent on data preparation, feature evaluation and selection, preparation of training and testing sets, and model evaluation. To maximize the productivity of a limited number of data science resources, the capability should exist for at least semi-automated support for model preparation and evaluation tasks as well as that for ongoing model performance evaluation and retraining after model deployment. Deployed models should not require ongoing human interaction for operational use but should support automated scheduling or dynamic use.

Because needs and capabilities expand and evolve, the data warehouse approach should include the flexibility to easily leverage additional capabilities and techniques as needed to address emerging and evolving needs such as incorporation of additional libraries, functions, and services including, but not limited to, R (statistical programming language), Python (general open source programming language with), and TensorFlow (open source machine learning library).

Availability of analytics tools alone will provide little value unless paired with availability and development of analytics skills. The Indiana Management Performance Hub and other programs pair centralized data science resources with evangelization and training efforts to raise analytic maturity statewide. The State of Tennessee provides a shared pool of four data scientists with different specialties for use on agency and inter-agency analytics efforts. It is recommended to include at least one data scientist on the data warehouse program staff to support analytics use cases.

There are a few successful examples of the use of advanced analytics in Arkansas state government, but adoption is not widespread. Predictive and prescriptive analytics were developed and tested successfully during a proof of concept (POC) project with the Arkansas Department of Corrections, Arkansas Department of Community Corrections, Arkansas Parole Board, and Arkansas Crime Information Center to leverage data and analytics to reduce recidivism. A predictive recidivism model was developed in less than 90 days with good accuracy compared to current instruments and peer models in other states. The project also produced a prescriptive model that recommends optimal interventions for successful community re-entry and made extensive use of spatial analytics that brought new insights from existing data. Based on the results of this effort, success in other states, and the success of smaller-scale efforts in Arkansas, this approach is considered to be feasible. While difficult to implement, analytics have the potential for tremendous value.

It is recommended that the statewide data warehouse include staffing and support for varied and extendable analytics capabilities including at a minimum support for predictive and prescriptive analytics, common machine learning algorithms, and spatial analytics.

Integration

Objective: *Integrate data for improved cross-agency analysis and reduced duplication of data and efforts*

Once data sharing concerns are addressed through controls on the privacy, security, and compliance of data access, interagency data sets must still be integrated with each other to gain a broader perspective than is possible using any individual system alone.

The two biggest challenges to data integration are:

- The lack of semantic consistency (using the same names and identifiers for the same people, places, events, things, etc.) increases the difficulty of matching data across systems.
- Security, privacy, and compliance concerns when personally identifiable or sensitive information must be disclosed by one or more parties in order to establish the required data relationship, even if the ultimate data product will only be disclosed in deidentified or aggregate form.

Recommended Approach: *Statewide Master Data Management*

Integrating data from multiple systems together is frequently challenging due to a lack of standard, consistent identifiers for the same entity (person, place, location, thing) across systems. Data regarding organizations frequently suffer from lack of a standard identifier. This can result in many records for the same organization in even the same source and impede analysis such as spend on a single vendor or payments to a single organization.

One effective method of addressing this issue is establishment and use of common identifiers across systems. In some cases, multiple agencies are served by the same information system such as the Arkansas Administrative Statewide Information System (AASIS), allowing for centralized control over the consistency of names and identifiers. When multiple information systems are used to describe the same entity, there must be more planning and coordination to establish and maintain consistent naming and globally unique identification of entities.

For entities with a limited, relatively static range of values, the decision can be made for multiple systems to use the same authoritative source for encoding data. One common example used in many Arkansas information systems is the Federal Information Processing System (FIPS) code for US states. One benefit of this approach is that national standards may be used by other states, facilitating the possibility of interstate integration and comparison of data or required for Federal reporting.

One successful example of standards-based interagency data integration and reporting in Arkansas is the effort to comply with Federal reporting requirements for the American Recovery and Reinvestment Act of 2009 (ARRA). The act resulted in the award of 218 Federal grants with award amounts totaling over $3.7B to Arkansas government organizations, but it also made continued funding contingent upon compliance with quarterly Federal reporting requirements. A key aspect of these requirements was that very little submission of free form text was allowed. Reporting of grant programs, recipient organizations, vendors, grant activity types, and locations were all supported only through specified coding standards such as FIPS codes for counties and state, DUNS numbers for organizations, or NAICS codes for industries. Upon report submission, the validity of codes was checked, and invalid or conflicting codes resulted in a failed submission. Arkansas was a centralized reporting state, meaning that the Department of Finance and Administration Office of Intergovernmental Services (IGS) was responsible for all Federal reporting to the White House Office of Management and Budget for all state agencies. By coordinating on standards for encoding and reporting data, quarterly reports for 28 state agencies using 3 different information systems (plus AASIS) were successfully integrated and submitted without issue over a 4-year span. This semantic consistency also allowed for consistent national aggregation and reporting and was beneficial for supporting Arkansas' own ARRA transparency portal.

Common identifiers can enable data from across sources to be effectively integrated but doesn't help with duplication of data and effort when the same data is being created and maintained in multiple systems. It also introduces the challenge of maintaining and sharing a consistent list of valid identifiers across all sources when dealing with large, dynamic lists such as people, organizations, and addresses.

The commonly used and recommended approach for addressing the coordination of consistent identifiers across systems is Master Data Management (MDM). Master data management is a technology-enabled discipline comprised of the processes, governance, and tools required to create, maintain, integrate, monitor, and share master and reference data. Master data are the key business entities or "nouns" of an organization that are widely used across departments, processes, and systems (people and roles, organizations, locations). Reference data are a set of standard permissible values that are not specific to the enterprise but are used by multiple systems to facilitate data sharing (industry codes).

The technology component of master data management is a master data "hub" that is used to coordinate the exchange or "harmonization" of codes and values between subscribing systems. If one system can be determined to be the authoritative system of record, its values can be made available for reference from, or pushed to, subscribing systems that require the same data, reducing duplicative efforts to maintain the same data. When multiple systems need to maintain different, potentially overlapping lists of the same type of entity (location), MDM can be used to synchronize all subscribing systems into a single master list with a combined "golden record" consisting of the most complete and timely information from the combination of all systems and establishing a common unique identifier.

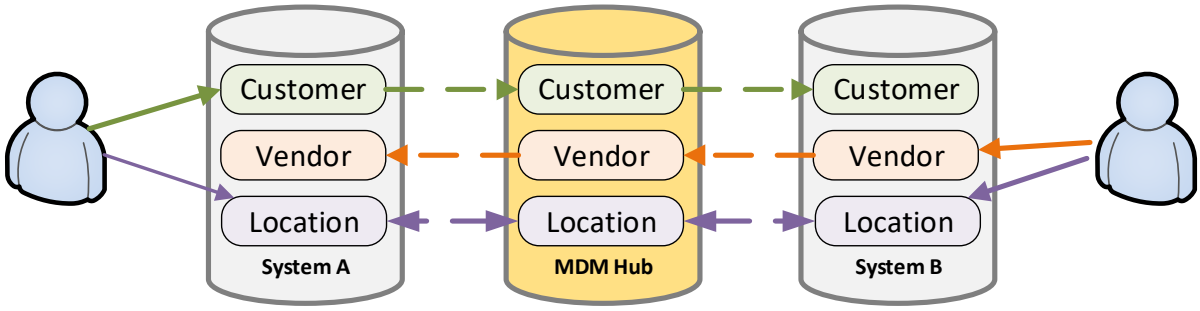


Figure 9 - Master Data Management

The key benefits of this approach include:

- Each record is only created and maintained in one place and reused globally, reducing duplicate effort
- All systems use the same codes and names, making reporting, integration, and reconciliation easier
- Consolidated "golden records" can be more complete, timely, and accurate than in any individual system
- Provides a single location for stewardship of master data, typically with support for data quality rules
- Increases operational efficiency through easier, more comprehensive access to data
- Minimizes the impact of changes by serving as an intermediate broker between systems
- Serves as a key enabler for logical data warehousing by storing the relationships between systems and entities

Another key component of MDM is entity resolution, which is the process of linking records for the same entity.

The Arkansas Department of Information Systems and Arkansas Department of Parks and Tourism both have successful implementations of MDM systems providing valuable business benefits. Due to the maturity and prevalence of this technology, existing successes in Arkansas, specific requirements to provide MDM in Act 912 of 2017, and additional recommendations from the CDO Gap Analysis, the feasibility and value of this approach is considered to be very high.

It is recommended that the statewide data warehouse include capabilities for statewide master data management to support the coordination of consistent identifiers across agency systems.

Recommended Approach: *Centralized Broker*

Master data management can help address the problem of semantic consistency, but implementing statewide master data management, particularly for citizen data, requires access to personally identifiable information from multiple agencies in order to perform the matching process required to maintain a unique list of records with a consistent identifier. Similarly, the integration of data related to various citizen services requires access to private and sensitive data from multiple agencies, even if the data distributed to consumers is deidentified and aggregated enough to be considered open. Agencies can share their own deidentified, aggregate data, but for interagency data integration efforts, the requirement for one party to have access to two or more sets of personally identifiable information has been a frequent and persistent barrier.

The State of Indiana addressed this issue by establishing in the Indiana Management Performance Hub (MPH) a group which can serve as a trusted broker between and on the behalf of all agencies. MPH acts as an agent of the agencies they serve to integrate, analyze, and distribute data on behalf of and subject to governance defined by the providing agencies. MPH was established by Governor's executive order (14-06 of 2014) and later codified by Act 1470 of 2017:

> *Sec. 13. The MPH is considered to be an agent of the executive state agency sharing government information and is an authorized receiver of government information under the statutory or administrative law that governs the government information. Interagency data sharing under this chapter does not constitute a disclosure or release under any statutory or administrative law that governs the government information.*
> http://iga.in.gov/legislative/2017/bills/house/1470#document-af23f3bf

This successful approach could be applied in Arkansas by leveraging a centralized group to act as agents on behalf of providing agencies in integrating and sharing data in compliance with established governance policies. The data should be stored in a secure zone in the state data center that meets the physical, network, multi-factor authentication, and other requirements of the most sensitive data hosted. Support staff should be required to have the same background checks, training, and certifications (CJIS, HIPAA, etc.) as the agencies they represent when integrating, analyzing, and sharing data on their behalf. The security, governance, architecture, and operating procedures for the data integration layer should be designed and overseen by the state Chief Data Officer, Chief Privacy Officer, and Chief Information Security Officer to ensure stringent security, privacy, and compliance. The group should be located within the Department of Information Systems along with the CDO, CPO, and CISO.

The feasibility of this approach is demonstrated by success using this approach in the State of Indiana.

It is recommended that the integration layer of the statewide data be managed by a centralized group housed at the Department of Information Systems and acting as agents of providing agencies.

Recommended Approach: *Data Integration Layer*

Data should be integrated using current best practices for balancing timeliness, performance, agility, and ease of use. In order to meet the widest array of operational and analytical data integration needs, the data integration layer should:

- Include support for a data warehouse as needed for performance or historical retention
- Include support for a data lake for agility and exploratory analysis
- Use a data hub approach to logically integrate data via virtualized integration views over master data hubs, data warehouse, data lake, and federated local data sources

This pattern has been used successfully at DIS and is considered to be a feasible option for statewide integration.

It is recommended that the statewide data warehouse leverage the data hub approach by logically integrating data from a centralized data warehouse, data lake, master data hub, and local data sources.

Timeliness

Objective: *Enable real-time assessment, support, alignment, and automation of decisions, programs, and resources*

Some of the initial barriers that led to the development of separate reporting databases and data warehouses were:

- Source systems were designed for efficiently processing individual transactions (inserts, updates, deletes) and not for mass query and analysis. They were frequently either too slow for efficient analysis or using them for analysis was detrimental to the availability and performance of the transaction system.
- Early business intelligence tools typically only supported connection to a single data source, but decision support and analysis needs required the combination of data from multiple data sources.

These factors, along with the need to retain historical data, required a second copy of the source data to be maintained that could be optimized for analysis, integrated into a single consolidated data store, and isolated from impacting the source systems. Since the batch extract, transform, and load (ETL) technology available at the time was also detrimental to the performance of the source systems, reporting databases or data warehouses were typically loaded overnight when they would have the least impact to business users.

This approach enabled for consolidated, performant data access that met many business needs, but resulted in some decision latency due to the nightly, weekly, or monthly batch processes required. It also resulted in a redundant volume of data to store, maintain, secure, and keep consistent with the source.

The value of some data decreases with time, and some business processes such as operational support benefit greatly from applying decision support and analytics to real-time data. Advancements in the technologies and approaches for hosting source system data and connecting source data with consumers have now matured to the point that real-time access is a feasible and integral component of a modern data and analytics infrastructure.

Recommended Approach: *Federation*

The primary approach used for enabling real-time access to distributed data sources is Data Federation. As previously defined in the logical data warehouse section, Data Federation is a form of data virtualization where the data stored in multiple sources (of the same or different type) are made accessible to data consumers by using on-demand data integration, rather than executing data movement and physically storing integrated data. This is essentially a reversal of the extract, transform, and load process which distributes queries to the data and applies necessary integration and transformation on demand as opposed to moving and transforming the data to a separate data store in a batch process.

In order to minimize impact on the source system, several approaches are used by the federation layer to avoid sending computationally expensive queries to transactional data source. Leveraging advances in database and hardware technology, some source systems are also becoming robust enough to handle greater analytic workloads and are even being designed for hybrid transactional and analytical use.

The Arkansas Longitudinal Data System developed by the Arkansas Department of Education is a successful example of batch federation for analysis of integrated inter-agency data sets. The Arkansas Department of Information Systems successfully uses federation for providing distributed real-time access to operational data. Because this is a mature, common technology, the feasibility of this approach is considered to be high.

It is recommended that the statewide data warehouse include support for federated, real-time access to disparate data sources when appropriate.

Quality

Objective: *Manage enterprise data as a state asset*

In the 90s, data quality was often corrected into the data downstream in the data warehouse, which is the equivalent of inspecting quality into the product in manufacturing by checking for and discarding or repairing defective products. Ensuring quality data products, much like manufacturing, is more efficiently and effectively done as close to the source of creation as possible. Only the data creator, at the moment of data creation, truly knows what real-world person, item, or event they are trying to represent.

The most effective place to address data quality is at the source and through business rules and restricting entry to valid domain at point of entry. This is not always practical for some systems but is a best practice that can be adopted through attrition over time. A less invasive and more immediate way to start measuring and managing data at the local agency-level is through the use of data quality management tools for profiling, measuring, and monitoring the quality of local data assets.

When age is first profiled in many Arkansas data sources, it typically results in individuals who are negative or hundreds of years old. Self-service data quality management tools can enable data stewards to profile the data to identify and measure any issues with completeness, obvious inaccuracies (people with negative ages), problems with consistency of representation, etc. These stewards can then establish data quality rules and dashboards for measured improvements on data quality problems and potentially identify controls to mitigate future issues at the source.

Data quality management and stewardship portals are also frequently the source for collecting and sharing metadata with consumers and establishing data lineage that can be used to assess the impact of potential data change.

While data quality management tools can help identify and monitor local data quality issues to speed and ensure correction, a more proactive approach is the use of agency-level master data management. The benefits of master data management have been discussed in the integration section in relation to statewide master data management. However, there are many agency-specific data elements that may never need to be included in the statewide master data hub. Local, agency-level master data management can yield substantial gains in improving the quality of agency data, increasing the ease of access to key agency data.

Finally, Data Governance is needed over all aspects of the data warehouse program, including a comprehensive data catalog and shared metadata repository.

It is recommended that the statewide data warehouse include support for local data quality management tools and stewardship portals, agency-level master data management, and comprehensive data governance.

Summary of Recommendations

The following summarizes how the identified program objectives map to each group or layer of recommended approaches:

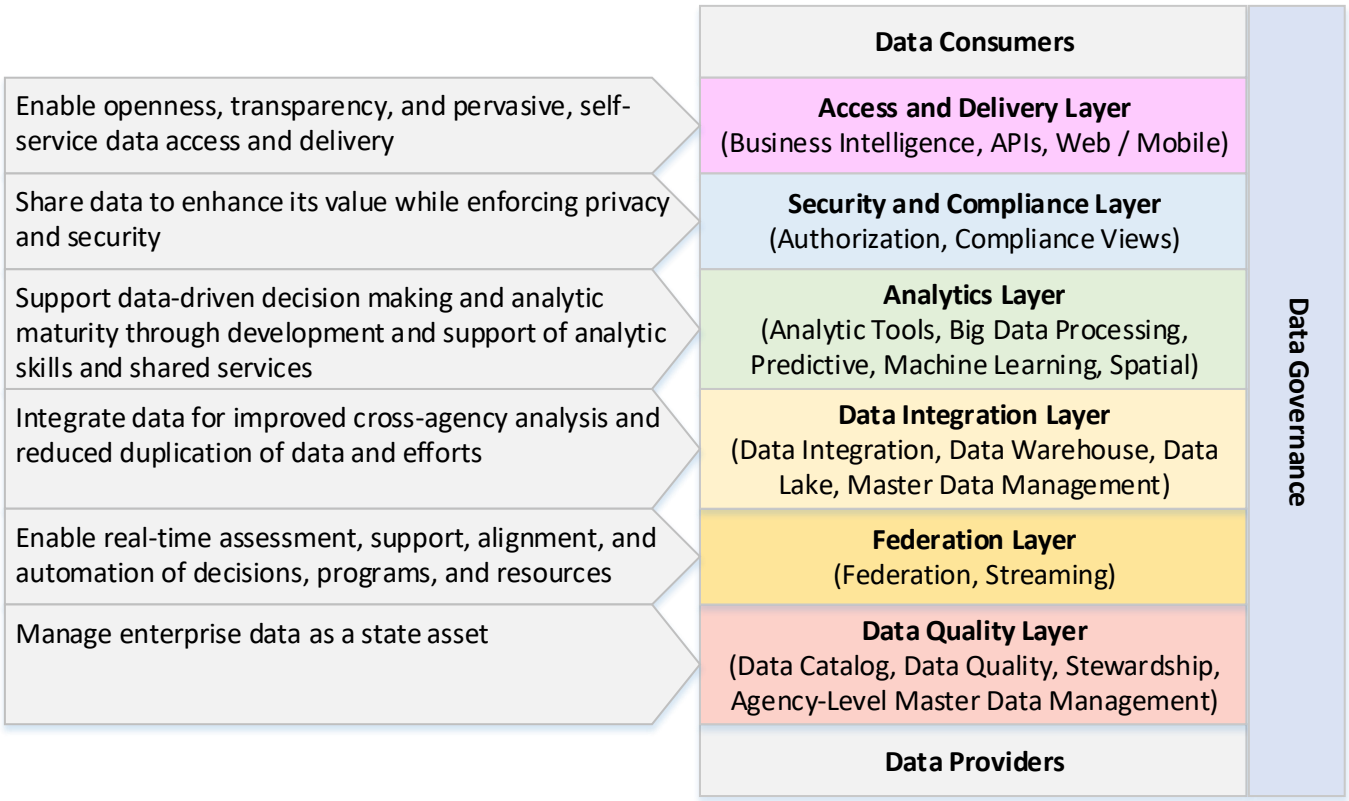


Figure 10 - Program Objectives Mapped to Recommended Approaches

Some dependencies and rationale considered when ordering the layers from provider to consumer include:

- Widespread access via Business Intelligence tools, database connectivity, Application Programming Interfaces, mobile and web applications, etc. should be on top of the security and compliance layer to allow for decentralized development and delivery via a wide variety of tools but leveraging centralized management of data access in compliance with security, privacy, regulatory, and other policies and applicable laws.
- Analytics should have the capability to be performed on identified and integrated individual-level data if needed and permissible and then presented with applicable anonymization, aggregation, filtering. Therefore, the analytics layer should be located between the data integration layer and the security and compliance layer.
- Integration should have the capability to be performed dynamically across both federated or streaming real-time data and persisted data from the centralized master data management hub, data warehouse, and data lake. Therefore, the data integration layer should be located between the analytics and federation layers.
- Data quality should be managed as close to the source as possible and with self-service access by the actual agency-level data stewards. The data quality layer should be provided directly to data providers with the capability of being used with local data sources.
- Data governance should encompass all aspects of the statewide data warehouse. This includes the collection, integration, and provisioning of metadata from providers to consumers as well as centralized change management for shared data and services.

The Recommended Arkansas Data Hub (Statewide Data Warehouse) Approach

The following diagram illustrates in more detail how the recommended approaches detailed in this study fit together to comprise a holistic approach to delivering upon the identified needs. The full "data hub" architecture expands on the traditional data warehouse to enable compliant, governed data sharing for both operational and analytical needs.

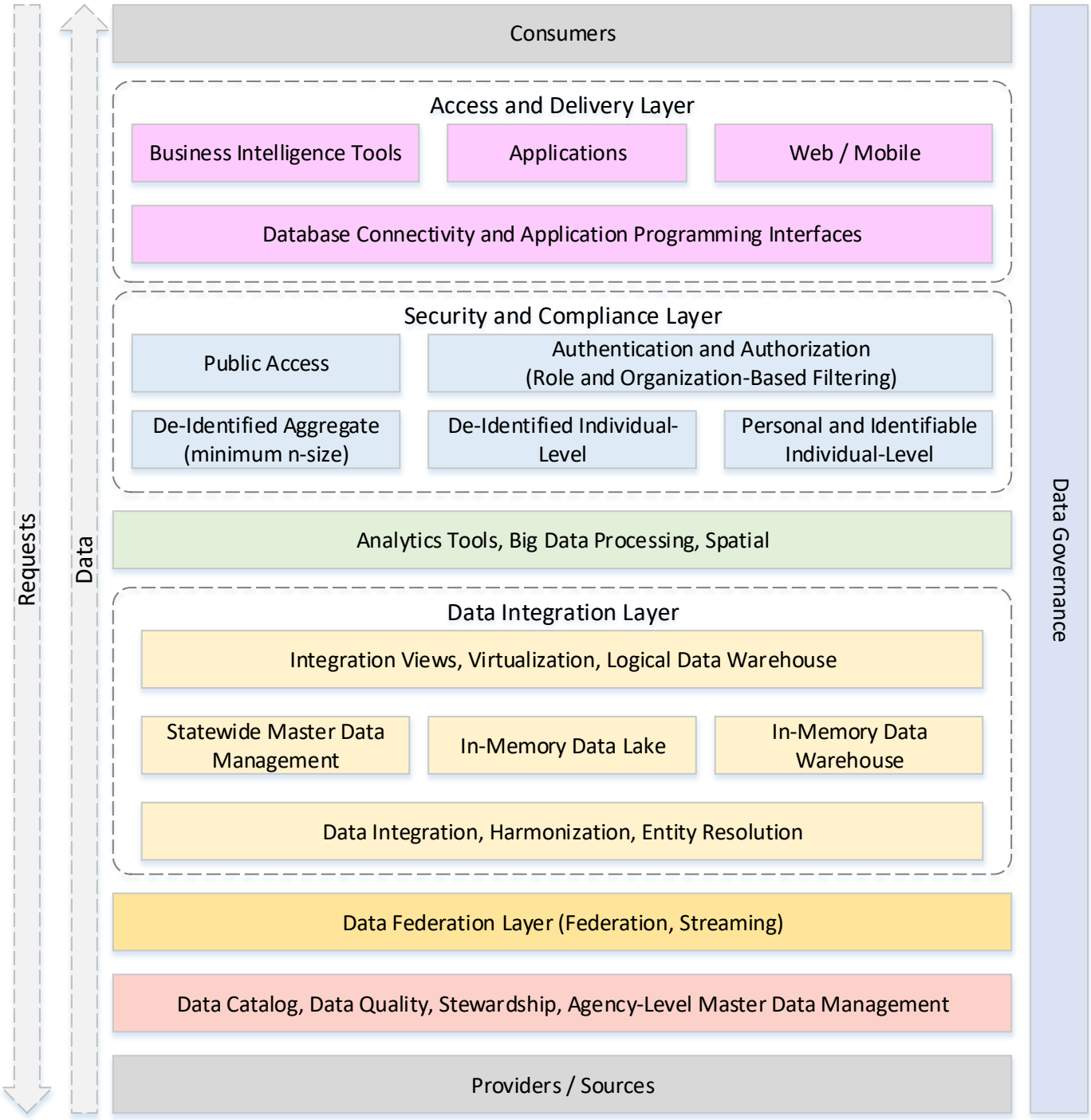


Figure 11 - Recommended Arkansas Data Hub Approach

Costs and Benefits

Costs and Funding Models

Costs

The annual costs for establishing and supporting a statewide data warehouse program are estimated at **$3.9M per year for years 1-5 and $2.8M per year for years 6 and beyond**. The lower annual cost after year 6 is due to the capital expenditures for software purchases being fully amortized (assumes a 5-year useful life). Hardware cost is also based on amortization over a 5-year useful life but assumes replacement at end of useful life.

These costs were based on conservative estimates for scaling established state data platforms, staffing the program based on the current Arkansas classification and compensation plan, and adding recommended additional capabilities and capacity based on costs for representative interoperable technologies.

Software costs were estimated based on purchase and annual software support costs for representative software platforms and products to address the recommended approaches to be included in the program scope. Hardware costs were based on platform sizing equivalent to that used by the Indiana Management Performance Hub.

Staffing costs were estimated based on full or partial resource allocations for program staff including the CDO, CPO, Data Warehouse Lead, Database Administration Lead, Data Scientist, Data Engineers, Data Warehouse Specialists, Platform System Administrator, Database Administrator and Project Manager.

Total annual costs for **other statewide data warehouse programs studied ranged from $8M to $9M per year**. The Indiana Management Performance Hub is the closest analogue to the proposed program. It has similar hardware and software sizing, but a larger staff size (20 full time employees plus contract staff for data science).

Funding Models

Other statewide programs use **a combination of funding strategies** including federal and private **grants, general revenue** appropriations, and usage-based **chargeback**. The Indiana Management Performance Hub's $9M annual cost is covered by a $6M appropriation of general funds, $1.7M in database management funds, and $1.3M from the Department of Insurance fund. The $8M annual operating cost for Michigan's statewide data warehouse is recovery by chargeback to agencies served.

There are Federal grants for inter-agency data efforts being leveraged by other states to start and support statewide data warehouse efforts. Relevant grant funding opportunities should be monitored and leveraged as appropriate for helping to start Arkansas inter-agency data sharing and analysis efforts.

Benefits

The financial benefits of this initiative are expected to exceed the costs. Some examples of financial impact of enhanced data sharing and integration from other states include:

- The 2018 Annual Report published by the Indiana Management Performance Hub estimated a **$40M return on investment over 18 months** ($4.50 return for every $1 invested taxpayer dollar) as quantified by serviced agency project owners.
- The State of Michigan is **saving $1M per day** by linking data across programs in 21 different agencies and extending access to 10K users.
- The Texas Workforce Commission implemented an early detection program for unemployment insurance fraud leveraging interagency data integration and **avoided $71.84M in costs** in FY15-FY17.
- Additional financial ROI examples are cited in the 2015 Arkansas Legislative Audit (ALA) Special Report on the "Potential Benefits of a Centralized Data Warehouse for the State of Arkansas" including:
    - The State of Washington recoups more than **$10M per year** in fraudulent tax refunds.
    - The State of Georgia detected **$25M over 2 years** in fraudulent tax returns.
    - The State of New York **increased collections by $100M** and **reduced fraudulent refunds by $1.2B** in 2010.

An Arkansas proof of concept project exploring the use of advanced analytics to reduce the risk of recidivism resulted in a prototype with a projected cost savings of **$8.1M over 3 years** with a 1% reduction in recidivism.

Beyond financial benefits, there are many notable examples of how integrating, sharing, and analyzing state data can yield improvements to citizen safety, health, and quality of life. The Indiana MPH 2018 Annual Report includes details on MPH projects that improved policy and operational support in combatting the Opioid epidemic, optimizing Medicaid, increasing government transparency, projecting recidivism risk, enabling agency performance management, and aligning education and workforce development.

Next Steps

The recommended next steps towards implementation of a statewide data warehouse program include the following:

- Formalize a **multi-department data sharing agreement** for compliant, secure and efficient data sharing between departments
- Determination of initial and sustaining program **funding approaches**
    - Identify and leverage **potential grants** for program startup
- Develop a **program charter** to formalize the scope of the program
    - **Discussion and agreement** on data warehouse program components and architectural approaches
    - Determine **prioritization mechanism** for implementing and addressing use cases
- Implementation of a **data hub** (value-driven broker for cross-agency data sharing and analytics) to act as an agent of individual agencies in integrating and providing secure, compliant access to and analysis of inter-agency data