Special Report

Arkansas Legislative Audit

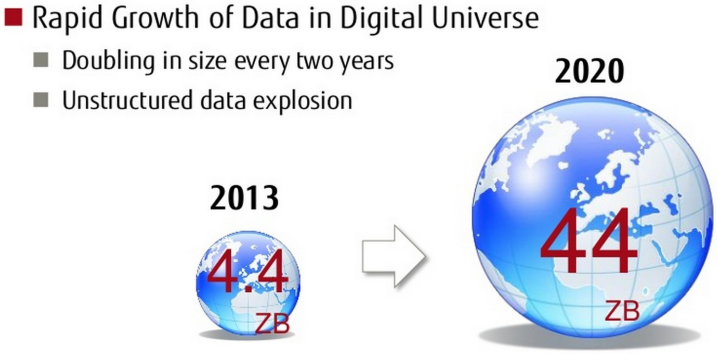
Potential Benefits of a Centralized Data Warehouse for the State of Arkansas

November 6, 2015

INTRODUCTION

This report is issued to inform the Legislative Joint Auditing Committee (LJAC) of potential benefits of centralized data warehousing and to recommend steps that can be taken to introduce centralized data warehousing to Arkansas state government. Separate from the centralized *processing* of data (as achieved in the Arkansas Administrative Statewide Information System [AASIS]), **centralized data warehousing** is the collection, storage, and streamlining of data in a single repository. A centralized data warehouse allows state and local entities to share data as appropriate and authorized and allows data to be more easily accessed and safeguarded, as well as more efficiently distributed to those making government policy decisions.

As technology has evolved from "pen and pad" to iPad®, the amount of data produced worldwide has grown and continues to grow at a staggering rate (see image at right). This massive amount of data that comes from a variety of sources and is too large and complex to be efficiently stored, managed, or utilized using conventional means is known as **big data**.



ZB = Zettabyte. See **Exhibit I on page 2** for definition.

Source: *EMC Digital Universe with Research and Analysis by International Data Corporation (IDC) 2014 (http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm)*

Big data obtained from mobile and web technologies have long been used by businesses to follow consumer habits and enhance marketing efforts. For example, when individuals search for a product online and an ad for that product then appears on their social media page, they have seen big data analytics at work. **Big data analytics** is the process of examining big data to uncover trends, connections, and other useful information. Despite such use in the realm of business for quite some time, state governments have only recently begun harnessing the potential of the data within their own systems. As stated in the September 2014 issue of *State Legislatures Magazine*,

> *Although the term 'big data' sounds vaguely sinister – like a relative of Big Brother or Big Government – it is an unfair rap. At least in the case of state governments, it is being used to increase public safety, uncover fraud, save money, create efficiencies, and improve health and human services, among other things.*

In large measure, centralized data warehousing can make finding the proverbial needle in the haystack of big data both possible and practical, creating valuable information assets for state government.

ARKANSAS LEGISLATIVE AUDIT
500 Woodlane Street, Suite 172, Little Rock, AR 72201
Phone: 501-683-8600 • Fax: 501-683-8605
www.arklegaudit.gov

Report ID: SPSA01215 Report Date: October 20, 2015

Exhibit I: Understanding Data Size

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit," after the binary code (1 or 0) that computers use to store and process data.
Byte (B)	8 bits	Enough information to create an English letter or number in computer code; the basic unit of computing.
Kilobyte (KB)	1,000 or 2^{10} bytes	From "thousand" in Greek. One page of typed text is 2KB.
Megabyte (MP)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB.
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB.
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB.
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour.
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*.
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence in 2010 was forecast to be around 1.2ZB.
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine.

Note: The prefixes are set by an intergovernmental group, the International Bureau of weights and measures. Zetta and Yotta were added in 1991; terms for larger amounts have yet to be established.

Source: *Adapted from* The Economist*, February 27, 2010*

OBJECTIVES

The objectives of this report are to:

- Provide background information regarding the State's current information technology (IT) infrastructure related to data sharing.
- Explore benefits the State might recognize from the creation of a centralized data warehouse.
- Outline the steps needed to implement an effective centralized data warehouse for Arkansas state government.

SCOPE AND METHODOLOGY

Arkansas Legislative Audit (ALA) staff reviewed relevant reports and records and interviewed Information Officers of selected states about their experiences with centralized data warehousing. ALA staff also administered surveys at 20 of Arkansas's large state agencies to gather information regarding their IT needs and costs related to data storage, data security, and application development. In addition, interviews were conducted with personnel at these agencies and at the Arkansas Department of Information Systems (DIS).

The methodology used in preparing this report was developed uniquely to address the stated objectives; therefore, this report is more limited in scope than an audit or attestation engagement performed in accordance with *Government Auditing Standards* issued by the Comptroller General of the United States.

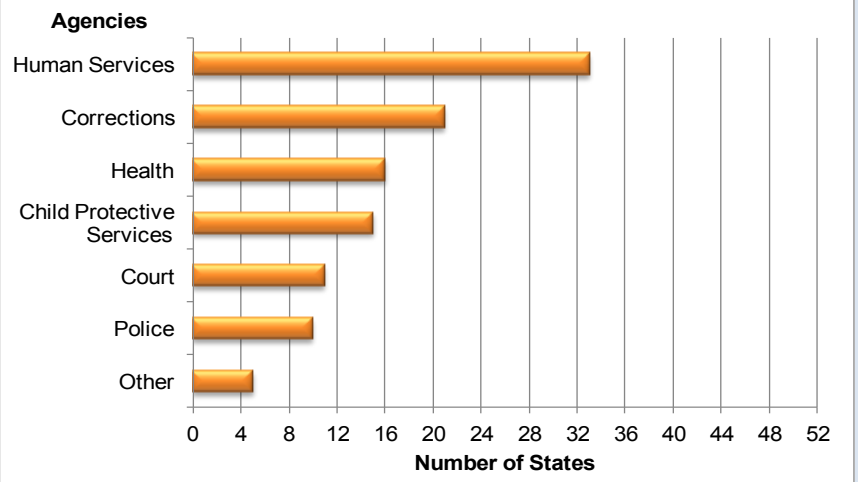
BACKGROUND

The State of Arkansas currently has a decentralized IT infrastructure in which state and local government entities maintain individualized IT systems that use a variety of formats and computing platforms. As a result, data collected and stored by each entity are not accessible to other entities or standardized into a common format. The State's current IT infrastructure has created multiple "silos" of information across and even within state and local government entities, making data-gathering, data analytics, and coordination of services a significant challenge.

Arkansas is not the only state facing this challenge. As shown in **Exhibit II**, in a 2010 survey of all 50 states, the District of Columbia, and Puerto Rico, a majority reported that sharing of students' educational data with other state agencies occurs infrequently; if sharing does occur, it is most likely with human services agencies. The states also reported that data sharing most often occurs in one direction, from other agencies to education, and rarely vice versa.

Exhibit II: Nationwide Sharing of Student-Level Data Among State Agencies

Number of state education agencies in all 50 states, the District of Columbia, and Puerto Rico that can link data between K-12 and other agencies.



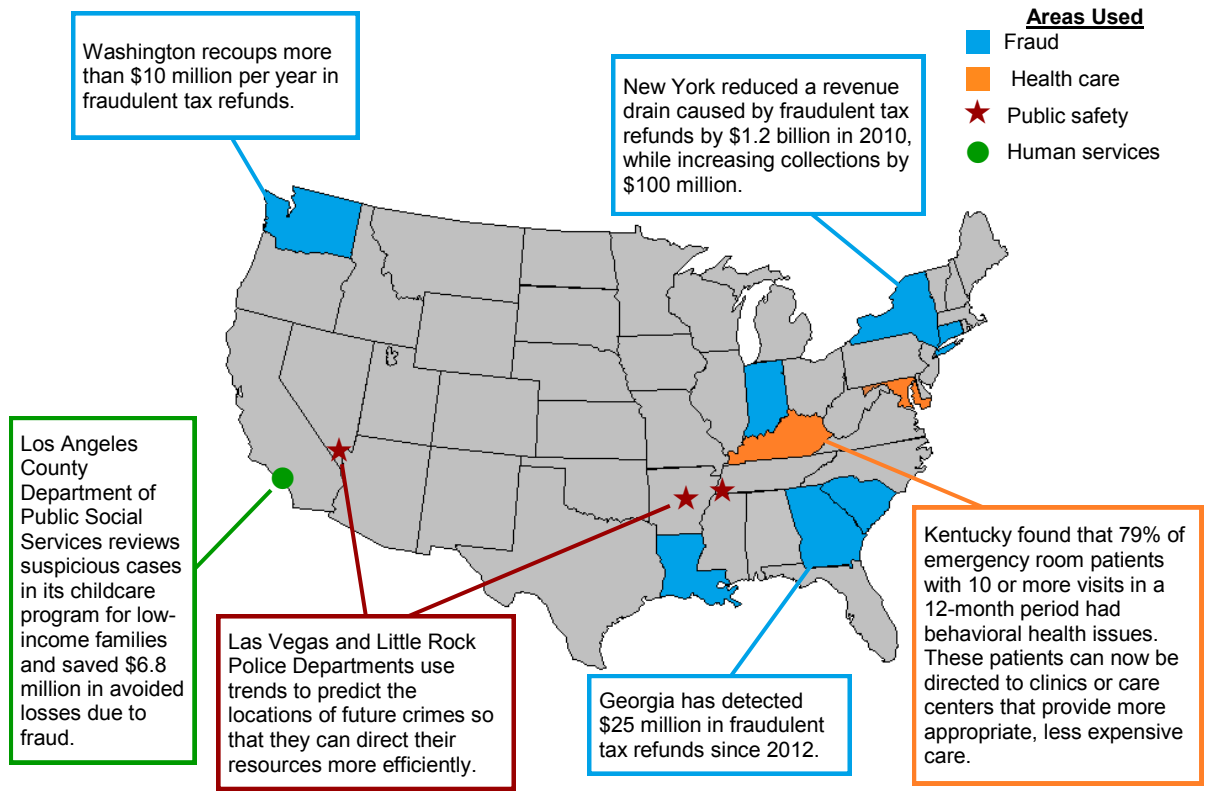
Source: *"Linking Data across Agencies: States that are Making it Work," March 2010, Data Quality Campaign and Forum for Youth Investment (http://forumfyi.org/files/States.That.Are.Making.It.Work.pdf)*

According to a survey published in September 2014 by the National Association of State Chief Information Officers (NASCIO), 41.2% of states were still investigating opportunities for big data, and 21.6% described the status of big data in their states as "no activity at this time."

However, some states, counties, and municipalities are already reaping benefits of using big data in innovative ways. New York, Indiana, Connecticut, Georgia, Louisiana, and South Carolina are using data analytics to uncover tax fraud, saving or recouping millions of dollars for their states, while Maryland and Kentucky have used data analytics to conserve state dollars while improving health care. Notably, big data analytics have also yielded reduction of fraud in the Los Angeles County childcare program for low-income families and public safety improvements in metropolitan areas like Las Vegas, New York City, and Memphis (see **Exhibit III**). In fact, the Little Rock Police Department has also used historical data to predict where crimes are likely to occur based on data trends.

Additional benefits could potentially be realized by enhancing or replacing existing programs being used to detect fraud by state agencies through the creation of a centralized data warehouse.

Exhibit III: States' Use of "Big Data"



Source: *"Big Data = Big Benefits,"* State Legislatures Magazine, *September 2014*

POTENTIAL BENEFITS OF A CENTRALIZED DATA WAREHOUSE

A data warehouse offers a variety of benefits for a range of audiences, from the general public to the General Assembly, including the following, which are discussed in the sections that follow:

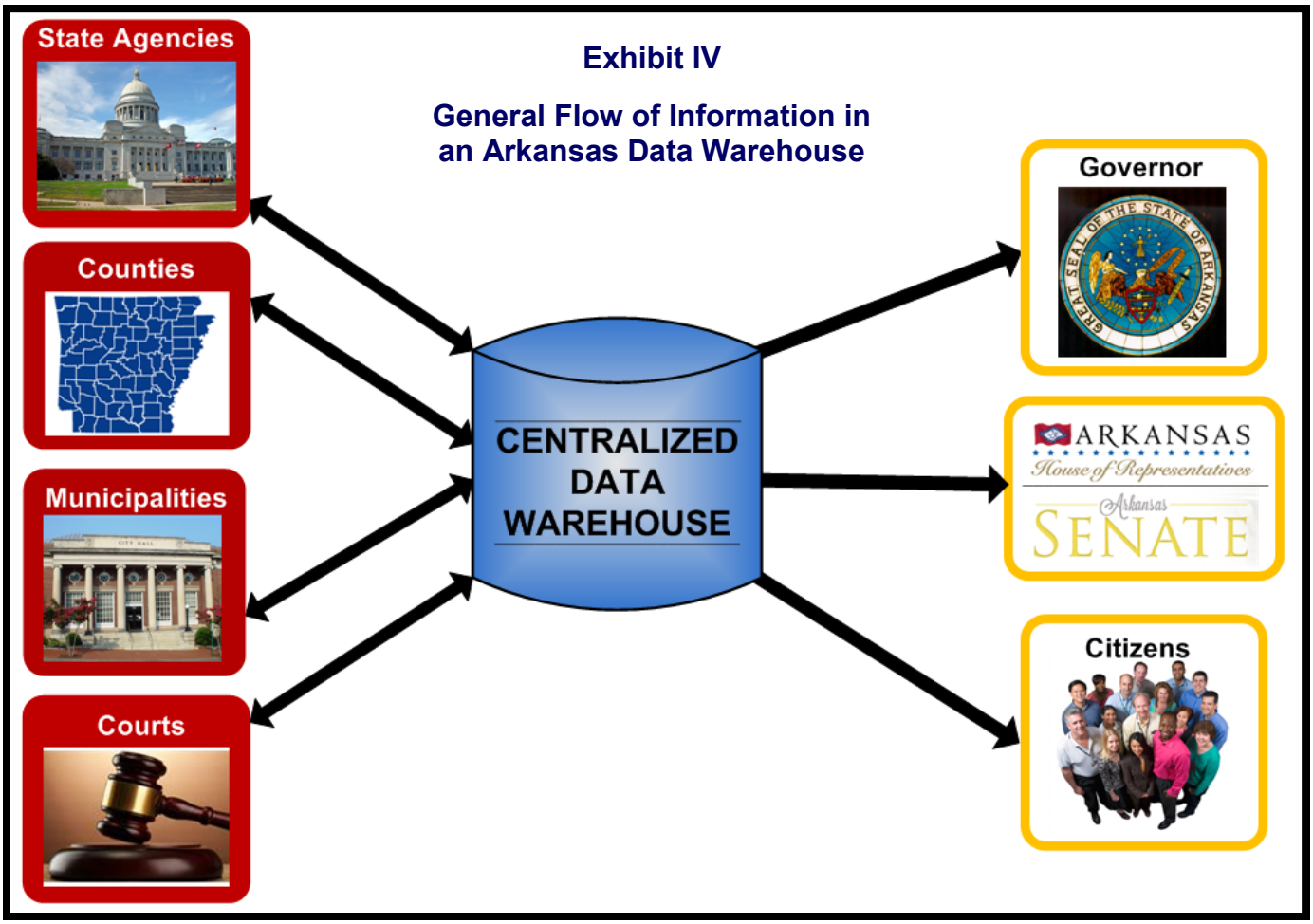
1. Appropriate and authorized access to large data sets for reporting and analytics.
2. Long-term reduction in costs for IT security, infrastructure, and backup.
3. Improved quality and accuracy of data.
4. Sharing of data among state and local entities.
5. Greater efficiency through reduction of duplicate efforts.

1. Direct Access to Large Data Sets

Since a centralized data warehouse gathers data from all state and local entities in one location and in a common format, comprehensive data can be shared, as appropriate and authorized, among entities and made accessible to stakeholders, as illustrated in **Exhibit IV**.

Exhibit IV

General Flow of Information in an Arkansas Data Warehouse



The centralization of massive data sets offers multiple benefits to Arkansas:

- The ability to share and cross-reference information among multiple state entities.
- A "big picture" view of state activities.
- Potential reduction of fraud, waste, and abuse through identification of duplicated and overlapping efforts, irregularities in high-risk transactions, and fraud trends and patterns.
- Analysis of public service performance and spending.
- The ability to predict future needs of the citizenry with more accuracy and confidence.
- Improved transparency and accountability.

By 2020, the digital universe will grow exponentially – to more than 5,200 gigabytes for every man, woman, and child on earth.[1]

More specifically, big data analytics could be used to calculate the potential economic outcomes of changes in state and federal legislation or policy by:

- Evaluating the impact of changes in the economy by measuring changes in economic activity.
- Assessing state and local government revenue collections, trends, and anomalies.
- Estimating the future needs of the citizenry with more accuracy and confidence.
- Analyzing economic incentives represented by tax credits and rebates.

In addition, public service performance and spending could be further analyzed, allowing the State to:

- Negotiate better vendor contracts by identifying vendors not currently on statewide contracts.
- Identify bulk purchase opportunities.
- Monitor state procurement card usage.
- Verify policy compliance.
- Increase tax compliance and identify possible abusive income tax transactions.
- Identify questionable Medicaid claims.

Through the use of big data analytics, legislators and other government officials could access pre-defined reports in a standardized format, search for specific information, or obtain data for more detailed analysis, all with reduced time and effort. For example, with data centralized in one location and in a common format, big data analytics could be utilized to provide information indicating how services provided by Arkansas Rehabilitation Services are distributed around the State, how school spending compares to school performance across districts, or to what degree certain public health services are utilized.

These benefits are not limited to legislators, however. Cities across the nation are using big data to detect disease outbreaks and reduce traffic congestion and to provide citizens with information regarding sources of city revenues and differences in city zoning. Additionally, states are using data analytics to plan for workforce demands based on retirement trends.

[1]John Gantz and David Reinsel, "The Digital Universe in 2020," International Data Corporation (IDC), December 2012.

2. Long-Term Reduction in Costs for IT Security, Infrastructure, and Backup

Centralized data warehousing allows the concentration of IT security in one location and safeguards the integrity of critical information, potentially reducing costs and decreasing security lapses and database vulnerabilities. Based on unaudited figures reported to ALA staff, in fiscal year 2015, 20 Arkansas agencies spent $4.4 million on data security, $32.6 million on application development, and $1.3 million on backup solutions. Of the 20 agencies, 18 currently maintain offsite backups under the control of another state agency (e.g., DIS) or a private, out-of-state vendor. It should be noted that private vendors may limit access to data and make it cost-prohibitive to obtain data regarding state systems.

In 2012, the State of South Carolina experienced a data breach that exposed 3.8 million Social Security numbers, 3.3 million bank account numbers, and proprietary information for nearly 700,000 businesses.

In addition, reducing multiple data destination points to a single backup location decreases the risk of unauthorized access to data. The State's current decentralized IT infrastructure leaves data vulnerable to security breaches. As such breaches become more common nationwide, data security must be a priority. Data breaches can be very costly in terms of not only lost dollars but also lost public confidence in government. The centralizing of the backup process should also allow the simultaneous collection of information in a centralized data warehouse with minimal processing costs. By concentrating IT security in one location, a centralized data warehouse could minimize dollars being paid by the State to vendors to protect, manage, and access multiple database platforms.

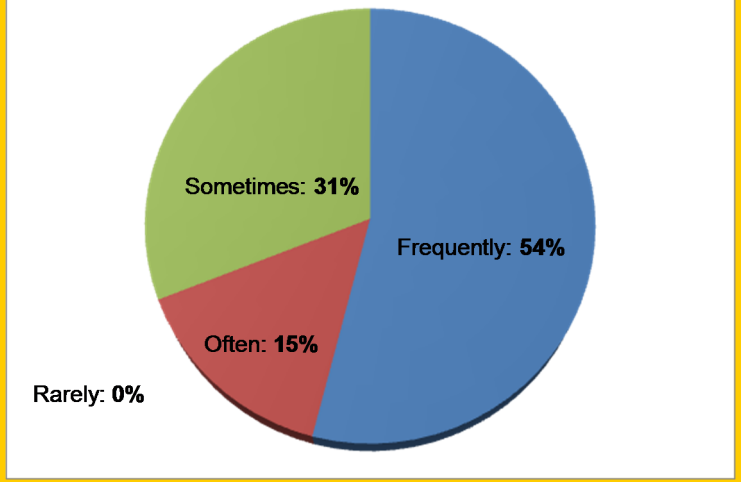
3. Data Quality and Accuracy

Making big data accessible, secure, and in a useable format is only part of the challenge since results of analytics are only as good as data quality. As noted in a 2012 NASCIO report, "The higher the quality of the data, the more powerful are the conclusions drawn from the data analytics."[2]

Data quality and accuracy remain a challenge for many states. In 2015, *Governing Magazine* interviewed over 75 public officials in 46 states whose job duties include data analysis, asking them all the same question: "In your work, how often do you run into problems with data integrity, accuracy, availability, or timeliness?" Among respondents, 69% said they frequently or often encounter problems. Notably, none identified data integrity issues as rare (see **Exhibit V**).

Exhibit V: Data Integrity Survey Results

How often public officials encountered issues with data integrity, accuracy, availability, or timeliness



Note: Based on a survey of 75 public officials in 46 states who perform data analysis as part of their job duties.

Source: *"The Causes, Costs and Consequences of Bad Government Data,"* Governing Magazine, *June 24, 2015*

[2] *Is Big Data a Big Deal for State Governments? The Big Data Revolution – Impacts for State Government – Timing is Everything.* National Association of State Chief Information Officers, 2012.

Centralized data warehousing creates the ability to cross-check data among governmental entities, which should enhance the integrity of data collected by state and local entities. For example, since separate entities provide unemployment benefits, collect taxes, and provide state and federal benefits, it is important to cross-reference identifying information among agencies to ensure benefits are being paid only to those eligible. While some matching is currently being done, the process is time-consuming: The data must be gathered from multiple locations and consolidated into a usable format before it can be analyzed to answer questions. Centralized data warehousing should create the ability to identify all names and addresses an individual has provided to state agencies in a much faster, more efficient manner. Such abilities should allow fraudulent activities to be identified before services are rendered or benefit claims are paid and move Arkansas from the "pay-and-chase" model of providing services – where the State attempts to recover fraudulent claims after they have been paid – to a proactive model that identifies and prevents fraudulent activities before they occur.

4. Data Sharing

Lack of data standardization is a primary obstacle in sharing data among entities. A centralized data warehouse organizes data in a common, consistent format that is accessible across multiple entities, which facilitates data sharing, as appropriate and authorized.

According to the State's 2014 Comprehensive Annual Financial Report (CAFR), Arkansas state government costs total $19 billion per year (combined federal and state funding). As legislators and agency directors increasingly work to allocate tax dollars effectively and increase efficiency in government operations, they need information to evaluate these costs. The vision for the centralized data warehouse is that it will provide secure, accurate, consistent, and organized data that can become information assets shared among entities. Ark. Code Ann. § 25-4-102 declares information and information resources to be a strategic asset of the State. Such a warehouse would be a shared state resource that would increase the yield from existing systems and deliver information in a form that would be useful for everything from day-to-day operational decisions to long-range strategic planning.

A centralized data warehouse would facilitate the administration and sharing of data while also providing data security, meeting confidentiality requirements, linking and expanding state information from existing systems, and supporting analysis and reporting.

5. Reduction of Duplicate Efforts

Another benefit of centralized data warehousing is the reduction of duplicate efforts. For example, efforts to verify data could be reduced to checking a single database, rather than multiple ones located at various entities. Whether the information to be verified is income eligibility, employment status, or vaccination records, this information could be accessible from one data center, eliminating the need for multiple entities to obtain the same information from other entities.

In addition, as legacy IT systems are replaced at the state and local level, how the new systems will provide data to a centralized data warehouse should be considered, and the interface should be standardized as much as possible.

IMPLEMENTING A CENTRALIZED DATA WAREHOUSE

A centralized data warehouse could be introduced to the State through the following steps:

1. **Appoint a Chief Data Officer to lead the project.** This individual must clearly understand the needs and expectations of stakeholders and the objectives of a centralized data warehouse, as set forth by the General Assembly. This position could be a stand-alone position or be housed within a state agency able to meet the diverse centralized data warehousing needs of multiple state and local entities (e.g., DIS).

2. **Conduct a feasibility study** to identify the IT requirements for a centralized data warehouse, determine the cost and expected benefits of the project, look at ways to reappropriate existing state funding, and develop a clear strategy for development and successful implementation.

3. **Ensure data security** by controlling access to the centralized data warehouse and ensuring encrypted transmission and storage of data.

4. **Utilize currently-available facilities** for the project. Arkansas Data Center West, a 9,600-square-foot, physically-secure, environmentally-protected facility operated by DIS, was established in 2013 as a secondary backup facility for the State (see **Exhibit VI**). Currently, utilization is at 15% and steadily growing. This location provides an opportunity for efficient use of state resources and backup by both state and local governmental entities to an in-state facility.

5. **Address known challenges**. State and local entities must adopt a culture of data sharing and address legal challenges if a centralized data warehouse is to be implemented successfully. Currently, entities may reluctantly agree to share data, or personnel may not be comfortable with the idea of "their" data being available to others. However, these data are a state asset and should be available as appropriate to those who have a defined need. It should be noted that federal regulations control access, use, and sharing of state data assets.

Exhibit VI: Arkansas Data Center West

Source: *Arkansas State Data Center West brochure (http://www.arkansas.gov/dis)*

CONCLUSION

Sound decision-making is critical in state government, and the best decisions are made when all relevant data are available for consideration. One of the best possible sources of that data is a well-designed centralized data warehouse. In this era of big data, the State generates enormous quantities of information from diverse sources, and a centralized data warehouse could allow realization of its full value as a strategic state asset, as defined by Ark. Code Ann. § 25-4-102, by making information transparent and more usable at higher levels of integrity. Building a centralized data warehouse in Arkansas should be a statewide initiative requiring technical expertise and cooperation from state and local entities working in close partnership.

RECOMMENDATIONS

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.
- Known challenges are addressed as soon as possible.