DATA LAKES FOR ADVANCED ANALYTICS

2024 NASCIO AWARD SUBMISSION

ENTERPRISE IT MANAGEMENT INITIATIVES

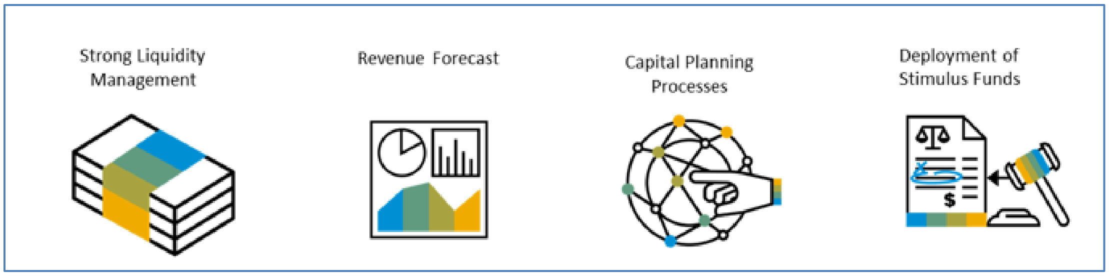
JUNE 2023 - MARCH 2024

CONTACT DETAILS
COMMONWEALTH OF PENNSYLVANIA
OFFICE OF ADMINISTRATION / OFFICE OF INFORMATION TECHNOLOGY
INTEGRATED ENTERPRISE SYSTEM OFFICE (IESO)
ROBYNE LYONS - IESO CIO
RA-OAIES_PMO_TEAM@PA.GOV

EXECUTIVE SUMMARY

The Data Lakes for Advanced Analytics project utilizes the Commonwealth's data effectively by creating and leveraging data lakes to mitigate the risk and issues associated with reliance upon fragmented and disparate data sets. A data lake differs from a data warehouse in that it serves as a central data repository, consisting of raw data, that helps to eliminate data silo issues, whereas a data warehouse consists of pre-processed data. Leveraging data from a single source promotes the ability to apply new insights to the combined data and provide a holistic view to business users across the Commonwealth. The real-time analytical reporting enabled by this project allows users to make critical business decisions quickly and efficiently and provides accountability and transparency to key stakeholders. Critical success factors are real-time data access, real-time key performance indicators (KPIs), and real-time insights.

Examples of key processes business owners seek for analytical purposes include:


Strong Liquidity Management
Revenue Forecast
Capital Planning Processes
Deployment of Stimulus Funds

After deploying the solution in March 2024, the project has empowered business users to make critical business decisions with insights derived from advanced analytics. The use of harmonized data standards and embedded machine learning analytics has helped to modernize the Commonwealth's approach to data and analytics. Additionally, solutions are persona-based to ensure usability for the end user, (e.g., executive level, manager level, transactional level, etc.).

PROJECT DESCRIPTION

Idea:

Like many large organizations, the Commonwealth of Pennsylvania has vast amounts of data that are siloed in disparate applications, systems, and storage locations. These applications and systems include, but are not limited to, the Commonwealth's enterprise resource planning (ERP) system, travel system, various procurement systems, and agency applications, as well as external third-party data (e.g., purchasing card data).

Although the siloed data may function for a particular use case or individual agency task, creating a holistic view of the data is far more difficult since the data is often duplicated and altered across different systems and agency applications. The current approach results in different versions of data being available in different locations and applications, which significantly hinders accountability and transparency to key stakeholders.

Key business challenges:

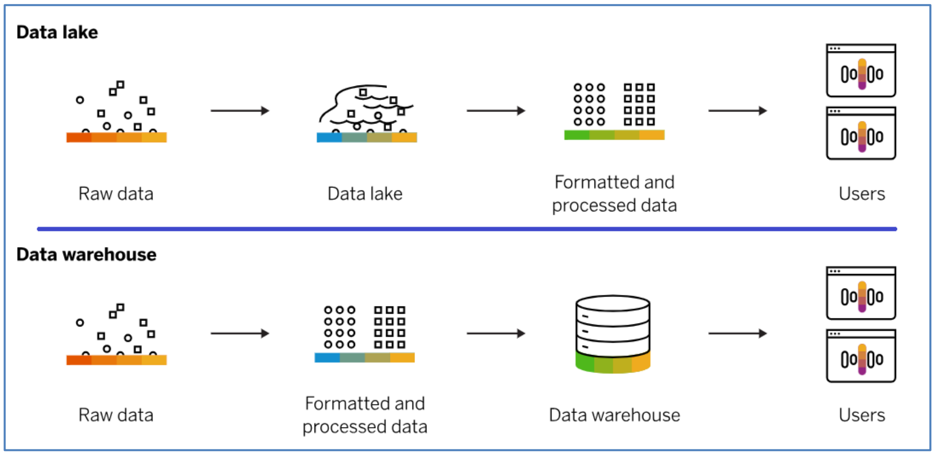
- Fragmented data - Data stored in multiple locations, including duplicated data, or versions created for specific use cases.
  - Lack of data management and orchestration for extensive volumes of data across various systems and applications without a single, trustworthy, and centralized source.
- Lack of enterprise-wide data standards, policies, and metadata management.

  - Siloed reports and dashboards often rely on ad hoc data, leading to stale and untrustworthy metrics. (e.g., Expenditure Dashboard A utilizes a monthly data set and Expenditure Dashboard B utilizes real-time data, resulting in inconsistent information.)
  - Limited capabilities to provide a strategic information management approach, limited digital adoption by business users, and lack of advanced analytical capabilities.
  - Lack of predictive analytics requires business users to use non-intelligent technologies that can lead to poor outcomes.
- Extended time needed to create reports and dashboards due to manual efforts to gather, define, and compile data.
  - No ability to provide multi-dimensional data transparency (e.g., different views of Commonwealth spending include commitments, obligations, invoices, purchase orders, etc.)
  - Lengthy development cycles result in business owners creating localized solutions for reporting.

Faced with these challenges, the Commonwealth of Pennsylvania launched an initiative to create a modern data strategy and architectural approach to deliver intelligent analytical applications to empower business users to make quick and effective data-driven decisions. The Data Lakes for Advanced Analytics initiative focused on strengthening the Commonwealth's commitment to government accountability, transparency, and cost-efficiency.

There are four fundamental principles defined as part of this project realization:

1. **Data goals**: A lean, efficient, and simplified data approach with a heavy focus on self-service capabilities across key functional domains. The domains include on-premises and software as a service (SaaS) data sources; real-time data acquisition with no data redundancy / persistency; technology agnostic data discovery; and advanced analytic capabilities.
2. **Collaboration**: Achieve the highest level of collaboration between business and IT to achieve digital, data, and analytics goals.
3. **Customer-centric approach**: Focus on customer needs and overall experience with the solution.
4. **Greater than 90% Availability**: Data immediately accessible across domains.



The Data Lakes for Advanced Analytics solution included the creation of individual data lakes for functional and cross functional areas such as Finance, Procurement, Travel, and Human Resources. The resulting data lakes are tied together by key data elements across each of these functions (e.g., a document number may be used as a key data element to link a procurement and a financial transaction. Similarly, an HR record can be used to tie together an individual travel request and the eventual invoice). The data lakes include data from the Commonwealth of Pennsylvania's on-premises ERP system and the cloudbased travel management system. The Commonwealth's third-party purchasing card (P-Card) data was automated to replicate daily from the bank servers. The various development efforts were accomplished via different components of SAP Business Technology Platform (BTP). Key solution highlights are as follows:

- Multiple data lakes with key data elements were created for Financial, Procurement, Travel, and Human Resources to provide live and scalable data for business analytics.
- Key data elements within the data lakes were used to maintain the data referential integrity to ensure it is the single system of truth for data within the Commonwealth's ERP system. This data orchestration provides a platform to significantly reduce future development cycles from months to days.
- By utilizing data directly from the Commonwealth's ERP system, the solution minimizes the risk of data duplication. This also minimizes maintenance overhead and reduces the potential for inconsistencies.
- The solution eliminates the need for complex data transformation, simplifying the data extraction process and increasing efficiency.
- The solution groups data from different lines of business, adding a layer of data governance.

  The solution provides advanced analytical capabilities such as predictive analytics, geospatial, and highlight anomalies within the data. This empowers business with real-time, consolidated insights through dashboards and reports.

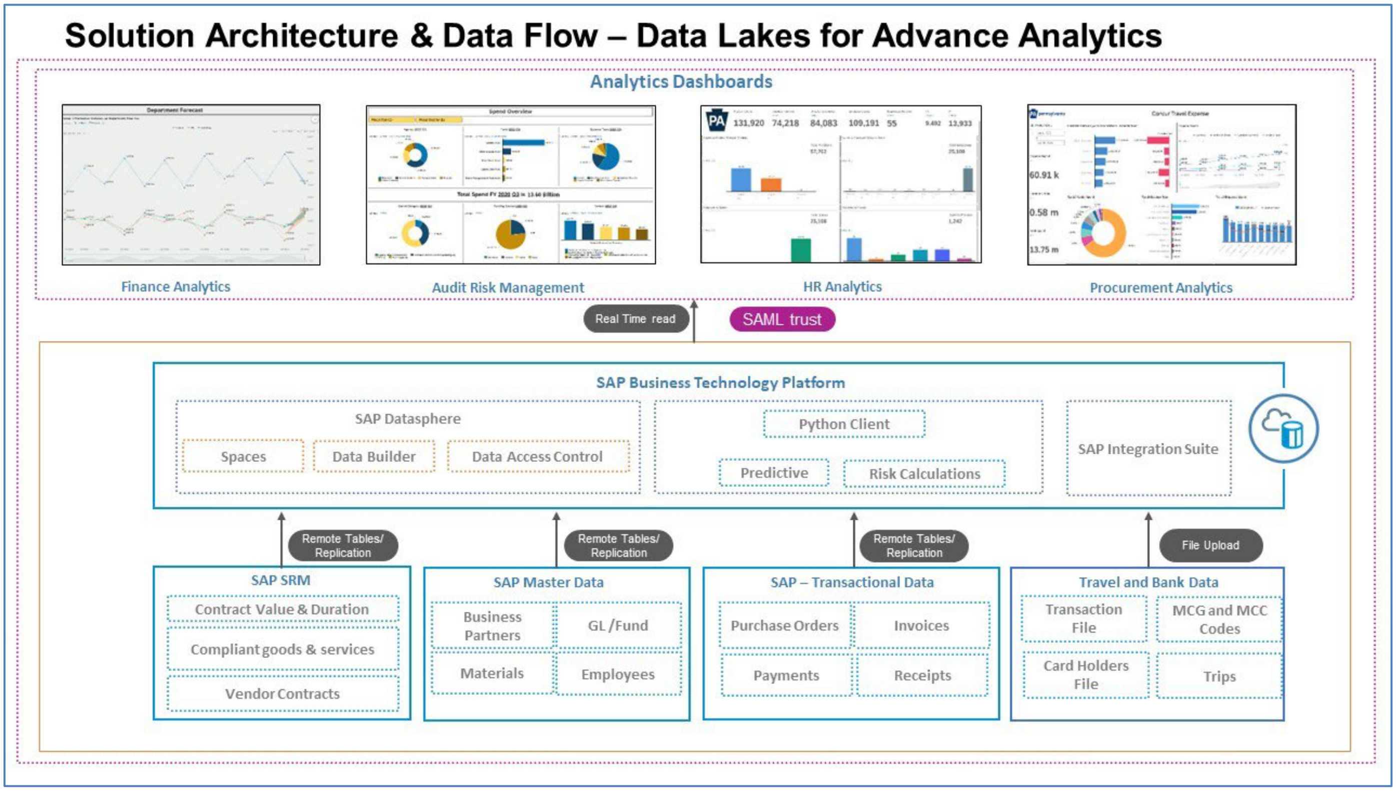
Implementation

How did we do it?

Implementation consisted of the following:

1. Defining core analytical processes and cross functional process dependencies
2. Defining the project team setup, collaboration alignments, Agile methodologies, and planning
3. Increasing digital adoption and business change management
4. Architecture definitions
   a. Governance and roadmap definitions
   b. IT architecture
   c. Tool selections
   d. UX design and analytical wireframes
   e. Design thinking and incremental iterations
   f. Self-service capabilities and ease of use
5. Infrastructure requirements management and license setup
6. Data validation, production setup, and business onboarding

Below is the high-level architecture along with the outcomes delivered by the program.



Solution Architecture & Data Flow – Data Lakes for Advance Analytics

Who was involved?

Various agencies and business users were involved in the initiative and contributed to its overall success. The Integrated Enterprise System Office (IESO) served as the primary sponsor with the department's technical advisor, The Solutions Co. (TSC). Subject matter experts from each of the following functional areas were instrumental in the overall approach and validation efforts:

- Finance
  - Controlling and Budget Execution
  - Accounts Payable
  - Accounts Receivable
- Human Resources
  - Payroll
  - Travel
  - Time Management
  - Core HR

Procurement

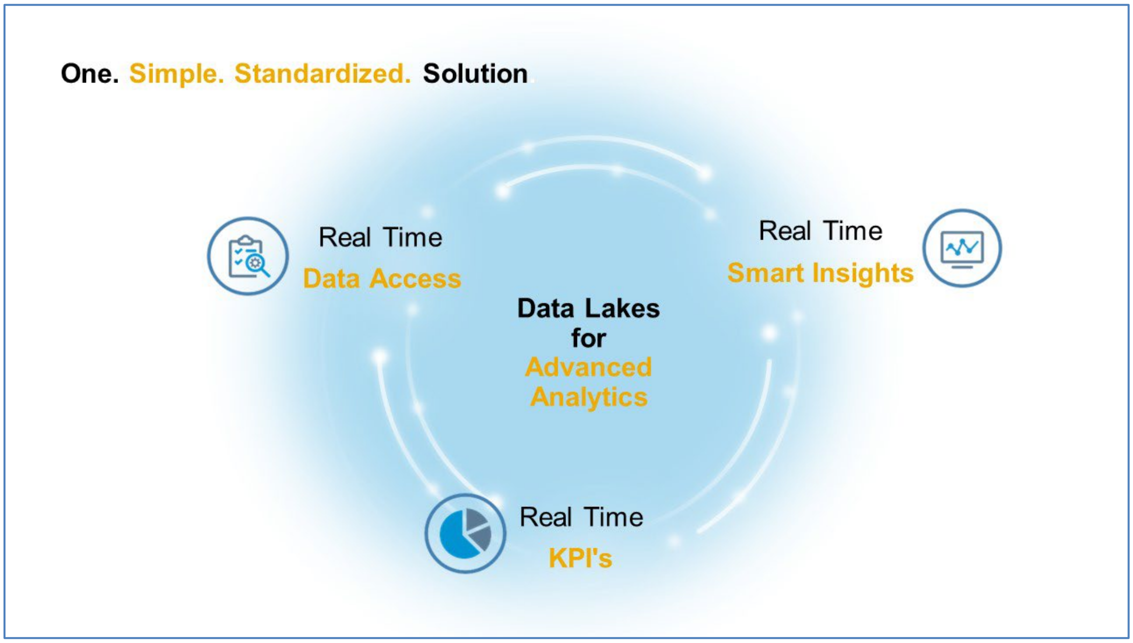
Impact

Outcomes

The overall implementation led to:

- State of the art data architecture
- Simplified, flexible, and scalable data lakes implementation across the business functions and business operations
- Better management of extensive volumes of data across various ERP and third-party systems

- Creation of a centralized source of trustworthy data for streamlined reporting and sophisticated analytics
- Elimination of complex data transformations, siloed implementations, and manual efforts
- Simplification of data transfer process and increased IT efficiency



What did the project make better?

Various Commonwealth of Pennsylvania agencies now have a central data repository consisting of real-time data to ensure that a single system of truth is utilized for analytical purposes. Eliminating dependencies on the collection and comparison of data from disparate systems has allowed a shift in business focus to a more analytical mindset. Non-technical business users are also beginning to recognize and take advantage of faster and more informed decision-making.

How do you know?

Business owners are now able to spend less time collecting, reviewing, and manipulating data and more time applying business knowledge to analytical results, significantly helping to drive an informed decision-making process. The project's utilization of harmonized data standards, persona-driven dashboards, and embedded machine learning has reduced the time required to produce desired analytics by 70%. Business use case examples include:

- **Procurement Approval Data:** Commonwealth staff previously utilized an 18-step manual process to view procurement approval data across the organization. The Data Lake / Analytics approach was able to condense the process into five user friendly steps by leveraging cross functional data lakes tying both Human Resource and Procurement data into a single model. The data lake solution allows users to easily access specific procurement actions and users with real-time results.
- **Procurement Role Data**: The previous process to view users with access to specific procurement roles required significant manual efforts to join Role and User Data to Organizational Data. Steps included a "point in time" extract of over 15 procurement roles and associated position data from the ERP and then merging the position data with Organizational HR data. The Data Lake / Analytics approach resulted in the delivery of a business friendly, less-than-five-clicks-solution within a week. The new approach leveraged multiple data lakes and joined each of the data sets in a user friendly, searchable format to provide real-time results.

- **Spend Analysis Dashboards:** The Commonwealth's ERP system provides functionality and logic for financial management and budgeting purposes but does not natively deliver intuitive identification of actionable procurement cost savings and other spend management opportunities related to overall spend. The Data Lake / Analytics approach transformed supplier, agency, and contract data into a series of three real-time dashboards focused on SAP Accounts Payable - General Ledger line item postings, SAP purchase order line item data, and P-Card transaction data. The interactive dashboards provide visibility and analytical insights on Commonwealth procurement spend that previously would have taken significant efforts to compile.
- **Purchasing Card Analytics:** The Data Lake / Analytics approach significantly streamlined manual steps required to answer basic business questions regarding overall usage of the Commonwealth's purchasing cards. Previous solutions approached this data with a data warehouse / operational reporting mindset where available reports were pre-filtered and targeted a specific use case. This approach resulted in multiple inquiries to obtain the desired information. By leveraging the Data Lake / Analytics approach, users are now able to explore the entire purchasing card data lake with user friendly, searchable options without the constraints and limitations of predefined filters.
- **Human Resources Data Lakes:** HR data lakes were created to streamline the fulfilment of frequent requests for data by various agencies. The HR data lakes consist of data elements pertaining to HR actions, employee details, job, organization, and position data. The IESO HR team notes that the enablement of data lakes provides business owners with a curated set of data elements for use in various applications. Regarding time savings, the team now receives fewer requests for custom data interfaces, freeing up resources to work on other business requests. Additionally, development time has been reduced from months or weeks to same-day turnaround for basic requests.
- **Governor's Budget Office (GBO) Views:** Several financial management reports were being shared via spreadsheet and/or emails and did not consist of real-time data. Real-time data becomes a more significant risk when combining data from various SAP modules. The Data Lake / Analytics approach utilizes the SAP Analytics Cloud (SAC) solution to provide access to combined data models, significantly improving GBO's ability to make decisions and answer stakeholder questions without needing significant effort and time.

What now?

The Commonwealth of Pennsylvania's initial investment in the tools and technology necessary to establish the core set of data lakes has proven successful in paving the way towards a simplified data landscape. The approach has significantly reduced development cycle times by delivering analytics projects with measurable outcomes within weeks. The availability of the Commonwealth of Pennsylvania's enterprise-wide data lakes continues to transform government operations and services by providing better business outcomes and new insights. The Data Lakes for Advanced Analytics initiative continues to grow and focus on the critical success factors for real-time data access, real-time key performance indicators (KPIs), and provide real-time insights.