Analytics and Data Sharing

Ohio's Journey to Enterprise Analytics and Data Visualization

Category: Data management, analytics & visualization

State: Ohio

Project initiation: July 2016

Project completion: November 2018

Contact:
Derek Bridges
Department of Administrative Services, State of Ohio
derek.j.bridges@das.ohio.gov

NASCIO 2019 STATE IT RECOGNITION AWARDS

Executive Summary

Ohio is taking on one of state governments' toughest challenges: sharing data cross-programs and cross-agencies. The State of Ohio had the need to implement modern technology in order to advance their analytical maturity into predictive and prescriptive analytics. Data Warehouses require much more front end work, and generally do not have the capability to ingest massive amounts of raw data and transform to insights rapidly. Yet, both processes and systems are required to become modern and agile, and leverage next generation technology to meet the demands of incoming analytical work. Ohio's goal is to deploy and successfully manage the change associated with introducing new and evolving technology. **Therefore, sharing data and making data actionable through better data and insights to tackle complex problems with solutions that improve Ohioans' health, security and well-being.**

To eliminate data silos and share information across agencies, the state and agency security and privacy leaders partnered to create a data sharing protocol, templated data-sharing agreements, and project governance recommendations. The protocol helped to break down barriers to data sharing and address other overall enterprise challenges. They designed the platform with state and federal regulations in mind, including HIPAA, FERPA, PHI, 1075, CGIS, and PCI.

> *The modern architecture exists to make the State's data **actionable**, **increasing efficiencies** and **improving citizen services***

Data privacy considerations were at the forefront from the very beginning. The state of Ohio's privacy office and legal counsel partnered to establish standard processes for data sharing agreements. To address security concerns, the state Chief Information Privacy Officer (CISO) partnered with agency security, IT, and Chief Information Officer (CIO) representation. Finally, Ohio developed processes and training to provide clarity, process, and responsibilities for data stewardship, stipulating that control of the data would reside with agency sources.

The holistic analytics platform technology supports state of Ohio agencies comprehensively, including a big data platform that allows for the collection, sharing, and applied advanced analytics of Ohio's data across State agencies. **The platform includes a hybrid approach allowing agencies to choose whether to run their workloads in the cloud or on the State's on-premise data platform**. Agencies are then able to choose from the variety of ingestion patterns, focusing on what makes the most sense for an integrated architecture. The modern architecture exists to make the State's data actionable, increasing efficiencies and improving citizen services. Ohio is one of only five states to earn an "A" from industry experts for its work in advanced technology.

The strategy for self-service capabilities enables state agencies to source, manage, understand, and apply their data insights without the need for IT intervention. The platform brings agencies the ability to quickly leverage capabilities in the areas of streaming, natural language relational search, big data engineering, unstructured data processing and parsing, fuzzy matching, machine learning, artificial intelligence, and applied advanced data science. This allows agencies to quickly realize business value by leveraging pre-built capabilities and products. Further, it enables agency IT to focus their efforts on converting platform capabilities to agency value through integration, security, governance, and automation activities rather spending resources setting up IT infrastructure and redundant capabilities.

Concept

Ohio's big data technology was implemented to modernize and improve service delivery at the State of Ohio. **The goal: to make state government a more effective and efficient leader in using technology to improve customer service and make more efficient use of the State's tax dollars**. Ohio is taking an iterative approach to modern data science, placing the citizen at the center of focus and encourages the kind of collaboration and action that will make Ohio a more attractive place to live and work.

Ohio's analytics platform leverages true big data science including advanced analytical data models and modern embedded analytics data visualizations. From a platform perspective, the technology securely accommodates agency projects and workloads in all of Ohio's agencies through the use of edge nodes while maintaining a common data repository. When an agency is onboarded to the platform, edge nodes are created to support them. This allows agencies and multiple external data science experts to perform analytics on the platform in a secure manner but while acting upon common data promoting governance and a single version of the truth. In addition the integrated platform consists of capabilities for self-service data ingestion and de-identification, collaborative data science program development, self-service data science and data transformation, modernized data exploration, profiling, preparation, and visualization. In addition to the technologies deployed within the core platform, edge nodes support a "bring-your-own-tool" approach so that data analysts and data scientists can work with the platform in their preferred analytical toolset.

Because of Health Insurance Portability and Accountability Act (HIPAA) security regulations for agency data, de-identification is required to conduct these research projects that share data across agencies. Repeatable data de-identification workflows anonymize data that is being used for large, cross-agency data analytics projects. Some of these de-identification workflows are complex, with a lot of transformations and large volumes of data, which are pushed to the big data platform for high performance processing.

The data analytics platform provides visual analytics capabilities for agency users to data shared and enriched on the platform. The hybrid platform exists in the state's data center as well as having public cloud options and is available to authenticated users who desire to share data or execute an analytic use case. Flexibility was an objective when standing up the platform. The State wanted to easily integrate state agency systems with the enterprise platform, and provide a number of tools to accomplish this objective.

Big data science technology allows the State to be much more agile when it comes to developing workflows to quickly prepare data and to support data analytics initiatives, such as the large data science research projects on the State's data analytics platform. Agencies have the ability to quickly build prototypes that can quickly identify valuable insights within days/weeks, instead of providing hand written requirements to IT that typically lead to slow turnarounds in delivery.

From a solution delivery perspective, the State administers the enterprise analytics environment including a full Cloudera Hadoop distribution as well as additional technologies deployed within the platform including Tableau, Alteryx, and StreamSets among others. In addition to administration and maintenance, the State is centrally developing solutions on the platform including ingesting numerous data types, streaming, visual analytics, analytical models, security, data governance and audit, and operationalizing solutions which can be utilized by all of Ohio's 160+ boards, agencies, and commissions.

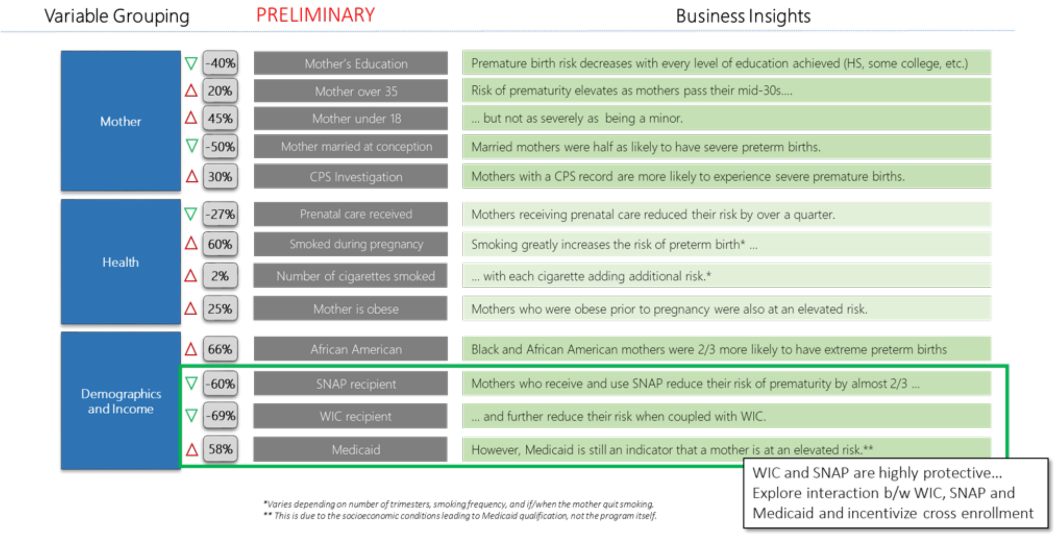
Significance

There is an extraordinary opportunity in performing sharing, unification, and analysis of the underlying data sets by placing them under the lens of advanced analytical tools. The big data platform allows agencies to turn raw data into actionable insights that agencies can use to identify important trends and address real-life scenarios and issues. Additionally the platform was designed to take action in agency transactional systems and programmatic touch points with citizen constituents by providing analytical answers from the “black box” algorithms running within the platform.

Insights alone do not benefit citizens. **Citizens benefit when their state agencies do something with those insights**. Ohio’s state agencies can now compare data based on performance, evaluate outcomes throughout the state, use prescriptive and predictive analytics to anticipate trends, and more. Together, these are helping to drive the creation and augmentation of new programs and initiatives that directly benefit the lives of Ohioans.

With new, modern technology available, agencies can glean actionable and operationalized data analytics from the massive amounts of information at their disposal. With the introduction of this platform, the state of Ohio has encouraged its agencies to move beyond simple research to using their data to truly make a difference for public good.

One example is how Ohio is utilizing the platform to reduce infant mortality. The Ohio Department of Health is utilizing technology to expand and enhance predictive profiling models to determine those at risk for infant mortality and design targeted interventions based on this data. The State has successfully and securely linked over 30 datasets from 4 agencies to form a 360-view of the individual. Key accomplishments include: gaining new insights in program utilization, creation of several models to predict characteristic of mothers most likely to benefit from interventions, and creation of an Online State Health Assessment (SHA) that presents a comprehensive and actionable picture of health and wellbeing in Ohio. Below is a preliminary predictive model that shows a list of indicators that are significantly tied to infant mortality – leading indicators of positive and negative outcomes. For example, mothers who are cross-enrolled in WIC and Medicaid are associated with positive long-term outcomes for both mothers and babies. The Ohio Department of Health is working across agencies to address programmatic enrollment changes for more mothers and babies.



Variable Grouping			PRELIMINARY	Business Insights
Mother	▽	-40%	Mother’s Education	Premature birth risk decreases with every level of education achieved (HS, some college, etc.)
	△	20%	Mother over 35	Risk of prematurity elevates as mothers pass their mid-30s....
	△	45%	Mother under 18	... but not as severely as being a minor.
	▽	-50%	Mother married at conception	Married mothers were half as likely to have severe preterm births.
	△	30%	CPS Investigation	Mothers with a CPS record are more likely to experience severe premature births.
Health	▽	-27%	Prenatal care received	Mothers receiving prenatal care reduced their risk by over a quarter.
	△	60%	Smoked during pregnancy	Smoking greatly increases the risk of preterm birth* ...
	△	2%	Number of cigarettes smoked	... with each cigarette adding additional risk.*
	△	25%	Mother is obese	Mothers who were obese prior to pregnancy were also at an elevated risk.
Demographics and Income	△	66%	African American	Black and African American mothers were 2/3 more likely to have extreme preterm births
	▽	-60%	SNAP recipient	Mothers who receive and use SNAP reduce their risk of prematurity by almost 2/3 ...
	▽	-69%	WIC recipient	... and further reduce their risk when coupled with WIC.
	△	58%	Medicaid	However, Medicaid is still an indicator that a mother is at an elevated risk.**

WIC and SNAP are highly protective...
Explore interaction b/w WIC, SNAP and
Medicaid and incentivize cross enrollment

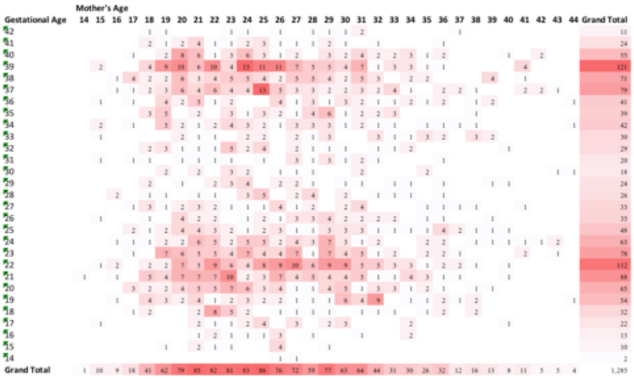
*Varies depending on number of trimesters, smoking frequency, and if/when the mother quit smoking.
** This is due to the socioeconomic conditions leading to Medicaid qualification, not the program itself.

Impact

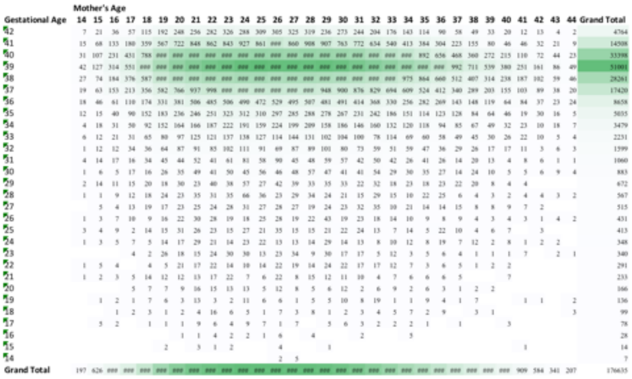
Ohio has consistently ranked among the highest in the country for infant mortality rates. It is a problem the State has taken very seriously which is why infant mortality project to utilize advanced data science capabilities of the analytics platform. The Ohio Department of Health in combination with three other state agencies have effectively and efficiently shared their state data assets, including Ohio Department of Job and Family Services, Ohio Department of Mental Health and Addition Services, and Ohio Department of Medicaid. The agencies are utilizing technology to enhance predictive and profiling models with a goal to determine those at risk for infant mortality and design targeted interventions based on this data.

When the Ohio Department of Health reviewed the infant mortality data, it began to uncover inconsistencies in how the data were collected and classified. Some of the information was incomplete or inaccurate; there were blind spots in the data. Identifying these blind spots and the inconsistencies in reporting can help enhance analysis and data collection, and ultimately change the way data are collected and programs are administered. Below, is work in progress, providing the Ohio Department of Health with a visualization to tell the story about which gestational ages interventions should be focused on. For people less familiar with the research and data behind the causes of infant mortality, visualizations can be powerful for ensuring understanding of an issue such as infant mortality.

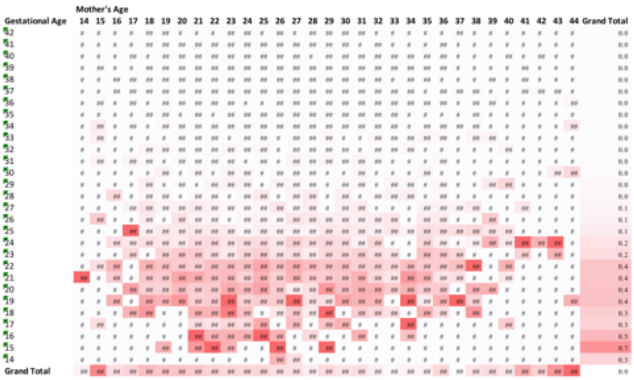
1. Infant Mortalities, OEI Counties 2013-16



2. Live Births, OEI Counties 2013-16



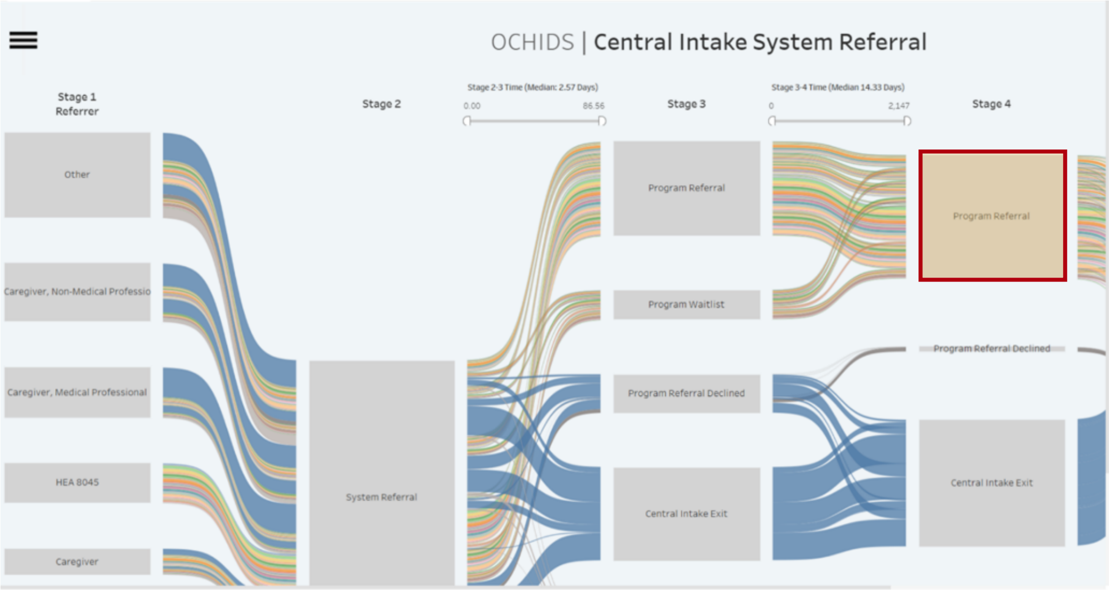
3. Infant Mortality Rates (%), OEI Counties 2013-16



Twenty-one state services are within the scope of the infant mortality project spanning the 4 participating agencies. These services include over 1.5 million enrolled Ohioans. The outputs of the analytics are presented to decision-makers and program experts by embedding the visualizations.

Impact con’t

In addition to understanding the data, The Ohio Department of Health began to explore why programs might be not be effective and/or efficient, and how can State programs for mothers be improved so that help is available before it is too late. As pictured below, what we can see is a program intake journey map of mothers referred to various home visitation providers. The Ohio Department of Health learned that while home visitation programs proved to be successful if mothers are referred and enrolled, the challenge is actually reaching the mother to offer enrollment. The diagram shows that 50% of mothers referred made it into a home visitation program, while the other 50% fall out. If mothers are not enrolled and participating, home visitation programs cannot be effective. Through customer journey mapping, Ohio is using analytics for various referrals and various groups to truly understand what is working, what is not, and why. Program decision-makers can examine and visualize data in order to understand the current state, to inform policy, and drive policy changes. The state has historically spent time understanding what will lead to more positive outcomes for mothers and babies, but this information has unveiled to the Ohio Department of Health that the types of outreach communication methods used to reach their customers is key. The Ohio Department of Heath exploring new and modern communication channels and methods to reach their customers, and understand if interventions will reach those in need.


OCHIDS | Central Intake System Referral
Stage 1
Referrer
Stage 2
Stage 2-3 Time (Median: 2.57 Days)
0.00
86.56
Stage 3
Stage 3-4 Time (Median 14.33 Days)
0
2,147
Stage 4
Other
Caregiver, Non-Medical Professio
Caregiver, Medical Professional
HEA 8045
Caregiver
System Referral
Program Referral
Program Waitlist
Program Referral Declined
Central Intake Exit
Program Referral
Program Referral Declined
Central Intake Exit

The Infant Mortality Rate project combines data from around 30 data sets from 4 agencies. The data sets were sourced from a wide variety of sources ranging from database systems, file based integrations, and API based sources. Data sets were in diverse formats including comma separated, pipe delimited, custom formatted, JSON , XML and SAS formatted files. The total size of the data set was around 1.4 TB including replication around the cluster. The average file size was around 1GB with the biggest file of around 30 GB. 14 Databases were created on the Hadoop ecosystem. Data was ingested on-demand and on-periodic frequency using Sqoop data pipelines, Hadoop commands, and by utilizing StreamSets. SAS files were translated using R to conform to the standard csv format for easier analysis by the data science team. To handle schema changes, files were stored in AVRO format and to promote faster analysis, curated datasets were stored in Parquet formats. Data from the analytics platform was integrated in an automated fashion into Tableau to present the results visually to the public. Tableau was embedded/integrated into state websites using a Javascript API.

Impact con't

Ohio is realizing its goal to turn data into actionable information. Multiple agency programs are analyzing critical societal problems for Ohio and the country, including issues like infant mortality, opioid use disorder, transparency, operational effectiveness, and program efficiency. Operationally, data storage costs have come down dramatically and Ohio is quickly eliminating data silos across the state through an enterprise roll out of a single big data platform and supporting services program.

Ohio has undertaken a large-scale educational and promotional initiatives to get additional agencies on board with the platform. The analytics implementation team routinely engages with agency representatives to identify their challenges and goals and develop manageable statements of work (SOWs). Agencies are guided through the procurement process and trained on the technology and data analytics. Further, the implementation team stages data for agencies. Promotion is handled through a variety of channels to reach the widest audience, including newsletters, conferences, and one-on-one engagement conversations with agency leaders, program experts, and technologists.

To date, multiple state agencies have partnered to release additional analytic project SOWs. Multiple agencies have on-boarded to the platform for executing workloads to drive additional value for their agencies and constituents. The platform has succeeded in reducing administrative overhead and streamlining access to data for users across the state's agencies. Users are able to get the information they need and data warehouse storage costs have been reduced.

The opportunities to improve performance and policymaking based on actionable data are only going to increase as the data from state agencies itself increases. Ohio's analytics platform was designed to be dynamic; its flexible architecture allows the platform to grow as enterprise needs grow, demand increases, and policy questions become more complex. Once the data is staged, it becomes highly iterative in nature, since it does not have to be spun up again. The time to deliver actionable information is possible in months, rather than years, thus state agencies are getting more accurate, insightful information, more quickly.

Each question that is input will derive more answers. As more data is shared and the system is used, advanced models can be created which will drive better outcomes. Ultimately, agencies throughout the State will benefit from this intelligence. More importantly, so will their constituents.