MANAGEMENT PERFORMANCEHUB

2020 NASCIO IT AWARD

Enhanced Research Environment Ushers in New Era of Secure Collaboration and Machine Learning

CATEGORY: EMERGING AND INNOVATIVE TECHNOLOGIES

NASCIO 2020 STATE IT RECOGNITION AWARDS

STATE
Indiana

PROJECT INITIATION DATE
April, 2019

PROJECT COMPLETION DATE
March 19, 2020

CONTACT
Graig Lubsen
Dir. of Communication &
External Affairs
glubsen@iot.IN.gov

EXECUTIVE SUMMARY

Technology and resource allocation has equipped state governments to carry out the purpose of improving outcomes for their citizens in ways of which previous generations of leaders likely never dreamed. But to measure the long-term impact and make further improvement in serving citizens through programs like Women, Infants and Children (WIC), for example, governments must fully leverage the data they maintain for citizens. WIC aims to improve the health of pregnant women, new mothers, and their infants, but siloed, heavily- protected data makes it nearly impossible to know if WIC participants achieved intended outcomes like longer gestation periods and lower infant mortality. Oftentimes, private sector organizations—like grocery vendors and distributors—assist in delivering services for these programs as well, but also lack the context to fully understand outcomes.

The Indiana Management Performance Hub (MPH) was created in 2014 to address this need. MPH's mission is two-fold: to support state agencies to help manage performance, saving Indiana residents' lives, time and money; and to support external partners through facilitating governmental and non-governmental partnerships around data. Yet internal and external entities often lack both the sophistication and infrastructure to effectively analyze sensitive datasets while caring for the regulations and security measures associated with them. In April 2019, MPH cast a vision for a secure, scalable environment where researchers and analysts could "meet" to collaboratively share and analyze data. There were no existing solutions that met MPH's requirements, so it enlisted KSM Consulting to build what would come to be known at MPH as the Enhanced Research Environment(ERE).

MPH deployed the ERE on March 19, 2020, just 14 days after Indiana confirmed its first case of COVID-19. ERE would prove crucial to the State of Indiana's COVID-19 response for the same reasons it would revolutionize sensitive data-sharing and analysis for the State of Indiana: it unlocked secure collaboration and facilitated complex modeling and machine learning. ERE is also responsible for enabling the daily, operational dashboard that Indiana Governor Eric Holcomb used to inform the State's response, as well as coronavirus.IN.gov, a public data visualization portal created to support knowledge sharing and analysis of COVID-19-related data for the entire state.

The COVID-19 project within the ERE has more than 50 concurrent users from organizations including Indiana State Department of Health, Indiana University, and Regenstrief Institute, far surpassing unique user key performance metrics. The ERE largely enabled Indiana's speedy COVID-19 response and has since contributed to driving decision making among state leadership when determining next steps to protecting Hoosiers and reopening the State.

PROJECT NARRATIVE

Concept

The key requirements for the environment included:

- Prevent unauthorized access to data by segregating an individual's access to specific projects within the environment
- Monitor user behavior and use of the environment
- Enhance transparency surrounding the analysis and use of sensitive data, including administrative oversight and approval of any exports from the environment
- Enable the joint analysis of data from multiple sources by multiple analysts in a secure manner
- Not only facilitate, but accelerate advanced analysis through the use of best-in-class data science and analytical tools

- Enable sharing of both analysis and code in a secure manner
- Ability to scale to a nearly limitless number of users and volume of data through dynamically flexing to meet computational and user needs in near real time

Once all parties were aligned on requirements for the environment, existing platforms were researched, and it was ultimately determined that no available solutions met MPH's requirements.

The State of Indiana Enhanced Research Environment, under the leadership of MPH, was then built in a highly agile, iterative fashion. The team took an infrastructure-as-code-based approach to create an elastic cloud environment largely built on Kubernetes. The rationale for this approach is as follows:

Elastic cloud environment

- Consumption-based
- Able to scale up or down as needed
- Infrastructure as code
- Brings the application software development approach to server infrastructure
- Version control
- Ability for secondary approval process

Kubernetes/containers

- Ability to decouple tools and applications from their packages and dependencies, such that each container can be spun up independent of another container in order to not run into conflict with other user software that's being installed or other version dependencies
- Capability for massive parallel processing using Apache Spark ERE was deployed on March 19, 2020.

The key features include:

- Windows-based desktop application containing Linux-based open-source tools
- Designed with the ability to be supported on multi-cloud environments so it can be stood up on AWS or Azure if needed
- More than 10 ready-to-use open source data tools, including Jupyter Notebooks and R Studio
- Flexibility and capability for "bring-your-own-license" (BYOL), so data users can install familiar tools they would typically use for analysis, including SAS software, SSPs, IBM, Power BI, and Tableau, and Microsoft Office products
- Code repository so analysts can iterate on and keep track of their code as they develop
- Ability to support an unlimited number of users and projects

Significance

The Enhanced Research Environment plays a critical role in MPH's effectiveness. MPH commenced gathering project requirements in April 2019. A number of external researchers were interviewed, answering questions like, "What are you hoping to gain from a research environment?" and "What tools do you need within the platform?"

MPH was able to begin using the ERE on March 19, 2020, just 14 days after Indiana's first confirmed case of COVID-19. In the following days, ERE would prove crucial to the State of Indiana's COVID-19 response for the same

reasons it will revolutionize sensitive data sharing and analysis for the State of Indiana:

Privacy and security - The ERE is a hardened environment in which data can flow in, but cannot flow out without approval from MPH. This allows an analyst or researcher to upload his or her own dataset into their user space within the environment and marry it up with state data to quickly iterate and produce output. MPH then reviews export requests to ensure that privacy laws, contractual requirements, and security protocols are met. This is a significant improvement over traditional SFTP-type data sharing. The ERE also allowed the State to respond to working remotely— physical presence in a secure facility isn't required for data analysis.

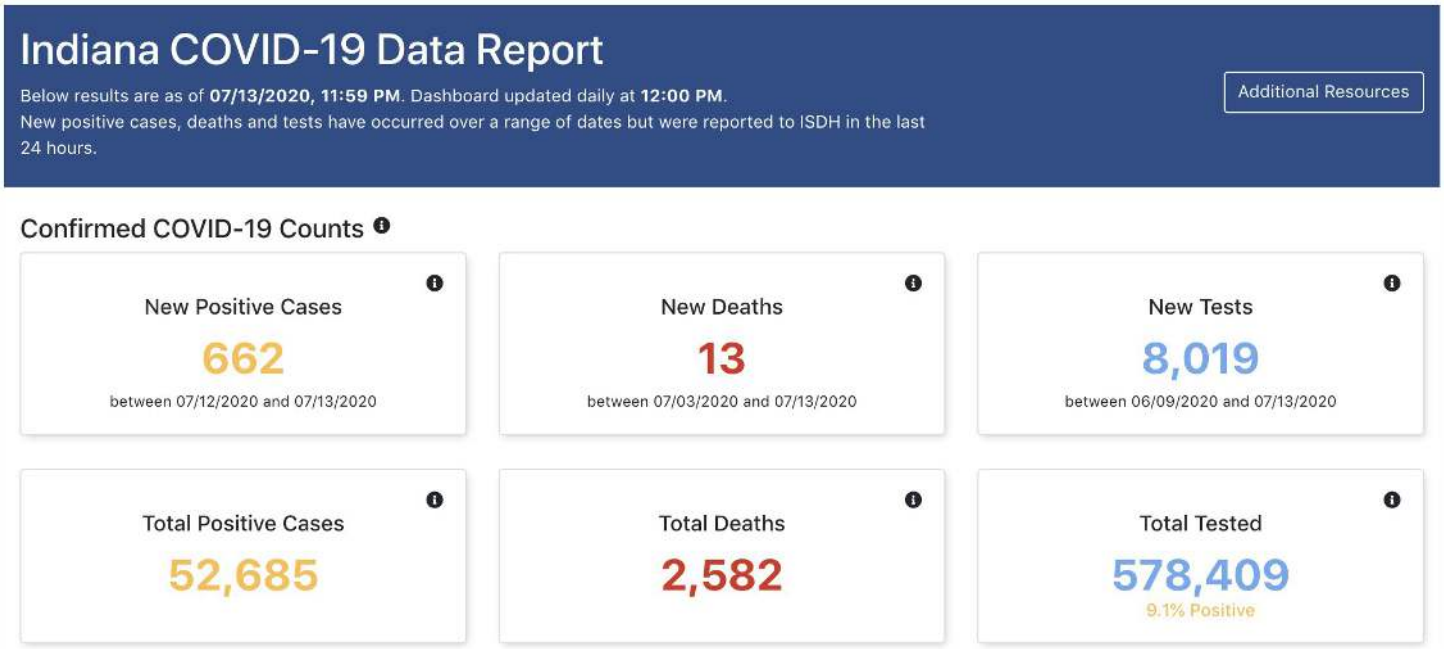
Flexibility and scalability - In mid-March, when the COVID-19 crisis was rapidly evolving, the State needed to obtain and analyze five years of statewide health records to understand endemic signals of flu and predict how it would trail down and free up beds. Those records constituted half a terabyte of data from the Indiana Transparency Portal and required a level of computational horsepower that would produce results in 12 hours. Using a piece of physical hardware with a finite amount of RAM and processing power, a similar job would have taken significantly more time to run in a datacenter. Because computational demands for analysis are widely varying and it's impossible to know the size and frequency of data uploads, the ERE was developed in the cloud to achieve the burst ability, scalability, and flexibility needed to scale up and down. We adjusted hundreds of gigabytes of mobility data from a company without even a question of, "do we have enough space?" Due to its scalable infrastructure, ERE theoretically enables an unlimited number of researchers to collaborate on a single project. As such, a lengthy back-and-forth process is eliminated by enabling collaborative use of code and immediate sharing of results between researchers within ERE who are working together on the same project.

Facilitation of complex infectious disease modeling and machine learning - To inform the COVID-19 response, analysts used natural language processing to engineer features from doctors' handwritten notes to determine how to decouple the pandemic signal from the endemic signal. To understand the endemic signals, the analysts characterized influenza-like illnesses from the chief complaint narrative and the syndromic surveillance system to model epidemic and endemic signals separately. The ERE provides access to both the statistical programming languages and GPU processing power needed to facilitate machine learning.

Impact

Without the ERE, Indiana's COVID-19 response would have been much more difficult to stand up. Prior to the ERE's deployment, there was no centralized location where data could have been quickly consumed and analyzed. If data-sharing had been relegated to email, most of the computation would have been pushed to local environments, which would have been less secure, less accurate, and a lot slower. The ability to quickly engage third-party partners like Regenstrief Institute, Indiana Health Information Exchange (IHIE), and Indiana University researchers was also critical to the speed of the response in the early days of the outbreak when every minute mattered. The ERE enabled analysts and researchers to access critical datasets within hours, rather than weeks.

At the time of writing, there are 17 different datasets being ingested from eight different entities—from government agencies to statewide health partnerships—with more than 56 users from multiple organizations collaborating in a shared workspace. At the project's outset, onboarding five unique users was a key success metric. This goal was surpassed by 900 percent.



Randomized testing

Early in the COVID-19 outbreak, testing shortages across the United States inhibited the response. As more tests became available, MPH supported researchers from Indiana University—Purdue University Indianapolis (IUPUI) and the Indiana State Department of Health (ISDH) in the design and execution of a random-sample testing research study to assess how widely COVID-19 had spread across the state. Between April 25 and May 1, researchers tested more than 3,600 randomly- selected residents and 900 volunteers for viral infections and SARS-CoV-2 antibodies. The results revealed “a general population prevalence of about 2.8 percent of the state’s population,” according to an IUPUI press release. As a result of the ERE, MPH was able to produce both the representative randomized group for the state and contact information for all 3,600 people in less than 24 hours. Typically, for a university, the effort would have taken one to two months at a minimum.

National Guard Response

Data analysis and modeling from the ERE assisted with forecasting when and where the National Guard would need to deploy resources as COVID-19 cases were increasing in various parts of the state. Based on information within ERE from hospitals and healthcare providers around hospital bed usage, ventilators, PPE equipment availability, and viral swabs, the Governor’s office was able to inform and direct where the National Guard was needed to provide supplies, stand up field hospitals, and provide additional resources so supply never outweighed demand.

Public and Operational Dashboarding

The ERE enabled MPH to pull data from across the State to both provide real-time, actionable information to decision- makers from across government, and facilitate the kind of transparency the state knew it needed to provide the public information about the situation as it was evolving. In addition to producing a daily, operational COVID-19 dashboard for the governor, the ERE powers the public-facing COVID-19 dashboard at https://www.coronavirus.IN.gov/.

Resource Usage

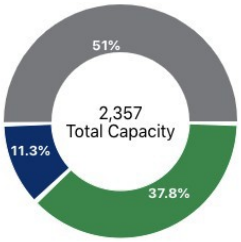
Today | By Day

Today's Statewide ICU Bed Usage

51.0%
ICU Beds in Use - Non-COVID

11.3%
ICU Beds in Use - COVID

37.8%
ICU Beds Available

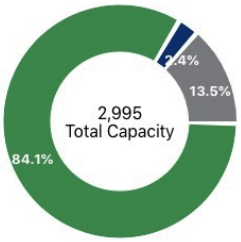


Today's Statewide Ventilator Usage

13.5%
Ventilators in Use - Non-COVID

2.4%
Ventilators in Use - COVID

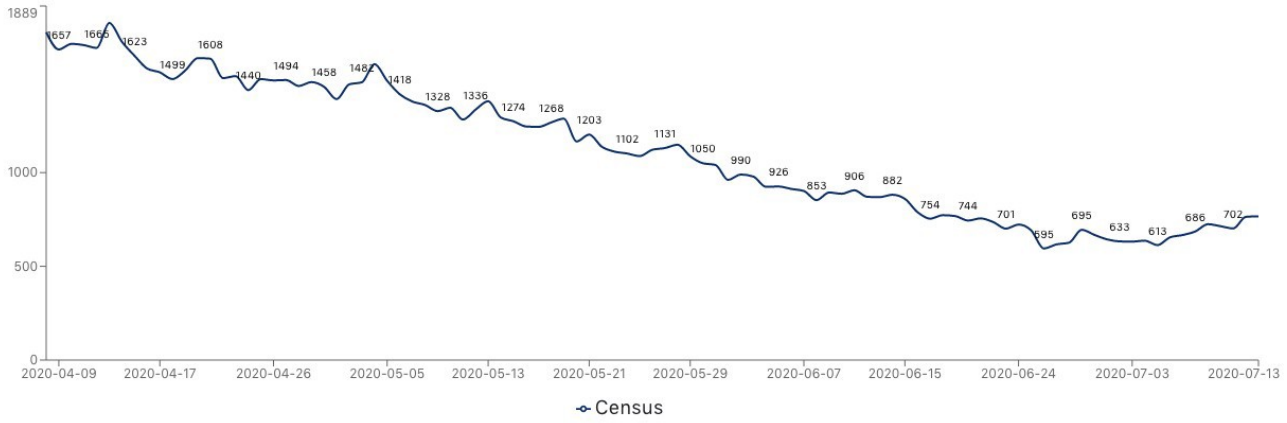
84.1%
Ventilators Available



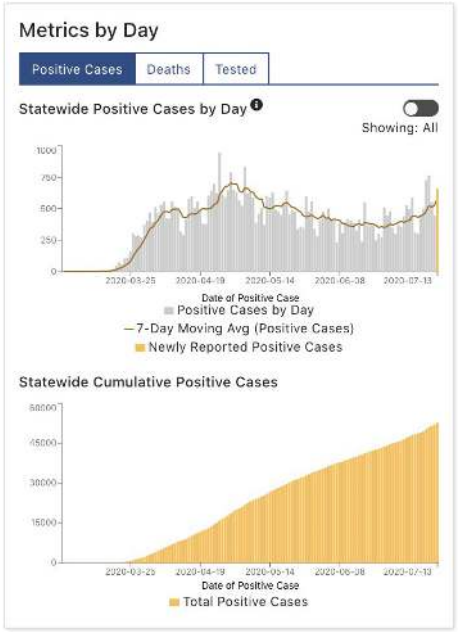
Hospitalizations

Census | Admissions

Statewide COVID-19 Hospital Census



Predictive modeling



ERE has enabled MPH to develop advanced predictive models for disease propagation and resource utilization. The results of these models were communicated to policymakers and helped build a science and data-driven action plan for fighting COVID-19. An ensemble of predictive models was developed to drive near and long-term aspects of COVID-19. Examples of such models include an initial SEIR model to predict COVID-19 spread in the immediate future (the next 24-72 hours), a general time-series based model to predict hospitalizations in near term driven by ILI syndromes, a long-term COVID-19 hospitalization model based on statistical methods such as Richards curve fit, and statistical analysis of length of stay in hospitals to understand the hospitalization metrics. Results of these models were used to calculate PPE usage and need of the healthcare facilities. ERE also provided an environment to install and use agent-based predictive models (FRED) to develop what-if scenarios to better inform policymaking decisions. The modeling team used hospitalization data, mobility data coupled with clinical expert research (FSSA, ISDH) to understand and inform the Governor's five-stage "Back on Track Plan."

Probabilistic Record Linkage

ERE enables record linkage that runs and processes 900 million records every Thursday for all Indiana citizens. It attempts to match people across disparate datasets, like BMV, the health department, department of workforce development, etc., to build context around each Indiana resident and measure the performance of government programs.

[1]Grefenstette JJ, Brown ST, Rosenfeld R, Depasse J, Stone NT, Cooley PC, Wheaton WD, Fyshe A, Galloway DD, Sriram A, Guclu H, Abraham T, Burke DS. FRED (A Framework for Reconstructing Epidemic Dynamics): An open-source software system for modeling infectious diseases and control strategies using census-based populations. BMC Public Health, 2013 Oct;13(1),940. doi: 10.1186/1471-2458-13-940. PubMed PMID: - https://fred.publichealth.pitt.edu24103508- https://fred.publichealth.pitt.edu