

THE FOUNDATION FOR DATA INNOVATION:

The Enterprise Data Hub

Industry Perspective

cloudera®



THE FOUNDATION FOR DATA INNOVATION:

THE ENTERPRISE DATA HUB

Centrally managing data has long been a goal for IT managers. Now, with emerging technology and improvements to the way data can be stored and hosted, centrally managing data is finally a reality for organizations. As such, more and more agencies have been adopting an architecture – known as the enterprise data hub (EDH) – as the spot to consolidate and store data.

In this report, GovLoop explores the power of the EDH for public sector organizations. We spoke with the following four industry experts to help us understand how the EDH is transforming data management for government:

- » **Matthew Carroll**, General Manager of 42six
- » **Joey Echeverria**, Public Sector Chief Architect at Cloudera
- » **Erin Hawley**, Director, National Security Programs, Cloudera
- » **Webster Mudge**, Senior Director of Technology Solutions at Cloudera

With the EDH, data can be stored in its original fidelity, integrated with existing infrastructures and supporting the flexibility needed for various kinds of workloads – such as batch processing, interactive SQL, enterprise search and advanced analytics.

Today this can be done with the proper data protections, governance, security and management needs required by your agency. Powered by Apache Hadoop™ (see Figure 1), agencies can now capitalize on the ability to leverage their data in transformative ways.

By adopting an EDH architecture, an organization has made the commitment to become information driven. This means they understand that data is the key to remaining economically viable in an increasingly competitive world.

“We even talk about [data] as a raw material, like steel and electricity,” said Webster Mudge, Senior Director of Technology Solutions at Cloudera.

But to leverage this raw material to improve decision-making, organizations must overcome numerous obstacles. One of the major challenges is centrally managing data from disparate locations and in various file types.

That’s why so many agencies are looking to the EDH as a means to centralize their data. In doing so, agencies have witnessed significant reductions in the costs of data management – partly due to the reduced duplication of data – and can now have information available to multiple users. The key to the EDH is that it is an open, scalable data architecture that is shared by multiple computing frameworks. This architecture is what has separated EDH from previous data management models. The EDH is driving many benefits for government agencies, such as:

1. Shared resources, systems, memory and data sets.
2. Minimal data movement.
3. Mitigated synchronization issues.
4. Very straightforward data acquisition.
5. Centralized multipurpose workloads.

6. Centralized processing.
7. No incremental costs for new computing.
8. Easy provision of new data.

These benefits are helping agencies to unlock new insights from information. But an often-overlooked benefit of the EDH is that the architecture consolidates not only contemporary data, but also historical data. This gives government the ability to capture data in real time and measure against former trends. This is because the EDH architecture allows organizations to quickly and easily integrate with existing legacy systems.

“What you’re able to do with the enterprise data hub is bring in data sources over time,” said Joey Echeverria, Public Sector Chief Architect at Cloudera.

“[With an EDH,] legacy applications that connect directly with relational databases can also connect directly to an EDH, so that you can transfer data in the places where they exist now into an EDH, and

integrate with the existing tools that you use for accessing that data.”

With the ability to collect multiple kinds of data from disparate systems, the EDH is a powerful tool to help agencies unlock insights from their data.

With the EDH, now the computing power is being brought to the data, improving application development, access to software and information sharing.

“Before this EDH architecture you had different computing environments and you were moving the data around from one application to the next, moving data to the compute,” said Mudge. “The fundamental shift here is that the data is now the center, and the computing is being brought to the data.”

“There are obvious cost advantages in centralizing your management of data, both from de-duplication, and just from being able to use cost effective platforms, without scaling at such steep cost curves,” said Echeverria.

FIGURE 1:

POWERING THE EDH - APACHE HADOOP



The solution at the center of an EDH is Apache Hadoop, a 100% open source platform to store and process data. Hadoop removes the need for organizations to rely solely on expensive, proprietary hardware to store and process data. Yet Hadoop alone lacks the proper governance, data protection, and management solutions needed for the public sector. With Cloudera’s commitment and investments in the open source community, many challenges - such as previously mentioned governance, auditing, data protection, and centralized management with Hadoop - have been addressed.

Two examples of the solutions that Cloudera has created to overcome solution gaps in Hadoop are Cloudera Navigator and Cloudera Manager. Both of these have been specifically designed to complement Hadoop and the EDH as a whole. These solutions are providing government agencies with a full suite of solutions needed to capitalize fully on the power of Hadoop. With Cloudera Navigator and Manager, Cloudera has become the lead engineering force behind Hadoop-powered analysis and storage.

Cloudera’s enterprise data hub offers the ability to:

- » Meet compliance regulations by providing immediate access to data and archiving content digitally.
- » Complement existing data warehouses to help manage costs and improve performance.
- » Support a movement towards self-service business intelligence within the agency.
- » Offer a consolidated approach to searching across data systems.
- » Give you advanced analytic solutions that can quickly and efficiently predict and spot fraud and anomalies.

FACILITATING CONVERGED ANALYTICS & DATA GOVERNANCE

The EDH is also facilitating an effective and efficient data ecosystem and is essential to supporting the adoption of converged analytics strategies. The basis of converged analytics is to provide an organization the ability to gain a holistic view of its data by employing many different kinds of computing simultaneously against a single shared data set without duplication or data movement.

“[Converged analytics] is really critical, because you just don’t know and can’t dictate what exactly is going to be the right mix of analytics at that time,” said Mudge. “And you don’t necessarily want to dictate, because that’s going to get in the way of actually solving the problem.”

To truly leverage data today, organizations must be able to aggregate different data types, run reports against data, all while maintaining data integrity and fidelity.

Yet, with this kind of data freedom, administrators are challenged on how to best maintain the right level of business control. To overcome these trials, Cloudera has created Cloudera Navigator, a native data management application for Hadoop. Navigator is designed to provide data management capabilities for users, administrators, and auditors, helping them to govern and explore the data within the EDH. Specifically, Navigator is guided by four core principles, which Cloudera defines as:

- » **DATA AUDIT AND ACCESS CONTROL** – Verify appropriate user/group permissions and report on data access for files, records and metadata.
- » **DISCOVERY AND EXPLORATION** – Find out what data is available and how it’s structured so that you can use it effectively.
- » **LINEAGE** – Trace the progression of data sets back to their original sources to verify reliability of results.
- » **LIFECYCLE MANAGEMENT** – Ensure the correct placement and retention of data based on value or policies.

Cloudera has also created Cloudera Manager, a system management application for Hadoop and the enterprise data hub. Cloudera Manager provides an agency the ability to deploy, configure, and manage an enterprise data hub, ensure quality of service and performance, and streamline operations – all at scale. Cloudera Manager:

- » Provides a cluster-wide, real-time view of nodes and services running.
- » Provides a single, central console to enact configuration changes across your cluster.
- » Provides per-node services templates for rapid provisioning and rolling updates to maintain service level agreements (SLA) and quality-of-service (QoS) measures.
- » Incorporates a full range of reporting and diagnostic tools to help you optimize performance and utilization.

With the capabilities of Cloudera Manager, IT operators get end-to-end administration for Hadoop and their EDH. Ease-of-use is critical for success with distributed systems like Hadoop, especially as the platform grows to handle new data and new usage.

“THERE ARE OBVIOUS COST ADVANTAGES IN CENTRALIZING YOUR MANAGEMENT OF DATA, BOTH FROM DE-DUPLICATION, AND JUST FROM BEING ABLE TO USE COST EFFECTIVE PLATFORMS, WITHOUT SCALING AT SUCH STEEP COST CURVES.”

- Joey Echeverria, Public Sector Chief Architect at Cloudera

EDH IN ACTION: DEFENSE INTELLIGENCE AGENCY CASE STUDY

The Defense Intelligence Agency (DIA) provides military intelligence to warfighters, defense policymakers and force planners to the Department of Defense and the Intelligence Community. The DIA is responsible for the planning, management and execution of intelligence operations during peacetime, crisis, and war.

Tasked with this mission, it is essential that the DIA has the ability to provide people information safely, securely and efficiently, no matter what the situation. This comes as quite the challenge, as the agency:

- » Manages **400 plus apps** within the enterprise.
- » Has over **1,000 active data sources** consuming data on the order of terabytes daily.
- » Supports over **230,000 daily users** with mission and business needs.
- » Has a network that's **deployed worldwide** at every combatant command.

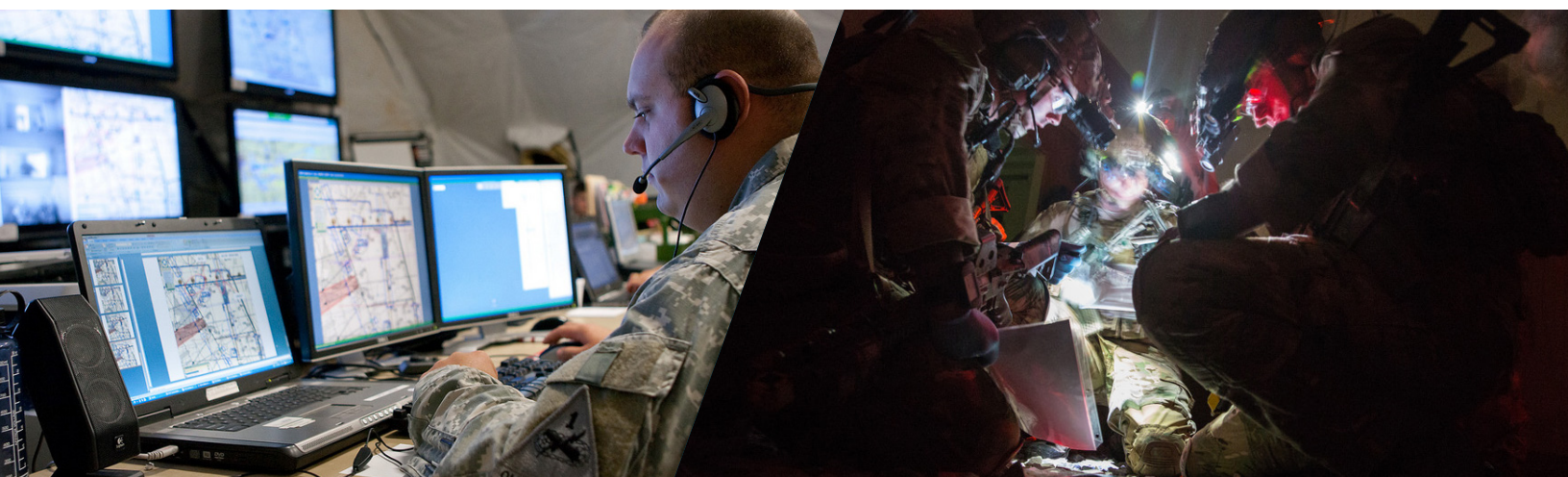
At the DIA, Matthew Carroll, general manager of 42six, which is a CSC company, led the adoption of an EDH. During [a recent online training with Gov-Loop](#), Carroll provided the following statistics on the early return on investments the DIA has witnessed:

- » Distributed query engine is able to cut duplication in storage, cut down data explosion by 100x fold, expecting a **savings of \$15-20M per year**.
- » Deployment model and SDK has cut down time to delivery of prototypes on the cloud from **six months to thirty days**.
- » Applications can be **accredited within a week** from an average of four to six months.
- » Systems and security personnel can be removed from contracts forecasting a **savings between \$25-35 million per year**.
- » Re-use of existing database investments has decreased transition costs between **\$20-30 million over the next two years**.

Although these are significant improvements, the DIA faced numerous challenges to get to this point. Carroll noted that the agency had to navigate the budget to transition over 400 custom apps. App migration is not an easy task, and they also needed more time spent on security efforts.

The EDH provided the necessary framework and architectures to help the DIA transform how they achieve their mission.

"The EDH is creating the necessary interfaces for specific tools for each use case," said Carroll. "And it contains the ability to aggregate all the data in a way so that it becomes discoverable, and indexable in a way that makes sense for application developers – without fundamentally changing the data itself, promoting reuse and rapid development for multiple applications."



"Monitoring at the Tactical Operations Center", "Searching Documents", US Army Flickr

HOW YOU CAN GET STARTED

In nearly every government program, data has been a key factor to deliver better outcomes.

"Agencies have leveraged [the EDH] across the board, whether it's the intelligence community, Department of Defense, or civilian agencies," said Erin Hawley, Director, National Security Programs, Cloudera.

In particular, Hawley identified an example from one client from the defense community. Tasked with need to archive flight mission data, the agency had invested significant hours physically going to archives facilities to find previous mission data records, and bring the information back to the team for manual entry.

"It was so expensive for them to do it that way, it was cheaper to fly the missions again," said Hawley.

By leveraging an EDH architecture, the agency was able to store its archival data at minimal cost, yet also make the data immediately available for analysis and avoid the hidden costs of data access that they faced before. This case study is a testament to the radical shift that agencies are going through in terms of data management.

To start thinking about how your agency your can leverage an enterprise data hub; here are three key starting points.



ASK THE RIGHT QUESTIONS

One of the keys to success with the EDH is that agencies must work hard to understand how data can improve their business operations. This requires leadership and a commitment to becoming a data driven agency. Preliminary questions that organizations should explore include:

- » What's important to our business operations? Our mission operations? Where is there overlap and separation?
- » What kinds of data are we capturing? What kinds will we have to capture?
- » Can we identify data sets that might have value to the organization if viewed, analyzed, or correlated differently?

- » What kinds of insights from data will help us achieve our mission?
- » What's the current status of data in our agency? Is our culture data- or information-driven?

By starting with these questions, agencies can develop deeper insights around data. "We've seen organizations combine datasets that previously we thought to be completely unrelated, and they end up being able to find new insights," said Echeverria.



START SMALL AND ITERATE

Successful organizations have tackled a specific problem through data management, and then reached deeper into their agency to find solutions. "Like anything, don't try to boil the ocean. You need to start off with something small," said Hawley. By starting small and iterating, agencies can learn the best practices and strategies to drive organizational change.

"There's all these new technologies coming out that can take advantage of many machines and we can shift through billions and trillions of records in an unprecedented amount of time," said Carroll. "And that's where you start to get away from the cost, but you get to delivery of new capabilities, new types of analytics and new recommendations."



WORK WITH EXPERTS

The enterprise data hub continues to be an emerging trend for government, and it requires partnership with experts. Moving away from the traditional relational database and siloed information is a cultural and technological shift for agencies.

"[Cloudera] has the original founders of Hadoop. Our whole organization is built around making this platform an open architecture that we can leverage," said Hawley.

In today's world, data is the key to innovation, and through progressive data management architectures like the EDH, agencies can turn their data into insights, and transform the way they operate.

"Data should be at the heart of your decision-making," said Mudge. By taking on this philosophy, organizations will continue to drive innovation and transform the business of government.

ABOUT CLOUDERA

Cloudera is revolutionizing enterprise data management with the first unified platform for big data: the enterprise data hub. Cloudera offers enterprises one place to store, process and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data.

Founded in 2008, Cloudera was the first and is still today the leading provider and supporter of Hadoop for the enterprise. Cloudera also offers software for business critical data challenges including storage, access, management, analysis, security and search.

With over 15,000 individuals trained, Cloudera is a leading educator of data professionals, offering the industry's broadest array of Hadoop training and certification programs. Cloudera works with over 700 hardware, software and services partners to meet customers' big data goals. Leading organizations in every industry run Cloudera in production, including finance, telecommunications, retail, internet, utilities, oil and gas, healthcare, biopharmaceuticals, networking and media, plus top public sector organizations globally.

Learn more at: www.cloudera.com

The Cloudera logo consists of the word "cloudera" in a bold, blue, sans-serif font, followed by a registered trademark symbol (®).

ABOUT GOVLOOP

GovLoop's mission is to "connect government to improve government." We aim to inspire public sector professionals by serving as the knowledge network for government. GovLoop connects more than 100,000 members, fostering cross-government collaboration, solving common problems and advancing government careers. GovLoop is headquartered in Washington D.C. with a team of dedicated professionals who share a commitment to connect and improve government.

For more information about this report, please reach out to Pat Fiorenza, Senior Research Analyst, GovLoop, at pat@govloop.com.

1101 15th St NW, Suite 900
Washington, DC 20005

Phone: (202) 407-7421 | Fax: (202) 407-7501

www.govloop.com

Twitter: [@GovLoop](https://twitter.com/GovLoop)



cloudera®

