The Public Sector Improves Security and Efficiency with Hadoop®

A Modern Data Architecture for Government Agencies

A Hortonworks White Paper
FEBRUARY 2015
## Contents

<table>
<thead>
<tr>
<th>Area</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT Efficiency</td>
<td>• Improve storage efficiency with Hadoop as an active archive</td>
</tr>
<tr>
<td></td>
<td>• Offload ETL workloads to liberate data warehouse capacity</td>
</tr>
<tr>
<td></td>
<td>• Combine high performance computing (HPC) with Hadoop</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Good Government</td>
<td>• Prevent fraud, waste and abuse</td>
</tr>
<tr>
<td></td>
<td>• Deliver timely, accurate public reporting</td>
</tr>
<tr>
<td></td>
<td>• Analyze citizen engagement with social media data</td>
</tr>
<tr>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>
Apache™ Hadoop® in the Public Sector

Within the public sector, many agencies have made the strategic decision to deploy Apache Hadoop in order to deliver data-driven capabilities and enable significant improvements in the security and efficiency of their IT infrastructures—all of which allows them to deliver better outcomes in public services.

Hadoop enables efficient categorization and analysis of large data sets, freeing up resources for higher-value activities in the public sector. For example, the Internal Revenue Service leverages Hadoop data analysis\(^1\) to automate revenue audits, allowing them to streamline the process and drive down costs. Hadoop based “robo-auditors” trace the digital footprint of tax evaders from Internet and social media sites and provide this information to investigators. This maximizes the use of their time and increases the efficiency and accuracy of revenue investigations.

Hadoop delivers a secure storage and processing framework that is efficient and cost effective, providing government agencies new options for optimizing their IT infrastructures and developing new analytic capabilities. This is driving developers and systems administrators within the Intelligence, Defense and Civilian agencies to employ Enterprise Hadoop as part of their investment in a modern data architecture.

Building a Modern Data Architecture with Enterprise Hadoop

To realize the value of your investment in big data, use the blueprint for Enterprise Hadoop to integrate with your enterprise data warehouse (EDW) and related data systems. Building a modern data architecture enables your organization to store and analyze the most important data at massive scale and extract critical insights from all types of data from any source.

Read more at: http://hortonworks.com/hdp

\(^1\) http://www.datanami.com/2013/04/15/irs_to_utilize_big_data_to_improve_returns/
Hortonworks Data Platform Provides Enterprise Hadoop

Hortonworks Data Platform (HDP) is powered by 100% open source Apache Hadoop, with a centralized YARN-based architecture that allows multiple applications to share a common cluster and data set. HDP also provides all of the Apache Hadoop related projects necessary to integrate Hadoop alongside an EDW as a part of a modern data architecture. HDP comes complete with efficient data management and versatile data access capabilities, as well as robust data governance and integration, security and operations components that are critical for widespread adoption and enterprise deployments.

![Hortonworks Data Platform 2.2](image)

Figure 2: Five core capabilities—data management, data access, governance & integration, security, and operations.

Uncovering the Value of Data with Hadoop and HDP in the Public Sector

Hortonworks partners with public sector organizations to help them quickly derive value from HDP and Hadoop by cutting costs through increased efficiency and uncovering new opportunities with advanced analytic applications. Highlighted below are some real-world examples from the public sector that illustrate ways in which Hadoop has been deployed to deliver efficient, data-driven capabilities.
IT Efficiency

With its efficient use of commodity hardware and its ability to scale linearly, open-source Apache Hadoop is an ideal platform upon which to offload many data processing workloads. Through the use of Hadoop, agencies are able to improve the efficiency of their IT systems, allowing them to retain more of their valuable data for longer. Moreover, agencies can leverage Hadoop’s computational framework to create new, advanced analytical applications to better extract insights from their data.

IMPROVE STORAGE EFFICIENCY WITH HADOOP AS AN ACTIVE ARCHIVE

Challenge

In 2013, a fiscal standoff in the U.S. Congress forced a budget sequestration that ultimately shut down the United States Federal Government. When the Federal Government reopened, agencies were faced with sudden, dramatic budget cuts. The cuts pressured leaders to slash spending in all areas, particularly in IT. Because of this pressure, a major federal IT consulting firm was facing work backlogs with twenty of its federal clients. Those agencies had been told that they needed to reduce data storage costs rapidly and still meet all of their mandates and mission objectives.

Solution

Apache Hadoop was designed at Yahoo! to economically store and process huge amounts of data in a distributed platform that could scale linearly. Today, Hortonworks’ public sector customers leverage this economic efficiency to drive down storage costs, in some cases by as much as two orders of magnitude.

The consulting firm analyzed their clients’ existing infrastructures, and they noticed that all data, regardless of its relevance and perceived value, was stored in the same database infrastructure. This infrastructure made sense for the most valuable subset of data; however, much of the data was less valuable and the cost to store that data within the existing database platform outweighed its expected value.

Ultimately, the consulting firm approached its federal clients with this unique value proposition: reduce your storage costs by moving infrequently accessed data to Hortonworks Data Platform (HDP). HDP’s linear-scale architecture, based on commodity equipment, is ideally suited for providing economical data storage. Because HDP is integrated with other leading data center and analytic technologies, analysts could continue to access the data seamlessly and securely, without having to learn new skills or replace existing processes.

Impact

By adding Hadoop to the modern data architecture as a complimentary storage repository, the consulting firm was able realize significant increases in the efficiency of its clients' IT infrastructures while also delivering dramatic cost savings. This allowed them to surpass their budget reduction goals and free agency resources for other uses.

As an example of the potential cost savings, consider Hortonworks private sector customer TrueCar, which achieved a similar reduction in its data storage costs by deploying Hortonworks Data Platform. As reported in the Wall Street Journal in June 2014, “[TrueCar] said that their previous cost for storing a gigabyte of data (including hardware, software, and support) for a month in a data warehouse was $19. Using Hadoop, they pay 23 cents a month per gigabyte.”

2 http://blogs.wsj.com/cio/2014/06/04/hadoop-hits-the-big-time/
Challenge
Traditionally, data storage platform technologies like RDBMS, EDW, and MPP were designed for schema-on-write use cases. First, data managers defined their data structure, or schema, that determined how data would be ingested and stored. Data that did not match the pre-defined schemas needed to be processed with expensive, slow ETL tools to prepare it for storage—or it was simply discarded.

Today, the vast majority of new data created, such as text, voice, clickstream, social, geo-location, image, or video data, are considered “semi-structured” or “unstructured,” meaning that they can have highly variable schemas. This unstructured data must still be stored and associated with existing structured data, but legacy storage platforms require ETL preparation in order to merge the data.

Solution
With its schema-on-read architecture, Hadoop empowers agencies to capture and access all of their data, regardless of format. Simply put, the efficiencies created by Apache Hadoop have shifted the landscape for data storage and processing. Today, it is often more expensive and complicated to determine what data NOT to store in traditional data architectures than it is to capture and store all of the data in Hadoop.

With Hortonworks Data Platform (HDP) government agencies can maintain data fidelity, capturing all potentially relevant data (tabular, social, geographic, clickstream, audio and video) in one multi-faceted set. They can store that dataset efficiently in Hadoop without running it through costly transformations with ETL tools. When necessary, data administrators can use Hadoop to perform ETL tasks currently done on their more expensive platforms and then use the resulting data to feed operational data stores (ODS) and purpose-built analytic tools without losing the fidelity of the original raw data.

Using Hadoop to power your data infrastructure frees more EDW resources to perform high-value functions like analytics and operations, reducing costs and improving efficiency. Just as importantly, ETL processes no longer force constraints on the nature of data that is ingested and stored. Now agencies can capture all potentially relevant data and analyze it in real-time, batch, or through interactive query.

Impact
One federal agency, with operations spanning multiple overseas locations and an infrastructure footprint of thousands of servers, used Hadoop to improve the efficiency of its EDW. By ingesting system logs into Hadoop they were able to transform the data, through aggregation, sorting and joining, before loading it into the EDW. The pre-processing of this data with Hadoop significantly decreased the agency’s data footprint, thereby increasing the efficiency of the EDW.

HDP’s deployment flexibility and partnerships with other datacenter and analytic technologies permitted the agency to modernize its data architecture by seamlessly integrating Hadoop with prior investments. This allowed its analysts to continue using their preferred BI tools to perform interactive queries and searches on that data. Their existing processes were uninterrupted by the architectural improvement, but now the analysts benefit from access to more and better data at a lower cost.
COMBINE HIGH PERFORMANCE COMPUTING (HPC) WITH HADOOP

Challenge
High performance computing (HPC) uses leading edge computing systems, often supercomputers, to perform computations that are either too large or too complex for traditional computers to perform. One example of a workload for HPC systems is a Monte Carlo simulation, which is an algorithm that is used to obtain the distribution of an unknown probabilistic entity through repeated random sampling.

Government agencies use Monte Carlo simulations to model phenomena with significant uncertainty, such as calculating risk or modeling the dispersion of fluids (i.e. a chemical release). To obtain an effective distribution of numerical results these simulations are run hundreds, even thousands of times, generating vast amounts of output data that is typically stored in highly-tuned parallel file systems. However, overhead for HPC datacenter operations is expensive, and those high-cost resources are often utilized below capacity.

Solution
Apache Hadoop clusters can be ideal long-term storage companions to HPC clusters in public sector datacenters. Although HPC environments are not always fully utilized, the facilities and highly-trained technical staff operate regardless of utilization levels. Hadoop workloads can complement HPC workloads and improve the utilization of resources in a hybrid HPC-Hadoop data center.

Most importantly, the results of HPC simulations can serve as yet another input to a Hadoop data lake, allowing other tools access to HPC results for analysis and visualization. Moreover, YARN-powered Hadoop clusters can be dynamically tuned to allocate resources in support of evolving compute and storage needs. This allows Hadoop to drive efficient use of infrastructure by consuming unused cycles on HPC systems, or it can economically store the results or intermediate output data generated by HPC simulations. Both scenarios can reduce the effective cost of storage and processing without diminishing accuracy.

Impact
Government agencies that implement a hybrid HPC-Hadoop datacenter can gain the same type of flexibility and efficiency that Hortonworks enterprise customers see when they combine a highly specialized RDBMS platform with Hortonworks Data Platform (HDP). More importantly, the efficiency of a hybrid architecture can be improved by offloading targeted portions of the processing and storage workloads from HPC systems to Hadoop clusters—without sacrificing the performance of existing modeling and simulation workloads for which the HPC systems were designed. This approach can drive down operational costs while increasing the organization’s ability to consume and derive value from these analytic results across the rest of the infrastructure.
Good Government Use Cases

Hadoop drives improved efficiency and better outcomes in the delivery of public services, such as social aid programs and public works, through its ability to improve resource utilization and transparency. Civilian agencies use Hortonworks Data Platform for both the optimization of existing data systems as well as the creation of new analytic capabilities.

**PREVENT FRAUD, WASTE AND ABUSE**

**Challenge**
Any agency that processes benefits and grants must deal with the inevitable scenario of fraudulent claims in which applicants try to claim benefits to which they are not entitled. The United States Social Security Administration (SSA) described the challenge in a 2012 report:

“Nationally, in Fiscal Year 2011, there were more than 103,000 allegations of Social Security fraud, with more than 7,000 criminal investigations resulting in 1,374 convictions and more than $410 million in recoveries, fines, restitution, judgments, settlements, and savings.”

In many instances, claims audit systems relied on by federal agencies still run on decades-old technologies. This outdated technology is compounding the problem by making it difficult, or even impossible, to identify and respond to these events in a timely manner. Today, a typical batch process might take days to complete and often suffers from access to incomplete or stale data. This cripples an agency’s ability to perform comprehensive data discovery and makes it easier for perpetrators of fraud to avoid detection.

**Solution**
Hadoop addresses this challenge by enabling data discovery at Internet scale, allowing government agencies to securely consume data that was once too big to store and process. Authorized government employees can collect, blend and refine data from sources previously unavailable at scale (i.e. tweets and electronic copies of local newspaper articles or social media posts). Once additional sources of structured and unstructured data are joined with the current data, claims that used to appear legitimate can be rapidly exposed as fraudulent and flagged for review or even stopped in their tracks. Moreover, much of this analysis can be automated, increasing security and minimizing the latency added by human-in-the-loop analysis.

**Impact**
At one major federal Civilian agency, Hortonworks Data Platform’s efficiency reduced processing time from thirty hours to approximately two hours. Now fraud analysts can spend more time investigating results rather than waiting for processing to complete. Naturally this led to an increase in the rate of successful investigations.

Moreover, the agency has already tripled the amount of data included in its daily processing runs and plans to go further. Because Hadoop is a “schema-on-read” system, rather than the traditional “schema-on-write” paradigm, the agency now plans to ingest data from additional legacy systems while also including more upstream contextual data such as social media and online content in its analysis. All of this will make it easier to identify and stop fraud, waste, and abuse.

---

Challenge

National statistics agencies are challenged with compiling vast amounts of economic data from hundreds of sources into an aggregated set of metrics. This data covers a wide range of topics, from industrial output and inventory levels, to inflation and unemployment. Much of this data is reported monthly and the financial markets respond to what they regard as particularly good or bad news. Politicians, investors, the media and the general public eagerly await and closely scrutinize these reports, making accuracy and timeliness of reporting absolutely essential.

One national statistics agency in particular conducts a massive, monthly ETL process to ingest more than 200 million rows of data from comma-separated files submitted by state affiliates. Each month the agency has ten days from the time they receive the data to the deadline for publishing the report.

Solution

For this agency, the initial data validation process has historically taken more than eighteen hours. With only a three-day window dedicated to the completion of this ingestion and verification phase, even a minor statistical error could put the team at risk of missing the publishing deadline.

Hadoop is able to address this challenge through its ability to scale performance and capacity linearly, and through its support for cloud deployment. Predictable performance gains through horizontal scale allow organizations to efficiently size their infrastructure to meet peak demands, while cloud deployment options allow them to dynamically adjust resource allocations to accommodate bursts in activity and effectively control costs.

Impact

The agency began with Hortonworks Data Platform (HDP) in a cloud deployment and through scale and other optimizations was able to reduce data processing time from eighteen hours to less than one hour. This beat even their internal goal of three hours.

The efficiency gained through the modernization of this aspect of the data processing pipeline effectively returned time to other parts of the pipeline, which helped to drive the improved accuracy of the reporting. Moreover, these improvements were achieved without any disruption to the agency’s operations, and were invisible to the business analysts and end users.

Ultimately the agency chose to deploy HDP on-premise after the initial cloud proof-of-concept. However, this scenario clearly illustrates the flexibility of Hadoop and the Hortonworks Data Platform to enable linear scale and dynamic allocation in both cloud and on-premise deployments.
Challenge
Governments have always sought ways to better understand the views of their constituents. Typical outreach, such as speaking at community events or attending a town hall meeting, might be effective for a few high-quality interactions with a small number of citizens, but most constituents do not have the time or motivation to attend in person. The loudest citizens with the free time to attend these events are heard above those who aren’t present. This can lead to an inaccurate or incomplete public perception.

In one country, the national education authority wanted to better understand the views of its constituents. Employing a high-touch approach to understanding public sentiment would have required the agency to enlist hundreds of outreach workers to canvas the country around the clock. Instead, the agency decided to leverage the power of publicly available information by manually analyzing social media posts related to topics of the highest public concern. Each day a team of analysts manually combed through blogs, tweets, and chat sessions to produce a weekly report on what the citizens were saying about education.

Solution
Hadoop’s schema-on-read architecture allows agencies to bring together data from disparate sources and derive immediate value, without waiting for time-consuming ETL processes. Furthermore, its linear scalability allows the infrastructure to seamlessly grow to meet the demand of exploding data volumes without sacrificing performance.

The Hadoop architecture also includes a flexible computational framework that allows agencies to take advantage of existing analytic capabilities, such as sentiment analysis and influence scoring, as well as to rapidly develop new capabilities, leveraging cutting edge open source and commercial analytic tools and techniques.

Impact
Realizing the limitations of their manual analysis process, the agency decided to bring all of the social media data into Apache Hadoop to perform more timely sentiment analysis and trend-spotting. Today, the team uses an HDP cluster to produce daily reports for government leaders that are much more thorough, objective, and quantitative than the original ad hoc reports.

The sentiment analysis provides valuable real-time data on sudden changes in public opinion. The raw data also provides a detailed window into local opinions, down to the individual. Now citizens who are particularly active on social media—or key influencers on the topic of the day—sometimes receive an invitation to a one-on-one meeting with their representative.

One topic of particular importance to this government is the country’s obesity epidemic. About one quarter of all adult citizens in the country is clinically obese. Now, as the government wrestles with educational and public health policies to fight obesity, the education agency is able to use the daily sentiment reports from HDP for rapid feedback on its school health programs.

Big data sentiment analysis provides policymakers unprecedented:
Reach - It represents opinions from a broad cross-section of constituents
Confidence - More data from more sources is the foundation for better analysis
Timeliness - Daily reports allow policymakers to respond quickly
Precision - Highly granular analyses of specific issues and locations
Hortonworks Data Platform Provides Enterprise Hadoop

Hortonworks Data Platform (HDP) is powered by 100% open source Apache Hadoop. HDP provides all of the Apache Hadoop related projects necessary to integrate Hadoop alongside an EDW as part of a Modern Data Architecture. It ships with efficient Data Management and versatile Data Access capabilities, along with three capabilities enterprises require for widespread adoption: Data Governance & Integration, Security and Operations.

![Hortonworks Data Platform 2.2](image)

Figure 2: Five core capabilities—data governance & integration, data management, data access, security, and operations.

Why Hortonworks for Hadoop?

Founded in 2011 by 24 engineers from the original Yahoo! Hadoop development and operations team, Hortonworks has amassed more Hadoop experience under one roof than any other organization. Our team members are active participants and leaders in Hadoop development, designing, building and testing the core of the Hadoop platform. We have years of experience in Hadoop operations and are best suited to support your mission-critical Hadoop project.

For an independent analysis of Hortonworks Data Platform and its leadership among Apache Hadoop vendors, you can download the Forrester Wave™: Big Data Hadoop Solutions, Q1 2014 report from Forrester Research.

About Hortonworks

Hortonworks develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities. Hortonworks Data Platform deeply integrates with existing IT investments upon which enterprises can build and deploy Hadoop-based applications. Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop.

For more information, visit [www.hortonworks.com](http://www.hortonworks.com).

* http://info.hortonworks.com/ForresterWave_Hadoop.html