U.S. CENSUS EMBRACES THE DIGITAL AGE WITH ENTERPRISE DATA LAKE

330M

Residents served

Impact

- Facilitates capture, analysis and management of petabytes of data
- Enables faster, better decisions through real-time analytics
- Improves operational efficiencies and supports information security
- Enables data sharing across agencies

In January, the 2020 U.S. Census officially begins, with the population count in remote parts of Alaska. By April 1, the process begins for the rest of the U.S. population. The census, conducted every 10 years, is the most important initiative of the U.S. Census Bureau, the Federal government's largest statistical agency and the nation's leading provider of quality data about its people and economy. The data collected by the census determines the number of seats each state has in the U.S. House of Representatives, and it is used to distribute more than $675 billion in federal funds to local communities. This funding supports education, healthcare, infrastructure improvements, and more.

The 2020 census requires counting an increasingly diverse and growing population of about 330 million in more than 140 million housing units. And, for the first time in 2020, the census will be conducted largely online instead of by mail. Online collection will result in petabytes of data that must be stored, analyzed, and secured.

Enterprise Data Lake Provides Big-data Processing Capability

To meet this challenge, Census leadership established the Census Enterprise Data Lake (EDL) initiative. The EDL provides big-data processing capability to fulfill petabyte-scale data management and analytics while satisfying security and privacy requirements and controlling costs. The initiative is transforming how the agency processes demographic and economic data using open-source technology and high-performance cloud infrastructure.

"The EDL will support the processing of big datasets quickly and easily with large, dynamically scalable compute and storage capabilities throughout the enterprise," says Kevin Smith, chief information officer at the U.S. Census Bureau. "The data lake also provides a centralized repository to consolidate operational paradata, response data, and cost data from multiple modes of data collection. It provides a single place to analyze all operational data and make informed decisions during operations."

The Census Bureau chose Cloudera as the data platform for the 2020 census to help mine, process and extract insights that can be used to inform important decisions at all levels of government. The platform leverages the entire technology stack and professional service offerings. Cloudera DataFlow (CDF®) is used to ingest data and provide real-time analytics. Hortonworks Data Platform (HDP®) serves as the data lake and repository for the massive amount of data collected. Hadoop Distributed File System, Apache Ranger, Apache Atlas and encryption of data at rest and data in motion are used to enable data sharing, as well as security and data governance policies.

This is a hybrid deployment, with much of the workloads running in AWS GovCloud. Transient clusters will be launched on demand to process survey data and store the resulting products in the data lake.

> “The 2020 census is a massive undertaking to ensure that everyone living in the United States is counted. The Enterprise Data Lake enables us to efficiently process the vast amounts of data we anticipate generating in 2020 via massively parallel processing and taking advantage of cloud-based infrastructure. ”
>
> Kevin Smith, Chief Information Officer, U.S. Census Bureau

Massive Undertaking Improves Data Quality

The census requires years of research, planning, and development of methods and infrastructure to ensure an accurate and complete count. The Census Bureau must build an accurate address list of every housing unit, motivate people to respond, analyze the data, and release the results. Each stage requires significant data processing to organize data into actionable intelligence.

Prior to the 2020 census, all data was collected by paper survey, and then the data was transported to the U.S. Census Bureau and input manually. There was significant opportunity for error. Process changes for 2020, including reuse of administrative data, are designed to shrink the margin of error from collected data.

The public will benefit in other ways as well. Responding to census questionnaires will take less time and effort because the platform enables reuse of responses, analyzes the quality of respondent data more quickly, and enables data to be easily corrected based on administrative records, Smith notes. Redundant data collection will be reduced, while the amount of data that supports the bureau’s mission will be increased.

Data Innovation Accelerates With Enterprise Data Lake

Information security is a critical imperative in all aspects of census operations. The EDL enables security, privacy, and policy controls for all types of sensitive data and code at an enterprise level. As a result, the bureau can effectively manage and secure multiple, large datasets via automation and use metadata to monitor, link and aggregate datasets through the survey lifecycle until the final products are disseminated, Smith says.

The ability to tag datasets enables governance of data through authorization policies – controlling access to personally identifiable information – and also enables lineage tracking of the datasets, allowing a data analytics process to be repeatedly executed on a set of source datasets and producing the same result with each execution. Prior to the EDL initiative, it was difficult to know with certainty how a dataset was transformed from one state to another.

In addition, authorization policies can be created so that multiple users can access the data without making copies of the original. As a result, data scientists will be able to share data and insights more easily within the bureau and across agencies, while adhering to policies for security and data governance. Because of this new capability, the Census Bureau is able to help other agencies derive insights from the data to ensure that resources are provided to those who need them and the government can plan for future needs through insight into the patterns of population growth and change.

“The EDL supports the Census Bureau’s long-standing leadership in data analytics and technology, accelerating data innovation, realizing benefits through standardization and using cloud computing and open-source technology,” Smith observes. “In addition, the EDL reduces infrastructure costs and drives efficiencies in our business operations.”

Cloudera, Inc.   395 Page Mill Road Palo Alto, CA 94306 USA cloudera.com