RESEARCH NOTE
PROGRAM: INFRASTRUCTURE
DOCUMENT NUMBER: U212 DECEMBER 2020

NUCLEUS RESEARCH

THE VALUE OF DATA MODERNIZATION WITH CLOUDERA

ANALYST
Nick Grizzell

THE BOTTOM LINE

Nucleus evaluated several Cloudera customers to understand the value delivered from Cloudera's suite of streaming, data engineering, data warehousing, machine learning, and analytics tools. Cloudera Data Platform (CDP) is an integrated data platform built on a foundation of different open-source projects for hybrid and multi-cloud deployments. With CDP and the Cloudera Data Warehouse, customers experienced a wide range of benefits including an 88 percent reduction in script runtime, a 75 percent reduction in time spent coding, and a 47 percent reduction in the development lifecycle.

OVERVIEW

Cloudera delivers an integrated data platform that leverages different open-source projects to bring an enterprise data cloud for any data, anywhere, from the Edge to AI. The solution runs on any combination of public and private clouds and provides end-to-end data lifecycle services. Customers interviewed explored CDP and specifically Cloudera Data Warehouse services to support their digital transformation efforts.

The Cloudera Data Platform (CDP) is Cloudera's enterprise data cloud, enabling businesses to manage and secure the end-to-end data lifecycle. CDP facilitates collecting, analyzing, experimenting, and predicting data to enable critical decision-making and provide actionable insights. With enterprise data infrastructures reaching extreme levels of complexity, CDP enables businesses to extract value from large-scale, distributed, and ever-changing data. Cloudera supports a wide range of use cases, including marketing automation, customer experience management, churn prevention, and customer retention to name a few. Companies can connect through the IoT processes like predictive maintenance, business analytics, smart cities, and industrial IoT. Furthermore, Cloudera's platform can be used to protect businesses with regulatory compliance, risk modeling and analysis, fraud detection, and cybersecurity.

Cloudera continues to prepare for the complex structure of multi-cloud, and intercloud deployments with data management spread across diverse deployment environments. CDP ensures consistent data security, compliance, and control across any cloud deployment and supports the adoption of cloud-native data services for private, hybrid, and public clouds. CDP Public Cloud is connected via a Shared Data eXperience (SDX) to provide enterprise-class security and governance across the individual analytic services such as Cloudera Data Hub, Data Warehouse, and Machine Learning. Cloudera Data Hub supports Apache Spark, Hive, Impala, HBase, Phoenix, NiFi, Kafka, and Flink to allow users to streamline the management of data clusters across the data lifecycle using a form factor similar to traditional big data platform. CDP further utilizes Data Visualization, Data Engineering, and DataFlow tools to enable a scalable, real-time data pipeline and ETL process monitoring tools through a series of simplified dashboards, reports, and charts. Users can collect, report, and model enterprise data to deliver cloud-native data services in any cloud.

Cloudera Data Warehouse packages up projects already in use, such as Apache Impala and Hive, into a cloud-native service. Cloudera Data Warehouse is fully integrated with CDP, including Cloudera Machine Learning and Data Hub, enabling the delivery of multi-function analytics and machine learning in any cloud. The service runs on Kubernetes, which offers cloud optimization tools to support flexible and auto-provisioning capabilities surrounding runtimes and scalability. The Cloudera UI simplifies data warehousing processes with sectioned environments, database catalogs, and virtual warehouses. A database catalog is an instance of the Hive meta store, including references to the cloud storage where the data

is located. An environment can have multiple database catalogs. Any new catalogs are set as isolated instances of a Hive meta store if a user wants a standalone data warehouse without any data from the tables in the environment. If a user makes a change in the default database catalog, they will see those changes reflected in their environment. The virtual warehouses run queries allowing users to create a Hive or Impala cluster with standard sizes such as XSmall, Small, Medium, and Large, along with options for autoscaling to simplify users' process. Cloudera Data Warehouse's agility extends to self-service workload management tools to identify the over-utilization of resources. Cloud burst technology will move data to the cloud and return resources when peak demand is over. Tools such as HUE allow users to browse through the databases and tables in a data warehouse, as well as a full history of the queries that have been run on a specific virtual warehouse.

KEY BENEFIT AREAS

Nucleus identified the following specific benefits from the experiences of Cloudera customers. Companies realized benefits across multiple segments, including reduced project development time, reduced maintenance and monitoring times, improvements to customer insights, reduced manual processes, scalability, and enterprise-scale performance.

- **Reduced Development Lifecycle.** Without a cloud data platform like Cloudera providing connections between data and applications, developers can spend a significant amount of time setting up queries, data pipelines, database processes, and executables when implementing new solutions. The process is time-consuming and can lead to multiple errors that further compound the development lifecycle. Developing detailed reports on customer insights often requires multiple data analysts to extract benefits and insights from customer behavior. With Cloudera's capabilities, users can have truly automated databases without the need to create indices or maintain them. Analysts and managers can spend more time gaining insights into customer behavior and retrieving actual results instead of spending time managing and maintaining the solution to generate a report. Companies can have all different data sources within one data warehouse, meaning analysts and data engineers can begin looking for links between data sources and disparate business units to tie data together into a measurable and understandable report. With automated processes and integration with SparkSQL, HBase, and Kafka, one company was able to test new features in

Reduce the time to deploy new projects by 47 percent

applications and reduce the time-to-live projects by 47 percent from two months to just over one month.

- Increased Employee Efficiency. CDP Public Cloud eliminates hardware and software installation needs while also handling ongoing maintenance, management, and tuning. Additionally, the platform supports automated processes, scalability, reduced maintenance, and reduced monitoring to allow users to redirect efforts within a company to drive projects and revenues. The projects associated with extracting customer insights are often too time-consuming and require frequent reworks and updates to stay on top of continuing trends and customer behaviors. Efficiency is key for these projects, and developers need every tool to help maintain current projects and move on to new ones. The platform enabled one company to implement accelerated querying to the business structure to quickly gain insights into the current data set. Cloudera allowed a company to use an old code base with small changes to the library instead of rewriting the entire code reducing the time spent writing code for solutions by 75 percent from four days to one day. Another company reduced the time to run a script to build ZooKeeper, Apache Impala, and Accumulo databases by 88 percent, from two hours to just 15 minutes. Due to the ease of deployment, the developers and system administrators can quickly build and test applications to focus on more value-driven processes.
- Scalability. For many organizations, growth is a positive factor but can increase costs surrounding employees, infrastructures, and maintenance. As the company expands to new landscapes, users will adopt more applications and data management tools, leading to an increase in workflow processes. For enterprise-level organizations, data integration and migration are extreme tasks considering the landscape of platforms, applications, and environments used. Additionally, enterprise-level companies can deploy new use cases quickly to new cloud environments. Cloudera provides a central scalable, flexible, secure environment for handling workloads from batch, interactive, to real-time analytics. Users can deploy Cloudera to facilitate growth through marketing automation, churn prevention, and customer retention to deliver scalable storage and distributed computing.

Reduce the time spent writing code for solutions by 75 percent

CUSTOMER PROFILES

FINANCIAL SERVICES COMPANY

The company is a small investment bank centered around advisory-based financial services on behalf of individuals and corporations. Further functions surround arranging debt financing for corporations by finding large-scale investors for corporate bonds and examining their financial statements for accuracy.

The company deployed a legacy solution to store large amounts of data and process the data in real-time. The legacy solution lacked internal performance and integration capabilities for external solutions. For example, the legacy solution lacked support for SparkSQL and HBase, which were crucial for the company's operational structure.

With the overall complexity of the data architecture increasing, the company looked towards Cloudera Data Warehouse to address its need for a modernized solution. Immediately after deployment, the company noted the ease of use and setup as engineers could quickly diagnose problems with clusters through the GUI. Furthermore, Cloudera offered integration with SparkSQL and HBase to address the needs of the banking solutions along with Apache Kafka integration to help test new features in applications and reduce the time-to-live of projects by 47 percent from two months to just over one month. Additionally, Cloudera Data Warehouse allowed the company to use old code base with small changes to the library as opposed to rewriting the entire code reducing the time spent writing code for solutions by 75 percent from four days to one day.

INFORMATION TECHNOLOGY SERVICES COMPANY

Reduce the time to run a script by 88 percent

The company is a small information technology services company focused on providing operational structure consulting in government and private sectors through proven proprietary methodologies. The company performs short-term cybersecurity assessment as well as manage a fully outsourced business process over a period of multiple years. Further functions include software development and systems integration to support agile methodologies and connect existing systems with modern technologies.

The company uses Cloudera Manager to rapidly deploy ZooKeeper, Apache Impala, and Accumulo clusters via a Python script. With the combination of these capabilities, the company can enable its customers to easily monitor, deploy, and manage cloud services using a simplified web UI. Cloudera's easy-to-use web UI offers full capabilities surrounding starting and stopping clusters and services and monitoring tools to maintain the cluster,

services, and physical host hardware as well. With the simplified web UI, the company does not need to worry about enabling its non-technical customers as most are comfortable with the platform. APIs allowed the company to create scripts to automate deployment processes, accelerating the time-to-live of projects. Cloudera Manager saves the company time and effort by having an automated script to deploy solutions throughout the existing environment. The company reduced the time to run a script to build ZooKeeper, Apache Impala, and Accumulo database by 88 percent from two hours to just 15 minutes. Due to the ease of deployment, the developers and system administrators can quickly build and test applications allowing them to focus on more value-driven processes.

LOOKING AHEAD

Growth in data volumes, diversity, and velocity is exponentially increasing as we approach the limits of existing information management infrastructures. Companies are still looking towards traditional hardware-based solutions requiring costly upgrades for databases and data warehouses. The additional cost of a legacy data architecture is increasing costs and limiting growth by stunting scalability when dealing with hundreds of terabytes of data. Furthermore, the relational system's incompatibility with unstructured data only further compounds the complexity of traditional database and data warehouse solutions.

As more companies look to modernize the existing infrastructure, upper-level management will look towards platforms like Cloudera to implement innovative approaches to handling growth in both big transaction data (data warehouses, databases, and enterprise applications) and big interaction data (social media, mobile devices, and website data). Cloudera proves to be a solution for integrating traditional structured, multi-structured, and unstructured data to gain insights without sacrificing overall capabilities and performance. Cloudera will help companies make sense of data that drives modern organizations by discovering various patterns and uncovering the connection between critical business processes. Cloudera can handle massive data growth and produce insights as a company continues to grow, which will prove to be essential across the Big Data industry.