Just add data: Cazena serves up an 'instant' cloud data lake

SEPTEMBER 16 2020

By James Curtis

The company's data lake service is getting a makeover, with an emphasis on time – the time it takes to set up a data lake and the time it takes to start analyzing data. Instant Cloud Data Lake is targeted at enterprises wanting a cloud-based analytics platform that requires minimal resources to manage.

451 Research®
Now a Part of

S&P Global Market Intelligence

Introduction

Cazena was founded in 2014 by individuals from the former data warehousing appliance upstart Netezza. From the beginning, the company's goal has been to help enterprises reap the benefits of big-data processing and analytics without needing in-house expertise to configure and manage the various products. As such, Cazena's focus has been on delivering a managed data lake as a service, leveraging an array of technologies from open source software to the company's proprietary tools. Cazena is refining its data lake service with what it is calling the Instant Cloud Data Lake, a type of turn-key service targeted at enterprises wanting a cloud-based analytics platform that requires minimal, if any, resources to manage.

451 TAKE

The concept of the data lake has certainly evolved over the past 10 years. Where data lakes often consist of multiple open source products that are used to store raw data from multiple sources (including unstructured data), data warehousing systems are codified systems with defined data schemas. But to operate and manage a data lake requires an ecosystem of products and tools. This ecosystem has its benefits, but it also adds complexity to managing a data lake. It's precisely this that gets to the heart of Cazena's new Instant Cloud Data Lake service. Cazena's managed data lake service is meant to address the time it takes to stand up a data lake and also the complexity of managing the environment. As such, Cazena recently launched its service along with expanding its existing relationship with AWS by embedding analytics services (EMR, Athena, Glue, Redshift and SageMaker, for instance), thus driving the company's broader vision to create a flexible data lake platform allowing third-party and other products to be added to the platform. That platform vision is worth noting, particularly for enterprises that are considering a data lake or have struggled to manage their current data lake.

Details

Cazena's new Instant Cloud Data Lake service is considered an evolution, or generation 2 version, of the company's existing, managed cloud service that is available on AWS and Azure. There is a deliberate emphasis on 'instant' because Cazena wants to greatly simplify the data lake deployment model and plans to lean on its historical roots to accomplish this. Where Netezza brought the appliance model to data warehousing, Cazena wants to bring a similar mentality to the deployment and management of data lakes.

The driving force behind the new service is based on some trends and perceived gaps in the market, Cazena has observed. These challenges include the length of time it takes to deploy a data lake and getting it into production. Costs can be another factor, particularly in finding not only skilled resources to manage but also to find developers to create applications that sit atop the data lake. There is also the trend that data lakes are moving away from on-premises deployments to cloud environments leveraging cloud object storage.

This last point is supported by research. According to 451 Research's Voice of the Enterprise survey data (see figure below), object storage will be a primary landing spot for data processing and analytics over the next two years. In fact, 71% of respondents 'completely' or 'mostly' agree, while 29% 'mostly' or 'completely' disagree.

From an architectural perspective, the Instant Cloud Data Lake service includes proprietary and patented parts to drive the service for so-called 'faster analytics consumption.' More specifically, there is an orchestration component that manages the service's underlying hardware provisioning, such as compute and storage resources. Security, identity management, data governance and compliance are also handled by the orchestration layer. The core of the system provides self-service analytics capabilities that are provided by an updated analytics console. The console supports use cases for BI and data engineering/ETL, as well as for machine learning (ML) workloads. Depending on the workload, users can leverage a variety of processing engines, including Spark, Presto, Hive and search. Data gets ingested into the platform via streaming or batch load, whether from the cloud or on-premises systems and gets persisted to cloud object storage. The final component is the service operations layer. Because the service is fully managed, cloud operations, DevOps and support are all handled by the operations layer, which enables Cazena to provide SLA guarantees without administrative efforts from users.

Where Cazena differentiates itself from other data lake services is its focus on the time it takes to stand up a data lake (hence, 'instant') and the time it takes to manage the data lake. Cazena has historically relied on partnerships and integrations with third-party/open source products and this will continue to be its strategy. For instance, from the service analytics console, there are integrations with the likes of DataRobot, ThoughtSpot and Amazon SageMaker.

Object Storage as Primary Location for Data Processing and Analytics, in Next Two Years

Source: 451 Research's Voice of the Enterprise: Data & Analytics 2H 2019.

Object storage will be a primary data platform for data processing and analytics at my organization in the next two years (n = 482)

23%	48%	21%	8%

■ Completely Agree ■ Mostly Agree ■ Mostly Disagree ■ Completely Disagree