REPORT REPRINT

Better together: Cloudera SDX data governance layer leverages strengths from Hortonworks

NOVEMBER 14 2019

By Paige Bartley

Enhanced governance and security have always been benefits of pursuing commercial support for open source. But when Cloudera acquired Hortonworks, one question was how their governance initiatives would be combined. With Cloudera Data Platform now available, the SDX governance layer borrows strengths from both companies and their respective open source projects.

Introduction

With the general availability of Cloudera Data Platform (CDP) announced in late September, Cloudera made it clear that it was combining the strengths of the former Hortonworks technology assets into the revamped offering. Given the historical parallels between these former competitors, there was significant overlap in some of the technological capabilities and open source projects that the respective organizations had pursued over time, and the general availability announcement of the new platform was an opportunity to clarify the company's vision for data governance functionality. Cloudera SDX, first established in 2017 and short for 'shared data experience,' will merge with Hortonwork's DataPlane Services to be the official branding and presence of the combined company's governance abstraction layer moving forward, although the layer itself borrows key elements and strengths from the offerings and projects of both organizations.

Cloudera SDX meshes closely with the Cloudera Data Catalog – fleshed out from the technology assets of the former Hortonworks Data Steward Studio – to provide navigability and governance control for data across environments, whether that be multi-cloud, hybrid or on-premises. So the SDX layer is poised as an enabler for a broad spectrum of users and use cases, ranging from compliance control to data exploration.

451 TAKE

Data governance and security requirements have always been a major motivator for the adoption of commercial support for open source projects, as was the case with Hadoop – at least initially – with both Cloudera and Hortonworks. SDX and DataPlane Services were introduced from each company, respectively, to address the growing complexity of unified data governance across diversifying cloud and hybrid data environments; so when the brands joined forces, one of the biggest lingering questions was how they would combine their parallel investments in governance functionality.

The current iteration of SDX did have to choose 'winners' and 'losers' with overlapping functionality, potentially creating tension for long-time customers as they progress on their upgrade paths toward CDP. However, Cloudera posits these decisions will ultimately facilitate a governed ecosystem where personas across IT and lines of business can get the custom compute environments they need without straying toward unmanaged solutions. The challenge in fending off shadow IT, of course, will be full end-user adoption of the platform: something that remains to be seen as Cloudera continues to flesh out functionality for less-technical personas such as data stewards and compliance teams.

Context

Data governance is increasingly viewed as an enabler of business value by organizations rather than a cost center; according to 451 Research's Voice of the Enterprise, Data & Analytics, 2H 2019 survey results, 71% of respondents 'completely' or 'mostly' agreed that data governance is seen as an enabler of business value within their organization. But for governance to become an enabler rather than a burden, it needs to be made accessible to multiple stakeholders and seamlessly effective across data environments. This is exactly what Cloudera is aiming to do with its SDX layer.

When SDX was first announced in 2017 by Cloudera prior to the merger with Hortonworks, it was positioned as a unified governance layer for uniformly managing data across clusters and environments to enable a single point of control for data being used in various workloads and use cases. DataPlane Services, introduced at nearly the same time by Hortonworks, was similarly positioned, but with an emphasis on extensibility and tools for business persona involvement in governance, typified by the Data Steward Studio offering. Both SDX and DataPlane Services integrated with their respective vendors' endorsed Apache projects related to security: Sentry and Ranger, respectively. With the combination of the companies, the challenge was deciding how to merge these technology assets into a cohesive governance layer that would enable both IT and line-of-business users to work together, seamlessly leveraging data without departmental or role friction.

As Cloudera advances deeper into supporting MLOps with its newly announced Cloudera Machine Learning managed offering, the importance of SDX will only increase because data science workflows depend on governed and high-quality data sources. However, well-managed data sources alone are not enough; they need to be navigable by a variety of enterprise stakeholders. With an architectural rebuild and re-implementation of SDX, including a refined data catalog, the Cloudera data ecosystem aims to make data across clouds and other environments centrally accessible by those who need it, facilitating collaboration.

The announcement of CDP in late September, in conjunction with the Strata Data Conference in New York City, largely resolved the details around combination of Cloudera and Hortonworks technology assets, and fleshed out SDX as a concrete architectural component borrowing elements both from the original respective offerings from each company. Some of the key details are as follows.

Developments

Integration of governance and security: standardizing on Apache Ranger

For Cloudera, some difficult decisions needed to be made regarding 'winners' and 'losers' when there was obvious overlap with Hortonworks functionality. And security integration with SDX, for use cases such as managing role-based (as well as attribute-based) access to data, is a critical piece of the data governance puzzle, especially under modern data privacy and protection mandates such as GDPR and CCPA. Originally, Cloudera and Hortonworks as separate business entities had each latched on to separate Apache projects – Sentry and Ranger, respectively – to help fulfill these needs. However, these two projects were very similar in their objectives and capabilities. Cloudera today has made the decision to standardize on Apache Ranger for security functionality: the project that arguably had slightly more advanced and fine-grained controls for data. Cloudera suggests that although the initial switchover to Ranger for past users of Sentry may pose a minor organizational speedbump, there will be a tool available to assist in the process, and businesses may find they need fewer policies in Ranger to accomplish the same things that were formerly done in Sentry. Centralized administration of security policies across clusters is the goal, with the ability to not only control for role-based access at the column level, but also to set controls for time-based rules and data masking.

Extensibility of SDX: incorporating the adaptability of Hortonworks DataPlane Services

When Hortonworks first introduced DataPlane Services as its unified governance layer, also in 2017 – almost simultaneously with then-competing Cloudera's announcement of SDX governance functionality – the differentiation emphasis was on extensibility and governance support of Hortonworks' streaming data capabilities. DataPlane Services was designed to grow and adapt over time, supporting different use cases and functions. Now that Cloudera has absorbed the technology assets associated with Hortonworks DataFlow for streaming data, the updated Cloudera SDX governance layer is incorporating the highlights of the DataPlane Services offering: most notably its extensibility, points of entry and friendly UX for line-of-business users, as well as integration with streaming data management functionality.

Cloudera Navigator and Hortonworks' Apache Atlas: charting a new course

In principle, prior to the acquisition by Cloudera, Hortonworks was philosophically more of an open source purist. While gentle ribbing regarding the number of Apache 'zoo animals' that Hortonworks wrangled at any given time may have become a trope, the fact remained that the company tended to support open source projects for functionality whenever possible. Cloudera, on the other hand, initially supported more of an 'open core' model that used proprietary software as a governance and management wrapper for open source features. So when it came to needs such as data lineage and metadata, the two companies diverged somewhat, as typified by their respective supported projects and/or offerings: the proprietary Cloudera Navigator and Apache Atlas as supported by Hortonworks. Today, with Cloudera adopting a pure open source model, the two philosophies are being merged into the singular SDX layer: picking the strengths of each respective product and project. For instance, the new Cloudera Data Catalog borrows significant aesthetic and UX elements from Cloudera Navigator, making it more easily accessible to business personas – such as data stewards and compliance professionals – that are invested in the management and control of data. However, its underlying technical capabilities for handling metadata in a standardized way are very much inherited from the open source Apache Atlas, benefiting from the standards – and capabilities for real-time capture of metadata -- that it established.

A modern multi-cloud and hybrid data catalog: borrowing functionality from Hortonworks

The Cloudera Data Catalog associated with CDP aims to be an enterprise-wide catalog that can establish a unified metadata layer across diverse multi-cloud and hybrid environments. Architecturally, within CDP, the data catalog sits within the Cloudera Control Plane, which runs vertical to the CDP stack and offers single-view capabilities for multiple management and navigation purposes. The company has emphasized that the data catalog is a core part of strategy; without the ability to evaluate the attributes of data and understand context, it is impossible to assign levels of accuracy or gauge eventual output; the catalog is the portal that will allow users to navigate and interpret these informational resources. From a user experience perspective, the catalog front end looks a lot like the former Cloudera Navigator product, aiding in usability; however, under the hood, it has been merged with the technology of Hortonworks Data Steward Studio to provide richer governance capabilities for data. Users with appropriate permissions can pattern match, identify, profile and detect certain data types: creating policies for sensitive or other relevant types of data. Profiling, in fact, can even be done automatically without dependence on individual users. Additionally, because Data Steward Studio profilers will match with IBM's StoredIQ, the governance reach is further increased via business partnership. The merging of investments continues with Apache Atlas' functionality, as nearly anything in Atlas – glossary, taxonomy – can be leveraged in the catalog, and integration with Ranger allows for continuous policy enforcement.

Competition

Prior to Cloudera and Hortonworks joining forces, MapR was the third major player in the commercial Hadoop distribution competitive landscape. These days, the company has since been acquired by Hewlett Packard Enterprise, ostensibly to bolster its Intelligent Data Platform. So while the MapR assets might nominally compete with Cloudera from a technological perspective, Cloudera's partnership with HPE makes this overt rivalry less likely. What is more likely is potential overlap and competition between the respective companies' AI/ML offerings.

A number of newer software providers aim to tackle the challenge of administering role-based and policy-based access controls for data, much like Ranger, to ensure the right people have the right access to the right data at the right times. Immuta, Okera and PlainID all offer central controls for administering access to data that may reside in diverse repositories and locations. With an emphasis on facilitating self-service and remaining invisible to the business end user consuming the data, these products aim to accelerate initiatives such as data science and analytics while ensuring control for use cases such as data privacy and compliance.

The data catalog market, and closely related governance market, is sprawling with enormous overlap. Perhaps most relevant are the largest providers with cloud ecosystems and accompanying catalog/governance capabilities: Amazon, IBM, Google, Microsoft and Oracle – although Cloudera's close partnership with IBM makes direct competitive pressure less likely in this scenario. Data 'catalog' definitions may vary by vendor, but all cloud providers listed offer a metadata layer for their cloud, and provide ways to navigate and exert (at least basic) governance controls on the data held there.

A thriving pack of governance and data catalog providers exists outside of the public cloud providers. Informatica, with its Enterprise Data Catalog (EDC) and Axon product, is a good example of a company providing unified navigation and governance controls for data spread across multi-cloud and hybrid environments. ASG Technologies provides rich data lineage, catalog and metadata management capabilities in its Data Intelligence offering. Alation, Collibra and Waterline Data are all known primarily for their catalogs. Analytics platform and data prep specialist Alteryx offers a data catalog as part of its Alteryx Connect data management and collaboration module, as well as its planned Server.NEXT offering. Likewise, Qlik has data catalog functionality in its analytics ecosystem via its Podium Data assets. Unifi Software, focused on self-service management and enablement, recently broke off its data catalog into an optional stand-alone offering. And Zaloni, a data lake management provider, also has a catalog, potentially putting it in competition with Cloudera.

SWOT Analysis

STRENGTHS

Cloudera has taken notable strides in combining the best governance and security features formerly associated with the Hortonworks brand; not a trivial feat, given the scope of the product portfolios. A focus on improved ease of use and catered user experiences for relevant enterprise stakeholders and personas emphasizes the collaborative nature of modern data consumption and governance.

WEAKNESSES

Although organizations are embracing multi-cloud and hybrid strategy, there are several other vendors eager to provide security, governance and particularly catalog functionality across these environments. Cloudera is somewhat late to the data catalog game and will face an uphill battle against providers leveraging a departmental 'land and expand' strategy that appeals directly to data-hungry business end users.

OPPORTUNITIES

Proliferating data protection and privacy regulations, such as GDPR and CCPA, are putting pressure on organizations to control data across multi-cloud and hybrid environments. Businesses are additionally starting to genuinely view data governance as an enabler of business value. If Cloudera can demonstrate that its capabilities can help accelerate both 'reactive' compliance efforts and 'proactive' data-driven initiatives, it will have a strong value proposition.

THREATS

The major hyperscalers – AWS, Google and Microsoft – all have incentive to provide tools for businesses to easily manage data natively in their cloud environments, and some organizations may decide that these features are 'good enough' for all intents and purposes, especially if the pricing is highly competitive. With a growing web of business partnerships that provide overlapping governance functionality, Cloudera needs to carefully manage relationships to avoid 'frenemy' situations.