Case Study

Data Discovery for Data Science

CASK

Challenge

Once an organization has established a process for ingesting data into a data lake, there must be an easy way discover, inspect and track datasets within the data lake. One organization's data scientists and business analysts found it difficult to locate the datasets in their data lake and discover them based on business or technical metadata. This led to a lengthy process of having to regenerate datasets and redo the same work over and over, thereby wasting time, which required additional costly cluster capacity as well as staff resources. Some of the additional issues were:

- The time it took to discover datasets needed for ideation was days or even weeks
- Due to the difficulty of locating datasets, cluster resources were over-utilized to re-create datasets already present on the cluster
- Multiple instances of the same datasets created confusion about the most recent and authoritative copy
- IT spending was unpredictable as more data scientists and analysts were added to the team

With the Cask Data Application Platform (CDAP) and Cask Tracker, all datasets generated were tracked and indexed by both technical and business metadata, allowing data scientists and data analysts to discover all datasets and inspect tags, schema and properties. They were then able to use the automatic lineage tracking capabilities to determine the source of the dataset and understand the transformations applied. Operational metadata helped the organization identify the type and frequency of processing applied on the datasets, in addition to answering the question who performed it; also, the audit data captured helped determine the freshness and activity-level of the datasets.

Use Case: Data Lake

Benefits

Rapid Time to Value

Time to discovering datasets on a data lake was reduced from days or weeks to minutes or hours.

Seamless integration between Cask Tracker and Cask Hydrator took the company from the data discovery phase to ideation and pipeline creation in minutes, which previously required hours and sometimes days.

Lower Cost of Ownership

Having an easier method to discover datasets in the data lake lowered the utilization of cluster resources in terms of compute and storage.

Ultimately IT spending became much more predictable.

Business Empowered with Self-Service Analytics

Lineage and audit capabilities allowed users to obtain authoritative answers to source, transformation and freshness of data, increasing transparency and their trust in the quality and nature of datasets.

Collaboration between data engineers, scientists and analysts improved.

About CDAP

The first unified integration platform for big data, Cask Data Application Platform (CDAP) lets developers, architects and data scientists focus on applications and insights rather than infrastructure and integration. CDAP accelerates time to value from Hadoop through standardized APIs, configurable templates and visual interfaces. It enables IT organizations to broaden the big data user base within the enterprise with a radically simplified developer experience and a code-free self-service environment. CDAP is 100% open source, and along with its extensions Cask Hydrator for data pipelines and Cask Tracker for data discovery and metadata, it seamlessly integrates with existing MDM, BI and security and governance solutions.

About Cask

Cask makes building and running big data solutions on-premise or in the cloud easy with Cask Data Application Platform (CDAP), the first unified integration platform for big data. CDAP reduces the time to production for data lakes and data applications by 80%, empowering the business to make better decisions faster. Cask customers and partners include AT&T, Cloudera, Ericsson, Lotame, Salesforce, and Tableau, among others. For more information, visit the Cask website at cask.co and follow @caskdata.