Case Study

Data Lake

Challenge

Building an enterprise data lake requires a reliable, repeatable and fully operational data management system, which includes ingestion, transformations, and distribution of data. It must support varied data types and formats, and it must be capable of capturing the data flow in various ways. The system must do the following:

- Transform, normalize, harmonize, partition, filter and join data
- Interface with anonymization and encryption services external to the cluster
- Generate metadata for all data feeds, snapshots and datasets ingested, and make it accessible through APIs and web services
- Perform policy enforcement for all ingested and processed data feeds
- Track and isolate errors during processing
- Perform incremental processing of data being ingested
- Reprocess data in case of failures and errors
- Apply retention policies on ingested and processed datasets
- Setup common location format (CLF) for storing staging, compressed, encrypted and processed data
- Filter views over processed datasets
- Monitor, report and alert based on thresholds for transport and data quality issues experienced during ingestion. This helps provide the highest quality of data for analytics needs
- Annotate datasets with business/user metadata
- Search datasets using metadata
- Search datasets based on schema field names and types
- Manage data provenance (lineage) as data is processed/transformed in the data lake

Use Case: Data Lake

Benefits

Lower barrier to entry to Hadoop

The company's non-Hadoop developers were able to build an end-to-end data ingestion system without training, saving time and resources.

Rapid Time to Value

Developers were able to build the data lake and get it to customers faster.

CDAP's ingestion platform standardized and created conventions for how data is ingested, transformed and stored, allowing faster on-boarding.

Business Agility

Developers provided a self-service platform for the rest of the organization, enabling departments to use data to make better business decisions.

Scalability

CDAP was installed in eight clusters with hundreds of nodes.

Using Cask Tracker, data lake users were able to quickly locate and access datasets and metadata, data lineage and data provenance. This allowed them to efficiently utilize their clusters, aided them in data governance and auditability and improved data quality.

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.