Cask Hydrator

SELF-SERVICE DATA LAKES
ON HADOOP
Introduction

A recent Gartner survey on Hadoop cited the two biggest challenges in working with Hadoop: “Skills gaps continue to be a major adoption inhibitor for 57% of respondents, while deciding how to get value from Hadoop was cited by 49% of respondents.” A recent study by the consulting firm, PwC supported the second point, “There is a clear business need to unlock the value tied up in information.”

In spite of these objections, enterprises increasingly need to ingest high volumes of data from a wide variety of both structured and unstructured sources in high velocity environments. When we talk with enterprise data project managers, they say, “Our biggest challenge is getting the data from the source into an application.” It doesn’t matter whether they are getting data from a batch process or a streaming source. We hear it over and over again, “I have a source (or sources) of data and I need to get it into a data lake so we can get to the value.”

When an organization decides to roll out Hadoop, they face additional challenges when creating solutions on Hadoop. In a typical scenario, an enterprise must engage costly Hadoop engineers to build data pipelines that have a high degree of complexity. The multiple technologies used in these projects add another layer of complexity, which typically equates to increased costs, longer time-to-production, and a high cost of maintenance. As a result, these pipelines can be difficult to test, often require a great deal of maintenance, and demand significant tuning to achieve performance goals, all of which lead to a high TCO.

Ideally, a solution would sit on top of Hadoop that would allow users to quickly drag-and-drop the appropriate components to build and test data pipelines. Such a solution would translate complex, loosely connected technologies, programming and scripts into easy to build and maintain data pipelines on Hadoop. As a result, the skills gap and business need objections would be obviated, and expensive and rare engineering resources would no longer be necessary to build data pipelines on Hadoop. Such a solution would also reduce the time-to-production from months to weeks, thereby cutting costs and delivering business value faster.

Also known as Cask Hydrator, this solution is an extension to the open source Cask Data Application Platform (CDAP). The solution simplifies the process of developing, running, automating, and operating data pipelines on Hadoop. As a result, Cask Hydrator allows users to rapidly build and run streamlined data refineries to support many use-cases.

Cask Hydrator empowers organizations to:

- Blend, transform, and analyze data across multiple sources to create an on-demand analytical view across key customer interactions. This provides powerful insights into buyers, brand performance, products, and services.
- Target high-value clients with personalized promotions and custom recommendations by aggregating data from multiple data stores and applying data science.
- Reduce the strain on the data warehouse by offloading simple use-cases, less frequently used data, and its corresponding transformation workloads to Hadoop.


“Seizing the information advantage: How Organizations can unlock value and insight from the information they hold,” published by PwC, September 1, 2015, p. 2.
Self-Service

Hydrator provides easy-to-use interfaces and integrated capabilities that enable a broad range of users to solve the challenges of transforming, aggregating, and enriching both large scale and varied data without involvement of infrastructure experts. It lets you deliver governed, managed, on-demand data to data scientists, data engineers, analysts, and end-users in an agile fashion. Hydrator empowers more users in the enterprise to apply the power of data to their roles. Other capabilities that drive self-service include:

- Empowering users to architect big data pipelines to create complete and accurate analytical solutions.
- A drag-and-drop graphical user interface that provides visual data pipeline creation, which eliminates manual programming and scripting from the process.
- Automatic creation and publishing of datasets to drive faster and more reliable analytics.
- Consistent support for batch and real-time data pipelines.
- Minimal support needed from IT for organizations and business users with reliable, repeatable, governed data pipelines.
Comprehensive

Hydrator includes the necessary capabilities to deliver end-to-end data ingestion and processing pipelines in production. With the ability to extend using plug-ins, Hydrator can access any data type and perform any type of processing. Separation of pipeline creation from execution enable the use of any processing technology, including MapReduce, Spark or Spark Streaming. Additional features include:

- A rich library of pre-built Plugins to access, transform, blend, and aggregate data from relational sources, big data sources, and more.
- Integrations with advanced analytics tools like Spark ML to operationalize predictive intelligence while reducing the build time.
- Powerful orchestration capabilities that coordinate and combine transformations, including notifications and alerts.
- An integrated enterprise scheduler that coordinates workflows and a debugger for testing and tuning job executions.
- Seamless integration with visualization and data services, making datasets immediately available for reports and other applications.
- Robust support for relational data sources, NoSQL data stores, cloud services and others.

Enterprise Ready

Hydrator goes beyond data ingestion and achieves scalable and flexible management of end-to-end data pipelines for the enterprise. Dynamic containers or templates are reusable allowing users to leverage them time and time again to dramatically reduce time and effort spent on subsequent pipeline projects. The application is both distribution and deployment agnostic; it enables users to move data and applications from on-premises to the cloud and back, changing to a different distribution of Hadoop, or switching technologies for running data pipelines. Other key advantages for the enterprise are:

- Robust administration features including SLA monitoring, job restart, error handling and restart, and centralized operations to provide auditing access.
- Enterprise-grade security that includes access and version controls as well as LDAP, JASPI, and Active Directory integration.
- Enhanced Data Management through integration with other extensions like Cask Tracker to track data and Metadata at all times.
- Processing and consistency guarantees allow users to architect big data pipelines with complete and accurate analytical results.
- Best-in-class enterprise grade customer support and rapid user time-to-train.
Conclusion

Data Pipelines in Hadoop -- Simple to Build, Simple to Run, Simple to Operate

In its simplest form, a data pipeline built in Cask Hydrator’s visual drag-and-drop interface graphically represents the source, one or more transforms, and a sink. Once defined, validated, and saved, it can run as an active program. If built to ingest data from a streaming source, it sits and waits, perhaps for a millisecond, until a transaction appears in the source. It immediately applies all transforms and then loads that transaction to the sink. The pipeline then continues to the next transaction with no intervention from a human operator. If built to process batch data, it can run immediately or be scheduled as an ongoing process in line with the demands from Operations.

The extensible Cask Hydrator, an open source extension to CDAP, enables the construction of ETL ingestion pipelines that handle batch or streaming data to solve real business challenges. A number of typical examples of how customers are using Hydrator are in the use cases below. In every one, a legacy pipeline used complex technologies, custom Hadoop APIs, and expensive specialized programmers. In each case, the performance of the legacy pipeline lagged in one or more of these areas: missed the target SLAs significantly; required costly ongoing maintenance; required frequent restarts due to failures; lacked the ability to test and validate the pipeline; or failed to separate the pipeline development process from operations.

Hydrator empowered in-house Java developers or those with no programming experience to build complex batch or streaming pipelines rapidly, bridging the skills gap. They only required a few hours of training in all of the use cases cited. The pipelines were constructed, tested, and validated, then operationalized within days or weeks versus the legacy pipelines which had required many months of development.
Use-cases

**Data Lake**

Building an enterprise data lake requires a reliable, repeatable and fully operational data management system, which includes ingestion, transformations, and data distribution. It must support varied data types and formats, while capturing the data flow in multiple ways. The system must provide multiple capabilities from data ingestion through error and failure tracking to critical post-ingestion steps.

During the initial data ingestion process, the system needs to transform, normalize, harmonize, partition, filter and join data. Additional key functionality during this step includes:

- 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 webservice.
- Enforce policies for all ingested and processed data feeds.
- Perform incremental processing of data being ingested.

Robust error and failure tracking capabilities during ingestion and analytical processes insure high quality and high reliability results. The system must:
- Track and isolate errors during processing.
- Reprocess data in the event of failures and errors.
- Apply retention policies on ingested and processed datasets.
- Setup common location format (CLF) for storing and staging compressed, encrypted and processed data.

Once the data ingestion process is complete, the management system needs to make it easy for users to run experiments and create new insights from their data. It needs to:

- Provide user-determined filtered views of processed datasets.
- Continue to monitor, report, and alert based on thresholds for transport and data quality issues experienced during ingestion. This delivers the highest quality data to meet analytical requirements.
- Annotate Datasets with business/user metadata.
- Search Datasets using metadata, schema field names, or types.
- Manage data provenance (lineage) as data is processed/transformed in the data lake.

**The Cask Hydrator Solution** — The team of Java developers, all non-Hadoop engineers, built an end-to-end, self-service ingestion platform using Hydrator in conjunction with CDAP. This enabled the rest of the organization to ingest, process, and catalog data. The solution delivers native support for incremental processing, reprocessing, tracking metadata, workflow, retention, snapshotting, monitoring, and reporting. As a result, Hydrator dramatically accelerated their time-to-market. The ingestion platform also standardized and created conventions for how to ingest, transform and store data on the cluster.

**Results**

- Platform users were able to on-board at a much faster rate than previously, delivering rapid time-to-value.
- CDAP is installed in 8 clusters with 100s of nodes.
- Data Lake users were able to locate Datasets faster and could more quickly access metadata, data lineage and data provenance. This allowed them to efficiently utilize their clusters and also aided them in data governance, auditability, and improving the data quality of Datasets. CDAP Tracker provided this set of capabilities.
Information Security Analytics and Reporting

The Challenge — The customer, a Fortune 50 financial institution, created a pipeline that aggregates batched data onto a secured Hadoop cluster to create daily aggregates and reports. The current system performed multiple transformations, which created new datasets. The customer faced multiple issues:

- The data pipeline was inefficient, took 6 hours to run, and required manual intervention almost on daily basis.
- Reports were not aligning correctly with day boundaries.
- Any points of failure required reconfiguring and restarting the pipeline, a time-consuming and frustrating task.
- The pipeline required major setup and development time to add new sources.
- The team could not test and validate the pipeline prior to deployment. As a result, testing was conducted directly on the cluster, an inappropriate use of resources.

The Cask Hydrator Solution — The customer’s data development team created independent parallel pipelines that moved the data from SQL Servers into Time Partitioned Datasets. Transformations were performed in-flight with the ability to handle error records. After completing the initial transfers, another pipeline combined the data into a single Time Partitioned Dataset and fed it into an aggregation and reporting pipeline.

Results

- In-house Java developers with limited knowledge of Hadoop built and ran the complex pipelines at scale within two weeks following four (4) hours of training.
- The new process reduced complexity and simplified pipeline management, now requiring only two (2) hours to run without any manual intervention.
- The visual interface enabled the team to develop, test, debug, deploy, run, automate, and view pipelines during operations. Simultaneously, the new pipeline reduced inappropriate utilization of the cluster.
- Transforms were performed in-flight with the ability to handle error records.
- Tracking tools made it easy to rerun the process from any point of failure.
In-Flight Brand Sentiment Analysis of the Full Twitter Firehose

The Challenge — A Fortune 500 company in the E-Commerce sector built a data pipeline that ingested their Twitter stream in real-time. The data was cleansed and transformed prior to conducting multi-dimensional aggregation and sentiment analysis on the marketing campaign based on Tweets. The results were updated twice daily to HBase. However, the legacy pipeline suffered on two fronts: first, latency in the existing pipeline delayed the decision making process due to the velocity of data (5000+ tweets/sec). Second, the existing data movement process proved to be costly in time and money.

The Cask Hydrator Solution — The company’s in-house team of Java developers built a real-time pipeline in two weeks using the drag-n-drop visual interface in Cask Hydrator. They developed a sentiment analysis transform using the API and then included it in the pipeline. Further, they added multidimensional aggregations without needing code creation using the Cube Plugin.

Results

- The analysis of Tweets in real-time allowed the business to make faster decisions on their campaigns.
- The new pipeline eliminated latency between aggregation and the availability of results, producing quicker and better automated decisions while cutting costs.
- The new pipeline cleansed, transformed, analyzed and aggregated tweets at the rate of the full Twitter firehose.
- Consolidated the infrastructure into a single Hadoop cluster.
- In-house Java developers were able to build the pipeline and sentiment analysis plugin with a 4-hour learning curve.
- Seamless transparency through the custom dashboards provided easy operational insights and aggregated logs for debugging.
Encrypted and Data Masking

The Challenge — The customer, a Fortune 50 company in the Telecom sector, developed a legacy custom data pipeline that performed format-preserving encryption and data masking on a KerberOS Hadoop cluster. The pipeline extracted data from Teradata to HDFS, performed transformations, and loaded the results back into Teradata on a daily basis. This pipeline, built by a third-party service, was operationally unstable and required constant, costly intervention to keep it running.

The Cask Hydrator Solution — The self-service, code-free interface allowed the in-house team to reproduce and replace the existing pipeline. The new process performed the extraction, encryption, masking and reload to and from Teradata in-flight. It created a copy of the data on HDFS so the team could run complex ad-hoc queries using Hive.

Results

- Using the code-free drag-n-drop visual interface, the in-house team built the pipeline in 5 days.
- They were able to easily achieve scale with Hydrator to monitor and achieve their SLAs.
- IT gained immediate insights into the performance of the data pipeline and were able to easily handle failure scenarios.
- More complex Ad-Hoc Queries were offloaded from Teradata, thereby reducing the overall cost.
Data Cleansing and Validating 3 Billion Records

The Challenge — The customer, a Fortune 500 company in the Financial sector, had custom-built a data pipeline to perform data validation and correction transforms. The pipeline was constructed using multiple complex technologies. Example transforms performed on the 3 billion records included:

- Standardization, verification and cleansing of USPS codes
- Domain set validation, null checks, length checks
- Regular expression validation (validate email, SSN, dates, etc.)

The legacy pipeline ran overnight, required multiple teams to keep it operating, and costly experts to maintain it.

The Cask Hydrator Solution — In-house Java programmers developed, tested, and ran the replacement pipeline using the drag-n-drop visual interface. The new pipeline only required limited coding other than configuring the plugins in the pipeline.

Results

- The in-house team built, tested, and deployed the pipeline in 3 days.
- Processing the three billion records took less than 65% of the time versus the prior custom built pipeline.
- The development team only required the standard four hour training on Hydrator before launching the project.
- The new pipeline eliminated the need for costly Hadoop experts, improved performance and reduced the number of technologies involved.

Learn More About Cask Hydrator

- Request a demo at http://cask.co/request-a-demo
- Or visit the Hydrator product page at http://cask.co/product/hydrator