Cask Data Application Platform

EDW Optimization:

NETEZZA OFFLOAD TO CLOUDERA WITH CASK

The Ever-Growing Netezza Data Warehouse

In principle, data growth is a good thing. When more data is available for analysts and business users, their reporting and analytical capabilities increase. However, there are some practical drawbacks:

- Expensive data storage
- Increasing data volume and variety
- Inability to meet SLAs or support increasing concurrent users
- Expensive database administration

The Data Warehouse and Apache® Hadoop

Hadoop has massive horizontal scalability along with system-level services that allow developers to free up storage and processing power from premium platforms like Netezza. Estimates indicate managing data in Hadoop can range from $1,000 to $2,000 per terabyte of data, compared to $20,000 to $40,000 per terabyte for appliance-based data warehouses.

But Hadoop is not a complete ETL solution. While Hadoop offers powerful utilities and virtually unlimited horizontal scalability, it does not provide the complete set of functionality users need for enterprise ETL. In most cases, these gaps must be filled through large number of lines of complex manual coding, slowing Hadoop adoption and frustrating organizations eager to deliver results.

Cask's Approach for Netezza Offloading

At Cask, we have formulated the following steps to overcome the challenges of offloading data and ELT workloads from Netezza to Hadoop:

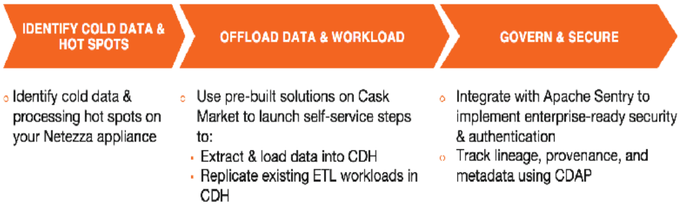
IDENTIFY the data and transformations that can bring highest benefits.

OFFLOAD the identified workloads and data to Hadoop quickly by using CDAP's self-service tools to replicate existing transformations.

GOVERN & SECURE by using CDAP's integration with Apache Sentry for enterprise-ready security & authentication, and CDAP's metadata management for lineage, and provenance information.

Identifying Cold Data & Processing Hot Spots

There are a couple of common places where you can locate cold data and workload hot spots. The first is the ELT workload which is also called "staging area" on Netezza. These workloads can take up precious CPU cycles while costing significant resources and IT costs. The second is by identifying cold data, usually in the form of massive fact tables, or voluminous multi-structured data like log or textual data.



IDENTIFY COLD DATA & HOT SPOTS

- Identify cold data & processing hot spots on your Netezza appliance

OFFLOAD DATA & WORKLOAD

- Use pre-built solutions on Cask Market to launch self-service steps to:
  - Extract & load data into CDH
  - Replicate existing ETL workloads in CDH

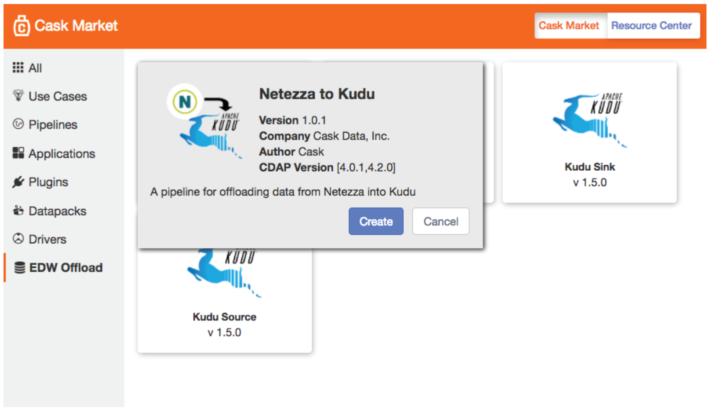
GOVERN & SECURE

- Integrate with Apache Sentry to implement enterprise-ready security & authentication
- Track lineage, provenance, and metadata using CDAP

Offloading Data & Workloads

CDAP (Cask Data Application Platform) simplifies data integration, app dev, data governance, and security & operations on Hadoop so you can focus on application logic and insights instead of infrastructure and integration. Use Cask's Unified Integration Platform for Big Data, CDAP, to operationalize your Netezza offload to Hadoop project.

Cask Market is Cask's "Big Data App Store" with push button deployment of pre-built applications, pipelines, and plugins. We have provided "Netezza Offload to Hadoop" solution in the Cask Market, that consists of pre-built pipelines and step-by-step



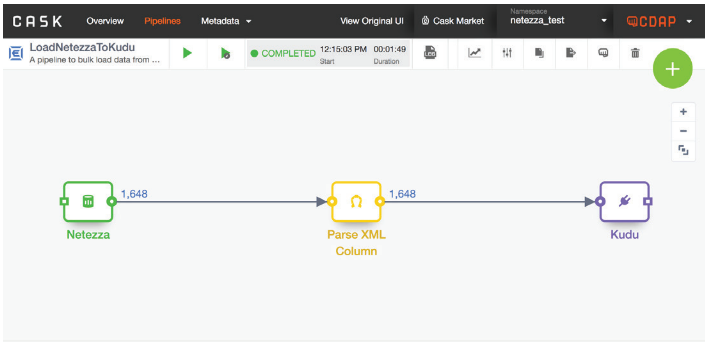
Netezza to Kudu

Version 1.0.1
Company Cask Data, Inc.
Author Cask
CDAP Version [4.0.1,4.2.0]

A pipeline for offloading data from Netezza into Kudu wizards to quickly configure and deploy new entities on the platform. The solution supports both the data and workload offload scenarios. CDAP provides a visual pipeline tool that allows you to build these pipelines without writing any Scoop, MapReduce, or Spark jobs. These pipelines will run on your existing hardware and take full advantage of the horizontal scalability of Hadoop.

Scenario #1: Offload Voluminous Data

After you identify cold or voluminous data, you can leverage this pre-built pipeline for offloading data from Netezza. CDAP provides drivers to connect to your Netezza tables to bulk load into Kudu. During the data flow, you wrangle unstructured data in-flight to convert it into a queryable dataset. For example, you can parse multi-structured XML data into columns before storing it on Kudu. The resulting dataset can be mined for additional business insights.

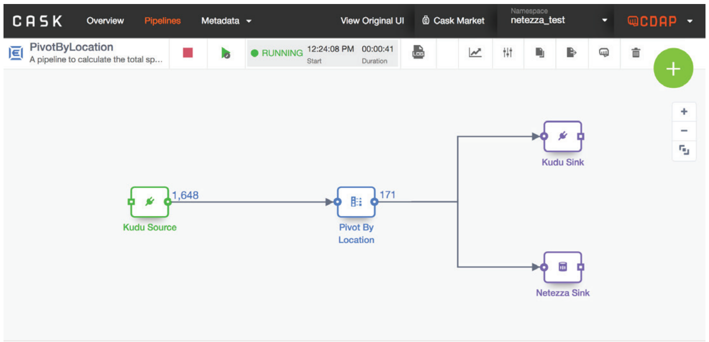


COPYRIGHT © 2017 CASK DATA, INC.

Scenario #2: Offload Workloads

One of the key differences between running queries in a Data Warehouse like Netezza versus a system like Hadoop, is that Hadoop works very well with denormalized data. However, data transformations can become very complex because joins, sorts, and aggregations are very expensive operations requiring large lines of complex code and performance tuning if hand coded. Building sophisticated CDC (Change-Data-Capture) data flows is even more difficult.

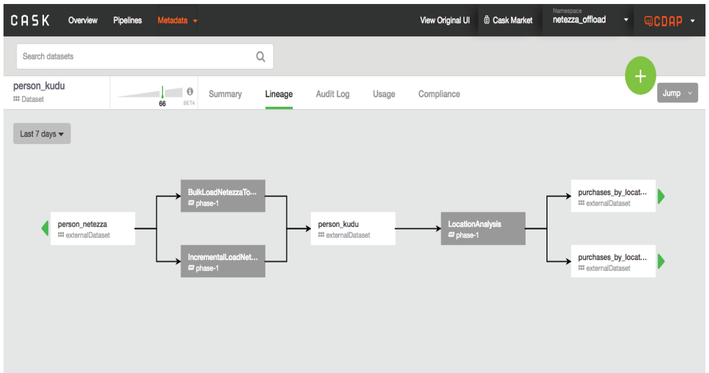
CDAP fills the gaps between Hadoop and ETL needs. The Netezza Offload solution package comes with pre-built pipelines which can be edited in a studio environment that consists of drag-and-drop sources, transforms, analytics, sinks, and actions. The bulk load pipeline reads data from your Netezza database, transforms fields in the data using the wrangling tool, and writes it to a denormalized table in Kudu. After this step, you can run the incremental loading pipeline to merge new data using the CDC process to Hadoop. The ability to source (or sink) data using a multitude of real-time and data-at-rest adapters make it easy to drop-in and connect to prototype and legacy data sources. Even if a source type does not exist, CDAP makes it easy to create a new driver and drop it into the pipeline without having to change the process model. This abstraction allows the solution to be upgradeable even if it is already deployed.



COPYRIGHT © 2017 CASK DATA, INC.

Secure & Govern

After you have migrated data and workloads to Hadoop, CDAP ensures extensive security and sophisticated governance across your datasets. This is done by automating the collection of technical, operational, and business metadata from ingestion to data delivery. Additionally, CDAP provides access controls that match permissions, roles, and groups in Netezza. Integration with Apache Sentry and KMS (encryption) services ensures compliance and mitigates risk.



Summary

The Cask solution for Netezza offload to Cloudera consists of pre-built pipelines, drivers, transformation logic, and best practices that can mitigate the challenging task of moving data and workloads to Hadoop. CDAP provides self-service tools that can support custom scenarios, users, and data sources.

Cask Data, Inc. | 150 Grant Avenue, Palo Alto, CA 94306 | 650.469.328 | cask.co