

Reference Architecture

Revelytix Loom and Hortonworks Data Platform

Executive Overview

Loom provides enterprise data management for Hadoop. Loom fills the gaps between what enterprises expect from production IT and what Hadoop provides. Loom provides a robust, centralized metadata management system and an extensive framework for automatically detecting data and generating metadata. Loom “datasets” provide a consistent, uniform, actionable format for Hadoop data. Loom tracks all transformations so that dataset lineage is always known. Loom exposes Hadoop data and Loom metadata externally through simple, RESTful APIs.

Customer Benefit

Loom customers benefit by massive improvement in the productivity of their data scientists, data engineers, and other data workers using Hadoop. Loom drastically reduces the time spent by these valuable employees finding, understanding, and preparing data for analysis.

Use Case Overview

Loom is a horizontal application that provides benefits across a broad spectrum of use cases. Any customer who is interested in using Hadoop as an information management platform, whether it is for advanced analytics or for ETL, will benefit from using Loom.

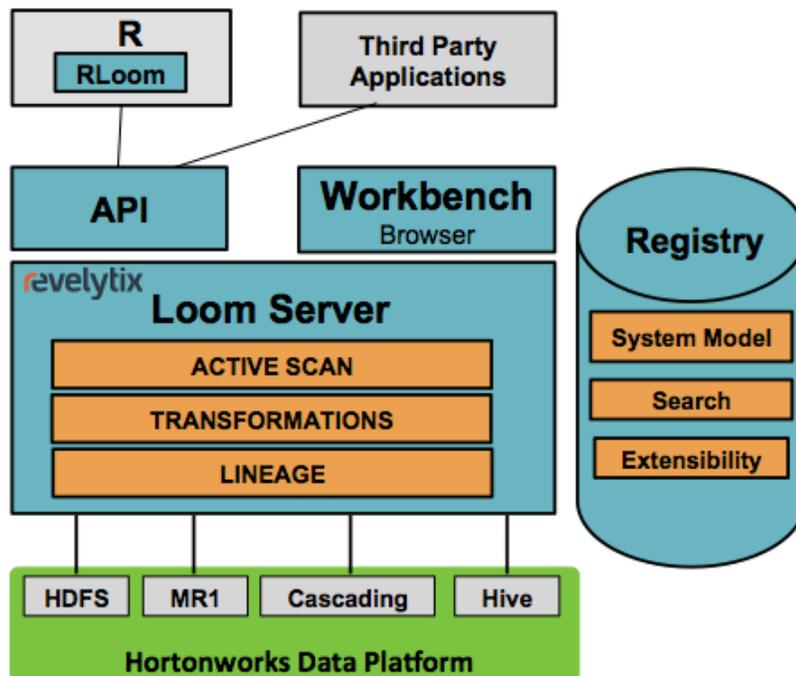
Company Overview

Revelytix is a commercial software company providing tools for enterprise information management. The founders and engineering team have been together for 14 years, eight at Metamatrix (sold to Red Hat in 2006) and six at Revelytix. For the first few years at Revelytix we built complex data management software for the Department of Defense. With the emergence of Hadoop, we have ported some of our core technology to that platform, into Loom, which is now our only generally available product and the sole focus of our company.

Solution Overview

The solution involves running Loom alongside a Hadoop cluster of any size, with Hive installed. Loom's Activescan feature crawls the cluster and scheduled intervals to discover and parse new files and new Hive databases. These are automatically registered in Loom as Sources. Users can preview data and schemas for Sources and interactively define formatting parameters. Activescan is pluggable, so that specific Source recognizers and parsers can be defined for any file formats. Activescan currently supports delimited text files, log files, and Hive databases as Sources. Once Sources are parsed correctly, they are converted to Datasets. Datasets are copied into Hive and become Loom-managed.

Datasets have known, formal schemas and row and column level statistics generated by Activescan. Datasets are also actionable; they can be transformed using HiveQL through the Loom Workbench. Loom tracks the execution of all transformations and automatically generates lineage metadata. Lineage graphs for Datasets show detailed relationships between Datasets, through transformation executions. The Loom Workbench is used by data scientists, data engineers, and other Hadoop users to track, manage, and transform Hadoop-based data. The Loom API exposes all this functionality to third-party tools, so that users can make use of other products for data loading or transformation, while still tracking and managing data and transformations through Loom. This enables Loom to serve as the central dataset management platform for a cluster.



Patterns of Use – Hortonworks Data Platform and Loom

In this “Data Refinery” use case, Hadoop is being used to distill large quantities of data into something more manageable. The resulting data is loaded into the existing data systems to be accessed by traditional tools – but with a much richer data set.

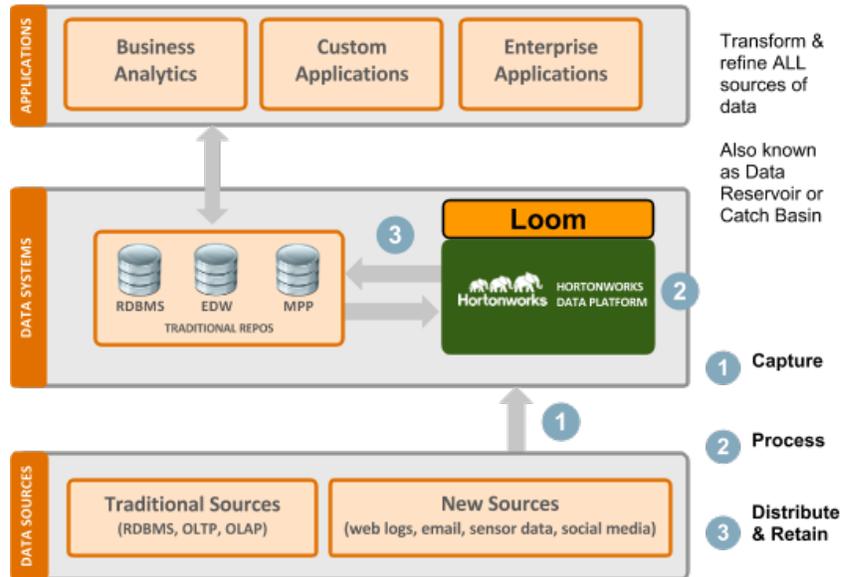


Figure 1: Operational Data Refinery

In the “Data Exploration” use case, organizations are capturing and storing a large quantity of new types of data (sometimes referred to as a data lake) in Hadoop and then exploring that data directly and iteratively. So rather than using Hadoop as a staging area for processing and then putting the data into the EDW – as is the case with the Refinery use case – the data is left in Hadoop and then explored directly.

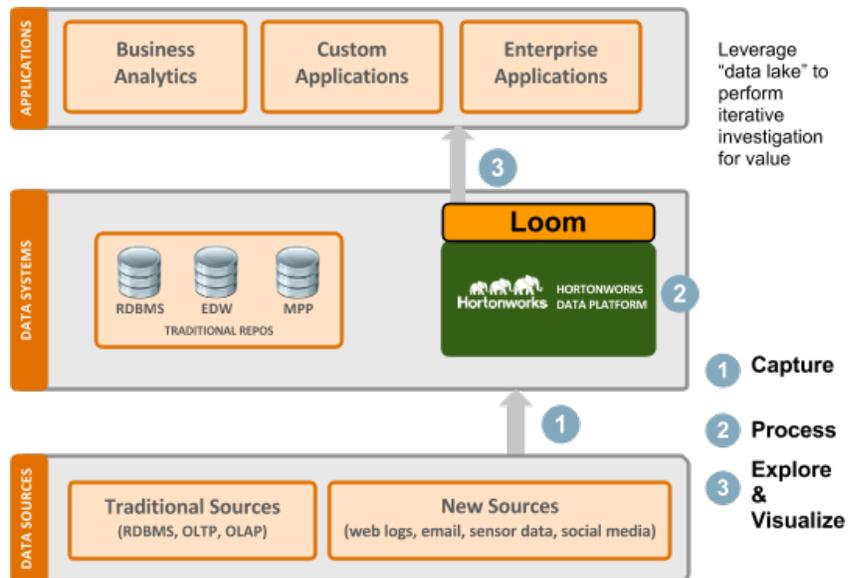


Figure 2: Big Data Exploration and Visualization

Summary

Loom provides essential capabilities enabling Hadoop and Hortonworks Data Platform to be enterprise-ready analytics platform. Loom automates many data management tasks and collects detailed lineage metadata. Loom's API provides a simple, robust access point into Hadoop data for third-party applications.

Next steps

Please visit www.revelytix.com for more details on Loom.

Visit www.Hortonworks.com for more information on Hadoop and Hortonworks Data Platform.

Contact Us

Revelytix: bmeindl@revelytix.com for more information.

Hortonworks: <http://hortonworks.com/about-us/contact-us/>