HITACHI
Inspire the Next

WHITE PAPER

Build a Streamlined Data Refinery

An Enterprise Solution for Blended Data That Is Governed, Analytics Ready and Available On Demand

By Hitachi Vantara

August 2019

Introduction

As the volume and variety of data has exploded in recent years, putting that data to business use has been a complex undertaking. Extracting value from big data has been a substantial challenge in and of itself, but it has been compounded by new preferences for data consumption. Basic batch reporting doesn't cut it anymore. Information consumers want analytics that they can explore in their favorite format on demand, often in the context of the other software applications they use every day.

All of this leaves IT organizations strained just to keep up. New data visualization tools have helped line-of-business groups (LOB) "help themselves." However, demands have only partially been met. At best, business users have access to a subset of data but they are stuck between the competing dilemmas of not trusting the data and not being able to wait for it to be stamped with the IT seal of approval. At worst, users simply cannot access the data they want when they want it: This is often the case if they require big data or unstructured data to meet their goals.

In light of these circumstances, many organizations can benefit from a streamlined data refinery architecture. This option represents a flexible, economical way to process and automate delivery of information to large numbers of users for a variety of analytic purposes. A streamlined data refinery is powered by on-demand orchestration for blending traditional data and big data, and it is a first step toward governed data delivery (see below). In this white paper, we will explore the relevant use cases where a streamlined data refinery can deliver substantial value, and break down the solution architecture to its component parts.

The Business Problem

Given the need to deliver more and more diverse data to users in ever-narrower time windows, we have identified the following use cases where a premium is placed on timely delivery of custom data blends that are fully governed and analytics ready. This is, of course, not an exhaustive list.

On-Demand Data for Business Analysts and Researchers

These roles often need to go beyond traditional SQL-based approaches to retrieving data from individual databases. Researchers often need "deep slices" of information from unwieldy sources, including machine and sensor data, weblog data, and unstructured text, which are often archived in Hadoop. One solution is providing an easy ability to request custom datasets on demand and dropping blended big datasets off in a convenient location (that is, FTP server or collaboration portal) and in a ready-to-use format (that is, Excel or CSV). Further, datasets can be staged in an analytical database like HP Vertica to offload complex query workloads from Hadoop.

Controlled Delivery of Datasets to Auditors and Regulators

Companies in highly regulated industries like financial services, healthcare and energy are pressured to prove they are in compliance with government regulations. This often requires them to combine data from multiple sources, run statistics and prove that their data management practices meet specific standards. For example, the Dodd-Frank Act mandates capital reserve ratios for bank holding companies and savings and loans. "Stress tests" must be run on a variety of bank operational data sources to assess compliance – and results of these tests must be auditable.

Forensic Analysis In Response to Exceptional Business Events

The scale of big data often prevents organizations from "pre-integrating" it into a data warehouse using traditional extract, tranform, load (ETL) approaches. Organizations increasingly rely on predictive analytics to screen for anomalies (such as financial fraud or network security threats) and to generate alerts that indicate the need for detailed forensic research by analysts. This can be optimized and accelerated by automating the preparation of analytic datasets for end users.

Governed data delivery is defined as the delivery of blended, trusted and timely data to power analytics at scale regardless of data source, environment, or user role. It lays the groundwork for seamless end user exploration and analysis of validated data blends from across the organization.

Data Blending as a Service for Customers and Partners

Data products are an emerging source of revenue for software as a service (SaaS) vendors, and many traditional corporations are incorporating analytics into their customer- and partner-facing applications to boost stakeholder relationships. In addition to providing raw data feeds to third parties, companies can offer data blending as a value-added service. In this scenario, users upload data to a site where it can be combined with the hosting organization's data and then returned as an enriched dataset.

Among these use cases, there are three common primary needs that present opportunities to drive additional productivity and business value for the organization in question (as shown in Table 1).

TABLE 1. NEEDS AND OPPORTUNITIES TO

Need	Functional Description	Associated Value
On-Demand Orchestration	Users need to be able to easily request complex datasets, and the resultant data delivery must happen in a just-in-time fashion, which requires the triggering of data processing, blending and modeling on demand. An automated process must be implemented to facilitate this on-demand orchestration.	This accelerates time to value in analytics projects and simplifies the process for end users by hiding complex details of underlying systems, empowering the business to respond to changing conditions quickly. Further, IT saves time by addressing many requests with one automated process for end-user self-service.
Proper Data Governance	This analytics-ready data must adhere to all relevant governance rules, ensuring that data is trusted, compliant, up to date and properly combined.	This minimizes risk and ensures confidence in business decisions made based on all enterprise data.
Blended Data in Format of Choice	Analytics users require the blending and enrichment of multisource data, delivered in a consumable way, whether that is in an ad hoc analysis tool or in a specific file format and location.	This makes users more productive in getting insight from raw data.

The Streamlined Data Refinery Solution

A streamlined data refinery architecture meets all of the core requirements of the use cases described above, providing for a user-driven, trusted data delivery process. At its core, the design pattern accommodates an on-demand process of user-initiated data requests, blending and refining of any data, automatic analysis schema generation, and publishing of analytic datasets in the format of choice. It consists of several key components.

Scalable Data Processing Hub

Usually Hadoop, this store is meant to house and manage a variety of structured and unstructured data from across the organization. In the diagram, Hadoop serves as the landing zone for data across the web, social media and transactional systems, as well as machines and sensors.

High-Performance Database

The database chosen must facilitate high-performance queries for analytics and visualization. When scale is required, an analytical database such as HP Vertica is a solid choice.

Pentaho Data Integration

Hitachi Vantara's Pentaho platform has a highly scalable data integration engine that is managed through its intuitive end-user interface. This engine provides the "glue" between the different data sources and stores in this architecture. The entire process outlined here can be triggered, on demand, via Pentaho Data Integration (PDI):

- **Blending and Orchestration:** PDI ingests data from virtually any data source, including both traditional systems and big data stores. It then processes, cleanses and blends the data in the required combinations to drive insight.
- **Automatic Modeling and Publishing:** As part of the data orchestration process, PDI automatically creates an online analytical processig (OLAP) schema and publishes it to the Pentaho Business Analytics server for end-user exploration and visualization.
- **Governance:** PDI's robust functionality enables IT to quickly and easily validate data sources being blended at the source. This allows for the right measure of control, without creating unnecessary frictions to end-user data access.

Architected Blends

Data developers leverage the power of Pentaho Data Integration to create a data blending process that users can execute at run time. This partnership between IT and business ensures governed data blends on demand through self-service data requests.

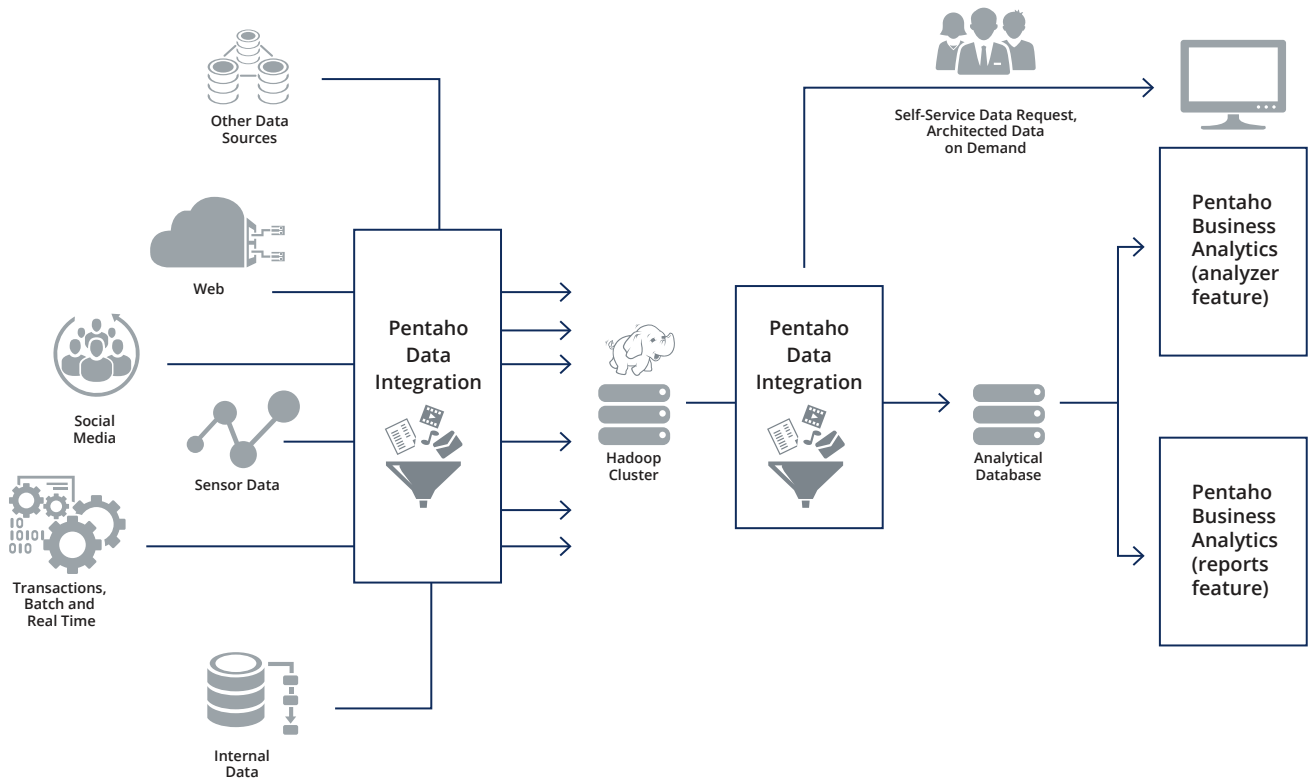
Self-Service Data Request

Users can request analytics-ready data delivery on demand via a web-based interface, created with Pentaho's CTools framework for 100% custom analytics user experiences. Through such an online interface, users can enter parameters (that is, data fields, source systems, time ranges and so forth) quickly and easily. They can also select whether data should be populated as a governed data source in Pentaho Business Analytics or in another format (Excel, CSV and so forth) in a different target location.

Pentaho Business Analytics

Represented by the analyzer and reports features shown in Figure 1, Pentaho Business Analytics is a flexible toolset for data exploration, visualization and consumption. Users leverage Pentaho here to access the automatically generated data models for interactive analysis.

Figure 1. Streamlined Data Refinery Architecture Diagram



Customer Example: Financial Regulatory Body

Goal

Empower analysts to identify suspicious patterns among billions of market transaction records per day.

Pentaho Solution

Users explore summary data with the ability to request detailed datasets on the fly for drill down through multi-dimensional models.

Architecture

The architecture leverages Hadoop with Amazon Elastic MapReduce and Hive; it uses Amazon RedShift as a high-performance analytical database in the cloud.

Advantages of the Streamlined Data Refinery

In addition to delivering unprecedented access to diverse, governed data for analytics on-demand, Pentaho's streamlined data refinery solution architecture provides a number of other benefits.

Datasets Are "Virtualized" and Managed Through Logical Service Endpoints

This means that the implementation is hidden from users, allowing IT to "replatform" or refactor the underlying data infrastructure without affecting LOB users.

Security Is Centrally Enforced

All requests are made through a common application (the Pentaho user console), which means that access can be revoked simply by disabling a user account or removing their membership from a role. Additionally, Pentaho-controlled access can be configured to map to existing enterprise security schemes.

Storage, Data Transformations and Query Serving (SQL and OLAP) Can Be Implemented Using Products That Match Existing Skills and Infrastructure

PDI jobs and transformations are flexible, allowing IT developers to run workloads in Hadoop (MapReduce or YARN), in dedicated PDI clusters or on single PDI servers. Similarly, Pentaho's OLAP engine (Mondrian) can work with a large number of analytical databases. Moreover, the infrastructure can evolve to take advantage of new storage and processing options without affecting the availability of the logical service endpoints. Pentaho's adaptive big data layer helps facilitate this "future-proofing."

The Backlog of Requests for Custom Data Feeds Can Be Reduced

By implementing parameterized request forms as part of a streamlined data refinery, IT departments can offload the selection and filtering of raw data to analysts and researchers. This is directly analogous to self-service interactive reporting, except that the data can be used with any number of company-standard reporting, analysis and statistical tools.

Figure 2. The Streamlined Data Refinery's User-Driven, Governed Process

Conclusion

In this discussion, we highlighted three core data delivery needs that are only being met on a limited basis in the market today:

- Orchestrate on-demand processing, blending and modeling of user-requested datasets to accelerate time to value in complex analytics initiatives.
- Ensure proper data governance during the delivery process, such that risk is minimized and confidence is increased in data-driven decisions.
- Provide blended and enriched data in the end-user format of choice, so that business users can be more productive in deriving insight from diverse data.

Indeed, these challenges cut across a variety of sector-specific use cases discussed. These cases include deep data exploration by researchers, forensic analysis of unexpected events, compliance assurance in regulated industries, and delivery of data to key customers and partners as a service.

The streamlined data refinery provides a well-defined solution architecture to address these needs. It both leverages existing organizational competencies and ensures that the on-demand data delivery process can quickly adjust to changes in the data environment.

Learn more about Pentaho Data Integration and Pentaho Business Analytics at HitachiVantara.com

We Are Hitachi Vantara

DataOps is the data practice for the AI era, connecting data consumers with data creators to accelerate collaboration and digital innovation. We are analytics, industrial expertise, technology and outcomes rolled into one great solution partner. Get Your DataOps Advantage.

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
hitachivantara.com | community.hitachivantara.com

Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
hitachivantara.com/contact

P-032-C BTD August 2019