Build a Streamlined Data Refinery

A comprehensive solution for blended data that is governed, analytics-ready, and on-demand

Introduction

As the volume and variety of data has exploded in recent years, extracting valuable information from public and private sources (both on premises and cloud) has been a complex undertaking. Moreover, extracting insights 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 to access more and more data 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 shared IT service agencies strained just to keep up. New data visualization tools have created niche solutions, for some agencies to "help themselves"; however, demands have only partially been met. At best, non-technical users (leadership) 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 data to arrive in another manor. 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. Additionally, many of these types of solution do not meet open format guidelines/mandates, thus can only be utilized as one piece of a solution, rather than a complete platform.

In light of these circumstances, many organizations can benefit from a Streamlined Data Refinery architecture, which 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 brief, 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 Data 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 FIELD SPECIALISTS

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

CONTROLLED DELIVERY OF DATA SETS TO JOINT AGENCIES & OVERSIGHT COMMITTEES

Federal agencies are under considerable pressure and expend a great deal of time and energy 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. Having an open set of standards and APIs allows for transparency, without compromising security or data quality. For example, an agency that supports field operatives can show exactly how intelligence data was compiled and their sources when justifying troop allocations and strategy.

FORENSIC ANALYSIS IN RESPONSES TO EXCEPTIONAL GLOBAL EVENTS

The scale of Big Data often prevents agencies from "pre-integrating" it into a data warehouse using traditional ETL approaches. Hence, many are increasingly relying 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 any agency.

'DATA BLENDING AS A SERVICE' FOR SHARED SERVICES & INFRASTRUCTURE

Data products are an emerging source of revenue for SaaS vendors, and perhaps not surprisingly, are becoming key components of shared government services. Through budget and data sharing some services are incorporating analytics into externally and internally facing applications to boost cooperation and improve efficiency. In addition to providing raw data feeds to other agencies can offer data blending to bolster analytic insights. In this scenario, users upload data to a site where it can be combined with the hosting agency's data and then returned as an enriched data set.

Among these use cases, there are three common primary needs that present opportunities to drive additional productivity and business value for the agency in question. They are as follows:

NEED	FUNCTIONAL DESCRIPTION	ASSOCIATED VALUE
On-Demand Orchestration	Users need to be able to easily request complex data sets, 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 agencies 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 tactical decisions made based on multiple agencies' data and share layers.
Blended Data in Format of Choice	Analytics users require the blending and enrichment of multi-source 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 data sets 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 any agency. In the diagram, Hadoop serves as the landing zone for data across the web, social media, transactional systems, and machines/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

Pentaho's highly scalable data integration engine, managed through its intuitive end user interface, provides the 'glue' between the different data sources and stores in this architecture. The entire process outlined here can be triggered on-demand via PDI:

Blending & Orchestration: PDI ingests data from virtually any data source, including both established systems and Big Data stores – and then processes, cleanses, and blends the data in the required combinations to drive insight.

Automatic Modeling & Publishing: As part of the data orchestration process, PDI automatically creates an 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 – allowing 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 ensures governed data blends on demand through self-service data requests.

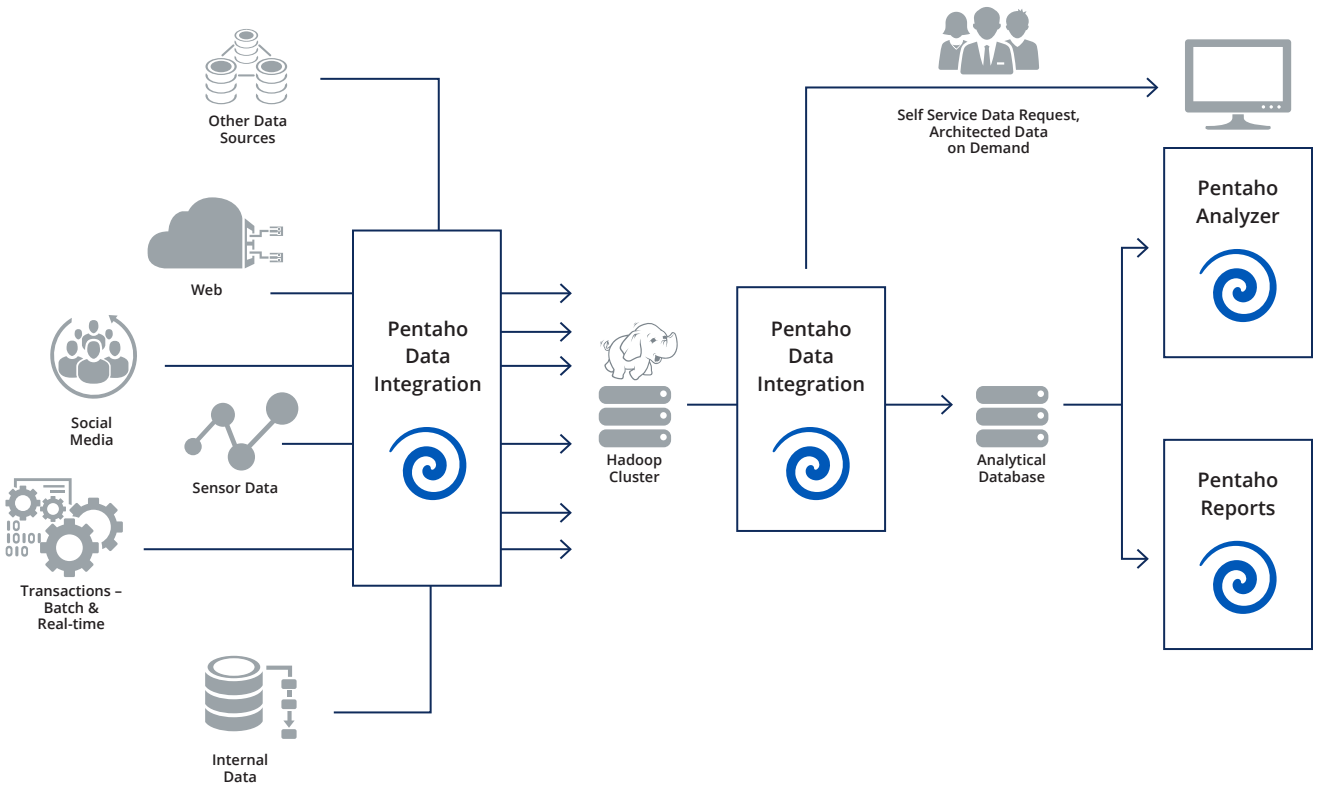
SELF-SERVICE DATA REQUEST

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 (i.e. data fields, source systems, time ranges, etc.) 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, etc.) in a different target location.

PENTAHO BUSINESS ANALYTICS

Represented by Analyzer and Reports in the diagram, 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.

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 data sets on the fly for drill-down through multi-dimensional models.

ARCHITECTURE

Leverages Hadoop with Amazon Elastic MapReduce and Hive; 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.

DATA SETS ARE "VIRTUALIZED" AND MANAGED THROUGH LOGICAL SERVICE ENDPOINTS

This means that the implementation is hidden from users, allowing IT to "re-platform" or refactor the underlying data infrastructure without affecting end 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, 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

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 data sets in order 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 users can be more productive in deriving insight from diverse data.

Indeed, these challenges cut across a variety of sector-specific use cases discussed, including '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 in a fashion that both leverages existing departmental competencies and ensures that the on-demand data delivery process can quickly adjust to changes in the data environment.

Learn more about Pentaho Business Analytics

pentaho.com/contact
+1 (866) 660-7555.

Global Headquarters

Citadel International - Suite 340
5950 Hazeltine National Drive
Orlando, FL 32822, USA
tel +1 407 812 6736
fax +1 407 517 4575

US & Worldwide Sales Office

353 Sacramento Street, Suite 1500
San Francisco, CA 94111, USA
tel +1 415 525 5540
toll free +1 866 660 7555

United Kingdom, Rest of Europe, Middle East, Africa

London, United Kingdom
tel +44 (0) 20 3574 4790
toll free (UK) 0 800 680 0693

FRANCE
Offices - Paris, France
tel +33 97 51 82 296
toll free (France) 0800 915343

GERMANY, AUSTRIA, SWITZERLAND
Offices - Munich, Germany
tel +49 (0) 322 2109 4279
toll free (Germany) 0800 186 0332

BELGIUM, NETHERLANDS, LUXEMBOURG
Offices - Antwerp, Belgium
tel (Netherlands) +31 8 58 880 585
toll free (Belgium) 0800 773 83

ITALY, SPAIN, PORTUGAL
Offices - Valencia, Spain
toll free (Italy) 800 798 217
toll free (Portugal) 800 180 060

Be social with Pentaho: