pentaho®
A Hitachi Group Company

Big Data at Cloud Scale

Pushing the limits of flexible and powerful analytics

Cloud Computing and Big Data

Two of the most disruptive technology trends over the last 10 years have been the growth of cloud computing and the emergence of Big Data systems. These developments have changed the way technology organizations operate and deliver value to their stakeholders.

At a basic level, cloud computing has allowed enterprises to optimize IT operations by significantly reducing the need to invest in on-premise hardware and software, not to mention the staff required maintain these systems. The cloud affords businesses a new level of flexibility, as they can acquire applications, infrastructure, and computing power in a way that is much more closely matched with the timing and duration of their project needs.

Further, by pooling infrastructure across many customers, cloud vendors are able to provide services that are highly elastic and scalable. This means it is much more financially and operationally manageable for enterprises to address unanticipated peaks and troughs in infrastructure needs. Overall, cloud adoption continues to show momentum, as the public IT cloud services market is expected to grow five times faster than the IT industry as a whole.[1]

At the same time, Big Data technologies have enabled organizations to generate value from data assets like never before. Historically, data that was high in volume, diverse in structure, and rapidly changing posed difficult challenges for enterprises that were used to working with traditional relational database technology.

However, new technical paradigms such as schema-on-read, massively parallel processing, and MapReduce have provided ways to reduce the overhead required to get raw data into a data store and drastically increased the speed and efficiency of processing large amounts of data. They have also made unstructured and semi-structured data much more accessible for businesses.

These innovations have begun to unleash actionable analysis on a variety of previously challenging data sources, including web logs, documents and text, and machine sensors. Even, "dark" data (data locked in corporate warehouses with little analytic access) has been given new life through these new technologies. As open source Big Data technologies have matured into commercially supported products, we have seen several platform categories start to gain rapid adoption:

- **Hadoop distributions:** Frameworks for large scale data storage and high performance processing across a distributed file system with the MapReduce paradigm as well as the more recently released MapReduce 2, also known as the YARN data operating system for cluster resource management; ideal for high volume unstructured data

- **NoSQL stores:** Non-relational databases with a very flexible structure; ideal for extremely rapid data ingestion and large numbers of reads based on key values

- **Analytic databases:** Databases designed for high performance analytics, leveraging techniques like compression, column-based storage, and high-speed "bulk" inserts of structured data; ideal for complex queries and OLAP analysis

[1] IDC press release, "IDC Forecasts Worldwide Public IT Cloud Services Spending to Reach Nearly $108 Billion by 2017 as Focus Shifts from Savings to Innovation," 9/3/2013

Two Worlds Converging

Big Data systems help organizations solve hard problems, but they normally require a significant up-front and ongoing IT investment. This includes a potentially large number of server machines as well as employees with skills that may be hard to come by, such as Java MapReduce programming. At the same time, the sheer amount of data in more ambitious projects (at times greater than 1 Petabyte or 1,000,000 Gigabytes) may lead teams to re-think whether keeping everything in-house is the best strategy. Finally, the time element is also important – procuring, installing, configuring, and testing the required technology doesn't happen over night.

On some level, it makes sense that enterprises would turn to cloud providers who have expertise in managing and maintaining extremely scalable and flexible computing and storage infrastructure. While on-premise data systems are by no means going away, research indicates that "cloud platforms are ideal deployment options for elastic and transient workloads built in modern application architectures." This suggests that organizations can effectively push the limits of analytics at scale by tapping into Big Data systems hosted on cloud infrastructure.[2]

In fact, a recent survey of enterprise decision makers reported that over a quarter of organizations have already started utilizing public cloud resources for Big Data analytics projects and another quarter plan to do so going forward.[3] While many of these early cloud projects involve high volumes of structured data, there are several key technology components that are already enabling extraction of value from massive, diverse data on cloud infrastructure:

- **Cloud Analytical Databases:** These cloud-based services, such as Amazon RedShift, are elastic data warehouses optimized for analytics with existing BI tools. In addition to leveraging enhancements like massively parallel processing and columnar storage to boost performance, this type of analytical database also includes management and monitoring of the solution by the provider. Users are able to avoid many of the costs related to setting up and managing a traditional data warehouse.

- **Hosted Hadoop Services:** Hadoop clusters can also be hosted in the cloud, which avoids the need for on-premise infrastructure and reduces reliance on in-house Hadoop-specific staffing to support Big Data use cases. Given on-premise start-up costs and cluster hardware expansion over time, it's easy to see where the cloud can provide value. Some Hadoop cloud offerings also include managed services, like job troubleshooting, software installation, testing, and more.

- **Data Integration and Analytics:** While adoption of 'cloud BI' tools has increased, Pentaho is unique in that it provides a cloud-deployable platform that supports end-to-end data integration and business analytics for Big Data stores, including the cloud analytical databases and hosted Hadoop services discussed above. This data can be blended with a variety of other cloud-based data for further insight.

The next section discusses a sample solution architecture, illustrating how these different technologies can be leveraged to drive business results in practice.

[2] Forrester, "The Public Cloud Market Is Now In Hypergrowth," 4/24/2014

[3] GigaOm Research, "How enterprises will use the cloud for big data analytics," 11/10/2014

Sample Solution Architecture

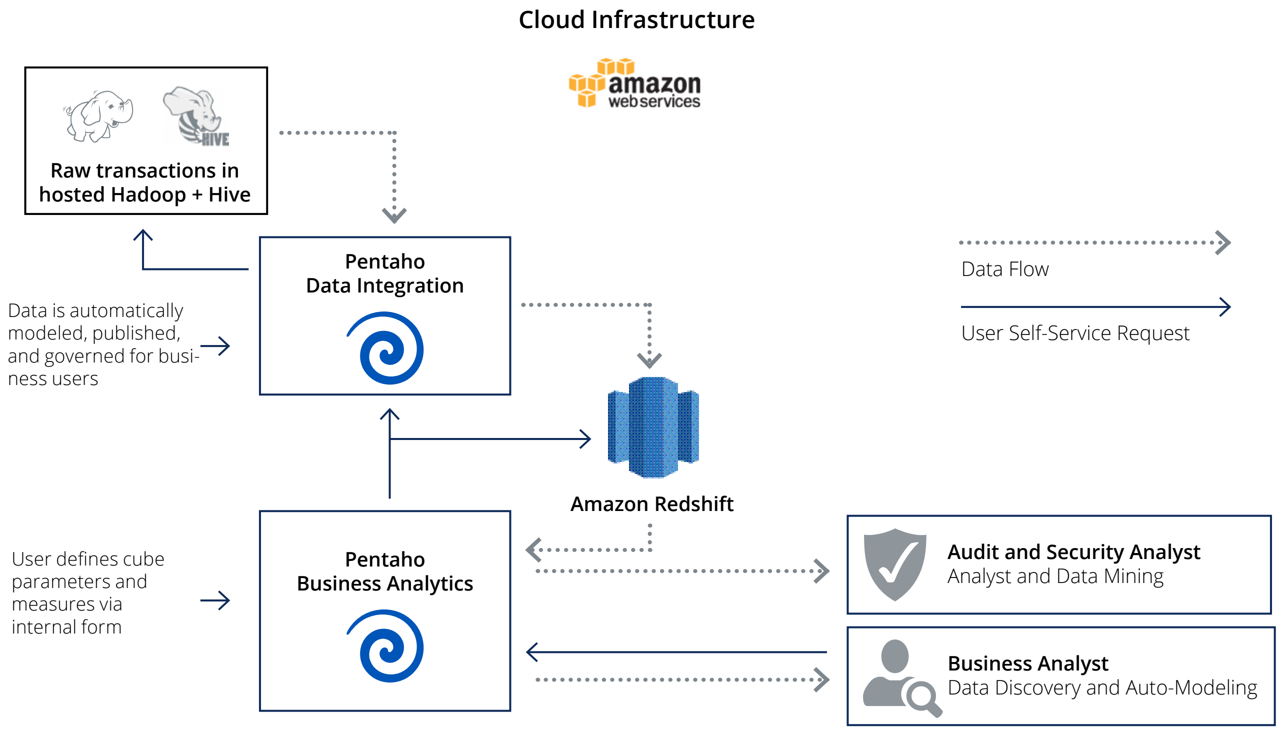
In this example, a regulatory organization has implemented a cloud-based data refinery solution in order to facilitate claims analyst access to up to 50 billion daily records of diverse structure. The goal is to enable more granular identification of potential violations and access to individual transactions to support claims. Amazon Web Services elastic computing and storage resources have been leveraged to control the cost of supporting these activities from an IT perspective.

Raw record data is first staged in a hosted Hadoop distribution, and a summary of that data is made available to Amazon Redshift via the Hive relational data warehouse layer, using Pentaho Data integration for orchestration. End users have the ability to drill down into detailed underlying data by selecting a set of parameters, such as dates, from a simple Pentaho form interface.

Upon submission of the selection, Pentaho Data integration triggers in-cluster Hadoop transformations to pull the desired data set and stage it in Redshift. PDI also automatically creates and publishes a multi-dimensional analysis model for this data set, logging the whole process for audit and administrative purposes.

From here, the claims investigator uses Pentaho Business Analytics for ad hoc analysis and visualization of the extracted data set in order to better identify regulatory risks. The end-to-end solution facilitates navigation of up to 500 TB of daily transaction data for precise drill-down and business insight.

At the same time, infrastructure characteristics, like Hadoop processing power or Redshift storage needed at a given time, are elastic. This of course translates to minimized fixed project costs – which would be substantially higher in a case where the entire solution architecture was hosted on premise.


Cloud Infrastructure
amazon web services
HIVE
Raw transactions in hosted Hadoop + Hive
Pentaho Data Integration
Data is automatically modeled, published, and governed for business users
Data Flow
User Self-Service Request
Amazon Redshift
User defines cube parameters and measures via internal form
Pentaho Business Analytics
Audit and Security Analyst
Analyst and Data Mining
Business Analyst
Data Discovery and Auto-Modeling

Conclusion

Early adopters are already illustrating how the cloud can expand on the value proposition of Big Data, delivering elastic and cost-effective solutions for integrating and analyzing data at unprecedented scale. However, 'taking Big Data to the cloud' doesn't eliminate the challenges of blending and orchestrating multiple complex data sources to drive value-added analysis. Only the right end-to-end data integration and analytics platform can translate these visionary architectures into proven solutions.

Raw record data is first staged in a hosted Hadoop distribution, and a summary of that data is made available to Amazon Redshift via the Hive relational data warehouse layer, using Pentaho Data integration for orchestration. End users have the ability to drill down into detailed underlying data by selecting a set of parameters, such as dates, from a simple Pentaho form interface.

Pentaho helps deliver on the promise of Big Data in the Cloud with the following unique capabilities:

- Flexible data transformation and orchestration for cloud-based Big Data stores, including hosted Hadoop distributions and Amazon Redshift
- Drag & drop ETL design for Big Data, including MapReduce workflow
- Auto-modeling and auto-publishing of analysis models for Amazon Redshift and other analytical databases
- The full spectrum of cloud-friendly end-user analytics, including visualization, ad hoc analysis, reporting, and dashboards

Case Study - Nasdaq

BUSINESS CHALLENGE

Nasdaq manages several billion rows of financial information each business day, and needed a modern, cost-effective way to make this information readily useful for several lines of business.

PENTAHO SOLUTION

Nasdaq leverages Pentaho's end-to-end platform to transform large complex data sets, integrate with Amazon Redshift, and empower end users with automatically generated reports as well as self-service analytics and dashboards to effectively manage several lines of business.

VALUE ADDED

With Pentaho, NASDAQ OMX created a cloud-based solution that manages huge volumes of data efficiently and cost effectively so the business can derive more useful information. Now a single development team replaces work previously done by a mix of development, system admins, and database administrators. The new solution cost represents over 50% savings relative to previous solution.

> "Our legacy systems were extremely slow, and lacked required data governance for data at scale. With today's big data solution in the cloud, we're not only able to scale beyond previous capabilities, but do it in a much more cost-effective way with flexible deployment options and higher data confidence."

MICHAEL WEISS
SENIOR SOFTWARE ENGINEER, NASDAQ OMX

pentaho®
A Hitachi Group Company

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:

For the latest information, please visit our web site at **pentaho.com**.