HITACHI
Inspire the Next

WHITE PAPER

Big Data at Cloud Scale

Push the Limits of Flexible and Powerful Analytics

By Hitachi Vantara

November 2019

Contents

Executive Summary

The growth of cloud computing and the emergence of big data systems have been two of the most disruptive technology trends over the last few years. These developments have changed the way technology organizations operate and deliver value to their stakeholders.

Cloud computing has allowed enterprises to optimize both IT operations and the rapid creation of new services. This is achieved by significantly reducing the need to invest in on-premises hardware, software and technical skill. At the same time, big data technologies have enabled organizations to generate value from data assets like never before. With the right unified, end-to-end, data integration, business intelligence and machine learning orchestration platform, organizations can quickly deliver big data processing in the cloud and on the premises.

This white paper covers:

- How open source big data technologies and platform categories have gained rapid adoption.
- Key technology components that are enabling extraction of value from massive, diverse data on cloud platforms.
- Sample solution architecture, which illustrates how the different technologies can be leveraged to drive business outcomes.
- NASDAQ Case Study, which describes how the company employed a cloud-based solution with Hitachi Vantara's Pentaho platform to manage huge volumes of data and drive business insight.

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-premises 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 defining schema on read when accessing data, massively parallel processing, microservices and stream processing have provided many new opportunities. These options include the abilities to reduce the overhead required to get raw data into a data store, to deal with data in motion, and to make robust and flexible architectures. They drastically increase the speed and efficiency of processing large amounts of data. Making unstructured and semistructured data much more accessible for businesses combined with these new paradigms make whole new generations applications, business models and efficiencies available.

These innovations have also 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 silos 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, especially for next-generation applications and analytics.

- Apache Hadoop based distributions: Frameworks for large-scale data storage and high-performance processing across a distributed file system, ideal for high volume unstructured data.
- Not-only-SQL (NoSQL) stores: NoSQL databases are agile and can include geographical distributed scale-out architecture. The main types of NoSQL stores are document databases, graph stores, key value stores, wide column stores and multimodal stores.

Two Worlds Converge

Big data systems help organizations solve hard problems, but they normally require a significant upfront and ongoing IT investment. This type of venture includes a potentially large number of server machines as well as employees with skills that may be hard to come by, such as Java or MapReduce skills. At the same time, the sheer amount of data in more ambitious multi-petabyte projects may lead teams to rethink 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 overnight.

On an infrastructure-as-a-service (IaaS) 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.

[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

While on-premises 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] Now, more advanced platform-as-a-service (PaaS) versions of data processing engines, Hadoop-as-a-Service or NoSQL-as-a-Service have enabled far better integration with other cloud-based application stacks.

A 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 business intelligence (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.
- **Hadoop and NoSQL services:** Hadoop services can also be hosted or run as a platform in the cloud, which avoids the need for on-premises infrastructure and reduces reliance on in-house Hadoop-specific staffing to support big data use cases. Given on-premises startup 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, Hitachi Vantara's Pentaho platform is unique. Pentaho 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 or platform Hadoop services discussed above. This data can be blended with a variety of other cloud-based data for further insight. An extract, transform and load (ETL) job can be created within the tool but executed through push-down processing using either Spark or MapReduce without recoding the task. Dealing with streaming data from Apache Kafka, connecting to Amazon S3 using Identity and Access Management, connecting to Google Cloud Storage, Google BigQuery, Microsoft Azure storage and many other approaches are simplified. Even file type such as ORC, Avro and Parquet are catered for.

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

Sample Solution Architecture

In this example, shown in Figure 1, 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.

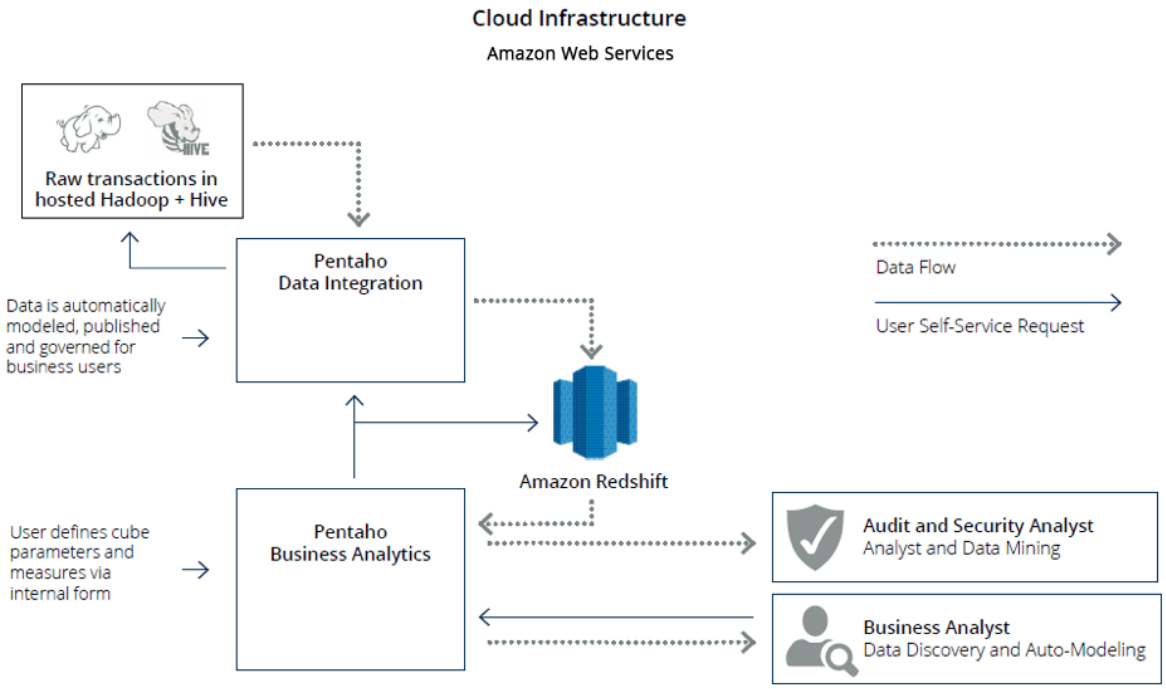
Upon submission of the selection, Pentaho Data Integration triggers in-cluster Hadoop transformations to pull the desired data set and stage it in Redshift. Pentaho Data Integration also automatically creates and publishes a multidimensional analysis model for this data set, logging the whole process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

[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

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 500TB 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 premises.

Figure 1. Cloud-Based Data Refinery solution



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. The platform also provides end users with 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 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

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. 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-and-drop ETL design for big data, including MapReduce workflow.
- Automodeling and autopublishing 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.

For more information regarding Hitachi Vantara's Pentaho solutions for big data in the cloud visit hitachivantara.com or contact your Hitachi Vantara representative.

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

WP-581-B BTD November 2019