Big Data Analytics for Retail with Apache™ Hadoop®

A Hortonworks and Microsoft White Paper

Contents

The Big Data Opportunity for Retail

Increasingly connected consumers and retail channels have served to make a wide variety of new data types and sources available to today's retailer.

Traditional structured types of data—isuch as the transactional and operational data found in a data warehouse—iare now only a small part of the overall data landscape for retailers. Savvy retailers now collect large volumes of unstructured data such as web logs and clickstream data, location data, social network interactions and data from a variety of sensors.

The opportunities offered by this data are significant, and span all functional areas in retail. According to a 2011 study by McKinsey Global Institute, retailers fully exploiting big data technology stand to dramatically improve margins and productivity.

> *"In the coming years, the continued adoption and development of big data levers have the potential to further increase sector-wide productivity by at least 0.5 percent a year through 2020. Among individual firms, these levers could increase operating margins by more than 60 percent for those pioneers that maximize their use of big data. Such a boost in profitability would be especially significant in a sector where margins are notoriously tight."*[1]

In order for retailers to take advantage of new types of data, they need new approaches to data storage and analysis.

To unlock the promise of big data, retailers from across the industry have turned to Apache Hadoop to meet these needs and to collect, manage and analyze a wide variety of data. In doing so, they are gaining new insights into the customer, offering the right product at the right price, improving operations and supply chain management, and enabling innovative new business models.

"According to a 2011 study, retailers that fully exploit Big Data technology stand to dramatically improve margins and productivity."

1) Mankiya, et. al.. "Big Data: The next frontier for innovation, competition and productivity." McKinsey Global Institute, May 2011. <http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation>.

The Data Deluge, and other Barriers

Modern retailers are increasingly pursuing omni-channel retail strategies that seek to create a seamless customer experience across a variety of physical and digital interaction channels, including in-store, telephone, web, mobile, and social media. The emergence of omni-channel retailing, the need for a 360-view of the customer and advances in retail technology are leading today's retailers to collect a wide variety of new types of data.

These new data sources—including clickstream and web logs, social media interactions, search queries, in-store sensors and video, marketing assets and interactions, and a variety of public and purchased data sets provided by 3rd parties—have put tremendous pressure on traditional retail data systems.

Retailers face challenges collecting big data. These are often characterized by "the three Vs":

Volume, or the sheer amount of data being amassed by today's enterprises.

Velocity, or the speed at which this data is created, collected and processed.

Variety, or the fact that the fastest growing types of data have little or no inherent structure, or a structure that changes frequently.

These three Vs have individually and collectively outstripped the capabilities of traditional storage and analytics solutions that were created in a world of more predictable data.

Compounding the challenges represented by the Vs is the fact that collected data can have little or no value as individual or small groups of records. But when explored in the aggregate, in very large quantities or over a longer historical time frame, the combined data reveals patterns that feed advanced analytical applications.

In addition, retailers seeking to harness the power of big data face a distinct set of challenges. These include:

Data Integration. It can be challenging to integrate unstructured and transactional data while also accommodating privacy concerns and regulations.

Skill Building. Many of the most valuable big data insights in retail come from advanced techniques like machine learning and predictive analytics. However, to use these techniques analysts and data scientists must build new skill sets.

Application of Insight. Incorporating big data analytics into retail will require a shift in industry practices from those based on weekly or monthly historical reporting to new techniques based on real-time decision-making and predictive analytics.

Fortunately, open technology platforms like Hadoop allow retailers to overcome many of the general and industry-specific challenges just described. Retailers who do so benefit from the rapid innovation taking place within the Hadoop community and the broader big data ecosystem.

"When collected, stored and explored in the aggregate, Big Data can reveal patterns that feed advanced analytical applications, often translating into actionable—and profitable—business insights for retailers."

Hadoop in Retail

Apache Hadoop is an open-source data processing platform created at web-scale Internet companies confronted with the challenge of storing and processing massive amounts of structured and unstructured data.

By combining the affordability of low-cost commodity servers and an open source approach, Hadoop provides cost-effective and efficient data storage and processing that scales to meet the needs of the very largest retail organizations.

Hadoop enables a shared "data lake," allowing retail organizations to:

Collect Everything. A data lake can store any type of data, including an enterprise's structured data, publicly available data, and semi-structured and unstructured social media content. Data generated by the Internet of Things, such as in-store sensors that capture location data on how products and people move about the store, is also easily included.

Dive in Anywhere. A data lake enables users across multiple business units to refine, explore and enrich data on their terms, performing data exploration and discovery to support business decisions.

Flexibly Access Data. . A data lake supports multiple data access patterns across a shared infrastructure, enabling batch, interactive, real-time, in-memory and other types of processing on underlying data sets.

A Hadoop-based data lake, in conjunction with existing data management investments, can provide retail enterprises an opportunity for Big Data analytics while at the same time increasing storage and processing efficiency, which reduces costs.

Big data technology and approaches have broad application within the retail domain. McKinsey Global Institute identified 16 big data use cases, or "levers," across marketing, merchandising, operations, supply chain and new business models:

Function	Big data lever
Marketing	Cross-selling Location-based marketing In-store behavior analysis Customer micro-segmentation Sentiment analysis Enhancing the multichannel consumer experience
Merchandising	Assortment optimization Pricing optimization Placement and design optimization

> "Hadoop provides cost-effective and efficient data storage and processing that scales to meet the needs of the very largest retail organizations."

Function	Big data lever
Operations	Performance transparency Labor inputs optimization
Supply chain	Inventory management Distribution and logistics optimization Informing supplier negotiations
New business models	Price comparison services Web-based markets

With so many possible applications, it can be challenging for retailers to know where to start with big data. The following use cases represent particularly compelling business opportunities for retailers new to big data or just starting out with Hadoop. These scenarios are often funded in conjunction with an effort to offload demanding Extract, Transform and Load (ETL) processing from the Enterprise Data Warehouse (EDW). The resulting margin lift from these business cases, combined with the cost savings gained by reducing load on the EDW, are delivering improved financial results to retailers of all sizes:

Single View of the Customer. Combine historical sales data from structured systems with new, unstructured and semi-structured data from social media, web logs, and location data. This single view delivers a complete consumer profile, allowing the retailer to better attract and retain customers through a more thorough understanding of shopping and buying patterns.

Recommendation Engine. Apply historical spending behavior and purchase affinity to recommend suitable upsell and cross-sell opportunities. This can be deployed online, over in-store kiosks, via mobile devices, and through customer service scripting. Ultimately, recommendation engines drive higher revenue, increased margin, and improved customer satisfaction.

Basket Analysis. Support marketing efforts by analyzing item-level cash register receipt detail (i.e. T-Logs). Gain insight into product affinity, article pairing, and opportunities to build promotions around item families.

Price Optimization. Collect online competitive price information in order to optimize prices in real-time.

Inventory Optimization. Ensure that the right product is available at the right time, preventing out-of-stock events and ensuring customer satisfaction. Blend supply chain data on item movement with promotions that move product you have on the shelves. Use this data to ensure a consumer will always get the product she wants after a recommendation or promotion.

CASE STUDY: OMNI-CHANNEL PERSONALIZATION WITH A 360° VIEW

In today's data-rich world, retail enterprises need to rethink traditional approaches to knowing the customer. Traditionally, retailers analyze customer interactions on a channel-by-channel basis: store sales are tallied from cash register receipts; web data is analyzed to determine site effectiveness; call center data is used to report phone sales and common consumer requests or complaints. Few retailers have visibility into sales and service interactions that span channels, and so they don't actually understand the full breadth of an individual customer's experience.

Modern retailers want a single, 360° view of their consumers, but data silos and fragmentation can block that single view. Without it, retailers have a hard time accurately calculating lifetime customer value (LCV) and they are limited in their ability to create effective promotions and offers. Without an intimate understanding of their customers, retailers can miss upsell and cross-sell opportunities.

Pier 1 Imports is a retailer committed to engaging customers across multiple channels, including their network of over 1,000 stores and their e-commerce web site. To meet its sales and marketing goals, Pier 1 Imports needed to combine online and in-store behavioral data to predict customer purchase intent across both channels.

The company developed a web-based solution that collects e-commerce site data and point-of-sale store data in Hadoop, specifically using Microsoft Azure HDInsight. That data is then made anonymous and processed with predictive analytic models using Azure Machine Learning. The system creates reports and visualizations that can be accessed by Pier 1 Imports' data analysts.

By integrating the data from disparate sources, the company was able to gain new insights into customers' shopping behavior and product preferences, allowing it to create more personalized marketing campaigns.

> "Pier 1 is a very data-rich company. We needed a way of leveraging these multiple sources to better match our customers to the products they're looking for."
>
> **—ERIC HUNTER**
>
> Executive Vice President of Marketing, Pier 1 Imports

CASE STUDY: ANALYZING BRAND SENTIMENT TO DRIVE GROWTH

Retailers lack a reliable way to track and analyze brand health and engagement. It's difficult to explore the impact of advertising, competitor moves, product launches and news stories on a brand. Canned social media dashboards are not enough, and internal brand studies can be slow, expensive and flawed.

BlackBall, a Taiwanese restaurant chain with 60 stores throughout Taiwan and Malaysia, recognized the influence of social media engagement on regional demand, but found it very difficult to make the connections between the disparate sources of data it would need to correlate in order to make decisions. Connecting social inputs with point-of-sale data would allow BlackBall to predict regional demand, enabling it to deliver more targeted promotions, improve product distribution and reduce waste.

After looking at a variety of BI tools, BlackBall implemented a hybrid cloud solution based on Microsoft SQL Server running on-premises and the Azure HDInsight Hadoop as-a-service offering running on the Azure platform.

BlackBall's Hadoop solution lets it constantly monitor customer feedback on social media and correlate this data with sales, allowing the company to plan more strategically. The Hadoop solution provides insights that allow the company to develop new products, execute more effective promotions, and ensure that its stores are equipped with the staff, ingredients and supplies needed to meet customer demand.

CASE STUDY: CREATING NEW BUSINESS OPPORTUNITIES WITH BIG DATA PRICING OPTIMIZATION

A fundamental task of the retailer is to ensure that each consumer is offered the right product at the right price, and motivated to buy using the right promotion. Retailers have traditionally relied on historical sales reports and the intuition of marketers to accomplish this. Today's data driven environment offers retailers the opportunity to dynamically optimize pricing and promotion on an individual consumer basis, allowing the retailer to maximize sales and profits.

Santa Monica, California based TrueCar is building entirely new business models based on its ability to aggregate data from numerous sources and use this data to help consumers get the right price for the right car. Before moving to Hadoop, the company held data in over 230 discrete databases. TrueCar enriches this data via hundreds of proprietary processes that allow it to offer consumers local price estimates with a high level of confidence through its mobile apps, over 400 branded partner sites, and its flagship site TrueCar.com.

As the company grew, it found its ability to innovate and scale hindered by an outmoded data architecture. A large part of the time required by company developers to field new pricing applications was spent in building the pipeline to fetch and pre-process required data. To increase its agility, TrueCar moved all of its mission-critical product and pricing data—including data on approximately 8,400 dealers, 8 million vehicles and 250 million car images—to a data lake.

In moving to the Hortonworks Data Platform (HDP), TrueCar realized three profound benefits. First, with all data in one Hadoop-based data lake, company developers were empowered to innovate more quickly. New applications can more easily use company data, and Hadoop's more efficient processing model allows applications to run full historical analyses. Second, the company now stores and processes its data at a cost of 23 cents per gigabyte. A traditional platform would have cost 19 dollars per gigabyte, without any processing capability. And finally, the company can better serve customers (both buyers and dealers) by storing and processing enough data to make trustworthy price estimates with a system that scales easily as more cars and dealers come into the system.

TrueCar stores and processes data at

23¢ per gigabyte

Instead of a traditional platform's $19 per gigabyte

"TrueCar can better serve customers (both buyers and dealers) by storing and processing enough data to make trustworthy price estimates with a system that scales easily as more cars and dealers come into the system."

Microsoft and Hortonworks Bring Apache Hadoop to Retail

Hortonworks and Microsoft have partnered to bring the benefits of Hadoop to retail. Through this partnership, Hortonworks delivers enterprise-grade solutions that integrate deeply with Microsoft tools and applications while providing deployment flexibility. Unlike purely on-premises or purely cloud implementations, Microsoft and Hortonworks offer both the control of on-premises deployment and the elasticity of Hadoop in the cloud. The partnership also opens up a wealth of cloud computing and advanced analytics opportunities accessible through Azure.

Interoperability, Flexibility and Portability. Through their joint engineering partnership, Microsoft and Hortonworks have built an HDP Windows distribution that was contributed back to the open source community through the Apache Software Foundation. This gives organizations choice of operating systems (Windows or Linux).

Cloud Scale. Microsoft and Hortonworks offer on-premises Hadoop customers the scale and redundancy of the Azure cloud. Customers can backup their on-premises data, elastically scale resources during peak demand, and easily reduce those resources when demand wanes

Accessibility to Cloud Analytic Services. Customers have access to advanced analytics and data services available via the Azure cloud. They can then process new types of data with HDInsight or use Azure Machine Learning to predict future trends without needing a seasoned team of data scientists. Customers can also use Azure Data Factory to orchestrate and curate their data through simple, fault tolerant data pipelines. All of this is available as part of the Cortana Analytics Suite, a big data and advanced analytics solution in Azure without having to install hardware or maintain software.

> "The partnership between Microsoft and Hortonworks brings the benefits of Hadoop to retail, offering the control of an on-premises deployment, coupled with the elasticity of Hadoop in the cloud."

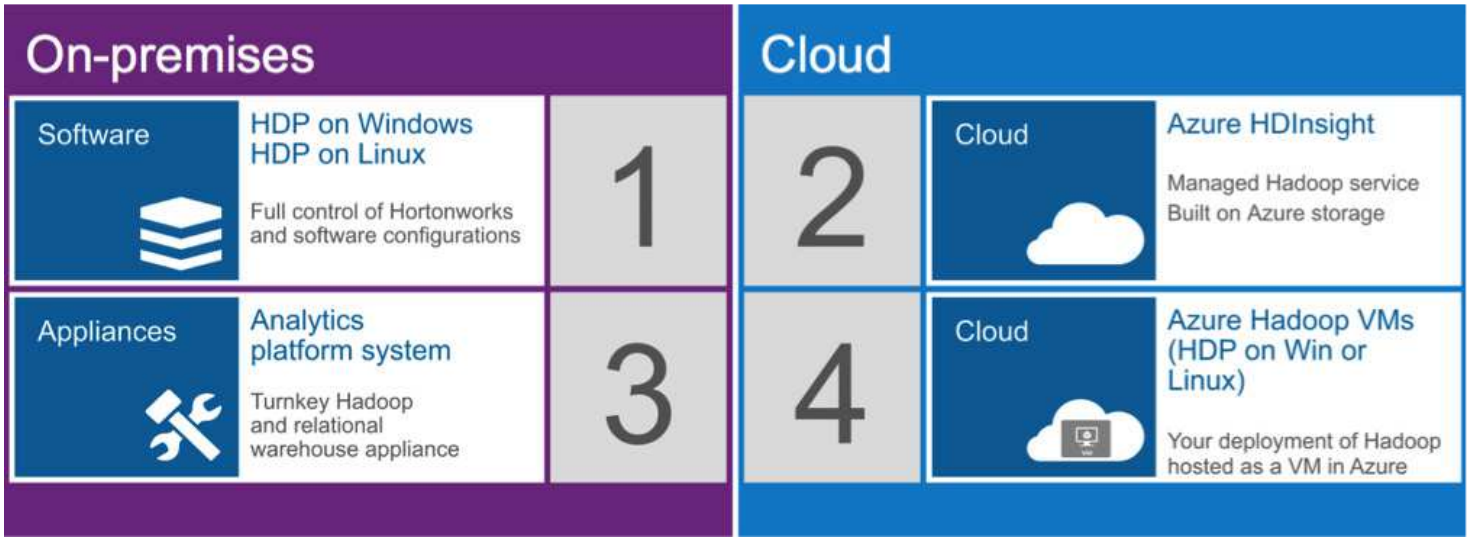

Figure 1

About Hortonworks

Founded in 2011 by 24 engineers from the original Yahoo! Hadoop development and operations team, Hortonworks has amassed more Hadoop experience under one roof than any other organization.

Hortonworks is the leader in Open Enterprise Hadoop and develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities. The Hortonworks Data Platform provides an open platform that deeply integrates with existing IT investments and upon which enterprises can build and deploy Hadoop-based applications. Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop. For more information, visit www.hortonworks.com.

About Microsoft

Founded in 1975, Microsoft (Nasdaq: "MSFT") is the worldwide leader in software, services and solutions that help people and businesses realize their full potential. Microsoft offers the Azure cloud computing platform, a growing collection of integrated services—analytics, computing, database, mobile, networking, storage, and web—for moving faster, achieving more, and saving money. Within Azure is HDInsight, Microsoft's Hadoop distribution powered by the cloud. HDInsight was architected to handle any amount of data, scaling from terabytes to petabytes on demand. You can spin up any number of nodes at anytime without buying new hardware or time-consuming installation and set up. HDInsight is part of Cortana Analytics, Microsoft's fully managed big data and advanced analytics suite that enables you to transform your data into intelligent action. Find out more at www.microsoft.com/cortanaanalytics.