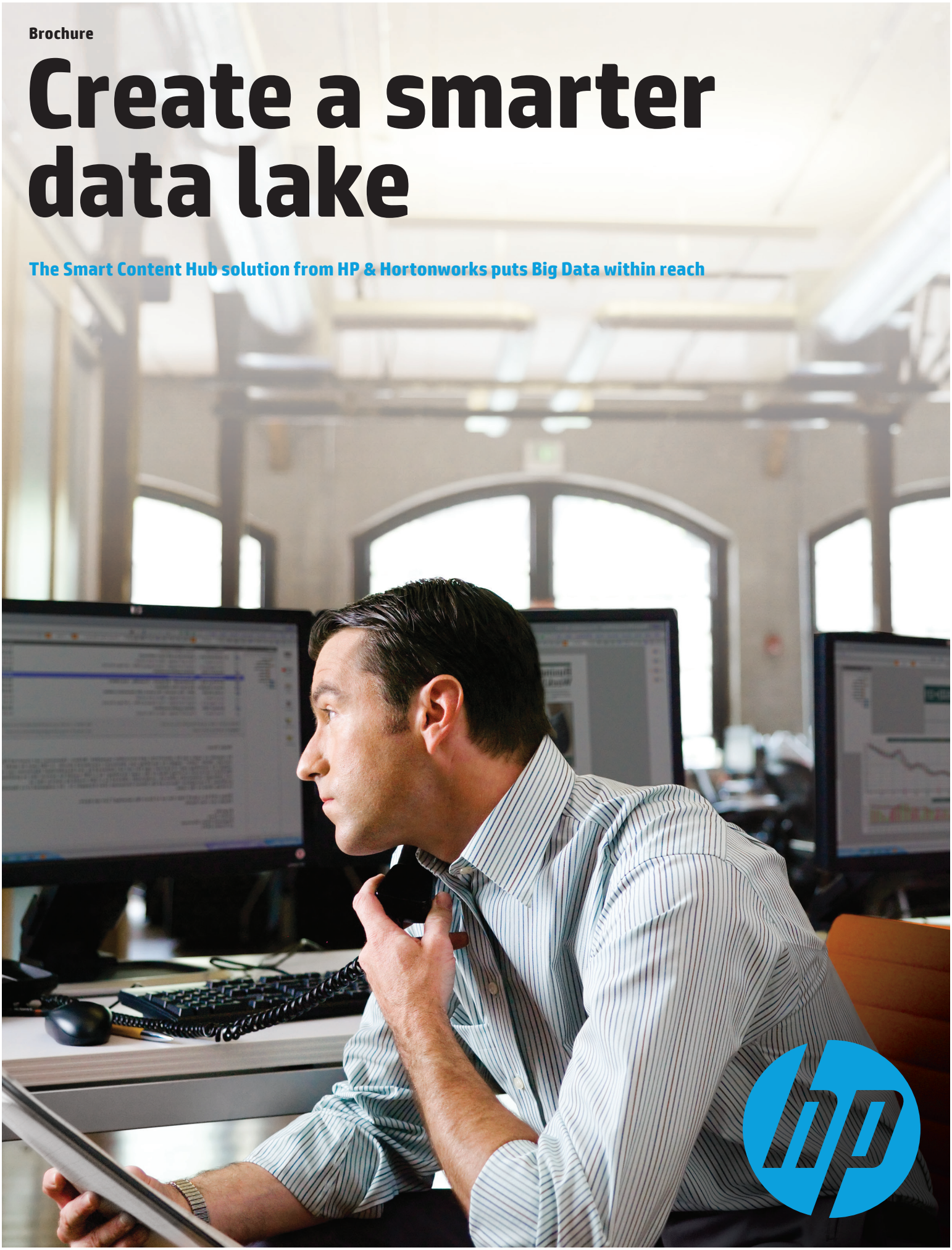
Brochure

Create a smarter data lake

The Smart Content Hub solution from HP & Hortonworks puts Big Data within reach



When you can efficiently organize, discover, and analyze all of your enterprise information, you greatly increase your rate of success in today's competitive environment. For over a decade, database and data warehouse platforms have enabled organizations to effectively manage transactional data. More recently, Apache Hadoop has empowered organizations to cost-effectively capture and analyze newer and faster growing data types such as machine data, log data, and unstructured data, allowing you to more effectively compare and correlate transactional and behavioral data.

Highlights

- Empowers business analysts and other knowledge workers to use Hadoop, not just programmers and data scientists
- Bi-directional and intelligent data connectors to automate the data flow to/from the Smart Content Hub without complex programming. Out-of-the-box support for 400+ data sources and 1000+ file formats.
- Advanced search, analytics, and machine learning, including entity extraction, sentiment and behavior analysis, and image and video analytics
- Enterprise-class security with support for role and user-based access controls
- Architected for the Hadoop platform with high scalability and reliability
- Single, easy-to-use interface to manage, search, and run analytics on information in Hadoop, along with other repositories

Despite the emergence of new storage and processing technologies for handling big data, much of the content within an enterprise has remained fractured and out of reach. Information is spread across multiple repositories, on-premise and in the cloud, with each repository having its own security roles and authorization models—which then need to be respected when their contents are brought into a data lake (a storage repository that holds a vast amount of raw data in its native format). Once the information is in the data lake, actually using the data is complicated by the huge variety of file formats such as archives, binaries, or difficult to consume image, audio, and video files. All of these challenges limit enterprises in their ability to correlate information and derive insights.

The Smart Content Hub solution from HP and Hortonworks addresses key information challenges by enabling a shared infrastructure that transparently synchronizes information with existing systems and offers an open, standards-based platform that allows you to perform search and deep analysis from a single interface.

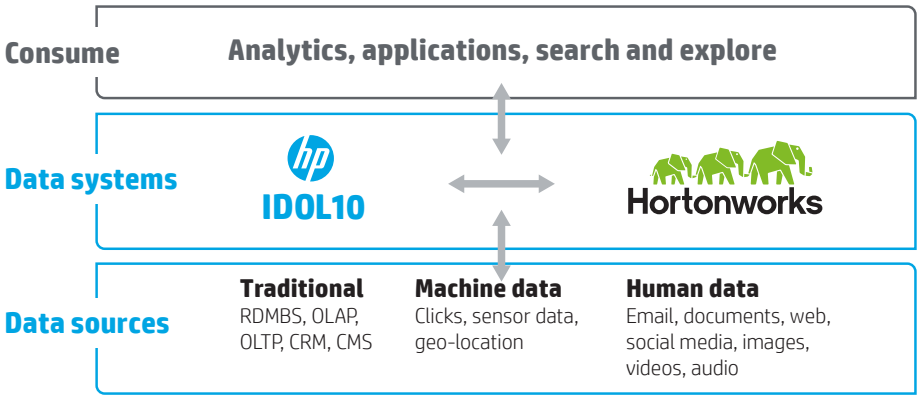
The solution architecture

The Smart Content Hub is comprised of two key technologies: HP IDOL (Intelligent Data Operating Layer) and the Hortonworks Data Platform (HDP).

- A Gartner Magic Quadrant leader in contextual search and analytics, HP IDOL understands text, images, audio, video, and structured data, putting information within reach so you can act on all your data. IDOL offers intelligent ETL with over 400 secure data connectors, drives policy-based synchronization of information between systems, and provides best-in-class information discovery and analysis capabilities.
- HDP, 100% open source Apache Hadoop, offers a scalable, cost effective, and secure platform for managing all enterprise information. HDP's open source data platform enables fit-for-purpose applications to analyze a shared collection of information without the need to duplicate data or infrastructure.

With Smart Content Hub, you can extend your enterprise data warehouse and deliver improved accessibility and insight.

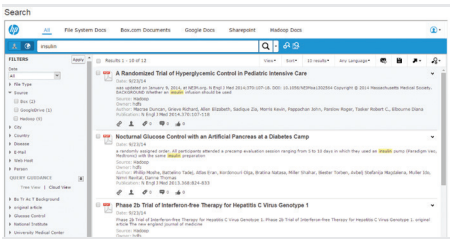
Smart content hub solution architecture



Key benefits of the Smart Content Hub solution

- **Derive insights from 100% of your data**: Text, images, audio, video, and many more data types are automatically consumed and enriched, making it possible to integrate this valuable content and the corresponding insights into various line-of-business applications.
- **Extend the enterprise data warehouse**: Enterprise data warehouses have data that is rigidly formatted, highly cleansed, and suitable for BI-type queries. The Smart Content Hub provides you with access and analysis capabilities to all types of data.
- **Right-size existing content management systems**: Content is spread across multiple systems within an enterprise. Several of these systems are well past their effective lifespan, but remain in place as a source of information. Organizations can sunset these legacy systems and consolidate information into a Smart Content Hub with HDP. Additionally, you can speed up and extend the life of active systems by archiving data into a Smart Content Hub.
- **Democratize and enable rich multi-dimensional content analysis**: Analysts, business users, and data scientists can all search and analyze Hadoop data with ease. The open architecture allows you to benefit from the latest innovations without having to move data or establish separate infrastructure.
- **Gain a unified view of your data**: You can search and analyze data in Hadoop alongside other enterprise and cloud repositories, to gain a single, unified view of the information landscape.

HP IDOL works seamlessly with Hortonworks to facilitate the search and analysis of Hadoop data, as well as data from other repositories, through a single portal. In the below example, a search for "insulin" returned documents from Box, Google Drive, and Hadoop in one consolidated view. HP IDOL has also automatically extracted metadata and entities to further refine the search results. The "Query Guidance" in the bottom left pane displays the related concepts that are present in the search results.



Shine a light on dark data

Massive volumes and varieties of human and machine data are generated every day. This data often sits idle, untapped of its potential, in legacy systems and is most often referred to as dark data. You can put this data in Hadoop, but traditional technologies typically have difficulty even understanding binary files, such as embedded .zip files, .pst files with binary attachments, PDF files, and many others. Rich media content such as audio and video are not understood past their generically descriptive metadata.

The Smart Content Hub from HP and Hortonworks allows you to search and analyze Hadoop data, as well as data from other repositories, through a single portal, giving you unprecedented access and understanding of all your enterprise data. HP IDOL provides advanced contextual search and high-performance analytics with functions like conceptual search, clustering, categorization, and classification of complex content. You can also extract rich metadata and text for SQL-based analytics with Hive/PIG.

Following are a sampling of IDOL functions:

- **Entity extraction**: IDOL automatically identifies and extracts terms in documents that lend themselves to key fields such as the names of companies or people, locations, addresses, and telephone numbers. IDOL 10 offers hundreds of entities out-of-the-box across numerous languages.
- **Sentiment analysis**: You can gain an understanding of what customers, business partners, or investors really think of your company or your products. IDOL recognizes positive and negative comments, as well as emotions and opinions expressed by people across sources such as news feeds, logs, and social networks, as well as other sources.
- **Clustering**: IDOL can take a large set of data and automatically partition it so that similar information, even of varying formats, is clustered together. Each cluster represents a concept area, making it easier for you to identify inherent themes and emerging trends.
- **Image analytics**: IDOL provides numerous image functions, including OCR, the ability to detect subtle patterns in images, the identification of the same images from different angles, and recognition and analysis of objects such as faces, bodies, gender, age range, expressions, and clothing. IDOL can also extract key data fields from scanned documents.

Democratize data analysis

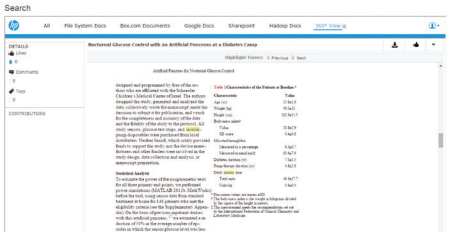
The key to realizing the full value of content is the ability to enable a shared infrastructure that supports a multitude of analysis. The Smart Content Hub offers enterprises an open platform to enable fit-for-purpose content analysis. You can give analysts and data scientist the richest set of analytics functions to have at their disposal through the Smart Content Hub. Instead of command line navigation, we provide folder-based navigation. You can search and better understand the content using a standard interface, and view files with a near-native viewing capability that preserves the basic formatting of the document, including tables and graphics.

Build, enrich, and maintain a clean Hadoop data lake

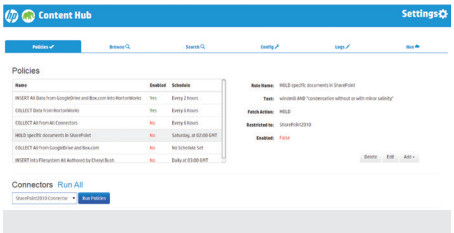
Many businesses like yours are trying to use Hadoop as their data lake—a place to store practically unlimited amounts of data of any format or schema—for its massive scalability and cost effectiveness. But wouldn't it be better if you could spend more time analyzing data and less effort getting data into Hadoop file systems? IDOL connectors make it simple to move data in and out of your Hadoop cluster, while staying in sync with the source systems. A few capabilities to help you streamline processes:

- Quickly choose which repository you want to connect to, and set up rules and policies for customized data ingestion
- Ingest a copy of the content or the analytic remnant of the content
- Push the data into a remote repository, collect data to a local directory, delete matching files from the repository, and even mark files as undeletable to comply with legal hold requirements
- Automatically propagate updates from the original repository to the data lake
- Extract and tag information from the data as it moves into Hadoop so that you can run native Hadoop jobs and analysis on the extra metadata

The Smart Content Hub allows you to preview any document (PDF is shown below) without the need to open the native application. The searched term ("insulin") is highlighted throughout the document.



Easily push data into and out of Hadoop from a variety of different repositories using a wide range of customizable parameters. With a simple click, you can take files from a remote repository, extract the text and metadata, and load it into Hadoop.



Maintain persistent enterprise-level security

When dealing with data from multiple repositories and various users, it is critical that sensitive information is protected and access rights are respected. The Smart Content Hub maintains source-based information security in your data lake. All different security models from different repositories are bound to the index, so that users can only see what they are entitled to see. User-level access controls include a rich set of policies that allow you to enforce user rights with every data movement, and provide secure data management.

About Hortonworks

Hortonworks is the only 100% open source software provider to develop, distribute, and support an Apache Hadoop platform explicitly architected, built, and tested for enterprise-grade deployments. Hortonworks Data Platform provides an open and stable foundation for enterprises and a growing ecosystem to build and deploy big data solutions. Hortonworks is the Hadoop of choice by many premier data center managers and provides an excellent companion to HP IDOL for information processing and analytics on Hadoop data.

About HP

HP creates new possibilities for technology to have a meaningful impact on people, businesses, governments and society. With the broadest technology portfolio spanning printing, personal systems, software, services and IT infrastructure, HP delivers solutions for customers' most complex challenges in every region of the world.

More information about HP (NYSE: HPQ) is available at **http://hp.com.**

Sign up for updates
hp.com/go/getupdated

Share with colleagues