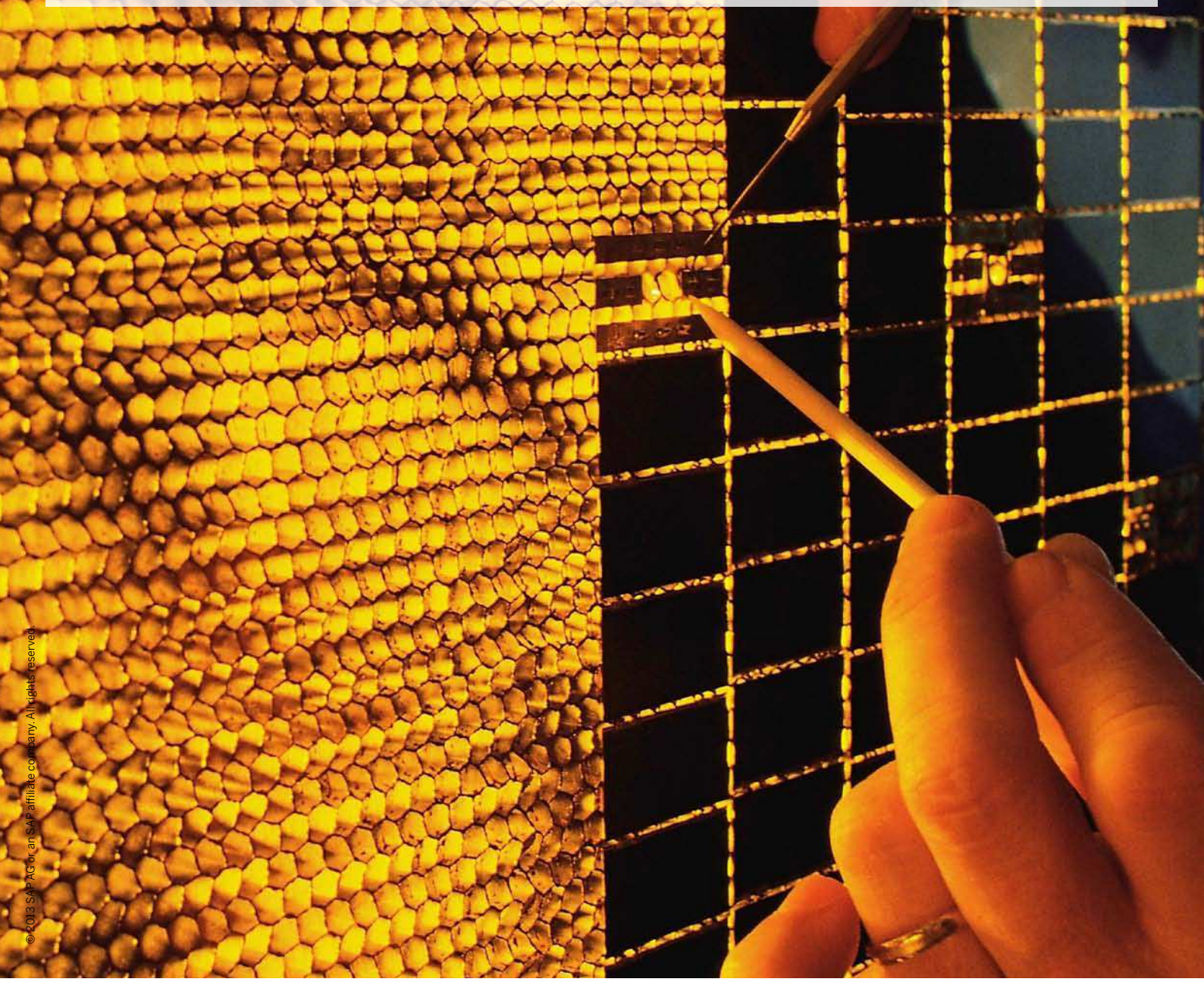
SAP Real-Time Data Platform

Combining SAP® Real-Time Data Platform with Hortonworks Data Platform

Reference Architectures, Best Practices, and Case Studies for Analyzing "Big Data"



SAP®

Hortonworks

Table of Contents

Faster, Better Insight at a Lower Total Cost of Ownership

SAP® Real-Time Data Platform in combination with Hortonworks Data Platform delivers an IT architecture for analyzing massive data sets in real time. This paper outlines three reference architectures as well as best practices to ease implementation.

Enterprises have access to vast amounts of data, and in many cases they have accumulated this data for years for regulatory and compliance reasons. But there are many challenges when deriving value from this data. Even as enterprises work through these challenges and start to expose value in greater amounts of transactional information, there is also a growing desire to expand the collection of unstructured data and to use new advanced analysis techniques across all forms of data. To achieve this, a new data management architecture is required.

SAP provides a best-in-class portfolio of databases, information management solutions, analytic tools, and analytic applications. Now SAP has partnered with Hortonworks to enable integration of Apache Hadoop into SAP Real-Time Data Platform using the Hortonworks Data Platform to facilitate business intelligence (BI) and analysis of "Big Data."

With SAP and Hortonworks software, enterprises can retain and process more data, join new and existing data sets, and lower the total cost to perform this net-new analysis. While enterprises can build petabyte-scale environments for transactional data with analytic databases, these solutions are not always well suited to handling nontraditional data sets, such as text, images, machine data, and online data. Hadoop can help enterprises handle nontraditional data when they embrace the following principles:

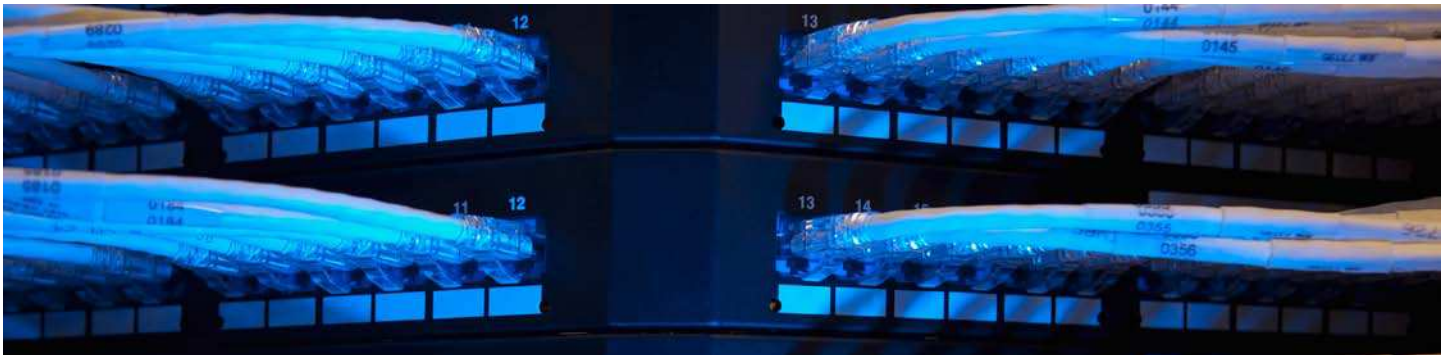
- **Retain as much data as possible**. Traditional data warehouses age and will eventually store only summary data over time. Analyzing detailed records is often critical to uncovering business insights.
- **Archive data at low cost**. It is not always clear whether much of the data stored will be of value in future analysis. Therefore, it is hard to justify expensive processes to capture, cleanse, and store that data.
- **Access all of the data efficiently**. Data needs to be readily accessible. Apache Hadoop clusters can provide a low-cost solution for storing massive data sets while still making the information readily available. Hadoop is designed to efficiently scan all the data, which is complementary to databases that are efficient at finding subsets of data.
- **Apply basic data cleansing and data cataloging**. Categorize and label all the data inside the Hadoop platform with enough descriptive information (or metadata) to make sense of it later and to be able to integrate it with transactional databases and analytic tools. This will greatly reduce the time and effort of harmonizing data with other data sets, and it will avoid the eventuality that valuable data is accidentally rendered useless in the future.

USE CASE: HIGH-TECH MANUFACTURER OPTIMIZES CUSTOMER EXPERIENCE

In this example, a major hardware manufacturer has operated on the combination of the SAP ERP application, Oracle RAC, and SAP Sybase® IQ software for years. The company's business processes, from customer relationship management (CRM) to inventory management, manufacturing, and fulfillment, all run on SAP software. Oracle RAC supports the system's transactional data flow, and SAP analytics solutions are used to analyze and report on data stored in SAP Sybase IQ. This two-database architecture helps improve throughput by separating out transactional and analytic workloads.

The challenges that the business sought Hadoop to address are as follows:

- Its data does not fit neatly in a relational format. The company gathers more than one hundred million surveys each year. The most valuable data is in the "comments" field in these forms, and to date, the business has not analyzed this information since it is somewhat unstructured.



- The business cannot view data across departments. Customer training data, for example, is not typically joined across departments with the call center's CRM application to help tailor a support call to the customer's expertise.
- Even if custom solutions are built to handle free-form, unstructured data like comment fields, and custom logic associates training and certification data with CRM data, there is no model to deal with the next unstructured data set or join together previously unrelated data in a powerful manner.

The company chose to implement Hortonworks Hadoop Data Platform to refine previously unstructured data sets and to begin to explore the relationships among previously unrelated data. Within the first half of the year, these explorations proved valuable. Today, the company enriches the view of the customer over time and across systems to improve customer satisfaction, which leads to improved retention and repeat business.

New business capabilities include:

- Automatic support escalation – Presenting the certification level of individuals calling customer support and automatically escalating experienced customers to advanced technical support
- Improved customer records – Creating a more accurate notion of customer master ID that crosses systems and process boundaries. Heuristics are in fact used to determine the relationship of a form being filled out on the Web site to a customer record in the SAP database.
- Better customer insight – Accurately assessing customer satisfaction over time and across individuals, teams, and departments to help target future sales and product efforts
- Improved customer support – Gathering usage information from devices when authorized and associating devices with the support organization to shift customer support to a fully proactive model. Now devices can be fixed or patched before they exhibit issues, and problem devices can be replaced without significant risk to long-term customer relationships

The following table illustrates how Hadoop helped unlock the value in the manufacturer's deep customer data. Note that the reference architecture patterns listed in the table are defined in detail in the next section of this paper.

With SAP and Hortonworks software, enterprises can retain and process more data, join new and existing data sets, and lower the total cost to perform this net-new analysis.

Use Case	Challenge	SAP® and Hortonworks Solutions	Business Impact	Architectural Pattern
Accurate customer satisfaction levels via analysis of free-form feedback data	Free-form text fields are not easily analyzed using Structured Query Language	• Natural language processing (NLP) framework parses free-form text into indexed, queryable data • Machine-learning framework rates customer's satisfaction based on keyword analysis	• Lowered cost of feedback by reducing need to call customers for feedback • Accurate product requirements captured	SAP: Data warehousing Hortonworks: Refinery
Better routing of customer support calls	Customer info stored at an account level, and training and certification is stored for the individual	• Accurately associate individuals with customer accounts on large scale • Analyze call log data to compute phone number to customer relationships	• Improved customer satisfaction • Reduced per-call duration and costs	SAP: Data warehousing Hortonworks: Refinery, enrichment
Association of Web site visitor with CRM customer ID	Web user identity is obscured, not easily identified from transactional data	• Data-clustering analysis determines association between Web user, customer login, and IP address information • Association between Web visitor and customer ID (many-to-one relationship) is stored in fast, Web-scale database	• Anonymous Web visitors no longer anonymous • Web site personalization, showing products customer owns first, and so on	SAP: "Big Data" and real-time analytics Hortonworks: Refinement, exploration, enrichment
Gathering device sensor data	Volume of semistructured data from more than 10 million devices every minute is not easily stored in traditional database	• Store the raw volume of sensor data • Group and insert the data into database at summary level • Compute device health	• Answer questions about feature usage and priority • Make device data available to call center • Lower field-warranty and repair costs	SAP: Scheduled reporting Hortonworks: Refinement, enrichment

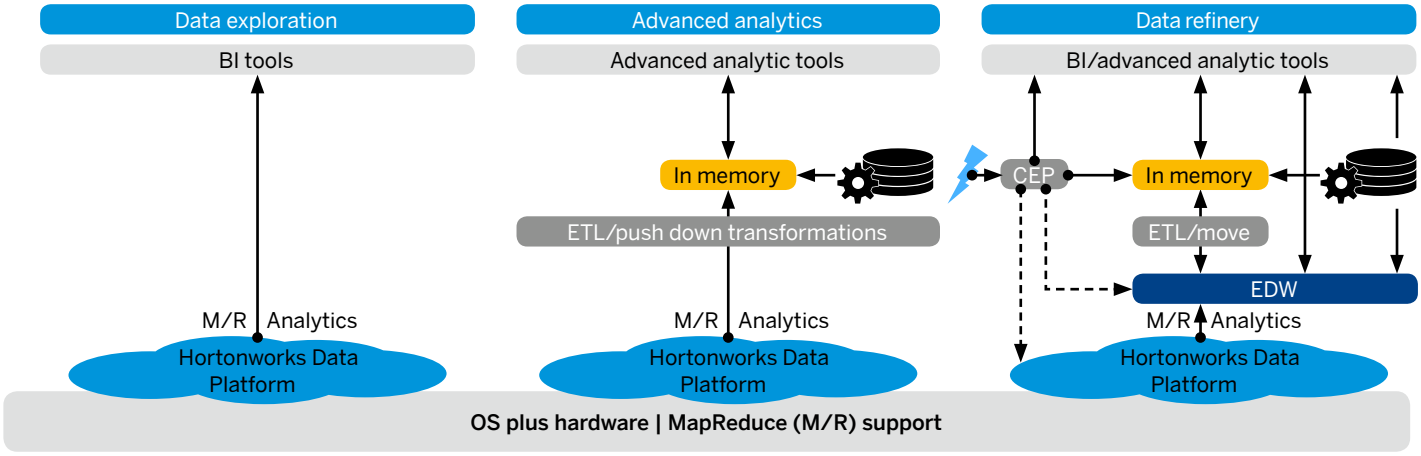
Reference Architectures

Enterprises can combine Hortonworks Data Platform with SAP solutions by following one of three reference architectures:

- Data exploration
- Advanced analytics
- Data refinery

See the figure for details.

Figure: Three Scenarios for "Big Data" Implementation



BI = business intelligence
CEP = complex event processing
EDW = enterprise data warehouse
ETL = extract, transform, and load

SAP AND HORTONWORKS OFFERINGS FOR DATA EXPLORATION

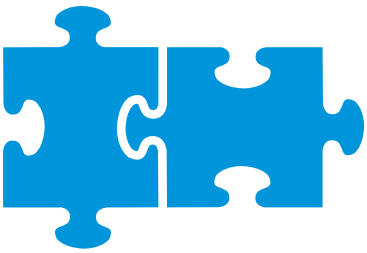
Before doing the work to produce a scheduled report, BI analysts may first want to explore data stored in Hadoop, as well as other data sources. The SAP BusinessObjects™ BI platform provides a robust, powerful, and standards-based way to access Hadoop environments, using the HIVE interface. HIVE provides a Structured Query Language (SQL)–like interface to the Hadoop Distributed File System (HDFS) and can be configured as a data source through semantic layers in the SAP BusinessObjects BI platform. Hortonworks Data Platform provides you with not just Hadoop but also the connectors and interfaces to connect to the SAP BusinessObjects BI platform.

Sometimes, however, before you can look at Hadoop data via tools in SAP BusinessObjects BI and use SQL for queries, you must look at the data first in Hadoop-native form. From statistics and mathematical analytics products like Revolution Analytics to machine-learning tools like Mahout, some analysis is best augmented with Hadoop-native tools first. With an approach using integrated solutions from SAP and Hortonworks, you get the best of both worlds: SQL query and reporting as well as Hadoop-native analysis.

Key benefits of Hadoop for data exploration in SAP landscapes:

- Improve the quality of reports by first exploring the validity of the data at large scale
- Reduce the cost of extract, transform, and load coding by eliminating wasted development cycles
- Store all raw data inside Hadoop without expensive preprocessing
- Store dashboard snapshots indefinitely
- Create prototype reports to validate that key performance indicators are accurately being tracked

The power of Hadoop has been brought to bear through Hortonworks Data Platform and SAP solutions to provide an end-to-end approach for analyzing business data at low total cost of ownership.



SAP AND HORTONWORKS OFFERINGS FOR ADVANCED ANALYTICS

Depending on the analysis performed, data scientists will want to be able to complete their analysis in Hadoop using MapReduce algorithms, or in memory by exporting relevant data from Hadoop into the SAP HANA® database using SAP Data Services software.

Using SAP Data Services with the MapReduce framework, business analysts can define queries to push down text transforms to the Hadoop cluster. Using the native processing capabilities of Hadoop, only relevant text data is extracted and loaded rapidly into SAP HANA for deeper analysis with structured data.

Key benefits of Hadoop for advanced analytics in SAP landscapes are as follows:

- Improve end-user experience by reducing the number of interactions required to deliver value
- Feed otherwise complex analytical models straight into transactional and online application logic in application-native form
- Report against columnar and application-native data without first converting to relational form
- Enrich the view of the business to longer time windows and unstructured data types
- Report against data sets that were formed or computed at runtime, such as product recommendations that are typically unavailable in data warehouses
- Reduce the amount of data moved between systems to generate reports
- Increase the amount of data being summarized beyond that which the warehouse can handle

SAP AND HORTONWORKS OFFERINGS AS A BIG DATA REFINERY

In the refinery model, data starts upstream in SAP software from your enterprise resource planning system, which contains customer data, product data, supply chain information, and process execution data. Often this data is stored in an enterprise data warehouse, such as the SAP NetWeaver® Business Warehouse application powered by SAP HANA or SAP Sybase IQ software with native Hadoop connectors. Upstream data can also come from Weblogs or sensor network data on the manufacturing floor. In addition, relevant data can be extracted from large volumes of data processed efficiently in Hadoop. As a result, key insights can be discovered by joining data from disparate data sets and emitting outputs in an easy-to-consume form for analytics environments.

Key benefits of Hadoop as a refinery in SAP landscapes are as follows:

- Report over large data sets and time windows moving on relevant data
- Improve cost of retention
- Retain more data in original format, outside the warehouse at low cost
- Forgo the unnecessary notion of aging or rolling up data
- Convert the data on an as-needed basis to a format specific to various different calculations and reports (schema on read)
- View otherwise unstructured Hadoop data as structured and queryable using existing SAP tools and functionality
- Transform complex multimedia or text data before loading into your warehouse

Using Hadoop and Engaging Hortonworks and SAP

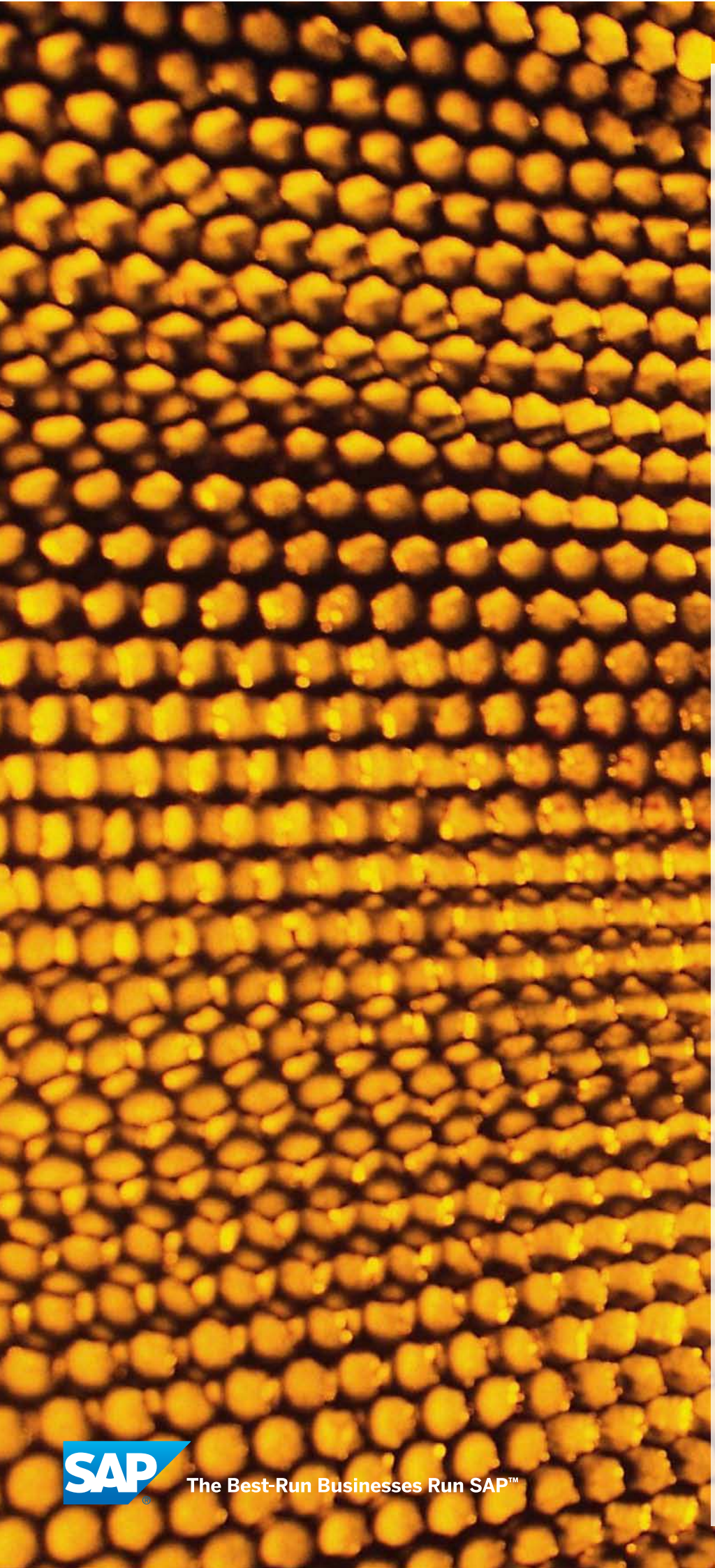
The power of Hadoop has been brought to bear through Hortonworks Data Platform and SAP solutions to provide an end-to-end approach for analyzing business data at low total cost of ownership. As we have illustrated, unifying these platforms provides new ways to:

- Refine and transform data for storage and analysis in traditional warehousing environments
- Enable corporations to explore new aspects of new types of data
- Enrich online applications so they can be more responsive to end-user demands

With SAP Real-Time Data Platform and Hortonworks Data Platform, you can gain valuable insight from Hadoop. Use the following table to identify the patterns covered in this paper and accelerate your adoption.

Learn More
To find out more, please visit www54.sap.com/solutions/big-data.html.

Analytics Function	Value Add of Big Data		
	Refine	**Explore**	**Enrich**
Data exploration	Refine reports with information from new sources, never before considered	Explore root source of data and patterns and run what-if alternate scenarios	Join rich context about users, products, and processes from online systems into reports
Data refinery	Use Hadoop where extract, transform, and load has become too slow or expensive on large, complex data	Before setting up reports and tables in warehouses, use Hadoop to find the most valuable views; avoid aging out or dropping data without increasing warehouse costs	Take advantage of robust data models stored in application-native form, which are more expressive than SQL tables and, with Apache HCatalog, can be easily joined to SQL analytics logic
Advanced analytics	Retain raw streams on disk to be added to complex analytics later	Retain the entire business history, enabling you to study the relative value of past data before investing in advanced analysis	Add large time-windows and time-series data that otherwise is too big to stream via Hadoop



www.sap.com/contactsap

CMP24469 (13/02)

SAP®

The Best-Run Businesses Run SAP™