Enterprise Data Warehouse

# Meeting the New Data Challenge with SAP and Cloudera

## How Data Warehousing Environments Are Evolving to Meet New Data Requirements

© 2013 SAP AG or an SAP affiliate company. All rights reserved.

# Table of Contents

3 **Meeting the New Data Challenge with SAP and Cloudera**

Integrating with Existing Data Management Infrastructure

4 **Enterprise Architectures Need to Evolve**

Data Marts Encompassing Structured and Unstructured Stores

Need for Capabilities to Integrate

New Applications Need "Big Data" and Real-Time Functionality

6 **A New Platform for Data**

Scalability

Flexibility

Economy

7 **The Analytic Data Hub**

Use Any Kind of Processing

Integrate with Existing and New Relational Technologies

8 **Benefits of a Combined Platform**

Improving Operational Efficiency

9 **The Combination of SAP and Cloudera Offerings**

An Innovative Approach for Forward-Thinking Organizations

© 2013 SAP AG or an SAP affiliate company. All rights reserved.

# Meeting the New Data Challenge with SAP and Cloudera

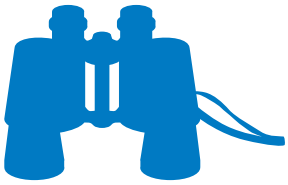
Modern enterprises face a daunting challenge: explosive data growth at an unprecedented scale. In 2010 enterprises created 800 petabytes of new data. In 2011 that volume ballooned to 1.2 zettabytes and more than doubled to 2.7 zettabytes in 2012. The vast majority of this new data is unstructured and generated from social media streams, log files, sensor and telemetric readings, images, and other nontransactional data sources. Enterprises are struggling to produce meaningful business value from these growing unstructured archives. Moreover, this data is often produced outside the typical boundaries of the enterprise.

## INTEGRATING WITH EXISTING DATA MANAGEMENT INFRASTRUCTURE

Enterprises are today at a crossroads in the evolution of their enterprise data architecture, given these core issues. Organizations need to address three dimensions when evaluating a solution:

- What is the best way to generate high-quality insights from poorly accessible data archives within the enterprise?
- What is the best way to extend enterprise practices to include new forms of data found both inside and outside the organization?
- What is the best way to adapt practices so that the enterprise is able to profit from current investments and establish new baselines for future business?

Elastic storage with integrated distributed compute – a combination both reliable and robust against component and compute failure – makes Apache Hadoop an attractive option to explore. However, Hadoop executes these core capabilities in isolation. To bring these capabilities into the enterprise, organizations need to gracefully integrate the technology into existing business intelligence (BI) and analytic environments.



To lead practical improvements in the IT landscape, organizations should gain a broad perspective of the new data requirements in the enterprise and how data warehousing environments are evolving to meet them.



© 2013 SAP AG or an SAP affiliate company. All rights reserved.

# Enterprise Architectures Need to Evolve

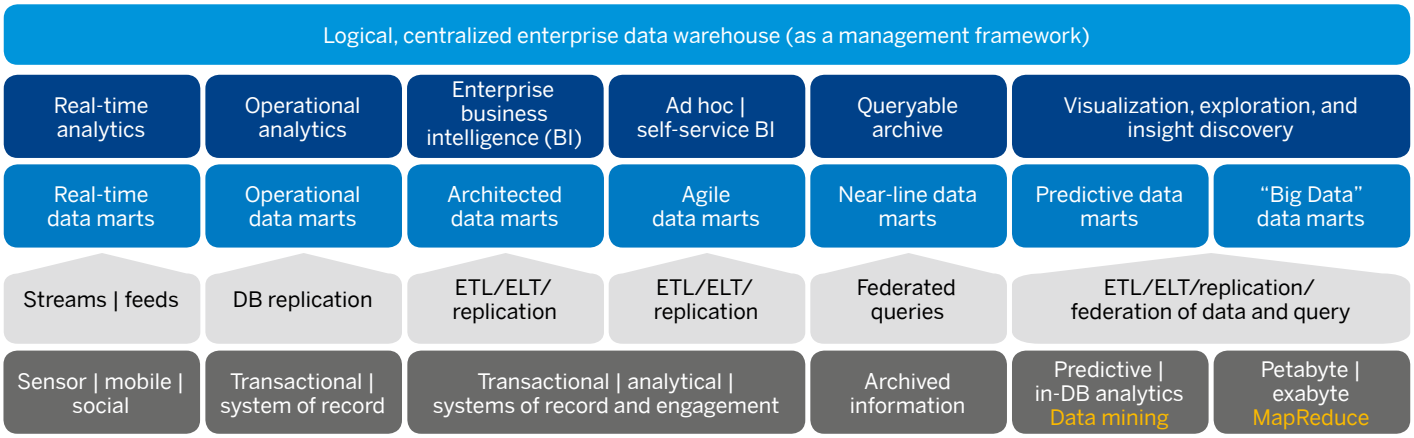
To lead practical improvements in the IT landscape, organizations should gain a broad perspective of the new data requirements in the enterprise and how data warehousing environments are evolving to meet them. In this section, we will outline current and emerging data warehousing patterns and how SAP and Cloudera are embracing these new architectures.

## DATA MARTS ENCOMPASSING STRUCTURED AND UNSTRUCTURED STORES

A particularly useful way to examine the data warehouse ecosystem is to catalog the existing and evolving architectural patterns by the speed at which insights are needed (see Figure 1). The data mart architectures are organized from left to right based on the speed at which insight can be offered to the business:

- **Streaming and real-time analytics** are data marts that focus on generating insights directly from the incoming stream, prior to or during storage. Complex event processing (CEP) engines and specialized NoSQL/key-value systems are well suited for these low-latency scenarios where a priori patterns signal alerts that in turn trigger an appropriate response.
- **Operational analytics** has transitioned to a type of data mart that is generated via real-time replication. While this data is available in reports from operational and transactional systems, enterprises that have high transactional throughput often seek to improve latency by isolating analysis in a replicated analytical system dedicated to this activity.
- **Architected data marts** are the fundamental, value-generating BI models that often require careful planning between business and IT to generate high-quality reports. Since the quality requirements are substantial and global consolidation of information is paramount, organizations focus significant effort on cleansing, transforming, and harmonizing the data. These data warehouses are the central corporate memory and are often referred to as the single source of truth.
- **Agile data marts** serve the savvy business users' growing need to explore data and modeling options in a "search-and-experiment" paradigm. Within these data marts, IT offers a larger and more flexible "universe" for analytics yet still maintains control of the primary definitions and overall management. Business users are free to explore the data within the data warehouse and "mash up" with data from outside the warehouse.

**Figure 1: A Logical, Centralized Enterprise Data Warehouse**




ETL = extract, transform, and load
ELT = extract, load, and transform



© 2013 SAP AG or an SAP affiliate company. All rights reserved.

Distributed as a single platform, the Cloudera Distribution including Apache Hadoop bundles Hadoop and its eco-system into an enterprise package well suited for the new era of Big Data management.

- **Queryable archives** are essentially near-line stores that are augmenting the current range of data marts. These scalable systems allow for historical analytics and trending analysis while supporting the mandatory compliance and regulatory requirements for past business.
- **"Big Data" data marts** and **predictive data marts** are emerging data marts and are the focus of this paper. While many enterprises have used alternative technologies to implement these patterns, Hadoop combined with analytic databases offer a way for organizations to address these solutions as first-class data mart scenarios.

## NEED FOR CAPABILITIES TO INTEGRATE

In order to bring the value of Hadoop to the enterprise, IT organizations need clear connectivity models. The connectivity models are grouped into three types:

- Allow direct connectivity from clients into the Hadoop cluster, thereby enabling Hadoop as a first-class BI data source.
- Integrate into the enterprise information management tool set and offer the rich capabilities of data movement, transformation, and governance afforded by these systems, thereby establishing Hadoop as a first-class data source available to managed data marts.
- Offer mechanisms to federate queries on the fly from a data warehouse into Hadoop, thereby extending the SQL paradigm to naturally embrace Hadoop as a near-line queryable store. A similar native mechanism to federate data is to set up on-the-fly, transient tables for the lifetime of a query.

## NEW APPLICATIONS NEED BIG DATA AND REAL-TIME FUNCTIONALITY

Armed with these Hadoop integration means, enterprises can define innovative analytical applications and data marts. Before exploring potential applications, let us review the current data mart models in light of the Hadoop connectivity options:

- **Streaming and real-time data marts** – Although the core function of streaming data marts works on the principle of "data in motion," organizations need to design and tune the event patterns, and this exercise benefits from large data samples. In this capacity, Hadoop can act as a vast data store for the raw events during CEP query modeling.
- **Enterprise BI** – Since these data marts are typically heavily engineered systems, Hadoop is best incorporated as a complementary approach to data management that shares the processing burden with the BI data mart. Until enterprise BI practices fully embrace Hadoop's approach to data management, these data marts might initially consider Hadoop simply as another source system, albeit one containing refined results.
- **Agile data marts** – Hadoop today is largely viewed as a batch processing system. However, recent developments within the Hadoop community, such as Cloudera Impala, a low-latency query engine for Hadoop using familiar Structured Query Language (SQL) operators, show promise for direct and iterative interactions between a BI client and Hadoop.
- **Queryable archives** – By definition, the queryable archives can immediately benefit when large volumes of transactional information are housed within elastic storage like Hadoop and provisioned within the query federation support mechanics.
- **Big Data data marts and predictive data marts** – These are natural fits for Hadoop and analytic databases. In these contexts, data center design readiness and the integration of data mart practices are paramount. In addition, the integration of unstructured and structured data demands smart engineering capabilities and planning.



© 2013 SAP AG or an SAP affiliate company. All rights reserved.

# A New Platform for Data

While enterprises are considering a new building block – Hadoop – to modernize their data warehousing environment for Big Data, many organizations are still unfamiliar with the platform and its various systems and tools. Developed by Cloudera architect Doug Cutting, Hadoop is open source software that enables distributed storage and parallel processing of huge amounts of varied, multistructured data using industry standard servers. Hadoop also includes a rich and diverse ecosystem of supporting systems and applications.

Distributed as a single platform, the Cloudera Distribution including Apache Hadoop (CDH) bundles Hadoop and its ecosystem into an enterprise package well suited for the new era of Big Data management. CDH offers a set of capabilities that make it the standard platform for Big Data processing and storage.

## SCALABILITY

The core of Hadoop includes a system that marries fault-tolerant, high-bandwidth clustered storage (Hadoop Distributed File System, or HDFS) with fault-tolerant distributed processing (MapReduce). HDFS and MapReduce are the core components of processing power and storage capacity. Both are designed to scale linearly as additional servers and data are added to the Hadoop cluster. For example, the same MapReduce code that runs on 10 GB of data can run on 10 PB of data without any changes by the developer; all Hadoop requires is additional servers that contain both data storage and compute. Deploying Hadoop provides organizations with near limitless data storage and processing.

## FLEXIBILITY

The underlying storage in Hadoop, HDFS, is a flexible file system that can accept any data format and store this raw information in perpetuity. Hadoop supports pluggable serialization for efficient and reliable storage in the original raw format. As a result, if an application needs to reprocess the data or read the data using a different schema, the original data is both local and in its high-fidelity, native state, as Hadoop does not enforce any normalization or restructuring of data for storage.

MapReduce and other compute frameworks like Impala allow developers to decide how to parse data and handle errors in data quality each time data is read. Thus, the application both avoids costly retrievals from the source systems and gains full format control while maintaining data precision. This process of formatting the data at query time is known as "schema on read." It is one of the primary advantages Hadoop offers over traditional systems that require data to be formatted first to a "schema on write" before processing and storage.

## ECONOMY

Hadoop employs industry-standard hardware, with the cost per terabyte of Hadoop-based storage running 10 to 100 times less than a traditional relational data warehouse system. Hadoop uses servers with local storage, thereby optimizing for high I/O workloads. Servers are connected using gigabit networking, which lowers overall system cost and still allows near-limitless storage and processing, thanks to Hadoop's scale-out features. Through its use of local storage and standard networking, Hadoop can reduce the total hardware requirements since a single cluster provides both storage and processing.

0/1

Most EDW environments are optimized for BI tasks rather than advanced analytic techniques or for handling nontraditional data sources.



© 2013 SAP AG or an SAP affiliate company. All rights reserved.

# The Analytic Data Hub

Using CDH as a unified data refinery delivers immediate benefits to all other data management systems, including data warehouses. Hadoop's schemaless design accommodates new data formats easily, and its low-cost scale-out storage is fast and reliable. Both features make Hadoop an ideal data hub for an analytic data warehouse.

Moreover, by combining all the data together into the high-fidelity storage of Hadoop, enterprises gain better coordination and aggregation of their varied formats and sources at the early stages of processing. For example, an enterprise may collect transaction logs, parts catalogs, customer details, and manufacturing sensor data into a single Hadoop cluster and perform an initial cleansing, transformation, and aggregation to the data prior to use by downstream systems, thus saving both network bandwidth and processing capability of the receiving systems.

## USE ANY KIND OF PROCESSING

The underlying facilities in Hadoop (HDFS and MapReduce) provide scalable data storage and primitives for data processing. While any type of data transformation can be written in MapReduce, this is not always the most efficient abstraction for processing data. Today, most transformations are authored in Apache Hive, Apache Pig, or third-party commercial solutions.

Apache Hive is an SQL-like language, officially known as Hive Query Language (HQL). HQL is accessible to the SQL developer, making it organizationally efficient for people with existing SQL skills to write data transformations in Hadoop. Apache Pig is a data flow system with a language (Pig Latin) designed for expressing data-processing logic in a sequential fashion. While foreign to SQL developers, it comes naturally to those who write Python or Perl-based data transformation code, or to those who write in stored procedures such as T-SQL.

## INTEGRATE WITH EXISTING AND NEW RELATIONAL TECHNOLOGIES

Hadoop is not simply an extract, transform, and load (ETL) tool, but it's a platform, and one that supports running ETL processes in parallel. The core components of Hadoop (HDFS and MapReduce) as well as the evolving ecosystem of supporting tools and systems, like Sqoop, Hive, and Impala, offer a rich and flexible foundation for future processing techniques and integration points.

The partnership of SAP and Cloudera offers enterprises compelling solutions to the challenges of Big Data and enterprise data management and analytics.



© 2013 SAP AG or an SAP affiliate company. All rights reserved.

# Benefits of a Combined Platform

Hadoop is a strong complement to existing BI and analytic systems. The combination of these technologies offers enterprises tremendous opportunities to maximize IT investments and expand business capabilities by aligning IT workloads to the strengths of each system.

## IMPROVING OPERATIONAL EFFICIENCY

Enterprises face a growing list of technical and business demands that place strains on their existing infrastructure and impede their operational latitude. These demands take various forms, such as:

- Operational and historical data
- Operational and business analytics
- Reporting and data exploration and analysis
- Data processing and transformation
- Application services

In the past, the best system to meet these operational and analytical workloads has been the enterprise data warehouse (EDW). However, most EDW environments are optimized for BI tasks rather than advanced analytic techniques or for handling nontraditional data sources. Furthermore, the relational model of the EDW is not the ideal methodology for some advanced analytic workloads. These characteristics can result in degraded performance, cumbersome operational mechanics, and costly scale-out requirements if the EDW is forced to accommodate the needs and requests of the business in their entirety.

Enterprises can shift to Hadoop and CDH many of the tasks required to satisfy the increased demands, such as large-scale data processing and transformation and the exploration and analysis of historical data. The now-unburdened EDW is free to focus on its standard workloads, like current operational analytics and real-time reporting, and yet still benefit from the processing and output of the Hadoop cluster.

This architectural pattern has several benefits. These include a lower cost to store massive data sets via industry-standard hardware, faster data transformations of large data sets by profiting from the proximity of data to compute, and a reduction of the data load into the EDW. This results in faster overall ETL processing and greater EDW capacity and agility. In short, each system – EDW and Hadoop – focuses on its strength to achieve the business goals. Hadoop stores the data and renders intermediate results while the analytic database completes the analysis and provides the real-time business services.



© 2013 SAP AG or an SAP affiliate company. All rights reserved.

# The Combination of SAP® and Cloudera Offerings

SAP and Cloudera offer enterprises a rich set of opportunities and applications, and the integration of Hadoop into SAP® Real-Time Data Platform is well under way.

SAP and Cloudera recommend that enterprises initially focus on integrating Hadoop into their reporting infrastructures as an enhanced data refinery. Figure 2 shows a scheme that offers three stepwise opportunities that enterprises can employ to bring the Hadoop advantage to their workloads.

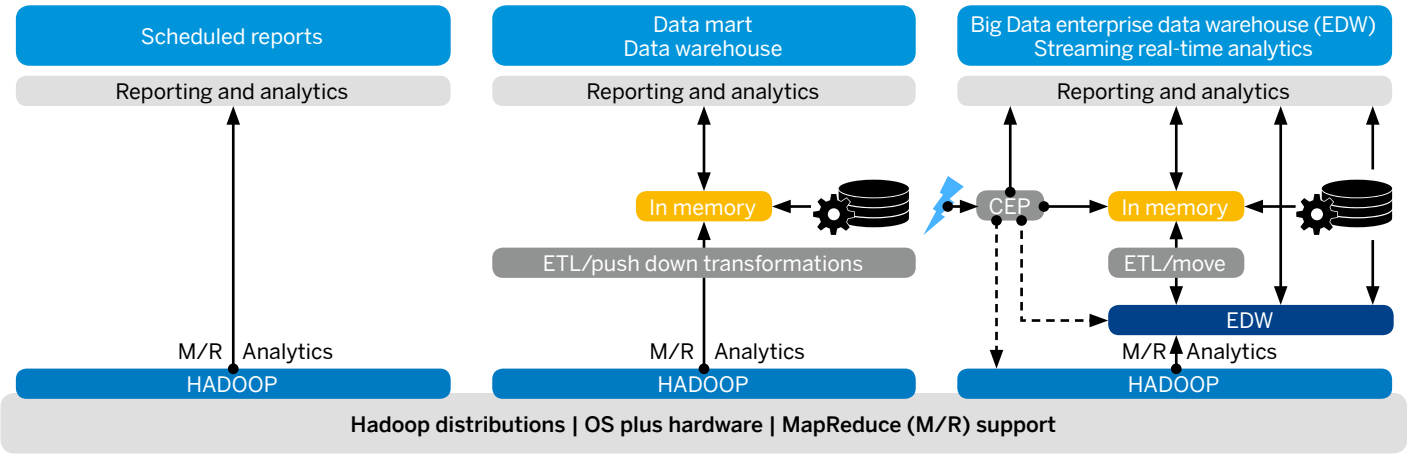
**Data exploration.** Analysts using the SAP BusinessObjects™ BI suite can model and explore data in Hadoop using the GUI tools. The SAP BusinessObjects BI suite automatically generates HiveQL queries to a Hadoop cluster through Hive, which in turn executes MapReduce jobs. To the end user, the Hadoop cluster is simply another data source accessible through the BI tools.

**Advanced analytics.** Connecting the SAP HANA® platform with Hadoop is a powerful combination for advanced analytics. Savvy users can tap into the broad reservoir of data housed within Hadoop, power the majority of the necessary ETL and preprocessing analytics with the cluster's MapReduce framework, and ultimately export the results into SAP HANA for final analytics and direct user manipulation. Administrators may employ SAP Data Services software for data movement and modeling from the native Hadoop tools.

Furthermore, business analysts may use SAP Data Services in conjunction with the MapReduce framework during data transformation. Users can "push down" text transforms from SAP Data Services into the Hadoop cluster for text-related processing. This scenario offers users the native processing capabilities of Hadoop while benefiting from the ease of use and familiarity of SAP Data Services.

**Big Data EDW and streaming real-time analytics.** In this advanced scenario, the combined offerings from SAP and Cloudera span the gamut of possibilities for enterprises seeking maximum value from Big Data. Building upon the data movement and ETL features of the advanced analytics scenario, enterprises can expand their capabilities to include further Big Data computing features with SAP Sybase® IQ software and its native Hadoop connectors. And they can incorporate real-time signals from streaming events with SAP Sybase Event Stream Processor.

**Figure 2: Scenarios for Big Data Implementation**




CEP = complex event processing
ETL = extract, transform, and load



© 2013 SAP AG or an SAP affiliate company. All rights reserved.

These three opportunities, which use Hadoop as the enhanced data refinery, benefit from Cloudera's unparalleled data management platform for CDH, Cloudera Enterprise. With Cloudera Enterprise, administrators gain the advantage of a central Big Data system from a single vendor with:

- A single set of nodes
- A single data repository and metadata model
- A single security and governance framework
- A single, centralized management console
- A single set of dependencies

Cloudera Enterprise and its management system, Cloudera Manager, deliver granular visibility into and control over every part of the CDH environment. Cloudera Manager empowers administrators to improve cluster performance, enhance quality of service, increase compliance, and reduce operational costs. In short, the Cloudera Enterprise portfolio grants organizations the power of Big Data and the economics of open source software with the assurance of 24x7 support and no proprietary vendor lock-in.

## AN INNOVATIVE APPROACH FOR FORWARD-THINKING ORGANIZATIONS

The partnership of SAP and Cloudera offers enterprises compelling solutions to the challenges of Big Data and enterprise data management and analytics. With SAP and Cloudera, forward-thinking enterprises can connect existing business intelligence applications and next-generation analytical systems, like the SAP BusinessObjects BI suite and SAP HANA, directly to their Cloudera-powered Big Data sources.



© 2013 SAP AG or an SAP affiliate company. All rights reserved.

www.sap.com/contactsap

CMP24465 (13/02)
© 2013 SAP AG or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.

Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP AG and its affiliated companies ("SAP Group") for informational purposes only, without representation or warranty of any kind, and SAP Group shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP Group products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and other countries.

Please see http://www.sap.com/corporate-en/legal/copyright/index.epx#trademark for additional trademark information and notices.

The Best-Run Businesses Run SAP™