informatica | cloudera Ask Bigger Questions

White Paper

Data Warehouse Optimization with Hadoop

A Big Data Reference Architecture Using Informatica and Cloudera Technologies

This edition published January 2014

Table of Contents

Executive Summary

Traditional data warehouse environments are being overwhelmed by the soaring volumes and wide variety of data pouring in from cloud, mobile, social media, machine, sensor, and other sources. And the problem will only worsen as big data continues to grow. IT organizations that need to address performance degradation in warehouses approaching their capacity are already considering costly upgrades. However, an upgrade is not the most effective way to manage an excess of seldom-used data. Nor does it save valuable CPU cycles currently consumed by the need to execute compute-intensive extract, load, and transform (ELT) jobs. To keep pace with exploding data volumes, the data warehouse itself needs to evolve.

One emerging strategy is data warehouse optimization using Hadoop as an enterprise data hub to augment an existing warehouse infrastructure. By deploying the Hadoop framework to stage and process raw or rarely used data, you can reserve the warehouse for high-value information frequently accessed by business users. This white paper outlines a new reference architecture for this strategy, jointly developed by Informatica and Cloudera to help organizations speed time to value, maximize productivity, lower costs, and minimize risk.

Leveraging complementary technology from Informatica, the leading provider of data integration software, and Cloudera, the leader in enterprise analytic data management powered by Apache™ Hadoop®, this reference architecture supplies a blueprint for augmenting legacy warehouses to increase capacity and optimize performance. It enables organizations to better capitalize on the business value of big data.

The Need for Data Warehouse Optimization

Today's information-driven business culture challenges organizations to integrate data from a wide variety of sources to improve customer acquisition and retention, increase operational efficiencies, strengthen product and service delivery, and enter new markets. To meet these goals, enterprises demand accessible, timely, actionable data—plus analytical environments that can scale to the tremendous growth in data volume, variety, and velocity while also handling diverse types of enterprise data processing workloads.

Today's data warehouses, however, aren't up to the challenge of meeting these new demands. As data volumes and business complexity increase, the traditional "scale up and scale out" approach of adding infrastructure using high-end servers or appliances with expensive shared storage (e.g., SAN or NAS) has become impractical and far too costly. IT and business leaders must rethink their warehousing strategies to address the inadequacies of existing systems. These inadequacies include the following five issues:

Low-value data that consumes warehouse space. Over time, many warehouses have become bloated with both raw data staged for preprocessing and rarely accessed data that provides little business value.

Inadequate data for business demands. Because of capacity and performance constraints, some warehouses contain only summary data, not the granular and detailed information that the business needs. Users become frustrated when they are unable to access the data necessary to address business questions.

In-warehouse transformations that impair performance. Running data transformations within a warehouse on data staged for preprocessing (i.e., ELT) consumes valuable CPUs, hindering query performance and further diminishing a warehouse's business value.

Network performance that bottlenecks in grids. In grid computing environments, the network can become a bottleneck when large data volumes are pushed to CPU workloads, limiting how much data can be processed in a reasonable time.

Limitations in multi-structured data and schema flexibility. Warehouses based on relational databases are not built to handle the multi-structured datatypes from new big data sources, while schema changes can trigger disruptions and delays.

Real-World Use of Data Warehouse Optimization

Augmenting legacy data warehouses with Hadoop-based solutions optimizes data warehouses, helping to deliver timely data while lowering costs, increasing performance, and meeting the business demands of terabyte- and petabyte-scale big data in virtually every industry—finance, telecommunications, retail, Internet, utilities, oil and gas, healthcare, pharmaceuticals, media and entertainment, and the public sector.

Leading organizations are already putting the Informatica®/Cloudera solution to work:

A **large global financial services and communication company** is cost-effectively scaling the access, storage, and processing of hundreds of terabytes of data from 18 diverse sources. The company, which processes 650 million transactions a year, is improving customer service across 25 global call centers, reducing fraud, identifying trends to guide business decisions, and enhancing product development.

A **large U.S. government agency** facing a fivefold increase in data volumes in the next few years found that 60 percent of its warehouse data was dormant and that 40 percent of CPU capacity was consumed by ELT. Rather than spend millions on infrastructure, the healthcare-related agency uses Informatica and Cloudera to manage data integration across multiple sources, processing up to 25 billion records a day.

A **global media and entertainment company** faced data delivery delays and exorbitant costs in its traditional warehouse as its data increased 20 times annually. With data warehouse optimization, the company anticipates reducing data storage costs by up to 100%, speeding data delivery from 48 hours to 15 minutes, and gaining a 360-degree customer view.

Informatica and Cloudera Converge on Big Data

Cloudera Chief Architect Doug Cutting founded the Apache Hadoop project in 2006 to offer enterprises a way to store and process large volumes of data for Web indexing. The use of Hadoop was compelling for data warehouse optimization which emerged later. The open-source Hadoop framework enables fault-tolerant, distributed parallel processing and storage of huge amounts of multi-structured data across highly available clusters of low-cost commodity servers. Hadoop is ideally suited for large-scale data processing, storage, and complex analytics, often at just 10 percent of the cost of traditional systems.

In the early days of Hadoop, developers had to hand code data integration workloads in new languages. Although Hadoop enabled enterprises to reduce infrastructure costs, the limited availability and high cost of Hadoop developers diminished the value proposition. In 2010, Informatica and Cloudera formed a partnership to create data integration tools for Hadoop. Today, Informatica offers a set of Hadoop-native tools for codeless development and execution of ETL, data integration, and data quality flows.

With Informatica and Cloudera technology, enterprises have improved developer productivity up to five times while eliminating errors that are inevitable in hand coding. Informatica's visual development environment lets enterprises reuse existing Informatica skills for Hadoop big data projects with no further training, substantially increasing return on investment. If organizations need additional resources, they can draw on a global pool of more than 100,000 developers trained on Informatica, as well as comprehensive training and professional services from both Informatica and Cloudera to help reduce cost, speed time to value, and improve project quality. By leveraging the technologies and skills available from Informatica and Cloudera, enterprises can optimize data warehouses using Hadoop as an enterprise data hub:

- Cost-effectively scaling out infrastructure to support unlimited data volumes
- Leveraging commodity hardware and software to lower infrastructure costs
- Using existing and readily available skilled resources to lower operational costs
- Supporting virtually all types of data from both internal and external sources
- Enabling agile methodologies with schema-on-read, rapid prototyping, metadata-driven visual development environments, and collaboration tools
- Integrating with existing and new types of on-premise and cloud infrastructure

Seven Key Processes for Data Warehouse Optimization

As part of their jointly developed reference architecture, Informatica and Cloudera have identified seven fundamental processes for IT architects to consider in mapping out and implementing a data warehouse optimization architecture in two phases. Informatica and Cloudera technologies support each of these steps (see Figure 1).

1. **Offload ELT processing and infrequently used data to Hadoop.** This step alleviates the CPU burden on data warehouses consumed by in-warehouse data transformations in ELT models, and frees space by offloading low-value or infrequently used information.
2. **Batch load raw data to Hadoop.** Instead of feeding data from source systems into the warehouse, raw transactional and multi-structured data is loaded directly into Hadoop, further reducing impact on the warehouse.
3. **Replicate changes and schemas for data.** Entire schemas can be replicated to Hadoop, offloading processing from OLTP and mainframes and operational data stores . Users can further optimize performance and reduce latency by choosing the option of change data capture to move only newly updated information. Because Hadoop doesn't impose schema requirements on data, unstructured information previously unusable by the warehouse can be leveraged in Hadoop.
4. **Collect and stream real-time machine and sensor data.** Data generated by machines and sensors, including application and Web log files, can be collected in real time and streamed directly into Hadoop instead of being staged in a temporary file system—or worse, in the warehouse.

5. **Prepare data for analysis.** Within Hadoop, data can be profiled to better understand its structure and context. Multi-structured and unstructured data (such as Web logs, JSON, sensor data, call detail records, and FIX, HL7, and other industry-specific information) can be parsed to extract features and entities, and data quality techniques can be applied. Prebuilt transformations and data quality and matching rules can be executed natively in Hadoop, preparing data for analysis.
6. **Execute real-time interactive queries.** Cloudera's Impala enables users to run native, real-time SQL queries directly against Hadoop data, sidestepping real-time query limitations of Hive and MapReduce to explore, visualize, and analyze data to discover interesting patterns and trends.
7. **Move high-value curated data into the warehouse.** After data has been cleansed and transformed in Hadoop, high-value data can be moved from Hadoop to the warehouse for direct access by the enterprise's existing BI reports, applications, and users.

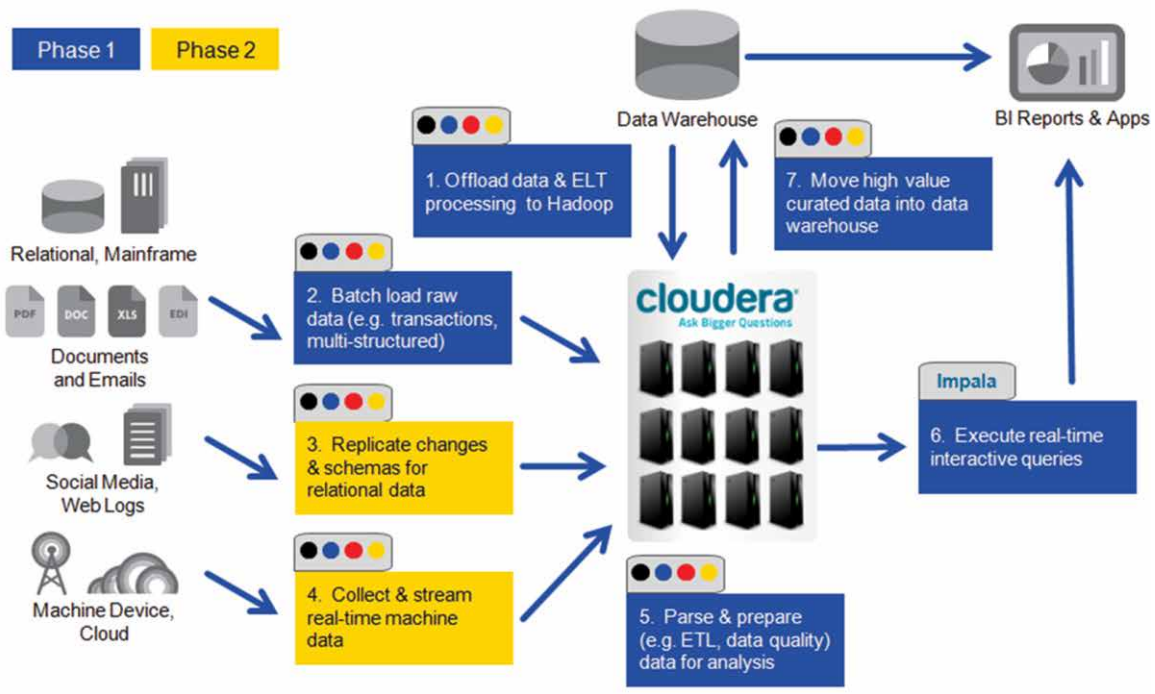

Figure 1. Data warehouse optimization process flow.

A Reference Architecture for Data Warehouse Optimization

At the core of the reference architecture are the Informatica data integration platform, including PowerCenter Big Data Edition and powered by the Informatica Vibe™ embeddable virtual data machine, and CDH, Cloudera's enterprise-ready distribution of Hadoop (see Figure 2). To learn more about Informatica Vibe, please see the white paper, "Informatica and the Vibe Virtual Data Machine: Preparing for the Integrated Information Age." The reference architecture also uses complementary technologies of the Informatica Platform and Cloudera Enterprise, including CDH, the Cloudera Manager management console, and the Impala query tool.

These integrated technologies provide a proven platform for data warehouse optimization, incorporating all these necessary features:

- Universal data access
- Flexible data ingestion
- Streamlined data pipeline design and execution
- Scalable data processing and storage
- End-to-end data management
- Real-time interactive queries
- Enterprise accountability, control, and governance

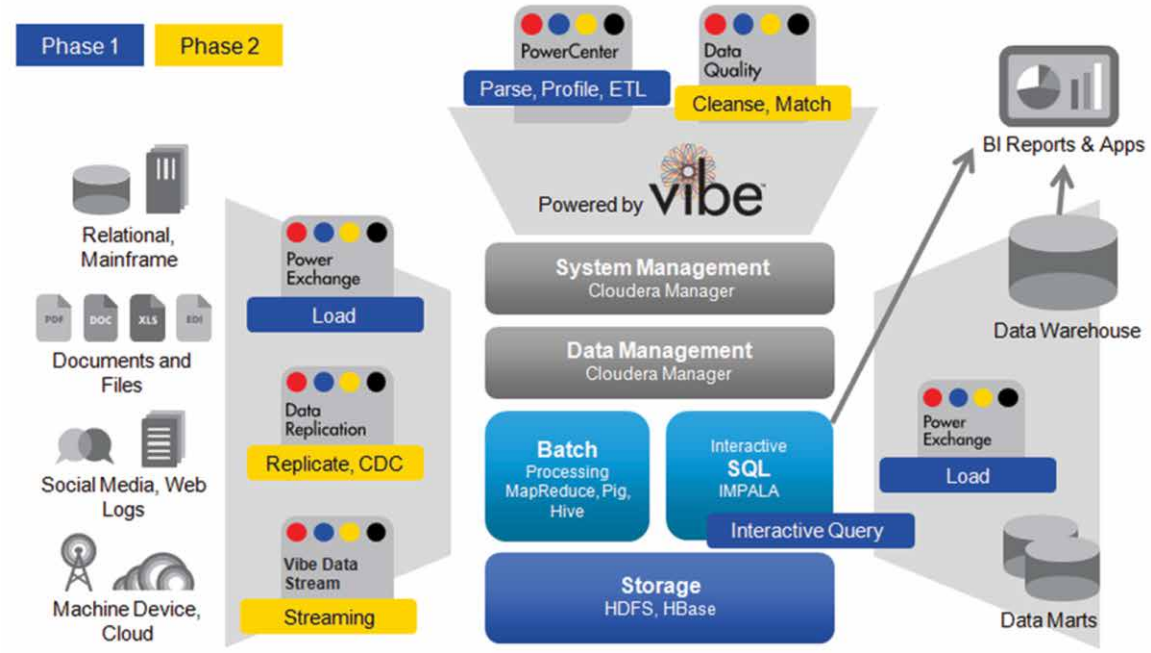

Figure 2. The Informatica/Cloudera data warehouse optimization architecture.

Universal Data Access

To access all the necessary data for processing and move it into Hadoop, developers often resort to building custom adapters and scripts that require expert knowledge of source systems, applications, and data structures. This labor-intensive hand coding is time-consuming and costly to maintain as software versions change. If these adapters lack direct connectivity between the source systems and Hadoop, the data may need to be staged temporarily, increasing storage costs. Custom coding also can't always deliver the needed scalability, performance, and reliability, increasing the risk of noncompliance and system downtime.

Informatica PowerExchange® addresses these issues by accessing data from virtually any source and delivering it directly to Hadoop — or, conversely, delivering data from Hadoop to an enterprise warehouse and applications. In addition to relational data, PowerExchange can access a broad range of other datatypes, including mainframes, NoSQL databases, social media, machine data, email, Word, PDF, HL7, HTML, and other sources, with no need for developers to custom-code access or address data format differences. It furnishes all these necessary capabilities:

- Universal and secure access to virtually any data source or target
- Native connectivity for optimal performance
- Flexible modes of delivery (batch, micro-batch, real time, CDC, replication, streaming)
- High availability and parallel processing using partitioning and grid computing
- Easy wizard-driven setup and a single point of administration

Flexible Data Ingestion

Flexible data movement is a prerequisite for meeting unique business demands, latency requirements, and service-level agreement (SLA) targets. Informatica supplies a range of technologies to optimize movement of various datatypes between sources, while Hadoop offers unprecedented performance and eliminates the need to temporarily stage copies of data.

Batch and micro-batch data load. Conventional data loads enable IT to schedule periodic movements through Informatica PowerCenter® or PowerExchange to Hadoop, with options for partitioning and parallel processing for greater performance and minimal operational impact.

High-speed data replication and CDC. Log-based Informatica Data Replication and CDC software non-invasively replicates data to Hadoop in real time as changes occur in source systems, or to replicate entire schemas or subsets of schemas. Replication lacks the robust transformation capabilities of conventional ETL, but offers the advantages of real-time updates and high-speed data replication with no disruption to operational systems, often in support of operational BI.

Real-time data collection and streaming. Informatica Vibe Data Stream for Machine Data gives enterprises a new way to capture data from the "Internet of Things" for delivery to Hadoop or other targets (see Figure 3). Capable of streaming millions of records per second, the solution enables use of previously hard-to-access machine, sensor, and Web information across industries.

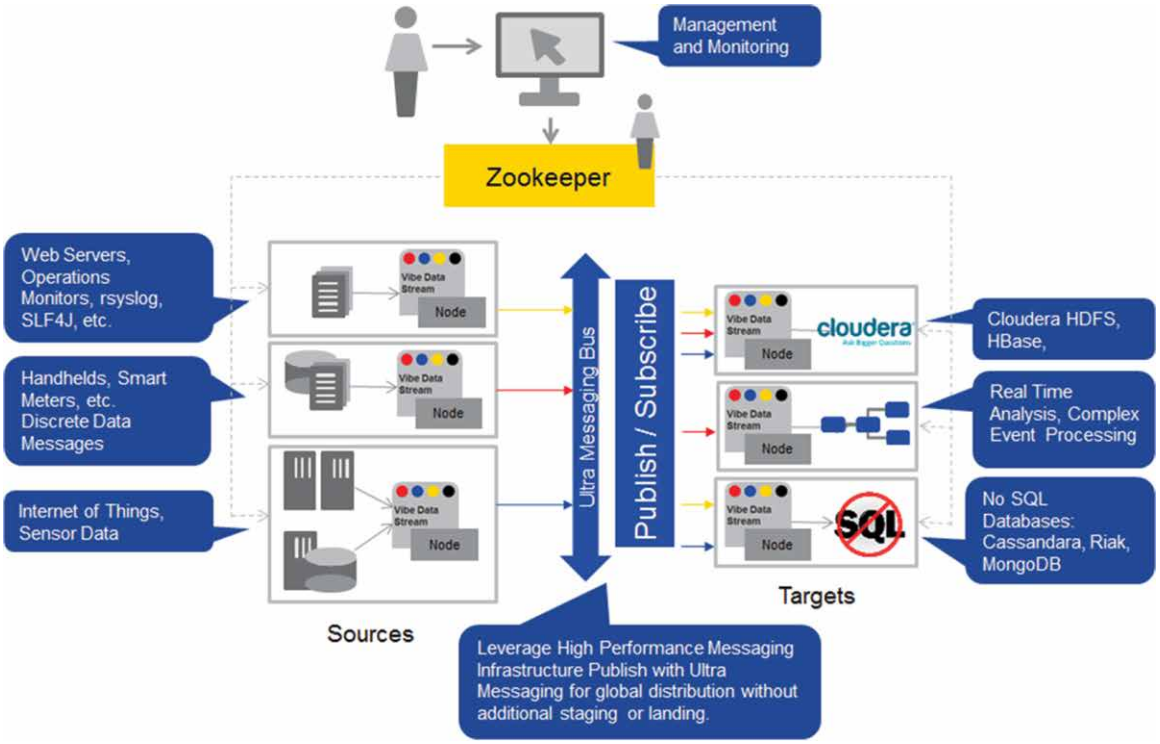

Figure 3. Real-time data collection and streaming into Hadoop.

Streamlined Data Pipeline Design and Execution

The visual user interface of Informatica PowerCenter Big Data Edition simplifies the design and execution of data pipelines, or mappings, by eliminating the need to know MapReduce, Hive, Pig, or other technologies. Developers can rapidly devise data pipelines to extract, transform, load, and cleanse data to and from Hadoop and execute pipelines without manual intervention.

When these data pipelines are deployed and run, Informatica optimizes the end-to-end flow from source to target by generating Hive query language (HiveQL) scripts. Transformations that don't map to HiveQL (e.g., name and address cleansing routines) are run as user-defined functions via Informatica Vibe libraries residing on each Hadoop node.

Because design is separated from deployment, existing PowerCenter mappings can be run on Hadoop. Informatica executes all mapping logic natively on Hadoop. The entire end-to-end data pipeline is optimized for best performance by rearranging mapping logic for maximum efficiency, using HiveQL as a MapReduce translation layer and the user defined function (UDF) framework to execute logic that does not translate to HiveQL.

Instead of using the Hive server for execution, Informatica achieves maximum performance by only using optimized HiveQL scripts as a mechanism to translate SQL-like transformations into MapReduce and to execute UDFs for non-SQL logic such as expression transforms and data quality rules. In other words, the entire data pipeline is executed natively on Hadoop via MapReduce.

For data that may not reside on the Hadoop cluster (e.g., reference data, lookups, dimension keys, etc.), Informatica streams the data from the source (e.g., RDBMS) into Hadoop for processing. Resulting data sets can then be delivered to the target system, whether on Hadoop or another system (e.g., data warehouse or data mart).

As Figure 4 illustrates, the data pipeline has a relational source, a flat file source in a Linux file system, and a flat file source in Hadoop. The target of the data pipeline is a flat file in Hadoop. In the case of a relational source, Informatica uses native connectors to join to the relational database and temporarily stage the data on the fly in Hadoop. In the case of a flat file source on the Linux file system, a bit-by-bit copy of the file is temporarily staged in Hadoop.

Data is moved from the sources and processed before the results are written to the target. After the data is written to the Hadoop target result file, all temporary data staged on Hadoop is removed automatically. The same processing and staging of source data happens whether the sources are relational, flat files, application data, or semi-structured data in JSON/XML formats.

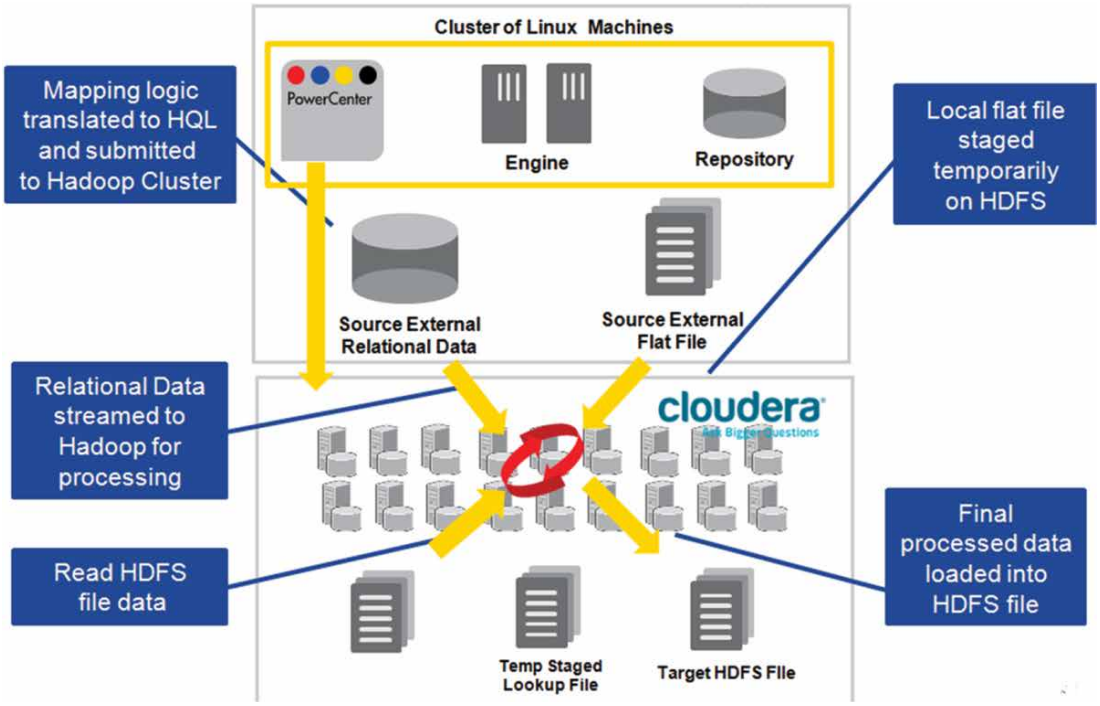

Figure 4. Simplified data pipeline execution in a Hadoop environment.

Scalable Data Processing and Storage

The enterprise-ready CDH (Cloudera's distribution including Apache Hadoop) delivers the core elements of Hadoop—scalable data processing and storage—as well as capabilities for security, high availability, fault tolerance, load balancing, compression, and integration with software and hardware solutions from Informatica and other partners. CDH combines storage and computation in a single system to deliver the necessary flexibility and economics for big data that traditional solutions cannot.

Ideally suited to serve as an enterprise data hub for data warehouse optimization, the 100 percent open source distribution is the world's most complete and widely deployed distribution of Apache Hadoop. CDH incorporates core Hadoop Distributed File System (HDFS), MapReduce, and more than a dozen other leading open source projects. CDH enables enterprises to:

- Unify storage and computation within a single set of system resources
- Store data in any format, free from rigid schemas
- Process data in parallel and in place with linear scalability
- Deliver data in real time to the users and applications that need it
- Integrate the system with existing data management and analysis tools
- Bring flexible computational frameworks to a single pool of data, including batch processing, interactive SQL, interactive search, and machine learning

With an average of 8,000 downloads per month, more enterprises have downloaded and are using CDH than all other distributions combined. Cloudera customers include Samsung, Allstate, Expedia, Monsanto, eBay, CBS Interactive, Experian, Orbitz, Trend Micro, and the U.S. Social Security Administration.

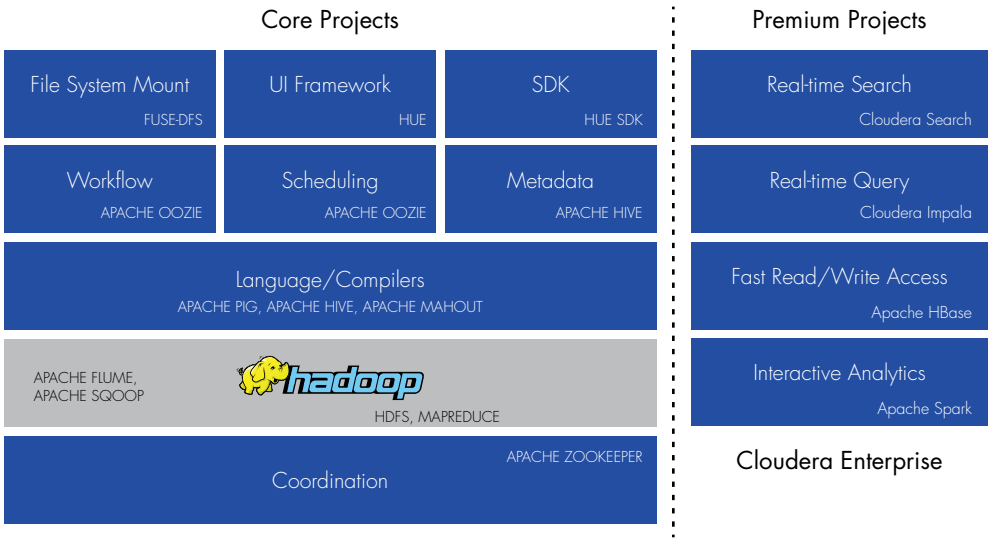

Figure 5. Data processing and storage at scale with CDH.

End-to-End Data Management

The Cloudera and Informatica technologies used for data warehouse optimization through an enterprise data hub operate with minimal administrative overhead (see Figure 6). By delivering centralized management facilities for visibility and control over a Hadoop cluster, they improve performance, enhance quality of service, increase compliance, and minimize costs.

Cloudera Manager is the industry's first and most sophisticated management application for Apache Hadoop, designed to make administration of Hadoop simple and straightforward, at any scale. With Cloudera Manager, administrators can:

- Manage multiple Hadoop clusters, easily deploying, configuring, and operating clusters with centralized, intuitive administration for all services, hosts, and workflows
- Monitor Hadoop clusters with a central view of all activity in the cluster through heatmaps, proactive health checks, and alerts
- Troubleshoot, diagnose, and resolve cluster issues with operational reports and dashboards, events, intuitive log viewing and search, audit trails, and integration with Cloudera Support
- Integrate with existing enterprise monitoring tools through SNMP, SMTP, and a comprehensive API

Informatica Administrator is a complementary browser-based tool that supplies centralized control over data services configuration, deployment, monitoring, and management. Informatica Administrator enables administrators to:

- Monitor the status of data pipeline task executions and workflows on a Hadoop cluster, and check the status of corresponding Hadoop jobs associated with each data pipeline
- Manage the Informatica data pipelines and cancel a data pipeline running on a Hadoop cluster
- View the status of all completed data pipeline runs and trace the corresponding Hadoop MapReduce jobs run for the pipeline execution

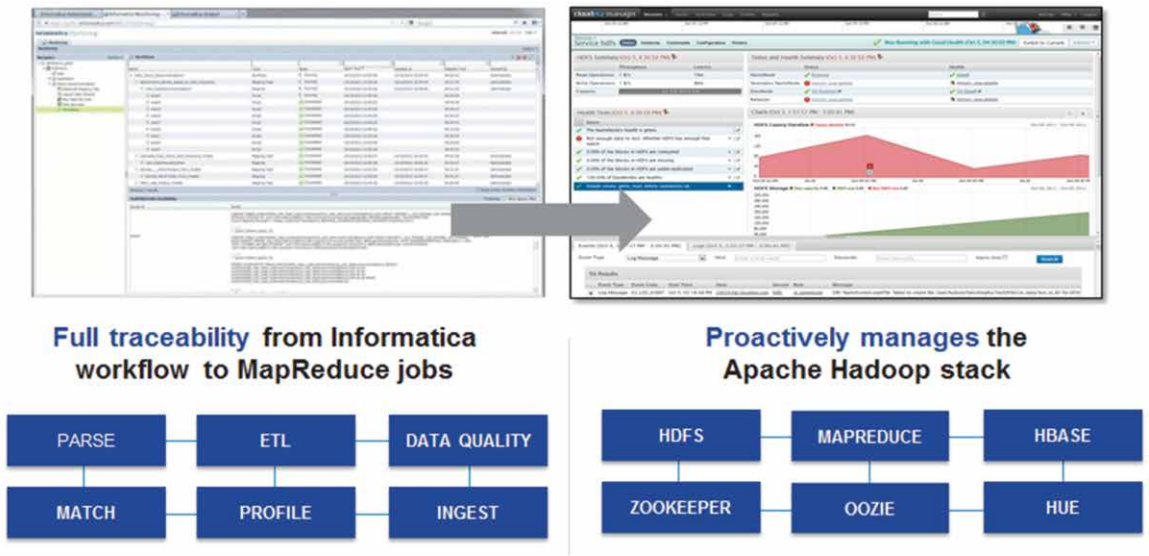

Figure 6. Informatica and Cloudera management facilities provide complete control over an enterprise data hub environment.

Real-Time Interactive Queries

Cloudera Impala gives users real-time, native SQL query capabilities directly against Hadoop data, complementing traditional BI queries against a data warehouse as well as the somewhat limited query functionality of MapReduce and Hive. Cloudera testing has found that Hadoop queries on Impala are between 4 to 100 times faster than the same queries on Hive, meaning that users typically get answers in seconds, not minutes or hours.

The use of Impala against CDH delivers deeper analysis because of the greater volumes of granular data that can be stored in a Hadoop cluster. For instance, using Impala makes it practical to let analysts query all transactions by all customers over any period of time—a data set so large that it would usually be available only in summary form in a data warehouse. The scale-out Hadoop infrastructure makes it possible to execute brute-force queries in parallel over multiple nodes, using standard ODBC or JDBC interfaces.

Viable use cases continue to exist for MapReduce and Hive (e.g., for batch data transformation workloads) as well as traditional data warehouse frameworks (e.g., for complex analytics on limited, structured data sets). Impala complements those approaches, supporting use cases where users need to interact with very large data sets to get focused results quickly.

Implemented in C++ and bypassing the overhead of MapReduce, Impala includes three key query components:

- **Query planner:** Clients submit queries to the planner, which turns the request into a collection of plan fragments for execution
- **Query coordinators:** Clients Iinitiate execution on remote Impala daemons
- **Query execution engines:** Clients execute plan fragments and return results

Queries are submitted to a single Impala process, at which point the query planner in the Impala process turns the query into a set of plan fragments and the coordinator initiates execution of these plan fragments on Impala processes local to the data. Intermediate data is streamed between Impala processes, and the final results are returned to the client.

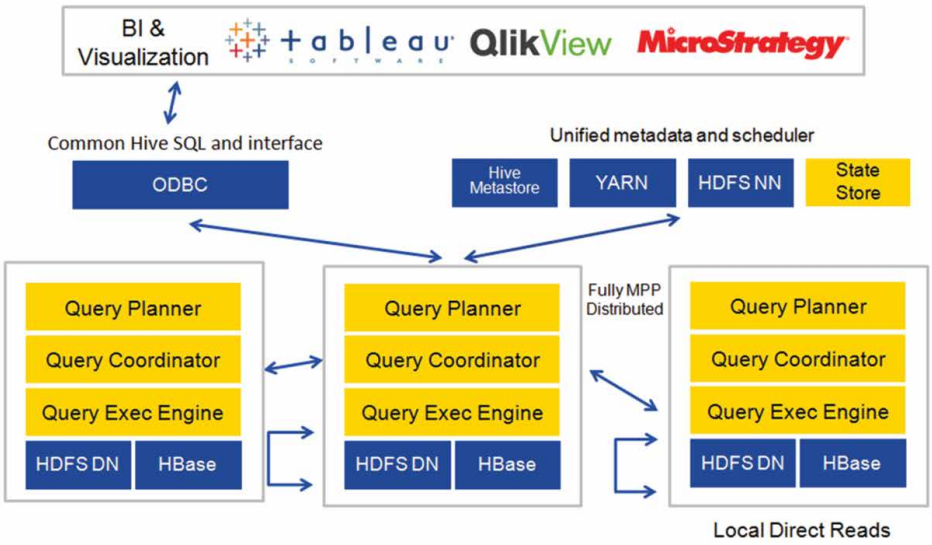

Figure 7. Cloudera Impala executes real-time interactive queries against Hadoop.

Enterprise Accountability, Control, and Governance

Hadoop holds the promise to deliver unprecedented scalability at an affordable cost. But organizations need confidence that Hadoop can also support requirements for enterprise accountability, control, and governance for compliance with corporate and governmental regulations and business SLAs. The reference architecture includes four key components for governance—Cloudera Navigator, Informatica Metadata Manager and Business Glossary, Informatica Dynamic Data Masking, and Informatica Data Archive. These tools deliver the ability to manage metadata to find, search, discover, explore, and trace the lineage of data across systems as it is transformed from source to target. They also enable administrators to securely manage the information lifecycle through archiving, data retention, and the masking of sensitive data.

Cloudera Navigator. The first fully integrated data management tool for Hadoop, Cloudera Navigator provides data governance capabilities such as verifying access privileges and auditing access to all data stored in Hadoop (see Figure 8). The software tracks access permissions and actual access to all data objects in HDFS, HBase, Hive, and Hive metadata.

With it, administrators can answer questions such as who has access to which data objects, which data objects were accessed by a user, when was a data object accessed and by whom, what data was accessed using a service, and which device was used to access it. Cloudera Navigator allows administrators to configure, collect, and view audit events and enables audit data to be exported to third-party tools.

Cloudera Navigator is fully integrated with Cloudera Manager, with all Navigator functions accessible through the Cloudera Manager interface. Forthcoming versions of Navigator will enhance this functionality by offering:

- Information discovery to extract or search metadata on all data stored in Hadoop
- Data lineage to track the lineage of all data assets stored in Hadoop
- Impact analysis to determine the impact of changes to data assets
- Data lifecycle management to automate the data lifecycle, from ingress to egress, and define data retention policies for data stored in Hadoop

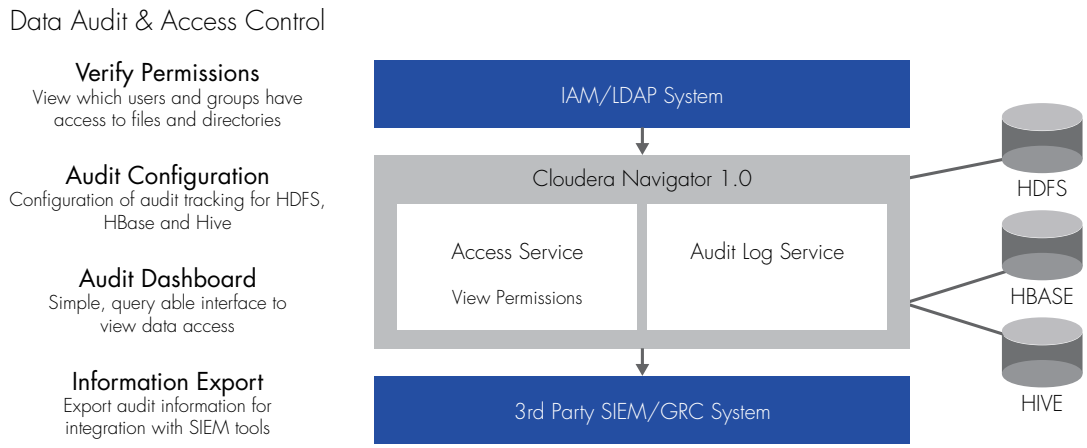

Figure 8. Key functionality in Cloudera Navigator.

Metadata Manager and Business Glossary. Informatica Metadata Manager provides the visibility and control needed to manage change, reduce errors caused by change, and ensure data integrity. Key features include:

- End-to-end data lineage across system boundaries with a dynamic visual map of all data flows in an integration environment
- Impact analysis for viewing all upstream and downstream impacts of proposed changes to the environment before they are implemented
- Data integration metadata catalog for capturing and storing enterprise metadata in an enterprise repository
- Metadata connectors to automatically gather metadata from databases, enterprise applications, mainframes, files, Informatica PowerCenter, data modeling tools, and BI tools
- Audit trail to visually track data from source to target, including all transformations and reference data
- Built-in collaboration tools to enhance productivity among developers and among developers and business users

Geared for business users and to promote IT-business collaboration with a common business vocabulary, Informatica Business Glossary shares a central repository with Informatica Metadata Manager and supplies:

- A business-friendly user interface for business users to create, annotate, review, and approve a common business vocabulary
- Common vocabulary of business terms to ensure clear communication, boost productivity, and provide clear definitions for audit purposes
- Data lineage to furnish the underlying detail about how business terms were derived and where they are used
- Business term ownership to ensure accountability for the accuracy, documentation, and freshness of the business vocabulary
- Built-in collaboration tools to enhance productivity among developers and among developers and business users
- Built-in review and approval of business terms to speed up term creation and review

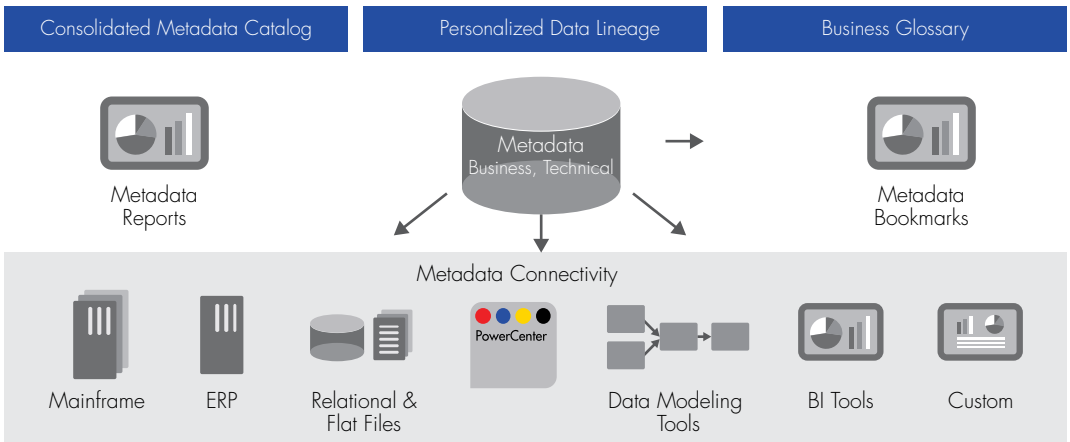

Figure 9. The Metadata Manager/Business Glossary environment.

Dynamic Data Masking. To help ensure the privacy and security of data in a Hadoop environment, Informatica Dynamic Data Masking provides real-time capabilities to prevent unauthorized users from accessing sensitive information. It allows an IT organization to apply data masking rules based on a user's authentication level. Informatica Dynamic Data Masking protects sensitive information and blocks, audits, and alerts users who access sensitive information.

In a Hadoop environment, Dynamic Data Masking can be set up as a security layer for HiveQL requests. For instance, if a user makes a HiveQL request for data he or she is not authorized to see, Dynamic Data Masking applies a HiveQL rewrite in real time to mask the results set. All types of masking or scrambling functions are supported, with no appreciable performance overhead. With it, administrators can:

- Dramatically decrease the risk of a data breach
- Easily customize data masking solutions for different regulatory or business requirements
- Protect personal and sensitive information while supporting offshoring, outsourcing, and cloud-based initiatives
- Secure big data by dynamically masking sensitive data in Hadoop

Data Archive. Informatica Data Archive software can leverage the cost-effective Hadoop infrastructure to store archived data in Hadoop using the same extreme compression, data immutability, and accessibility via SQL, ODBC, or JDBC it uses to store data on regular file systems.

For example, IT users may have one or more structured applications and databases they would like to archive or retire. To do this, an archive operator or developer uses Data Archive to discover, define, and classify the business entities and business rules for archiving data from the source system. Once the business entities and business rules have been defined, the archive operator can define an archive job to be executed. Defining a job is accelerated with prebuilt accelerators for many packaged applications from Oracle, SAP, Amdocs, and others.

Informatica archives data to an optimized format with up to 98 percent compression while maintaining direct, easy access to the archived data via standard reporting applications, Informatica data discovery portal and data visualization tools, and SQL 92 for queries. Retention policies and purge schedules can be applied on the archived data and legal holds enforced to support e-discovery processes and ensure that data is locked down when it's relevant to a legal case, even though the retention has expired. Enterprises use Informatica Data Archive to:

- Improve application performance by smart partitioning active data
- Lower IT costs by archiving inactive data and legacy applications
- Maintain easy and appropriate levels of access to archive data
- Increase IT operational efficiencies with streamlined maintenance processes
- Ensure regulatory compliance with complete enforcement of data retention and disposition policies

The Safe On-Ramp to Data Warehouse Optimization

Traditional data warehouse environments are unsustainable in the era of big data. Data warehouse optimization is the safe on-ramp for organizations to meet the challenge of evolving data warehousing.

The Informatica/Cloudera reference architecture provides a sustainable model that can scale to accommodate the epic growth in data volumes, variety, and velocity. It equips enterprises with the foundational capabilities necessary to drive business value from big data:

Cost-effective storage and processing. Hadoop can be up to 10 times less expensive than traditional data platforms and obviates the need to add warehouse infrastructure.

Greater warehouse capacity. Staging and offloading data to Hadoop to preprocess and refine data sets for analysis helps preserve the data warehouse for high-value curated data.

Warehouse performance and scalability. Hadoop enables you to quickly scale out data storage and processing capacity to handle petabyte-scale data sets.

The ability to leverage existing resources and skills. More than 100,000 Informatica-skilled developers can supplement your in-house Informatica resources.

Increased developer productivity. Informatica eliminates the need for hand coding and increases developer productivity on Hadoop by up to five times.

Datatype flexibility. Hadoop can access, ingest and process all datatypes and formats, including legacy mainframe, relational, social, machine, and other sources, while ensuring data quality.

End-to-end data pipeline. The entire data pipeline can be executed natively on Hadoop, including loading, parsing, feature extraction, profiling, integration, cleansing, and matching to prepare data for analysis.

Low-latency interactive queries. Cloudera Impala enables real-time interactive queries and searches to explore and visualize data on demand.

End-to-end Hadoop cluster management. The joint solution ensures that business SLAs are met and that workloads are easy to administer, schedule, and support.

As leaders in their respective fields, Informatica and Cloudera deliver a proven and cost-effective joint solution to augment warehousing systems with a Hadoop-based enterprise data hub.

About Informatica

Informatica Corporation (Nasdaq:INFA) is the world's number one independent provider of data integration software. Organizations around the world rely on Informatica to realize their information potential and drive top business imperatives. Informatica Vibe, the industry's first and only embeddable virtual data machine (VDM), powers the unique "Map Once. Deploy Anywhere." capabilities of the Informatica Platform. Worldwide, over 5,000 enterprises depend on Informatica to fully leverage their information assets from devices to mobile to social to big data residing on-premise, in the Cloud and across social networks.

About Cloudera

Cloudera is revolutionizing enterprise data management with the first unified Platform for Big Data: The Enterprise Data Hub. Cloudera offers enterprises one place to store, process and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Founded in 2008, Cloudera was the first and still is the leading provider and supporter of Hadoop for the enterprise. Cloudera also offers software for business critical data challenges including storage, access, management, analysis, security and search. Cloudera works with over 800 hardware, software and services partners to meet customers' big data goals.

Put potential to work.™

Worldwide Headquarters, 100 Cardinal Way, Redwood City, CA 94063, USA Phone: 650.385.5000 Fax: 650.385.5500
Toll-free in the US: 1.800.653.3871 informatica.com linkedin.com/company/informatica twitter.com/InformaticaCorp

IN09_0114_02609