Unifying the Enterprise Data Hub and the Integrated Data Warehouse

TERADATA

cloudera

CONTENTS

Encompassing All of the Big Data Universe

Businesses have access to more data now than at any time in recorded history. Companies today must adjust to meet the new strains and demands big data places on them. And they must begin to do so now, as big data is going to grow and become more integral to success in the next decade. According to IDC[1], the amount of data generated will be in excess of 44 zettabytes by 2020. With that type of growth, businesses that are not evaluating their current data architecture risk being overwhelmed by the volume and speed of data. Even worse is being left behind by competitors who are better equipped to use data as a way of driving decision-making.

According to IDC, the amount of data generated will be in excess of 44 zettabytes by 2020

The term big data is misleading—there's more to this phenomenon than size alone. In addition to unprecedented data volumes, organizations now face a range of new data types and a pace of data creation they often aren't prepared to handle. Big data is highly complex and differentiated; it originates from a variety of sources. Yet at the same time, big data also encompasses traditional structured financial, operational, and customer data.

A New Breed of Big Data Solutions

In the face of this, a new breed of big data solutions has hit the market. Considering that IDC[2] also forecasts big data hardware, software, and services spend to grow from $12.6B in 2013 to $32.4B in 2017, it goes without saying that It can be difficult for businesses to navigate the many data management options available today. As a result, the technology buyer winds up evaluating one technology versus another rather than looking at how a multi-faceted approach may be the right strategy for the business.

It turns out the ideal solution is one built on multiple components, allowing the enterprise to turn big data hype into productive solutions fit to their specific needs. To handle big data and create systems that help the business extract full value from it, companies have to implement comprehensive solutions that enable flexibility, speed, and movement between data repositories, yet still ensure that governance, security, and reliability are not thrown by the wayside.

This logical data warehouse is a dramatic expansion of the prior blueprint for a data warehouse. Together, Teradata® and Cloudera are championing the rapid expansion of data warehouse concepts into a fresh, new way of delivering analytic value to the business. In simple terms, this means combining the data warehouse and the enterprise data hub into a single ecosystem. But it's more than that.

[1] IDC, The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things Digital Universe Survey, April 2014

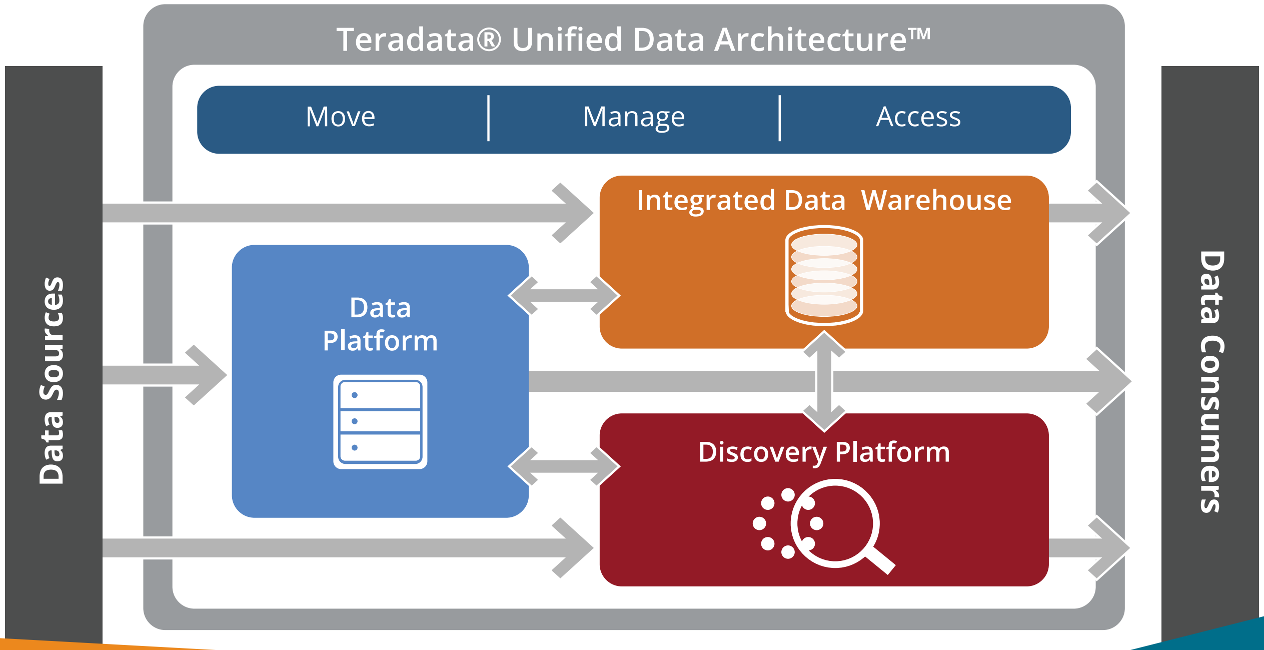
[2] IDC, Worldwide Big Data Technology and Services 2013–2017 Forecast, Feb 2014

The Ideal Structure

An ideal data management structure integrates multiple workloads across multiple repositories holistically. This type of overarching architecture goes by many names: Gartner refers to it as the logical data warehouse; the 451 Group labels it the total data warehouse. Regardless of what a business calls it, the idea behind a collection of platforms is that an architectural ecosystem is required to exploit all your data for competitive advantage. For simplicity, we use Teradata Unified Data Architecture™ to illustrate the hybrid structure because of its similarities to Gartner's logical data warehouse.

The Teradata Unified Data Architecture is a useful blueprint because it organizes data-centric workloads into major service level agreements (SLAs). The SLAs are the requirements, expectations, and workloads of the corporation. The Unified Data Architecture recognizes and elevates the data platform and discovery platform as peer workloads in the integrated data warehouse. The single version of the truth is simply getting bigger than ever. It's no longer contained in a single repository, but spread across a large ecosystem of services. Here are the workloads and service level agreements included:

- Data platform. The first major SLA supported is a landing zone for raw data of unknown value for long periods of time at an economical cost. The second SLA served is data staging and refinement. The third major SLA is self-service exploration analytics using search and analytics tools.
- Discovery platform. The primary SLA is to enable quick use of sophisticated algorithms across data and analytics engines for discovery. This means tools for data wrangling, collections of analytic algorithms, ease of use, and flexibility to explore and fail forward.
- Data warehouse. The primary SLA is defined by a logical data model that mirrors business processes and subject areas in the data itself. Data elements are rationalized across the enterprise, put into rigorously defined structures, and transformed for intersecting use cases. Other major SLAs served by the data warehouse include that the data must be persisted, time variant, and optimized for response time.



The Unified Data Architecture reveals the need for different subsystems to fulfill specialized roles: IT is too complicated for a single repository to perform every conceivable role. The architecture prescribes multiple repositories, data virtualization, and distributed processes in combination to solve a wide set of business objectives. However, when inserting specific products and tools into the architecture, we find overlap in functionality across platforms. This can cause angst, as IT architects, programmers, and CIOs must learn how to select a platform product for a given workload. Although each platform product has strong differentiation, it takes time to learn their best-fit strengths. But this is also a benefit. It's great to have multiple options for some workloads, especially if you are not ready to deploy all types of platforms.

So, what does this all add up to? First, there is more usable data for the enterprise, leading to additional insights and discoveries. It also means more accurate decision making thanks to the blend of tools and analysis. Ultimately, these hybrid capabilities lead directly to competitive advantages. Second, there are new multi-structured data types supporting previously underserved business user communities. For example, analyzing weblogs, sensor data, text, and email leads to additional ways to calibrate lead generation, churn analysis, supply chain optimization, risk, and fraud detection. Third, it means finding the best fit for a workload without putting the business at risk. Analytic workloads need different repository technologies, similar to the way corporations have multiple business intelligence tools for different tasks.

The Enterprise Data Hub: Refining Raw Data

It's easy to visualize the enterprise data hub—based on Apache™ Hadoop®—as a refinery of raw iron ore. Each scoop of rock (raw data) from the mine contains iron to be extracted and turned into something useful. By smelting, enriching, and transforming the iron, we can produce sheet metal (for example, reports). With additional transformations, the data hub also produces finished goods. In other cases, it feeds sheet metal (transformed data) to downstream manufacturing plants. In this way, the enterprise data hub is the beginning of an information supply chain.

Companies can capture the data and later determine its relative value to the business

At its core, Hadoop is designed to be an ideal first-touch processing engine and repository for big data. Apache Hadoop and many of its surrounding projects represent a viable solution for rapid processing of large amounts of unstructured data. An enterprise data hub builds on Hadoop to include the reliability, governance, and data security capabilities the enterprise requires. It also serves as a landing zone for data whose value is not yet understood.

An enterprise data hub is a place to store data for as long as it's needed, in its original structure, at full fidelity. It allows for discovery and initial investigation of raw data. Companies can capture the data and later determine its relative value to the business. The ability to keep this raw data is important, as it provides companies with the flexibility to repurpose it or look for additional signals later. The goal is to use the speed of Hadoop to transform data as quickly as possible so that it can be consumed or made more widely available for other downstream uses.

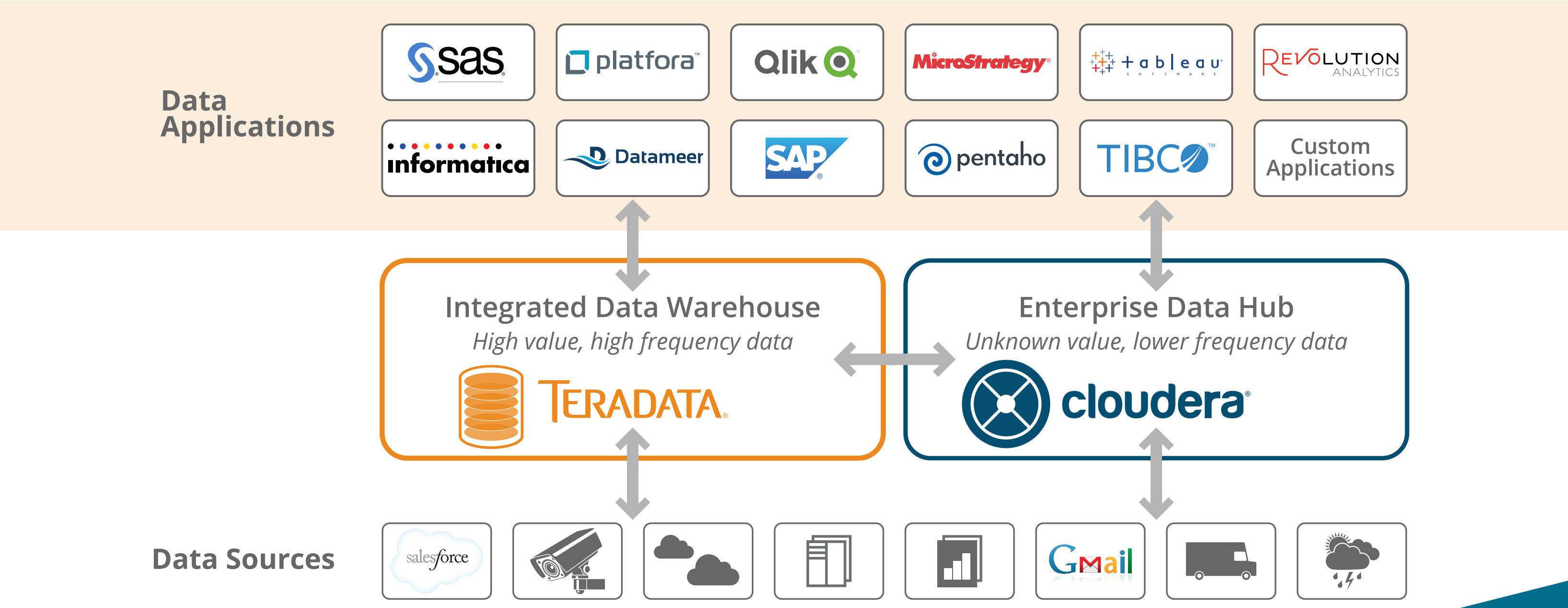
The Integrated Data Warehouse: Integrating Data

The integrated data warehouse consumes data from the data hub as well as directly from operational applications. This step in the information supply chain includes extensive cleansing, validation, normalizing, deduplicating, and rationalizing inconsistent semantics. This is what's called schema-on-write. This form of data integration makes the data easier to use and more understandable by diverse populations of users. Robust data governance, vetting, metadata lineage, and standardization eliminate data reconciliation hurdles for the user community.

Extensive data integration from multiple operational systems quickly attracts hundreds to thousands of concurrent users to the data warehouse. This in turn drives demand for high levels of system availability, administration, and security. With hundreds of users and dozens of concurrent batch jobs, the data warehouse has a clear mandate for strong workload management to prioritize and enforce service-level policies. Large user populations also demand exceptional query response times, necessitating advanced indexing, OLAP, and cost-based optimization of queries. With all of these features in place, the most advanced implementations evolve into active data warehouses that add near real-time data loading and sub-second tactical queries. This means external brokers, suppliers, and consumers can directly access the integrated data.

Cloudera and Teradata: Better Together

The hybrid architecture blends the enterprise data hub and the integrated data warehouse. From the business user's point of view, it is one large ecosystem. This simplifies the business user's tasks and focuses IT developers on a more versatile value delivery model. The combined strengths of Teradata and Cloudera help CIOs and CTOs consolidate data silos into a shared infrastructure where best practices can be applied. Data marts can be spawned and later consolidated in one repository or another. Further, data virtualization is making great strides in simplifying access to hybrid independent repositories.



Similar to a mutual fund portfolio, this hybrid blend of technologies mitigates risks while paying consistent dividends. The Cloudera enterprise data hub performs the schema-on-read role within the hybrid architecture. An enterprise data hub satisfies the needs of new varieties of data, archival of data, exploration, and search. The Teradata Integrated Data Warehouse brings integrated, optimized data to wide populations of users of various proficiencies. The integrated data warehouse satisfies complex OLAP queries and in-database analytics, even while hot and cold data migrate throughout the hardware for better performance.

This blended portfolio is not limited to technologies. The strength of this collaboration also lies in the best practices of Cloudera and Teradata's people—the laboratories, professional services, sales, and marketing teams, as well as business partners. Unifying the vision and skills across these communities makes it easier for our joint customers to learn and capitalize on their data assets. Cloudera and Teradata working together on-site with a shared architecture vision advances the CIO's agenda while reducing architecture confusion.

The Cloudera Advantage

Cloudera is the industry leader in products that support and improve Hadoop. While Hadoop offers a large data repository, Cloudera supplies governance, security, and management as well as the ability to communicate with the data warehouse, enabling Hadoop to function as part of an enterprise-ready data solution. One major value proposition behind building an enterprise data hub is the economic model. It enables organizations to store more data than could ever be stored before. It provides access to a lot more raw material for analytics. This expands to discovering insights, where even the question might not have been obvious. By having access to all of the raw data and all of the new analytics, you can start to mine for new nuggets of information that can help direct business strategy.

Cloudera, in conjunction with the largest group of Hadoop committers and contributors, is focused on developing innovative solutions that make Hadoop more enterprise ready. The enterprise data hub builds on these open source tools, bringing in Cloudera intellectual property in data security and management, again making Hadoop more enterprise ready. With Cloudera Manager, companies can control every aspect of the data hub. Cloudera Manager is a full-service administration tool for the enterprise data hub, greatly reducing deployment time and enabling businesses to deploy, configure, manage, and monitor their cluster from a single location.

Cloudera Navigator is the leading end-to-end governance solution for Hadoop-based systems. Through a single user interface, it provides visibility for administrators, data managers, data scientists, and analysts to secure, govern, and explore the large amounts of diverse data that land in Hadoop. By enforcing comprehensive security and unified auditing with data lineage across Hadoop, companies can feel confident that their enterprise data hub is ready for compliant data exploration.

Cloudera Impala is an interactive SQL query engine designed for parallel processing. Impala enables analysts and data scientists to directly interact with any data stored in Hadoop, using their existing BI tools and skills through a high-performance SQL engine.

Cloudera has more customers running production-level Hadoop clusters than any other vendor. This includes more than half the Fortune 50 as well as the top federal defense agencies.

The Teradata Advantage

Numerous Teradata innovations contribute to the overall hybrid architecture success. Just a few of the important innovations include:

- **Teradata QueryGrid™** lets business analysts ask any question, transparently pulling together data from both platforms into a single answer set. QueryGrid is the switching strategy between repositories and workflows.

- **Teradata Aster Discovery Platform** provides an abundance of tools for the data scientist. It pulls weblogs from Hadoop, sessionizes web clicks, and joins the results to consumer profiles from the data warehouse. Using Teradata Aster nPath on Hadoop data, clients can correlate a consumer's journey through websites, call centers, email, and other touch points to find churn or fraudulent events.

- **Teradata Loom** provides tracking, exploring, cleaning, understanding, and transforming of HDFS files. Loom monitors the lineage of every HDFS file from the moment it is loaded through its usage lifecycle. Operations managers and programmers can easily evaluate 50 million HDFS files.

- The **Teradata Database** has a cost-based optimizer that handles 20–50 table joins of hundreds of terabytes. It enables you to move beyond simple data marts to a comprehensive cross-organization view of the business. Furthermore, the Teradata Database's workload management is the gold standard for helping administrators manage performance and throughput service level objectives.

Teradata Professional Services' data warehousing skills and best practices are immediately transferable to the Cloudera enterprise data hub. A shortened list of Teradata Professional Services offerings include data lineage, data cleansing, governance, ETL, security, ingestion, reporting and dashboards, predictive analytics, and when-to-use-which platform selection. These capabilities are strengthened by Teradata's acquisition of Think Big Analytics. Think Big brings Hadoop ecosystem skills including ingestion, MapReduce, streaming, cascading, NoSQL integration, search, and machine learning.

Teradata customers are achieving exceptional results. One Teradata system powers a single 60TB database supporting 185 applications and 14 million queries a day, most of those in real time. Teradata systems also run a 36PB production data warehouse with over 1,000 users. One vehicle manufacturer uses QueryGrid to combine Hadoop sensor data with data warehouse equipment schedules, parts inventories, and staffing. Using predictive analytics, this helps them eliminate false positives from vehicle repairs, saving considerable costs and labor. This manufacturer also pointed out that "QueryGrid provides a bridge for collaboration between the MapReduce and SQL teams." It should come as no surprise that Teradata has been a leader in the Gartner Data Warehouse DBMS Magic Quadrant for 14 of the last 15 years.

Conclusion

Whether it's called a Unified Data Architecture or a logical data warehouse, this new architecture blends established and emerging technologies. Both Teradata and Cloudera technologies are experiencing rapid evolution and innovation. This helps corporations gain access to the most relevant capabilities so they can exploit all of their data, regardless of its type, structure, size, or source into a logical view that is easily queried. This is essential because data is only as good as it is relevant. If users cannot find the answer they're looking for fast, companies stagnate and fall behind their competition.

Incorporating an enterprise data hub and integrated data warehouse into a larger architecture assures business users that they are seeing the sharpest picture their data can provide. Additionally, this architecture offers the economical solution the enterprise requires. Ultimately, business users will have better data, find it easier to query what they have, and be able to do so at a lower cost than in the past. This is the full data orchestration that can power enterprises in the big data age.

Co-Authors

Daniel T. Graham | *Technical Marketing Director*

With over 30 years in IT, Dan joined Teradata Corporation in 1989 where he was the senior product manager for the DBC/1012 parallel database computer. He then joined IBM where he wrote product plans and launched the RS/6000 SP parallel server. He then became Strategy Executive for IBM's Global Business Intelligence Solutions. As Enterprise Systems General Manager at Teradata, Dan was responsible for strategy, go-to-market success, and competitive differentiation for the Active Enterprise Data Warehouse platform. He currently leads Teradata's technical marketing activities.

Clarke Patterson | *Senior Director, Product Marketing*

Clarke Patterson is the Senior Director of Product Marketing at Cloudera. In this role he is responsible for product and solutions marketing activities supporting Cloudera's Platform for Big Data. Clarke joined Cloudera after spending almost three years in a similar role at Informatica. Prior to Informatica he held product management positions at IBM, Informix and Red Brick Systems. Clarke brings over 17 years of leadership experience to Cloudera having lead teams in product marketing, product management and engineering. He holds a Bachelor of Science degree from the University of Calgary and an MBA from Duke University's Fuqua School of Business.

QueryGrid is a trademark and Teradata and the Teradata logo are registered trademarks of Teradata Corporation and/or its affiliates in the U.S. and worldwide.

Cloudera is a registered trademark of Cloudera Corporation in the United States and in jurisdictions throughout the world. All other company and product names may be trade names or trademarks of their respective owners.

Copyright © 2014 by Teradata Corporation and Cloudera. All rights reserved. Produced in USA.