Best Fit Engineering in Analytic Architecture

Stephen Brobst, CTO, Teradata Corporation
Scott Gnau, CTO, Hortonworks
Sanjay Krishnamurthi, CTO, Informatica

Table of Contents

Introduction

Data is the lifeblood of most industry-leading companies. In addition to transaction data from core banking systems, billing systems, and ERP systems, leading organizations have begun to focus on the huge volumes of interaction data—originating from sources as diverse as social media, sensors, and Web sites—for analytic purposes. The ability to convert this data into valuable insights that inform business tactics and strategies, with the right data delivered at the right time, separates industry leaders from laggards.

Yet managing this data requires a new approach to technology infrastructure—one that can efficiently process data, support innovative analytic tools, and optimize data integration processes. The flexibility and proven delivery capability of a unified data architecture (UDA) can help companies combine multiple data technologies into a cohesive, agile ecosystem for extracting the value from data assets. A UDA can help organizations reliably and consistently access, refine, and deliver data as actionable information to business users, rapidly converting huge data volumes into real business value.

This paper will help an enterprise maximize the business value of its data assets by reviewing best practices for deploying a UDA. The importance of a best-of-breed approach using technologies from Hortonworks, Informatica, and Teradata to optimize the value extracted from data is proposed as a critical success factor for maximizing return on investment and to become data centric in today's data-driven economy.

informatica Hortonworks TERADATA

The Right Processing on the Right Technology

A key tenet of good engineering is to select the right tool for the problem at hand. No one single technology solves all problems in the best way. A well-engineered architecture will have multiple technologies integrated into a single ecosystem to support different aspects of an analytic solution. We refer to the framework for organizing components to support the analytic workload into a single ecosystem as the UDA. There are four distinct components of the UDA: (1) the data lake, (2) a discovery platform, and (3) an integrated data warehouse (IDW), and (4) a data integration platform. The key is to understand which parts of an analytic workload should be allocated to each technology within the UDA and when these subsystems should interoperate as well.

The **data lake** infrastructure houses the *raw material* to be transformed to insights and business value. Capturing *all data forever* is an implicit goal of the data lake infrastructure because any data that is thrown away will never be harnessed for enterprise value. Of course, *all and forever* is an exaggeration wherein practical implementation will be relative to infrastructure budgets and value density of the data. Given this general approach toward capturing data, it is clear that three key characteristics of a successful data lake environment will be: (1) scalability, (2) low cost per terabyte of data stored, and (3) support for both record-oriented and non-traditional "unstructured" and "semi-structured" data types.

The target use case for a **discovery platform** is the data R&D tasks performed by data scientists and business analysts in pursuit of new insights from data. A discovery platform is focused on agility in uncovering new patterns and relationships from the raw materials stored in the data lake. Integration between the discovery platform and data lake is critical. The requirement is to provide extreme agility in bringing content from the data lake into the discovery platform. The schema-on-read approach is typically used for this step. High-performance access to Hadoop using Teradata QueryGrid™ allows easy provisioning of content into the data R&D environment using skill sets consistent with those of a data scientist and business analysts.

The **integrated data warehouse** (IDW) is where the results from data R&D undertaken in the discovery platform are productized for deployment to the enterprise. Great data science without productization of the data insights will

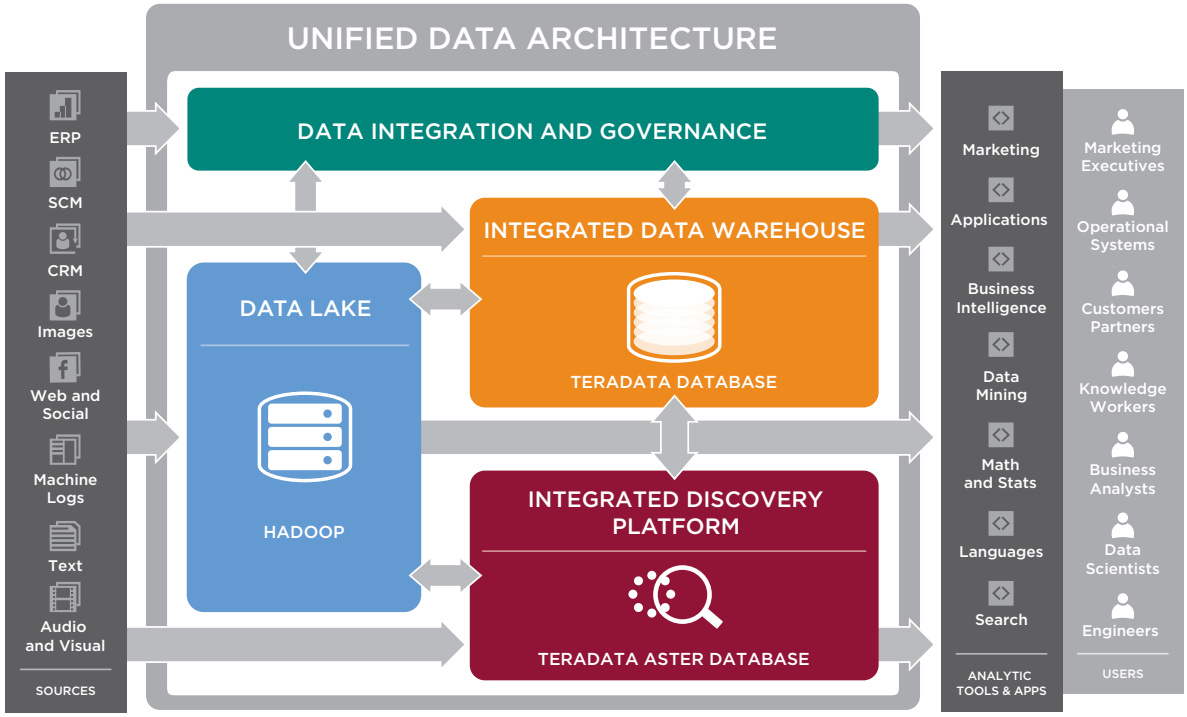

Figure 1. Unified Data Architecture.

not yield business value. The requirements for productization of data insights include much more emphasis on data quality certification, auditability, and integration than is the case for the data R&D environment. An integrated data warehouse will typically use relational database technology and must deliver on much more stringent service level requirements than the discovery R&D environment. An integrated data model where data can be stored once and re-used many, many times is a critical success factor for effective IDW deployment.

Data integration (DI) is the process of ingesting, transforming, and distributing refined data. Data integration identifies and rationalizes inconsistent semantics in the data so business users can extract the greatest meaning from data. Using metadata lineage tracking, users can determine where specific datum came from for compliance and internal reporting purposes. Data governance is ensured by data stewards that profile and monitor data quality for analytical or master data management initiatives. As new sources of increasingly diverse types of data are used by enterprises, the need for data integration and governance increases. Furthermore, data integration is a process that is pervasive throughout the UDA. In this paper we use data integration to illustrate the flexibility and breadth of the UDA.

Next, we review the roles of major components in the UDA.

Data Lake Environment

With continued growth in the scope and scale of new applications built using Apache™ Hadoop® within the enterprise, the vision of an enterprise data lake enabled by a modern data architecture can become a reality.

The need for the data lake arose because new types of data needed to be captured and exploited by the enterprise. As this data is increasingly available, the data lake emerged as companies needed the capability to:

- Capture and store large amounts of raw data at scale for a low cost.
- Store many dissimilar types of data in the same repository.
- Perform large-scale transformations on the data.
- Define the structure of the data at the time it is used (schema-on-read).
- Perform new types of data processing and single subject analytics.

Many kinds of big data are found in the data lake, such as clickstream data, server logs, social media, machine and sensor data, geolocation coordinates, video, audio, and text. Some of this data (sensors, clickstreams) has been managed in isolation for many years. But most of it

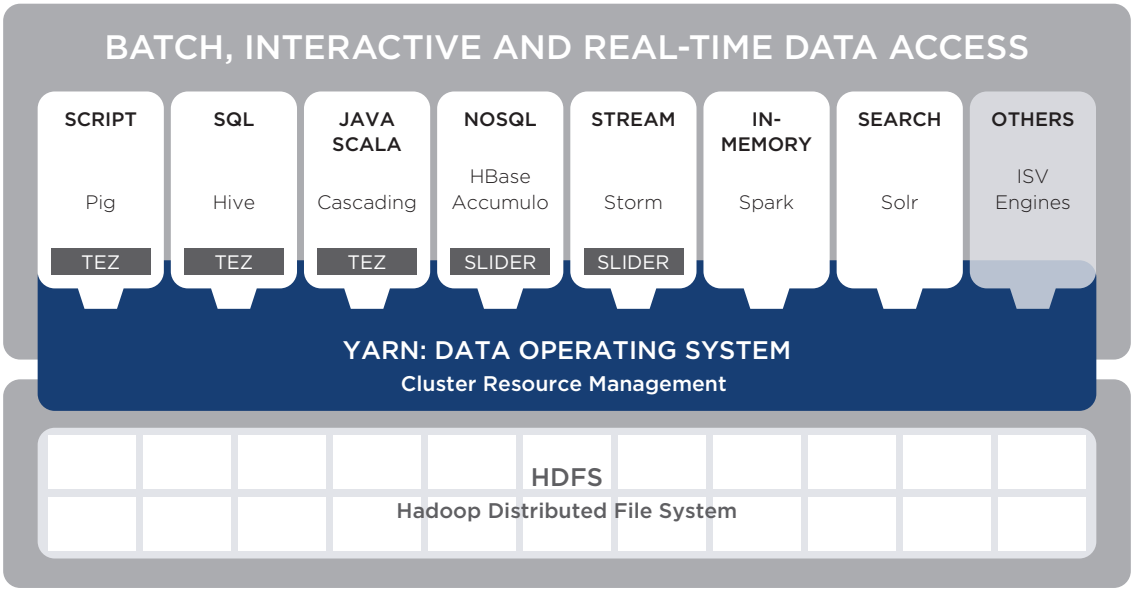

Figure 2. Data Lake Subsystems.

was simply discarded for the lack of a way to convert it into business value affordably. Today, whenever the data volume grows unwieldy, IT professionals start building data lakes.

Hadoop 2.0 enables true data lake architectures. The introduction of YARN in particular adds a pluggable framework that enables new data access patterns in addition to MapReduce. Familiar languages like SQL can now be used in addition to MapReduce, and new programming constructs, such as cascading, offer an efficient alternative to MapReduce for interested developers.

Access to all the data in the data lake is provided by both SQL and NoSQL technologies. The leading SQL-in-Hadoop project is Apache Hive®. Hive has the widest adoption in the Hadoop community because of its strong SQL compatibility and its recent performance enhancements when coupled with Apache Tez. As the Hive community marches towards sub-second data access, Hortonworks is also contributing Hive innovations that can be leveraged by Spark. SparkSQL leverages the Hive metastore to read a table's schema and then reads data directly from HDFS. With SparkSQL, there is a JDBC/ODBC server included and it can be used to submit SQL queries remotely. This allows SparkSQL to use modern versions of Hive to access data for tasks such as machine learning, modeling, and joins while conserving resources.

One of the unique capabilities of the data lake is the data files themselves are agnostic to the access method used. Different tools and languages such as Hive, Apache Pig, and Spark can all access the same HDFS files directly. This allows developers to choose the best fit data access method. More important, it allows technologies to coexist, sharing the same data. This adds both flexibility and investment protection.

The data lake is therefore both a source of raw data and a data service to downstream applications. Data scientists and data miners will frequently self-provision raw and refined data to the discovery platform for deep exploration. After data integration processing, the transformed raw data feeds into the IDW. This is a primary benefit of the UDA: the right tool for the problem at hand exploits the strengths of each platform.

Data centricity requires enabling analytics at different degrees of data integration. There is no one single approach to modeling and integrating data. Prior to the big data trend, there was a single approach to data integration whereby data is normalized, persisted in a database, and only then can value be created. The idea is that by absorbing the costs of data integration up front, the costs of extracting insights decreases. This approach can be called tightly coupled integration. This is still an extremely valuable methodology, but is no longer sufficient as a sole approach to manage all enterprise data.

Now with big data, using only one approach to integration undermines the value of newer data sets that have unknown or under-appreciated value. New methodologies are essential to cost effectively manage and integrate the data. We call these newer approaches loosely coupled and non-coupled integration.

Loosely-coupled data is a methodology whereby effort to apply structure and rules is deferred as late as possible—often at runtime. This avoids unnecessary data preparation. Only the bare minimum of data rationalization in the form of a key occurs. Data is treated as raw materials stored close to original form.

Data in its purest raw form is non-coupled. There are no additional keys defined during acquisition or prior to consuming the data to aid in integration. Integration of non-coupled data with loosely and tightly coupled data is done through expertly written end user code or through data virtualization that creates the needed linkages via keys on the fly.

It is the use of all three of these techniques—tightly, loosely, and non-coupled integration—that enables companies to leverage all their data in an efficient, agile, and cost effective manner.

informatica Hortonworks® TERADATA.

Discovery Platform

The primary goal of the discovery platform is to support innovation with the data assets of the enterprise. Innovation requires agility and the flexibility to use a wide range of analytical tools and techniques. While SQL and its variants are powerful and easy-to-use for manipulating traditional data types that can be rendered in table-like formats, it is clear that additional capabilities are required for advanced data science. Use of imperative programming constructs, as well as extensions into graph processing, text processing, advanced statistical methods, and so on, are required. Moreover, interactive response times when experimenting with data is crucial for the creative process when iteratively discovering new patterns and relationships in data.

Traditional systems development life cycle (SDLC) models for deployment do not work in a data R&D setting. It must be low overhead to provision new data into a discovery platform to facilitate agility in experimenting with new data. Integrated data models, data quality certification, and auditability are not typically requirements for conducting data R&D. Instead, the focus is on self-provisioning of experimental data sets into the discovery platform with minimum (bureaucratic or technical) overhead in doing so (think hours, not days). A data scientist knows how to use tools to wrangle the less than perfect data into a form that exposes patterns and relationships of interest to the business (or not). A data scientist typically spends up to 80 percent of their time trying to find, access, and prepare data for analysis. This is where metadata plays a crucial role in streamlining data R&D by facilitating search, provenance, collaboration, and reuse.

The tools that a data scientist uses in a discovery platform are very different than the tools used by a typical business analyst. A business analyst wants to obtain the answers to business questions and will use query and reporting tools to get those answers. On the other hand, a data scientist is not actually interested in the answer to a business question. A data scientist's job is to discover new questions rather than answers to existing questions. The tools used for this discovery are focused on exposing patterns and relationships in the data—rather than answers to business questions. Data visualization tools and advanced statistical analysis tools are much more useful to a data scientist than reporting and drill-down tools.

If analysis with an investigational data set does not expose any business value, the cost is relatively low for discarding it and retrieving a new set of data with which to experiment. More often than not, the investigational data set will be retrieved from the data lake infrastructure within the enterprise. The data set might also be retrieved from an outside data source or internal data source that has not yet been acquired into the analytic ecosystem of the enterprise, but will usually be staged in the data lake for retention. The often-voiced complaint about duplicating data to the discovery platform is mistaken: discovery platforms hold the data temporarily, then reuse that space for other work. The data set investigated may also need to be combined with existing data in an integrated data warehouse to expose value without re-creating dimensional or other data that has been previously curated in the analytic ecosystem.

Data that proves to have value will generally be promoted from the data R&D environment into the integrated data warehouse where data productization takes place. It is important that discipline is enforced to ensure that the discovery platform does not become a graveyard for lost or black market data. Black market data arises where data science workloads are completely ungoverned and of questionable accuracy. Data in the R&D area should either be promoted or discarded (but still kept in the data lake, of course). Running *black market* production out of the data R&D labs is generally considered a poor practice. Reliable audits are done in the IDW, not the discovery platform. Paying attention to the processes for promoting good ideas from data R&D into data products is a critical success factor for extracting full business value from innovation with data.

informatica Hortonworks TERADATA

Integrated Data Warehouse Environment

The integrated data warehouse (IDW) is where data is deployed as product for casual users and knowledge workers to drive decisions in the value chain of an enterprise. The decisions may be aligned to either strategic initiatives (strategic BI) or operational execution of the strategy (operational BI). A fully leveraged IDW will be deployed with access to its data products from across multiple lines of business and channels within the enterprise. There will be multiple subject areas of data (i.e., customer, channel, inventory, and suppliers) that have been integrated and deployed for access throughout the enterprise. Reuse of data using a service-oriented model for deployment (bring the processing to the data rather than replicating data for each analytic application) is essential for agility and leveraging investment in data assets.

In order for the data product to be consumable for the masses of knowledge workers, it must be cleansed, integrated, and prepared for ease of access by a wide range of business intelligence tools. Data integration tools play an essential role in readying the data for exploitation by knowledge workers. For data to be consumable, it must meet certain standards of quality—which will vary by application area. Certification of data quality, while not as important in the discovery platform, is essential in the IDW. In fact, for many uses of data as a product in the context of financial reporting and risk scoring, the requirements will be both for data quality certification as well as auditability of results - including lineage related to sources of data.

In addition to quality, there will be service level expectations in other areas related to the productization of data. Data freshness, system performance, and availability will be characteristics of an IDW that must be aggressively managed. If a store manager depends on a daily report that must be created at 6 AM before a store opens, then the data must be loaded into the IDW with a data freshness service level aligned to this requirement. And the IDW must support appropriate response times for data access—even when many, many knowledge workers are accessing the data at the same time. Workload management and efficiency with concurrent query execution are essential to meeting performance requirements when deploying data as a product. As an enterprise becomes more and more data-driven in its execution, the requirements for high availability for its IDW will escalate. Without access to data for decision making, the organization becomes paralyzed because the fuel for its execution has been shut off. For this reason, 7x24 availability is increasingly the requirement of IDW deployment in data-driven organizations.

Combinations of SQL and NoSQL database technologies will likely be used for data productization. The IDW is typically a relational database engine due to aggressive requirements for scalability and service levels. Operational data stores and data marts are likely to be deployed in a large range of technologies which include relational engines, but are not exclusively relational. JSON document stores, for example, are increasingly used to support operational data store (ODS) implementations for tactical decision-making with late-binding capability.

Unifying the Data

Ideally, no business user should be required to know the details of the technologies where their data is stored, nor should they login to multiple platforms to search for the data. Such fishing expeditions often lead to less than optimal choice of data sets for analysis and misinterpretation of data content. Furthermore, while most corporations have some or all of the UDA elements in use, there needs to be product and service offerings that simplify analytic ecosystems for business users.

informatica Hortonworks TERADATA

Several technologies exist and are emerging to make the various UDA environments behave as if they were one system. There are different technologies welding together the various UDA platforms. What makes the UDA a single architecture instead of technology islands includes:

UDA Unifier	Description	Example
Orchestrated data access	Pull data from remote platforms and combine it with local data	Teradata and Teradata Aster Databases QueryGrid; Informatica Data Services
SQL language	SQL is the language of the business user. All UDA platforms offer SQL access.	Hadoop Hive and Spark SQL Teradata Aster and Teradata Databases
Remote processing	Send process steps (Python, Perl, UDFs) to remote systems	Teradata Database remote processing; Informatica push-down to Teradata Database
Virtualized processes	Easy to deploy or relocate programs on any UDA platform	Informatica Big Data Edition and PowerCenter, powered by Informatica Vibe
Data relationships	Match and link disparate data sets	Informatica Big Data Relationship Management
Multi-Systems management tools	Monitor performance, jobs, and errors across platforms	Teradata Viewpoint and Teradata MultiSystem Manager

For example, some customers use the data lake to hold raw sensor data. Programmers sift through the sensor data to provide result sets to engineers. Meanwhile, the business users aim their favorite BI tool at Teradata Database which then issues a federated query to Hadoop Hive. The sensor data answer returned by Hive is joined to repair schedules in the data warehouse and delivered to the business user. To the business user, all they need is their favorite business intelligence tool, no programming is required. This saves labor costs at all levels while accelerating the ability to quickly get analysis done.

The following explains how one vital workload—data integration—operates in and across the UDA. Many of the functions above are used in data integration steps within the UDA.

Data Integration: Best Fit within the UDA

Industry analysts estimate that up to 80 percent of the work involved in big data projects involves accessing, integrating, and preparing the data for analysis. These data integration tasks include finding, gathering, parsing, normalizing, standardizing, and cleansing the data. This is what is often referred to as data preparation.

The UDA provides choices as to where to best deploy data integration tasks. For example, it often makes sense to stage large amounts of raw data in the data lake. Data aimed at sophisticated analytics should be copied to the discovery platform scratch pad. Data that is refined for extensive sharing, analytics, and applications will be integrated into the data warehouse. Therefore, determining the best fit UDA location for data integration workloads is a logical place to start.

Some factors to consider are the type of data, the size of the data, where the data is currently located, the type of transformations required, who needs to access the data, and for what purpose (e.g. data exploration, reporting, and archival).

informatica Hortonworks TERADATA

The Data Pipeline

From source data to the delivery of actionable information, it is essential to optimize the entire data pipeline. This includes data access, load or ingest, parsing, integration, and delivery. Some or all of these tasks can be performed in the data lake, the data warehouse, discovery platform, or a separate DI server. The business users must clearly state when the data arrives and when it must absolutely be available for analysis. While they often say the data must be available in real time, handing them the IT costs for such an effort often reveals they can easily tolerate hours or overnight delivery of the data.

Thus, for each data source and business objective, it is mandatory to define the end-to-end pipeline service level goals. Optimizing only one or two segments of the pipeline often creates bottlenecks at end points, which cannot scale or provide end-to-end high availability. These delays inevitably create service level disruption and dissatisfaction among internal customers. DI vendors and solutions should support different delivery modes, including batch, real-time, federation, and changed-data-capture modes. DI solutions should also support web services, real-time alerts and notifications, or topic subscriptions. Real-time data is often transformed in-flight and in-memory to avoid unnecessary hops and lost time in staging areas. While most data integration tasks are batch oriented, a good architecture will contain products to meet multiple DI delivery modes.

Many new data types are not structured in a traditional row/column tabular format. Many of them are in more complex and dynamic formats such as JSON, web logs, social data, machine and sensor data, unstructured text, and industry standards like FIX, SWIFT, HL7, HIPAA, and EDI. To derive value from these data sets, intelligent parsing capabilities can extract relevant data elements, placing them in the row/column format if required. Advanced data integration tools must include parsing and extraction from complex data formats whereas do-it-yourself development can be daunting and expensive over time.

Big Data Governance

Historically, the number one reason business users abandon a data mart or data warehouse is that the data quality is untrustworthy. After struggling with many reports and the BI Competency Center to repair the data, the users go back to their spreadsheets leaving the data mart to die from neglect. Worse, no one wants to be asked by auditors or lawyers to explain how a corporation failed to meet regulatory compliance requirements by using poorly governed data. Thus, data governance is important whether an enterprise is dealing with big data or little data.

It may be acceptable that some data is less refined as long as the users and management understand the level of governance and costs associated. For example, the discovery platform often has to deal with data that has never been seen before using agile development methodologies that discourage extensive data cleansing. In some situations, getting an answer with 70-80 percent confidence now is preferable to waiting a month for programmers to improve data quality. Each corporation should define the levels of trust (e.g. completeness, accuracy, conformity, etc.) for data sets and be able to offer a cost estimate for transforming the data to the next higher level of trust. The recommended approach to governing data is to implement a universal metadata catalog to discover data domains and relationships, track sensitive data, assess data quality, and recommend data transformations so that data is fit for use.

The best time to have these discussions with the users is long before errors appear in reports and analytics. One approach is to assign a confidence factor to critical files and answer sets, a practice sophisticated data miners have used for many years.

Unlike the cleansed data warehouse, the data lake and discovery platform may not need extensive data preparation investments for all the data files. For example, some files in the data lake require trustworthy level governance, but many do not. Yet, one purpose of the data lake is to retain

informatica Hortonworks TERADATA

full fidelity to the original raw state. In the discovery platform, considerable responsibility for data cleansing and conditioning sits on the shoulders of data miners and data scientists. This too is not as extensive as the data quality conditioning needed for the IDW. Regardless, programmers and data miners will need tools and skills to ensure appropriate data quality results.

Similarly, there is the constant concern of "where did this data come from?" Tracking data lineage occurs at two levels: 1) the original file and its derivatives; 2) the data elements within the files, either columns or name-value pairs. When an executive exclaims that an analysis or report value makes no sense, the programming staff often races to trace the aggregate number back to its origins. They search for mistakes made in IT, or proof that the data is what it is supposed to be. These concerns are heightened in the UDA, especially when a data lake is present. The data lake can have hundreds of millions of files, many with 5-10 derivative files. Tracing a data element from the source system to the raw file to the 5th derivative requires a lot of metadata and traceability tooling.

Data Integration in a Big Data World

Advanced data integration solutions have evolved to include more sophisticated intelligence about data and a universal metadata catalog encompassing new data structures, semantics, and data relationships. Self-describing data objects become increasingly important with schema-on-read where the burden of understanding the structure is left to the data consumer. Schema-on-read applied to new data structures alters the normal use of metadata, yet metadata must still be captured and exploited. Semantic metadata also expands to capture the business meaning of data, especially as relates to automated discovery and provisioning in the data lake. Best practices in capturing metadata relationships go beyond data provenance and lineage to encompass a social graph of who is using the data sets and a graph of related data domains. Knowledge worker collaboration can increase where usage relationships show data sets in common. The evolution to data intelligence functionality provides a smarter, wider context for data discovery, governance, and security.

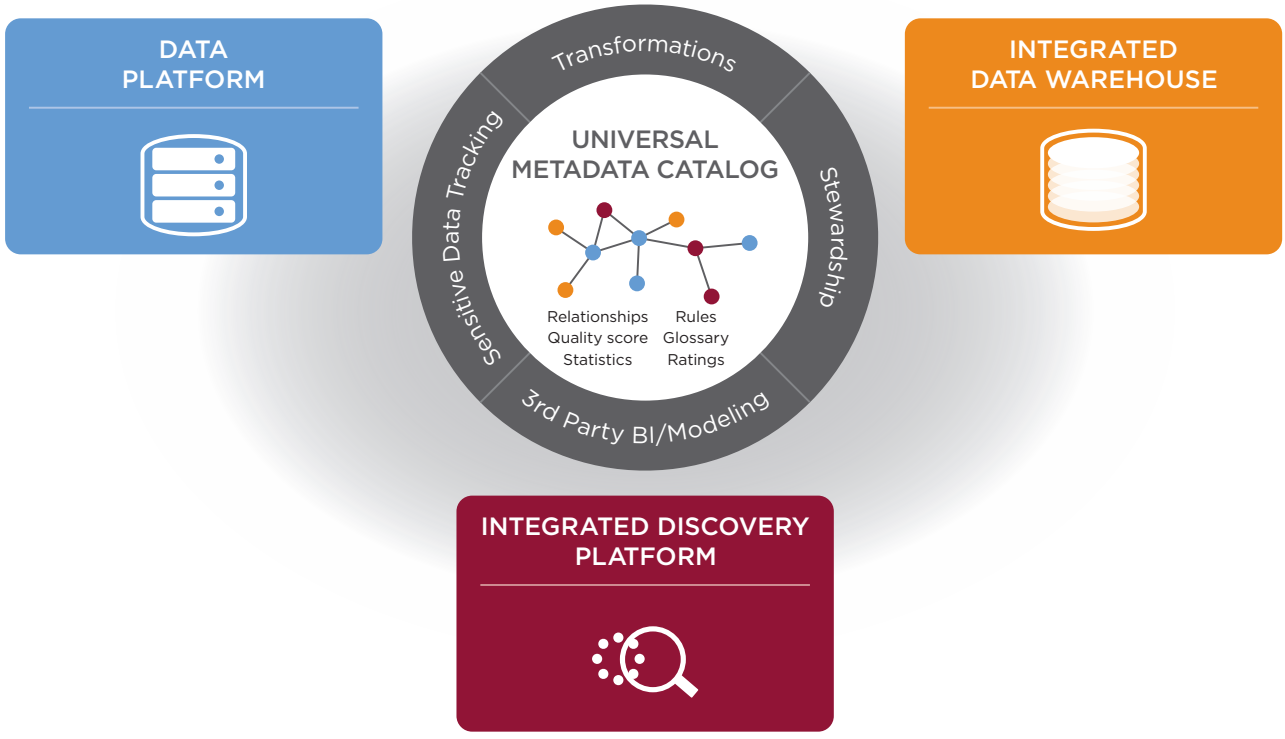

Figure 3. Data Intelligence and the Universal Metadata Catalog.

Flexibility in UDA workload placement protects your investment in the future. Informatica data integration processes and mappings are built once in a graphical user interface. They can be redeployed and redirected to new targets with minimal rework. This reuse of skills and processes is a compelling reason for enterprises selecting advanced and agile data integration solutions.

With all this flexibility built into the UDA, visionary organizations will optimize the data integration pipeline such that specific steps run on different platforms. Imagine a new collection of data to be transformed and cleansed. However, the data lake programmers cannot currently accept this workload while there are idle CPU cycles and storage available in the data warehouse. In this case, the data gets captured in the data lake using the data warehouse for extract-load-transform steps. Or, the data lake may perform filtering and transformations, passing the results to the data warehouse for final cleansing and house-holding tasks. Being able to place data integration pipeline steps where they best fit today and easily change them in the future protects investments in tools and labor.

Conclusions

The Unified Data Architecture provides a holistic design pattern for vendors and data centers alike to build out their big data strategies. It consists of the data lake, the discovery platform, the integrated data warehouse, and a data integration platform. Inside and across all of these environments, data integration is taking place in bulk and in real time.

The UDA provides business users and IT considerable flexibility in placing workloads on the best fit platform. While all platforms can perform some or all of a given task, there are trade-offs and optimizations to be considered. Good engineering matches the requirements to the best fit trade-offs, selecting the right tool for the problem at hand. The key is for the UDA developer to leverage what is known about the data (i.e., metadata) to determine which technology solves a specific problem in the best way.

A critical best practice is starting with the business users to clarify their requirements. This should be an ongoing journey, educating and including the business users in many of the architecture decisions. As requirements unfold, they should be mapped to the various platforms strengths as well as the labor pool available to implement those technologies. We recommend using flexible data integration platforms that can migrate workloads and data across the UDA. In summary, optimize and exploit the advantages of each UDA platform.

informatica Hortonworks TERADATA

Stephen Brobst

Chief Technology Officer, Teradata

Stephen performed his graduate work in Computer Science at the Massachusetts Institute of Technology where his Masters and PhD research focused on high-performance parallel processing. He also completed an MBA with joint course and thesis work at the Harvard Business School and the MIT Sloan School of Management. Stephen has been on the faculty of The Data Warehousing Institute since 1996. During Barack Obama's first term he was also appointed to the Presidential Council of Advisors on Science and Technology (PCAST) in the working group on Networking and Information Technology Research and Development (NITRD). He was recently ranked by ExecRank as the #4 CTO in the United States (behind the CTOs from Amazon.com, Tesla Motors, and Intel) out of a pool of 10,000+ CTOs.

About Teradata

Teradata helps companies get more value from data than any other company. Our big data analytic solutions, integrated marketing applications, and team of experts can help your company gain a sustainable competitive advantage with data. Teradata helps organizations leverage all of their data so they can know more about their customers and business and do more of what's really important. With more than 10,000 professionals in 43 countries, Teradata serves top companies across consumer goods, financial services, healthcare, automotive, communications, travel, hospitality, and more. A future-focused company, Teradata is recognized by media and industry analysts for technological excellence, sustainability, ethics, and business value. Visit teradata.com.

Scott Gnau

Chief Technical Officer, Hortonworks

Scott has spent his entire career in the data industry, most recently as president of Teradata Labs where he provided visionary direction for research, development and sales support activities related to Teradata integrated data warehousing, big data analytics, and associated solutions. He also drove the investments and acquisitions in Teradata's technology related to the solutions from Teradata Labs. Scott holds a BSEE from Drexel University.

About Hortonworks

Hortonworks develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities.

The Hortonworks Data Platform provides an open platform that deeply integrates with existing IT investments and upon which enterprises can build and deploy Hadoop-based applications.

Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop.

Sanjay Krishnamurthi

Senior Vice President and Chief Technology Officer, Informatica

Sanjay Krishnamurthi is senior vice president and chief technology officer for Informatica. Krishnamurthi is responsible for the company's technology and product strategy in addition to the architecture of the Informatica products. Previously Krishnamurthi was the chief architect for Informatica leading the architecture and direction of the Informatica Platform during the past decade. Krishnamurthi graduated with a Bachelor of Technology from Indian Institute of Technology, Madras. Additionally, he has a Master's in Computer Science from University of Wisconsin, Madison, where his research focused on performance of distributed and parallel database systems.

About Informatica

Informatica is the world's number one independent provider of data integration software. Organizations rely on Informatica to realize their information potential and drive their 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. Vibe harnesses data in every application, every process, for every person and in every device in the world. Organizations around the globe depend on Informatica to fully leverage their information assets from devices to mobile to social to big data residing on-premises, in the Cloud and across social networks.

QueryGrid and Unified Data Architecture are trademarks, and Aster, Teradata, and the Teradata logo are registered trademarks of Teradata Corporation and/or its affiliates in the U.S. and worldwide. Apache is a trademark and Hadoop is a registered trademark of the Apache Software Foundation. Informatica is a trademark or registered trademark of Informatica Corporation in the United States and in jurisdictions throughout the world. Hortonworks is a trademark of Hortonworks Inc. in the United States and other countries.

Copyright © 2015 by Teradata Corporation All Rights Reserved. Produced in U.S.A. 08.15 EB9037

informatica Hortonworks TERADATA