The Data Lake Design Pattern: Realize Faster Time to Value with Less Risk

By Chad Meley, Vice President of Product and Services Marketing

07.16 EB9461 BIG DATA

Table of Contents

Introduction

As data volumes and varieties of data have exploded, organizations have been searching for ways to economically harness and derive value from this goldmine of information. This “dark” data from new sources—such as web, mobile, and connected devices—is all too often deleted, but it contains valuable insight just waiting to be discovered.

These massive volumes of data along with new forms of analytics have made it necessary to find a new way to manage and derive value from data—with data lakes emerging in response to this need.

A data lake is a collection of long-term data containers that capture, refine, and explore any form of raw data at scale, enabled by low-cost technologies that multiple downstream facilities can draw upon, such as data marts, data warehouses, and data products like recommendation engines. A new data lake design pattern compliments traditional design patterns such as the data warehouse.

The Value in Data Lakes

Organizations are starting to realize value from data lakes in the four areas described here. The first two areas lead to better corporate effectiveness, while the last two improve IT efficiencies.

New insights from data of unknown or under-appreciated value

Prior to the big data trend, there was a single approach to data integration whereby data was made to look the same or normalized in some sort of persistence, such as a database, and only then was value created. While still true and valuable, this is no longer sufficient as a sole approach to manage all data in the enterprise, and attempting to structure all of this data undermines the value—particularly newer data where the value and extent of reuse is unknown. For this reason, dark data is often never captured. Yet many times, a data scientist digs through dark data and finds a few facts worth repeating every week. That is, they find gold nuggets.

New Forms of Analytics

One of the most under-talked about aspects of big data innovation is how new technologies like Apache® Hadoop®, Apache® Spark™, and other innovations enable the parallelization of procedural programming languages and how that has enabled an entirely new breed of analytics. These new forms of analytics can be efficiently processed at scale, like graph, text, and machine learning algorithms that get an answer, then compare that answer to the next piece of data, and so on until a final output is reached.

Corporate Memory Retention

Archiving data that has not been used in a long time can save storage space in the data warehouse. Until the data lake design pattern came along, there was no other place to put colder data for occasional access except the high performing data warehouse or the offline tape back-up. With virtual query tools, users can easily access the data in conjunction with the data in the data warehouse through a single query.

Data Integration Optimization

The industry has come full circle on how to best squeeze data transformation costs, moving from the standard architecture for data integration being ETL to ELT. Now, the data lake offers greater scalability than traditional ETL servers at a lower cost, forcing organizations to rethink their data integration architecture. Organizations employing best practices are rebalancing the hundreds of data integration jobs by using the data lake, data warehouse, and ETL servers to complete data integration, as each has its own set of capabilities and economics.

Data Lakes vs. Data Warehouses

Data lakes and data warehouses are both design patterns, but they are actually polar opposites. Data warehouses are an approach based on structuring and packaging data for the sake of quality, consistency, reuse, ease of use, and performance with high concurrency levels. Data lakes go the other direction, complementing data warehouses with a design pattern that focuses on original raw data fidelity and long-term storage at a low cost while providing a new form of analytical agility.

In a world with growing data volumes, the new challenge is that business analysts must manually find and reconcile fragmented, replicated, incomplete, and inconsistent data across the organization. As a result, analysts face delays in accessing needed data and quickly sharing it with one another. And, with the exponential increase in data volume and data proliferation across systems, business analysts run the risk of delivering inaccurate reports and predictions because of data that is insufficient, incomplete, inconsistent, inaccurate, or insecure.

Data Lake Pitfalls

On the surface, data lakes appear fairly straightforward—offering a way to manage and exploit massive volumes of structured and unstructured data. But, they are not as simple as they seem, and failed data lake projects are not uncommon across many types of industries and organizations. Early data lake projects faced challenges because best practices had yet to emerge. More recently, a lack of solid design is the primary reason data lakes don't deliver their full value.

As data lakes have been maturing, five common pitfalls have emerged:

- **Proliferation of data silos and clusters** – there is a notion that data lakes have a low barrier to entry and can be done makeshift in the cloud. It's not uncommon for there to be a new Hadoop cluster generated each time it becomes difficult to solve concurrency or security problems. This leads to redundant data and inconsistency with no two data lakes reconciling, as well as synchronization problems.
- **Conflicting objectives for data access** – there is a balancing act between determining how strict security measures should be versus agile access. Plans and procedures need to be in place that align all stakeholders.
- **Limited commercial-off-the-shelf tools** – many vendors claim to connect to Hadoop, but the offerings lack deep integration and most of these products were built for data warehouses, not data lakes. At this point in the still maturing data lake area, successfully loading data requires mostly custom development.
- **Lack of end user adoption** – users have the perception—right or wrong—that it's too complicated to get answers because it requires premium coding skills or they can't find the data they need.

The Data Lake Design Pattern

Being successful with a data lake requires planning, and a data lake design pattern is that plan. How the plan gets implemented varies from workload to workload and organization to organization.

A data lake design pattern offers a set of workloads and expectations to help guide a successful data lake implementation. As data lake technology and experience have matured, an architecture and set of corresponding requirements have evolved to the point where leading data lake vendors have agreement and best practices for implementations. Technologies are critical to the outcome, but before anything gets built, it needs a plan (design pattern). This approach of creating a design pattern before building a data management environment has served the data warehousing market well for decades.

It's important to note: Design patterns are independent of technology. A data lake can be built on multiple technologies or combinations of them. While Hadoop is what most people think of first, it is is not required. Data lake technologies also include Amazon S3, Cassandra, and the Teradata Integrated Big Data Platform.

Data Lake Solutions

Teradata designs and builds enterprise-class data lakes, drawing on more than 35 years of experience in developing and implementing enterprise data warehousing design patterns, coupled with Think Big—the world's first, pure-play big data services firm, focused on generating business value from big data.

The Teradata data lake solution involves services, products and reusable intellectual property (IP).

Teradata offers quick hitting services using baseline templates, best practices, established frameworks, and IT accelerators to reduce risk and achieve faster value. Data lake services from Think Big encompass three areas:

Architecture

Teradata architecture services deliver value through a variety of components that include high level architectural design and technology accelerator recommendations and organizational readiness, training requirements, and workload considerations. Additional components include cluster, configuration and performance optimization; data management and ingest blueprints and recommendations; metadata gap assessment and management; security assessment and planning; governance models; and data export assessment and planning.

Foundation

Teradata further delivers value through a number of data lake foundation components, from high-level architectural design and linkages to detailed designs for areas covered, a cluster optimization run book, and data management and ingest services to revise existing ingest patterns and refit data previously ingested. Foundation services also include metadata management, security implementation, and data archive implementation.

Analytics

Data lake Analytics is a custom offering drawing from a range of potential components to meet specific needs, such as modeling and materialization, which involves mapping data into optimal analytics readiness on the target platform. There is also publishing implementation to select the self-service user interface and user-defined data export format definition; defining source data encryption and obfuscation criteria, as well as secure downstream export implementation to provide proper user roles with secure access.

Teradata products, such as Teradata Listener, Teradata Aster Analytics on Hadoop, engineered appliances, and Presto, make data lakes easier and faster to build and use. Teradata self-service tools and accelerators deliver reliable data ingest and powerful multi-genre analytics for the data lake.

Teradata Listener™

Teradata Listener is an intelligent solution for ingesting and distributing extremely fast moving data streams throughout the analytical ecosystem. A self-service dashboard that can be accessed by multiple users simplifies data ingestion and makes configuration of data sources and targets a simple task, eliminating the need for programming. A Teradata Listener cluster of servers can scale horizontally to meet growing demands of multiple data streams in the enterprise. And, Listener is built leveraging proven open source projects like Kafka, Cassandra, Elastic Search, and Mesos, along with modern software engineering based on Docker, micro services, and RESTful APIs.

Teradata Aster Analytics on Hadoop

Teradata Aster Analytics for Hadoop is a multi-genre advanced analytics solution that provides more than 100 powerful, business-ready analytics such as path, pattern, machine learning, text, graph and statistics at scale. These various functions can be intelligently combined in a multi-genre manner to address virtually any use case across verticals. Furthermore, integrations with R and Spark empower various user personas to extend their advanced analytic implementations and to operationalize critical insights in the same environment. Aster Analytics for Hadoop is also a native Hadoop application and a first-class citizen of Apache Hadoop YARN allowing users to scale Aster instances to support a variety of use cases from experimental sandboxes to production analytic systems accessing the same data in HDFS. Organizations can now unleash the magic of Aster Analytics for Hadoop to quickly and easily uncover non-intuitive insights accelerating time to value.

Presto

Offering exceptional production performance, Presto is an open source distributed SQL on Hadoop query engine for running interactive analytic queries against data sources of all sizes, ranging from gigabytes to petabytes. Presto leverages standard ANSI SQL and has been architected from the ground up for high performance interactive query processing against Hadoop and other data sources—making the data lake accessible. Teradata is contributing to Presto's open source development and providing commercial support to help increase Presto adoption and make it easy to install and maintain.

Engineered Appliances

Performance hurdles, prolonged implementation periods, and reliability issues—are solved by the Teradata Appliance for Hadoop when compared to solutions that are not preconfigured. Teradata does the hardware and software integration plus plenty of testing so you don't have to do it. The Teradata appliance is delivered ready-to-run and optimized for enterprise-class big data storage and discovery.

Get deep strategic insights from massive amounts of data with the Teradata Integrated Big Data Platform—the lowest cost per terabyte in the Teradata Workload-specific Platform Family. Analyzing multi-structured data began with Teradata Database 14 when name-value-pair functions and regular expressions enabled Teradata sites to process web logs using popular business intelligence tools. The Teradata Integrated Big Data Platform supports workloads such as deep history analytics, storage of massive amounts of multi-structured data, and a raw data landing zone for transformations.

eBay Analyzes Click Streams to Enhance Auction Sites and Buyer Searches

Auction website eBay began a project several years ago to store all customer data because approximately 85 percent of the analytics questions users ask are new or unknown. Imposing structure and throwing out data would mean users could not ask questions they didn't know. Conversely, if eBay stored everything, there would be 100 million hours of data per month and users would not have been able to analyze all of it. The auction site needed a product that could handle hundreds of petabytes of raw customer journey data, but would be easy to maintain by a team of five people, yet could be accessed easily by analysts. The company worked with Teradata to develop a custom appliance built with several hundred user-defined functions. The data lake system eBay has developed can run ad-hoc queries in 32 seconds, as opposed to queries that would have taken 30 minutes in Hadoop. The system is proving its value in 'A/B testing' of unstructured weblogs on the eBay site. This allows eBay to test ideas on the site and assess what works, such as testing whether site visitors prefer larger pictures in search results. Power users transform weblogs into buyer preferences, which are joined to consumer profiles. The ultimate result is that eBay auction sites and buyer searches have been enhanced for a better visitor experience.

Data Lake IP Accelerators

Teradata offers three data lake IP accelerators:

- **Dashboard Engine** – makes it easier to access data in Hadoop through BI tools such as Tableau at speed and scale by serving as an analytic engine that provides pre-aggregated and pre-calculated data from Hadoop (stored in HBase) to enable sub-second responses.
- **Buffer Server** – is used to move data from local servers into Hadoop, with the file system serving as the lowest common denominator for batch dumping into a directory.
- **Pipeline Controller** – is an orchestration framework for processing data, including movement of files from local servers into Hadoop. Pipeline Controller can navigate the directory, find the data, and transfer it.

Teradata also has extensive data lake partnerships to provide trusted advice without bias, including with all the major Hadoop distributions (Cloudera™, Hortonworks™, MapR™, IBM™), plus a host of partnerships with the leading providers of metadata, data integration, security and beyond.

Perhaps most importantly, Teradata captures and reuses existing data lake IP and brings this IP to every project for faster and more reliable implementations. This recycled IP includes predefined reference semantic data models for the access layer that are not limited by platform, as well as gathering unstructured and structured data and pre-calculating the combinations of dimensions and facts into aggregated result sets for consistently fast response times.

The path to implementing a successful data lake can be challenging. By following a trusted route that experts have already mapped out, that path can be much easier.

For More Information

To find out more about how Teradata can put the power of data lake best practices to work in overcoming your big data challenges, contact your local Teradata representative or contact us through **Teradata.com**.

About Teradata

Teradata helps companies get more value from data than any other company. Our big data analytic solutions 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.

About the Author

Chad Meley is Vice President of Product and Services Marketing at Teradata. Chad understands trends in the analytics and big data space and leads a team of technology specialists who interpret the needs and expectations of customers while also working with Teradata Labs engineers, consulting teams and technology partners such as Cloudera and Hortonworks.

Prior to joining Teradata, he led Electronic Arts' Data Platform organization that supported Financial Analysis, Game Development Insights, and Marketing Analysis and CRM. Chad has held a variety of other roles within data warehousing, business intelligence, Marketing Analysis, and CRM while at Dell and FedEx.

Chad can be reached on Twitter at @chad_meley

10000 Innovation Drive, Dayton, OH 45342 **Teradata.com**