IBM®

Future-Proof Your Hybrid Data Strategy

With an Enterprise-Grade Data Lake

Data lakes are emerging as the next generation of hybrid data management solutions, meeting the challenge of the increasing volume, velocity and variety of today's data being driven by artificial intelligence, Internet of Things (IoT), cloud, mobile and other new technologies.

Historically, relational databases and data warehouses have provided the foundation for data management. However, today's data warehouses are unable to process semi-structured and unstructured data such as streaming audio and video, social media, clickstream, logs, sentiment, etc. Additionally, they are unable to accommodate the growing audience of users, including data scientists, analysts, line-of-business owners and developers who are seeking to eliminate their reliance on IT for immediate and ad hoc access to their data. Because of these limitations, businesses are turning to data lakes to leverage more types of data and make it more accessible throughout the organization.

When built and implemented correctly, data lakes provide a secure, governed storage repository where users can self-serve, drive advanced analytics, implement machine learning and develop more relevant applications. Data lakes drive user self-service, and federate data to break down organizational silos and enable real-time analytics supporting a 360-degree view of the customer, processes and operations for better predictions/decisions at the right time. Data lakes also provide massive scalability and cost efficiency for unlimited amounts of raw, unformatted data. However, when designed and implemented poorly, the data lake is no more than a "data swamp"—disorganized and unusable with little value to the organization.

To get the most value and benefit from your data lake and avoid the creation of a data swamp, make sure to avoid the following pitfalls:

NO BUSINESS CASE

The first step in developing a solid business case is to know how the data lake fits into the overarching hybrid data management strategy of your organization. This entails understanding the capabilities and challenges of both your enterprise data warehouse (EDW) and/or data mart, in addition to a potential data lake. It is important to consider that in most cases the data lake will not replace but integrate with and augment your EDW.

Developing use cases for the EDW and the data lake will strengthen your business case. Before data is added to the EDW, it is cleansed and processed, and this structured data is highly secured and governed by IT. Outputs include pre-defined reports, production with historical comparisons, customer analysis including segmentation, KPI calculations, profitability analysis, and more. The limitation of the EDW for the business is the time and cost of mining data.

Data lakes are built to accommodate a new world of streaming, semi-structured and unstructured data. Characteristics of the data lake include user self-service, ad hoc and real-time data query, analysis and decisioning. Use cases revolve around un-planned data exploration by new user groups, cause/effect analysis, and pattern analysis, to name a few. For customer service, this might equate to the suggestion of the next best or location-based offers, real-time fraud detection and supply chain management. For IoT, stream processing provides continuous process automation from operational systems that merge and compare performance to historical data. This drives use cases ranging from in-home "smart" thermostats to measuring, monitoring, and shutting down factory equipment.

Data lakes and the EDW should be viewed as complementary technologies and this should be called out in your business plan. Additionally, the business plan should include cost analysis for each, a planned migration strategy, assessment of existing skill sets, next steps and proposed timelines.

WRONG TECHNOLOGY CHOICES

One of the most critical steps to building and successfully managing a data lake is choosing the right platform and related technologies. Apache Hadoop® is growing in popularity as the platform of choice. It is a highly scalable, open source software storage repository designed to process very large data sets across hundreds to thousands of computing nodes operating in parallel. Since it is community built, adopters benefit from continuous improvements and cost efficiencies.

There are limitations when implementing the free version of Apache Hadoop for the enterprise. Some of these limiting factors include the lack of enterprise-grade security, access control, compliance, and the ability to control, manage and track the data throughout the lifecycle. It is important that your organization assess if it has the right skill set to ensure that Apache Hadoop meets the required standards of your organization. Due to the variety of the data and the number of new users, there is a heightened need for data management, provisioning and data governance. This requires considerable add-ons to the free downloadable version of Apache Hadoop.

IBM and Hortonworks have partnered to offer an enterprise-grade Hadoop distribution with data integration and advanced querying tools, Hortonworks Data Platform (HDP) and Hortonworks Data Flow (HDF) in conjunction with IBM Db2 Big SQL. This solution offers massive scalability, security and governance, and the ability to federate both data-at-rest and data-in-motion across the organization, spanning relational databases and Hadoop, whether on premise or in the cloud. Users benefit from self-service data access, the ability to do ad hoc and real-time queries for predictive analytics and better data driven decisions.

INADEQUATE FOUNDATION FOR DATA GOVERNANCE, COMPLIANCE, SECURITY AND AUDITING

Many early adopters of data lakes believed that conventional methods of data preparation, management, governance, and security would work the same as they would in a traditional data warehouse. Unlike in the data warehouse, data in the lake is not cleansed or formatted when ingested. Since it is composed of raw data, ingestion, governance, security and management become even more critical.

Governing, securing and managing the data lake is complicated because of the immense variation and quantities of data but also the variety of users (data scientists, line-of-business owners) wanting self-service access and ownership of their data. This necessitates a plan to ensure that each user group can easily find, understand and duplicate their data while maintaining overall security and governance.

The EDW is typically owned and only accessed by central IT. Ownership in a data lake can be divided in many ways:

- **Co-ownership—**A line of business (LOB) may determine user access and dictate the proper security and compliance needed to protect their sensitive data, while central IT would ensure adherence to overarching standards and processes, communicate best practices and perform timely updates and audits.
- **IT ownership—**When IT has full control of the data lake, it implements standard governance, metadata formats and best practices across the data lake.
- **Line of business—**When a line of business has partial or full control of the mechanisms that operate the data lake, it has the responsibility for data classification and identification of the different data types that need to be abstracted through services and metadata. This creates a view of the data lake that makes sense to the business and can be modified as needed.

LACK THE RIGHT TOOLS FOR DATA INTEGRATION AND ANALYTICS

One key advantage of data lakes is the ability to federate disparate structured, semi-structured and unstructured data from sources across your organization. Having access to this broad range of data drives organizations to more accurate analytic predictions and decisions. IBM and Hortonworks offer Hortonworks Data Flow (HDF) in conjunction with IBM Db2 Big SQL, helping organizations drive data integration and fuel advanced analytics.

To optimize the value of data lakes, a real-time and enterprise-grade streaming platform that connects on-premise deployments with cloud is necessary. Available through IBM, the HDF platform offers the only end-to-end platform that collects, curates, analyzes and acts on data in real time. HDF integrates with Apache NiFi/MiNiFi, Apache Kafka, Apache Storm and Druid. This allows users to collect and manipulate big data flows securely and efficiently while giving real-time operational visibility, control, and management.

Once data is collected, IBM Db2 Big SQL supports ad hoc and complex queries, high performance, security, and SQL compatibility. Db2 Big SQL uses a single database connection or query to connect to disparate sources such as HDFS, RDMS, NoSQL databases, object stores and WebHDFS.

CONCLUSION

Data lakes represent the next evolution of hybrid data management built to capture new formats in addition to the growing volume and velocity of data. To stay competitive, companies are capturing and analyzing customer sentiment expressed on social media, data streaming from IoT sensors, typed physician notes, weather data, audio from call center interactions, email correspondence, surveillance video and much more. They are using this data to proactively improve customer experience, detect/prevent fraud, correct operational failures and improve processes.

When planned, designed, implemented, governed and secured correctly, data lakes help organizations to integrate data sources, streamline ingestion and preparation, provide real-time data access, reduced costs and improved analytics.

Register here for a no cost trial of Db2 Big SQL Sandbox to get started today: http://ibm.biz/db2-big-sql-trial-dbta. This personal desktop environment is preconfigured with sample data, a tutorial and an exercise to help you test, and try new software features.

IBM
www.ibm.com