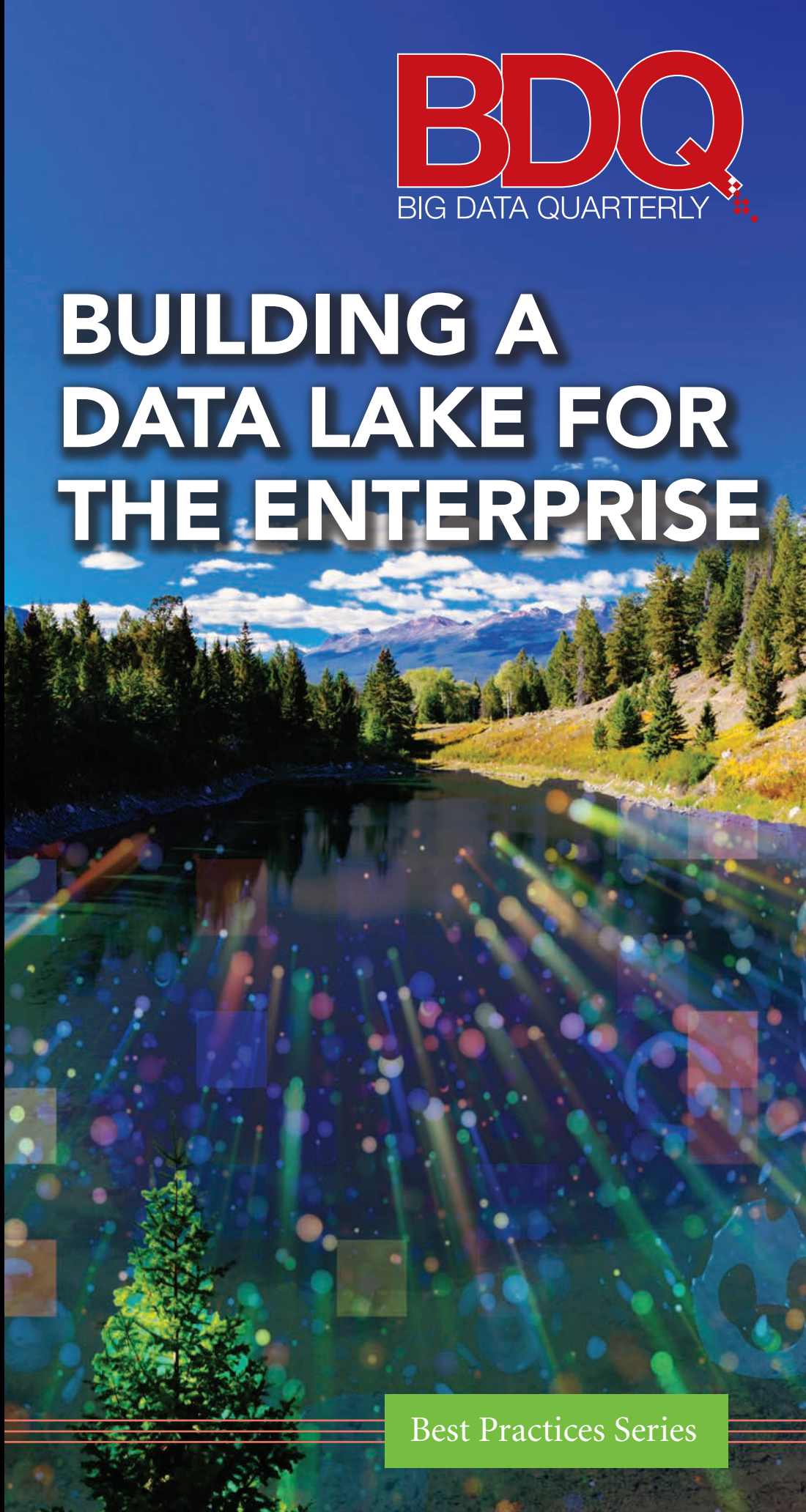
BDQ
BIG DATA QUARTERLY

BUILDING A DATA LAKE FOR THE ENTERPRISE



Best Practices Series

BUILDING A DATA LAKE FOR THE ENTERPRISE

Best Practices Series

In today's environment, businesses are feeling the heat of competition from all sides, in many cases from disruptors within their industries who are wiping their markets clean with technology and data. To compete, businesses need innovation; and to innovate, they need data.

There's a fire hose of data moving through enterprises, and ongoing analysis from Unisphere Research shows that many organizations already have data well into the petabyte range. However, not enough of this growing pile of data is making it over to the tools and platforms that decision makers use or to any accompanying decision-making applications. A total of 31% say the majority of their data actually makes it to the analysis stage.

Big data is enabling many types of business opportunities—from predictive analytics to the Internet of Things. IoT in particular is making it critical to be able to take lots of data feeds, pull the points that are of material importance to customers, and engage with them in real time. Other emerging initiatives include artificial intelligence, machine learning, and robotic process automation. Most of the data that is generated is not captured. As it moves through the enterprise and is discarded, it ends up in silos and locked away in proprietary databases. It may even end up on tapes stored in a basement.

Why isn't more data being made available to help the organization? There are still limitations on the data that is moving into the analytics stage; it still gets shepherded through the extract, transform, and load model—or ETL—which is not going away anytime soon. ETL is a technology and methodology that has worked well for enterprises for more than 2 decades, especially for bringing data from enterprise sources into one place where it can be examined, such as within a data warehouse.

However, the ETL methodology is expensive, and with more data flowing in, those costs will invariably rise. Aggregating, transforming, and reloading data between systems takes time and resources, requiring investments in people, skills, and systems. The time it takes to move data from one system to another doesn't work in a world demanding real-time, or near real-time responsiveness. Current ETL systems

may require days, or even weeks, to move data between sources and analytic or decision-making systems.

Is an ETL-centric infrastructure ready for the challenges ahead? What is needed is a way to have critical data ready and waiting when and where it is needed. That's where the data lake comes in. A survey by Unisphere Research and Radiant Advisors found that close to four in 10 enterprises, 38%, support these environments. These are still the early days of data lake adoption, and streams are just starting to flow in to fill the basins being created within enterprises. The Unisphere Research-Radiant Advisors survey found the bulk of the data lakes do not yet exceed 50TB, with more than one-third still less than 5TB.

In the past, the focus for data lakes was on storing, staging, or transforming data, or creating a place where data scientists could use advanced coding and analytics to extract insights. However, it has been recognized that too much value in that data is being kept away from the business. New BI and analytics tools which are native—or uniquely suited—to the data lake are becoming standards for organizations that want the business to gain insights directly. Data lakes help address the challenge of managing widely disparate data sources and the inertia it fosters within enterprises.

Data lakes help address the challenge of managing widely disparate data sources, and the inertia it fosters within enterprises.

Data lakes have connotations of fluidity, resembling a body of water in its natural state. Data flows from the streams—in other words, the source systems—to the lake. Taking the watery analogy a step further, enterprises need to dip their toes cautiously into data lakes—not because the technology isn't sufficient to support these vast data stores, but because the people and organizations that use them may not be ready. There are issues that need to be considered in terms of data lake governance, regulatory compliance, security, and access.

If data is akin to water, then it's very expensive and difficult to move once it lands someplace. A typical data warehouse architecture is a highly controlled and highly directed environment. A data lake architecture, on the other hand, is much more diverse. A data warehouse still remains a key player—it doesn't get taken out of the picture—but there are more options for working with data that surges into and through enterprises. ETL technologies can't keep up with the data volumes and data variety in today's enterprise. Data warehousing was revolutionary in its day, but lately, the volume and variety of data have grown beyond enterprises' ability to pre-process it, to pre-structure it, or to answer questions. The ETL-based, relational data warehousing model that worked fairly well over the past 25 years just simply is not built for today's data analytics-driven world.

In today's forward-looking enterprises, a movement is being seen from relational to frameworks such as Hadoop or Spark. These newer systems can handle the exploding variety and volume of data—but enterprises can't drag along the same techniques they've been using over the past 25 years in this new data architecture landscape.

The rise of the data lake means thinking differently about the way data serves the business and the way data and IT managers serve their users. The older mode, or data warehouse thinking, is vanishing, but the implications are much broader. For starters, data lakes are also becoming more operational and real time in nature. The core of most data lakes is the Hadoop ecosystem, which now supports and enables more real-time and streaming analytics using projects such as Apache Kafka and Apache Kudu.

Importantly, the data lake architecture helps introduce fresh thinking on the possibilities data provides to innovation. Many innovations that may be driving companies 1, 2, or 3 years from now may not have been put into action—or even conceptualized—yet.

Innovators are going to need data to make their ideas work. Innovation can't be stifled by locking it away in a silo somewhere, out of reach, limited or encumbered within an ETL mechanism, or, even worse, its existence a mystery. Data bottlenecks cannot be allowed to get in the way of innovation. Decision makers and innovators need to be able to process and visualize data and to look at the many, many different variables and attributes. Before, a data model could only support a limited number of dimensions—10, 20, maybe 50 dimensions. Now, innovators may have thousands of dimensions they want to look at. Businesses these days thrive on event-driven architecture, even in real time, and every event that gets captured becomes another attribute.

It's important to look at the data lake as an enabler of innovation. There are applications and services that have yet to be written or that have yet to be conceived by today's entrepreneurs, intrapreneurs, employees, and scientists. It's not known what they will use to get at the data they need to make things happen—and they probably don't even know yet, either.

—Joe McKendrick

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

ORACLE®

Stretching the Limits of Data Lakes with Streaming Data

Autonomous cars, virtual, and augmented reality, financial transaction monitoring, and infrastructure maintenance sensors—these are all examples of new technologies that are changing how we live our lives and how business is done. But they're also just a few examples of the types of the technology that will stream data at increasingly higher levels in just the next few years.

Being able to work with and analyze that streaming data will not only open up new doors for your business, but will govern how well you're able to capitalize on these new technologies. Today, the data lake is being hailed as a miracle, a sort of a catch-all for data. Of course, there are different kinds of data lakes for different data needs. But the truth of the matter is that all data lakes will have to have a certain set of capabilities to be successful in the near future.

Let's break down this concept of streaming data and its connection to data lakes. And for the sake of this example, let's take autonomous cars. Autonomous cars will change the way we live our lives, but they'll also change the way we handle our data. Right now, autonomous cars are in the testing phase—in different countries, on various roads, with different traffic and weather conditions. But when they appear on the road en masse, the data they generate and use will be in the petabyte, or even the exabyte, range. Much of that data will be streamed in real time, which will put even greater demands on data lakes.

The autonomous car's data needs will be truly enormous. According to some sources, the average self-driving car with GPS, LIDAR, cameras, and more will produce about 4,000 GB of data per hour of driving.

Here's a look at why they will produce and need so much data:

1. Spatial and Mapping Data

Self-driving cars need accurate maps to perform well. As a result, autonomous cars will require a broad mix of geospatial technologies to be successful. For example, the cars will need a 3D model of the area they're navigating. These maps will have to be extraordinarily detailed and updated regularly to safely ferry people around through lane changes, night driving, and more. That data will move both ways: Updates will get pushed to the cars, and anything new that the car encounters will update the map.

2. Normal Operation

Under normal driving conditions, cars generate massive amounts of data from the various sensors that scan the environment and manage the operation of the vehicle. The car will have to sense everything from bikes or deer in the road to approaching cars and pedestrians to different weather conditions, and decide what to do about them. This data has to be sent, received, and analyzed in a split-second in order to avoid a wreck.

And what about wear and tear to the car's components? Ideally, data would be sent back to the manufacturer to ensure parts are behaving optimally and to the car's owner to remind them when it's time for preventive maintenance. The data could also be used for liability and diagnostics investigations after a crash.

3. Human Behavior

The driver's behavioral data can also be recorded. What is the driver's personal driving style when he's operating the vehicle? How should the autonomous car adjust to match it? Will the car be expected to "help" the driver if it senses danger?

As drivers rely more heavily on their cars' autonomous capabilities, they won't need to pay as much attention to the road and will look for more entertainment options. Not all of this will come from their personal mobile devices. That data will need to be streamed to and from the car.

WHY STREAMING DATA IS ESSENTIAL

In the case of the autonomous car, timely and accurate data isn't just important—it's essential. There are currently over 10,000 fatalities and as many as 500,000 injuries related to traffic accidents per year in the US alone. Autonomous cars could come close to eliminating these numbers, but only if the data is timely and accurate.

Speed has always been a necessary part of big data. But new use cases, like the autonomous car, put new demands on big data. It's not just that you need the data fast or faster—it's that now, you need it in real time.

To complicate matters further, a lot of the data is constantly changing. It's coming in streams (hence, the phrase "streaming data"). It's dynamic. And it's not data that can be handled in batches.

For example, if the car's sensors are detecting hazardous conditions, it can't sit on that data for a few hours or even a few seconds. It needs to be able to use that data to make decisions immediately.

Companies using autonomous cars will need a data lake that can handle streaming data. While your particular use case might not generate as much data as an autonomous car (or have the potential to be a life-or-death situation), you'll likely need to consider some of the same factors to make your data lake successful.

WHAT DOES A SUCCESSFUL DATA LAKE NEED?

Low Latency: If you need messages to go back and forth between the data center and another point quickly (say, in a hundredth of a millisecond), you're going to need a data lake that can process a high volume of data with very little delay.

Availability: It's hard to think of a use case where you don't need your data to be available. But streaming data takes the importance of availability to a new level. Can your data lake ensure that your data won't be lost and guarantee that all messages will come in?

Volume: You've heard it time and time again. Data is being generated at never-before-seen volumes. You've heard this so often because it's true, and it's only getting worse. Expect to have to handle petabytes or even exabytes of data.

Analytics: You can perform simple analytics on data in the data lake, but there are times when you want to perform analytics in the event stream as the problem is happening. Successfully running, say, a machine learning model in real time, typically also requires access to a wide range of data sitting outside of the stream.

STREAMING DATA USE CASES

Streaming data is critical to many different use cases. Here are just a few examples.

Infrastructure Maintenance: From monitoring bridges to rail tracks, there are an increasing number of ways to use data in the infrastructure world. Smart cities are being built out with real-time data streams on items as diverse as traffic flow, power consumption, and air temperature. All these events will need to be analyzed and acted upon in real time.

Virtual or Augmented Reality: Virtual reality and 360°-viewing experiences are growing in popularity, and data will be key. In addition to the core graphics or visuals, monitoring people's reactions, acknowledging when participants are getting fatigued or bored, and keeping the user relaxed and able to focus will add to the data stream. Real-time responses will be essential to maintain the experience.

Financial Transactions: Every purchase or online transaction is an opportunity for fraud, costing consumers billions of dollars a year. The data behind these actions flows in a fast-moving stream, and with the right analytics on that stream, fraud can be stopped in its tracks.

Improving Customer Experience: Your customers are constantly generating information as they move, shop, and interact with your company. Harnessing that stream of data and applying the right analytics to improve the customer experience is important for all companies and critical to survival for some.

These are just a few of the wide range of different use cases that depend on streaming data. Take a look at your own big data problems, and you should see the importance of streaming data there as well.

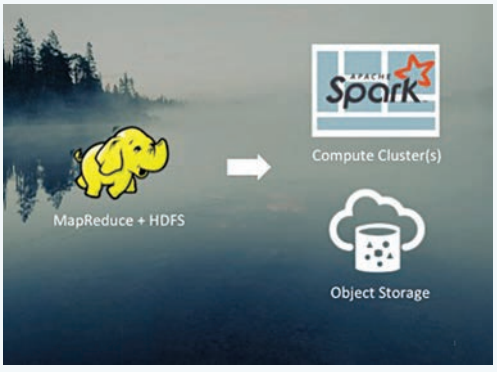
CONCLUSION

Being able to work with and analyze streaming data in the data lake opens up new doors for you. Whether you're working with the autonomous car or something else entirely, there's a lot more that can be done with streaming data.

Oracle provides a data lake that can handle streaming data. If you'd like to build a self-functioning data lake that also handles streaming data, we have a free trial you can experiment with. The trial will lead you through the process of building a small but functional data lake based on publicly available data.

ORACLE
Big Data Journey
https://go.oracle.com/bigdatajourney

THE NEW DATA LAKE



Traditionally, data lakes have been built on-premises, based on Hadoop. The cloud offers a new way. Replacing MapReduce and Hadoop Distributed File System with Apache Spark and cloud-based object storage, results in a far superior data lake. The right integration with Apache Kafka can address streaming requirements. This new approach can deliver better agility, improved stability and reliability, and lower cost of ownership.

Learn more at
https://blogs.oracle.com/bigdata/what-is-object-storage

collibra®

Stop Drowning In Your Data Lake

The big data landscape is changing. Large players like Amazon, Google, and Microsoft are joining the big data game. And their weapon of choice is cloud infrastructure. Their interest and investment signal that the data lake market is big—and the potential is even bigger. But their entry into the market—coupled with the increasing desire of organizations to move data lakes to the cloud—signals maturity. And, as the market matures, so does the conversation. It's shifting from volume to usage.

Until recently, dumping vast amounts of raw information into a data lake was an acceptable practice. Data junkies would plumb the depths of the lake to find the raw data needed to solve complex data puzzles. Today, that's no longer enough.

Now, business users skim the data lake looking for information to solve customer problems. But discovering a raw data table floating on the surface of the lake is not what they were fishing for. Instead, they want refined data sets, complete with quality scores, documented lineage, and established usage policies. Unfortunately, most data lakes fail to deliver what business users need, so they struggle to:

- Find the data they need
- Understand what it means
- Trust that it's right

When you don't meet these basic business needs, the business suffers. Report quality is called into question. Adoption of self-service BI and analytics tools languishes. Or worse, the lake becomes a stagnant pool of unusable information.

As a data leader, you must save the business from drowning in the data lake. Moving your data lake to the cloud is a good first step, but without solid governance practices in place, your users will remain under water.

THREE BEST PRACTICES FOR GOVERNING YOUR DATA LAKE

1. Clearly define data policies *before* data enters the lake: Dumping data into your data lake with no clear purpose isn't effective. Clearly-established data policies help everyone know which data sets belong in the lake—and which do not. Aligning your data policies to business objectives ensures that the data entering your lake has purpose, which increases the likelihood of adoption.

2. Document data lineage: Failing to include data lineage at the time of ingestion causes big problems downstream. Think about it: Omitting lineage from the raw data means that any data set, report, or algorithm that uses that data also fails to have lineage. Including data lineage from the start gives business users information about the data, including its relationship to other data; its origin, format, and use; and how it is organized, classified, and connected.

3. Link to your established business glossary: It's common knowledge that organizations have multiple views of the truth—and that each must be trustworthy. By connecting your data lake to your established business glossary, users can understand the context of the data, in familiar, well-understood language.

BUILD TRUST WITH A DATA CATALOG

A data catalog encourages cross-organizational collaboration to understand the data's meaning and use, making it consumable for a broader audience of data users. It includes data and data sets, glossaries and definitions, reports, metrics, dashboards, algorithms, models, and more. It defines rules and operating models for how data should be ingested into the data lake. And it allows users to add context, which helps everyone understand the data. A data catalog also helps users determine which data is fit for purpose and which data is unusable, incomplete, or irrelevant.

At the core of a data catalog is trust: trust that the right data is entering the data lake; trust that it is the right data for the data consumer to use; trust that data is accurate and of high-quality; and trust that the usage of the data aligns with provider and organizational policies.

When evaluating data catalogs, you must ask the tough questions. Does the data catalog provide transparency into the data journey? Is it easy for users to know if the data, algorithm, definition, or report they want to use is still certified—or if has changed? Can users tell if they can share data—or if it must remain internal? Finally, does it allow you to put data sharing agreements in place to answer these questions—and more?

With the right data catalog, your data lake becomes more than a data repository. It becomes a valuable business asset that delivers the information your users need to accelerate decision making, discover better business processes, deliver an amazing customer experience, and inspire ground-breaking discoveries.

LEARN MORE AT
www.collibra.com/data-catalog

ARCADIA DATA

Native BI for Data Lakes:
Delivering on the Promise of Data Agility

Many businesses today rely on the same data warehouse infrastructure that they've relied on for years. Those same businesses have also turned to data lakes as a cost-effective way of quickly gaining deeper insights across large volumes of disparate data sets. Unfortunately, the success rates of these data lakes have been disappointing. Many of these data lakes were not able to deliver faster value, particularly to business users.

Why isn't the data lake living up to its promise? It turns out that the data lake architecture is not the problem—it's the choice of tools. The data landscape continues to change, but not all tools have kept pace.

To elaborate, we know that data has evolved over the years. Complex/unstructured data, real-time processing, data volume/variety are part of the evolution. Platforms have changed as well. "Schemaless," real-time events, "schema-on-read," and extract/load/discover/transform (ELDT) are now part of our vernacular. Platforms such as Apache Hadoop, NoSQL databases, and Apache Kafka have emerged to power data lakes.

If platforms have changed, then business intelligence (BI) tools should change as well. Many businesses, however, still try to use traditional, data warehouse-centric BI tools on their data lakes. Inefficient processes remain, including data movement to BI-specific servers, heavy data modeling, and slow feedback loops between IT and business analysts. These are the same time-to-insight challenges that exist when using a data warehouse. This architecture is no better than that of your existing data warehouse. Never mind that when you use traditional BI tools with a data lake, you compromise on speed, scale, and data granularity. Traditional BI tools limit the value of the data lake.

Instead of dooming your data lake, you should embrace a separate BI standard for your data lake. Keep your existing BI tools for your remaining data warehouse workloads, but use a separate BI standard for your data lake. Choose "native" BI tools that were architected for data lakes. These tools are deemed "native" because they run within the data lake cluster. In other words, the query engine is distributed and runs on each data node. Analytics are performed where the data lives, and data movement to a separate BI-specific cluster is eliminated. In addition, the benefits of a native BI platform include unified security, seamless semantic modeling, and query processing optimizations that result in reduced administration, improved self-service, and high speed/scale.

If faster time-to-insight on large and diverse data sets is what you hope to achieve from your data lake, then using the right tools will be a key factor in achieving that objective. Arcadia Enterprise, the flagship product from Arcadia Data, is a native BI platform for data lakes. It reduces the time for the analytical lifecycle by keeping data in the data lake. This means you can do data discovery and semantic modeling in short feedback loops without any IT intervention. This essentially reverses the order of tasks in a data warehouse lifecycle, in which discovery is done well after significant data modeling is completed by the IT team, first on the data lake for semantics, then again on the dedicated BI server for performance tuning. That flow leads to long delays, while Arcadia Enterprise puts discovery and semantic modeling in adjacent tasks to greatly reduce the time to build visualizations. And with Smart Acceleration™, Arcadia Enterprise can make recommendations on how to optimize queries with pre-computed structures known as "Analytical Views." The combination of Analytical Views and Smart Acceleration ensures high end-user concurrency without the delays of performance modeling that are typical of OLAP cube-oriented environments.

EXPLORE NATIVE BI FOR DATA LAKES TODAY

Are you ready to explore how your analysts can get real value quickly from your data lake?

Visit www.arcadiadata.com, email hello@arcadiadata.com or call +1 (415) 680-3535.

ARCADIA DATA
www.arcadiadata.com

Beachbody 'Pumps Up' Data Architecture to Help Clients Live Healthier Lifestyles

Currently there are over two billion people globally who are either overweight or obese, which translates into about one-third of the world's population. The market for health and fitness products and services is therefore understandably huge, totaling over $80 billion per year. Beachbody, a worldwide leader in health and Fitness, was founded in 1998 with the mission to help people achieve their goals and enjoy healthy, fulfilling lives. Beachbody's formula of world-class fitness, nutrition, motivation and support has proven again and again to deliver results for its 23 million customers.

KEEPING PACE WITH DATA GROWTH

Beachbody averages more than five million monthly unique visits across its digital platforms. As the fitness market moved online, Beachbody started streaming its fitness videos in 2015 via Beachbody On Demand (BOD), featuring more than 500 streaming workouts. Beachbody wanted to be able to answer such questions as "How can we better target and retarget customers leveraging master customer data?", "How can we personalize offers for each customer?", and "What indicators determine if a subscriber is about to cancel?"

"Those bigger questions span multiple data sources or larger volumes of data," says Eric Anderson, Executive Director, Data for Beachbody. "We're moving away from questions that can be answered with 10 million rows to ones that rely on billions of rows. The database technology we were using simply couldn't keep up with the pace of the business and the data growth that accompanied it."

The problem was, it took months to acquire a new data source that could provide business visibility into new product offerings. "We were running a conventional on-premises Oracle data warehouse to store our corporate information," says Anderson. "But it was too slow and wouldn't scale quickly or cost effectively."

Anderson says the decision to use a cloud architecture was easy to make. "Creating on-premises the type of data lake we wanted would have been not only very difficult, expensive, and time-consuming, but would have required us to hire a lot more people with the right big data skills."

After evaluating several vendors of integration and big data solutions, Beachbody selected Talend.

Beachbody's IT team, Hortonworks, and Talend worked together to upgrade the company's analytics architecture by adding a scale-out Hadoop cluster in combination with a cloud data lake on AWS S3 so Beachbody could move closer to real-time access to essential information.

WHY TALEND?

Beachbody selected Talend Real-time Big Data for a range of reasons, including its ability to work across heterogeneous environments, built-in connectors to AWS, native Spark processing, and broad out-of-the-box connectivity to traditional on-premises data sources. "We knew we needed the flexibility of a big data integration tool to ingest and analyze our data, and Talend met all our criteria," says Anderson.

In order to get to insights more quickly, Beachbody used Talend Real-time Big Data to accelerate its ability to load massive amounts of data into its cloud data lake. With Talend and AWS, Beachbody easily ingested dozens of critical data sources into the cloud in less than six months. The company originally estimated that this type of project would take over a year if done on-premises.

Beachbody also used Talend Data Preparation to deliver self-service data access. "One of our guiding principles is open systems," says Anderson, "and a big part of that is self-service. Our IT team can't scale infinitely, so we knew we needed to provide self-service access. We now do that by sourcing information at any scale and any frequency, and making that available to our users."

STRENGTHENING COACH AND CUSTOMER ENGAGEMENT

Since creating its cloud data lake, Beachbody enjoys near-real-time data access to essential information about its customers, coaches and more. This enables the company to make faster decisions based on more-comprehensive data.

The data lake can be accessed by line-of-business users via self-service analytics tools using Beachbody's Bring Your Own Tool (BYOT) approach. The tools enable users to analyze website activity, logs from Beachbody On Demand, call-center records, and external data on customer acquisition and spending, as well as sales and financial transaction data.

Asked to summarize the principal benefits of the Talend/AWS solution, Anderson says, "The ones that really stand out are faster time-to-market, increased effectiveness for our digital marketing campaigns, and decreased customer churn. All of those can have a major impact on our continuing success."

TALEND www.talend.com
BEACHBODY www.beachbody.com

Progress®

Connected Enterprise Data Lakes

This article discusses best practices for the data access layer connecting data lakes with enterprise data sources in various data integration patterns that include data movement and on-demand access through standard SQL and REST interfaces.

ENTERPRISE DATA SOURCES

Large organizations building data lakes access a variety of enterprise applications to provide business context to big data sets, such as those collected from sensors in industrial IoT. These data lakes represent data collection that intersects both types of bimodal IT strategies. Enterprise applications often store data in databases such as Oracle, SQL Server, Teradata or IBM DB2. Many of these databases also have popular counterparts hosted in the cloud, such as Oracle Autonomous Data Warehouse Cloud (ADWC), Microsoft Azure SQL or IBM Db2 Warehouse on Cloud. There are several other databases popular in the enterprise, but the key to relating big data sets with business value is a secure hybrid data access pipeline to load data and share insights with key business stakeholders for the data lake initiative, as well as other groups that can benefit from the data analytics.

LOAD DATA

There are a number of open source projects to load enterprise data lakes, such as Apache Sqoop, Apache NiFi and others. There are also commercial data integration solutions from trusted enterprise vendors such as Oracle, IBM, Informatica and others to load these data sources.

The underlying data access layer for any data load into big data platforms is primarily based on JDBC running on Linux platforms. This introduces challenges in hybrid data architectures where data sources may reside across firewall boundaries or only be accessible from web APIs; or the data sources may not be compatible with supported security protocols such as Windows Authentication from Linux.

When exploring open source technologies to load data into your lake, performance can become critical as the project proves to have business value. The best practice is to also evaluate a commercial JDBC data access layer engineered for moving large data sets with enterprise security, such as those from Progress DataDirect. Five of the top six commercial vendors in the Gartner Magic Quadrant for data integration tools embed this technology.

SHARE INSIGHTS

The next step in establishing ROI from a big data analytics project is to democratize the insights produced to other lines of business. According to a recent survey of the Apache Sqoop Developer Community, 65% of respondents are exporting data to external systems from the Hadoop ecosystem. Very large enterprises may have analytics platforms running on Microsoft, Oracle, SAP and IBM among others, and exporting large data sets from Hadoop into these warehouses and marts can place stress on service level agreements for acceptable load times to support business operations.

It is recommended to discuss data movement requirements with various stakeholders for your big data analytics project to understand export targets. If you are working with exporting large data sets, one suggestion is to evaluate bulk enabled JDBC drivers from Progress DataDirect to meet performance requirements with existing Apache Sqoop jobs. On the other hand, if you're looking to expose analytical data sets on-demand for hybrid data access, it is recommended to consider a standard such as OData, which is a supported interface for analytical tools such as Tableau, Power BI, Qlik, Spotfire and others; or even cloud applications such as Salesforce or Microsoft Dynamics.

SECURITY IN DATA ACCESS

Integrating enterprise data sources can be challenging and often subject to regulations, such as the collection of personal data with GDPR. From a technology perspective, we're seeing more stringent policies in global organizations to comply with this regulation, making it imperative to authenticate and secure data in transit using existing security policies rather than make exceptions that introduce significant risk.

Each database has its own security strategies, and if we just look at authentication, for example, SQL Server supports Windows Authentication, which presents challenges from big data environments running across worker nodes in a given cluster on Linux.

Hybrid data architectures introduce additional complexity with an increasing number of firewall boundaries and data being exposed through API strategies that limit the surface area of details that are key to big data analytics. The same Apache Sqoop developer survey showed 37% of organizations run jobs to import/export data in the cloud.

The recommendation for the data access layer is to fully evaluate open source solutions to ensure they support security required to run in production environments, or consider a secure commercial solution from Progress DataDirect. These solutions include a security vulnerability response policy where high risk vulnerabilities (CVSS 8+ or industry equivalent) will be patched within 30 days.

PROGRESS
www.progress.com/jdbc

CAMBRIDGE SEMANTICS
THE SMART DATA COMPANY®

Deliver Data Lake Management and Analytics On Demand

Enterprise Semantic Layer is the Key to Successful Data Lakes

The data-driven business must treat data as an asset, not as a business inhibitor. Leading companies seek insights across their enterprise data, finding new ways to grow revenue. Competitive companies execute on business commitments with speed, flexibility, and lower cost in the face of changing requirements and dynamic data environments. Standing in their way is the complexity, size, and diversity of data—leading to long wait times for insights and one-off manual solutions for the needs of the day. Forward-looking organizations are investing in a Semantic Layer for their enterprise—a Rosetta Stone to make all data accessible and understandable by the business.

Unfortunately, most attempts to deliver the benefits of a Semantic Layer just make the problem worse. Relational technologies, including data warehouses, have proven to lack the required flexibility and scalability—resulting in more data silos. Data Lakes, based on Hadoop or Cloud storage, have therefore proliferated into data swamps—without the required management and governance capabilities.

Until now, no technology has been able to deliver a Semantic Layer at enterprise scale—with security, governance, and accessibility. Finally, there is a production technology with the breadth of functionality and performance to make it not only possible but practical—for organizations to truly treat enterprise data as an asset.

Anzo Smart Data Lake® 4.0 is the first end-to-end platform, from data ingestion to dashboard analytics, for delivering a true Enterprise Information Fabric, a Semantic Layer at enterprise scale. As the only big data management and exploratory analytics platform based on knowledge graphs available today, ASDL provides a dramatically superior approach to enterprise data management, exploration, and analytics. This single platform empowers IT departments and business users alike to flexibly manage, explore, and analyze all of their enterprise data assets with speed of thought performance, at unprecedented big-data scale, and at a fraction of the implementation time and costs compared with any other technology or approach.

The Enterprise Knowledge Graph at the core of ASDL is the most pure and future-proof data representation possible. The nodes and edges of the graph flexibly capture a high-resolution twin of every data source—structured or unstructured. The graph can help users answer any question quickly and interactively, allowing users to converse with the data to uncover insights. In addition to making everyday big data analytics problems easy, the graph unlocks new possibilities where graph is particularly well-suited. The graph, based on open standards is a platform for continuous improvement. Within the graph, sources are quickly linked and harmonized using business rules, text analytics, and even machine learning

Once sources are linked, the data flows into AnzoGraph, a massively parallel distributed native graph database built to query and interactively analyze trillions of relationships. The database runs 111x faster than its nearest competitor. Most graph databases are designed for, and excel at, "point" queries and insertions, such as finding data on specific entities. That is a Graph Online Transaction Processing (GOLTP) use case. AnzoGraph is designed for interactive analysis of broad swaths of data, accumulated over weeks or years of transactions, possibly from many disparate GOLTP and other database sources. In other words, GOLTP is designed for finding patterns, trends, anomalies, and other insights and discovery of big data.

ABOUT CAMBRIDGE SEMANTICS

Cambridge Semantics Inc., The Smart Data Company®, is a big data management and enterprise analytics software company that offers a universal semantic layer to connect and bring meaning to all enterprise data. Its software, the Anzo Smart Data Lake®, allows IT departments and their business users to semantically link, analyze and manage diverse data—whether internal or external, structured or unstructured, with speed, at big data scales, and at a fraction of the implementation costs of using traditional approaches.

CAMBRIDGE SEMANTICS

www.cambridgesemantics.com

Turn Data Into Intelligence in Record Time With Database Replication

Many organizations today rely on data analytics to glean critical business intelligence and gain a competitive edge. But quality data analytics requires more than a consolidation of data in a central location; it also requires up-to-the-minute data. Snapshots simply don't cut it—a snapshot of the data at any point in time becomes out of date as soon as it is taken, and refreshing snapshots is time-consuming and resource-intensive.

For the highest quality data analytics, you need a data replication solution that will remove the problem posed by the continual need for up-to-the-minute data. SharePlex® is a logical data replication solution for SQL Server, Oracle and various databases. It provides one or more near real-time copies of production data without impacting the primary database's performance and availability. The data is replicated from the source database to one or more targets. The target database can be Oracle, SQL Server, Hadoop, Postgres, or Azure to name a few.

By providing a near-real time copy of your Oracle or SQL Server databases, SharePlex enables you to offload reporting from your production databases, eliminating the impact reporting or any other extraction process can have on database performance. Moreover, SharePlex can receive data from transactions on one or multiple databases, so you can consolidate all or subsets of your data to a data warehouse or data store for analytics, reporting or dashboarding.

SharePlex runs continuously on the database server itself, replicating data in real time with sub-second latency and no perceivable impact on source database performance. SharePlex facilitates better business intelligence and analytics by ensuring that only the most recent and most relevant data is analyzed, since replication can be limited to specific schemas, tables, or columns or rows within tables.

Here are four simple ways to get from data to intelligence faster:

1. HETEROGENEOUS DATA REPLICATION

You can easily replicate data from Oracle and SQL Server in near real-time to a single centralized target or distribute the data to multiple database targets. To reduce overhead and speed replication, choose a solution that runs continuously, but only captures and transfers changed data.

2. DATA DISTRIBUTION AND CONSOLIDATION FOR NEAR REAL-TIME DATA WAREHOUSING

One of the biggest database performance drains occurs when end users query, report or extract data from production transactional databases. You should use a data replication solution to offload reporting from production machines and consolidate data from various Oracle or SQL Server databases to a data warehouse. This not only improves performance and reporting, but it also helps reduce costs by enabling you to use a more affordable data storage option.

3. DATA INTEGRATION

A good data integration solution will let you easily combine all or a subset of your Oracle and SQL Server data with other structured or unstructured targets to allow reporting from an integrated data set. Think SQL Server, Hadoop etc.

4. ARCHIVING AND PURGING OF DATA

You should be able to easily set your own criteria for archiving policies. As you replicate data to the secondary server, consider archiving and purging data as needed by the business. That way you are only analyzing current and relevant data.

Data analytics is the key to success and growth for many organizations today, and SharePlex provides a powerful platform to transform your data into insights and intelligence faster. SharePlex enables you to easily maintain one or more near real-time copies of your production data in for all your analytics and data warehousing needs.

Look for a comprehensive, easy-to-install, easy-to-use, impact-free toolset that enables near real-time replication, simplifies data access and improves database performance. SharePlex is that toolset. Download your free 30-day trial today and see for yourself.

QUEST
www.shareplex.com