The Modern Data Lake

OPERATIONALIZING BIG DATA ANALYTICS FOR EVERYONE

CONTENTS

OVERVIEW

Companies have embraced the concept of the data lake or data hub to serve their data storage and data-driven application needs. However, gaps remain in the maturity and capability of the Hadoop stack, leaving organizations struggling with how to reap the benefits of these data lakes and how to create analytic applications that deliver value to end users.

For data lakes to succeed, organizations need to learn and understand the differences between these big data scenarios:

I. Data discovery and exploratory analysis

II. Analytic applications and operationalization analytics across the enterprise

This white paper examines these two scenarios in detail, where and when each one is appropriate, and how to step from one to the other. It also covers examples where enterprise-grade interfaces, self- service data discovery, application development, and business intelligence or analytics capabilities are required to reach the unique objectives of each.

WHAT IS DATA LAKE?

The data lake concept centers on landing all analyzable data sets of any kind in raw or only lightly processed form into the easily expandable scale-out Hadoop infrastructure to ensure that the fidelity of the data is preserved. Instead of forcing data into a static schema and running an ETL (Extract, Transform, Load) process to fit it into a structured database, a Hadoop-first approach enhances agility by storing data at its raw form. As a result, data is available at a more granular level without losing its details, and schemas are created at a later point. This process is also referred to as 'schema-on-read.'

The data going into a lake might consist of machine-generated logs and sensor data (e.g., Internet of Things or IoT), customer behavior (e.g., web clickstreams), social media, documents (e.g., e-mails), geo-location trails, images, video and audio, and structured enterprise data sets such as transactional data from relational sources and systems such as ERP, CRM or SCM.

The economics of Hadoop versus a traditional data warehouse has positioned data lakes as less grandiose data stores which function as feeder systems to other data warehouses, analytic dashboards, or operational applications. Some treat them as initial landing zones and use them to figure out what data should be processed and sent downstream. However, the data stored in data lakes is at a micro-granular level, and not ready for business users or downstream applications.

Another reason for data lakes' rudimentary use is their lack of enterprise-grade features required for broad and mission-critical usage. This includes lack of security, multi-tenancy, SLAs, and data governance capabilities that are core parts of existing data warehouses today. Therefore, while data lakes provide an economical and fast way to do detailed data discovery, it is critical to consider the longer term architectural journey on Hadoop as an analytical repository.

REMEMBER THAT 1ST GENERATION DATA LAKES ARE EXPLORATORY?

Data lakes are created to store historical and micro-transactional data – what in the past was not sustainable in data warehouses due to volumes, complexity, storage costs, latency, or granularity requirements. This level of detail in data offers rich insights, but deducting meaning from it is prone to error and misinterpretation.

For example, Hadoop can be used to store customer interactions with an application or website. While the data that represents the interactive nature of the customer experience has record-by-record details by capturing each click, it might be missing customer demographics, identification and prior activity. In this case, other data management tools are needed to add schemes around the most important elements of this data. For example, mapping web cookies to

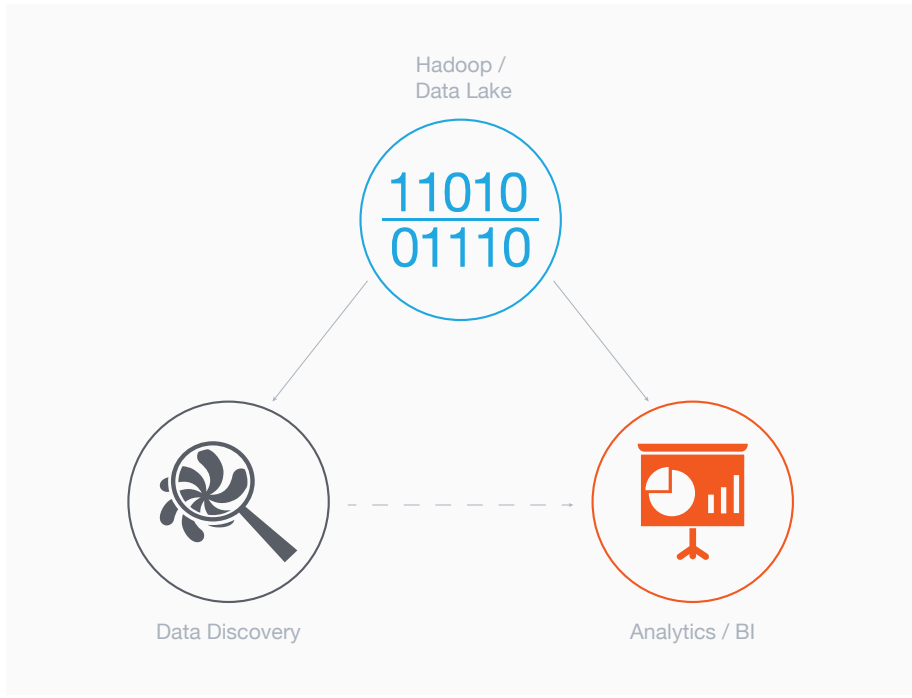
customer IDs provides additional dimensions about the visitor such as their age, location and prior purchases. In the same scenario, enriching the IP address of the visitor clickstream data can reveal geo-location and further segmentation of data.

Discovering patterns and analyzing data in the data lake leads to insights, but also to further questions. Data discovery is a process for extrapolating what data, level of detail and insights should be presented in customer-facing or business applications, and what other pieces of information are needed to enrich the data for a more complete picture.

In general, data lakes often lead to one of the following scenarios:

- One-off analysis and hypothesis validation– typically against a single data set
- Creating analytics applications and operationalizing across the enterprise

In the sections below we will explore each one of these scenarios.



1) One-off analysis, hypothesis validation and data discovery

Take the use case of connected devices. Logs collected from Internet of Things (IoT) devices typically show device failure, reasons for failure, time, intervals, frequency, and sequence of events that led to device failure. Engineers and product managers in manufacturing are keen to learn about the operating characteristics of their devices to diagnose issues and enhance their products in order to prevent future outages.

In this category, the analysis is typically one-off and has a discovery nature. The insights are consumed by a selected few and work as an input or recommendation into future development or other processes.

Spotting certain patterns, anomalies and trends in data offers data analysts a set of cues about what is at risk, what is the recommended set of actions to prevent future failure, and what possible levers they can pull to change certain outcomes.

A pure data discovery use case is less concerned with shortcomings of a data lake – such as consistency of analysis across different individuals, always-on availability, data privacy, protection and security, backup, recovery, and performance. The usage is typically one-off and the users are generally a few trusted analysts.

2) Creating analytic applications and operationalizing across the enterprise

Making data lakes work in analytic scenarios requires more than just directing a data discovery tool on top of a Hadoop cluster. Experienced IT professionals object to creating a single point of failure for an entire organization. Putting all data in one place and relying on it as a golden master, without data protection, backup, recovery, and proper governance introduces risk and liability.

Corporate data needs access, audit and authorization processes. Additionally, Hadoop does not have enough performance for large numbers of users doing concurrent interactive queries.

Integration and metadata creation are also crucial for this type of data lake implementation. Fully integrating the data lake with the rest of the enterprise data adds context, breadth of insights, and relevance. While data discovery can be the first step in understanding which data types matter to the enterprise and whether the data in its raw form is correct and consistent or has gaps, additional work needs to be done to make data business-ready.

Take the use case of analyzing website visits and customer buying journeys. For every visitor, the number of visits may be updated hourly, daily, monthly, or even by the minute. While the spacing of events and identifying period of inactivity might be useful for an analyst, it is not that valuable for a marketing manager, who wants to tie in hourly sales to web site activity.

In this case, other data management tools are needed to create a scheme around the data. For example, defining the start and end of a session, bucketing website visits to hourly intervals, eliminating activities by internet bots, and calculating a rolling hourly time each visitor spends on the website. These schemes create an analytic view into the data that arms marketing managers with enough perspective to take actions.

Additionally, when more than a few users are accessing the analysis - like a group of marketers - consistency across analysis is important. Data Discovery tools create data silos as users can make unrestricted changes to the data, definitions, and results. Creating governance around data helps everyone trust the data, so they can focus on decision-making instead of data debates.

To turn a data lake into an analytic application that serves an entire organization or customer base, you must follow these four principles:

Data cataloging and metadata management: To make data business-ready, you need to create a catalog or inventory of all your data, so business users can search data in simple business terms such as ‘revenue’ or ‘conversion rate.’ With high volumes of new data added every day, automating this process is critical.

Governance and multi-tenancy: Authorizing and granting access to subsets of data requires security and data governance. Delineating who can see which data and at what granularity level requires multi-tenancy features – something that data lakes don’t offer today. Without these capabilities, data is only at the fingertips of few data scientists instead of the broader organization and business users.

Operations: For a data lake to become a key operational business platform, it needs high availability, backup, and constant recovery.

Self-service and end-user data mashups: To infuse a data lake with value, you must offer a rich and intuitive user interface, and encourage information sharing and self-service. In many cases, business users want to blend their own data with the data from the data lake. Providing search-based analytics, guided navigation, and end-user data mashups is essential for making data lakes accessible.

The non-analytic applications of data lakes

It is worth mentioning another breed of applications built on data lakes – transactional applications – which are applications that your business or product run on. These applications often require real-time or streaming data and are mission-critical by nature. For example, heartrate monitors, stock-exchange apps, fitness and wearable devices, and other applications in this category monitor certain signals in the data, detect motions, gestures, and changes in order to infer engagement levels or send alerts upon a particular event. Micro-transaction data from these applications or devices is often stored in data lakes.

When designing these applications, discovery tools can be used to test data dependency scenarios. For example, based on the data variance of a certain data type, an app can set a threshold to notify and inform users about a particular change.

Another set of applications are those that facilitate a business process, such as a supply chain or purchasing process. These applications offer work flow, but they often embed analytics and visualizations in their user interface to better inform and guide business users.

COMPARISON OF DATA LAKE DEPLOYMENTS FOR DATA DISCOVERY VS. BUSINESS ANALYTICS SCENARIOS

While exploratory and discovery analysis may be a good way to get started quickly as you learn about the relevance of your key data relationships, you also need to invest in the infrastructure and architecture of your analytics in order to make the entire analytic journey into production scenarios that offer value now and in the future.

The table below offers a clear delineation between data discovery on Hadoop vs. Analytics on Hadoop.

	DATA DISCOVERY	OPERATIONALIZING ANALYTICS
Use	Discovery Exploration E.g. what are all possible device failure patterns? Did we have inventory shortage on <date>?	Mission critical apps Repetition E.g. which devices are failing & why? What are my inventory levels?
Users	Data Scientists / Analysts 5-10% of the population	Business Users 90-95% of the population
Outcome / Action	Hypothesis analysis Recommendations Pattern recognition	Business decisions Knowledge sharing Operations
Data is in...	Raw bits	Business terms
Access Time	Immediately (given skills)	1-2 weeks to define your data model
Shelf Time	Weeks-months	Years
Risks vs. Rewards	Quick exploration but un-governed, leaves room for misinterpretation by non-data people	Governance, trust, but requires defining metadata on data

When should you leverage data discovery on Hadoop?

- You don't know the nature of your data and this is your first opportunity to examine it
- You want to recognize the gaps and quality issues in your data
- Your users are a few data scientists and you don't have business users that want access to the data
- The shelf-life of your analysis is short lived

When should you think beyond data discovery on Hadoop?

- Business users need information and insights from your data lake
- Data at its raw form does not have business context
- You need to blend data from the data lake with other sources
- Insights need to be embedded into other applications or predictive models

CONCLUSION

In many cases, data lakes are created to store historical, micro-transactional event data, but most enterprises must bring to bear the operational intelligence aspects of a data lake. While data discovery tools give you a head start in identifying the gaps in your data or creating one-off analysis, operationalizing big data requires further data management and analytic engines. To maximize the value of data lakes, organizations must think ahead architecturally and balance experimentation and the use of pure data discovery activities with creating enterprise applications that add context, consumability, and availability of data to the entire enterprise.

ABOUT BIRST

Birst is the global leader in Cloud Business Intelligence (BI) and Analytics for the Enterprise. Birst's Networked BI platform redefines the way BI is delivered and consumed, eliminating analytical silos to dramatically improve the speed, alignment and economics of BI across the enterprise. Built on top of Birst's next-generation, multi-tenant cloud architecture, Networked BI enables centralized and decentralized BI applications to be transparently connected via a shared analytical data fabric, delivering local execution with global governance. Today, Birst serves thousands of organizations across the globe by making trusted enterprise business data a part of everyday operational decision making. Learn more at www.birst.com and join the conversation @BirstBI.