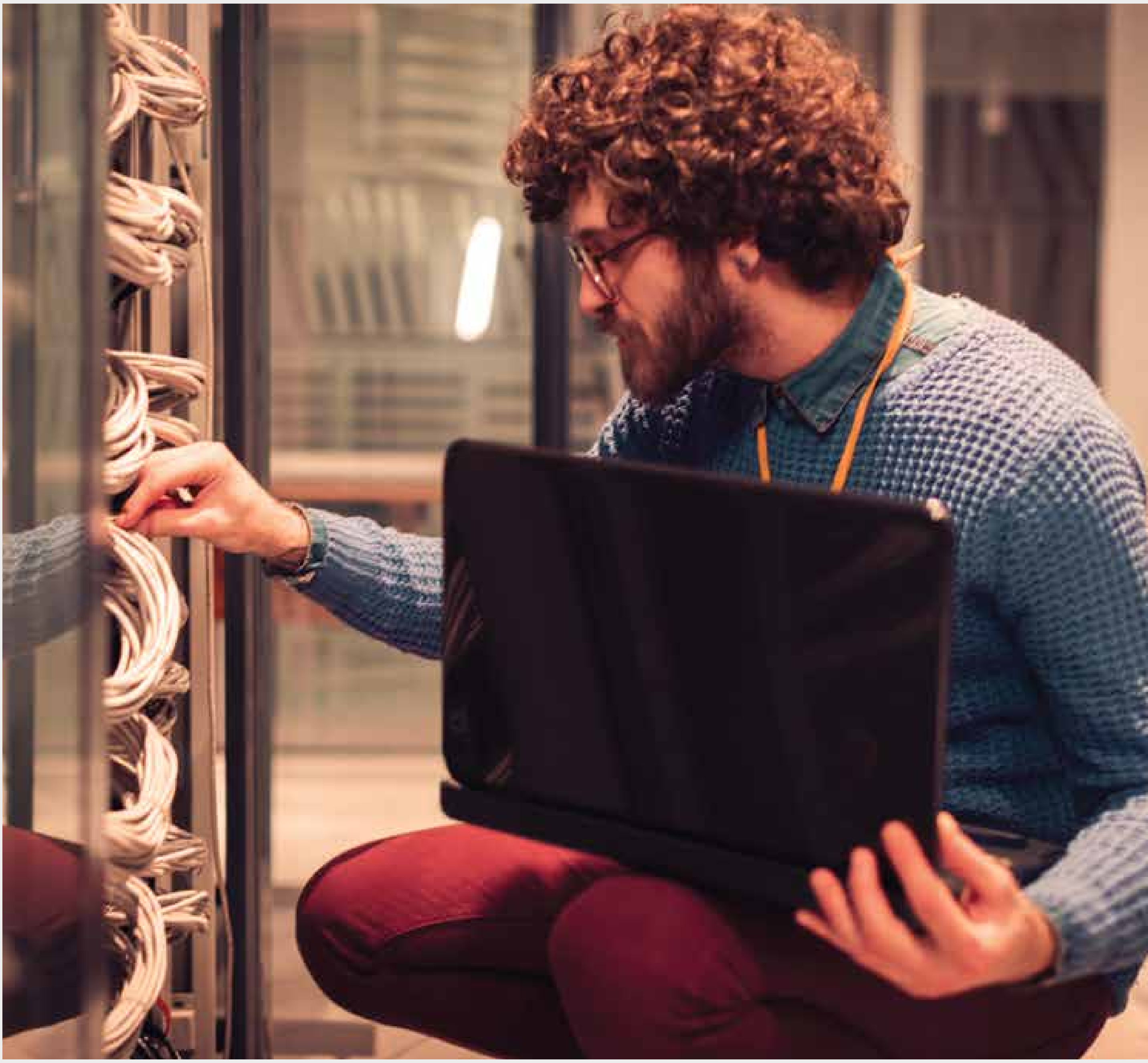
Five best practices to keep your data lake healthy

The volume and variety of data available to organizations in the digital age is enormous, giving rise to the need of expanding their data-management infrastructures considerably and rapidly. Data lake, a new class of data-management system, holds a strong potential to achieve success in this area.

The humongous growth in computer processing power, increase in cloud storage capacity, its usage and advanced network connectivity are transforming the existing data flood into an infinite stream of comprehensive information about customers' preferences and buying habits, product attributes, sales and ecommerce data, email and web analytics, survey and feedback. These datasets are acquired from a range of sources and across all formats, including Internet-of-Things (IoT) devices, social-media sites, sales and CRM systems, etc.

Today, with the increased growth in number of data sources and the pervasive nature of analytics, a huge demand for strong analytics capabilities is surfacing, just to keep pace. To gain maximum value out of data in this extremely competitive environment, organizations are emphasizing not just on using analytics, but are transforming themselves to build an insights-powered enterprise.



If the data in a lake is not well curated, then it may turn into a data swamp, flooding an organization with information.

Growing relevance of data lakes

In industry parlance, data lake refers to a repository of data that brings together variety of data available, such as structured (relational), semi-structured (XML, JSON, etc.) and unstructured (Images, audio and video). It provides a way to store all the raw data together in a single place at a low cost. Distributed file systems such as Hadoop file system (HDFS) often serves this purpose. The adoption of Hadoop as the platform created the first version of data lake. The use of cloud based data lake i.e. the latest version of data lake, gained momentum in the market because of the challenges in storage flexibility, resource management, data protection deep-rooted in Hadoop-based data lakes.

Industry experts anticipate that in the coming years, majority of large organizations will experience more data flowing from data lakes to data warehouses. This prediction shows the change in the mindset of organizations which were typically driven by enterprise data warehouse as the de facto standard for their data need. As the popularity of data lakes grows, organizations face a bigger challenge of maintaining an infinite data lake. If the data in a lake is not well curated, then this data lake may turn into a data swamp, flooding an organization with information that is difficult to locate and consume.

How to keep data lakes relevant

Digital transformation requires identifying authentic and accurate data sources in an organization to truly capitalize on increasing volumes of data and generate new insights that propel growth while maintaining a single version of truth.

Following are some of the ways to keep data lakes dynamic, immaculate and viable:

1. Identify and define organization's data goal: One of the most important preemptive steps in avoiding an organizational data swamp is to set clear boundaries for the type of information organizations are trying to collect, and their intent of what they want to do with it. Amassing a lot of data should not be the sole aim of organizations. They need to have a clarity on what they want to attain from the data they are collecting. An enterprise with a clear data strategy shall reap benefits in terms of avoiding data silos, incorporating a data driven culture to maintain customer centricity, scale up and meet demands of the modern day data environment.

2. Incorporate modern data architecture: The old data architecture models are not sufficient and may not satisfy the needs of today's data - driven businesses in a cost effective manner. These workflows give a primer to ensure modern data architecture:

- Data profiling: Today, organizations are facing an immense challenge in retaining and re-using the huge amount of unstructured data stored by them. Data profiling allows building understanding of current data assets and their condition (data quality and lineage). It helps in scanning, classifying, indexing structured and unstructured data residing in different sources. This provides the organization an information database that can be better managed and yields maximum value in future. It is recommended that profiled data should be loaded into different curated zones (loading, staging, certified and consumption) for ease of understanding and to avoid creating a data swamp.

- Data cataloguing: There is a huge possibility that data collected in one business section may provide value to other business sections in other scenarios. However, business users may not always have the requisite visibility or awareness about which data exists and its corresponding ownership. As a result, data is seldom used beyond its context, and plenty of opportunities to extract maximum value from data are lost. One approach to resolve this drawback is to make a data catalog for the organization. A data catalog will help in managing and maintaining the existing datasets and KPI glossary to foster easy search of all the available data within the organization for users to seek maximum benefit out of it.

- Data backup and data archival: Data backup protects organization's active and inactive data on the cloud whereas, data archive solutions are aimed for continuous data possession at minimal cost for longer periods. All cloud vendors have tools and technologies to achieve a robust archival process for the organization.

3. Build sound data governance, privacy and security: Data governance and metadata management is a critical step to keep a healthy and effective data lake strategy. A well-curated data lake contains data that's clean, easily accessible, trusted and secure. As a result, this high-quality data can be easily consumed with confidence by the business users. It is of utmost importance to establish responsibility for data.

4. Leverage automation and AI: Due to the variety and velocity of data coming into data lake, it is important to automate the data acquisition and transformation processes. Organizations can leverage next generation data integration and enterprise data warehousing (EDW) tools along with artificial intelligence (AI) and machine learning that can help them classify, analyze and learn from the data at a high-speed with better accuracy.

5. Integrate dev/ops: DevOps processes will go hand on hand with building and maintaining a healthy data lake. Ensure clear guidelines are established on where and how data is collected to prevent "data wildness," and make sure those standards are always followed. Take time to evaluate sources as "trustworthy," and take preventive steps to ensure it stays that way. A little work on the front end will prove highly valuable when it comes to putting data to use.

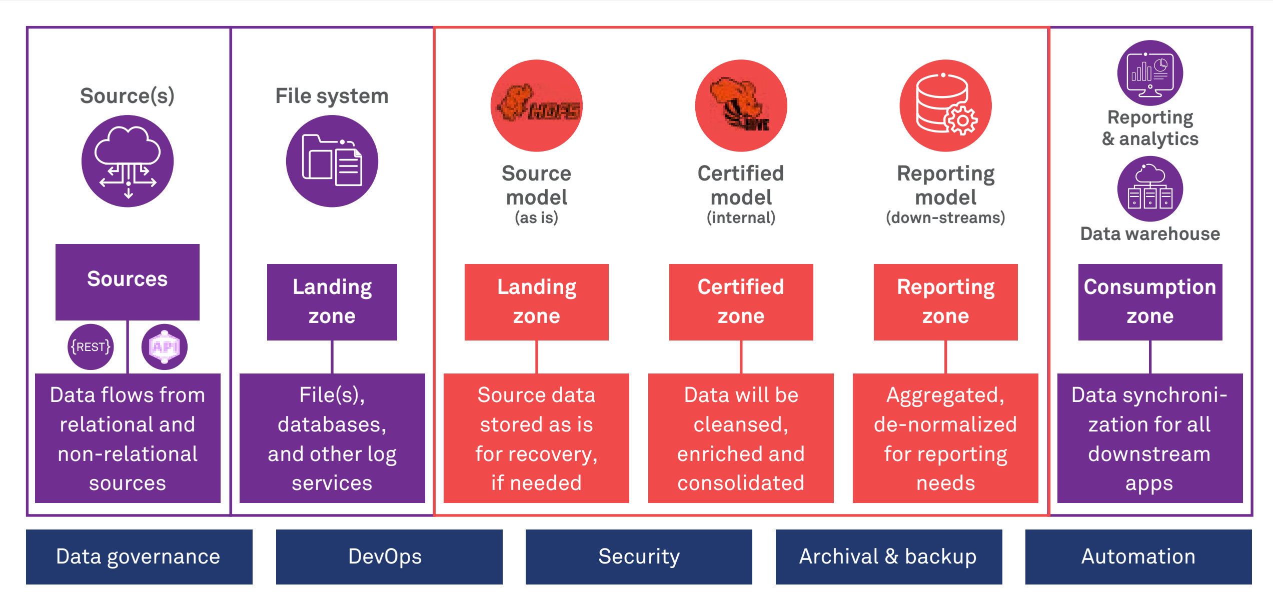

Source(s)
File system
Source model
(as is)
Certified model
(internal)
Reporting model
(down-streams)
Reporting & analytics
Data warehouse
Sources
{REST}
API
Landing zone
Landing zone
Certified zone
Reporting zone
Consumption zone
Data flows from relational and non-relational sources
File(s), databases, and other log services
Source data stored as is for recovery, if needed
Data will be cleansed, enriched and consolidated
Aggregated, de-normalized for reporting needs
Data synchronization for all downstream apps
Data governance
DevOps
Security
Archival & backup
Automation

Figure 1: Reference architecture to avoid data lake failure

The reference data architecture, given in Figure 1, represents an efficient process that includes data zones starting from the data foundation layer to providing insights in conjunction with the essential measures viz. data governance, security, data archival and backup, automation and DevOps to keep a data lake pristine and clean. A large medical device manufacturing company, struggling to garner value from its data, leveraged a data lake solution based on this architecture. Over a short span of a year, the new data lake became a potent digital transformation tool for the company and is being used across the organization for building more than 40 analytics use cases.

Conclusion

The use of big data has no boundaries. It can help businesses create new growth opportunities, outperform existing competitors and provide seamless customer experience. However, to get the best out of data and thrive in this digital world, enterprises should possess well-curated, good quality data lakes. A clear data strategy, modern data architecture, proper data governance combined with automation, security and privacy measures, and DevOps integration will not only prevent any data lake from becoming a data swamp but also transform it from simply being a data repository to a dynamic tool that will empower digital transformation across an enterprise.

About the authors

Neeraj Verma

Managing Consultant,
Data, Analytics & AI,
Wipro Limited.

Neeraj brings over 13 years of experience in envisioning and building large scale enterprise data warehouse solutions for fortune 100 Retail, CPG and Pharma clients across the globe. He has deep domain and analytics expertise in CPG and Retail.

Ajinkya Pawar

Consultant -Strategy & Planning,
Data, Analytics & AI,
Wipro Limited.

Ajinkya with a deep understanding of the digital and analytics industry, including market insights, competitive landscape and related ecosystems, provides advisory services and thought leadership for building value propositions and driving cohesive strategies.

Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035,
India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 175,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information, please write to us at **info@wipro.com**