- Information systems—management

Focus

What it Takes to Manage Data in the Digital Era

Effective data management helps to streamline business operations, enhance decision-making, drive innovation, and improve business efficiency and competitiveness. There are a number of open source tools that can be used to manage data efficiently, ensuring its security and privacy.



Today, enterprises are swamped with information from customer interactions, online transactions, blog posts, tweets, social media and sensor data. They must manage this data efficiently in order to make informed decisions, enhance customer experiences and drive innovation.

Data management is the process of acquiring, organising, storing and utilising data throughout the data life cycle in an enterprise. The goal is to ensure that the data is accurate, accessible, secure and aligned with business objectives. It involves a set of practices, processes, policies and technologies that govern data collection, storage, processing and sharing.

The growth of data management

- According to Statista, by 2027 the global market for Big Data will be worth US$ 103 billion.
- Global data creation is expected to reach 175 zettabytes by 2025, as per IDC.
- Gartner earlier predicted that the organisations that promote data sharing will outperform their peers on most business value metrics.
- A survey by Deloitte found that 67% of executives consider data management and analytics to be critical for their business's success.

Challenges in data management

Enterprises face quite a few challenges in managing data effectively and efficiently.

- Data is in silos, segmented across various platforms, channels, tools, and business units, making it challenging to access it. This leads to inefficiency and data inconsistency.
- Data accuracy, completeness, and timeliness is an issue.

- Bringing offline data online is a challenge.
- Managing regulatory compliance, establishing data security, and data privacy are all challenging tasks.

A poor data management strategy may lead to an enormous amount of data in a completely unmanageable format.

Next-generation data capabilities

To manage the huge and different types of data that enterprises collect and process, they must build the following data capabilities.

- **Data processing techniques**: Unstructured data processing covers entity extraction, concept extraction, sentiment analysis, NLP, ontology, etc. To automate data extraction, machine learning techniques must be leveraged.
- **Agile data delivery**: Agile development methodology can be used to reduce the overall cost of data delivery.
- **Digital 360-degree view of the customer and products**: Customers interact with enterprises through social media channels. The ability to create a single view of a customer across internal and external data is critical in a digital world. This applies to products too.
- **On-demand data services**: On-demand data services should be available for both online and offline use. The ability to query in a federated fashion for both online and offline access is important.
- **Unstructured metadata**: Data is being thrown into enterprise systems at a rate that doesn't allow smaller businesses to look at data structures and extract metadata. Automated metadata extraction based on ontology is critical.
- **Real-time analytics**: Multi-channel applications and decision management systems need to be used to capture interactions for digital processes in real-time scenarios.
- **Data archival**: Compliance and performance requirements drive the need for archival of both structured and unstructured data.

Principles of data architecture

Data architecture principles are a structured set of ideas that collectively define and guide the development of a solution architecture — from values to design and implementation, harmonising decision-making across an organisation. A few principles are listed below.

- **Data is an asset**: Data is an asset that has a specific and measurable value for the enterprise.
- **Data is shared**: Users have access to the data necessary to perform their duties; therefore, data is shared across enterprise functions and organisations.
- **Data trustee**: Each data element has a trustee accountable for data quality.
- **Common vocabulary and data definitions**: Data definition is consistent throughout the enterprise, and the definitions are understandable and available to all users.
- **Data security**: Data is protected from unauthorised use and disclosure.
- **Data privacy**: Data privacy and protection are paramount throughout the life cycle of a Big Data project. All data sharing will conform to the relevant regulatory and business requirements.
- **Data integrity and the transparency of processes**: Each party to a data analytics project must be aware of and abide by its responsibilities regarding the provision of source data, and the obligation to establish and maintain adequate controls over the use of personal or other sensitive data.
- **Fit for purpose tools**: The next-generation information ecosystem needs to have fit for purpose tools, as one technology will not meet the requirements of all the workloads and processing techniques.
- **Data governance:** As data resides around many repositories, integrated metadata lineage and PII (personally identifiable information) protection are important.

The key technology trends that are driving enterprise data management are depicted in Figure 1. These trends help to establish more dynamic, secure and user-friendly data management practices across the enterprise.

Key components of data management

Data collection: This is the process of collecting data from various sources (internal and external) to provide the insights required to solve various business problems. The main data sources today are social media, mobile devices, cloud computing, real-time sensor data, AI/ML, cybersecurity and predictive analytics.

As an example, for a trained AI model to carry out sentiment analysis, the collected data should include a large number of reviews, comments, and social media posts. Web scraping is the automated method of extracting data from various websites.

Notable open source data collecting tools are ODK (Open Data Kit), KoboToolbox, Jotform, and Fulcrum.

Data storage: This helps to determine where and how the data will be stored as an enterprise data repository. It covers both structured and unstructured data sources, which includes databases, data warehouses and data lakes.

Key open source players in this field are Hadoop, LakeFS, Cassandra, and Neo4J. These tools provide scalability, robustness and performance in managing large data and analysing large data sets in various applications.

Data processing: Data is pre-processed and undergoes cleansing, normalisation, and tokenisation to derive insights from it that support business processes.

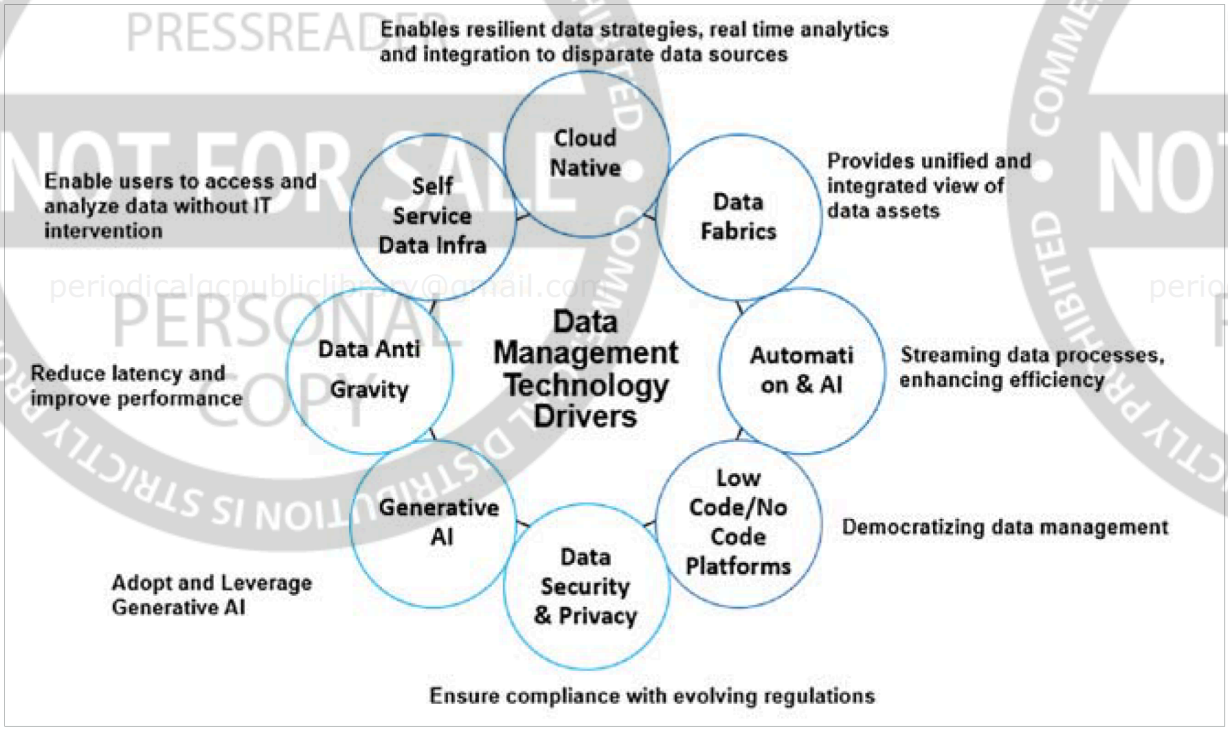


Figure 1: Data management technology drivers

- *Data cleansing* involves the identification and removal of inaccurate, incomplete, or irrelevant data. Duplicate data is removed and incorrect data values are corrected.
- *Data normalisation* helps transform data to a standard format for easy comparison and analysis.
- *Data tokenisation* is a process of breaking the text into individual words or phrases using natural language processing (NLP).

Top open source data processing tools that are widely used include Python, R, Matplotlib, Apache Spark, Apache Hadoop, Apache Kafka and Apache Airflow.

Data quality*:* This helps to ensure the accuracy, completeness, and consistency of data. Data quality management involves identifying and correcting errors, standardising formats, and maintaining a high level of data integrity.

Some of the best open source data quality tools are Cucumber, Deequ, dbt Core, MobyDQ, Great Expectations and Soda Core. These tools help automate the data validation process, as well as data cleaning and monitoring.

Data pipeline*:* This acts like a path to move data from one system to another. ETLs are a popular type of data pipeline. They make it easier for businesses to pull data from multiple sources into a single source. During the process, the data moves through three steps:

- *Extraction:* Pulling data from a database.
- *Transformation:* Manipulating the data with code for formatting and preparing it for analysis.
- *Loading:* Writing the data into the new location.

The leading open source ETL tools include Apache Airflow, Apache Kafka, Data Build Tool, Airbyte, Meltano, Singer and Mage. These tools help to ingest, process, store, transform and visualise data efficiently. They provide scalability, flexibility and active community support.

Data integration*:* At the enterprise level data may reside in various systems, such as databases, data warehouses, cloud storage, Excel sheets, etc, each with its own structure and semantics. Also, the data may be in different formats, including CSV, XML, JSON, or proprietary formats specific to ERP-centric applications (SAP, Salesforce, etc).

Data integration solutions deal with the aggregation of data from diverse sources and different formats to provide a single view. This eliminates data silos, enhancing cross-functional collaboration and decision-making.

Widely used open source data integration tools are Apache NiFi, Apache Kafka, Apache Flume, Fluentd, Sqoop, Apache Camel, Apache Flink, Logstash, Talend Open Studio and StreamSets. These tools can handle various data integration tasks ranging from simple batch jobs to complex streaming data flows.

Focus

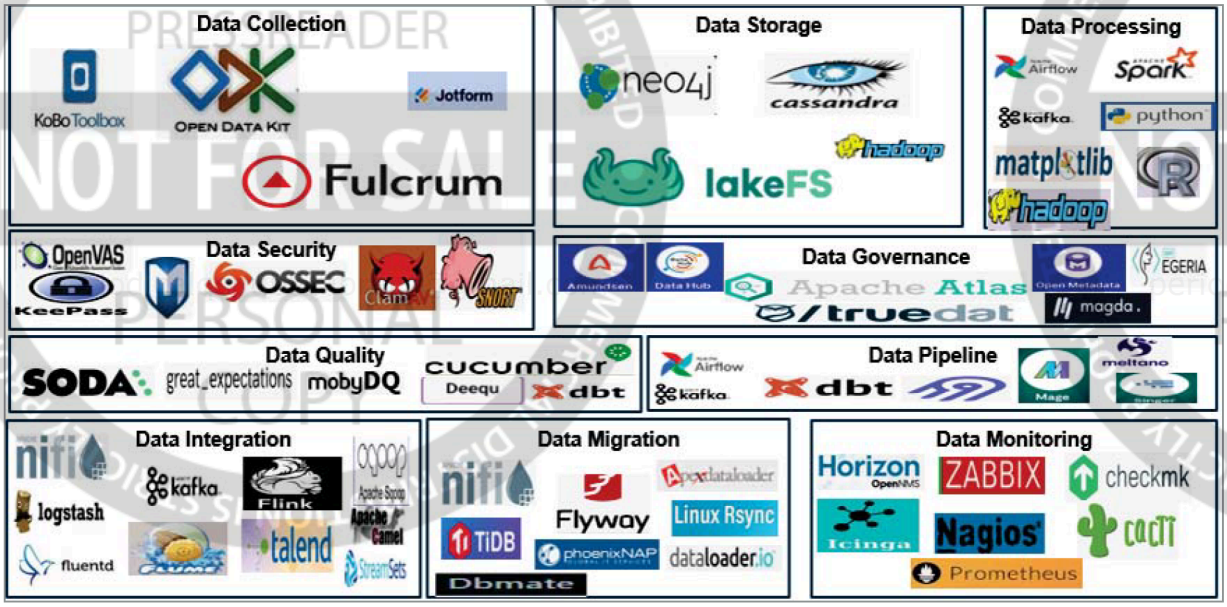


Figure 2: Open source data management ecosystem

Data migration: This is a one-time process of moving data from one database to another. This usually happens because an enterprise is adding a new system or data location. Migration can also mean shifting data formats or applications.

For example, if an enterprise is changing to a new CRM, it needs to migrate data from the current platform to the new one.

Well-known open source based data migration tools are rysync, phinx, Dbmate, Apache NiFi, Flyway and TiDB. These tools help in simple file transfers as well as complex database migrations. They provide scalability, ease of use and active community support.

Data security: Establishing proper security protocols helps to reduce the risk of data breaches. It also safeguards sensitive information. Encryption, access controls, authentication, and intrusion detection systems help protect data across the life cycle.

Open source data security tools that are widely used include Metasploit, OSSEC, OpenVAS, Snort, KeePass, and ClamAV. These tools can be integrated into various security strategies to protect a wide range of cyber threats.

Data governance: This covers automation of policies, guidelines, principles, and standards for managing data assets. It ensures data quality, accuracy, and compliance with regulatory requirements, building trust in the data. Data governance must be aligned at the enterprise level to realise the business objectives.

Some of the open source data governance tools are Amundsen, DataHub, Apache Atlas, Magda, Open MetaData, Egeria, and TrueDat. These tools offer features like metadata management, data cataloguing and collaboration to manage data assets effectively.

Data monitoring: This relates to monitoring the quality and relevance of the data, and updating it to improve the performance of applications.

Some of the best open source data monitoring tools are Prometheus, Zabbix, Nagios Core, Checkmk, Cacti, OpenNMS Horizon, Icinga, etc. These tools provide robustness, scalability and active community support.

Having good ways to manage data is important for businesses today. By implementing data management best practices, enterprises can ensure that they are managing their data well enough to maximise its value. END

Acknowledgements

The author would like to thank Santosh Shinde of BTIS, Enterprise Architecture division of HCL Technologies Ltd, for giving the required time and support to help write this article as part of Architecture Practice efforts.

By: Dr Gopala Krishna Behara

The author is an enterprise architect in the BTIS Enterprise Architecture division of HCL Technologies Ltd. He has a total of 28 years of experience in the IT industry.

Disclaimer: The views expressed in this article are that of the author and HCL does not subscribe to the substance, veracity, or truthfulness of the said opinion.

https://www.pressreader.com/india/open-source-for-you/20240701/page/38 1/1