Integrated Data Platforms: Unlocking Business Intelligence and Driving Innovation

Photo Source: iStock by Getty Images

A major challenge in all sectors is resolving the intricacies involved in consolidating independent data clusters into a modernized unified architecture.

Imagine being able to access data all in one place, anytime, and have the power to fulfill endless business needs.

This should become the goal of many companies in the age of digital transformation. The rise of cloud-based technology, big data, and machine learning tools has made this increasingly possible.

Integrated Data platforms can open a plethora of opportunities for intelligence, innovation, and modernization. With data easily accessible in real-time, businesses can make more informed decisions, predict customer behavior, manage risks, optimize operations, and eventually realize their business vision.

Visualize new data potentials and develop applications designed to strengthen and enhance business operations.

Typically, With the right data management solutions and data analysis tools, businesses can access and analyze all their data from a single platform. This not only improves efficiency and accelerates the decision-making process, but also enables businesses to unlock new insights, identify growth opportunities, and drive innovation to stay ahead of competitors.

The primary aim is to provide a unified and comprehensive view of data to all groups of people within a company such as managers, analysts, marketing teams, and business decision-makers.

Let's look at the key factors that are essential in planning for building an enterprise platform involving the process of defining, integrating, storing, and managing all data related to business operations and transactions.

Define Objectives

Clearly define the goals and objectives of the data platform. It could be to improve decision-making, streamline business operations, increase customer satisfaction, Application management, or predict trends.

Identify Data Sources and Create a Data Dictionary

Record all places from where the company collects data. It could be customer databases, Data documents, web analytics, revenue tracking, social media activity, competitors' performance data, market data, and more. Designate each data set by its function in the overall business process. This data dictionary will serve as a foundation to build the data model and lay out the logic for data relationships.

Prepare Data Model

Design dataset relationships to view communication connections between different data sets. The data model will outline the integrated view of the organization's data. Data modeling provides well- defined technique to organize data to meet business goals, effectively becoming a roadmap to strategize and guide an organization's IT infrastructure.

Data modeling tools are used to design and develop complex database structures and ensure that data objects, relationships, and their rules are accurately represented. Here are some commonly used data modeling tools:

- **ER/Studio**: Embarcadero's ER/Studio provides robust logical and physical modeling and allows for forward and reverse engineering.
- **Sparx Systems Enterprise Architect**: This tool provides well-rounded support for data modeling, including UML, BPMN, SysML, and many other designs.
- **IBM InfoSphere Data Architect**: IBM's tool allows for visual design, management, and analysis of relational database schemas, with a specific focus on IBM DB2.

- **Oracle SQL Developer Data Modeler**: This is a free tool that provides forward and reverse engineering capabilities for Oracle SQL and is also integrated with Oracle's SQL Developer.
- **Toad Data Modeler**: It supports a wide range of databases, has a user-friendly interface, and provides functionalities for both logical and physical modeling.
- **PowerDesigner (SAP)**: It supports model-driven architecture for data modeling, and can also handle metadata management.
- **Microsoft Visio**: While not strictly a database modeling tool, Visio includes templates and shapes for building Entity-Relationship diagrams.
- **CA ERwin Data Modeler**: ERwin enables users to visualize complex database structures, build and maintain databases, and work collaboratively on data models.
- **MySQL Workbench**: This free tool from MySQL includes database design and modeling along with other development features.
- **Navicat Data Modeler**: Navicat supports multiple databases, offers functionalities for both logical and physical modeling, and provides reverse/forward engineering capabilities.

Build Data Integration Framework

Develop a process to automatically gather and ingest data from multiple sources and formats into one consolidated and cohesive view. This can be achieved through ETL (Extract, Transform, and Load) processes. There are many ways to build data ingestion frameworks based on business model, architecture. However, data ingestion itself can be done in two ways, batch or streaming.

Data ingestion tools help in the process of importing, transferring, loading, and processing data for later use or storage in a database. Here are some popular data ingestion tools:

- **Apache NiFi**: It's an open-source tool for automating and managing data flows between different systems. It allows data routing, transformation, and system mediation logic.
- **Fluentd**: This is an open-source data collector that unifies data collection and consumption.
- **Logstash**: A part of the ELK Stack (Elasticsearch, Logstash, Kibana), which is primarily used for log and event data. It provides real-time insights from the ingested data.
- **AWS Kinesis:** This is a cloud-based service from Amazon that's used to process large streams of data in real time.
- **Google Pub/Sub**: This is a scalable event ingestion service provided by Google that allows real-time analytics from ingested events.
- **Apache Flume**: This is a distributed, reliable, and highly available service for efficiently collecting, aggregating, and moving large amounts of log data to a centralized data store.
- **StreamSets**: It's a platform designed for building, executing, operating, and protecting enterprise data flow pipelines.
- **Sqoop**: This is a tool designed to transfer data between Hadoop and relational databases.
- **Informatica PowerCenter**: Well known for its ETL capabilities, PowerCenter also supports data ingestion tasks.

These tools play a crucial role in managing the challenges of volume, velocity, variety, and veracity of big data and the choice of the tool depends on the specific requirement and the technology stack implemented in your organization.

Decide how and where the data will be stored.

It could be on-premises, cloud, or a hybrid solution depending on company size, budget, and business needs. A variety of Digital technologies can be used depending on Data usage from analytical systems to transactional processing. Data storage mechanism used should be robust, scalable, cost effective, performance efficient, reliable with fault tolerance. The right storage solution should be offer capabilities to enable automation, digitalization, apply machine learning models, Dev-Ops compliant.

These are some methods pertaining to Data Storage Platforms:

- **Databases**: This is perhaps one of the most common forms of data storage. Databases like MySQL, Oracle, PostgreSQL, and SQL Server help in storing structured data.
- **Data Warehouses**: Data warehouses such as BigQuery, Snowflake, Amazon Redshift, and Teradata help store data drawn from transactional systems, relational databases, and other sources.
- **Data Lakes:** They store both structured and unstructured data at any scale. Examples include Amazon S3, Microsoft Azure Data Lake Storage, and Google Cloud Storage.
- **Cloud Storage:** Services like Google Cloud Storage, Amazon S3, Microsoft Azure Storage provide a scalable environment to store data.
- **Distributed Storage:** Distributed storage systems like Apache Hadoop, Cassandra, and MongoDB store data across multiple nodes to ensure redundancy, speed, and to lower the risk of data loss.
- **On-Premises Storage**: This includes traditional data storage like hard drives and SSDs.
- **Hybrid Storage:** It uses a combination of on-premises and cloud-based storage to create a balanced infrastructure that optimizes cost, speed, security, and availability.
- **Object Storage:** It offers infinite scalability at a lower cost. Examples include Amazon S3, Google Object Storage Cloud Storage, and Azure Blob Storage.

Each of these platforms are suited to different types of data and use-cases, the right platform for an enterprise depends on their specific needs.

Design Data Transformation rules

Data transformation tools are software or services that convert, clean, and standardize data into a format that can be used for data analytics and reporting. The different types of Data Transformation involves Data cleansing, Data deduplication, application of business rules, master data management, validation, data quality checks.

Here are a few commonly used data transformation tools:

- **'ETL' Tools (Extract, Transform, Load):** These tools, like Informatica PowerCenter, Microsoft SQL Server Integration Services (SSIS), and IBM InfoSphere, help extract data from various sources, transform the data into an appropriate format or structure, and then load it into a final target, commonly a data warehouse.

- **'ELT' Tools (Extract, Load, Transform):** These are similar to ETL but the transformation is done after loading data into the target system. Examples include Google's BigQuery, Amazon's Redshift, and Snowflake.
- **Data Cleaning Tools:** OpenRefine, Google Cloud's Dataprep, and Trifacta Wrangler are examples of this type of tool that help clean up messy data, finding inconsistencies and making the data more usable.
- **Data Pipeline Tools:** Tools like Apache Beam, Fivetran, Stitch, and Airflow allow you to build data pipelines which can extract, transform, and load data in real-time or batch modes.
- **Scripting Languages:** Python, especially with pandas library, and R are often used for data transformation because of their flexibility and the extensive amount of libraries they offer for data manipulation.

Selection of the appropriate tool depends on the specific requirements, like the type and volume of data, the complexity of transformations, the target system, and the required performance.

Build Data security Mechanisms.

Data Security: Deploy methods to ensure the security of data both during transit and while at rest in the database. This includes robust access management system, sensitive data encryption, and frequent security audits. Modern data security tools leverage advanced technologies, like artificial intelligence, machine learning, and automation, to provide robust and proactive security measures for data protection. Some of these tools are:

- **Cloud-native Security Platforms**: Tools like Prisma Cloud by Palo Alto Networks, Google's Chronical, and IBM Cloud Pak for Security provide comprehensive security for multi-cloud and hybrid cloud environments.
- **AI and ML-powered Security Solutions**: Tools like Darktrace, and Cylance use artificial intelligence and machine learning to predict, detect, and respond to threats in real time.
- **Security Orchestration, Automation, and Response (SOAR) Tools**: Solutions like IBM Resilient, Splunk Phantom, and Swimlane enhance the efficiency of security operations by automating tasks and orchestrating responses to incidents.
- **Endpoint Detection and Response (EDR) Solutions**: Tools like CrowdStrike Falcon, SentinelOne, and Carbon Black provide real-time monitoring and protection for endpoint systems from various cyber threats.
- **Zero Trust Network Security Tools:** Solutions like Zscaler, Akamai's Zero Trust, and Cloudflare Access help enforce the zero trust model, which assumes no user or system is trusted by default, whether inside or outside the network.
- **Data Loss Prevention (DLP) Tools**: Modern DLP tools such as Symantec DLP, Forcepoint DLP, and McAfee DLP provide advanced features such as fingerprinting data, machine learning-based analytics and integration with cloud and other IT services.
- **Blockchain-based Data Security**: Blockchain technology can improve data security due to its decentralized, transparent, and immutable characteristics. For example, Guardtime uses blockchain to ensure the integrity of data.
- **Advanced Threat Protection (ATP) Solutions**: Tools like Microsoft ATP and Symantec ATP offer comprehensive, coordinated protection against sophisticated threats across endpoints, networks, and email.

Data security is an ongoing process and requires not only the use of the latest tools but also a commitment to best security practices, regular audits, and continuous staff education.

Data Visualization

Incorporate a dashboard that makes data easily digestible and visible to stakeholders. This tool will provide insights and metrics that will help in making business decisions. Here are a few commonly used data visualization tools:

- **Tableau**: This is a powerful tool to create interactive, real-time dashboards and access data sets from multiple sources. It offers robust reporting and sharing capabilities.
- **Power BI:** This tool by Microsoft allows users to create interactive reports and dashboards using a simple drag-and-drop interface. It also offers the ability to embed reports in other applications.
- **QlikView**: QlikView supports a variety of analytics and business intelligence functions, facilitating the creation of sophisticated reports and dashboards.
- **Looker**: Looker is a data platform that makes it easy to create, deploy, and iterate on data visualizations and to share these across an organization.
- **D3.js:** This is a JavaScript library that allows users to create unique and sophisticated datavisualizations for web applications.
- **SAS Visual Analytics**: It provides interactive reporting and dashboards, self-service data discovery, and is a part of the SAS Business Intelligence Suite.
- **Google Charts**: This is a straightforward tool for creating a variety of charts and graphs that can be used on websites.
- **Datawrapper**: An online tool p opular amongst journalists for creating simple charts or maps quickly.

These tools cater to different needs and vary in their complexity, required skill level, cost and versatility.

Real Time Data Processing

We are living in an era where real time data streaming is becoming extremely important. Almost every consumer-based applications require real time data updates. We need companies to make this transformative change and adopt real time techniques. Real Time analytics will make sure of making latest information available for consumers for accurate, timely decisions.

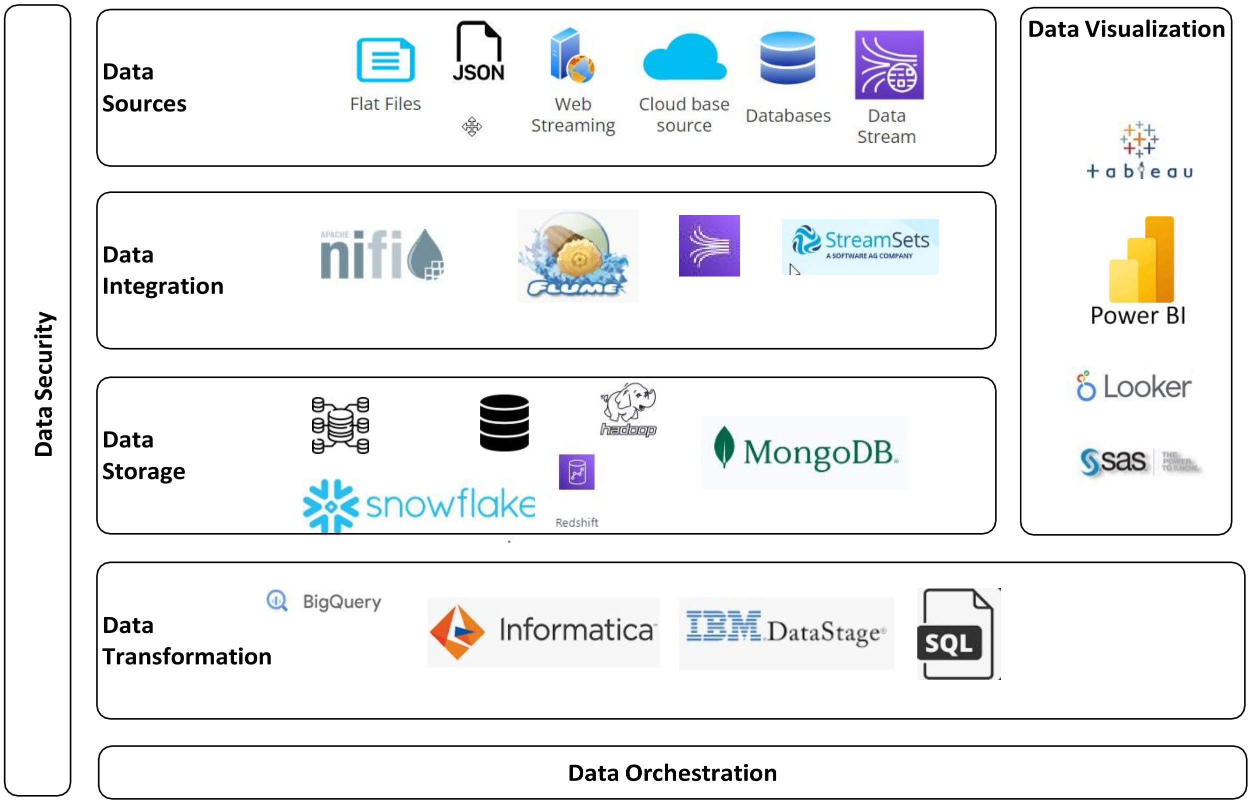
Real-time data processing tools are designed to process or analyze data as soon as it enters the system or database. These are often used in streaming applications where immediate insights are crucial. Here are some commonly used real-time data processing tools:

- **Apache Storm**: Known for its real-time processing capabilities, it enables users to smoothly process unbounded streams of data.

- **Apache Flink**: This open-source stream processing framework is mainly designed for real-time data analytics, batch processing, and machine learning.
- **Spark Streaming**: As part of the Apache Spark platform, it delivers high throughput for processing real-time data streams.
- **Kafka Streams**: Developed by LinkedIn, Kafka Streams is a client library for building applications and microservices that process streamed data in real time.
- **Google Cloud Dataflow**: It's a fully-managed service for executing Apache Beam pipelines within the Google Cloud Platform, performing both batch and real-time data processing tasks.
- **Amazon Kinesis**: Provided by Amazon Web Services (AWS), Kinesis is capable of processingmassive amounts of data in real time and can be used for real-time analytics, log and event datacollection, and more.
- **Apache Samza**: Developed by LinkedIn and incubated by Apache, Samza is designed to handle real-time data feeds at scale.
- **Azure Stream Analytics**: A real-time event data streaming service from Microsoft Azure that includes out-of-the-box integration with Event Hubs, IoT Hub, and Blob Storage.
- **Pulsar**: Apache Pulsar combines high-performance streaming with flexible queuing in a unified messaging model.

Each of these tools has its own strengths, and the choice between them depends on factors like the volume and velocity of data, the nature of the insights required, the existing technology stack, and the required reliability and fault tolerance.

Here is how modern data stack looks like in an Integrated Modernized Data Platform:



Other important factors to consider include:

User Training: Provide usage manuals or training to the employees on how to use the data platform and understand the information presented.

Maintenance and Upgrading: Assure regular performance tuning, keep up with the changing business goals, and meet new project requirements. Implement feedback loops for constant iteration on the design and function of the data platform.

Compliance: Make sure all the data collection, storage, and usage comply with the applicable legal and regulatory requirements.

In conclusion, moving towards a unified architecture for managing data is crucial for businesses in the age of digital transformation. This process involves defining clear objectives, identifying data sources, creating a data model, managing the ingestion, storage, scalability, transformation, security, and visualization of the data. Numerous tools are available to support each of these steps, each suited to different types of data and use cases. The right selection depends on the company's specific needs. This integration not only fosters efficiency and accelerates decision-making, but also opens opportunities for innovation, helping companies to stay competitive and realize their business visions.

A very commonly used African proverb says "It takes a village to raise a child". Similarly, it takes a team of data scientists, data architects, programmers, systems analysts, and end-users from various business divisions for a successful implementation of the data platform.