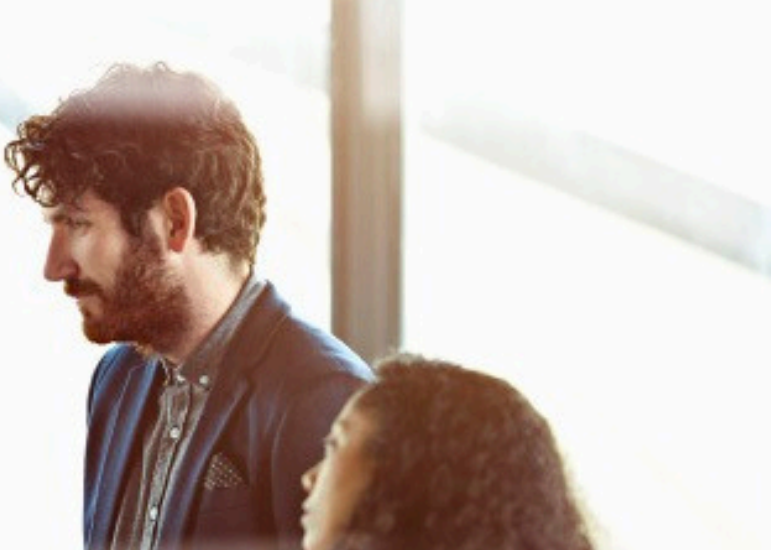
ADEPTIA

Large File Data Ingestion: A Powerful “Big Data” Solution to Handle Data Volume, Velocity, Variety and Veracity

EBOOK





Enterprises are navigating through a period of intense change where business operations are becoming primarily data intensive. Studies indicate that over 2.5 quintillions bytes of data is being generated by enterprise systems each day. This data growth in large part is spurred by digital economies where there are several connected technologies, processes, and systems spanning across different B2B operations:

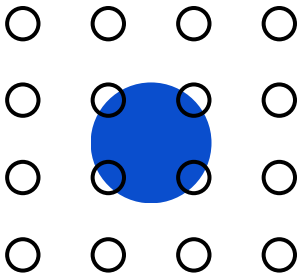
- **Customer data** — Different types of data sent by customers based on the services to be provided such as banking data, stock market transactions, employee benefits, insurance claims, etc.
- **Operations data** — Orders, sales data, point of sales data, pricing data, competitor analytics, online transactions data, etc.
- **Marketing data** — Prospect targeting, market segmentation, prospect contact lists, website log data, web traffic data, etc.

“Many companies take six to 12 weeks to implement data integration with new business customers.”

The Four V's

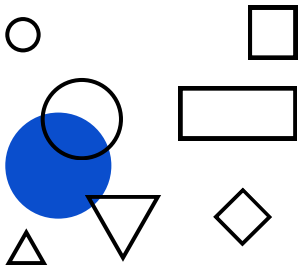
Volume

This is the size in GB, TB and Exabytes of data. Surging volumes and heaps of data being produced at astronomical rates. This volume of data cannot be processed with conventional methods.



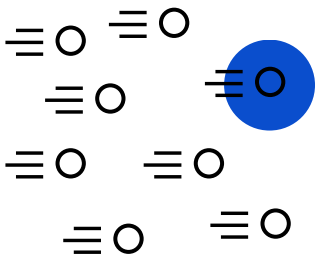
Variety

Semi-Structured, unstructured or heterogeneous data that can be too disparate for enterprise B2B networks. Data could be structured or unstructured data such as videos, pictures etc.



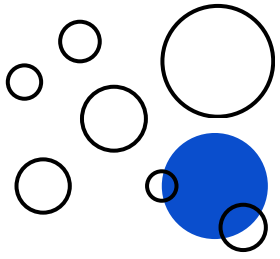
Velocity

Refers to the frequency of incoming data that needs to be processed. Fast moving data that hobbles the processing speed of enterprise systems causing downtimes and breakdowns.



Veracity

This refers to the quality or trustworthiness of the data. Analyzing loads of data that are not accurate and contain anomalies is of no use. Teams cannot use it for business operations.



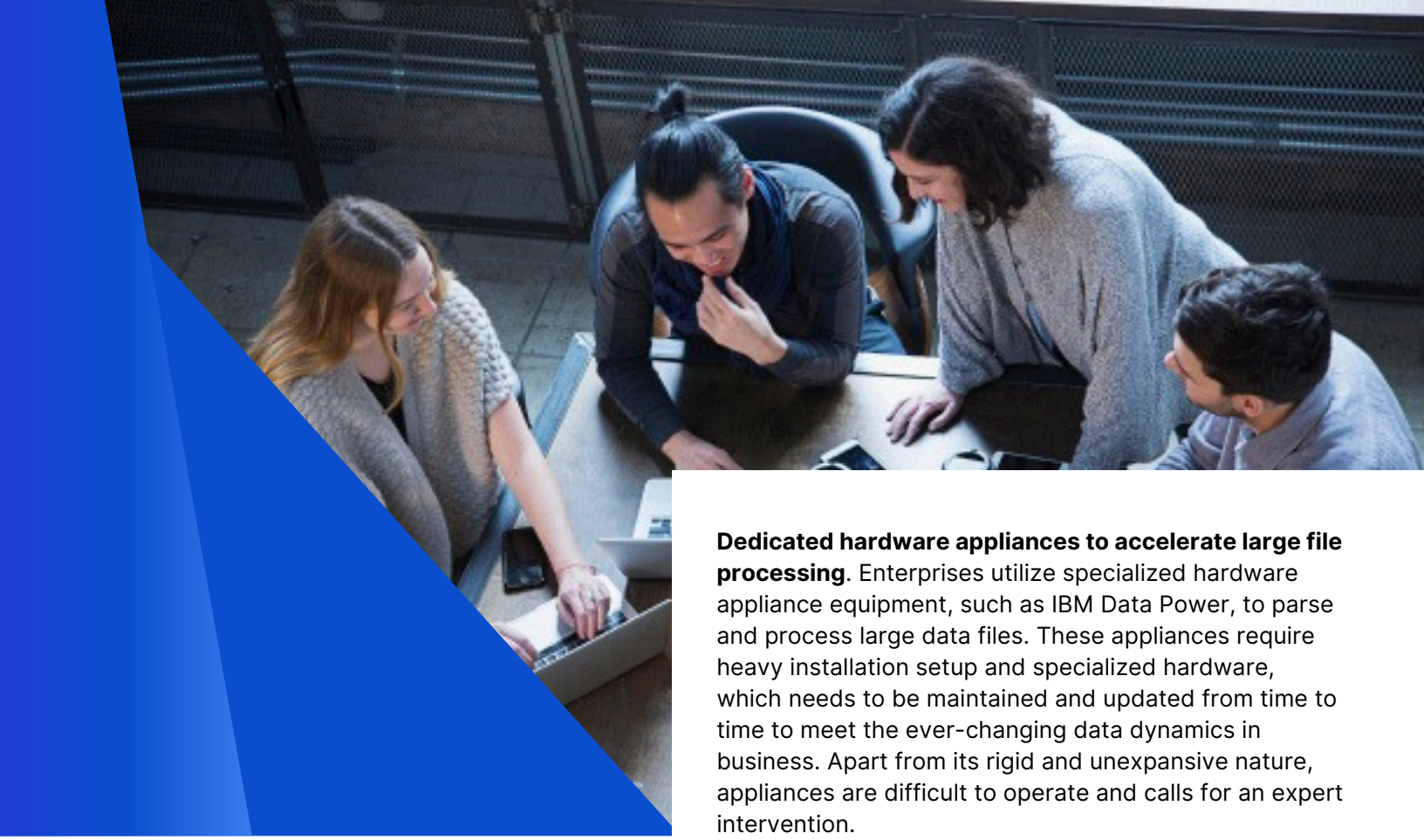
The Importance of Agility

Data is usually shared or moved in files and the processing of large or flat files leads to application failures and breakdown of enterprise data flows, causing incomprehensible information losses and painful delays in mission-critical business operations. Enterprises lose valuable time and money in discovering, extracting, preparing, and managing rogue data sets. They fail to adapt their business according to new market realities and capitalize on market opportunities.

The need for processing large data sets with agility is dire — where earlier businesses were handling data in kilobytes and megabytes, now they are talking about tens and hundreds of GBs of daily volume. There has to be a foundational technology that helps businesses to get quick access to data, prepare it for analytics, and deliver improved services to customers.

To solve this problem, Adeptia has developed its Large File Data Ingestion feature. This innovative “big data” solution is a total shift from the conventional ways of structured data processing, it resolves all multi-GB data processing hassles in a single interface. Near zero-second latency enables users to manage torrents of data and ensure high availability of data at all times. The solution makes it a lot easier to pull scattered pieces of puzzle together and extract value out of them in all four dimensions: Volume, Variety, Velocity, and Veracity of data. On its interface, business users get pre-built constructs to deal with real time data in action. They can deal with computational complexity & quickly process data in structured, unstructured, and heterogeneous formats.





Current Way of Handling Large File Data Exchange

Enterprises use different approaches for navigating large data sets, which don't necessarily conform to expectations of quality. Many of these approaches don't promise guaranteed smooth processing of data residing in mature and functional warehouses. These methods can be complicated, they defeat the objective and just recreates silos.

Dedicated hardware appliances to accelerate large file processing. Enterprises utilize specialized hardware appliance equipment, such as IBM Data Power, to parse and process large data files. These appliances require heavy installation setup and specialized hardware, which needs to be maintained and updated from time to time to meet the ever-changing data dynamics in business. Apart from its rigid and unexpansive nature, appliances are difficult to operate and calls for an expert intervention.

Custom coding. Most enterprises use the approach of writing specialized custom coding for processing multi-GB data. High level of expertise and skill set is required to write custom codes that split large files into smaller chunks and then merge the obtained results. This coding snowballs and becomes difficult to maintain and upgrade. Rigid coding fails to scale with new coding requirements and data files. It also requires manual oversight to ensure correct processing and handling of frequent errors and exceptions.

Big data software tools. Big Data tools are often nowadays used for collection of data sets but this form of data processing has major limitations. The purpose of this data collection is mostly for analytical purposes. Another relevant drawback associated with Big Data is that it requires heavy coding and expertise for moving data files and is just an extension of custom coding approach. Enterprises need training and expertise to use Big Data tools properly. New code needs to be written for processing new data files and big data tools typically only work with simpler, flat data formats like log files and not complex or hierarchical multi-record formats.



Adeptia’s Large File Data Ingestion Solution Fixes this Critical Problem

Large File Data Ingestion Solution provides key breakthroughs as compared to traditional solutions and costly appliances for multi-GB data processing. This feature can process multi-GB files, ingest and transform large colossal amount of data, and deliver this data in designated common formats.

Large File Data Ingestion Solution Processes using typical server hardware:

- 25GB XML file with complex transformation rules in 33 minutes
- 200GB XML with complex transformation rules in 4 hours
- 50 different XML files of 25GB concurrently with complex transformation rules in 10 hours
- 10 different text files of 5GB concurrently in less than an hour

Powerful features for end-to-end data management without expensive hardware and separate infrastructure:

- Data Quality Management: Enables enterprises to process, consolidate, and maintain a lineage of data. Ensures that enterprise data meets precise standards, compliances, and business rules.
- Business Data Lake Architecture: Makes data ready for big data projects & applications and aligns it with organizations business objectives. Hub & Spoke Model supports Big Data projects: real-time analytics, Data Lake, etc.
- Metadata Management: Meta-driven architecture manages both structured and unstructured data in EDI files or non-EDI files. It supports any-to-any data transformation and workflows for handling exceptions.
- Data Monitoring: Rich monitoring and tracking interface with real-time dashboards helps operational users to track and get all the data without the need of going to multiple systems.
- Data Protection: Industry-leading secure communications framework ensures full data security with encryption, guaranteed data delivery, VAN connectivity, etc. essential for high volume file transferring.
- Enterprise Data Governance: Flexible data governance strategy helps in shaping business data in new & unique ways. Enables enterprises to manage data lifecycle for compliances and other purposes.



"A self-service data integration solution helps to retain customers and improves their lifetime value through a superior customer experience."

Large File Data Ingestion and Streaming Vs Current Generation Data Handling Approaches

A unique feature about Large File Data Processing approach is that it is offered as a pure-software solution vs. the old approach of hardware appliance or custom-coded solution. Its easy-to-manage setup helps clients get up and running with high speed data processing over slower and manual old approach. The solution ensures architectural coherence, centralized management, security, automated error handling and, top-down control interface to reduce large file data processing time.

- Allows non-technical business users to process large files easily without manually coding or relying on specialized IT staff
- Reduces manual effort and cost overheads, ultimately accelerating delivery time
- Eliminates requirement of expensive hardware, IT databases, and servers
- Handles large data volumes and velocity by easily processing up to 100GB or larger files
- Handles data variety by supporting structured data in various formats, ranging from Text/CSV flat files to complex, hierarchical XML and fixed-length formats
- Handles data veracity by allowing data validation, cleansing and data integrity rules to be applied while processing the data

Understanding Large File Data Ingestion and Streaming in Detail

Large File Data Ingestion and streaming capability is built from the ground up to address the common big data problems. Business users get constructs to stream multi-GB data in heterogeneous formats so that it is accessible in a consistent manner. The solution offers business user capabilities to enable data streaming at speeds the future demands:

- Transforms large data sets easily, quickly and economically.
- Minimizes latency (delay) and downtimes: Business users can see results before they read the whole input.
- Hands data processing to business users: Minimizes resources when handling many small documents.
- Handles real-time data feeds: Streaming can be scaled for data that is infinite, that never stops.

Large File Data Ingestion capability provides streaming ability to process large Volume, Variety, Velocity, and Variety of data sets without the need of building a tree in memory. It provides a range of constructs for processing large sets of documents in streaming mode. At its heart resides a streaming mode for processing gargantuan data sets. The data which needs to be processed is read by a powerful transform engine which eliminates the need of building a tree representation of the document in memory. Streaming allows the engine to avoid rewriting streamable constructs into a non-streamable form and to start producing output before it has finished receiving its full input, thus reducing latency.

The data life cycle, from ingestion, manipulation can be tracked whenever there is a change. The platform allows people, processes and technologies to consume data and push data out to other sources, cloud stores or connected devices. Enterprises can get smart about their data and process it towards the stream of business opportunities.

The Large File Data Ingestion solution allows large file sizes of 100s of GBs running on a single server (with say 4 cores and 16 GB RAM) to be ingested, processed and transformed in a short amount of time without the need of correspondingly large amounts of memory.

To illustrate: Suppose there is a very flat XML document (no deep nesting) such as an employee file with data of 100,000 employees in it. There are 100,000 elements at level 2 in the tree. Large File Data Ingestion solution processes each employee in turn. It builds a tree for each employee and processes that tree in normal (non-streaming) way. It never builds a tree for the entire file, so the memory requirements are bounded to just the amount of space needed to hold a single employee element. This sort of streaming has got practical potential to stream 10s and 100s of GB of data.

Business Benefits of Large File Data Ingestion and Streaming

Large File Data Ingestion is built to meet today's scalability, efficiency, and zero-impact requirements of data processing. It is an enhanced functionality for processing and moving huge sources of real-time data to target systems or various users. Enterprises of all sizes can use this functionality for monetizing data and supporting enterprise initiatives. With this capability, the data processing power of an enterprise increases by 80X and they don't have to stop for processing data in batches. They can integrate and process data faster and meet analytics objectives without costly hardware appliances & infrastructure.

For enterprises dealing with the 4 Vs of data, the next round opportunity is coalescing around the large file data ingestion capability. This unique capability helps in scaling out for higher iterative data intensive work and feeding structured & unstructured data to data lakes. Data lake processing becomes more manageable, available, cost-effective, and persistent. Enterprises can synchronize cloud repositories & data warehouses and reduce batch processing loads.
Improve customer data onboarding with our Adeptia's Large file data ingestion and streaming, request a demo now.





© Copyright 2023 Adeptia.
All marks are the property of their respective owners.