Data warehouse augmentation, Part 2: Use big data technologies as a landing zone for source data

Shweta Jain (shweta.jain@in.ibm.com)
IT Architect
IBM

Sujay Nandi (sujnandi@in.ibm.com)
Senior IT Architect
IBM

10 June 2014

Use a Hadoop environment as the landing zone to pull in data from various sources, process it, and transfer the processed data to the existing data warehouse or other repositories. Explore scenarios of different ways to implement a landing zone. Learn about the architecture of the zone and the tools and techniques for integrating it with various environments.

View more content in this series

Other articles in this series

Part 1 of this series describes the current state of the data warehouse, its landscape, technology, and architecture. It identifies the technical and business drivers for moving to big data technologies and identifies the use cases for augmenting the existing data warehouse by incorporating big data technologies. This article further describes one of the use cases that are introduced in Part 1, using a Hadoop-based environment as a landing zone for big data. Multiple scenarios for storing, handling, preprocessing, filtering, and exploring big data are described. This article introduces the concept of a landing zone and describes its logical architecture and high-level components.

Landing zone

Hadoop provides the capability to configure a landing zone, which acts as a staging area for data. The landing zone accepts data from many structured and unstructured sources. It is responsible for several tasks:

- Provides early data exploration: Explore data to determine what data can be moved to the data warehouse and retained there to save on storage costs. Data at rest is often in this category.
• Handles streaming data: The volume of streaming data can large and not all of it must to be stored. The landing zone can accept, analyze, and process streaming data without storing it first. Use the landing zone to determine what streaming data must be saved, either in the Hadoop Distributed File System (HDFS) or in a data warehouse.
• Provides a preprocessing facility: The landing zone can transform the data before it loads the data in the data warehouse.

Scope and boundary of the landing zone

As shown in Figure 1, the landing zone exists between the data sources and the data warehouse. Real-time analytics processing is outside the landing zone.

Figure 1. Scope of the landing zone

Implementing a landing zone to handle big data

A big data platform provides the infrastructure to store and process the massive volume of data in a parallel and distributed way. The data includes structured, semi-structured, or unstructured data at rest or data in motion.

The volume of data is stored in distributed file systems such as HDFS and in document-oriented databases such as MongoDB. The data is processed by distributed processing technologies such as MapReduce. The data is fetched from the distributed storage by using SQL-like programming models such as Jaql, Hive, and Pig.

A landing zone to handle big data can be implemented for the following scenarios.
Big data as a data source to the data warehouse

Do you have data that can provide significant business insight if it were in the data warehouse?

Common big data sources for businesses are social media, mobile and smart devices, weather information, and information that is collected via sensors. Adapters are available to connect to these data sources. Data can be acquired directly from the data sources or through data providers. These providers own or acquire the data and expose it in specific formats, at required frequencies, and through specific filters.

As shown in Figure 2, the first step is to move the data into the landing zone. Convert the unstructured data into structured or semi-structured format and perform the initial preprocessing, extraction, and transformation of the data. Store the data in a storage repository that can accommodate it. The next step is to identify possible entities and use an entity identifier to get the entities of interest. Further parallel processing can be performed on the stored data so that the outcome of analysis can be consumed in various forms such as graphs, dashboards, and summary reports.

In this scenario, big data is acquired from new sources and can be moved to the existing data warehouse platform. This new data can be used to get more business insight from warehouse applications.
Do you have documents that must be manually searched?

Unstructured data from within the enterprise is another source of data for the landing zone. The large amounts of data that is stored in document management tools include information in the form of digital documents, images, video, and audio.

Unstructured data from these internal applications can be sent to the landing zone. As shown in Figure 3, the first step is to move the data into the landing zone. Convert the unstructured data into structured or semi-structured format, and perform the initial preprocessing, extraction, and transformation of the data. During preprocessing, existing entities from the data warehouse can be correlated with the entities from unstructured data so that data can be moved to the data warehouse. The next step is to identify possible entities and use an entity identifier to get the entities of interest. Further parallel processing can be performed on the stored data so that the outcome of analysis can be consumed in various forms such as graphs, dashboards, and summary reports.

In this scenario, unstructured data from within the enterprise is stored in storage repositories so that it can be explored and consumed. The filtered data is also moved to the data warehouse.
Move data from the data warehouse to big data storage

Does the storage limit of the data warehouse force you to purge or archive data that is still valuable?

Are there reports that require a significant amount of time to process large amounts of data? If these reports were available in real time, might they help the organization make faster decisions?

As shown in Figure 4, the first step is to move the data from the data warehouse into the landing zone. Perform the initial preprocessing, extraction, and transformation of the data and then store the data in a repository. Further parallel processing can be performed on the stored data so that the outcome of analysis can be consumed in various forms such as graphs, dashboards, and summary reports.

In this scenario, big data is acquired from an existing data warehouse and stored in big data storage cost effectively. The data that might be archived or purged because of the limitations of existing data warehouse is now available to be used for more business insight.
Figure 4. Move data from data warehouse to big data storage

Move staging data into big data storage

Is there data that needs complex processing before it can be stored in a data warehouse?

As shown in Figure 5, the first step is to move the data from the data warehouse staging area into the landing zone. Perform the initial preprocessing, extraction, and transformation (if needed) so that the data can be moved to the data warehouse. Store the data in big data storage, if necessary. Further parallel processing can be performed on the stored data so that the outcome of analysis can be consumed in various forms such as graphs, dashboards, and summary reports.

In this scenario, big data is moved from the data warehouse staging area to the big data platform before it is pushed to the data warehouse.

Large volumes of data that require time-consuming, complex processing can now be processed by using big data platform.
Process streaming data in the landing zone and move filtered data to the data warehouse

Consider whether streaming data might provide more business insight and help the organization respond to situations in a more timely manner.

Typically, streaming data comes into an organization at a high velocity. As shown in Figure 6, the first step is to move the streaming data into the landing zone. Convert the unstructured data into structured or semi-structured format, and perform the initial preprocessing, extraction, and transformation of the data. Store the data in a storage repository that can accommodate it.

In this scenario, streaming data is acquired from new sources in real time and preprocessed in the landing zone. Filtered data can then be moved to the data warehouse. This additional data can be used to get more business insights from warehouse applications.
Explore data in the landing zone and move filtered data to the data warehouse

Do you have big data that might provide more business insight to existing enterprise applications if it were in the big data platform?

As shown in Figure 7, the first step is to acquire data from external sources. Convert the unstructured data into structured or semi-structured format, and perform the initial preprocessing, extraction, and transformation of the data. Preprocess and analyze the data in the landing zone. Perform data exploration in landing zone to determine the data that can be moved to the data warehouse.

In this scenario, big data is acquired from new sources, analyzed and explored in the landing zone, and filtered data is then moved to the data warehouse. This additional big data can be used to get more business insight from warehouse applications.
Figure 7. Explore data in landing zone and move filtered data to the data warehouse

Architecture and logical components of the landing zone

The landing zone includes the following components:

- Integration environment
- Data exploration environment
- Hadoop environment: Distributed file system (includes the name node and data node) and MapReduce zone (includes the job tracker and task tracker)
- Streams environment

Integration environment

The integration environment, through a data integrator, provides an interface between various data sources, Hadoop, and the data warehouse. IBM® InfoSphere® Information Server provides this integration capability.

IBM DataStage®, a component of InfoSphere Information Server, is an extract, transform, and load (ETL) tool that can integrate and transform multiple, complex, and disparate sources of information. DataStage brings the data into IBM InfoSphere BigInsights™, which analyzes the new data regularly. DataStage also pushes data from Hadoop to the data warehouse.

The data integrator thus manages several kinds of integration:

- Integration with JDBC sources through a general-purpose Jaql module
- Integration with ODBC data sources
- Integration with IBM® DB2®
- Big SQL integration: ODBC drivers, LOAD statements from DB2, IBM Netezza data warehouse appliances, and Teradata. Big SQL enables access to all data in Hadoop via the JDBC/ODBC interface. Big SQL enables the IBM Cognos® Business Intelligence server to transfer many types of computations to BigInsights MapReduce processing.

Figure 8 depicts a scenario in which Hadoop contains the data integrator.

**Figure 8. Landing zone architecture (Hadoop with a data integrator)**

Figure 9 describes a separate integration environment outside Hadoop.
Data exploration environment

IBM InfoSphere Data Explorer serves as the data exploration component (shown in Figure 10), which provides the visualization and discovery capabilities to the landing zone. It enables users to discover and navigate data before and after preprocessing.
In the landing zone, Data Explorer provides several functions:

- An early discovery environment to identify the most suitable content to move to data warehouses, the data that must remain in Hadoop, and the data that can be discarded
- A flexible, compact architecture that is built on a position-based index, rather than a traditional index-based approach
- Multiple interfaces such as web, command line, API, and framework for easy administration and deployment
- Integration with the Hadoop environment that facilitates data exploration

**Hadoop environment**

As shown in Figure 11, the Hadoop environment is for large-scale processing of data, especially data at rest. It contains a storage framework and an operations framework. InfoSphere BigInsights provides the Hadoop environment.

**Figure 10. Data exploration environment**

![Data exploration environment diagram]

**Figure 11. Hadoop cluster**

![Hadoop cluster diagram]
Storage framework: Distributed file system

The Hadoop Distributed File System (HDFS) is optimized to support large files. It spreads data across nodes at load time. HDFS includes the following major components:

- Name node
- Data node
- Job tracker

Each block of data (or portion of a file) is replicated across three data nodes, by default, that are contained in racks that hold multiple data nodes. The first replica is placed on the same node as the client application that needs to read the data, if the client application is running on the same cluster. The second replica is placed on a different rack from the first rack. The third replica is placed on the same rack as the second rack, but on a different node.

Operational framework: MapReduce zone

The operational framework includes only the simple programming model MapReduce, which contains a task tracker for the MapReduce tasks. Each cluster includes one or more task trackers.

Streams environment

The streams environment is the entry point to the landing zone for the data that comes in from various streaming sources. The streams environment handles and processes the data in near real-time. Real time in this situation implies low latency, which is the delay from the time a packet of data arrives to the time the result is available.

IBM InfoSphere Streams serves as the stream environment. InfoSphere Streams processes data in memory, rather than accessing mass storage from a disk. As shown in Figure 12, the first set of analytics pushes the relevant data to the Hadoop environment. Because InfoSphere Streams is fast, scalable, and programmable, data analysis ranges from simple to the sophisticated. Complex analytics happen outside the landing zone.

Figure 12. Flow of data through the streams environment

Preprocessing tasks

Data that is loaded in the Hadoop environment can undergo many types of preprocessing:

- Initial data filtration using the exploration zone
- Data cleansing before the data is sent to the data warehouse or sent for transformation
- Data transformation before it is sent to the data warehouse
• Merging of structured, semi-structured, and unstructured data. For example, in the banking scenario that is shown in Figure 13, the following tasks occur:
  • Data from multiple sources, related to calls, are loaded into Hadoop.
  • To analyze the data in the system of record (SOR), source attributes are selected.
  • Semi-structured and unstructured data that is related to structured data is modeled.
  • A three-month rolling subset of detailed, structured data is loaded into SOR relational database management system (RDBMS) to support detailed, operational reporting.
  • Aggregates are created.

Figure 13. Preprocessing tasks include loading data through Hadoop

Integrating Hadoop with various environments

To include Hadoop in the data processing environment, integrate it with other components in the architecture.

Integrate with existing data warehouse and other traditional data sources

To integrate Hadoop with the relational, traditional data sources and data warehouse, use data integration of ETL processing, as shown in Figure 14.

Figure 14. Data integration using ETL

Integrate with the streams environment

As shown in Figure 15, integration of Hadoop with the stream environment takes place directly within Hadoop environment, by using the following tools in InfoSphere Streams:

  • HDFSDirectoryScan: Similar to DirectoryScan, except for HDFS
Integration with the data exploration zone

When data exploration is integrated with the Hadoop environment, content that is stored in distributed storage can be accessed and navigated. The integration enables users to discover and navigate data before it is loaded into the warehouse. The data exploration tool creates a user interface for information discovery and exploration of data in HDFS. It serves as the early discovery tool to identify the most suitable content to move into the warehouses and the content to retain in Hadoop.

InfoSphere Data Explorer uses a crawling and indexing model that uses a commit mechanism that ensures the consistency of InfoSphere Data Explorer indexes. Because it is built on XML, InfoSphere Data Explorer offers flexibility in integration. All configurations are saved in XML format, data that is ingested during querying or crawling is always converted into XML to enable further manipulation. All data is first produced in XML format before it is ultimately transformed into other formats, such as HTML for delivery or presentation.

InfoSphere Data Explorer servers can receive data in real time from the cluster of the Hadoop server. InfoSphere Data Explorer can also push relevant data to Hadoop from the applications that are created by using the InfoSphere Data Explorer Application Builder.

Summary

In the past, organizations had to extract data from many sources and move the data to a data warehouse or database before they queried, analyzed, discovered, and explored it. Often, these landing or staging areas were used as only temporary storage space before they passed the data to the permanent data warehouse or a database.

With Hadoop-based technologies, organizations can use the landing area more effectively. The landing zone receives data from many sources that include real-time, streaming sources. The landing zone serves as the preprocessing zone to cleanse the data and transform it before the zone sends it to the data warehouse. With the landing zone filtering the data, the data warehouse capacity and performance is improved.
Part 3 of this series describes the case of using big data technologies for historical data in the data warehouse. Given the volume of data that flows through organizations, it's important to implement an effective process of moving less frequently used data out of the data warehouse into an archive repository. Part 3 describes how to build active archives for historical data in a data warehouse by using big data technologies.
Resources

Learn

- "Big data architecture and patterns" series (developerWorks, September 2013): Classify big data problems, choose an architecture for a big data solution, and implement the solution with the IBM big data platform.
- IBM Big data warehouse augmentation: Video that describes various types of data warehouse augmentation scenarios.
- Architecting a big data platform for analytics (Mike Ferguson, Intelligent Business Strategies): Provides an end-to-end overview of architectural solutions and strategies for big data platform.
- IBM InfoSphere BigInsights tutorials: Provides information about the Hadoop platform from IBM.
- Enterprise Information Protection - the Impact of Big Data (Mike Ferguson, Intelligent Business Strategies): Provides the view on enterprise information security and privacy and the impact of big data on it.
- IBM InfoSphere Data Explorer architecture: IBM white paper provides an architectural view of the IBM® Data Explorer.

Get products and technologies

- Download InfoSphere BigInsights Quick Start Edition: Available as a native software installation or as a VMware image.
- Download InfoSphere Streams Quick Start Edition: Stream computing software, free to download. Quick to start.

Discuss

- Ask questions and get answers in the InfoSphere BigInsights forum.
About the authors

Shweta Jain

Shweta Jain is an accredited IT architect with IBM AIS Global Delivery with more than 10 years of industry experience. She specializes in building an architecture for SOA-based integration solutions that use industry standards and frameworks. She has experience in architecture, design, implementation, and testing of integration solutions that are based on the SOA framework, SOMA methodology, and the software development lifecycle based on methods. As an integration architect, she is responsible for architecting the BPM/EAI layer by using IBM tools, standards, processes, and methodologies; and incorporating industry standards for complex integration and transformation projects. She is driving some of the Technical Excellence initiatives in emerging trends of big data, M2M, IoT, and Web API.

Sujay Nandi

Sujay Nandi is a Certified Senior IT Architect with IBM in the Global Business Services (GBS) business unit. He specializes in the travel and transportation industry including railways, airlines, and shipping freight and logistics. Sujay has nearly 14 years of experience as an application architect. He has expertise in solution architecture definition, evaluation, selection of software products, and presenting on complex, custom application development, application integration, and application migration projects in multiple technologies that include Microsoft, IBM, Open System, Informatica, and Oracle. He also has experience with System Engineering Architecture during his work with Global Technology Services (GTS) in ensuring the availability, reliability, and optimal capacity of the system. Sujay was part of the team in driving the big data initiative in enterprise architecture and technology and other initiatives in large focused accounts.

© Copyright IBM Corporation 2014
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)