For organizations that have seen the benefits of big data, integrating Apache Hadoop into their data management strategy means struggling with new developer tools, new user interfaces, and a need for new skills in a completely new environment. They're learning to cope with questions such as:

- What use cases have they overlooked in developing their big data strategy?
- Which skills transfer effectively from existing data management projects to Hadoop-centered projects? Which ones are even necessary?
- How can they ensure the availability of resources skilled in such diverse disciplines as statistics, process engineering, distributed programming, and information architecture?

Moreover, they're asking these questions in the context of existing business intelligence (BI) and data warehousing strategies and tools that Hadoop must extend, reinforce, or replace – but it’s not clear which approach is best.

With all of these questions comes an array of confusing answers. Some technology vendors, for example, have assimilated Hadoop into their marketing messages more than their technology – even claiming that their tools run “in the cluster” when, in reality, they’re just running under Linux on one of the cluster’s nodes, without taking advantage of native...
Hadoop resource negotiation. Organizations are left wondering whether their big data integration choices are limited to complex native scripting on one hand and non-Hadoop tools on the other.

**Native Hadoop Data Management**

*iWay Big Data Integrator (iBDI)* helps organizations deal with these questions by supplying a modern, native approach to Hadoop-based data integration and management that ensures high levels of capability, compatibility, and flexibility.

- It provides a simplified, easy-to-use interface to perform data integration in Hadoop, including data ingestion, transformation, and data quality
- By eliminating Hadoop coding, you can purchase a system that enables a more rapid response to business requirements using less-expensive skill sets
- It runs under YARN, taking advantage of native Hadoop performance and resource negotiation
- It uses Hadoop standards, such as Avro, in support of a variety of big data integration use cases
- It includes additional sophisticated process management and governance
- It avoids composing, debugging, and using batch mode for long-running scheduled jobs, ensuring long-term stability in dynamic environments
- It supports contemporary Hadoop solutions

iBDI enables organizations to support any kind of use case by incorporating Apache Spark™ pipelines for transformation, cleansing, joining, and more; Sqoop for data replication, capture, and export; Kafka, Flume, and even Flafka for streaming and unstructured data acquisition; and iWay Service Manager for ingestion of non-Hadoop data (e.g., transactions, messages, alternate data sources) – all running natively in the cluster.

**iWay Service Manager** is an integration engine that enables organizations to create, compose, and manage services, respond to internal and B2B events, and interoperate with standards-based and proprietary technologies.

Skills learned outside the cluster (e.g., using iWay Service Manager for traditional real-time data integration) can easily be transferred into the cluster, and vice versa. iBDI's easy-to-use interface ensures that users don't need as much training or knowledge of Hadoop to be effective.

Existing BI and data warehouse infrastructure can be extended, reinforced, or replaced as desired using the same infrastructure. For example, the same data quality rules used outside of Hadoop on non-Hadoop data can also be applied to a data lake in Hadoop.
Key Capabilities

Spark Pipeline

iWay Big Data Integrator uses Spark-based pipelines, which operate as building blocks to create processing without programming. The pipelines utilize streaming, Hadoop-defined or RDBMS data sources all by the use of parameters, and not code. They also use compute stages like order by or cleanse, and then use the console to view an extract of the data. You can also take advantage of the flexible configuration options to deploy the pipeline as a real-time component or publish to run as a scheduled job.

With the Spark-based pipeline, many operations can be performed in memory on the Spark cluster, which is more efficient and saves resources, time and read/write operations. The pipeline in iBDI uses the DataFrames API of the Apache Spark Framework, which is an optimized engine for cluster computing.

Data Ingestion

Modern data integration often requires rapid ingestion of data into a data lake from varied data sources, while minimizing the impact on their performance. iWay Big Data Integrator hides the complexity of data ingestion, replacement, and de-duplication using pipelines or native Spark, Flume support, other techniques, and our iWay Service Manager – all without programming. Technologies include:

- Flat streams. Avro, JSON, and XML are supported in the Pipeline Spark Streaming Source
- Native Spark Streaming. A simple socket text stream reads data from a port and builds data frames to pass to the pipeline
- Kafka Streaming. Topics can be collected through Kafka, and then aggregated, enriched, and passed on for further processing by Kafka or other methods
- Nifi Streaming. Nifi pipeline listeners support additional protocols and file types that can be ingested by iWay Big Data Integrator
- Changed Data Capture (CDC). Data capture can be extended beyond typical Hadoop approaches to include automatic comparisons to previous captures of the same data to determine where changes have taken place
- Lambda Architecture. Kafka and Spark-based pipelines enable a Lambda Architecture pattern that provides a data warehouse solution with batch operations for bulk data retrieval and real-time queries on a streaming speed layer so that important up-to-date data is not missed

iWay Big Data Integrator supplies a modern, native approach to Hadoop-based data integration and management.
- Architectural Flexibility. iWay Service Manager can be used for structured data, Apache Flume for unstructured and streaming data, and Spark and Kafka for streaming.
- iBDI enables the transport of streaming data sources (e.g., network traffic data, social media data, log data, etc.) into HDFS using Apache Flume or Spark as the underlying technology. Spark streaming is supported with Avro schema, XML and JSON objects, native Spark streaming, Kafka streaming, and Nifi streaming.
- iSM can run natively within the cluster, providing an easy way for the Hadoop administrator to manage the ingestion of multiple sources of legacy data (e.g., transactions, message protocols, data sources, etc.). iWay Service Manager also provides native data transformation capabilities with a rich array of out-of-the-box functions.

**Data Transformation**

iBDI leverages the Hadoop ecosystem and native components to transform data from many sources to one or many targets. These targets can be sources for follow-on transformations.

An expansive library of operators ships with iBDI, including relational, arithmetic, logical, complex type constructors (e.g., Map), and complex type operators (e.g., array) and functions (mathematical, collection, type conversion, date, conditional, string). Aggregate functions, table-generating functions, grouping, and sorting are also included. Users can create custom functions as well.

iBDI provides the ability to transform data in Hive through the transformer tool. Hive tables can be transformed through the transformer in a pipeline or by creating an entirely new Hive table via the Mapper, which allows a source table to be mapped to an output schema. The source table can be joined to other tables. This enables rapid migration from non-Hadoop data to Hadoop, with automatic metadata generation to support easy follow-on usage.

With iBDI, the same tools that enable efficient Hadoop/Spark import can be used to export data into a variety of formats on the Hadoop cluster or into legacy sources.

**Learn More**

To find out more about iWay Big Data Integrator or schedule a demo, visit our website, or call 800-969-INFO.