Working with Big Data and Hadoop in PDI

Pentaho Data Integration (PDI) can operate in two distinct modes, job orchestration and data transformation. Within PDI they are referred to as jobs and transformations.

PDI jobs sequence a set of entries that encapsulate actions. An example of a PDI big data job would be to check for existence of new log files, copy the new files to HDFS, execute a MapReduce task to aggregate the weblog into a click stream and stage that clickstream data in an analytic database.

PDI transformations consist of a set of steps that execute in parallel and operate on a stream of data columns. Through the default Pentaho engine, columns usually flow from one system where new columns can be calculated or values can be looked up and added to the stream. The data stream is then sent to a receiving system like a Hadoop cluster, a database, or even the Pentaho Reporting Engine.

You can also run transformations using the Spark engine. Pentaho uses the Adaptive Execution Layer (AEL) to run transformations in different engines. AEL builds a transformation definition for Spark, which moves execution directly to the cluster, leveraging Spark's ability to coordinate large amounts of data over multiple nodes. See Adaptive Execution Layer for details.

See Big Data Tutorials for examples of how to use PDI jobs and transforms in typical big data scenarios. PDI job entries and transformation steps are described in Transformation Step Reference and Job Entry Reference.

PDI's Big Data Plugin

The Pentaho Big Data plugin contains all of the job entries and transformation steps required for working with Hadoop, Cassandra, and MongoDB.

PDI can be configured to communicate with most popular Hadoop distributions. See the Set up Pentaho to Connect to Hadoop Cluster section for more information.

For a list of supported big data technology, including which configurations of Hadoop are currently supported, see the section on Supported Components.

Using PDI Outside and Inside the Hadoop Cluster

PDI is unique in that it can execute both outside of a Hadoop cluster and within the nodes of a Hadoop cluster. From outside a Hadoop cluster, PDI can extract data from or load data into Hadoop HDFS, Hive and HBase. When executed within the Hadoop cluster, PDI transformations can be used as Mapper and/or Reducer tasks, allowing PDI with Pentaho MapReduce to be used as visual programming tool for MapReduce.

These videos demonstrate using PDI to work with Hadoop from both inside and outside a Hadoop cluster.

Loading Data into Hadoop from outside the Hadoop cluster is a 5-minute video that demonstrates moving data using a PDI job and transformation: http://www.youtube.com/watch?v=Ylekzmd6TAc
Use **Pentaho MapReduce** to interactively design a data flow for a MapReduce job without writing scripts or code. Here is a 12 minute video that provides an overview of the process: [http://www.youtube.com/watch?v=KZe1UugxXcs](http://www.youtube.com/watch?v=KZe1UugxXcs).

- **Pentaho MapReduce Workflow**
- **PDI Hadoop Job Workflow**
- **Hadoop to PDI Data Type Conversion**
- **Hadoop Hive-Specific SQL Limitations**
- **Big Data Tutorials**