



Guidelines: Pentaho Tools and Integrated Authentication Methods



Change log (if you want to use it):

Date	Version	Author	Changes

Contents

Overview.....	1
Before You Begin.....	1
Pentaho Tools – Integrated Authentication Options.....	2
Pentaho Analyzer	2
Background Information.....	2
Authentication Methods for JDBC.....	2
Pentaho Reporting and Interactive Reporting	2
Background Information.....	3
Authentication Methods	3
Streamlined Data Refinery.....	3
Background Information.....	3
Authentication Methods	3
Data Integration Calls from the BA Server.....	3
Related Information.....	4

This page intentionally left blank.

Overview

This document gives a brief description of the different authentication methods recommended for the Pentaho Tools when using them in a Big Data clustering environment.

Our intended audience are system admins or anyone with a background in security who is interested in authentication.

The intention of this document is to speak about topics generally; however, these are the specific versions covered here:

Software	Version(s)
Pentaho	6.x, 7.x, 8.0

The [Components Reference](#) in Pentaho Documentation has a complete list of supported software and hardware.

Before You Begin

This document assumes that have knowledge of Pentaho tools and authentication and that you have already [installed](#) and [configured](#) Pentaho. More information about related topics outside of this document can be found in [Pentaho Documentation](#).

Pentaho Tools – Integrated Authentication Options

There are diverse ways that Big Data uses a Big Data cluster to authenticate Pentaho Client Tools, depending on use cases and tools.

You can find details on these topics in the following sections:

- [Pentaho Analyzer](#)
- [Pentaho Reporting and Interactive Reporting](#)
- [Streamlined Data Refinery \(SDR\)](#)
- [Data Integration Calls from the BA Server](#)

Pentaho Analyzer

This tool uses Java Database Connectivity (JDBC) connections to communicate with the database using a model based on Mondrian's online analytical processing (OLAP) engine. This is required to analyze and pre-cache some dimension members and other cardinality queries, even if a user is doing a simple field request, turning one query into multiple structured query language (SQL) statements.

Background Information

We recommend using Parquet format with [Impala on Cloudera](#) distribution to improve your query performance speed. Using Hive as the underlying data source could result in delays due to multiple queries being issued during MapReduce processing.

Authentication Methods for JDBC

JDBC connections create one connection per user. If more than one is required, we recommended using IDBDataSourceService.

- No Authentication
- Kerberos Authentication
- User/Password Authentication

Pentaho Reporting and Interactive Reporting

Pentaho Interactive Reporting uses Metadata query language, where SQL statements are built based on the Table and Fields relationships.

Background Information

Pentaho Reporting combines many data sources such as JDBC, PDI, and others. Developers have full control over interaction with the data sources, SQL statements parameters, processes, data transformation, extensions, and mappings.

Authentication Methods

We focus on two main authentication methods:

- JDBC-related connections, which carry the same options as the Analyzer section.
- PDI Sources, which use shim configurations that are based on the underlying data source implementation. They can connect through JDBC into Hadoop's distributed file system (HDFS) sources where Kerberos tickets are used.

Streamlined Data Refinery

Streamlined data refinery (SDR) is a data integration process that works on user's request in the background while moving, segmenting, and changing data.

Background Information

Streamlined data refinery is a combination of technologies that include a front-end, built with CTools, that uses source metadata and controls for users to select their desired data. It includes an auto-template-based data model, with metadata for Analyzer and Dashboards to use with high performance.

Most of the processes in SDR require PDI as a data source that is used as a data processor and JDBC for Analyzer.

Authentication Methods

We focus on two main authentication methods:

- JDBC-related connections, which carry the same options as in the above Analyzer section.
- PDI sources, which use shim configurations that are based on the underlying data source implementation. They can connect through JDBC into HDFS sources where Kerberos tickets are used.

Data Integration Calls from the BA Server

BA Server includes a Big Data plugin that can be found in the Pentaho solutions system folder:

<<pentaho-solutions>>/system/kettle/plugins/pentaho-big-data-plugin

The configuration is described in [Pentaho Big Data Integrated Authentication Best Practice](#).

Related Information

Here are some links to information that you may find helpful while using this best practice document:

- [Pentaho Documentation – Components Reference](#)
- [Best Practice - Pentaho Analyzer with Impala as a Data Source](#)
- [Best Practices- Pentaho Big Data Integrated Authentication](#)