Pentaho Big Data Analytics

Comprehensive, unified solution that supports the entire big data lifecycle

Conectores Pentaho Big Data

Community VS Enterprise

Agosto 2014

Stratebi Business Solutions

www.stratebi.com

info@stratebi.com

Índice

1. Resumen

En Pentaho hay dos versiones: la versión Community y la versión Enterprise. Llevan muchos años en Pentaho realizando aportaciones a la comunidad Big Data y todos los conectores que tienen, a excepción de Splunk, se encuentran en la versión Community. La lista siguiente muestra un pequeño resumen de conectores y disponibilidad:

Conector	Versión Community	Versión Enterprise
Avro Input	✓	✓
Cassandra Input	✓	✓
Cassandra Output	✓	✓
CouchDB Input	✓	✓
Hadoop Input	✓	✓
Hadoop Output	✓	✓
HBase Input	✓	✓
HBase Output	✓	✓
MapReduce Input	✓	✓
MapReduce Output	✓	✓
MongoDB Input	✓	✓
MongoDB Output	✓	✓
Splunk Input	X	✓
BigData Instant View	X	✓

Así mismo la versión Enterprise de Pentaho incluye un wizard que permite a un usuario poco experimentado en Pentaho conectarse a una fuente de datos Big Data y explorarla en tres/cuatro pasos. No obstante es una característica secundaria para los profesionales de Pentaho que requieren realizar ETLs complejas con Pentaho Data Integration.

Este documento explora los conectores y el significado de la tecnología de la que se nutre.

2. Introducción

Big Data irrumpió en Pentaho hace un par de años y desde entonces están desarrollando diferentes proyectos en torno a esta categoría tecnológica. En Pentaho tenemos dos tipos de versiones: la versión Community y la Enterprise. Ambas se diferencian en que la versión Community es gratuita mientras que la versión Enterprise es de pago. Debido a este modelo de negocio que tiene Pentaho, tendremos diferentes conectores Big Data gratuitos y otros que sin embargo serán de pago.

3. Objetivo

Este documento tiene como objetivo esclarecer de que conectores Big Data dispone Pentaho para tener una base sólida. La versión que hemos estudiado en este documento de Pentaho es la 5.1.0.

4. Pentaho Community Edition

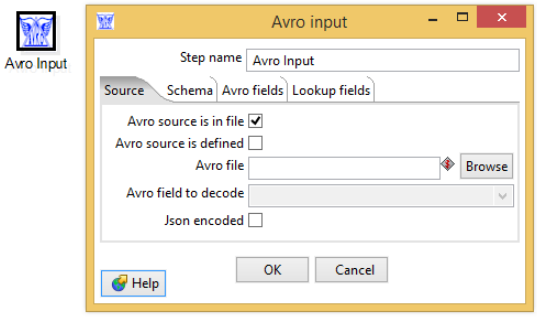
Pentaho CE es la versión gratuita de Pentaho. Esta incluye conectores en Pentaho Data Integration que son gratuitos.

4.1.Avro

Avro, es un sistema de serialización de datos. En los proyectos en Hadoop, suele haber grandes cuantidades de datos, la serialización se usa para procesarlos y almacenar estos datos, de forma que el rendimiento en tiempo sea efectivo. Esta serialización puede ser en texto en plano, JSON, en formato binario.

Con Avro podemos almacenar y leer los datos fácilmente desde diferentes lenguajes de programación. Está optimizado para minimizar el espacio en disco necesario para nuestros datos.

En Kettle tenemos Avro Input:



1 Avro Input

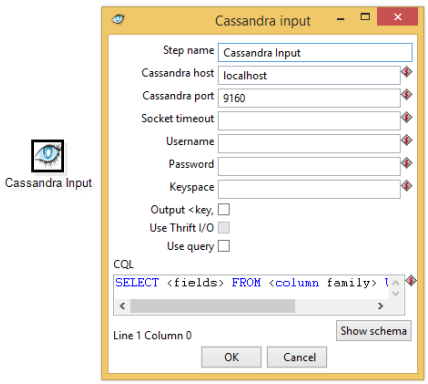
4.2.Cassandra

Cassandra es una base de datos distribuida, con la que podemos obtener un alto rendimiento en entrada/salida de datos y además es extremadamente escalable. Además está creada para ser tolerante a fallos, veremos que esto más tarde. Se dice que Cassandra es una solución de bases de datos post-relacional.

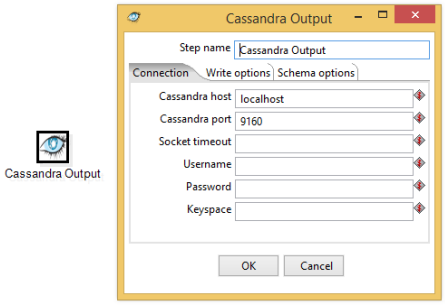
Digamos que Cassandra sirve como un datastore en tiempo real para aplicaciones online/transaccionales y como base de datos es muy buena para un alto número de lecturas/escrituras.

Es el producto de la unión de ideas de BigTable y de Dynamo, y se forjó finalmente como concepto en Facebook.

En Kettle tenemos tanto Cassandra Input (leer) como Cassandra Output (escribir):

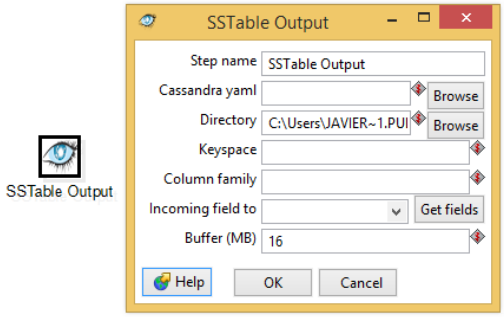


2 Cassandra Input



3 Cassandra Output

Por otra parte en Kettle tenemos otro paso llamado SSTable Output, que nos permite sacar una tabla al completo de Cassandra (se vuelca su estructura de ficheros de Cassandra).



4 SSTable Output

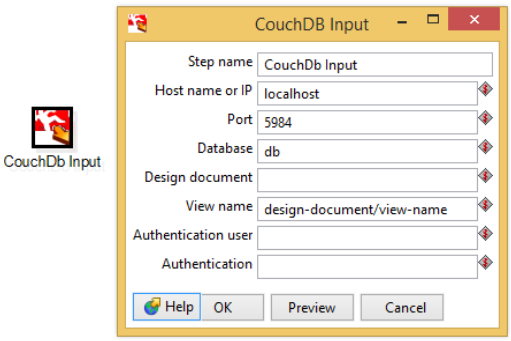
4.3.CouchDB

Es un gestor de bases de datos de código abierto, cuyo foco está puesto en la facilidad de su uso y en ser "una base de datos que asume la web de manera completa".

Se trata de una base de datos NoSQL que emplea JSON para almacenar los datos, JavaScript como lenguaje de consulta por medio de MapReduce y HTTP como API.Una de sus características más peculiares es la facilidad con la que permite hacer replicaciones.

CouchDB fue liberada por primera vez en 2005, transformándose en un proyecto Apache en 2008.

En Kettle tenemos CouchDB Input

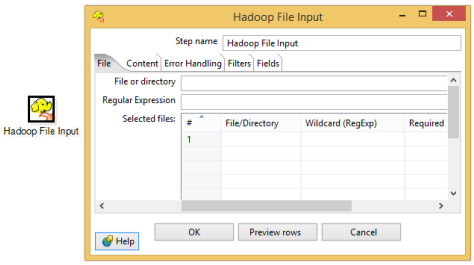


5 CouchDB Input

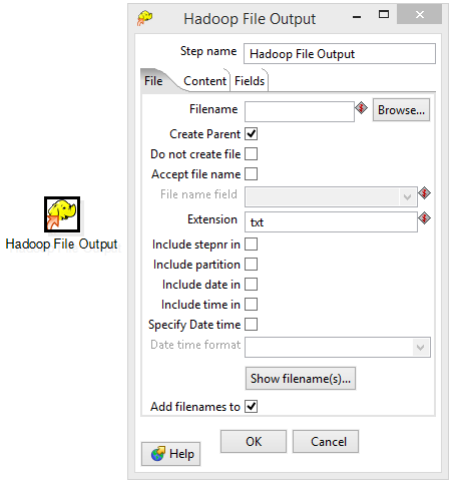
4.4.Hadoop

Apache Hadoop es un framework que permite el procesamiento de grandes volúmenes de datos a través de clusters, usando un modelo simple de programación. Además su diseño permite pasar de pocos nodos a miles de nodos de forma ágil. Hadoop es un sistema distribuido usando una arquitectura Master-Slave, usando para almacenar su Hadoop Distributed File System (HDFS) y algoritmos de MapReduce para hacer cálculos.

En kettle tenemos tanto Hadoop File Input como Hadoop File Output:



6 Hadoop File Input

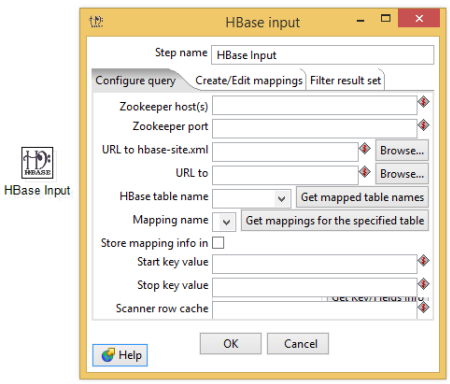


7 Hadoop File Output

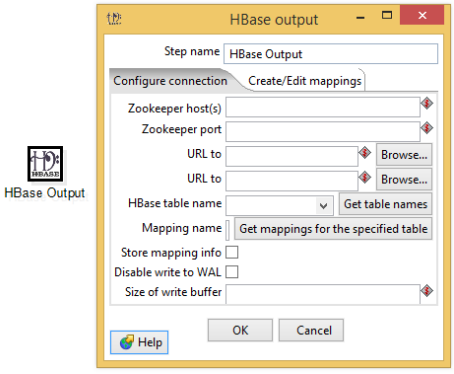
4.5.HBase

HBase, se trata de la base de datos de Hadoop. HBase es el componente de Hadoop a usar, cuando se requiere escrituras/lecturas en tiempo real y acceso aleatorio para grandes conjuntos de datos. Es una base de datos orientada a la columna, eso quiere decir que no sigue el esquema relacional. No admite SQL.

En kettle tenemos tanto input como output:



8 HBase Input



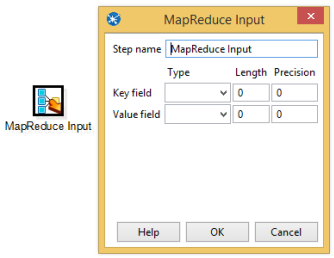
9 HBase Output

4.6.Map Reduce

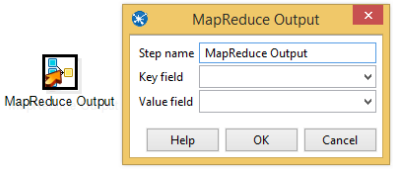
MapReduce es un proceso batch, creado para el proceso distribuido de los datos. Permite de una forma simple, paralelizar trabajo sobre los grandes volúmenes de datos, como combinar web logs con los datos relacionales de una base de datos OLTP, de esta forma ver como los usuarios interactúan con el website.

El modelo de MapReduce simplifica el procesamiento en paralelo, abstrayéndonos de la complejidad que hay en los sistemas distribuidos. Básicamente las funciones Map transforman un conjunto de datos a un número de pares key/value. Cada uno de estos elementos se encontrará ordenado por su clave, y la función reduce es usada para combinar los valores (con la misma clave) en un mismo resultado.

Un programa en MapReduce, se suele conocer como Job, la ejecución de un Job empieza cuando el cliente manda la configuración de Job al JobTracker, esta configuración especifica las funciones Map, Combine (shuttle) y Reduce, además de la entrada y salida de los datos.



10 MapReduce Input

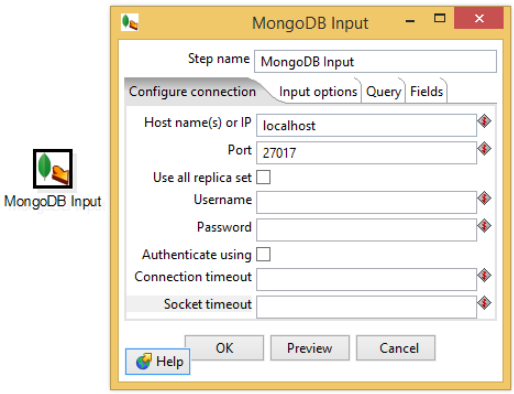


11 MapReduce Output

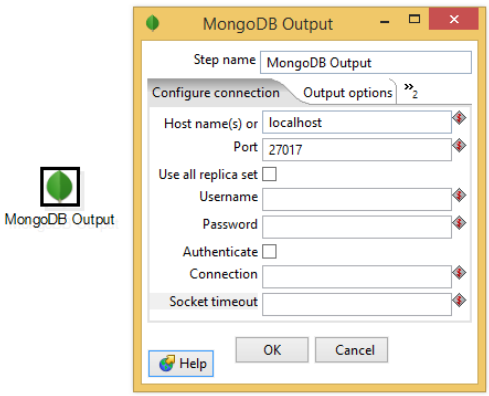
4.7.MongoDB

MongoDB forma parte de la nueva familia de sistemas de base de datos NoSQL. En vez de guardar los datos en tablas como se hace en las base de datos relacionales, MongoDB guarda estructuras de datos en documentos tipo JSON con un esquema dinámico (MongoDB llama ese formato BSON), haciendo que la integración de los datos en ciertas aplicaciones sea más fácil y rápida.

En Kettle tenemos tanto para leer como para escribir en MongoDB:



12 MongoDB Input



13 MongoDB Output

Hay varios sitios de referencia dentro de pentaho community de Big Data:

http://www.pentahobigdata.com/ecosystem/community -> es el frontal web en que nos muestran cómo pentaho puede conectarse con sistemas big data.

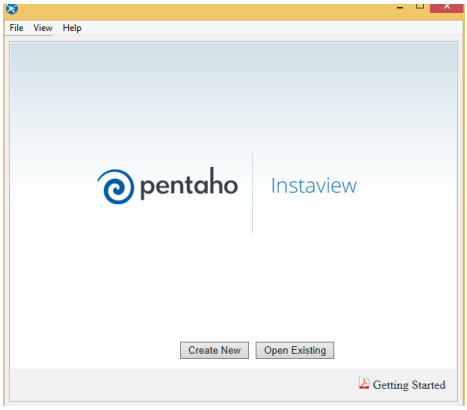
http://wiki.pentaho.com/display/BAD/Pentaho+Big+Data+Community+Home -> es la web de la comunidad, orientada a desarrolladores.

5. Pentaho Enterprise Edition

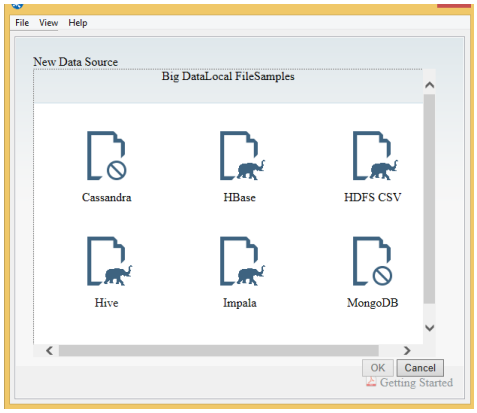
La versión Enterprise tiene las siguientes características adicionales:

5.1.Big Data Instant Analytics

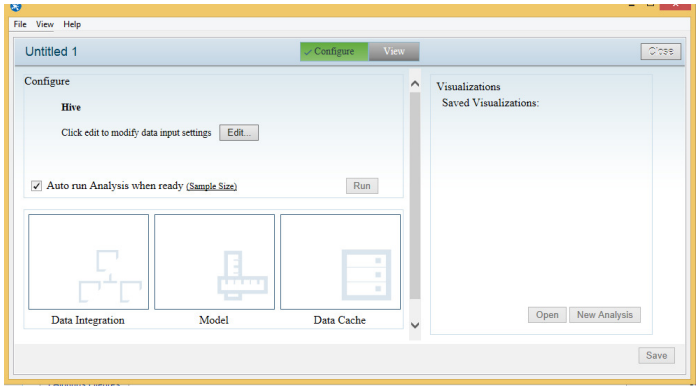
Es un wizard en el que puedes seleccionar la fuente "big data" de la cual deseas realizar una ETL. Es realizar pasos con Kettle a través de un wizard para acabar realizando un report. Tiene un conector que no trae la version community: Splunk. Esto se debe a que Splunk es de pago y en Pentaho han decidido que sea de pago el conector.



14 Instant Analytics Paso 1



15 Instant Analytics Paso 2



16 Instant Analytics Paso 3

5.2.Model Editor/Scheduler

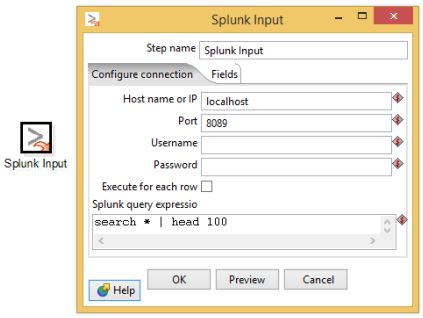
Esto es algo característico de la versión Enterprise de Pentaho, no sólo para temas de Big Data, aunque puede venir bastante bien para inspeccionar los data sources rápidamente a través de análisis e informes.

5.3.Splunk

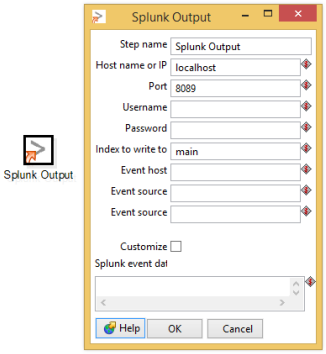
Splunk es un software para buscar, monitorizar y analizar datos generados por máquinas (Big Data) de aplicaciones, sistemas e infraestructura IT a través de un interfaz web.Splunk captura, indexa y correla en Tiempo Real, almacenándolo todo en un repositorio donde busca para generar gráficos, alertas y paneles fácilmente definibles por el usuario.

El objetivo de Splunk es hacer los datos de estas máquinas (este Big Data) accesible a toda la organización, permitiendo la identificación de patrones, realización de medidas, diagnosis de problemas y provisión de inteligencia (Business Intelligence) a cualquier parte del negocio. Splunk es una tecnología que escala a nivel horizontal usada para gestión de aplicaciones: security, compliance, business y web analytics. Splunk tienen más de 3,700 clientes licenciados en 74 países, incluyendo más de la mitad de las Fortune 100.

En la versión de Kettle Enterprise tenemos:



17 Splunk Input



18 Splunk Output