UNIVERSIDAD NACIONAL DE ASUNCIÓN
Facultad Politécnica
Proyecto Centro de Innovación TIC Paraguay-Corea

Curso Básico de Introducción a Big Data

Prof. Ing. Richard Jiménez, <rjimenez@pol.una.py>

Unidad I: Pentaho Kettle

HERRAMIENTA DE INTEGRACIÓN DE DATOS

*"**Pentaho Kettle** es una herramienta robusta, flexible y escalable para la gestión e integración de datos en entornos empresariales. Su facilidad de uso, junto con su capacidad para manejar diversas fuentes de datos, lo convierte en una solución ideal para proyectos de Business Intelligence, Big Data y Data Warehousing".*

¿Qué es Pentaho Kettle?

Pentaho Kettle, actualmente denominado Pentaho Data Integration (PDI), es una herramienta de integración de datos ETL (Extract, Transform, Load) utilizada para extraer, transformar y cargar grandes volúmenes de datos desde diversas fuentes hacia destinos específicos, como bases de datos, data warehouses o sistemas analíticos.

Es una solución de código abierto desarrollada por Pentaho, que posteriormente fue adquirida por Hitachi Vantara. Su flexibilidad y capacidad de integración con diversas tecnologías la hacen una opción popular para la gestión de datos empresariales.

Historia de Pentaho Kettle

Pentaho Kettle, actualmente conocido como Pentaho Data Integration (PDI), es una herramienta de integración de datos ETL (Extract, Transform, Load) que ha evolucionado significativamente desde sus inicios hasta convertirse en una de las soluciones más utilizadas en el ámbito del Business Intelligence y Big Data.

Orígenes y Creación (2001 - 2006)

Pentaho Kettle fue creado originalmente por **Matt Casters** en 2001 como un proyecto de código abierto llamado **KETTLE** (acrónimo de "Kettle Extraction, Transport, Transformation and Loading Environment"). La intención de Kettle era proporcionar una solución flexible y escalable para la integración de datos, evitando las limitaciones de herramientas propietarias.

El éxito del proyecto atrajo la atención de **Pentaho**, una empresa enfocada en Business Intelligence de código abierto, que adquirió Kettle en 2006 e integró la herramienta dentro de su suite de soluciones analíticas. A partir de ese momento, Kettle pasó a llamarse oficialmente Pentaho Data Integration (PDI).

Crecimiento y Expansión (2006 - 2015)

Tras su adquisición por Pentaho, PDI experimentó un crecimiento significativo, consolidándose como una de las herramientas ETL más populares en el ámbito empresarial. Se añadieron características avanzadas como:

- Soporte para procesamiento en entornos distribuidos.
- Integración con sistemas de almacenamiento masivo como Hadoop y NoSQL.
- Compatibilidad con servicios en la nube.
- Automatización y optimización del procesamiento de datos.

Durante esta etapa, Pentaho fortaleció su presencia en el mercado de Business Intelligence y Analítica de Datos, compitiendo con soluciones como Informatica PowerCenter y Talend.

Adquisición por Hitachi Vantara (2015 - Presente)

En 2015, Pentaho fue adquirida por Hitachi Data Systems, posteriormente integrada en Hitachi Vantara. Esta adquisición permitió mejorar la escalabilidad y capacidades de PDI dentro de un ecosistema más amplio de soluciones para Big Data y transformación digital.

Desde entonces, Pentaho Kettle ha seguido evolucionando con mejoras en rendimiento, soporte para arquitecturas en la nube y funcionalidades avanzadas de Machine Learning e Inteligencia Artificial. A pesar de la adquisición, PDI sigue siendo una herramienta open-source, con una comunidad activa que contribuye a su desarrollo.

Pentaho - Hitachi Vantara

Pentaho es una plataforma de inteligencia empresarial (BI) que ofrece herramientas para la integración, análisis y visualización de datos. Es conocida por su enfoque en procesos como minería de datos, generación de informes y análisis avanzado. Pentaho permite a las empresas optimizar la gestión de datos y tomar decisiones basadas en información precisa.

Sitio oficial: https://pentaho.com/

Hitachi Vantara, por otro lado, es una subsidiaria de Hitachi Ltd. que se especializa en soluciones de infraestructura de datos, análisis y gestión. Ofrece herramientas para optimizar el valor de los datos, desde almacenamiento hasta análisis avanzado, con un enfoque en la sostenibilidad y la innovación tecnológica.

Sitio oficial: https://www.hitachivantara.com/

Ambas empresas están orientadas a potenciar el uso estratégico de los datos en las organizaciones.

Conclusión

Pentaho Kettle ha pasado de ser un proyecto independiente para convertirse en una de las principales herramientas ETL del mercado. Su evolución ha estado marcada por la integración con tecnologías emergentes y su adopción en entornos de Big Data,

consolidándose como una solución clave para la gestión e integración de datos en empresas de todo el mundo.

Fundamentos de Pentaho Kettle

1. Arquitectura y Componentes

Pentaho Kettle se compone de varios módulos esenciales:

1. **Spoon:** Interfaz gráfica que permite diseñar transformaciones y trabajos ETL sin necesidad de programación.
2. **Pan:** Utilidad en línea de comandos para ejecutar transformaciones de datos.
3. **Kitchen:** Herramienta de línea de comandos para la ejecución de trabajos ETL.
4. **Carte:** Servidor liviano que permite ejecutar tareas ETL en entornos distribuidos y de alto rendimiento.

2. Principales Funcionalidades

- **Extracción de datos:** Soporte para múltiples fuentes como bases de datos SQL, archivos planos, XML, JSON, servicios web, APIs, y sistemas Big Data (Hadoop, Spark, etc.).
- **Transformación de datos:** Permite realizar limpiezas, agregaciones, normalización, enriquecimiento y combinación de datos.
- **Carga de datos:** Transferencia eficiente de datos hacia distintos destinos, como bases de datos, data lakes o sistemas analíticos.
- **Integración con otras herramientas:** Compatible con sistemas de BI, almacenamiento en la nube y herramientas de Machine Learning.
- **Automatización de procesos:** Mediante la creación de flujos de trabajo programados y ejecución en entornos distribuidos.

Casos de Uso Comunes

- Migración de datos entre bases de datos heterogéneas.
- Integración de fuentes de datos para análisis y reporting en Business Intelligence.
- Limpieza y enriquecimiento de datos antes de su almacenamiento en Data Warehouses.
- Procesamiento de datos en tiempo real para aplicaciones de Big Data.