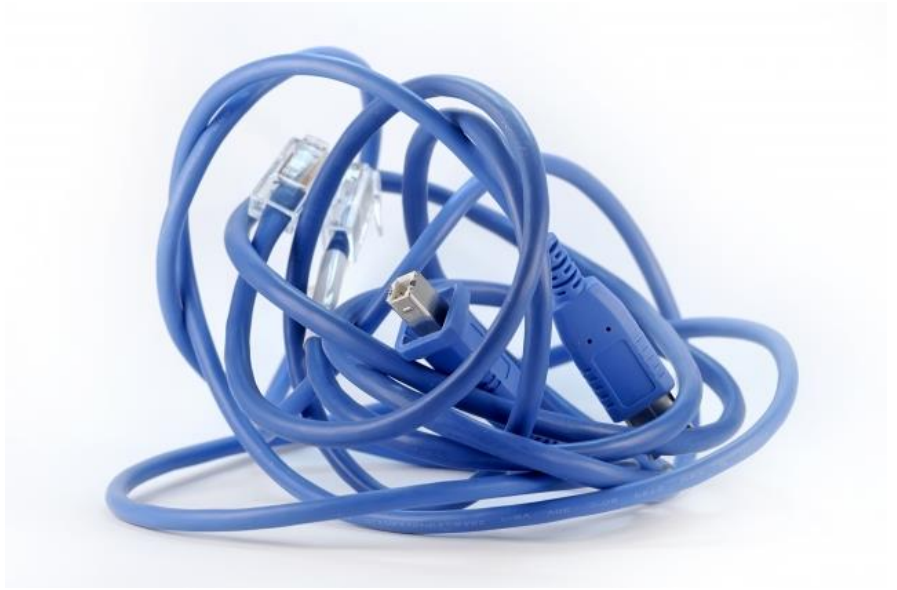
Almacenes y Minería de Datos

Prácticas

DEMO Talend Open Studio



Joaquín Ángel Triñanes Fernández
Instituto de Investigacións Tecnolóxicas
Joaquin.Trinanes@usc.es Ext: 16001 Externo: 881816001

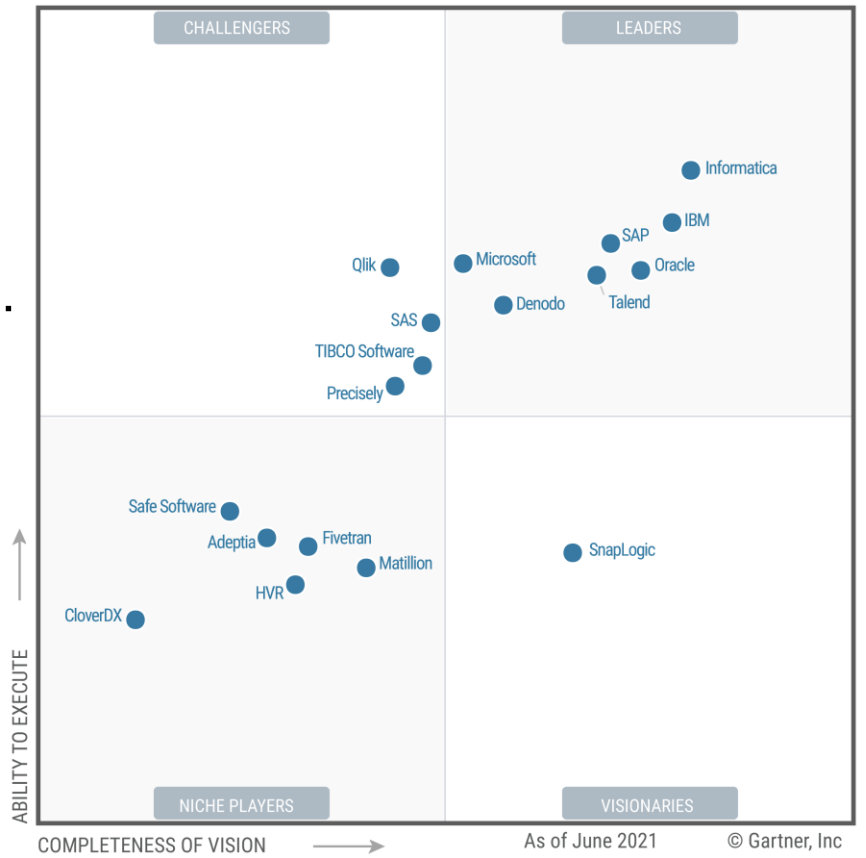
Sistemas Distribuidos

- Múltiples ordenadores que trabajan juntos de forma coordinada
- Ventajas:
  - Escalabilidad. Fácil de expandir horizontalmente para server a más usuarios
  - Baja latencia. Podemos situarlo geográficamente cerca al usuario
  - Tolerancia a fallos. Si un nodo falla, los demás pueden seguir operando. Se mantiene la disponibilidad aún en el caso de fallo.
  - Concurrencia. Podemos soportar cargas de trabajo muy superiores a las de los sistemas tradicionales
  - Compartir recursos. Pôdemos acceder a recursos (hardware, software) disponibles en cualquiera de los nodos.
- Retos:
  - Alta Complejidad - Gestión compleja
  - Seguridad - ACID vs BASE

DW

- Las fuentes para un DW pueden ser:
  - Sist. Operacionales - Inteligencia de negocio - Redes sociales
  - Datos históricos, …
- Los redirigimos a procesos ETL para poblar el DW
- Existen multiples herramientas ETL:
  - Microsoft SSIS - Stitch Data -Informatica -Talend OS
  - Pentaho DI - IBM DataStage
- Podéis crear vuestras propias herramientas en Python, C++,…
- ETL
  - Debemos asegurarnos de la calidad de los datos
  - Las mismas entradas deben generar las mismas salidas (idempotencia)
  - Debemos tener alertas y herramientas de monitorización
  - Cargar los datos de forma incremental

Figure 1: Magic Quadrant for Data Integration Tools



Source: Gartner (August 2021)

Talend

- Provee múltiple conectores para responder de forma ágil a las necesidades del negocio.
- Herramienta gráfica basada en Eclipse con capacidad de monitorización. Flujo de datos visual.
- Amplia comunidad de desarrolladores. Buen soporte de la comunidad.
- Provee control centralizado para desplegar en multiples nodos
- Conexión a múltiples fuentes, incluyendo sistemas legados
- Manejo de multiples formatos.
- Compatibilidad con múltiples protocolos
- Integración con múltiples BBDD
- Responde a necesidades ETL analíticas y ETL de integración operacionales

Talend Open Studio

- Herramienta gratuita
- Comunidad muy active
- Job (trabajo): Componentes relacionados entre sí
- Puedo en caso de error (ifOk/ifError). Cada componente tiene su propio OnComponentError.
- Puedo generar estadísticas (tStatCatcher) y usar ficheros log (tLogCatcher)
- Puedo modificar el Código.
- Puedo generar binarios

Talend

- Talend OS es una herramienta ETL de código abierto para la integración de datos y Big Data. Requiere Java.
- 64-bit. Recomendamos al menos 3GB RAM y 5GB disco
- Bajamos Talend Open Studio for Data Integration de https://www.talend.com/download/
- https://www.talend.com/products/data-integration-manuals-release-notes/
- Ayuda- Quick Tour
- Izquierda: El repositorio donde encontramos los datos y metadatas que van a ser usados por los Jobs
- En el área de trabajo tenemos la vista de diseño y el código asociado.
- A la derecha tenemos la paleta de conectores

Talend

- Vamos a importar y exportar datos
- Llevamos a cabo el mapeado usando tMap. Permite varias operaciones:
  - Multiplexado y desmultiplexado
  - Transformar cualquier campo
  - Concatenar campos e intercambiar objetos
  - Usar restricciones para filtrar datos
  - Operaciones para rechazar datos
- tMap usa conexiones entrantes para crear esquemas de entrada en el Map Editor
- Puedo crear un join desplazándome de una tabla a otra
- Puedo usar Java para transformar los datos
- Podemos crear Jobs que contienen varios subjobs

Talend

- Ejercicio:

Partimos de los ficheros del Rexistro de buques 2019-2021 en formato csv y XML. Vamos a convertir el fichero csv en XML y viceversa, comparándolos. Podemos ordenarlos por uno de los campos.

Importamos el csv a un esquema normalizado en PostgreSQL, rechazando en el proceso los datos que tengan una eslora menor de 4 metros y quedándonos con los datos de un puerto.

Vamos a calcular el total de arqueo por provincia usando talend y comprobamos los resultados comparándolos con PostgreSQL

Ahora vamos a utilizar esa BBDD como entrada para otro trabajo en el que vamos a importarlo a un esquema desnormalizado que simule el fichero original.

Vamos a ejecutarlo desde la línea de comandos