Lectura interesante: https://www.alooma.com/blog/open-source-etl-tools-comparison

Requerimiento general: herramienta de ETL para conectarnos a bases de datos (principalmente MySQL y PostgreSQL) y a archivos de Excel, extraer información, transformarla y depositarla en un warehouse de datos bajo un horario programado.

Casos de uso:

1. Tenemos una base de datos MySQL que está en producción para el sistema de una clínica de salud. Necesitamos extraer datos de esta base de datos en dos horarios distintos: un monitoreo “semi-tiempo real” y un historial cada 24 horas. Necesitamos transformar la información y depositarla en tablas “planas” en un warehouse para luego conectarle Metabase al warehouse.
2. Tenemos varios potenciales clientes que utilizan SAP Business One (Microsoft SQL Server – Windows o SAP Hana – Linux). Necesitan que extraigamos, transformemos y almacenemos datos en un warehouse que creemos para ellos. Luego conectaríamos Metabase al warehouse. Otros clientes también utilizan SAP By Design.

**Otros requerimientos:**

1. Open source (por lo menos con una versión “community” con acceso gratis) (crítico)
2. Conexiones a bases de datos relacionales (MySQL y PostgreSQL como arranque) (crítico)
3. Capacidad de instalarlo en un servidor web propio para que los datos no pasen a través de un servidor que no es propiedad de Kemok (crítico)
4. Capacidad de instalarlo en un servidor Linux (crítico)
5. Capacidad de establecer el horario en el que se ejecutarán las tareas de ETL (crítico)
6. Capacidad de establecer distintos horarios para distintas tareas (por ejemplo, extraer datos de inventario una vez al día pero datos de ventas varias veces en el día) (crítico)
7. Cuenta con documentación? Preferiblemente una documentación clara? (Crítico)
8. Capacidad de definir si se busca extraer toda la data de una tabla o únicamente la data nueva (útil)
9. Facilidad de implementación y administración mediante una interfaz gráfica (útil)
10. Con una comunidad activa en foros (por ejemplo Stack Overflow) y en el código fuente (por ejemplo evaluando el número de commits y contribuidores en Github) (útil)
11. Capacidad de conectarse a otros servicios web o plataformas que no sean bases de datos relacionales (ideal)
12. Capacidad de construir cubos analíticos para visualización de datos (ideal)
13. Acceso al código fuente (ideal)
14. Monitoreo de procesos (éxito, fallos, tiempos de ejecución, etc.) (ideal)

**Puntos que vale la pena evaluar:**

1. El lenguaje de la herramienta y el lenguaje que utilizaríamos para programar sus actividades (Java? Perl? Según hemos encontrado)
2. Reviews / comentarios en línea: cuál ha sido evaluado por otras personas?
3. Sitio web del proyecto (ideal si contiene algún roadmap de hacia dónde se dirige el producto)
4. Bajo qué tipo de licencia utilizaríamos el software? Nos obliga esa licencia a hacer público nuestro software en caso usemos la herramienta? (Ideales: MIT, Apache. Otros: LGPL. Otros: GNU GPL)
5. Hay vulnerabilidades de seguridad conocidas respecto a la herramienta?
6. Qué tan establecido / desarrollado está el proyecto? Ya tienen por lo menos una versión estable?
7. Qué tan frecuentemente publican nuevas actualizaciones?
8. Qué estándares de manejos de datos utiliza la herramienta?
9. Hay empresas (o una empresa en específico) que ofrece soporte pagado?