OPEN SOURCE ALTERNATIVES TO SSIS

Microsoft SSIS or SQL Server Integration Services is a data migration and integration tool that comes as a part of Microsoft SQL Server database,it is generally used to extract, integrate, and transform data. In a nutshell,SSIS is an Extract, Transform and Load (ETL) solution.SSIS is actually a successor to another program called **Data Transformation Services** (**DTS**) that existed as a part of SQL Server 7 and 2000.Some of the key tasks that can be performed using SSIS are-

- Analyzing
- Cleansing
- Loading
- Transformation
- Aggregation
- Merging
- Extraction

With its ease of use and versatility ,SSIS is a widely popular choice for a multitude of organisations. However, like most of the data integration and ETL tools in the market it requires licensing and involves some hidden costs.

With the advent of data explosion,ETL tools have and will continue to be in demand.With this current situation in play, It would be beneficial to explore some open source ETL tools in the market.

Talend Open Studio

Talend Open Studio is a free open source version of Talend's commercial set of tools. It provides a rudimentary set of features for managing data, and can work very well as an ETL tool for deriving business intelligence.Key capabilities of Talend's Open Studio include-

- Facilitate connections to packaged applications like ERP and CRM, databases, mainframes, files, Web Services, to accommodate data from a wide variety of sources and format.

- Writing to data warehouses, data marts, OLAP applications - for analysis, reporting, dashboarding, scorecarding, and more.
- Built-in advanced features for ETL like string manipulations, Slowly Changing Dimensions, automated lookup handling, bulk loads support, etc.

Even Gartner rated Talend as one of the emerging leaders in the data integration sphere.

Pentaho Data Integration(Kettle)

Pentaho is one of leading business intelligence platforms that make it possible for the organizations to easily access data, prepare, and analyze through user friendly and intuitive interfaces.It offers exceptional data integration, OLAP, data mining, reporting, and ETL (Extract, Transform, and Load) capabilities.Pentaho kettle is an open source tool and comes as 2 editions-community and subscription.

It can be used for

- Data migration between different databases and applications
- Population of Data warehouses containing changing dimensions
- Integration of real-time ETL as a data source
- Data cleansing with steps

It allows developers to easily create a data pipeline using a simple interface creator without necessitating the use of code. It utilises a metadata-driven approach and a shared repository that enables remote ETL execution in a quick and an efficient manner.

Apache Airflow

Apache Airflow is a platform that allows users to author, schedule, and monitor workflows with the use of code(programatically). It is completely open source and is especially useful in building complex data pipelines. It was initially developed to overcome the issues that come with long-running cron tasks and hefty scripts, but ever since it has grown to become one of the most powerful open source data pipeline platforms out there.

Some of the benefits of Apache Airflow are-

- **Dynamic:** You can perform any operation on Airflow that can be performed on Python..
- **Extensible:** Airflow has a number of readily available plugins for interacting with most common external systems. Plugins are also customisable and can be created according to need
- **Scalable:** Airflow can run upto thousands of different tasks per day.

KETL

KETL is a platform ready ETL software allows the handling of complex manipulation of data quickly and efficiently while leveraging an open source data integration platform.It is designed for the development and deployment of data integration efforts.KETL features are noteworthy and in fact, compete with major commercial products available today.

Some key features include:

- Facilitates the integration of security and data management tools.
- Highly scalable and can be extended across multiple servers and CPU with any volume of data.
- Does Not necessitate the presence or usage of third party schedule, dependency, or notification tools.

The aforementioned tools are just some of the open source tools available in today's data integration/ETL sphere with new technologies emerging faster than ever. Open source ETL tools are slowly but steadily gaining popularity among organisations in today's data driven world.Open source tools provide a lot of the required capabilities at zero or low costs when compared to their licensed counterparts and are hence worth consideration.

CITATIONS

https://www.astronomer.io/guides/intro-to-airflow.

https://www.comparitech.com/net-admin/what-is-microsoft-ssis/#Advantages_and_Disadvantages_of_SSIS.

https://www.datasciencecentral.com/profiles/blogs/10-open-source-etl-tools.

https://www.spec-india.com/blog/pentaho-data-integration-kettle.

“KETL - production ready ETL platform.” *LinuxLinks*, https://www.linuxlinks.com/ketl/. Accessed 10 December 2021.

“Talend Open Studio for Data Integration download.” *SourceForge*, https://sourceforge.net/projects/talend-studio/. Accessed 10 December 2021.