A comparative study of leading ETL tools used in Asia Pacific Region

IS 549 – Data Warehousing
Spring 2021

Submitted By
Natapong Sornprom
Data Science

Submitted to
Professor Sultan Raziuddin
College of Computing and Digital Media
DePaul University

Abstract

Today's business environment demands companies to manage many, yet complementary, information systems. ETL tools become extremely beneficial in data integration and data warehousing in this digital age. The data warehouse receives input via ETL. Extraction, Transformation, and Loading (ETL) are three terms used to describe their essential processes. This tool is used to move data from one system to another. On the other hand, Enterprises have a difficulty in selecting the ideal ETL process and ETL tool for their business, since one incorrect step or choice can result in a slew of financial and time losses, not to mention the amount of arduous effort put in by employees. The business can pick from a range of ETL solutions, but without investigating or understanding their features, the decision-making process will be distorted. In this study, I attempted to give a comparative assessment of some of the most extensively used ETL technologies in the Asia Pacific area. This study is to familiarize people with the product's primary benefits and disadvantages.

1. Introduction

To enhance leadership abilities, every organization needs relevant data to grow and prosper. The concept of business insight based on employing a vast volume of company data to obtain important data comes to be. Business intelligence (BI) is used to assist organizations in managing their data. For that, a company would require a data warehouse, which would have to store all of the data from numerous sources and types of data. A lot of people are interested in the BI because of the recent trends. The above statement accurately describes its many impacts on the company, including boosting the value and performance of the company. Because BI managers and IT teams stay on top of BI's progress, the rest of the organization does as well. Many corporations believe that a strong digital strategy has a substantial influence on profitability. By not using business intelligence tools, businesses put themselves at a competitive disadvantage. Business intelligence strategy is regarded to be needed.

Making use of many data sources, bringing them together in a targeted information distribution center, and processing them all for advanced structure are crucial to implementing a wide-ranging strategy. A data warehouse is where the data is stored and organized. A big part of Business Intelligence is using data warehouses to help make better decisions based on the information they provide. Previous databases had one common problem: data had been stored in a way that was ideal for reporting and other forms of analysis, but not the greatest for queries. By combining hardware and software efforts, the data warehouse is used to acquire data as well as to deliver it to customers in the manner they choose. Once the data is in a data warehouse, it is feasible to utilize it to create data marts, which have a distinct purpose for handling certain organizational data. It will also allow IT to monitor security, while making reporting simpler for end users and more information accessible for usage. A single database regulates access to users and permissions instead of administering security for an extended number of databases. The Extract, Transform, and Load (ETL) process involves the acquisition of different data, the formatting of that data, and the importation of it into a database or data warehouse. ETL (Extract, Transform, Load) activities use up to 80% of the time devoted to BI projects. Great speed is

essential to handle large amounts of data and keep an up-to-date database. The report maker will be able to use different reporting technologies that enable consumers to have more control over their own data. With the birth and rise of the data warehouse, it has become an integral part of Business Intelligence.

In order to continue to develop, the industry as a whole is witnessing numerous and diversified submarkets converge at both the vendor and technological levels. Desire drives this. More and more organizations are seeing the multitude of data integration problem kinds, which are matched by different architectures and delivery methods, which is shaping a greater diversity of data delivery systems. Also, the merchants' habits impact the business. This Vendor Market Report also highlights the expansion of vendors in certain data integration submarkets as well as acquisition activity, which brings suppliers from diverse submarkets together. A high-quality data integration solution that uses common design tooling, metadata, and runtime architecture might lead to a steady demand for other data integration solutions that use other data integration methodologies. Although there are various ETL technologies, each provides a set of features exclusive to it. Pricing and learning curves also vary, and varied degrees of skill are required to use them. Deciding on the right equipment for any certain job entails several considerations. Practice, however, demonstrates that end-to-end tools and technologies are not widely understood by end users, and this results in mandatory training to aid users in finding them. The study's aim is to evaluate seven prominent ETL tools in various situations to determine which ones work best. In this paper, the features of various tools are compared. The study results may help consumers in picking out the most suitable ETL software to their places.

Here are some of the contributions presented in this report:

- An equal number of technical and non-technical criteria were utilized to compare the seven ETL tools. Other relevant criteria include, but are not limited to, such aspects as system features, platform requirements, training time, simplicity of use, and others.

- Offering a collection of ETL software alternatives

The remaining sections of the paper are organized as follows. The ETL process is addressed for a brief time, after which the discussion shifts to various ETL technologies utilizing generic criteria, such as the level of technical and non-technical expertise required. ETL (Extract, Transform, Load) tools are required for BI (Business Intelligence) evaluation. An organization's successful attainment of its strategic objectives relies on using the suitable ETL technology.

2. Background

There are numerous ways and technologies to pick when it comes to business intelligence. The basis of business intelligence includes online analytical processing and enterprise information systems. These components help to make decisions promptly. Both of these components may be utilized to develop a BI application potentially. Analytical BI includes a series of ideas and technologies that allow collections of data with either a model or a method. In Figure. 1 you can get an overview of the BI ideas and technology with their process focused goals and their direction.

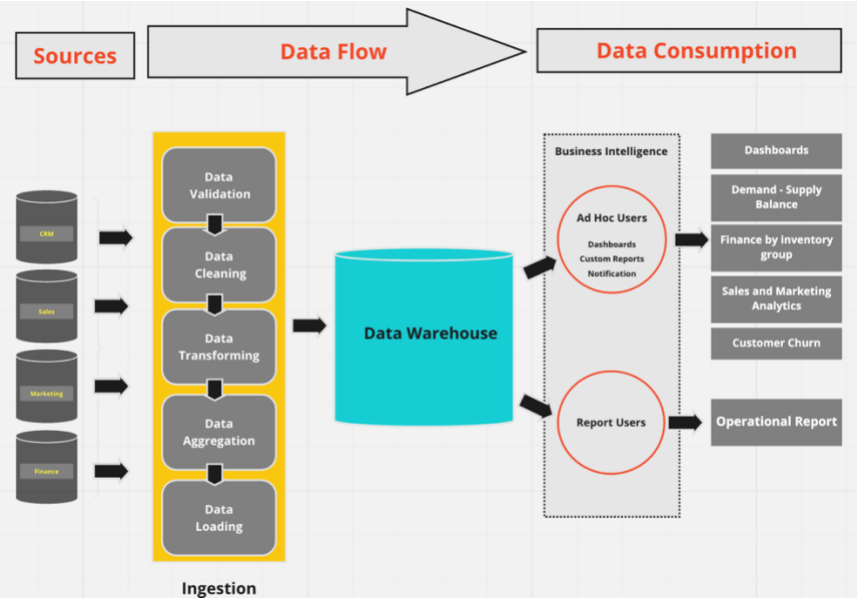

Figure 1: Data flows from ingestion to BI and consumption

2.1 Classification of ETL tools

a) **An ETL process coded by hand:**

Prepared in-house using Perl, COBOL, C, and PL/SQL, for extraction, transformation, and loading into a variety of destination databases. Due to the complicated nature of these applications, documentation was extensive and difficult. A developer must employ many programming languages to do the extract, transform, and load (ETL) job. This developer may employ Perl scripts, along with SQL Loader and PL/SQL Bulk to complete the process. As stated previously, hand coded ETL tools have the benefit of being able to manage all of the information that is generated on their own and giving developers total flexibility to use those tools in new ways to meet new demands. On the other hand, this imposes additional limits, because these systems must be changed often to handle enormous amounts of new data created from several different sources. Additionally, changes done at the metadata table need to be adjusted, which means modifications need to be made to hand-coded tables as well. Finally, the hand-coded ETL often operate at a snail's pace due to their single-threaded nature, but current tool-based tools operate quickly due to their multithreaded structure and run-on fast engines.

b) A Tool Based ETL:

Many companies built these tools so that businesses might purchase them instead of hand-coding, with all of the costs and slowness. Extract, transform, and load (ETL) programs evolved from extraction workstations to a target database and now offer fully working GUIs with additional features and performance. Innovative software solutions with today's ETLs allow the extraction and transformation of diverse data sets, integrate databases or flat files, enable multidimensional designs, surrogate key creation, varied transformation capabilities, and are native on both SQL and OS platforms. Even though the two data warehouse metadata repositories are distinct, it's possible for an organization to have various internal metadata repositories. They rid ETL operations of the complications that stem from the costs of building and sustaining sophisticated routines and transformations. Additionally, these applications are helping developers by providing user-friendly GUIs, enabling them to do their work without going through any

training. In addition to these functions, the tools include capabilities such as monitoring, scheduling, bulk loading, and incremental aggregation.

To expand on this, current ETL tools may be categorized into four main groups.

- **Pure ETL Tools:** The database and the Business Intelligence tool with which it will be utilized have no effect on these items. A business strategy focused on profit maximization and eliminating profit loss as much as possible will keep a company from incorporating new capabilities and transitioning databases, regardless of integration processes.
- **Database Integrated:** These software packages are accessible when you purchase the database software and part of the capability is included inside the database software, which incorporates the extract, transform, and load (ETL) process.
- **Business Intelligence integrated:** This is the same supplier of the BI software, which is what these items are made by. BI applications may support a wide range of use cases, and in many cases, they may be integrated with other software products that are offered separately by the same vendor.
- **Niche Product:** This group consists of items that do not fall into any of the aforementioned divisions but are nonetheless relatively capable of carrying out the related ETL processes.

2.2 The ETL processes and common concepts

An important part of the Data Warehouse architecture is extraction, transformation, and loading (ETL). Data extraction is done in the process, after which data is processed to meet business requirements and loaded into the data warehouse. Even if you are working with snippets, an ETL process may be written in any computer language. There are several ETL solutions on the market that may help a company choose one that best suits their needs and requirements. These tools have grown through time, and now they include a variety of additional features. As a result of these advancements, users will be able to do things

like data profiling, data quality control, monitoring and cleaning, real-time and on-demand data integration in a service-oriented architecture, and metadata management. Moreover, the functional needs of an enterprise data warehouse are also adjustable in ETL technologies.

- **Extraction:**

Data extraction is the initial phase in the ETL process. Therefore, it places special emphasis on gathering data from disparate systems. These sources are called source systems because they might be internal, external, structured or unstructured, which means they may be of any sort. In addition, several systems may be used as inputs, such as mainframe programs, flat files, ERP programs, relational databases, non-relational databases, CRM tools, or even message queues. These multiple sources may store their data in a variety of ways, such as distinct internal representations. This makes Extraction a challenging task. There should be an extraction tool which is capable of learning how different data storage types are handled, communicative databases, reading and understanding multiple file formats utilized in an organization, better at interpreting documents in diverse forms and prioritizing the most important data extraction, and bring it into DW.

- **Transformation:**

By carrying out an extensive sifting, removal, and clean-up of information, this process makes certain that all the information in question is ready for inspection and will also ready any extra information which might be accessed by using a query table or administrator, or which will be blended with other information to yield a completely revamped state. The change stage involves the process of determining if information is good, followed by a process of deleting any questionable data (where possible). Other techniques commonly used in the course of change include arranging, separating, clearing the copies, institutionalizing, and translating.

- **Loading:**

After all of the retrieved data has been transformed and cleansed, it is loaded into fact and dimension tables in the data warehouse for usage in various analytical applications. It is done often to keep data structures from accumulating in a pile. It must be done one

of two ways: (1) Fetching the current operating data from the database, (2) If changes have happened in the operational database, get the latest database updates. While many prefer to employ incremental loading, it may be claimed that incremental loading is the ideal way to data warehouse refreshing since it typically decreases the amount of data that has to be extracted, converted, and loaded by the ETL system. tasks that need incremental loading of data have to have access to the data source from the last load, which has been altered since then. in order to better carry out these changes, so-called Change Data Capture (CDC) procedures at the source can be used, assuming they are available. The requirement to do extra loading of data possibly necessitates ETL jobs.

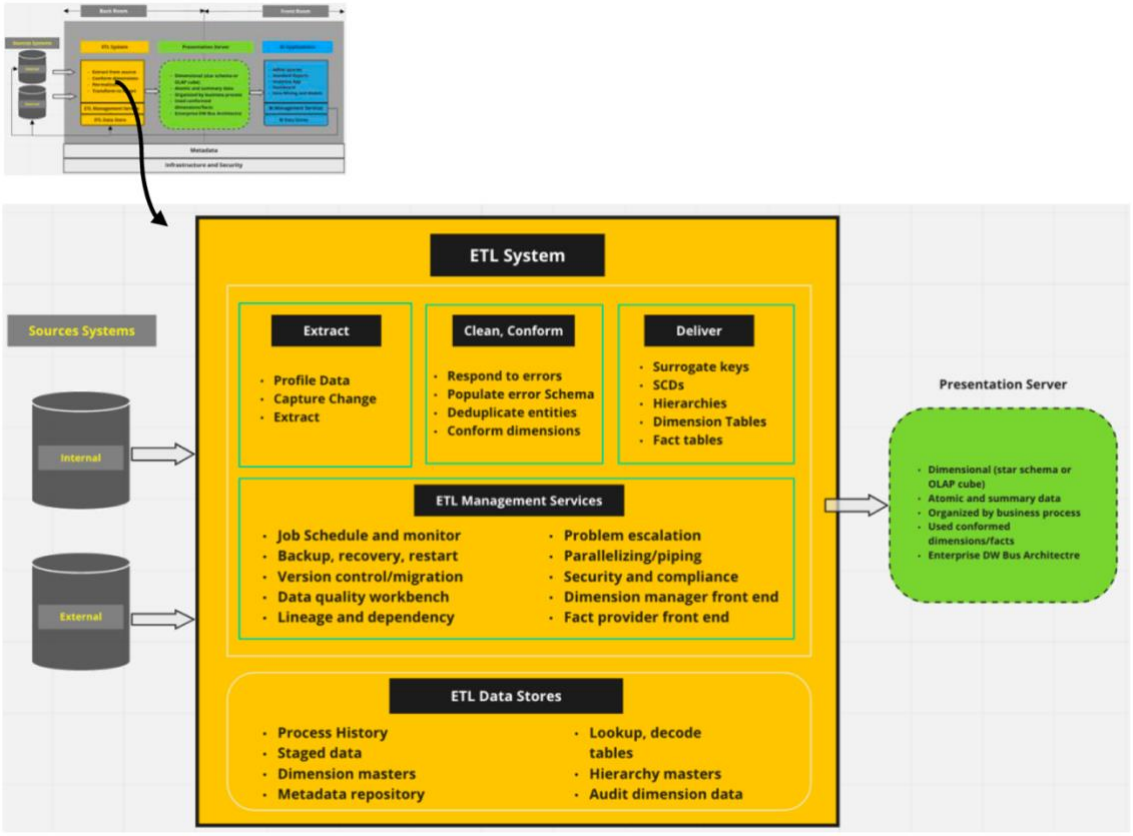

Figure 2: ETL System

2.3 The leading ETL tools

A number of ETL tools are available for data integration and transformation; a few are listed below.

I. Microsoft SQL Server Integration Services:

This is a powerful and versatile component of the Microsoft SQL Server database software that enables you to do a broad variety of data transfer, data transformation, and data integration operations. It may be used to manage and update SQL Server database multidimensional cube data in order to solve difficult business challenges. SSIS was initially offered alongside Microsoft SQL Server 2005 and eventually replaced Data Transformation Services, a capability of SQL Server since Version 7.0. The efficient handling of all forms of data in the short space of time offered by SSIS may be attributed to the fact that it utilizes many different source types and numerous destination types. This SSIS has a broad array of transformation objects prebuilt in, such as aggregate, auditing, cache transformation, data conversion, data mining query processing, dimension processing, export and import column, fuzzy grouping, pivot, row sampling, and term extraction. The modular, object-oriented paradigm supports the saving, loading, and usage of pre-programmed packages for execution of code. SSIS is frequently used in corporations with big data sets to handle that data. Such data processing requires a Microsoft Server OS and system software, which therefore adds hefty maintenance and OS costs.

II. Informatica Power Center:

This enterprise data integration platform can be described as an embedded data exchange component working as a unit for business-to-business data exchange, cloud data integration, complex event processing, data migration, data masking, data quality, data replication, data virtualization, master data management, and ultra-messaging. Basically, the components of Informatica Power Center include three primary groups. (1) Informatica Power Center Client Tools: This is the development software, which is installed at developer ends to help throughout

the mapping process. (2) Informatica Power Center Repository: The repository is where all the metadata for your application is stored, and that is where all of the Informatica tools are based. It functions as a data inventory. (3) Informatica Power Center Server: The server is in charge of the processing of all of the data, as well as putting this data into the system.

III. Oracle Data Integrator:

Oracle Data Integrator delivers a single, complete data warehousing solution, and as part of an information-centric architecture or as part of a SOA framework, it is particularly useful in a business intelligence setting. To make it easier, it also incorporates all the disparate parts of data integration: data mobility, data synchronization, data quality, data management, and data services. The active integration platform offered by Oracle Data Integrator (ODI) supports a wide range of integration types, including data-based, event-based, and service-based. The ODIs primary unifies disparate systems and unifies their disparate systems by increasing the amount of data that they handle efficiently, and by real-time processing of data using their advanced Changed Data Capture (CDC) framework. These functions also apply to the Oracle SOA Suite. The expanded version of this item ensures the consistency and accuracy of data by including sophisticated data integrity management capabilities.

IV. Talend Open Studio:

Data Integration and Big Data may be achieved using Talend Open Studio, which is completely free and open source. It is a Java-based tool that both developers and job creators use. To build ETL or ETL Jobs, you only need to drag & drop components and link them together. We won't have to write a single line of code since the program will generate Java code for the task. The following are some key features to note with Talend Open Studio:

- All of the 900 components required for data integration and synchronization are already built-in, with in-built connectors that allow the tasks to be

converted to Java source code, along with all the capabilities needed for data integration and synchronization.

- Because the tool is absolutely free, we may benefit from large financial savings.
- Terms of service is widely accepted in the business world since numerous major firms have implemented it for data integration in the previous 12 years.
- Data Integration has a well-connected network in the Talend community.
- New features continually being added to the tools, and the documentation is simple to understand and is well designed.

V. Pentaho Data Integration:

PDI is a component of the Pentaho Open-Source Business Intelligence package. It comprises software to manage data warehouses, data integration and analysis tools, management software, and data mining tools. Many users have found that using Pentaho Data Integration is simple and quick to learn. With PDI, the development focuses on stating what actions to do, and not on explaining how to accomplish them. Pentaho allows administrators and ETL developers design their own data manipulation processes using a graphical user interface that offers an almost limitless number of choices and does not need any coding whatsoever. A centralized, shared repository known as PDI helps remote ETL execution, lowers costs, and enhances cooperation.

VI. IBM Information Server:

An information server is a unified software platform made up of a collection of key functional modules that allow enterprises to combine data from different sources and offer reliable and full information at the time and in the format necessary. In the same way that an application server is a software engine that distributes programs to client computers, an information server distributes consistent data to consuming programs, business processes, and portals. It enables a high level of flexibility, while at the same time following a clear direction in regard to the market with an identified target market focus. Overall, the clients of the

Information Server are pleased with the levels of pleasure that they receive and the number of different projects the organization offers. Since it is so simple to use, but as a result, it is quite difficult to use because it is burdened with gigabytes of data and version 8.x demands a lot of processing power.

VII. Amazon Web Services:

Although it is just new and largely unproven, AWS Glue might make it far easier for business teams to work together, as data extraction, transformation, and loading (ETL) is a daunting task that frequently hinders teamwork. In this way, the super-fast, parallel processing of Apache Spark is combined with the simple data structuring of Hive metastore stores to provide an unparalleled link across diverse AWS resources. AWS Glue has been under development for two years, and it has had several important features being added to it on a frequent basis throughout the years. You should bear in mind that glue is not a silver bullet, and it may not function as easily or unobtrusively as you would want. Nevertheless, with the correct technique and situation, it may be a viable solution. The goal of glue is to deliver processing and data setup in one location, with as little of a foundation as possible. The Glue data catalog provides an additional means of integrating sources such as file-based and conventional data with Glue activities, such as the capability to identify existing schemas using crawlers. Connecting our data catalog to the Hive metastore allows us to share the information in AWS Athena with other AWS Athena users, who we consider existing customers. Python shell, Spark, or most recently, Spark Streaming, a beta feature that requires you to turn it on, are the tools that Glue can utilize to handle the type of massive data outlined above. The level of difficulty for a given work varies greatly depending on the language used to write it.

3. Methodology

Though there are numerous tools on the market, the ETL tools used within this comparison research are only the big players which are widely used in the Asia Pacific area according to Gartner review in 2021. In this way, only equals should be evaluated. To find out which product is best, the goal must be to look for solutions that are effective in every area of business. I explored several sources to get information on a universal one, therefore I investigated academic journals, articles, books, and also reports that are produced from time to time to eliminate any prejudice. This has resulted in the development of the general criteria that will help me assess ETL products. The criteria are organized into two separate groups: one focused on their technical abilities and the other on their general perspectives. I think this offers benefits to a variety of backgrounds, from people who are interested in technology and business to those who focus on each field alone.

Table 1: Technical Criteria

No	Criteria
1	Mechanism Feature
2	Data Quality
3	Scheduler
4	Grid Application
5	Usability and Reusability
6	Tool base
7	Complier
8	GUI

Table 2: Non-Technical Criteria

No	Criteria
1	Establishment
2	Latest Version
3	Platform Support
4	Big data Support
5	Source as joined table
6	Ease of use
7	Product Feedback
8	Cost

Table 3: The Products that are evaluated by above criteria

No.	Vendor	Product
1	Microsoft	SQL Server Integration Services (SSIS)
2	IBM	Information Server
3	Informatica	Informatica Power Center
4	Amazon Web Services	AWS Glue
5	Talend	Talend Open Studio
6	Hitachi Vantara	Pentaho Data Integration
7	Oracle	Oracle Data Integrator

3.1 ETL tools summary

The summary below is a representation of the survey that I took, in which I gathered information on the advantage and downside of Tool. Although, it should be noted that they will be excluded from many of the criteria that are provided in the beginning, they will be able to give some perspective in making decisions in the future.

Table 4: The prominent advantage of selected ETL tools

Products	Advantage
SQL Server Integration Services (SSIS)	• Encompasses all aspects with all types of data. • It eases the implementation details to have fewer overall constraints. • Amazing customer help and good distribution support; it is cheap as well.
Information Server	• Tool has the greatest flexibility and strength in the market. • Appeals to the highest level of pleasure among the customers.
Informatica Power Center	• Resonant with our track record of accurately logging information, which makes learning straightforward, and the capacity to meet present-day needs for data integration. • concentrate on business-to-business data sharing.
AWS Glue	• This will let you figure out the data structure, code generation, and configuration, all in a single process. • It may also automate many other tasks, as is the case with AWS Glue. • As consumers just pay for resources utilized, it is cheaper.

Talend Open Studio	• Talend is super quick and lengthy, therefore there are no concerns when you use it to run millions of records all in one job. • Talend displays a clear depiction of the changes in data counts among several systems. • By referencing the property files in Talend, we may integrate a variety of environment-specific data with apparent ease.
Pentaho Data Integration	• The tools that come as part of the integrated ETL features are simple to understand and are capable of importing and transforming data you have. • It is quick and simple to master the built-in ETL tools and use them to import any type of data. • It just took a few minutes to set up, and it was up and running with dashboards almost immediately.
Oracle Data Integrator	• A tight connectivity to all of Oracle data warehousing and application. • All the tools are combined into one program that may be found in one single environment.

Table 5: The prominent disadvantage of selected ETL tools

Products	Disadvantage
SQL Server Integration Services (SSIS)	• Limiting of window, increasing complexity. • A vague plan and an unclear vision.
Information Server	• Really tough to understand. • It may be extensive and drawn out.

	• It is driven by a need for a huge quantity of memory and processing.
Informatica Power Center	• The decrease in the importance of these technologies has been reduced through many collaborations. • In practice, however, field experience is scarce.
AWS Glue	• To personalize the resultant ETL job, the engineers that need to do so must know Spark well. • Requests the usage of stream and batch processes that are independent from one other. • It lacks integrations with third-party products outside of the AWS environment.
Talend Open Studio	• Deduplication and fuzzy matching cannot be completed since it does not contain the necessary components. • The components that make up the Talend Open Studio product do not lend themselves to unit testing.
Pentaho Data Integration	• The lack of popularity of the technology could turn some people off, as fewer developers are available. • Without a thriving user community, it was unable to take use of tools. • The absence of a local office made it difficult to solely connect with the representative.
Oracle Data Integrator	• Single-task execution systems. • Mostly only employed for batch driven projects. • Future use of this might be unpredictable.

3.2 Comparative study of selected ETL tools on technical criteria

Criteria	SQL Server Integration Services	Information Server	Informatica PowerCenter	AWS Glue	Talend Open Studio	Pentaho Data Integration	Oracle Data Integrator
Mechanism Feature	message queuing + logging + triggers	Logging + triggers	logging	Logging + triggers	message queuing + triggers	message queuing + logging + triggers	message queuing + logging + triggers
Data Quality	Yes, Data Profiling task	Yes	Yes	Yes, DataBrew console	Yes	Yes	Yes, Oracle Data Profiling
Scheduler	Applicable	Applicable	Applicable	Applicable	Applicable	Applicable	Applicable
Grid Facility	No - Removed in 2012	Yes	Yes	Yes	Yes	Yes	Yes
Reusability	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tool base	Engine based and code generated	Engine based and code generated	Engine based	Visual and code-based interfaces	Code generated	Engine based	Code generated
Complier	Yes	Yes	Half	Yes	Yes	Yes	Half
GUI	Eclipse Add-ins	Eclipse RCP-based GUI	GUI tool Informatica PowerCenter	Visual and code-based interfaces	Eclipse based GUI	Design tool based on SWT	ODI 12c is non-GUI

3.3 Comparative study of selected ETL tools on non-technical criteria

Criteria	SQL Server Integration Services (SSIS)	Information Server	Informatica PowerCenter	AWS Glue	Talend Open Studio	Pentaho Data Integration	Oracle Data Integrator
Product Launched	Year 2005	Year 2006	Version 5.1 in July 2001	Initial release in 2017	Year 2005	Acquired in Year 2015	Year 2006
Latest Version	Version 3.12.1, Released on 3/17/21	Version 11.7.1, as 3/27/19	Version 10.5, Released on 3/21	May 10, 2021	Version 7.3.1 in 2021	Version 3.2.0 - stable	Version 12.2.1.2.0
Platform Support	Windows, Unix, Linux Ubuntu	Windows, Unix, Linux, OS X	Windows, Unix, Linux and the mainframe	Windows, Unix, Linux	Windows, Unix, Linux	Windows, Unix, Linux	Windows, Unix, Linux
Big data Support	Hadoop, HDFS, Hive, PIG, Azure Feature Pack	Hadoop, Hive, HBase, PIG, Spark, Information Server Cloud	Hadoop, Spark, PythonTx, other cloud environments	Hadoop, Hive, HBase, PIG, Spark	Hadoop, Cassandra, MongoDB, Hive, and PIG	Hadoop, NoSQL, and Analytic databases	Hadoop, Hive, HBase, PIG, Spark, Oracle SQL
Source as joined table	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ease of use	High user friendly	Less user friendly	High user friendly	High user friendly	High user friendly	Less user friendly	High user friendly
Product Feedback	4.3/5	4.6/5	4.4/5	4.4/5	4.5/5	4.0/5	4.6/5
Cost	Enterprise plan costs $14,256 per core.	Depending on the plan; small is $19,000 per month	Starting from $3.50/hour or from $24,528/year	$0.44 per DPU-Hour, billed per second, with a 10-minute minimum	Monthly: $1,170/user, and annual subscription: $12,000/user	$100 to $1,250 per month depending on scale	$900 per named user plus $198 for update license and support

3.4 Criteria description

This is how I've done the comparison of the tools, with an overview of all of the tools compared and a breakdown of the details of all of the tools. This should be enough to expose the most important facts. This is how it is described, and this is how it compares.

No.	Technical Criteria	Description
1	Mechanism Feature	Highlighting how the data is identified as it changes through the processes.
2	Data Quality	Whether or not the product has DQ application which usually equipped in their GUI.
3	Scheduler	If the tool has the capability to plan jobs based on interdependencies.
4	Grid Application	Is it possible to conduct an ETL process on a grid of PCs or servers?
5	Reusability	The tools should be able to break down the process into little chunks, enabling users to create custom functions, and then employ these functions in the process flow.
6	Tool base	Indicating whether the products are coding generated or engine base.
7	Complier	How simple is it to find errors, and if there are any, are they marked in the code with a single click?
8	GUI	Can this GUI be adjusted following customer's usability? Like new custom tabs.
No.	**Non-Technical Criteria**	**Description**
1	Establishment	The year when the product has been sold.
2	Latest Version	The latest version of that product.
3	Platform Support	The number of platforms supported by the ETL product, such as Windows and Linux, is indicated by this criterion.
4	Big data Support	Does the product support other big data manipulation like Hadoop and etc.?
5	Source as joined table	Does the product provide table-joining application in graphical format?
6	Ease of use	Showing the ease with which, the product can be used, the speed with which it can be learned, and the number of training days necessary.
7	Product Feedback	The rating from customer referring the Gartner review.
8	Cost	The expected price for products and services

4. Conclusion

ETL tools are created and used to save time and money while developing a new data mart or data warehouse. ETL tools are fundamental to Business Intelligence because they enable the transfer or transformation of data from one format to another or data mobility. They assist in extracting data from disparate heterogeneous databases, transforming it into a uniform standard format through the use of multiple procedures, and eventually loading it into a data mart or data warehouse. This study examines ETL tools by placing equal emphasis on their technical and non-technical aspects. Eclipse-based ETL tools, and open-source IDE supporting to big data were shown to be common technological characteristics of the products. Notably, most tools support Hadoop, Hive, and Pig and they are also supporting the similar operating system of Window. While SQL Server Integration Services, Talend Open Studio, Pentaho and Informatica offer excellent products, none has a perfect business model. Informatica provides a far larger range of products, but it is also more expensive. This, along with the fact that Oracle Data Integrator, IBM Information Server, and Pentaho are all also capable of working with both small and large systems, serves as an explanation for their abundance of users. But the fact is, Pentaho's recent success has made business adoption more likely for these types of open-source products. On the other hand, not all ETL tools are guaranteed to have all features and even if there is a compromise, one tool may not have the capability of the other. Finally, it can be said that nowadays there is no one tool which incorporates all these capabilities, thus we're definitely in the market of the development of ETL tools. The business requirements might heavily influence the selection of the technologies.

5. Limitations and Future Options

The comparison study was done through the collection of research reports, publications, journals, and website content. In fact, no testing was done on the actual items, only the assumptions. My evaluations were limited since I was unable to use all tools on my own. This is why the evaluation criteria I used relied on the information I obtained from the reports. The way I evaluate the expertise of other market researchers is based on two things: First, I don't have a solid background in computer systems, and furthermore, the financial investment required to gain such knowledge is in the thousands. Most importantly, ETL is just one aspect of the overall BI offering provided by the vendors. This price was also not mentioned due to the fact that many firms haven't disclosed their prices, which is a key factor in staying competitive. This includes many additional criteria, and space is limited, so I can't include them all. The long-term vision includes putting significant weight to the criteria and assessing the weighted aspects of all the ETL products instead of only focusing on the most famous ones in the certain region.

6. Reference

Amine, A., Ait Daoud, R., & Bouikhalene, B. (2016). Efficiency Comparaison and Evaluation between Two ETL Extraction Tools. *Indonesian Journal of Electrical Engineering and Computer Science*, *3*(1), 174. https://doi.org/10.11591/ijeecs.v3.i1.pp174-181

Barahama, A. D., & Wardani, R. (2021). Data analysis and data warehouse design based on Pentaho data integration (kettle) to support the determination of student learning achievement. *IOP Conference Series: Materials Science and Engineering*, *1098*(5), 052089. https://doi.org/10.1088/1757-899x/1098/5/052089

Big Data, RDBMS and HADOOP - A Comparative Study. (2016). *International Journal of Science and Research (IJSR)*, *5*(3), 1455–1458. https://doi.org/10.21275/v5i3.nov162167

Biswas, N., Sarkar, A., & Mondal, K. C. (2019). Efficient incremental loading in ETL processing for real-time data integration. *Innovations in Systems and Software Engineering*, *16*(1), 53–61. https://doi.org/10.1007/s11334-019-00344-4

Brownlee, J. (2020, March 18). *Step-By-Step Framework for Imbalanced Classification Projects*. Machine Learning Mastery. https://machinelearningmastery.com/framework-for-imbalanced-classification-projects/

Comparison of the E-LT vs ETL Method in Data Warehouse Implementation: A Qualitative Study. (2020, November 19). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9354284

Funmilola A, O. (2017). Analysis and Comparison of Open Source Business Intelligence to Commercial Business Intelligence for SME in UK. *Data Research*, *1*(1), 44.

Gartner, Inc. (n.d.). *Data Integration Tools Reviews 2021 | Gartner Peer Insights*. Gartner. Retrieved June 3, 2021, from https://www.gartner.com/reviews/market/data-integration-tools

Gill, R., & Singh, J. (2014). An Open Source ETL Tool - Medium and Small Scale Enterprise ETL(MaSSEETL). *International Journal of Computer Applications*, *108*(4), 15–22. https://doi.org/10.5120/18899-0190

Kherdekar, V. A., & Metkewar, P. S. (2016). A Technical Comprehensive Survey of ETL Tools. *International Journal of Applied Engineering Research*, *11*(04), 2557. http://www.ripublication.com/ijaer16/ijaerv11n4_64.pdf

Mithrakumar, M. (2019, November 11). *How to tune a Decision Tree?* Https://Towardsdatascience.Com/. https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680

Oni, S. (2019, October 1). *A Comparative Study of Data Cleaning Tools*. Https://Www.Igi-Global.Com/Article/a-Comparative-Study-of-Data-Cleaning-Tools/237137. https://www.igi-global.com/article/a-comparative-study-of-data-cleaning-tools/237137

Prasad, B. R., & Agarwal, S. (2016). Comparative Study of Big Data Computing and Storage Tools: A Review. *International Journal of Database Theory and Application*, *9*(1), 45–66. https://www.semanticscholar.org/paper/Comparative-Study-of-Big-Data-Computing-and-Storage-Prasad-Agarwal/ae14d49bce0d8fdfb94b3ede22c807e71134a03e?p2df

Wikipedia contributors. (2020, December 10). *Oracle Data Integrator*. Wikipedia. https://en.wikipedia.org/wiki/Oracle_Data_Integrator