SQL ETL project

You need to design a data warehouse/data store ETL/ELT system to manage data for a B2B e-commerce site. The data storage system has three defined data sources that feed the data to the system :

- B2B platform database
- Log data from the webserver
- Marketing lead spreadsheet file

B2B platform

C**ompanies** may have many end **Customers**. The B2B platform allows them to buy **Products** from qualified **Suppliers** and sell them to end **Customers**.

End **Customers** are people who are identified by their document number, full name, date of birth.

Companies are identified by CUIT number (a unique identifier), name

Suppliers are just a different type of company, and they expose a list of **Products** and default prices.

Each **Company** can define its own price list (catalog) using **Products** from many **Suppliers**. They can also place **Orders** to the platform indicating which end customer has to receive the goods

Weblog data

The clients are accessing B2B website from various devices and geo-locations. This data is represented in the combined log format.
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\""
combined • %h is the remote host (ie the client IP)
• %l is the identity of the user determined by ident (not usually used since not reliable) •
%u is the username determined by HTTP authentication
• %t is the time the request was received.
• %r is the request line from the client. ("GET / HTTP/1.0")
• %>s is the status code sent from the server to the client (200, 404, etc.) •
%b is the size of the response to the client (in bytes)
• Referer is the Referer header of the HTTP request (containing the URL of the page from which this request was initiated) if any is present, and "-" otherwise. • User-agent is the browser identification string.

For our project, the relevant data of this data source is the **client IP**, **username**, **time, and user-agent**. You can omit other data as non-relevant data.

The target datastore is the foundation for these reports :

- What are the most popular used devices for B2B clients (top 5)
- What are the most popular products in the country from which most users log into •
- All sales of B2B platform displayed monthly for the last year

Requirements

For the project solution, please prepare the following:
1) Database implementation with generated data for the B2B database source
2) Weblog generated via script in a language of your choosing
3) A target database which represents the data warehouse or data mart, you can choose relation or NoSQL solution
4) ETL/ELT* process with transformations that will :

- Fill the initial load of the target datastore
- Be restartable if the jobs or sub-job fails
- Handle erroneous data
- Track ETL/ELT metadata (when did the load start, break, finish)
- Transform the data into a readable data format for reporting
- Demonstrate the ability to transform large data size

* Use any ETL/ELT tool or hand code the ETL/ELT process