INFO-H-415 : Advanced Database
Extract-Load-Transform

Lecturer : Esteban Zimanyi
Teaching-Assistants : François Picalausa, Serge Boucher
`http://cs.ulb.ac.be/public/teaching/infoh415`

Academic Year 2010–2011

Log aggregation

After a server outage, we would like to analyze the Apache log file coming from a web server. These log files are provided on the course lab page. Using the provided data, you will create and modify jobs in Talend Open Studio as follows:

A simple spam filter the `access.log` file shows frequently recurring words such as `viagra` and `cialis`. Create a new job that finds the unique IP addresses having used those words in the `file` field of the log.

Write those addresses, along with the country of origin in a text file named `ip_blacklist`.

Too frequent to be honest Some IP addresses are repeated over and over in the log. Add the addresses that are repeated more than 50 times to the `ip_blacklist` text file.

Sorting the log out The `access.log` file contains useful lines (i.e. that can be used to do analytics on the website usage) and spam lines. Create a new job that will tag each line of the log file, to indicate whether it is genuine or spam.

Export the whole log into a database. The destination table is `Apache_log`. This tables comprises 4 fields: host (string), date (date), file (string), spam (boolean). Also export spam lines into an excel file, with the host, its country of origin, and the file. This file will be used to find new keywords to be fed into the filter.

Including errors Supplemental The `error.log` is also of interest. Export it into the database, in the same job.

One job to rule them all Create a new job that will call the construction of the blacklist, as well as the tagging of the log.

Data warehouse import

Read the Northwind datawarehouse import document. This document describes requirements of an ETL solution to import data from the OLTP Northwind database into a corresponding datawarehouse. An incomplete solution is available as an Microsoft Integration Services project.

- Explore the loading of the DimGeography dimension. Explain what each step does.
- Load the Shippers.
- Load the Suppliers, referring to the Geography dimension.
- Load the time dimension from the corresponding Excel file

- Load the sales facts.
- **Supplemental** Load the employee facts.