Mothership – Nature

README me FIRST

Objective:-

Design and build a sample Proof of Concept (POC) Work-Flow to gather publicly available Data Sets from different sources across the Internet, using an automated or semi automated method. The Data Sets are then to be saved, parsed and then inserted into the database of an associated Graphical Information System (GiS). This data can then be further processed for correlations according to a any given computational algorithm to yield results that can be visualized using a presentation platform or choice. Else can be used as the base for generating selective reports.

I must be stated here that the approach taken herein is conducive with testing a POC, and as such does not represent the level of design, implementation and consideration detail that would otherwise be taken in respect of a fully working production system.

Method:-

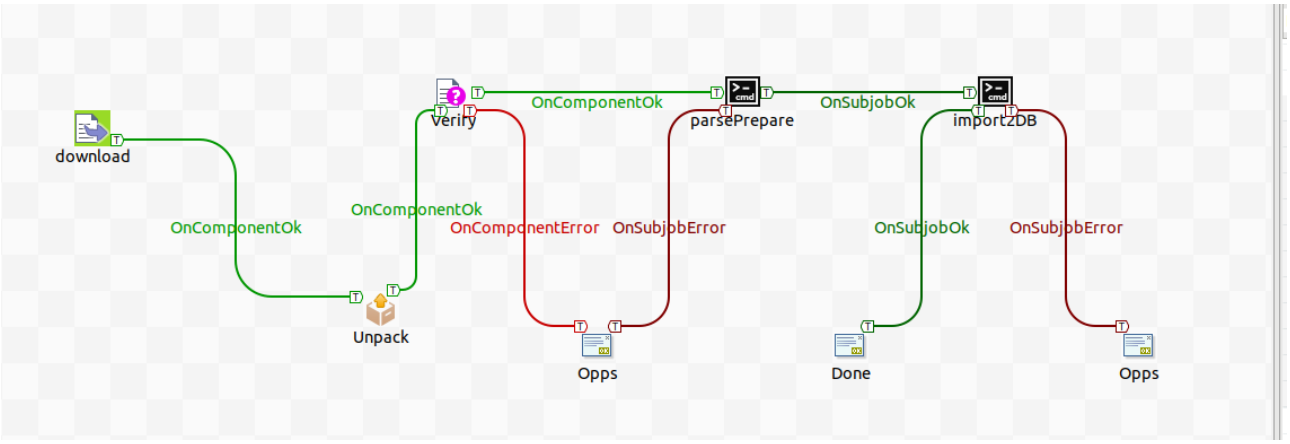
From a range of selective Internet searches, a small sample list of sources are selected for the necessary data. The respective URL for each target is noted, stored and used as the input source for the *"Fetch File"* process. For example – Geographic data representing the boroughs (Buurten) of Den Haag are presently located at this address:- *Buurten Den Haag*. The sub process makes a http connection to the remote address and downloads the associated file. This file is stored locally in a download directory and then unzipped (where approiate) into an extraction directory that is reflective by name of the contents.

The preferred file format is the Esri standard Shape File. This is made up of a bundle of Geo-Spatial files along with their associated attribute data. Once extracted, the data bundle is parsed and inserted in to the Geo-Spatial database (PostGiS) using an open source software application called shp2pgsql. Once inserted – the magic of interpreting the combined datasets can begin. From this point forward it is the domain of the data-scientist.

Technology Stack

- Database Server:- PostgreSQL
- GiS Server:- PostGiS
- GiS Deskop Tool:- QGiS
- GiS data parser:- Shp2pgsql - *installed alongside PostGiS*
- ETL:- Talend Open Studio (Java Based)
- OS of choice:- Linux
- Scripting Languages:- Python & Bash
- Visualization suit:- Tableau public

System m Overview:-



The associate directory structure for this POC is as follows

/mothership/
nature/
dataset/
downloads/
extracted/
reference/
PipelineCode
scripts/

Prerequisites:-

It is necessary to install the software outlined above under section called Technology-Stack. There after a a database user called “nature’ must be created and granted permission to created and drop tables.

Flow:-

A java program is launched from the relevant subdirectory of PipelineCode, that proceeds to orchestrate the fetching, parsing and insertion of the data into the database. Data is first brought into the download folder, from where it is unzipped and saved into sub-folders of the extracted directory. Thereafter the contents is verified. After verification, the shp2pgsl utility is called with associated parameters. It parses the shape-files and generates a database scripting file. Then the psql utility of PostgreSQL is then called with the database scripting file as one of the parameters. This psql utility will connected to the database and execute the contents of the database scripting file. The default behavior is to drop the table if it already exist and then re-create a new. Finally a Geo-Spatial index is created (if applicable) on the new table.

From the then on the dataset is ready for onwards computation.