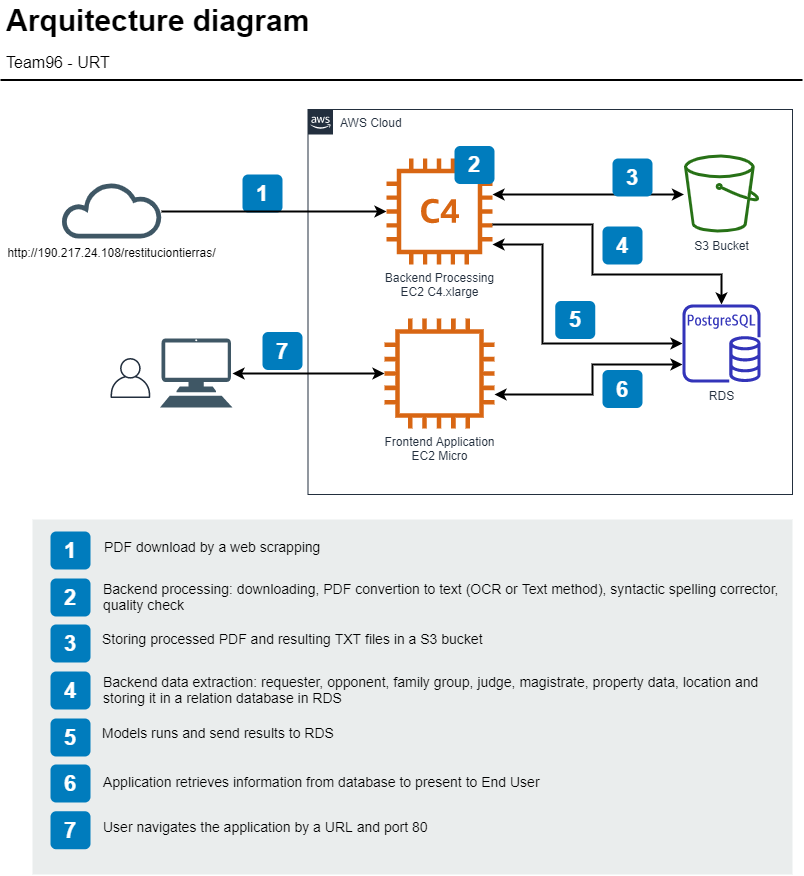
****

**Installation guide on Ubuntu 18.04 LTS**

This guide is a step by step process in order to install SmartScan in a Virtual Machine with Ubuntu 18.04 LTS with CLI commands. Following this process, you will have the solution available for you.

**Solution Architecture**

First, the following is the architecture of the solution (developed in AWS, but you can install in a similar environment):

****

**Latest version of SmartScan can works without S3 bucket, it will keep the files locally in the hard disk drive in Ubuntu.**

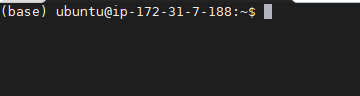
You need 2 virtual machines:

* For backend process (like scraping, downloading, converting to txt, extracting parts of the document, predict if apply some categories of interest, etc.), we recommend a C4 (or equivalent)
* For frontend (application side) you can use a micro instance

**Backend installation**

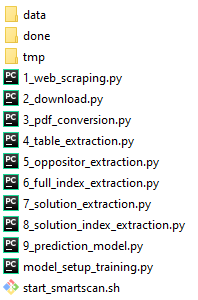
Step by Step process

You should have a machine (virtual or physical) with an Ubuntu 18.04 installed on, and with root access (sudo permission). Copy all the sentences in a CLI environment like this:



Please follow the next step in the indicated order. If the machine already accomplish, please jump to the other step:

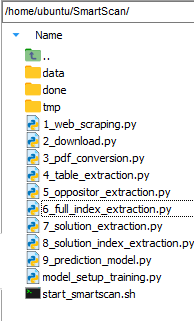
1. Install Python:
   * sudo apt-get install software-properties-common
   * sudo add-apt-repository ppa:deadsnakes/ppa
   * sudo apt-get update
   * sudo apt-get install python3.8
2. Install Miniconda:
   * wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86\_64.sh
   * sh Miniconda3-latest-Linux-x86\_64.sh
3. Install Python libraries needed (in this exact order)
   * conda install -c conda-forge tika
   * pip install PyMuPDF
   * conda install -c conda-forge pytesseract
   * conda install -c conda-forge opencv
   * conda install -c conda-forge nameparser
   * pip install roman
   * pip install ipypb
   * conda install -c anaconda pandas
   * conda install -c anaconda nltk
   * conda install -c conda-forge matplotlib
   * conda install -c anaconda seaborn
   * conda install -c conda-forge wordcloud
   * conda install -c anaconda scikit-learn
   * conda install -c anaconda psycopg2
   * conda install -c conda-forge gdal
   * pip install pyproj
   * conda install -c conda-forge fiona
   * pip install geopandas
   * conda install -c anaconda unidecode
   * conda install -c cogsci pyspellchecker
   * pip install symspellpy
   * conda install smart\_open
   * pip install gensim
   * conda install folium -c conda-forge
   * conda install -c anaconda ipython
   * conda install -c anaconda sqlalchemy
   * conda install -c conda-forge spacy
   * conda install -c plotly plotly
   * sudo apt install awscli
   * sudo apt install zip unzip
   * sudo apt-get install chromium-chromedriver
   * conda install -c conda-forge scrapy
   * conda install -c conda-forge selenium
   * conda install -c anaconda beautifulsoup4
   * conda install -c anaconda html5lib
4. Make the initial downloads for nltk library
   * python
     + import nltk
     + nltk.download('maxent\_ne\_chunker')
     + nltk.download('words')
     + nltk.download('punkt')
     + nltk.download('stopwords')
     + nltk.download('wordnet')
     + nltk.download('averaged\_perceptron\_tagger')
   * quit()
5. Configure tesseract for Ubuntu
   * sudo apt install tesseract-ocr
   * sudo apt-get install tesseract-ocr-spa
   * sudo apt install libgl1-mesa-glx
6. Install java for make tesseract run into Ubuntu VM
   * sudo apt install default-jre
7. Create a folder to the backend content:
   * mkdir SmartScan
8. Load content of Backend folder into the previous folder. Structure should seems like:



1. Give full permissions to start\_smartscan.sh
   * chmod +777 start\_smartscan.sh
2. Setup your connection to database. Each .py files contents a variable named engine, you should put there your connection string to database. Be sure to give proper permissions to Virtual Machine to access database:



At the end of the installation, your backend should looks like that:



**Database installation**

You should have an instance of an SQL database, we recommended Postgres because it was used during develop and it performed great. Please create a new database with the name of your preference, setup users and passwords to access full the new database, the, please run the attached SQL script to create all the objects.



After run, you will have all the structures created, just need to start publishing new PDFs. Optionally, you can upload external information on mt\_dane\_censo\_2018, mt\_dane\_estrato and mt\_dane\_geo if you have it, it will help to standardize codes of states, towns and municipality.

**Frontend installation**