BIG DATA PROJECT REPORT

IMPLEMENTATION OF A DATALAKE ARCHITECTURE FOR THE CONSOLIDATION AND ANALYSIS OF PUBLIC DATA

Prepared by : Aaraba Anass and Omar Farouk Lafdil

supervised by : Mr. Benelallam Imade

ABSTRACT :

In this project, we implemented a Big Data architecture in the form of a data warehouse lake using a selection of technologies, including Hadoop, Hbase, Spark, web scraping, and AI models. The goal of this project was to consolidate and analyze public data from the website "marchespublics.gov.ma". To achieve this, we used web scraping techniques to extract various document formats, including .pdf, .png, .jpg, .jpeg, .doc, .docx, .ppt, and .zip files. We then applied data cleansing and transformation techniques to convert these documents into .txt files and used AI models and natural language processing tools to extract relevant data. The resulting data was stored in .csv files and loaded into the Hadoop ecosystem using HDFS commands. We also configured Anaconda and a virtual environment to run a Jupyter notebook, and used PySpark to conduct further analysis. This project demonstrates the effectiveness of using a diverse set of tools and techniques to build a robust Big Data architecture for the consolidation and analysis of public data.

GENERAL PLAN :

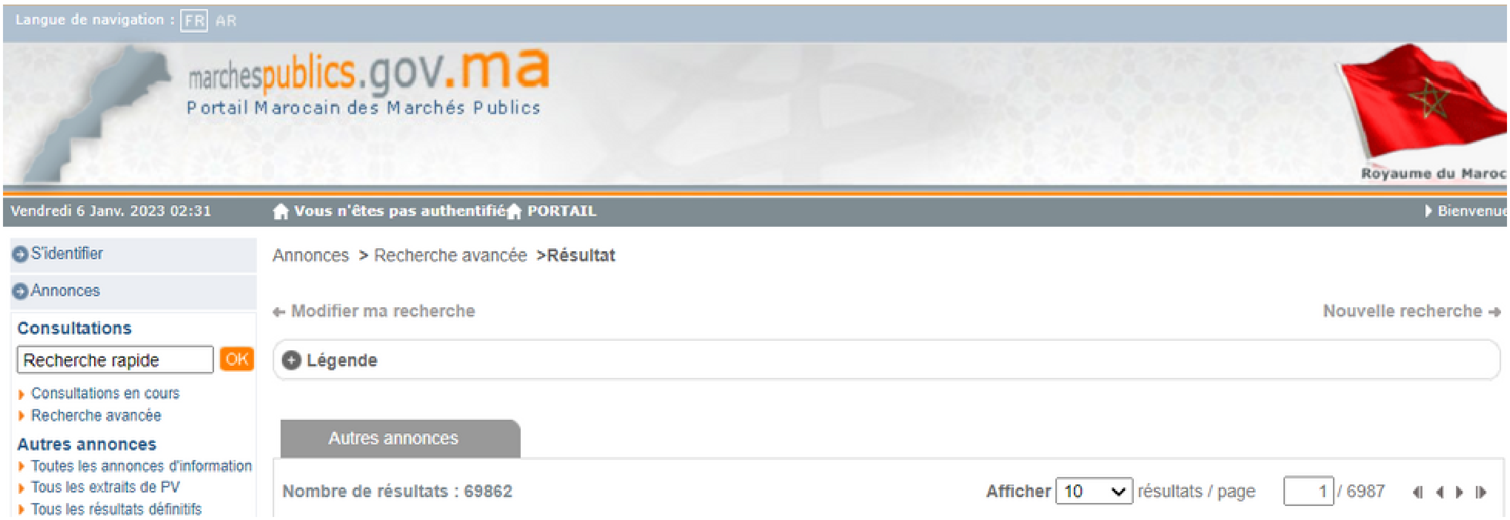
INTRODUCTION :

The goal of this project was to build a Big Data architecture in the form of a data warehouse lake using a selection of technologies, including Hadoop, Hbase, Spark, web scraping, and AI models. The objective was to consolidate and analyze public data from the website "marchespublics.gov.ma" in order to extract all types of data, such as the subject, the project owner, the contract winner, and some relevant data. To achieve this, we applied web scraping techniques to extract various document formats from the website, including .pdf, .png, .word, and .zip files. We then used data cleansing and transformation techniques to convert these documents into .txt files and applied AI models and natural language processing tools to extract relevant data. The resulting data was stored in .csv files and loaded into the Hadoop ecosystem using HDFS commands. We also configured Anaconda and a virtual environment to run a Jupyter notebook and used PySpark to conduct further analysis. In this report, we will describe the data, methods, and results of this project, as well as the implications and recommendations for future work.

THE DATA:

1 / DATA SOURCES

The data for this analysis was collected from the website "marchespublics.gov.ma", which is a public portal for procurement and tendering information in Morocco. The website contains a range of documents and information related to various procurement processes, including .pdf, .png, .word, and .zip files. We used web scraping techniques to extract these documents from the website and used data transformation techniques to convert them into .txt files for further analysis. Unfortunately, we had a capacity issue, so we limited the period to 3 months, from January to March 2019.

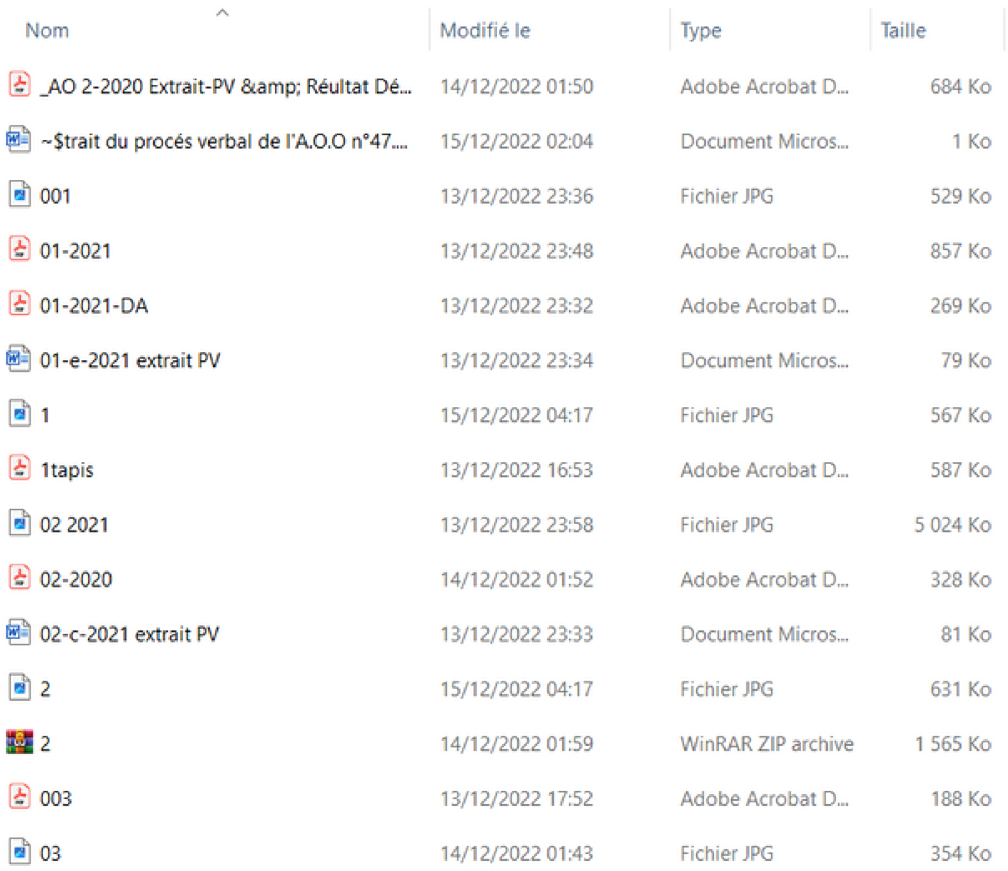


Nombre de résultats : 69862

THE DATA:

2 / DATA EXTRACTION

To extract the data from "marchespublics.gov.ma", we used web scraping techniques using the BeautifulSoup library and web scraping using Selenium webdriver. We wrote a script to scrape the website and extract the relevant documents and information. This process involved navigating to the "https://www.marchespublics.gov.ma/index.php?page=entreprise.EntrepriseAdvancedSearch&AvisExtraitPV" link, configuring the download directory, and removing duplicate files. We encountered some challenges during the data extraction process, such as dealing with CAPTCHAs. To address these challenges, we used the undetected_webdriver to ensure that CAPTCHAs don't recognize our bot. Overall, the data extraction process required more than 7 hours and a storage capacity of 3GB.



Nom	Modifié le	Type	Taille
_AO 2-2020 Extrait-PV &amp; Réultat Dé...	14/12/2022 01:50	Adobe Acrobat D...	684 Ko
~$trait du procés verbal de l'A.O.O n°47....	15/12/2022 02:04	Document Micros...	1 Ko
001	13/12/2022 23:36	Fichier JPG	529 Ko
01-2021	13/12/2022 23:48	Adobe Acrobat D...	857 Ko
01-2021-DA	13/12/2022 23:32	Adobe Acrobat D...	269 Ko
01-e-2021 extrait PV	13/12/2022 23:34	Document Micros...	79 Ko
1	15/12/2022 04:17	Fichier JPG	567 Ko
1tapis	13/12/2022 16:53	Adobe Acrobat D...	587 Ko
02 2021	13/12/2022 23:58	Fichier JPG	5 024 Ko
02-2020	14/12/2022 01:52	Adobe Acrobat D...	328 Ko
02-c-2021 extrait PV	13/12/2022 23:33	Document Micros...	81 Ko
2	15/12/2022 04:17	Fichier JPG	631 Ko
2	14/12/2022 01:59	WinRAR ZIP archive	1 565 Ko
003	13/12/2022 17:52	Adobe Acrobat D...	188 Ko
03	14/12/2022 01:43	Fichier JPG	354 Ko

THE DATA:

3 / DATA PREPROCESSING AND CLEANSING

This part includes multiple data processing steps in order to convert every file to a .txt file :

1- UNZIP THE .ZIP FILES

- The first step was to unzip the .zip files in order to maximize our data lake. In this step, we used the zipfile and glob libraries

2- DEALING WITH .PDF FILES

- The next step was dealing with .pdf files by converting every .pdf file to an image using the pdf2image library.
- After that, we stored every converted image to a directory, then they were deleted.

3- DEALING WITH IMAGES

- The next step was converting every image stored in the images directory and other .png, .jpg, .jpeg... files to .txt files.
- This process is done by using a cvttext.py script that use PIL and pytesseract libraries.
- This process was done using a cvttext.py script that used the PIL and pytesseract libraries.

4- DEALING WITH WORD FILES

- The next step was dealing with every .doc or .docx file by converting them to image files using the aspose.words API, which handles every version of Word and converts it to a .png image.
- Then we dealt with every image to get a .txt file for data mining purposes.

THE DATA:

4 / DATA MINING

This part includes multiple data mining steps in order to insert all valuable data into a .csv file.

1- EXTRACTING VALUABLE DATA

- The first step was to extract the data we needed for our analysis, including the subject, the project owner, the winner of the contract, the price of the contract, the competitors, and their offers.
- The output was stored in a .txt file in a directory for text mining.
- For this step, we used the OpenAI API to extract all useful data.

2- DATA EXTRACTION

- This step was about extracting the data from the output of every OpenAI API request and finally generating a data frame containing the specific data we needed
- In this step, we used various text mining libraries.

3- CREATION OF A .CSV FILE

- The next step was taking the data frame and converting it to a .csv file.
- For some financial reasons, we were not able to pursue requests for every file at the same time. There is a rate limit in the OpenAI API on the number of requests per minute.

4- MERGING DATA IN ONE .CSV FILE

- After completing the process of extraction for every file in the data lake, we obtained multiple .csv files
- Finally, we combined the .csv files into one .csv file.

THE DATA:

5 / FINAL DATASET CHARACTERISTICS

Our final data set is a .csv file containing the valuable data we sought to use in our analysis. Our data set includes:

1- OBJET AND MAITRE_OUVRAGE COLUMNS

- The first column is about the subject of every minutes of the open call for tenders document.
- The second column is about the project owner

2- GAGNANT AND OFFRE_GAGNANT

- These two columns are about the winner of the project and it's offer in moroccan dirham currency.

3- CONCURRENTS COLUMN

- This column is about the competitors of the winner.
- This column contains a list of the competitors.

4- OFFRES_CONCURRENTS COLUMN

- Finally, this column contains the competitors' offers in Moroccan dirham currency.
- This column also contains a list of the offers.

THE DATA:

6 / DATA ANALYSIS USING APACHE HADOOP

In order to achieve our analysis goals, we followed multiple steps

1- DATA TRANSFER TO HADOOP SYSTEM

- The first step was transferring our data set into Cloudera Hadoop using a Vmware virtual machine.

2- HDFS STORING

- The next step was connecting to the Hadoop system using the 'hdfs' command.
- Then, we imported the csv file into the Hadoop system using the '-put' command.
- Finally, we checked that the file had been imported into the Hadoop file system.

3- ANACONDA3 INSTALLATION

- Due to some issues with the Spark default installation, we were not able to interact with the pyspark shell.
- So, we installed the Anaconda3 version that is compatible with our Unix system.
- After installing Anaconda3 and setting up the environment, we were able to start a Jupyter-Notebook.

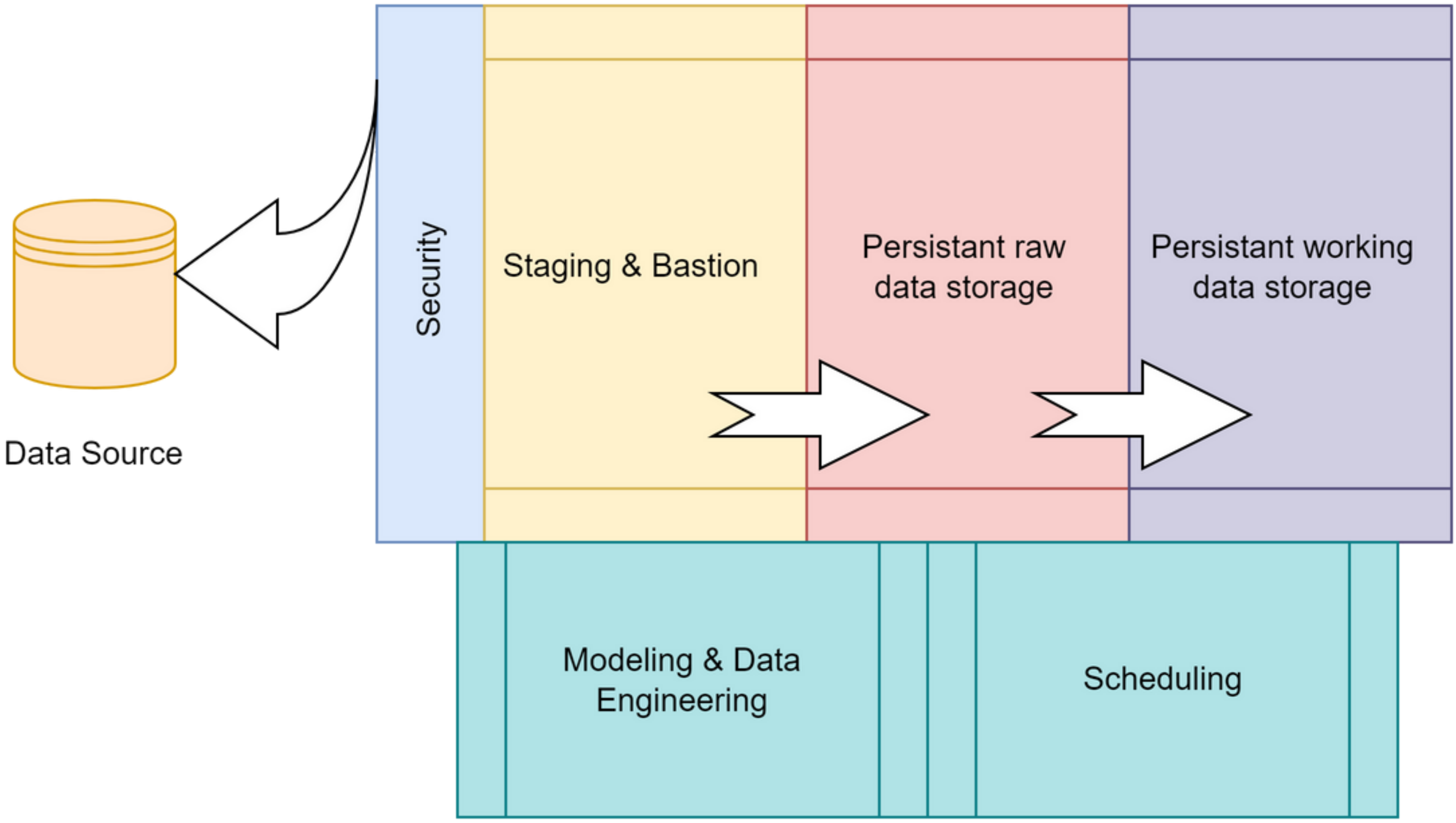
4- PYSPARK INSTALLATION

- After starting the Jupyter notebook, we installed Pyspark using the 'pip' command.
- After installing Pyspark, we were able to begin the data analysis.

BIG DATA ANALYTICS ARCHITECTURE

1 / LOGICAL ARCHITECTURE

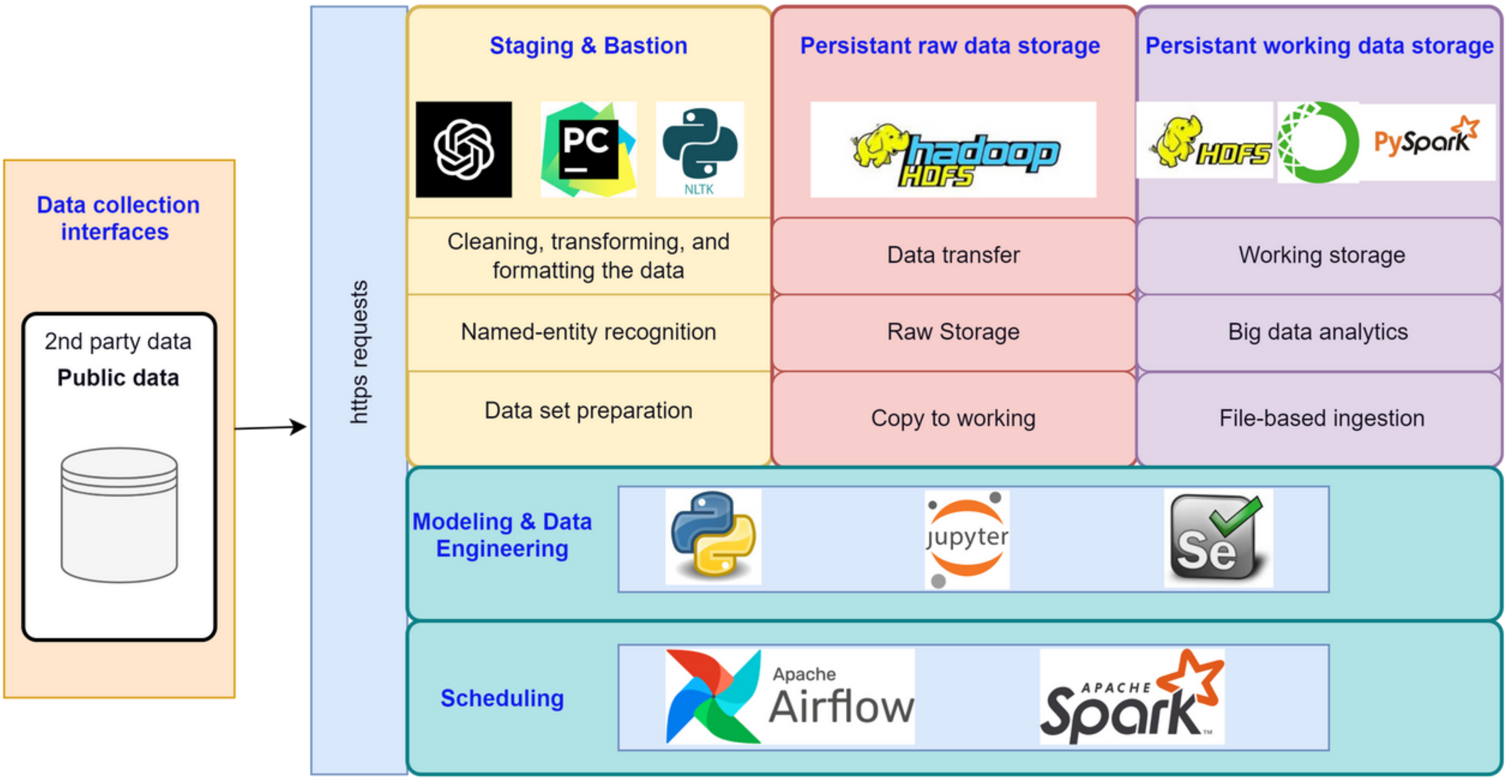
In our big data analysis, we implemented the following logical architecture:


Data Source
Security
Staging & Bastion
Persistant raw data storage
Persistant working data storage
Modeling & Data Engineering
Scheduling

BIG DATA ANALYTICS ARCHITECTURE

2 / TECHNICAL ARCHITECTURE

In our big data analysis, we implemented the following technical architecture:

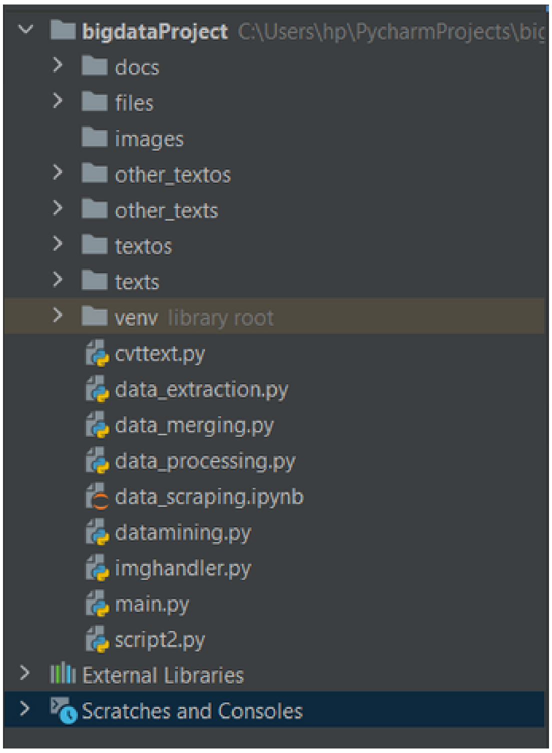


PROJECT DEMONSTRATION

1 / DATA SCRAPING

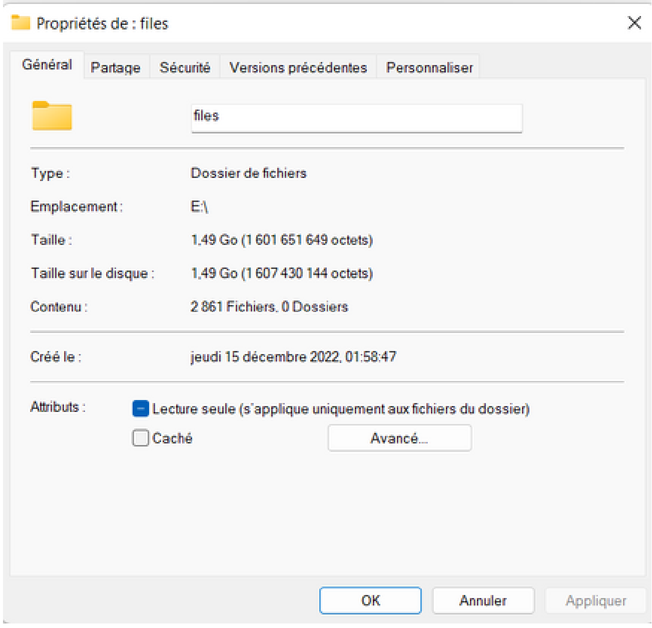
The first task of our big data analytics was scraping data from the public website:

This Pycharm project contains all the work from data scraping to moving the data to the Hadoop ecosystem.
The file in charge of scraping the data is "data_scraping.ipynb".



```
bigdataProject  C:\Users\hp\PycharmProjects\big
    docs
    files
    images
    other_textos
    other_texts
    textos
    texts
    venv  library root
    cvttext.py
    data_extraction.py
    data_merging.py
    data_processing.py
    data_scraping.ipynb
    datamining.py
    imghandler.py
    main.py
    script2.py
External Libraries
Scratches and Consoles
```

The data lake resulting from the data scraping process includes 1.49 GB of data in multiple formats (.pdf, .png, .jpg, etc.).



Propriétés de : files

files

Type : Dossier de fichiers

Emplacement : E:\

Taille : 1,49 Go (1 601 651 649 octets)

Taille sur le disque : 1,49 Go (1 607 430 144 octets)

Contenu : 2 861 Fichiers, 0 Dossiers

Créé le : jeudi 15 décembre 2022, 01:58:47

Attributs : ■ Lecture seule (s'applique uniquement aux fichiers du dossier)

□ Caché

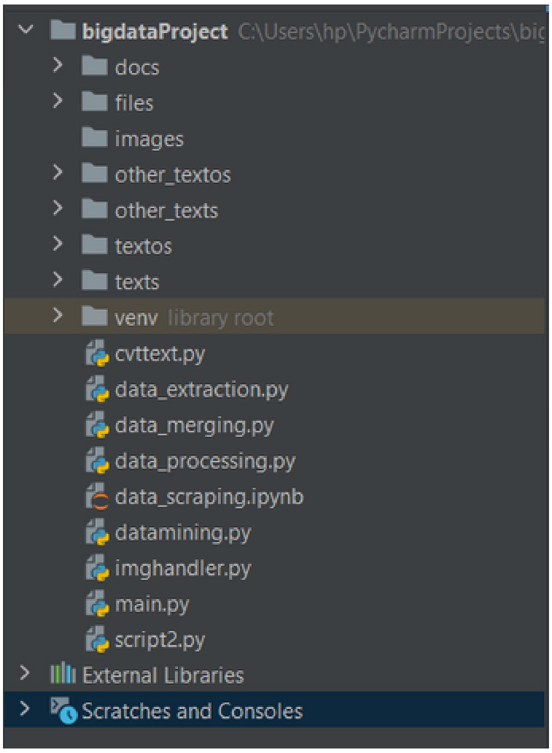
Avancé...

PROJECT DEMONSTRATION

2 / DATA PREPROCESSING AND CLEANSING

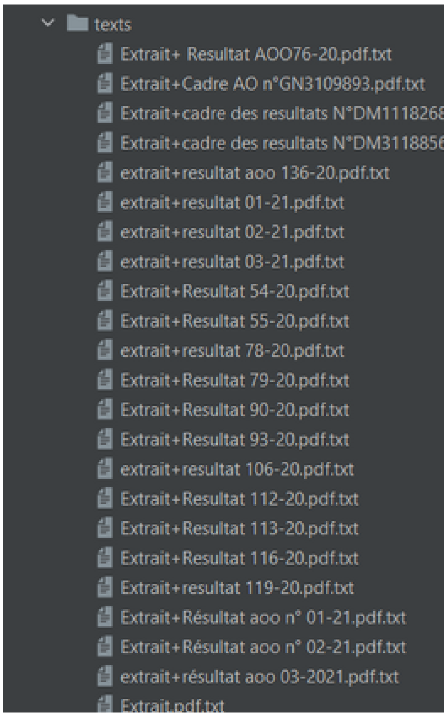
The next task of our big data analytics was processing and cleansing the data scraped in the first task.

This Pycharm project contains all the work from data scraping to moving the data to the Hadoop ecosystem. Multiple python scripts are responsible for handling this task, such as cvttext.py, script2.py, imgha.dler.py, and data_processing.py which is in charge of orchestrating this process.



```
bigdataProject  C:\Users\hp\PycharmProjects\big
    docs
    files
    images
    other_textos
    other_texts
    textos
    texts
    venv  library root
    cvttext.py
    data_extraction.py
    data_merging.py
    data_processing.py
    data_scraping.ipynb
    datamining.py
    imghandler.py
    main.py
    script2.py
External Libraries
Scratches and Consoles
```

The result of this process is a data lake that includes data in .txt format.



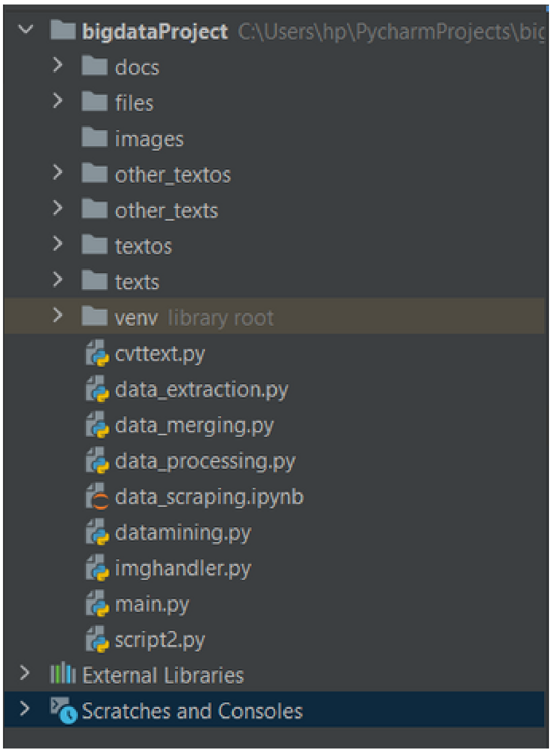
```
texts
    Extrait+ Resultat AOO76-20.pdf.txt
    Extrait+Cadre AO n°GN3109893.pdf.txt
    Extrait+cadre des resultats N°DM1118268
    Extrait+cadre des resultats N°DM3118856
    extrait+resultat aoo 136-20.pdf.txt
    extrait+resultat 01-21.pdf.txt
    extrait+resultat 02-21.pdf.txt
    extrait+resultat 03-21.pdf.txt
    Extrait+Resultat 54-20.pdf.txt
    Extrait+Resultat 55-20.pdf.txt
    extrait+resultat 78-20.pdf.txt
    Extrait+Resultat 79-20.pdf.txt
    Extrait+Resultat 90-20.pdf.txt
    Extrait+Resultat 93-20.pdf.txt
    extrait+resultat 106-20.pdf.txt
    Extrait+Resultat 112-20.pdf.txt
    Extrait+Resultat 113-20.pdf.txt
    Extrait+Resultat 116-20.pdf.txt
    Extrait+resultat 119-20.pdf.txt
    Extrait+Résultat aoo n° 01-21.pdf.txt
    Extrait+Résultat aoo n° 02-21.pdf.txt
    extrait+résultat aoo 03-2021.pdf.txt
    Extrait.pdf.txt
```

PROJECT DEMONSTRATION

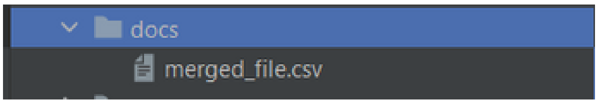
2 / DATA MINING

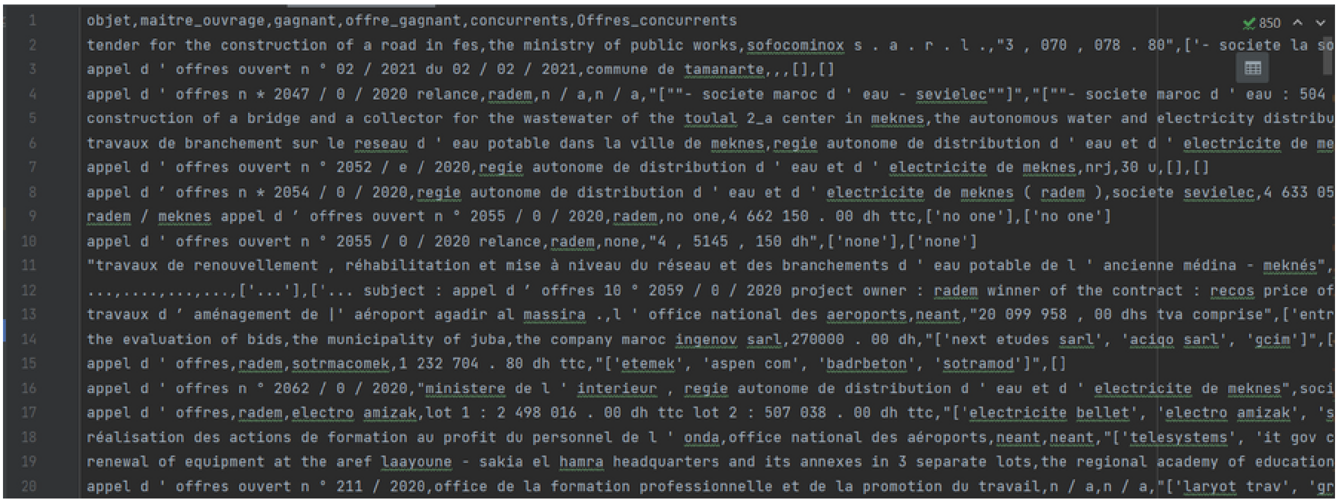
The next task of our big data analytics was about the data mining in order to extract the valuable data we need:

This Pycharm project contain all the works from the data scraping to moving the data to the Hadoop ecosystem: Multiple python scripts are in charge of handling this task such as datamining.py, data_extraction.py and data_merging.py



The result of this process is our data warehouse in a .csv format. this .csv file contains all the extracted data we need for our analytics.





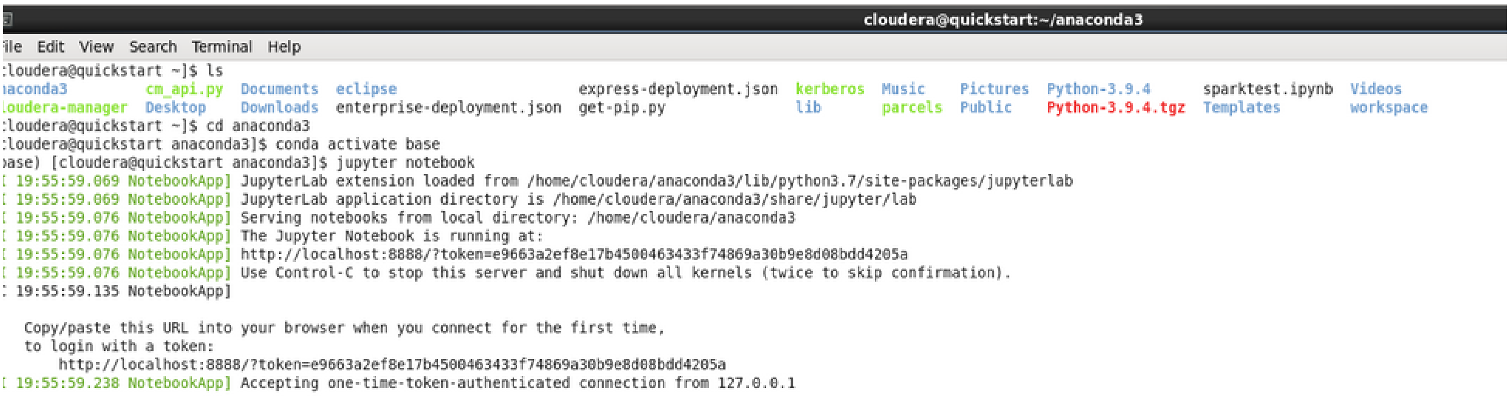
```
objet,maitre_ouvrage,gagnant,offre_gagnant,concurrents,Offres_concurrents
tender for the construction of a road in fes,the ministry of public works,sofocominox s . a . r . l .,"3 , 070 , 078 . 80",['- societe la so
appel d ' offres ouvert n ° 02 / 2021 du 02 / 02 / 2021,commune de tamanarte,,,[],[]
appel d ' offres n * 2047 / 0 / 2020 relance,radem,n / a,n / a,"[""- societe maroc d ' eau - sevielec""]","[""- societe maroc d ' eau : 504
construction of a bridge and a collector for the wastewater of the toulal 2_a center in meknes,the autonomous water and electricity distribu
travaux de branchement sur le reseau d ' eau potable dans la ville de meknes,regie autonome de distribution d ' eau et d ' electricite de me
appel d ' offres ouvert n ° 2052 / e / 2020,regie autonome de distribution d ' eau et d ' electricite de meknes,nrj,30 u,[],[]
appel d ' offres n * 2054 / 0 / 2020,regie autonome de distribution d ' eau et d ' electricite de meknes ( radem ),societe sevielec,4 633 05
radem / meknes appel d ' offres ouvert n ° 2055 / 0 / 2020,radem,no one,4 662 150 . 00 dh ttc,['no one'],['no one']
appel d ' offres ouvert n ° 2055 / 0 / 2020 relance,radem,none,"4 , 5145 , 150 dh",['none'],['none']
"travaux de renouvellement , réhabilitation et mise à niveau du réseau et des branchements d ' eau potable de l ' ancienne médina - meknés",
...,....,...,....,['...'],['... subject : appel d ' offres 10 ° 2059 / 0 / 2020 project owner : radem winner of the contract : recos price of
travaux d ' aménagement de |' aéroport agadir al massira .,l ' office national des aeroports,neant,"20 099 958 , 00 dhs tva comprise",['entr
the evaluation of bids,the municipality of juba,the company maroc ingenov sarl,270000 . 00 dh,"['next etudes sarl', 'acigo sarl', 'gcim']",[
appel d ' offres,radem,sotrmacomek,1 232 704 . 80 dh ttc,"['etemek', 'aspen com', 'badrbeton', 'sotramod']",[]
appel d ' offres n ° 2062 / 0 / 2020,"ministere de l ' interieur , regie autonome de distribution d ' eau et d ' electricite de meknes",soci
appel d ' offres,radem,electro amizak,lot 1 : 2 498 016 . 00 dh ttc lot 2 : 507 038 . 00 dh ttc,"['electricite bellet', 'electro amizak', 's
réalisation des actions de formation au profit du personnel de l ' onda,office national des aéroports,neant,neant,"['telesystems', 'it gov c
renewal of equipment at the aref laayoune - sakia el hamra headquarters and its annexes in 3 separate lots,the regional academy of education
appel d ' offres ouvert n ° 211 / 2020,office de la formation professionnelle et de la promotion du travail,n / a,n / a,"['laryot trav', 'gr
```

PROJECT DEMONSTRATION

2 / DATA ANALYSIS USING APACHE HADOOP

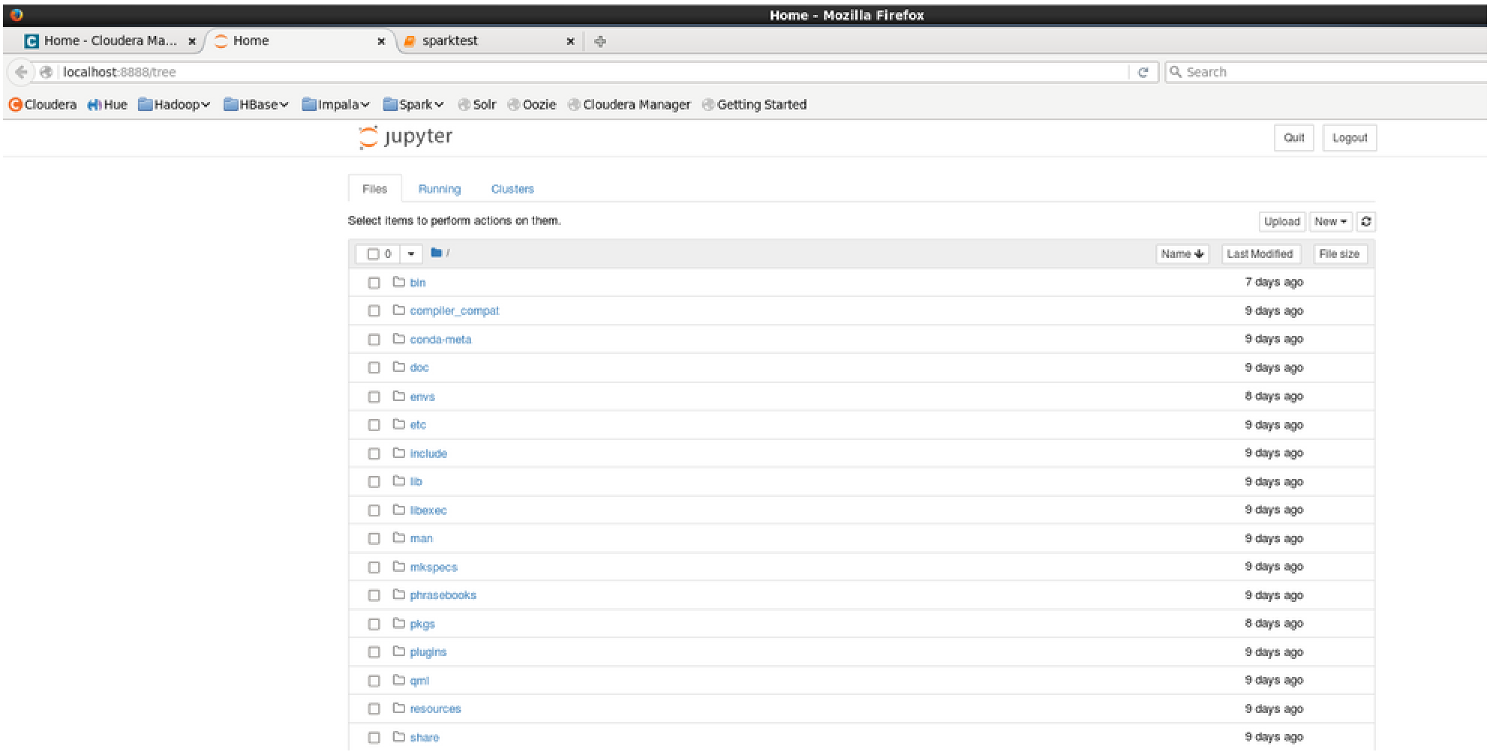
The next task of our big data analytics was about the analytics using Pyspark in Apache Hadoop:

First we prepare the Anaconda3 environement to performe the analytics with Pyspark using a Jupyter-notebook:


cloudera@quickstart:~/anaconda3
ile Edit View Search Terminal Help

```
:loudera@quickstart ~]$ ls
naconda3          cm_api.py  Documents  eclipse                    express-deployment.json  kerberos  Music    Pictures  Python-3.9.4      sparktest.ipynb  Videos
loudera-manager   Desktop    Downloads  enterprise-deployment.json get-pip.py               lib       parcels  Public    Python-3.9.4.tgz  Templates        workspace
:loudera@quickstart ~]$ cd anaconda3
:loudera@quickstart anaconda3]$ conda activate base
)ase) [cloudera@quickstart anaconda3]$ jupyter notebook
[ 19:55:59.069 NotebookApp] JupyterLab extension loaded from /home/cloudera/anaconda3/lib/python3.7/site-packages/jupyterlab
[ 19:55:59.069 NotebookApp] JupyterLab application directory is /home/cloudera/anaconda3/share/jupyter/lab
[ 19:55:59.076 NotebookApp] Serving notebooks from local directory: /home/cloudera/anaconda3
[ 19:55:59.076 NotebookApp] The Jupyter Notebook is running at:
[ 19:55:59.076 NotebookApp] http://localhost:8888/?token=e9663a2ef8e17b4500463433f74869a30b9e8d08bdd4205a
[ 19:55:59.076 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
: 19:55:59.135 NotebookApp]

  Copy/paste this URL into your browser when you connect for the first time,
  to login with a token:
      http://localhost:8888/?token=e9663a2ef8e17b4500463433f74869a30b9e8d08bdd4205a
[ 19:55:59.238 NotebookApp] Accepting one-time-token-authenticated connection from 127.0.0.1
```



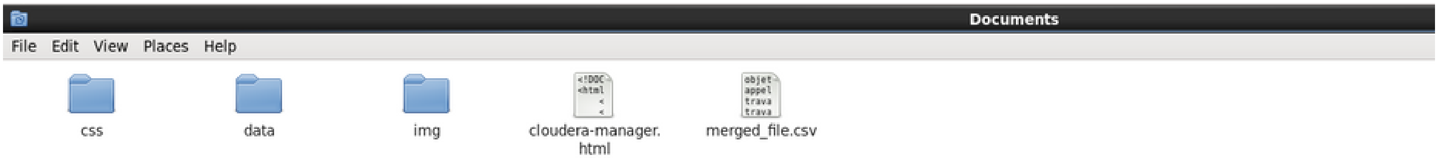
Select items to perform actions on them.

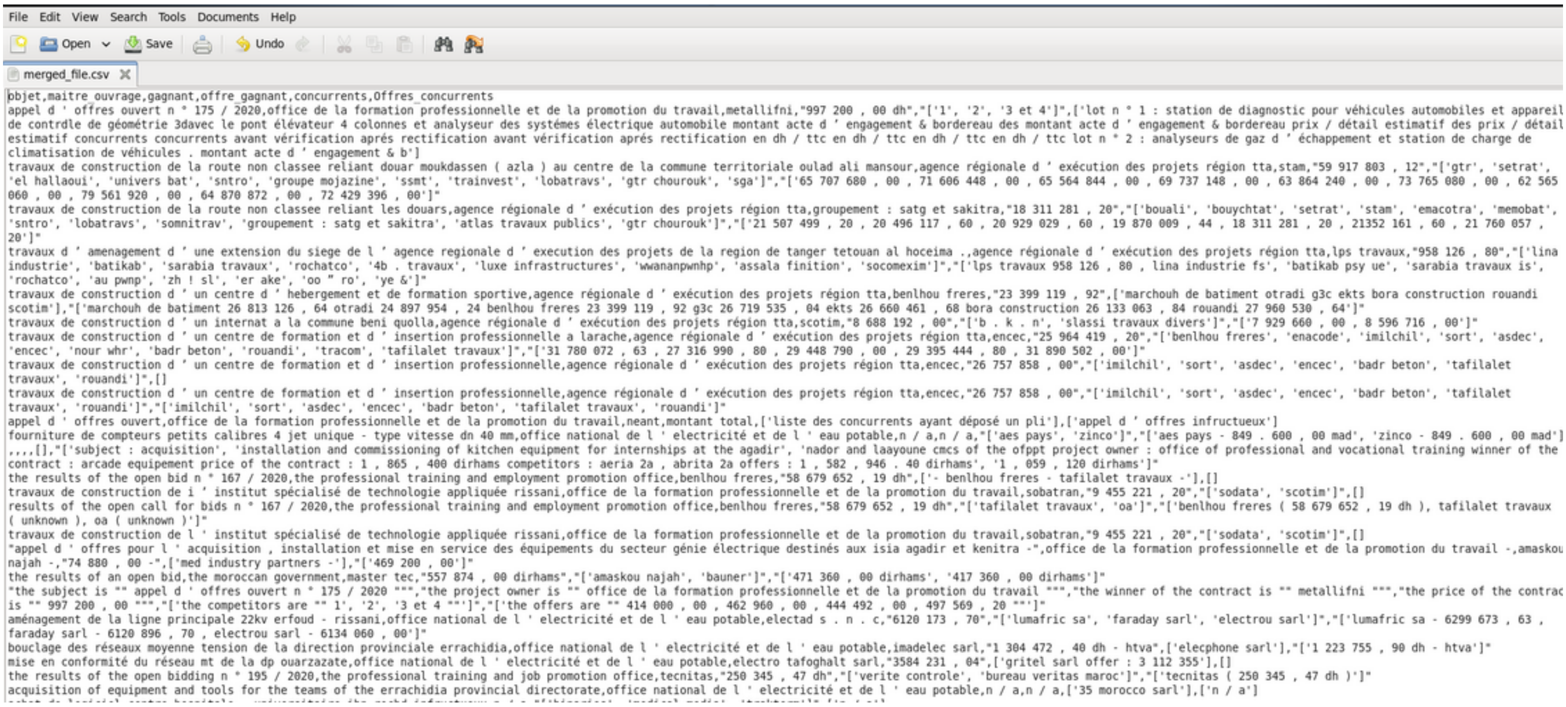
Name	Last Modified
□ bin	7 days ago
□ compiler_compat	9 days ago
□ conda-meta	9 days ago
□ doc	9 days ago
□ envs	8 days ago
□ etc	9 days ago
□ include	9 days ago
□ lib	9 days ago
□ libexec	9 days ago
□ man	9 days ago
□ mkspecs	9 days ago
□ phrasebooks	9 days ago
□ pkgs	8 days ago
□ plugins	9 days ago
□ qml	9 days ago
□ resources	9 days ago
□ share	9 days ago

PROJECT DEMONSTRATION

2 / DATA ANALYSIS USING APACHE HADOOP

Next we copy our .csv file using WinSCP to transfer the CSV file from our Windows machine to the Cloudera Hadoop VM. Then we store our data into the Hadoop Distributed File System :

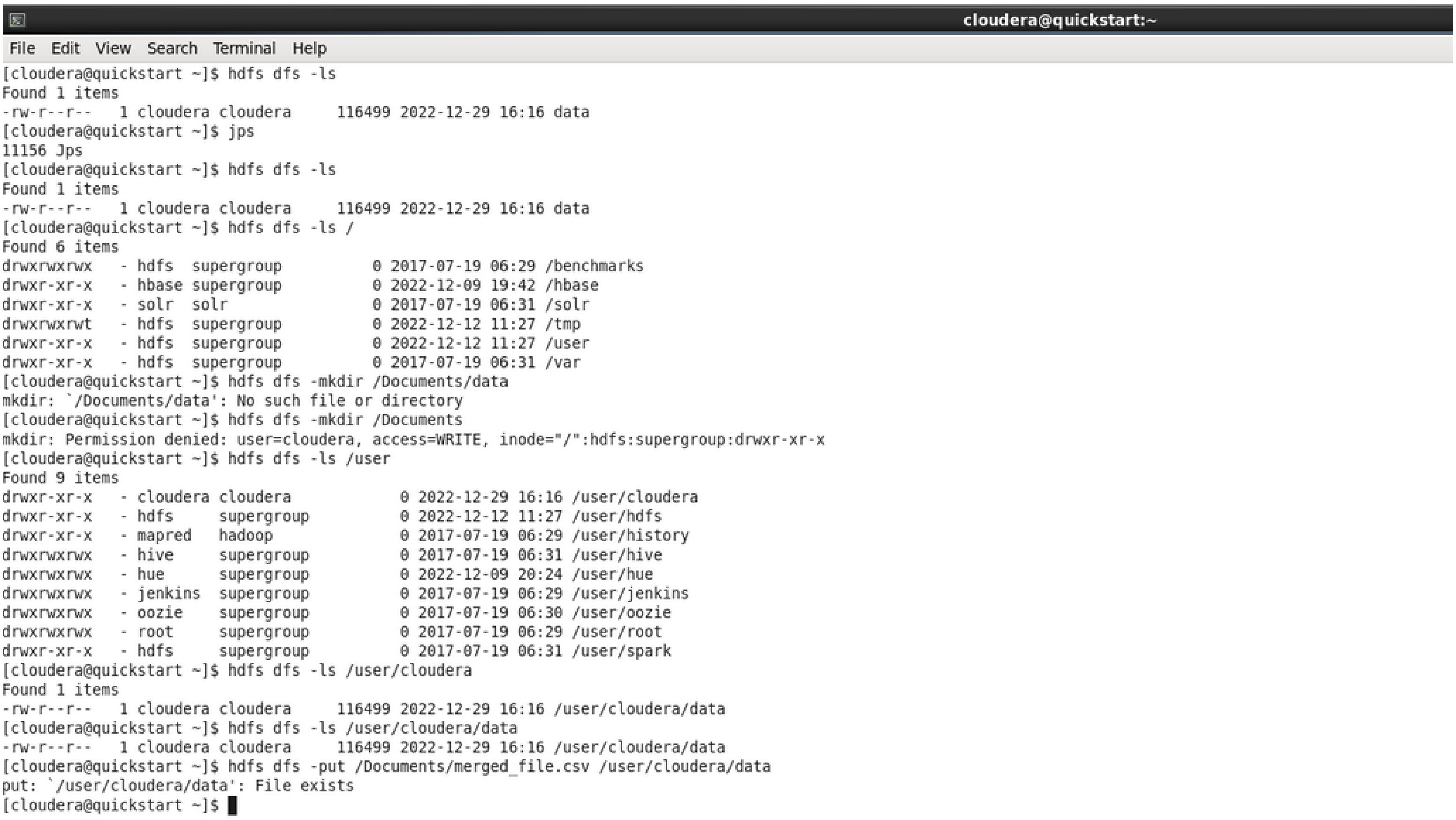




```
objet,maitre_ouvrage,gagnant,offre_gagnant,concurrents,Offres_concurrents
appel d ' offres ouvert n ° 175 / 2020,office de la formation professionnelle et de la promotion du travail,metallifni,"997 200 , 00 dh","['1', '2', '3 et 4']",['lot n ° 1 : station de diagnostic pour véhicules automobiles et appareil
de contrdle de géométrie 3davec le pont élévateur 4 colonnes et analyseur des systémes électrique automobile montant acte d ' engagement & bordereau des montant acte d ' engagement & bordereau prix / détail estimatif des prix / détail
estimatif concurrents concurrents avant vérification aprés rectification avant vérification aprés rectification en dh / ttc en dh / ttc en dh / ttc en dh / ttc lot n ° 2 : analyseurs de gaz d ' échappement et station de charge de
climatisation de véhicules . montant acte d ' engagement & b']
travaux de construction de la route non classee reliant douar moukdassen ( azla ) au centre de la commune territoriale oulad ali mansour,agence régionale d ' exécution des projets région tta,stam,"59 917 803 , 12","['gtr', 'setrat',
'el hallaoui', 'univers bat', 'sntro', 'groupe mojazine', 'ssmt', 'trainvest', 'lobatravs', 'gtr chourouk', 'sga']","['65 707 680 , 00 , 71 606 448 , 00 , 65 564 844 , 00 , 69 737 148 , 00 , 63 864 240 , 00 , 73 765 080 , 00 , 62 565
060 , 00 , 79 561 920 , 00 , 64 870 872 , 00 , 72 429 396 , 00']"
travaux de construction de la route non classee reliant les douars,agence régionale d ' exécution des projets région tta,groupement : satg et sakitra,"18 311 281 , 20","['bouali', 'bouychtat', 'setrat', 'stam', 'emacotra', 'memobat',
'sntro', 'lobatravs', 'somnitrav', 'groupement : satg et sakitra', 'atlas travaux publics', 'gtr chourouk']","['21 507 499 , 20 , 20 496 117 , 60 , 20 929 029 , 60 , 19 870 009 , 44 , 18 311 281 , 20 , 21352 161 , 60 , 21 760 057 ,
20']"
travaux d ' amenagement d ' une extension du siege de l ' agence regionale d ' execution des projets de la region de tanger tetouan al hoceima .,agence régionale d ' exécution des projets région tta,lps travaux,"958 126 , 80","['lina
industrie', 'batikab', 'sarabia travaux', 'rochatco', '4b . travaux', 'luxe infrastructures', 'wwananpwnhp', 'assala finition', 'socomexim']","['lps travaux 958 126 , 80 , lina industrie fs', 'batikab psy ue', 'sarabia travaux is',
'rochatco', 'au pwnp', 'zh ! sl', 'er ake', 'oo " ro', 'ye &']"
travaux de construction d ' un centre d ' hebergement et de formation sportive,agence régionale d ' exécution des projets région tta,benlhou freres,"23 399 119 , 92",['marchouh de batiment otradi g3c ekts bora construction rouandi
scotim'],"['marchouh de batiment 26 813 126 , 64 otradi 24 897 954 , 24 benlhou freres 23 399 119 , 92 g3c 26 719 535 , 04 ekts 26 660 461 , 68 bora construction 26 133 063 , 84 rouandi 27 960 530 , 64']"
travaux de construction d ' un internat a la commune beni quolla,agence régionale d ' exécution des projets région tta,scotim,"8 688 192 , 00","['b . k . n', 'slassi travaux divers']","['7 929 660 , 00 , 8 596 716 , 00']"
travaux de construction d ' un centre de formation et d ' insertion professionnelle a larache,agence régionale d ' exécution des projets région tta,encec,"25 964 419 , 20","['benlhou freres', 'enacode', 'imilchil', 'sort', 'asdec',
'encec', 'nour whr', 'badr beton', 'rouandi', 'tracom', 'tafilalet travaux']","['31 780 072 , 63 , 27 316 990 , 80 , 29 448 790 , 00 , 29 395 444 , 80 , 31 890 502 , 00']"
travaux de construction d ' un centre de formation et d ' insertion professionnelle,agence régionale d ' exécution des projets région tta,encec,"26 757 858 , 00","['imilchil', 'sort', 'asdec', 'encec', 'badr beton', 'tafilalet
travaux', 'rouandi']",[]
travaux de construction d ' un centre de formation et d ' insertion professionnelle,agence régionale d ' exécution des projets région tta,encec,"26 757 858 , 00","['imilchil', 'sort', 'asdec', 'encec', 'badr beton', 'tafilalet
travaux', 'rouandi']","['imilchil', 'sort', 'asdec', 'encec', 'badr beton', 'tafilalet travaux', 'rouandi']"
appel d ' offres ouvert,office de la formation professionnelle et de la promotion du travail,neant,montant total,['liste des concurrents ayant déposé un pli'],['appel d ' offres infructueux']
fourniture de compteurs petits calibres 4 jet unique - type vitesse dn 40 mm,office national de l ' electricité et de l ' eau potable,n / a,n / a,"['aes pays', 'zinco']","['aes pays - 849 . 600 , 00 mad', 'zinco - 849 . 600 , 00 mad']
,,,,[],"['subject : acquisition', 'installation and commissioning of kitchen equipment for internships at the agadir', 'nador and laayoune cmcs of the ofppt project owner : office of professional and vocational training winner of the
contract : arcade equipement price of the contract : 1 , 865 , 400 dirhams competitors : aeria 2a , abrita 2a offers : 1 , 582 , 946 . 40 dirhams', '1 , 059 , 120 dirhams']"
the results of the open bid n ° 167 / 2020,the professional training and employment promotion office,benlhou freres,"58 679 652 , 19 dh",['- benlhou freres - tafilalet travaux -'],[]
travaux de construction de i ' institut spécialisé de technologie appliquée rissani,office de la formation professionnelle et de la promotion du travail,sobatran,"9 455 221 , 20","['sodata', 'scotim']",[]
results of the open call for bids n ° 167 / 2020,the professional training and employment promotion office,benlhou freres,"58 679 652 , 19 dh","['tafilalet travaux', 'oa']","['benlhou freres ( 58 679 652 , 19 dh ), tafilalet travaux
( unknown ), oa ( unknown )']"
travaux de construction de l ' institut spécialisé de technologie appliquée rissani,office de la formation professionnelle et de la promotion du travail,sobatran,"9 455 221 , 20","['sodata', 'scotim']",[]
"appel d ' offres pour l ' acquisition , installation et mise en service des équipements du secteur génie électrique destinés aux isia agadir et kenitra -",office de la formation professionnelle et de la promotion du travail -,amaskou
najah -,"74 880 , 00 -",['med industry partners -'],"['469 200 , 00']"
the results of an open bid,the moroccan government,master tec,"557 874 , 00 dirhams","['amaskou najah', 'bauner']","['471 360 , 00 dirhams', '417 360 , 00 dirhams']"
"the subject is "" appel d ' offres ouvert n ° 175 / 2020 ""","the project owner is "" office de la formation professionnelle et de la promotion du travail ""","the winner of the contract is "" metallifni ""","the price of the contrac
is "" 997 200 , 00 ""","['the competitors are "" 1', '2', '3 et 4 ""']","['the offers are "" 414 000 , 00 , 462 960 , 00 , 444 492 , 00 , 497 569 , 20 ""']"
aménagement de la ligne principale 22kv erfoud - rissani,office national de l ' electricité et de l ' eau potable,electad s . n . c,"6120 173 , 70","['lumafric sa', 'faraday sarl', 'electrou sarl']","['lumafric sa - 6299 673 , 63 ,
faraday sarl - 6120 896 , 70 , electrou sarl - 6134 060 , 00']"
bouclage des réseaux moyenne tension de la direction provinciale errachidia,office national de l ' electricité et de l ' eau potable,imadelec sarl,"1 304 472 , 40 dh - htva",['elecphone sarl'],"['1 223 755 , 90 dh - htva']"
mise en conformité du réseau mt de la dp ouarzazate,office national de l ' electricité et de l ' eau potable,electro tafoghalt sarl,"3584 231 , 04",['gritel sarl offer : 3 112 355'],[]
the results of the open bidding n ° 195 / 2020,the professional training and job promotion office,tecnitas,"250 345 , 47 dh","['verite controle', 'bureau veritas maroc']","['tecnitas ( 250 345 , 47 dh )']"
acquisition of equipment and tools for the teams of the errachidia provincial directorate,office national de l ' electricité et de l ' eau potable,n / a,n / a,['35 morocco sarl'],['n / a']
```

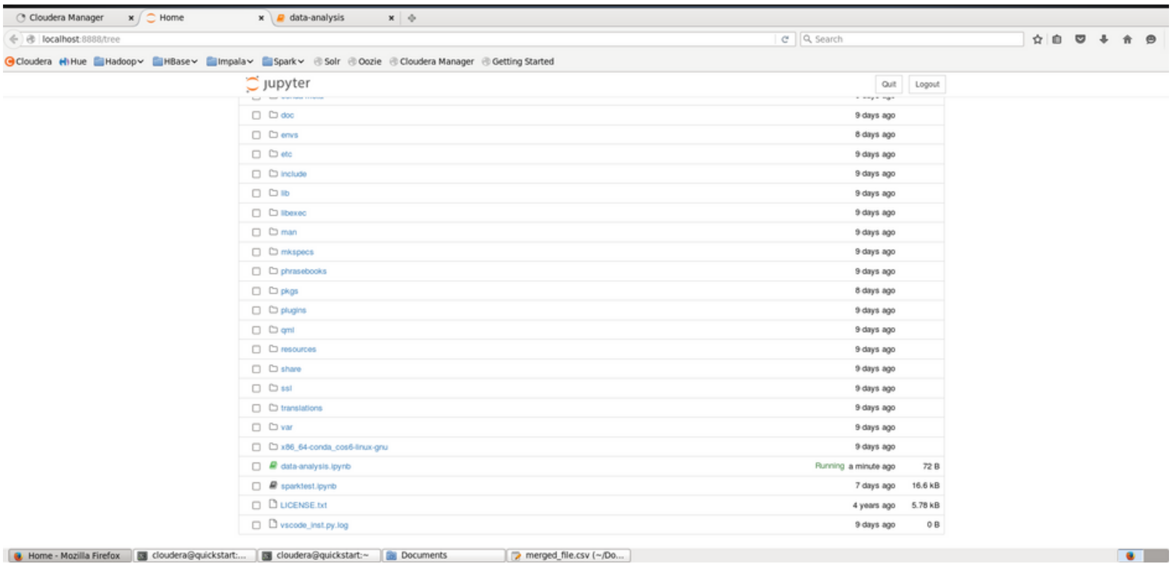
PROJECT DEMONSTRATION

2 / DATA ANALYSIS USING APACHE HADOOP



```
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 1 items
-rw-r--r--   1 cloudera cloudera     116499 2022-12-29 16:16 data
[cloudera@quickstart ~]$ jps
11156 Jps
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 1 items
-rw-r--r--   1 cloudera cloudera     116499 2022-12-29 16:16 data
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx   - hdfs  supergroup          0 2017-07-19 06:29 /benchmarks
drwxr-xr-x   - hbase supergroup          0 2022-12-09 19:42 /hbase
drwxr-xr-x   - solr  solr                0 2017-07-19 06:31 /solr
drwxrwxrwt   - hdfs  supergroup          0 2022-12-12 11:27 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2022-12-12 11:27 /user
drwxr-xr-x   - hdfs  supergroup          0 2017-07-19 06:31 /var
[cloudera@quickstart ~]$ hdfs dfs -mkdir /Documents/data
mkdir: `/Documents/data': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -mkdir /Documents
mkdir: Permission denied: user=cloudera, access=WRITE, inode="/":hdfs:supergroup:drwxr-xr-x
[cloudera@quickstart ~]$ hdfs dfs -ls /user
Found 9 items
drwxr-xr-x   - cloudera cloudera            0 2022-12-29 16:16 /user/cloudera
drwxr-xr-x   - hdfs     supergroup          0 2022-12-12 11:27 /user/hdfs
drwxr-xr-x   - mapred   hadoop              0 2017-07-19 06:29 /user/history
drwxrwxrwx   - hive     supergroup          0 2017-07-19 06:31 /user/hive
drwxrwxrwx   - hue      supergroup          0 2022-12-09 20:24 /user/hue
drwxrwxrwx   - jenkins  supergroup          0 2017-07-19 06:29 /user/jenkins
drwxrwxrwx   - oozie    supergroup          0 2017-07-19 06:30 /user/oozie
drwxrwxrwx   - root     supergroup          0 2017-07-19 06:29 /user/root
drwxr-xr-x   - hdfs     supergroup          0 2017-07-19 06:31 /user/spark
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera
Found 1 items
-rw-r--r--   1 cloudera cloudera     116499 2022-12-29 16:16 /user/cloudera/data
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/data
-rw-r--r--   1 cloudera cloudera     116499 2022-12-29 16:16 /user/cloudera/data
[cloudera@quickstart ~]$ hdfs dfs -put /Documents/merged_file.csv /user/cloudera/data
put: `/user/cloudera/data': File exists
[cloudera@quickstart ~]$ 
```

Finally we have proceeded to the data analytics using Pyspark in a Jupyter-notebook using the data we have stored in the hdfs :



□ 🗀 doc		9 days ago	
□ 🗀 envs		8 days ago	
□ 🗀 etc		9 days ago	
□ 🗀 include		9 days ago	
□ 🗀 lib		9 days ago	
□ 🗀 libexec		9 days ago	
□ 🗀 man		9 days ago	
□ 🗀 mkspecs		9 days ago	
□ 🗀 phrasebooks		9 days ago	
□ 🗀 pkgs		8 days ago	
□ 🗀 plugins		9 days ago	
□ 🗀 qml		9 days ago	
□ 🗀 resources		9 days ago	
□ 🗀 share		9 days ago	
□ 🗀 ssl		9 days ago	
□ 🗀 translations		9 days ago	
□ 🗀 var		9 days ago	
□ 🗀 x86_64-conda_cos6-linux-gnu		9 days ago	
□ data-analysis.ipynb	Running	a minute ago	72 B
□ sparktest.ipynb		7 days ago	16.6 kB
□ LICENSE.txt		4 years ago	5.78 kB
□ vscode_inst.py.log		9 days ago	0 B

CONCLUSION

In this report, we analyzed a large dataset of procurement information from "marchespublics.gov.ma" using a range of big data tools and techniques, including web scraping, data transformation, AI modeling, and data analysis.

These findings have several practical and theoretical implications for understanding procurement processes in Morocco.

While our analysis provides valuable insights into the procurement processes on "marchespublics.gov.ma", it is important to note that there were several limitations and challenges that we faced such as the storage limite.

Overall, this report provides a detailed analysis of procurement processes on "marchespublics.gov.ma" using big data tools and techniques, and offers valuable insights into the procurement landscape in Morocco. Our findings contribute to the field of big data analysis and have practical and theoretical implications for understanding procurement processes in the region.