Projet BigData NoSQL

Sujet

Le projet de cette année concerne l'analyse et l'exploitation d'annonces d'appartements à louer sur Airbnb dans la ville de Bordeaux.

2 objectifs se dégagent dans ce projet :

- proposer une analyse de données « descriptive » pour synthétiser les données (statistiques descriptives, visualisation de données)

- proposer un modèle de prédiction du tarif nuitée (l'objectif métier serait la recommandation d'un tarif pour un nouveau propriétaire qui veut mettre son appartement à louer sur le site)

L'analyse des données brutes vous permettra de constater les éléments suivants :

- Les types des variables sont hétérogènes :

    - Il faudra traiter des variables « textuelles » : titre de l'annonce, résumé de l'annonce par exemple.

    - D'autres variables plus classiques sont également présentes: numériques, catégorielles.

- Il faudra gérer la problématique des données manquantes

En plus de la problématique "Machine Learning", il faudra scripter et implémenter les transports de données décrits dans les étapes suivantes :

Etape 0

Les données doivent être installées sur HDFS, sur une VM Hadoop (ex : HortonWorks Sandbox de Cloudera) téléchargée et installée sur un Virtual Box local.
Vous proposerez une première exploration et projection des données (en utilisant l'outil et le format de votre choix, par exemple ceux abordés lors du cours de DataViz) afin de visualiser des caractéristiques de ces dernières permettant une première analyse visuelle.

Etape 1

Ces données sont rapatriées en local, via un script dédié, qui les récupère depuis HDFS.

Etape 2

Ces données sont poussées sur une VM dans le cloud AWS. La nature des données étant sensible, vous devez au préalable vous poser les questions de sécurité à mettre en œuvre sur la VM. Par ailleurs, les données poussées sur la VM doivent être chiffrées en amont.

Etape 3

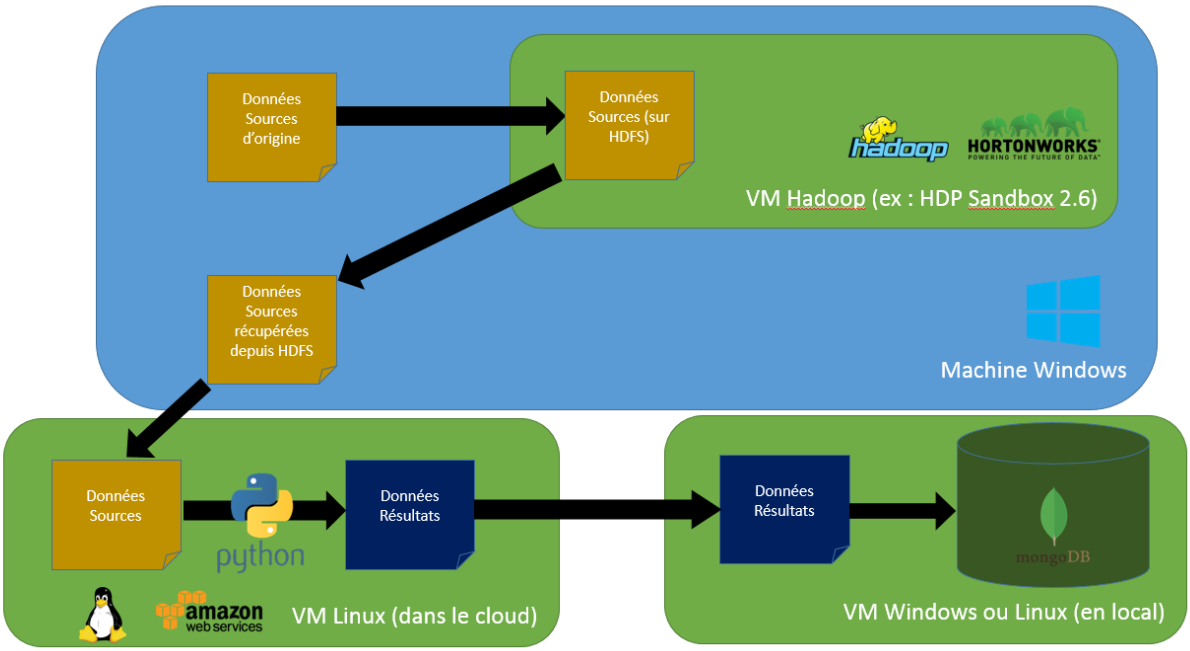
Un modèle d'apprentissage est appris à l'aide d'un des algorithmes de machine learning vu en cours ainsi que des données fournies. L'algorithme est exécuté sur cette VM dans le cloud AWS. Le choix de l'algorithme utilisé devra être justifié.

Etape 4

Vous utiliserez une partie des données pour créer un fichier predict.csv contenant uniquement les informations sur les individus, c'est-à-dire sans la catégorie de poste (assurez vous que ces données n'ont pas été utilisées au moment de la phase d'apprentissage). Vous exécuterez sur la VM AWS votre modèle appris afin de prédire la catégorie de poste de chaque individu du fichier predict.csv. Un fichier de résultats sera créé en concaténant chaque ligne de predict.csv avec la prédiction de votre modèle. Ce fichier sera sauvegardé sur le FileSystem de la VM AWS, au format CSV.

Etape 5

Les résultats de l'étape 4 sont alors récupérés et chargés dans une base NoSQL (par ex. MongoDB, choix le plus approprié à faire par l'équipe) s'exécutant en local (sur une VM ou sur l'OS local), via un script dédié (langage ou outil à déterminer par l'équipe).
Vous utiliserez D3.js afin d'afficher graphiquement vos résultats (tout ou partie) et d'appuyer votre analyse et le message que vous souhaitez faire passer.



Modalités

Equipes

Le projet se fait par équipes de 3 (6 équipes de 3).
Lors des soutenances, les questions pourront être posées indifféremment aux différents membres du groupe. Cela signifie que chacun doit avoir une maîtrise a minima des différentes parties du projet.

Livrables

Soutenance intermédiaire

Le 30/03/2022 auront lieu les soutenances intermédiaires : présentation par l'équipe :

- de la compréhension du sujet, des exigences et des attentes des différentes étapes
- d'un 1er niveau d'analyse des données
- du travail réalisé
- le tout devant le jury des évaluateurs. 20 min (10 de présentation + 10 de questions) par soutenance.

Soutenances finales

Le 31/05/2022 (après-midi) auront lieu les soutenances finales : présentation par l'équipe du travail réalisé devant le jury des évaluateurs. 30 min (20 de présentation + 10 de questions) par soutenance.

Livrables finaux

Le 31/05/2022 (avec la soutenance finale) devront être remis les livrables finaux :

- Document de soutenance
- Repository Github avec les codes sources réalisés pendant le projet (scripts, notebooks…)
- ReadMe file (dans le repo Github) expliquant le projet
- Tout type de documentation technique permettant de mieux comprendre le travail réalisé

Notation

- Soutenance intermédiaire (sur 20)
- Soutenance finale (sur 20)
- Livrables finaux et gestion de projet (satisfaction Product Owner) (sur 20)