

|  |
| --- |
| Rapport projet DW :  Listing sur Airbnb |
|  |
| Encadrée par : Prof. Hilal Imane  Réalisé par : - Alaoui Mdaghri Ahmed - Khoulti Wissal |



# Introduction générale

|  |  |
| --- | --- |
| Onze ans plus tard, le site d'Airbnb répertorie plus de six millions de chambres, appartements et maisons dans plus de 81 000 villes à travers le monde. En moyenne, deux millions de personnes se reposent la tête dans une propriété Airbnb chaque nuit – un demi-milliard depuis 2008.  Londres, Paris et New York ont ​​le plus grand nombre d'annonces, mais des logements Airbnb sont disponibles à Mandalay, Oulan-Bator et Brazzaville. | |
| *« Some residents in areas with a big Airbnb presence claim the business is hollowing out communities»* | |
| Qui veux dire que l'extraordinaire succès d'Airbnb n'a pas été accueilli sans réserve. Certains résidents des zones à forte présence d'Airbnb affirment que l'entreprise creuse les communautés en forçant les loyers et en limitant la disponibilité pour les personnes à la recherche de locations à long terme, et en important un grand nombre de touristes qui manifestent peu d'intérêt pour la courtoisie envers leurs voisins temporaires.  En partant de ces faits on voudra créer un entrepôt de données regroupant les différents référencements et aussi les différentes villes et leurs pouvoir d’achat et après visualiser des index portant sur le prix de logement et le pouvoir d’achat d’une zone aussi visualiser l’emplacement des référencements selon des critères ou bien aussi le nombre de référencement par nombre de personne. Conception de l’entrepôt de données *La collecte de données*  La collecte de données ce fera d’après deux sources Airbnb[Inside Airbnb](http://insideairbnb.com/get-the-data.html) et aussi Kaggle[*World Cities Population and Location*](https://www.kaggle.com/i2i2i2/cities-of-the-world)*,* [*Cost of Living Indices*](https://www.kaggle.com/debdutta/cost-of-living-index-by-country)***.***   * Pour les données provenant d’Airbnb on aura besoin de scrapper les données du site web avec les différents noms de ville et le listing appropriés. * En ce qui concerne Kaggle on pourra télécharger les données directement en ayant un compte Kaggle.   *Les bases de données*  Les deux bases de Kaggle sont les suivant :   * Cost\_of\_living\_index :      * Cities15000 :     Les données scrappées de Airbnb suivent le format « ville » «date listing » :    Avec chaque lisiting sous le format suivant : | |
| ***Les outils utilisés***  *Python*  Utiliser pour le data scrapping depuis Airbnb et l’extraction de données depuis kaggle pour ensuite être insérer dans Snowflake  *Snowflake*  Snowflake est un entrepôt de données construit au-dessus de l'infrastructure cloud Amazon Web Services ou Microsoft Azure, et permet au stockage et au calcul d'évoluer indépendamment.  *DBT*  dbt est un outil de ligne de commande open source qui aide les analystes et les ingénieurs à transformer plus efficacement les données de leur entrepôt.  *Metabase*  Metabase est un outil de business intelligence open-source. La métabase vous permet de poser des questions sur vos données et affiche les réponses dans des formats logiques, qu'il s'agisse d'un graphique à barres ou d'un tableau détaillé. Vous pouvez enregistrer vos questions et les regrouper dans de beaux tableaux de bord. |
| ***Conception du DW***  Notre DW va contenir deux bases de données RAW et ANALYTICS  Avec 3 Warehouses LOADING, REPORTING et TRANSFORMING  Dans notre premier warehouse LOADING on va pouvoir alimenter notre base de données RAW avec les données reçu après opération de collecte  LE deuxième warehouse servira pour la transformation des données de leurs formats brute en leurs formats final bien structurer pour réaliser notre DASHBOARD  Le troisième servira pour les données qui nous servirons pour réaliser notre reporting et aussi stocker sous la base de données ANALYTICS  *Le DW suivra le schéma suivant :*    *Avec les deux tables autres que les dimensions et la table de fait étant des tables relatifs à la visualisation, par exemple la table flagged\_analytics contient house\_flag qui prend une valeur booléenne selon la valeur de vérité du fait que le listing propose une maison ou non, d’une autre part most\_recent\_analytics regroupe tout listing récent.*  *Sous Snowflake on crée les bases de données + Warehouses pour être alimenter depuis le code python après extraction de données des sources*    La base de données RW contient les tables :    Après alimentation de notre base de données avec les éléments brute on passe à l’ETL à l’aide de l’outils DBT pour réaliser les transformations suivantes    Après exécution le résultat est le suivant dans la base de données ANALYTICS :    Et finalement passer à l’outils METABASE pour réaliser le DASHBOARD suivant : | |
|  | |
|  | |