```
                                                                    ┌──────────────────────┐
                                                                    │   K-means            │
  ┌─────────┐                                                       │   clustering         │
  │Database │                                                       │   25000 clusters     │
  └─────────┘                                                       └──────────────────────┘
       │                                                                        │
       ▼                                                       ┌────────────────┼────────────────────────┐
  Cleaning/Removing                                            ▼                ▼                         ▼
  non-useful                                              ┌─────────┐      ┌─────────┐              ┌─────────┐
  data                                                    │Cluster 1│      │Cluster2 │              │Cluster n│
       │                                                  └─────────┘      └─────────┘              └─────────┘
       ▼                                                       │
  ┌─────────┐  Splitting into 4                                ▼
  │ Clean   │  equal sets for                            ┌──────────┐   ┌──────────────┐   ┌──────────────┐
  │ data    │  parallel processing                       │Images of │   │VGG model     │   │FInd pairs of │
  └─────────┘                                            │all products│→ │returns 64    │ → │imageswith least│
       │                                                 │in cluster│   │dimensional   │   │Euclidean     │
  ┌────┬────┬────┬────┐                                  └──────────┘   │embeddings of │   │distance      │
  ▼    ▼    ▼    ▼                                                      │each image    │   └──────────────┘
┌────┐┌────┐┌────┐┌────┐                                               └──────────────┘           │
│ D1 ││ D2 ││ D3 ││ D4 │                                                                           ▼
└────┘└────┘└────┘└────┘                                                                      Hard compare
  ▼    ▼    ▼    ▼                                                                             mrp,brand, color,
['title']['title']['title']['title']                                                          size, seller
  │    │    │    │                                                                                 │
┌─────────────────────────────────┐  Using Doc2Vec                  Above threshold of matching    │
│Vectorise Vectorise Vectorise     │  model trained                       │
│          Vectorise               │  befrehand                           ▼
└─────────────────────────────────┘                                ┌──────────┐
       │                                                           │Duplicate │
       ▼                                                           └──────────┘
  ┌─────────┐
  │ n rows  │                                                      ┌──────────┐
  │500 columns│                                                    │Store in JSON│
  └─────────┘                                                      │format    │
       │                                                           └──────────┘
       ▼
  ┌─────────┐
  │PCA  n x 10│
  └─────────┘
       │
       ▼
  ┌─────────┐
  │t-SNE n x 3│
  └─────────┘
```