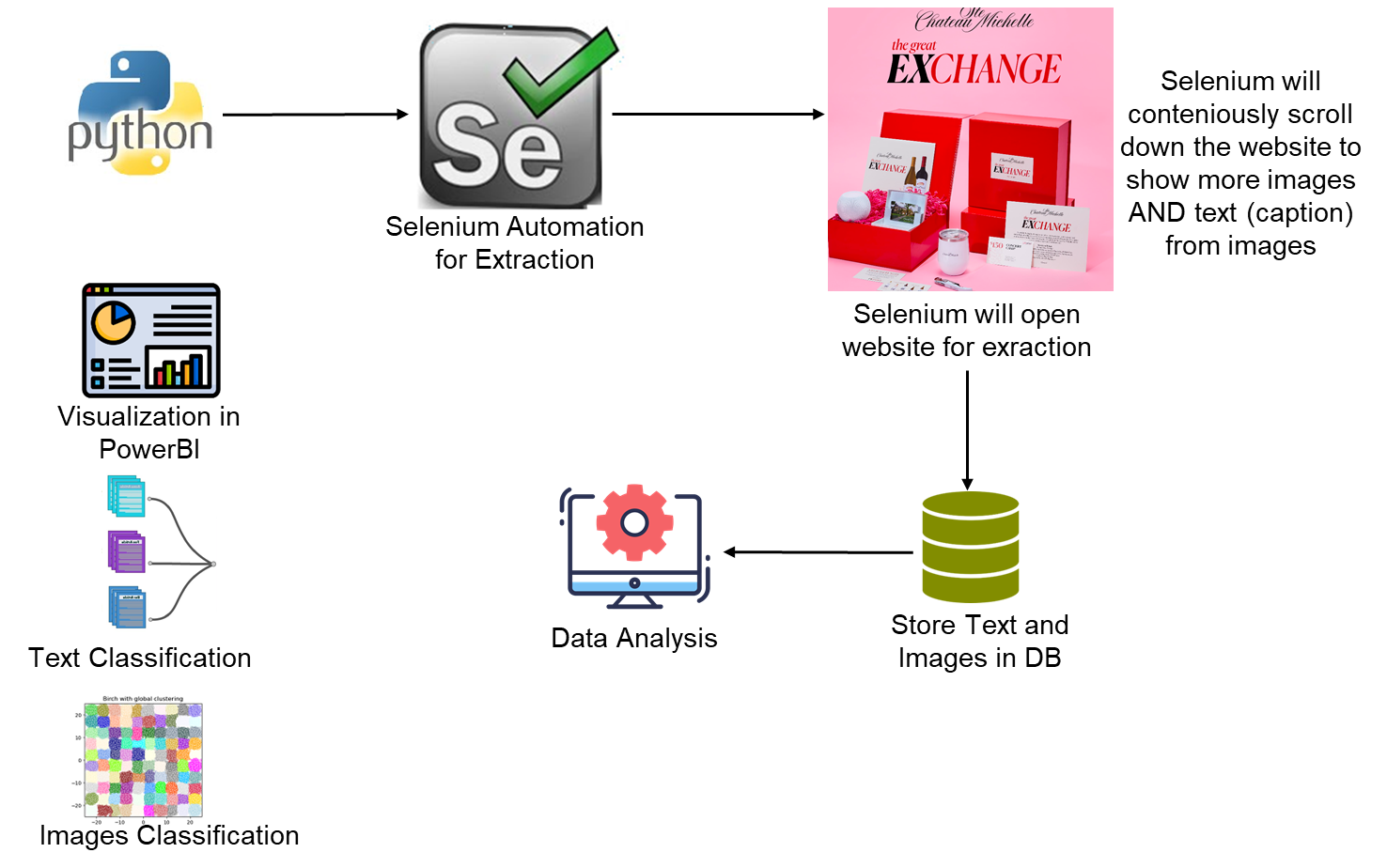
**Proposal for Images/ Text Extraction and Classification/ Clustering**

Following image is the over-all proposed system along with sub-modules.



There are following main modules in this system.

* **Data Scrapping**
  + The website is dynamic, like you have few images and text in first page, as you scroll down the page, you will see more images. So we need to automation of the page to scrap data.
  + **Technical Requirements**
    - Python Selenium Framework (for automation and scrapping data)
  + **Expected Output**
    - Image caption and images in CSV file
* **Text Classification**
  + **Method 1**
    - If we have labelled data (currently not available – we have to manually label the text in categories like JEWELARY, CLOTHES, WATCHES, etc.)
    - We will use NLP + Machine learning to classify text in to categories.
    - We will label the text according to ML algo.
  + **Method 2**
    - If we don’t have labelled dataset, we can use clustering machine learning algorithm like K-Means Clustering.
    - We will have different clusters.
    - We have to manually check and name each cluster.
* **Image Classification**
  + **Method 1**
    - If we have labelled data (currently not available – we have to manually label the image in categories like JEWELARY, CLOTHES, WATCHES, etc.) – there are free libraries but not very strong. Accurate libraries are paid, hence finding an open source accurate library needs sometime to explore.
    - We will use Computer Vision and Machine learning to classify images in to categories.
    - We will label the images according to ML algo.
  + **Method 2**
    - If we don’t have labelled dataset, we can use clustering machine learning algorithm like K-Means Clustering.
    - We will have different clusters.
    - We have to manually check and name each cluster.
* **Visualization Module**
  + Create PowerBI dashboard
    - Main Topics
    - Frequency of items

