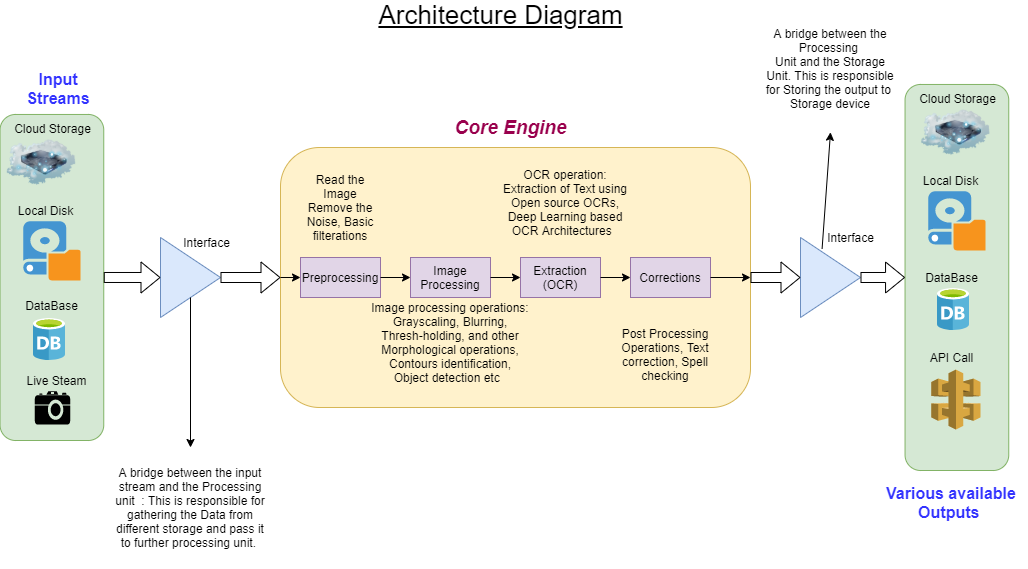
***Text Extraction using Computer Vision Techniques***

**Architecture Diagram:**



**Description:**

**Input Stream**: The data can be available in any of the places mentioned in the diagram.   
Cloud Storage, Local Disk, RDBMS, or it could be the live stream images coming from various sensors.

**Interface:** This is a bridge between input Streams and the processing unit. It is responsible for reading the data from various storages and pass it to next stage

**Core Engine:** It is responsible for extracting the data from the image**.** It involves following processing units  
1. *Preprocessing:* Basic Image processing operations, Noise removal

2. Image Processing: more advance Image Processing operation like blurring , thresh-holding, Object Localization, object detection, Morphological oprations like dilation, erosion, closing , opening, Contours identification etc are applied to get the ROI from the Image.

3. Extraction(OCR): Once the ROI is detected or image is properly processed, it is then forwarded to OCR extraction for extracting text from processed image. Pytesseract is open source OCR platform which gives good result of processed image (less noisy). User can build their own custom Deep neural network architecture to do this task.

4. Correction: Once the text is extracted it is then fed to next unit for text cleaning, spell checking. It requires the technical data corpus in background to perform the Spell-Checking operation. To make the spell checking perfect, user have to create the domain specific text corpus. Right now, the generalized English text corpus is used for spell checking. The file is stored inside template folder(name: big.txt).

***Output:*** Once the text is corrected, user can store it in various options, Cloud, Local disk, RDBMS, or it can be used as an API call to another process and can be consumed there.