## Question 4: Exploiting semi-structured and unstructured data

Being in the healthcare field, Dell Seton Medical Center will have a load of unstructured data. Specific unstructured data sources for Dell Seton Medical Center include written text, radiological images, and doctor audio files. Radiological images are important data since they can be used in the future for the prediction of various diseases. Mainly written text is doctor's prescription. This data is quite useful to store to automate the process of ordering medicines and later run analysis on it as well. Most importantly, the ability to analyze unstructured data plays a pivotal role in the success of big data in healthcare settings since 80% of health data is unstructured. Dell Seton Medical Center will need to run image recognition and image segmentation models to extract information out of the radiological images. Dell Seton Medical Center will need to run NLP models to extract written text information. After unstructured data has been gathered across multiple healthcare units, it can be stored in a Hadoop distributed file system and NoSQL database that can maintain it until it can be called up in response to users' requests. NoSQL databases support the storage of both unstructured and semi-structured data from multiple sources in multiple formats in real-time.