PROPOSAL FOR PATENT DATABASE USAGE

On reading earlier posts by fellow interns and doing a little research by myself ,as a data engineer ,I would propose that the clustering algorithm (and NLP) should be executed by the Spark Execution Engine preferably by deploying it on Google Cloud Platform and creating a Dataproc instance(GCP version of Spark Framework).

The major benefits of utilizing,running and maintaining this database and its proposed (NLP or clustering) code on Cloud are given as follows :-

- By providing important credentials, all project members can work simultaneously and also test their codes together.This shall improve speed and thus the latency.
- By using Spark Streaming (or Apache Storm,Flint),many of these datasets can be parallelly updated (because of new patents added on individual websites) and also new training models thus can successfully undergo new iterations to improve the efficiency of the clustering algorithm.(Thus streaming such a huge data and its regular timely updated dataset can be extremely useful)
- The connection speed for these cluster would be much better than any other cluster and the results to these analysis could be easily made available in public domain by giving public links of buckets stored in the Cloud Storage.
- Security (an important issue) will be at its apex point because of the security provided by the Google Datacentre itself.
- Service and Maintenance : No maintenance required in case of cluster breakdowns.(it becomes responsibility of Google to manage that)

The database and code can also run on Databricks which is nearly 10X faster than Spark and also provides for streaming of data.

Seeing may be in future if the workload for these patent websites increases to huge level,it would be better to make use of distributed file system and their framework which can handle the big data generated everyday.