GDELT Storage Plan
——————————————————

- Schema-on-read (ELT) approach
- Azure Blob storage as data lake w/ HDFS as underlying format. (Databricks?)
- Distributed Spark setup, deployed on YARN for data processing (Spark references HDFS locations natively)
- Apache Kafka w/ Zookeeper - For handling real-time data feeds. Used to publish and subscribe to GDELT topics.
- Apache Parquet - compression schemes specified on per column level, stores data schema with the data itself.
- Apache Avro - serializes data in compact, binary format.
- Elasticsearch - for full-text search of GDELT data.
- Kibana - for analytics and visualization.
- Accumulo - NoSQL database, based on Google's Bigtable design, offers bulk loading, parallel reading, iterators, and cell level security.