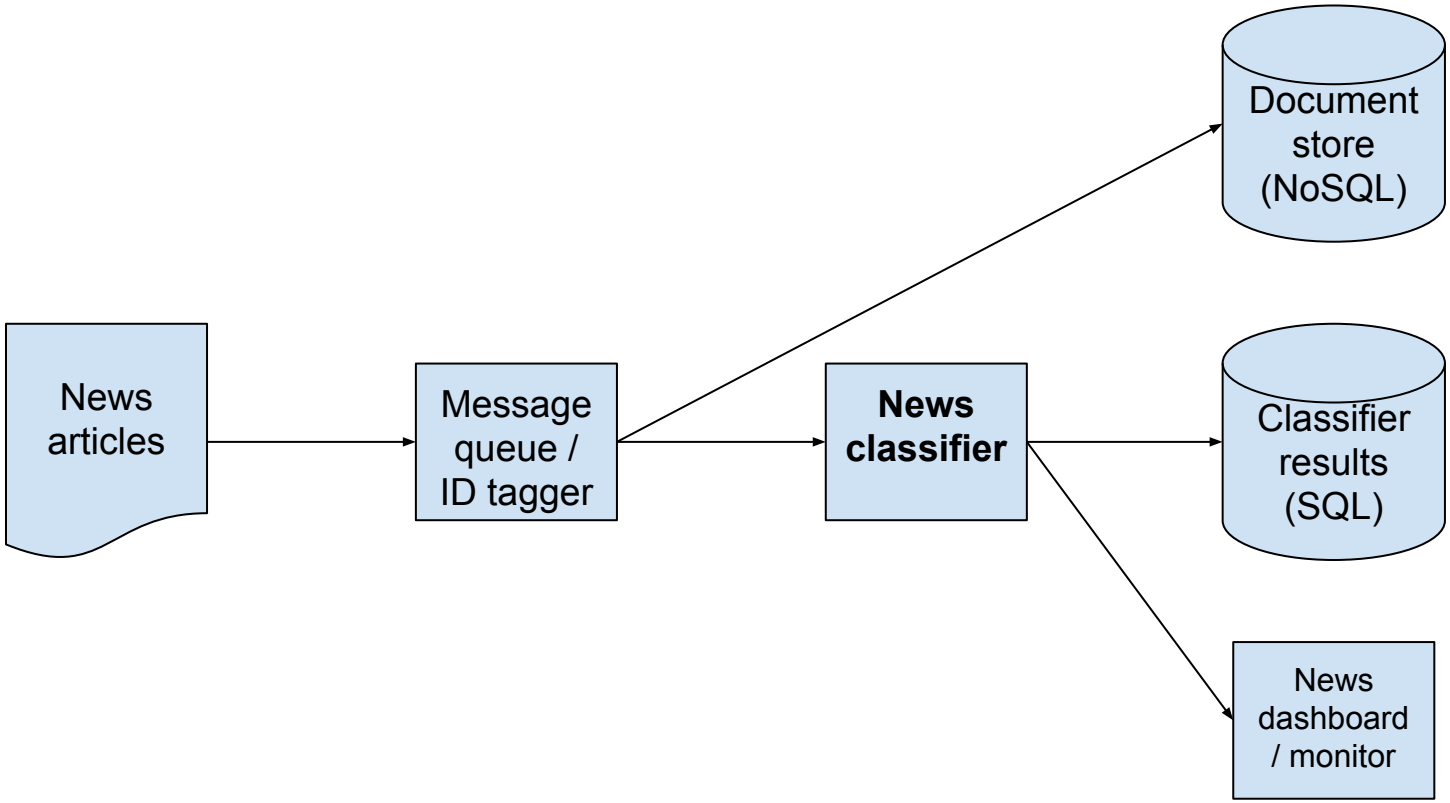

This architecture assumes that news articles would be received in a streaming manner from a wide variety of sources, especially news websites and Twitter. It also assumes that there is a downstream need for both real-time awareness and retrospective analysis.

The news classifier could be put into a Docker container and scaled up to run across multiple machines if needed, but its heavier cost is for training. To that end, it could actually be refactored as an Apache Spark analytic. This would make training faster and more scalable, and could easily be converted to a streaming analytic for inference. Ideally, we would add to the 20 newsgroup dataset with other pertinent labeled news articles over time.