The pipeline that I am discussing here is the Yelp real-time pipeline. This pipeline can be described in terms of the basic components of pipelines as follows:

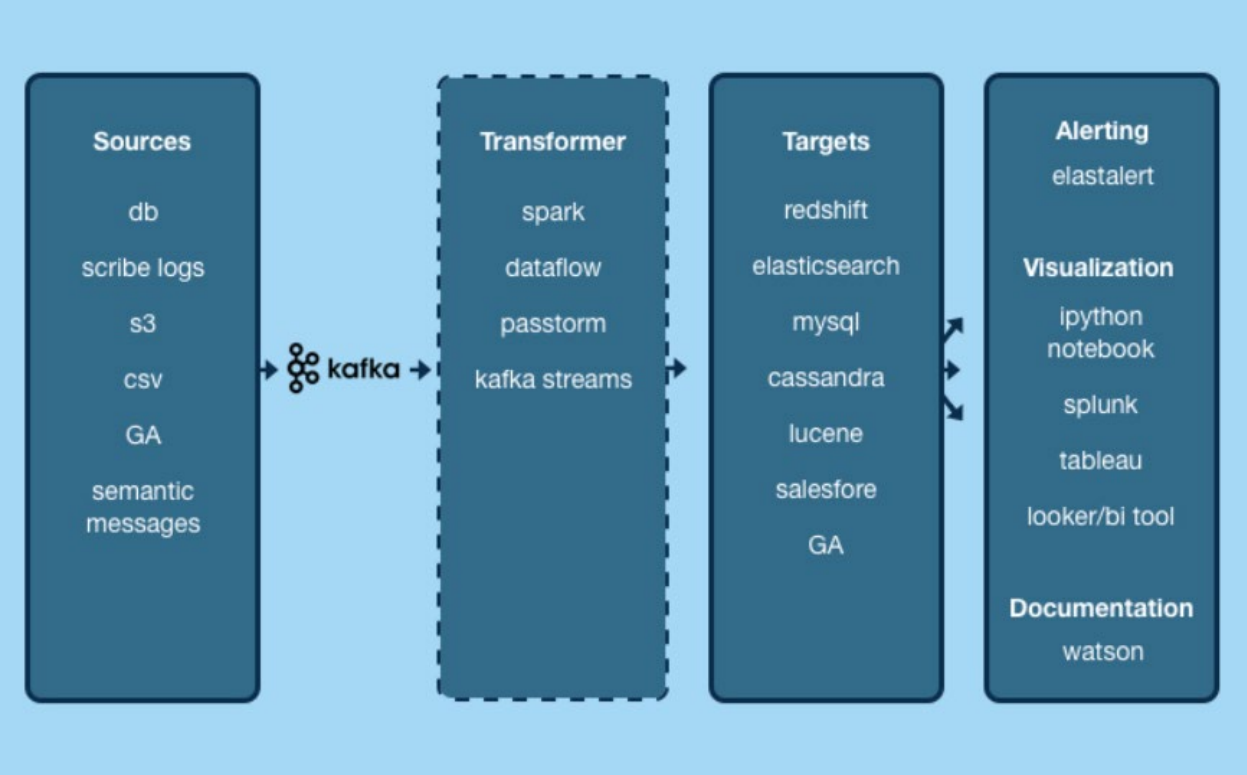
Data Sources and Types: The major source of data on yelp is the number of reviews that are up on it every single day. The input files could be db, s3 or csv files. There are about 80 million reviews on yelp within a day. The data although seems to be small, is actually very large in size. This large data must be stored in appropriate formats for use and preservation. This process is discussed in the next section.

Data Storage Technologies: The size is data is very large, hence a few advanced techniques must be used to store the data. This ensures that the data is stored in a safe and an easily accessible method, as well as proper compression techniques are applied to it. The data for yelp is stored using Apache Kafka file systems. The file can be accessed as a JSON object. The kafka file is stored in binary object format, hence the data contained in the file could be anything without any restrictions.

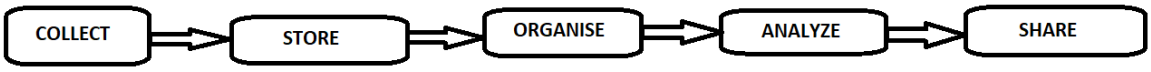
Data Transformation Tools and Techniques: All messages are guaranteed to be published with a pre-defined schema, and the schemas are guaranteed to be registered with the schema store. Data Pipeline producers and consumers deal with data at the schema level, and topics are abstracted away. Schema registration is idempotent, and registered schema are immutable. The data transformation is done using the tools such as spark or kafkastream. The services are performed before the data is sent for actual analysis.

Other Major Components: The yelp pipeline also supports other components of the pipeline such as Visualization and Documentation. Documentation is majorly done via Watson. Visualization on the other hand can be done using a lot of available tools including (but not limited to) PowerBL, tableau and iPy Notebook.

The General Pipeline Architecture of Yelp Pipeline can be shown as follows:



What Steps are followed in the chosen Pipeline:



The pipeline follows all the major steps that are mentioned above. The data is collected from the review that are put up on the website. These reviews are converted into a binary object format before being saved into the kafka file format. Then, a JSON object is created for easy access to the data so that only the necessary parts of the data are retrieved at a very fast time scale. The conversion of data to binary object format helps to store the data more effectively while taking up less space. The data is then made available to the user to apply the proper analysis technique to get the results for the desired queries. The data is then processed upon and the conclusions are drawn. The conclusions from the data can either be stored as a document using the Watson technology or can be viewed visually using the tools such as powerBL, Tableau or iPy Notebook.

What Steps are not included in the Pipeline:

The intricate steps of filtering the data in order too detect and mitigate the fallacies is an important step during any data processing operation, which has not been given proper importance in this pipeline. Interpreting the data using various models and comparison between their accuracy rates should have been a major part of this pipeline. Also, the data pipeline focuses very less towards preservation of the data for further use or re-use for different applications. The re-use of data for other applications is one of the necessary steps for data analysis, which is not given proper thought to, in this pipeline model.