Literari.ly

Ankith Gunapal, Sandip Palit and Adam Lenart

December 13, 2016

1 Summary

Application idea Literari.ly is a literary assistant that helps users evaluating the popularity of a queried expression. Popularity is defined by (i) the observed frequency of the expression in published volumes (books and journals) and (ii) the rating that these books received. Moreover, the rating and the number of volumes which contain the queried expression can be partitioned into categories. Furthermore, literari.ly makes future predictions about the frequency with which the queried expression is going to be used.

Data sources The Google N-grams dataset is used for describing the time trends of the queried expression. The N-grams dataset includes terrabytes of data of English and several other languages 1 to 5-gram expressions. Currently, we focus on a 3.5 TB English bigram set of the data and use a 15 GB subset of it for the examples. The ratings and categorizations of the user queries stem from Google Books and Goodreads APIs. The Google Books API contains categories for the expression which occurs in the meta-information of the book as well as ratings of these books by Goolge Books users. Similarly, the Goodreads API holds information on the ratings of the books by Goodreads users. These two data sources can be joined by common ISBN13 number.

Architecture The data intensive processes, stored in HDFS, run on Apache Spark and the smaller tables containing aggregates from a PostgreSQL database are loaded into either Python or R for serving the user. R was chosen for visualizing data on its Shiny interface on the internet using an Amazon AWS instance and for forecasting the future popularity of the queried expression. Python was used to interact with Goolge Books and Goodreads APIs. Search history is visualized by a Kibana dashboard.

Services used Google bigrams data is stored in an Amazon S3 bucket, queries from Google Books and Goodreads are stored in an Amazon PostgreSQL RDS and the visual interface of the application along with a Kibana dashboard runs on Amazon AWS instances.

2 Architecture

The application relies on three data sources: Google N-grams, Google Books API and Goodreads API.

Google N-grams data contain information on expressions with different gram lengths stored in compressed CSV files separated by the the starting two letters of the first expression from aa to zz. Presently, we use a 3.5 GB subset of the data, the English bigrams. These tables are stored in an Amazon S3 bucket, currently a 15 GB part of it, the bigrams starting with the "ab" sequence are loaded into HDFS on an Amazon EC2 instance.

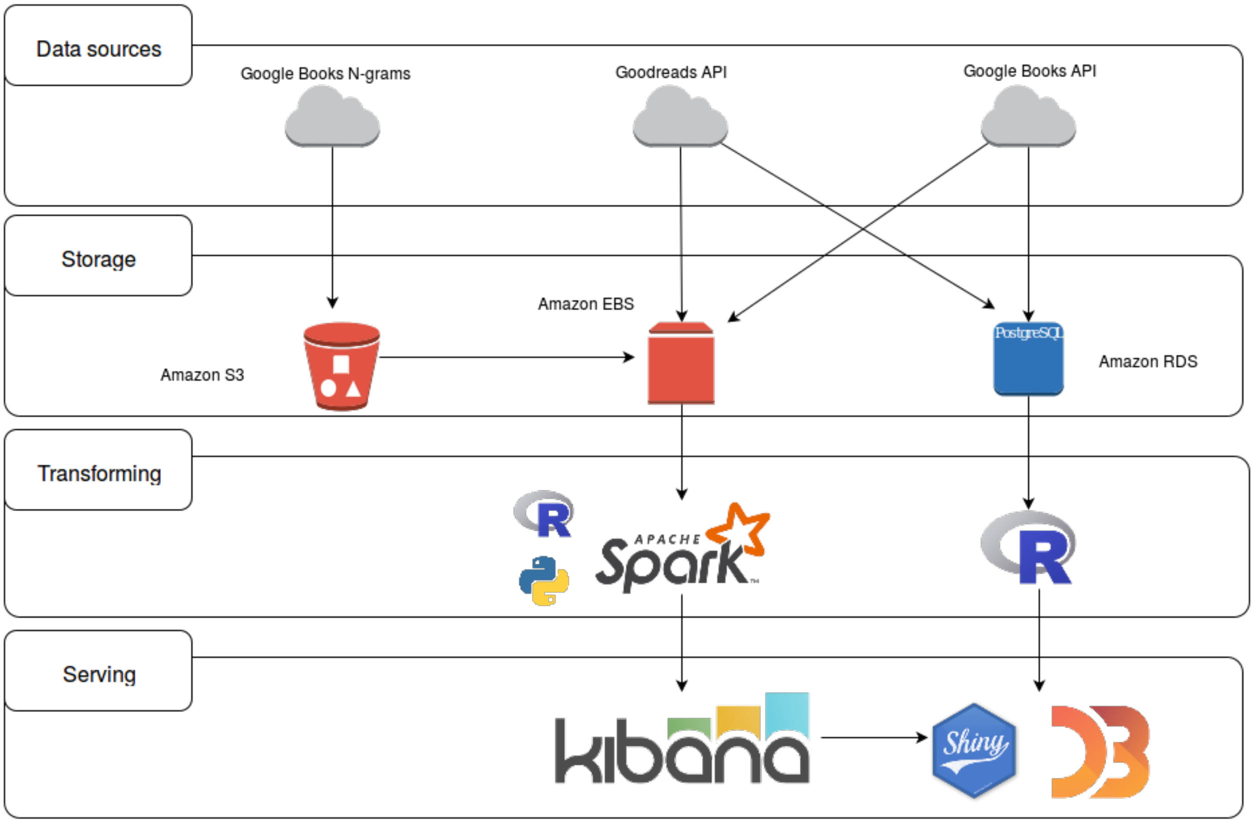

Figure 1: High-level overview of the architecture of the application