KTH Royal Institute of Technology
Stockholm

School of Electrical Engineering and
Computer Science

Data-Intensive Computing - ID2221

Project Proposal

Author
Emil Ståhl

Author
Selemawit Fsha

September 30, 2021

Project Proposal - ID2221

Emil Ståhl and Selemawit Fsha Nguse

September 30, 2021

1 Problem statement

In this project, we are going to analyze Wikipedia traffic and compare the data over time. How is the traffic changing over time? Day-to-day as well as year over year. Furthermore, we are going to compare the Wikipedia traffic data with data obtained from Google trends. Is there any correlations between the two?

1.1 Potential extensions

If time allows, we are going to extend this project by analyzing the complete Wikipedia database of articles. How do articles link to each other?

2 Data

In this project, we are going to make use of the Wikimedia pageviews dataset.[1] For obtaining Google trends data, we are making use of the pytrends library for Python.[2] The Wikipedia data-set of articles is available at meta.wikimedia.org.

3 Tools

The tools utilized for this project include:

- Scala
- Python
- Spark
- Spark Streaming
- Kafka

[1] https://dumps.wikimedia.org/other/pageviews/

[2] https://pypi.org/project/pytrends/

- HBase
- Neo4j

4 Methodology

The approach to this project can be summarized by the tasks below:

1. Retrieve the data and set up Kafka to read from the source
2. Use Spark Streaming to fetch data continuously from the Kafka message broker
3. Analyze the data with Spark
4. Write it back to Kafka
5. Store in distributed database