Ciência de Dados em Larga Escala

Inês Dutra and Zafeiris Kokkinogenis

DCC-FCUP
room 1.31
ines@dcc.fc.up.pt
zafeiris.kokkinogenis@gmail.com

March 7th, 2024

Practical Class # 2: Stream processing

- Task 1: basic use of Google Pub/Sub (Publisher/Subscriber streaming model)
    - Go to the Pub/Sub basic tutorial and try it on.

Practical Class # 2: Stream processing

- Task 2: basic use of Google Pub/Sub: copying text
  - Go to python streaming with GCP
    - repeat the exercise that uses "cat" of some file and pull the messages (you will need to run the commands in two different shell)
    - Try the command line to run the DirectRunner with `streaming_wordcount.py`. Notice that this program reads from a topic and writes to another topic. If you want to see results (counters), you may need to change the program and write to a text file

Practical Class # 2: Stream processing

- In order to run the DirectRunner as in the example below:

  ```
  python3 -m apache_beam.examples.streaming_wordcount \
     --input_topic projects/cdle2324/topics/input-topic \
     --output_topic projects/cdle2324/topics/output-topic \
     --streaming
  ```

  You need to follow the steps:
  - download your credential key from your project in the GCP (json format). Instructions here.
  - Initialize the environment variable `GOOGLE_APPLICATION_CREDENTIALS`:

    ```
    export GOOGLE_APPLICATION_CREDENTIALS= \
    <pathtoyourkeyfile/keyfilename.json>
    ```
  - you can obtain the path to your topic channels by using `gcloud pubsub topics list`
  - don't forget to create your topics
- now you are ready to execute the python line above

Practical Class # 2: Stream processing

- Task 3: streamed wordcount
  - Will you be able to change your own `wordcount.py` program (written last practical class) to run using stream processing?
  - follow instructions given here
    (do not use a very large file)

Practical Class # 2: Stream processing

- Task 4: profiling
  - Go to this site, apply the techniques to your own wordcount program and try to understand the sources of overhead