Project #2

Assignment #4 – Data Processing with Spark Streaming

7th December 2023

Advanced Infrastructures for Data Science

Pedro Neves, pedroneves@dei.uc.pt, 2023/2024

Master in Data Science and Engineering (MDSE) Course

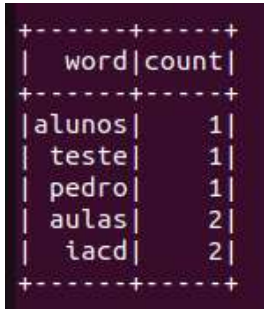
Task #1 – Obtain and sort the Word Frequency in the Command Line

Assignment #4

Context: Stream words in command line

Objective: What is the word frequency streamed in the command line?

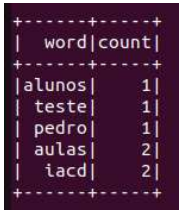
- Type some lines of text, and the word count will be updated in real-time in the console where the PySpark application is running



Task #1 – Obtain and sort the Word Frequency in the Command Line

Assignment #4

- PySpark Streaming job **hints**
  - Sets up a streaming source using the socket format, which listens for incoming data on a specified host and port (e.g., localhost:9999)
  - Read data from the socket (localhost:9999) and perform a word count in real-time
  - Streaming query must output the results in complete mode and to the console



- Tool to feed data into the socket
  - **netcat** server to feed data into the socket
  - netcat -l 9999

Good
Work