UPPSALA
UNIVERSITET

DEI - Project

Group 7

Contents

- What dataset have you chosen to work with, and what is the scientific background of the data?
- What is your tentative plan for the architecture and computational experiments?
  What technology will you work with, and how will you design your scalability studies?
- What questions will you answer using the specific dataset?
- Present any preliminary results.

Dataset chosen: Reddit comments

- **Why?**
  - Large dataset in JSON format (18-19 GB)
  - Analyze engagement of users online to find trolls using an algorithm

What questions will you answer using the specific dataset?

- Find out who’s the most frequent reddit author in every subreddit and what does this person write? (AKA who shitpost the most!)
- The idea is the most active person who posts lots of unrelated posts is not engaging in the community in a “good” way

Dataset Format

- **author: string (nullable = true)**
- body: string (nullable = true)
- normalizedBody: string (nullable = true)
- **content: string (nullable = true)**
- content_len: long (nullable = true)
- **summary: string (nullable = true)**
- summary_len: long (nullable = true)
- id: string (nullable = true)
- **subreddit: string (nullable = true)**
- subreddit_id: string (nullable = true)
- title: string (nullable = true)

Semi-structure and good for dataframes

Source: https://zenodo.org/records/1043504#.Wzt7PbhXryo

Tentative plan

- **What is your tentative plan for the architecture and computational experiments?**
  - Use hadoop to store the data.
  - Use Apache Spark to execute code up to 4 worker nodes.
- **What technology will you work with, and how will you design your scalability studies?**
  - Work with apache spark
  - Will design scalability by changing the number of cores that can be used (strong scalability)
  - Theoretically can horizontally scale as well by take worker nodes offline or add more worker nodes

Technologies and design plans

- Apache Spark for large scale data processing
- Hadoop stores the data on a distributed system
- Time testing for evaluation
- Tentatively add other metrics to measure resource utilization to assess scalability such as adding more drivers to try and overload the system

Improving scalability

- Adding more worker nodes
- Change number of drivers

Improving local performance

- Changing number of cores

Preliminary Results

- We have the hadoop cluster up. 1 datanode holding all the data
- 4 worker nodes for the spark cluster (2 cores, 8 cores, 8 cores, 16 cores)
- Right now the spark worker node doesn’t run/ execute actions
- We are developing code on the course cluster which then can be moved over to our own cluster

UPPSALA
UNIVERSITET