Distributed Data Engineering

Bitcoin Tweets using DynamoDB
(Assignments 1)

Group Management

- Up to 4 students per group
- Group must be registered on the excel sheet shared on the Google Classroom
  - After registration group cannot be changed!

Use Case: Bitcoin Tweets

- The tweet data is available on the following link.
  - https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets
- The data consists of the following fields
  - User name
  - User location
  - User description
  - User created
  - User followers
  - User friends
  - User Favorites
  - User verified
  - Date
  - Text
  - Hashtags
  - Source
  - Is Retweet

Queries to Support

- All tweets of a user
- All tweets by the users from the same location
- Top k users with most followers
- Tweets by top k users with the most followers
- Top k tweets with the most matching tags
- Delete all posts of user with followers less than a threshold

Marks Distribution

- No user interface is required!!

- Design Database - 30 marks
  - Primary/Secondary index
  - Detailed presentation with the design justification to be submitted.

- Insertion of data – 10 marks
  - Write script to insert data

- Retrieval queries – 60 marks
  - See previous page
  - 10 marks for each query

Technology Stack to Use

- NoSQL Database
  - DynamoDB (Cloud or local installation)

- No restriction on the use of language
  - NodeJS, Python, Go, Java

Submission Assignment 1

- Final date of submission
  - **Monday, 14th April 2025**

- Submit only Google Classroom
  - Source code / Documentation
  - Group submission!

Questions?