

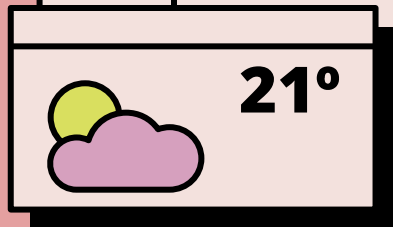
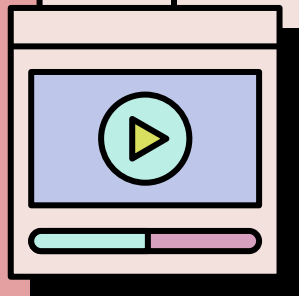
COMP 6231 Distributed System Design, Master of Applied Computer
Science, Concordia University

Reddit Sentiment Analysis

Involves evaluating and understanding the emotional tone and opinions
expressed within Reddit posts and comments.

>>>>>

~~~~~  
.....





# Table of contents



## Introduction

01

Analyzing a Big Dataset with Cloud Services, Apache Spark, and AWS, with a Focus on AWS EKS

## Dataset

02

Analyzing a diverse dataset of posts from various subreddits, we discern sentiments, temporal patterns, & overall sentiment.

## System & services

03

AWS S3, AWS EKS, AWS EMR, and Apache Spark.

## Distributed Characteristics

04

Auto Scaling, Elasticity, High Availability, Data Redundancy, Security and Compliance



01

.....

>>>>

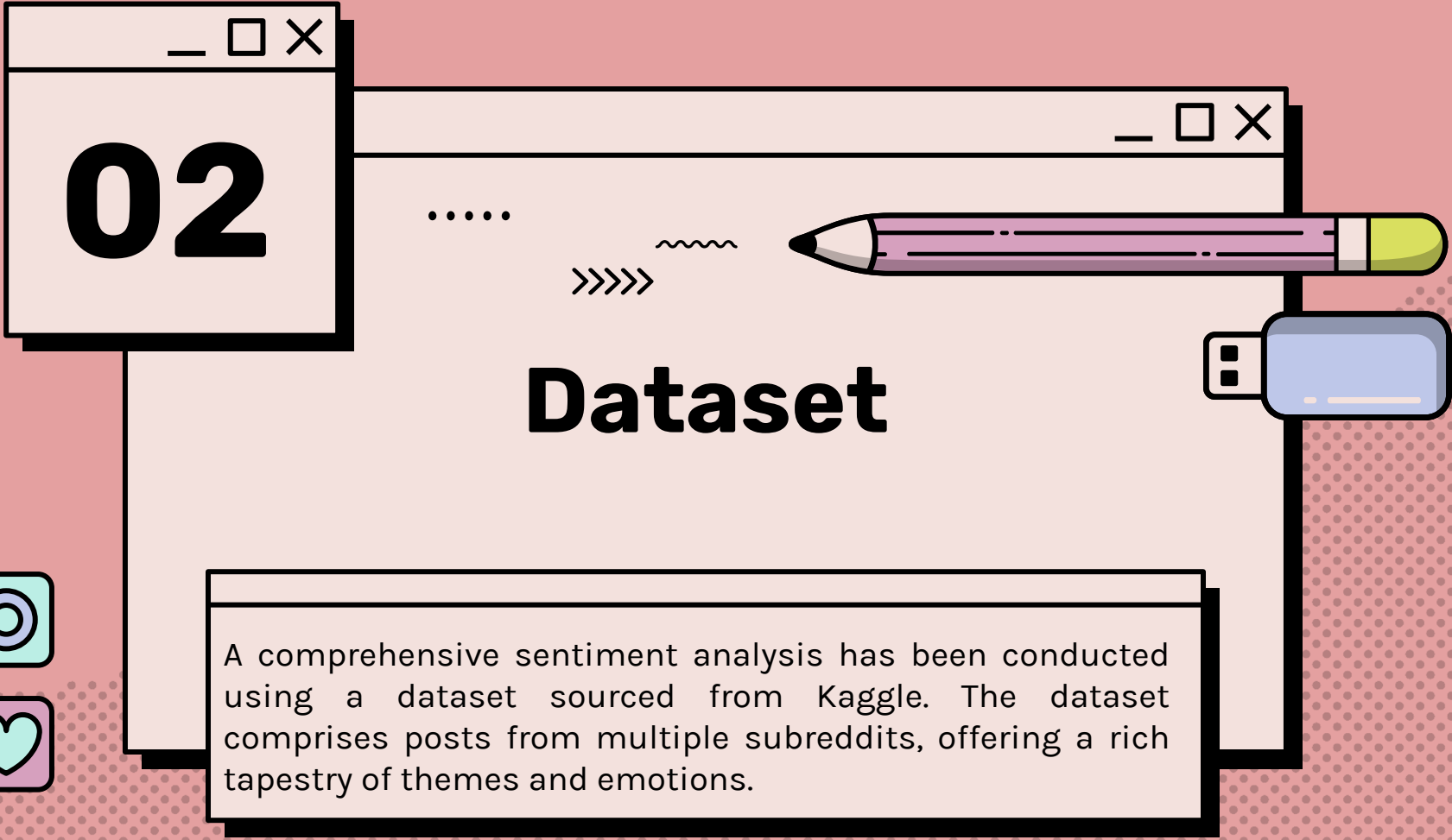


# Introduction



In this project, we analyze a large dataset from Kaggle consisting of Reddit Submissions using cloud services like Apache Spark and AWS (AWS EKS, AWS S3, AWS EMR). The data from Kaggle, totaling nearly 2 GB, is employed.

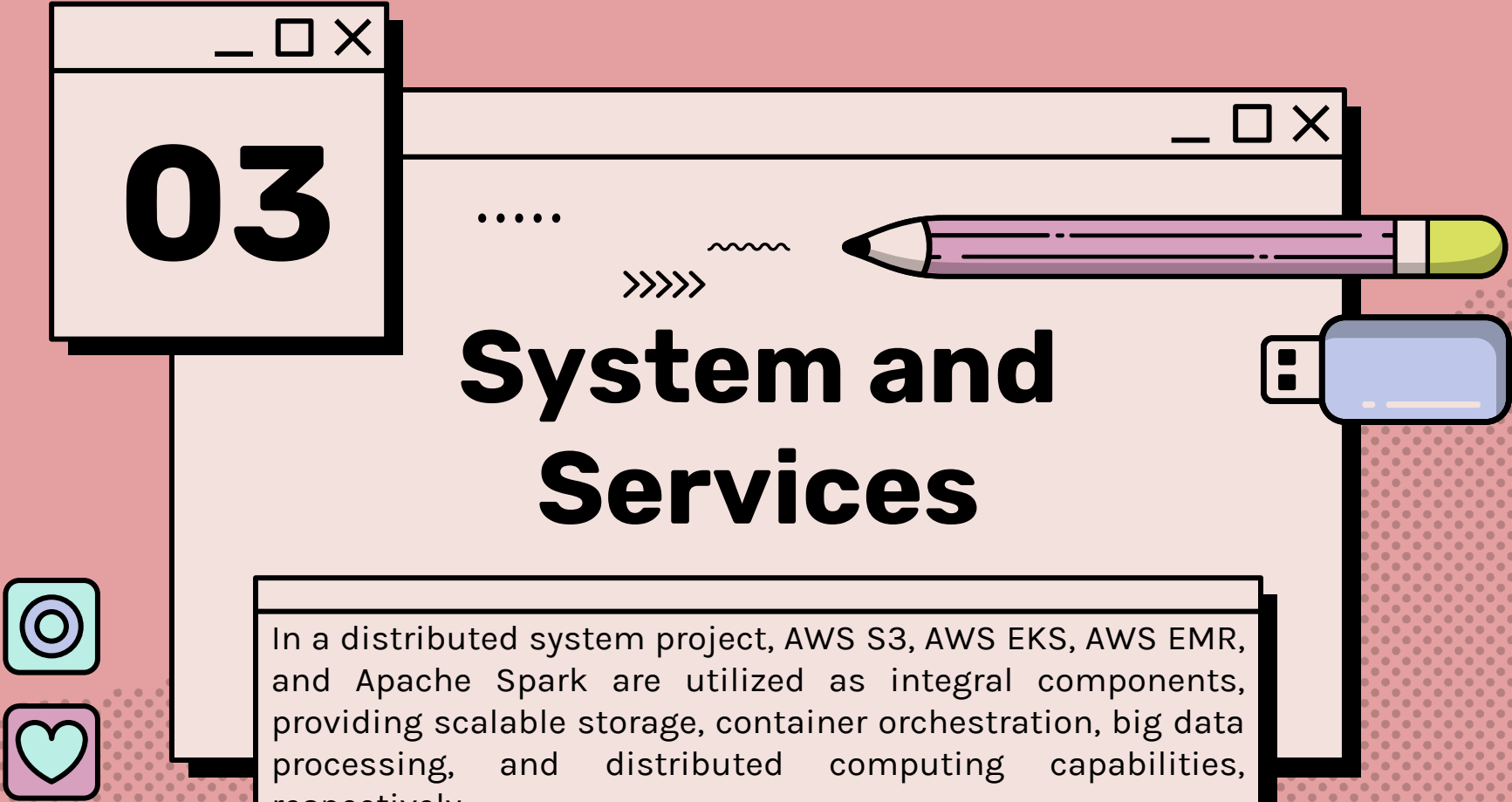
The project focuses on AWS EKS, exploring its distributed computing features, including Auto Scaling, scalability, fault tolerance, replication, and availability. We also cover the system and services used, while outlining the demo scenario, including the distributed system architecture and project results.



02

# Dataset

A comprehensive sentiment analysis has been conducted using a dataset sourced from Kaggle. The dataset comprises posts from multiple subreddits, offering a rich tapestry of themes and emotions.



03

# System and Services

In a distributed system project, AWS S3, AWS EKS, AWS EMR, and Apache Spark are utilized as integral components, providing scalable storage, container orchestration, big data processing, and distributed computing capabilities, respectively.



**AWS S3:** Storage powerhouse in the cloud.

**AWS EKS:** Seamless Kubernetes orchestration

**AWS EMR:** Elastic big data processing.

**Apache Spark:** Distributed analytics powerhouse.



04

# Distributed Characteristics

Auto-scalability, High availability, Elasticity, Data Redundancy, and Security and compliance are key characteristics defining the robustness and adaptability of a distributed system.



**Auto-scaling:** EKS (Elastic Kubernetes Service) supports Kubernetes auto-scaling, a feature that dynamically adjusts the number of nodes within your clusters based on the current workload demands.

**High Availability:** It is a part of Fault Tolerance. High availability in EKS is achieved by spanning Kubernetes clusters across multiple AWS Availability Zones. This design ensures that if one zone experiences issues, such as hardware failures or network disruptions, the other zones continue to function seamlessly.

**Elasticity:** AWS EMR (Elastic MapReduce) systems exhibit elasticity by dynamically scaling computing resources up or down based on demand.





**Data redundancy:** AWS provides robust data redundancy for its EMR systems. This redundancy is vital in preventing data loss and ensuring the continuous operation of systems, contributing significantly to their fault tolerance.

**Security and Compliance:** This involves data encryption, secure network configurations, and compliance with required standards (ex: healthcare). AWS EKS has also improved its control plane scaling and update speed, which indirectly enhances the security and compliance aspect by ensuring the system is up-to-date with the latest security patches and features.

# Meet the team

- **Vidhi Sagathiya (40232374)**
- **Anurag Agarwal (40232644)**
- **Kalinga Swain (40226333)**
- **Jimi Mehta (40225526)**
- **Muqaddaspreet Singh Bhatia (40276333)**



**Thanks!**