## Introduction

This project will use Twitter tweets to analyse the sentiment for the UK Government based on tweets using the keyword ‘ukgov’

**Mention big data?**

## Data collection

The data was gathered from Twitter using the snscrape library. Tweets with the keyword ‘ukgov’ were scraped between the date range 23 Oct 2021 – 23 Oct 2022 and a dataframe with the following info: Datetime, tweet id, text, Username, reweetCount, likecount, hashtags.

## Data Storage and processing

**The tweets csv was stored on the …..**

**The tweets are imported into a mysql database, before being imported into PySpark for analysis.**

## Comparative Analysis

This is a comparative analysis between MySQL and MongoDB. MySQL is a relational database, and MongoDB is

NoSQL database. The ycsb library was used to perform a benchmark test on both MySQL and MongoDB. The results were gathered for both databases based on operation counts of 1000, 10000, 10000, 100000, 1000000, 10000000.

**Runtime**

When looking at the runtime performance, MongoDB performed the best. Figure 1 shows MongoDB is more than 2 times faster than MySQL

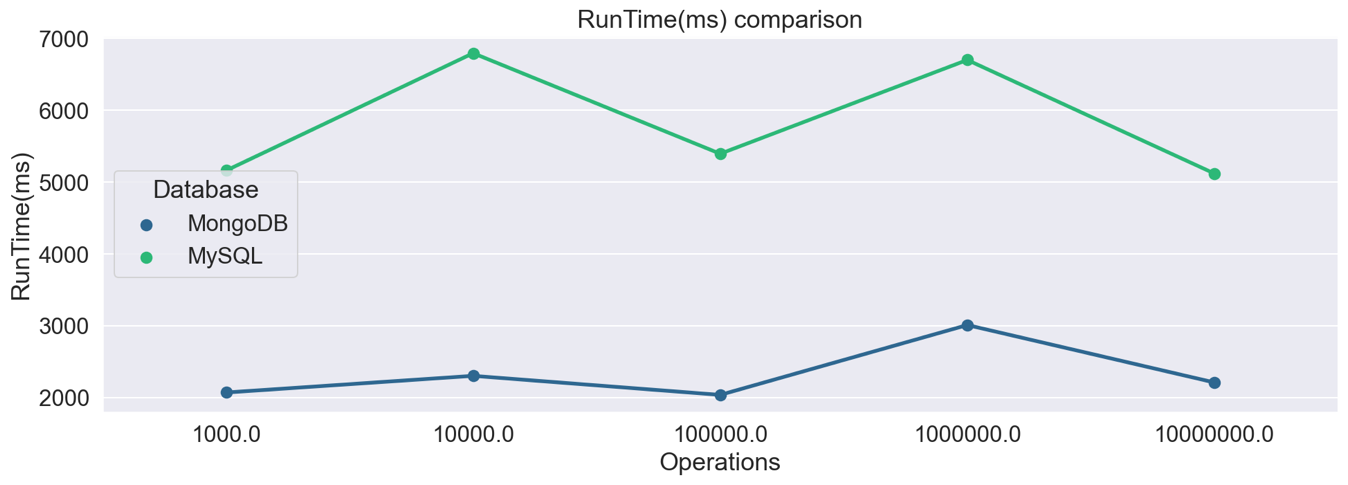
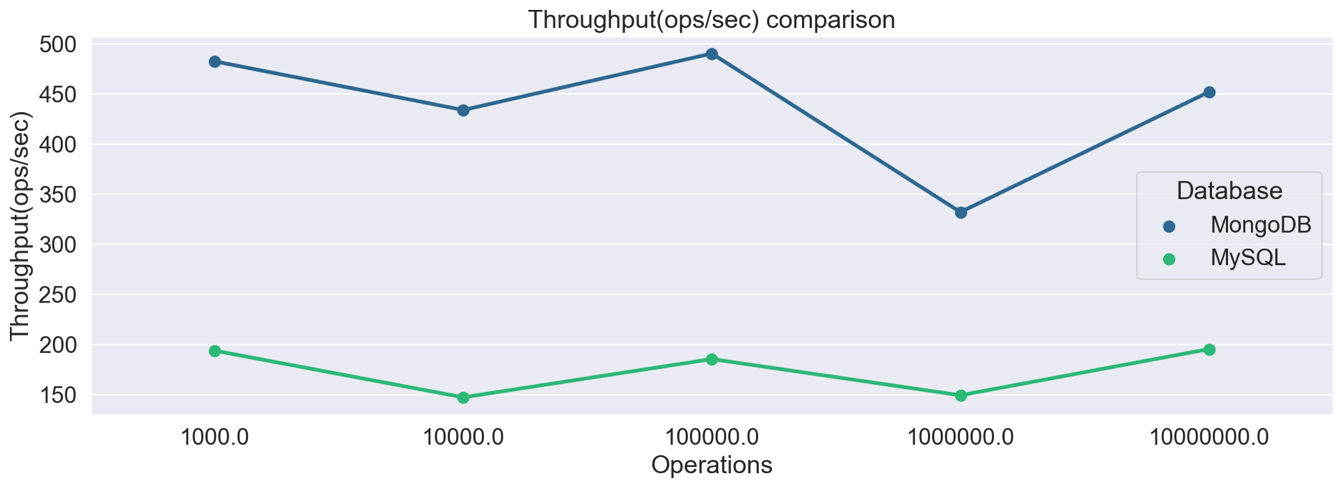


Figure 1: RunTime(ms) comparison

**Throughput**

Throughput is a measure of how many units of information a system can process in a given amount of time

When analysing both database's throughput (ops/sec), MongoDB was able to process more than double the information that MySQL could.



**Insert**

The benchmark test for insert looks are how long it takes to insert a

Chart, line chart

Description automatically generated

MongoDB is a document-oriented NoSQL database used for handling big data. MongoDB uses collections and documents.

Partition-tolerant, consistency

Integrates with Mapreduce

Schema-less database, highly scalable deep query-ability, faster access to data

## Sentiment analysis over time

Looking at the sentiment over different periods

## Forecasting Sentiment