**[Project Title]**

**Project Deliverable 1: Requirements Analysis**

**9/21/2022**

Group 03

*Contributors:* Arunabh Bhattacharya,

Kathleen Tiley,

Benjamin Weiss, and

Jonathan Williams

Table of Contents

[Overview 2](#_Toc114223828)

[Project Description 2](#_Toc114223829)

[Requirement for a Database System 2](#_Toc114223830)

[Web-Based User Interface Functionality 2](#_Toc114223831)

[Data 2](#_Toc114223832)

[Software 2](#_Toc114223833)

[Trend Analysis and Queries 2](#_Toc114223834)

[Trend Analysis Goals 2](#_Toc114223835)

[Trend Queries 2](#_Toc114223836)

# 

# Overview

## Project Description

Businesses across many industries use database systems to organize data collected to determine trends, obtain better insight into their market, and make more informed business decisions. Electronic commerce websites often use data obtained through their platform like sales, and reviews/ratings, to guide critical decisions like product inventory, determining customer recommendations, among many others. Yet, the effectiveness of this analysis is dependent on user engagement in their specific platform, limiting trends that can be seen from user data. Online communities can be an alternative rich source of information for businesses to supplement data collected to make informed decisions. Social media platforms have high user engagement and are a great candidate for collecting data on online communities. Goodreads.com is an online social media platform website where readers can rate, recommend, and share opinions on different books. Easily organized data like book ratings, books read by users, and user-specific ratings, could be used to obtain better insight into the community’s sentiment toward different books.

## Requirement for a Database System

A database system used should have user-specific information to evaluate user sentiment toward specific products, in this case, books. Data collected should be computationally easy to analyze, where we would prefer a clear score rating opposed to a full written text review. Book specific information would also be beneficial as we could try and see the relationship between book specific traits and their perceived quality by the user base.

## Web-Based User Interface Functionality

The web-based interface will use queries to interact with a database and generate figures to illustrate the trends in the data. Successfully generating the figures in the website will be dependent on the communication between the database and the website.

## Data

In 2017 a Ph.D. researcher at the University of California San Diego, Mengting Wan, scraped the largest currently available data from Goodreads.com. Wan collected information about the books on the website, like associated genres, synopsis, and user reviews. Additionally, Wan collected information on user-book interactions and reviews organized by individual users. Wan has made this data openly available online through an open-source project on github.com [github.com/MengtingWan/goodreads]. The data is currently posted in a JSON file format, which will be parsed and cleaned to successfully transfer the data to the Oracle SQL database.

## Software

We will need to use Oracle SQL databases to make queries on the data and tease out possible trends related to user sentiment on books. The front end of the website will be developed using a Javascript framework which will be used to design all user interface and generate the figures on the website based on the information in the database.

# Trend Analysis and Queries

## *Trend Analysis Goals*

## *Trend Queries*

Some potential trend queries on the data could be

* What was the average number of books read during a certain period, ordered by genre?
* Which genres had the highest ratings during a certain period?
* Which genres received more reviews during a certain period?
* What was the percent of shelved books by genre?
* What was the ratio of read and shelved books by genre?
* Does book genre have a significant effect on its average rating?
* Do longer books, books with more pages, get reviewed less often?

**Phase 1**

**The summarized requirements from Prof:**

propose and understand an appropriate project topic, identify its main data management needs, explore and motivate its potential for interesting queries, and analyze the needed user functionality.

1. What are the main functions that the web-based user interface should provide?
2. **How do the different functions work together? Sometimes there are dependencies between different functions.**
3. Which real-world data are needed to support the functions identified before?
4. Can such real-world data be found in the Internet?
5. **What (colloquial) queries are important for the application?**
6. Which public domain and/or proprietary software is needed to perform the task? (The database system used must be CISE Oracle.)

(PDF file) that presents a clear and structured description and motivation of the project topic and requirements that the software solution should later fulfil.

The focus of this project is supposed to be on the database part and not so much on the application part. focus on database queries that evaluate large volumes of stored data. Of course, the application part must be highly functional, and the different user functions must cooperate nicely together. However, a fancy layout design of the user interface is not required but appreciated.

It is important that each group demonstrates in their deliverable that their application would really benefit from database support and that new information (such as trends) can be derived from the stored data. A simple retrieval of data from the database (that is, search) or the pure connection of different tables (that is, joins) are not sufficient. As an example, let us assume that a group selects a sales application as their project topic and stores many daily sales numbers in their database. Of course, one can search for sales data of interest in the database and display them in the user interface. But searching only identifies an interesting subset of all data stored in the database. DBMS are specialized for search tasks, and the respective SQL queries are relatively simply structured. This project aims at more interesting queries that, first, derive new information which is not explicitly stored in the database but can be derived from the data in the database by computations and, second, represent trends (see Section 3).

Arunabh-

Since we are required to do trend analysis and compute the changes in data over a period, based on the datasets, I could think of the following:

For point 5, some of the colloquial queries could be:

* What was the average number of books read during a certain period, ordered by genre?
* Which genres had the highest ratings during a certain period?
* Which genres received more reviews during a certain period?
* What was the percent of shelved books by genre?
* What was the ratio of read and shelved books by genre?