# HW 1 – Apache Cassandra Executive Summary

**1. Dataset Description and Selection:**

* Dataset: Book-Crossing dataset was selected for its real-world application and large size, ideal for testing Cassandra's capabilities.
* Contents: Three CSV files - Users, Books, and Ratings.

**2. Database Schema Design**:

* Design was tailored for specific queries like book ratings, author ratings, and user activity.
* Tables: books, users, books\_rated\_by\_user, users\_by\_rated\_books.
* books\_rated\_by\_user: Facilitates queries about user activity and book preferences.
* users\_by\_rated\_books: Enables analysis of user demographics based on book ratings.

**3. Key Insights from Data:**

* User Activity: Identified the most active user groups.
* Top Books and Authors: Analysis of books and authors with the highest average ratings.
* Queries Example: 'Find average rating of books by each author'

**4. Challenges and Solutions:**

* For easier readability and interfaces we decided to integrate all code to one Jupyter notebook. CQL commands were given to the cluster using Cassandra Driver, a python-Cassandra API.
* During data ingestion, our original script has loaded one row at the time in a for loop. As we have a very large number of rows in our dataset (about 270K in the largest table) that took a long time. The solution was to use

**5. Visualizations:**

Top 10 Books Chart: Shows books with the highest normalized scores and rating counts.

Book Distribution by Publisher: Bar chart illustrating the number of books per publisher.

User Ratings Analysis: Pie chart depicting average ratings per user.