

INF2006 Cloud Computing & Big Data

Assignment 2

Big Data Analytics

Objective

The objective of this project is to analyze large-scale social media data to identify trends, popular topics, and user engagement patterns. The project will involve data ingestion, storage, processing, and visualization, utilizing Hadoop for distributed storage and batch processing, and Spark for real-time data processing and analytics.

Project Description

This project is designed to provide you with the flexibility to define your own scope and direction, allowing you to tailor it to your interests and background. This open-ended approach encourages innovation and creativity, enabling you to explore diverse datasets or application scenarios that align with your academic or professional goals. You may choose from a selection of provided datasets or propose your own, provided they are relevant and suitable for the project's objectives.

While the project is open-ended, a structured task list has been provided to guide your project's scope and development. This framework ensures that your work remains focused and aligned with the learning outcomes of the module, while still allowing room for originality and exploration.

By combining flexibility with a clear framework, this project aims to foster both independent thinking and practical application of big data concepts using Hadoop and Spark.

Project Framework/Requirement:

1. **Dataset and application scenarios:** Identify a suitable dataset for your project. You may select from the provided datasets below or propose your own dataset based on the application scenario your team plans to explore. If you choose to collect your own data (e.g., through web crawling or APIs), ensure it aligns with the project's objectives and is of sufficient scale and quality for analysis.

Provided Datasets:

Below is a list of publicly available datasets that you can use for your project:

- ✓ **Twitter Dataset (Stanford):** A collection of tweets with metadata such as timestamps, user information, and tweet content.
 - Link: [Stanford Twitter Dataset](#)
- ✓ **Reddit Comments Dataset:** A dataset containing Reddit comments from various subreddits, including metadata like timestamps, user information, and upvotes.
 - Link: [Reddit Comments Dataset](#)
- ✓ **Yelp Dataset:** A dataset containing business reviews, user data, and check-ins from Yelp.
 - Link: [Yelp Dataset](#)
- ✓ **Amazon Product Reviews Dataset:** A dataset containing product reviews and metadata from Amazon.
 - Link: [Amazon Product Reviews Dataset](#)
- ✓ **Common Crawl Dataset:** A large-scale web crawl dataset suitable for text analysis and trend detection.
 - Link: [Common Crawl Dataset](#)

- ✓ **Google Books Ngrams Dataset:** A dataset containing n-grams from Google Books, useful for text analysis and linguistic trends.
 - Link: [Google Books Ngrams Dataset](#)
- ✓ **US Census Dataset:** A dataset containing demographic and economic data from the US Census.
 - Link: [US Census Dataset](#)
- ✓ **NASA Exoplanet Archive:** A dataset containing data on exoplanets discovered by NASA missions.
 - Link: [NASA Exoplanet Archive](#)

2. Data Storage:

- **Task:** Store the collected data in a distributed file system.
- **Tools:** Use Hadoop HDFS for storing large volumes of data.
- **Deliverable:** Data stored in HDFS, ready for processing.

3. Data Processing with Hadoop:

- **Task:** Perform batch processing on the stored data to clean, filter, and transform it.
- **Tools:** Use Hadoop MapReduce for tasks like removing duplicates, filtering irrelevant data, and aggregating metrics (e.g., count of posts per user).
- **Deliverable:** Cleaned and processed data stored back in HDFS.

4. Data Analysis with Hadoop or Spark:

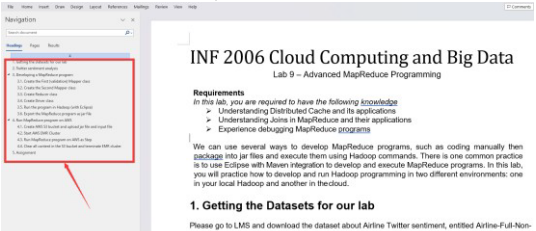
- **Task:** Perform analytics on the processed data to identify trends and patterns.
- **Tools:** Use Hadoop or Apache Spark for tasks like sentiment analysis, trend detection (e.g., most frequent hashtags), and user engagement analysis.
- **Deliverable:** Analytical results such as sentiment scores, trend graphs, and engagement metrics.

5. Data Visualization:

- **Task:** Visualize the results of the analysis to provide insights.
- **Tools:** Use visualization libraries like Matplotlib, Seaborn, or Tableau.
- **Deliverable:** Interactive dashboards or static reports showing trends, sentiment analysis, and user engagement.

This project not only covers the technical aspects of Hadoop and Spark but also provides a practical application of big data technologies in a real-world scenario, making it an engaging and educational experience for your students.

Timeline and Deliverables

Thursday, 3 April 2025, 23:59pm	Final Submission	<p><u>Final Report:</u></p> <p>Create a final report document (no more than 15 pages, an additional appendix can be added if needed). The report should include titles and headings for easy navigation to each section (see below screenshot).</p> 
---------------------------------	------------------	--

		<p>The report should include the following sections:</p> <p>a. Introduction: Introduce your project, explaining the context, objectives, functions developed and what the project aims to achieve.</p> <p>b. Team Member Contributions: List each team member's name and describe their individual contributions to the project. Mention that every member has contributed to the coding.</p> <p>c. Design Overview: Describe how each program works, the data flow, and any key design decisions.</p> <p>d. Task Implementation: List down all the tasks/functionalities developed by your team. For Hadoop or Spark design, <i>you shall list how the MapReduce jobs are designed, like the number of MR jobs, what does Map, Partition, Reduce phases do in each job etc.</i></p> <p>e. Screenshots: Include screenshots or captures that show the successful execution of each task and the corresponding results. Label each screenshot appropriately.</p> <p>f. Limitations: Address any limitations of the project. Be honest about areas where improvements could be made.</p> <p>Compile all source code files into a single folder. Each code file should start with a comment line indicating the contributor's name.</p> <p>Zip the final report document and the source code folder into a single zip file.</p> <p>Name the zip file according to your project's name or a related identifier.</p> <p><u>Submission:</u></p> <p>Go to the xSITE Dropbox ('Project 2 Final Submission') and upload the zip file containing your final report and source code.</p> <p>Double-check that you've included all the required elements in your final report and that it's well-organized and formatted.</p>
--	--	--

		Confirm that your submission adheres to any additional guidelines or instructions provided by your instructor or institution.
Thursday, 3 April 2025, 23:59pm	Project video	Submit a video with the presentation and demonstration of your project. Keep your video within 6 minutes. Faces should be shown during the presentation.

Assessment criteria

Your assignment will be assessed according to the criteria listed in the mark scheme in Table 1 (Group Assessment) and Table 2 (Individual assessment). Your mark for this assignment will be computed as follows:

$$\text{Group_mark} * \text{Individual_contribution}$$

Table 1 Group Assessment

Criteria	Weight
Quality of the project: <ul style="list-style-type: none"> • Functionality: The project includes a comprehensive set of features. • Efficiency of Implementation: The MapReduce jobs are designed and implemented efficiently. • Code Completeness and Accuracy: All required analyses are correctly and efficiently implemented using well-designed MapReduce/Spark algorithms. 	60%
Report <ul style="list-style-type: none"> • Your report should be professionally written and clearly present your projects. • Clearly specify the contributions of each teammate. • Include key implementation ideas, screenshots of successful program execution, and the results for each task. • Highlight the unique features as well as any limitations of the project. 	20%
Presentation /Demonstration <ul style="list-style-type: none"> • Conduct a clear presentation and demonstration 	20%

Table 2 Individual Contribution Assessment

Criteria
A peer evaluation will be conducted. The purpose is to recognize individual effort and contributions in a team project by peer members. It is important to recognize that it is not peer grading but really recognizing and acknowledging differential peer contributions by members in the project.

Late Submission

A penalty of 20% per day for each deliverable will be imposed for late submission unless extension has been granted by the lecturers prior to the submission date. Request for extension will be granted on a case-by-case basis. Any work submitted more than 4 days after the submission date will not be accepted and no mark will be awarded.

Plagiarism

SIT's policy on copying does not allow you to copy software as well as your assessment solutions from another person. It is not acceptable to copy other person's work. It is the students' responsibility to guarantee that their assessment solutions are their own work. Meanwhile, you must also ensure that others don't obtain access to your work. Where such plagiarism is detected, both of the assessments involved will receive ZERO mark.