UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

ASSIGNMENT COVER PAGE

Student Number									Surname	Initials
U	2	3	8	2	5	8	4	8	Dutywa	Y

Module Code	**MIT**	**8**	**0**	**5**
Assignment number	**2**			
Date of Submission	**10 October 2023**			
Name of Lecturer	**Prof Stacey Baror**			

Declaration: I declare that this assignment, is my own work and that I have referenced all the sources that I have used.

The University of Pretoria commits itself to produce academic work of integrity. I affirm that I am aware of and have read the Rules and Policies of the University, more specifically the Disciplinary Procedure and the Tests and Examinations Rules, which prohibit any unethical, dishonest or improper conduct during any form of assessment. I am aware that no student or any other person may assist or attempt to assist another student, or obtain help, or attempt to obtain help from another student or any other person during tests, assessments, assignments, examinations and/or any other forms of assessment.

I declare that this assignment, submitted by me, is my own work and that I have referenced all the sources that I have used. In addition, I assure that, when using IT/ AI- supported writing tools, I have listed these tools in full in a section titled: “Overview of tools”;, with their product name, my source of supply (e.g. URL) and information on the functions of the software used as well as the scope of use. I also declare that I have not copied any text directly from any AI tool. Exempt from this are those IT/ AI-supported writing tools that were classified as not necessary to declare by the module lecturer.

MARK	

Analysis Report: PubMed Knowledge Graph Dataset

Dataset: PubMed Knowledge Graph

Table	Size
OA01_Author_List	10 GIG

Map-Reduction and Visualisation Report

Data Preparation

- Split the large dataset into smaller manageable chunks.
- Store the data in HDFS (Hadoop Distributed File System) for distributed processing.

MapReduce Algorithm

- The algorithm would consist of two main phases:
  - Map phase and
  - Reduce phase.

Map Phase

- In the Map phase
  - Read each record (line) from the input file(s) and
  - Extract relevant information.

Reduce Phase

- In the Reduce phase, data will be grouped for each key.

Output

Results will be stored in HDFS.

Example: MapReduce algorithm for counting the number of authors per publication ("PMID" and "LastName"):

Java

```
// MapReduce Mapper
Mapper(LongWritable key, Text value, Context context):
    // Parse the input line to get PMID and LastName
    String[] tokens = value.toString().split(",");
    String PMID = tokens[0];
    String LastName = tokens[1];

    // Emit PMID as the key and 1 as the value
    context.write(new Text(PMID), new IntWritable(1));

// MapReduce Reducer
```

```
Reducer(Text key, Iterable<IntWritable> values, Context context):
    int sum = 0;

    // Sum the values (number of authors) for each PMID
    for (IntWritable value : values) {
        sum += value.get();
    }

    // Emit PMID as the key and the total count of authors as the value
    context.write(key, new IntWritable(sum));
```