DT228A/B - Programming for Big Data - Hadoop/Map-Reduce Assignment **Due Before : 5th March @ 23:00**

Overview

This assignment requires you to compile a set of data, load this data into hdfs and to write a map-reduce process that will extract and present the data as outlined in the following sections.

Background

The Wikimedia Foundation, Inc. (http://wikimediafoundation.org/) is a nonprofit charitable organization dedicated to encouraging the growth, development and distribution of free, multilingual, educational content, and to providing the full content of these wiki-based projects to the public free of charge. The Wikimedia Foundation operates some of the largest collaboratively edited reference projects in the world; you are probably most familiar with Wikipedia which is a free encyclopedia and is available in over 50 languages (see https://meta.wikimedia.org/wiki/List_of_Wikipedias for a list of languages).

Information on all the projects that are the core of the Wikimedia Foundation available at http://wikimediafoundation.org/wiki/Our_projects.

Aggregated page view statistics for Wikimedia projects is available at http://dumps.wikimedia.org/other/pagecounts-raw/. This page gives access to files that contain the total hourly page views for Wikimedia project pages by page. Information on the file format is given on this page view statistics page.

Required Tasks

The task of this assignment is twofold:

1. Use HDFS and MapReduce to identify the popularity of ***Wikipedia*** projects by the number of pages of each Wikipedia site which were accessed over an *x* hour period. Your job should allow you to directly identify from the output the most popular Wikipedia sites accessed over the time period selected. You can choose whichever *x* hour period you wish from the files available on the page view statistics page, with the constraint that *x>=6*.
2. Use HDFS and MapReduce to identify the average page count per language over the same period, ordered by page count.

Deliverables

You will be required to document your approach for processing the data and producing the required outputs using map-reduce only.

Your report (saved as a PDF document) should contain the following:

- Explanation of the steps you performed for loading the data sets into HDFS
- Detaild design, including diagrams and detailed explanations of each part of the process
- Explanations of any design decisions (evaluating alternatives) and any assumptions made
- Well written and fully commented Java code for the map-reduce process
- Examples of the output files from the map-reduce process illustrating the data produced at each stage.

The output files from the map-reduce process should be included. If these are not included then your assignment mark will be reduced by 30%.

Submission Details

The assignment is due by 5th March @23:00

You should create one document/report containing all the material for each item listed in the deliverables. Convert this document into a PDF. It is this PDF document that should be submitted. All images should be imbedded in this document.

In addition to the report the output files from the map-reduce process should be submitted. You will need to extract these files from HDFS.

The Report and the Output Files should be ZIPPED (**only zip format will be accepted**) and it is this ZIP file that should be submitted on WebCourses.

You will need to submit your assignment on WebCourses. You cannot submit your assignment via email.

Marking Scheme

The marking scheme for this assignment is:

- 10% Explanation of the steps you performed for loading the data sets into HDFS.
- 25% Design and structure of the map-reduce process.
- 40% Well written and fully commented Java code for the map-reduce process.
- 15% Extent of use of map-reduce features and scalability.
- 10% Output files from the map-reduce process.

The documentation for your assignment must contain your name, your student number, your class, course (**DT2??**) and year information, assignment, lecturer name and your **Failure to give this information will incur a 10% penalty.**

The assignment most be performed **individually.**

Each submission must be original work as plagiarism will result in a **zero** mark (0%).
DIT Plagiarism Policy : http://www.dit.ie/media/documents/campuslife/plagiarism.doc

There will be a 10% penalty deduction will be applied for each day the assignment is late.

The output files from the map-reduce process should be included. If these are not included then your assignment mark will be reduced by 30%.

There is no penalty for submitting early.

Assignment feedback will be provided on Webcourses.