WST510/CS542 Web Architecture

Assignment #3: Learning to Speak Pig Latin

Due date: 11:55pm on Friday, June 6th, 2014

Welcome to assignment #3! In this assignment, you will be ***rewriting*** your MapReduce implementation of Wikipedia PageRank in Apache Pig. Pig Latin is a fast-prototyping language built on top of Hadoop that enables you to describe a series of different MapReduce tasks without the headache of having to implement and debug Map and Reduce interface functions line-by-line at each step. Also, you might be glad to learn that Pig allows 2 different execution modes, one of which is local execution mode, where you get to execute your Pig script in a single node configuration, i.e. possibly in your own PC.

This assignment is composed of two steps: Part I to calculate the PageRank of the Wikipedia corpus, ***same as Part II of the previous programming assignment***, but in Pig Latin. We use the same Wikipedia XML corpus as the input data, but this time, with reduced volume to avoid heavy computation. In Part II, you will retrieve the execution plan of your Pig script using the 'EXPLAIN' command, convert the plan into graphical DAG of MapReduce tasks using '-dot' option, and explain how your query is planned out. The idea is to let you experience firsthand the difference between running a chain of MapReduce jobs and running a Pig script, while also getting the idea of how Pig scripts translate to individual MR tasks.

While Apache Pig abstracts away the loading, filtering, grouping, element iteration, set operations, and others, you may need to implement custom, unsupported functions in the form of UDFs (user-defined functions). By itself, Apache Pig does not support iteration of Pig scripts. Some workarounds exist, however, such as using a different language that embeds Pig scripts. You may choose any method you like, as long as the deliverables are correct and the core implementation uses Pig Latin.

Part I: PageRank

Same algorithm, in different language, using less data, producing same deliverables.

Part II: Plan

Using 'EXPLAIN' command with the '-dot' option, the query execution plan can be dumped into three DAGs, each representing logical, physical, and execution plan, respectively. Fed to the right applications, these '.dot' files can be transformed into common image formats, graphically describing how Pig script is executed. Using this information, possibly combined with the text output of 'EXPLAIN', you will:

1. Explain the plan step-by-step
2. Illustrate the difference between implementing in MapReduce and in Pig
3. Add your personal observations

in a single PDF document, within 5 pages with graphical plans embedded. Priority of this task is on explaining the plan and your observation clearly without losing detail.

Deliverables

File named '**a3p2-[your_student_id].zip**' containing the deliverables of Part I and Part II.

Cluster Information

Due to some issues, the AWS accounts may need some time to be available again. Meanwhile, you may test your script in a single node setup, using only small subsets of input data. The link to data will be provided to you soon.

Useful links

Pig execution modes: http://pig.apache.org/docs/r0.11.1/start.html

Running Pig: https://wiki.apache.org/pig/RunPig