Lab 3

Matrix Multiplication

CS429 - Introduction to Big Data Analysis

Lab Instructor: Nguyễn Đình Thảo

Table of Contents

Matrix-Vector Multiplication

Problem Statement

Suppose we have an $n \times n$ matrix **M**, whose element in row i and column j is denoted m_{ij} . Suppose we also have a vector **v** of length n, whose j th element is v_j . Then the matrix-vector product is the vector **x** of length n, whose i th element x_i is given by

$$x_i = \sum_{j=1}^{n} m_{ij} v_j$$

The matrix M and vector **v** each will be stored in a file of the distributed file system. We assume that the row-column coordinates of each matrix element will be discoverable, either from its position in the file, or because it is stored with explicit coordinates, as a triple (i , j , m_{ij}). [*Mining of Massive Datasets, section 2.3.1, p.31*]

Example

Pseudocode

Mapping

Input: list of triples (i , j , m_{ij}); vector **v**
For each matrix element m_{ij} , create a key-value pair of (i , $m_{ij} \times v_j$)
Output: key-value pairs of (i , $m_{ij} \times v_j$)

Reducing

Input: key-value pairs of (i , $m_{ij} \times v_j$)
For each given key i , sum all values associated with key i
Output: key-value pairs of (i , x_i)

Implementation

See the notebook ***Lab 3.ipynb***.

Matrix Multiplication

Problem Statement

If **M** is a matrix with element m_{ij} in row i and column j, and **N** is a matrix with element n_{jk} in row j and column k, then the product **P=MN** is the matrix **P** with element p_{ik} in row i and column k, where

$$p_{ik} = \sum_j m_{ij} n_{jk}$$

It is required that the number of columns of **M** equals the number of rows of **N**, so the sum over j make sense. [*Mining of Massive Datasets, section 2.3.9, p.38*]

Example

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \cdot \begin{pmatrix} 7 & 8 \\ 9 & 10 \end{pmatrix} = \begin{pmatrix} 25 & 28 \\ 57 & 64 \\ 89 & 100 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \cdot \begin{pmatrix} 7 & 8 \\ 9 & 10 \end{pmatrix} = \begin{pmatrix} 25 & 28 \\ 57 & 64 \\ 89 & 100 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \cdot \begin{pmatrix} 7 & 8 \\ 9 & 10 \end{pmatrix} = \begin{pmatrix} 25 & 28 \\ 57 & 64 \\ 89 & 100 \end{pmatrix}$$

...

Pseudocode

Below is the pseudocode of the 2-pass Matrix Multiplication, which is based on the MapReduce algorithm provided in [*Mining of Massive Datasets, section 2.3.9, p.38*].

Input: matrix **M** as list of triples (i , j , m_{ij}); matrix **N** as list of triples (j , k , n_{jk})
For each matrix element m_{ij} , create a key-value pair of $(j, ("M", i, m_{ij}))$
For each matrix element n_{jk} , create a key-value pair of $(j, ("N", k, n_{jk}))$

For each key j , examine its list of associated values:

    For each pair of $("M", i, m_{ij})$ and $("N", k, n_{jk})$:

        Create a key-value pair of $((i, k),\ m_{ij} n_{jk})$

Sum all the pairs with same keys (i, k)

Output: key-value pairs $((i, k), p_{ik})$

Lab Assignment

Implement the provided pseudocode for Matrix Multiplication in PySpark. Please provide the input matrices for testing along with your jupyter notebook.
Name your notebook file as **<your studentID>_lab3.ipynb** and submit it to the course moodle.