Pràctica Hadoop

Autor: Àlex Pardo
Adaptat per Lluís Garrido

Maig de 2016

Índex

1 Introducció

Hadoop, tal com heu vist a la sessió de teoria dedicada a aquesta tecnologia, és tot un framework de paral·lelització de processos, dissenyat per a facilitar aquesta tasca als usuaris. Les seves principals propietats són que és fàcilment escalable, treballa amb grans volums de dades de forma eficient i és resistent a fallades.

Amb tot això, cal dir que tot i que qualsevol problema es podria paral·lelitzar amb aquesta eina, és aconsellable fer-ho en el cas de tenir grans volums de dades i de voler fer operacions seqüencials amb elles, atès que està optimitzat per a això. Per exemple, en el cas de voler fer una cerca sobre un fitxer, o de treballar sobre molt poques dades, estaríem desaprofitant el potencial de Hadoop.

En aquesta pràctica haureu de treballar amb les dades mundials de vols per a diversos anys. En el vostre cas, fareu servir les dades dels anys 2007 i 2008. Podeu descarregar les dades de la web: http://stat-computing.org/dataexpo/2009/the-data.html.

2 Tasques a realitzar

Tal com hem dit, Hadoop és útil per a treballar amb grans volums de dades i realitzar operacions seqüencials amb elles. Això és el que farem en aquesta pràctica.

2.1 Les dades

Les dades consisteixen en fitxers CSV amb diverses columnes (detallades a la web esmentada a la secció 1) on a cada fila hi consten les dades d'un vol. Entre tots els camps ens interessa els de:

- Any
- Aeroport de sortida
- Si el vol ha estat cancel·lat
- Si el vol ha estat endarrerit

Els fitxers estan agrupats per anys. Dins, a la primera fila, trobareu les descripcions de les columnes. Podeu eliminar aquesta fila a mà si us resulta problemàtic.

Com hem dit, anem a treballar amb volums grans de dades. En aquest cas, cada fitxer ocupa aproximadament uns 600Mb.

2.2 Càlculs

Per tal que veieu com funciona Hadoop, en aquesta pràctica, extraureu estadístiques d'aquestes dades. Fer aquestes operacions amb eines com Python o Matlab pot ser molt costos i inclús impossible amb d'altres com Microsoft Excel.

Concretament us demanarem una sola dada: **Quin és l'endarreriment mitjà anual per a cadascún dels aeroports?**

Aquesta dada és tan simple d'obtenir com: acumular tots els minuts per a cada aeroport i dividir pel nombre de vols que surten d'aquell aeroport. El problema és que, Hadoop, per tal de balancejar la càrrega de treball, ens divideix els fitxers en parts més petites, de forma que el que tindríem serien mitges parcials de les nostres dades. Com solventar aquest problema és la tasca que us correspon a vosaltres.

3 Resultats

Com a resultats, se us demanarà que doneu el valor obtingut per a un aeroport i un any concret, escollit per nosaltres al moment de la prova (3 de Juny).

4 Altres consideracions

- Al Campus Virtual tindreu disponible un exemple on es calcularà el nombre de vols cancelats, agrupat per any i aeroport.
- El processament l'haureu de fer a la sala IA a les hores de classe atès que el cluster està apagat la resta dels dies.
- Aconsellem que realitzeu el codi en Pyhton, seguint l'exemple dels bigrames i comptatge de paraules explicat a classe.
- Per tal de testejar el vostre codi fora de l'aula, podeu fer servir canonades entre l'entrada i el resultat[1]:

  $ cat input_file | python map.py | sort | python reduce.py
- Heu de fer una única execució per a tot el problema, és a dir, no podeu executar el vostre mateix codi a cada any per separat.
- Disposareu de tres dies (18, 25 de maig i 1 de Juny) per a realitzar la pràctica, amb aproximadament 3 hores cada dia.

[1]Hadoop realitza, per defecte, una ordenació dels resultats després de fer el map. Es pot aprofitar aquest fet perquè el reduce sigui més senzill