Map/Reduce Poređenje performansi – Apache Pig i Java

Pig Latin je jezik jednostavne sintakse i višeg nivoa apstrakcije nego Java. Njegovim stvaraocima je cilj bilo olakšavanje posla programerima, olakšavanje implementacije MR zadataka, ali i kreiranje alata koji će obezbediti visok stepen čitljivosti. Pored toga, Pig je morao da pruži podršku ljudima koji imaju malo ili nikakvo tehničko znanje o MR algoritmu i platformi – ljudima koji prate i predviđaju ponašanje tržišta, upravljaju poslovnim procesima, analiziraju zahteve korisnika i klijenata... Njegova sintaksa je upravo takva: jednostavna za čitanje i lako razumljiva. Neko ko ne zna ništa o programskom jeziku Java i MR, iz prvog pogleda na neki Pig skript može da shvati šta se tu zapravo dešava.

Ovakvi preduslovi njegovom kreiranju su mali kamen spoticanja kada su u pitanju performanse izvršavanja jer se Pig obezbeđivanjem apstrakcije udaljio od mašine. Naravno, Pig kompajler radi veoma dobar posao što znači da se uz poznavanje njegovih mogućnosti i optimizacijom mogu postići veoma dobri rezultati.  
Sa druge strane, svaki Pig skript mora biti preveden u Java MR prilikom pokretanja pa se postavlja pitanje koliko će to dobro i efikasno Pig kompajler uraditi, koliko će mu vremena za to trebati i da li će se pojaviti još neka problematična mesta tokom izvršavanja.

U svrhu boljeg razumevanja brzine oba pristupa, izvedeno je poredjenje performansi MR zadatka implementiranog u programskom jeziku Java sa ekvivalentnim Pig skriptom. Dakle, obe implementacije obavljaju isti posao, rade nad istim skupom podataka i proizvode iste rezultate.

* Test okruženje

VM Ubuntu 12.04 64bit  
4 GB RAM  
Intel i5 3230M CPU @ 2.60GHz

Hadoop cluster - Cloudera Hadoop 2.0.0-cdh4.0.1 single node cluster  
Java version – OpenJDK 1.6.0\_27 64bit

* Test podaci

Test podaci su podeljeni u tri grupe, po količini podataka. S obzirom da su potrebna dva ulazna fajla, jedan koji sadrži informacije o studentima i drugi koji sadrži informacije o proseku ocena, oba fajla su napravljena u tri veličine.

|  |  |
| --- | --- |
| ulazni podaci | veličina |
| S - 100000 slogova | 5.96 MB |
| 1.03 MB |
| M - 10000000 slogova | 614.84 MB |
| 121.78 MB |
| L - 50000000 slogova | 3.04 GB |
| 651.3 MB |

* Vreme izvršavanja

|  |  |  |
| --- | --- | --- |
| ulazni podaci \ vreme | Pig MR | Java MR |
| S - 100000 slogova | 37.086s | 27.890s |
|
| M - 10000000 slogova | 2m 22.135s | 2m 4.485s |
|
| L - 50000000 slogova | 11m 37.209s | 9m 23.491s |
|

Run time 1 (in seconds)

Run time 2 (in seconds)

* Counters, spilling, etc...

Spilling (number of records)