Proiect Proiectare Software – Hadoop/ MapReduce

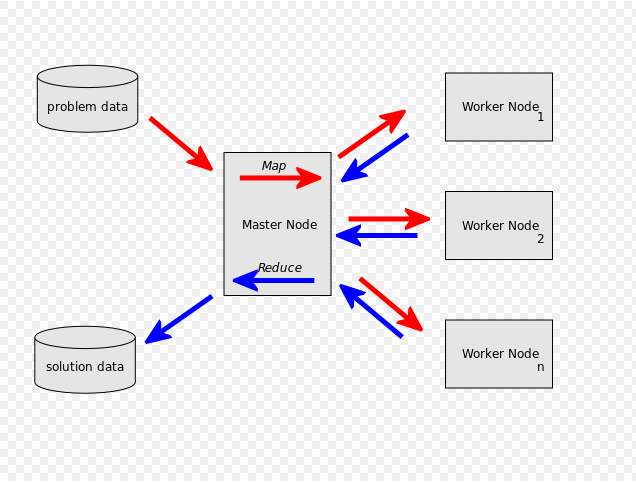
1. Introducere Hadoop – MapReduce

Big data reprezintă o colecție de seturi de date complexă și de dimensiune foarte mare și care nu poate fi procesată cu ajutorul aplicațiilor standard fiind necesare aplicații capabile să ruleze în mod paralel pe un număr foarte mare de servere. Printre dificultățile întâlnite în procesarea acestor date se numără: captura, curățarea, stocarea, căutarea, partajarea, transferul, analiza și vizualizarea. Din cauza complexității sistemelor ”big data”, a fost necesară dezvoltarea unei arhitecturi specializate.

Aceasta problema a aparitiei “big data” a fost intampinata pentru prima data de catre Google, constatand ca este imposibil sa mai detina controlul asupra unei asemenea cantitati imense de informatie. Astfel, inginerii din spatele Google, au dezvoltat un proiect prin intermediul caruia au reusit sa rezolve problema asa numita “big data”. Dupa publicarea unui document descriptiv al acestui proiect, a aparut un open-source soft, numit Hadoop. Apache Hadoop este framework de dezvoltare de software pentru cloud computing, care permite scrierea și executarea aplicațiilor distribuite ce procesează cantități mari de date sub o licență liberă. Aplicațiile se pot executa pe sute de sisteme de calcul independente și pot prelucra petabytes de informații. Fondatorul Hadoop este Doug Cutting, care a preluat informații din documentele publicate de Google și a implementat [Google File System](http://ro.wikipedia.org/w/index.php?title=Google_File_System&action=edit&redlink=1) și [Map reduce](http://ro.wikipedia.org/wiki/Apache_Hadoop#JobTracker_.C8.99i_TaskTracker:_Motorul_MapReduce) în [Nutch](http://ro.wikipedia.org/w/index.php?title=Nutch&action=edit&redlink=1" \o "Nutch — pagină inexistentă).

Pentru stocarea datelor Apache Hadoop utilizează sistemul de fișiere HDFS (Hadoop Distributed File System) (Figura 1) care creează copii ale datelor și le distribuie pe mai multe noduri din cadrul unui cluster pentru a permite acces rapid și fiabil la informațiile stocate. Unele distribuții de Hadoop oferă alternative la HDFS cum ar fi NFS (pentru MapR) și Apache Cassandra (pentru DataStax). Aceste sisteme de fișiere oferă avantaje precum performanțe mai bune și posibilitate de accesare a informațiilor prin intermediul unor sisteme de operare ca Windows, OS X și Linux.

Apache MapReduce este un model de programare pentru procesarea unor seturi foarte mari de date utilizat în special pentru procesare distribuită în cadrul unor clustere de calculatoare. Aplicația este împărțită în fragmente mici, fiecare putând fi executat sau reexecutat pe orice nod din cluster.



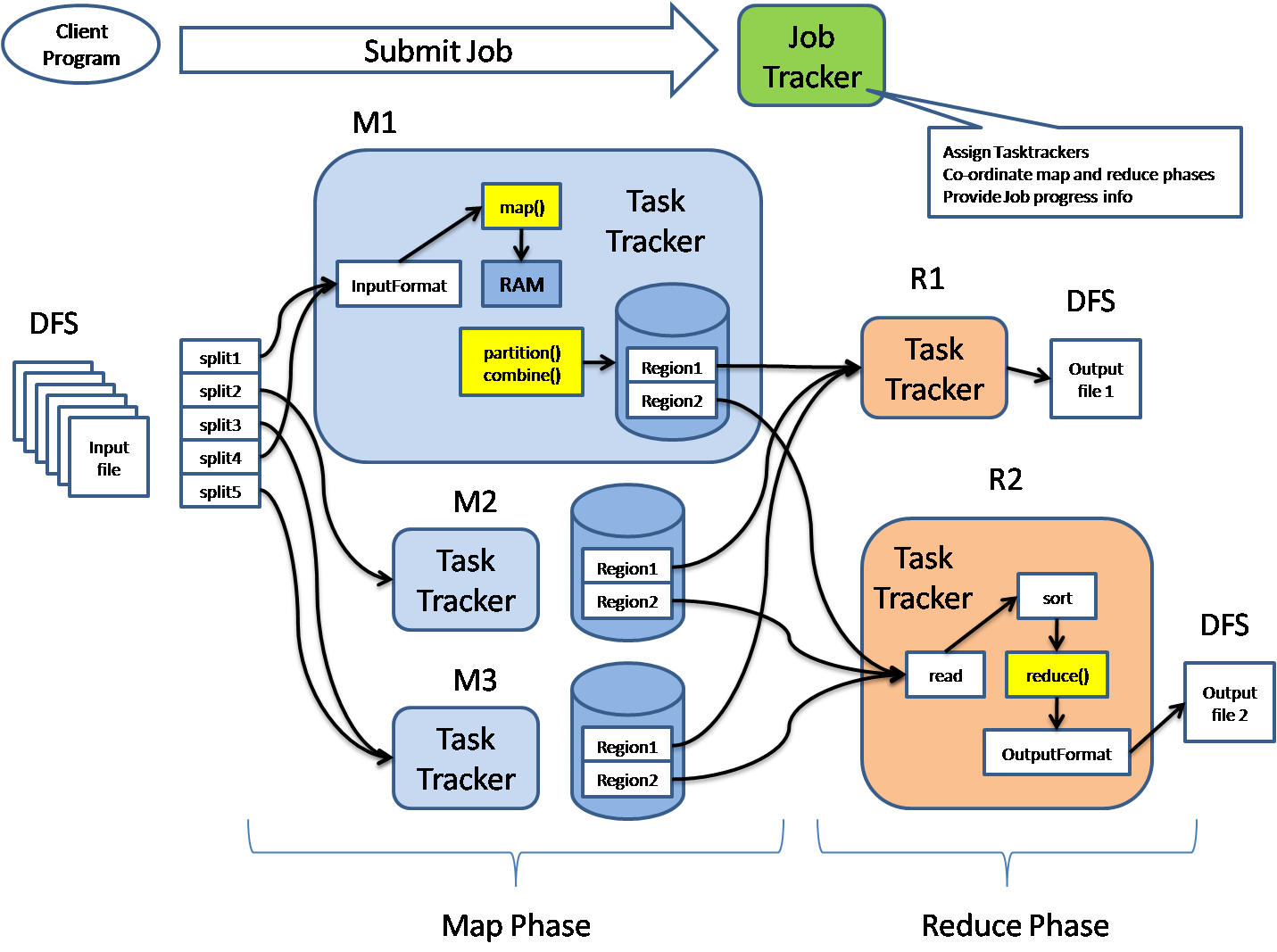
1. Descrierea proiectului

In cadrul proiectului, dupa studierea a ceea ce inseamna acest intreg environment prezentat in partea de introducere succint, am hotarat ca una din incercari sa fie o problema de altfel tipica, si de incepatori, de management al unei cantitati mari de date din domeniul meteo, dorind, ca prin introducerea unui input.txt, cuprinzand datele meteorologice culese de-alungul unui intreg an, sa aflu prin tehnica MapReduce, diferite date, de exemplu Temperatura maxima, sau Temperatura medie.

In fisierul de input, informatia pe care am utilizat-o a fost doar cea referitoare la valoarea temperaturii – aflata in fiecare linie a documentului la intre pozitiile 88-92, si informatiile referitoare la an – aflate la pozitiile 15-19. Acest lucru l-am gandit in sensul unei posibile dezvoltari ulterioare, de exemplu de a avea informatii culese din mai multi ani.

Cele doua clase statice, Map si Reduce, contin fiecare o metoda, map si respective reduce, prin intermediul carora se realizeaza operatiile de mapare a intregii informatii din fisierul input.txt, si de reducere, operatie, rin intermediul careia este posibila aflarea rezultatului dorit – temperature maxima.

In metoda run, sunt prezente setarile necesare pentru declansarea Job-ului, in cadrul frameworkului existand roluri de JobTracker si TaskTracker.



Cum functioneaza exact aceste joburi?

Clientul inainteaza in primul rand, un job. Acest job ajunge la JobTracker. In background, totodata va fi copiata informatia in cadrul unui fisier de date, ajungand in final in cadrul HDFS. Aici apare cea mai importanta calitate a frameworkului, si anume calitatea de a duce operatiile la date, si nu datele la operatii. TaskTrackerul este cel care este practci aproape de informatie, si urmatorul pas va fi cel de realizare al unui plan de executie, JobTrackerul dand submit catre TaskTracker. Acesta, pe langa realizarea computatiei effective, in cicluri repetitive, are si rolul de a asigura o alta calitatea majora a frameworkului si anume toleranta mare la esec.

1. Dificultati intampinate

Una dintre dificultatile majore a fost setarea si instalarea intregului environment. O prima incercare a fost prin intermediul unui virtual machine, VirtualBox, si prin utilizarea Hortonwork Sandbox. Din pacate insa, incercarea a esuat, datorita cerintelor reprezentate de un sistem de operare pe 64 biti. O a doua incercare a fost in cadrul Ubuntu, unde instalarea a reusit, insa setarea unui cluster a esuat. Dupa executarea comenzii --jps --, pentru a vizualiza daemonurile care erau active pe sistem, cele e JobTracker, TaskTracker, nu erau active. O ultima incercare a fost una online, insa datorita unor taxe de inregistrare nu am continuat.

1. Rezultate

In urma maparii si a reducerii datelor de intrare, in fisierul de output, rezultatul este valoarea temperaturii maxime. Datorită volumelor din ce in ce mai mari de date din cadrul organizațiilor, stocarea și procesarea acestora cu ajutorul sistemelor de baze de date clasice a devenit din ce în ce mai dificilă. Astfel a fost necesară dezvoltarea unor tehnologii care să stocheze cât mai eficient datele și care sa aplice tehnici de procesare paralelă pentru reducerea timpului necesar extragerii informațiilor dorite.

Printre problemele legate de integrarea ”big data” se numără varietatea surselor de date, calitatea datelor ce urmează a fi integrate și vizualizarea datelor. Una dintre cele mai utilizate tehnologii de integrare ”big date” este MapReduce, fiind utilizată de companii precum Google, Yahoo sau Facebook deoarece oferă o serie de avantaje: suport pentru seturi foarte mari de date distribuite în clustere de computere și posibilitatea de procesare atât a datelor structurate cât și a celor nestructurate.