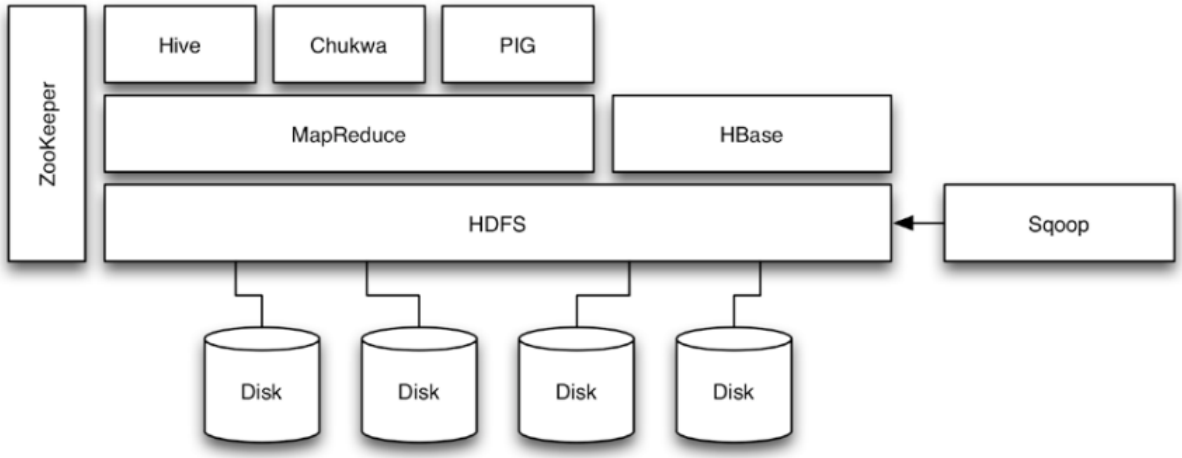
MINISTERUL EDUCAȚIEI ȘI CERCETĂRII
UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA
FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE

Apache Hadoop

Hădărean Andreea Roxana

Grupa 30242

Apache Hadoop reprezinta un framework open source Java ce are ca principal rol implementarea retelelor scalabile si de incredere. Acesta include unele subproiecte, precum MapReduce, Pig, ZooKeeper, Hbase si Hive.



Figură 1:Componentele Hadoop [1]

Acest framework se compune din instrumente si functionalitati ce permit serializarea datelor, accesul la sistemul de fisiere si comunicarea intre procese. Exista doua tipuri de configuratii: singulare si de tip cluster. Acestea includ in marea majoritate a timpului HDFS, optimizat pentru un debit mare de date de intrare. Se remarca existenta a doua componente esentiale: HDFS si MapReduce.

Rolul componentei HDFS este acela de a oferi acces rapid si scalabil la nivelul sistemului distribuit de fisiere. HDFS se bazeaza pe consistenta, o tranzactie neconsiderandu-se completa pana in momentul in care datele nu sunt scrise in cel putin doua volume configurabile. De asemenea, componenta inglobeaza toate discurile fizice sau sisteme de fisiere. MapReduce are ca principal rol management-ul nodurilor si

urmarirea modului in care se executa fiecare task. De asemenea, acesta stocheaza toate metodele de mapare si reducere scrise in Java.

Framework-ul Apache Hadoop inglobeaza de asemenea si o alta serie de framework-uri: Chukwa, Hive si HBase. Chukwa are ca principa rol colectarea datelor pentru a monitoriza, a afisa si a analiza inregistrarile de la nivelul sistemelor distribuite de dimensiuni mari. Hive se ocupa cu stocarea datelor, permitand de asemenea stocarea lor si transformarea folosind un limbaj de interogare similar cu SQL. Hbase reprezinta baza de date NoSQL care stocheaza in timp real, extrage si permite cautarea in tabele de dimensiuni mari, ruland deasupra HDFS.

Totodata, framework-ul pune la dispozitie si un set de utilitare. Pig reprezinta un set de instrumente pentru a programa procesele de analiza a datelor, oferind un limbaj pentru programare, transformarea datelor si paralelizarea proceselor. Sqoop permite importarea si exportarea datelor din bazele de date relationale in Hadoop sau Hive sau invers. Pentru configurarea, sincronizarea evenimentelor si organizarea nodurilor intr-o retea construita folosind Hadoop este utilizat ZooKeeper.

Pentru sistemele distribuite construite cu ajutorul framework-ului Hadoop se poate realiza deploy-ul in trei moduri: local, pentru implementare si testare, bazat pe thread-uri pe o singura masina viruala, pseudo-distribuit, pe mai multe masini virtuale, existand un singur nod si distribuit, pentru productie, existand mai multe noduri. Fiecare nod poate executa fie cod MapReduce, fie instructiuni HDFS, executat de un proces ce ruleaza in fundal.

Procesele ce ruleaza in fundal se executa pe trei nivele diferite a sistemului distribuit: master, slave si user application. La nivelul Master sau Name Node, se executa procesul de management al sistemului de fisiere, se monitorizeaza executia task-urilor , se replica blocuri de date si se mentin uniform distribuite. Aceasta componenta poate rula si pe un alt calculator. La al doilea nivel, Slave sau Data Nodes, sunt stocate blocurile de date in sistemul local de fisiere, sunt stocate meta-date, sunt transmise aceste date catre Master. Ultimul nivel, user applcation, executa implementarea contractelor pentru Java, ofera parametrii specifici executiei applicatiei si seteaza configuratia parametrilor ce sunt folositi in celelalte doua nivele.

O aplicatie construita cu ajutorul Hadoop consta intr-unul sau mai multe job-uri. Un job contine un fisier de configurare si una sau mai multe clase Java. Datele trebuie sa existe in HDFS. Date de intrare pot avea mai multe formate, precum KeyValueTextInputFormat sau MultiFileInputFormat. Datele de iesire se mapeaza pe formatul datelor de intrare.

In concluzie, Hadoop reprezinta un framework scalabil, flexibil, rapid si tolerant la esec, util cand se doreste stocarea si procesarea datelor. Eficienta acestuia a fost dovedita si prin faptul ca este folosit de companii renumite, precum Google si Facebook.

Bibilografie

[1] E. Ciurana and M. Kalali, "Getting Started with Apache Hadoop," *DZone Refcardz*, pp. 1-6, 2018.