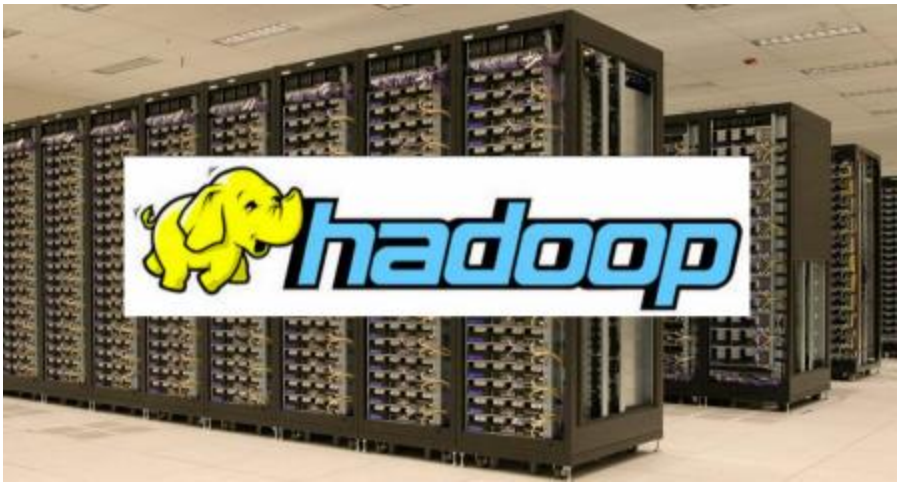
Hadoop Nedir ?

Hadoop , büyük veri kümeleri ile birden fazla makinada paralel olarak işlem yapmamızı sağlayan Java ile yazılmış açık kaynak kodlu kütüphanedir .

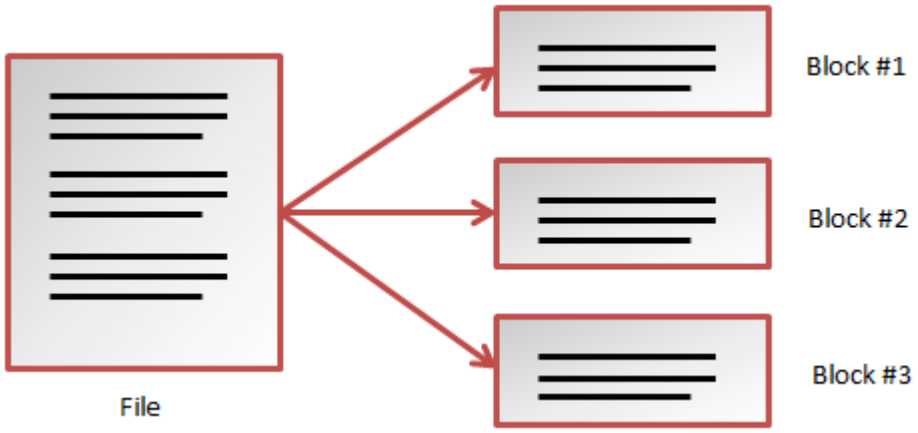


Hadoop büyük verileri birden fazla makinada saklar ve yönetir .

Hadoop Büyük Verileri Nasıl Saklar ? HDFS Nedir ?

Hadoop içerisinde büyük verileri sakladığımız bileşene **HDFS** (Hadoop Distributed File System) denir .

Büyük verileri HDFS sistemine yüklediğimiz zaman , Hadoop bu verileri bloklara ayırır .



Farklı bloklara ayrılan veriler Hadoop Cluster üzerinde farklı node lara dağılır .

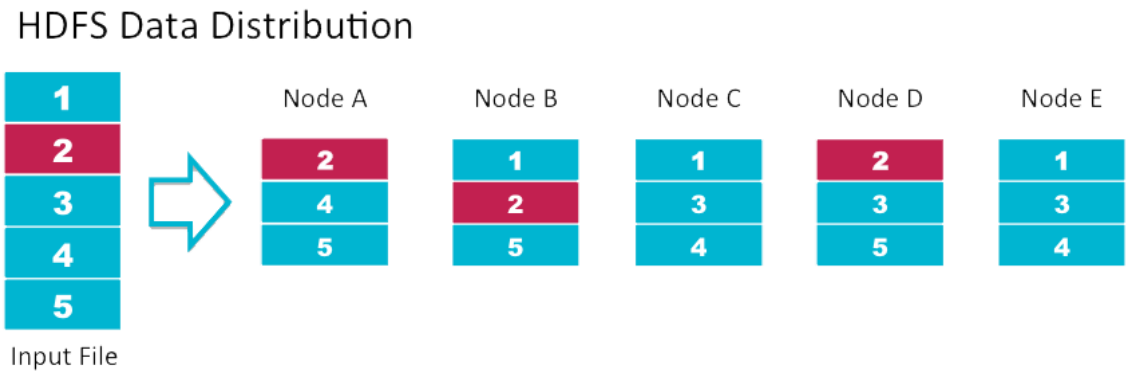
Şimdilik her bir node u farklı bir makina olarak düşünebiliriz .

Alttaki şekilde görüldüğü gibi **Input File** içerisindeki bloklar farklı node lara dağıtılmıştır .

Burada dikkat etmemiz gereken en önemli husulardan bir tanesi her bir blok **çoklanarak** kaydedilmiştir .

Mesela **2** numaralı blok 3 farklı (Node A , Node B , Node D) node üzerine dağıtılmıştır. **(Replication factor)**

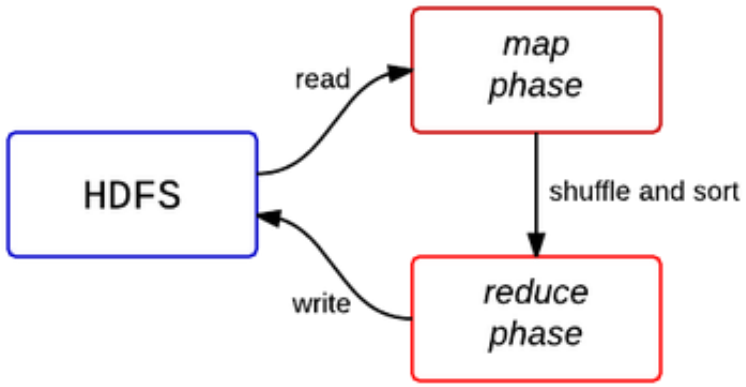
Bunun asıl nedeni ise node lardan bir tanesi zarar gördüğünde veya sistemden çıktığında veri kaybının yaşanmasını engellemek



Hadoop Verileri Paralel Olarak Nasıl İşler ? MapReduce Nedir ?

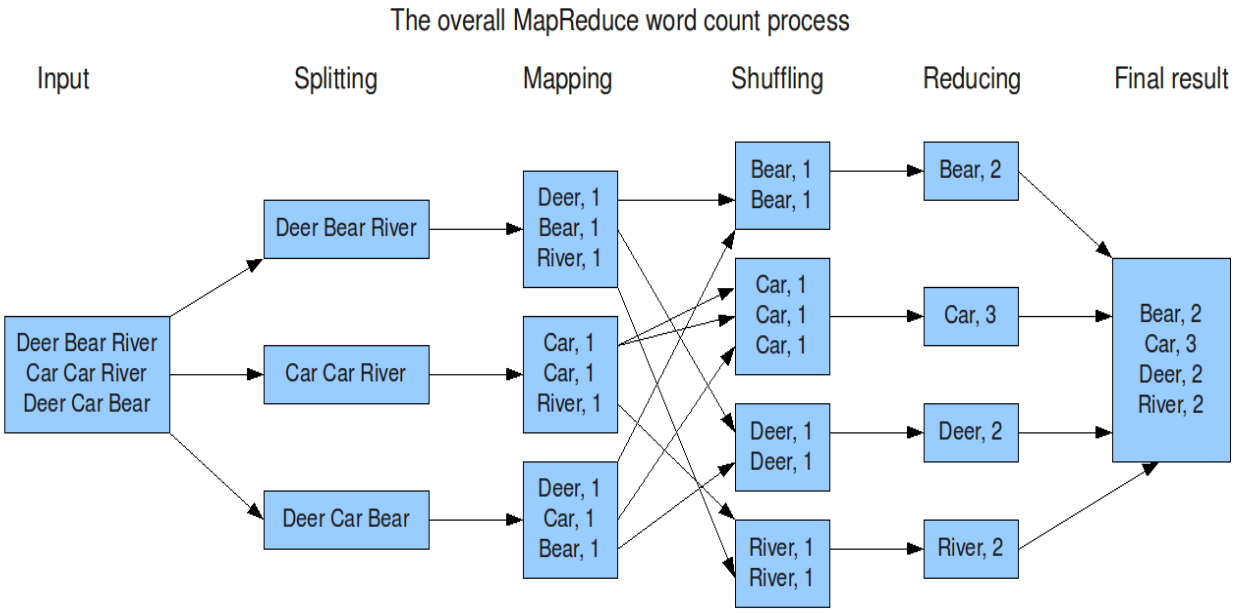
Hadoop içerisinde büyük verileri paralel olarak işleyebileceğimiz bileşene **MapReduce** denir .

Veri kümeleri HDFS üzerinden yüklendikten sonra **Map** ve **Reduce** fazları işletilir . Bu kodlamaları Java , Pig ve Hive .. ile geliştirebiliriz



Örnek olarak bir text dosyasının içerisindeki kelime sayısını bulan MapReduce programını inceleyelim .

MapReduce şu adımlardan oluşacaktır ;



mapreduce nedir

- *Spliting : Veriler 64 MB lık bloklara ayrılır . Bu değer değiştirilebilir*
-
- *Mapping : Burada her bir kelime key(word) ve value(1) şeklinde bölümlere ayrılır .*
-
- *Shuffling : Map işleminden çıkan sonuçları Reducer a yönlendirir . Amacımız word-count uygulaması oldugu için aynı kelime grubu aynı Reducer a yönlendirilir .*
-
- *Reducing : Gelen sonuçlar üzerinden toplama işlemi yapılır ve sonuçlar istediğiniz kaynaklara yazılır (HDFS , SQL , NoSQL)*

Özet

Genel olarak özetlemek gerekirse çok yüksek trafikte akan bir veriniz olduğu zaman (Örnek günlük 100 milyon +) verileri HDFS üzerinde saklayabilir ve MapReduce ile verilerinizi analiz edebilirsiniz .

Alternatif olarak diğer NoSQL (Mongo , ElasticSearch) saklama yöntemlerini yada Apache Spark gibi paralel veri işleme yöntemlerini tercih edebilirsiniz .

Buna ihtiyaçlarınıza göre karar vermelisiniz . Her sistemin kendine göre artı ve eksileri vardır

Bu işlemleri Java da paralel processing ile yaparım diyorsanız çok büyük development maliyetinin altına girmiş olursunuz .

Hadoop un size sağladığı (replication factor , MapReduce health check) gibi işlemleri kendiniz yönetmek durumunda kalırsınız