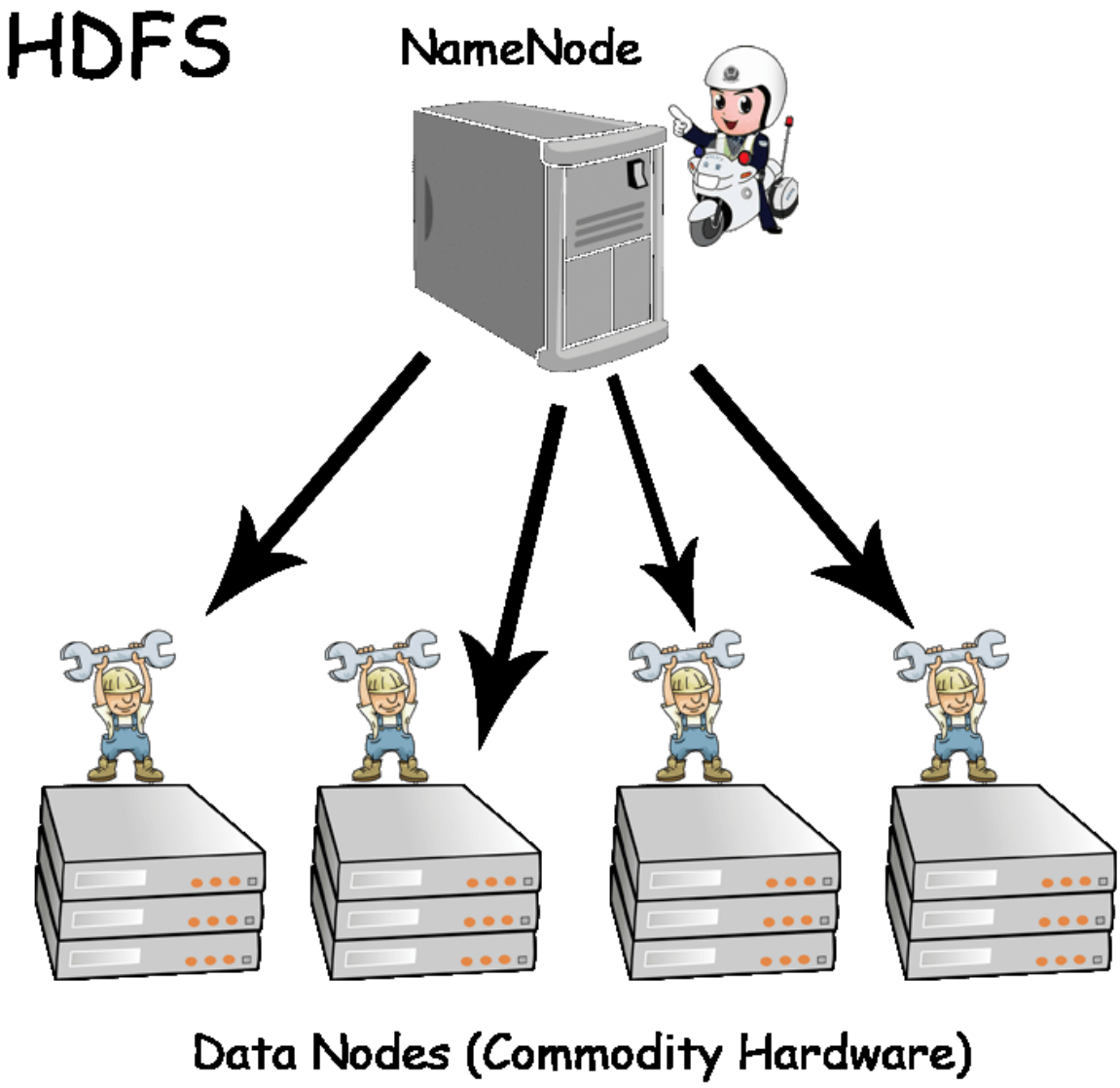
Hadoop HDFS Mimarisi

Bu bölümde Hadoop HDFS(Hadoop Distribted File System) Mimarisi hakkında bilgiler vereceğiz



hdfs-hadoop

Son yıllarda toplanan verilerin devasa boyutlara ulaşmasından dolayı verileri tek bir makine yerine birden fazla makineda saklama çözümleri ortaya atılmıştır.(**distributed filesystems**)

Bu noktada Hadoop projesinde, verileri birden fazla makinede saklayan HDFS kütüphanesi geliştirilmiştir

Genel Özellikleri

- Petabyte seviyesindeki büyük verileri saklayabilir
- Pahalı bir donanım satın almanıza gerek yoktur. Günlük hayatta kullandığımız bir kaç makine ile Hadoop cluster kurabiliriz
- HDFS içerisine veriler kopyalandıktan sonra birçok kaynak üzerinden aynı anda erişim sağlanabilir
- Verilere hızlı bir erişim sunar(**Low-latency data access**)
- Veriler küçük dosya blokları halinde saklanır

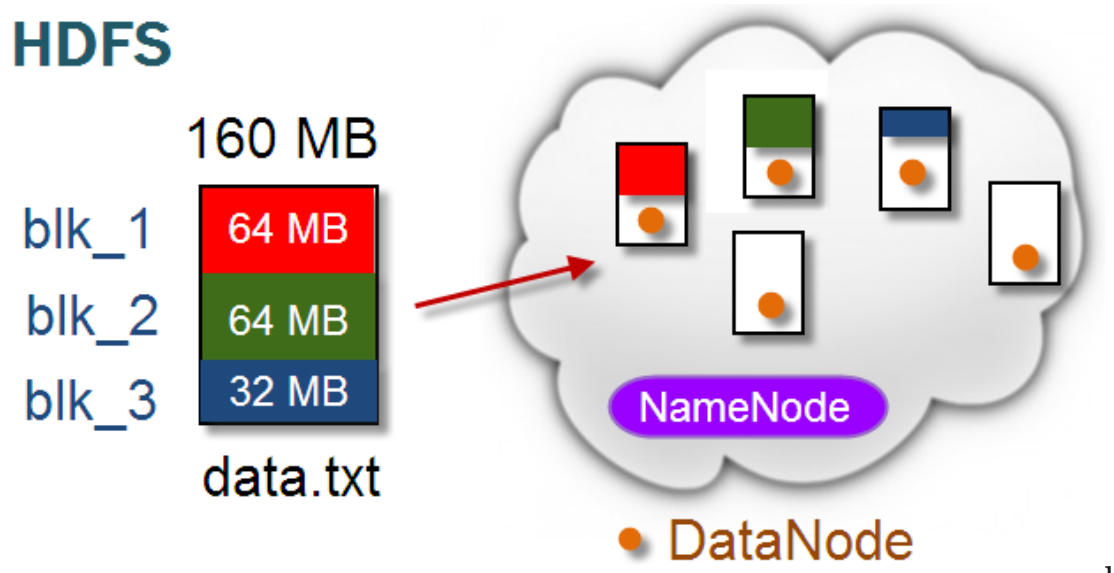
Şimdi HDFS bileşenlerini inceleyelim

Bloklar(Blocks)

HDFS üzerinde veriler bloklar halinde saklanır. Varsayılan değeri ise 128 MB'dir ve kullanıcı tarafından değiştirilebilir

Blok yapısının avantajları şunlardır;

- Yüklemek istediğimiz dosya boyutları disk boyutundan büyük olabilir.Böylece verileri bloklara ayırarak farklı diskler üzerinde kopyalanmasını sağlamış oluruz
-
- Disklerden bir tanesinde hata meydana gelirse, o disk üzerindeki bloklar farklı makinelere kopyalanır
-
- Blok boyutu çok küçük olduğu (512 byte gibi) durumda disk üzerinde arama işlemi maliyetli olduğu için Hadoop işlemleri yavaşlar.Bu yüzden blok boyutu ihtiyaca göre 64 MB,128 MB verilebilir
-
- Ayrıca blok boyutu çok büyük olursa(1 gb gibi) birden fazla disk üzerinden okuma yapmak yerine tek bir diskten okuma yapılacağı için yine Hadoop işlemleri yavaşlar


HDFS
160 MB
blk_1
blk_2
blk_3
64 MB
64 MB
32 MB
data.txt
NameNode
DataNode

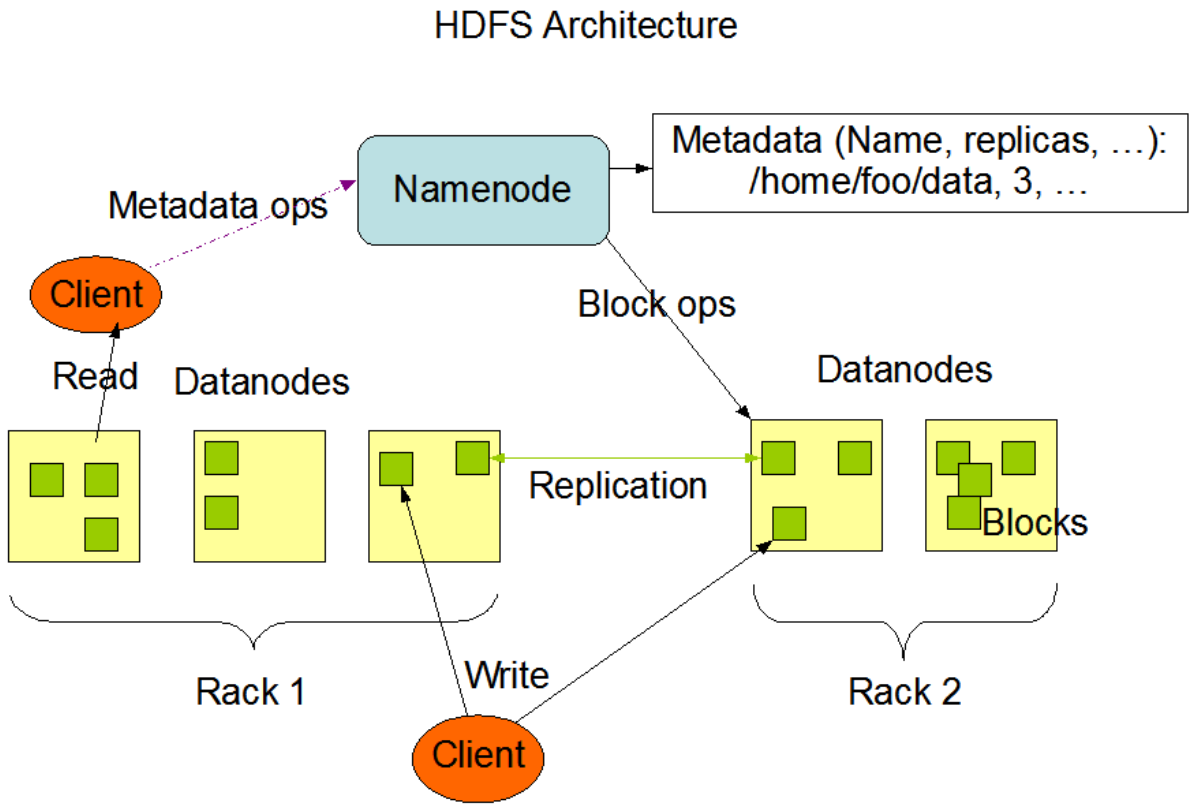
hdfs-block

Namenode ve Datanode

Namenode genel olarak verilerin nerede saklandığı bilgisini tutar. Hadoop sisteminde master olarak düşünebiliriz.Datanode ise verilerin saklandığı makinalardır(slave)

Datanode makineleri belirli periyotlarda hangi blokları sakladıklarını Namenode'a bildirir.

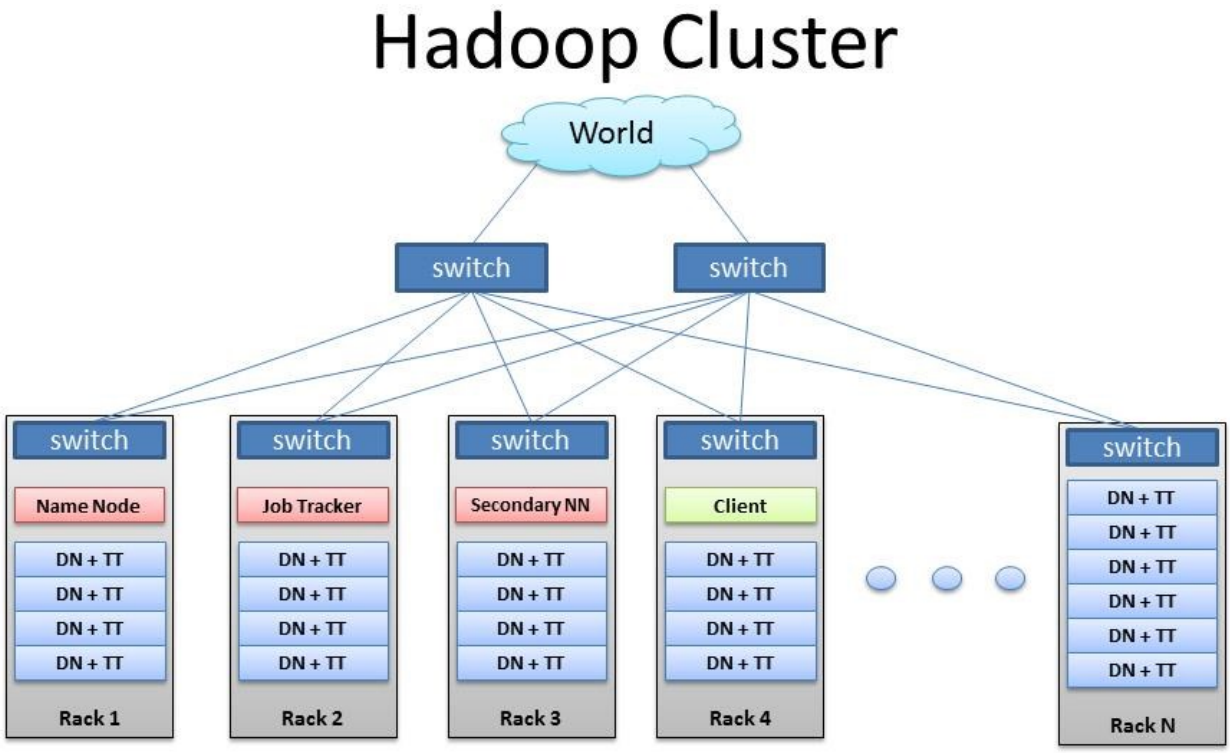
Eğer namenode makinesinde bir hata meydana gelirse Hadoop içerisindeki verilere erişemeyiz.



namenode-datanode

Hadoop Rack Nedir?

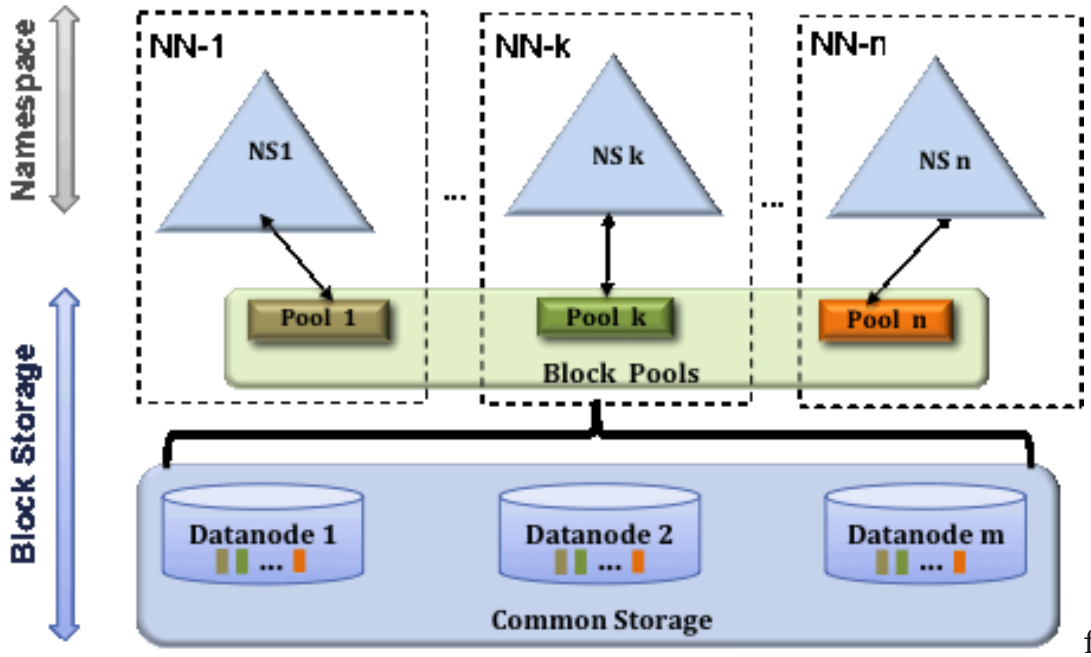
Hadoop cluster kurulurken belirli sayıdaki node aynı switch'e bağlanır.Bu yapıya **rack** denir



hadoop-rack

HDFS Federation Nedir ?

HDFS Federation ile birden fazla NameNode oluşturabiliriz.Bu sayede Hadoop cluster içerisindeki uygulamaları birbirinden izole edebiliriz.Mesela 10 makinalık bir Hadoop Cluster olduğunu düşünelim.7 makineyi bir müşterimize,3 makineyi diğer müşterimize ayırabiliriz.Bu sayede birbirlerinin kaynaklarını etkilemezler



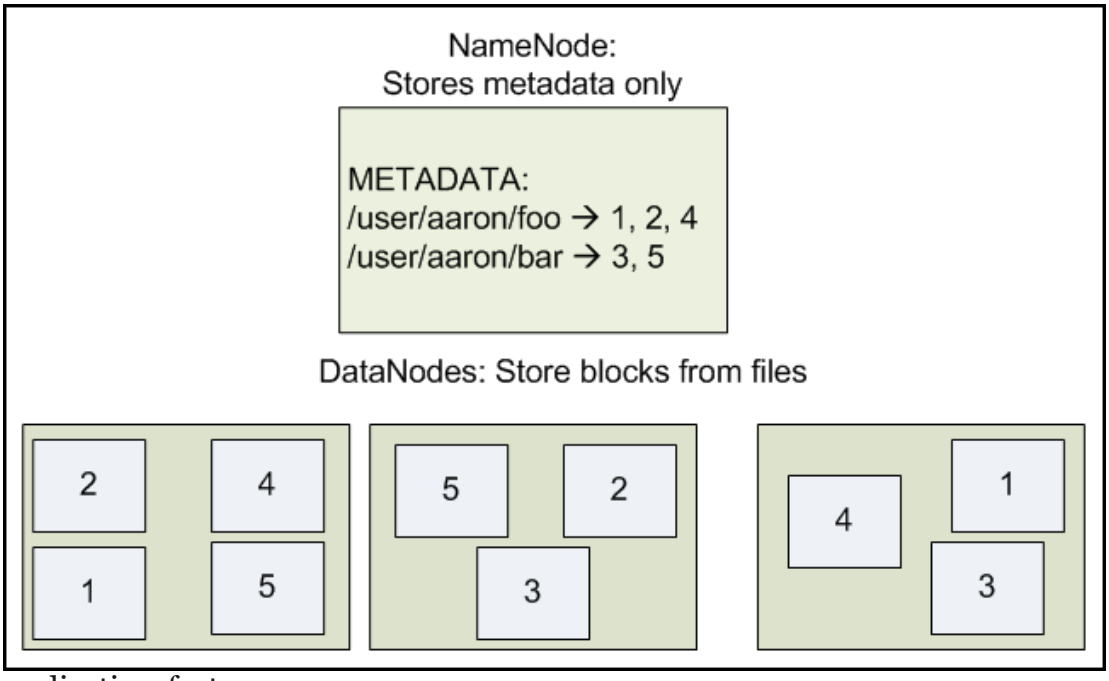
federation

Hadoop Replication Factor Kavramı

HDFS sistemine bir dosya kopyaladığımız zaman Hadoop bu dosyaları bloklara bölerek DataNode makinelerine gönderir.Ayrıca makine sorunlarında veri kaybı yaşanmaması için dosyaların kopyaları farklı makinelerde saklanır

hdfs-site.xml dosyasında şu şekilde konfigurasyon yapılır

```
1 <property>
2    <name>dfs.replication<name>
3    <value>3<value>
4    <description>Block Replication<description>
5 <property
```



replication-factor