Big Data

Big Data bu özelliklerden bir ya da birkaçına sahip olmalıdır.

Volume, Velocity, Variety, Verification, Value. 5V kuralı olarak da geçer. Bazı kaynaklarda 3V veya 11V olarak da geçer. Fakat standartlaşmış olan 5V kuralıdır.

Volume: Verinin hacmidir. Veriler her geçen gün hızla artıyor. Örneğin bir uçakta ortalama 5000 tane sensör var. Buda bir uçuşta terabaytlarca veri ürettiği anlamına geliyor. Bir ayda Hexabytelar Petabytelar seviyesinde veri üretmiş oluyor.

Velocity: Verinin hızı demek. Örneğin Google’da bir dakikada 4.000.000 arama yapılıyor. Dakikada yapılan işlemler çok hızlı bir şekilde olduğu için işlenme hızının da hızlı olması gerekiyor.

Variety: Verinin çeşitliliği demek. Verilerin belirli bir yapısı yoktur, genellikle değişkendir. Resimler, ses dosyaları, text dosyaları örnek verilebilir.

Verification-Veracity(Doğrulama): Veriler içerisinde anlamsız kayıtlar olabilir. Anlamsız kayıtlar analizlerimizin sonuçlarını etkilediği için bu kayıtları temizlememiz gerekir.

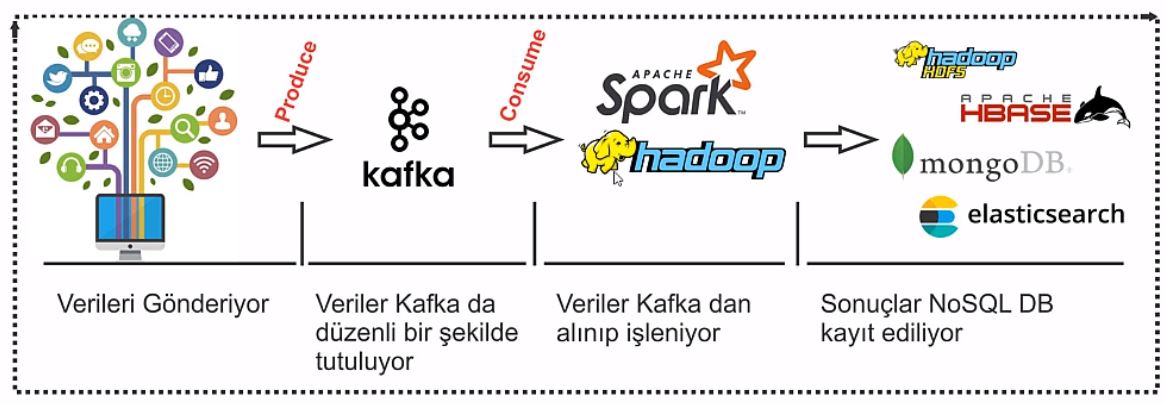
Value: Verinin değerli olması. Büyük Verinin üretilmesi ve işlenmesi katmanlarında elde edilen verilerin şirketimiz için artı değer sağlıyor olması gerekiyor.

Big Data Kütüphaneleri

Hadoop ve Spark Big Datanın popüler kütüphanelerindendir. Hadoop’un iki bölümü var. HDFS ve MapReduce bölümü. HDFS de verilerimiz ve dosyalarımız dağıtık bir şekilde tutulabiliyor. MapReduce ise bu tutulan verilerin Map ve Reduce işlemlerinden geçirilerek işlenmesinden oluşuyor. Spark’ın Hadoopdan farkı; Sparkda bir dosyalama sisteminin bulunmaması. Spark RAM üzerinde çalışıyor buda ona hız katıyor. Spark’ın web sitesinde Hadoopdan 100 kat daha hızlı olduğu savunuluyor.

* Hadoop(HDFS) gönderdiğiniz dosyaları bloklara ayırarak saklar.
* Hadoop(MapReduce) gönderdiğimiz dosyaları bloklara ayırarak işler.
* Makinelerin birbirine paralel olarak bağlanmasına Cluster adı verilir.
* Ana makine Master diğer makineler ise Slavdır.
* Cluster de saklama işlemleri çoklanarak kayıt edilir(Replication Factor).
* Hadoop(HDFS) dosyaların kopyalarını farklı makinelere dağıtır.

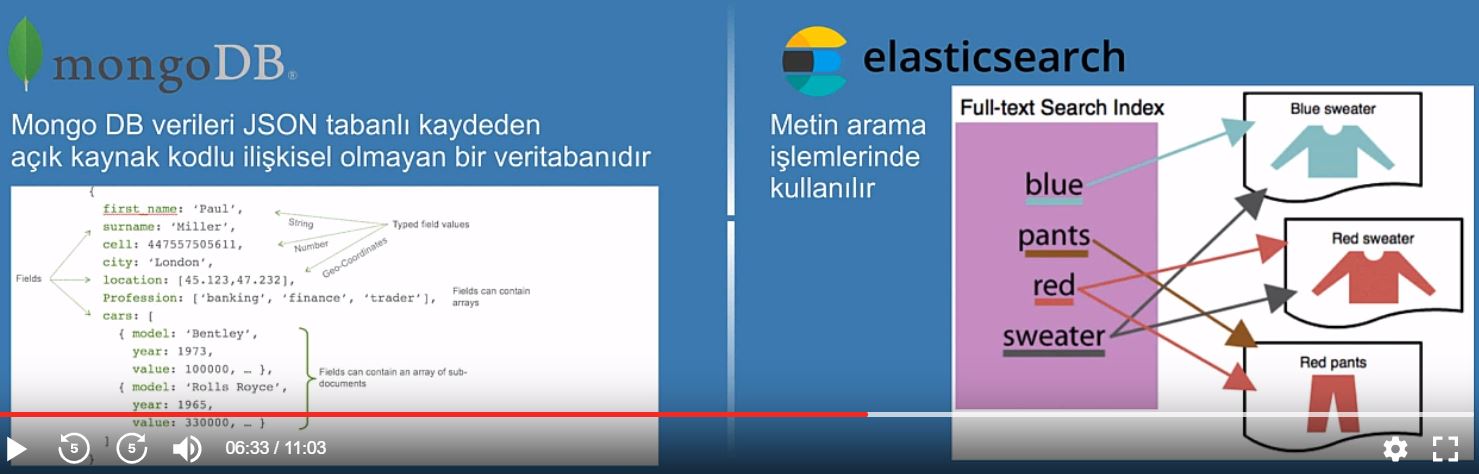
Apache Kafka akan verileri bir queue(mesaj kuyruğu) içerisine atarak Hadoop, Spark, Elastic Search gibi diğer sistemlere transfer etmemizi sağlar.



NoSQL(Not Only SQL): NoSQL kavramı, yıllardır bilişim dünyasında vazgeçilmez bir yere sahip olan RDBMS’ye altarnatif olarak ortaya çıkan, aslen internetin gün geçtikçe artan verisini depolayabilmek ve yüksek trafiğe sahip sistemlerin ihtiyacına cevap verebilmek amacıyla ortaya çıkmış ölçeklendirilebilen sistemlere genel olarak verilen addır. Örnek: Apache HBASE, Amazon Dynamo DB, Mongo DB, Cassandra. 4 tane NoSQL veritabanı çeşidi var. Document, Wide Column, Key Value, Graph DB.

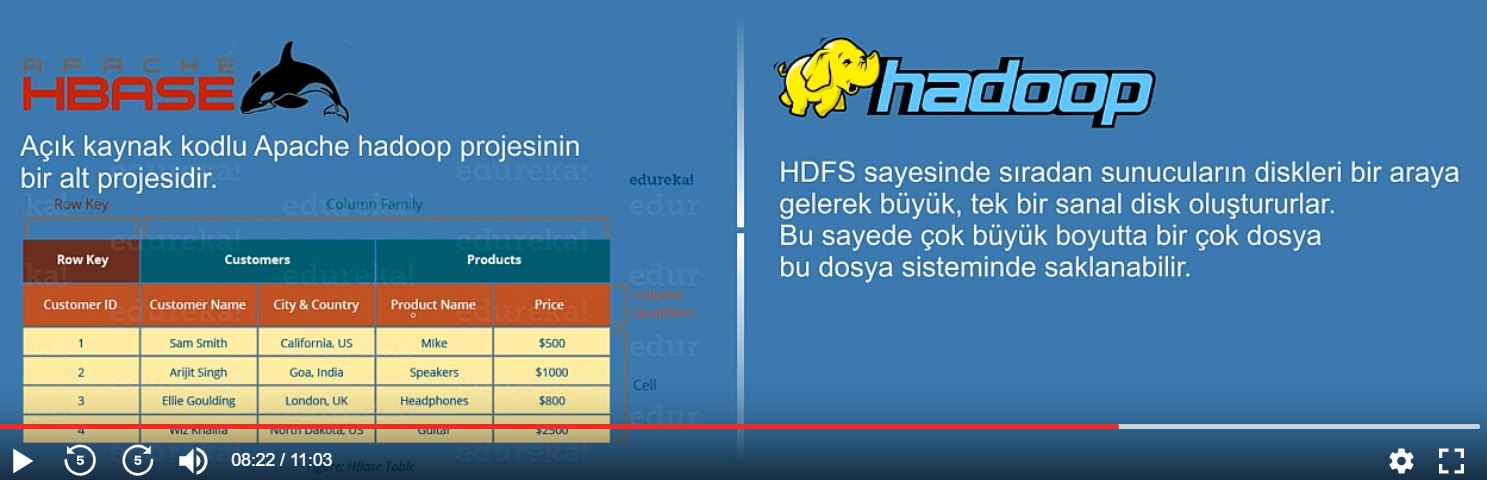
Doküman (Document) Tabanlı: Veriler belge şeklinde kaydedilir. Mesela bir JSON verisi bir doküman olarak kabul edilebilir.

Doküman Tabanlı NoSQL veritabanına örnek olarak:



Wide Column Tabanlı: Buradaki yapı klasik veritabanlarına benzer. Tablolar, kolonlar ve satırlar vardır. En büyük farklı ise yapıyı belirlerken süper column olacak alanları belirleriz ve bu alanların altına her kayıt için farklı kolon gelebilir.

Wide Column Tabanlı veritabanlarına örnek:



Key Value Tabanlı: Bu yapıda her bir key karşılığında bir array yada veri yapısı bulunur. Amacımız keyler üzerinden hızlı bir şekilde verilere ulaşabilmektir.

Grap DB: Bu yapının genel amacı veriler arasındaki ilişkileri saklayabilmektir. Her bir veri birer Node olabilir ve buradaki asıl amacımız Nodeler arası ilişkileri yönleriyle birlikte tanımlayabilmektir.

Hadoop Nedir?

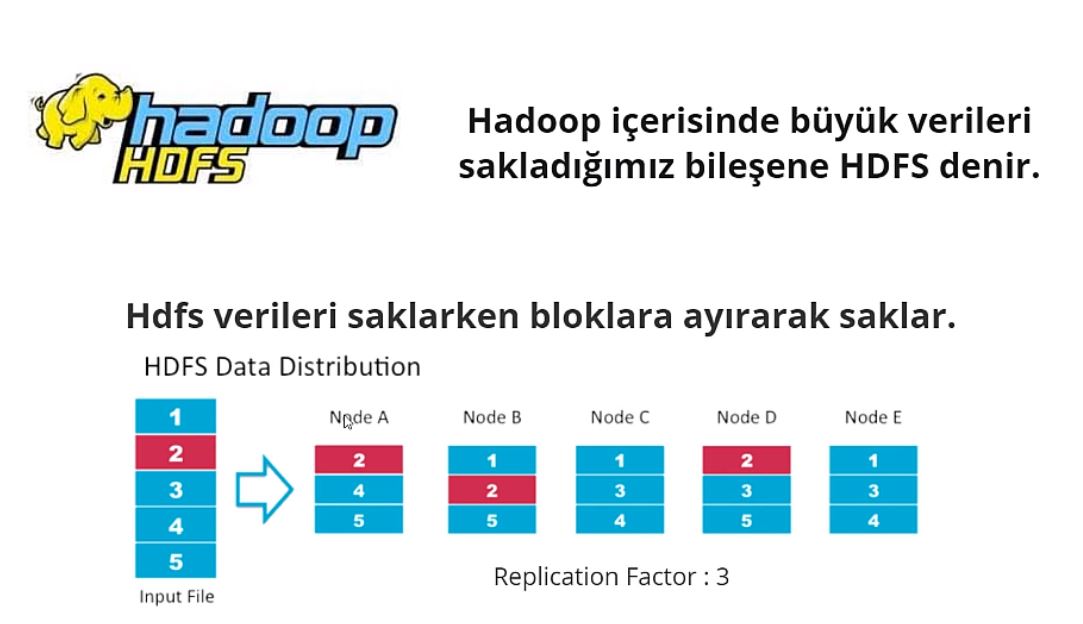
Hadoop büyük veri kümeleri ile birden fazla makinede paralel olarak işlem yapmamızı sağlayan Java ile geliştirilmiş açık kaynak kodlu bir kütüphanedir. Büyük veriden kastımız Exabytelar ve petabytelar seviyesindeki veri. Hadoop 4 modülden oluşur. Bunlar Hadoop Common, HDFS(Hadoop Disturbed File System), Hadoop Yarn ve Hadoop MapReduce.

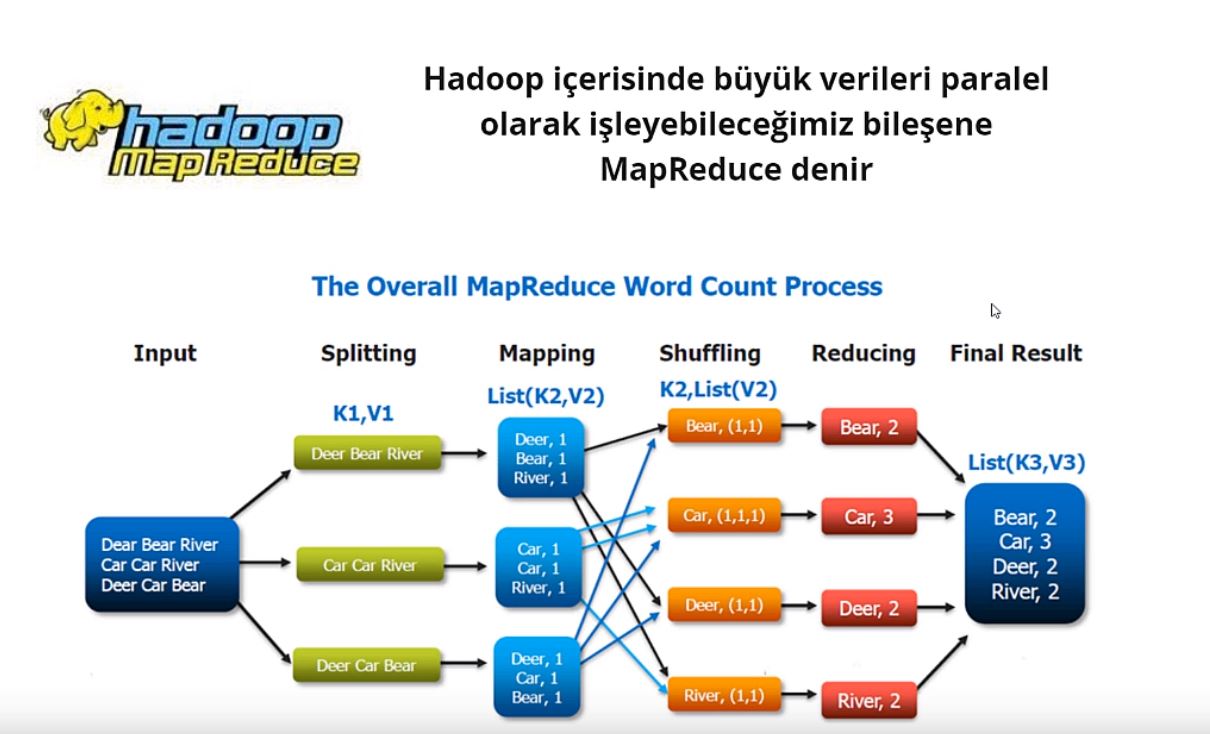
Hadoop Common bazı modüllerin Hadoop’a erişebilmesi için gerekli olan bir kütüphane.

HDFS(Hadoop Disturbed File System): Dosyaların ve verilerin dağıtık bir şekilde tutulması.

Hadoop Yarn: Hadoop üzerinde çalışan uygulamaların kaynak yönetimini kontrol ediyor. Ne kadar sistem kaynağı kullanması gerekiyorsa onu hesaplıyor. Örneğin işlemci ve RAM kullanımını hesaplıyor.

Hadoop MapReduce: HDFS üzerinde tutulan verileri işlememize yarayan Map ve Reduce araçlarıydı.



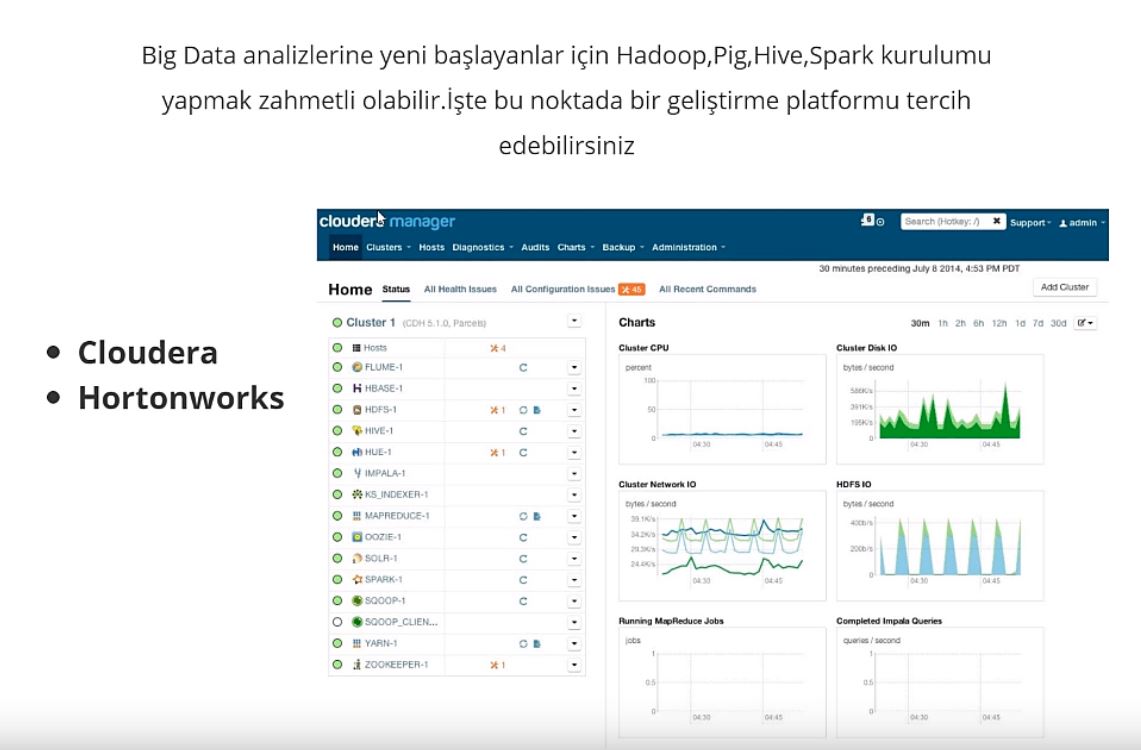




Hadoop Kurulum Modları

Hadoop 3 farklı mod ile çalıştırılabilir. Bunlar Standalone Mode, Single Node Cluster, Multiple Node Cluster.

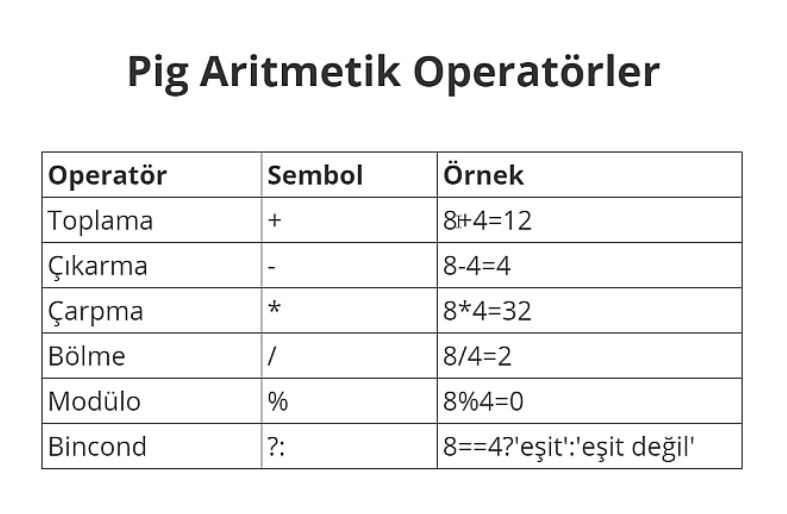


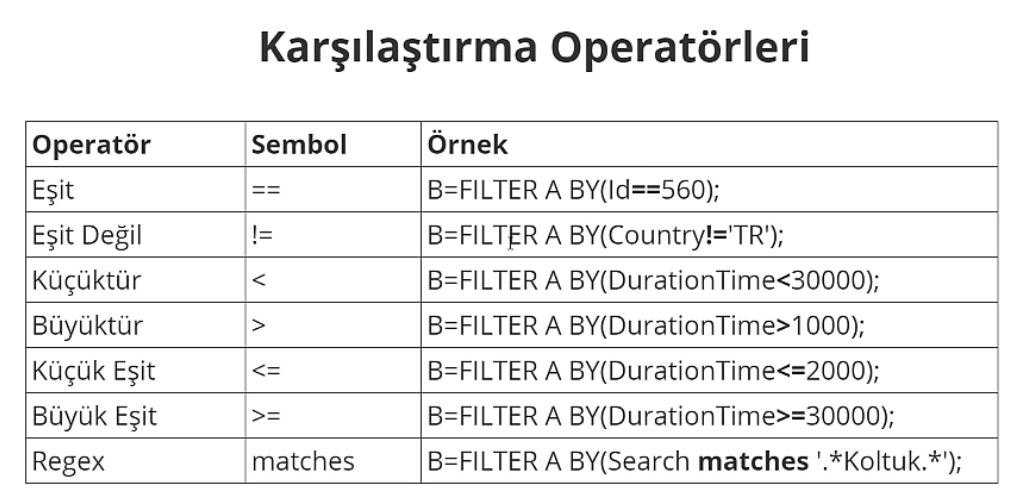


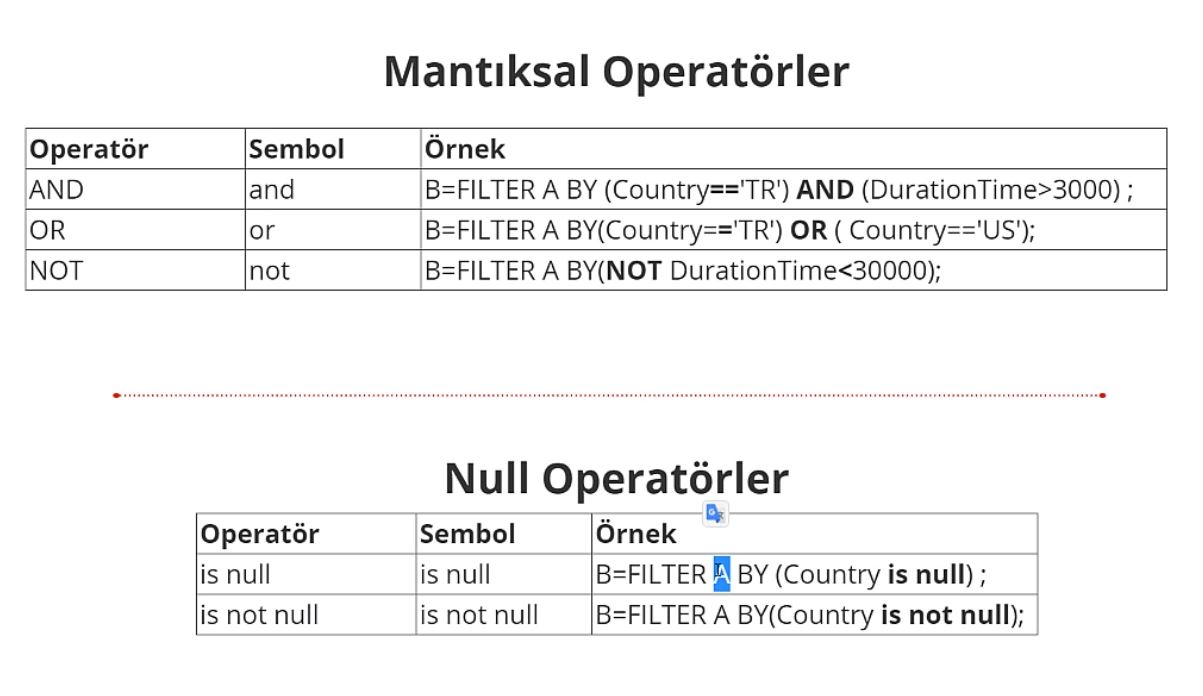
Apache Pig

Büyük verileri MapReduce ile analiz edebiliriz. MapReduce geliştirme yöntemleri şunlarıdır; Java, Python, Scala MapReduce. Apache Pig, Apache Hive. Apache Pig’in bir Script dili var. SQL’e benziyor. Buradan kodlar yazıyoruz.

Apache Pig Operatörler







Fonksiyonlar

