Bu bölümde Apache Spark MLib kütüphanesi(machine learning library) inceleyeceğiz

spark-mlib
MLib , Apache Spark projesi içerisinde makine öğrenmesi(machine learning) algoritmalarını içeren kütüphanedir

Genel özellikleri şu şekildedir

Ölçeklenebilir(scalable)

Apache Spark ile MLib metodlarını birden fazla makinede çalıştırabiliriz.Böylelikle büyük ölçekli verileri hızlı bir şekilde analiz edebiliriz

ölçeklenebilir makine öğrenmesi

Dil Desteği

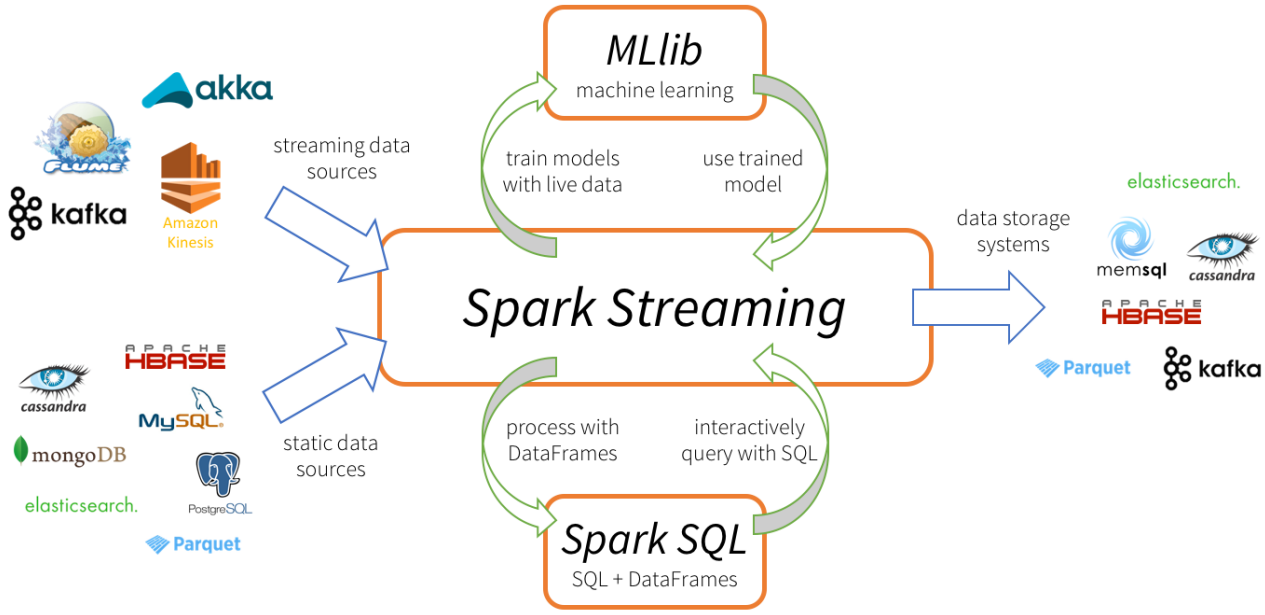
Scala,java,python ile geliştirme yapılabilir

Spark SQL desteği

Spark SQL kullanılarak SQL tabanlı işlemler yapılabilir

Spark Streaming

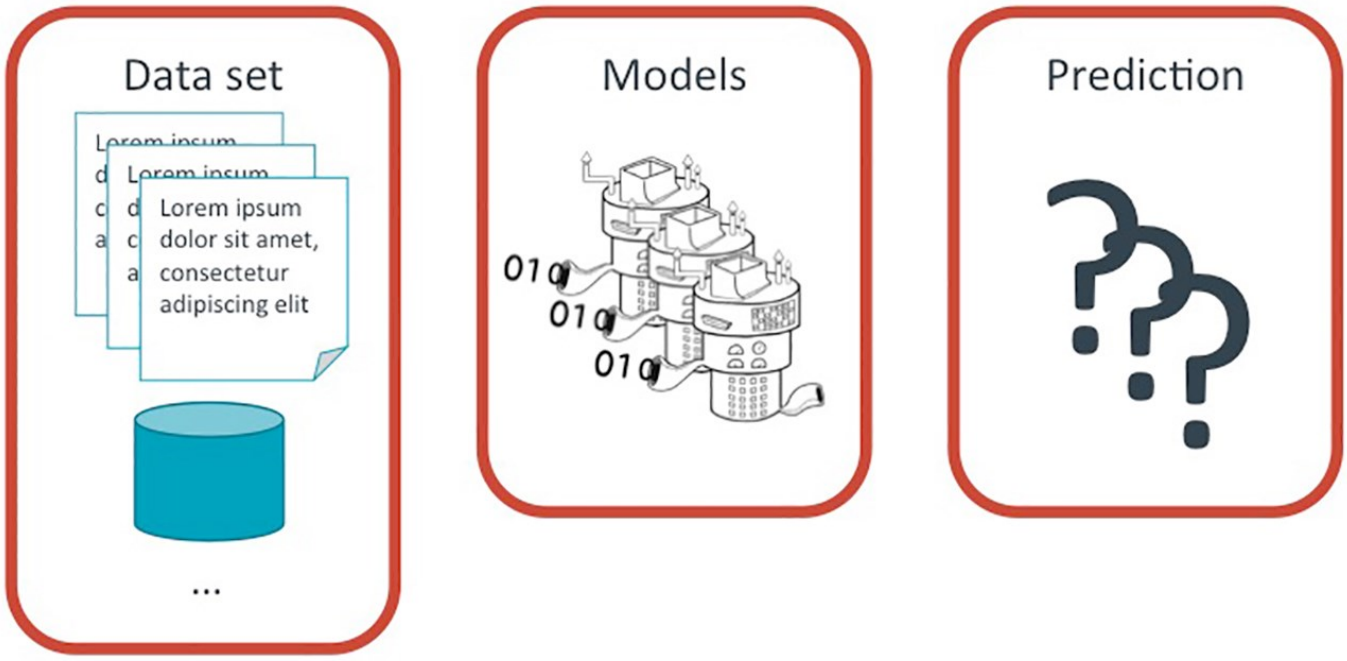
Spark Streaming ve Mlib kullanarak verileri anlık olarak analiz edebiliriz



spark-streaming-mlib

Cloud(Bulut) desteği

YARN, EC2, ve Mesos gibi ölçeklenebilir ortamlarda çalıştırılabilir



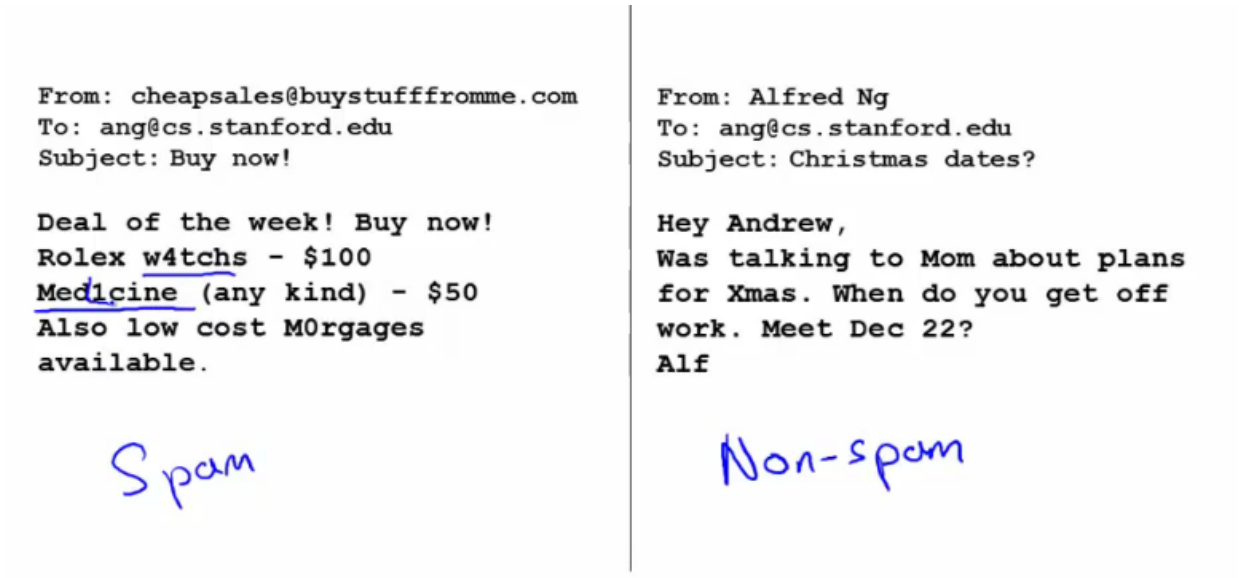
spark-mlib

Şimdi temel MLib kütüphanelerini inceleyelim

Classification(Sınıflandırma)

Veriler önceden belirlenmiş özelliklere göre sınıflandırılır ve bir model oluşturulur. Daha sonra yeni gelen verilerin özelliklerine göre hangi sınıfta olduğu tahmin edilir

Örnek verirsek mail içerisindeki gelen kelime gruplarına göre bir mailin spam olup olmadığını tahmin edebiliriz



spam-email

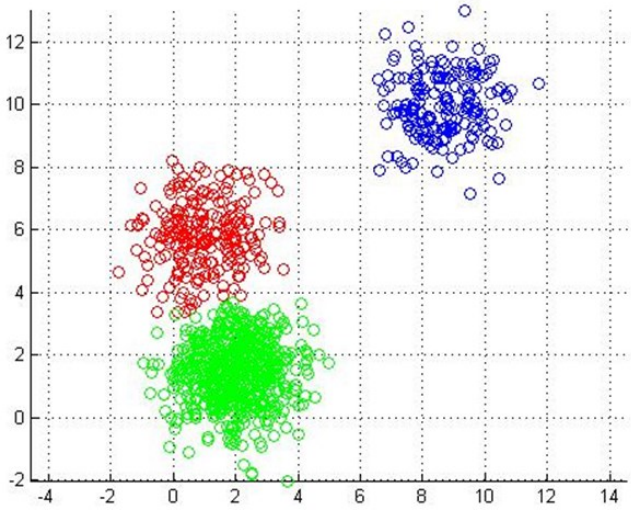
Clustering

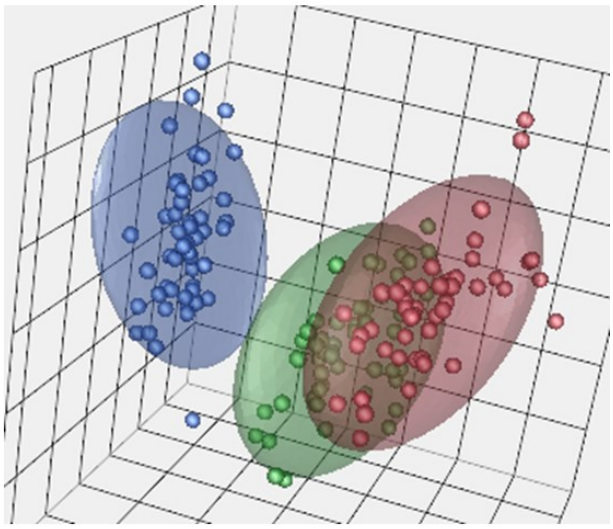
Veriler birbirlerine olan benzerliklerine göre kategorilere ayrılır .

Mesela bir network üzerindeki veri akışını incelediğimizde şu sonuçlar çıksın

- 100 kullanıcı 80-90 MB arası download yapıyor
- 150 kullanıcı 20-30 MB arası download yapıyor
- 1 kullanıcı 3 GB download yapıyor

Bu örnekte verilerin birbirlerine olan benzerliklerine göre kategorilere ayırdık.Burada anomaly detection yapmak istediğimiz durumlarda 3 Gb download yapan kullanıcı anormal bir davranış göstermiştir diyebiliriz

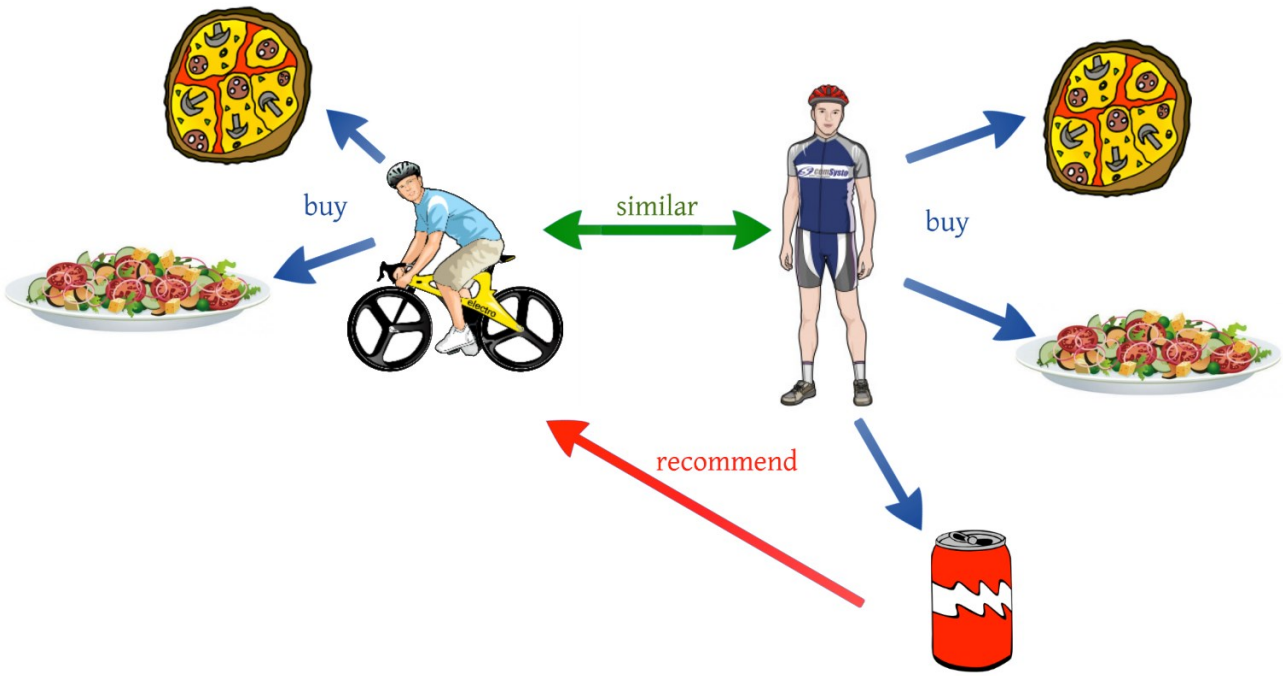




kmeans3d

Collaborative Filtering

Genellikle öneri sistemlerinde kullanılır . Mesela film öneri sisteminde , benzer filmleri iyi puanlayan izleyicilere benzer filmler önerilir



Collaborative-Filtering

Decision Tree

Verilerin özelliklerine göre bir karar ağacı oluşturulur ve bu karar ağacına göre sınıflandırılır . Sonrasında yeni bir veri geldiği zaman bu karar ağacına sorularak gelen verinin sınıfı belirlenir

Alttaki örnekte Titanic içerisinde kurtarılan yolcular için bir karar ağacı oluşturulmuştur

“sibsp” -> Gemideki eşler ve kardeşlerin sayısı