Apache Spark

Hangi big data teknoloji olursa olsun disrubuted(dağıtılmış ) bir mimari var

Spark ta bunlardan biri .buranın çıkış noktası 1 tane çok güçlü makine yerine n tane makineyi entegre etme düzeni.1 tane makine yerine n tane makineyi birbirine entegre edeyim bunlardan clustered bir sistem oluşturuyum ve gücümü katlayayım.

İntroduction spark

Large scale spark ta mevcut

2 tür scalability var 1 dikey büyüyebiliyoruz.2- yatay büyüyebiliyoruz

Biz dikeyi tercih etmiyoruz(databaseler buna örnek database lerde colon sayısını arttırmıyoz ama satır sayısı artabilir)

Yatay büyüme daha ucuz (big data sistemleri)

Yatay büyüme şöyle diyelim yukarıdaki mantık 1 tane güçlü makine yerine n tane standart makine gibi mesela dünyanın en hızlı okuyan adamı dakkada 150 sayfa okusun (bu en güçlü makine diyebiliri)

Standart kişilerde dakkada 100 sayfa okusun ben 2 standart kişiyle 1 dakkada 200 sayfa okudum daha ucuza en fazla okuyan adam pahalı onun gibi okuyanda yok onları bir araya getirelim

Bilgisayar dünyasındada böyle 64 gb bilgisayar alcan diyelim çok pahalı 16 gb nin 4 katı fiyatına eşit değil ondan daha pahalı işte ama şöyle diyosun 4 tane 16 gb li bilgisayar alayım bunlardan bir cluster oluşturayım diyon clustered olmanın bir diğer avantajıda bir makine gitse bile sistem devam eder clustered based demek bu

Harddiskler var bunlar veri tutuyorlar sadece ucuzlar çünkü sadece veriyi tutuyor ama işleytemiyor

Ama laptoplar öyle değil laptoplarda hem hardiskk var veriyi yutcak hemde ram cpu var veriyi işlicek

Sparkın en önemli özelliği inmemory process yapması pig hive diskten işliyo veriyi

İnmemory şöyle açıklayabiliriz veriyi diske gidip gelmeden yada sadece diskten 1 kere okuyarak memoryde işleyebilmesi spark ile inmemery data processing yapabiliyoruz

Select \* from tabloadı diye bi sorgu olsun önce veriyi diskten okur memoryde process eder ben tekrardan aynı sorguyu çalışıtırırsam memoryden direkt çıkıyı getirir

Pig böyle değil diskten okuyo diskten yazıyo map reduce işi

Development Apı:

Spark ortamında geliştirme yapabilmek için scala java python r dilleri ile gelişltirme yapabiliriz

Ben pyspark yapıyorum

Spark üzerinde bir çok kütüphane var

Örneğin spark sql : sql biliyorsan kolay öğrenirsin

Spark streaming var real time işler anında işlem yapıyom oluşan datalara

Ml modülü var

Sparkı kullanıcaksak direkt kendi bilgisyarımada kullanabilirim işte jupyter notebooka entegre ederim gerekli pysark pyspark.sql leri import edebilirim

Google cloud da dataproc u var sadece spark cluster kurabilirim yada on premise sistemde zeppelinde hiç bir şey kurmadan import edmeden her şey gömülü gelmiş çünkü

Pyspark diyerek pythona bağlanabilirim spark –shell diyerek scala %djbc(hive) diyerek hive da ki tabloya hql(hive query language) atabilirim

Spark big data clusterında çalışır kurulu geliyor ama sparkın çalışması için

1-Drıver

2-SparkContext

3-Yarn

4-Hdfs

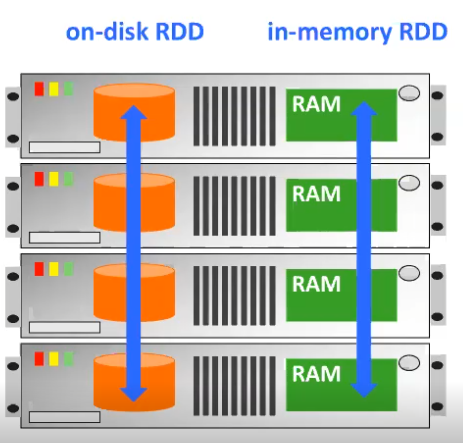
5-executors

Tüm spark komutları şu şekilde çalışır REPL read evaluate print loop

Spark taki kavramlar

Resilent Distributed Dataset(RDDS)(her türlü datayı işleyebilirim structured unstructered semi structured)

DataFrame (sql çıktısı structured data olcak adı mert soyadı Türkoğlu)



Diskteki veriler üzerinden çalışması momorydeki veirler üzerinden çalışması

Atıyom 2.diskten veriyi okuyamadım problem çıktı diğer dsikten veriyi okumaya devam edebilirim distrubuted kısmı bu oluyo

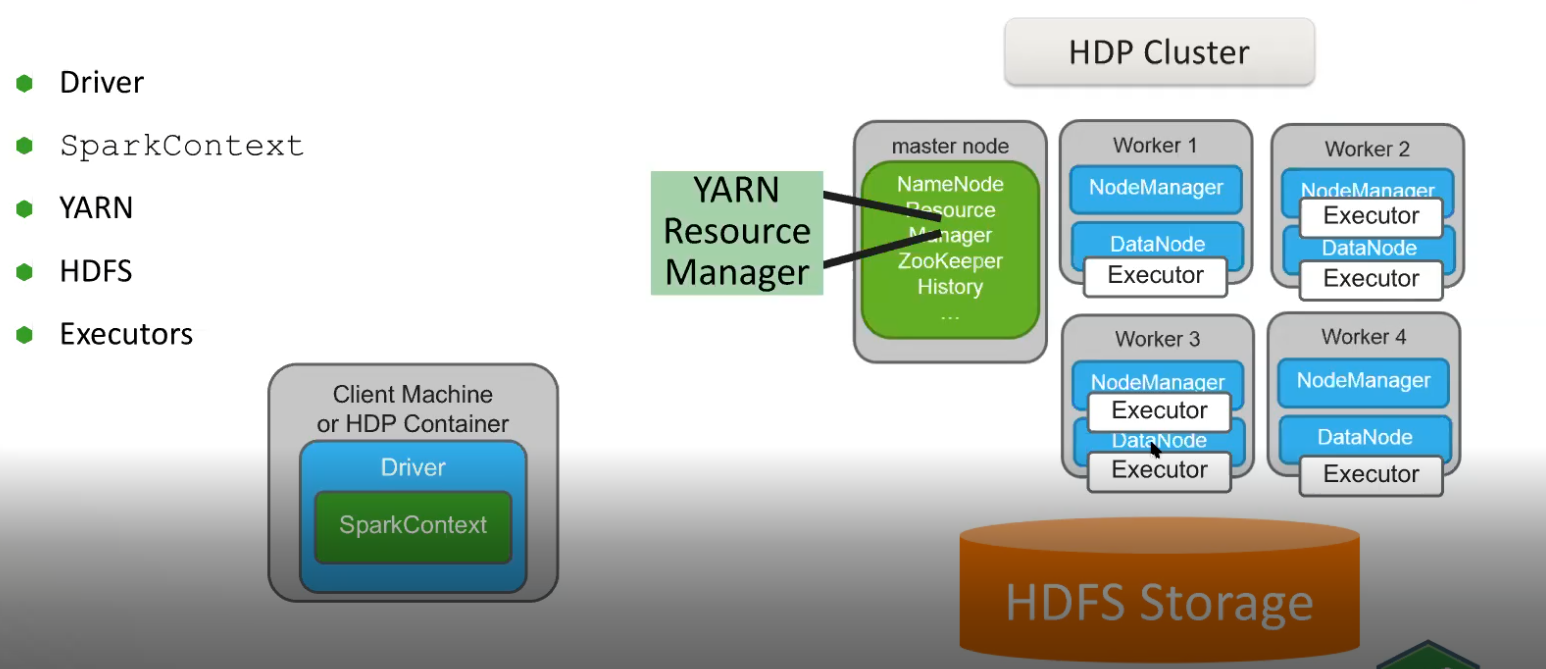
Sparkın rdd si veya dataframe in en büyük oluyo horizontal scability (yatay büyüme)

Herhangi bir nosql teknolojisi öğrenelim en büyük avantajı yatay büyüme

Herhangi bir rdbms i öğrenelim bunlar bunlar yatay büyümeye çok uzak

3000 tl ye 4 tane telefon alalım eğer 4 tane telefonu yan yana horizantal scability yapabilirsek ıphone 12 den faha hılı çalışır sadece daha iyi çalışmak değil bir ıphone 12 tek telefon gitti mi gitcek paralel processing olcak işin bir kısmını biri diğer kısmını biri bu şekilde

Spark nasıl çalışcak



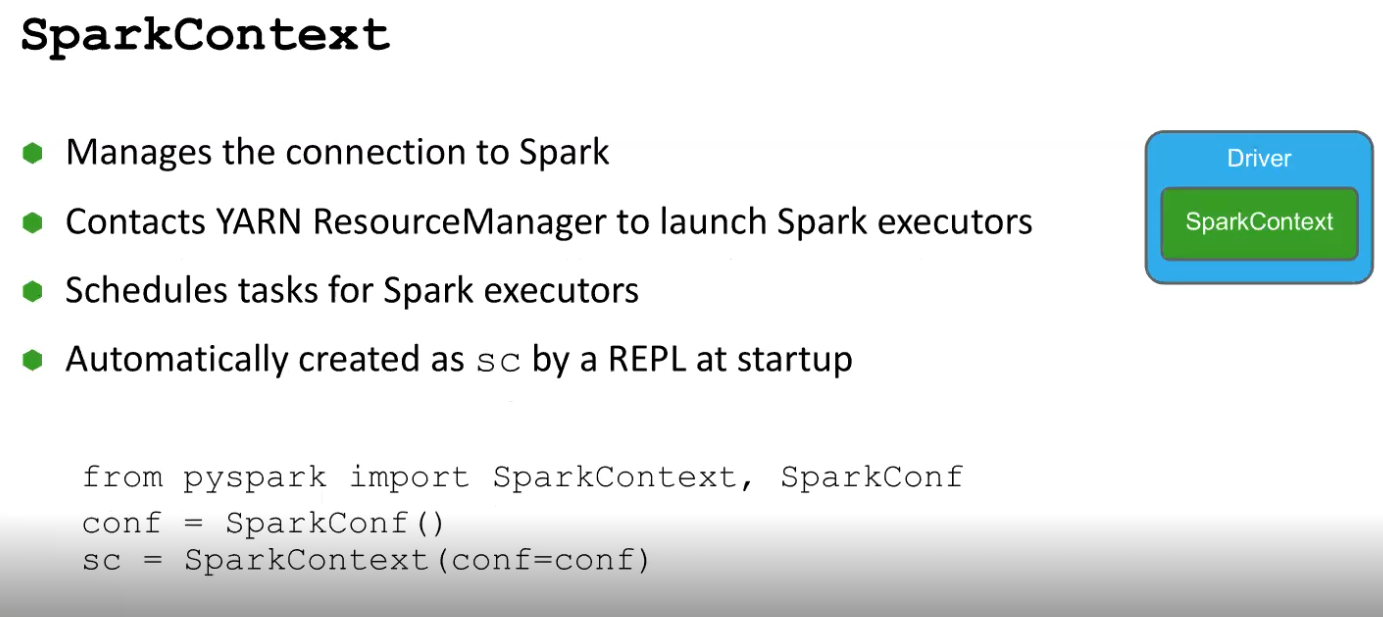
Eğer ben sparkı kendi makineme client makinesine kuruyorsam SparkContext ve driver gerekli

Eğer cloudera cluster veya hotronworks clusterda isem spark kurulu zaten zeppelin de bir şey yazmaya gerek yok(SparkContext SparkSession SparkConf) yarn hadoop clusterdaki servislerin birbirine entegre şekilde çalışmasını sağlıyo kaynak yönetimini sağlıyor dengeli çalışmam lazım biri 4 ile 5 te process yapcak ona göre kaynak ayırıyom

Cloudera Cluster veya hortwonworks clusterda isem data HDFS te tutuluyor big data cluster

Google cloud ise datam bucket de tutuluyor amazonda isem datam s3 de tutuluyor

Executorda işte spark ne kadar çalışcak executorlar da onu belirliyor işi paralel bir şekilde dağıtıyor



Sparkı kendi makinemde çalıştırmak için yukarıdaki ifadeyi yazcam

sc sparkın bütün konfigürasyonların bulunduğu alan

zeppelinde bunları yazmaya gerek yok

Spark executors da ne kadarla kaynakla çalışacaksın