T.C.
ESKİŞEHİR OSMANGAZİ UNIVERSİTESİ - BİLGİSAYAR MÜHENDİSLİĞİ FAKÜLTESİ
2015 - 2016
BÜYÜK VERİ - FİNAL RAPORU

Önbilgi

Stanford Üniversitesinin sayfasından sunulan Gowalla veriseti üzerinde çalışacağız. Bu verisetinde iki adet adet tablo bulunmaktadır.
İlk tablodaki kolonlar : [user] [check-in time] [latitude] [longitude] [location id]
İkinci tablodaki kolonlar ise : [user] [user]

1 - Top 100 checkin yapan kullanıcı sıralı listesi

Bu problemi çözmek için önce veriyi "\\t+" regex komutuyla yani tab boşluğuna göre sütun şeklinde bölüp sadece "userid" olan sütunu almam gerekiyordu. Bunu take() fonksiyonuyla yaparak sadece ilk sütunu aldım (take(1)). Bu fonksiyon içine kaç yazarsanız o kadar sütunu alıyor. 1 değerini girerek sadece birinci sütunu almasını sağladım. En baştaki değeri alıyor böylece. Eğer take(2) deseydik ilk iki sütunu alacaktı. Daha sanra map-reduce işlemini yaptırdım. Sparkda sortByValue adında bir fonksiyon olmadığı için burdaki kolonları yer değiştirip sortByKey fonksiyonunu kullandım. Bunun için reduce işleminden sonra swap işlemiyle key ve value değerlerinin yerlerini değiştirerek map fonksiyonunu çağırdım. Yani yeni key benim reduceden gelen value değerim oldu. Daha sonra sortByKey(false) diyerek büyükten küçüğe sıralamasını sağladım. Çünkü default olarak küçük büyüğe sıralama yapıyor bu fonksiyon. İlk yüzü alabilmek içinse take(100) komutunu kullandım ve böylece ilk yüz değere ulaşabildim.

2 - Top 100 checkin yapılan yer sıralı listesi

Bu problemi çözmek için önce veriyi "\\t+" regex komutuyla yani tab boşluğuna göre sütun şeklinde bölüp sadece "locationID" olan sütunu almam gerekiyordu. İlk soruda yaptığım gibi take() fonksiyonuyla bunu yapamadım çünkü take(4) dersem dört sütunu alacaktı. Bu yüzden "line.split(**"\\t+"**)(4)" yaparak , her satırda split sonrası dördüncü elemanı al dedim. Bu şekilde sadece "locationID" kolonunu almış oldum. Daha sonra üstteki sorudaki gibi map-reduce işlemini yaptım. Swap fonksiyonuyla key ve value değiştirerek, sortByKey(false) fonksiyonunu kullanarak büyükten küçüğe doğru sıraladım. Daha sonra yine take(100) ile en çok checkin yapılan ilk yüz yerin id bilgisini aldım.

3 – Bi kişinin checkin yaptığı yerde en çok checkin yapan kişilerin sıralaması

Bu problemi tam olarak anlayamadığım ve sizden mail olarak tam geri dönüş alamadığım için iki farklı şekilde çözdüm.

a) İlk önce tüm veriyi okutup burdan user ve location bilgisi alıp sonra diğer çıkardığım locationa göre sıra listesiyle join etmek istedim ama memory yetmediği için kapattı. Ben de random veri aldım ve bu veriden user ve location bilgilerini mapledim. Yani [LocationId,UserId] şeklinde bi verim oldu. Daha sonra ana veriyi(394mb) yeniden alarak burda [LocationId, UserId] ikilisi oluşturdum ve map-reduce işlemi yaptım. Bu şekilde bi yere en çok giden kullanıcıların listesi oluşturulmuş oldu. Swap işlemi uygulayarak yine bunları sortByKey(false) fonksiyonuyla sıklıklarına göre büyükten küçüğe sıraladım. Join işlemiyle birlikte bunları elimdeki bu iki veriyi birleştirdim. Collect() ile ArrayList haline getirdim ve ekrana yazdırdım. Böylece çıktı olarak [UserId, LocationID,CheckinYaparUserID,Count] şeklinde aldım.

b) Burda ise emailde ilk emailde bahsettiğiniz gibi bi kullanıcı id girip bunun gittiği yerlerde checkin yapan kullanıcıları sıraladım. Bunun için filter fonksiyonunu kullandım. A şıkkında yaptığım işlemlerden sonra filter ve contains methodları ile bir kullanıcının gittiği yerlerde checkin yapan kullanıcıları checkin sayılarına göre sıraladım.