Hive Partition ve Bucketing Uygulaması

Görev 1: Github repoda bulunan **u.data** ve **u.item** veri setlerini Hive'a tablo olarak yükleyiniz.

Adım 1: datasets klasörüne veri setlerini indiriniz ve inceleyiniz.

Adım 2: Beeline ile Hive'a bağlanıp movielens adında yeni bir veri tabanı oluşturunuz.

Adım 3: u.data verisine uygun **ratings** adında bir tablo oluşturunuz.

Adım 4: Localden **ratings** tablosuna veriyi yükleyiniz ve tabloyu inceleyiniz.

Adım 5: u.item verisine uygun **movies** adında bir tablo oluşturunuz.

Adım 6: Localden **movies** tablosuna veriyi yükleyiniz ve tabloyu inceleyiniz.

Görev 2: İş kullanıcıları bazı sorgulamalar yapmayı ve bu sorguların mümkün olduğu kadar kısa süre içinde sonuçlanmasını talep etmektedir. İş kullanıcılarının söz konusu ihtiyacını karşılamak üzere Hive üzerinde gerekli veri organizasyonunu yapınız.

Adım 1: Aylık olarak en popüler (en çok oylanan, en yüksek ortalama puanı alan) **filmler** belirlenmek istenmektedir. Buna göre tabloyu tasarlayıp (partition ve bucketing), oluşturunuz.

Adım 2: Dinamik Partitioning ayarlayınız.

Adım 3: İki tablonun verilerini oluşturduğunuz tabloya yükleyiniz.

Adım 4: Oluşturduğunuz tabloyu inceleyiniz.

- Gözlem sayısına bakınız.
- Partitionları listeleyiniz.
- Tablo özelliklerini listeleyiniz
- Review Year ve Review Month olarak kaç unique değer olduğunu inceleyiniz.

Görev 3: İstenen analizler için gerekli sorguları oluşturup yorumlayınız.

Sorgu 1: 1998 yılının Nisan ayında en çok puanlanan 20 filmini bulunuz.

Sorgu 2: 1998 yılının Nisan ayında oylanan filmlerden en yüksek ortalama puana sahip 20 filmi bulunuz.

miuul

miuul.com