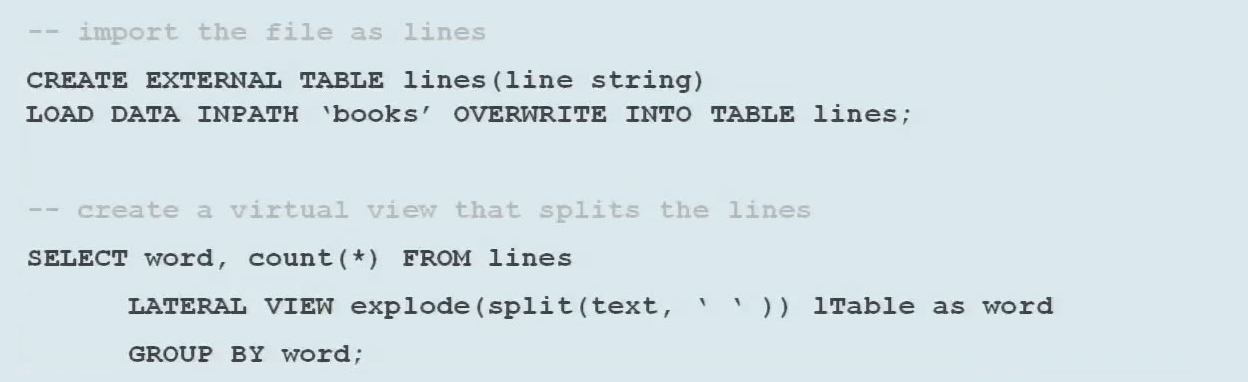
# Apache Hive Nedir?

Büyük verileri Hadoop üzerinde paralel olarak işlemek için MapReduce programı yazmamız gerekir.MapReduce programı Java,Apache Pig veya Apache Hive ile geliştirilebilir.



Bu bölümde Hadoop MapReduce geliştirme yöntemlerinden Apache Hive’ı inceleyeceğiz.Apache Hive , SQL biçimde sorgular geliştirmemizi sağlayan Apache Hadoop tabanlı veri ambarı(datawarehouse) kütüphanesidir.

## **Hangi durumlarda Apache Hive tercih edilmeli?**

Büyük veriler üzerinde basit analizler yapmak için Java MapReduce yada Apache Pig yazmak yerine hızlı bir şekilde Apache Hive tabanlı SQL sorguları geliştirilebilir. Mesela 1 milyar log üzerinde şöyle bir analiz yapalım

Sitemizi hangi ülkeden kaç tane tekil kullanıcı ziyaret etmiş

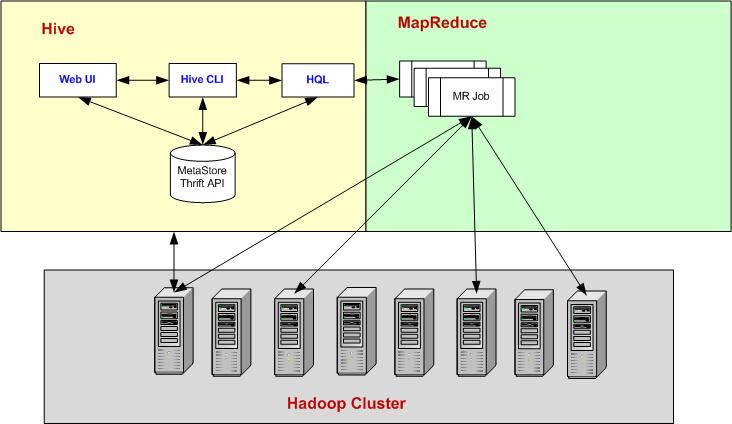
Bu durumda şöyle bir hive kodu işimizi görecektir

*select country,count(distinct user\_id) from data where log\_date >= '2015-04-17' and log\_date <= '2015-04-18' group by country;*

Ayrıca benzer sorguları sürekli big data developer ‘ a yaptırmak yerine iş birimine yaptırabiliriz.

## Hive Mimarisi

Web arayüzünden yada Hive client üzerinden çalıştırdığımız HQL(Hive Query Language) sorguları daha önce girilmiş şema bilgileri kullanılarak arka planda MapReduce job’larına çevrilir.Hive kurulduğu zaman beraberinde Thrift server ile gelir.Bu servisin amacı ise Hive üzerinde belirli portlara erişmeyi sağlamaktır

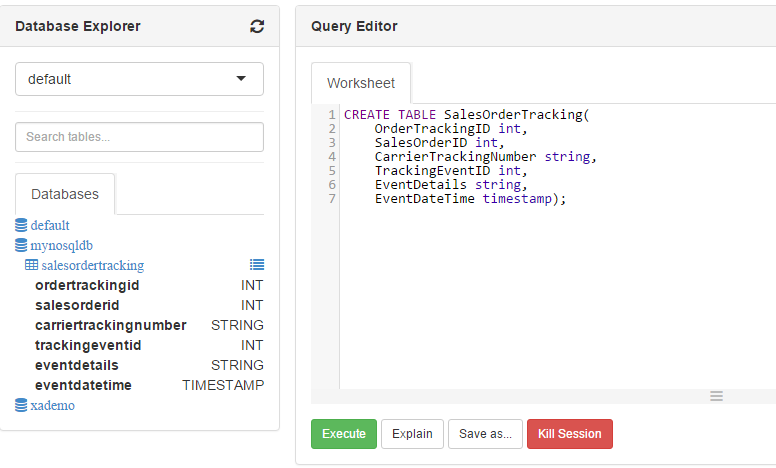


## Hive MetaStore Nedir?

Şema bilgilerinin ve bu şema altında yer alan tablolar,kolonlar,kolon tiplerinin saklandığı bölümdür



## Apache Hive ile Veritabanı ile Tablo Nasıl Oluşturulur?



### Hive Veritabanı Nasıl oluşturulur?

Hive ile oluşturduğumuz tablolar bir veritabanı altında görünür. Bu nedenle öncelikle bir veritabanı oluşturmamız gerekir. Veritabanı 2 şekilde oluşturulur

*hive> CREATE DATABASE [IF NOT EXISTS] my\_db;*

ya da

*hive> CREATE SCHEMA my\_db;*

test için

*hive> SHOW DATABASES;*

*default*

*my\_db*

Cloud bir sistem üzerinden rahatlıkla veritabanı oluşturabilirsiniz



### Hive Tablo Nasıl oluşturulur?

Hive ile tablo oluşturarak aslında HDFS üzerinde var olan bir veriyi bir şema bilgisi tanımlamış oluruz. Veriler yine hdfs sistemine eklenir yada silinir,fakat hive şema bilgisini saklayarak SQL ile kolayca sorgulama yapmamızı sağlar

HDFS üzerinde alttaki gibi bir veri kümemiz olsun

*1,US,5,https://www.google.com.tr/#q=apache+pig,apache,pig*

*2,US,25,https://www.google.com.tr/#q=apache+hive,apache,hive*

*3,TR,20,https://www.google.com.tr/#q=apache+hadoop,apache,hadoop*

*4,EN,10,https://www.google.com.tr/#q=apache+oozie,apache,oozie*

*5,PL,16,https://www.google.com.tr/#q=apache%20flume,apache,flume*

*6,US,24,https://www.google.com.tr/#q=apache+spark,,spark*

*7,US,36,https://www.google.com.tr/#q=apache+kafka,,kafka*

*8,EN,48,https://www.google.com.tr/#q=+storm,apache storm in hadoop ,storm*

Bu veri üzerinden şu şekilde tablo oluşturabiliriz

*CREATE EXTERNAL TABLE IF NOT EXISTS weblog (*

*id int,*

*country String,*

*duration int,*

*url String,*

*keyword1 String,*

*keyword2 String)*

*ROW FORMAT DELIMITED*

*FIELDS TERMINATED BY ','*

*LINES TERMINATED BY '\n'*

*STORED AS TEXTFILE;*

Bu script ile Hive tablosu oluşturabiliriz.Burada dikkat etmemiz gereken konu **EXTERNAL**anahtar kelimesini eklemezsek tabloyu sildiğimiz zaman HDFS altındaki loglarda silinir.Bu yüzden geçici(temporary) bir tablo kullanmıyorsak **EXTERNAL**anahtar kelimesini eklemeliyiz

!!!!! Tablo oluşturduktan sonra HDFS üzerindeki verilerle tablo arasında bir ilişki kurmamız gerekir

*LOAD DATA INPATH '/user/hue/dizin/log5.txt'*

*OVERWRITE INTO TABLE tabloAdi;*

Daha sonra işlemleri test etmek için basit SQL sorguları yazalım

*Select \* from tabloAdi*

### **Tablo Yapısındaki Değişiklikler(Alter)**

Hive içerisindeki tablo yapılarında değişiklik için alttaki script’leri kullanabiliriz

*ALTER TABLE weblog RENAME TO new\_weblog (tablo ismi değiştirme)*

*ALTER TABLE weblog ADD COLUMNS (access\_count int) (yeni kolon ekleme)*

*ALTER TABLE weblog DROP COLUMN id (kolon silme)*

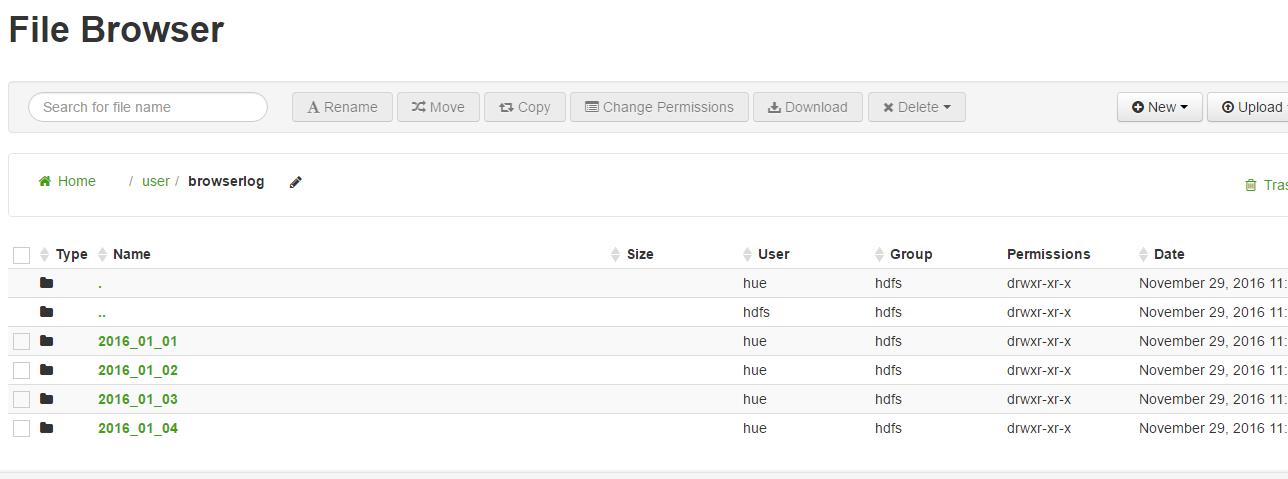
*ALTER TABLE weblog CHANGE id new\_id (kolon adı değiştirme)*

*ALTER TABLE weblog CHANGE id String (kolon tipi değiştirme)*

*ALTER TABLE weblog REPLACE COLUMNS(id int,url STRING,duration int); (burada belirtilmeyen kolonlar silinir)*

### **Hive Partition nedir?**

HDFS içerisinde veriler genellikle klasörler altında atılır.Örnek verirsek browser üzerinden logların toplandığı bir sistemde günlük klasörler vardır. 1 Ocak tarihinde gelen log’lar 2016\_01\_01 klasörünün altına, 2 Ocak tarihinden gelen loglar 2016\_01\_02 altına atılabilir.Bunun en önemli nedeni ise sorgulama performansını artırmaktır



Bu noktada alttaki gibi bir hive sorgusu yapmak isteyelim

*select \* from weblog where log\_date = '2016\_01\_01'*

Böyle bir sorgu yapabilmek için log\_date olarak isimlendirilmiş bir partition(bölümlendirme) eklememiz gerekmektedir

Partition şu şekilde eklenir

*ALTER TABLE weblog ADD PARTITION (log\_date='2016\_01\_01') location '/user/browserlog/2016\_01\_01';*