Data Engineering Masterclass

Gün 1

Modül-1: Data Collection

İçerik: API'lar, loglama, sensory data, web scraping.

Anahtar sözcükler: JSON, XML, HTTP, HTML, DOM, grep, RegExp.

Araçlar: Postman, log4j, python-logging, BeautifulSoup, Jsoup, Selenium

Konu anlatımına geçelim.

Hepimizin bildiği gibi veri bizim işimiz ve eğer veri hazır şekilde sağlanmıyorsa veriyi toplamamız gerekir. Eğer veri sağlanıyorsa; o alınan veriyi düzenleme, ekleme, silme ya da verinin formatını değiştirmek gerekebilir. Örneğin sensörden bir veri elde edildiğinde o elde edilen verinin formatı kullanılacak proje için uygun olmayabilir. Bu gibi durumlarda veriyi düzenlemek gerekir. Veri, uğraşma işi olduğunu buradan anlayabiliriz.

Modül 1- Data Collection kısmında data nasıl toplanır bu kısım ile ilgileneceğiz.

Data Toplama Nasıl Olur?

Information retrieval(Bilgi Alma), web-scraping, alınan API dataları ya da başkalarına sağlanan API dataları data toplamaya birer örnektir.

API - Sensory Data - Web Scraping

API

İki sistemin arasında nasıl konuşacağını belirlediği bir yöntemdir. Belirli istekler var ve bu tipte akış ve veri sunar. Belirli bir kuralları olan çerveveli bir veri sunacağını söyler. Bu veriyi belirli bir rate(limit) içerisinde sunar. Karşı tarafta kurallar vardır. Karşı ne kadar bilgi sunuyorsa onunla yetinilir ve kurallı bir yapıdır.

Belirli sorgulara karşı belirli bir data parçası geçmektedir, tüm sistemin veri akışını sağlamak için değildir.

API, kontrollü, yavaş ve kurallıdır.

API ile sağlanan veri formatları .xml veya json olabilir. JSON oldukça popüler, şu an genel akım json üzerinden çalışıyor. .xml ise yapısal olarak değişmez.

Sensory Data

Gerçek zamanlı ya da sensörlere bağlı şekilde bir veri alıp işlem yapmak gerektiğinde sensör kullanılır. Yani dış dünyadaki bazı birimleri değişimini (yağmur, sıcaklık, nem vb.) bulmak için ufak bilgisayarlara sensör denilebilir, genellikle amaca yönelik sadece görevini yapan pil ömrü yüksek olan mini cihazlardır. Dışarıdan bilgiyi alma özelliğine sahiptir. Örneğin sıcaklık ile ilgili çalışacaksam sıcaklık ölçen sensör ile çalışılmalıdır. Pil ömrü uzun olmalıdır.

Web Scraping

Her yerden her insanın eriştiği, büyük verilerin olduğu yerden veri toplama işlemidir. Web sitelerinden bilgi çıkartmanın bilgisayar programı tekniğidir. Veri genellikle dağınık

biçimde web'de bulunur. Web scraping'in dezavantajı ise belirli bir protokolun olmamasıdır. Bunun için birçok Kütüphane, Framework var;

- LXML
- Selenium
- Requests
- Mechanize
- Beautiful Soup 4
- Scrapy | A Fast and Powerful Scraping and Web Crawling Framework

Veri çekme işlemini yaptık. Peki ya sonra web sitesi değişirse ne olacak?

Bu sorunun cevabı olarak kullanılan iki yöntem vardır. Bunlardan birincisi:

Değişikliklerde **call** denilen bir sistem kullanılabilir. İki taraf için yüklü bir sistem olduğundan dolayı istenilen bir yöntem değildir. Bir websitesi için yazılan scraping scriptleri her gün değişmez bu yüzden büyük bir problem yaratmayacaktır.

İkincisi ise:

Subscribe yönteminde webhook gibi yöntemler kullanılabilir fakat karşı tarafın da sizi tanıyor olması gerekmektedir.

Loglama Nedir?

Loglama, data toplama yöntemi değildir bir kavramdır.Kullanımı açısından bize kolaylık sağlar. Kelime anlamına gelecek olursak **Loglama,** log kayıtları aracılığı ile dijital hareketlerin saklanması işlemine denir. Log tutma anlamına da gelmektedir. Gelen sorunları tespit etmek ve çözmek için sistemin geçmiş kayıtlarını tutarak yani geçmişteki izlerini takip ederek kayıt tutma işlemidir.

Kesintisiz çalışırken sorun olabilecek yerlere ya da üzerinde fikir yürütülebilecek bilgiler loglanır. Loglama önemlidir.

Her satırın loglanmasının anlamı yoktur. Bu yöntem oldukça dikkat dağıtıcı ve kullanışsız olabilir ve logları takip edemez hale geliriz. Buna çözüm olarak loglamanın seviyeleri vardır.

Loglamanın seviyeleri(önceliklendirilmesi):

- Low level
- Critical level
- Warning level
- İnfo level
- Debug level gibi seviyeler vardır.

Bunlardan kısaca bahsedecek olursak:

- Debug Level: Sorunların teşhisi için debugging gibi düşünebiliriz. Ayrıntılı bilgilere ihtiyacımız vardır.
- Info Level: Beklenen çıktılarımızdır. En çok kullanılan seviyedir. Çünkü akış görülmek istenir.
- Warning Level: Yazılım hala çalışıyordur ama uyarı çanları çalmaktadır.

- Error Level: Yazılım ciddi bir sorunla karşılaştı ve görevini yerine getiremedi.
- Critical Level: Programın işlevini yerine getiremeyecek bir sorunla karşılaşmasıdır.

Loglama yaparken mutlaka tutulması gereken bilgiler:

- Timestamp: Tarih, zaman damgası bulunması ve ne zaman olduğuna dair bilgi vermesi açısından önemlidir.
- Logging Level: Hatanın derecesi, nedeni veya olayın ne olduğuna dair seviyenin belirtilmesi gerekmektedir.
- API Bilgisi: Sensor ID, hangi fonksiyon, genel bilgiler içermelidir.
- Logun içeriği: value, logun içeriği json, plain text, xml olabilir.

Keywords:

POSTMAN: API'ları paylaşmak, test etmek, dokümante etmek, monitör etmek

için kullanılır. En öne çıkan özelliği tüm bunlar için çok kullanışlı bir arayüz

sunmasıdır.

Log4j: Java uygulamalarında kullanılacak loglama kütüphanesidir.

python-logging: Log4j'in python versiyonu

BeautifulSoup: BeautifulSoup, HTML veya XML dosyalarını işlemek için oluşturulmuş bir kütüphanedir. HTML'den veri çıkarmak için kullanılabilen ayrıştırılmış sayfalar için bir ayrıştırma ağacı oluşturur ve bu, web'den veri toplama için yararlıdır.

Jsoup: HTML belgelerinde saklanan verileri ayrıştırmak, ayıklamak ve değiştirmek için tasarlanmış açık kaynaklı bir Java kütüphanesidir.

Selenium: Selenium, bilgisayarınıza yükleyeceğiniz bir driver yardımı ile

ekrana chrome, firefox gibi bir tarayıcı açarak, gerçek bir insan gibi istediğiniz

tüm işlemleri programlama dili yardımıyla çalıştırmanızı sağlayan bir araçtır.

Birinci gün bitti. Keyifli geçen bir dersti umarım sizde okurken yeni bilgiler edinirsiniz 😊