Day 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

Data Toplama

Information retrieval, web-scraping, alınan API dataları bunlara örnektir.

API

İki sistemin arasında nasıl konusacağını belirleyen bir yapıdır. Belirli spesifik tipte akış ve veri sunar. Belirli bir rate içerisinde olmaktadır. Belirli sorgulara karşı belirli bir data parçası geçmektedir, tüm sistemin veri akışını sağlamak için değildir.

> 💡 Farklı formatlarda dönüşü olabilir. **.xml** veya **json** olabilir. JSON oldukça popüler, şu an genel akım json üzerinden çalışıyor.

Sensörlere ufak bilgisayarlar denilebilir, genellikle amaca yönelik sadece görevini yapan pil ömrü yüksek olan mini cihazlardır. Üzerindeki datayı merkeze alınarak kullanılır. *Aslında **edgedeki** cihazlardan bahsediliyor.*

Web Scraping

Görece API'ye göre verinin elde edilmesi daha zordur. Veri genellikle dağınık biçimde web'de bulunur, veri toplama prosesi kullanıcının bu rotaları tanımlayarak gerçekleştirmesiyle sağlanır. Dezavantajı ise belirli bir protokolun olmaması, ***challenging but fun!***

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

Loglama

Genellikle problemler belirli bir zaman diliminden itibaren veya geçmişten gelir. Anlık sorunları ve sistemin ne yaptığını en ince detayın kadar görmek için **loglama** kullanılır.

🔥 **Pro Tip:** Her satırın loglanmasının anlamı yoktur. Bu yöntem oldukça dikkat dağıtıcı ve kullanışsız olabilir. Bunun için loglamanın seviyeleri vardır.

Loglamanın seviyeleri

Low level, critical, warning, info gibi seviyeler vardır.

- **Debug:** Sorunların teşhisi için debugging gibi düşünebiliriz. Ayrıntılı bilgilere ihtiyacımız vardır.
- **Info:** Beklenen çıktılarımızdır.
- **Warning:** Yazılım hala çalışıyordur ama uyarı çanları çalmaktadır.
- **Error:** Yazılım ciddi bir sorunla karşılaştı ve görevini yerine getiremedi.
- **Critical**: Programın işlevini yerine getiremeyecek bir sorunla karşılaşmasıdır.

🔥 **Pro Tip:** *grep* komutuyla dosyaların içerisindeki belirli kelimeyi arar. *(e.g. 2 şubat tarihinde bir problem oldu ve onun bulunması için kullanılabilir.)*

Loglama yaparken:

- **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.

Logları tek büyük bir dosyada depolamak giderek büyüyen bir dosya olacağından dolayı mantıklı değildir. Databasede tablo olarak saklanabilir böyle bir durumda select süresi giderek uzayacaktır. Sistemi optimize edebilmek için aktif olanlar veya olmayanlar yahut arşivlenmiş gibi farklı parçalara bölmek zaman açısından yararlı olacaktır.

Bu gibi büyük çaplı verilerde hızlı işlemler yapabilmek için **Hadoop** gibi özelleşmiş sistemler bulunmaktadır. Temelde depolanan ve güncel olarak kullanılan iki parçaya bölmek mantıklı olacaktır.

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.
- **Jsoup:** BS4'un java versiyonu
- **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.

End of the first day!