thinktech
STM Teknolojik Düşünce Merkezi

El-Dorado

"Dünyadaki verilerin toplamının 2012 yılı itibariyle tahmini 2.8 zettabayt olduğu (1 zettabayt: 1 trilyon gigabayt) olduğu tahmin edilmektedir[1]." Bu ifade ilk okunduğunda çoğu henüz işlenmemiş olan bu veri hazinesini bildiğimiz veri tabanlarında tablo yapısında (daha teknik adı ile ilişkisel yapıda) tutulmuş olarak tasavvur edebiliriz. Oysa bu veri hazinesi sanıldığı kadar kolay bir yerde değil. Bu mistik hazineyi bir zamanların El-Dorado'suna benzetebiliriz.

El-Dorado 16'ncı yüzyılda İspanyol denizcilerinin dillerinde dolaşan bir efsanedir. Efsane birçok denizcinin altın tutkusuyla aklını başından alarak, El-Dorado adlı yeri keşfetme umuduyla, And dağları civarındaki bölgeleri sürekli aramasına neden olmuştur. Bu efsane, insanların servet umuduyla sonundaki gerçekliği düşünmeden kapıldıkları heyecanı ve "sorgulamama" durumunun kitlesel bir norma dönüştüğü durumları özetleyen önemli bir metafora dönüşmüştür. Aslında, o zamanlar El-Dorado adında bir altın imparatorluğu hiç olmamıştır ancak daha sonra El-Dorado adı Kaliforniya'da bir kasabaya ve Cadillac'ların bir serisine verilmiştir.

Büyük veri için El-Dorado ise organizasyonların verileri içerisinde bulacaklarına inandıkları hazineleri ifade eder. And Dağları civarında El-Dorado mevcut değilse de büyük veri için El-Dorado umudu aslında zayıf değil: "2008 yılında Facebook'da depolanan toplam veri miktarı insanlığın o güne kadar ürettiği veri miktarından daha fazlaydı. İnsanlık tarihinde yazılan kitaplar, yazılı metinler ve kayıtlı tüm verilerden daha fazlasının artık sürekli olarak üretildiği günümüzde oluşan bu verileri anlamlandırmak, analiz etmek ve karar vericilerin hizmetine sunmak artık "veri bilim" olarak adlandırılan bir alanın ortaya çıkmasına zemin hazırlamıştır[2]."

İçerisinde El-Dorado olsun olmasın, günümüzde işlenmemiş verilerin çoğunun, işlenmesi kolay olmayan metin, video, resim ve hatta el yazısı formatında olduğunu söyleyebiliriz. Tablo formatında olmayan bu verilere yapılandırılmamış veri (unstructured data) deniyor.

Yüzde 25 ya da Samanlıkta Çuvaldız Aramak

2.8 zettabayt civarındaki devasa veri setini neyse ki tek bir veri bilimci işlemeyecek. Toplamda bu miktarı teşkil eden veri setlerinin veri bilimci veya organizasyon başına düşen miktarını işlemenin zorluğu sadece astronomik veri miktarlarından değil bu verilerin yapılandırılmamış olmasından da kaynaklanıyor. Bu soyut teknik ifadenin alan dışından da tam olarak anlaşılması faydalı olacaktır; nitekim artık veri hepimizin hayatında.

1 https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
2 https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

Durumu kabaca şöyle açıklayabiliriz: Yapılandırılmış veri (structured data) için en iyi örnek klasik excel veya access dosyalarında tutulan tablo şeklindeki verilerdir. Bu veriler toplanmadan önce yapı (schema) belirlidir. Örneğin basit bir telefon defterinde ad, soyad ve telefon verisi vardır. Bu yapı belli olduğundan veri tabanında aynı kolon altında aynı tipte toplanmış veriler görülür. Aynı tipteki bu veriler sayısal ise onları kolayca toplayabilir, metinsel ise kategorize edebilirsiniz. Excel'de formüllerle yapılan işlemler veritabanlarında SQL sorgularıyla yapılır ve bu çok pratiktir.

Yapılandırılmamış verilerde ise genel olarak bir şema yoktur. Bu durumu her satırında ad, soyad, telefon kolonlarının dağınık bir şekilde yer aldığı, kimi satırlarda e-posta, doğum günü gibi bilgiler yer alırken kimi satırlarda sadece ad ve soyad bilgisinin yer aldığı bir excel'e benzetebiliriz. Özünde bu bile iyi bir senaryodur. Telefon defterinde yer alan ad, soyad ve telefon bilgilerinin metin halinde yazılmış adli soruşturmalar arşivi içinden derlenmesi gerektiğinde de bu metinler için bu kez yapılandırılmamış veri sıfatı söz konusu olmaktadır.

Yapılandırılmamış veri olarak yukarıda anılan metinsel veriler söz konusu olduğunda bile geleneksel veri tabanı uzmanlarının kullandığı araç ve yetenek setleri hatta paradigmaları tamamen devre dışı kalabilmektedir. Doğrudan canlı bir örnek olarak Google Cloud'unda[3] kullanılan farklı veri setlerinden Expando modeli örnek gösterilebilir[4]. Expando anlattığımız dağınık excel dosyasına benzetilebilecek bir şemasız veri yapısıdır. Expando modelindeki esnek duruma benzer olarak NoSQL veri tabanları ekosistemindeki neredeyse her bir teknoloji farklı bir paradigma sunmaktadır. Bu paradigma ise klasik yapılara alışkın uzmanların yetenek setlerini en azından gelecek için erozyona uğratmaktadır.

Peki bunca zahmet niye? 2.8 zettabaytlık küçük! küresel veri setindeki verilerin sadece yüzde 25'lik değerli kısmı için! Söz konusu verilerin geri kalan yüzde 75'lik kısmının organizasyonel olarak çok değerli olmadığı ifade edilmektedir[5].

Bu durum ise samanlıkta iğne kadar değilse de daha büyük şeylerin arandığına işarettir. El-Dorado hikâyesi kadar dramatik olmasa da elimizdeki geri kalan işlenmemiş verilerin birer kayıttan ibaret olduğunu tekrar hatırlamakta fayda var.

Yottabayt'a Varmadan, Büyük Veri Biliminin Birimleri

Disket "çağında" kilobayt ve megabayt güncel birimler iken ve gigabayt bir efsane iken, bugün gigabayt ve hatta terrabayt güncel birim haline geldi. Büyük veriden söz eden yazılardaki efsanevi birimler aşağıdaki gibidir:

1 Petabayt = 1024 Terabayt
1 Eksabayt = 1024 Petabayt
1 Zettabayt= 1024 Eksabayt

Bu birimlerin de üstünde Yottabayt adlı bir birim daha var. 1 Yottabayt 1024 Zettabayt ediyor ancak henüz yottabayt o kadar yaygın değil.

3 https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
4 https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
5 https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf