BÜYÜK VERİ: ÖNEMİ, YAPISI VE GÜNÜMÜZDEKİ DURUM[1]

BIG DATA: ITS IMPORTANCE, STRUCTURE AND CURRENT STATUS

Korcan DOĞAN
Uzm., Ankara Üniversitesi, Dil ve Tarih-Coğrafya Fakültesi, Bilgi ve Belge Yönetimi Bölümü, dogank@ankara.edu.tr

Sacit ARSLANTEKİN
Prof. Dr., Ankara Üniversitesi, Dil ve Tarih-Coğrafya Fakültesi, Bilgi ve Belge Yönetimi Bölümü, arslantekin@ankara.edu.tr

Makale Bilgisi
Gönderildiği tarih: 29 Nisan 2016
Kabul edildiği tarih:14 Haziran 2016
Yayınlanma tarihi: 23 Haziran 2016

Article Info
Date submitted: 29 April 2016
Date accepted: 14 June 2016
Date published: 23 June 2016

Anahtar sözcükler
Büyük veri, Veri, Veri madenciliği, Metin madenciliği, Verinin değeri, Veri işleme teknikleri

Keywords
Big data, Data, Data mining, Text mining, Value of data, Data processing techniques

DOI: 10.1501/Dtcfder_0000001461

Öz

Büyük veri, günümüzde son derece popüler bir kavram haline gelmiştir ve yeni bir devrin başlangıcı olarak yorumlanmaktadır. Büyük veri ile birlikte dünyada büyük bir dönüşüm gerçekleşirken, kurum ve kuruluşların veriye olan bakışı ve veriden sağladığı fayda farklı bir noktaya gelmiştir. Kurum ve kuruluşlarında bu dönüşümün dışında kalmaları mümkün görünmemektedir. Büyük veri, yeni ve popüler bir kavram olmasına rağmen, yerli literatürde kuramsal nitelikte çok fazla araştırmayla karşılaşılmadığı ortadadır. Kuramsal yönü işleyen az yayın olmakla birlikte, teknoloji ile ilgili pek çok kurum ve kuruluş ise bu konuya son derece önem vermekte ve çok büyük yatırımlar yapmaktadırlar. Çalışmada büyük veri kavramsal olarak ele alınmış, pek çok kavramla olan ilişkisi, büyük veri teknolojileri ve büyük veri işlenirken kullanılan yöntemler aktarılmış, büyük veri ile ilgili farkındalık yaratan kuruluşlarla ve dünyada büyük verinin kullanım alanları ile ilgili farklı örnekler verilmiştir.

Abstract

Nowadays big data has become a popular concept and it is interpreted as beginning of a new era. While a huge transformation occurs with the creation of big data concept, institutions and organizations' point of view and benefits gained from the data have changed and come to a different point. It seems that it is impossible for institutions and organizations staying out of the said transformation. Although big data is a new and popular concept, there are no so many studies having corporate nature in local literature. Despite a few publications having corporate base, numerous institutions and organizations operating in technology field place a significant importance to this concept and make big investments. In the study big data has been discussed conceptually and its relation with other concepts, big data technologies and methods used for processing of big data have been explained and different examples regarding institutions creating awareness for big data and usage area of big data around the world have been given.

Giriş

Gelişen bilgi ve iletişim teknolojilerinin kapsamında kabul edilen internet teknolojileri; web sayfaları, bloglar, sosyal medya uygulamaları, sensörler ve daha pek çok veri toplayan cihaz ve uygulamalar sayesinde her an bilimsel olan veya olmayan verileri toplar hale gelmiştir. Toplanan bu veriler, pazarlama, halkla ilişkiler, bankacılık, güvenlik vb. pek çok alanın yanında araştırmacıların yaptıkları araştırmalarda da kullanılabilir nitelik taşıyabilmektedir. Nitekim bu veri yığının değerinin anlaşılması sonucunda, bu veriyi toplama, işleme, kullanıcılara hazır hale getirme, erişime sunma, saklama, analiz etme gibi aşamalarda pek çok farklı teknikler de kullanılabilmektedir.

Bu verilerin günümüzde hız, çeşitlilik, kapasite (hacim) açısından büyük artış göstermesi ve bu artışa teknolojinin de destek vererek, yeni çözümler üretmesi ile birlikte "Büyük Veri" kavramı ortaya çıkmıştır.

[1] Yüksek lisans tezinden üretilmiştir.

Yaşanan gelişmelerle birlikte büyük veri kavramına hazırlıksız yakalanan kurum/kuruluşlar ve bireyler için karşılaştıkları bu durum; gerek bünyelerinde bulundurmaları bakımından, gerekse işlem ve hizmetlerinde kullanmaları bakımından son derece önemli hale gelmiştir. Bu veriyi kullanabilenler, diğerlerine göre avantajlı hale gelmekte, iş yapılarını geliştirmekte, ar-ge ve uygulama faaliyetlerini daha rasyonel gerçekleştirebilmektedir.

Konuyla ilgili oluşturulan yeni teknikler sayesinde âtıl durumda bulunan verilerin ilişkilendirilerek katma değer yaratan enformasyon haline getirilebilmesi sağlanmıştır.

Veri nedir?

Çalışmamızın dayandığı kavram "veri"dir. Bu nedenle burada verinin bazı tanımları üzerinde durmak gerekir.

Şeker (22), veriyi, *"Tanım itibariyle, herhangi bir işleme tabi tutulmadan, gözlem veya ölçüm yöntemleri ile ortamdan elde edilen her türlü değerdir"* şeklinde tanımlamaktadır.

Bir başka kaynağa göre veri, veri tabanında bulunan enformasyon için kullanılan genel bir terimdir (Prytherch 195).

Yılmaz (98) ise veriyi, *"tek başına anlam ifade etmeyen veya kullanılamayan, bununla birlikte enformasyona ve bilgiye temel oluşturan ilişkilendirilmeye, gruplandırılmaya, yorumlanmaya, anlamlandırılmaya ve analiz edilmeye gereksinim duyulan ham bilgi"* şeklinde tanımlamıştır.

Veri, araştırmalardan, gözlemlerden, internetten, sosyal medyadan, sensörlerden vb. çok farklı ortamlardan elde edilen genel bir terimi ifade etmektedir.

Veriler aşağıda verildiği gibi gruplara ayrılarak nitelendirilebilir (Jeffery 6):

- Yapılandırılmış, Yapılandırılmamış, Yarı yapılandırılmış
- Statik, dinamik, akan
- Güvenli / açık, özel / halka açık
- Ücretli / ücretsiz
- Açık hükümet verisi
- Açık veri
- Büyük veri

Verinin bu kadar değişik şekillerde gruplandırılması ve eskiden bu grupların çok azının yönetilebiliyor olması, büyük veri konusunun temelini oluşturmaktadır.

Veri Tabanı ve Veri Tabanı Yönetim Sistemi

Yukarıda belirtilenler doğrultusunda verilerin derlenmesi, gerektiğinde ilişkilendirilip bir enformasyon haline dönüştürülebilmesi için belirli bir düzen ve sistematik doğrultusunda kayıt altına alınmaları gerekir. Verilerin elde edilerek kullanılabilir diye varsayılanları, veri tabanlarında kaydedilip işlenerek erişilebilirken, günümüzde özellikle büyük veri kavramıyla birlikte veri ambarları da oldukça önem kazanmıştır.

Nitekim Uysal (31), veri tabanının tanımını, *"birbiri ile ilişkili veriler topluluğudur; veri tabanı sadece veriler yığınını değil, bunlar arasındaki ilişkileri de saklar"* şeklinde yapmıştır.

Veri tabanı, belirli bir konuda dizgeli biçimde düzenlenmiş ve bilgisayar ortamında korunan verileri ifade etmektedir (Türkçe Bilim Terimleri Sözlüğü 1216).

Veri tabanı, belirli bir amaç için ya da son kullanıcıların belirli bir kümesi için verilerin bilgi alanları ile organize edilerek gruplanmasını ve verinin saklama, gruplama, erişime sunulması ve raporlama gibi manipülasyonlarına olanak veren araçları sağlar. Veri tabanı, bibliyografik veri veya sayısal, istatistiksel veri içerebilir. Bu bilgiler doğrultusunda veri tabanı, *"iş, veri veya diğer materyal dermesinin : (a) sistematik veya metodolojik yollarla düzenlenmesi ve (b) elektronik ve diğer yollarla bireysel olarak erişilebilmesi"* (Prytherch 196) olarak da tanımlanmaktadır. Veri tabanını oluşturmak, yönetmek, verileri saklamak, gerektiğinde sorgulama yapmak ve daha pek çok başka işlemi yapabilmek için bir yazılıma gereksinim duyulur. Bu yazılım "veri tabanı yönetim sistemi" olarak karşımıza çıkar.

Veri tabanı yönetim sistemi, veri tabanın oluşturulması ve idame ettirilmesi için gerekli olan yazılım paketleridir (Prytherch 196).

Veri tabanı yönetim sistemi, verilerin güncelleştirilmesini, saklanmasını, erişimini düzenleyen ve çok sayıda kullanıcıya, birden çok bilgisayar sistemine hizmet verebilen veri tabanlarının kurulmasını ve işletilmesini sağlayan sistem olarak tanımlanabilir (Türkçe Bilim Terimleri Sözlüğü 1216).

Bir başka yazıda ise *"veri tabanı yönetim sistemi terimi, tam olarak bir veri tabanını ve bu veri tabanı üzerindeki yönetimle ilgili bütün yazılımları kapsamaktadır"* (Şeker 23) şeklinde tanımlanmaktadır.

Bazı durumlarda iş süreçlerimizde kullanılacak veri tabanlarında kayıt altına alınmış veriler çeşitli karar alma mekanizmalarında yeterli olmayabilir. Böylesi durumlarda dışarıdan da büyük miktarda veri girişine olanak veren sistemlere gereksinim duyulabilir. İşte tam burada veri ambarları devreye girmektedir.

Veri Ambarı

Kurumların stratejik seviyede karar destek sistemlerinin yaratılması için özel bir alt yapının tasarlanması gerekmektedir. Veri ambarı bu tür gereksinimlere yanıt vermek üzere hazırlanan bir ortam olarak tanımlanabilir ve karar destek sistemlerinin teknik alt yapısını oluşturmaktadır. Veri ambarı, birbirleriyle bütünleşik olmayan uygulamaların bütünleştirilmesi açısından bir olanak sağlamaktadır. Veri ambarı, kurum içinde ya da kurum dışında üretilen verilerin özellikle anlık sorgular için hazır bulundurulmasını sağlamaktadır. (Ören, Üney ve Çölkesen 879,883)

Çalışmaları ile ilgili hacim olarak çok büyük miktarlarda verileri bulunan fakat tek bir noktadan yönetilmesini sağlayacak yeterli araçları bulunmayan kurum ve kuruluşlar, planlama stratejilerinde bu nedenle problemler yaşayabilmektedir. Satış, envanter, maliyet profilleri, satış ekibi geri bildirimleri, kullanıcı memnuniyeti vb. dışarıdan elde edilen verilerle, kurum ve kuruluş içi bilgiler entegre edilebilir ve veri ambarında birleştirilebilir. Bütün veriler tek bir ana kaynağa yerleştirilir ve bunlar şirkete bütün faaliyetlerinde tutarlı ve güvenilir bir görüş sağlanmasında kullanılabilir (Prytherch 196).

Daha önce belirtilen veri türlerinden yapılandırılmış ve yapılandırılmamış veri ve bazen de kısmî-yapılandırılmış veri konusu, büyük veri konusunda önemli bir yer tutmaktadır. Hangi verinin büyük veri olup olmadığı konusunda da bu kavramlar önemli rol oynamaktadır.

Yapılandırılmış veri tipi, belirli kurallar ve sistemler doğrultusunda depolandıkları için kolay erişilebilir, düzenlenebilir, kategorize edilebilir vb. yapıdadır.

Bu veri tipi, tablolar üzerinde satır ve sütunlar halinde düzenlenmiş verilerdir. Bu tür veriler kağıt üzerinde olabileceği gibi, bilgisayarlarda istatistik paket programlarının satır ve sütunlarından oluşan matris şeklindeki yapılara da kayıt edilmiş olabilir (Oğuzlar 1).

Veri tabanı yönetim sistemlerinde yapısı gereği belirli bir düzen dâhilinde depolanan veriler, yapılandırılmış verilerdir. Yaygın olarak yapılandırılmış verilerden bahsedilirken, veri tabanı yönetim sistemlerinde depolanan veriler anlaşılmaktadır.

Yapılandırılmamış veriler ise mektup, doküman, kitap gibi kağıt üzerinde bulunan veya e-mail, web sayfaları gibi elektronik ortam metinlerinden; fotoğraf gibi durağan ya da film gibi hareketli görüntülerden ve/veya seslerden oluşur (Oğuzlar 1-2).

Yukarıdaki açıklamalara göre yapılandırılmış veri ile yapılandırılmamış veri arasındaki temel farkın, yapılandırılmış verinin üzerinde her türlü işlem ve sorgulamanın yapılabileceği, ilişkilerinin kolaylıkla kurulabileceği bir veri tabanı yönetim sistemi üzerinde bulundurulması olduğu görülebilir.

Geleneksel bilgi kayıt ortamlarına oranla elektronik ortamlardaki kaynakların metin, ses, resim, video, vb. içeriğinden oluşan daha farklı ve fazla türe sahip olması, bu verilerin daha etkin bir biçimde kimliklendirilmesi sorununu ortaya çıkarmıştır. Bu verileri kimliklendirme, düzenleme, yönetme, tarama ve erişim işlevlerini sürdürmek için kullanılan en önemli araç, üstveri'dir (Bayter 4).

Arslantekin (376), üstveri ile ilişkili olarak, yapılandırılmış ya da yapılandırılmamış verinin her birini karmaşık uzman dizinleme mekanizmaları ile taramanın oldukça kolay olduğunu, bununla birlikte bu iki veri şeklinin ikisinin de, hepsinden önemli olan üstveri ile bağlanmadığı takdirde, eldeki tüm bilgi birikimini kapsayan bir tarama gerçekleştirmenin zor olacağını vurgulamıştır.

Verinin Değeri

Verinin önemini anlatmak için ortaya konulabilecek en belirgin kriter, verinin yapılan çalışmalara kattığı değerdir. Bu nedenle verinin değerine ilişkin bazı kavramlara yer vermek gerekir. Bunlar; verinin gerçek değeri ve opsiyon değeri arasındaki ilişki, verinin amortisman değeri ve veri simsarlığı gibi kavramlardır

Verinin Gerçek Değeri ve Opsiyon Değeri Arasındaki İlişki

Rotelle'nın, Ann Winblad'tan aktardığına göre, *"Veri, yeni petroldür"* (Rotella) ya da Michael Palmer'ın söylediği gibi *"Veri, sadece ham petroldür, rafine edilmezse, değeri vardır ama kullanılabilir değildir"* (Palmer) ifadeleri gelecekte veriye sahip olanlar ve veriyi iş süreçlerinde daha iyi kullananların çok daha değerli kurumlar/firmalar/kişiler olacaklarını ve bu şekilde rakiplerine karşı ciddî fark yaratacaklarını kastetmektedir. Bu cümleler günümüzde veri ile ilgili en popüler

cümlelerdendir ve büyük veri ve veri yönetimi ile ilgili (Huang; Jeffery) pek çok kaynakta geçmektedir.

Andres Weigend ise "veri, yeni petroldür" sözüne aşağıdaki şu eklemeyi yapmıştır: *"Şirketler artık veriyi optimizasyon için kullanır hale geldi. Veri aslıda petrol gibidir. Şirketlerin veriyi üretmesi lâzım, saklaması lâzım ve analiz ederek rafine etmesi gerekiyor. Bu süreçleri yerine getiren şirketler, sadece ürettikleri bu veriyi ürün olarak kullanılabilir. Bu yeni bir oyun anlamına geliyor"* ("Artık Güç Tüketicide").

Verinin gerçek değeri, okyanusta yüzen buzdağının görünen parçası olarak tanımlanabilir. İlk bakışta sadece ufak bir kısmı görünürken, büyük bölümü deniz seviyesinin altında gizlidir. İnovatif şirketler bu değeri ortaya çıkarabilirler (Schönberger ve Cukier 110).

Eğer veri, geleceğin çok büyük değerlerinden biri ise, kurumlar bu doğrultuda yeni iş modellerini düşünmek durumundadırlar ve yeni modellerin değerli kaynağının yararlanılabilir hale getirmesi gerekmektedir (Rotella).

Veri'nin opsiyon değeri, verinin kullanılabileceği bütün olası biçimlerin toplamını ifade etmektedir. Görünüşte sonsuz olan bu potansiyel kullanımlar, opsiyonlar gibidir. Veri'nin değeri, bu seçeneklerin toplamıdır. Büyük Veri Çağı'nda, veri, ana değeri kullanılmaya başlandıktan sonra uzun süre değer vermeye devam eden sihirli elmas madenine benzetilmektedir. Veri'nin değeri, yeniden kullanarak, veri kümelerini birleştirerek, vb. yöntemlerle tekrar ortaya çıkarılabilir (Schönberger ve Cukier 111). Veri'nin değeri için en önemli nokta görünüşte sınırsız olan yeniden kullanım potansiyeli, yani verinin opsiyon değeridir. Bilgiyi toplamak önemli ama yeterli değildir, verinin sahip olduğu potansiyel, yalnızca sahip olunmasında değil, kullanımında yatmaktadır (Schönberger ve Cukier 129).

Veri'nin Amortisman Değeri

Dijital veri depolamanın maliyeti düştüğü için, kurumların, veriyi aynı ya da benzer amaçlarla verinin opsiyon değerinden de yararlanarak yeniden kullanmak üzere saklamak yönünde güçlü ekonomik nedenleri bulunabilmektedir. Fakat bu faydanın da bir sınırı olması doğaldır. Çoğu veri, zamanla faydasının bir kısmını kaybetmektedir. Bu nedenle kurumlar, veriyi sadece verimli olmaya devam ettiği sürece kullanma yönünde bir dürtüye sahiptirler ve bunun sonucu olarak da sürekli verilerini incelemeleri ve değerini kaybeden bilgiyi ayıklamaları gerekmektedir. Buradaki zorluk hangi verinin artık faydalı olmadığını bilebilmektir.

Bu kararı sadece zamana dayandırmak nadiren uygun olmakla beraber, genellikle kurumlar ilgisiz verileri ayıklamak için komplike modeller geliştirmektedirler. Bu şekilde, eski verinin faydası daha iyi değerlendirilebilir ve dolayısı ile veri için daha doğru amortisman oranları modellenebilir (Schönberger ve Cukier 117-118).

Büyük Veri Kavramı

Verinin, günümüzde organizasyonlar için çok büyük avantajlar ve fırsatlar sunmakta olduğu daha önce belirtilmişti. 2012 yılında Davos'taki Dünya Ekonomik Forumu'nda tıpkı para, altın gibi varlıklara ek olarak, yeni bir ekonomik değer olarak "veri" den bahsedilmiştir. Bir değer olarak kabul edilmesine karşın, verinin ekonomik değerini bulmak oldukça zordur. Bir başka deyişle verinin kişi, kurum kuruluş vb.lerine ekonomik katkısını rakamlarla ifade edebilmek oldukça güçtür. 2011 yılında Amerika Birleşik Devletleri'nde 17 endüstri sektöründen 15'indeki şirket başına düşen veri miktarı, Birleşik Devletler Kongre Kütüphanesi'nin sakladığı 235 Terabayt veriden daha çoktur. Wal-Mart Mağazaları Şirketi, her saat bir milyondan fazla müşterinin veri işlemini veri tabanında saklamak zorunda kalmaktadır ve sakladığı veri miktarı 2,5 petabayta ulaşmıştır. Bu rakam Kongre Kütüphanesi'nin elinde bulundurduğu veri miktarının yaklaşık olarak 167 katıdır. Yine bunlara benzer olarak 2010 yılında 5 milyar cep telefonu kullanılmıştır ve 30 milyar adet Facebook içeriği paylaşılmıştır (Johnson 51-52).

Günümüzde pek çok kuruma yalnızca kendilerine ait operasyonel veri tabanları yetmemektedir. Dış kaynaklardan alınan verilerle çeşitli analizler yapılarak yeni bilgilerin üretilmesi ve bu bilgilerin kurum için süreçlerde kullanılması ihtiyacı doğmuştur. Yaygın olan ve alışılmış veri tabanı yönetim sistemleri ise dış kaynaklardan gelen bu verilerin kurum içi enformasyonun yönetiminde kullanılması konusunda yeterli desteği verememektedirler. Çünkü dış kaynaklardan alınan veriler hem kendi operasyonel veri tabanlarına kolaylıkla aktarılabilir nitelikte hem de yapılandırılmış durumda olmayabilir. Bu nedenle günümüzde pek çok büyük teknoloji şirketi büyük veri konusunda çok büyük yatırımlar yapmaktadır.

Gürsakal (20), büyük veri kavramının ilk kez Ağustos 2000'de Francis X. Diebold tarafından, *Makroekonomik Ölçümler ve Kestirim İçin Büyük Veri Dinamik Faktör Modelleri* (Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting) isimli bildiri ile Seattle'da 8. Dünya Ekonometri Kongresi'nde ortaya atıldığını belirtmektedir. Francis X. Diebold ise (3), büyük veri kavramının ilk

defa Silicon Graphics (SGI)'den John Mashey[1] tarafından 1998'de *Büyük Veri ve Altyapı Gerilimi Dalgası* (Big Data and the Next Wave of InfraStress) isimli sunumunda kullanıldığını belirtmektedir. Halen Gartner'ın bir parçası olan Meta Group isimli şirket ise, 2001 yılında büyük veriyi niteleyen hacim, hız ve çeşitliliği konularından bahsetmiştir (Laney). Bunun sonucunda hemen her ortamda büyük veri, 3V ile anılmaya başlanmıştır. Büyük verinin bu terimler ile nitelendirilmesi Meta Group'un literatüre önemli bir katkısıdır.

Schönberger ve Cukier (14), büyük verinin kesin bir tanımının bulunmadığını ifade etmektedirler. İlk başlarda bilgi hacminin çok büyümesi durumunda, incelenen veri miktarının, bilgisayarların işlem için kullandıkları belleğe uygun olmadığı düşünülmekteydi. Mühendislerin, verinin tümünü analiz etmek için kullandıkları araçları yenilemeleri gerekiyordu. Günümüzde ise Google'ın MapReduce'u ve bunun Yahoo'da ortaya çıkan açık yazılımdaki karşılığı sayılabilecek Hadoop gibi yeni işlem teknolojilerinin kaynağı olan bu araçlar, günümüzde eskisinden çok daha fazla verinin yönetilmesine olanak tanımaktadır. Bu teknolojilerde verinin düzenli sıralara ya da klâsik veri tabanı tablolarına yerleştirilmesi gerekmemektedir.

Büyük veri ile birlikte eskiden asla ölçülemeyen, saklanamayan, analiz edilemeyen ve paylaşılamayan şeylerin büyük çoğunluğu verileştirilmeye başlanmıştır (Schönberger ve Cukier 25).

Büyük veri genel olarak kullanılan programların saklama, yönetme ve işleme kapasitesinin ötesindeki veri kümelerini anlatmak için kullanılan bir terimdir. Büyük verinin devasa boyutları ile bundan fayda sağlamak için gereken analizlerin karmaşıklığının birleşmesi, yeni sınıf teknolojilerin ve bunları yönetecek araçların gelişmesine neden olmuştur. Aslında büyük veri, genelde, hem yönetilen verinin türünü, hem de onu depolamak ve işlemek için kullanılan teknolojiyi anlatmaktadır. Bu teknolojilerin büyük bir kısmı, Google, Amazon, Facebook ve LinkedIn vb. şirketlerin inanılmaz büyük sosyal medya verisi ile uğraşırlarken, kendileri için geliştirdikleri teknolojiden doğmuştur. Bu şirketler, doğası gereği, düşük maliyetli hazırda bulunan donanım ve açık kaynaklı yazılımlara önem vermektedirler (Cackett 14).

[1] Mashey'in bu sunumuna https://www.usenix.org/legacy/event/usenix99/invited_talks/mashey.pdf adresinden ulaşılabilir.

Büyük veri, genellikle birbirlerinden farklı veri kaynaklarından toplanan geniş veri dermelerinin analizi, işlenmesi ve depolanması ile ilgili bir alandır. Büyük veri çözümlerinin ve uygulamalarımın karakteristik, yani kendine özgü olması gerekmektedir. Geleneksel veri analizi işlemleri, depolama teknolojileri ve teknikleri yetersiz kalmaktadır. Spesifik olarak büyük veri, çoklu ilişkisiz veri kümelerinin birleştirilmesi, büyük miktarda yapısal olmayan verinin işlenmesi, gizli enformasyonun kısıtlı zaman içinde toplanması gibi farklı gereksinimlere işaret etmektir (Erl, Khattak ve Buhler 19).

Monino ve Sedkaoui (XI), terim olarak büyük veriyi, *"büyük veri terimi, organizasyon için kullanılan verinin hacmi kritik seviyeye ulaştığında ve bunun için yeni teknolojik depolama, işlem ve kullanım yöntemleri yaklaşımları gerektiğinde kullanılmaktadır"* şeklinde tanımlamıştır.

McKinsey Global Institute, 2011 yılında büyük veri kavramını, tipik ve geleneksel veri tabanı yazılımlarının yapamayacağı şekilde, bunların kabiliyetlerinin ötesinde, veri kümelerini alan, saklayan, yöneten, erişime sunan ve analiz eden araçları tanımlamak için kullanmıştır (Manyika, Chui ve Brown 1). Bu tanım, konusu büyük veri olan Kord Davis ve Doug Patterson'ın (4), *Ethics of Big Data* gibi başka yayınlarda da kullanılmış olduğundan konu açısından önem taşımaktadır.

Gülle (581), "*Büyük Veri ya da İçgörü*" başlıklı yazısında büyük veriyi, kurum ve kuruluşların stratejilerinde öncelikli olarak yer alan "öngörü" modelinin yanı sıra "içgörü"ye de bir model olarak odaklanılmasının gerekli olduğu düşüncesini belirtmiştir.

Ancak unutmamak gerekir ki, büyük verinin yalnızca verinin hacmi nedeniyle büyük olduğu söylenemez. Büyük verideki "büyük" kelimesi verinin işlenme sürecindeki önemini ve etkisini de kapsamaktadır. Açık veri ile bu miktar devamlı olarak artmıştır (Monino ve Sedkaoui XXXIV).

Büyük Veriyi Niteleyen Unsurlar

Görüldüğü üzere büyük veri tanımları çok büyük kesinlik taşımamaktadır. Belki büyük veri zamanla değişecek ve bugünün büyük verisi gelecekte aynı anlama gelmeyebilecektir. Bu yüzden büyük veri kavramının tanımlamasında yardımcı olması için genellikle verinin hacmi, hızı ve çeşitliliğini ifade eden "3 V" (volume, velocity, variety) notasyonu, yaygın olarak onu diğer veri türlerinden ayıran kavramlar olarak kullanılmaktadır. Nitekim büyük veri kavramının tanımlanmasındaki değişim günümüzde bile kendini göstermektedir. Literatürde

3V'ye "verinin değerini" (value) ekleyerek 4V ile tanımlayanlar da bulunmaktadır (Cackett 14).

Verinin hacmi, verinin büyüklüğü ve boyutunu ifade etmektedir. Verinin boyutunu rakamsal bir şekilde belirtmek genelde çok kısıtlayıcı olmaktadır. Teknoloji ilerledikçe, rakamlar hızlı bir şekilde değişmektedir ve kısıtlayıcı rakamlar artık geçerliliğini yitirmektedir. Bu yüzden verinin göreceli miktarını belirtmek daha faydalı olmaktadır. Eğer ilgilenilen verinin miktarı daha önce kullanılan verinin üstündeyse muhtemelen büyük veri ile uğraşılmaktadır. Bu bazı kurum/kuruluşlar için onlarca terabayt olurken, bazıları için onlarca petabayt olabilmektedir (Cackett 14).

Günümüzde artık pek çok cihaz veri üretebilir hale gelmiştir. Gerek bireylerin ve kurumların sakladığı verilerin, gerekse de internet dünyasında saklanan verinin büyüklüğü her geçen gün artmaktadır. Veri depolama birimlerinin fiyatlarındaki hızlı düşüş, saklanan verilerin oranının geometrik şekilde hızlanarak artmasında önemli bir etken olmaktadır. Eskiden yalnızca operasyonel veri tabanlarının kullanılması yeterli olurken şimdi ise, bilgi ve iletişim teknolojilerinde yaşanan gelişmeler doğrultusunda, veri ambarlarında toplanan bütün veri işlenip analiz edilebilir hale gelmiştir.

Verinin hızı, elde edilen veri ile ilgili gerçek zamanlı (anlık) olarak harekete geçilebilmesini ifade etmektedir. Her ne kadar gerçek veri analizinin verinin geldiği dönemle aynı anda tamamlanması mümkün olmasa da; uygulamaya geçmedeki gecikmeler kaçınılmaz olarak yapılması istenen ve beklenen çalışmaların verimliğini kısıtlamakta, müdahale etmeyi zorlaştırmakta ve optimal olmayan süreçlere yol açmaktadır. Örneğin, coğrafik konum olarak müşterinin nerede olduğuna dayanarak yapılan bir indirim/promosyon teklifi; müşteri o noktadan geçtikten sonra müşteriye ulaşırsa, başarılı olma şansı çok düşebilecektir (Cackett 14).

Günümüzde bilgi ve iletişim teknolojilerinde yaşanan gelişmeler, verinin üretildiği anda kullanılmasına olanak vermektedir. Hızla akan veriye en hızlı tepkiyi verip, daha veri akarken müdahale etmeyi, işlemeyi ve analiz etmeyi olanaklı hale getirmiştir. Verinin bu hızına yetişebilen firmalar daha veri yeni yaratıldığı anda yanlış yapılan bir işleme müdahale edebilmekte; bu veriler ortaya çıktığı anda kurumlar kendi analiz süreçlerine katabilmekte; karar destek sistemlerindeki analiz süreçlerine aynı anda bu veriler eklenip kullanılabilmektedir.

Verinin çeşitliliğinin söz dizimi (syntax) ve semantik (anlamsal) olmak üzere 2 boyutu vardır. Geçmiş dönemlerde bu iki boyut, hangi verinin güvenilir bir biçimde veri tabanlarında yapılandırıldığının ve analizin içeriği için ne kadar güvenilir olduğunun derecesini belirlemekteydi. Modern ETL[2] araçları görsel olarak gelen sanal sözdizimi verilerini çok başarılı bir şekilde işleyebilirken, serbest metin gibi semantik olarak zengin verilerin çözümünde daha başarısızlardır. Bu yüzden birçok organizasyon, enformasyon yönetim sisteminin veri kapsamını daha dar bir veri düzeni ile sınırlamışlardır. Bu sınırlamayı organizasyonların daha kapsayıcı, ek değer yaratması takip etmiştir ve bu muhtemelen Büyük Veri yaklaşımının en çekici olan özeliklerinden biridir (Cackett 15).

Günümüzde veri, önceden olduğu gibi sadece yapılandırılmış veriden değil, aynı zamanda yapılandırılmamış verilerden de oluşmaktadır. Bu veriler bilinen ve yeni kabul edilen ortamların yanında gittikçe artar duruma gelmiştir. Hattâ bazı büyük web siteleri kullanıcının imlecini nerelerde gezdirdiğinin ve web kullanım bilgilerini bile veri olarak saklamaktadırlar. Bu çeşitlilik hem kendi arasında, hem de alt dallarıyla her geçen gün hızla büyümektedir. Büyük veri ile birlikte günümüzde tüm bu çeşitler iş süreçlerinde de kullanılabilir hale gelmiştir.

Çalışmanın başlarında büyük verinin oluşumunda pek çok noktadan veri toplandığından bahsedilmişti. Bu doğrultuda günümüzde veri sağlayan pek çok araçtan söz edilebilir. Bu bağlamda da karşımıza bir başka yeni kavram olan "Nesnelerin İnterneti" çıkmaktadır.

Nesnelerin İnterneti

Nesnelerin interneti, her gün kullanılan nesnelerin içine çipler, sensörler ve iletişim modülleri yerleştirilerek, kısmen çevrimiçi ağ oluşturmakla, ama bundan da daha çok insanları çevreleyen her şeyi verileştirmekle ilgili bir kavramdır. Dünya bir kere verileştirildiğinde, bilginin potansiyel kullanımları temel olarak sadece kişinin marifetleri ile sınırlı olabilecektir. Verileştirme, insanın kavrayışında temel bir zenginleşmeyi temsil etmektedir. Büyük veri ile birlikte, bundan böyle dünya temel olarak bilgiden oluşan bir evren olarak görülebilecektir (Schönberger ve Cukier 103-104).

[2] ETL: Extraction,transforming ve load işlemlerinin kısaltması . Şirketlerin networklerinde farklı yerlerde / veri tabanlarında olan bilgilerin oradan alınması (extraction), temizlenip belli bir formata dönüştürülmesi (transforming) ve veri madenciliği yapılacak veri tabanına yüklenmesini (load) belirtir.

Nesnelerin interneti kavramı, sensörlerinin kablolu ya da kablosuz bağlantıları yoluyla, iletişim kurabilen aygıtların birbirlerine bağlanabilme kabiliyetlerini tanımlamak için kullanılmaktadır. Bu aygıtlar; termostat, arabanız, doktorunuzun sağlık durumuzu kontrol edebilmesi için yuttuğunuz hap şeklinde bir ilaç vb. olabilmektedir. Bu birbirine bağlı aygıtlar, verileri iletmek, derlemek ve analiz etmek için interneti kullanmaktadırlar (The White House).

Birbirlerine bağlanan ve birbirleriyle haberleşen bu cihazların artması, veri patlamasına neden olmakta, veri patlaması sonucu büyük veri oluşmakta ve bununla birlikte günümüzde verinin analiz edilmesi ile ilgilide önemli sorunlar doğurmaktadır. Büyük veriyi depolamak ve erişime sunmak kadar, analizi içinde yeni yaklaşımlar ve yöntemler geliştirilmektedir. Bu yöntemlerin en başında veri madenciliği ve metin madenciliği gelmektedir.

Veri Madenciliği

Veri madenciliği elde edilen büyük verinin analizi için en önemli yöntemlerden biridir.

Veri Madenciliği İngilizce-Türkçe Ansiklopedik Bilişim Sözlüğü'nde, *"Doğal Dillerin Semantik yapısına dayanarak elektronik metin belgeleri içinde saklı kalmış ilintileri, örüntüleri, stratejik bilgileri, modelleri vb. bulup ortaya çıkarmayı amaçlayan araştırma tekniği"* (Sankur 831) şeklinde tanımlanmaktadır.

Veri madenciliği, veriden bilgi keşfi olarak da tanımlanabilir. Veri madenciliğinde otomatik ve kısmî otomatik metotlar kullanılarak büyük miktarlı veriden bilgi çıkarımı hedeflenir. Veri madenciliği veriden modeller geliştirmek, ilgi çekici yapılar veya yinelenen temalar bulmak vb. için istatistik, yapay zekâ, bilgisayar bilimi gibi çeşitli bilim dallarından algoritmalar kullanmaktadır. Veri içindeki kullanışlı enformasyonun ve mümkün, anlamlı ve kullanışlı ilişkilerin bulunabilmesi için veritabanı enformasyonunu analiz edebilen bütün teknolojileri bir araya getirmektedir (Monino ve Sedkaoui XIII).

Veri madenciliği, veri ambarlarında saklanan, yararlı olabilecek, aralarında bilinmeyen ilişkilerin olduğu verilerin keşfedilerek, bu verilerin hem anlaşılır hem de kullanılabilir bir şekle dönüştürülmesine yönelik geliştirilmiş yöntemler topluluğudur (Oğuzlar 5-6).

Arslantekin (372-373), veri madenciliğini, *"büyük miktarda veriden anlamlı bilgi çıkarma sanatıdır…toplanan büyük yığın halindeki veriler arasında örnek kalıpların tanımlanması, eğilimlerin belirlenmesi ve gerekli ilişkilerin kurulması işlemlerine ait bir süreçtir"* şeklinde belirtmiştir.

Cackett ise (10), veri madenciliğini, büyük miktarlı veriden, otomatik ve yarı otomatik yöntemler kullanmak suretiyle bilinmeyen enformasyonun çıkarılması şeklinde tanımlamıştır. Bazı yayınlarda geniş veri kümelerindeki bilgi keşfi olarak anılmaktadır ki bu tanım veri madenciliğinin geçmişi şeklinde düşünülebilmektedir.

Veri madenciliği, *"ilişkileri ve modelleri analiz etme amacı ile gizli olan enformasyonu keşfetmek için ayrıntılı teknikler kullanarak, enformasyonun veri tabanından ve veri kümelerinden çıkarılması işlemidir"* (Prytherch 195) şeklinde de tanımlanmaktadır.

Veri madenciliğini büyük veri bağlamında değerlendirecek olursak, elde edilen büyük verinin içindeki gizli olan enformasyonun önceden güvenilirliği kanıtlanmış istatistiksel tekniklerle ortaya çıkarılmasıdır şeklinde tanımlayabiliriz. Büyük verinin analizinde en temel yöntemlerden birisi olması bakımından veri madenciliği bu konuda son derece önemlidir.

Metin Madenciliği

Özellikle web ortamından alınan metin halindeki büyük verinin analiz edilmesinde metin madenciliği çok önemli bir istatistiksel tekniktir. Kimi zaman metin halindeki veriler sosyal medyadan elde edildiğinde buna sosyal medya madenciliği, web üzerinden elde edildiğinde buna web madenciliği gibi tanımlar yapılsa da ve her birinin kendine özgü yöntemleri olsa da, bunların geneli temel olarak veri madenciliği ve metin madenciliğine dayanmaktadır.

Metin Madenciliği, İngilizce-Türkçe Ansiklopedik Bilişim Sözlüğü'nde *"istatistik yöntemlere dayanarak büyük veri hacimleri, veri tabanları, örün sunucuları içinde gizli kalmış yönsemeleri, stratejik bilgileri ilişkileri keşfetmeye yönelik açınsama çalışması"* (Sankur 208-209) şeklinde tanımlanmaktadır.

Metin madenciliği geniş hacimdeki metin içeriklerinin ana eğilimlerini çıkarmak ve farklı konulardaki uğraşıları istatiksel değerlemek için süreçleri otomatikleştirmeyi mümkün kılan bir tekniktir (Monino ve Sedkaoui XVI).

Metin madenciliğini oluşturan temel alanlar istatistik, veri madenciliği, doğal dil işleme, web madenciliği ve bilgi erişimidir (Oğuzlar 20).

Doğal dil metinlerindeki örüntülerle ilgilenen ve yeni eski veya bilinmeyen enformasyonu keşfetmek amacıyla doğal dil işleme, veri madenciliği ve bilgi erişim tekniklerinin uygulayan teknolojidir. Metin madenciliği, birbirleriyle ilişkili metinlerden oluşan kaynakları bir araya getirmek için, bunları analiz etmek ve tanımlamak için, anahtar varlıkları ve bunların özelliklerini içeren nitelikleri çıkarmak için ve çıkarılmış nitelikleri birleştirmek için, yeni nitelikleri şekillendirmek için veya değerli içgörü kazanmak için kullanılabilir (Prytherch 688).

Yukarıdaki tanımlardan da anlaşılacağı üzere, her tanım farklı noktalara değinebilmektedir. Hattâ bazı tanımları herkesin anlamasında güçlük çekilmesi de söz konusudur. Metin madenciliği, "pek çok farklı ortamdan olabileceği gibi özellikle web ortamından elde edilen yapısal olmayan metin türündeki verilerin içindeki gizli enformasyonun çıkarılmasını sağlayan, bu enformasyonun çıkarılması sırasında metinlerin içindeki ilişkileri bulan ve bunları çeşitli kalıplarda ifade edilmesine olanak sağlayan, bu kalıplara dayanarak geleceğe yönelik tahminlerde bulunulmasına olanak veren istatiksel analiz yöntemidir" şeklinde de tanımlanabilir.

Belirttiğimiz bu madencilik teknikleri beraberinde büyük veri ile ilgili bazı yapıları da gündeme getirmiştir. Bunlar arasında Doğal Dil İşleme, NoSQL, Google Map Reduce, Hadoop sayılabilir.

Doğal Dil İşleme (Natural Language Processing - NLP)

Büyük veri teknolojilerine yatırım yapan kurum ve kuruluşlar için elde edilen metinlerin analiz edilmesinde en önemli aşamalardan biri de doğa dil işlemedir. Bu kurum ve kuruluşların, çok büyük metin verilerine sahip olmaları ve bu metinleri analiz etme gereği duymaları nedeniyle zamanla doğal dili en iyi işleyen şirketler oldukları da söylenebilir.

NLP doğal dillerin kurallı yapısının çözümlenerek anlaşılması veya yeniden üretilmesi amacını taşımaktadır. Bu çözümlemenin insana getirdiği kolaylıklar ise yazılı dokümanların otomatik çevrilmesi, soru-yanıt makineleri, otomatik konuşma ve komut anlama, konuşma sentezi, konuşma üretme, otomatik metin özetleme, bilgi sağlama gibi birçok başlıkla özetlenebilmektedir. Bilgi iletişim teknolojilerinin yaygın kullanımı, bu başlıklardan üretilen uzman yazılımların hayatımızın her alanına girmesini sağlamıştır. Örneğin, tüm kelime düzeltme yazılımları bir imlâ düzeltme aracı taşır. Bu araçlar aslında yazılan metni çözümleyerek dil kurallarını denetleyen doğal dil işleme yazılımlarıdır (Oğuzlar 13).

NLP, yapay zekâ (bilgi gösterimi, planlama, akıl yürütme vb.) biçimsel diller kuramı (dil çözümleme), kuramsal dilbilim ve bilgisayar destekli dilbilim, psikoloji gibi pek çok farklı alanlarda geliştirilmiş kuram, yöntem ve teknolojileri bir araya getirmektedir (Oğuzlar 12).

Dilin insan tarafından işlenmesinin algoritmaları kesin olarak bilinmemekle birlikte, her ne kadar çeşitli sistemler bir araya getirmeye çalışılsa da, istenilen yapıya henüz ulaşılamamıştır. Bu yüzden metin anlama sistemlerinin çoğunda problemi birkaç alt göreve ayırdıktan sonra, onları ayrı olarak çözen geleneksel böl ve yönet stratejisi kullanılmaktadır (Oğuzlar 14).

Büyük veri ve doğal dilin işlenmesi iki temel konu altında incelenebilir:

- Büyük veri ile birlikte, çok büyük miktarda veri ile çalışıldığında bilinen pek çok doğal dilin işlenmesi yöntemi değişmiş ve büyük veri, doğal dilin işlenmesi yöntemleri için kaynak oluşturmuştur. Doğal dilin işlenmesi konusunda günümüzde yaşanan önemli gelişmelerde büyük veri konusunun önemli bir etkisi olduğu söylenebilir.
- Büyük veri ile ilgili çalışmalarda, büyük verinin işlenerek anlamlı yeni bilgi ve öngörüler oluşturulabilmesinde doğal dil işleme yöntemleri bir araç olarak kullanılmaktadır.

Bütün bu çalışmalar web üzerindeki milyarlarca verinin işlenmesini, akıllı algoritmalar ile işlenerek anlamlı yeni bilgi ve öngörüler oluşturabilmesini kolaylaştırmaktadır.

NoSQL

Son on yılda, genellikle NoSQL (not only SQL) (SQL ve daha fazlası) sistemleri olarak adlandırılan, veri yönetim sistemlerinin yeni bir sınıfı ortaya çıkmıştır ve günümüzde de oldukça hızlı bir gelişme göstermektedir.

Veri tabanlarına erişmenin en yaygın dili uzun süre SQL (Structured Query Language) ya da Türkçe karşılığı ile Yapısal Sorgulama Dili olmuştur. Son yıllarda önceden belirlenmiş kayıt yapısı gerektirmeyen NoSQL'e büyük bir kayma yaşanmaktadır. NoSQL, farklı tip ve büyüklüklerde veriyi kabul etmekte ve başarı ile bu veriler içinden arama yapılmasına olanak tanımaktadır. Bu veritabanı tasarımları, yapısal dağınıklığa izin vermesinin karşılığında daha fazla işlem ve depolama alanı gerektirmektedir. Yine de bu durum, depolama ve işleme maliyetleri düştüğü için günümüzde pek çok kuruluşun gücünün yetebileceği bir değişim olarak düşünülebilir (Schönberger ve Cukier 52-53).

Büyük veri ile birlikte günümüzde geleneksel veri tabanı çözümlerinin yeterli olamaması nedeniyle büyük veri uygulamalarının hemen hemen hepsinde, NoSQL sistemler kullanılmaktadır.

Google MapReduce

MapReduce bir programlama modeli ve aynı zamanda büyük veri kümelerinin işlenmesi ve oluşturulmasıyla ilişkili uygulamadır (Dean ve Ghemawat 1).

Google MapReduce programlama modelini birçok farklı amaç için kullanmıştır. Google MapReduce başarısını çeşitli nedenlere bağlamaktadır. Birincisi bu modelin kullanımı paralel ve dağıtılmış sistem tecrübesi olmayan programcılar için bile kolaydır. İkincisi büyük çeşitlilikteki problemler, MapReduce hesaplamaları gibi kolayca ifade edilebilir. Örnek olarak MapReduce veri üretmek için, Google'ın ürünü olan arama motoru servisi için, sıralama için, veri madenciliği için, makine öğrenimi için ve daha pek çok diğer sistem için kullanılabilir. Üçüncü olarak Google, MapReduce'un binlerce makineyi içeren büyük makine kümelerine ölçeklenebilir bir uygulamasını geliştirmiştir. Bu uygulama makine kaynaklarını efektif kullanılmasını sağlar ve bu nedenle Google'da karşılaşılan geniş sayısal problemler için kullanımı uygundur (Dean ve Ghemawat 12).

Hadoop

Hadoop, depolama sistemi ve dağıtılmış işleme araçlarını içeren büyük veri altyapısı şeklinde tanımlanabilir (Monino ve Sedkaoui XIII).

Hadoop, Google MapReduce'un en büyük rakibidir. Büyük veriyi işleme konusunda günümüzde çok popülerdir ve büyük veri ile gelen değişimin simgesi olmuştur. Büyük veriyi daha küçük kümelere bölmekte ve bunları başka makinelere paylaştırmaktadır. Verinin temiz ve düzenli olmadığını yani verinin işlenmeden önce temizlenemeyecek kadar büyük olduğunu varsaymaktadır. Temel veri analizi, veriyi analiz edileceği yere taşımak için ETL denen bir işlem gerektirirken, Hadoop bu işlem yerine veri miktarının çok büyük boyutlarda olduğunu, bu nedenle taşınamayacağını ve olduğu yerde analiz edilmesi gerektiğini kabul etmektedir. Hadoop' un çıktıları ilişkisel veri tabanları kadar kesin değildir. Örneğin banka hesaplarının ayrıntılarını sorgulamak için pek güvenilir değildir. Ama kesin yanıtların gerekmediği birçok alanda ilişkisel veri tabanlarından çok daha hızlıdır. Buna da bir örnek vermek gerekirse; Visa, Hadoop kullanarak yaklaşık 73 milyar işlemi içeren 2 yıllık kayıtları için gereken işlem süresini yaklaşık bir aydan 13 dakikaya indirebilmiştir (Schönberger ve Cukier 53-54).

Schönberger ve Cukier (55), bu veri dağınıklığı ve büyüklüğü içinde dağınıklıkla yaşamanın karşılığında, geleneksel metotlar ve araçlarla kendi kapsamlarında çok değerli hizmetler alındığını belirtmişlerdir. Bunun yanında bütün dijital verinin sadece %5' inin yapısal olduğunu; dağınıklık kabul edilmezse, web sayfaları ve videolar gibi geriye kalan % 95 yapısal olmayan verinin belirsiz kalacağını; içgörülerden yararlanılamayan evrene bir pencere açılacağından da bahsetmektedirler.

Hadoop günümüzde büyük veri konusunda teknoloji geliştiren ve büyük veriyi kullanan pek çok kurum ve kuruluş için vaz geçilmez olmuştur.

Dünyada Büyük Veri'nin Kullanım Alanları ile ilgili Örnekler

Dünyada büyük verinin farklı alanlardaki bazı başarılı uygulamalarına örnekler verilmesi konunun daha rahat anlaşılmasını sağlayabilir.

PriceStats - Billion Prices Project

2007 yılının Ekim ayında Alberto Cavallo, çevrimiçi fiyatlardan faydalanarak, Arjantin, Şili, Brezilya ve Kolombiya için günlük enflasyon rakamları hesaplayan *Kazanılmış Veri ve Değişmez Fiyatlar: Sıklık, Tehlikeler ve Senkronizasyon* (Scraped Data and Sticky Prices: Frequency, Hazards, and Synchronization) isimli doktora tezini yayınlamıştır. 2008 yılının Mart ayında ise Alberto Cavallo ve Roberto Rigobon veri dermelerini 50 ülkeyi kapsayacak şekilde genişleterek, çevrimiçi verileri kullanan akademik araştırmayı yürütmek için Massachusetts Institute of Technolgy (MIT) akademik girişimi ile *Milyar Fiyat Projesi* (Billion Prices Project)'i başlatmışlardır. 2014 yılında Türkiye'nin de enflasyon rakamları bu siteye eklenmiştir (PriceStats).

2012 yılında "The Economist" dergisiü, Arjantin'in resmi enflasyon rakamları yerine bu proje kapsamında hesaplanan enflasyon verilerini kullanmaya başlamıştır (PriceStats). Bu olayla birlikte PriceStats'ın büyük veriyi kullanarak hesapladığı rakamların, kimi otoritelerce ülkelerin resmi istatistikleri yerine daha güvenilir bulunarak kullanılması ise üzerinde düşünülmesi gereken önemli bir noktadır.

Walmart

Walmart 2004 yılında Teradata'nın sayısal çözümleme uzmanları ile hangi müşterinin hangi ürünü aldığı, toplam maliyetleri, alışveriş sepetlerinde başka neler olduğu, günün saatleri ve hatta durumları gibi verileri içeren devasa veri tabanlarını incelemişlerdir. Bu incelemeyi yaparken şirket, bir kasırga öncesinde sadece el feneri satışlarının değil aynı zamanda şekerli bir Amerikan gevreği olan

Pop-Tarts satışlarının da arttığını fark etmiştir. Sonrasında hızla girip çıkan müşterileri için mağazanın ön tarafındaki kasırga malzemelerinin yanına Pop-Tarts'ları depolayarak satışları önemli ölçüde artırmıştır. Geçmişte, verinin toplanması ve fikirlerin test edilmesi için merkezdeki bir çalışanın önceden içine doğması gerekirken, günümüzde Walmart bu kadar büyüklükte veriye ve daha iyi araçlara sahip olduğu için, korelasyonları çok daha hızlı ve ucuz şekilde ortaya çıkarıp bunları şirket operasyonlarında kullanarak büyük faydalar sağlayabilmiştir ve günümüzde de fayda sağlamaktadır (Schönberger ve Cukier 61).

Barnes & Noble

E-kitap okuma cihazları, onları okuyan insanların yazınsal tercihleri ve alışkanlıkları üzerine çok miktarda veriyi ele geçirebilmektedirler. Okuyucuların bir sayfayı ya da bir bölümü okumalarının ne kadar sürmekte oldukları, göz ucu ile mi bakıp geçtikleri ya da gerçekten mi okudukları, bir pasajın altını çizdiklerinde, bir not aldıklarında bunların kaydedilmesi, vb. olaylar örnek olarak verilebilir. Bu tür veri toplama becerisi, uzun süre yalnız bir eylem olan okumayı bir çeşit müşterek deneyime dönüştürmektedir. Bir kere toplandığında veri dışatımı yayıncılara ve yazarlara önceden hiç bilmedikleri ve beğenmedikleri nicel şeyleri söyleyebilmektedir. Bu bilgiyi ticari olarak değerlendiren e-kitap firmalarının kitapların içerik ve yapısını iyileştirmek için yayıncılara sattıkları düşünülebilir. Barnes&Noble'ın Nook e-kitap okuyucusundan aldığı veri üzerinde yaptığı analiz, okuyucuların kurgu dışı uzun kitapları yarısında bırakmaya meyilli olduklarını ortaya koymuştur. Bu keşif şirketin "Noon Snaps" isimli bir seri yaratmasına ilham kaynağı olmuştur. Bunlar sağlık ve güncel gelişmeler gibi gündemdeki konular üzerine kısa çalışmalar şeklindedir (Schönberger ve Cukier 121).

Bu örnekte önemli olan konulardan biri de Barnes and Nobles'dan hizmeti aldıktan sonra bile son kullanıcının, eseri okumasının her aşamasında veri üretmeye devam ettiğinin ortaya konmasıdır.

Dünya Kupası 2014 Brezilya (World Cup 2014 Brazil)

Bilindiği gibi 2014 yılında FIFA Dünya Kupası şampiyonu Almanya olmuştur (Brazilian Federal Government). Turnuvada SAP ve Almanya Futbol Federasyonu (DFB), kupadaki oyuncu performanslarını artırmak için büyük veriyi akıllı kararlara dönüştürecek bir sistem için inovatif bir işbirliği yapmışlardır. SAP HANA platformunda çalışan bu çözüm, antrenmanların, hazırlıkların ve turnuvaların analizi kolaylaştırmak için ve oyuncu - takım performanslarını artırmak için

tasarlanmıştır. Oliver Bierhoff, bu konuda 10 dakikada 10 oyuncunun, 7 milyondan fazla veri noktasından veri ürettiğini belirtmektedir. İlgili çözüm ile bir sonraki maçın antrenmanlarının ve hazırlığının yapılmasında bu büyüklükteki verinin analizi gerçekleştirilebilmiştir (SAP SE). Bu analizler Almanya'ya kupayı getirmede büyük rol oynamıştır.

Futbol dünyasında başarının gelmesinde payı olan bu çözüm hem büyük verinin kullanım alanlarının çeşitliliği ve geliştirilmesi konusunda hem de spor dünyası açısından çok önemli bir gelişmedir. Maç içinde toplanan verinin, enformasyona dönüşmesinde ve bu enformasyonun karar süreçlerinde kullanılarak önemli bilgiler elde edilmesinde büyük veri uygulamalarında ne kadar önemli olduğu bu örnekte değerlendirilebilir.

Sonuç

Büyük verinin günümüzde bu kadar önemli ve üzerinde durulan bir konu haline gelmesinin temel nedeni; çeşitli devletlerin, toplulukların, kurum ve kuruluşların yaptıkları işler ve verdikleri hizmetler yanında, internet ve benzeri teknolojilerin kullanımı ile yaygınlaşan uygulamalar sırasında oluşan ve günümüze kadar değerlendirilmeyen verilerin öneminin anlaşılmasıdır. Belirtilen bu organizasyonlar büyük verinin işlenmesiyle kendileri için büyük fayda sağlayabilecek enformasyon üretebileceklerinin farkına varmışlardır. Bu farkındalık sonucunda günümüzde bu konuya çok büyük yatırımlar yapmaktadırlar. Bu sonuç büyük veri kavramının literatürde ve güncel medyada çokça yer almasına neden olmuştur.

Büyük veri ile birlikte yalnızca teknoloji alanında değil, düşünme, algılama biçimlerimizde, araştırma yöntemlerimizde, daha pek çok farklı alanda büyük değişiklikler yaşanmaktadır. Kurumlar, kuruluşlar ve bireylerin de bu değişimlerin dışında kalamayacağı sonucuna varılmıştır.

Araştırma kapsamında elde edilen bu sonuçlara göre yapılması gerekenler şu şekilde sıralanabilir:

- Özellikle araştırma kurumları büyük verinin etkin kullanımı konusunda öncü olmalı, üniversitelerin ve bilimle ilgili diğer kurum ve kuruluşların büyük veri ile ilgili teknoloji ve uygulama geliştirilmesine destek vermeleri gerekmektedir.

- Büyük veriyi kendi meslekleri doğrultunda değerlendirmek isteyenler için, gerekli eğitimi almaları konusunda gerekli ortam sağlanmalı, bu konuya eğitim programlarında yer verilmelidir.

• Büyük veri konusunu ve araçlarını anlatan uygulamalı seminerler ve hizmet içi eğitimleri düzenlenmelidir.

• Verilecek eğitimlerde büyük veriyi işleme sonucunda elde edilecekler somut olarak ortaya konmalı, böylelikle eğitim alacak kişilerin konunun önemini daha gerçekçi kavramalarına olanak sağlanmalıdır.

• Büyük veri konusunda yurt dışı çalışmalar örnek alınıp incelenmeli, benzeri modeller yurtiçinde özellikle üniversiteler öncülüğünde gerçekleştirilmelidir.

Şüphesiz bu konu bir takım yatırımları gerektirmektedir. Bu durumda malî destek devlet kurum ve kuruluşları tarafından sağlanabilir.

KAYNAKÇA

Arslantekin, Sacit. "Veri Madenciliği ve Bilgi Merkezleri." *Türk Kütüphaneciliği* 17.4 (2003): 369-380.

"*Artık Güç Tüketicide*." Hürriyet 27 Mayıs 2015. Web. 12 Nisan 2016.

Bayter, Mustafa. *Türkçe Web Belgelerinin Kataloglanması: Bir İşbirliği Modeli Önerisi*. Yayınlanmamış Doktora Tezi, Ankara Üniversitesi, Ankara, 2008. Web. 1 Haziran 2016.

Brazilian Federal Government. *All results*. 13. 07. 2014. Web. 5 Nisan 2016.

Cackett, Doug. *Information Management and Big Data, A Reference Architecture*. White paper. Redwood Shores: Oracle Corporation, 2013. Web. 20 Nisan 2016.

Davis, Kord ve Doug Patterson. *Ethics of Big Data: Balancing Risk and Innovation*. Sebastopol: O'Reilly, 2012.

Dean, Jeffrey ve Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters." *Google, Inc.* Web. 6 Nisan 2016.

Diebold, Francis X. "A Personal Perspective on The Origin(S) and Development of "Big Data": The Phenomenon, The Term, and The Discipline." Web. 1 Mayıs 2016.

Erl, Thomas, Wajid Khattak ve Paul Buhler. *Big Data Fundamentals, Concepts, Drivers & Techniques*. Indiana: Arcitura Education Inc, 2016.

Gülle, M. Tayfun. "Büyük Veri ya da İçgörü." *Türk Kütüphaneciliği* 27.4 (2013): 581-582.

Gürsakal, Necmi. *Büyük Veri*. Bursa: Dora, 2013.

Huang, Lung. “Be Mine: Data-driven Valentine' Day Wishes.” *AdExchanger.* Web. 10 Nisan 2016.

Jeffery, Keith. “Data is the New Oil.” *Best Practices for Data Management & Sharing.* Dü. The Joint Research Centre (JRC). Ispra, Italy, 2014. Web. 13 Mart 2016.

Johnson, Jeanne E. “Big Data + Big Analytics = Big Opportunity.” *Financial Executive* July/August 2012: 50-53. Web. 12 Nisan 2016.

Laney, Doug. “3D Data Management: Controlling Data Volume, Velocity, and Variety.” Web. 30 Nisan 2016.

Manyika, James ve diğerleri. “Big data: The Next Frontier for Innovation, Competition and Productivity.” Web. 10 Mayıs 2016.

Monino, Jean-Louis ve Soraya Sedkaoui. *Big Data, Open Data and Data Development*. 3. London: ISTE Ltd, 2016.

Oğuzlar, Ayşe. *Temel Metin Madenciliği*. Bursa: Dora, 2011.

Ören, Tuncer, Tuncer Üney ve Rifat Çölkesen. *Türkiye Bilişim Ansiklopedisi*. İstanbul: Papatya, 2006.

Palmer, Michael. “Data is the New Oil.” Web. 10 Nisan 2016.

PriceStats. *History: PriceSstats*. Web. 5 Mart 2016.

Prytherch, Ray. *Harrod’s Librarians’ Glossary and Reference Book : A Dictionary of Over 10,200 Terms*. 10. Hampshire: Ashgate Publishing Limited, 2005.

Rotella, Perry. “Is Data The New Oil?” 2 April 2012. *Forbes.* Web. 4 Mayıs 2016.

Sankur, Bülent. *İngilizce - Türkçe Ansiklopedik Bilişim Sözlüğü*. İstanbul: Pusula, 2004.

SAP SE. *SAP News Center: SAP and the German Football Association (DFB) Turn Big Data Into Smart Decisions to Improve Player Performance at the World Cup in Brazil*. 11 June 2014. Web. 5 Nisan 2016.

Schönberger, Viktor Mayer ve Kenneth Cukier. *Büyük Veri - Yaşama, Çalışma ve Düşünme Şeklimizi Dönüştürecek Bir Devrim*. Çev. Banu Erol. İstanbul: Paloma, 2013.

Şeker, Şadi Evren. *İş Zekası ve Veri Madenciliği*. İstanbul: Cinius, 2013.

The White House, Executive Office of the President. "Big Data: Seizing Opportunities, Preserving Values." 1 May 2014. *The White House Web Site.* Web. 1 Nisan 2016.

Türkçe Bilim Terimleri Sözlüğü: Sosyal Bilimler. Ankara: TÜBA, 2011.

Uysal, Mithat. *Access 2003 ile Veri Tabanı Yönetimi*. 1. İstanbul: Beta, 2006.

Yılmaz, Malik. "Enformasyon ve Bilgi Kavramları Bağlamında Enformasyon Yönetimi ve Bilgi Yönetimi." *Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi* 49.1 (2009): 95-118.