AJIT-e: Online Academic Journal of Information Technology
2019 Kış/Winter – Cilt/Vol: 10 - Sayı/Num: 36
DOI: 10.5824/1309-1581.2019.1.004.x
http://www.ajit-e.org/?menu=pages&p=details_of_article&id=324

Received : 01.11.2018 *Editorial Process Begin: 26.01.2019* *Published: 15.02.2019*

Büyük Verinin Akademik Çalışmalarda Kullanımı Üzerine Mukayeseli Bir Veri Tabanı Araştırması[1]

Serkan Bayrakcı, Marmara Üniversitesi İletişim Fakültesi, Arş. Gör. *serkan.bayrakci@marmara.edu.tr, ORCID: 0000-0002-3817-1927*

Muhammed Akif Albayrak, Marmara Üniversitesi İletişim Fakültesi, Arş. Gör. *muhammed.albayrak@marmara.edu.tr, ORCID: 0000-0002-1946-1638*

ÖZ *Web 2.0'ın hayatımıza girmesi ve dijital teknolojiye dayanan ürünlerin kitlesel olarak üretilebilmesiyle birlikte farklı türdeki verilerin bir arada üretilmesi, depolanması ve paylaşılması mümkün hale gelmiştir. Dijital dünyada giderek artan devasa miktardaki veriye ve bu verinin analiz sürecine büyük veri denilmektedir. Bu çalışmanın amacı büyük veri mefhumuna kavramsal bir çerçeve çizerek, akademik çalışmalarda büyük veri kullanımının tarih içindeki gelişimini, büyük veri yılı olarak adlandırılan 2012 yılı öncesindeki ve sonrasındaki farklılıkları ortaya çıkartıp, mevcut durumu serimlemektir. Ayrıca büyük veri konusunda dünyada ve Türkiye'de yapılan çalışmaların birbirlerine paralellik gösterip göstermediklerini ortaya çıkarmak çalışmanın bir diğer amacını oluşturmaktadır. Çalışmada, yöntem olarak tarama araştırması benimsenmiştir. Bu bağlamda EbscoHost ASC ve Yükseköğretim Kurulu (YÖK) tez veri tabanlarında yapılan tarama araştırmasıyla birlikte; 2012 sonrasında büyük veriyle alakalı akademik çalışmaların 2012 öncesine oranla keskin bir artış gösterdiği gözlenmiştir. Bununla birlikte "veri madenciliği" ile ilgili akademik çalışmaların sayısındaki artışları nispi düşüşü ise veri madenciliği alt dallarının daha çok spesifikleşmesiyle açıklanabilmektedir. Ayrıca paralel işleme modellerinden "Map Reduce" ve doğal dil işleme uygulamalarından "fikir madenciliği" yöntemlerinin son yıllarda ivme kazandığı da gözlenmiştir. Bu durum Web 2.0'dan Web 3.0'a yani Etkileşimli Web'den Semantik Web'e geçiş sürecinde olduğumuzu belgeler niteliktedir.*

Anahtar Kelimeler: *Büyük Veri, Veri Madenciliği, Veri Analizi, Tarama Yöntemi*

[1] Bu makale araştırma kısmı daha sonra güncellenmek suretiyle Serkan Bayrakcı'nn "*Sosyal Bilimlerdeki Akademik Çalışmalarda Büyük Veri Kullanımı"* başlıklı yüksek lisans tezinden üretilmiştir.

AJIT-e: Online Academic Journal of Information Technology
2018 Kış/Winter – Cilt/Vol: 10 - Sayı/Num: 36
DOI: 10.5824/1309-1581.2019.1.004.x

A Comparative Database Survey on the Use of Big Data in Academic Studies[2]

ABSTRACT *With the introduction of Web 2.0 into our lives and mass production of digital technologies, it has become possible to produce, store and share different types of data together. The concept of big data has been used to define the increasing size of data, the increasing velocity at which it is produced and transmitted, the increasing variety of formats of these data and the analysis process of the data. The purpose of this study is to draw a conceptual framework for the concept of big data, to examine the use of big data in academic studies, its change in years and examine the change before and after the 2012 is called big data year. Another purpose of the study is to reveal that whether the studies in the world and Turkey about big data show parallels to each other or not. In this study, database survey research was adopted as a method. With the quantitative research that has been conducted in EBSCOhost ASC database designed by Academic Search™ Complete and Council of Higher Education (YÖK) thesis database to see the state of big data and big data analysis in academic studies and the results of these databases have compared to each other. As a result of the research, it is obvious that big data and analysis techniques have severely increased after 2012 compared to before 2012 in academic studies. Drop in the increase rate of academic studies about "data mining" can be explained with rise of the specification on sub-branches of data mining. Especially in recent years Map Reduce in artificial neural networks, sentiment analysis and naturel language processing methods gaining acceleration has been observed. This condition shows that we are in the way towards semantic web from interactive web, which is prospering to web 3.0 from web2.0.*

Keywords: *Big Data, Data Mining, Data Analysis, Database Survey Method*

Giriş

Kullanıcıların içerik üretebilmelerine, çeşitli formatlardaki içerikleri paylaşabilmelerine ve bunlara ek olarak birbirleriyle etkileşimde bulunabilmelerine imkân sağlayan Web 2.0'ın hayatımıza girmesiyle birlikte, üretilen veri miktarında ve veri çeşitliliğinde muazzam bir artış meydana gelmiştir. Özellikle sosyal medya, farklı formattaki verileri aynı anda barındırabilme olanağı sağladığından, veri miktarındaki ve veri çeşitliliğinde bu artış her geçen gün katlanarak devam etmektedir. Web 1.0 ile sadece metin biçimleri üretilebiliyorken günümüzde yazı, ses, fotoğraf ve video gibi farklı türdeki verilerin bir arada üretilmesi, depolanması ve paylaşılması sıradanlaşmıştır. Aynı zamanda dijital teknolojiye dayanan ürünlerin kitlesel olarak üretilebilmeleriyle birlikte kamera, sensör, telefon, uydu sistemleri gibi cihazlar yardımıyla çok büyük miktarlarda veri üretilir olmuştur. Sürekli artan ve büyük bir yığın haline gelen veri miktarı da veri analizini içerisinden çıkılması zor bir duruma sürüklemiştir. Web 2.0 ile ortaya çıkan bu veri yığını ve bu veri yığınının analiz süreci literatürde büyük veri olarak tanımlanmaktadır. Bilgisayar bilimlerinden finansa, ulaşımdan altyapıya, sağlık hizmetlerinden enerji sektörüne, savunma sanayisinden sosyal bilimlere kadar çok geniş bir yelpazede kendine kullanım alanı bulan büyük verinin analizi için yeni teknikler ve yazılımlar da geliştirilmektedir. Dolayısıyla, bilimsel çalışmaların araştırma

[2] The article, the research part is updated and produced by Serkan BAYRAKCI's master's thesis titled "Big Data Usage in Academic Studies in Social Sciences".

temalarına yeni bir sorun alanı dâhil olmuştur. Bu çalışmada da söz konusu minval dahilinde, birçok disiplinde öneminin arttığı ifade edilebilecek büyük veri mefhumuna kavramsal bir çerçeve çizilmeye çalışılmıştır. Yapılan literatür taramasıyla birlikte, akademik çalışmalarda büyük veri kullanımının tarih içindeki gelişimi ve mevcut durumu ortaya konulmak istenmiştir. Dünyanın en geniş akademik bilgi bankası olan EbscoHost tarafından tasarlanan Academic Search™ Complete (ASC) veri tabanı taranarak, büyük verinin ve analiz tekniklerinin akademik çalışmalarda kullanımı incelenmiştir. Ayrıca, Yükseköğretim Kurulu (YÖK) tez veri tabanında "büyük veri" kavramı taranarak, Türkiye'de büyük veri ile ilgili olarak yapılan akademik çalışmaların sonuçlarıyla, EbschoHost ASC veri tabanındaki akademik çalışmaların sonuçları mukayeseli biçimde değerlendirilmiştir.

Bir Kavram Olarak "Büyük Veri"

Büyük veri kavramı ilk olarak 1990'larda John Mashey tarafından, büyük veri ambarlarının yönetimi ve analizi olarak kullanılmıştır. Kavram ilk olarak bilgisayar biliminde Weiss ve Indurkhya tarafından, ekonometri ve istatistik bilimlerinde ise Diebold tarafından 2000 yılında güncel anlamıyla kullanılmıştır (Diebold, 2012). Büyük veri, 2008 yılına kadar hem akademik literatürde hem de bilişim sektöründe bilinirliği sınırlı kalmış bir kavram olarak karşımıza çıkmaktadır. 2008 yılının Haziran ayında Wired dergisinde yayınlanan Petabayt Çağı (The Petabyte Age) başlıklı yazıyla birlikte popüler olan büyük veri kavramı, bu yazıda "bilimi, tıbbı, işletme yönetimini ve teknolojiyi değişime uğratan devasa miktarda veriyi tutma, depolama ve anlama kabiliyeti" olarak ifade edilmektedir (Wired, 2008). Büyük veri, teknik anlamda bir değer elde edilebilecek komplike veri yığınının maksadına uygun araçlarla ve makul bir sürede analiz edilmek üzere işlenmesini ifade etmektedir. Ancak teknik anlamının dışında büyük veri, kamusal faydayı da bünyesinde barındırmaktadır (Kaya, 2017).

Büyük veri üretiminin geldiği devasa boyutu anlamak için aşağıdaki bilgilere bakmak faydalı olacaktır. WhatsApp'ta günlük ortalama 34 milyar mesaj gönderilmekte (Smith, 2017), Facebook'ta her dakika 510 bin yorum, 293 bin durum güncellemesi yapılmakta ve 136 bin fotoğraf yüklenmektedir (Schultz, 2017). Üretilen bu veriler sağlık alanından politikaya, reklamcılık sektöründen altyapı çalışmalarına kadar birçok alanda kullanılmaktadır. Büyük veri kavramı bilimsel çalışma alanlarında kendine çok hızlı bir şekilde yer bulmuştur. Farklı disiplinlerin çatıları altında farklı anlam içerikleriyle tanımlandığı ve kullanıldığı görülmektedir. Bu farklılığın temel sebebi teknolojik gelişmeler olarak görülse de tüm alanlarda ortak kabul gören büyük veriye ait özellikler; "hacim" "hız" ve "çeşitlilik" 'tir. Bunların yanında "doğruluk" ve "değer" unsurları da büyük verinin özellikleri arasında nitelendirilebilmektedir.

Sosyal Bir Olgu Olarak Büyük Verinin Unsurları

Büyük verinin farklı disiplinlerde farklı anlam ve özellikleri olsa da en çok kabul edilen özellikleri; veri hacmi (volume), veri hızı (velocity) ve veri çeşitliğidir (variety) ki; bunlar 3V ile ifade edilir (Laney, 2001). Çeşitli kaynaklarda, doğruluk (verification) ve değer (value) unsurlarının da bunlara ilave edildiğine ve büyük verinin boyutlarının 5V şeklinde ifadelendirildiğine rastlamak da mümkündür. Bu beş kavram için standart bir tanımlama bulunmamakla birlikte, çeşitli kaynaklarda büyük veri

DOI: 10.5824/1309-1581.2019.1.004.x

bileşenleri/unsurları/elementleri/özellikleri/boyutları gibi farlı isimlerle ifade edilebilmektedirler. Şekil 1'de büyük verinin bu beş boyutu gösterilmektedir.

Şekil 1: Büyük Veri Boyutları
Kaynak: Silva A. http://andressilvaa.tumblr.com/post/87206443764/big-data-refers-to-5vs-volume (Erişim: 12.11.2017)

Literatürde genel kabul gördüğü şekliyle beş V olarak ifadelendirilen; veri çeşitliği, veri hızı, veri hacmi, veri doğruluğu ve veri değeri şeklinde sınıflandırılan bu unsurları, daha detaylı olarak incelemek, araştırma konusunun daha iyi aydınlatılması bakımından faydalı bulunmaktadır.

a) Veri Çeşitliliği

Veri çeşitliliği, büyük verinin yapısındaki farklılığın ve zenginliğin ölçüsüdür. Veri yapılandırılmış, yapılandırılmamış ve yarı yapılandırılmış olarak; sayı, metin, resim, video, ses ve diğer farklı formatlarda bulunabilir. Büyük verinin bu çeşitli formatları aynı anda bulundurması, analizinde zorluklara sebep olmaktadır (Kaisler, Armour, Espinosa, & Money, 2003). Open Data Center Alliance (2012) 'nın raporuna göre, büyük veri öncesinde yapısal olmayan veri ya yok sayılırdı ya da en iyi ihtimalle verimsiz olarak kullanılırdı. Ama NoSQL yapısı kullanılarak tasarımlanan veri tabanlarındaki yapısal olmayan veriler, veri madenciliği

yöntemleri, Hadoop ve MapReduce gibi yeni tekniklerle yönetilebilir, işlenebilir ve analiz edilebilir hale getirilmiştir. Veri çeşitliliği kaynaklarını; sosyal medya paylaşımları, fotoğraf, ses ve video belgeleri, GSM operatörlerinden alınan bilgiler, hastane kayıtları, DNA diziliş belgeleri, iklim algılayıcıları, hava durumu sensörleri ve mobese kayıtları şeklinde sıralamak mümkündür (Işıklı, 2014).

b) Veri Hacmi

Büyük veri özelliklerinden ilk akla gelen, veri hacminin büyüklüğüdür. Büyük veri kavramındaki "büyük" ifadesi de aslında verinin hacminden gelmektedir (Zadrozny & Kodali, 2013). Web 2.0 ve sosyal medya ile birlikte günlük bazda üretilen ve işleme konulan veri miktarının artışı dikkat çekicidir. Birçok şirket dünyadaki enformasyon miktarını ölçerek dijital evreninin büyüklüğünü belirlemeye çalışmaktadır. Örneğin IDC şirketinin yapmış olduğu çalışmada, dijital evrende şu an yıllık ortalama 16,3 Zetabayt (1 ZB=1 trilyon GB) veri üretilirken, 2025 yılında bu miktarın en az 10 kat artacağı öngörülmektedir (Cave, 2017).

c) Veri Hızı

Büyük veriyi farklı kılan en önemli özellik veri üretiminin dinamik doğasıdır. Küçük veri genellikle belirli bir zaman ve mekânda yapılan sabit çerçeveli çalışmalardan oluşur. Boylamsal çalışmalarda veri, belirli zaman aralıklarıyla (her ay, her yıl gibi) elde edilmektedir. Ancak büyük veri ise gerçek zamanlı olarak ya da diğer bir deyişle anlık ve sürekli biçimde üretilir. Aralıklı bir veri akışından ziyade veri selinden elde edilir ve veri, hızıyla birlikteyken işlenir ve analiz edilir. Bundan dolayı veri yığınlarından akan veriye doğru kayan bir ilgi vardır (Zikopoulos, 2012).

Bu veriyi analiz eden şirketler kişilere konum tabanlı reklam ve mesaj göndermekte ve bu durum giderek yaygınlaşmaktadır. Verilerin anlık olarak izlenmesi, ölçümlenmesi ve analiz edilmesi pazarlama sektörü açısından hayati önem taşımaktadır. Çünkü günümüz rekabet ortamında kurumlar, veriyi ne kadar hızlı analiz edebilirlerse, kurumsal pazarlama stratejilerini de o kadar hızlı biçimde geliştirirler ve aynı zamanda hedef kitlelerini belirleyebilirler. Böylece kişilerin ihtiyaç duydukları bir ürünü/hizmeti anlık olarak belirleyerek, kişilere ilgili ürünleri/hizmetleri gerçek zamanlı olarak sunabilir ve pazar paylarını genişletebilirler.

d) Veri Doğruluğu

Büyük verinin doğruluğunun ölçüsünü belirleyen iki temel kriter bulunmaktadır. Bunların birincisi, büyük veriyi oluşturan kaynağın yüksek bir güvenilirliğe sahip olup olmadığıyla ilgiliyken, ikincisi de verinin hedef kitleye uygun olup olmadığına ilişkindir. Günümüzde doğruluk, mevzu bahis büyük veri olduğunda, temin edilmesi en güç değerlerden biridir. Büyük verinin doğruluğunu ve kalitesini etkileyen temel unsurlar, veri hacmi ve veri miktarıdır. Çünkü veri miktarı artışı ve veri kaynaklarının çeşitlenmesine koşut biçimde, verinin güvenilirliği ve kalitesi de azalmaktadır. Günümüzde çeşitli büyük veri analitikleri, bu tür durumlarla da baş edebilmek için sürekli gelişim halindedirler (Marr, 2014).

DOI: 10.5824/1309-1581.2019.1.004.x

e) Veri Değeri

Büyük verinin en önemli özelliklerinden bir diğeri de değerdir. Ruffatti (2013) 5V olarak nitelendirilen bu özelliklerin en anlamlısının değer olduğunu belirterek değeri "veriden anlam çıkarma" olarak ifade etmektedir. Elde edilen verinin içerisinde taşıdığı anlamı belirten veri değeri; büyük veriyi anlamlandırmaya olanak sağlamaktadır. Veri değeri unsuru, verinin içindeki soyut anlamı ortaya çıkararak, veriden fayda ve verim kazanılmasını sağlamaktadır (Ruffatti, 2013). Büyük verinin diğer tüm boyutlarının, aslında büyük verinin sahip olduğu değeri ortaya çıkarmak için hizmet ettiğini vurgulayan Swoyer (2012, s. 2); kurumların, büyük veriden fayda elde edebilmeleri ve veriyi kurumsal karar alma süreçlerine uygulayabilmeleri için, veri değerini ortaya koymak zorunda olduklarını ve büyük verinin özelliklerinden olan veri hacmi, veri hızı ve veri çeşitliliğinin esasen tek başlarına veride yatan değeri ortaya çıkarmada yetersiz kaldıklarını belirtmektedir. Büyük veri analizinde bu değeri ortaya çıkarmanın farklı yöntemleri bulunmakla birlikte; bu araştırma kapsamında, bu yöntem ve tekniklerden en sık tercih edilenler veri tabanlarında taranmış ve tarama sonuçları analiz edilmiştir.

Büyük Verinin Akademik Çalışmalarda Kullanımı Üzerine Mukayeseli Bir Veri Tabanı Araştırması

İnternet teknolojileriyle birlikte ortaya çıkan büyük veri, önemini her geçen gün biraz daha artırmakta ve verinin etki alanı günden güne genişlemeye devam etmektedir. Günümüzde büyük veri ile ilgilenen şirketler, verideki gizli bilgiyi keşfetmek, veriden öngörüler çıkarmak ve daha yerinde kararlar alabilmek için veri analizi yöntemlerini ve tekniklerini geliştirmek adına, hatırı sayılır miktarlarda sermaye yatırımlarında bulunmaktadır. Twitter'ın makine öğrenimi ve yapay zekâ firması olan Whetlab ile Magic Pony'yi satın almasının yanı sıra, Amazon'un bulut bilişim şirketi Sqrrl'i bünyesine katması büyük veriye yapılan yatırımlara birer örnek niteliğindedir. Gelişmekte olan bu alan hiç şüphesiz, akademi çevrelerinin de dikkatinden kaçmamış ve kısa zamanda akademik araştırmalarda yerini almıştır. Akademik çalışmalarda büyük verinin kullanımı üzerine odaklanan bu araştırmada, dünyanın en geniş akademik bilgi bankası EbscoHost tarafından tasarlanan Academic Search™Complete (ASC) veri tabanı taranmıştır.

Büyük veri analiz yöntem ve tekniklerinde en sık kullanılanların kavramsal düzeyde, EbscoHost ASC veri tabanında taratılması neticesinde edinilen bulgular mukayeseli bir değerlendirmeye tabi tutulmuştur. Buna ek olarak büyük verinin Türk akademyasındaki yerini görebilmek adına da YÖK'ün tez veri tabanı taranarak; başlıklarda, özetlerde ve dizinlerde "büyük veri" veya "big data" kavramları geçen yüksek lisans ve doktora tezleri sayıları ve bunların yıllara göre dağılımları tespit edilmiş, edinilen bulgular da bu minvalde yorumlanmıştır.

Araştırmanın Amacı

Araştırmanın amacı, akademik çalışmalarda büyük verinin kullanımını, literatür taraması sonucunda elde edilen bulgular üzerinden yorumlamaktır. Bu bakımdan, araştırmada cevabı aranan sorular şu şekildedir:

- Büyük veri ne zamandan beri kullanılmaktadır?
- Büyük veri analiz yöntem ve tekniklerinden hangileri, daha yaygın biçimde tercih edilmektedir?
- Büyük veri analiz yöntem ve tekniklerinde ortaya çıkan yeni gelişmeler nelerdir?
- Son yıllarda hangi teknikler daha fazla gelişim göstermektedir?
- Büyük veri yılı olarak adlandırılan 2012 öncesinde ve 2012 sonrasında, büyük veriye ve büyük verinin analizine yönelen akademik çalışmalarda kayda değer bir gelişme gözlenmekte midir? Eğer bir gelişme gözlenmekte ise; 2012 yılının öncesindeki durum ile sonrasında ortaya çıkan durum arasındaki farklılıklar ve benzerlikler nelerdir?
- Son yıllarda veri madenciliği tekniklerinde gözlemlenen spesifikleşme eğilimleri nelerdir?
- Web 3.0 teknolojilerinin geliştirilmesiyle, büyük veri analiz yöntem ve tekniklerindeki değişimler arasında nasıl bir korelasyon kurulabilir?
- Türkiye'de, büyük veriye ilişkin olarak yürütülen tezlerin durumu nedir ve Türk akademik yazınındaki büyük veri araştırmalarıyla uluslararası literatürde yer alan araştırmalar arasında bir paralellik görülebilmekte midir?

Araştırmanın Metodu

Bu çalışmada, araştırma metodu olarak tarama araştırması tercih edilmiştir. Araştırma kapsamında, literatürde büyük veri yılı olarak kabul edilen 2012 yılı baz alınarak, büyük veri ve büyük veri analiz yöntem ve teknikleri; 2012 öncesindeki yıllar ile 2012-2014 ve 2015-2017 aralıklarındaki üçer yıllık zaman dilimlerini kapsayan bir dönemsellikte olmak suretiyle, akademik çalışmaların başlıklarında/özetlerinde/anahtar kelimelerinde/metin içerisinde taranmıştır. Tarama sonuçları da 2012 yılı öncesinde ve sonrasında olacak şekilde bir sınıflandırmayla kaydedilip, mukayeseli olarak değerlendirilmiştir.

Çalışma kapsamında taranan EbscoHost ASC veri tabanı 7700'den fazlası hakemli olmak üzere, 9000'e yakın dergiyi tam metin olarak kullanıma sunmaktadır. Ayrıca, tüm akademik disiplinlerden, 13000'den fazla dergide yayınlanan makalelerin indeksleri ve özetleri yer almaktadır. ASC'de yer alan tam metin makaleler, 1887'ye kadar uzanmaktadır ve veri tabanı her gün güncellenmektedir. 1400'den fazla dergi için taranabilir atıf bilgileri yer almaktadır. Bu bilgi bankasının seçilme nedeni, dünyanın en geniş bilgi bankası olan Ebsco tarafından desteklenmesi, dünya üzerindeki en kapsamlı multi–disipliner tam metin bilgi bankası olmasıdır.

Veri tabanı taraması yapılırken, büyük veriyle ilgili kavramlar İngilizce olarak taranmıştır. Zira veri tabanında yer alan Türkçe çalışmaların sayısı oldukça azdır. Tarama esnasında kullanılan 20 adet İngilizce terimin Türkçe karşılığı ve kategorisi Tablo 1'de gösterilmiştir.

DOI: 10.5824/1309-1581.2019.1.004.x

Tablo 1: Araştırmada Taranan Kavramlar ve Türkçe Karşılıkları

Taranan İngilizce Kavramlar	**Türkçe Karşılıkları**	**Kategoriler**
"Big Data"	Büyük Veri	Büyük Veri
"Data Mining"	Veri Madenciliği	Veri Madenciliği
"Linear Discriminant Analysis"	Lineer Diskriminant Analizi	Sınıflandırma
"Decision Trees"	Karar Ağaçları	
"k-Nearest-Neighbor"	k-En Yakın Komşu Algoritması	
"Artificial Neural Networks"	Yapay Sinir Ağları	
"Support Vector Machine"	Destek Vektör Makinesi	
"Hierarchical Clustering"	Hiyerarşik Kümeleme	
"Partitioning Clustering"	Bölümleyici Kümeleme	Kümeleme
"Density-based Clustering"	Yoğunluk Temelli Algoritma	
"Grid-based Clustering"	Izgara Temelli Algoritma	
"Subspace Clustering"	Alt Uzay Arama Algoritma	
"Association Rules"	Birliktelik Kuralı	Birliktelik Kuralı
"Message Passing Interface(MPI)"	Mesaj Geçirme Arayüzü	Yapay Sinir Ağları
"MapReduce"	MapReduce	
"Dryad"	Dryad	
"Text Mining"	Metin Madenciliği	Metin Madenciliği
"Natural Language Processing"	Doğal Dil İşleme	Doğal Dil İşleme
"Sentiment Analysis or Opinion Mining"	Fikir Madenciliği	Fikir Madenciliği

Tarama sürecinde EbscoHost ASC veri tabanının gelişmiş arama seçeneği üzerinden gidilmiş ve taramalarda başlangıç noktası olarak veri tabanı otomatik ayarlarında kayıtlı olan 01.01.1963 yılı kabul edilmiştir. Ayrıca arama sonuçlarının yer aldığı tablolarda akademik çalışmaların ilk olarak hangi tarihte veri tabanına girdikleri de belirtilmiştir.

Araştırmada, 2012 yılı öncesi ile 2012 yılı sonrası (2012-2014 ve 2015-2017 tarih aralıklarında üçer yıllık periyotlar halinde bir dönemselleştirme yapılarak) iki ayrı sınıflandırma oluşturulmuştur. 2012 yılının referans noktası olarak alınmasının nedenleri; New York Times'ın 2012 yılının Şubat ayında "Büyük Veri Çağı" (The Age of Big Data) adıyla yayınlanan özel sayısı ile Dünya Ekonomik Forumu (World Economic Forum) tarafından yine 2012 yılında yayınlanan "Büyük Veri Büyük Etki" adındaki raporun yanı sıra, 2012 yılının büyük veri yılı olarak adlandırılması şeklinde sıralanabilir. Araştırma, 2012 yılı sonrası için 1 Ocak 2012 ile 31 Aralık 2017 tarih aralığını kapsayan bir dönemle sınırlandırılmıştır.

Araştırmanın Sınırlılıkları

Araştırma sürecinde Google'da "çevrimiçi akademik veri tabanı" şeklinde bir arama yapıldığında, ilk üç sırada EbschoHost, Jstore ve Oxford Journals veri tabanlarının yer aldığı görülmüştür. Araştırmaya Jstore ve Oxford Journal veri tabanları da dâhil edilmek istenmiş olmasına rağmen; Marmara Üniversitesi Kütüphanesinin Jstore veri tabanında sınırlı sayıda koleksiyona abone olması ve Oxford Journal veri tabanında büyük veriyle ilgili sınırlı sayıda akademik çalışmaya rastlanması gibi içeriksel nedenlerin yanı sıra, yine Oxford Journal veri tabanında "sadece başlıklarda arama", "sadece özette arama" ve "sadece metinde arama" gibi

özel arama seçeneklerinin bulunmayışı gibi teknik nedenler, araştırma sürecinin yalnızca EbscoHost ASC çevrimiçi veri tabanıyla sınırlı tutulmasını gerekli kılmıştır.

Veri analiz tekniklerinden en çok kullanılanlar bu çalışma bağlamında ele alındığı için, veri tabanı araştırmasında da sadece bu kavramların kullanılması, tüm veri analiz tekniklerinin dâhil edilememesi bu araştırmanın diğer sınırlılıklarındandır. Ayrıca, YÖK veri tabanında yer alan çalışmaların başlıklarında, özetlerinde ve dizinlerinde arama yapılabilirken, aramanın metnin içinde yapılamaması da araştırmanın diğer bir teknik sınırlılığıdır.

Araştırmanın Bulguları ve Değerlendirme

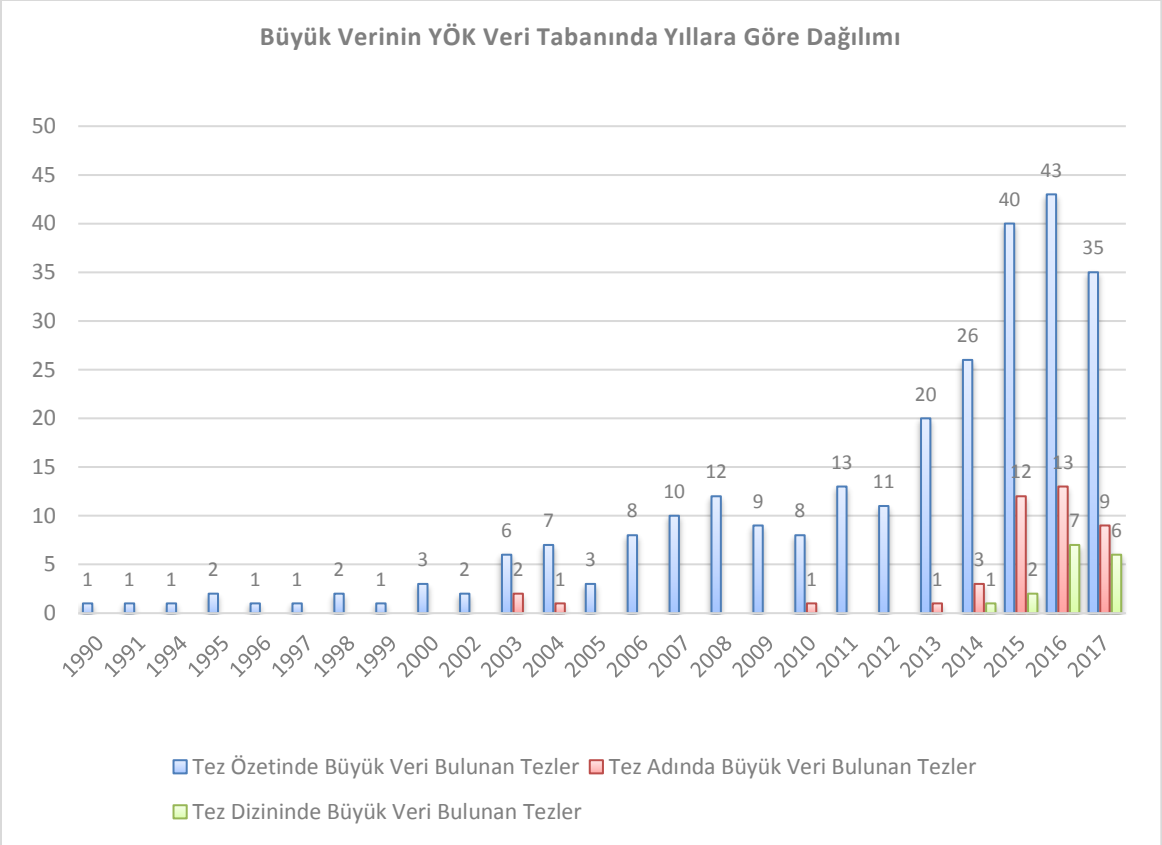
Daha önce de bahsedildiği üzere tarama araştırması, 2012 yılı sonrasını kapsayan tarih aralığının iki eşit dönemsellikte sınıflandırılması şeklinde gerçekleştirilmiştir. Bunun başlıca iki nedeni bulunmaktadır: Birincisi, (her yılın tabloda gösterilmesiyle) tablonun fazla karmaşık görünmesi bakımından, daha okunulabilir bir tablolaştırma için birtakım sadeleştirmelere gitme zorunluluğun ortaya çıkması gibi teknik bir nedenden kaynaklanıyorken; ikincisi de, tablonun daha işlevsel olması bakımından, 2012 yılı sonrasındaki dönemin artış trendini elden geldiğince net bir şekilde gösterebilmek için dönemselleştirmeye gitme zorunluluğun ortaya çıkması gibi fonksiyonel bir nedenden kaynaklanmaktadır.

Araştırma kapsamında, akademik çalışmaların başlıklarında, özetlerinde, anahtar kelimelerinde ve tam metin içinde taranmasıyla elde edilen bulguların değerlendirilmesi aşağıda verilmiştir.

Akademik Çalışmaların Başlıklarında Geçen Kavramlar Üzerine Yürütülen Taramalar Sonucunda Edinilen Bulguların Değerlendirmesi

YÖK tez veri tabanında büyük veri kavramına, başlığında yer veren ilk tez çalışması 2003'te, EbscoHost ASC veri tabanındaki ilk çalışma ise 1992'de yapılmıştır. YÖK tez veri tabanında büyük veriye başlığında, özetinde ve tez dizininde yer veren tez çalışmalarının yıllara göre dağılımları Grafik 1'de gösterilmiştir.

DOI: 10.5824/1309-1581.2019.1.004.x



Grafik 1: Başlığında Büyük Veri Kavramı Geçen Tezlerin Yıllara Göre Dağılımı

YÖK tez veri tabanında büyük veri kavramına tez özetlerinde ilk kez 1990 yılında rastlanmışken; tez başlıklarında 2003 yılında ve dizinlerinde ise 2014 yılında rastlanmıştır. 2012 yılına kadar başlığında, özetinde ve dizininde büyük veri kavramına yer veren tez sayılarında düzenli bir artış ya da azalış gözlenmezken; 2012 yılından itibaren üç bölümde de her yıl düzenli artışın gerçekleştiğini gözlenmiştir. Bir önceki yıla oranla 2017 yılında gerçekleşen kısmi azalışın nedeni ise; 2017 yılında YÖK veri tabanında tezlerin tamamının henüz yayınlanmamış olması olarak değerlendirilmektedir. Tez özetlerinde büyük veri kavramı kullanılan tez sayıları 2012 yılı baz alındığında; 2013 yılında %82, 2014 yılında %136, 2015 yılında %264 ve 2016 yılında %291 artış gösterdiği gözlenmiştir. Dolayısıyla, bilinirliği her geçen yıl artarak tez çalışmalarında kullanılmış olması bu durumun bundan sonraki yıllarda da devam edeceğinin sinyalini vermektedir.

YÖK tez veri tabanında büyük veri veya big data kavramına başlığında yer veren tezlerin türleri ve bunların hangi dilde kaleme alındıklarının yıllara göre dağılımları Tablo 2'de verilmiştir.

Tablo 2: YÖK Veri Tabanında Yer Alan Tezlerin Türlerinin ve Yazım Dillerinin Yıllara Göre Dağılımı

Tez Başlığında Büyük Veri Kavramı Bulunan Tezler		**Yazım Dili ve Frekansı**		**Toplam**
		İngilizce	**Türkçe**	
Yıl	**2003**	1 Yüksek Lisans	1 Yüksek Lisans	**2**
	2004	1 Yüksek Lisans	0	**1**
	2010	1 Yüksek Lisans	0	**1**
	2013	1 Yüksek Lisans	0	**1**
	2014	1 Doktora 1 Yüksek Lisans	1 Yüksek Lisans	**3**
	2015	6 Yüksek Lisans	1 Doktora 5 Yüksek Lisans	**12**
	2016	1 Yüksek Lisans	3 Doktora 9 Yüksek Lisans	**13**
	2017	1 Doktora 4 Yüksek Lisans	4 Yüksek Lisans	**9**
Toplam		**18**	**24**	**42**

2015 yılına kadar başlığında "büyük veri" veya "big data" ifadeleri yer alan tezlerin %75'inden fazlası İngilizce dilinde yazılmışken; sonrasında Türkçe tezlerde önemli bir artış gözlenmiş ve hatta 2017 yılında başlığında "büyük veri" veya "big data" ifadelerine yer veren tezler içerisinde İngilizce yazılanların oranı %42'lere düşmüştür. YÖK veri tabanında ilk olarak 2003 yılında İngilizce ve Türkçe dillerinde yazılmış birer yüksek lisans tezi, 2014 yılında ise ilk doktora tezi İngilizce dilinde yazılarak yer aldığı gözlenmiştir (Erişim 17.02.2018). Toplamda dördü Türkçe, ikisi de İngilizce olmak üzere altı doktora tezi ve yirmisi Türkçe, on altısı da İngilizce olmak üzere otuz altı tane yüksek lisans tezi yer almaktadır.

EbscoHost ASC veri tabanında yer alan çalışmaların başlıklarında yürütülen taramalarda edinilen bulgular Tablo 3'te gösterilmiştir. Tabloda 2012 sonrasındaki akademik çalışmaların toplam içindeki oranı ve ilgili kavramlara başlığında yer veren akademik çalışmaların veri tabanına ilk giriş yılı belirtilmiştir.

DOI: 10.5824/1309-1581.2019.1.004.x

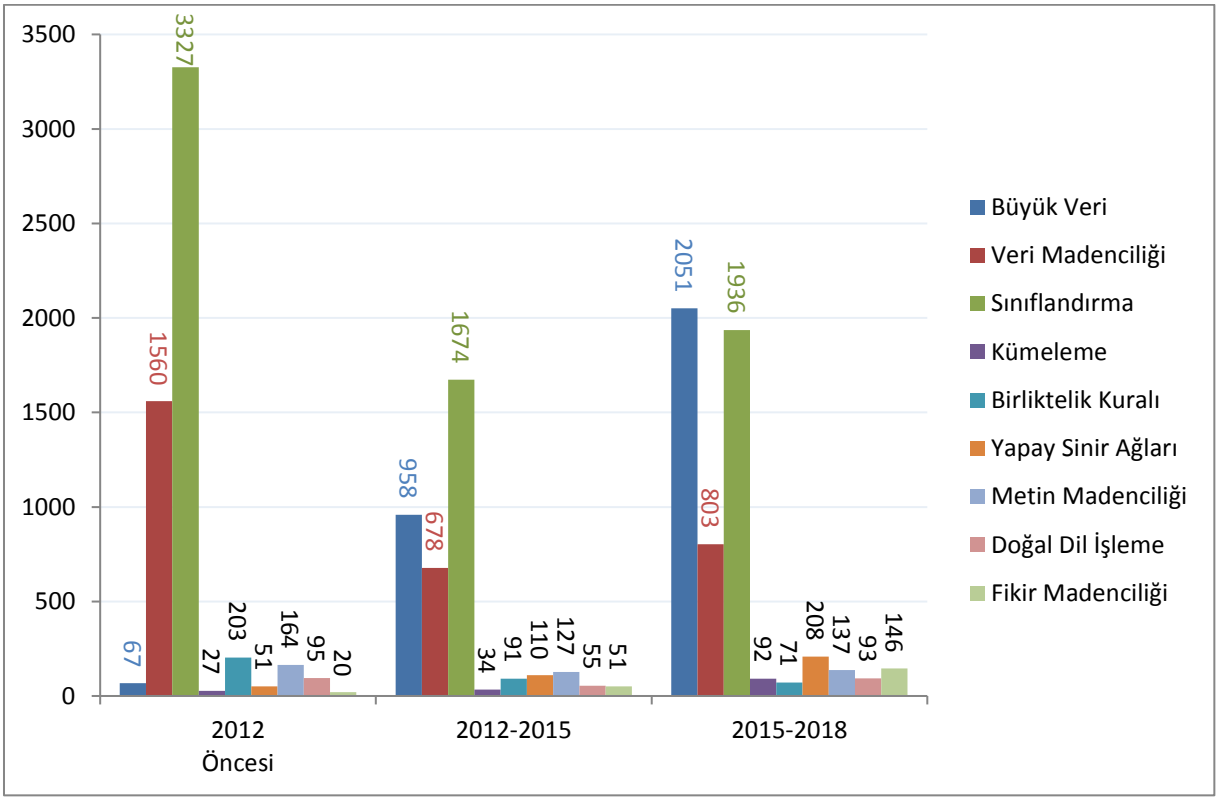
Tablo 3: Akademik Çalışmaların Başlıklarında Geçen Kavramlar ve Frekansları

Kavramlar	2012 Öncesi	2012-2014	2015-2017	2012 Sonrası	Toplam	2012 Yılı Sonrasının Toplam İçindeki Oranı (%)	Veri Tabanına İlk Giriş Yılı
Büyük Veri	67	958	2051	3009	3076	97,8	1992
Veri Madenciliği	1560	678	803	1481	3041	48,7	1994
Lineer Diskriminant Analizi	150	82	77	159	309	51,5	1983
Karar Ağaçları	252	89	99	188	440	42,7	1978
k-En Yakın Komşu Algoritması	87	56	99	155	242	64,0	1997
Yapay Sinir Ağları	1774	744	757	1501	3275	45,8	1990
Destek Vektör Makinesi	913	615	794	1409	2322	60,7	1997
Hiyerarşik Kümeleme	151	88	110	198	349	56,7	1979
Bölümleyici Kümeleme	1	1	1	2	3	66,7	2009
Yoğunluk Temelli Algoritma	10	8	30	38	48	79,2	1999
Izgara Temelli Algoritma	1	1	2	3	4	75,0	2010
Alt Uzay Arama Algoritma	15	24	59	83	98	84,7	2007
Birliktelik Kuralı	203	91	71	162	365	44,4	1996
MPI	17	7	2	9	26	34,6	2001
MapReduce	26	102	197	299	325	92,0	2007
Dryad	8	1	9	10	18	55,6	2009
Metin Madenciliği	164	127	137	264	428	61,7	1999
Doğal Dil İşleme	95	55	93	148	243	60,9	1986
Fikir Madenciliği	20	51	146	197	217	90,8	2008
Toplam Akademik Çalışma Sayısı	5514	3778	5537	9315	14829	62,8	

Yukarıdaki tabloda görüleceği üzere, büyük veri analiz tekniklerinin EbscoHost ASC veri tabanında 14.829 tane akademik çalışmanın başlığında kullanıldığı gözlenmiştir. 9.315 tanesi 2012 yılından sonra yazılan bu çalışmalar, toplamın %62,8'ini oluşturmaktadır. Eğer %62,8 oranı referans noktası olarak kabul edilirse; 2012 yılından sonra MapReduce %92 oranına, fikir madenciliği %90,8 oranına, alt uzay arama algoritması %84,7 oranına ve büyük verinin kendisi de %97,8 oranına ulaşarak, daha fazla akademik çalışmanın başlığında kullanıldığı gözlenmiştir.

Büyük veri analiz tekniklerinden özellikle fikir madenciliği ve paralel işleme modeli olan MapReduce kavramlarının, 2012 sonrasındaki ve 2012 öncesindeki akademik çalışmaların başlıklarında kullanım oranında rakamsal açıdan ciddi bir fark olduğunu kabul etmek gerekir. Fikir madenciliği başlıklı akademik çalışma sayısı 2012 yılı öncesinde toplam 20 iken, 2012-2014 tarih aralığında 51'e, 2015-2017 tarih aralığında ise 146'ya yükselmiştir. MapReduce 2012 yılı öncesinde toplam 26 iken, 2012-2014 tarih aralığında 102'ye, 2015-2017 tarih aralığında ise 197'ye yükselmiştir.

Büyük veri analiz tekniklerinin EbscoHost ASC veri tabanında yer alan akademik çalışmaların başlıklarında taranmasıyla ortaya çıkan durum Grafik 2'te gösterilmiştir.



Grafik 2: Akademik Çalışmaların Başlıklarında Geçen Kavramların Yıllara Göre Dağılımı

Taramanın yapıldığı EbscoHost ASC veri tabanında, büyük veri analizinde kullanılan yöntem ve tekniklerden karar ağaçları ve hiyerarşik kümeleme, ilk olarak sırasıyla 1978 ve 1979 yıllarında akademik çalışmaların başlıklarında yer aldığı tespit edilmiştir. Veri madenciliği kavramı ilk kez 1994'te, yapay sinir ağları da 1990'da akademik çalışmaların başlıklarında yer almıştır ve her iki kavram da büyük veriyle alakalı kavramlar içinde tüm yıllarda en çok kullanılan kavramlar olduğu tespit edilmiştir. Üç yıllık artış oranlarına göre kıyaslandığında; veri madenciliğinin ve yapay sinir ağlarının diğer kavramlar kadar hızlı artış göstermediklerini gözlenmiştir. Bu durumun nedeni; veri madenciliğinin de yapay sinir ağlarının da esasen çatı vaziyeti görmeleri ve bunların alt dallarının da akademik çalışmalarda daha fazla ilgi odağı haline gelmeleri olarak yorumlanabilir.

DOI: 10.5824/1309-1581.2019.1.004.x

Akademik Çalışmaların Özetlerinde Geçen Kavramlar Üzerine Yürütülen Taramalar Sonucunda Edinilen Bulguların Değerlendirmesi

EbscoHost ASC veri tabanında bulunan akademik çalışmaların özetlerinde yürütülen taramaların sonucunda elde edilmiş olan bulgular Tablo 4'te gösterilmiştir.

Tablo 4: Akademik Çalışmaların Özetlerinde Yapılan Tarama Sonuçları

Kavramlar	2012 Öncesi	2012-2014	2015-2017	2012 Sonrası	Toplam	2012 Yılı Sonrasının Toplam İçindeki Oranı (%)	Veri Tabanına İlk Giriş Yılı
Büyük Veri	56	1546	3798	5344	5400	98,9	1996
Veri Madenciliği	5127	2454	2972	5426	10553	51,4	1990
Lineer Diskriminant Analizi	1636	860	1182	2042	3678	55,5	1985
Karar Ağaçları	879	393	514	907	1786	50,8	1978
k-En Yakın Komşu Algoritması	586	495	683	1178	1764	66,8	1994
Yapay Sinir Ağları	5125	1687	1951	3638	8763	41,5	1988
Destek Vektör Makinesi	2236	2828	4595	7423	9659	76,9	1999
Hiyerarşik Kümeleme	1579	930	1191	2121	3700	57,3	1975
Bölümleyici Kümeleme	7	8	5	13	20	65,0	2007
Yoğunluk Temelli Algoritma	26	41	82	123	149	82,6	1998
Izgara Temelli Algoritma	3	5	5	10	13	76,9	2005
Alt Uzay Arama Algoritma	29	29	86	115	144	79,9	2002
Birliktelik Kuralı	557	238	245	483	1040	46,4	1963
MPI	373	147	193	340	713	47,7	1993
MapReduce	50	204	440	644	694	92,8	2007
Dryad	14	17	23	40	54	74,1	1982
Metin Madenciliği	606	401	554	955	1561	61,2	1998
Doğal Dil İşleme	671	328	616	944	1615	58,5	1965
Fikir Madenciliği	55	134	319	453	508	89,2	2003
Toplam Akademik Çalışma Sayısı	**19615**	**12745**	**19454**	**32199**	**51814**	**62,1**	

2012 yılı öncesinde ve sonrasında büyük veri analiz ve tekniklerine özetinde yer veren akademik çalışmaların sayısındaki artışlara bakıldığında, büyük veriden sonra en yüksek artışın MapReduce, fikir madenciliği ve kümeleme analizi algoritmalarından yoğunluk temelli algoritmalarda gerçekleştiği görülmektedir. Veri tabanında yer alan çalışmalardan ilk defa 2007 yılında çalışmanın özetine giren MapReduce, 2012 yılına kadar 50 çalışmanın özetinde kullanılmış, bu sayı 2012-2014 tarih aralığında 4 katına çıkmış, 2015-2017 tarih aralığında ise yaklaşık 9 katına kadar çıkmıştır. Özetinde büyük veri analizine ilişkin kavramlara yer veren çalışmaların %62,1'i 2012 yılı sonrasında yapılmıştır.

EbscoHost ASC veri tabanında yer alan akademik çalışmaların özet kısımlarında büyük veri analiz tekniklerinden veri madenciliği, yapay sinir ağları ve destek vektör makinesi terimleri en sık kullanılanlardır. Ancak çalışmaların ilk yayınlanma yılları farklıdır. Özetinde ilk kez; yapay sinir ağları, veri madenciliği ve destek vektör makinesi kavramlarını bulunduran akademik çalışmalar sırasıyla 1988, 1990 ve 1999 yıllarında yer almıştır. Ayrıca akademik çalışmaların özetlerinde birliktelik kuralı 1963'te ve doğal dil işleme 1965'te yer alarak bu kavramlardan en eski tarihli olanları olduğu gözlenmiştir.

Akademik Çalışmaların Anahtar Sözcüklerinde Yürütülen Taramalar Sonucunda Edinilen Bulguların Değerlendirmesi

Büyük veri analizi yöntem ve teknikleri EbscoHost ASC veri tabanında yer alan akademik çalışmaların anahtar sözcüklerinde taratıldığında; 2012 öncesinde büyük veri kavramına yer veren akademik çalışma sayısı yalnızca (2011 yılında yayınlanan) bir çalışmayla sınırlıyken, 2012-2014 tarih aralığında 383 tane çalışmaya rastlanmakta ve bu rakamın 2015-2017 tarih aralığında ise beş katına çıktığı görülmektedir. Öte yandan, akademik çalışmaların anahtar kelimelerinde kavramların veri tabanına ilk giriş yılları incelendiğinde; doğal dil işleme kavramı ilk kez 1969 yılında yer alarak, en eski tarihte veri tabanına giren kavram olarak dikkat çekmektedir.

EbscoHost ASC veri tabanında bulunan akademik çalışmaların anahtar kelimelerinde taratılmasıyla ortaya çıkan bulgular Tablo 5'te gösterilmiştir.

Tablo 5: Akademik Çalışmaların Anahtar Kelimelerinde Yürütülen Tarama Sonuçları

Kavramlar	2012 Öncesi	2012-2014	2015-2017	2012 Sonrası	Toplam	2012 Yılı Sonrasının Toplam İçindeki Oranı (%)	Veri Tabanına İlk Giriş Yılı
Büyük Veri	1	383	1849	2232	2233	100,0	2011
Veri Madenciliği	2612	1704	2056	3760	6372	59,0	1997
Lineer Diskriminant Analizi	623	317	399	716	1339	53,5	1982
Karar Ağaçları	354	199	304	503	857	58,7	1988
k-En Yakın Komşu Algoritması	181	149	198	347	528	65,7	2002
Yapay Sinir Ağları	2181	1079	1383	2462	4643	53,0	1995
Destek Vektör Makinesi	1408	1459	1990	3449	4857	71,0	2001
Hiyerarşik Kümeleme	258	148	271	419	677	61,9	1973
Bölümleyici Kümeleme	2	4	0	4	6	66,7	2004
Yoğunluk Temelli Algoritma	16	17	48	65	81	80,2	2005
Izgara Temelli Algoritma	5	4	3	7	12	58,3	2007
Alt Uzay Arama Algoritma	27	31	70	101	128	78,9	2003
Birliktelik Kuralı	272	136	138	274	546	50,2	2000
MPI	83	23	41	64	147	43,5	1998
MapReduce	20	129	297	426	446	95,5	2007
Dryad	1	3	2	5	6	83,3	2009
Metin Madenciliği	279	400	349	749	1028	72,9	2002
Doğal Dil İşleme	301	291	403	694	995	69,7	1969

DOI: 10.5824/1309-1581.2019.1.004.x

Fikir Madenciliği	41	110	256	366	407	89,9	2006
Toplam Akademik Çalışma Sayısı	**8665**	**6586**	**10057**	**16643**	**25308**	**65,8**	

EbscoHost ASC veri tabanında anahtar kelimelerin taratılmasıyla oluşan sonuçlar incelendiğinde, 2012 yılı öncesinde ve 2012 yılı sonrasında, MapReduce kavramını anahtar kelime olarak kullanan akademik çalışmalarda dikkate değer bir artış gözlenmiştir. 2012 öncesinde yalnızca 20 çalışmada yer alabilen kavramın 2012-2014 yılları arasında 129 ve 2015-2017 yılları arasında da 297 çalışmanın anahtar sözcükleri arasında yer aldığı tespit edilmiştir. Kavrama anahtar sözcükleri arasında ilk kez 2007 yılında yayınlanan bir çalışmada yer verilirken, MapReduce anahtar kelimeli çalışmaların sayısının %95'inin 2012 yılından sonra yayınlandığı görülmektedir. MapReduce kavramı yanında fikir madenciliği kavramını anahtar kelimelerinde kullanan akademik çalışmaların %90'ı 2012 yılından sonra yayınlandığı görülmektedir.

EbscoHost ASC veri tabanında yer alan çalışmaların anahtar sözcükleri arasına ilk kez 2002 yılında giren k-en yakın komşu algoritması; 2002-2012 yılları arasında 181, 2012-2014 yılları arasında 149 ve 2015-2017 yılları arasında da 198 çalışmada kullanılmak suretiyle toplam 528 çalışmanın anahtar sözcükleri arasında yer almıştır. Diğer yandan 2012 yılı öncesinde yılda ortalama 16 akademik çalışmanın anahtar kelimeler listesinde bulunan k-en yakın komşu algoritması, 2012 sonrasında yılda ortalama 57 tane akademik çalışmada yer alarak, diğer sınıflama tekniklerinden hatırı sayılır miktarda öne geçmiştir. Kavramın kullanımının, 2012 yılından sonra 3 kat yaygınlaşmış olduğunu da ayrıca ifade etmek gerekir.

Ayrıca, EbscoHost ASC veri tabanında anahtar kelimelerde küme analiz teknikleri incelendiğinde, 677 çalışmada hiyerarşik kümeleme kullanıldığı görülmektedir. Bu çalışmaların 258 tanesi 2012 yılı öncesine aitken, 419 tanesi de 2012 yılı sonrasında gerçekleşmiştir.

Akademik Çalışmaların Metin İçlerinde Yürütülen Taramalar Sonucunda Edinilen Bulguların Değerlendirmesi

Büyük veri analizi yöntem ve tekniklerini EbscoHost ASC veri tabanında yer alan akademik çalışmaların metinlerinde taratıldığında; büyük veri kavramına metin içinde yer veren akademik çalışma sayısı 2012 yılı öncesinde yalnızca 514 iken, bu rakamın 2012-2014 yılları arasında 12 katına çıktığı, 2015-2017 yılları arasında ise 30 katına çıkarak toplamda 22.081 rakamına ulaştığı görülmektedir. 2012 yılı sonrasında büyük veriyi metin içine alan akademik çalışma sayısının, tüm veri tabanında metin içinde büyük veri bulunan çalışmaların sayısına oranı ise %97,7'dir.

Büyük veri analiz tekniklerinin tamamı, EbscoHost ASC veri tabanında 201.486 tane çalışmanın metni içinde kullanılmıştır. Bu akademik çalışmaların 74.538 tanesi son üç yılda yayınlanmış ve toplam sayının üçte birini oluşturmaktadır. Eğer %62,6 oranı referans noktası olarak kabul edilirse, MapReduce, dryad, fikir madenciliği, yoğunluk temelli algoritma, alt uzay algoritması ve k-en yakın komşu algoritması'nın 2012 yılından sonra daha popüler hale geldiklerini ifade etmek mümkündür.

EbscoHost ASC veri tabanında yer alan çalışmaların metin içlerinde yürütülen taramalar sonucunda ortaya çıkan bulgular Tablo 6'da gösterilmiştir. Ayrıca 2012 sonrasındaki akademik çalışmaların toplam içindeki oranı ve veri tabanına ilk giriş yılları da verilmiştir.

Tablo 6: Akademik Çalışmaların Metin İçlerinde Yürütülen Tarama Sonuçları

Kavramlar	2012 Öncesi	2012-2014	2015-2017	2012 Sonrası	Toplam	2012 Yılı Sonrasının Toplam İçindeki Oranı (%)	Veri Tabanına İlk Giriş Yılı
Büyük Veri	514	6175	15392	21567	22081	97,7	1959
Veri Madenciliği	24462	11947	13976	25923	50385	51,4	1973
Lineer Diskriminant Analizi	3623	1796	2378	4174	7797	53,5	1973
Karar Ağaçları	5307	2389	2900	5289	10596	49,9	1969
k-En Yakın Komşu Algoritması	1772	1250	1823	3073	4845	63,4	1992
Yapay Sinir Ağları	11067	10992	12885	23877	34944	68,3	1968
Destek Vektör Makinesi	7971	5476	8513	13989	21960	63,7	1998
Hiyerarşik Kümeleme	9203	4578	6361	10939	20142	54,3	1971
Bölümleyici Kümeleme	91	37	35	72	163	44,2	1992
Yoğunluk Temelli Algoritma	144	117	236	353	497	71,0	1998
Izgara Temelli Algoritma	44	19	30	49	93	52,7	2003
Alt Uzay Arama Algoritma	135	103	205	308	443	69,5	1999
Birliktelik Kuralı	1713	712	724	1436	3149	45,6	1934
MPI	1555	466	458	924	2479	37,3	1993
MapReduce	145	640	1212	1852	1997	92,7	2005
Dryad	638	889	2083	2972	3610	82,3	1926
Metin Madenciliği	2614	1547	1892	3439	6053	56,8	1989
Doğal Dil İşleme	4176	1923	2586	4509	8685	51,9	1958
Fikir Madenciliği	270	448	849	1297	1567	82,8	1921
Toplam Akademik Çalışma Sayısı	**75444**	**51504**	**74538**	**126042**	**201486**	**62,6**	

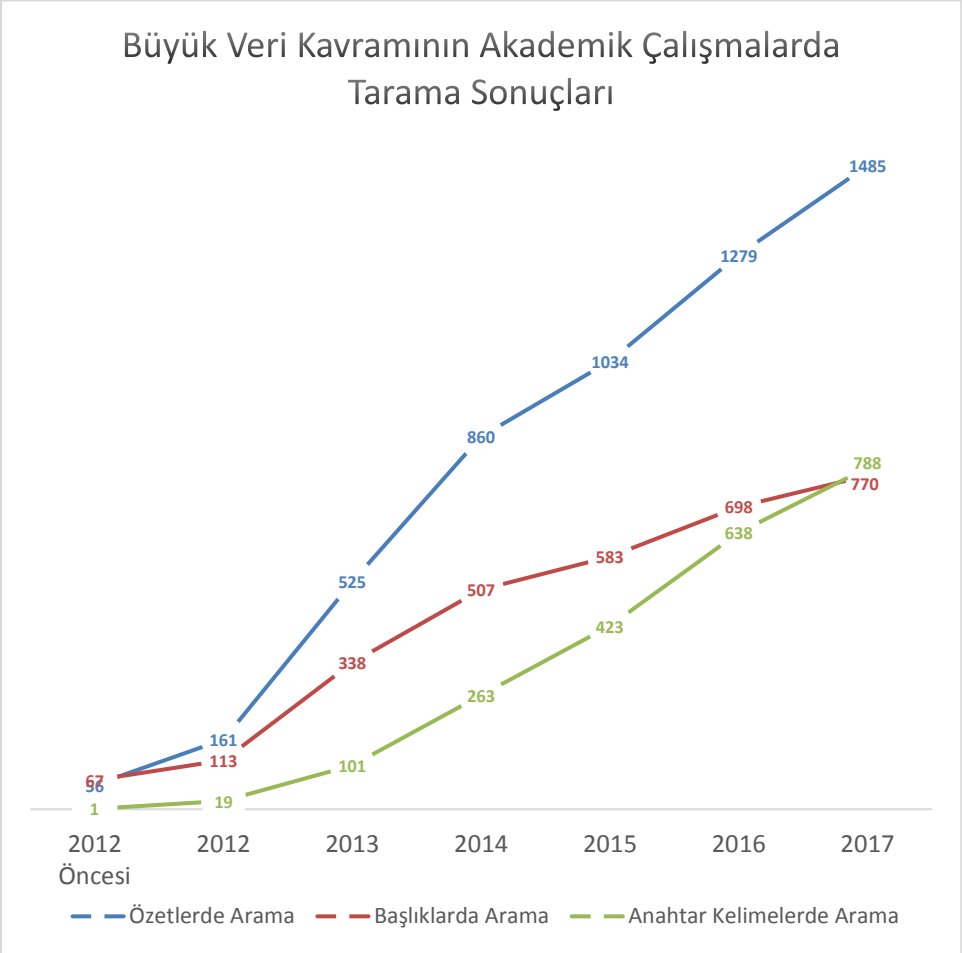
Fikir madenciliğinin üst başlığı konumunda bulunan ve dilin bilgisayarlar tarafından anlamlandırılmasına yönelik olan doğal dil işleme EbscoHost ASC veri tabanında 2012 yılı öncesinde 4.176, 2012 yılı sonrasında ise 4.509 tane olmak üzere toplamda 8.685 akademik çalışmanın metninde yer almıştır. Özellikle 2012 öncesinde 4.176 tane akademik çalışma metninde yer alan doğal dil işleme, fikir madenciliğini metninde bulunduran akademik çalışmalarının sayısının (270) yaklaşık 15 katı olduğu görülmüştür. 2012-2014 yılları arasında metninde doğal dil işleme kavramını bulunduran akademik çalışmaların sayısı (1923), fikir madenciliğini bulunduranların sayısının (448) yaklaşık beş katına ve son üç yılda ise bu oran üç katına gerilediği gözlenmiştir. Bu oranlar, dili anlamlandırmada 2012 yılı öncesinde daha çok doğal dil işlemenin, daha sonraki yıllarda da spesifik olarak fikir madenciliğinin kullanılmaya başlandığını göstermektedir.

Metin içinde, kümeleme analiz tekniklerinden hiyerarşik kümeleme, diğer kümeleme yöntemlerine nazaran daha fazla kullanılmıştır. Bölümleyici kümeleme toplamda 163,

DOI: 10.5824/1309-1581.2019.1.004.x

yoğunluk temelli algoritma 497, ızgara temelli algoritma 93 ve alt uzay algoritması toplamda 443 çalışmanın metninde kullanılmışken, hiyerarşik kümeleme toplamda 20.142 çalışma metni içinde kullanılmıştır. Son yıllardaki artış oranına bakıldığında ise yoğunluk temelli algoritma ile alt uzay algoritmasının daha fazla kullanıldığı görülmektedir.

EbscoHost ASC veri tabanında yer alan çalışmaların, başlıklarında, özetlerinde ve anahtar kelimeleri arasında büyük veri kavramının taratılması sonucunda edinilen rakamsal bulguların yıllara göre dağılımı Grafik 3'te gösterilmiştir.



Grafik 3: Büyük Veri Kavramının Akademik Çalışmalarda Tarama Sonuçları

EbscoHost ASC veri tabanında "büyük veri" kavramına başlığında yer veren akademik çalışma sayısı 2012'de 113 iken anahtar kelimelerde yer verilme sayısı 19'dur. Bu rakamlar her yıl artış göstererek 2017 yılında neredeyse eşitlenmiştir. Bu durum büyük veri kavramının başlıklarda yer almasa da anahtar kelimelerde daha fazla sayıda yer bulduğunu göstermektedir. Özetinde büyük veriye yer veren akademik çalışma sayısı da diğerleri gibi her yıl artarak devam etmiştir.

Sonuç

Teknolojinin gelişmesiyle birlikte dijitalleşme ve internet teknolojileri baş döndüren bir hızla büyümeye başlamıştır. Hem bireysel hem de toplumsal anlamda çeşitli dönüşümler

yaşanmaktadır. Bu dönüşümler sosyal bilimlerden fen bilimlerine kadar birçok alanı etkilemektedir. Bu etki alanının giderek artmasının nedeni ise teknolojik gelişmelerle birlikte ortaya çıkan büyük veri kavramıdır. Büyük veriyle, farklı formatlardaki, sürekli artış içinde olan devasa miktardaki veriyi ve bu verilerin analizi ifade edilmektedir.

Büyük verinin sahip olduğu temel özelliklerden biri olan "veri hacmi"; verinin miktarını, "veri hızı"; üretildiği anda yayılabileceğini, "veri çeşitliliği"; yapısında farklı formatlarda verileri barındırıyor olmasını, "veri doğruluğu"; büyük verinin güvenilirliğini ve "veri değeri" ise büyük verinin içerisinden çıkarılacak anlamı ifade etmektedir.

Araştırmanın yapıldığı EbscoHost ASC veri tabanında 1992-2007 yılları arasında başlığında on iki tane büyük veri ifadesini bulunduran akademik çalışma varken, sadece 2008 yılında on üç çalışma yapılmış olması dikkate değerdir. Bunun olası nedenlerinden birisi 2008 yılında Wired dergisinde yayınlanan Petabyte Çağı (The Petabyte Age) başlıklı yazıdır. Bu yazıda büyük veri kavramı; "bilimi, tıbbı, işletme yönetimini ve teknolojiyi değişime uğratan devasa miktarda veriyi tutma, depolama ve anlama kabiliyeti" olarak ifade edilmektedir (Wired, 2008). Bu ve benzeri bazı uluslararası yayınların büyük veri kavramına yer vermelerinin ardından, büyük veriye olan ilginin arttığını ve aynı zamanda hem teknik, hem akademik, hem de toplumsal düzlemde ilgi odağı olarak büyük verinin türlü veçheleriyle masaya yatırıldığı görülmektedir. Aynı şekilde, 2012 yılı Şubat ayında New York Times tarafından yayınlanan "Büyük Veri Çağı" (The Age of Big Data) başlıklı yazı, Dünya Ekonomik Forumu tarafından (World Economic Forum) yine 2012 yılında yayınlanan "Büyük Veri Büyük Etki" (Big Data Big Impact) başlıklı rapor ve 2012 Nisan ayının "Matematik, İstatistik ve Veri Seli İçin Farkındalık Ayı" olarak ilan edilmesi gibi nedenler, 2012 yılında büyük verinin ve büyük veriye ilişkin kavramların, akademik çalışmalardaki sayısında da ani artışa yol açmıştır.

2012 yılı öncesinde başlığında veri madenciliği kavramı bulunan akademik çalışma sayısının, 2012 yılı öncesindeki toplam akademik çalışma sayısına oranı %28 iken, 2012-2014 yılları arasında bu oran %18'e ve son üç yılda ise %14,5'e düşmüş ve bu durum çatı konumunda bulunan veri madenciliğinin konumunu değiştirmiştir. Yani 2012 sonrasında veri madenciliği altındaki metin madenciliği gibi çeşitli tekniklerle ilgili akademik çalışmaların, artık veri madenciliği ifadesini başlıklarında kullanmak yerine, spesifik tekniğin adını kullanmaya başladıkları görülmektedir. Veri madenciliği altındaki tekniklerle ilgili yapılan akademik çalışmaların artmasıyla birlikte, artık genel olarak veri madenciliği şemsiyesi değil; kullanılan tekniklerin her birinin zamanla kendi alt dallarını da içeren birer çatı olarak akademik çalışmalarda yer almaya başladıkları görülmektedir.

Paralel işleme modellerinden MapReduce 2004'te Google Inc. tarafından geliştirilerek, büyük veri analizinde en sık kullanılan açık kodlu yazılım çerçevesi olan Hadoop'un temel bileşeni olmuştur. EbscoHost ASC veri tabanında da MapReduce ilk defa 2007'de akademik bir çalışmanın konu başlığında yer almıştır. Bununla birlikte 2012 yılına kadar paralel işleme alanında en çok kullanılan model MPI iken, Google ve Hadoop'la birlikte MapReduce, MPI'ın önüne geçmeye başlamıştır. Her ne kadar mevcut durumda MPI daha ziyade akademik çalışmalarda kullanılmış olsa da, MapReduce ivme kazanmış bir halde akademide yer almaktadır ve gelecekte daha fazla çalışmada yer alacağı tahmin edilmektedir. Burada teknolojinin işlem hızını artırması nedeniyle mevcut imkânların daha da kolaylaştırılması söz konusudur. MapReduce, paralel işleme modeli olarak karşımıza çıkar. Bunun amacı, büyük miktarlardaki veriyi anlık olarak daha küçük eş işlemcilere ayırmak ve anında analiz etmektir.

DOI: 10.5824/1309-1581.2019.1.004.x

Özellikle pazarlamadaki rekabetçi ortamın, veriyi anlık olarak elde edip, analiz etmeyi ihtiyaç haline getirmiş olması temelde yapay sinir ağları eseri olan, MapReduce tekniğini her geçen gün daha cazip hale getirmektedir.

Dikkat çeken diğer bir durum ise 1972 yılında doğal dil işleme kavramının akademik bir çalışmanın özetinde yer almış olmasıdır. "MUSE: A Model To Understand Simple English" başlıklı ilgili çalışma, İngilizceyi anlamak için geliştiren bir modeli anlatmaktadır. Bu durum bilgisayarların insan dilini anlaması için yapılan çalışmaların, esasen bilgisayarın icadından itibaren devam ettiğini belgeler niteliktedir. Doğal dil işleme uygulamalarından en çok bilineni; fikir madenciliği ilk kez 2008 yılında bir akademik çalışmanın başlığında yer almış, fakat özellikle 2012 yılından sonra akademik akademik çalışmalarda kullanımının hızlı bir artış gösterdiği de gözlenmiştir. Cho'ya (2008) göre Web 2.0'ın devamı niteliğindeki yeni nesil web; diğer deyişle akıllı web olarak ifade edilen Web 3.0'ün gelişimi de bu anlamda değerlendirilmelidir. Çünkü Web 3.0, semantik web, doğal dil işleme, veri madenciliği ve yapay zekâ gibi teknolojileri kullanarak makinelerin anlamasını sağlamaktadır. Daha üretken ve sezgisel bir kullanıcı deneyimi sağlayan Web 3.0 kullanıcıya göre şekillenebilme özelliklerine sahiptir. Web 3.0 teknolojilerini ilk kullanan şirketlerden biri olan Nova Spivack's Twine, 2010 ile 2020 arasında web'in semantik web olacağını ifade etmektedir. 2010 yılında Apple tarafından geliştirilen Siri, kişisel akıllı asistan olarak doğal dil işlemeye ve semantik web'e dair verilebilecek en belirgin örnektir. Yeni nesil internet dönemi 2010'dan itibaren başlamıştır ve bu durum akademik çalışmalarda da doğal dil işleme ve fikir madenciliği kavramlarının kullanımının artmasına neden olmuştur.

EbscoHost ASC veri tabanında başlık, özet, anahtar kelime ve tüm metin içinde yapılan taramalar sonucunda büyük veri, k-en yakın komşu algoritması, yoğunluk temelli algoritma, ızgara temelli algoritma, MapReduce, metin madenciliği ve fikir madenciliği teknikleri genel olarak akademik çalışmaların tüm bölümlerinde 2012 sonrasında dikkate değer şekilde artış göstermiştir. Tüm teknikler içinde "algoritma" ifadesi geçen bu dört tekniğin 2012 sonrasında gösterdiği artış dikkate alındığında, büyük veri analizinin bilgisayar ve matematik bilimlerinin kesişiminde konumlandığını ifade etmek son derece mümkündür. Bunun bilgisayar kısmını yazılım ve donanım oluştururken, matematik kısmını ise istatistik teknikleri ve mantık oluşturmaktadır.

2015-2017 dönemindeki yıllık ortalamanın diğer tüm dönemler içindeki toplam payına bakıldığında ise: büyük verinin %72, Dryad'ın %69, fikir madenciliğinin %65, MapReduce'ın %64, yoğunluk temelli algoritmanın %62 ve alt uzay arama algoritmasının %61 oranına ulaşarak son yıllarda çok daha popüler çalışma konuları arasına girdikleri gözlemlenmektedir. Hatta MPI'ın %39, bölümleyici kümelemenin %42, birliktelik kuralının ve ızgara temelli algoritmanın ise %48 oranlarında olup geriye kalan tüm kavramların %50'nin üzerinde olduğu dikkat çekmektedir.

Ayrıca büyük veri analizinde, 2012 yılı sonrasında özellikle dil ve anlam üzerine yürütülen çalışmaların önem kazanmış olduğu, metin madenciliği ve fikir madenciliği kavramlarının akademik çalışmalarda daha sık kullanılmasından anlaşılmaktadır. Bu durumun iki temel nedeni bulunmaktadır. Bunlardan ilki Web 3.0'ın ortaya çıkması ve gelişmesiyken; diğeri ise temelde dijital pazarlama yöntemlerinin gelişmiş olmasıdır. Çünkü günümüzdeki rekabet

ortamında kurumlar; markaları, ürünleri ve kendileri hakkında internette neler konuşulduğunu takip etme ihtiyacı hissetmektedirler. Şirketler, sosyal medya kullanıcılarının ve müşterilerinin fikirlerini öğrenmek ve onların ihtiyaçları doğrultusunda doğru zamanda ve doğru ürünü onlara sunarak, onların ürünlerini satın almalarını sağlamak ve kâr elde etmek zorundadırlar. Bunun temel yolu da kişilerin; ne dedikleri, ne yedikleri, nerelere gittikleri, ne giydikleri, ne dinledikleri vb. eylemlerini takip etmek, ölçümlemek ve analiz etmektir. Bu yüzden kullanıcı içeriklerinin daha hızlı analiz edilmesi ve anlık olarak ölçümlenmesi, pazarlama stratejileri açısından çok önemlidir. Bunun yanında, risk yönetiminde ve müşteri merkezli sonuçlar elde etmede büyük veriden öngörüler oluşturmak için, dilbilim ve anlambilimle ilgili olan metin madenciliği ve fikir madenciliği gibi tekniklerin akademik çalışmalarda da son yıllarda giderek önem kazandığı gözlemlenmektedir.

Kaynakça

Cave, A. (2017, Nisan 13). *What Will We Do When The World's Data Hits 163 Zettabytes In 2025?* 1 18, 2018 tarihinde Forbes.com: https://www.forbes.com/sites/andrewcave/2017/04/13/what-will-we-do-when-the-worlds-data-hits-163-zettabytes-in-2025/#a630829349ab adresinden alındı.

Cho, A. (2008, Temmuz 22). *What is Web 3.0?* suite.io: https://suite.io/allan-cho/wy92cm1 adresinden alındı.

Diebold, F. X. (2012, Ağustos). *A Personal Perspective on the Origin(s) and Development of "Big Data":The Phenomenon, the Term, and the Discipline.* Şubat 3, 2015 tarihinde http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf adresinden alındı.

Işıklı, Ş. (2014). "Büyük Veri, Epistemoloji ve Etik Tartışmalar." *Online Academic Journal of Information Technology* 5.14, 7–13.

Kaisler, S., Armour, F., Espinosa, J., & Money, W. (2003). Big Data: issues and challenges moving forward. *Proceedings of the 46th IEEE Annual Hawaii international Conference on System Sciences (HICC 2013)*, (s. 995-1004). Grand Wailea, Maui, Hawaii.

Kaya, M. B. (2017). Büyük Verinin Hukuki Boyutları. Ş. Sağıroğlu, & O. Koç, *Büyük Veri Ve Açık Veri Analitiği : Yöntemler ve Uygulamalar* (s. 181-192). Ankara: Grafiker Yayınları.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety.

Marr, B. (2014, Mart 6). *Big Data: The 5 Vs Everyone Must Know*. Temmuz 20, 2014 tarihinde Linkedin: https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know adresinden alındı.

Open Data Center Alliance. (2012). Open Data Center Alliance: Big Data Consumer Guide. Open Data Center Alliance. Mart 12, 2015 tarihinde http://www.opendatacenteralliance.org/docs/Big_Data_Consumer_Guide_Rev1.0.pdf adresinden alındı.

Ruffatti, G. (2013, Mart 7). Value is the most meaningful V for Big Data. SpagoWorld: http://blog.spagoworld.org/2013/03/value-is-the-most-meaningful-v-for-big-data/ adresinden alındı.

Schultz, J. (2017, Ekim 10). *blog.microfocus.com*. Aralık 15, 2017 tarihinde How Much Data is Created on the Internet Each Day?: https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/ adresinden alındı.

Silva A., A. (2017, Aralık 11). My Favorite Infographics and Stats. Enjoy it!!! Big Data Refers to 5V's - Volume,Variety,Velocity,Value and Veracity: http://andressilvaa.tumblr.com/post/87206443764/big-data-refers-to-5vs-volume adresinden alındı.

Smith, C. (2017, Kasım). *65 Amazing WhatsApp Statistics and Facts*. Aralık 15, 2017 tarihinde Dmr: https://expandedramblings.com/index.php/whatsapp-statistics/2/ adresinden alındı

Swoyer, S. (2012, Temuz 24). Big Data -- Why the 3Vs Just Don't Make Sense--TDWI. Temmuz 19, 2015 tarihinde tdwi.org: http://tdwi.org/Articles/2012/07/24/Big-Data4th-V.aspx?Page=1 adresinden alındı

Weiss, S., & N., I. (1998). *Predictive Data Mining: A Practical Guide.* Morgan Kaufmann Publishers, Inc.

Wired. (2008, Haziran 23). *Wired*. Şubat 2, 2015 tarihinde http://www.wired.com/: http://archive.wired.com/science/discoveries/magazine/16-07/pb_intro adresinden alındı

Zadrozny, P., & Kodali, R. (2013). *Big Data Analytics Using Splunk: Deriving Operational Intelligence from Social Media, Machine Data, Existing Data Warehouses, and Other Real-Time Streaming Sources.* Apress.

Zikopoulos, P. C. (2012). *Understanding big data.* New York et al: McGraw-Hill.