

Kamuda Büyük Veri ve Uygulamaları

Doç.Dr.İzzet Gökhan ÖZBİLGİN¹

gozbilgin@thk.edu.tr

Özet: Son yıllarda özellikle Kamu kurumlarında çeşitli analiz ve istatistikler oluşturmak amacı ile saklanan veri miktarları hızla artmaktadır. Ayrıca Kamu kurumları sadece kendi bünyelerindeki veriyi değil, sosyal medya kullanıcılarının verilerini de gerektiğinde karar verme süreçlerine dahil etme ihtiyacındadır. Bu nedenle günümüzde hacim olarak geleneksel donanım çözümleri ile saklanamayacak kadar çok fazla yer tutan, kurumların hizmet kalitesini artırmalarına yönelik süreçlerde kullanılan Büyük Veri kavramı giderek yaygınlaşmaktadır. Bu çalışmanın amacı, büyük verinin tanımını yapmak, nasıl saklanıp, ne şekilde analiz edilebileceğini ortaya koymak ve hangi alanlarda kurumlara fayda sağlayabileceğini örneklerle açıklamaktır.

Anahtar Sözcükler: Büyük Veri, e-Devlet, HADOP, NonSQL

Abstract:

The amount of data that stored especially in public organizations has been rapidly increasing in recent years in order to create statistics and business intelligence analysis. In addition, government administrations need not only their structured data but also the data of social media users to process decision -making analysis. Therefore Big Data has been wide used for any collection of so large and complex data which process by used non-traditional hardware solutions to improve the quality of service. The aim of this study is to make Big Data definition, explain how it is stored, and detailed case studies.

Keywords: Big Data, E-Government, HADOP, NonSQL

¹ Türk Hava Kurumu Üniversitesi, Bilgisayar Mühendisliği Bölümü, Ankara

1. Giriş

Günümüzde daha fazla bilgiye dayalı karar alma ihtiyacı hem özel sektörde hem de kamuda bilgi işlem bölümlerinin sorumluluklarını arttırmaktadır. Eskiden sadece merkez veya taşra teşkilatının sisteme girdiği anlık işlem verilerinin saklanması ve basit olarak raporlanmasından sorumlu olan bilgi işlem birimleri karar verme sürecinde aktif rol alan analizleri gerçekleştirmekle yükümlü hale geldiler. Geleneksel donanım ve yazılım çözümleriyle saklanması efektif olmayan, alışlagelmiş analiz araçlarının incelemede yetersiz kaldığı büyük hacimdeki verilerin saklanması ve analizi için yeni teknolojiler ortaya çıkmaktadır. Bu teknolojiler ile büyük veriyi hem uygun maliyetlerle, hem de hızlı erişilebilir şekilde saklamak ve yönetmek mümkün olmaktadır. Bu çalışmanın kapsamında, büyük veri kavramı açıklanmış, örnek dünya uygulamaları araştırılmış ve ülkemizin büyük veri konusuna bakış açısı incelenmiştir.

2. Büyük Veri

Büyük Veri 21. yy hemen başında sahneye çıkmıştır. Bu verinin ilk aktif kullanıcıları ise yeni nesil internet şirketleridir. Bunların başında Google, eBay, LinkedIn ve Facebook bulunmaktadır. Bu firmalar tamamıyla büyük veri üzerine kurulmuş sistemlerde çalışmakta ve diğer firmalar gibi geleneksel IT altyapılarına sahip değildirler. Bu tip firmalar haricindeki kurumlar ise bugüne kadar tamamıyla geleneksel bilgi teknolojisi üzerinde çalışmıştır. Öncelikle büyük verinin ne olduğunu ortaya koymak gereklidir.

Büyük veri hali hazırda kullanılan veritabanı yönetim sistemleri veya geleneksel veri işleme uygulamaları ile işlenemeyen büyük ve karmaşık veri kümelerine verilen addir. Bu zorlukların içinde verinin yakalanması, saklanması, aranması, paylaşılması, transferi, analizi ve görselleştirilmesi bulunmaktadır.

Gartner'a göre büyük veri; maliyet-etkin ve yaratıcı süreçlerle gelişmiş öngörü, karar verme ve işlem otomasyonu sağlayacak yüksek hacimli, yüksek akış hızlı ve çok çeşitli veri kaynaklarıdır.

Colin White'a göre "büyük veri" maliyet veya teknoloji sınırları nedeniyle daha önceden desteklenemeyen veri yönetimi ve iş yükleridir. Bu konuda üç teknolojik unsurla büyük verinin çalışabilir hale geldiğini ifade etmektedir:

- Analitik ilişkisel veritabanı yönetim sistemleri (HADOP)
- İlişkisel olmayan ("NoSQL") dosya ve veritabanı sistemleri
- "Stream" işleme sistemleri

Büyük verinin öne çıkan 3 ana özelliği mevcuttur. Bunların birincisi verinin büyüklüğüdür. Tahmini olarak 2020 yılında 40 zetabyte veri oluşturulmuş olacaktır. Şu anda dünya üzerinde 6 milyar mobil telefon bulunmakta ve veri üretmektedir. İkinci öne çıkan özellik ise verinin akış hızıdır. New York ticaret borsasında her seansta 1 TB veri oluşmaktadır. Modern araçların üzerinde yaklaşık 100 civarı sensör bulunmakta ve devamlı veri üretmektedir. Üçüncü özellik ise verinin çeşitliliğidir. Twitter üzerinde günlük 400 milyon tweet atılırken, Facebook üzerinde aylık 30 milyar içerik paylaşılakta, Youtube üzerinde aylık 4 milyar saat video seyredilmektedir.

Büyük verinin özelliklerinden biri verinin hacminin yüksek olmasıdır. Birçok faktör veri hacmini yükseltmektedir. Uzun yıllar boyunca sistemlerde tutulan veri sadece işlem bazlı veriydi. Artık sosyal medyada devamlı artan yapısal olmayan veri, sensör kullanımının artması ve makineler arası iletişim verisi (M2M) toplanmaya başlanmıştır. Geçmişte büyük hacimli veriler saklama problemi yaratırdı. Saklama

maliyetlerinin düşmesiyle artık büyük veri kümeleri üzerinde nasıl analiz yapılacağı problem haline gelmiştir.

Veri bugüne kadar benzeri görülmemiş bir hızda artmaya başlamıştır ve akan bu veriyi zamanında yakalamak gerekmektedir. RFID etiketleri, sensörler ve akıllı ölçüm cihazlarından sel gibi akan verilerin gerçek zamanlıya yakın saklanıp işlenebilmesi sorunu ortaya çıkmıştır. Veri akış hızına zamanında cevap verebilmek birçok organizasyonun problemleri arasındadır.

Veri günümüzde çok çeşitli biçimlerde ortaya çıkmaktadır. Geleneksel veritabanlarında bulunan yapısal ve numerik veriler, yapısal olmayan metinler, e-mail'ler, video, ses, hisse senedi verileri, finansal işlemler bunların bir kısmıdır. Bu kadar çeşitli verinin yönetimi, birleştirilmesi ve analiz edilebilmesi birçok kurumun şu anda en çok uğraştığı süreçtir.

Bunun yanında büyük veriye ilişkin iki özellikten daha bahsedilebilir. Bunlardan biri verinin akış hızının ve çeşitliliğinin çok hızlı değişebilmesidir. Bir diğeri ise bu veri yapılarının gerçekten karmaşık olmasıdır. Verinin kullanılabilir ve anlaşılabilir hale getirilmesi işlemsel veri için harcanan emekten daha fazladır.

Büyük verinin bir kısmı eski verilerin şeklinin değişmesiyle, bir kısmı ise tamamen yeni kaynaklardan türemektedir. Genel olarak veri kaynaklarına bakılacak olursa aşağıdaki liste ciddi bir yer tutmaktadır.

Medya: Dijital olarak kayıt altına alınan medya dosyaları (video, resim ve ses) bir veri kaynağı olarak karşımıza çıkmaktadır. Yardım masası ses kayıtları, şehir içi- kurum içi kameralar bunlara örnek sayılabilir.

Sağlık: Sağlık sektöründe elektronik görüntülemeler, dijital test sonuçları resim ve benzeri biçimlerde sistemlerde tutulmaktadır.

Sensör verisi: GPS cihazları, RFID etiketleri, akıllı ölçüm cihazlarının ürettikleri veriler bu kapsama girmektedir.

Uygulama ve İnternet Logları: Kurumların sahip olduğu uygulama ve internet sitelerinin ürettiği izleme kayıtları son yıllarda hızla artmaktadır.

Doküman: Özellikle vatandaş ile direkt temas halinde olan kurumlarda ciddi miktarda doküman birikebilmektedir. Bu dokümanlar da büyük veri kapsamında değerlendirilebilir. Dokümanlar arasında "Word", "Excel", "PDF", "e-mail"ler ve "text" dosyalar sayılabilir.

Sosyal Medya: Twitter, Facebook gibi yoğun kullanılan platformlardan genel veya kurumun kendisini ilgilendiren verilerdir.

3. Kamuda Büyük Veri

Kamuda Büyük Veri için kamusal fayda sağlayacak bilgi çıkarımı yapılabilecek her türlü saklanabilen veya saklanamayan ölçekte, karmaşık yapıda, yapısal olmayan veriye denir. Kamuda genel olarak büyük veri ihtiyacı; vatandaşlarımızın güvenliğini garantilemek, sağlığını korumak, konforunu arttırmak, refahını sağlamak ve geleceğini garanti altına almak içindir.

Kamuda büyük veri kaynaklarına aşağıdaki örnekler verilmiştir:

- Vatandaşların yarattığı tüm veriler (Kamu'dan hizmet alımı, ikamet bilgileri, nüfus bilgileri, seçim/referandum sonuçları, eğitim/tecrübe bilgileri, resmi beyanlar, telekomünikasyon kanalları üzerinden iletişim, seyahat/ulaşım, harcama, vergilendirme vs.)
- Kamu Kurumların yarattığı tüm veriler (sağlanan hizmetler, çalışan bilgileri, bütçesel bilgiler, ülke kaynakları/fiziksel durumu, makroekonomik veriler vs.)

- Diğer ülke vatandaşlarının/kurumlarının yarattığı veriler (seyahat/ulaşım, telekomünikasyon kanalları üzerinden iletişim, ticari faaliyetler, sağlanan/alınan hizmetler, politik-siyasi-askeri aktiviteler vs.)
- Toplanan uydu fotoğraflaması, kamera sistemleri, havadan görüntüleme (LiDAR), uzaktan algılama, CDR (konuşma kayıtları)

4. Büyük Veri Teknolojileri

Veri ambarlarının olgunlaştığı günümüzde, süreçler ve iş akışları kurumsal veri manzarasını değiştirmeye başladı. 20 yıl önce çok az uygulama veri kaydetmekteydi. Zamanımızdaysa hemen hemen tüm işlemler bilgisayarlar ile desteklenmektedir. Her iş adımının verisi artık elde edilebilir durumdadır. Bu veriler arasında işlemlerin (transaction) yanı sıra makine sensör verileri ve insanların konuşmaları mevcuttur.

Kurumlar bu verileri toplamakta ancak bu veriler dağıtık durmakta ve kullanılmamaktadır. Bu noktada Apache Hadoop, MapReduce, NoSQL ve grafik işleme motorları gibi altyapılar yeni yetenekler sunmaktadır.

4.1. HADOP

Hadoop, basit sunucuların oluşturduğu bir küme üzerinde büyük verileri işlemek amaçlı uygulamaları çalıştıran bir platformdur. Hadoop Distributed File System (HDFS) olarak adlandırılan bir dağıtık dosya sistemi ve MapReduce özelliklerini bir araya getiren, Java ile geliştirilmiş açık kaynaklı bir altyapıdır. Hadoop, HDFS ve MapReduce bileşenlerinden oluşan bir yazılımdır. HDFS sayesinde sıradan sunucuların diskleri bir araya getirilerek büyük, tek bir sanal disk altyapısı oluşturulur. Bu sayede değişken boyutlu bir çok dosya bu sistemde saklanabilir. Bu dosyalar bloklar halinde

(genelde 64MB) birden fazla ve farklı sunucu üzerine (genelde 3 kopya) dağıtılarak RAID benzeri bir yapıda yedeklenir. Bu sayede veri kaybı önlenmiş olur. Ayrıca HDFS çok büyük boyutlu dosyalar üzerinde okuma işlemi (streaming) imkanı sağlar, ancak rastlantısal erişim (random access) özelliği bulunmaz.

4.2. NoSQL Veritabanları

NoSQL veritabanları günümüzde kullanılan ilişkisel veritabanlarına alternatif olarak değişen iş ihtiyaçlarına yönelik yazılmış sistemlerdir. Bu tip veritabanlarının kullanım nedeni tasarım basitliği, yatay büyüme ve erişilebilirlikte daha fazla kontrol olarak sayılabilir. Veri yapıları ilişkisel veritabanlarından farklı olduğu için veritabanı içinde gerçekleştirilen bir kısım işlem daha hızlı olmaktadır. NoSQL veritabanları ilişkisel veritabanları içinde çözmenin zor olduğu problemleri adreslemektedir. Büyük veri ve gerçek zamanlı web uygulamaları sayesinde bu veritabanlarının kullanımı artmaktadır.

4.3. Stream İşleme Sistemleri

Veri analiz yöntemleri ağırlıklı olarak “batch” veya “mini-batch” olarak adlandırılan daha çok sabit veriyle yapılan analizlerdir. Akan veri üzerinde analizler gerçekleştirilebilmek için bu sürece özel farklı bir altyapı gerekmektedir. “Stream Processing”, “Complex Event Processing (CEP)” gibi isimlerle anılan bu sistemlerin ortak özelliği daha işlem veya etkileşim gerçekleşirken yapılan analizler ve karar verme yeteneğidir. Bu tip ürünlerin genel amacı akan veri ve sistemlerdeki ilgili verileri o an içinde analiz ederek yapılan işlemin fayda zarar hesabının yapılması, fırsat veya tehlikelerin ortaya çıkartılmasıdır.

5. Dünyada Kamu Büyük Veri Uygulamaları

5.1. Trafik Yoğunluğu Takibi Projesi [1]

Sektör: Ulaştırma

İçerik: Hollanda İstatistik Bürosu, tüm yollarda bulunan sensörlerden gelen verileri toplayarak yolların kullanım oranlarını ortaya çıkartmıştır. Sensörler önünden geçen her aracın tipini (araba, kamyon ve benzeri) ve hızını algılayarak merkezi sisteme bildirmektedir. Yapılan bu çalışmayla ulaşımda alınması gereken önlemler ortaya çıkmaktadır. Projede toplanan veri miktarının yüksekliği sebebiyle büyük veri teknolojileri tercih edilmiştir.

5.2. Sosyal Medya Analizi [1]

Sektör: İletişim

İçerik: Hollanda İstatistik Bürosu, ülke halkının %70'nin kullandığı Twitter ve benzeri sosyal medya sitelerinden topladığı verilerle halkın genel olarak ne üzerine konuştuğunu analiz etmiştir. Bunun yanısıra duygu analizi yaparak genel olarak halkın mutluluk düzeyini ortaya koymuştur. Bunun yanında ayı analizlerde ekonomik durum ve benzeri konular üzerinde halkın düşüncesi meydana çıkmıştır.

Proje: Prematüre Bebek Takibi [2]

Sektör: Sağlık

İçerik: Ontario Üniversitesi her gün, erken doğan bebeklerden (prematüre) yaklaşık 100 milyon adet veri toplayarak, analizini en hızlı şekilde gerçekleştiriyor. Bunun sonucunda, hasta muayenesi sırasında erken teşhis edilen değişimler, bir hastalık durumuyla ilişkilendirilebiliyor.

5.3. Akıllı Şebeke Analizi [3]

Sektör: Enerji

İçerik: Tennessee Valley Authority, sayısı 1.5 trilyon olan akıllı şebeke verilerinin analizi için bir sistem geliştirdi. Sonuç olarak kurum, güç şebekesi arızaları üzerine yapılan analizler ile verimliliği arttırmaktadır. Doğal kaynakları koruyan üst düzey analizlerle tahminlemeler gerçekleştiriyor.

5.4. Görüntüleme Tani Hatalarının Azaltılması [4]

Sektör: Sağlık

İçerik: Asya Sağlık Bürosu, hasta görüntüleme verilerini Hadoop üzerinde tutup analiz ederek radyoloji ve patoloji uzmanlarının hem daha hızlı hem de daha az hata yaparak teşhis koymalarını sağlamıştır.

5.5. Suç Önleme Projesi [5]

Sektör: Güvenlik

İçerik: New York Polisi 911 kayıtlarını, yakalamaları, suçlu bilgilerini ve coğrafi verileri gerçek zamanlı olarak işleyerek günler sürebilen analizleri dakikalar içinde tamamlayarak suç oranını azaltmaya başlamıştır.

5.6. Su Kaynaklarının Takibi [6]

Sektör: Çevre

İçerik: Beacon Enstitüsü, Hudson Körfezi'ne yerleştirdiği sensörlerle topladığı biyolojik, fiziksel ve kimyasal verileri meteorolojik verilerle birleştirerek araştırmacı, kamu ve eğitimcilerle sunmaktadır. Toplanıp analiz edilen bu verilerle olası çevre felaketleri ve anlık değişimlerin daha hızlı fark edilmesi sağlanıyor.

5.7. Suç Önleme [7]

Sektör: Güvenlik

İçerik: Amerika'da "Önleyici Polis Hizmetleri" olarak adlandırılan ve Seattle, Los Angeles gibi şehirlerde uygulanan yapılandırmalar 4 aylık bir süreçte cinayet oranını yüzde 12 gibi bir miktarda düşürmüştür. Yüzde 26 gibi bir düşüş ise hırsızlık üzerinde gerçekleşmiştir. Vancouver polis bölümünün benzer uygulaması bir hizmet, suçun nereye yönlendiğini göstermiş, hatta birçok durumda engellenmezse gerçekleşeceği durumlarda sonlanmasını sağlamıştır. Mülki suçlar şehir genelinde 1000 yerleşimde %24 oranında düşmüş, şiddetli suç oranlarında 2007 senesinden 2011 e kadar %9 azalmıştır. İşlenen büyük veri; sadece suç içeren verilere sahip olmakla birlikte, eğilim yönünü ve suçun açığa çıkabileceği bölgeyi tam olarak bulabilmiştir. Seneler içinde gerçekleşen suçların haritalandırılması ve suçlu hareketlerinin takibi sonucunda polis, oluşabilecek suçlara karşı bir öngörüye kavuşmuştur.

6. Türkiye’de Kamu Büyük Veri Uygulamaları

Kamu sektörü büyük veri konusunda araştırmalara başlamıştır. Öncü bir kısım bakanlık ve kamu kurumları şu anda büyük verinin sağladığı imkanlar ile neler yapılabileceğini görmeye çalışmaktadır. Şu an için aktif olarak büyük veri üzerine kurgulanmış bir kamu projesi bulunmamaktadır. Ancak Kalkınma Bakanlığı 2014-2018 Onuncu Kalkınma Planı 412. maddesinde “Açık kaynak kodlu yazılımlar, büyük veri, bulut bilişim, yeşil bilişim, mobil platform, nesnelerin interneti gibi ürün, hizmet ve yönelimler değerlendirilerek kamu için uygun olabilecek çözümler hayata geçirilecektir.” hedefi yazılmıştır.

Ayrıca Kalkınma Bakanlığı 2014-2018 Bilgi Toplumu Stratejisi ve Eylem Planı Taslağı 50. Eylemi “Kamuda Büyük Veri Pilot Uygulaması Gerçekleştirilmesi” 2014-2016 yıllarında tamamlanması planlanmıştır. Bu eylem ile büyük verinin ekonomik değere dönüşmesi sağlanması ve sosyal güvenlik, sağlık, vergi, güvenlik gibi alanlar başta olmak üzere kamuda büyük veri uygulamalarının geliştirilmesi hedeflenmiştir. Ayrıca bu eylemin hayata geçirilmesiyle kamu kurum ve kuruluşları tarafından büyük veri alanında pilot uygulamalar hayata geçirilecek ve kamu verisi kullanılarak sosyal güvenlik alanında büyük veri pilot uygulamasının geliştirilmesi ve başarı örneklerinin oluşturulması bu teknolojilerin Türkiye’de yaygınlaştırılmasına öncülük edecektir. Sosyal Güvenlik Kurumu (SGK) sorumlu kurum olarak tespit edilmiş olup TUBİTAK ve TÜRKSAT işbirliği yapılacak kuruluşlar olarak yazılmıştır.

SGK, toplamış olduğu büyük miktardaki yapılandırılmış ve yapılandırılmamış veriler üzerinde çeşitli analizler yaparak verimliliği artırmak, kayıp-kaçak oranlarını düşürmek ve hizmet kalitesini yükseltmek için büyük veri konusunda çalışmalara başlamıştır.

50. Eylemin uygulama adımları aşağıdaki şekilde belirlenmiştir:

- Sosyal güvenlik alanında büyük veri uygulamaları belirlenecektir.
- Büyük veri uygulama alanları ile ilgili fayda ve maliyet analizleri yapılacaktır.
- Fayda ve maliyet analizine göre öncelikli alanlar ve gereksinimler belirlenecektir.

Büyük verinin kamudaki kullanım alanları hakkında ilgili kamu çalışanlarının eğitilmesi ve büyük veri alanındaki bilinç artırılacaktır.

7. Sonuç:

Kalkınma Bakanlığı 2014-2018 Bilgi Toplumu Stratejisi ve Eylem Planı Taslağında büyük verinin ekonomik değere dönüşmesi sağlanması ve sosyal güvenlik, sağlık, vergi, güvenlik gibi alanlar başta olmak üzere kamuda büyük veri uygulamalarının geliştirilmesi hedeflenmiştir. Büyük veri çalışmalarının kamu kurumlarında geliştirilmesi gerekmektedir.

Büyük veri konusundaki etkinliklerin artırılması ve Türkiye özelinde iyi uygulama örnekleri oluşturularak bu örnekler üzerinden kamu kurum yöneticilerine, büyük veri çalışmalarının önemi ve sağlayacağı faydalar anlatılmalıdır.

Birlikte çalışabilirlik ve verilerin kaynağının tespiti, kamu genelinde büyük çapta gerçekleştirilecek büyük veri çalışmalarının anahtarını oluşturmaktadır. Kamu genelinde birlikte çalışabilirlik çalışması tamamlanmalı ve bu veri kaynaklarının, büyük veri çalışması açısından nasıl değerlendirilebileceği kurumlar arası bir organizasyon ile tartışılmalı ve kurumlar arası büyük veri projeleri başlatılmalıdır.

Büyük veri alanında kamu-özel sektör-üniversite işbirlikleri, yenilikçilik ve Ar-Ge destekleri yetersizdir. Bu desteklerin artırılması yüksek maliyetler gerektiren Büyük veri çalışmaları için teşvik edici bir rol oynayacaktır.

Büyük veri kaynaklarının etkin bir şekilde birlikte kullanılmasını düzenleyecek fakat bireylerin mahremiyetini koruyacak şekilde denge kuracak bir düzenleme bulunmamaktadır. Bu açıdan çalışmaların mevzuat ile ilgili sorunları gerekli yasa ve yönetmeliklerle giderilmelidir.

8. Kaynaklar

(İnternet Siteleri-Son Erişim 10 Aralık 2015)

[1]http://www.unece.org/fileadmin/DAM/stat/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf

[2] <http://www.cnbc.com/id/101032950>

[3]<http://blog.cloudera.com/blog/2009/06/smart-grid-hadoop-tennessee-valley-authority-tva/>

[4] [https://www-950.ibm.com/events/wwe/grp/grp004.nsf/vLookupPDFs/SriSrinivasan'sPresentation/\\$file/SriSrinivasan'sPresentation.pdf](https://www-950.ibm.com/events/wwe/grp/grp004.nsf/vLookupPDFs/SriSrinivasan'sPresentation/$file/SriSrinivasan'sPresentation.pdf)

[5] <http://www.dissentmagazine.org/blog/the-incident-state-coercion-in-the-age-of-big-data>

[6] <http://www.bire.org/>

[7]http://www.huffingtonpost.com/2012/07/01/predictive-policing-technology-los-angeles_n_1641276.html

