# FET316-BİLGİSAYAR MÜHENDİSLİĞİNDE İLERİ KONULAR

# 3ncü DERS – BÜYÜK VERİ (BIG DATA)

## GİRİŞ

Büyük veri (Big Data), dijitalleşen dünyada hemen her sektörde önemli bir rol oynamaktadır. Çeşitli veri kaynaklarından toplanan devasa miktarlardaki verilerin, geleneksel veri işleme araçlarıyla analiz edilmesi ve yönetilmesi oldukça zorlayıcıdır. Ancak, büyük verinin etkin bir şekilde işlenmesi, organizasyonların daha iyi kararlar almasına, verimliliklerini artırmasına ve yeni iş fırsatları yaratmasına olanak tanımaktadır. Bu ders notu, Büyük Veri'nin temel kavramlarından başlayarak, veri analizi, depolama, işleme ve güvenlik gibi konuları derinlemesine inceleyecektir.

## BÜYÜK VERİ VE TEMEL ÖZELLİKLERİ

Büyük veri, çok büyük boyutlardaki, yüksek hızda üretilen ve çeşitli biçimlerde (yapısal, yarı yapısal, yapısal olmayan) olan verilerin genel adıdır. Bu veriler, klasik veri işleme araçlarıyla analiz edilemeyecek kadar karmaşık ve hacimli olup, basit anlamda "3V" (Volume, Velocity, Variety) modeliyle tanımlanır. Ancak biz 5V formülü ile tanımlanmış olan dikkate alacağız.

- **1. Hacim (Volume):** Büyük veri yalnızca veri miktarıyla değil, aynı zamanda bu verilerin nasıl hızla büyüdüğüyle de ilgilidir. Örnek olarak: Google Search: Google, her gün yaklaşık 3.5 milyar arama yapıldığını hesaplıyor. Bu aramalar her gün terabaytlarca veri üretir. Google, bu veriyi anlamak ve kullanıcıya daha iyi hizmet verebilmek için büyük veri teknolojilerinden yararlanır. Instagram: Her gün 95 milyon fotoğraf ve video yükleniyor. Bu da verinin boyutunun hızlı bir şekilde büyümesine neden oluyor.
- **2. Hız (Velocity):** Veri, hızla üretilip işlenmek zorundadır. Bu, hızlı işlem gücü gerektiren bir süreçtir. Örnekler: Twitter: Her dakika 500.000'den fazla tweet atılmaktadır. Bu verinin hızlı bir şekilde işlenmesi, duygu analizi ve trend takibi gibi işlemler için çok önemlidir. Finansal Piyasalar: Yatırımcılar, borsada saniyeler içinde gerçekleşen ticaretleri takip etmek zorundadır. Anlık piyasa verileri ve algoritmik ticaret stratejileri, yüksek hızda veri işleme gerektirir.

- **3. Çeşitlilik (Variety):** Büyük veri yalnızca yapılandırılmış verileri değil, aynı zamanda yapılandırılmamış ve yarı yapılandırılmış verileri de içerir. Örnekler: E-ticaret Siteleri: Kullanıcıların arama geçmişi, satın alma geçmişi, ürün yorumları, anket sonuçları ve müşteri destek talepleri gibi çok çeşitli veriler bir araya gelir. Video İçeriği: YouTube gibi platformlarda her dakika 500 saatlik video yükleniyor. Bu videolar, metin, görsel, ses ve etkileşim verilerinin birleşimidir.
- **4. Doğruluk (Veracity):** Verinin doğruluğu, güvenilir bir analiz yapabilmek için kritik önemdedir. Örnekler: Sağlık Verileri: Bir hastanın sağlık durumu hakkında toplanan veriler bazen hatalı veya eksik olabilir. Örneğin, hastaların yanlışlıkla sağlıklı oldukları veya ölümle sonuçlanmamış bir hastalığı olduğu kaydedilebilir. Trafik Verisi: Akıllı şehirlerde kullanılan trafik sensörleri, hatalı veriler üretebilir. Bir yol çalışması nedeniyle oluşan trafik, sistemin yanlış raporlama yapmasına yol açabilir.
- **5. Değer (Value):** Büyük veriden elde edilen bilgi, stratejik kararlar almak için kullanılmalıdır. Örnekler: Netflix: Netflix, kullanıcıların izleme geçmişlerini analiz ederek içerik önerileri sunar. Bu, kullanıcıların tercihleri hakkında değerli bilgiler yaratır ve şirketin kişiselleştirilmiş öneri sistemini güçlendirir. Havayolu Şirketleri: Havayolu şirketleri, müşteri rezervasyon bilgilerini, hava durumu verilerini ve uçuş geçmişlerini analiz ederek, uçuş iptalleri veya gecikmeleri konusunda tahminler yapabilir ve buna göre önlemler alabilir.

Bu 5 temel özelliği sayesinde büyük veri, geleneksel veri yönetim sistemlerinin ötesinde yeni analiz ve işleme yöntemlerini gerektirir.

## BÜYÜK VERİNİN ÖZELLİKLERİ

Büyük verinin en belirgin özellikleri arasında yüksek hacim, çeşitlilik ve hız yer alsa da, bu veri türünün daha fazla karakteristiği de bulunmaktadır. Bunlar:

- **Doğruluk:** Büyük verinin işlenmesi sırasında, verinin doğruluğu da önemli bir faktördür. Verinin yanlış işlenmesi veya eksik olması, yanlış sonuçlara yol açabilir.
- **Zaman Duyarlılığı:** Gerçek zamanlı veri analizleri gerektiren uygulamalarda, verinin ne kadar hızlı analiz edilebileceği, sürecin başarısını doğrudan etkiler. Örneğin, finansal piyasalar ve sağlık alanındaki kritik anlık kararlar için veri analizi büyük önem taşır.
- **Değişkenlik:** Veriler sürekli olarak değişir. Bu dinamik yapıyı yönetmek için veritabanı ve analiz sistemlerinin esnek ve güncel olması gerekmektedir.

- **Bütünsellik:** Büyük verinin bir araya getirilmesi, tüm veri kümelerinin tutarlı ve bütünsel bir şekilde analiz edilmesini gerektirir. Veri setlerinin farklı kaynaklardan gelmesi ve çeşitli formatlarda olması, entegrasyon süreçlerini zorlaştırır.

## BÜYÜK VERİNİN KAYNAKLARI

- **a. Sosyal Medya:** Facebook: Her dakika, Facebook kullanıcıları "beğeni" ve "yorum" yaparak büyük bir veri akışı yaratır. Yıllık veriler, kullanıcının etkileşimini analiz etmek için kullanılabilir. TikTok: Her gün milyonlarca video, etkileşim ve yorum oluşturulur. TikTok'un algoritması, kullanıcıların izleme alışkanlıklarını takip ederek kişiselleştirilmiş içerikler sunar.
- **b. Sensörler ve IoT Cihazları:** Otomobiller: Akıllı araçlar, sürüş verilerini, hız bilgilerini, motor performansını ve yakıt tüketimini sürekli olarak toplar. Bu veriler, araç bakımı ve sürüş alışkanlıklarını analiz etmek için kullanılabilir. Akıllı Ev Sistemleri: Akıllı termostatlar (örneğin, Nest), kullanıcıların sıcaklık tercihlerine göre veriler toplar ve enerji tasarrufu sağlamak amacıyla bunları analiz eder.
- **c. E-Ticaret:** Amazon: Amazon, kullanıcıların alışveriş davranışlarını, beğenilerini, arama geçmişini ve satın alma geçmişini toplar. Bu veriler, kullanıcıyı daha iyi anlamak ve tavsiyelerde bulunmak için kullanılır. Alibaba: Alibaba, Çin'deki e-ticaret devlerinden biri olarak, her gün milyonlarca ürün satın alım verisi toplar ve bu verileri tedarik zinciri yönetiminde kullanır.
- **d. Müşteri Geri Bildirimleri ve Anketler:** Uber: Uber, sürücü ve yolcu geri bildirimlerini toplar. Bu yorumlar, sürücülerin performansını değerlendirmek ve müşteri memnuniyetini artırmak için analiz edilir. Airbnb: Airbnb, kullanıcı yorumlarını ve değerlendirmelerini toplar. Bu veriler, ev sahiplerinin performansını değerlendirmek ve kullanıcıların tatminini artırmak için kullanılabilir.
- **e. Makine Verisi:** Üretim Tesisleri: Sensörler, makinelerin çalışma koşullarını izler ve verileri analiz ederek makinelerin arıza yapmadan önce bakım yapılmasını sağlar. Petrol ve Gaz Sektörü: Petrol sondaj kulelerinde sensörler, kuyunun basınç, sıcaklık gibi verilerini toplar. Bu veriler, operasyonel güvenliği artırmak için kullanılır.

## BÜYÜK VERİNİN KULLANIM ALANLARI

- **a. Sağlık Sektörü:** Medikal Görüntüleme: Röntgen, MR, CT taramaları gibi medikal görüntüleme verileri büyük veri analitiği ile daha hızlı ve doğru bir şekilde analiz edilir. Yapay zekâ, kanser hücrelerini tespit etmek gibi görevlerde kullanılabilir. Pandemi Yönetimi: COVID-19 gibi salgınlarda, dünya genelindeki vaka sayıları, hastaneye

başvurular ve hasta bilgileri büyük veri teknolojileriyle analiz edilerek, salgının yayılma hızı tahmin edilir.

- **b. Finans Sektörü:** Dijital Ödemeler: Visa ve MasterCard gibi ödeme sistemleri, her işlemde büyük miktarda veri toplar. Bu veriler, ödeme dolandırıcılığı tespiti ve kredi risk değerlendirmesi için kullanılabilir. Blockchain: Kripto para işlemleri de büyük veri üretir. Blockchain üzerindeki her bir işlem kaydı, merkeziyetsiz ve şeffaf bir şekilde analiz edilebilir.
- **c. E-Ticaret:** Dynamic Pricing: Amazon ve eBay gibi platformlar, fiyatlandırmayı dinamik olarak değiştirir. Bu fiyatlandırma, büyük veriyi kullanarak rakiplerin fiyatlarını, talep miktarını ve diğer faktörleri göz önünde bulundurur. Toptan Satış: Costco ve Walmart gibi büyük perakendeciler, ürün satış verilerini ve müşteri alışveriş alışkanlıklarını analiz ederek, hangi ürünlerin daha fazla satıldığını ve hangi fiyatların daha cazip olduğunu öğrenebilir.
- **d. Akıllı Şehirler:** Akıllı Trafik Sistemleri: Trafik ışıkları ve sensörler, araçların yoğunluk durumunu anlık olarak ölçer. Bu veriler, trafik akışını optimize etmek için kullanılır. Atık Yönetimi: Akıllı şehirlerde, çöpleri toplayan araçlar, her çöp kutusunun doluluk oranını izler ve bu verilere göre toplama rotalarını optimize eder.

## BÜYÜK VERİ İŞLEME YÖNTEMLERİ

Büyük veri analizinin etkin bir şekilde yapılabilmesi için kullanılan çeşitli işleme yöntemleri mevcuttur. Bunlar, verilerin depolanması ve analizi için farklı teknolojiler ve araçlar gerektirir.

- **1. Hadoop ve MapReduce:** Hadoop, büyük veriyi dağıtık bir ortamda depolamak ve işlemek için kullanılan açık kaynaklı bir framework'tür. Hadoop, verileri kümeleme (clustering) yöntemi ile çok sayıda sunucuya dağıtarak, her bir sunucuda veriyi paralel bir şekilde işler. MapReduce, Hadoop ile veri işleme sürecini düzenleyen bir algoritmadır. Veriyi, "Map" fonksiyonu ile işleyip, "Reduce" fonksiyonu ile sonuca ulaşmak için kullanılır.
- MapReduce süreci şu şekilde işler:
    - **Map aşaması:** Veriler, daha küçük ve işlenebilir parçalara dönüştürülür.
    - **Reduce aşaması:** Map aşamasında işlenen veriler, birleştirilir ve sonuçlar elde edilir.
- **2. Spark:** Apache Spark, Hadoop'a kıyasla daha hızlı veri işleme yeteneği sunan başka bir açık kaynaklı platformdur. Spark, veriyi bellekte (in-memory) işleme yeteneğine sahip olduğundan, disk tabanlı sistemlere göre çok daha hızlı sonuçlar verebilir. Spark, özellikle

büyük veri analizlerinde düşük gecikmeli veri işleme gereksinimlerini karşılamak için kullanılır.

- **3. Veritabanı Sistemleri ve NoSQL:** Büyük veri analizi için kullanılan bir diğer önemli yaklaşım, geleneksel ilişkisel veritabanlarının ötesine geçen NoSQL veritabanlarıdır. NoSQL veritabanları, büyük veri kümelerini daha esnek bir şekilde saklamak ve işlemek için kullanılır. Bu veritabanları, yapılandırılmamış veya yarı yapılandırılmış verilerle çalışırken, yüksek erişilebilirlik ve ölçeklenebilirlik sağlar.
- Öne çıkan NoSQL veritabanlarından bazıları:
    - **MongoDB:** Belge tabanlı bir veritabanıdır ve JSON benzeri formatlarda veri depolar.
    - **Cassandra:** Büyük veri kümeleri üzerinde yatayda ölçeklenebilirlik sunar ve özellikle büyük veri gereksinimleri olan uygulamalar için uygundur.

## BÜYÜK VERİNİN ZORLUKLARI

- **a. Veri Güvenliği ve Gizliliği:** Google Analytics: Web sitelerinin kullanıcı verilerini toplarken, kişisel bilgilerin güvenliğini sağlamak çok önemlidir. Herhangi bir veri sızıntısı, kullanıcı güvenini zedeler. Medikal Veriler: Hastaların tıbbi geçmişi ve genetik bilgileri çok hassastır. Bu tür verilerin kötüye kullanılmaması için şifreleme ve gizlilik politikalarına dikkat edilmelidir.
- **b. Veri Kalitesi:** Yanlış Etiketleme: E-ticaret sitelerindeki bazı ürünler yanlış etiketlenebilir. Örneğin, bir ayakkabı ürünü, “kadın ayakkabısı” yerine “erkek ayakkabısı” olarak etiketlenirse, bu veri yanlış analiz sonuçlarına yol açabilir. Eksik Veriler: Bazı kullanıcılar, anketlere eksik yanıtlar verebilir. Bu durum, analizin doğruluğunu düşürür.
- **c. Veri Entegrasyonu:** Farklı Sistemlerden Gelen Veriler: Bir işletme farklı sistemlerden veri alabilir (CRM, ERP, vb.). Bu verilerin birbirine entegre edilmesi zordur çünkü farklı sistemler farklı veri formatları kullanabilir. Birden Fazla Kaynak: Farklı e-ticaret platformlarından veya sosyal medya hesaplarından gelen verilerin bir araya getirilmesi ve analiz edilmesi karmaşık olabilir.
- **d. Yüksek Maliyetler:** Veri Depolama: Devasa veri kümelerini depolamak ve işlemek için güçlü sunuculara ve büyük veri platformlarına yatırım yapmak gerekir. Bu da maliyetleri artırır. Yazılım ve Altyapı: Hadoop, Spark gibi büyük veri analiz araçlarını kullanmak için yüksek performanslı altyapılar ve yazılım lisansları gerekebilir.

Büyük veri, günümüz dünyasında kritik bir rol oynamaktadır ve hemen hemen her sektörde potansiyel olarak büyük faydalar sağlayabilir. Verinin doğru bir şekilde toplanması, işlenmesi ve

analiz edilmesi, stratejik kararlar almak ve işletmeleri daha verimli hale getirmek için oldukça önemlidir. Bununla birlikte, büyük veriyle çalışırken karşılaşılan zorluklar ve güvenlik sorunları da göz önünde bulundurulmalıdır.

## BÜYÜK VERİ DEPOLAMA VE YÖNETİM STRATEJİLERİ

Büyük veri için uygun bir depolama çözümü, verinin hem verimli hem de güvenli bir şekilde saklanmasını sağlar. Büyük verinin depolanması için en çok tercih edilen yöntemler şunlardır:

- **Dağıtık Dosya Sistemleri:** Hadoop HDFS (Hadoop Distributed File System), büyük veri kümelerinin dağıtık bir şekilde depolanmasını sağlar. Veriler birden fazla sunucuda parçalar halinde depolanır, bu da veri kaybı riskini azaltır.
- **Veritabanı Yönetim Sistemleri (DBMS):** NoSQL veritabanları, büyük veri yönetimi için esnek çözümler sunar. Ayrıca, verinin ölçeklenebilir bir şekilde saklanmasını sağlayan araçlar geliştirilmiştir.
- **Bulut Depolama:** Bulut teknolojileri, büyük veriyi depolamak için popüler bir alternatiftir. Bulut depolama hizmetleri, verinin uzaktan erişilebilir olmasını sağlar ve veri güvenliğini artıran çeşitli çözümler sunar.

## BÜYÜK VERİ VE MAKİNE ÖĞRENİMİ

Büyük veri analitiği, makine öğrenimi algoritmaları ile birleştiğinde daha etkili sonuçlar verebilir. Makine öğrenimi, veriden öğrenme ve tahminlerde bulunma yeteneği sağlar. Büyük veri kümesi üzerinde kullanılan makine öğrenimi algoritmaları, doğrusal regresyondan derin öğrenmeye kadar birçok farklı yöntemi içerebilir. Bu tekniklerin başarısı, veri kalitesi ve algoritma optimizasyonu ile doğrudan ilişkilidir. Örneğin, bir e-ticaret sitesinde kullanıcıların davranışlarını analiz ederek, gelecekteki alışveriş tercihlerini tahmin edebilirsiniz. Bu tür tahminler, kişiselleştirilmiş pazarlama stratejileri oluşturulmasında yardımcı olabilir.

Büyük veri, günümüzün veri odaklı dünyasında önemli bir yer tutmaktadır. Bu alandaki gelişmeler hem teknoloji hem de iş dünyası için büyük fırsatlar yaratmaktadır. Ancak, büyük verinin etkin bir şekilde kullanılabilmesi için veri güvenliği, işleme yöntemleri ve uygun depolama çözümleri gibi zorluklarla başa çıkılması gerekmektedir. Bilgisayar mühendisliği ve yazılım mühendisliği öğrencilerinin, büyük veri teknolojilerini anlaması, bu alanda uzmanlaşarak sektördeki gelişmeleri yakından takip etmesi önemlidir.