Büyük Veri Analitiği

Elektronik ortamların gelişmesi ve yaygınlaşmasıyla, kullanılan verilerin boyutu artmakta, farklı verilerin beraber işlenmesi/değerlendirilmesi kolaylaşmaktadır. Bu bağlamda verilerin işlenmesi/değerlendirilmesindeki bakış açıları değişim geçirmekte ve gelişmektedir. Veriler değiştikçe ve geliştikçe, kullanılan teknik ve teknolojiler ile yöntem ve çözümlerde değişiklik göstermektedir.

Verilerin yönetimi, saklanması, işlenmesi ve/veya aktarılması kadar, verilerden değer elde etmek, verileri bilgi ve değere dönüştürmek, kullanıma uygun hale getirmek de önemlidir. Bu aşamalar aslında verinin değişimi/gelişimini ifade etmektedir. Verilerin bilgiye, bilginin bilgeliğe, bilgeliğin ise kararlara dönüşümü bu değişim/gelişimin en önemli adımlarıdır. Her adım, farklı aşamaları, zorlukları ve kazanımları kapsamaktadır.

Büyük veri kavramı, ilk kez Michael Cox ve David Ellsworth tarafından 1997 yılında düzenlenen bir Konferans ile literatüre kazandırılmıştır. Çalışmada, veri setlerinin çok büyük olduğu, bilgisayar sisteminin kapasitesini ve harici kapasitelerini doldurduğundan bahsedilmiş ve karşılaşılan soruna "Büyük Veri Problemi" adı verilmiştir.

Büyük veri; verilerin saklanmasında, analiz edilmesinde ve yönetilmesinde klasik veri tabanı yönetim sistemlerin yetersiz kaldığı durumlarda karşımıza çıkan bir kavram olarak tanımlanabilir (Veri tabanı: yapılandırılmış bilgi veya verilerin depolandığı alan). Bu kavram, organizasyonlara göre değişiklik gösterebilir. Tanım olarak büyük veri; "farklı formatlarda, hızlı bir şekilde ve büyük hacimde üretilen veriler" şeklinde ifade edilebilir. Büyük veri bizlere fırsatlar sunarken, beraberinde yeni sorunlar da getirmektedir.

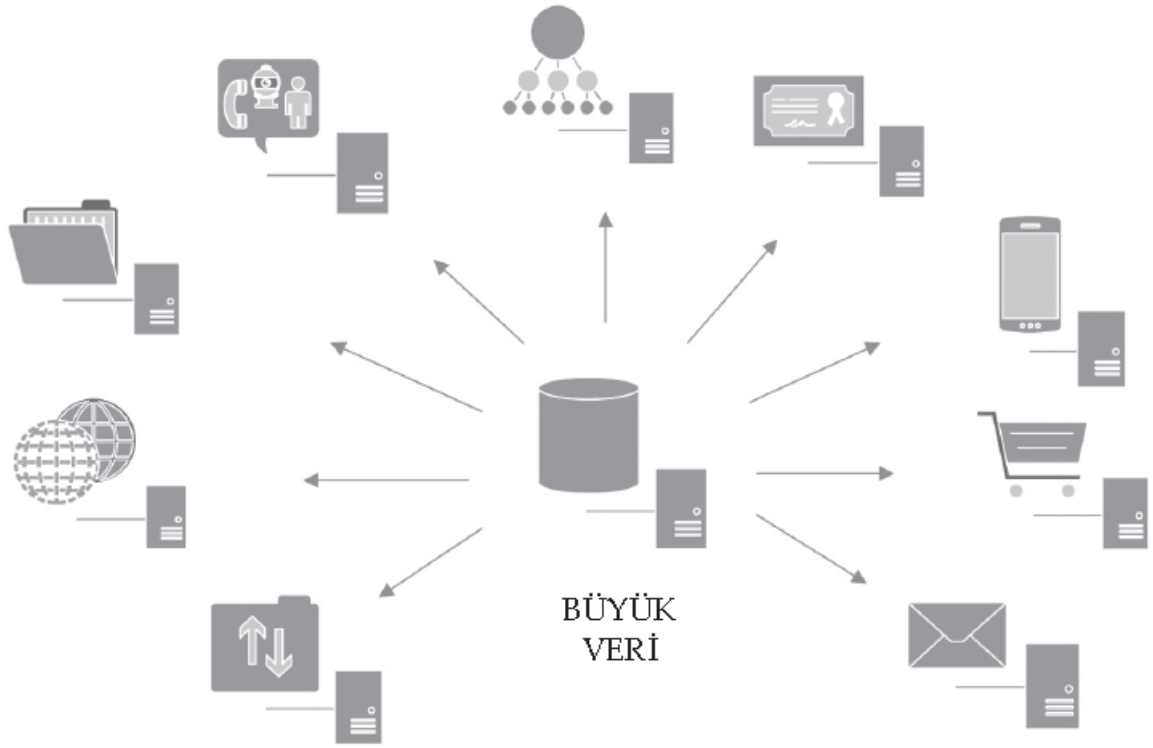
Veri Yapısal, Yarı Yapısal ve Yapısal Olmayan olarak 3 şekilde tanımlanmaktadır.

1. Yapısal veri: Yapısal veri, modellenmesi, girdi olarak sokulması, saklanması, sorgulanması, işlenmesi ve görselleştirilmesi kolay olan tüm veri türlerini ifade etmektedir. Genel olarak, belirli tür ve boyutlarda önceden tanımlı alanlarda sunulmakta, ilişkisel veri tabanlarında veya tablolarda yönetilebilmektedir. Katı bir yapıya sahip olan bu veri türünde, süreçlerin yüksek performanslı yetenekler veya paralel teknikler gerektirmemesinden dolayı faydalı bilgilerin elde edilmesi diğer veri türlerine kıyasla daha kolaydır.

2. Yarı yapısal veri: Yarı yapısal veya kendi kendini açıklayan veri, yapısal bir veri türünü yansıtmakla birlikte özünde sadece katı bir modeli barındırmamaktadır. Diğer bir ifadeyle yarı yapısal veri, yapısallığın tanımlandığı modellerin yanı sıra belirli öğeleri ve verideki farklı alanların hiyerarşik bir gösterimini tanımlamak adına kullanılan etiketler ve işaretler gibi çeşitli meta modelleri de bulundurmaktadır. Yarı yapısal verinin en çok bilinen örnekleri arasında XML (Extensible Markup Language - Metin İşaretleme Dili) ve JSON (JavaScript Object Notation – Veri Değişim Formatı) programlama dilleri yer almaktadır.

3. Yapısal olmayan veri: Yapısal olmayan veri, tanımlı bir format haricinde sunulan ve depolanan kayıt türleridir. Genellikle, kitaplar, makaleler, belgeler, e-postalar gibi serbest formatlardaki metinlerden ve resim, ses ve video gibi medya dosyalarından oluşmaktadır. Bu türdeki verinin katı bir şekilde sunulmasının zor olması, veri işleme süreçlerinde NoSQL (Not only SQL – İlişkisel Olmayan Veri Tabanı) gibi yeni mekanizmaların ortaya çıkmasına neden olmuştur.

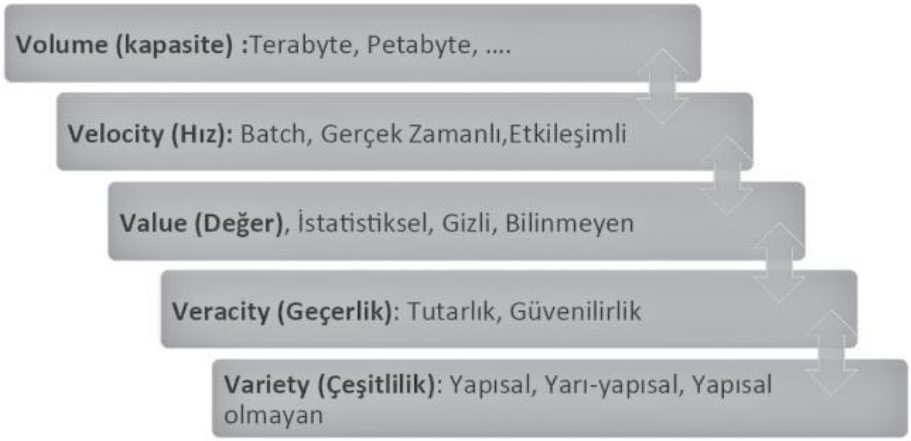
Büyük veriyi daha anlaşılır kılmak istersek elektronik ortamlarda yapılan işlemler, veri trafiği, uygulamalar, e-postalar, metinler, belgeler, videolar, sesler, resimler, tıklama akışları, sistem günlükleri, arama sorguları, sosyal ağ etkileşimleri, sağlık kayıtları, bilimsel veriler, devlet ve özel sektöre ait kayıtlar, sensörler ve akıllı telefonlardan beslenen yapıyı her boyutuyla anlamamız gerekmektedir.



Büyük veriyi oluşturan alanlar

Büyük Veri Bileşenleri

Büyük veri, kısaca 5V (volume, velocity, variety, veracity, value) olarak adlandırılan beş bileşen ile ifade edilmektedir.



Büyük veri bileşenleri

1) Hacim – Kapasite (Volume)

Verinin hacmi verinin boyutu ile doğru orantılıdır ve genel olarak gigabayt, terabayt, petabayt gibi çeşitli veri ölçü birimleri ile ifade edilmektedir.

Hacim büyük verinin bilinen en önemli özelliklerinden ve problemlerinden biri olarak karşımıza çıkmaktadır. Verinin hacminin problem oluşturma durumu organizasyonun büyüklüğüne göre değişebilmektedir. Veri sadece petabaytlarca verinin olması durumunda büyük veri olarak karşımıza çıkmamaktadır, eldeki veri miktarının istenilen zamanda analizinin mümkün olmadığı durumlarda büyük veri problemi olarak ele alınmaktadır.

Bayt Birimleri						
Yaygın önek				**İkilik önek**		
Ad	**Sembol**	**Ondalık**	**İkilik**	**Ad**	**Sembol**	**İkilik**
kilobayt	KB	10^3	2^{10}	kibibayt	KiB	2^{10}
megabayt	MB	10^6	2^{20}	mebibayt	MiB	2^{20}
gigabayt	GB	10^9	2^{30}	gibibyte	GiB	2^{30}
terabayt	TB	10^{12}	2^{40}	tebibayt	TiB	2^{40}
petabayt	PB	10^{15}	2^{50}	pebibayt	PiB	2^{50}
eksabayt	EB	10^{18}	2^{60}	eksbibayt	EiB	2^{60}
zettabayt	ZB	10^{21}	2^{70}	zebibayt	ZiB	2^{70}
yottabayt	YB	10^{24}	2^{80}	yobibayt	YiB	2^{80}

2) Hız (Velocity)

Veri değişken hızlarla üretilebilmektedir. Büyük hacimli durağan veriler büyük veri problemi oluştururken, özellikle gündelik hayatta sıklıkla kullanıldığımız telefon veya nesnelerin interneti cihazları, çeşitli makineler tarafından üretilen sensör verileri ve bunlara benzeyen veri kaynakları çok hızlı bir şekilde veri üretmektedir.

Bir akış içerisinde hızlı bir şekilde üretilen verinin gerçek zamanlı olarak analiz edilebilmesi ve yönetilebilmesi ise büyük verinin bir diğer problemi olarak karşımıza çıkmaktadır.

3) Çeşitlilik (Variety)

Verinin çeşitliliği, veri kaynaklarının farklılığından kaynaklanmaktadır. Üretilen veriler yapısal, yarı-yapısal veya yapısal olmayan formatlarda karşımıza çıkabilmektedir. Yapısal verilere ilişkisel veri tabanlarındaki veriler, yarı-yapısal verilere XML, JSON formatındaki veriler yapısal olmayan verilere ise ses, video, metin dosyaları örnek olarak verilebilir.

Bu farklı yapılarda verilerin bir arada kullanılıyor olması durumunda verideki bu çeşitlilik büyük bir problem olarak karşımıza çıkmaktadır. Özellikle verilerin çıkar-dönüştür-yükle işlemlerinde bu verilere özgü birçok yeni büyük veri teknolojilerinin kullanılması bir zorunluluk haline gelmektedir.

4) Geçerlik (Veracity)

Gerçek hayat problemlerinde kullanılan büyük veri içerisinde verinin doğruluğunu olumsuz olarak etkileyebilecek birçok faktör vardır. Bu faktörler genellikle karşımıza gürültü veya aykırılık olarak çıkmaktadır.

Doğruluğundan emin olunamayan veri üzerinde yapılacak analizler gerçek değerin ortaya konulmasını engellemektedir. Özellikle çeşitli sensörler aracılığı ile üretilen veriler sensör doğası gereği gürültüye çok meyillidir. Bundan dolayı bu ve benzeri durumların oluşması halinde büyük veriden büyük değer üretebilmek mümkün değildir. Büyük veriye özgü teknolojiler ile verinin doğruluğundan ve analize uygunluğundan emin olunmalıdır.

5) Değer (Value)

Veri analizini çeşitli organizasyonlarca önemli hale getiren veriden üretilen değerdir. Değer üretilemeyen herhangi bir veri anlamsızdır. Üretilen değer ise verinin içeriğine, üretilme amacına, uygulama alanına vb. faktörlere göre değişiklik göstermektedir. Var olan verilerin yukarıdaki özelliklere sahip olması durumunda bu veriden geleneksel yöntemler ile değer üretmek çok zor olmaktadır. Dolayısıyla bu verinin büyük veri bakışıyla ele alınıp büyük veri teknolojileri ile analiz edilme ihtiyacı doğmuştur.

Büyük Veri Kaynakları

Akıllı telefonlar, tablet bilgisayarlar, sensörler, tıbbi ekipmanlar, web trafiği kayıtları, sosyal ağlardaki etkileşimler ve eczacılık, meteoroloji, simülasyon gibi alanlarda çözümler sunan bilimsel araştırmalar gibi birçok kaynak, büyük verinin ortaya çıkmasını sağlamaktadır.

Bununla beraber web ortamının artan heterojenliği, web sayfaları üzerinde farklı medyalarda (metin, resim ve video), türlerde (ansiklopedi, haber, bloglar) ve konularda (eğlence, spor, teknoloji) büyük veri içeriğinin sağlanmasına neden olmaktadır.

Büyük veri çeşitliliğinin artmasında çok sayıda veri kaynağı etkili olmaktadır. Bu kaynaklardan bir kısmı tamamen yeni veri kaynağıyken, bazı veri kaynakları da mevcut verinin ayrışması (mevcut kaynakların sayısal ortama aktarılması) sonucu ortaya çıkmaktadır. Birçok endüstriyel alan, yeni veri üretimi ve mevcut verinin sayısallaştırılması şemsiyesi altına girmekte ve her biri ayrı bir büyük veri kaynağını oluşturmaktadır.

Taşımacılık, lojistik, perakendecilik, kamu hizmeti ve telekomünikasyon alanlarında kullanılan GPS alıcı vericileri, RFID etiket okuyucuları, akıllı sayaçlar ve telefonlarda yer alan sensörler vasıtasıyla veri toplanmaktadır.

Sağlık hizmetleri alanında, elektronik tıbbi görüntüleme ve raporlamalardan, kısa dönemli halk sağlığının gözlemlenmesinde ve uzun dönemli salgın hastalıkların araştırılmasında kullanılmak üzere veri toplanmaktadır.

Birçok devlet kuruluşu, nüfus sayımı, enerji kullanımı, bütçe raporları, kanunsal yaptırım sonuçları, seçim sonuçları gibi halka ait raporları sayısal ortama aktarmakta ve veri olarak halkın erişimine sunmaktadır.

Kitap, gazete, magazin, televizyon, radyo, film, sinema, müzik ve oyun gibi birçok alanda hizmet veren eğlence alanı; sayısal kayıt, üretim ve dağıtıma yönüyle kişi ve toplumların davranışlarını gözlemleyen geniş içerikte veri toplamaktadır.

Yaşam bilimleri alanında veri üretiminde belirli bir ücret temelinde yapılan gen sayımı, genetik çeşitliliği araştırmada ve potansiyel tedavi etkinliğini belirlemede analiz edilebilecek onlarca terabaytlık veriyi oluşturmaktadır.

Video görüntüleme alanında, alt yazılı televizyon teknolojisinden IP temelli televizyon kameralarına ve kayıt sistemlerine doğru ilerleme olmuş, IP temelli yeni teknolojik kamera verileri, güvenlik ve servis hizmetlerinin geliştirilmesi amacıyla kullanılmaktadır.

Araştırma kuruluşu Statista'nın istatistiklerine göre, 2016 yılı itibarıyla büyük veri ve analitiğinin dünya genelindeki pazar payında, bankacılık %13,1 ile en çok gelir sağlayan uygulama alanı olmuştur. Bankacılığı, %11,9 ile kesikli üretim, %8,4 ile süreç tipi üretim, %7,6 ile devlet hizmetleri ve %7,4 ile de profesyonel hizmetler takip etmiştir. Aynı yıl, büyük verinin tüm uygulama alanlarındaki toplam pazar değeri ise 130,1 milyar Amerikan doları seviyesine ulaşmıştır.

Büyük Verinin Avantajları ve Dezavantajları

Büyük veriyi kullanan organizasyonlar piyasadaki değişime hızlı biçimde adapte olarak daha etkin kararlar alabilmektedirler. Ayrıca doğru bilgiye zamanında ulaşarak daha fazla yenilik oluşturabilmekte ve bunun sonucunda organizasyonun üretkenliği ve finansal performansı artış göstermekte ve önemli bir zaman tasarrufu ve verimlilik sağlamaktadır.

Büyük Verinin Avantajı ve Dezavantajı

Avantaj	**Dezavantaj**
Daha iyi karar verme	Müşteri gizliliği tehdidi
Üretkenliği artırmakta	Yüksek fiyat
Maliyeti azaltmakta	Büyük verinin kötüye kullanımı
Müşterilere verilen hizmeti iyileştirmekte	Bilgi bombardımanı (Information overload)
Dolandırıcılığı tespit etmekte	Yeni sistemi öğrenirken zaman tüketimi
Geliri arttırmakta	Yeni donanıma yatırım yapılması (maliyetli)
Daha fazla yenilik sağlamakta	
Pazarda hızlı değişim sağlamakta	

Büyük Veri Analitiği

Organizasyonlar proje amaçlarına ve hedeflerine ulaşmak için ham madde olarak uygun veriye, geniş ve karmaşık veri kümelerinden ulaşmak durumundadır. Uygun veri elde edildikten sonra yapısal, yarı yapısal ve yapısal olmayan veriyi bir bütün olarak aralarında ilişkilendirecek, işleyecek ve projelerde stratejik kararlar verme aşamasında bilgiye ulaşmayı sağlayacak veri analitiğine ihtiyaç duyulacaktır. Teknolojik gelişmelerle birlikte, artık yapısal veriye ek olarak yarı yapısal ve yapısal olmayan veri türleri de kullanılmaya başlanmıştır.

Büyük veri analitiği, farklı türlerde içerik barındıran çok geniş ve farklı kayıtları işlemek adına geliştirilmiş analitik ve paralel tekniklerin kullanılmasıdır. Bu noktada büyük veri analitiği araçları, geleneksel veri tabanı teknikleri kullanılarak işlenmesi zor olan, hızla değişen ve çok miktardaki yapısal, yarı yapısal ve yapısal olmayan verinin bir bütün olarak analizi ile veriden değerli bilgiler elde edilmesini amaçlamaktadır.

Diğer bir ifadeyle büyük veri analitiği, karar verme aşamasında yol gösterici olacak bilgiyi elde etmek adına büyük veri kümelerinin analiz edilmesinde kullanılan bir tekniktir.

İnsan Kaynaklarında kullanılan büyük veriyi, büyük veri analitiği temelinde değerlendirdiğimizde işe alma, seçme, kariyer yönetimi gibi faaliyetleri önemli ölçüde kolaylaştırmaktadır. Büyük veri, organizasyonların işe alım çalışmaları için internetten daha geniş bir platform sağlamaktadır. Bu nedenle organizasyonlar, işe alımları sosyal ağ çalışmasıyla birleştirmektedir. Bunun sonucunda işe alımda büyük veri analizi; öz geçmiş ve uygulama bilgilerini sürekli olarak toplayabilme imkânı elde etmektedir.

Bununla birlikte, büyük veri açısından bilgi erişimi ve paylaşımı çok kullanışlıdır. Herhangi bir bireyin ağ üzerinden öğrenmek istediği bilgilerin herhangi bir zamanda veya herhangi bir yerde kolayca aranabilmesini sağlamaktadır. Dolayısıyla, büyük veri yaklaşımını izleyen organizasyonlar, günlük iş yükünü, işin belirli içeriğini ve her çalışanın görev başarısını kaydederek ardından bu verileri analiz edebilmektedir. Organizasyonların başarısı, büyük ölçüde işe aldıkları bireylerin kalite düzeyine bağlıdır. Modern insan kaynakları, teknolojiyi yalnızca yeni çalışanları işe alabilmek için değil, aynı zamanda çalışma ortamını değiştirmek, çalışanları motive etmek, onları daha iş birlikçi ve daha memnun ve etkili hale getirebilmek için bir araç olarak kullanmaktadır.

Buna ek olarak, organizasyonların modern insan kaynakları yönetimin yeni kurallarını benimsemesi önemlidir. Bu nedenle, insan kaynakları yönetiminin daha dijital hale gelmesi ve dijital dönüşüme dayalı modern başarılara daha bağlı olmasını gerektirmektedir. Özellikle önemli veri kaynaklarına dayalı olan analitiğin önemi, karar verme sureci için hayati bir girdi haline gelmektedir.

Büyük veri teknolojilerinin kombinasyonu ve web tabanlı kanallar aracılığıyla tüketici verilerine daha fazla erişim sağlanmaktadır. Böylece daha önce mümkün olmayan müşteri iç görüleri elde edilmektedir. Bu nedenle büyük veri iç görüleri, kurumdaki yöneticileri üstün bir konuma getirmektedir.

Mevcut araştırmalar, büyük verinin değer oluşturma potansiyelini belirlemede endüstrilerin rolünü vurgulamaktadır. Bazı endüstri sektörleri verilere sınırlı erişim, veri gizliliği ve koruma endişeleri nedeniyle daha az avantaj elde edebilmektedir. Dijital çağ aynı zamanda çalışanların kişisel ve profesyonel yaşamlarını üstlenmelerini sağlamaktadır. Büyük veri bileşenlerinin iyi biçimde yapılandırılırsa ve kullanılırsa, işe alım sürecinde bu bilgileri kullanan şirkete artı değerler getireceği belirtilmektedir.

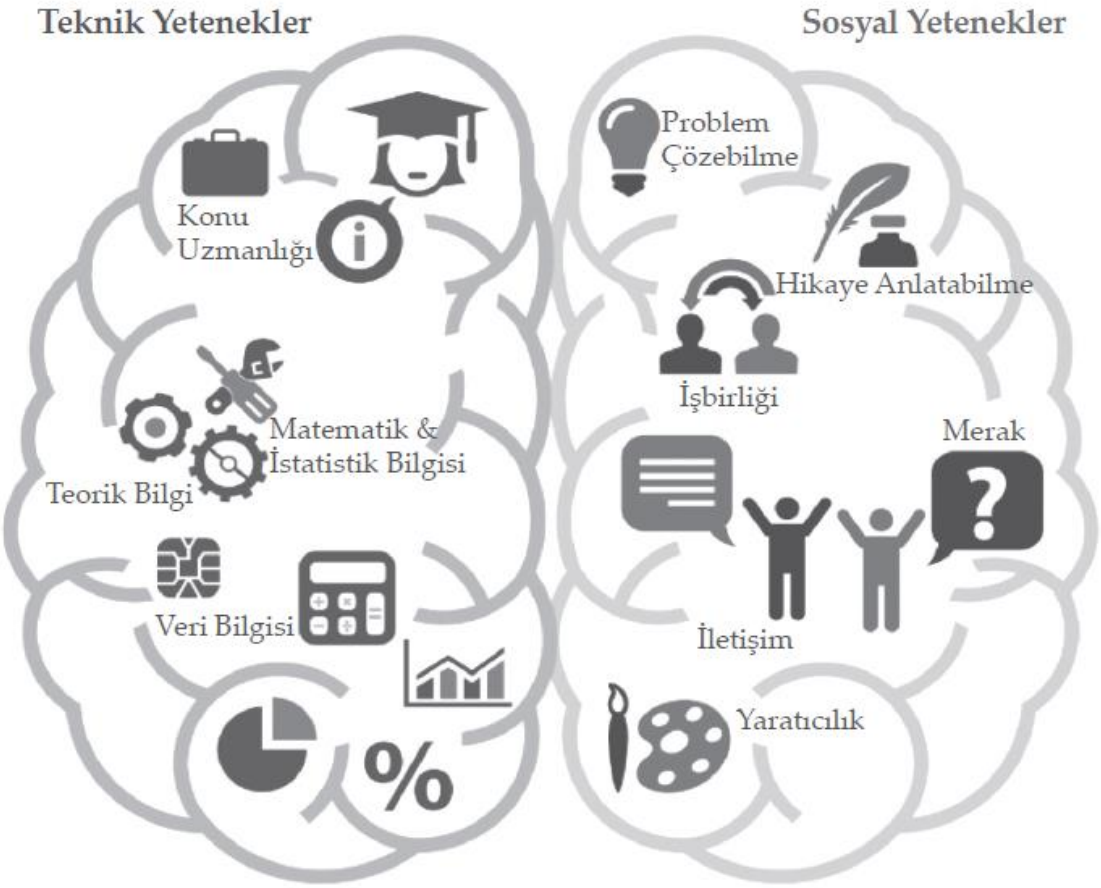
Büyük Veri Analitiğinde Cevap Bulunması Gereken Sorular

Büyük veri analitiğinde cevap bulunması gereken başlıca sorular şunlardır;

- Büyük veri boyutu ve çeşitliliği arttıkça, veri analitiğinde karşılaşılacak sorunlarla nasıl başa çıkılacaktır?
- Verinin tamamı depolanmalı mıdır?
- Verinin tamamı analiz edilmeli midir?
- Hangi büyük veri unsurlarının gerçekten önemli olduğuna nasıl karar verilmelidir?
- En iyi avantajı elde etmek için büyük veri nasıl kullanılmalıdır?

Cevap bekleyen bu sorular, büyük verinin analiz aşamasında çok büyük zorlukları da beraberinde getirmektedir. Büyük veri, yapısal, yarı yapısal ve yapısal olmayan veri türlerinden oluştuğu için büyük veri analitiğinde ileri kabiliyetlere gereksinim duyulmaktadır. Bununla birlikte veri üzerinde yapılması gereken analiz türü, elde edilecek sonuçlara da bağlıdır. Analiz aşamasında ya tüm büyük veri unsurları birleştirilir, ya da hangi büyük veri unsurunun elde edilecek sonuçla alakalı olduğu belirlenir

Bu noktada organizasyonlar, büyük veri analitiğinde kullanacakları iş zekâsı sistemlerini ve analitik girişimlerini genişletmek adına veri bilimcisine ihtiyaç duymaktadırlar. Veri bilimcisi istatistik, matematik, bilgisayar mühendisi, makine öğrenmesi, programlama ve veri işleme teknolojileri gibi konulara hâkim ve çalıştığı alana özgü bilgi birikimine sahip, iletişim yeteneği güçlü, kurumunun geleceği için gereken kritik sorunları tespit etme ve öngörülerde bulunabilme yeteneklerine sahip olan kişi/uzmanı ifade etmektedir.



Veri Bilimcisinin Sahip Olduğu Yetenekler

Bu sayede, saklı örüntüler ortaya çıkarılabilmekte, bilinmeyen gerçekler elde edilebilmekte, korelasyonlardan verilerin zenginleştirilmesi yapılabilmektedir. Bunun için güçlü platformlarda gelişmiş analiz algoritmalarına, bunların büyük veri ortamlarına uygulanmasına ve en önemlisi bunları gerçekleştirilebilecek platformlara ihtiyaç vardır.

Büyük Veri İşleme Platform ve Araçları

Veriler büyüdükçe, verilerin toplanması, saklanması, işlenmesi ve değerlendirilmesi için yeni algoritmalara, yaklaşımlara, metotlara ve en önemlisi de teknik ve teknolojilere ihtiyaç vardır.

Büyük veri ile uğraşırken karşılaşılan problemlerin üstesinden gelirken, depolama ve hesaplama süreçleri klasik yöntemlere göre farklılık göstermektedir. Değerli, gizli veya yeni bilgilerin keşfedilmesi için hem teknik ve teknolojiler için disiplinler arası çalışmalara hem de farklı metotların ve yeni yaklaşımların hem geliştirilmesine ve bu ortamlarda yeni çözümlerin geliştirilmesine ihtiyaç vardır.

Platform	**Lokal**		Hadoop, Spark, MapR, Cloudera, Hortonworks, InfoSphere, IBM BigInsights, Asterix
	Bulut		AWS EMR, Google Compute Engine, Microsoft Azure, Pure System, LexisNexis HPCC Systems
Veri Tabanı	**SQL**		Greenplum, Aster Data, Vertica, SpliceMachine
	IN-MEMORY		SAP HANA
	NoSQL	**Sütun Şeklinde**	HBase, HadoopDB, Cassandra, Hypertable, BigTable, PNUTS, Cloudera, MonetDB, Accumulo, BangDB
		Anahtar-Değer	Redis, Flare, Sclaris, MemcacheDB, Hypertable, Valdemort, Hibari, Riak, BerkeleyDB, DynamoDB, Tokyo Cabinet, HamsterDB
		Doküman Tabanlı	SimpleDB, RavenDB, ArangoDB MongoDB, Terrastore, CouchDB, Solr, Apache Jackrabbit, BaseX, OrientDB, FatDB, DjonDB
		Graf Tabanlı	Neo4J, InfoGrid, Infinite Graph, OpenLink, FlockDB, Meronymy, AllegroGraph, WhiteDB, TITAN, Trinity
Veri İşleme			MapReduce, Dryad, YARN, Storm, S4, BigQuery, Pig, Impala, Hive, Flink, Spark, Samza, Heron
Veri Ambarı			Hive, HadoopDB, Hadapt
Veri Birleştirme ve Transfer			Sqoop, Flume, Chukwa, Kafka, ActiveMQ
Sorgu Dili			Pig Latin, HiveQL, DryadLINQ, MRQL, SCOPE, ECL, Impala
İstatistik & Makine Öğrenmesi			Mahout, Weka, R, SAS, SPSS, Pyhton, Pig, RapidMiner, Orange, BigML, Skytree, SAMOA, Spark MLLib, H2O
İş Zekâsı			Talend, Jaspersoft, Pentaho, KNIME
Görselleştirme			Google Charts, Fusion Charts, Tableau Software, QlikView
Sosyal Medya			Radian6, Clarabridge

Kaynaklar

Aktan, E. (2018). Büyük veri: Uygulama alanları, analitiği ve güvenlik boyutu. Bilgi Yönetimi, 1(1), 1-22.

Sağıroğlu, Ş. (2017). Büyük Veri Ve Açık Veri Analitiği: Yöntemler ve Uygulamalar.

Küsbeci, P. (2021). Büyük Veri.