Module 4 - Understanding Data

> Bu notlar **IBM Data Science Professional Certificate** sertifika programının **What is Data Science** eğitimi üzerinden alınan bilgileri içermektedir. Söz konusu notların geliştirilmesinde internet kaynaklarından ve yapay zeka araçlarından da faydalanılmıştır.

Veri ve Veri Yapıları

1.1 Veri Nedir?

- Verinin, anlamlandırılmak üzere işlenmiş, düzensiz bilgilerin toplandığı bir kaynak olduğu ifade edilmektedir.
- Ham verinin; gözlemler, sayılar, semboller, karakterler, görüntüler gibi unsurlardan oluştuğu ve bu unsurların yorumlanarak anlamlı hale getirildiği belirtilmektedir.

1.2 Veri Sınıflandırması

Veriye, yapısına göre üç ana kategori atfedilmektedir:

a. Yapılandırılmış (Structured) Veri

- Verinin, önceden belirlenmiş şema veya veri modeline uygun şekilde düzenlendiği ifade edilmektedir.
- Genellikle satırlar ve sütunlar halinde tablolar halinde sunulan bu veriler, ilişkisel veritabanlarında (SQL) saklanabilmektedir.
- Kaynaklar arasında SQL veritabanları, OLTP sistemleri, Excel ve Google Spreadsheets, online formlar, GPS veya RFID gibi sensör verileri ile ağ ve web sunucu logları yer almaktadır.

b. Yarı Yapılandırılmış (Semi-structured) Veri

- Verinin belirli organizasyonel özelliklere sahip olmakla birlikte sabit bir şemaya bağlı kalmadığı belirtilmektedir.
- Veri, etiketler, elementler veya metadata kullanılarak hiyerarşik olarak gruplanabilmektedir.
- Kaynaklar olarak e-postalar, XML, JSON ve diğer işaretleme dilleri örnek gösterilmektedir.

c. Yapılandırılmamış (Unstructured) Veri

- Verinin, kolayca tanımlanabilir bir yapısının bulunmadığı, belirli bir format veya düzeni takip etmediği ifade edilmektedir.
- Web sayfaları, sosyal medya içerikleri, görüntüler, video ve ses dosyaları, belgeler (PDF, PowerPoint) gibi kaynaklardan elde edilen veriler bu kategoriye dahil edilmektedir.
- Bu veriler, genellikle NoSQL veritabanlarında veya dosya sistemlerinde saklanmakta, manuel veya özel analiz araçlarıyla incelenmektedir.

Veri Kaynakları ve Depolama Yöntemleri

2.1 İç ve Dış Kaynaklar

- **İç Kaynaklar:**

  Kurumların günlük faaliyetlerini destekleyen uygulamalardan elde edilen veriler; ilişkisel veritabanları (SQL Server, Oracle, MySQL, IBM DB2) ve veri ambarları gibi yapılar üzerinden sağlanmaktadır.

  Örneğin; perakende işlemleri veya müşteri ilişkileri yönetimi (CRM) sistemlerinden toplanan veriler analiz için kullanılmaktadır.

- **Dış Kaynaklar:**

  Kamu veya özel sektör tarafından yayınlanan demografik, ekonomik ya da sektör bazlı veriler; ayrıca satın alınabilen verisetleri de veri kaynağı olarak değerlendirilmektedir.

Bu veriler, genellikle düz dosyalar (flat files), elektronik tablolar veya XML belgeleri şeklinde sunulmaktadır.

2.2 Dosya Formatları ve Veri Transfer Yöntemleri

- **Düz Dosyalar (Flat Files):**

  Verinin, her bir kaydın bir satır olarak saklandığı, değerlerin belirli ayraçlarla (virgül, noktalı virgül, tab) ayrıldığı düz metin dosyalarıdır. CSV formatı en yaygın örnektir.

- **Elektronik Tablolar (Spreadsheets):**

  Satır ve sütun şeklinde düzenlenmiş, birden fazla çalışma sayfası içerebilen dosya formatları; Microsoft Excel (.XLS, .XLSX) gibi örnekler verilebilmektedir.

- **XML ve JSON:**

  XML, etiketler kullanılarak verilerin hiyerarşik şekilde organize edildiği bir format olarak; JSON ise anahtar-değer çiftlerine dayalı, esnek ve insan tarafından okunabilir bir veri formatı olarak kullanılmaktadır.

- **APIs ve Web Servisleri:**

  Uygulama programlama arayüzleri aracılığıyla veriye erişim sağlanmakta; RESTful API'ler, Twitter veya Facebook gibi platformlardan veri çekme işlemlerinde kullanılmaktadır.

- **Web Scraping:**

  Web sayfalarından belirli parametreler doğrultusunda veri çekme yöntemi, BeautifulSoup, Scrapy, Selenium gibi araçlarla gerçekleştirilmektedir.

- **Veri Akışları (Data Streams) ve RSS:**

  IoT cihazları, sensörler ve sosyal medya gibi kaynaklardan sürekli akan veri akışları; Apache Kafka, Spark Streaming ve Apache Storm gibi teknolojilerle işlenmekte, RSS beslemeleri de güncel veri akışı sağlanmasında kullanılmaktadır.

Veri Transferi, Esneklik ve Yeni Yaklaşımlar

- Veritabanları arası veri transferinde, farklı versiyonlama, özel karakter ve delimiter sorunlarının yaşandığı; bu nedenle veri transferinin, tek seferlik değil sürekli ve performanslı bir şekilde gerçekleştirilmesi gerektiği vurgulanmaktadır.
- Geleneksel ilişkisel veritabanlarının bazı durumlarda yetersiz kaldığı, özellikle yazma yoğun uygulamalarda (IoT, sosyal medya verileri) yeni nesil NoSQL çözümleri (Cassandra, HBase) gibi teknolojilerin benimsenmesine neden olduğu ifade edilmektedir.

Metadata ve Metadata Yönetimi

4.1 Metadata Nedir?

- Metadata, diğer veriler hakkında bilgi sağlayan veriler olarak tanımlanmaktadır.
- Üç ana metadata türü öne çıkarılmaktadır:
  - **Teknik Metadata:** Veritabanlarındaki veri yapılarının tanımlanması (tablolar, kolon sayıları, veri tipleri) gibi teknik ayrıntıları içermektedir.
  - **Süreç Metadata'sı:** Veri işleme süreçleri, veri hareketleri, kullanıcı erişim bilgileri ve sistem performansına dair bilgileri içermektedir.
  - **İş Metadata'sı:** Verinin nasıl elde edildiği, ne ölçtüğü ve diğer veri kaynaklarıyla olan ilişkileri gibi, iş açısından anlamlı bilgileri sunmaktadır.

4.2 Metadata Yönetiminin Önemi

- Metadata yönetimi, farklı kaynaklardan elde edilen bilgilerin entegre edilip, tüm organizasyon genelinde paylaşılmasını sağlamak amacıyla geliştirilen politika ve süreçleri içermektedir.
- Bir veri kataloğu oluşturulması, verinin keşfi, tekrarlanabilirliği, veri yönetişimi (data governance) ve erişimin sağlanması açısından kritik önem taşımaktadır.
- Popüler metadata yönetim araçları arasında IBM InfoSphere, CA Erwin, Oracle Warehouse Builder, SAS Data Integration Server, Talend Data Fabric ve diğerleri bulunmaktadır.

Veri Bilimi Ekosistemi ve Genel Özet

- Veri, yapılandırılmış, yarı yapılandırılmış ve yapılandırılmamış formlarda toplanmakta; her form, farklı analiz yöntemlerine ve araçlara ihtiyaç duyulmasına neden olmaktadır.
- Veri kaynaklarının çeşitliliği, iç ve dış kaynaklardan gelen verilerin; düz dosyalar, tablolar, XML/JSON, API'ler, web scraping ve sürekli veri akışları gibi yöntemlerle elde edilmesini sağlamaktadır.
- Verinin etkili analiz edilebilmesi için önce organize edilmesi, saklanması, aktarılması ve metadata yönetimi gibi altyapı süreçlerinin doğru biçimde uygulanması gerekmektedir.
- Ayrıca, veri transferinde karşılaşılan esneklik ve format uyumu sorunlarının aşılabilmesi için ilişkisel veritabanlarından NoSQL çözümlerine kadar farklı teknolojilerin kullanılması sağlanmaktadır.
- Metadata yönetimi, verinin hangi bağlamda oluşturulduğu ve nasıl işlendiğini anlamada ve veri yönetişimini desteklemede hayati bir rol oynamaktadır.

Veri Bilimi Uygulamalarının İşletme ve Toplum Üzerindeki Etkisi

5.1 İşletmelerde Rekabet Avantajı

- Veri biliminin, öneri motorları, rota optimizasyonu ve içerik stratejileri gibi uygulamalar aracılığıyla işletmelerin rekabet gücünü artırdığı belirtilmektedir.
- McKinsey, UPS, Netflix gibi örnekler üzerinden, veriye dayalı karar alma süreçlerinin; maliyet tasarrufu, operasyonel verimlilik ve stratejik konumlandırmada belirleyici olduğu vurgulanmaktadır.

5.2 Toplumsal ve Hayat Kurtarıcı Etkiler

- Sağlık sektöründe, tahmine dayalı analizler ve kişiselleştirilmiş tedavi önerileri sayesinde hasta sonuçlarının iyileştirildiği;

- Doğal afetlerde, erken uyarı sistemleri ve geniş veri setleri kullanılarak müdahale süresinin kısaltıldığı, can kayıplarının azaltıldığı ifade edilmektedir.

Nihai Teslimat ve Veri Odaklı Hikâye Anlatımı

6.1 Nihai Teslimatın Önemi

- Analitik sonuçların paydaşlara aktarılması için nihai teslimatın (final deliverable) tutarlı, anlaşılır ve etkili bir anlatı çerçevesinde sunulması gerekmektedir.
- Akademik raporlar, danışmanlık dökümanları veya sunumlar; içerik, grafik ve tablolarla desteklenen, stratejik öneriler içeren bir formatta hazırlanmalıdır.

6.2 Hikâye Anlatımının Rolü

- Veriye dayalı hikâye anlatımının, sadece sayısal bulguları değil; aynı zamanda verinin arkasındaki mantığı, bağlamı ve iş stratejisini de iletmeye yardımcı olduğu belirtilmektedir.
- “Geriye doğru çalışma” yaklaşımı ile nihai mesaj ve destekleyici veriler önceden belirlenmekte, bu sayede analiz sürecinin ve raporun kalitesinin artırılması sağlanmaktadır.

Genel Özet

Veri biliminin temelinde, ham verinin organize edilip anlamlandırılması yatmaktadır.

- Yapılandırılmış, yarı yapılandırılmış ve yapılandırılmamış veriler; farklı kaynaklardan temin edilmekte ve çeşitli araçlarla işlenmektedir.
- Veri kaynaklarının modern ekosistemi, iç (kurumsal uygulamalar, veritabanları) ve dış (kamu, özel verisetleri) kaynakları kapsamaktadır.
- Verinin transferi, esnekliği ve farklı formatlarla çalışma gereksinimleri; ilişkisel veritabanlarından NoSQL çözümlerine kadar yeni yaklaşımların benimsenmesini gerektirmektedir.

- Metadata yönetimi, verinin oluşum, işleniş ve dönüşüm süreçlerine dair bilgilerin korunması ve veri yönetişimi için kritik bir role sahiptir.
- Nihai teslimat süreci ve veriye dayalı hikâye anlatımı, analitik bulguların etkili biçimde paydaşlara aktarılmasını sağlayarak hem işletmelerin hem de toplumsal alanların yararına önemli içgörüler sunmaktadır.

Bu notlar, video transkriptlerinden elde edilen bilgilerin bütünsel olarak özetlenip, yapılandırılmış, bağlantıları güçlendirilmiş ve edilgen anlatım diliyle sunulması amacıyla hazırlanmıştır.