Büyük Veri Analitiği
(Big Data Analytics)

M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

Bu dersin sunumları, "Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011." kitabı kullanılarak hazırlanmıştır.

Genel bilgiler

Değerlendirme

Arasınav	: 25%
Ödevler	: 15%
Final Projesi	: 30%
Final Sınavı	: 30%

Ders kaynakları

- Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011.
- Real-Time Big Data Analytics: Emerging Architecture, Mike Barlow, O'Reilly Media, 2013.
- Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Jared Dean, Wiley, 2014.
- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services, 2015.

e-posta : akcayol@gazi.edu.tr
web : https://bigdata.gazi.edu.tr/akcayol

Genel bilgiler

Araştırma ödevleri

- Haftalık konu ile ilgili bir makale incelenerek detaylı rapor hazırlanacaktır.
- İncelenen makalede kullanılan **yöntemin/algoritmanın/yaklaşımın gerekçeleri** ile **elde edilen sonuçlar değerlendirilecektir.**
- İncelenen **makale son 3 yılda yayınlanmış** olacaktır.
- Makale **Q1, Q2 veya Q3 çeyrekliğinde yer alan bir dergide** yayınlanmış olacaktır (SCImago veya Scopus).
- Ödev dokümanı pdf formatında tek dosya olacak ve aşağıdaki dokümanları içerecektir:
  - Makalenin yayınlandığı yıl derginin yer aldığı çeyrekliği gösterir belge
  - İncelenen makalenin tam metni
  - Hazırlanan rapor
- Dosya adı '**DersKodu_ÖğrenciNo_ÖdevNo.pdf**' formatında olacaktır.

3

Genel bilgiler

Final Projeleri

- Proje bir yöntemin/algoritmanın bir alana/probleme uygulamasını içerecektir.
- Geliştirilecek uygulamanın yöntem/algoritma kısmında hazır araç, fonksiyon veya kütüphane kullanılmayacaktır.
- Hazırlanan projenin tüm dokümanları sıkıştırılmış tek dosya halinde elektronik olarak teslim edilecektir.
- Final proje raporu çıktı şeklinde teslim edilecektir.

4

Genel bilgiler

Ders içeriği

1. Büyük veri
2. Veri madenciliği
3. MapReduce
4. Benzer elemanların bulunması
5. Uzaklık ölçütleri
6. Data stream madenciliği
7. Link analizi
8. Öbekleme
9. Frequent itemsets
10. Birliktelik kuralları
11. Sıralı örüntüler
12. Information retrieval
13. Web arama
14. Web reklamcılığı

5

Konular

- Veri ve Bilgi
- Veri, Enformasyon, Bilgi, Anlayış, Bilgelik
- Büyük Veri
- Büyük Veri Analitiği
- Büyük Veri Kullanım Alanları
- Akış Verisi
- Akış Verisi Kaynakları

6

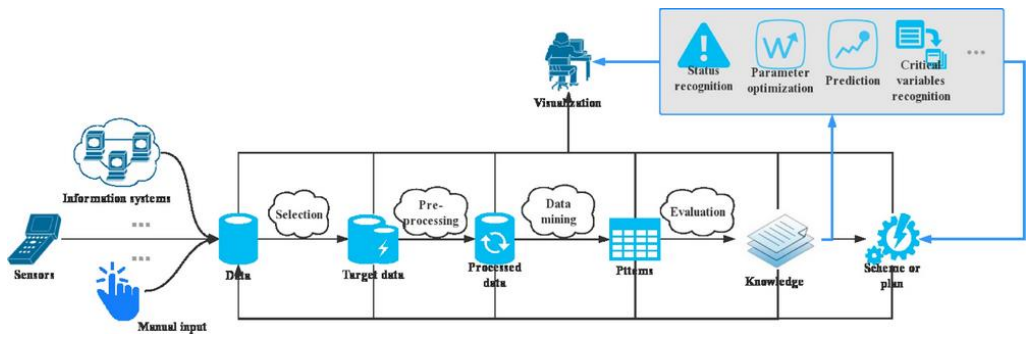
Veri ve Bilgi

- Bilgi, insanoğlu için vazgeçilmez unsurların başında gelir.
- Günümüzde **bilginin elde edilmesi, paylaşılması ve oluşturulması üzerinde teknolojik gelişmeler oldukça etkilidir.**
- Yeni teknolojilerin ortaya çıkması toplumsal yaşamın değişmesine, yeni ilişkiler ağının ortaya çıkmasına ve bilgilerin sürekli olarak yenilenmesine neden olmaktadır.
- Sözlük anlamıyla **bilgi; öğrenme, araştırma ve gözlem** yoluyla elde edilen her türlü **gerçek ve kavrayışın tümüdür.**
- Bilgi, önceden belirlenen bir dizi **sistematik kural** ve **prosedüre uygun bir biçimde işlenmiş enformasyondur.**

7

Veri ve Bilgi

- Veri ve bilgi arasındaki ilişki aşağıda görülmektedir*.



*Data and knowledge mining with big data towards smart production, Cheng, Ken Chen, Hemeng Sun, Yongping Zhang, Fei Tao, Journal of Industrial Information Integration, 9, 1-13, 2018.

8

Veri ve Bilgi

Türk Dil Kurumuna göre;

- **Veri (Data):** olgu, kavram veya komutların, iletişim, yorum ve işlem için elverişli biçimde gösterimi,
- **Enformasyon (Information):** haber alma, haber verme, haberleşme,
- **Bilgi (Knowledge):** veriye yöneltilen anlam, insan aklının erebileceği olgu, gerçek ve ilkelerin bütünü,
- **Anlayış (Understanding):** görüş ve inanış etmenlerinin etkisiyle beliren düşünme yolu, düşünüş biçimi, zihniyet, mantalite,
- **Bilgelik (Wisdom):** herkesin ulaşamadığı derin, kapsamlı, bütünsel bilgi olarak tanımlanmaktadır.

9

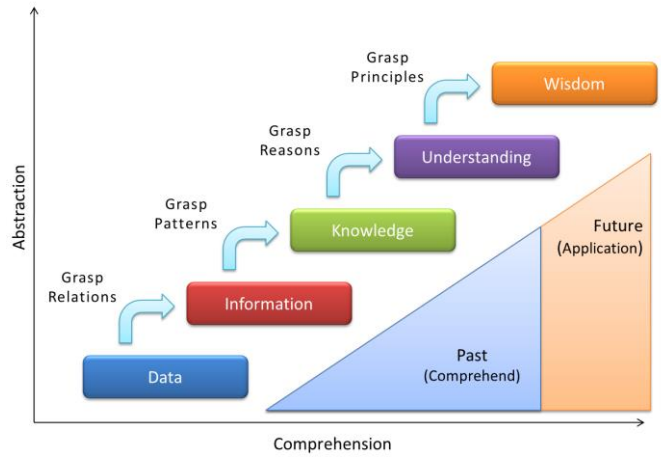
Konular

- Veri ve Bilgi
- Veri, Enformasyon, Bilgi, Anlayış, Bilgelik
- Büyük Veri
- Büyük Veri Analitiği
- Büyük Veri Kullanım Alanları
- Akış Verisi
- Akış Verisi Kaynakları

10

Veri, Enformasyon, Bilgi, Anlayış, Bilgelik

- Veri ve bilgelik arasındaki ilişki aşağıda görülmektedir*.



*https://medium.com/@Iyer/strive-to-get-higher-on-the-data-information-knowledge-understanding-and-wisdom-continuum-c5ccb96438

11

Veri, Enformasyon, Bilgi, Anlayış, Bilgelik

- **Veri (Data):** Sayılar, rakamlar, sözcükler, metinler, resimler, olaylar vb. biçiminde temsil edilen ham gerçekliklerdir. (Örn: 54000, 01/22/2006)
- **Enformasyon (Information):** herhangi bir konu ile ilgili bir bilinmeyeni giderme konusunda yardımcı olan tanımlayıcı ifadelerdir (Örn: Nazlı'nın bankada 54.000 TL'si var, Kemal'in doğum tarihi 01/22/2006).
- **Bilgi (Knowledge):** İşlenmiş enformasyondur (Örn: Nazlı'nın bankada biriken 54.000 TL'si beklediğinden fazladır).
- **Anlayış (Understanding):** Sonuç veya bilgi ile ilgili neden bulma veya kavrama sürecidir (Örn: Nazlı banka işlemlerine bakınca tanımadığı birisinin 4.000 TL yatırdığını farketti. Bu nedenle bankadaki parası yüksekmiş.).
- **Bilgelik (Wisdom):** Başka bir bakış açısıyla, değişen şartlar çerçevesinde ileriyi görebilme veya gözlem etkilerine göre prensipler ortaya koyma yeteneğidir (Bankaya para transferinde kişiden onay istenmelidir.).

12

Konular

- Veri ve Bilgi
- Veri, Enformasyon, Bilgi, Anlayış, Bilgelik
- Büyük Veri
- Büyük Veri Analitiği
- Büyük Veri Kullanım Alanları
- Akış Verisi
- Akış Verisi Kaynakları

13

Büyük Veri

- **Büyük veri,** kendine özgü özelliklere sahip olan ve genellikle yüksek hacimlerde ve çok farklı kaynaklardan elde edilen veridir.

- **Büyük veri analiz yöntemleri**, farklı kaynaklardan elde edilen düzenli veya **düzensiz verileri anlamlı ve işlenebilir hale dönüştürür.**
- **Dünyadaki verilerin %90'ı son 3-4 yılda oluşturulmuştur.**
- Sosyal medya, blog, fotoğraf, müzik, video, IoT, log dosyaları, …

14

Büyük Veri

- Büyük veri terimi ilk ortaya çıktığından itibaren farklı sayıdaki özellikler ile ifade edilmiştir.
- Büyük veri özellikleri **3V, 5V, 7V, 10V** ve hatta **42V** olarak ifade edilmiştir.
- Yaygın kullanılan 10V:
  1. **Volume**
  2. **Velocity**
  3. **Variety**
  4. **Variability**
  5. **Veracity**
  6. **Validity**
  7. **Vulnerability**
  8. **Volatility**
  9. **Visualization**
  10. **Value**

15

Büyük Veri

Volume

- **Büyük verinin en çok bilinen karakteristiğidir.**
- **YouTube'a** her bir dakikada 300 saatlik video yüklenmektedir.
- 2016 yılında 1,1 trilyon **fotoğraf** çekildiği tahmin edilmektedir.
- 2016 yıllında **cep telefonu** veri trafiğinin 6,2 exabyte olduğu tahmin edilmektedir (Byte, Kilobyte, Megabyte, Gigabyte, Terabyte, Petabyte, **Exabyte,** Zettabyte, Yottabyte, Xenottabyte, Shilentnobyte, Domegemegrottebyte, Icosebyte).
- **Twitter** kullanıcıları her bir dakikada 277.000 tweet atmaktadır.
- **Apple** kullanıcıları her bir dakikada 48.000 uygulama indirmektedir.
- **Facebook** kullanıcıları her bir dakikada 2.460.000 içerik paylaşmaktadır.
- Her bir dakikada 204.000.000 **e-posta** mesajı gönderilmektedir.
- **Google** her bir dakikada 4.000.000 arama sorgusu almaktadır.

16

Büyük Veri

Velocity

- Büyük verinin **üretilme, tüketilme, oluşturulma** ve **güncellenme hızını ifade eder.**
- Facebook günde 600 terabyte verinin geldiğini ifade etmektedir.
- Google her saniyede 40.000 sorguya cevap ürettiğini ifade etmektedir. Günde 3,5 milyar sorguya cevap verdiği söylenebilir.

Variety

- **Büyük verideki çeşitliliği ifade eder.**
- Büyük veride **yapılandırılmış**, **yarı yapılandırılmış** ve **çoğunlukla yapılandırılmamış** veri bulunur (ses, video, görüntü, sosyal medya güncellemeleri, log dosyaları, click verileri, makine ve sensör verileri vb.).

17

Büyük Veri

Variability

- **Büyük veride bazı farklı verilerde olabilir.** Bunlar, veride **tutarsızlıklara neden olabilir.**
- Bu verilerin **anomaly** veya **outlier** algılama yöntemleri ile bulunup yapılan analizlerin daha anlamlı hale getirilmesi gereklidir.

Veracity

- **Veri kaynaklarının güvenilirliğini ifade eder.**
- Büyük veride yukarıda bahsedilen özellikler artarken **verinin güvenilirliği ve doğruluğu düşer.**
- **Verinin kim tarafından oluşturulduğu**, **hangi metodoloji ile toplandığı**, **aynı türdeki kaynaklardan mı toplandığı**, **veriyi toplayanın özetleme yapıp yapmadığı**, **veri başka birisi tarafından değiştirildi mi** gibi sorulara cevap aranır.

18

Büyük Veri

Validity

- Verinin nasıl doğrulandığı ve geçerliliğinin nasıl test edildiğiyle ilgilenir.
- **Verinin analiz işleminden önce doğrulanması gereklidir.**

Vulnerability

- Büyük veri yeni güvenlik konularını da beraberinde getirir.
- **Verinin hack'lenmemesi**, kaynağından elde edildikten sonra **bozulmadan** ve güvenlik saldırısı sonucu **değişmeden** alınması gereklidir.

Volatility

- **Verinin, güncel olup olmadığı, kullanılabilir olup olmadığı ile ilgilenir.**
- Güncel veriyle istenen sonuçlar elde edilebilir.
- Kurumlarda veya büyük şirketlerde veri sürekli saklanır ve elde edilen büyük verinin önemli bir kısmı güncelliğini yitirebilir.

19

Büyük Veri

Visualization

- **Büyük verinin görselleştirilmesi sonuçların kolay anlaşılması ve analiz edilmesi için gereklidir.** Günümüzde büyük veri görselleştirmeyle ilgili hafıza kısıtları gibi teknik kısıtlar halen bulunmaktadır.
- Klasik grafik araçları ve yöntemleriyle büyük verideki milyarlarca noktanın görselleştirilmesi mümkün değildir.
- Bunun için kümeleme, ağaç haritaları, dairesel ağ diyagramları gibi görselleştirme yöntemlerinin kullanılması gereklidir.

Value

- **Büyük veriden anlamlı ve değerli bilgiyi çıkarmadıkça diğer bütün karakteristikleri anlamsızdır.**
- Anlamlı ve değerli bilgiyi elde etmek için veri madenciliği yöntemleri gibi karmaşık süreçlerin uyarlanıp kullanılması gereklidir.

20

Konular

- Veri ve Bilgi
- Veri, Enformasyon, Bilgi, Anlayış, Bilgelik
- Büyük Veri
- Büyük Veri Analitiği
- Büyük Veri Kullanım Alanları
- Akış Verisi
- Akış Verisi Kaynakları

21

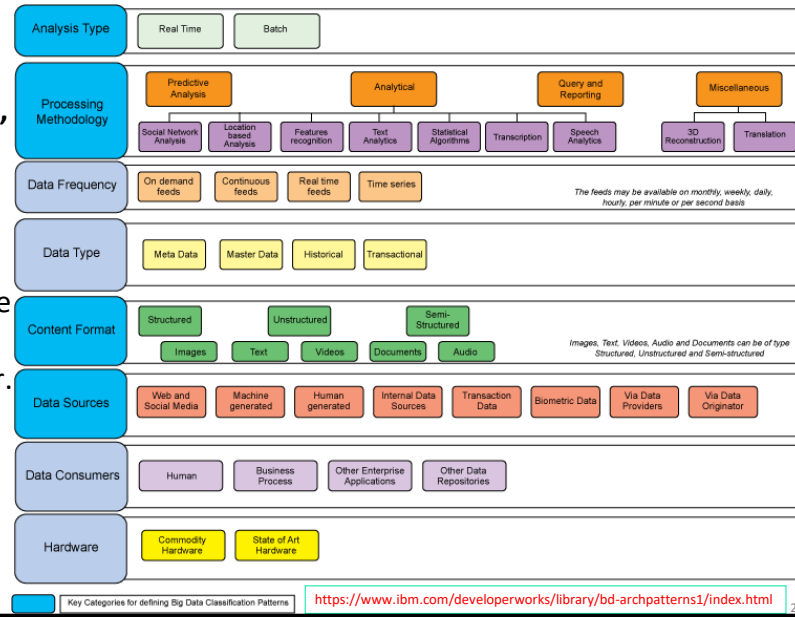
Büyük Veri Analitiği

- **Büyük veri analitiği, büyük ve çeşitli veri setleri üzerinde işlem yaparak gizli örüntüleri çıkarma, bilinmeyen ilişkileri keşfetme sürecidir.**
- Kullanılan yöntemlerle elde edilen bilgi, firmalara, kurumlara veya ticari girişimlere yönelik önemli bilgiler sağlamaktadır.
- Büyük veri analitiği uygulamaları veri bilimcilere modelleri tahmin etme, istatistikçilere ve diğer analiz alanında çalışan profesyonellere büyüyen verileri kolay analiz etme yeteneği kazandırır.
- **Büyük veri analitiği klasik yöntemlerle yönetilmesi çok zor olan çok büyük, yapılandırılmamış ve çok hızlı değişen veriyle uğraşır ve anlamlı örüntüler elde eder.**
- Büyük veri analitiği yöntemleri veriyi saklamak, veriyi elde etmek ve analiz etmek için gelişmiş teknolojiyi kullanır.

22

Büyük Veri Analitiği

Büyük veri; **analiz türü, işleme yöntemi, veri frekansı, veri türü, içerik biçimi, veri kaynağı, veri tüketicisi** ve **donanıma** göre sınıflandırılabilir.



https://www.ibm.com/developerworks/library/bd-archpatterns1/index.html

23

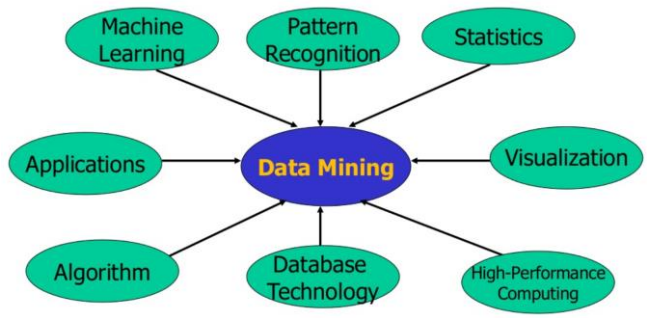
Konular

- Veri ve Bilgi
- Veri, Enformasyon, Bilgi, Anlayış, Bilgelik
- Büyük Veri
- Büyük Veri Analitiği
- Büyük Veri Kullanım Alanları
- Akış Verisi
- Akış Verisi Kaynakları

24

Büyük Veri Kullanım Alanları

- Büyük veri günümüzde, **veri madenciliği**, **makine öğrenmesi**, **örüntü tanıma**, **istatistik**, **görselleştirme**, **yüksek performanslı hesaplama**, **veritabanı teknolojisi**, **algoritma** gibi çok farklı disiplinlerde yaygın bir şekilde kullanılmaktadır.



25

Büyük Veri Kullanım Alanları

- Büyük veri uygulamaları farklı uygulama alanlarında başarılı bir şekilde kullanılmaktadır.

İşletme

- Özellikle büyük ölçekli işletmeler **müşteri analizi, müşteriye özel tavsiye, reklam veya öneri oluşturma, ürün dağıtımı** ve **lojistik optimizasyonu** gibi çok sayıdaki alanda büyük veri analiz yöntemlerini kullanmaktadır.

Perakende Satış

- **Personel gelir optimizasyonu, müşteri davranış analizi, müşteri ilişkileri analizi, ürün çeşitliliği, kampanya yönetimi** ve **fiyat optimizasyonu** gibi uygulamalarda büyük veri analiz yöntemleri kullanılmaktadır.

26

Büyük Veri Kullanım Alanları

Kamu

- **Verilere kolay ve güvenli erişebilirliği sağlama, gizlilik** ve **şeffaflık oluşturma, uygun ürün** ve **hizmetlerin sunumu, risk** ve **sahtekarlığı azaltmaya** yönelik alanlarda büyük veri analiz yöntemleri kullanılmaktadır.

Teknoloji

- **Gerçek zamanlı analiz** ve **işlem (menü) özelleştirme, işlem süresini azaltma, riskleri azaltma, otomatik sistemler ile karar verme** gibi alanlarda büyük veri analiz yöntemleri kullanılmaktadır.

Eğitim

- Eğitimde **öğrenci analizi**, **ders planlaması** gibi alanlarda büyük veri analiz yöntemleri kullanılmaktadır.

27

Büyük Veri Kullanım Alanları

Kişisel Konum Verileri

- **Konum tabanlı reklam, akıllı yönlendirme, acil müdahale** gibi alanlarda büyük veri analiz yöntemleri kullanılmaktadır.

Sağlık

- **Hastalık tespiti, hasta izlenmesi, kişisel DNA analizi** gibi alanlarda büyük veri analiz yöntemleri kullanılmaktadır.

Bankacılık

- **Geçmiş verinin, nakit hareketlerinin, öngörülebilen felaketlerin, soygunların** ve **müşteri davranışlarının** anlaşılmasında büyük veri analiz yöntemleri kullanılmaktadır.

28

Konular

- Veri ve Bilgi
- Veri, Enformasyon, Bilgi, Anlayış, Bilgelik
- Büyük Veri
- Büyük Veri Analitiği
- Büyük Veri Kullanım Alanları
- Akış Verisi
- Akış Verisi Kaynakları

29

Akış Verisi

- **Akış verisi geldiği anda işlem yapılmazsa (depolama, data process vs.) kalıcı şekilde kaybedilebilir.**
- Veriyi işleme hızından daha hızlı veri gelmesi durumunda da kaybedilebilir.
- Akış verisinde işlem yapan **algoritmalar akış verisini genellikle özetleyerek kullanırlar.**
- Akış verisi madenciliği algoritmaları, **faydalı örnekleri seçer** ve **istenmeyen örnekleri filtreler.**
- Özetleme yaklaşımında ise, sabit boyutlu bir pencere içerisindeki elemanlarla (belirli bir süre için geçmiş veri) özetleme yapılmaktadır.

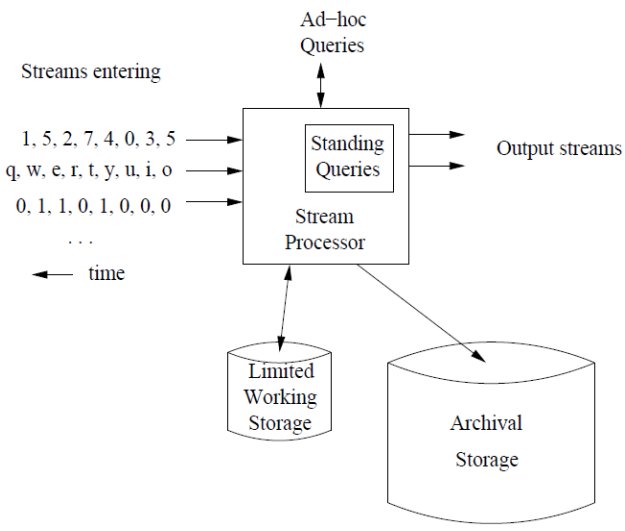
30

Akış Verisi

- Akış verisinin özetlenmesiyle birlikte **daha küçük alanda saklanması da sağlanmış olur.**
- Akış işlemcisi bir tür veri yönetim sistemi olarak görülebilir.
- Sisteme çok sayıda farklı stream'den veri gelebilir.
- Veri türleri, veri oranları ve veri gelme aralıklarının dağılımları farklı olabilir.

31

Akış Verisi



32

Konular

- Veri ve Bilgi
- Veri, Enformasyon, Bilgi, Anlayış, Bilgelik
- Büyük Veri
- Büyük Veri Analitiği
- Büyük Veri Kullanım Alanları
- Akış Verisi
- Akış Verisi Kaynakları

33

Akış Verisi Kaynakları

Sensor data

- **Bir okyanus yüzeyindeki ısı sensörü her saat ölçtüğü ısı değerini** reel sayı olarak bir istasyona gönderirse, **veri oranı çok düşük olduğundan günümüz teknolojisinde tüm veri ana hafızada tutulabilir.**
- GPS birimindeki **sensör yüzeydeki yükseklik değişimini** ölçüp bir istasyona gönderirse, bu durumda **veri oranı yüksektir** ve ancak ana hafızada veya ayrı bir diskte tutulabilir.
- **Bir okyanusun tüm davranışını ölçmek istersek**, milyonlarca sensör kullanılır ve **günlük birkaç terabyte veri alınabilir.**

34

Akış Verisi Kaynakları

Image data

- Uydulardan sürekli dünyaya ilişkin görüntüler alınıp yeryüzündeki istasyonlara gönderilir.
- Bu görüntü verilerinin boyutları günlük birkaç terabyte düzeyinde olabilir.
- **Şehirlerdeki güvenlik kameraları** uyduya göre düşük çözünürlüktedir, ancak her birisi **akış verisi oluşturur.**
- Londra'da 6 milyon kamera olduğu belirtilmektedir ve her birisi akış verisi oluşturur.

35

Akış Verisi Kaynakları

İnternet ve Web trafiği

- İnternet **anahtarlama düğümleri** (router) IP paketlerinden oluşan **stream'leri alır** ve **çıkış portlarına yönlendirme yapar.**
- Anahtarlama elemanlarının görevi sorgulama veya saklama değildir.
- Günümüzde anahtarlama elemanlarının kapasitesinin artırılmasına (DOS ataklarının algılanması, tıkanıklık denetimi yapılması) yönelik eğilim vardır.
- Web siteleri her gün milyonlarca sorgu almaktadır (Google her gün yüzlerce milyon arama sorgusu almaktadır, Yahoo milyarlarca click almaktadır.).
- Bu tür verilerden faydalı bilgiler elde edilebilir (sorgulardaki ani yükselme, click sayısındaki ani yükselme veya düşme).

36