POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Mahremiyet korumalı büyük veri yayınlama için kavramsal model önerileri

Conceptual model suggestions for privacy preserving big data publishing

***Yazar(lar) (Author(s))**: Yavuz CANBAY[1], Yılmaz VURAL[2], Şeref SAĞIROĞLU[3]*

ORCID[1]: 0000-0003-2316-7893

ORCID[2]: 0000-0002-2858-5448

ORCID[3]: 0000-0003-0805-5818

Bu makaleye şu şekilde atıfta bulunabilirsiniz(To cite to this article): Canbay Y., Vural Y. ve Sağıroğlu S., "Mahremiyet korumalı büyük veri yayınlama için kavramsal model önerileri", ***Politeknik Dergisi***, 23(3): 785-798, (2020).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: **10.2339/politeknik.535184**

Mahremiyet Korumalı Büyük Veri Yayınlama İçin Kavramsal Model Önerileri

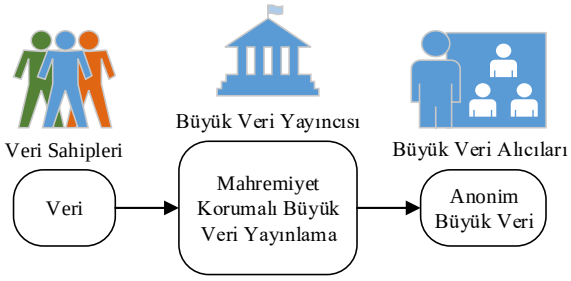
Conceptual Model Suggestions for Privacy Preserving Big Data Publishing

Önemli noktalar (Highlights)

- *Mahremiyet koruma modelleri (Privacy preserving models)*
- *Büyük veri yayınlama modelleri (Big data publishing models)*
- *Kavramsal model önerileri (Conceptual model suggestions)*
- *Mahremiyet korumalı büyük veri yayınlamada ilk kavramsal model önerileri (First conceptual model suggestions for privacy preserving big data publishing)*

Grafik Özet (Graphical Abstract)

Bu çalışmada, mahremiyet korumalı büyük veri yayınlama için kavramsal modeller önerilmiştir. In this paper, conceptual models are proposed for privacy preserving big data publishing.



Şekil. Mahremiyet korumalı büyük veri yayınlama mimarisi /**Figure.** The architectre of privacy preserving big data publishing

Amaç (Aim)

Bu çalışma, büyük veri mimarisine uygun mahremiyet korumalı veri yayınlama modellerinin geliştirilmesi amaçlamıştır. / This paper aims to develop privacy preserving big data publishing models.

Tasarım ve Yöntem (Design & Methodology)

Literatürdeki geleneksel veri için geliştirilen mahremiyet korumalı veri yayınlama modelleri araştırılarak büyük veri mimarisine göre yeniden tasarlanmıştır. / Existing privacy preserving data publishing models were reviewed and then adapted for big data architecture.

Özgünlük (Originality)

Bu çalışmada mahremiyet korumalı büyük veri yayınlama için ilk defa kavramsal modeller önerilmiştir. / In this paper, conceptual models for privacy preserving data publishing were proposed for the first time.

Bulgular (Findings)

Önerilen modellerin büyük verilerin yayınlanmasında başarı ile kullanılabilecek temel kavramsal modeller olduğu açıkça görülmektedir. / It can be clearly seen that the proposed models can be successfully employed for privacy preserving big data publishing.

Sonuç (Conclusion)

Bu çalışmada büyük veri mimarisine uygun kavramsal modeller önerilmiş ve başarıyla tasarlanmıştır. / In this paper, privacy preserving big data publishing models were proposed and successfully designed.

Etik Standartların Beyanı (*Declaration of Ethical Standards*)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler. / The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Mahremiyet Korumalı Büyük Veri Yayınlama için Kavramsal Model Önerileri

Araştırma Makalesi / Research Article

Yavuz CANBAY[1*], Yılmaz VURAL[2], Şeref SAĞIROĞLU[1]
[1]Mühendislik Fakültesi, Bilgisayar Müh. Bölümü, Gazi Üniversitesi, Türkiye
[2]Kişisel Verileri Koruma Kurumu, Türkiye

(Geliş/Received : 04.03.2019 ; Kabul/Accepted : 02.09.2019)

ÖZ

Teknolojinin gelişmesi ile beraber veri üretim ve işleme hızı artmış, bunun sonucu olarak hacim, hız, çeşitlilik ve değer gibi bileşenlere sahip büyük veri kavramı ortaya çıkmıştır. Büyük verilerden elde edilecek faydayı arttırmak için bu verilerin mahremiyetini koruyarak paylaşmak veya yayınlamak gerekir. Literatür incelendiğinde, büyük verinin mahremiyetini koruyarak yayınlanmasını kolaylaştıran herhangi bir modelin olmadığı tespit edilmiştir. Mahremiyet Korumalı Büyük Veri Yayınlama (Privacy Preserving Big Data Publishing – PPBDP) modellerinin oluşturulması, büyük veri mahremiyeti koruma sürecindeki tüm tarafların doğru bir şekilde yönlendirilmesi ve gereksinimlerinin doğru karşılanması, doğru alt yapı ve hizmetlerin oluşturulması adına önemlidir. Ayrıca, bu modelleri oluştururken maliyet ve güvenlik gibi faktörleri de göz önünde bulundurmak gerekir.

Bu çalışmada, mahremiyet korumalı geleneksel veri yayınlama modelleri araştırılmış, çeşitli kriterlere göre karşılaştırılarak mahremiyet risk seviyeleri değerlendirilmiş ve bu risk seviyelerini de dikkate alan büyük veri temelli yeni kavramsal modeller ilk defa önerilmiştir. Önerilen bu modeller senaryo temelli olarak oluşturulmuş, üstünlükleri ve dezavantajları sunulmuştur. Önerilen modellerin, büyük verilerin mahremiyetinin korunarak yayınlanması, mahremiyet risklerinin minimize edilmesi ve büyük veriden maksimum faydanın sağlanması gibi pek çok konuda katkılar sağlayacağı değerlendirilmektedir.

Anahtar Kelimeler: Mahremiyet koruma, büyük veri, mahremiyet korumalı büyük veri yayınlama modelleri, kavramsal model.

Conceptual Model Suggestions for Privacy Preserving Big Data Publishing

ABSTRACT

Recent developments in IT has increased the speed of data production and processing, as a result, big data concept with components such as volume, velocity, variety and value has emerged. In order to get more benefit from big data, it is necessary to share or publish the data by preserving or respecting privacy. The literature reviews report that there is no model that facilitates publishing big data by preserving privacy. Designing Privacy Preserving Big Data Publishing (PPBDP) models is important to direct all the parties and to meet the requirements of them correctly, and to create the right infrastructures and services. In addition, it is necessary to consider some factors such as cost and security when designing these models.

In this study, privacy preserving data publishing models were reviewed, compared based on various criteria and then evaluated based on privacy risk levels. Finally, big data architecture based new conceptual models were then established for the first time according to these evaluations and privacy risk levels. It is expected that the proposed models might contribute to the literature on some issues, such as publishing big data with preserving privacy, minimizing privacy risks and obtaining maximum benefit from the big data.

Keywords: Privacy preserving, big data, privacy preserving big data publishing models, conceptual model.

1. GİRİŞ (INTRODUCTION)

"Mahremiyet Hakkı" başlıklı makale [1] ile literatüre giren mahremiyet kavramı, günümüzde yasalarla korunan temel ve zorunlu bir ihtiyaç haline gelmiştir. Özellikle, teknolojinin her alana girdiği bir dönemde birçok siber saldırıya maruz kalan insanoğlunun mahremiyetinin korunması, bireyin kendi sınırlarını belirlediği gerçek ve sanal ortamlarda birey olabilme hakkı olarak tanımlanabilir.

Mahremiyet toplumdan topluma, kültürden kültüre ve hatta bireyden bireye değişiklik gösterebilen çok yönlü değişken bir kavramdır. Veri mahremiyeti her ne kadar ilk bakışta veri sahiplerinin mahremiyetini korumak olarak anlaşılsa da sadece bununla sınırlı olmayıp, verinin fayda sağlama özgürlüğünün de veri mahremiyeti koruma sürecine dâhil olduğu bilinmelidir.

Mahremiyet korumalı veri yayınlama (Privacy Preserving Data Publishing - PPDP) modelleri; veriden daha fazla değer üretilmesi amacıyla verinin mahremiyetini koruyarak yayınlanmasını sağlayan, veri mahremiyeti koruma sürecini ve bu süreçteki tarafları kapsayan yapılardır.

İlk defa 2012 yılında literatüre giren büyük veri kavramı, Gartner tarafından "daha iyi sezgi, karar verme ve süreç otomasyonunu mümkün kılan maliyet etkin, yenilikçi bilgi işleme biçimleri talep eden yüksek hacimli, yüksek hızlı ve/veya çok çeşitli bilgi varlıkları" [2] olarak

*Sorumlu Yazar (Corresponding Author)
e-posta : yavuzcanbay@gazi.edu.tr

tanımlanmıştır. Hacim, hız ve çeşitlilik açısından geleneksel mimari ve sistemlerle yönetilemeyen veri büyük veri olarak değerlendirilmekte, dolayısıyla yeni gereksinimleri de beraberinde getirmektedir. Yeni teknolojilerden sistemlere, yazılımlardan donanımlara, programlama modellerinden tasarımlara kadar pek çok farklılığı içeren büyük veri sistemleri, ülkelerin artık en çok yatırım yaptıkları alanlardan biri haline gelmiştir. Sağlık, eğitim, ülke yönetimi ve istihbarat gibi çeşitli alanlarda önemini arttıran büyük veri kavramı, politika belirleme, strateji geliştirme, seçim yönetimi, gelecek planlama, tahmin ve analiz yapma gibi pek çok alanda da büyük fırsatlar sunmaktadır.

Geleneksel veri alanından büyük veri alanına geçildiğinde fırsatların yanında yeni tehditlerin de ortaya çıktığı ve var olan tehditlerin ise karmaşıklaşarak mahremiyet risklerini daha da arttırdığı görülmektedir. Özellikle siber tehditlerin son zamanlarda yoğunlaştığı veri mahremiyeti alanı için, büyük veri alanında geliştirilecek çözümlerin yine bu alanın gereksinimlerine özgü olarak tasarlanması şarttır. Geleneksel verilerin mahremiyetini sağlama NP-Zor bir problem iken [3], büyük verilerde ise bu problemin daha da zorlaştığı değerlendirilmektedir.

Büyük verinin mahremiyetini koruyarak yayınlamak, analiz, araştırma, katma değer sağlama, gelecek planlamaları yapma, politikalar oluşturma gibi ülke menfaatine dokunacak çıktıların üretilmesinde büyük önem arz eder. Bunu gerçekleştirebilmek adına mahremiyet korumalı büyük veri yayınlama (Privacy Preserving Big Data Publishing – PPBDP) modellerinin oluşturulması önemlidir. PPDP modellerinin büyük veri alanına uyarlanması ile PPBDP modelleri elde edilebilir. Bu modellerin oluşturulmasında, veri mahremiyetinin korunmasına ek olarak maliyet ve güvenlik bileşenlerini de dikkate almak; doğru, güvenilir, güvenli ve faydalı modeller oluşturulmasına büyük katkı sunar. Büyük veri sistemleri yüksek maliyetli yapılar olduğu için böylesi sistemlerin farklı çatılar altında ayrı ayrı kurulması yerine tek bir merkezî yapıda kurulması maliyet açısından daha etkindir. Ayrıca bu yapılarda veri alıcısının sistemle doğrudan etkileşimi güvenlik riskini de arttıracağı için, bu aşamada kullanıcı etkileşimini dikkate almayan PPBDP modellerinin oluşturulması gerektiği bu çalışmada değerlendirilmektedir.

Bu çalışma altı bölümden oluşmaktadır. İkinci bölümde makale ile ilgili temel bilgiler sunulmuş, üçüncü bölümde temel veri yayınlama modelleri gözden geçirilmiş, dördüncü bölümde büyük veride mahremiyet problemi açıklanmış, beşinci bölümde önerilen büyük veri yayınlama modelleri tanıtılmış ve son bölümde ise sonuç ve tartışmalara yer verilmiştir.

2. TEMEL BİLGİLER (BACKGROUND INFORMATION)

Bu bölümde veri mahremiyeti, anonimleştirme ve büyük veri kavramları bu makale kapsamında önerilen modelleri daha iyi ifade etmek ve çıktılarını daha doğru anlamak adına aşağıda kısaca açıklanmıştır.

2.1. Veri Mahremiyeti (Data Prıvacy)

Veri mahremiyeti literatürde, "bilgisel seçici kontrol" [4] ve "muhatapların bilgilerinin doğru kullanımı ve muhatabın hangi bilgisinin, kiminle ve ne derecede paylaşılmasına karar verme mekanizması" [5] olarak tanımlanmıştır. Bu tanımlara ek olarak aşağıda sunulan tanımlar da konuyu daha iyi anlamaya yardımcı olacaktır;

- Veri üzerinde uygulanacak herhangi bir metot, teknik veya arka plan bilgileri ile veri sahiplerinin ifşa riskinin mümkün olduğu kadar minimize edilmesi,
- Veriden bir ya da daha fazla kişiye doğrudan veya dolaylı olarak erişilmesinin mümkün olduğu kadar önlenmesi,
- Verinin kiminle, hangi seviyede ve ne amaçla paylaşılacağına dair sınırların belirlenmesinde veri sahibinin seçici kontrolü ve
- Veriden kişiye ulaşmayı sağlayacak herhangi bir ilişkinin mümkün olduğu ölçüde ortadan kaldırılmasıdır.

Veri mahremiyeti tanımının doğru yapılması kişi, kurum ve kuruluşlarca bu kavramın özümsenmesini daha da kolaylaştıracaktır. Genellikle güvenlik ve gizlilik gibi kavramlarla karıştırılan mahremiyet kavramının yukarıda belirtilen tanımlar doğrultusunda bu kavramlardan ayrıştığı açıkça görülmektedir.

2.2. Anonimleştirme (Anonymization)

Anonimleştirme, mahremiyet gereksinimlerinin karşılanmasında sıklıkla kullanılan fayda temelli bir yaklaşımdır. Saldırganın veri sahibinin kimliğini tespit etmesini zorlaştırmak amacıyla, veri üzerinde genelleştirme, baskılama vb. teknikleri uygulayarak paylaşılan verileri ifşa saldırılarına karşı korur [6].

Anonimleştirilecek bir veri kümesinde öznitelikler, tam-tanımlayıcı, yarı-tanımlayıcı, hassas ve hassas-olmayan olmak üzere dört farklı sınıfta incelenir [7]. Tam-tanımlayıcılar; TC kimlik numarası, ad-soyad, pasaport numarası vb. gibi kişiyi doğrudan tanımlayan bilgilerdir. Yarı-tanımlayıcılar ise, kişiyi doğrudan tanımlama özelliği olmayan ancak bir araya gelmesi halinde kişiyi dolaylı olarak tanımlayabilecek öznitelik grubudur. Yaş, cinsiyet ve posta kodu özniteliklerinin oluşturduğu grup yarı-tanımlayıcılara örnek olarak verilebilir. Veri sahibine ait hassas bilgileri içeren öznitelikler hassas öznitelik olarak sınıflandırılmakta olup, hastalık, gelir, gider, din vb. bilgiler bu sınıfa örnek olarak verilebilir. Yukarıdaki sınıflandırmalar dışında kalan ve umuma açık öznitelikler ise hassas olmayan öznitelikler olarak sınıflandırılır.

Mahremiyet ihlalleri verinin toplanmasında, depolanmasında, paylaşılmasında ve işlenmesinde ortaya çıkabilmektedir [8]. Verinin toplanması ve depolanmasında güvenlik tedbirlerine ihtiyaç

duyulurken, paylaşılmasında ve işlenmesinde ise mahremiyet koruyucu önlemlere ihtiyaç duyulur. Verinin yayınlanmasında dikkat edilmesi gereken en önemli husus, saldırganların yayınlanan veriler ile başka kaynaklardan elde edeceği arka plan bilgilerini kayıt ve tablo düzeyinde birbirine bağlayarak yapacağı çıkarımlardır. Saldırganlar tarafından yapılan çıkarımlar sonucunda gerçekleşebilecek mahremiyet ihlalleri aşağıda özetlenmiştir [7];

- Kimlik ifşası: saldırganın, önceden yayınlanmış kamuya açık kimlikli veri kümesini, yayınlanan diğer kimliksiz verilerle kayıt düzeyinde eşleştirmesi sonucu kurbanın kimliğini ifşa etmesidir.
- Öznitelik ifşası: saldırganın, yayınlanan veri kümesindeki bilgilerin homojen dağılımından faydalanarak kurbanın hassas özniteliğini ifşa etmesidir.
- Üyelik ifşası: arka plan bilgisine sahip saldırgan, yayınlanan veri kümesinde, kurbana ait verilerin de olduğunu bilmesi durumunda veri bağlama yöntemleriyle kurbanın kimliğini ifşa etmesidir.

Yukarıda belirtilen mahremiyet ihlallerini en aza indirgemek adına literatürde sıklıkla kullanılan temel mahremiyet koruma modelleri aşağıda özetlenmiştir.

k-Anonimlik: yayınlanan veri kümesi içerisinde bir verinin en az *k*-1 tane veriden ayırt edilememesini sağlar ve yarı-tanımlayıcı öznitelikleri üzerinde işlem yapar. Saldırgan, kimliğini ifşa etmek istediği kurbanın yarı-tanımlayıcılarının değerini bilse bile, o kişinin verisini diğer *k*-1 tane veriden ayırt edemez. Bu şekilde kimlik saldırısı belirli bir seviyede engellenmiş olur [9].

l-Çeşitlilik: *k*-Anonimlik modeli her ne kadar kimlik ifşası saldırısına bir çözüm sunsa da, hassas öznitelik saldırısına karşı savunmasızdır. Kurbanın yarı-tanımlayıcı değerlerini bilen bir saldırgan, eşlenik sınıflardaki hassas öznitelik çeşitliliğinin az olması durumunda kişinin hassas verisine ulaşabilir ve bu şekilde hassas öznitelik saldırısı gerçekleştirebilir. Bu saldırıya karşı geliştirilen bir çözüm olan *l*-Çeşitlilik modeli, her bir eşlenik sınıf içerisinde en az *l* adet en iyi temsil edilmiş hassas verinin olmasını sağlar [10].

t-Yakınlık: tüm verideki hassas özniteliklerin dağılımı göz önüne alındığında, *l*-Çeşitlilik modelinin öznitelik saldırısına yeteri kadar çözüm sunmadığı görülür. Örneğin, bir hassas verinin tüm tabloda oranı %5 iken, bir eşlenik sınıfı içerisindeki oranı %50 ise bu durumda ciddi bir mahremiyet ihlali ortaya çıkabilir. *t*-Yakınlık modeli, eşlenik sınıflardaki herhangi bir hassas özniteliğin dağılımını tüm verideki dağılımına bir *t* değeri kadar yakın olmasını sağlar [7].

δ-Mevcudiyet: *k*-Anonimlik, *l*-Çeşitlilik ve *t*-Yakınlık modelleri kimlik ve hassas öznitelik saldırılarına karşı koruma sağlarken üyelik saldırılarına karşı koruma sağlayamaz. Üyelik bilgisinin keşfini zorlaştırarak mahremiyet riskini azaltmak amacıyla, Nergiz ve arkadaşları *δ*-Mevcudiyet modelini önermiştir [11]. Bu modeldeki temel yaklaşım yayınlanan veri kümesini, saldırganın arka plan bilgisini temsil eden genel veri kümesinin bir alt kümesi olarak modellemektir.

2.3. Büyük Veri Kavramı Ve Büyük Veri Teknolojileri (Big Data and Big Data Technologies)

Dijitalleşen dünyanın ürettiği önemli bir konsept olan büyük veri, sosyal medyadan güvenlik sistemlerine, sağlıktan finansa kadar pek çok alanda hayatın bir parçası haline gelmiştir. Veri büyük olunca üretilecek çıktıların değerinin de büyük olması beklenir [12]. Büyük veriden beklenen değeri üretmek için veriyi kendi yapısına uygun bir şekilde analiz etmek en doğrusudur.

Büyük veri, geleneksel sistemlerin sınırlarını aşan ve kabul edilebilir bir sürede işlenemeyen karmaşık verilerin yönetimi, analizi, depolanması, anlamlandırılması, görselleştirilmesi gibi konularda karşılaşılan zorlukların aşılması amacıyla ortaya çıkan bir kavramdır [13-15]. Hacim, hız, çeşitlilik ve değer bileşenlerine sahip büyük veri kavramına, zamanla değişen problemlerin ihtiyacına uygun olarak değer, değişkenlik, zafiyet ve görselleştirme gibi bileşenler de eklenmiştir [15-19].

Büyük veri analitiğinde Hadoop ve Spark literatürde en çok kullanılan anaçatı teknolojileridir. Hadoop, Hadoop Dosyalama Sisteminde (Hadoop File System - HDFS) tuttuğu verileri MapReduce yapısına uygun olarak işlenmesine; Spark ise geçici bellekte Esnek Dağıtılmış Veri Kümesi (Resillient Distributed Dataset – RDD) olarak tuttuğu verileri MapReduce yapısının uyarlanmış bir versiyonu ile işlenmesine imkân sağlar. Hangi teknolojinin neye göre tercih edileceği probleme bağlıdır. Örneğin; akan veri üzerinde bir analitik uygulaması yapılacaksa Spark'ı kullanmak Hadoop'a göre daha avantajlı iken, diskte duran bir veriyi işlemede Hadoop Spark'a göre daha avantajlı olabilir [20, 21]. Aşağıda büyük veri işlemede kullanılan temel teknolojiler ve yapılar daha detaylı açıklanmıştır.

Spark: Bilgisayar kümeleri üzerinde büyük veri analitiği yapılmasına imkân sağlayan, hızlı, etkili, hata-toleranslı ve ölçeklenebilir bir büyük veri işleme platformudur. Hızlı veri işlemenin yanı sıra akan veri işleme imkânı da sağlaması çoğu projede önemli bir tercih sebebidir. Geçici bellekte tutulan ve RDD adı verilen dağıtık kayıt dosyaları üzerinde işlem yaparak veri analizini hızlandırmayı amaçlar. Klasik MapReduce yapısını kendi amacı doğrultusunda uyarlayarak dağıtık işlem yapılmasını sağlar. Replika tabanlı çalışan bu yapı mevcut işlenen veride herhangi bir bozulma olması halinde replikasından yedekleme yaparak işlemlerin devam etmesini sağlar [22].

RDD: Spark için sunulan veri soyutlama yapısıdır. Verinin geçici hafızadaki soyutlanmış halini temsil eden dağıtık kayıt dosyalarıdır [22].

Hadoop: HDFS dağıtık dosya sistemi ve MapReduce dağıtık programlama modeli yapılarını barındıran, büyük veriyi analiz etme ve işlemede kullanılan, güvenilir, hızlı, ölçeklenebilir ve dağıtık hesaplamaya imkân sağlayan bir

açık kaynak sistemidir. Yapısal ve yapısal olmayan veri kümeleri için geliştirilmiştir. Büyük veri kümelerini küçük parçalara bölerek her bir parçanın bulunmuş olduğu düğüm üzerinde işlenmesini sağlar. Bilgisayar kümeleri arasında petabayt veya exabayt boyutlarına varan büyük verilerin, dağıtık ve paralel işlenmesine izin vererek bir tek makineden yerel işlemleri yürüten binlerce makineye işleri ölçeklendiren veya çoğaltan bir yapıdır [23, 24].

HDFS: Hadoop mimarisinde kullanılan dağıtık dosyalama sistemi olup uygulamalardaki tüm giriş ve çıkış verilerini depolayan yapıdır. Temel mimarisi efendi-köle örüntüsüne dayanır. Bir adet NameNode (efendi) ve çok sayıda DataNode (köle) düğümünden oluşur. NameNode, paylaşılan dosya sistemi için üst bilgi yapısını saklar ve giriş/çıkış işlemleri için DataNode'u yönlendirir. Sistemdeki düğümlerin çökmesi halinde sistemin devamlılığını sağlama adına NameNode kullanılarak düğümler tekrardan aktif hale getirilir. Veri çok sayıda parçalara bölünerek farklı düğümlere dağıtılır ve bir verinin birden fazla kopyası olduğu için herhangi bir düğümde sorun oluşması halinde diğer düğümlerde tutulan mevcut kopyalar kullanılır. Bir HDFS dosyasında okuma/yazma işlemleri sırasında dosya bloklara bölünür ve NameNode hangi bloğun hangi DataNode'da bulunduğu bilgisini tutar. Her bir DataNode başka bir DataNode ile veya NameNode ile iletişime geçebilir [24].

3. TEMEL MAHREMİYET KORUMALI VERİ YAYINLAMA MODELLERİ (PRIVACY PRESERVING DATA PUBLISHING MODELS)

Veri yayınlama, mevcut verinin çeşitli kişi, kurum veya kuruluşlarla çeşitli politika ve stratejilere göre paylaşılmasıdır. Bu paylaşımın temel amacı veriden elde edilecek faydayı olabildiği kadar yüksek seviyeye çıkararak veri sahibi veya veri yayıncısı olan kişi, kurum veya kuruluşlar için bir katma değer sağlamaktır.

ile iletişimin nasıl olacağı; veri alıcısı tarafında ise veri yayıncısı tarafı ile olan etkileşim gibi konular temel veri yayın senaryoları için önemli bazı kriterlerdir.

Günümüzde verilerin geleneksel (normal) veri veya büyük veri olarak sınıflandırılmasında dolayı, geliştirilen bir sistem veya modelin hangi sınıfa uyumlu olduğunu belirtmek de yerinde olacaktır. Çünkü bu veri sınıfları için geçerli alt yapı, sistem ve mimariler birbirinden tamamen farklıdır.

Geleneksel veri mimarisindeki veri yayınlama modelleri için örnek bir sistem alt yapısı Şekil 1'de gösterilmiştir. Verilen şekilde, veri sahiplerinden toplanan veriler merkezî mimariye sahip veri yayıncısı tarafında merkezî olarak depolanır, merkezî olarak anonimleştirilerek anonim veri kümesi oluşturulur ve veri alıcısına iletilir.

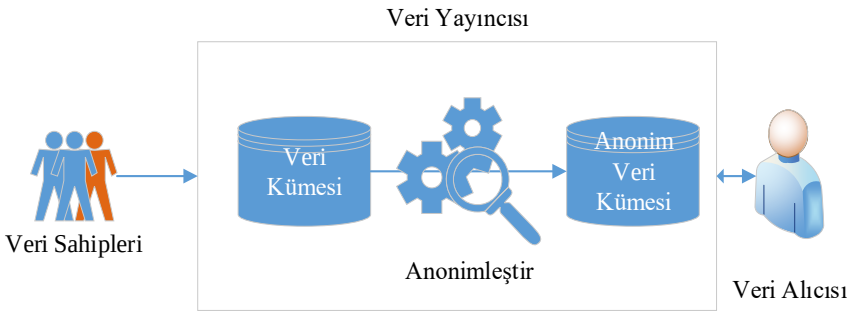
Bu bölümde, literatürde geleneksel veri mimarisi için geliştirilen temel PPDP modelleri açıklanmış, şekillerle örneklendirilmiş ve son olarak birbirleriyle mahremiyet riskleri açısından karşılaştırılmıştır. Literatürdeki temel PPDP modellerinin sınıflandırılması ise Şekil 2'de gösterilmiştir [25-28].

3.1. Tek Yayıncılı Veri Yayınlama Modelleri (Single Publisher Data Publishing Models)

Tek bir veri yayıncısının olduğu veri yayınlama modelidir. Veri yayınlama politikası sadece bir yayıncı tarafından oluşturulur ve yönetilir. Örneğin, bir A kurumu sahip olduğu çeşitli verileri analizler yapılması amacıyla belirli politikalar çerçevesinde yayınlayabilir. Bu durumda bu kurum tek veri yayıncısı olarak nitelendirilir. Bu model, yayınlanan veri sürümlerine göre farklı alt sınıflara ayrılmakta olup bunlar aşağıda özetlenmiştir.

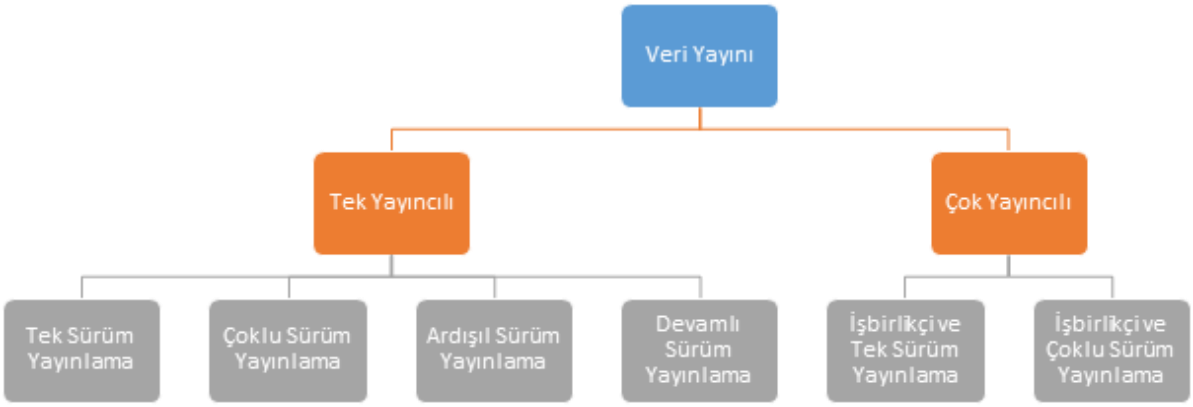
3.1.1. Tek Sürüm Yayınlama (Single Release Publishing)

Yayınlanacak veri kümesi sadece tek sürüm ve bir defaya mahsus olarak yayınlanır. Veri, yayınlandıktan sonra



Şekil 1. Geleneksel veri yayınlama modellerinde sistem alt yapısı (System infrastructure of traditional data publishing models)

Veri yayınlarken, mahremiyet koruma sürecindeki tüm tarafların gereksinimlerini en iyi şekilde karşılayan modellere ihtiyaç duyulur. Veri sahibi tarafında hangi verilerin paylaşılacağına karar verilmesi; veri yayıncısı tarafında hangi verilerin toplanacağı, hangi özniteliklerin hangi sınıflara atanacağı, nasıl bir mimari üzerinden bu işlemlerin gerçekleştirileceği, veri sahibi ve veri alıcısı ekleme, çıkarma, güncelleme gibi herhangi bir işleme tabi tutulmaz. 2010 yılına ait diyabet verileri, 2018 yılına ait kredi kartı harcama verileri gibi tek defaya mahsus yayınlanan veriler bu modele örnek olarak verilebilir. Kullanıcı etkileşimi içermeyip, veri yayıncısı kendi belirlediği stratejiye göre yayınlayacağı veriye ve özniteliklerine karar verir. Veri alıcısının ihtiyaç



Şekil 2. Temel mahremiyet korumalı veri yayınlama modellerinin sınıflandırılması (Classification of privacy preserving data publishing models) [25-28]

duyduğu veri alanından fazlasının yayınlanması durumu oluşabileceği için, böylesi bir durumda hem gereksiz bilgi paylaşımı yapılmış olur hem de yayınlanan veri saldırıya açık hale gelir. Yani saldırgan bilmesi gerekenden daha fazla öznitelik bilgisine sahip olacağı için çıkarım yapma ihtimali de artar. Bu ihtimali düşürmek için veri yayıncısının, yayınlayacağı verilerdeki öznitelikleri doğru bir şekilde belirlemesi gerekir [25].

Şekil 3'de bu model bir örnek üzerinde gösterilmiş olup, verilen şekilde orijinal veri tablosu T(Yaş, Meslek, Doğum Yeri, Hastalık) ile temsil edilirken, anonimleştirilen ve paylaşılan veri tablosu ise T'(Yaş, Meslek, Doğum Yeri, Hastalık) ile gösterilmiştir.

3.1.2. Çoklu Sürüm Yayınlama (Multiple Release Publishing)

Şekil 4'de gösterilen bu modelde veri alıcıları, bir veri tablosundan kendi amaçları doğrultusunda kullanmak istedikleri özniteliklere ait alt veri grubunu talep eder. Tek sürüm yayınlama modelinden farklı olarak tüm verinin alıcılara iletilmesi yerine, veri alıcılarının talepleri dikkate alınır ve bu şekilde belirli bir alt veri grubu iletilmiş olur. Örnek olarak; bir T(Yaş, Meslek, Doğum Yeri, Hastalık) veri tablosu üzerinde, bir araştırmacı kendi araştırmasında kullanmak üzere Meslek, Doğum Yeri ve Hastalık bilgilerini talep ettiğinde T^A(Meslek, Doğum Yeri, Hastalık) anonim veri tablosu araştırmacıya iletilirken, bir başka araştırmacı Yaş, Meslek ve Doğum Yeri bilgilerini talep ettiğinde ise ona da T^B(Yaş, Meslek, Doğum Yeri) anonim veri tablosu iletilir. Bu modelde hem ihtiyaç duyulmayan alanların yayınlanması engellenir hem de talep edilen özniteliklere göre veri anonimleştirilip paylaşılacağı için bilgi kaybı düşük olur [25].

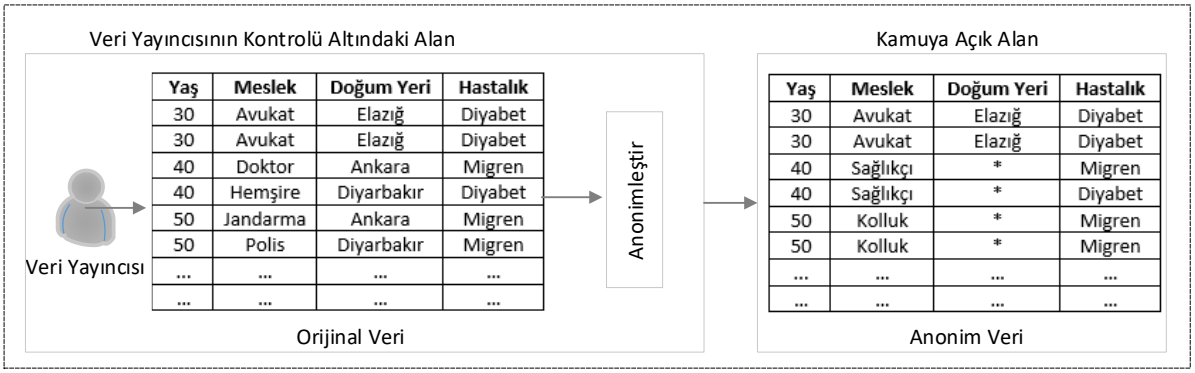
Bu modelde, bir saldırganın aynı anda birden fazla sürüme ulaşması halinde yarı-tamamlayıcılar üzerinden bağlantı kurarak hassas bilgi çıkarımında bulunması, önemli bir mahremiyet riski oluşturur. Ayrıca kullanıcının sistemle etkileşimi sonucu güvenlik zafiyetinin oluşma ihtimali de artar.

3.1.3. Ardışıl Sürüm Yayınlama (Sequential Release Publishing)

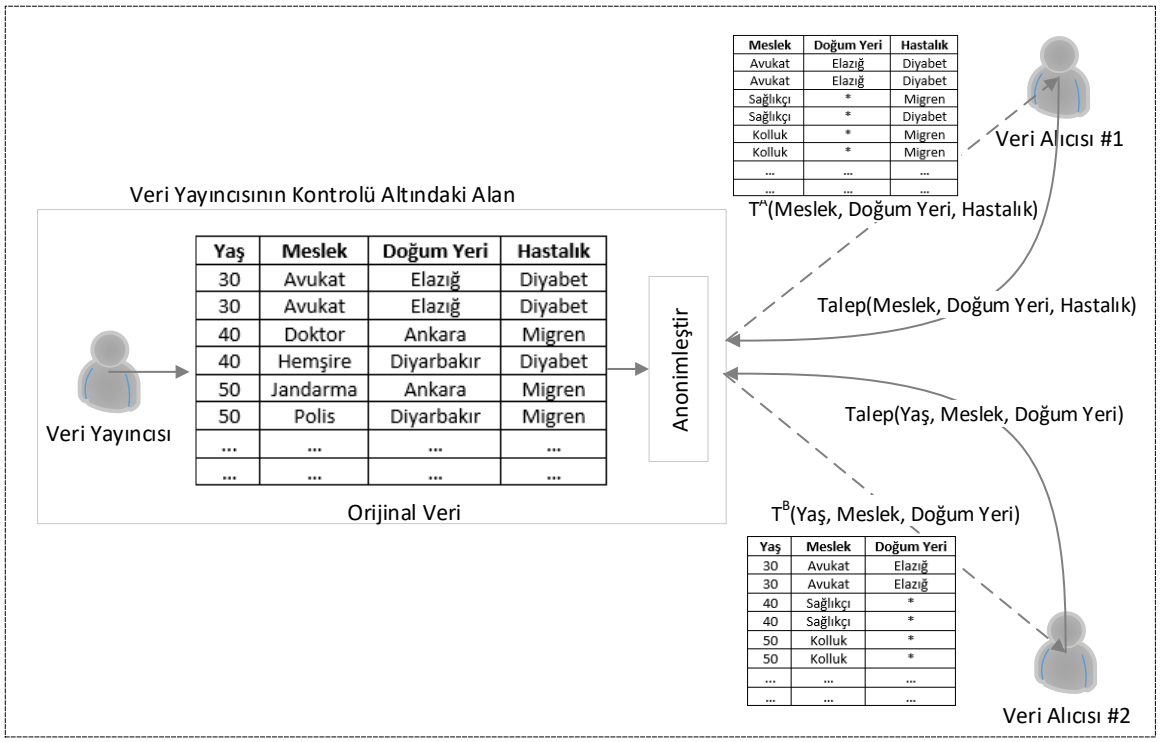
Dinamik veri yayınlama modellerinden biri olup Şekil 5'de bir örnek üzerinde gösterilmiştir. Bu modelde yeni bilgiler geldiği sürece veri ardışıl olarak yayınlanır. Kullanıcı etkileşimi söz konusu değildir. Yayınlanacak bir veri tablosu için kayıt sayıları aynı kalmak şartıyla, öznitelikleri barındıran sütunlar her bir sürümde değişir. T(Yaş, Meslek, Doğum Yeri, Hastalık) veri tablosu için; birinci sürümde T^A(Meslek, Doğum Yeri, Hastalık) anonim veri tablosunun, ikinci sürümde ise T^B(Yaş, Meslek, Doğum Yeri) anonim veri tablosunun yayınlanması bu modele örnek olarak verilebilir. Saldırgan, yayınlanan önceki sürümlerden faydalanarak yeni yayınlanan sürümdeki verileri birleştirip ifşa saldırıları yapabilir. Bunu önlemek için, veri yayıncısı bir sonraki yayınlayacağı sürümü öncekilerle birleştirerek anonimleştirebilir. Böylelikle sürümler arasındaki birleşim ilişkisi ortadan kaldırılmaya çalışılır [28].

3.1.4. Devamlı Sürüm Yayınlama (Continuos Release Publishing)

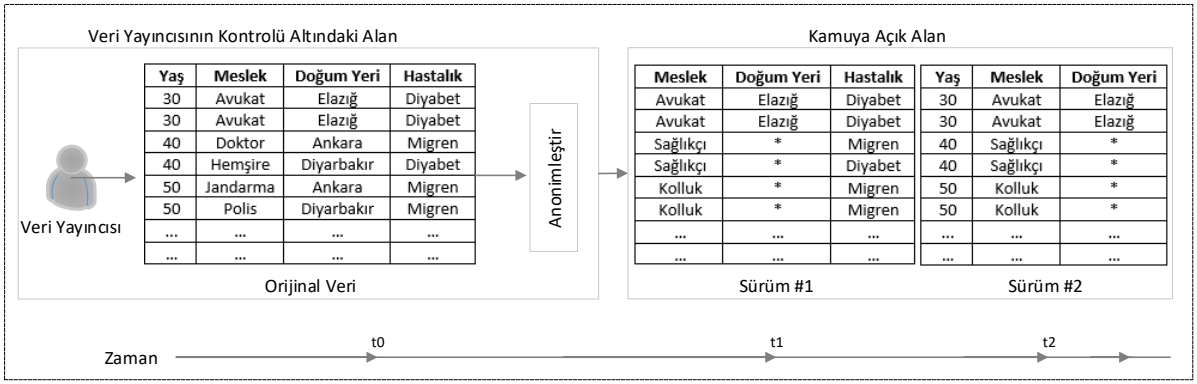
Dinamik veri yayınlama modellerinden biri olup Şekil 6'da gösterilmiştir. Veri yayıncısı yayınlamak istediği verideki öznitelikleri sabit tutarak her bir sürümde farklı kayıtlar ekler, siler veya güncelleştirir. Bu şekilde yayınlanan sürümler arasında farklılıklar oluşur. Ancak yine de saldırgan, sürümlerin zaman damgası bilgisine ve yarı-tanımlayıcı değerlerine sahipse bir mahremiyet ifşası gerçekleştirebilir [27].



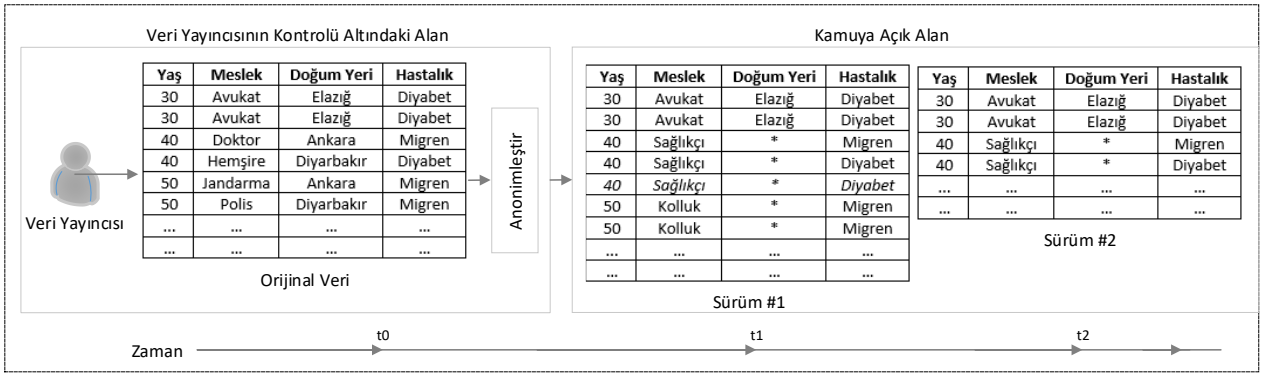
Şekil 3. Tek sürüm veri yayınlama modeli (Single release data publishing model)



Şekil 4. Çoklu sürüm veri yayınlama modeli (Multiple release data publishing model)



Şekil 5. Ardışıl sürüm veri yayınlama modeli (Sequential release data publishing model)



Şekil 6. Devamlı sürüm veri yayınlama modeli (Continuous release data publishing model)

3.2. Çok Yayıncılı Veri Yayınlama Modelleri (Multiple Publisher Data Publishing Models)

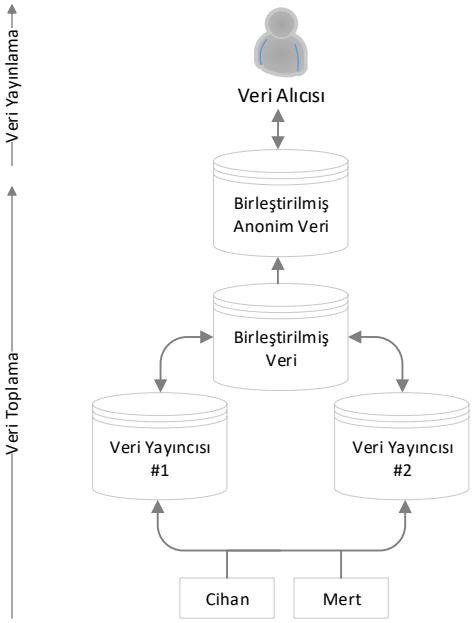
Her ne kadar veriler bazen tek yayıncılar tarafından yayınlansa da, çoğu zaman bu durum çoklu veri yayıncılarına doğru evrilebilmektedir. Yani veriler tek bir varlık tarafından yayınlanmaktan ziyade birden fazla varlık kendi verilerini yayınlamak isteyebilir. Böylesi modeller için, "birleştir-anonimleştir" ve "anonimleştir-birleştir" olmak üzere iki farklı yaklaşım geliştirilmiştir. Birleştir-anonimleştir yaklaşımında veri yayıncılarda tutulan veriler önce birleştirilir sonra anonimleştirilir ve yayınlanırken, anonimleştir-birleştir yaklaşımında ise yayıncılarda tutulan veriler önce anonimleştirilir sonra birleştirilir ve yayınlanır [25].

3.2.1. İşbirlikçi ve Tek Sürüm Yayınlama (Collaborative and Single Release Publishing)

Şekil 7'de gösterilen bu model birden fazla veri yayıncısından oluşur. Belirli amaçlar doğrultusunda aynı kişilere ait farklı verilerin birleştirilerek yayınlanması ilkesine dayanır. Örneğin; X hastanesi T^X(TC, Meslek, Doğum Yeri, Hastalık) tablosuna, Y hastanesi ise T^Y(TC, Yaş, Meslek, Doğum Yeri) tablosuna sahip olsun. Bu durumda iki tablonun birleşmesi ile T(TC, Yaş, Meslek, Doğum Yeri, Hastalık) tablosu elde edilir. Bu şekilde veri yayıncısında birleştirilen veri tablosu anonimleştirildikten sonra yayınlanabilir [25].

3.2.2. İşbirlikçi Ve Çoklu Sürüm Yayınlama (Collaborative and Multıple Release Publıshıng)

Veri yayıncısı ile veri alıcısı arasında bir etkileşim söz konusudur. Dağıtık halde bulunan veri yayıncılarda tutulan veriler, veri alıcısının talep ettiği öznitelikler çerçevesinde veri yayıncıları tarafından anonimleştirilir ve veri alıcısına iletilir. Bu modelde, kendi verilerini barındıran ve yayınlamak isteyen çok sayıda veri yayıncısı vardır. Veriler bu dağıtık yapılarda yatayda veya dikeyde bölümlenmiş olarak bulunabilir. Önemli



Şekil 7. İşbirlikçi ve tek sürüm veri yayınlama modeli (Collaborative and single release data publishing model)

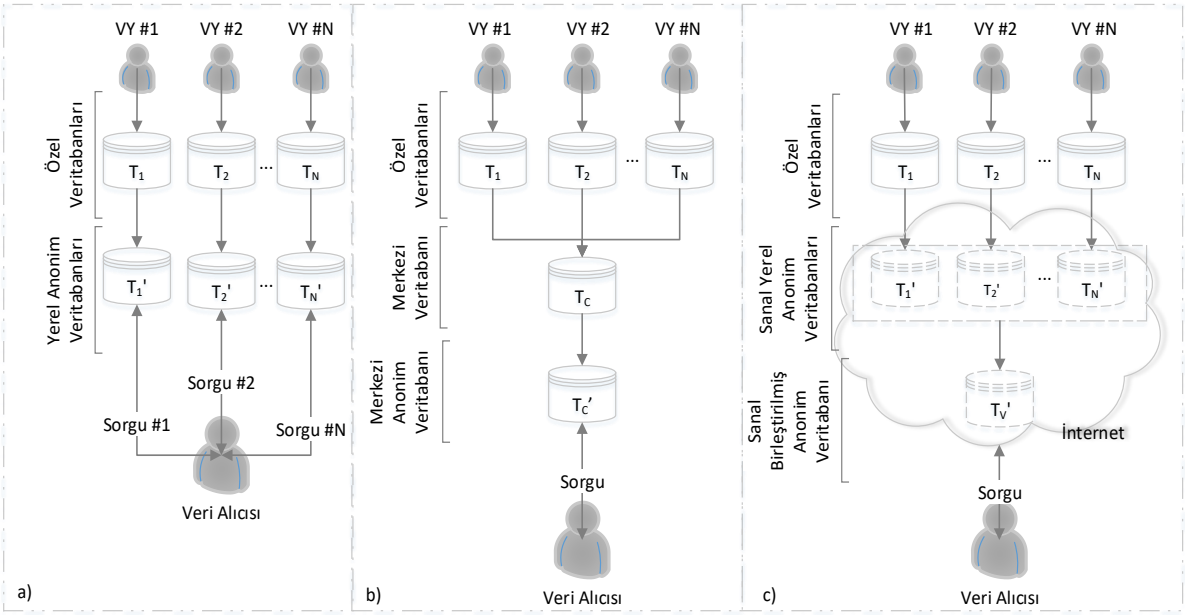
olan bu verilerin kullanıcı istekleri doğrultusunda anonimleştirilerek yayınlanmasıdır [29].

Bu modelde birleştirme, anonimleştirme ve verinin tutulduğu ortam üzerine literatürde çeşitli yaklaşımlar geliştirilmiş ve bunlar Şekil 8'de detaylı olarak sunulmuştur. Şekil 8.a'da her bir veri yayıncısı sahip olduğu verileri diğerlerinden bağımsız olarak kendi tarafında anonimleştirir. Veri alıcısı bir tablodan sorgu yapabileceği gibi birden fazla tablodan da sorgu yapabilir. Bu yaklaşımın dezavantajı, saldırgan farklı veri tablolarına aynı anda ulaşabileceği için çıkarım yapma ihtimali artar. Şekil 8.b'de ise veri yayıncıları verilerini merkezî bir yapıya iletir ve burada birleştirilen veriler anonimleştirilerek paylaşılır. Sorgular merkezî bir veri tabanı üzerinden gerçekleştirilir. Bu yaklaşımın dezavantajı ise güvenilir bir merkezî yapının bulunmasının zor olması ve böylesi bir merkezî yapının saldırıya uğraması halinde tüm verilerin doğrudan saldırganların eline geçme ihtimalinin olmasıdır. Şekil

8.c'de ise yatay bölümlenmiş veri tabanlarını barındıran veri yayıncıları sanal, bütünleşik ve anonim bir veri tablosu üretmek için dağıtık bir protokole dâhil olurlar. Yerelde anonimleştirilen her bir veri tabanı güvenli dağıtık protokoller kullanılarak birleştirilir. Böylelikle veri alıcısının bu birleşik veri tabanı üzerinden sorgu yapması sağlanır. Bu yaklaşımda sorgular doğrudan veri yayıncılarına gitmediği için güvenlik bir üst seviyede sağlanır. Ancak buna rağmen her bir veri tabanı kendi içerisinde anonimleştirildiğinden dolayı bilgi kaybı yüksek olur.

4. BÜYÜK VERİDE MAHREMİYET (PRIVACY IN BIG DATA)

Büyük verinin elde edilmesinden yayınlanmasına kadar geçen süreç, bir büyük veri mahremiyeti koruma sürecini temsil eder. Bu süreç gerek tarafları gerekse de bu tarafların rollerini ve sorumluluğu altında gerçekleşen işlemleri kapsar. Büyük veri mahremiyeti koruma sürecindeki taraflar büyük veri sahibi, büyük veri toplayıcısı/yayıncısı ve büyük veri alıcısı olarak üç gruba ayrılabilir.



Şekil 8. İşbirlikçi ve çoklu sürüm veri yayınlama modeli (Collaborative and multiple release data publishing model)

Çizelge 1. Tek sürüm veri yayınlama modellerinin karşılaştırılması (Comparision of single release data publishing models)

Yayınlama Modeli	Kullanıcı Sorgusu	Kimlik Saldırısı	Öznitelik Saldırısı	Üyelik Saldırısı	Mahremiyet Riski
Tek Sürüm	Yok	Düşük	Düşük	Düşük	**Düşük**
Çoklu Sürüm	Var	Orta	Orta	Orta	**Orta**
Ardışıl Sürüm	Yok	Yüksek	Yüksek	Yüksek	**Yüksek**
Devamlı Sürüm	Yok	Düşük	Düşük	Yüksek	**Yüksek**
İşbirlikçi ve Tek Sürüm	Yok	Düşük	Düşük	Düşük	**Düşük**
İşbirlikçi ve Çoklu Sürüm	Var	Orta	Orta	Orta	**Orta**

Yukarıda belirtilen tüm veri yayınlama modelleri Çizelge 1'de çeşitli kriterlere göre karşılaştırılarak mahremiyet riskleri açısından değerlendirilmiştir. Mahremiyet açısından kimlik, öznitelik ve üyelik saldırısı ihtimalleri; güvenlik açısından ise kullanıcı sorgusunun olup olmaması karşılaştırma kriterleri olarak belirlenmiş ve bu kriterler her bir model için değerlendirilmiştir. Buna göre; tek sürüm yayınlama modeli ile işbirlikçi ve tek sürüm yayınlama modelinin mahremiyet açısından en düşük riski barındırdığı sonucuna ulaşılmıştır.

Büyük veri sahibi, hassas bilgi içeren veri kümesinin toplandığı taraf olmakla beraber herhangi bir kişi, kurum veya kuruluş olabilir. Hassas bilgileri barındıran büyük verilere sağlık verileri, finans verileri, banka verileri, telekomünikasyon verileri, kredi kartı verileri, sigorta verileri ve eğitim verileri örnek olarak verilebilir. Bu verilerin ortak noktası ise bir kişiyi doğrudan veya dolaylı olarak tanımlayabilen veya ifşa riski taşıyan bilgiler içermesidir.

Büyük veri toplayıcısı/yayıncısı ise çeşitli varlıklardan büyük verileri toplayarak sahip olduğu teknik, teknoloji ve alt yapı ile bu verileri anonimleştirerek yayınlayan

taraftır. Veri sahiplerinden topladığı büyük veriyi dağıtık olarak depolar ve bu veriler üzerinde dağıtık mimariye uygun geliştirilen anonimleştirme tekniklerini uygulayarak dağıtık anonim veri kümelerini elde eder. Sonrasında ise bu verilerini birleştirerek büyük anonim veri kümesini oluşturur. Güvenilir bir merkezî yapı olabileceği gibi çeşitli anlaşmalarla hukuka uygun bir zeminin oluşturulduğu güvenilir üçüncü taraf yapısı da olabilir. Örneğin, Sosyal Güvenlik Kurumu (SGK) kendi bünyesinde güvenilir bir merkezî yapı iken, bir A firması güvenilir olduğu kabul edilen üçüncü taraf yapısı olabilir.

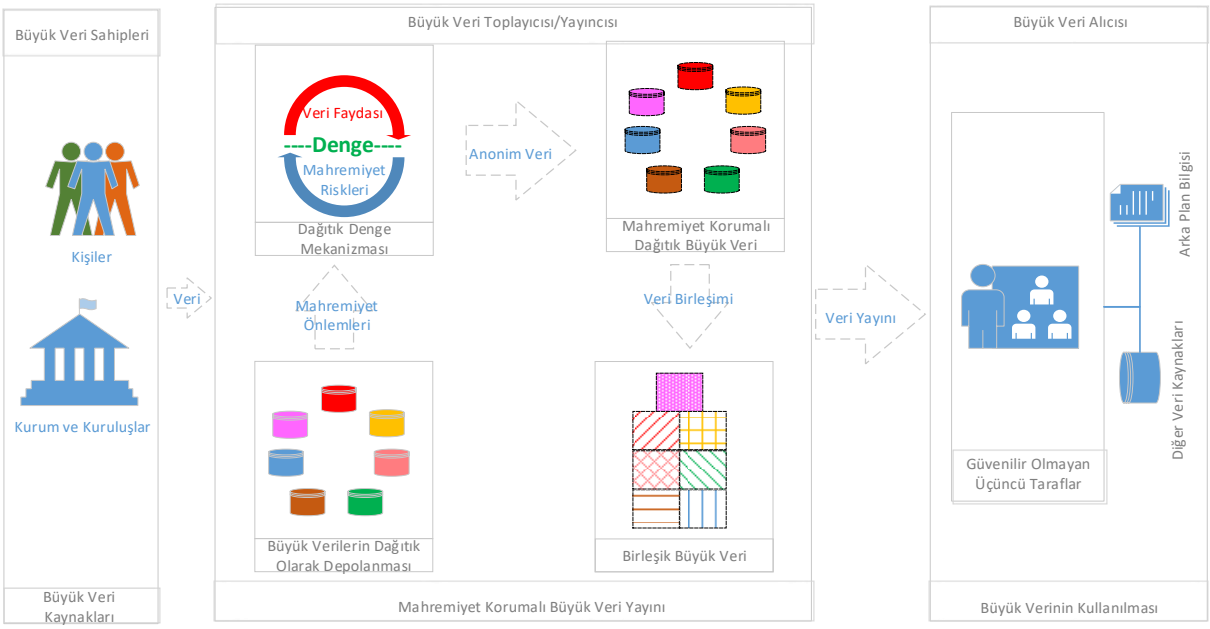
Büyük veri alıcısı büyük veriyi analiz ederek değer üretmesi beklenen ve güvenilir olmadığı kabul edilen üçüncü taraf yapısıdır. Büyük veri işleme alt yapı ve teknolojisine sahiptir. Aksi halde yayınlanan büyük veriyi işleme yetkinliğine sahip olamaz. Genelde beklenen durum, büyük veri yayıncısından aldığı anonim büyük veriyi analiz ederek değer üretmesidir. Ancak bu varlık eğer kötücül bir niyet taşıyorsa ve mahremiyet istenilen düzeyde korunmamışsa kimlik saldırısı, öznitelik saldırısı ve üyelik saldırısı gibi pek çok saldırıyı gerçekleştirebilme potansiyeline sahip olabileceği için güvenilir olmayan bir varlık olarak nitelendirilir.

Geleneksel veride mahremiyet koruma süreci, bu süreçteki taraflar ve rolleri Vural (2017)'dan [30] alınarak büyük veri konseptine uyarlanmış ve büyük veri mahremiyet koruma süreci bu çalışma kapsamında ilk defa oluşturulmuştur. Şekil 9'da gösterilen bu süreçte, büyük veri sahiplerinden toplanan veriler büyük veri toplayıcılarda anonim hale getirilir ve sonrasında ise büyük veri alıcılarına iletilir. Burada büyük verilerin temin edildiği taraflar büyük veri kaynaklarıdır. Mahremiyet korumalı büyük veri yayını yapan ve büyük veri toplayıcısı/yayıncısı rolünde olan taraf ise veriye çeşitli mahremiyet koruyucu yaklaşımlar uygulayarak veri faydası ile mahremiyet seviyesi arasında bir denge bulmayı amaçlar ve yayınlanacak olan büyük verinin mahremiyetini koruyarak büyük veri alıcıları ile paylaşır. Büyük veri alıcılar ise büyük veriden değer üreten taraftır.

Hassas ve kişisel bilgiler içerebilen büyük verinin paylaşılması sırasında gerekli tedbirler alınmadığında mahremiyet ihlallerinin ve diğer olumsuzlukların yaşanması kaçınılmazdır. Kişileri doğrudan hedef alan mahremiyet saldırılarının sonucunda, hassas bilgiler ifşa olur ve bu durum veri sahiplerinin mağduriyet yaşamasına yol açar. Özellikle sağlık, eğitim, ticari, sosyal medya gibi bireyin doğrudan dâhil olduğu alanlarda toplanan büyük veriler üzerinde gerçekleşebilecek mahremiyet ihlallerinin derinlemesine irdelenmesi gerekir.

5. ÖNERİLEN MAHREMİYET KORUMALI BÜYÜK VERİ YAYINLAMA MODELLERİ (THE PROPOSED PRIVACY PRESERVING BIG DATA PUBLISHING MODELS)

Literatür incelendiğinde, büyük veri kapsamında mahremiyet korumalı veri yayınlama modellerinin oluşturulmadığı veya önerilmediği görülmüştür. Ancak büyük veri konseptine uygun veri yayınlama modellerinin oluşturulması günümüz şartlarında bir ihtiyaçtır. Çünkü büyük veri konseptini içeren bir dijital dünyada mahremiyet gereksinimlerinin yanı sıra, verinin

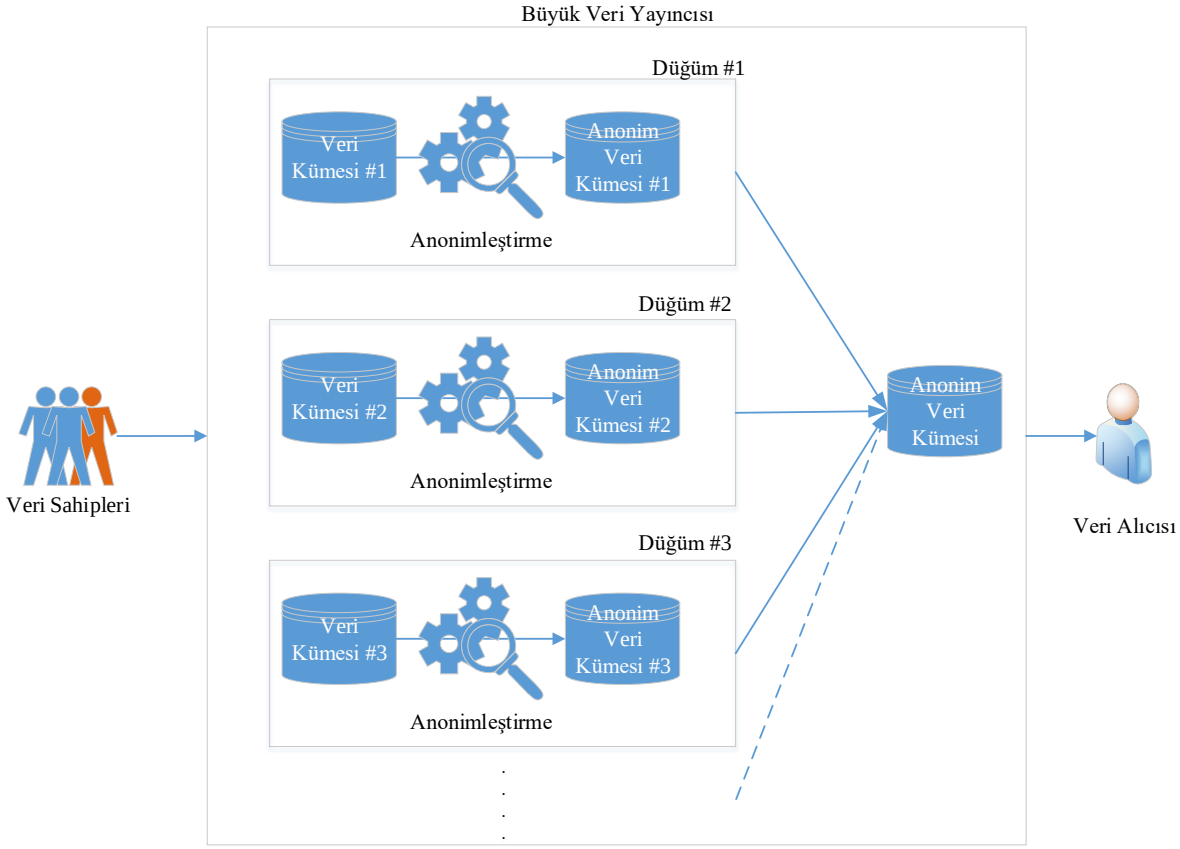


Şekil 9. Büyük veride mahremiyet koruma süreci (Privacy preserving process in big data)

alınmasından anonimleştirilip yayınlanmasına kadarki süreçte aktif rol alan tüm tarafların gereksinimlerinin

doğru anlaşılması, doğru alt yapı ve sistemlerin oluşturulması, maliyet açısından etkin ve güvenli

anonimleştirilerek birleştirilen verinin sunacağı faydaya göre daha yüksektir.



Şekil 10. Büyük veri yayınlama modellerinde sistem alt yapısı (System infrastructure of big data publishing model)

mekanizmaların geliştirilmesi büyük önem arz etmektedir.

Büyük veri mimarisine uygun olarak geliştirilecek veri yayınlama modelleri için örnek bir sistem alt yapısı Şekil 10'da gösterilmektedir. Verilen şekilde, veri sahiplerinden toplanan veriler dağıtık mimariye sahip büyük veri yayıncısı tarafında dağıtık olarak depolanır, dağıtık olarak anonimleştirilerek elde edilen anonim veri kümeleri birleştirilerek büyük anonim veri kümesi elde edilir ve son olarak büyük veri alıcısına iletilir.

PPBDP modellerinin oluşturulmasında dikkat edilmesi gereken iki önemli unsur vardır. Bunlardan ilki merkezî bir yayın modelinin olması gerekliliğidir. Bunun iki temel sebebi vardır;

- Büyük verileri yerelde geleneksel sistemlerle işlemek mümkün değildir. Bunun için merkezî bir yerde konuşlandırılan büyük veri sistemlerine ve teknolojilerine ihtiyaç vardır. Büyük veri sistemleri maliyeti yüksek yapılar olduğu için merkezî bir yerde bu sistemlerin kurularak verilerin bu merkeze çekilmesi maliyet açısından daha avantajlıdır,
- Merkezî bir yapıda birleştirilerek anonimleştirilen verinin sunacağı fayda dağıtık yapılarda

Dikkat edilecek ikinci unsur ise anonimleştirme işleminde kullanıcı etkileşiminin olup olmayacağıdır. Kullanıcı etkileşiminin olması durumunda cevap süresi önemli bir kriter olacağı için Spark gibi hızlı teknolojilere ihtiyaç duyulacaktır. Ancak kullanıcı etkileşimi olmayan bir model tercih edilecekse burada zaman kısıtı söz konusu olmayacağı için Hadoop teknolojisi de gerekli ihtiyaçları karşılayabilir. Ayrıca kullanıcı etkileşiminin olduğu bir sistemde çeşitli güvenlik riskleri de oluşabileceği için bu aşamada sadece veri yayıncısı özelinde temel modeller oluşturulmuş, kullanıcı etkileşimi dikkate alınmamıştır. Ek olarak, verinin farklı sürümlerini yayınlamak saldırganlara yayınlanan veriler üzerinde çeşitli çıkarımlar yapma imkânı da sunacağı için bu çalışmada sadece tek sürümlü modeller oluşturulmuştur.

Çizelge 1'de sunulan karşılaştırma dikkate alındığında; mahremiyet risk derecesi en az olan modellerin tek sürüm yayınlama ile işbirlikçi ve tek sürüm yayınlama olduğu görülmektedir. Bu değerlendirme dikkate alınarak, belirtilen iki ana model özelinde büyük veri yayınlama modelleri oluşturulmuştur. Her ne kadar diğer modeller de mevcut ihtiyaçlar doğrultusunda büyük veri kapsamında değerlendirilebilir olsa da, mahremiyet risklerinin düşük seviyede tutulması için çeşitli

önlemlere her zaman ihtiyaç vardır. Bu kapsamda önerilen büyük veri yayınlama modelleri ve bunların sınıflandırılması Şekil 11'de verilmiştir.

Şekil 11. Önerilen mahremiyet korumalı büyük veri yayınlama modellerinin sınıflandırılması (Classification of the proposed privacy preserving big data publishing model)

5.1. Tek Sürüm Büyük Veri Yayınlama Modeli (Single Release Big Data Publishing Model)

Büyük veri yayıncısının kendi bünyesindeki alt varlıklarından topladığı verileri birleştirip anonimleştirerek yayınladığı modeldir. Alt varlıklara ait veriler normal veya büyük veri formunda olabilir. Önemli olan bu verilerin merkezî bir yapıda birleştirilerek büyük veri formundaki verilerin elde edilmesidir. Bu modelde, dağıtık halde tutulan veriler merkezî yapıya aktarılır, veri yayıncısı bu verileri dağıtık olarak depolar ve büyük veri platformunda dağıtık anonimleştirme teknikleriyle anonim hale getirerek veri alıcısına iletir. Büyük veri sadece veri yayıncısı tarafından yönetilir ve tek sürüm olarak bir defaya mahsus yayınlanır. SGK'nın tüm kamu hastanelerinden hasta verilerini toplayarak belirli politika ve stratejilere göre tüm diyabet hastalarının 2018 yılına ait verilerini mahremiyet korumalı olarak yayınlaması bu modele örnek olarak verilebilir.

Önerilen tek sürüm büyük veri yayınlama modeli Şekil 12'de gösterilmiştir. Bu model her ne kadar işbirlikçi veri yayını yaklaşımını andırsa da, merkezî yapıdaki büyük veri tabanı ve işbirliğinde bulunan alt varlıklar tek bir kuruma ait olduğu için bu modelin tek veri yayıncılı veri yayınlama modeli sınıfına girdiği değerlendirilmektedir. Bu model yapılacak detaylı bir tasarıma göre daha çok özelleştirilebilir.

Önerilen model genel olarak aşağıdaki adımlardan oluşur;

1. Dağıtık varlıklarda tutulan veriler merkezî yapıda bulunan büyük veri tabanına güvenli olarak aktarılır,
2. Aktarılan veriler büyük veri platformunda dağıtık olarak depolanır,
3. Belirlenen özniteliklere göre yayınlanacak büyük veri dağıtık olarak anonimleştirilir,
4. Dağıtık olarak tutulan anonim veri kümeleri birleştirilir ve
5. Büyük anonim veri kümesi büyük veri alıcısına iletilir.

Önerilen model değerlendirildiğinde, avantajları ve dezavantajları aşağıda verilmiştir;

- Avantajlar:
  - Güvenilir bir merkezî yapı vardır,
  - Verinin üretilmesinden yayınlanmasına kadar geçen tüm süreçte veri güvenliği merkezî yapı tarafından sağlanır,
  - Büyük veri tek sürüm olarak yayınlanır ve böylece saldırganların çıkarım yapma ihtimali en aza indirgenir,
  - Merkezî yapıda gerçekleştirilen anonimleştirme, dağıtık yapılarda gerçekleştirilen anonimleştirmeye göre daha yüksek veri faydası sunar,
  - Kullanıcı etkileşimi olmadığı için daha güvenlidir,
  - Hangi özniteliklerin yayınlanacağı veri yayıncısının tasarrufunda olduğu için daha esnektir,
  - Hangi verilerin hangi amaç doğrultusunda yayınlanacağı merkezî yapı tarafından belirlenir,
  - Merkezî bir yapı olmasından dolayı maliyet açısından etkindir.
- Dezavantajı: ağda akan verilerin getirebileceği yük bir dezavantajıdır. Ancak veriden sağlanacak fayda göz önüne alındığında bu dezavantaj göz ardı edilebilir.

5.2. İşbirlikçi ve Tek Sürüm Büyük Veri Yayınlama Modeli (Collaborative and Single Release Big Data Publishing Model)

Önerilen bu model birden fazla veri yayıncısından oluşur ve belirli amaçlar doğrultusunda verilerin birleştirilerek yayınlanması ilkesine dayanır. Farklı veri yayıncılarının güvenli iletişim ortamı üzerinden ilettikleri veri tabanları üçüncü tarafa yapısında birleştirilerek büyük veri tabanlarında dağıtık olarak saklanır. Mahremiyet korumalı büyük veri yayını yapabilmek için verinin merkezî bir yapıda toplanarak işlenmesi gerekmektedir. Önerilen modelin çalışma prensibi şu şekilde örneklendirilebilir; A, B ve C Telekom firmaları kendilerinde hatları bulunan ortak 10 milyon abonenin iletişim örüntüsünün çıkarılarak daha iyi teklifler sunabilecekleri abonelerini belirlemek istesin. Bu durumda her bir firma kendilerinde tuttukları aynı özniteliğe sahip verileri güvenilir üçüncü taraf yapısına iletir. Bu veriler güvenilir üçüncü taraf yapısında yatayda birleştirilerek anonimleştirilir. Şekil 13'de gösterilen bu modelde, güvenilir üçüncü taraf yapısında büyük veri alt yapısı mevcuttur.

Önerilen model genel olarak aşağıdaki adımlardan oluşur;

1. Dağıtık halde bulunan veri yayıncılarda tutulan veriler, güvenli bir ortam üzerinden güvenilir üçüncü taraf yapısına aktarılır,
2. Aktarılan bu veriler birleştirilir,
3. Birleştirilen veriler büyük veri platformunda dağıtık olarak depolanır,
4. Belirlenen özniteliklere göre yayınlanacak büyük veri dağıtık olarak anonimleştirilir,
5. Dağıtık olarak tutulan anonim veri kümeleri birleştirilir ve
6. Büyük anonim veri kümesi büyük veri alıcısına iletilir.

Önerilen model değerlendirildiğinde, avantajları ve dezavantajları aşağıda sunulmuştur;

- Avantajları:
  - Dağıtık halde bulunan veri yayıncılarda tutulan verilerin merkezî yapıya iletilmesi işleminde tüm iletişim güvenli bir protokol üzerinden yapılır,
  - Veri tek sürüm olarak yayınlanır ve böylece saldırganların çıkarım yapma ihtimali en aza indirgenir,
  - Merkezî yapıda gerçekleştirilen anonimleştirme işlemi, dağıtık yapılardaki anonimleştirmeye göre daha yüksek veri faydası sunar,
  - Merkezî bir yapı olduğu için maliyet olarak etkindir ve
  - Kullanıcı etkileşimi olmadığı için güvenlidir.
- Dezavantajları:
  - Güvenilir bir merkezî yapı yoktur. Bundan dolayı güvenilir kabul edilen üçüncü bir yapıyı bulmak her zaman kolay olmayabilir,
  - Farklı veri yayıncıları farklı format ve özniteliğe sahip verileri topladıkları için standart olmayan bir veri kümesi ortaya çıkacaktır. Böylesi bir veriyi standart hale getirmek ise maliyetli bir işlemdir ve
  - Veri yayın politikası güvenilir üçüncü taraf yapısınca belirlenir.

6. SONUÇ VE DEĞERLENDİRMELER (RESULT AND EVALUATIONS)

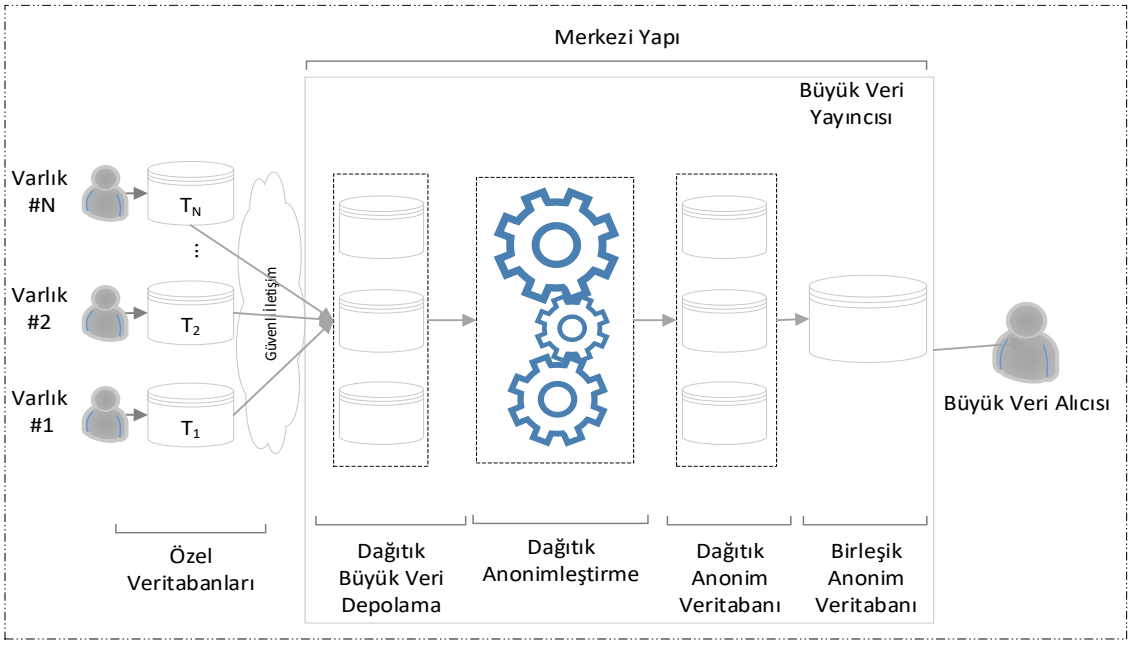
Mahremiyet korumalı büyük veri yayınlama modellerinin oluşturulması, büyük verilerin mahremiyetinin korunarak paylaşılmasıyla büyük veriden elde edilecek değerlerin maksimum seviyeye çıkarılması adına önemlidir. Bu çalışmada mahremiyet korumalı geleneksel veri yayınlama modelleri araştırılmış, bunlardan mahremiyet riski en düşük olduğu değerlendirilen iki model kullanılarak iki farklı mahremiyet korumalı büyük veri yayınlama modeli kavramsal olarak önerilmiş ve önerilen modellerin hem avantajları hem de dezavantajları sunulmuştur.

Bu çalışmanın ve önerilen PPBDP modellerinin sunduğu katkılar genel olarak aşağıda değerlendirilmiştir;

- Önerilen modellerin büyük verilerin yayınlanmasında başarı ile kullanılabilecek temel kavramsal modeller olduğu ve bu modellere ilave olarak çeşitli parametreleri de dikkate alan yeni modellerin geliştirilebileceği,
- Önerilen ilk modelin özellikle kamuda uygulanması daha uygun iken, önerilen ikinci modelin daha çok anonimleştirme hizmeti alımı şeklinde kullanımının uygun olacağı,
- Merkezî yapıda anonimleştirilen verinin sunduğu fayda dağıtık yapılarda anonimleştirilen verinin sunduğu faydaya göre daha yüksek olduğu için yeni modellerin geliştirilmesine ihtiyaç olacağı,
- Önerilen modellerin üstünlükleri olsa da çeşitli kısıtlarının da bulunduğu dikkate alınarak yeni modellerin geliştirilebileceği,
- Kavramsal modellerin uygulanması ve gerçekleştirilmesi sonrası elde edilen sonuçlara göre bu modellerin revize edilmesi gerektiği,
- Özellikle ülkemizde "Açık Büyük Veri" kavramını uygulamalarla desteklemek adına önerilen modellerin kamu kurum ve kuruluşlarınca dikkate alınarak kullanılabileceği,
- Pek çok kurumun barındırdığı büyük veri kümelerinin mahremiyetinin korunarak yayınlanması aşamasında önerilen modellerin referans modeller olabileceği ve
- Büyük veri bileşenlerinin mahremiyet açısından değerlendirilmesi ile gelecekte oluşturulacak yeni PPBDP modellerinin daha etkin, odaklı ve veri faydası yüksek modeller olacağı değerlendirilmektedir.

Kişisel Verileri Koruma Kanunu (KVKK) ve Genel Veri Koruma Yönetmeliği (GDPR) dikkate alındığında, kişisel bilgileri içeren verilerin paylaşılmasında verinin anonim hale getirilmesi bir zorunluluktur. Belirtilen yasa ve düzenlemeler veri mahremiyetini koruma ve mahremiyet korumalı veri paylaşımlarını daha dikkatli yapmayı gerektirdiğinden, önerilen kavramsal modellerin önemi açıkça görülmektedir.

2015-2018 Bilgi Toplumu Stratejisi ve Eylem Planı, 2016-2019 Ulusal e-Devlet Stratejisi ve Eylem Planı, 2013-2014 Ulusal Siber Güvenlik Stratejisi ve Eylem Planı, 11. Kalkınma Planı ve 2013 yılında yayımlanan Açık Yönetim Ortaklığı Girişimi Başbakanlık Genelgesi'nde belirtilen, mevcut kamu verilerinin anonimleştirilmesi, mahremiyetinin korunarak paylaşılması ve kamu verilerinin belirli bir kısmının kamuya açılması gerekliliğinin ortaya konulması ile



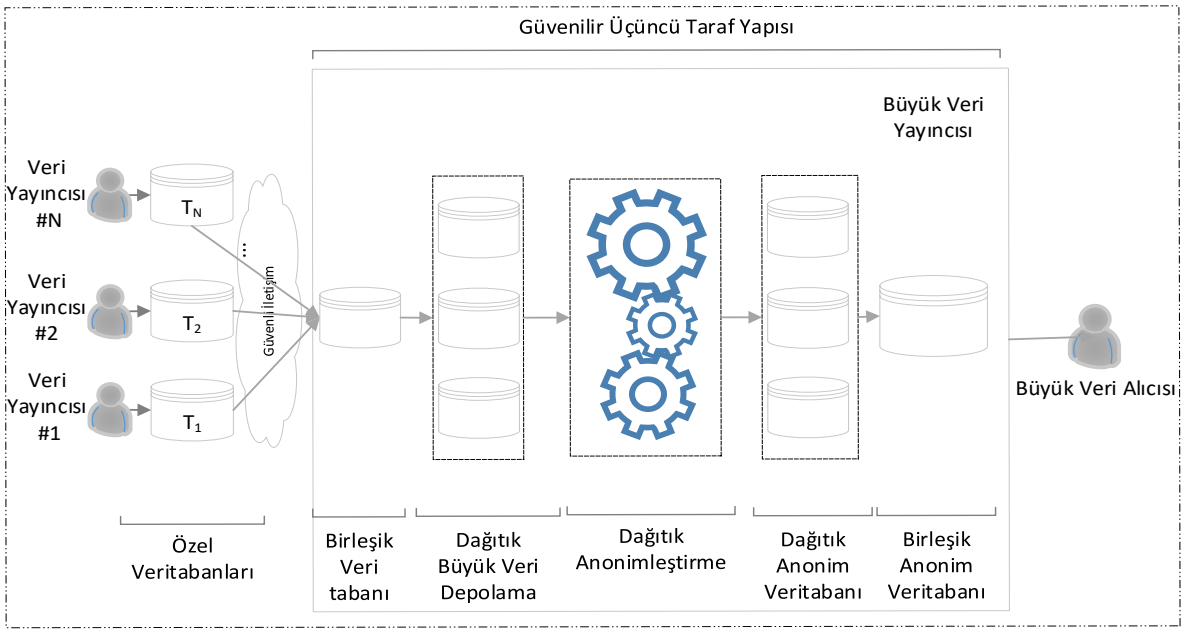
Şekil 12. Önerilen tek sürüm büyük veri yayınlama modeli (The proposed single release big data publishing model)

beraber bu çalışmada önerilen kavramsal modellere ihtiyacın olduğu görülmektedir.

Sonuç olarak; ülkemizde kişisel bilgiler içeren büyük verilerden daha fazla değer üretmek adına mahremiyet korumalı büyük veri yayınlama modellerinden mutlaka istifade edilmeli, kanun ve yönetmeliklerde belirtilen "istisnalar" ve "anonimleştirme" seçeneklerinin de bulunduğu unutulmadan ar-ge ve inovasyon çalışmaları yürütülmeli ve bu kapsamda büyük verilerin mahremiyetini koruyup yayınlayarak bu verilerden mümkün olabildiği kadar yüksek seviyede değer üretilmesi sağlanmalıdır.

TEŞEKKÜR (ACKNOWLEDGEMENT)

Gazi Üniversitesi Büyük Veri ve Bilgi Güvenliği Laboratuvarı (BIDISEC) ve Gazi Üniversitesi Bilimsel Araştırma Projeleri Birimi'ne bu çalışma kapsamında verdikleri destekten ötürü teşekkür ederiz.



Şekil 13. Önerilen işbirlikçi ve tek sürüm büyük veri yayınlama modeli (The proposed collaborative and single release big data publishing model)

KAYNAKLAR (REFERENCES)

[1] Warren S.D. and L.D. Brandeis, "The right to privacy", ***Harvard law review***, 193-220, (1890).

[2] Beyer M.A. and Laney D., "The importance of 'big data': a definition", ***Stamford***, CT: Gartner, (2012).

[3] Scott A., Srinivasan V., and Stege U., "k-Attribute-Anonymity is hard even for k=2", ***Information Processing Letters***, 115(2): 368-370, (2015).

[4] Chibba M. and Cavoukian A., "Privacy, consumer trust and big data: Privacy by design and the 3 C'S", ***IEEE ITU Kaleidoscope: Trust in the Information Society***, (2015).

[5] Jain P., Gyanchandani M., and Khare N., "Big data privacy: a technological perspective and review", ***Journal of Big Data***, 3(1): 25, (2016).

[6] Zhang X., et al., "A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud", ***IEEE Transactions on Parallel and Distributed Systems***, 25(2): 363-373, (2014).

[7] Fung B.C., et al., "Introduction to privacy-preserving data publishing: Concepts and techniques", ***CRC Press***, (2010).

[8] Mehmood, A., et al., Protection of big data privacy. ***IEEE access***, 4: 1821-1834, (2016).

[9] Zakerzadeh, H., Aggarwal, CC. and Barker, K., "Privacy-preserving big data publishing", ***International Conference on Scientific and Statistical Database Management***, Kaliforniya, ABD, (2015).

[10] Machanavajjhala, A., et al., "l-diversity: Privacy beyond k-anonymity", ***IEEE International Conference on Data Engineering***, Georgia, ABD, (2006).

[11] Nergiz, M.E., Atzori, M. and Clifton, C., "Hiding the presence of individuals from shared databases", ***ACM SIGMOD International Conference on Management of Data***, Beijing, China (2007).

[12] Chen H., Chiang R.H., and Storey V.C., "Business intelligence and analytics: From big data to big impact", ***MIS***, 36(4), (2012).

[13] Nandini K.S. and Pratheek T, "Providing anonymity using top down specialization on Big Data using hadoop framework", ***IEEE India Conference***, India, (2015).

[14] Patil H.K. and Seshadri R., "Big data security and privacy issues in healthcare", ***IEEE International Congress on Big Data***, Alaska, ABD, (2014).

[15] Victor N., Lopez D., and Abawajy J.H, "Privacy models for big data: a survey***", International Journal of Big Data Intelligence***, 3(1), 61-75, (2016).

[16] Zhang X., et al., "A MapReduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud", ***IEEE International Conference on Cloud and Green Computing,*** Karlsruhe, Germany, (2013).

[17] Li W. and Li H., "LRDM: Local Record-Driving Mechanism for Big Data Privacy Preservation in Social Networks", ***IEEE International Conference on Data Science in Cyberspace***, Changsha, China, (2016).

[18] Olaronke I. and Oluwaseun O., "Big data in healthcare: Prospects, challenges and resolutions", ***IEEE Future Technologies Conference***, Kaliforniya, ABD, (2016).

[19] Tanwar M., Duggal R. and Khatri S.K, "Unravelling unstructured data: A wealth of information in big data. in Reliability", ***IEEE International Conference on Infocom Technologies and Optimization***, Noida, India, (2015).

[20] Samadi Y., Zbakh M. and Tadonki C., "Comparative study between Hadoop and Spark based on Hibench benchmarks", ***International Conference on Cloud Computing Technologies and Applications***, Marrakech, Morocco, (2016).

[21] Lee M.S., et al., "Design of educational big data application using spark", ***IEEE International Conference on Advanced Communication Technology***, Bongpyeong, South Korea, (2017).

[22] İnternet: Apache Spark, http://spark.apache.org, (2017).

[23] Jam M.R., et al., "A survey on security of Hadoop***", IEEE International Conference on Computer and Knowledge Engineering, Mashhad***, Iran, (2014).

[24] Sogodekar M., et al., "Big data analytics: hadoop and tools", ***IEEE Bombay Section Symposium***, Baramati, India, (2016).

[25] Fung B., et al., "Privacy-preserving data publishing: A survey of recent developments", ***ACM Computing Surveys***, 42(4), 14, (2010).

[26] Jurczyk P. and Xiong L., "Distributed anonymization: Achieving privacy for both data subjects and data providers", ***Annual Conference on Data and Applications Security and Privacy***, Montreal, Canada, (2009).

[27] Byun, J.W., et al., "Secure anonymization for incremental datasets", ***Workshop on Secure Data Management***, Seoul, Korea, (2006).

[28] Wang, K. and Fung, B., "Anonymizing sequential releases", ACM SIGKDD ***International Conference on Knowledge Discovery and Data Mining***, Pensilvenya, ABD, (2006).

[29] Jurczyk, P. and Xiong, L., "Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers", ***DBSec***, 5645: 191-207, (2009).

[30 Vural, Y., "p-Gain: Privacy Preserving Utility-based Data Publishing Model", ***Ph. D. Thesis***, Department of Computer Engineering, Hacettepe University, (2017).