

E-Ticaret Trendlerinin Büyük Veri (Big Data) ile Analizi

Fatih Soysal

Özet

Büyük veri, dünya ekonomisini ve toplumsal değişimleri yönlendiren yeni bir itici güçtür. Dünyadaki veri birikimi, karar verme, sağlık yönetimi, şehir planlama, finans ve eğitim alanlarında yeni yaklaşımlar getirebilecek büyük teknolojik dönüşümler için bir dönüm noktasına ulaşmaktadır. Verinin hacmi, çeşitliliği, hızı ve doğruluğu gibi karmaşıklıklar artsa da, asıl etki Büyük Veri Analitiği teknolojileri aracılığıyla verideki 'değeri' ortaya çıkarma yeteneğimize bağlıdır. Büyük Veri Analitiği, farklı, karmaşık ve büyük ölçekli veri kümelerinden değer elde etmek için yüksek ölçeklenebilir algoritmaların ve sistemlerin tasarımında büyük bir zorluk teşkil etmektedir.

Anahtar Kelimeler: Analitik, Algoritmalar, Veri Depolama, Bulut Bilişim, Nesnelerin İnterneti (IoT), Büyük Veri İşleme

Giriş

Veri depolamanın olmadığı bir dünya düşünün: bir kişi veya kuruluşla ilgili her detay, gerçekleştirilen her işlem veya belgelenebilecek her unsur, kullanımdan hemen sonra kaybonsun. Bu durumda kuruluşlar, değerli bilgi ve bilgiyi çıkarma, ayrıntılı analizler yapma ve yeni fırsatlar ile avantajlar sağlama yeteneklerini kaybederler. Müşteri adları ve adreslerinden, mevcut ürünlere, yapılan satın alımlardan, işe alınan çalışanlara kadar her şey, günlük işleyişin sürekliliği için hayati önem taşır. Veri, her kuruluşun gelişiminin temel yapı taşıdır. Günümüzde teknolojiye ve internet alanındaki gelişmelerle birlikte sağlanan veri ve bilgi miktarındaki patlama düşünüldüğünde, depolama kapasitelerinin ve veri toplama yöntemlerinin artmasıyla büyük miktarda veri kolayca erişilebilir hale gelmiştir. Her saniye daha fazla veri üretilmekte ve değer elde etmek için depolanması ve analiz edilmesi gerekmektedir. Ayrıca, veriyi depolamanın maliyeti düştüğü için kuruluşların depolanan büyük miktardaki veriden olabildiğince fazla değer elde etmesi gerekmektedir. Bu tür verilerin boyutu, çeşitliliği ve hızlı değişimi, yeni bir büyük veri analitiği türünün yanı sıra farklı depolama ve analiz yöntemlerini de gerektirir. Bu kadar büyük miktardaki verinin uygun şekilde analiz edilmesi ve ilgili bilginin çıkarılması gerekir. Bu makalenin katkısı, büyük veri analitiği konusundaki mevcut literatürün bir analizini sunmaktır. Bu bağlamda, uygulanabilecek çeşitli büyük veri araçları, yöntemleri ve teknolojilerinden bazıları tartışılmakta ve bunların çeşitli karar alanlarındaki uygulamaları ve sağladığı fırsatlar betimlenmektedir.

Büyük Veri İşleme Araçları

Büyük veriyi işlemek için çok sayıda araç mevcuttur. Bu bölümde, büyük veriyi analiz etmek için güncel teknikler, özellikle üç önemli yeni araç olan MapReduce, Apache Spark ve Storm vurgulanarak ele alınmaktadır. Mevcut araçların çoğu toplu işlem, akış işleme ve etkileşimli analize odaklanır. Toplu işlem araçlarının çoğu, Mahout ve Dryad gibi Apache Hadoop altyapısına dayanmaktadır.

1. **Apache Hadoop ve MapReduce:** Büyük veri analizi için en yerleşik yazılım platformu Apache Hadoop ve MapReduce'dur. Hadoop çekirdeği, MapReduce, Hadoop Dağıtılmış Dosya Sistemi (HDFS) ve Apache Hive gibi bileşenlerden oluşur.
2. **Apache Mahout:** Apache Mahout, büyük ölçekli ve akıllı veri analizi uygulamaları için ölçeklenebilir ve ticari makine öğrenimi teknikleri sağlamayı amaçlar. Mahout'un kümeleme, sınıflandırma, örüntü madenciliği, regresyon, boyutsallığı azaltma, evrimsel algoritmalar ve toplu temelli işbirlikçi filtreleme gibi temel algoritmaları, MapReduce çerçevesi aracılığıyla Hadoop platformu üzerinde çalışır.
3. **Apache Spark:** Apache Spark, hızlı işleme ve sofistike analizler için oluşturulmuş açık kaynaklı bir büyük veri işleme çerçevesidir. Kullanımı kolaydır ve orijinal olarak 2009'da UC Berkeley'deki AMP Lab'da geliştirilmiştir. 2010'da bir Apache projesi olarak açık kaynaklı hale getirilmiştir. Spark, Java, Scala veya Python'da hızlı bir şekilde uygulama yazmanıza olanak tanır. MapReduce işlemlerine ek olarak SQL sorgularını, akış verilerini, makine öğrenimini ve grafik veri işlemeyi destekler. Spark, mevcut HDFS altyapısı üzerinde çalışarak gelişmiş ve ek işlevsellik sağlar.
4. **Dryad:** Büyük veri kümelerini işlemek için paralel ve dağıtık programları uygulamak üzere veri akış grafiğine dayalı popüler bir programlama modelidir. Bir bilgisayar kümesi kaynaklarını kullanarak programları dağıtık bir şekilde çalıştırmayı sağlar.
5. **Storm:** Storm, büyük akış verilerini işlemek için dağıtık ve hata toleranslı bir gerçek zamanlı hesaplama sistemidir. Toplu işlem için olan Hadoop'un aksine, özellikle gerçek zamanlı işlem için tasarlanmıştır.
6. **Apache Drill:** Apache Drill, büyük verinin etkileşimli analizi için başka bir dağıtık sistemdir. Birçok sorgu dili, veri formatı ve veri kaynağını destekleme konusunda daha fazla esnekliğe sahiptir.
7. **JasperSoft:** JasperSoft paketi, veritabanı sütunlarından rapor üreten açık kaynaklı bir yazılımdır. Ölçeklenebilir bir büyük veri analitik platformudur ve MongoDB, Cassandra, Redis gibi popüler depolama platformlarında hızlı veri görselleştirme yeteneğine sahiptir. JasperSoft'un önemli bir özelliği, veriyi çıkarma, dönüştürme ve yükleme (ETL) olmadan büyük veriyi hızla keşfedebilmesidir.
8. **Splunk:** Son yıllarda iş dünyasında makineler aracılığıyla çok fazla veri üretilmektedir. Splunk, makine tarafından üretilen büyük veriyi değerlendirmek için geliştirilmiş gerçek zamanlı ve akıllı bir platformdur. En güncel bulut teknolojileri ve büyük veriyi birleştirir.

IoT ve Büyük Veri Analitiği

Cihazlar tarafından toplanan verilerin ilgili ve faydalı hale getirilmesi için filtrelenmesi gerekir. IoT altyapısının doğası gereği, toplanan veride fazlalık yaygındır. Veri sürekli olduğu için değerli bilginin çıkarılması kolay değildir. Bu da, verinin hem güvende hem de anlamlı olmasını sağlayacak iyi bir protokol ve yazılım mekanizması gerektirir.

Büyük Veri için Bulut Bilişim

Bulut bilişim, sunucular, depolama, veritabanları, ağ, yazılım, analiz vb. bilişim hizmetlerinin esnek kaynaklar, daha hızlı inovasyon ve ölçek ekonomisi sağlama amacıyla internet ("bulut") üzerinden sunulmasıdır. Bulut bilişim, bilişim altyapısının soyutlanma ve kullanılma şeklini kökten değiştirmiştir.

Sorunlar ve Zorluklar

Büyük verideki zorluklar genel olarak üç türe ayrılabilir: veri zorlukları, işlem zorlukları ve yönetim zorlukları. Veri zorlukları, büyük verinin karakteristik özellikleriyle ilişkilidir. İşlem zorlukları, verinin işlenmesi sırasında karşılaşılan zorluklarken; yönetim zorlukları, güvenlik sağlama gibi veriyi ele alma ile ilgilidir.

- Hacim Zorlukları:** İç ve dış kaynaklardan gelen verideki benzeri görülmemiş artış, çok büyük miktarda veriyle sonuçlanmıştır. Bu durum, verinin depolanması gibi geleneksel araçlarla çözilemeyen zorluklar getirir.
- Çeşitlilik Zorlukları:** Verinin yapılandırılmış, yarı yapılandırılmış ve yapılandırılmamış formlarda bulunmasıyla ilişkilidir. Araştırmalar, verinin %95'inin yapılandırılmamış formda olduğunu göstermektedir. Analiz yapılabilmesi için uygun forma dönüştürülmesi büyük bir zorluktur.
- Hız Zorlukları:** Hız, cihazlar tarafından üretilen verinin akış süratini ifade eder. Veri toplu işleme veya gerçek zamanlı işleme yoluyla işlenebilir. Çevrimiçi alışveriş gibi alanlarda müşteri için değer yaratmak için gerçek zamanlı işleme gerekir.

Gelecek Çalışmalar için Öneriler

Pek çok büyük şirket bu trendleri tamamen benimsemiş veya benimseme eşiğine gelmiş olsa da, büyük veri analitiğinin geleceği artık yüksek maliyet duvarlarının ardına kilitli değildir. Veri mühendisleri ve bilimcileri, Fortune 500 bütçesi gerektirmeden veri yığınının altında gizli içgörülerini ortaya çıkarmak için yenilikçi yollar geliştirmektedir. Önümüzdeki dönemde, küçük ve orta ölçekli şirketlerin de iş stratejilerine büyük veri analitiğini dahil etmeleri beklenmektedir.

Yöntem

Büyük veri analitiği, gerekli bilgiyi keşfetmek amacıyla veriyi toplama, dönüştürme, temizleme ve modelleme sürecidir. Bir araştırmada, e-ticaret web sitelerinden ürün satın alan kullanıcıya ilişkin sonuçlara ulaşmada araştırmacıyı destekler.

Sonuç

Büyük Veri'nin bulunabilirliği, düşük maliyetli donanımlar ve yeni bilgi yönetimi ve analitik yazılımlar, veri analizi tarihinde benzersiz bir an yaratmıştır. Bu trendlerin birleşimi, tarihte ilk kez, şaşırtıcı derecede büyük veri kümelerini hızlı ve uygun maliyetli bir şekilde analiz etme kapasitesine sahip olduğumuz anlamına gelir. Bu yetenekler ne teorik ne de önemsizdir. Verimlilik, üretkenlik, gelir ve karlılık açısından muazzam kazanımlar elde etmek için gerçek bir sıçrama ve net bir fırsat temsil ederler. Büyük Veri Çağı buradadır ve hem iş hem de teknoloji profesyonelleri birlikte çalışmaya ve vaadi yerine getirmeye devam ettikleri sürece, bunlar gerçekten devrimci zamanlardır.

Kaynakça

1. Demirkan, H., Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems*, 55, 412–421.
2. Barton, D., Court, D. (2012). Making Advanced Analytics Work For You. *Harvard Business Review*, 90(10), 79–83.
3. McAfee, A., Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, 90(10), 60–68.
4. Wu, X., et al. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 28, 97–106.
5. Pääkkönen, P., Pakkala, D. (2015). Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems. *Big Data Research*, 2, 166–186.
6. Thusoo, A., et al. (2010). Data warehousing and analytics infrastructure at Facebook. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*.
7. Kreps, J., Narkhede, N., Rao, J. (2011). Kafka: a distributed messaging system for log processing. *Proceedings of the 6th International Workshop on Networking Meets Databases*.
8. Wu, L., et al. (2012). Avatar: OLAP for web-scale analytics products. *Proceedings of the 38th International Conference on Very Large Databases*.
9. Mishne, G. (2013). Fast data in the era of big data: Twitter’s real-time related query suggestion architecture. *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*.
10. Lin, J., Ryaboy, D. (2013). Scaling big data mining infrastructure: the Twitter experience. *ACM SIGKDD Explorations Newsletter*, 14, 6–19.
11. Lee, G. L., et al. (2012). The unified logging infrastructure for data analytics at Twitter. *Proceedings of the 38th International Conference on Very Large Databases*.
12. Busch, M., et al. (2012). EarlyBird: real-time search at Twitter. *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*.
13. Amatriain, X. (2013). Big & Personal: data and models behind Netflix recommendations. *Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining*.
14. Amatriain, X., Basilico, J. (2013). System architectures for personalized recommendations. *Netflix Tech Blog*.
15. Boulon, J., Konwinski, A., Qi, R. (2008). Chukwa: a large-scale monitoring system. *Proceedings of the Cloud Computing and Its Applications Workshop*.