Büyük veri anlamlandırmada panoramik yaklaşım

İlter Tolga Doğan[1], Yasemin Şahin Doğan[1], Cevat Şener[2]

[1] E-Kalite Yazılım, ODTÜ Teknokent, Ankara
{ilter,yasemin}@e-kalite.com.tr
[2] Bilgisayar Müh. Bölümü, ODTÜ, Ankara
sener@ceng.metu.edu.tr

Öz: Verinin hızla ve ivmelenerek artmakta olduğu çağımızda veriden anlam çıkaran yazılımlar da yeni altyapılar kullanmakta ve her geçen gün güçlenmektedir. Burada verilerin gösterimine yönelik yeni bir yaklaşım olan panoramik adı verilen gösterim şekli tanıtılmaktadır. Panoramik gösterim, fasetlerin yoğun kullanımı, verinin 3 boyutlu fasetlendirilmesi, ağaç yapısı yardımıyla fasetlerin ortak kullanımı ve verinin önceden belirlenen sekmelerle farklı perspektiflerde görünmesi olarak özetlenebilmektedir. Fasetlerin bu derece yoğun şekilde kullanılabilmesi için NoSQL olarak adlandırılan, varlık-bağlantı modelinden uzaklaşmış ve yatay ölçeklendirilebilir sistemlerin kullanılması gerektiği düşünülmüş ve bu düşünce örnek bir uygulamada gerçekleştirilmiştir. Bu bildiride büyük verinin panoramik gösterim biçimiyle gösterilmesinin veri anlamlandırılmasındaki yararları ve panoramik gösterim için hangi alt yapıların kullanıldığı anlatılmaktadır.

Anahtar Kelimeler: Faset, NoSQL, Panoramik gösterim, Yatay ölçeklendirme.

1 Giriş

Veriden yararlanma ve alınacak kararları verilerin analizi ve görselleştirmesi üzerine kurgulama bilgisayarlar tarihinden çok daha eski bir yöntemdir[1]. Veri analizi ihtiyacının kaynakların en verimli ve hızlı bir şekilde kullanılmasının gerektiği İkinci Dünya Savaşı'nda daha yoğun bir şekilde hissedildiği ve ilkel bilgisayarların da bu dönemde kullanılmaya başlandığı bilinmektedir [1]. Verinin analizi ve görselleştirmesi bu dönemde savaşla ilgili kritik konularda karar almada kullanılmıştır[2].

Veri analizi ve görselleştirmesinde atılım niteliğindeki ilerlemeler için ise mekanik yöntemler yetersiz kaldığından, bilgisayar teknolojisinin ilerlemesi gerekmiştir. Bilgisayar kullanımının giderek yaygınlaşması ve yoğunlaşması ile aynı zamanda veri de

[1] C.J. Minard tarafından 1869'da çizilmiş olan Napoleon'un Rusya seferinin görselleştirmesi elle yapılmış olsa da (http://en.wikipedia.org/wiki/File:Minard.png) verinin kolayca gözlenmesine olanak sağlayarak konunun daha hızlı anlaşılabilmesini sağlayan eski bir veri analizi uygulaması olarak görülebilir.

[2] II. Dünya savaşı'nda daha verimli savaş için sonradan ABD savunma bakanı olacak Robert McNamara'nın kayıp/yıkım oranını ölçerek ateş bombalarının kullanılması kararına farklı grafikler çizerek yardımcı olması gibi.

hızlanarak çoğalmaya başlamıştır. Bugünkü hızıyla Dünya'daki toplam veri her beş yılda bir 10 kat artmaktadır [2]. Bu toplanan veriden yararlanmak giderek daha zorlayıcı bir zorunluluk olmaktadır[3].

Bu bildiri ile farklı veritabanları ya da veri ambarlarında saklanan verileri *panoramik* adı verilen indekslemeye dayalı arama-raporlama araçlarından geçirme işlemine neden gerek duyulduğu, bu işlemin veriden yararlanmaya yönelik iş zekası alanında Web yazılımları yaratmaya yönelik uygulama çerçevesi, e-Dijital Analiz Platformu (e-DAP[4]) içerisinde nasıl uygulandığı ve geliştirilme sürecindeki sorunlar özetlenmektedir.

2 Panoramik modeli doğuran gereksinimler

Bu bölümde birçok yazılımın incelenmesi ve E-Kalite Yazılım'ın 10 yıllık arama-raporlama-çizelgeleme deneyiminin sonucunda panoramanın doğmasını sağlayan koşullar ve sorunlar listelenmektedir.

Veriden yararlanılması öncelikle verinin aranabilir olmasını gerektirir. Ancak bu konuda yeterince deneyim edinilmesi ve yüksek düzeyli bir yazılım geliştirilmiş olması durumunda veriden anlam çıkarma teknolojilerine geçilebilir. Her tür verinin aranabilir-raporlanabilir-çizelgelenebilir olmasını sağlayan bir yazılım çerçevesinin E-Kalite Yazılım tarafından geliştirilmiş olması iş zekası yönünde ilerlemek için önemli bir artı olmuştur [4]. Zamanla aranan verinin boyutlarındaki artış, verinin daha etkin yöntemlerle anlamlandırılmasını bir zorunluluk hâline getirmiştir. Az sayıda veri bir liste içinde bile anlaşılabilirken, veri çoğaldıkça gruplama ve görselleştirme önem kazanmaktadır. İstek üzerine yaratılan 3 boyutlu çizelgeler, veri anlamlandırma için bir ilk adım olarak düşünülebilir. Ancak daha kapsamlı bir veri anlamlandırma çözümü ile raporlamaların daha hızlı ve daha ayrıntılı gösterilmesi gereksinimi doğmuştur.

İVTYS'ler (varlık-bağlantı modelini kullanan ilişkisel veritabanı yönetim sistemleri) ile oluşturulan geleneksel yazılımlarda büyük tabloların bağlantırılması durumunda verinin anlamlandırılması için birçok sorun oluşmaktadır. Bu sorunlar aşağıdaki gibi özetlenebilir:

1. Tüm sonuçların listelenmesinin çok fazla sonuç için anlamlı olmaması;
2. Yalnızca genel toplam veriden bir anlam çıkarmak için yeterli olmaması;
3. Bazı arama alanlarının çok sayıda ve büyük tablolardan sonuç çekmesinden dolayı yavaşlık yaratması;
4. Bazı veri gösterim hücreleri için ayrıca sorgu yapılması gerekmesi.

Bu sorunların çözümü olarak ilk düşünülen yöntemler aşağıdaki gibiydi:

---

[3] "Every day I wake up and ask: How can I flow data better, manage data better, analyse data better?" says Rollin Ford, the CIO of Wal-Mart": Wal-Mart'ın CIO'su veriden yararlanmak zorunluluğunun ne kadar ön planda olduğunu bu sözlerle anlatmıştır [3].

[4] http://e-dap.com.

1. Sonuçların alt bölümlere (faset) ayrılarak kullanıcının yönlendirilmesi gerektiği düşünülmüştü.
2. Fasetlerin[5] oluşturulması için ayrı sorgular çalıştırılabileceği düşünülmüştü. Ancak faset sayısının artmasıyla bunun çok verimsiz olacağı görülmekle OLAP benzeri bir sistem kurulabileceği değerlendirildi.
3. Birkaç tablonun çaprazlanmasıyla elde edilecek önbellek tablolarının yaratılması gerektiği düşünülmüştü.
4. Aramaların da önbellek tablolar üzerinden yapılması gerektiği düşünülmüştü. Ancak önbellek tabloların yaratılması ve idamesinin başlı başına performans sorunlarına yol açabileceği görüldü.

Bu aşamada NoSQL modeli incelenmeye başlanmış ve NoSQL modelinin, indekslenmiş denormalize veride fasetlerin çok etkin bir şekilde kullanılabilmesini sağladığı görülmüştür.

Bahsedilen bu gereksinimler için Apache Solr kullanılmıştır. Bilindiği üzere Solr, Apache Lucene projesinin bir parçası olan tam metin arama, çok yönlü arama, devirgen kümeleme, veritabanı ile bütünleşik, Microsoft Word veya PDF benzeri belgeleri indeksleme gibi özellikleri olan açık kaynak kodlu oldukça esnek bir arama motorudur[6]. Solr indekslemeyle çalışır ve bu indeksleme aşamasında verinin çok boyutluluğunu yatay düzlemlerde tekrar tanımladığından İVTYS'lere göre çok daha hızlı bir şekilde arama isteklerine sonuç verebilir. Bu şekilde yalnızca büyük veride arama değil fasetler yaratmada da Solr en verimli şekilde kullanılabilir.

3 Panoramik Modelin Yazılım Tasarımı

Yatay ölçeklendirme (NoSQL) ile aramaya yenilikçi yaklaşım ve yukarıda bahsedilen 4 sorunun da çözümünü sağlayan arama modeline panoramik arama adı verilmiştir. Panoramik arama Şekil 1'de görülen NoSQL veri aktarım avantajlarını kullanarak fasetleme özelliğinin etkin kullanılmasıyla elde edilmiştir. İVTYS'de fasetlenmiş sınıflandırma oluşturmak için boyutların birbiriyle kombinasyonu sayısınca sorgu çalıştırmak gerekirken, bu NoSQL'de sistem içi araçlarla kolayca yapılabilmektedir. Bu sayede tek bir hiyerarşiye dayanan bir sınıflandırmanın sınırlarından çıkılıp çok boyutlu sınıflandırma yapılabilmektedir. Bunun e-DAP içinde kullanılma biçimine Panorama denilmiştir.

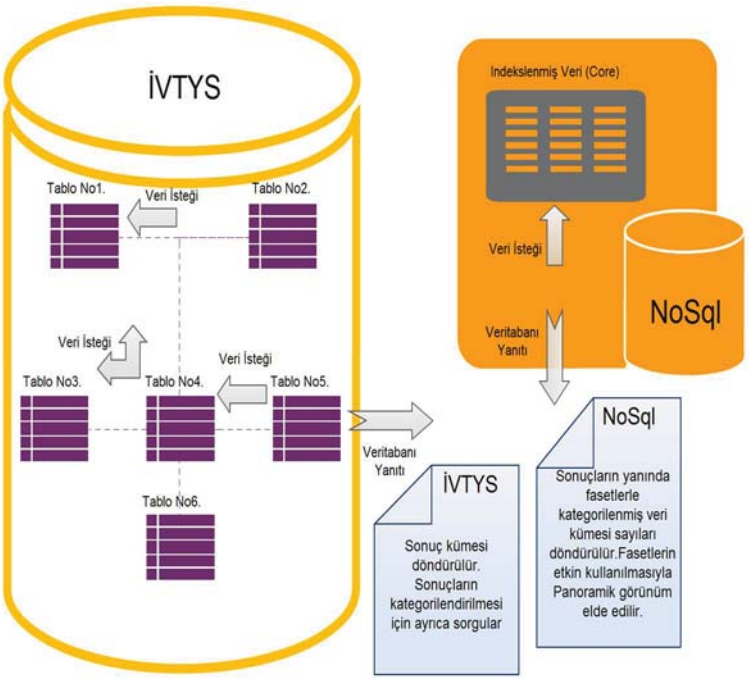
Panoramik aramanın klasik NoSQL sistemlerinden önemli ölçüde farkı vardır. NoSQL ile yapılan aramalar, genel olarak kullanıcıya tek bir arama alanı (anahtar kelime/*keyword*) sağlayarak aramayı tüm indeks (hızlı alınabilmek için optimize edilmiş veri yapısı) üzerinde tam metin (*full-text*) yapacak şekilde kurgulanmıştır. Burada kullanıcı farklı veri tiplerinde arama yapmak istese de (*text*/*float*/*date* vb) anahtar kelime alanına arayacağı değeri, değerin tipinden bağımsız olarak girer ve sonuç tüm indeks üzerinden getirilir. e-DAP bünyesinde çalışan Panoramik aramada

[5] Patent no:US6275821 Danish Mohamed Sherif [US]; Kimbrough Kris Walter (2001)

[6] Ön bilgi http://lucene.apache.org/solr sayfasında görülebilir.

ise farklı veri kategorilerinin aranabilmesi sonucun farklı perspektiflerden izlenebilmesi amaçlanmıştır.

Yine standart aramaların diğer bir limiti kullanıcıların *NOT*, *OR*, *AND* gibi operatörleri kullanmasına izin vermemesidir. Panoramik arama yapısında her arama alanının yanında o alanda birden fazla değeri kullanıcının operatörler ile birleştirmesine (*AND*, *OR*) ya da elemesine (*NOT*) izin verecek ikonlar yer almaktadır.



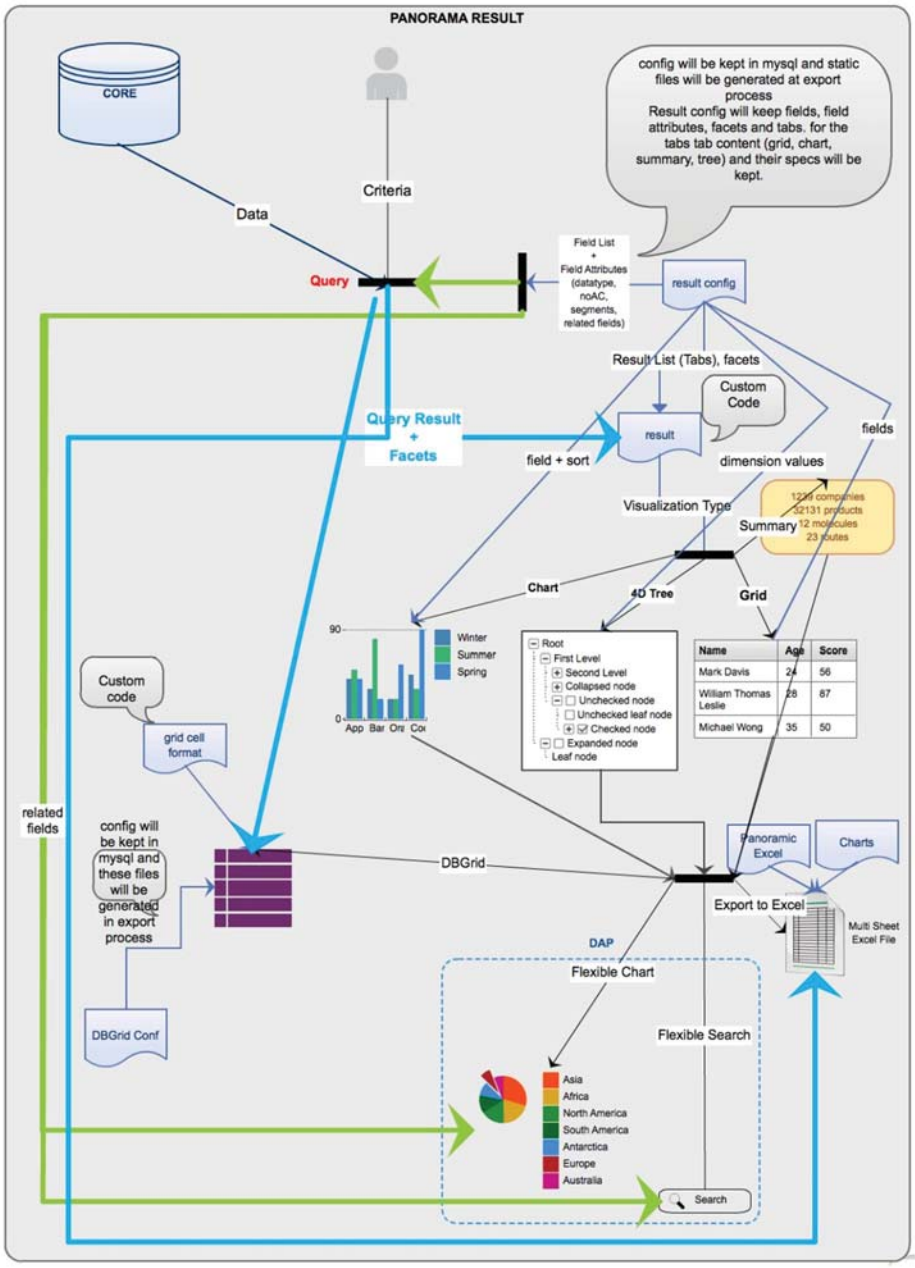
Şekil 1. NoSQL ve İVTYS veri aktarım farklılığı kavramsal şeması.

Bir NoSQL sistemi olan Solr üzerinden sağlanan performans artışı ile yenilikçi bir veri gösteriminin yapılabileceği düşünülmüştür. Bu gösterimde yalnızca arama sonuç listesi dökülmemesi sonuç kümesinin farklı perspektiflerden nelerden oluştuğunun gösterilmesi hedeflenmiştir. Ayrıca bu sayede kullanıcının isteyebileceği alt alanlar açık bir şekilde görülecek veri kümesini daraltma ve alt bölümlere inebilmesi kolaylaşacaktır.

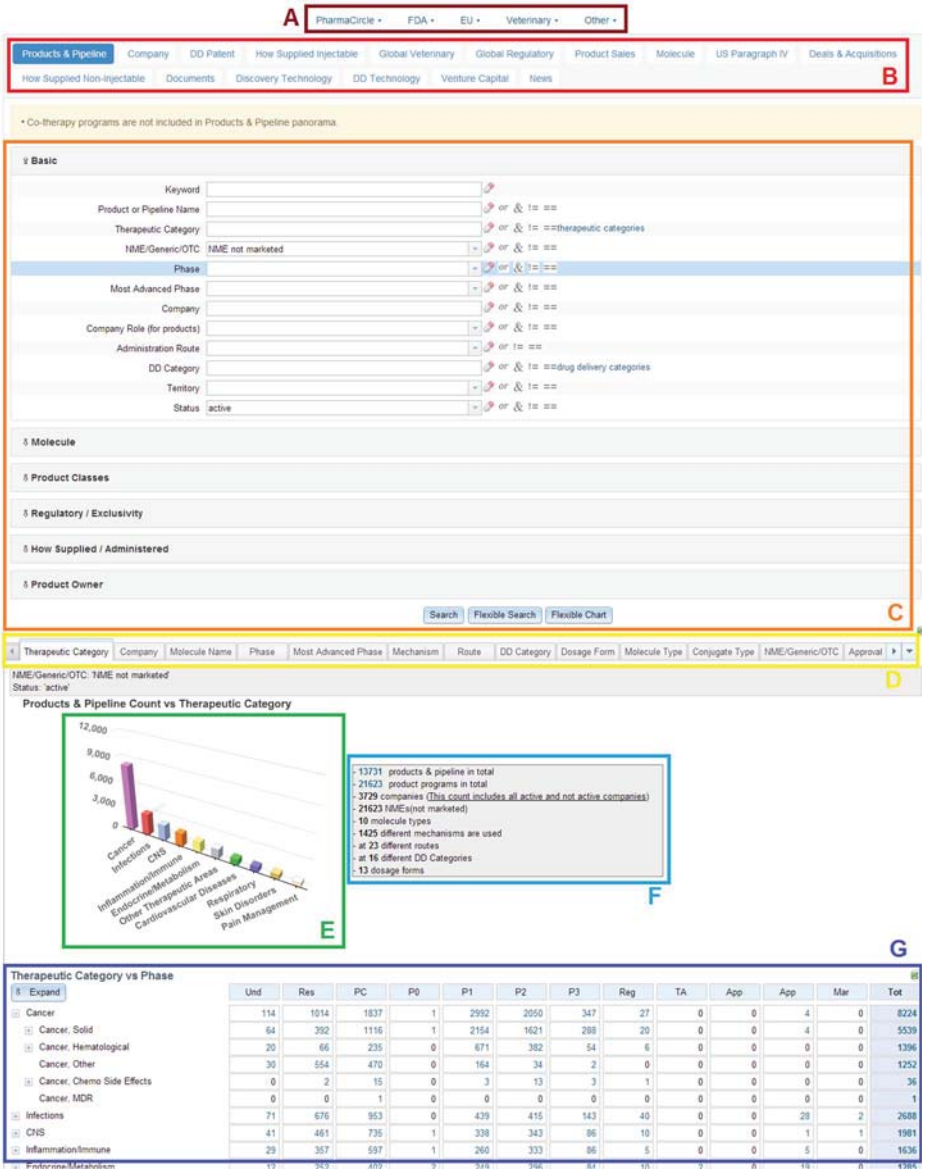
Panoramik aramanın indekslenmiş veri üzerinde çalışmasından dolayı İVTYS'lerden tamamen kopma düşünülmemiştir. İVTYS'lerin bir arama bölümüne ulaşmak için kullanıldığı hibrit bir sistem düşünülmüştür.

Sonuç bölümü kavramsal şeması panoramik ve detaylı/esnek aramalarının veri kaynağı ve sonuç gösterim formatları arasındaki ilişkiyi özetlemeye yönelik hazırlanmıştır (Şekil 2). Şemada da gözlemlenebileceği gibi panoramik aramalar veriyi NoSQL veri kaynaklarında sorgularken İVTYS aramalar (şekilde "*detailed*/*flexible*" olarak gösterilmiştir) verisini ilişkisel veritabanından çekmektedir. Panoramik arama ile İVTYS aramalar veri gösterim tipleri arasındaki fark da şemadan kolayca anlaşıla-

bilir. Sağlıklı bir arama sürecinin panorama ile başlayarak kullanıcının verinin genel durumu ile ilgili bilgi sahibi olmasından sonra detaylı/esnek ile tekil verilere yoğunlaşması beklenmektedir.



Şekil 2. Panorama Sonuç kavramsal şeması.



Şekil 3. Panorama görünümü.

4 Panoramik Modelin Gösterimi

Bu bölümde sonuçta oluşan ürünün farklı özellikleri ekran görüntüleriyle verilmiştir:

- **Arama Sonuçlarının Görselleştirilmesi:** Arama sonuçlarını kullanıcıların saniyeler içinde anlamlandırmasına 3 temel öğe yardım etmiştir: Sonuç Özet Bölümü

(Şekil 3.F), En Önemli Sonuçların Çizelgelenmesi (Şekil 3.E) ile 2 ve üstü Boyuta Sahip Grid ve Ağaçlar (Şekil 3.G). Her sekmede (Şekil 3.D) bu 3 öğede yer almakta kullanıcının baktığı anda genel resmi algılayabilmesi sağlanmaktadır. Bu öğelerin detaylı yenilikçi tarafları aşağıda verilmektedir.

- **Sonuç Özet Bölümü (Şekil 3.F):** Solr fasetlerinin etkin kullanımı sayesinde arama sonucunun hızlı analizine olanak sunacak şekilde özet sonuç bilgisi fasetlerden getirilmiş, dahası bu fasetlerle o perspektifte verisi olmayan sonuçların sayısı da ayrıca belirtilmiştir. Bu amacı bir örnekle somutlamak gerekirse, kanser ilaçları aramasında ilaçların formunu gösteren perspektifte kanser ilaçlarının kaçının bu bilgiye sahip olduğu detayı kullanıcıya özet bölümünde gösterilmiş bu sayede yaptığı analizin daha gerçekçi olmasına yardım edilmiştir.
- **En Önemli Sonuçların Çizelgelenmesi (Şekil 3.E):** Sonuç sekmelerinde sekmenin perspektifine konu olan veriye göre en önemli sonuçlardan (burada en önemliden kasıt sekmeye konu olan perspektifteki verilerden indekse baz varlığı en çok içerenlerdir; bazı ilaç olan indeksin firma sekmesinde en çok ilacı olan ilk 10 firmanın listelenmesi gibi) çizelge oluşturulup (fasetlerden gelen veriyi kullanarak) örneğin firma sekmesinde en çok ilaç üreten firmaları çizelge aracılığı ile hızlıca görmek mümkün olmuştur. Bunun İVTYS içerisinde yapılması oldukça deoptimize olacaktır.
- **İki ve üstü Boyuta Sahip Grid ve Ağaçlar (Şekil 3.G):** Sonuçlar verinin hiyerarşi içermesi ya da içermemesine göre grid ya da ağaçlarda görüntülenmektedir. Burada yenilikçi taraf hiyerarşi içeren ağaçları yapının aynı zamanda bir matris görünümü ile toplam sonuç sayısının da farklı kategoriler için gösteriminin yapılmasına olanak sağlamasıdır.
- **Arama Alanlarının Gruplanması (Şekil 3.A,B,C):** Çoklu arama alanlarını destekleme özelliğinin beraberinde çok fazla arama alanı içeren arama sayfalarının karmaşık ve arama yapılacak aranı hızlıca belirleyememe sorunlarını getirebileceği varsayımı ile arama alanlarının mantıksal gruplamasına izin verilmesinin kullanıcı dostu özelliğini pekiştirdiği görülmüştür.

5 Sonuçlar

Verinin değerinin her geçen gün daha iyi anlaşıldığı çağımızda daha iyi veri analizi yapabilmek için yenilikçi yöntem ve modellerin denenmesi ve kullanılması bir gereklilik olmuştur. Bu çalışmada da bir NoSQL yaklaşımı ile panoramik adı verilen veri görselleştirme modeline nasıl ulaşıldığı gösterilmiştir. Panoramik model özellikle veri yığınlarının kategorileştirilerek ayrı gösterildiği durumlarda güçlüdür. Buna somut bir örnek olarak ABD yan etki veritabanı (FAERS) sonuçların karşılaştırılması gösterilebilir. İVTYS modelindeki bu veritabanında içindeki veri sayısı 20 milyonu bulan 7 tablo bulunmaktadır ve birkaç tabloyu kullanan bir MySql sorgusu 3 dakika sürerken, indekslenen aynı veritabanının NoSQL biçiminde aynı işi yapan Panorama sorgusu 3 saniye sürmektedir. Öte yanan görece küçük veri kümelerini incelemek için panoramik modele gerek duyulmayacaktır.

Bununla birlikte, NoSQL'de indekslenen verinin tek boyuta düşürülmesi aşamasında veri bir perspektiften incelenebilmekte, ayrıntılı kategorizasyon indeksleme düzeyinde düşünülmemişse bazı bağlantılar birleştirilmekte yani veri bazı yönlerden daha basitleştirilebilmektedir. Ayrıca, Verinin NoSQL'e indeksleme işlemi sistem kaynaklarını yüksek oranda kullanabilmektedir.

Panorama şu anki hâliyle tek sunucuda çalışmaktadır. NoSQL yapısı buna uygun olduğundan, aşırı büyük veri için birden fazla sunucuda çalışmasını sağlayacak Apache Hadoop kullanılması düşünülmektedir.

Teşekkür: Projenin hayata geçmesinde sağladığı destekten ötürü TÜBİTAK TEYDEB 1507 programına teşekkür ederiz.

6 Kaynakça

1. Information Systems based on Logistics Perspective Project: Historical Perspective on Information Systems, http://www.uh.edu/~mrana/try.htm#HIS (2000)
2. Nye, J.S.: The Future of Power, s. 115, Public Affairs,(2011)
3. The Economist: Data, data everywhere, 25 Şubat 2010 (2010)
4. Doğan, İ.T., Doğan, Y.Ş. ve Şener, .: DAP: Gelişmiş Web tabanlı Arama, Raporlama ve Çizelgeleme oluşturma amaçlı Yazılım Çerçevesi, UYMS 2013, http://ceur-ws.org/Vol-1072/submission56.pdf (2013)
5. Bo, C.: Faceted Browsing of OPAC Based on Open-source Software Solr, New Technology of Library and Inf. Service, http://en.cnki.com.cn/Journal_en/I-I143-XDTQ-2007-11.htm (2007)
6. Powell, G.: Chapter 8: Building Fast-Performing Database Models, Beginning Database Design, Wiley (2005)
7. Jern, M.: Information Drill-down using Web Tools, Visualization in Scientific Computing, Springer Vienna (1997)