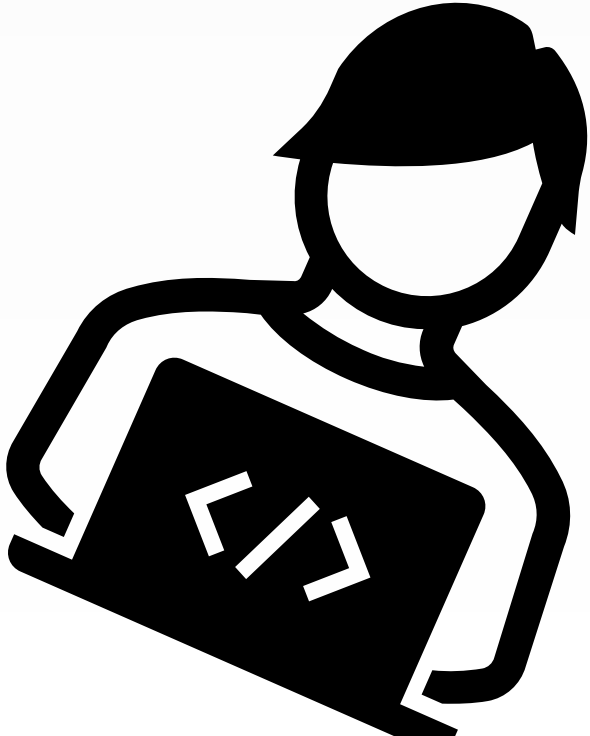
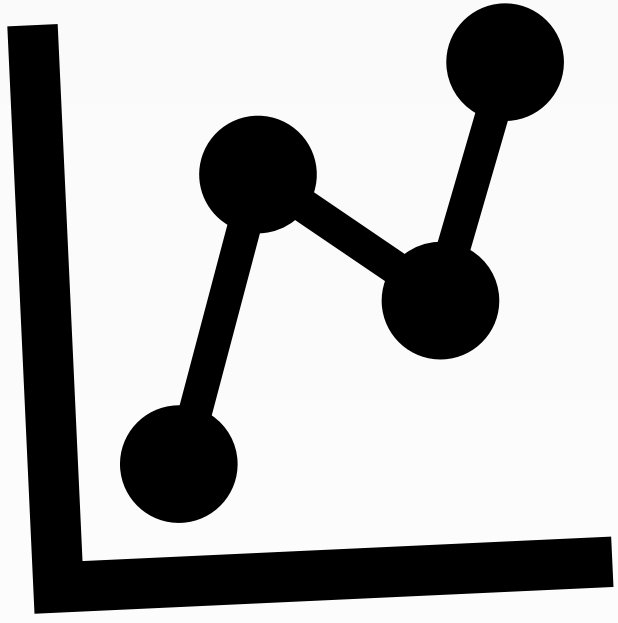




EKO 469-Veri Madenciliği

Hafta 5 – Veri Ambarları ve OLAP

Karar Destek Sistemleri

- **Karar destek sistemleri**, kurum içi ve dışı verilerin, karar verme sürecinde kullanılabilecek bilgiye dönüştürülmesi ve doğru karar verme olasılığının arttırılması ile ilgilenir.
- Günümüzde yöneticiler için karar alma ortamı eskisine oranla daha karmaşık hale gelmiş, yöneticilerin kararları üzerinde etkili olabilecek faktörler artmıştır. İç ve dış çevredeki değişiklikler ve kriz durumları karar verme sürecini zora sokmaktadır.
- Yönetici herhangi bir anda, daha önceden öngörülmemiş bir bilgiye gereksinim duyabilir. Bu gibi durumlarda yöneticiye hemen yanıt verebilecek, karar alma sürecini kolaylaştıracak bir sisteme ihtiyaç vardır.



Karar Destek Sistemleri

- Karar Destek Sistemleri (Decision Support Systems) ve Üst Yönetici Sistemleri (Executive Information Systems) bilgisayarların karar alma sürecindeki etkisini arttırmak amacıyla ortaya atılmıştır.
- Karar destek sistemleri basit karar modellerinin yanı sıra karmaşık matematiksel yada istatistiksel modelleri de içerebilir.
- Üst yönetici sistemlerde ise karmaşık modellere yer verilmez, çoğunlukla sistemin sorgulama olanaklarından yararlanır. Stratejik düzeydeki yönetici personel için hazılanır.

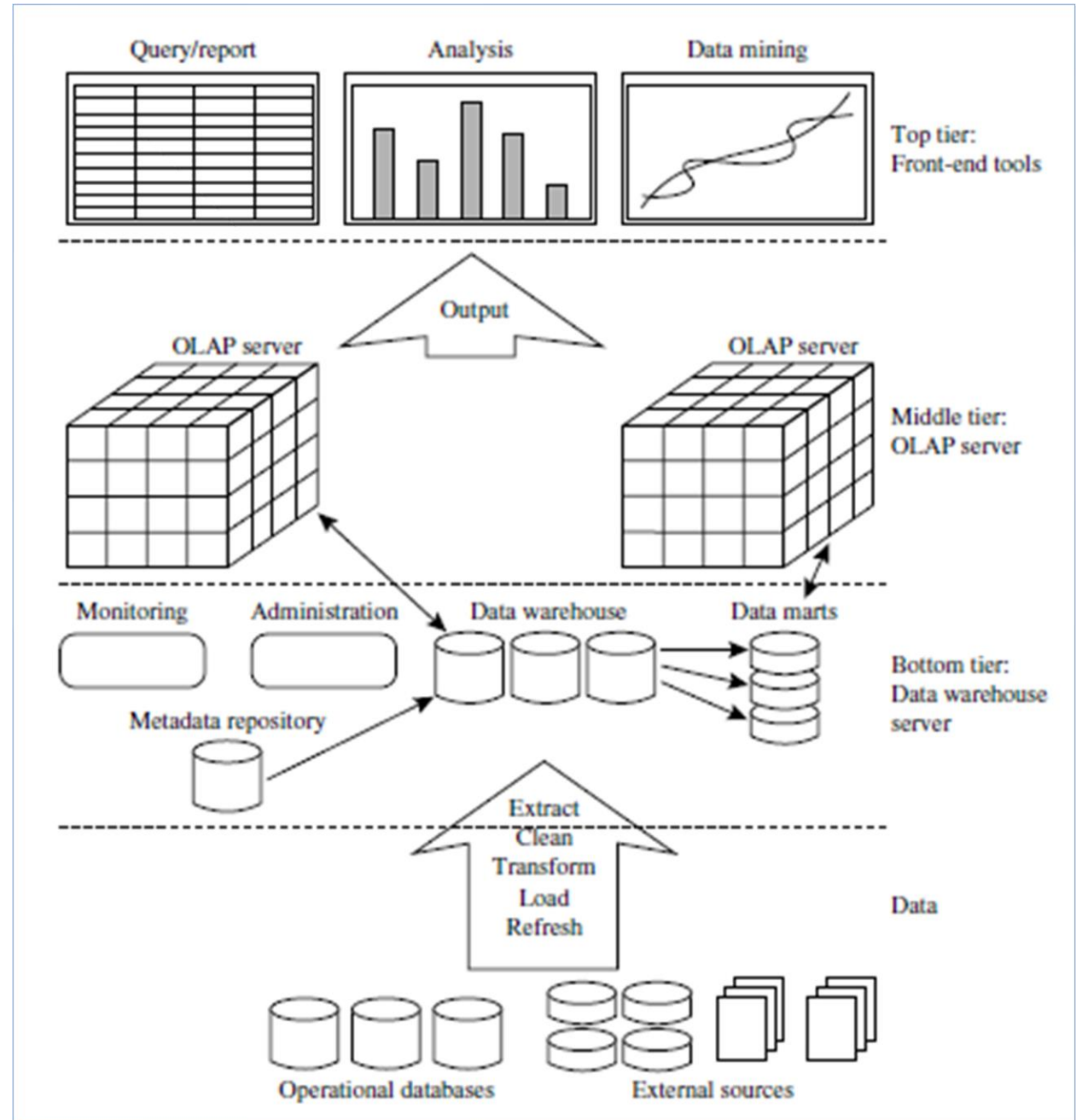


Karar Destek Süreci

- 90'lı yıllara kadar karar destek sistemleri elektronik tablo programlarının kullanımı seviyesinde kalmıştır. 90'lı yıllarda hem donanım hem de yazılım teknolojilerindeki gelişmeler karar destek sistemlerinin gelişmesi için ortam hazırlamıştır.
- Veritabanı sistemleri sağladığı avantajlara rağmen, karar destek uygulamalarındaki gereksinimleri karşılamakta zorlanmaktadır.
- Kurumların sahip olduğu verilerin boyutu bazen geleneksel veritabanı yöntemleri ile işlemek için çok büyük olabilmektedir.
- Bu gereksinime paralel olarak verinin farklı biçimde saklanması ve veriye hızlı erişimin sağlanabilmesine olanak tanıyacak başka yollar aranmaya başlanmıştır.
- Bu noktada **veri ambarı** kavramı karşımıza çıkmaktadır.



Üç katmanlı veri depolama mimarisi



ETL (Extract-Transform-Load) Süreci

- **ETL(Extract (Çıkar)-Transform (Dönüştür) –Load (Yükle)):** Büyük boyutlu veriler operasyonel sistemlerden alınıp veri ambarı veya data mart'lara yüklenilmesi süreçleridir.
- **Extract:** Veritabanından Veri ambarının ihtiyacı olan verilerin çekilmesidir.
- **Transform:** Veri ambarındaki verilerin belirli bir standartta olması bu aşamada sağlanır. Veriler veri ambarının yapısına uygun hale getirilir.
- **Load:** Veri kaynaklarından çekilen ve dönüştürülen veriler veri ambarına yüklenir.

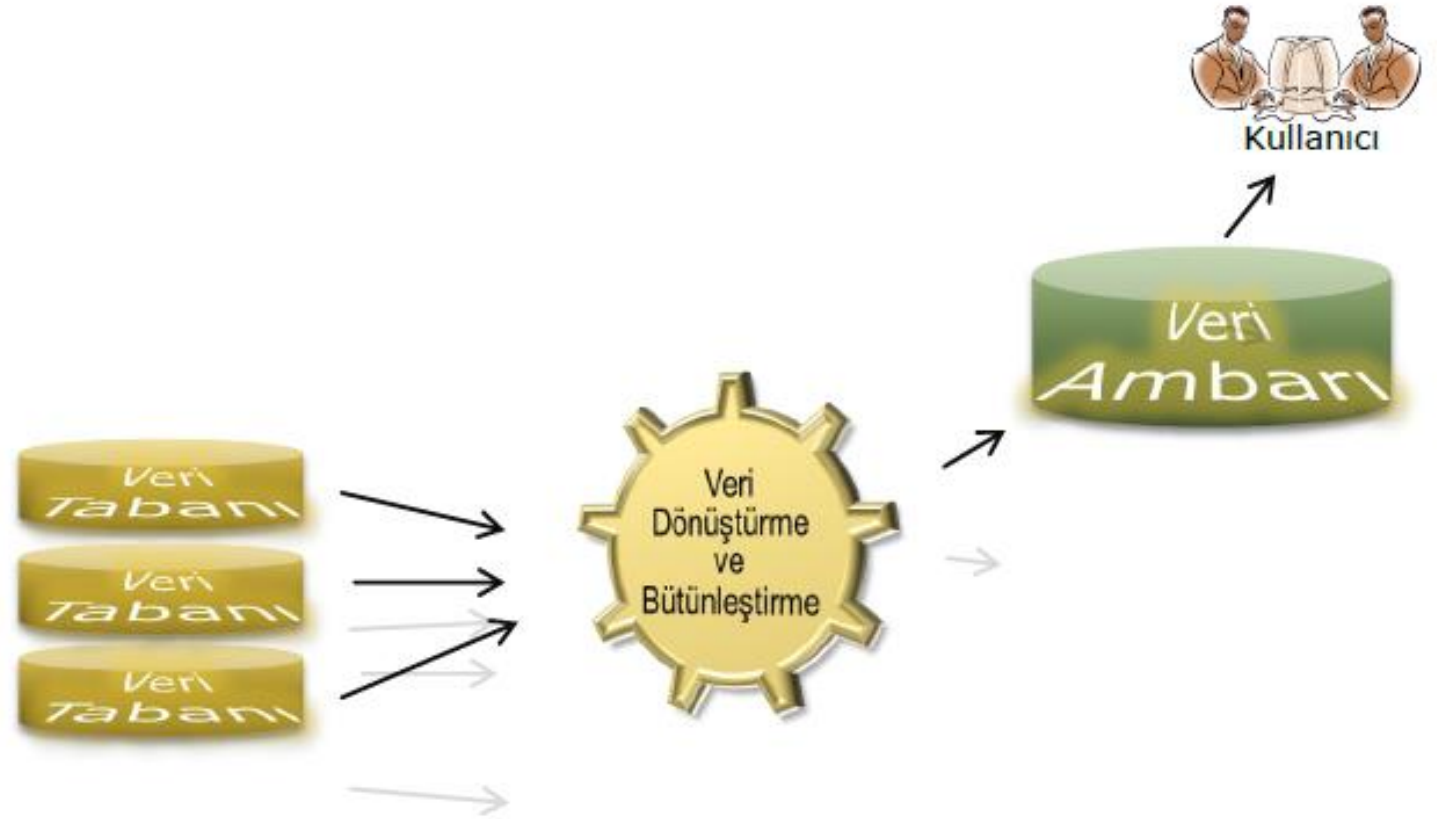
Bu işlemler için Microsoft DTS (Data Transformation Services) vb. araçlar kullanılabilir.

Veri Ambarı



- İşletmelerde kullanılan işlemsel veritabanları (transactional database) doğrudan veri madenciliği uygulamalarında kullanılmaz, bu verilerin veri madenciliği amacıyla kullanılabilmesi için uygun hale getirilmesi gerekmektedir.
- Belli bir döneme ait, yapılacak çalışmaya göre konu odaklı olarak düzenlenmiş, birleştirilmiş ve sabitlenmiş veritabanlarına **veri ambarları** denir.

Veri Ambarı Mimarisi





Veri Ambarı

- Veri ambarları bir işletmenin sahip olduğu verilerin karar destek amacıyla kullanılmasına olanak sağlar.
- Veri ambarları karar destek sistemlerinin teknik altyapısını oluşturur.
 - Konu odaklıdır
 - Bütünleşiktir
 - Zaman boyutu vardır, belirli bir döneme ve zaman dilimine aittir
 - Sadece okunabilir

Veri Ambarı

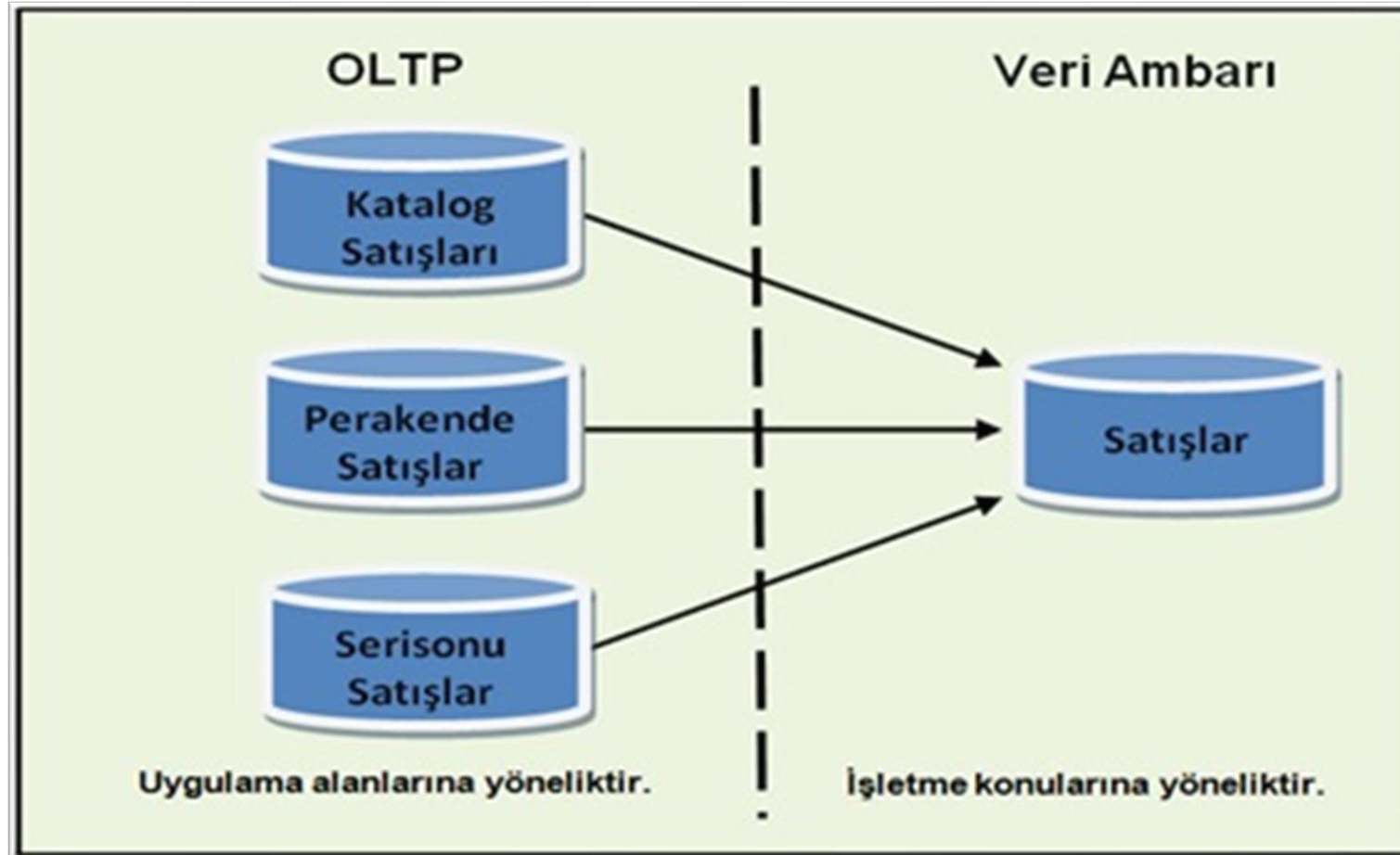
KONU ODAKLI

Klasik işlemsel sistemler, işletmelerdeki işlemler, süreçler yada fonksiyonlar üzerine yoğunlaşmıştır. Veri ambarı ise işletmedeki belli başlı konular üzerine yoğunlaşmıştır.

Örneğin işlemsel uygulamalar muhasebe, personel, stok vb sistemlere yada fonksiyonlara yönelik olabilir. Veri ambarı müşteri, satıcı, ürün, eylem(işlem) gibi konulara yönelik olarak tasarlanabilir.

Veri ambarları KDS sürecinde kullanılmayacak veriyi içermez.

KONU ODAKLI



Veri Ambarı

BÜTÜNLEŞİK

İşlemsel sistemlerde verinin tanımlanması farklı şekillerde tercih edilmiş olabilir.

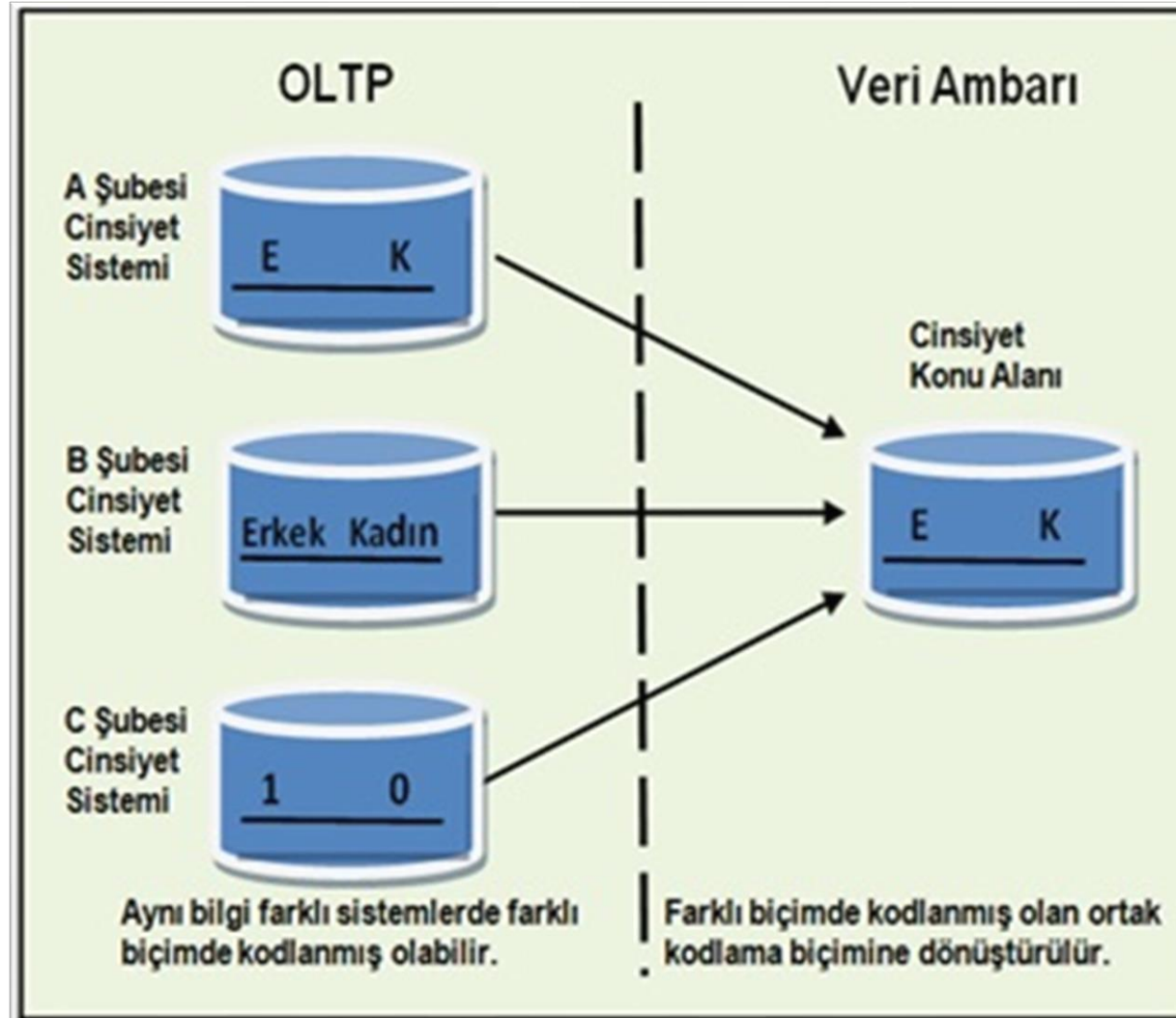
Örneğin, cinsiyet ile ilgili bir alan, bir uygulamada “E” ve “K” ;bir başka uygulamada 1 veya 0; bir başka uygulamada ise “Erkek” veya “Kadın” şeklinde ifade edilmiş olabilir.

Ölçü birimleri farklı (inch, cm, m...) tercih edilmiş olabilir.

Tarih gösterimi için farklı formatlar tercih edilmiş olabilir.

İşlemsel veri tabanından veri ambarına veriyi taşıırken, farklı tanımların tümünü ortak bir formata dönüştürmek gerekecektir.

BÜTÜNLEŞİK



Veri Ambarı

ZAMAN BOYUTU

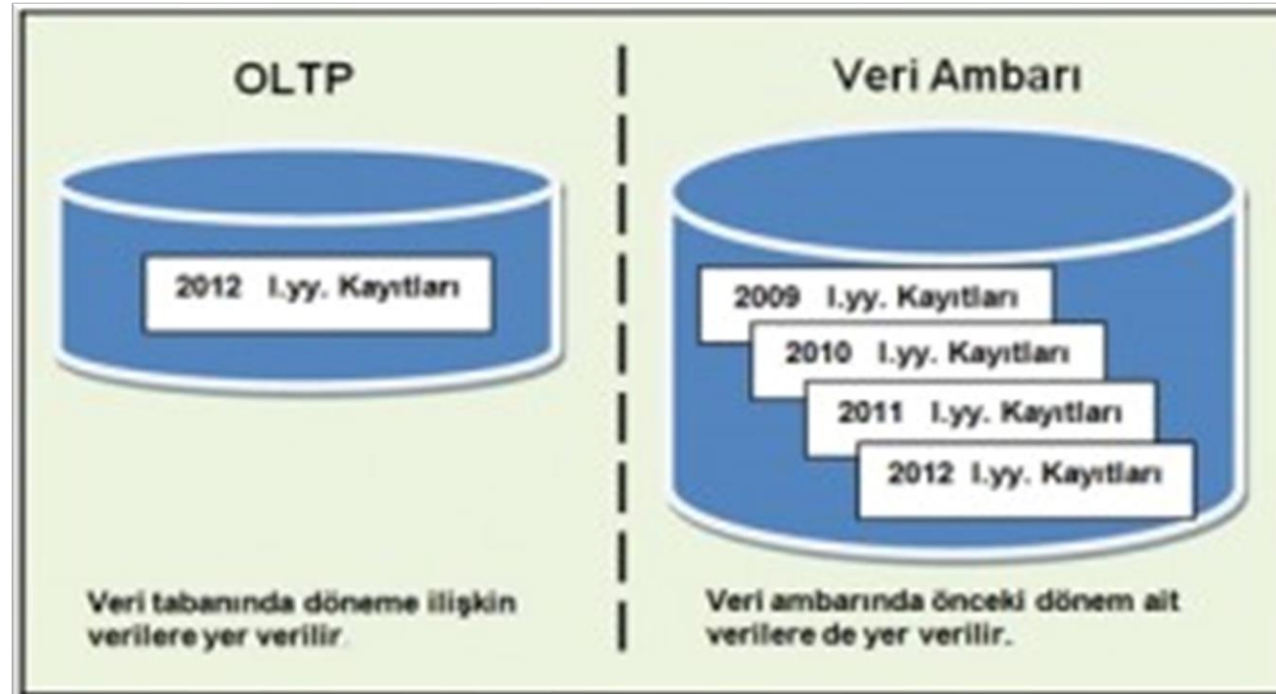
Veri ambarındaki tüm veriler zamanın belirli bir anına aittir. Sadece o andaki değerler ile değil; geçmişteki değerler ile de ilgilidir.

Veri, zaman içindeki bir noktayla birleştirilerek değerlendirilir. Örneğin mali yıl, ödeme dönemi, sömestre dönemi gibi unsurlar dikkate alınır.

Veri ambarı, verinin zaman içinde aldığı değerleri de çözümlemeye katacaktır.

İşlemsel veride zaman boyutu olmadığı için güncellenebilir bir özelliğe sahiptir. Veri ambarı ise geçmiş dönemlere ilişkin değerleri içerdiği için, işlemsel sistemlerdeki veri bir kez yansıtıldığında bir sonraki yansımaya kadar herhangi bir güncelleme yapılmayacaktır.

ZAMAN BOYUTU



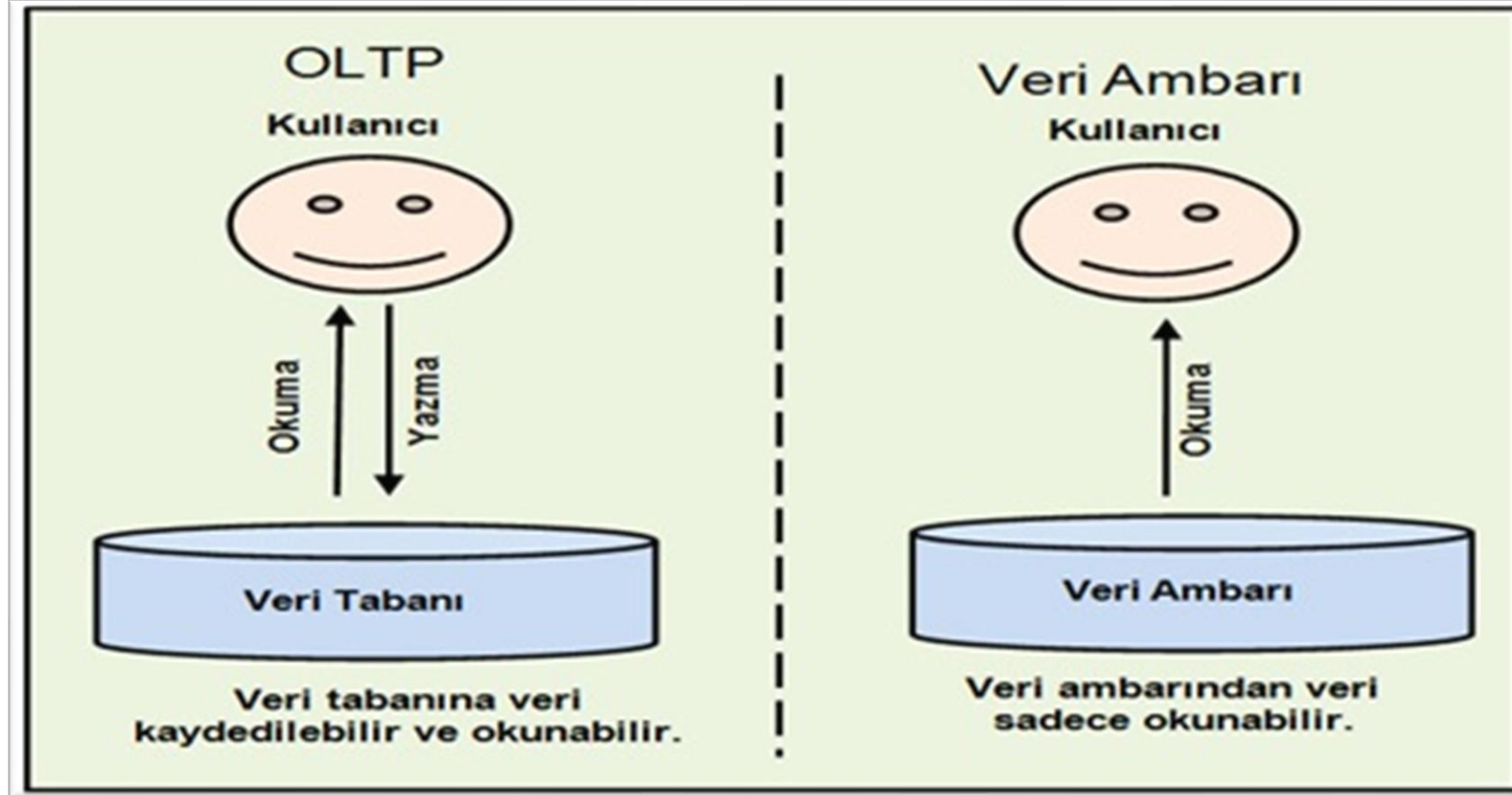
Veri Ambarı

SADECE OKUNABİLİR

Veri ambarındaki veri yönetimin gereksinimlerine yanıt vermek üzere tasarlandığı için günlük işlemlere tabi tutulmaz, yani silinemez ve güncellenemez.

Veri ambarına veri yükleme işlemi belirli zamanlarda yapılır ve bu aşamadan sonra kullanıcılar verilere erişebilirler.

ZAMAN BOYUTU



Veri Ambarı

Temel Özellikler

- İşlemsel ortamdaki veri, bir süzme işleminden geçirilerek karar destek sistemlerinin gereksinim duyacağı kadarı veri ambarına aktarılır, verinin tamamının aktarılması beklenmez.
- İşlemsel ortamdaki veri çok tazedir, veri ambarındaki veri daha eskidir.
- Veri ambarı özet bilgileri içerebilir, işlemsel sistemlerde özet veriye yer verilmez.
- Bütünleşme sürecindeki dönüşüm işlemleri, işlemsel veri ile veri ambarındaki veri arasında içerik farklılıkları oluşturabilir.

Veri Ambarı Veri Modeli

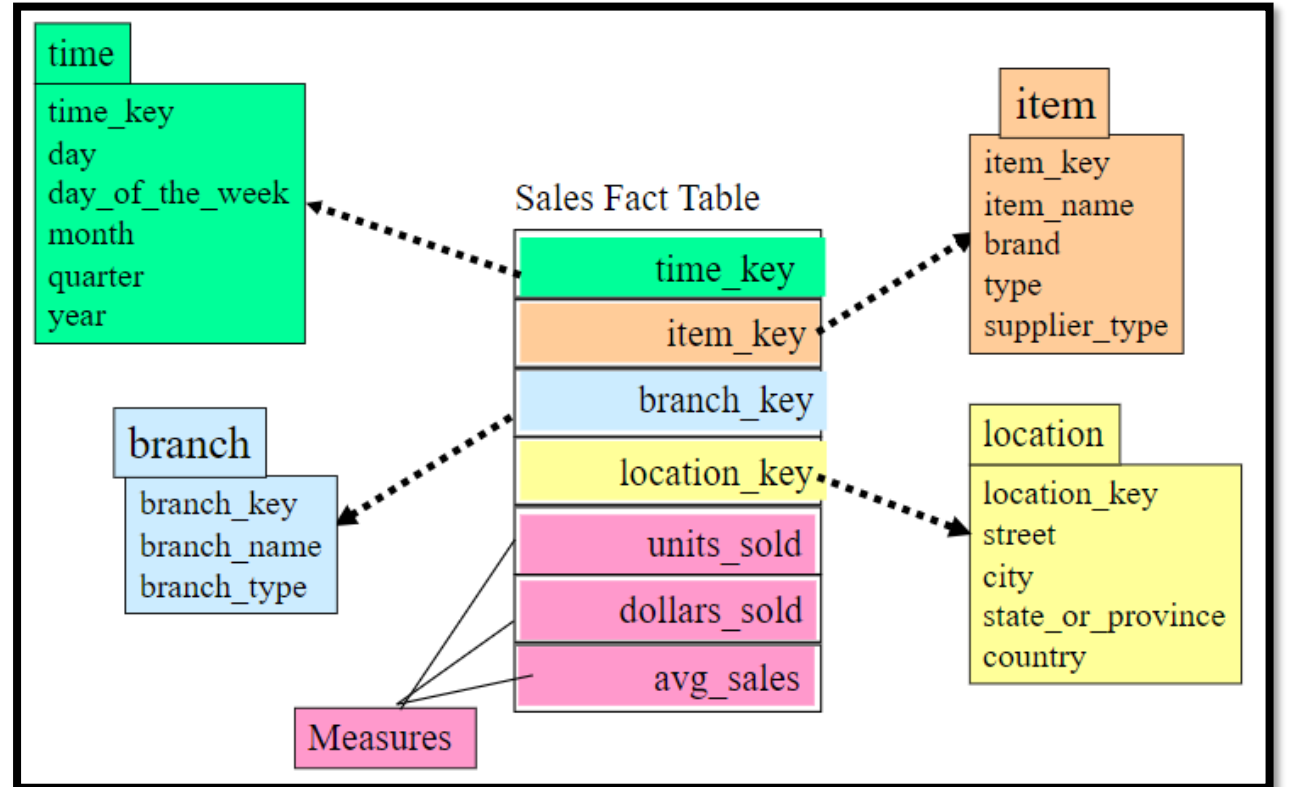
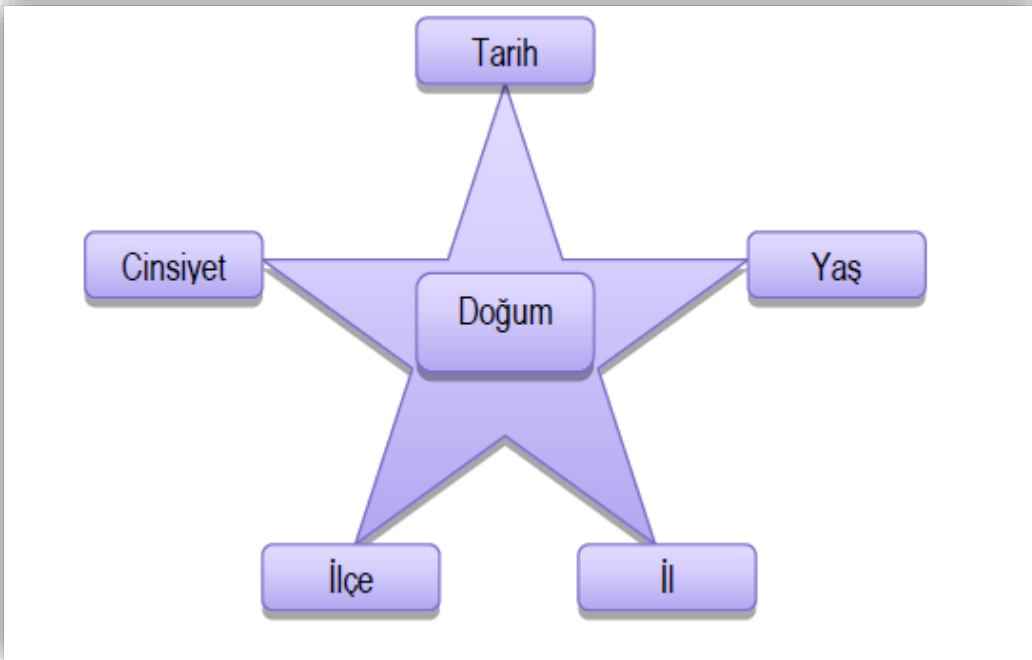
- Veri ambarları veriyi en kısa sürede çözümleyerek kullanıcıya ulaştırmayı amaçlar.
- Veri ambarı gündelik işlemleri yürütmek için değil, eldeki veriyi hızlı bir şekilde çözümlemek için tercih edilir.
- Veri ambarlarında karmaşık sorguların yüksek performans ile çalıştırılmasına imkan tanıyan bir veri modeli seçilmektedir.
- Veri ambarı bir veya birkaç ana tablodan oluşur. Bu ana tabloya (fact) bağlı olarak birçok boyut (dimension) tablosu veri modeli içinde yer alır. (Veri ambarında veriler ana tablolarda ve boyut tablolarında tutulur.)
- Bir işletmenin satış verilerini içeren bir veri ambarında yıl, ürün grubu, mağazalar birer boyut olarak düşünülebilir.

Veri Ambarı Veri Modeli

- Veri ambarının veri modeli, işletmenin gereksinimlerine dayalı olarak “boyutsal model” olarak düşünülür. Bir veri ambarında birden fazla boyut yer alabilir. Bu nedenle söz konusu modele “çok boyutlu model” denir.
- Veri Ambarları çok boyutlu bütünleşik yapılara göre düzenlenmektedir. Farklı boyutların toplanması ile de bir küp yapısı oluşturulmuş olmaktadır.
- Veri küplerinde mutlaka bir boyut, zaman boyutu olmalıdır.
- Veri ambarı tasarlanırken en fazla tercih edilen modeller:
 - Yıldız (Star) Şema
 - Kar Tanesi (Snowflake) Şema
 - Galaksi (Galaxy) yada Gerçek takımyıldızı (fact constellation) Şema
- Yıldız şema yapısı en sık tercih edilen şema modelidir.

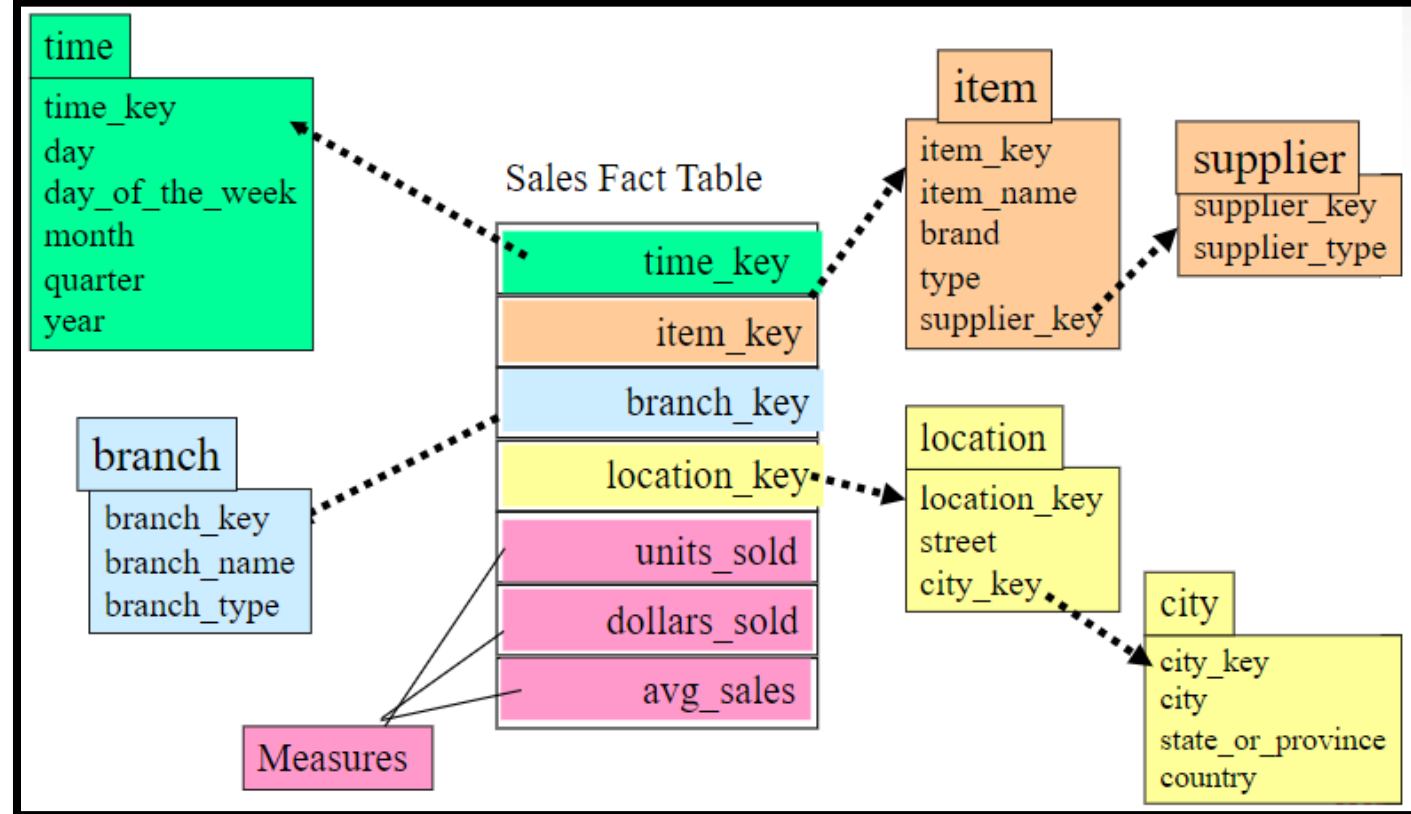
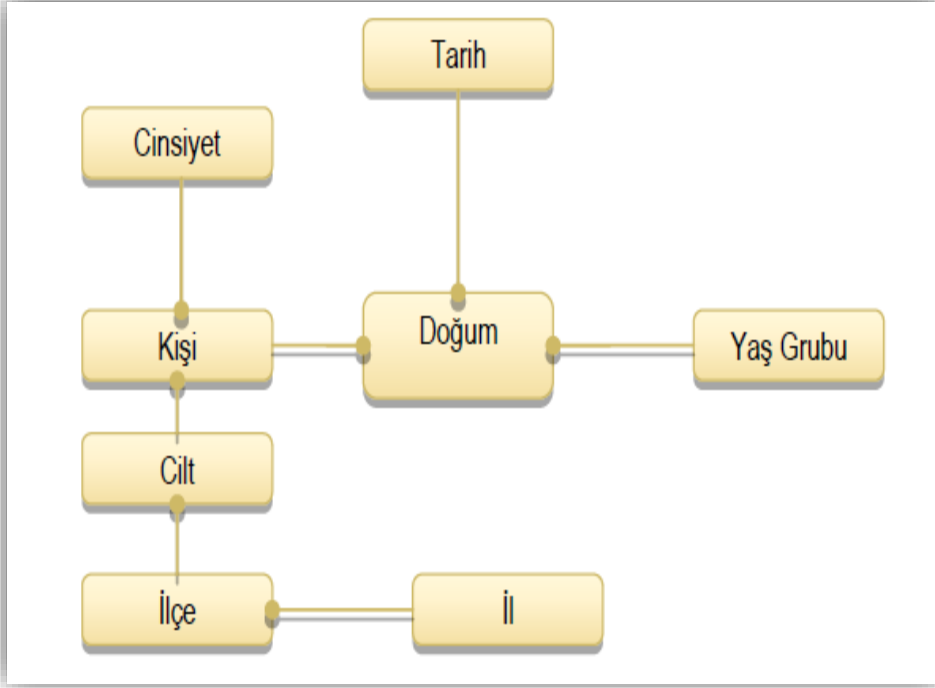
Yıldız (Star) Şema

Merkezde bulunan bir *olay tablosu* (fact table), çevresindeki *boyut tablolarını* (dimension table) birleştirir. En fazla tercih edilen modeldir.

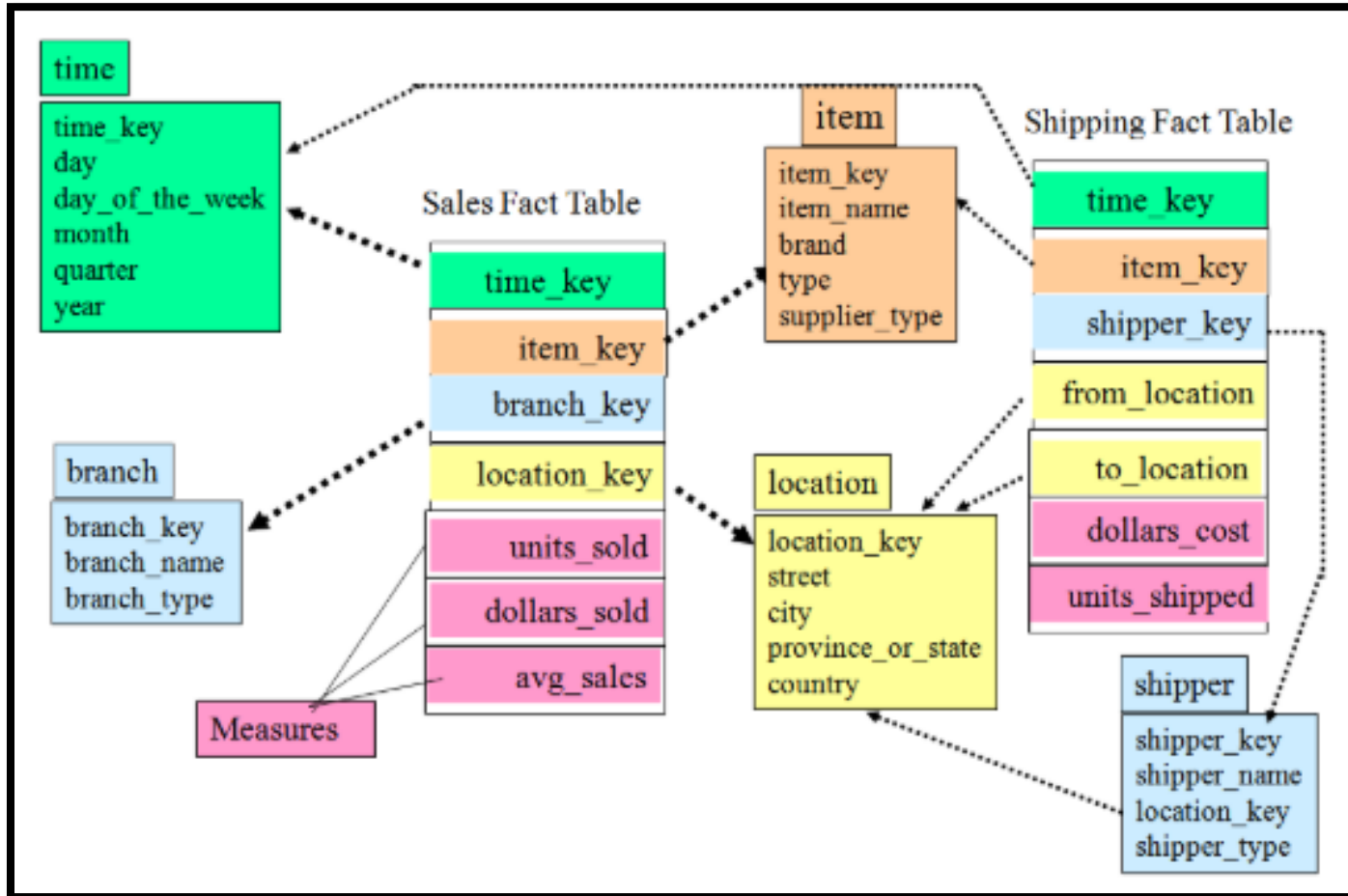


Kartanesi (snowflake) Şema

Yıldız şemaya benzer, yıldızın uç noktaları genişleyerek alt boyutlara açılır. Görüntüsü kartanesine benzediği için bu ismi almıştır. Bu şema yapısının kurulması çok zaman alması ve bilgiye ulaşılmasını zorlaştırması nedeniyle de pek fazla tercih edilmemektedir.



Galaksi (Galaxy) yada Gerçek takımyıldızı (fact constellation) Şema





MetaData

- Metadata (metaveri yada üst veri) veri ambarı içeriğinin ne olduğunu belirten veri bilgisidir. Veriye ilişkin veri olarak tanımlanabilir.
- Veri ambarı içindeki veriden ayrı bir yerde tutulur.
- Kullanımına göre ikiye ayrılır:
 - ❑ ***İşletme perspektifi açısından metadata:*** Veri ne anlama geliyor? Nerede bulabilirim? sorularının cevabını içerir. Bir sözlük gibi düşünülebilir.
 - ❑ ***Teknik perspektif açısından metadata:*** Teknik personel için gerekli bilgilerin bulunduğu kütüktür.

Metadata

Meta Veri Türü	Tanım	Dönüştürme	Yönetim
İş perspektifi (Çoğunlukla Yapısal Değil)	Ne anlama geliyor? Nerede bulabilirim?	Nasıl hesaplanmış? Kaynakları neler? Hangi iş kuralları uygulanmış?	Hangi eğitimler var? Yönlendirme takımında kimler var? En kolay şekilde nasıl ulaşılır? Bilgi ne kadar güncel?
Teknik perspektif (Çoğunlukla Yapısal)	Biçim Uzunluk Tanım aralığı Veri tabanı Katalog	Filtreler Topluluklar Hesaplamalar İfadeler	Kapasite planlama Boşluk atama İçerik& içerikselleştirmeme Disk kullanımı İş çizelgeleme



Veri Küpleri – Çok boyutlu veri modeli

- Veri küpü, verilerin modellenmesine ve birden çok boyutta görüntülenmesine olanak sağlar.
- Veri küpü, boyutlar (dimensions) ve olaylar (facts) tarafından tanımlanır.
- Her bir dilim bir tabloyu temsil edecek şekilde tablolar bir araya gelerek küp yapısını oluşturur.
- Küpleri genellikle 3-B geometrik yapılar olarak düşünmemize rağmen, veri ambarını depolamada veri küpü n boyutludur.
- Çok boyutlu bir veri modeli, merkezi bir temanın etrafında düzenlenir. Bu tema bir olgu tablosu ile temsil edilmektedir.

Veri Küpü oluşturma

- 2 boyutlu tabloda ürün-zaman boyutları için satış değerleri

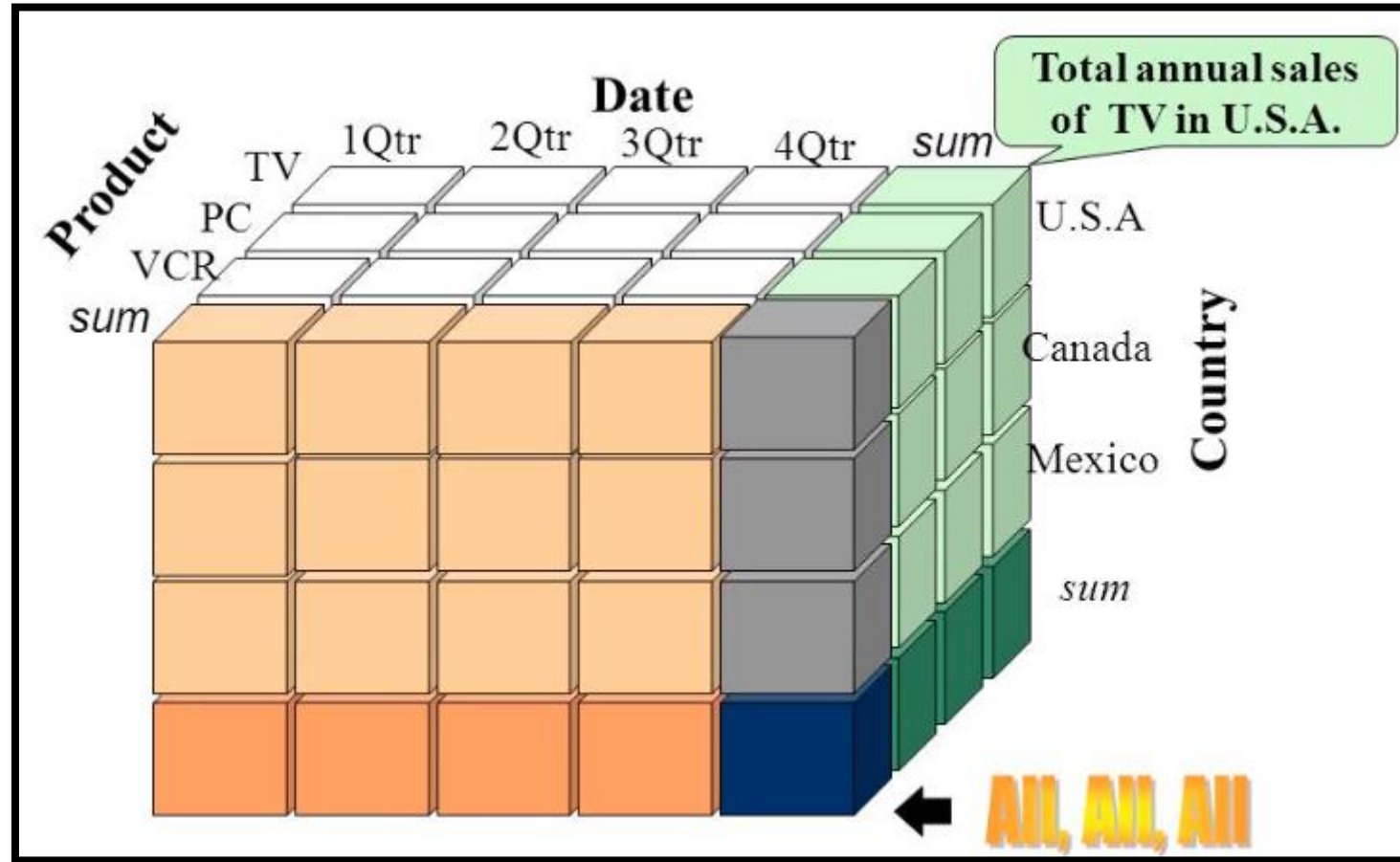
	TV	PC	VCR
1 st Qtr	1000	850	350
2 nd Qtr	1352	940	298
3 rd Qtr	1450	658	314
4 th Qtr	1500	965	365
USA			

Çok boyutlu yapıya geçiş

- 2 boyutlu tabloda ürün-zaman boyutları için satış değerleri

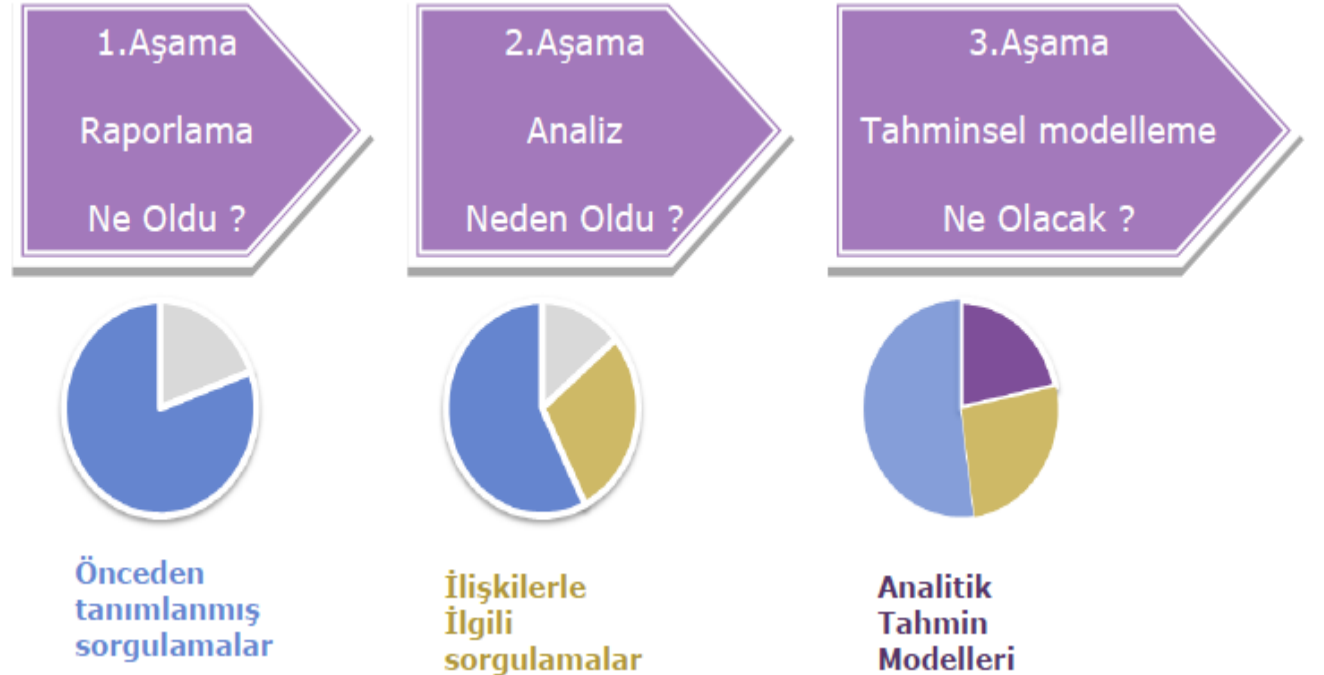
	TV	PC	VCR	TV	PC	VCR	TV	PC	VCR
1 st Q	1000	850	350	2600	750	425	1300	850	350
2 nd Q	1352	940	298	1752	860	236	1200	1000	400
3 rd Q	1450	658	314	1055	458	520	1150	555	510
4 th Q	1500	965	365	1350	1065	390	900	750	425
	USA			Canada			Mexico		

Veri Küpü



Veri Ambarı Sorgulama

Veri ambarları önceleri raporlama amaçlı sorgulamalar için kullanılırken daha sonra analiz aşamasına ve tahminsel modellemeler yapma aşamasına geçmek mümkün olmuştur.



Veri Ambarı Sorgulama

Raporlama aşamasında yapılan sorgulamalar :

- Geçen X aylık dönemde, müşteri ilişkilerinde en karlı dağıtım kanalı hangisidir?
- Ürün ve müşteri grupları için, aylık ve haftalık satış miktarları ne kadardır?
- Kimlerin bakiyeleri belirli bir miktarın altında?

Veri Ambarı Sorgulama

Analiz aşamasında yapılan sorgulamalar :

- Zaman içinde geline bu aşama, olaylar arasında bağlantı kurmak, birbirleriyle ilişkili ve birbirini etkileyen değişkenleri ortaya çıkarmak şeklinde gerçekleşmektedir.
- Örneğin, cinsiyet ile satın alma tutumu arasında nasıl bir ilişki olduğu, kadınların satın alma tutumlarının erkeklerden farklı olmasının sebeplerinin, belirli bölgelerde satışların artmasının veya azalmasının sebeplerinin neler olduğu türünde sorulara cevap bularak, ilişkilerin analiz edilmesini gerektiren sorgulamalardır.

Veri Ambarı Sorgulama

Tahminsel Modelleme aşamasında yapılan sorgulamalar:

- "Önümüzdeki günlerde, haftalarda hangi müşterilerimizi kaybetme riskimiz var ve bu riskin derecesi nedir?"
- Bu sorgulama sonucu elde edilen "çıkarımsal bilgi", "knowledge" olarak ifade edilen, keşifsel ve proaktif bir bilgidir.
- Bu bilgiyi elde etmekteki amaç, kaybetme riski olan müşteriye ayırdıktan sonra hangileri için bunun engellenmesi gerektiğini belirlemektir. Yani en azından, kaybetme riskini azaltmak için bir eylem planı, bir strateji geliştirmek ve proaktif davranmak için kullanılacak bir bilgidir.
- Bu aşama, özellikle, **veri ambarlarının** klasik veri tabanı yaklaşımından farkını en belirgin biçimde ortaya koyan aşamadır. Veri ambarları aslında ileri düzey kullanıcılar ve tahmin yapma ihtiyacı bulunan, stratejik kararları verenler için, modelleme yapabilmek ve eyleme yönelik bilgiler elde etmek açısından önemli fırsatlar sunmaktadır. Bu aşamada, veri madenciliğinin tahmin modelleri de devreye girmekte, müşterilerin tutum ve davranışları hakkın da geleceğe yönelik tahminler yapılabilmektedir.

OLTP (Online Transaction Processing)

- OLTP sistemler genellikle ilişkisel veritabanları üzerine kurulmuş, üzerinde sürekli işlem yapılan veritabanı sistemleridir. Kurumların günlük verileri genellikle bu ortamlarda işlenir.
- OLTP sistemler, kayıt okuma, kayıt ekleme, kayıt silme, güncelleme gibi işlemlerin yoğun olarak gerçekleştirildiği veri depolama sistemleridir.
- OLTP sistemlerin karar destek faaliyetlerinde kullanılması performans açısından tavsiye edilmez.

OLAP (Online Analytical Processing)

- OLAP(On-Line Analytical Processing) : Veriler üzerinde çok boyutlu analizler yapılmasına imkan sağlayan bir yaklaşımdır. OLAP veri ambarlarına ait sorgulama ve oluşturma işlemlerini de kapsar.
- Çok boyutlu tasarlanmış bir veri ambarında, siparişler ile ilgili bilgilerin yer aldığı tabloda, “zaman” boyutunun elemanlarına denk gelen gün, hafta, ay, yıl gibi sütunlar bulunabilir ve bu da OLAP’ın boyut modellemesine olanak sağlar.
- Veri ambarı ve OLAP terimleri bazı kaynaklarda birbirlerinin yerine kullanılmaktadır fakat bu yanlış bir yaklaşımdır. Veri ambarı özelleştirilmiş bir veritabanını belirtirken, OLAP eldeki veriler üzerinde çok boyutlu sorgular yapmayı kolaylaştıran bir analiz yaklaşımıdır.
- Veri ambarları ve OLAP birbirlerini tamamlarlar.

OLTP / OLAP farkı

- OLTP üzerinde SQL sorguları kullanılırken, OLAP üzerinde ise SQL diline benzeyen ve OLAP'la ilgili küp, slice, fact ve dimension gibi OLAP'a özgü konuları da destekleyen MDX (Multi Dimensional eXpression Language) sorgulama dili kullanılır.

KRİTER	OLTP	OLAP
Amaç	Günlük operasyonel kullanım (işlemler ve sorgular)	Karar destek, veri analizleri (sorgular)
Kullanıcı	IT personeli, müşteriler, ofis çalışanları vb (kullanıcı sayısı çok)	Yöneticiler ve analistler (kullanıcı sayısı az)
Veri Kaynağı	İlişkisel Veritabanı	Veri ambarı yada özel veritabanı
Raporlama	Rutin, periyodik raporlar	Özel amaçlı, çok boyutlu, geniş odaklı rapor ve sorgular
Erişim	Okuma/yazma/güncelleme	okuma
Uygulama Hızı	Hızlı (rutin sorgular)	Yavaş (yoğun, karmaşık sorgular)
Veritabanı Dizaynı	Varlık-iliski veri modeli ile dizayn edilmiş uygulama odaklı yapı	Genellikle yıldız yada kar tanesi modeli ile dizayn edilmiş yapı

OLAP küp bileşenleri

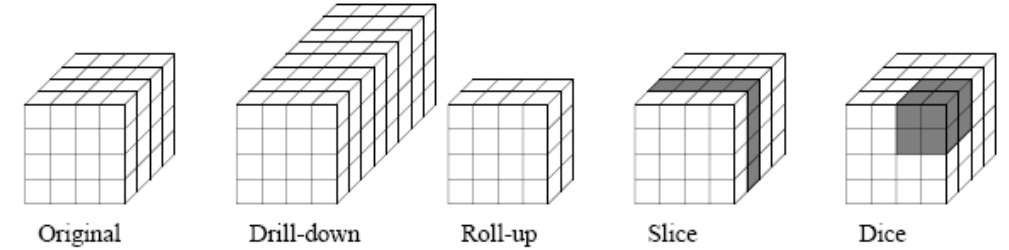
- **Küp:** Ölçüleri, çözümlemek istediğiniz boyutların her birinin düzey ve hiyerarşilerine göre derleyen bir veri yapısıdır. Küpler; zaman, coğrafya ve ürün türü gibi çeşitli boyutları satış veya stok rakamları gibi özet verilerle birleştirir. Küpler, matematikte bildiğimiz anlamda birer "küp" değildirler, çünkü eşit kenarlara sahip olmaları gerekmez.
- **Ölçü:** Bir küpte bulunan ve küpün bulgu tablosundaki bir sütunu temel alan değerler kümesidir ve genellikle sayısal değerler halindedir. Ölçüler, küpteki önceden işlenmiş, derlenmiş ve çözümlenmiş merkezi değerlerdir. Yaygın örnekler arasında satış, kar, gelir ve maliyet sayılabilir.
- **Üye:** Bir hiyerarşide, bir veya daha çok veri yinelenmesini temsil eden öğedir. Bir üye benzersiz veya benzer olabilir. Örneğin, bir zaman boyutunun yıl düzeyinde 2007 ve 2008 benzersiz üyeleri temsil eder, ancak ay düzeyinde Ocak benzer üyeleri temsil eder çünkü birden fazla yılın verilerini içeriyorsa zaman boyutunda birden fazla Ocak bulunabilir.
- **Hesaplanan üye:** Değeri çalışma esnasında ifade kullanılarak hesaplanan boyut üyesidir. Hesaplanan üye değerleri diğer üyelerin değerlerinden türetilebilir. Örneğin, Kar hesaplanan üyesi, Maliyetler üyesinin değerini Satışlar üyesinin değerinden çıkararak belirlenebilir.

OLAP küp bileşenleri

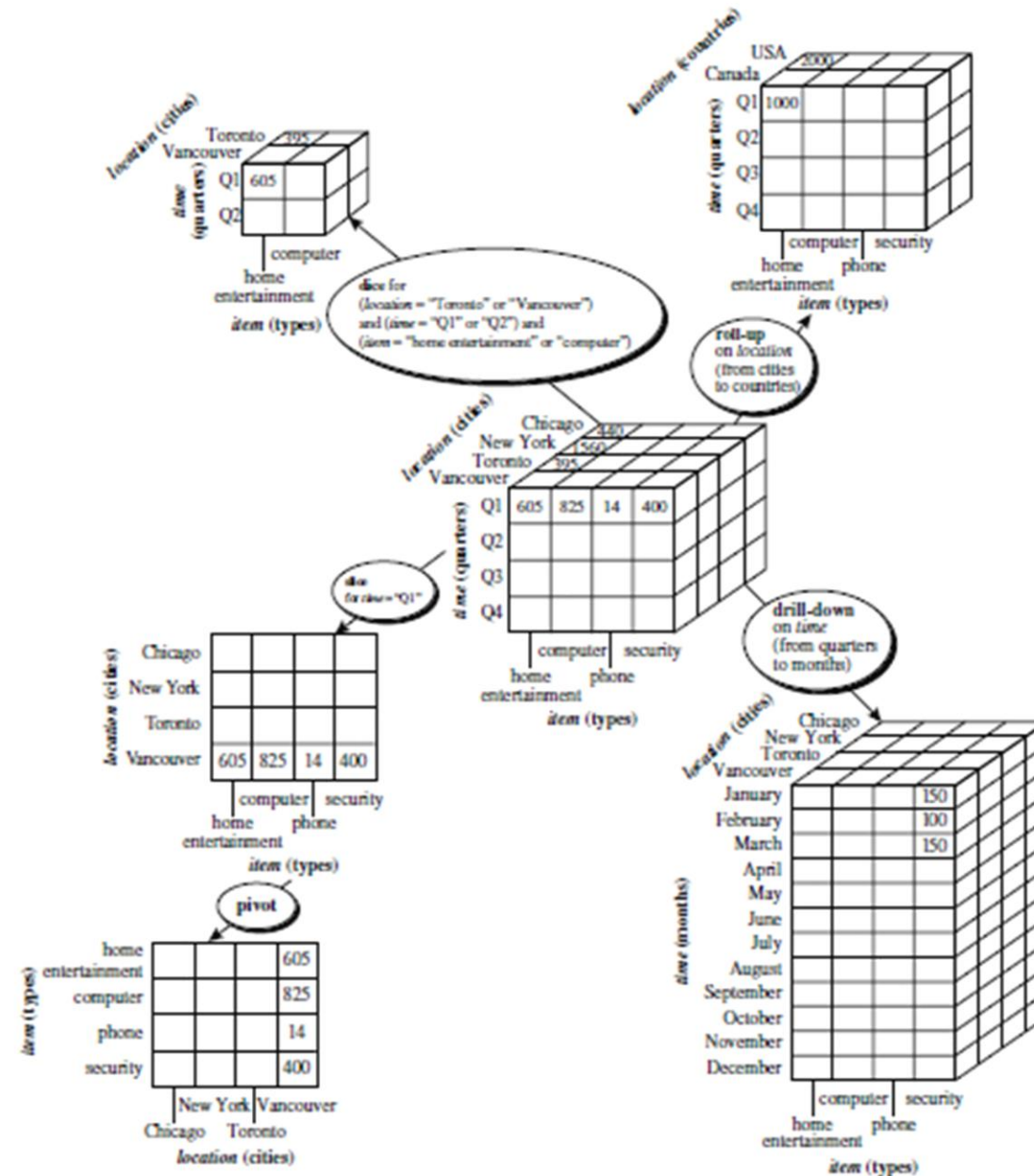
- **Boyut** Bir küpte, kullanıcının anladığı ve veri çözümleme için temel olarak kullandığı düzenlenen bir veya daha fazla düzey hiyerarşileri grubudur. Örneğin bir coğrafya boyutu Ülke/Bölge, İl/İlçe ve Şehir için düzey içerebilir. Ya da bir zaman boyutu yıl, çeyrek, ay ve gün için hiyerarşi düzeyleri içerebilir. Özet Tablo raporu veya Özet Grafik raporunda, her hiyerarşi, daha düşük veya daha yüksek düzeyleri göstermek için genişletebileceğiniz ve daraltabileceğiniz alanlar kümesi haline gelir.
- **Hiyerarşi** Boyutun üyelerini, her üyenin bir üst üyesi ve sıfır veya daha fazla alt üyesi olacak şekilde düzenleyen mantıksal ağaç yapısıdır.
- **Düzy** Bir hiyerarşi içinde, veriler alt ve üst ayrıntı düzeyleri olarak düzenlenebilir. Örnek olarak Zaman hiyerarşisinde Yıl, Çeyrek, Ay ve Gün düzeyleri verilebilir.

Bir OLAP küpü üzerinde şu işlemler yapılabilir:

- **Dice(Çevir):** Satış verisinin bölge-zaman yüzünü incelerken, ürün-zaman yüzüne geçebiliriz.
- **Slice and dice (Dilimle):** Bütün aralığı değil de belirli bir aralığı seçebiliriz; son 1 yıla ait dilim..
- **Drill Down:** Ayrıntı seviyesinde alta in; yıl bazından ay bazına geç (detaylandırma)
- **Roll Up:** Ayrıntı seviyesinde yukarı çık; şehir bazından bölge bazına çık (özetleme)
- **Pivot (rotate):** Alternatif bir veri sunumu sağlamak için veri eksenlerini görünümde döndüren bir görselleştirme işlemidir.



Çok boyutlu veri üzerinde OLAP işlemleri



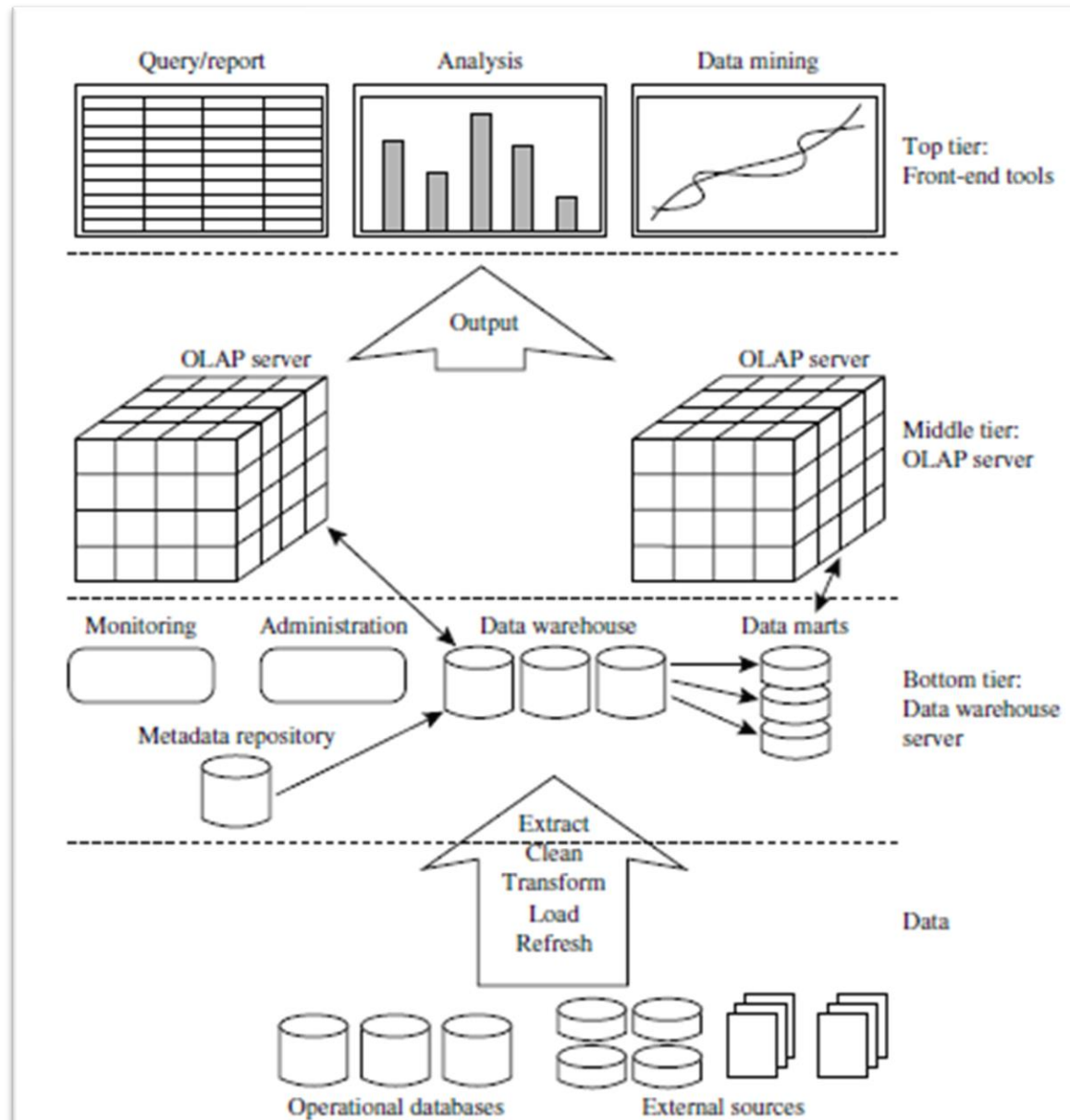
OLAP küp bileşenleri

Buradan OLAP küplerinin sadece 3 boyutlu olabileceğini çıkarmamalıyız. Daha az veya çok boyut da olabilir.

OLAP küpleri, tutuldukları yerlere göre farklı isimler alırlar.

- Eğer küpler çok boyutlu veritabanında tutuluyorsa **MOLAP (Multidimensional OLAP)**,
- Web üzerinden erişiliyorsa **WOLAP (Web OLAP)**,
- Uç birimde tutuluyorsa **DOLAP (Desktop OLAP)**,
- İlişkisel veritabanında tutuluyorsa **ROLAP (Relational OLAP)** adını alır.

Özet



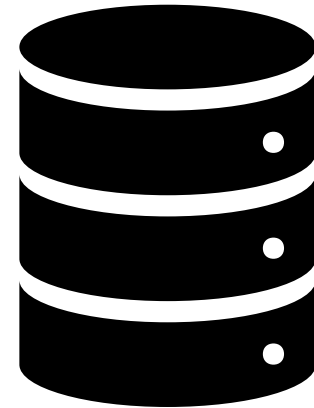
Ücretsiz veri setleri

<http://archive.ics.uci.edu/ml/index.php> (UCI Machine Learning Repository)

https://github.com/viktree/curly-octo-chainsaw/blob/master/BreadBasket_DMS.csv Market sepet analizi için veri

<https://github.com/awesomedata/awesome-public-datasets#machinelearning> Liste halinde farklı veri kaynakları

<https://www.kaggle.com/datasets>



Referanslar

- ***Veri madenciliği ders notları, Dr. Burcu Çarklı YAVUZ, SAÜ (Temel İçerik)***
- Veri Madenciliği Ders Notları, Doç. Dr. Nilüfer YURTAY, SAÜ
- Data Mining: Concepts and Techniques, J. Han, M. Kamber, J. Pei
- Veri madenciliği Kavram ve Algoritmaları, Gökhan Silahtaroğlu, Papatya Yayıncılık
- Veri Madenciliği Yöntemleri, Yalçın Özkan, Papatya Yayıncılık

