BTK AKADEMİ

# Veri Maratonu / 11. Gün

## Büyük Veri Ekosistemi

Talha KILIÇ

Fırat Üniversitesi Bilgisayar Mühendisliğinden mezun olan Talha KILIÇ, üniversite öğreniminin ilk yılından beri çeşitli şirketlerde Nesnelerin İnterneti, Bulut Bilişim ve Büyük Veri teknolojileri üzerine çalışmıştır. Birçok etkinlik de Büyük Veri teknolojileri ile ilgili sunumlar ve eğitimler vermiştir

Şu an Türkiye'de özel bir banka da büyük veri mühendisi olarak çalışan Talha KILIÇ, büyük veri projeleri geliştirmenin yanında online ve sınıf içi eğitim vermektedir. Bu eğitimlere 70000'den fazla öğrenci katılmış ve bir çok kişiye istihdam olmuştur



Talha KILIÇ

Büyük Veri Mühendisi / Eğitmen

@talhaklc

# Büyük Veri Ekosistemi

















## Eğitimin Amacı;

Katılımcılara Büyük Veri Ekosistemi içerisindeki en popüler teknolojiler hakkında bilgiler verirken, bu teknolojilerin kurulumunu, kullanımını, mimarisini ve örnek uygulamalarını gerçekleştirmektir. Bunlarla birlikte büyük verileri oluşturmak, aktarmak, saklamak, analiz etmek ve yönetmek için gerekli bilgi birikimini, iş hayatından tecrübe paylaşımları ile birleştirerek gerçek vaka analizleri ile pekiştirmektir

## Eğitim Gereksinimleri;

- Temel seviyede SQL bilgisi,
- Temel seviyede programlama bilgisi (Python,Java),
- Veri odaklı çalışma arzusu

**Eğitim Sonunda;**

- Büyük Veri Teknolojilerindeki temel kavramları, mimariyi, kurulumları ve terminolojiyi öğrenecek,
- En popüler Büyük Veri Teknolojilerini kullanacak,
- Yapılan Örnekler ile Kodlama Yeteneğiniz Geliştirecek,
- Cloud Sistemler Hakkında Bilgi Sahibi Olacak,
- NoSQL Veritabanlarına Giriş Yapabileceksiniz.

# Büyük Veri Nedir ?

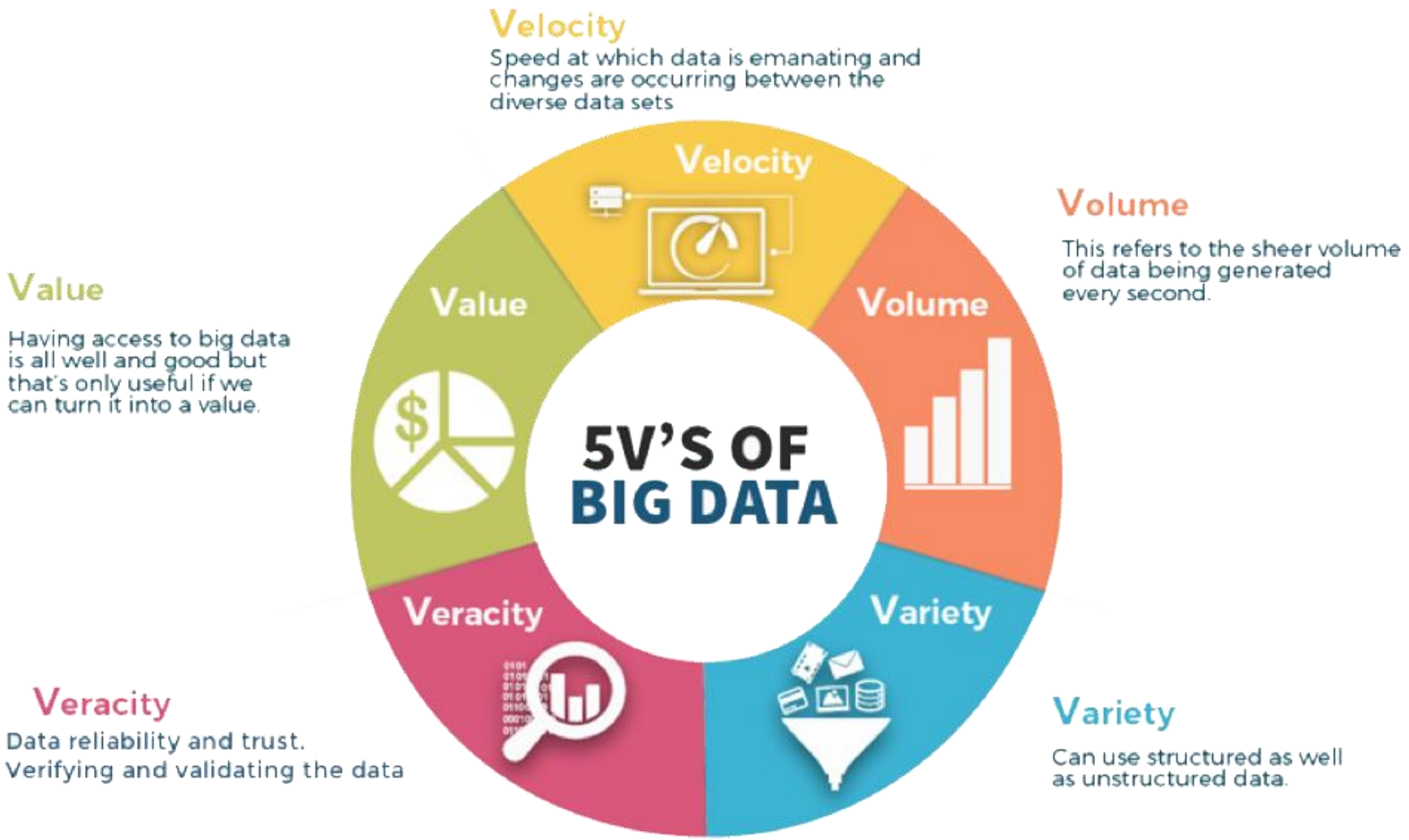
- Günümüzde son derece değerli ve popüler bir teknoloji olan **Büyük Veri**, teknolojinin ilerlemesiyle birlikte gün geçtikçe daha çok değerleniyor. Dünyada bir çok şirket, veri ile birlikte **dijital dönüşümü** gerçekleştirirken, bu dönüşüm Büyük Veri'yi yeni **bir devrin** başlangıcı olarak tanımlıyor.

- Büyük Veri 2000'li yılların başında analist Doug Laney ile popülaritesinde ivme kazanmaya başlamıştır. Büyük Veri kavramı büyük hacimli -yapılandırılmış ve yapılandırılmamış- verileri ifade etmek için kullanılmaktadır. Bu veriler o kadar büyük, hızlı ve komplikedir ki geleneksel yöntemler kullanarak işlemek mümkün değildir.

# Büyük Veri Nasıl Oluştu ?

- Sosyal Medya Verileri
- Mobil Telefon Uygulamaları
- Sensör Verileri (Akıllı ev, Otonom Araçlar vs)
- Log Verileri
- Gerçek Zamanlı Veri Analizi İhtiyacı
- Yapay Zeka Teknolojilerinin Veriye Bağımlılığı

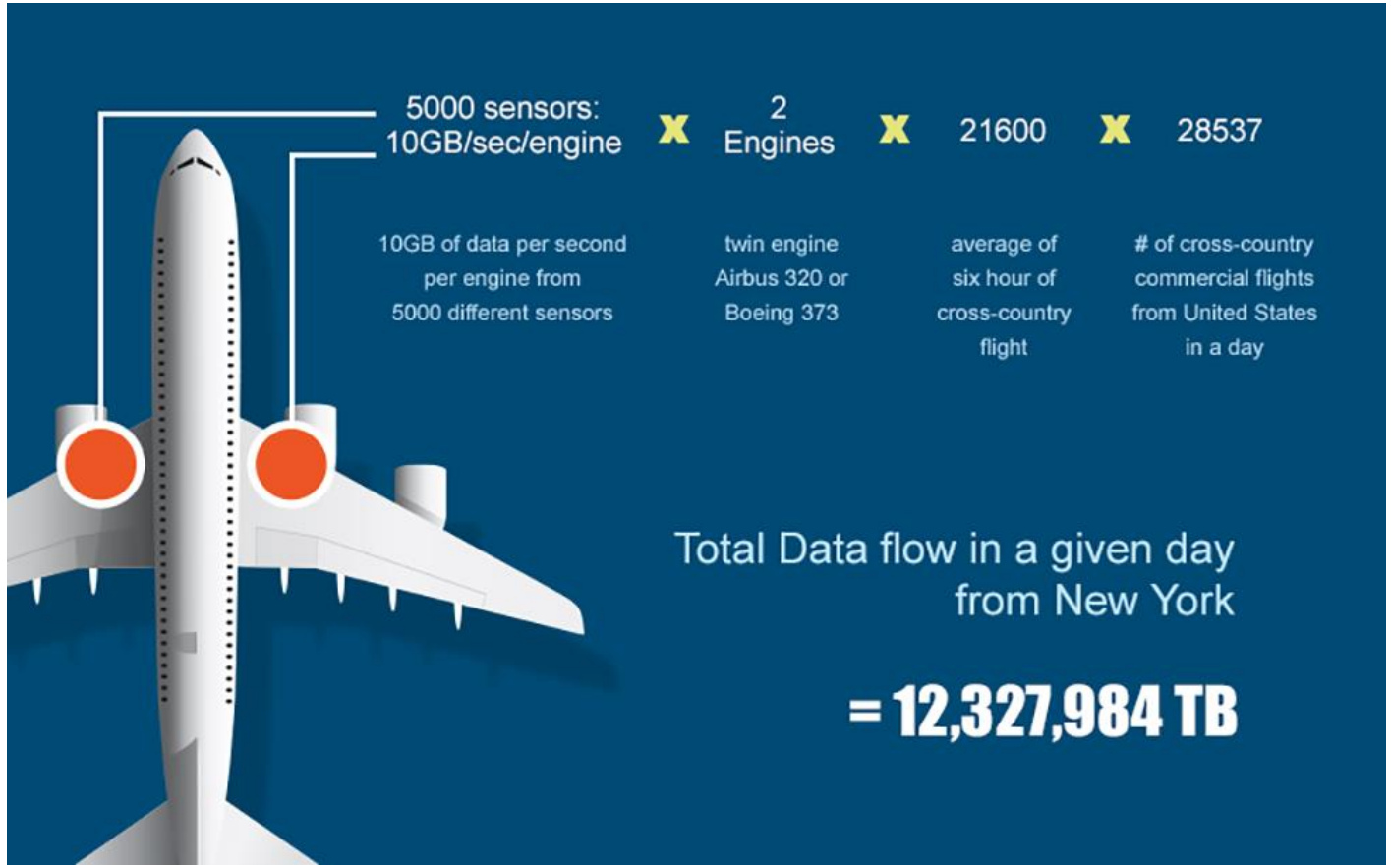
# Büyük Veri Bileşenleri



# Büyük Veri Özellikleri - Volume

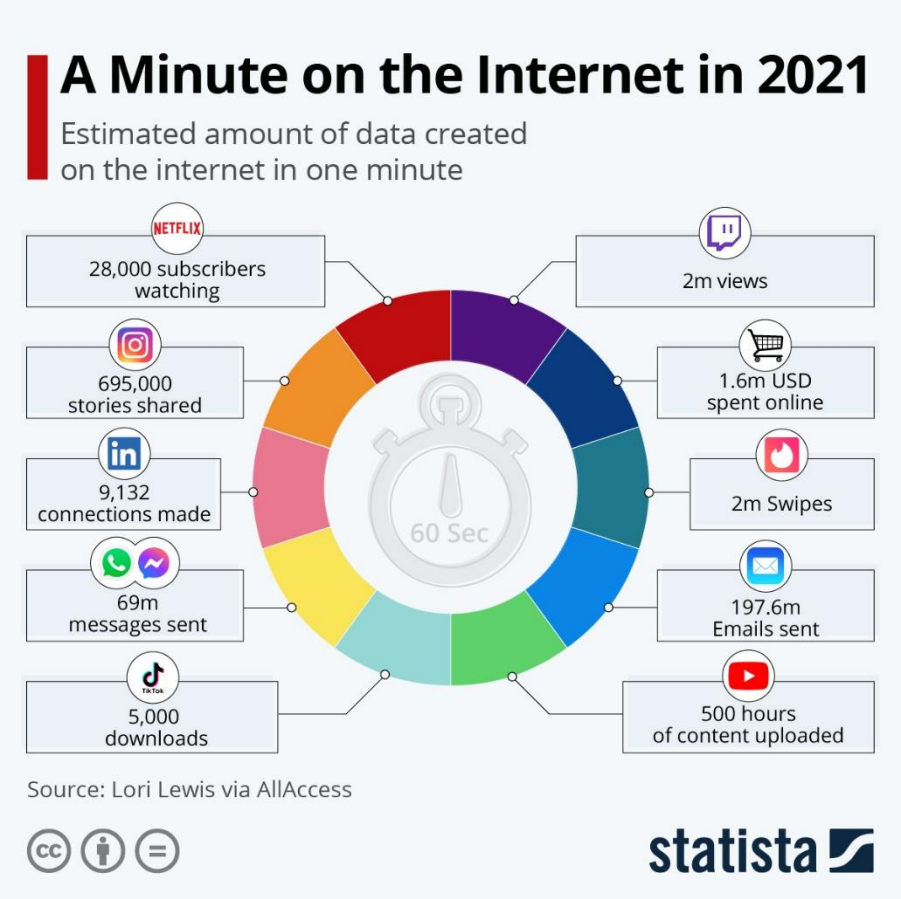
- Bir verinin ‘büyük veri’ olup olmamasının en önemli şartı ciddi boyutlarda olmasıdır. Verinin boyutu verinin değerini belirler

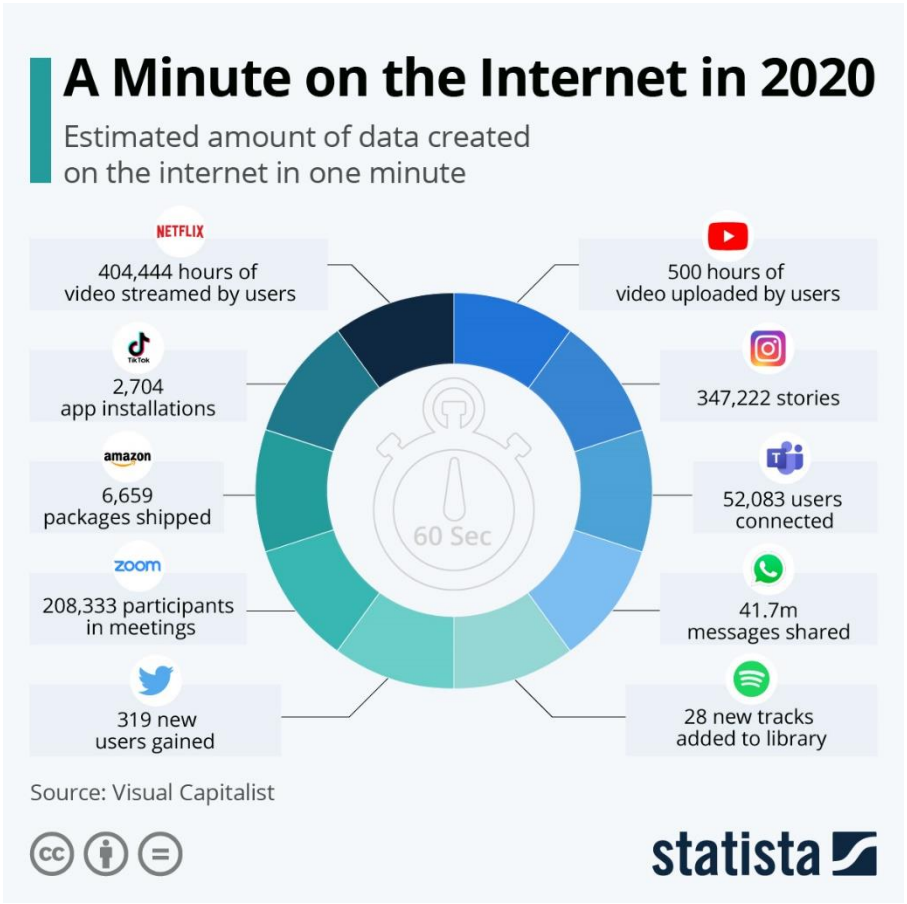
- Ticari işlemler, akıllı cihazlar (Iot), endüstriyel ekipman, video, sosyal medya gibi kaynaklardan toplanan verileri içermektedir. Geçmişte bu kadar büyük hacimli verinin depolanması sorun yaratabilirdi fakat günümüzde “Data Lakes” ve “Hadoop” gibi ucuz depolama platformları bu veri yükünün depolanmasına yardımcı olmaktadır.

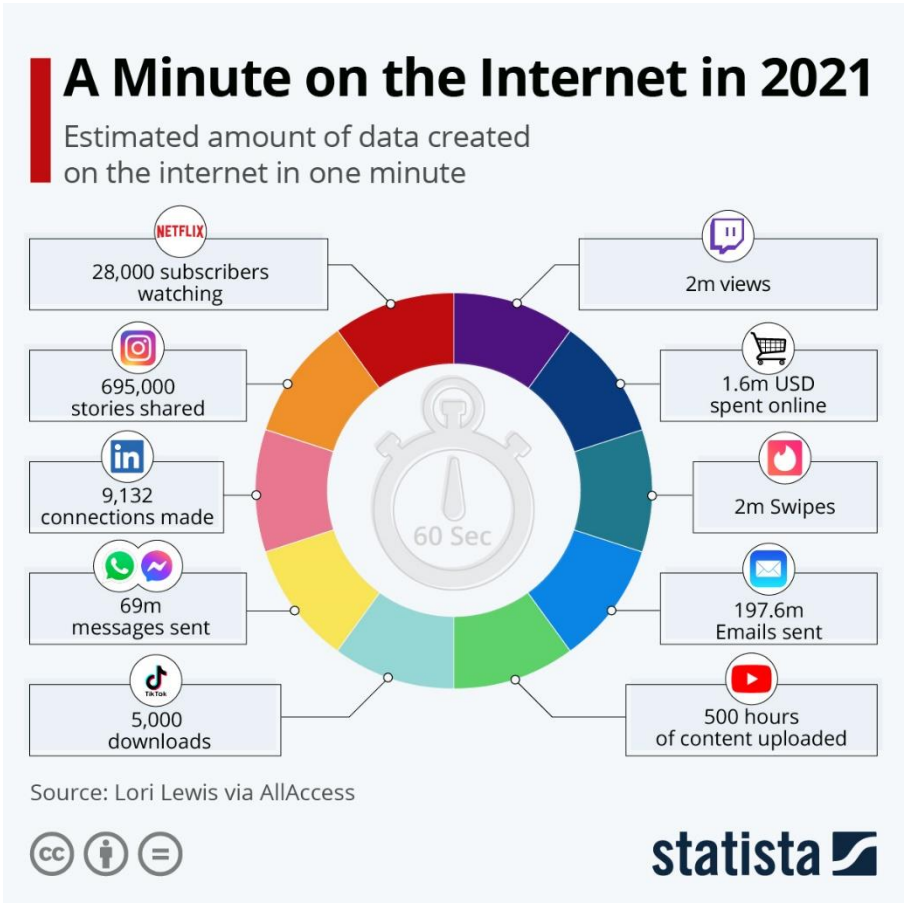


# Büyük Veri Özellikleri - Velocity

- Büyüme ve gelişme yolunda yatan talepleri ve zorlukları karşılamak için verilerin üretilme ve işlenme hızından bahsedilmektedir. Big Data genelde verilerin üretilmesi ile eş zamanlı olarak bir yandan da işlenmektedir. Her milisaniyede veri kaydedip ürettikçe, bu verileri de aynı hızla anlayabilmemiz gerekir. Trafiği izlemekten salgın yayılmaları izlemeye ve hisse senedi alım satımına kadar, zaman çok önemlidir. Bilgiyi anlamada birkaç saniyelik gecikme, yalnızca paraya değil, aynı zamanda hayata da mal olabilir.




A Minute on the Internet in 2020
Estimated amount of data created on the internet in one minute
NETFLIX
404,444 hours of video streamed by users
500 hours of video uploaded by users
TikTok
2,704 app installations
347,222 stories
amazon
6,659 packages shipped
52,083 users connected
zoom
208,333 participants in meetings
41.7m messages shared
319 new users gained
28 new tracks added to library
60 Sec
Source: Visual Capitalist
statista


A Minute on the Internet in 2021
Estimated amount of data created on the internet in one minute
NETFLIX
28,000 subscribers watching
2m views
695,000 stories shared
1.6m USD spent online
9,132 connections made
2m Swipes
69m messages sent
197.6m Emails sent
TikTok
5,000 downloads
500 hours of content uploaded
60 Sec
Source: Lori Lewis via AllAccess
statista

# Büyük Veri Özellikleri - Variety

- Veriler, geleneksel veri tabanlarındaki yapılandırılmış sayısal verilerden yapılandırılmamış metin belgelerine, e-postalara, videolara, seslere, hisse senedi verileri ve finansal işlemlere kadar her tür biçimde gelir.

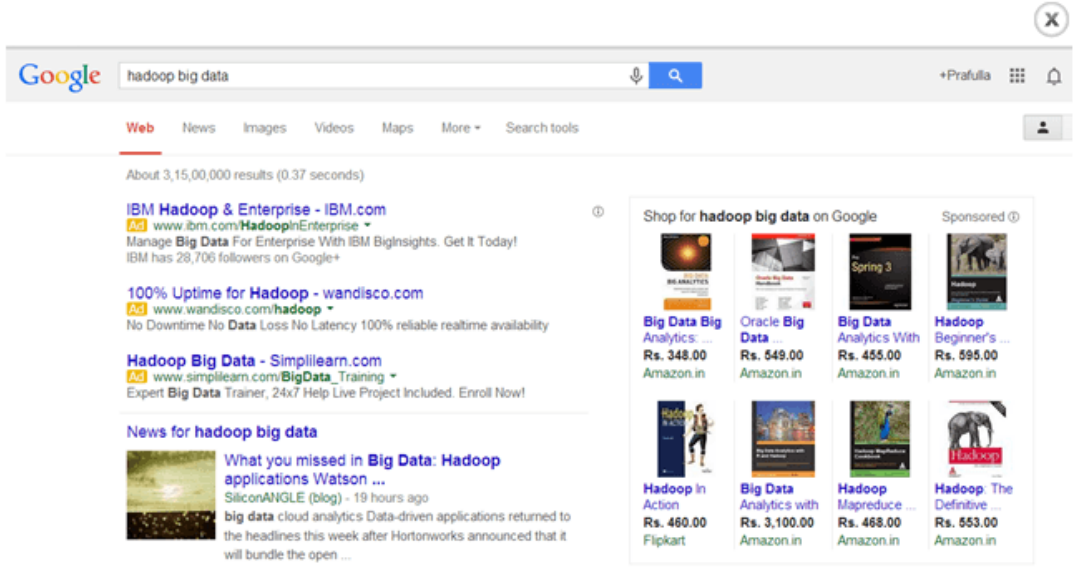
## Semi-Structured

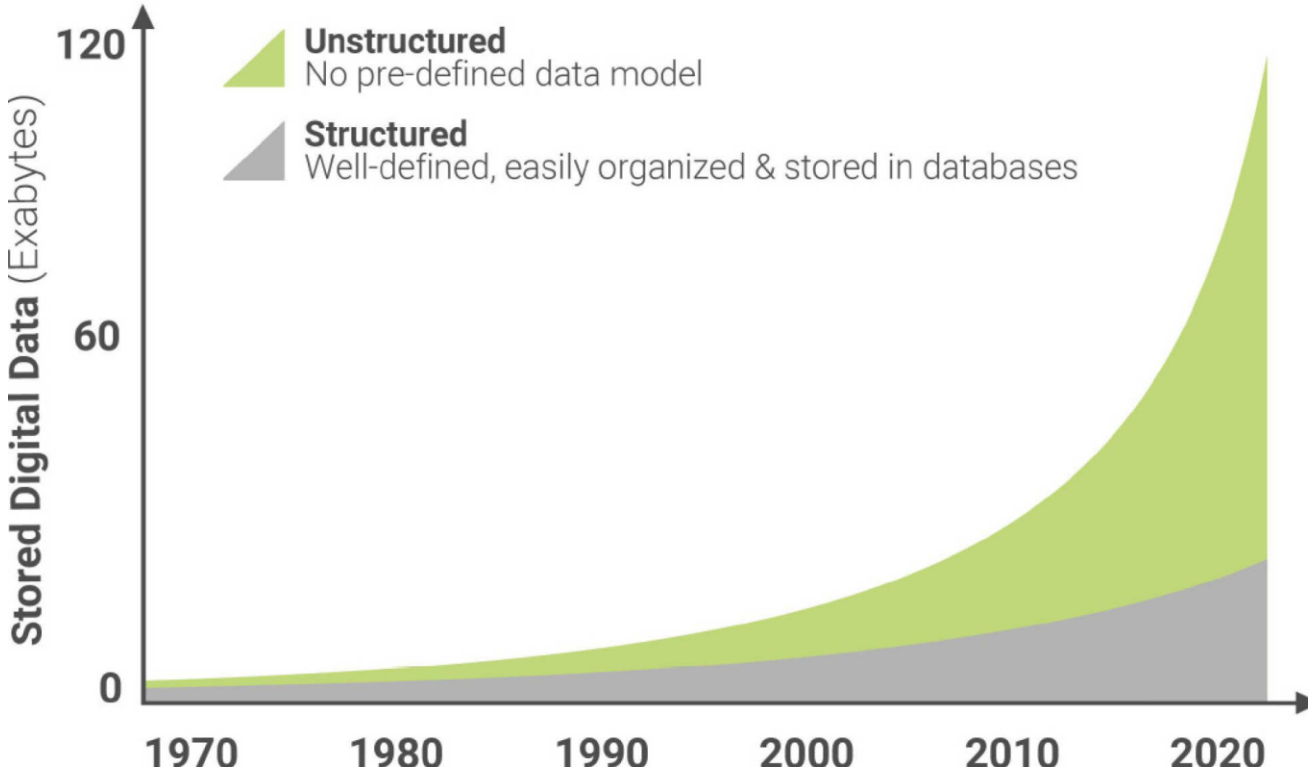
```
{
        "firstName": "John",
        "lastName": "Smith",
        "age": 25,
        "address": {
                "streetAddress": "21 2nd Street",
                "city": "New York",
                "state": "NY",
                "postalCode": 10021
        },
        "phoneNumbers": [
                {
                        "type": "home",
                        "number": "212 555-1234"
                },
                {
                        "type": "fax",
                        "number": "646 555-4567"
                }
        ]
}
```

## Structured

| Employee_ID | Employee_Name | Gender | Department | Salary_In_lacs |
|---|---|---|---|---|
| 2365 | Rajesh Kulkarni | Male | Finance | 650000 |
| 3398 | Pratibha Joshi | Female | Admin | 650000 |
| 7465 | Shushil Roy | Male | Admin | 500000 |
| 7500 | Shubhojit Das | Male | Finance | 500000 |
| 7699 | Priya Sane | Female | Finance | 550000 |

## Unstructured






Data Variety – Veri Çeşitliliği

# Büyük Veri Özellikleri - Veracity

- Verinin gerçek ve doğru olması olarak açıklanabilir. Büyük miktardaki verilere ait bilgilerin doğru olup olmadığına karar verilmesi için önemli bir yaklaşımdır.

  Araştırmalar 3 iş insanından 1'inin kendi verilerine güvenmediğini gösteriyor .Bu yüzden anlamlı ve doğru verileri bulunası için veri;
  Kategorize edilmeli,
  Analiz edilmeli ve
  Görsel hale dönüştürülmeli

## Büyük Veri Özellikleri - Value

- Büyük veri ile ilgili en önemli bileşen, değerdir. Elde edilen ve işlenen veriler, kuruma değer kattığı sürece anlamlıdır. Bu nedenle, büyük verinin analizinin ve simülasyonlarının doğru şekilde kurgulanması ve büyük veriyi kullanan kuruma fayda sağlaması öncelikli olarak ele alınmalıdır.

# Batch vs Real Time Veri Analizi

## Batch Veri Analizi

Belirlenen miktarda veri biriktiğinde ya da bir zaman planlandığında duran/biriken/saklanan veri üzerinde yapılan analizdir

İşlem Gerçekleşir → Veri Kaydedilir → Analiz için zaman beklenir / Schedule time

**Örnekler;**

- Fatura oluşturma
- Gün sonu raporları
- Csv, excel dökümanları
- Son 3 ayın ihracat verileri

## Real Time Veri Analizi

Akan verinin işlem gerçekleştikten hemen sonra analiz edilmesidir.

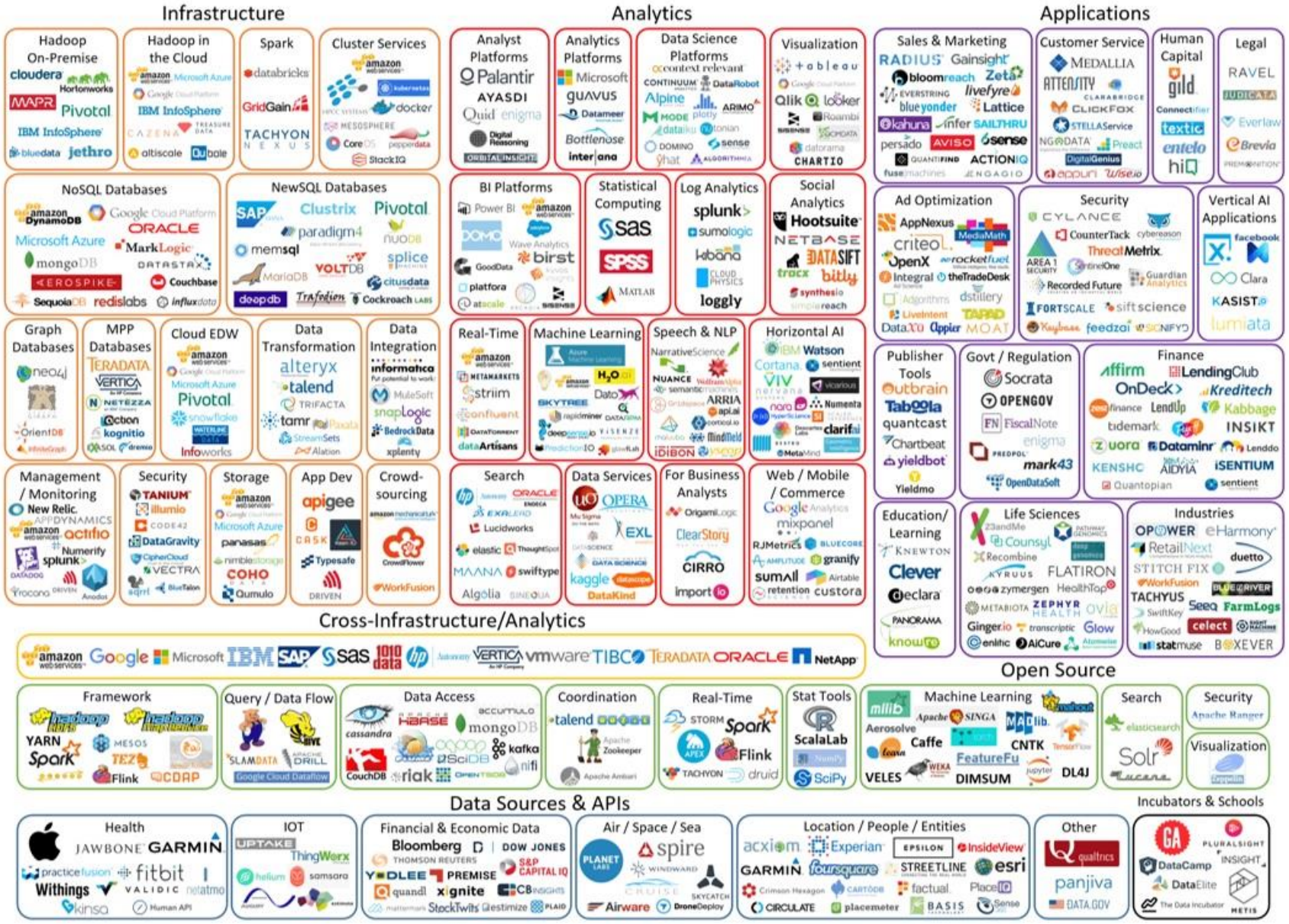
İşlem Gerçekleşir → Analiz edilir → Analiz Sonucu

**Örnekler;**

- Radar sistemleri
- E-Ticaret sipariş işlemleri
- ATM Sistemleri
- Anlık Dashboard'lar

Avantajlarını ve Dezavantajlarını tartışalım

# Big Data Tech Stack



## Büyük Veri Kullanım Senaryoları

### Kamu Kurumlarında

Türkiye Cumhuriyeti Cumhurbaşkanlığı himayesinde, içerisinde büyük veri ekibi bulunacak olan Dijital Dönüşüm Ofisi kurulması kararlaştırıldı. Türkiye Cumhuriyeti sahip olduğu büyük veriler sayesinde, verileri hızlı ve düzenli bir şekilde analiz edilecektir

- Terör ve bilişim suçları tespiti
- Trafik sorunları
- Doğal afet durumlarını yönetme
- Açık veri portalı
- Hastalık ön teşhis
- Kurum giderlerini tahmin etme

### E-Ticarette

- Müşteri profillerini belirleme
- Anlık olarak popüler ürünleri tespit etme
- Kullanıcı bazlı alışveriş zamanlarını tespit etme
- Kişileştirilmiş İndirimler
- Sosyal medya paylaşımlarını takip
- Ürün fiyatlarının doğru tahminlenmesi
- Site içi search kullanımının incelenmesi
- Harcama Tahminleri

## Büyük Veri Kullanım Senaryoları

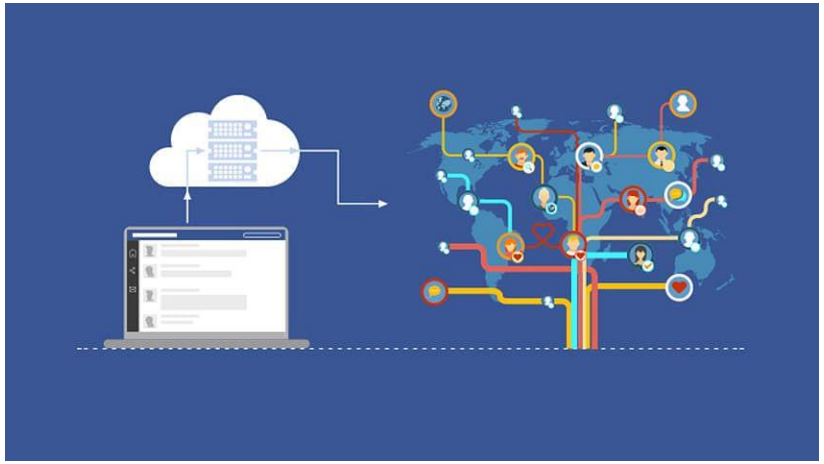
- Netflix başarısını veri toplama, veri analizi ve tahmin algoritmalarına borçlu olduğu üzerine makalesi bulunmaktadır


- UPS kargo 1 yılda taşıdığı ortalama 4 milyar kargo dağıtımı ve yönetimi için büyük veri sistemlerini kullanmaktadır

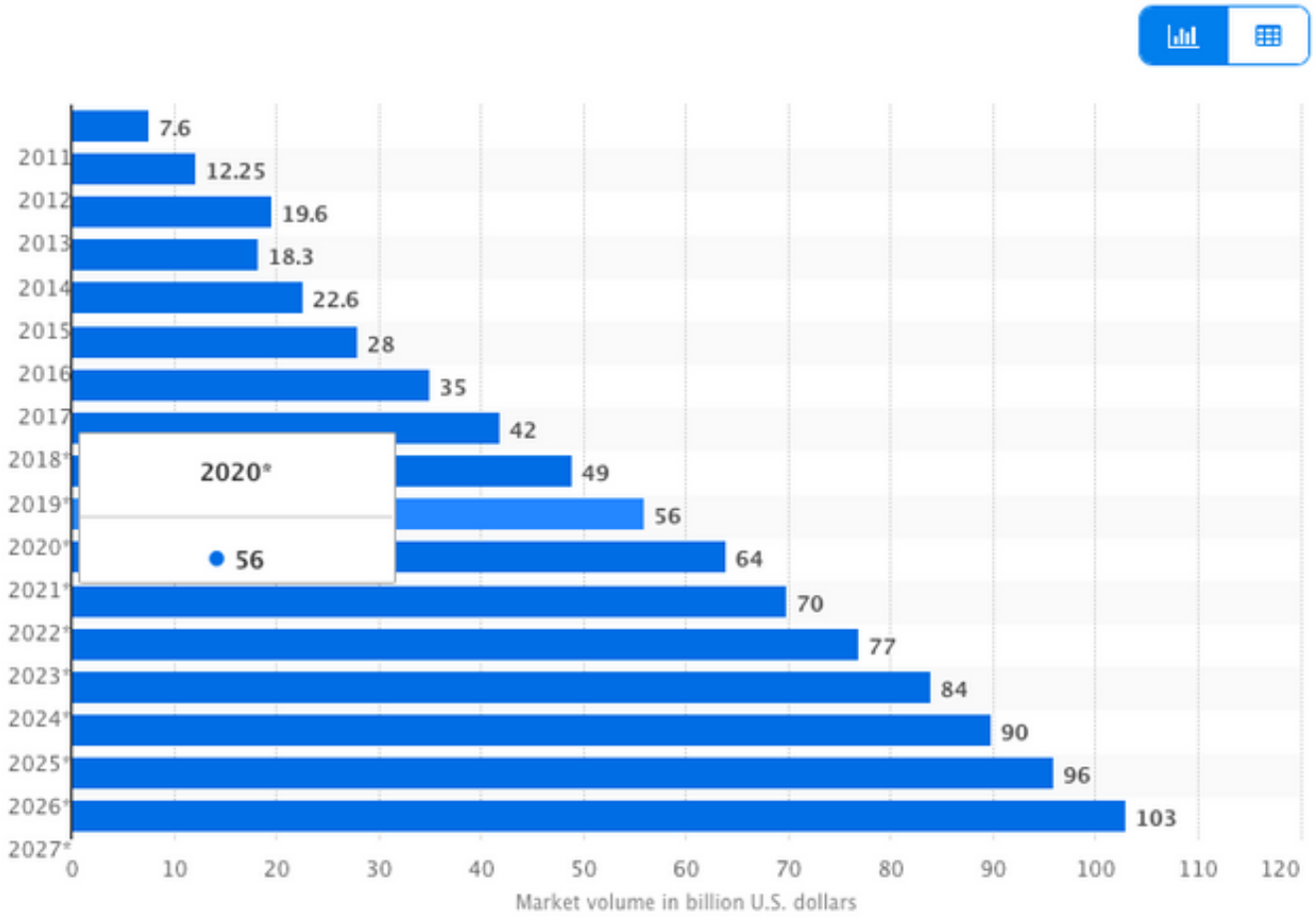
## Büyük Veri Kullanım Senaryoları

- Dünyanın en büyük tedarik zinciri Walmart, stok takibinden ürün tedariğine, müşteri davranış analizinden ürün fiyat tahminlemeye kadar bir çok alanda büyük veri sistemleri kullanmaktadır. Cadılar bayramında satılmayan kurabiyelerin tespiti real time veri analizine örnektir


- Dünyanın en büyük sosyal medya platformlarından birisi olan Facebook, kullanıcıların verilerini topladıktan sonra reklam çalışmalarında kullanılması için küçük ve orta ölçekli müşteriler ile çalışmaktadır



**Forecast of Big Data market size, based on revenue, from 2011 to 2027 (in billion U.S. dollars)**

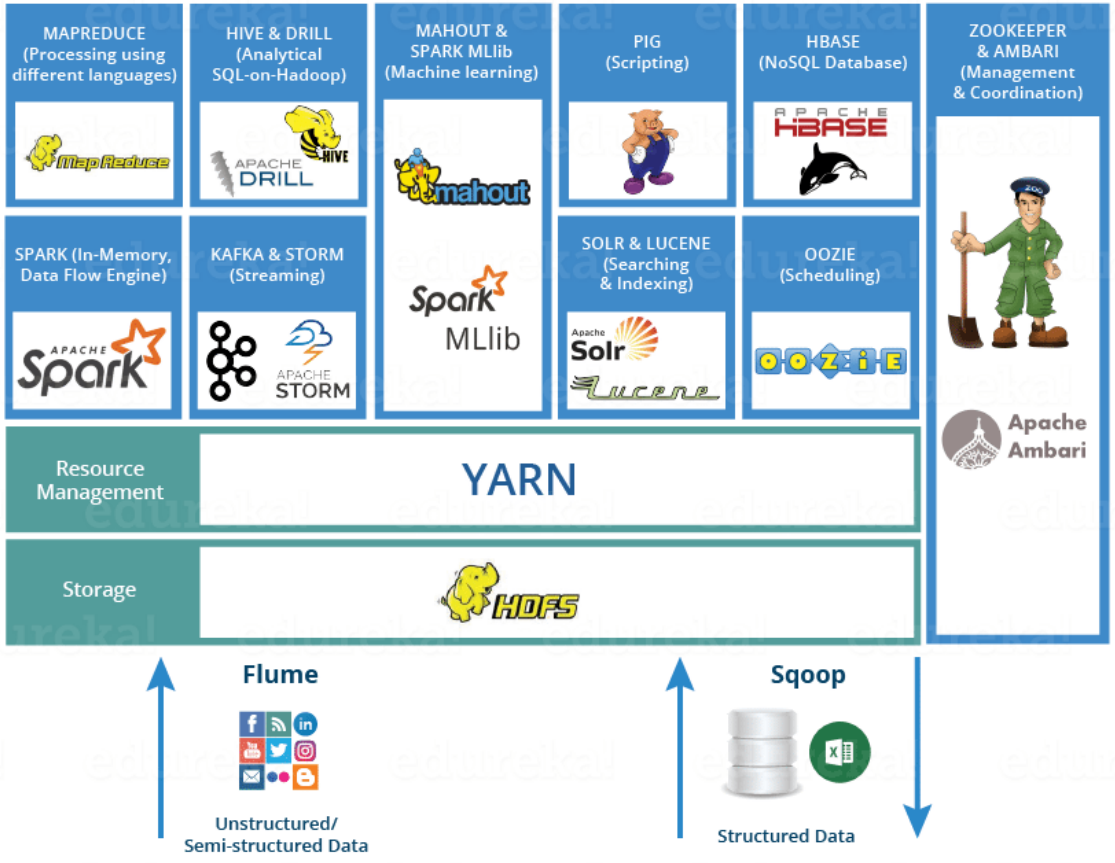


Break

# Apache Hadoop

- Apache Hadoop®, basit programlama modellerini kullanarak büyük veri kümelerinin yüksek düzeyde güvenilir, ölçeklenebilir ve dağıtık biçimde işlenmesini sağlayan bir açık kaynak platformudur.

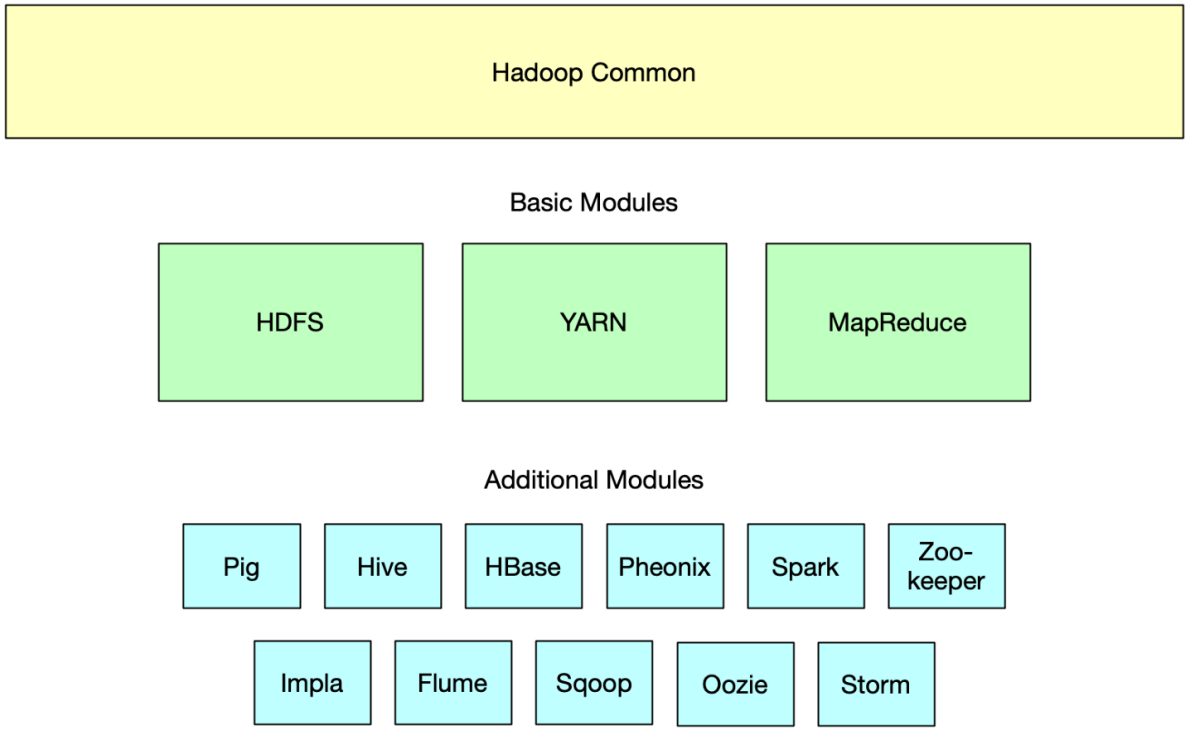
- Her türlü veri için devasa depolama, çok yüksek işlem gücü ve neredeyse sınırsız sayıda eşzamanlı görevleri yönetme yeteneği sağlar. Dağınık bir bilgi işlem ortamında büyük verileri verimli bir şekilde yönetmenizi ve işlemenizi mümkün kılar.



# Apache Hadoop

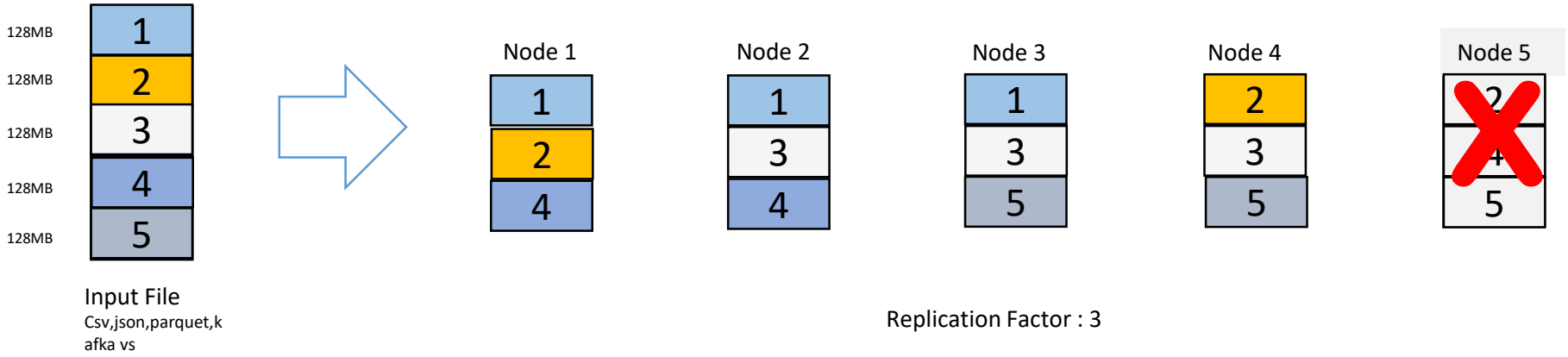
Hadoop dört ana modülden oluşur

- HDFS (Hadoop Distributed File System)
- MapReduce
- YARN (Yet Another Resource Negotiator)
- Hadoop Common



# HDFS

- Hadoop'un dosya sistemidir. Sıradan sunuculardan oluşan kümeler üzerinde büyük verileri işlemek amaçlı kullanılan, dağıtılmış bir dosya sistemidir.
- Geleneksel dosya sistemlerine kıyasla daha iyi veri çıkışı sağlar. Sıradan sunucu disklerini bir araya getirir ve büyük sanal bir disk oluşturur. Bu da çok büyük boyutlardaki dosyaların saklanmasını ve işlenmesini mümkün kılar.



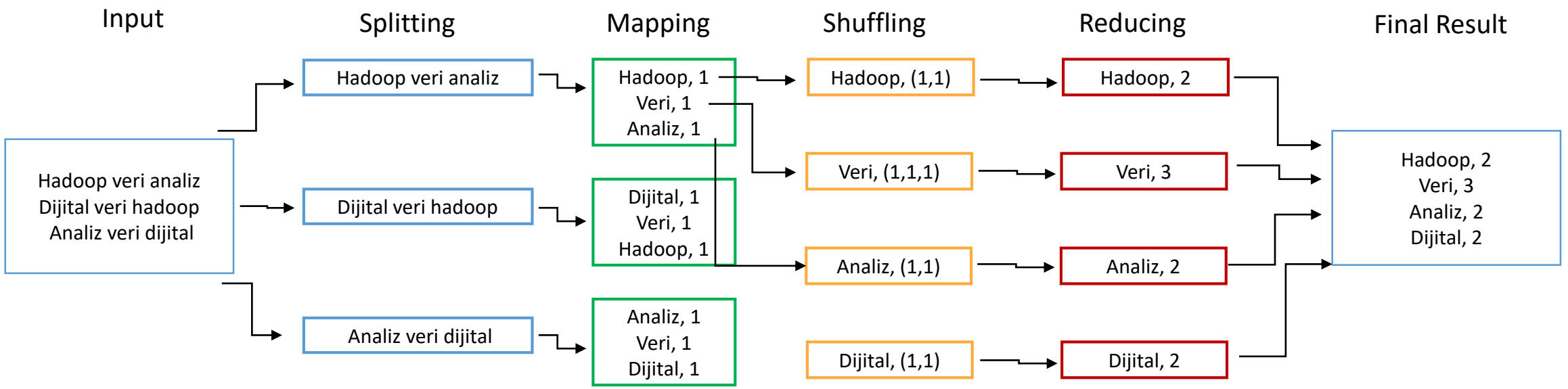

## HDFS

- Java ile yazılmış bir dosya sistemidir
- Bilinen dosya sistemleri ile çalışır (ext3, ext4, xfs)
- Büyük hacimde veriler için tasarlanmıştır
- Dosyalar read-only yöntemle tutulur (performans için)
- Büyük hacimli verilerde performanslıdır (100MB ve üzeri)

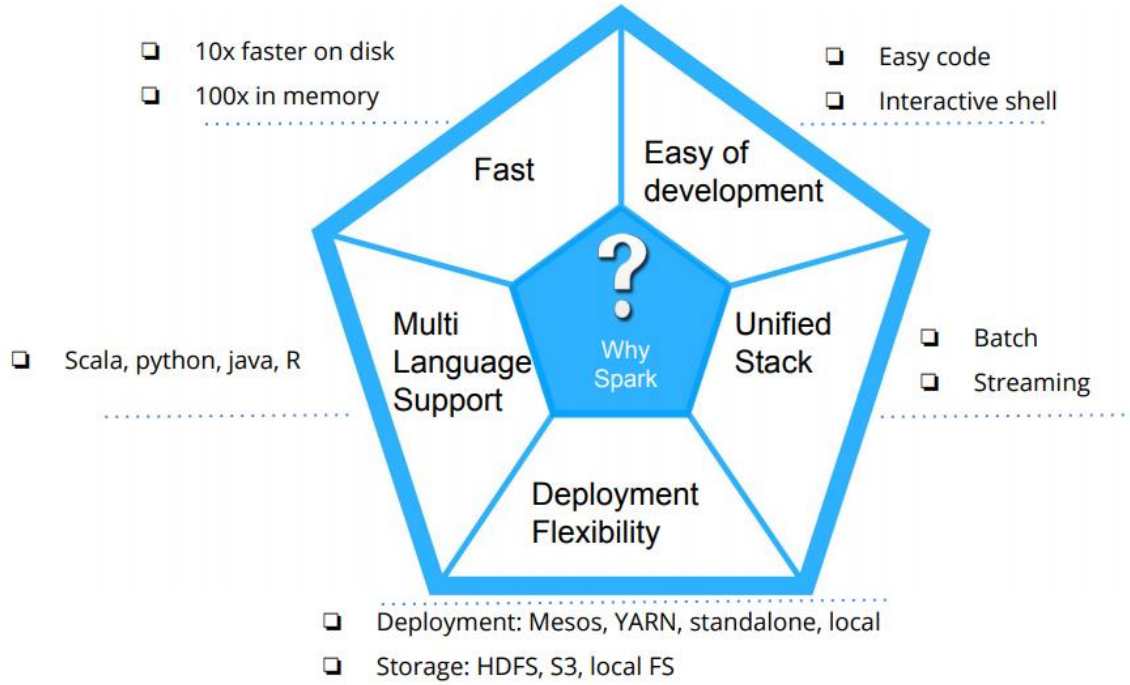
# MapReduce

- Programların eş zamanlı veri işlemesine yardımcı olur. İş parçacıkları küme üzerinde dağılarak aynı anda işleme yaparlar.

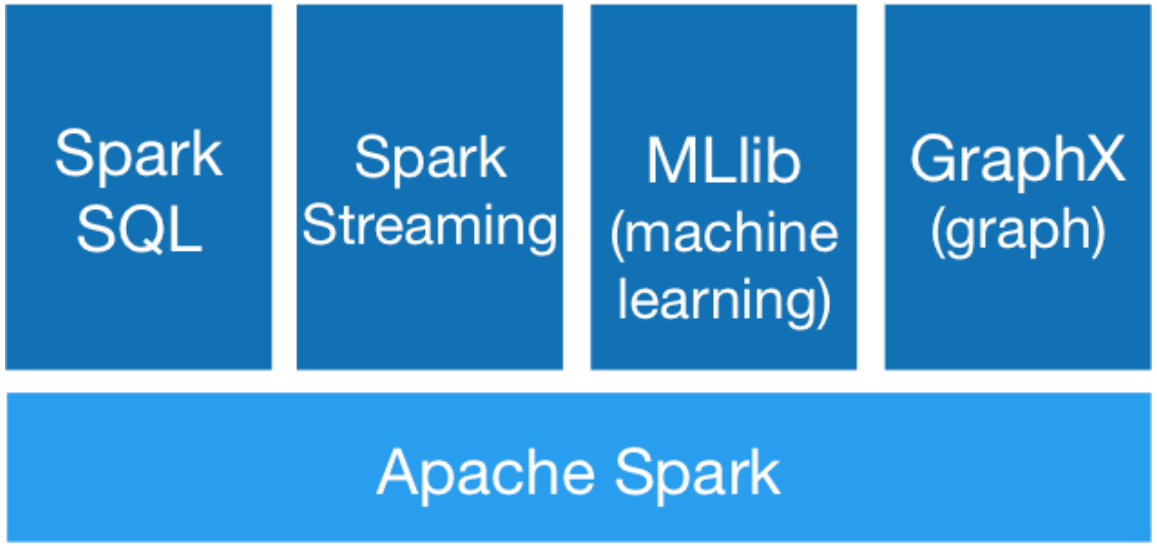
## MapReduce Kelime Sayma İşlemi



- Apache açık kaynak projesi ilk olarak AMPLab'da (California Berkeley Üniversitesi) geliştirilmiştir.
- Büyük ölçekli veri analizi uygulamalarını çalıştırmak için kullanılan açık kaynaklı bir paralel işleme motorudur
- RAM üzerinde çalışır. Bundan dolayı 100x daha hızlı olduğunu savunmaktadır



- Spark SQL ile SQL veya Hive Query Language aracılığıyla veri sorgulama desteklenir. RDBMS veritabanları için Spark SQL performans arttırıcı bir çözüm sunar
- Spark Streaming canlı veri akışlarının ölçeklenebilir, yüksek verimli, hataya dayanıklı akış işlemesini sağlayan Spark Core'un bir modülüdür
- Spark MLLib, Spark'ın makine öğrenmesi (ML) kütüphanesidir. Amacı, pratik makine öğrenimini ölçeklenebilir ve kolay hale getirmektir.



>

**DataFrame API**

```
flights.select("Origin", "Dest", "DepDelay”)
       .filter($"DepDelay" > 15).show(5)
```

**SQL API**

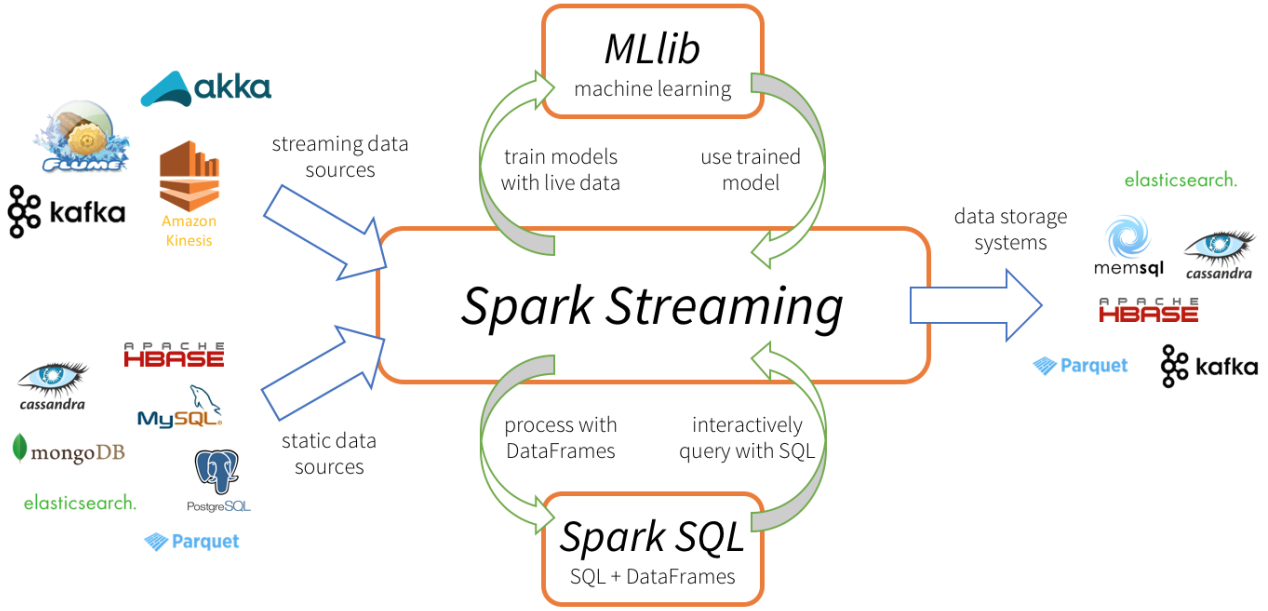
```
SELECT Origin, Dest, DepDelay
FROM flightsView
WHERE DepDelay > 15 LIMIT 5
```

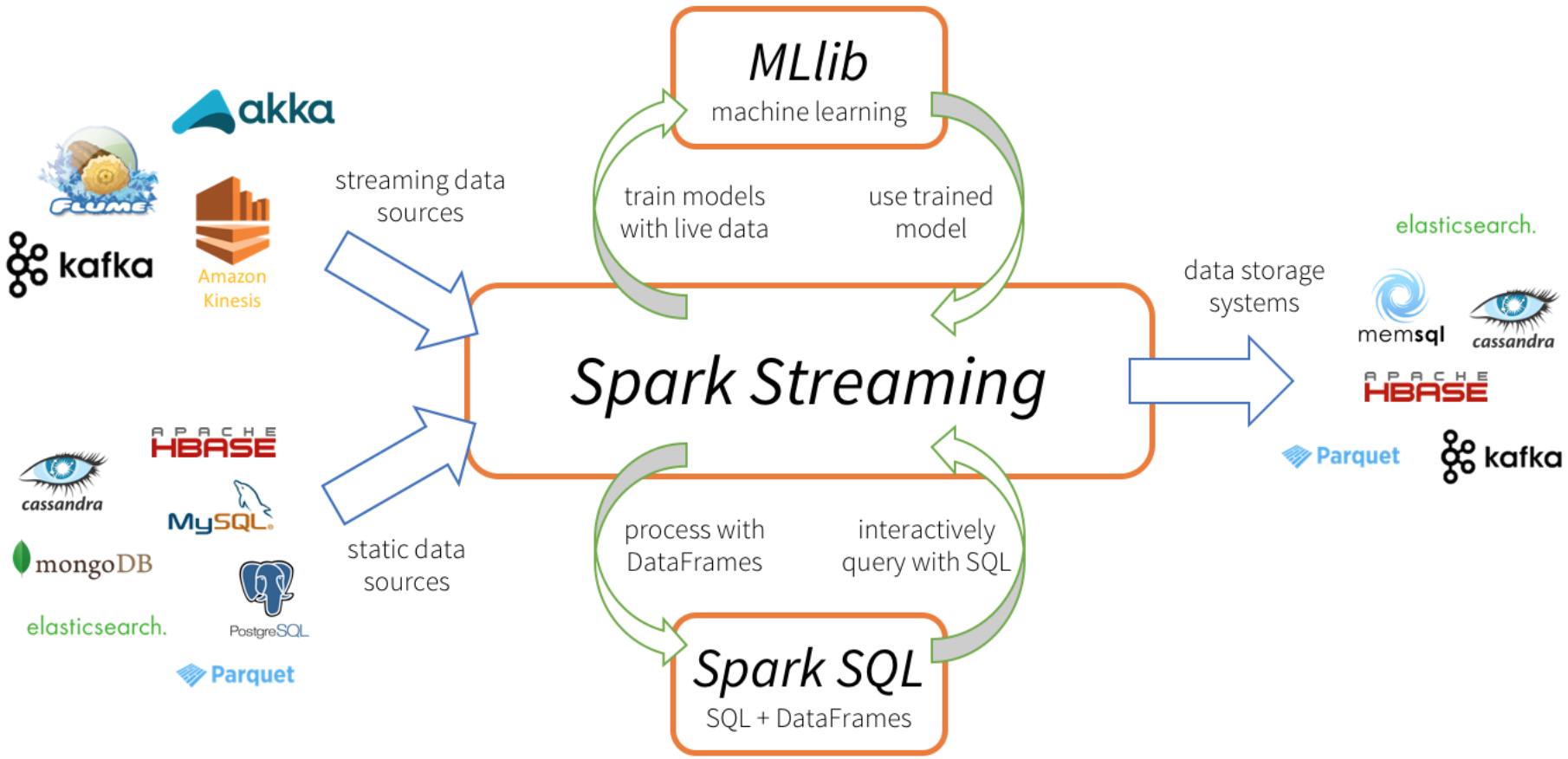
**Results**

```
+------+----+--------+
|Origin|Dest|DepDelay|
+------+----+--------+
|   IAD| TPA|      19|
|   IND| BWI|      34|
|   IND| JAX|      25|
|   IND| LAS|      67|
|   IND| MCO|      94|
+------+----+--------+
```

Bir Spark projesini 3 aşamaya bölebiliriz

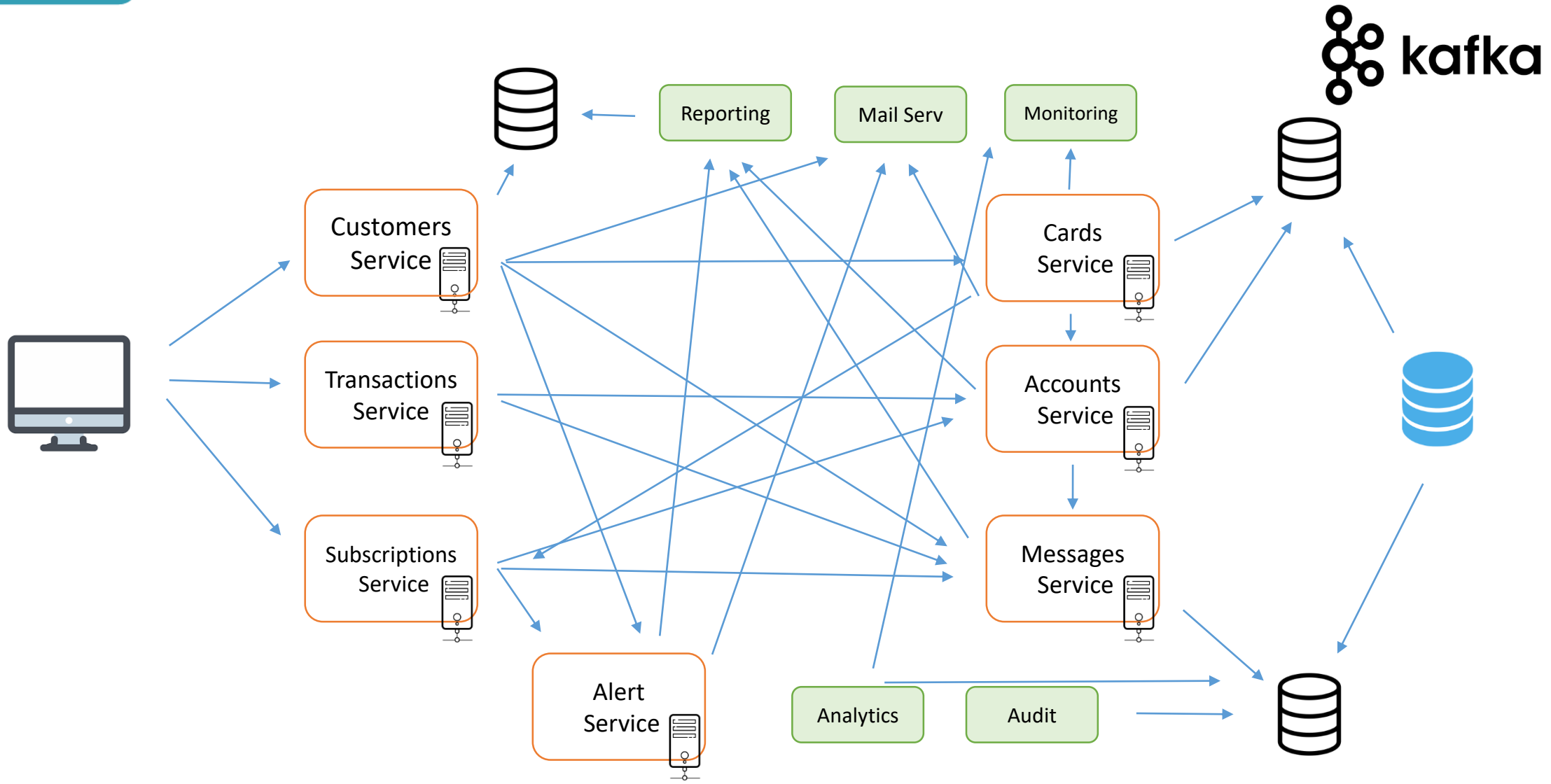
- Okuma: Apache Spark bir çok veri kaynağından static ve streaming olarak okuyabilir
- Analiz: Oluşturulan analiz senaryosuna göre Spark SQL, Spark Streaming ya da Spark MLLib modülleri tercih edilerek istenen analiz yapılır
- Yazma: Analiz sonuçları bir çok veritabanına istenen formatta yazılabilir



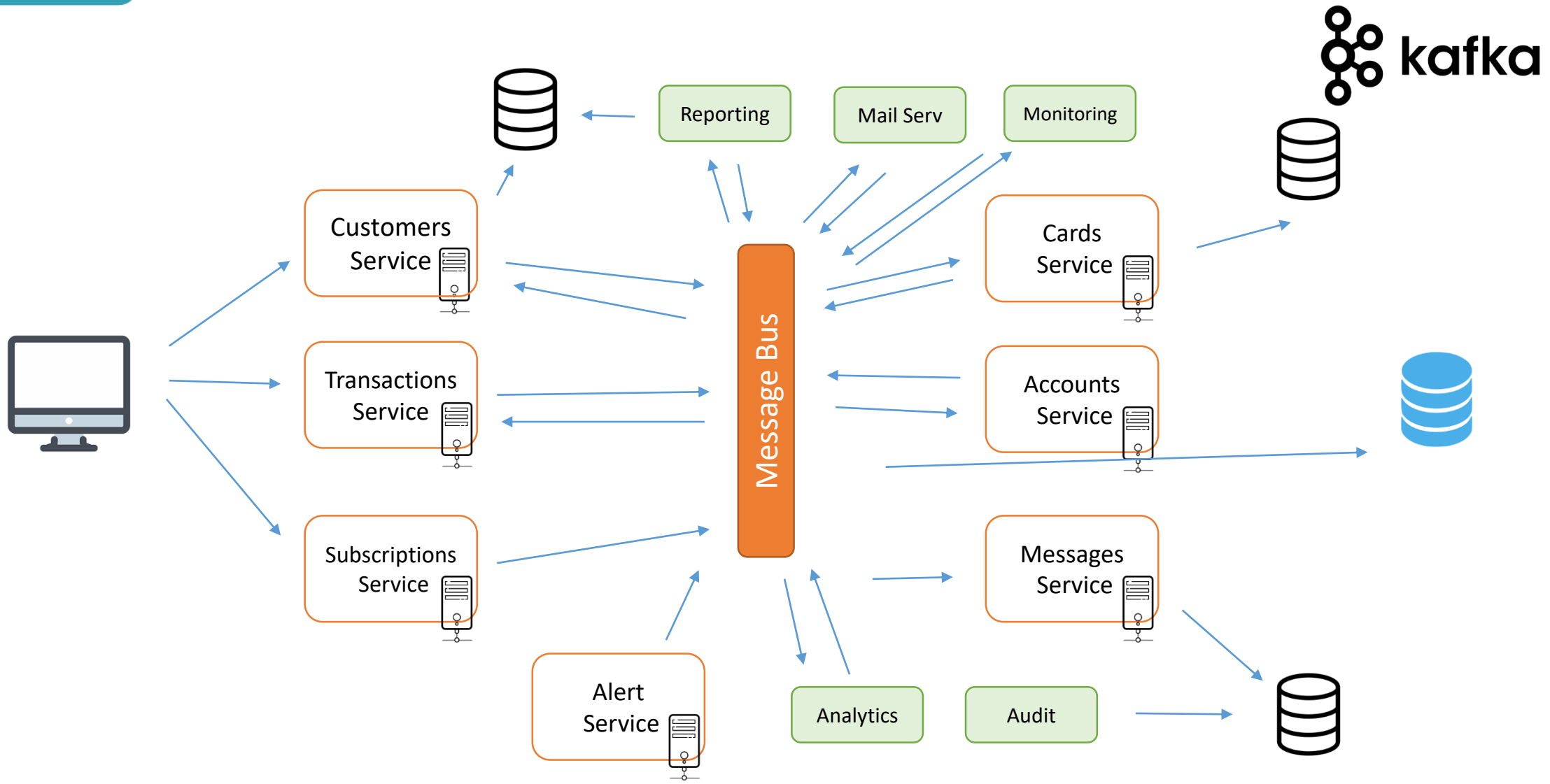

MLlib
machine learning
train models with live data
use trained model
streaming data sources
static data sources
Spark Streaming
process with DataFrames
interactively query with SQL
Spark SQL
SQL + DataFrames
data storage systems
akka
Amazon Kinesis
kafka
cassandra
HBASE
MySQL
mongoDB
elasticsearch.
PostgreSQL
Parquet
memsql
cassandra
HBASE
Parquet
kafka
elasticsearch.

kafka

- Apache Kafka, LinkedIn tarafından geliştirilmiş, şu an Apache yönetiminde açık kaynak olarak çoğunlukla Confluent şirketi tarafından bakımı ve geliştirimi yapılan bir projedir
- Dağıtık (distrubuted) bir veri akış (streaming) mesajlaşma platformudur
- Hataya dayanıklı, yatay olarak ölçeklenebilen, esnek bir mimariye sahiptir
- Mesajlaşma sistemi (messaging system) olabilir, etkinlik takibi(activity tracking) için, uygulama loglarını toplamak için, sağladığı API ile stream processing amacıyla kullanılabilir
- Resilient mimarisi ve retention sistemi bulunmaktadır
- Yüksek seviyede ölçeklenebilirdir (Yüzlerse broker-server oluşturabilir, saniyede milyonlarca event işlenebilir)

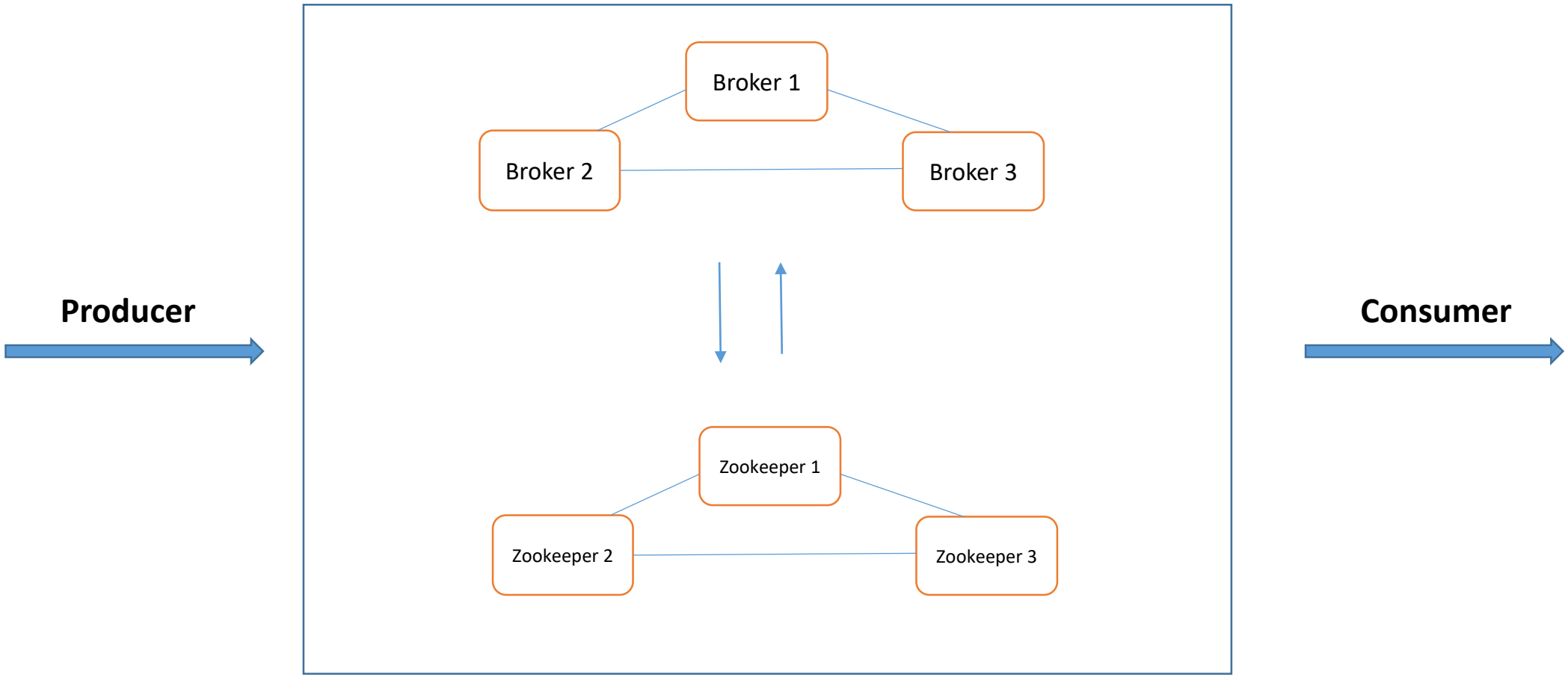



Apache Kafka'nın Tasarım Sebebi

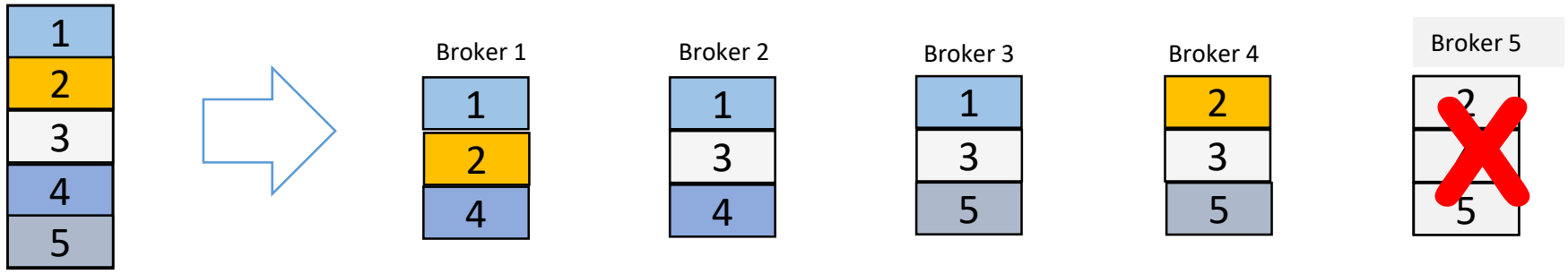



Apache Kafka'nın Çözümü

## Apache Kavramları

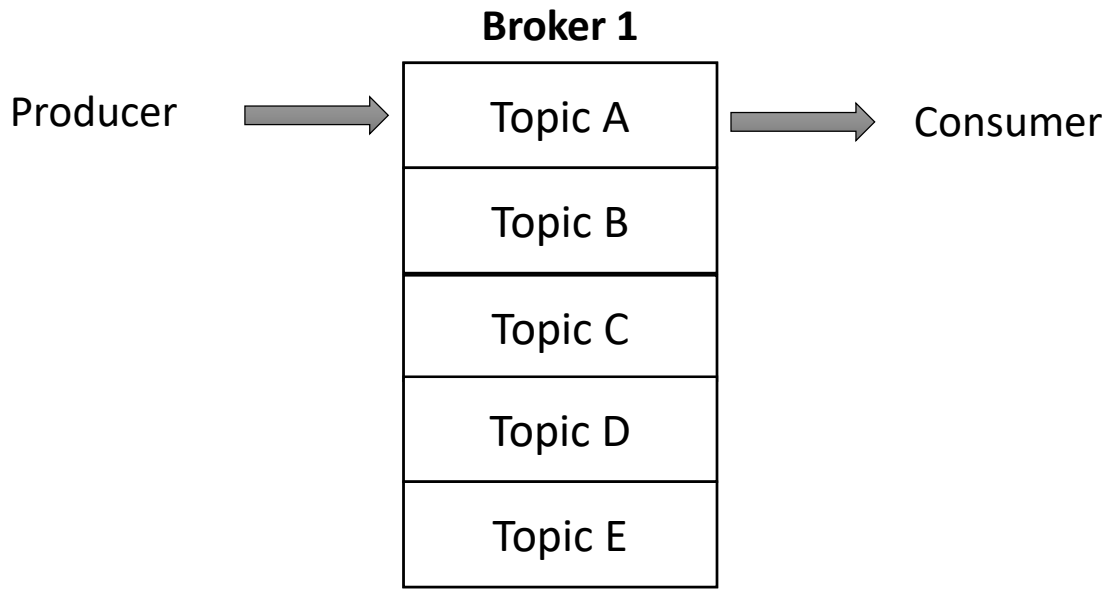


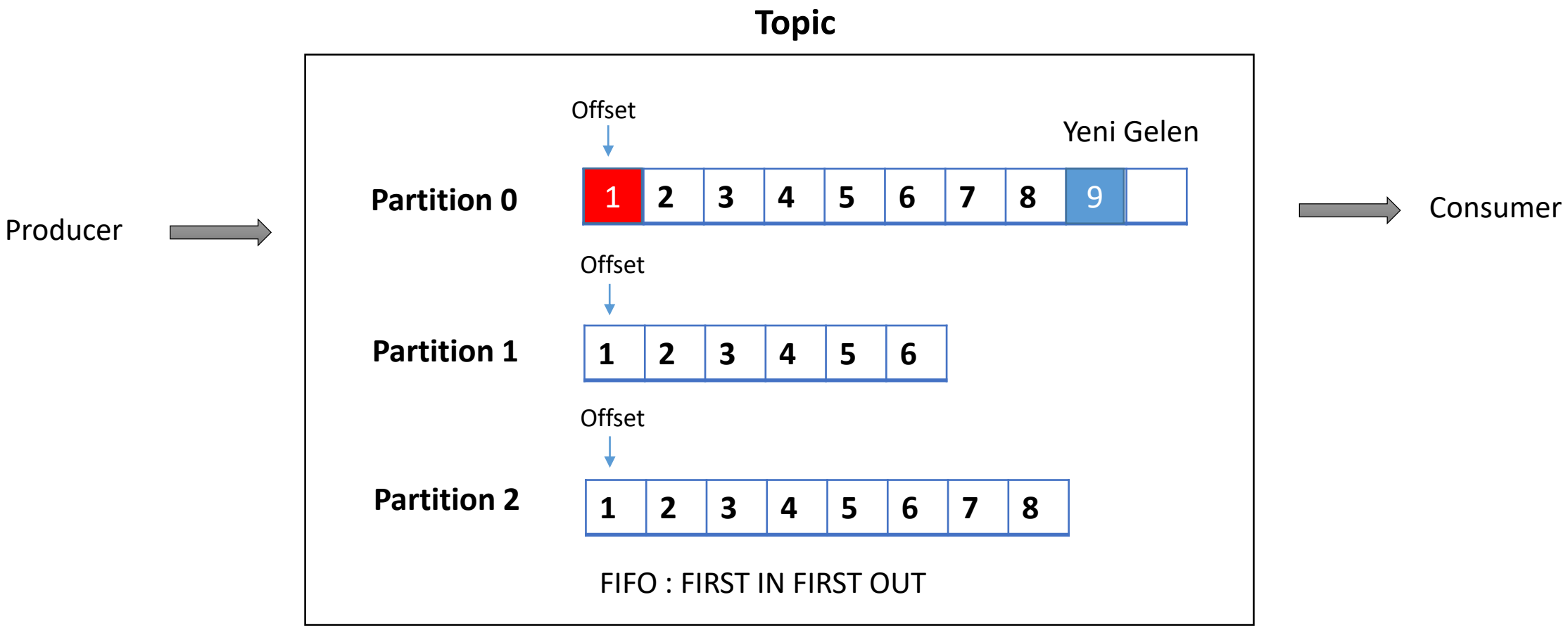
# Apache Kafka Replication




Replication Factor : 3

## Apache Kavramları




Topic
Producer
Offset
Yeni Gelen
Partition 0
1 2 3 4 5 6 7 8 9
Consumer
Offset
Partition 1
1 2 3 4 5 6
Offset
Partition 2
1 2 3 4 5 6 7 8
FIFO : FIRST IN FIRST OUT