
# TÜBİTAK BİLİŞİM VE BİLGİ GÜVENLİĞİ
# İLERİ TEKNOLOJİLER ARAŞTIRMA MERKEZİ

www.bilgem.tubitak.gov.tr

b3lab
BULUT BİLİŞİM VE BÜYÜK VERİ ARAŞTIRMA LABORATUVARI
CLOUD COMPUTING & BIG DATA RESEARCH LABORATORY

# BÜYÜK VERİ TARİHÇESİ VE TEMELLERİ

# Sunum Planı



- Büyük Veri Giriş
- Hadoop Temelleri
- Büyük Veri Ekosistemi
  - **Çekirdek Hadoop:** HDFS, MapReduce ve YARN
  - **Veri Entegrasyonu:** Flume, Sqoop, Kafka ve Nifi
  - **Veri Analizi:** Pig, Hive
  - **Veri İşleme:** Spark
  - **Veri Depolama:** HBase
  - **Veri Arama:** Solr
  - **Koordinasyon:** ZooKeeper
  - **İş Akışı Zamanlayıcısı:** Oozie
  - **Veri Güvenliği:** Ranger
  - **İnteraktif Veri Analitiği**: Zeppelin
  - **Veri Görselleştirme:** Superset

# Büyük Veri Giriş



## Bu bölümde değinilecek konular

- Büyük Veri Nedir?
- Verinin Geleceği
- Büyük Veri: 10V

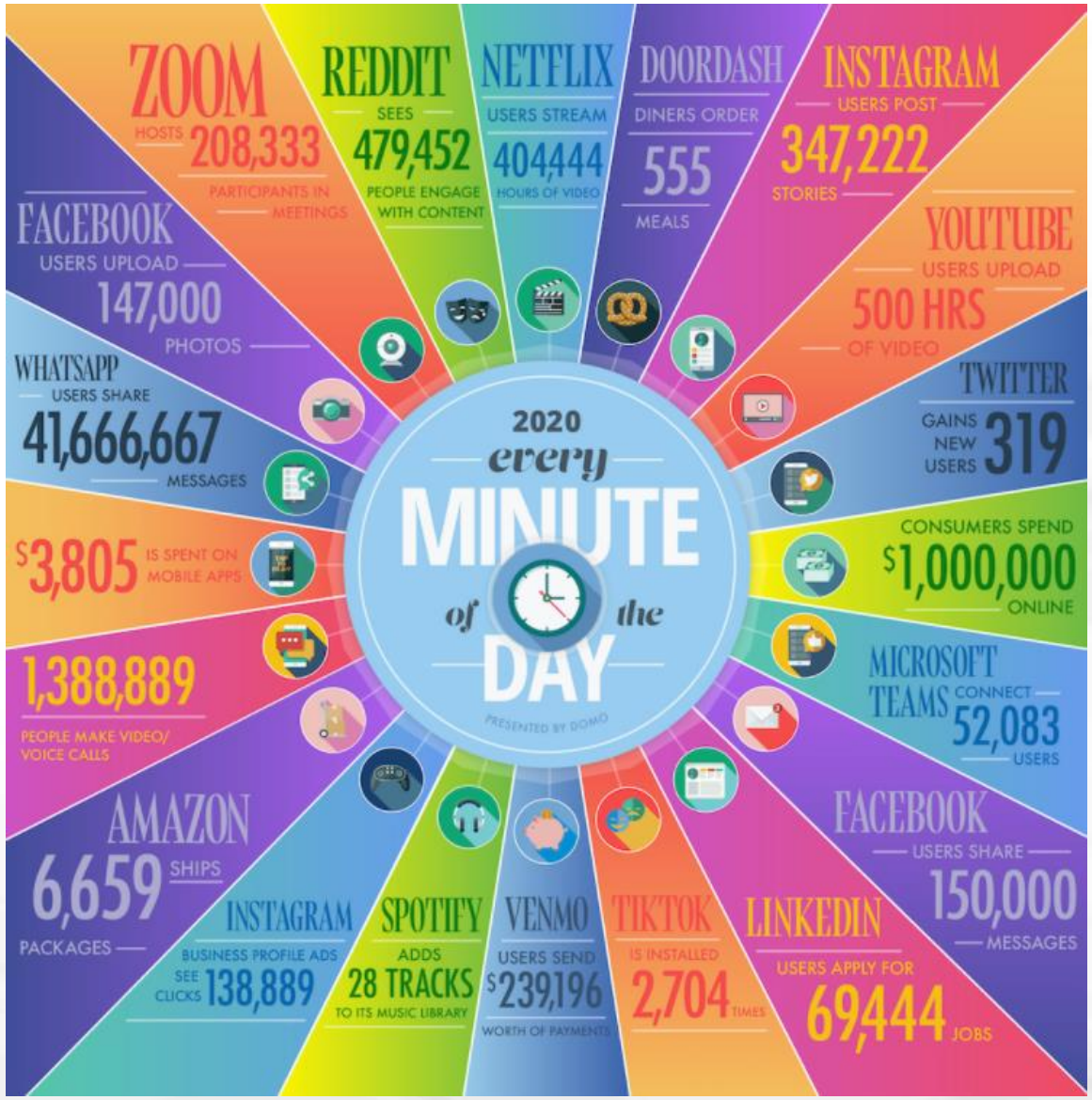
# Büyük Veri Nedir?

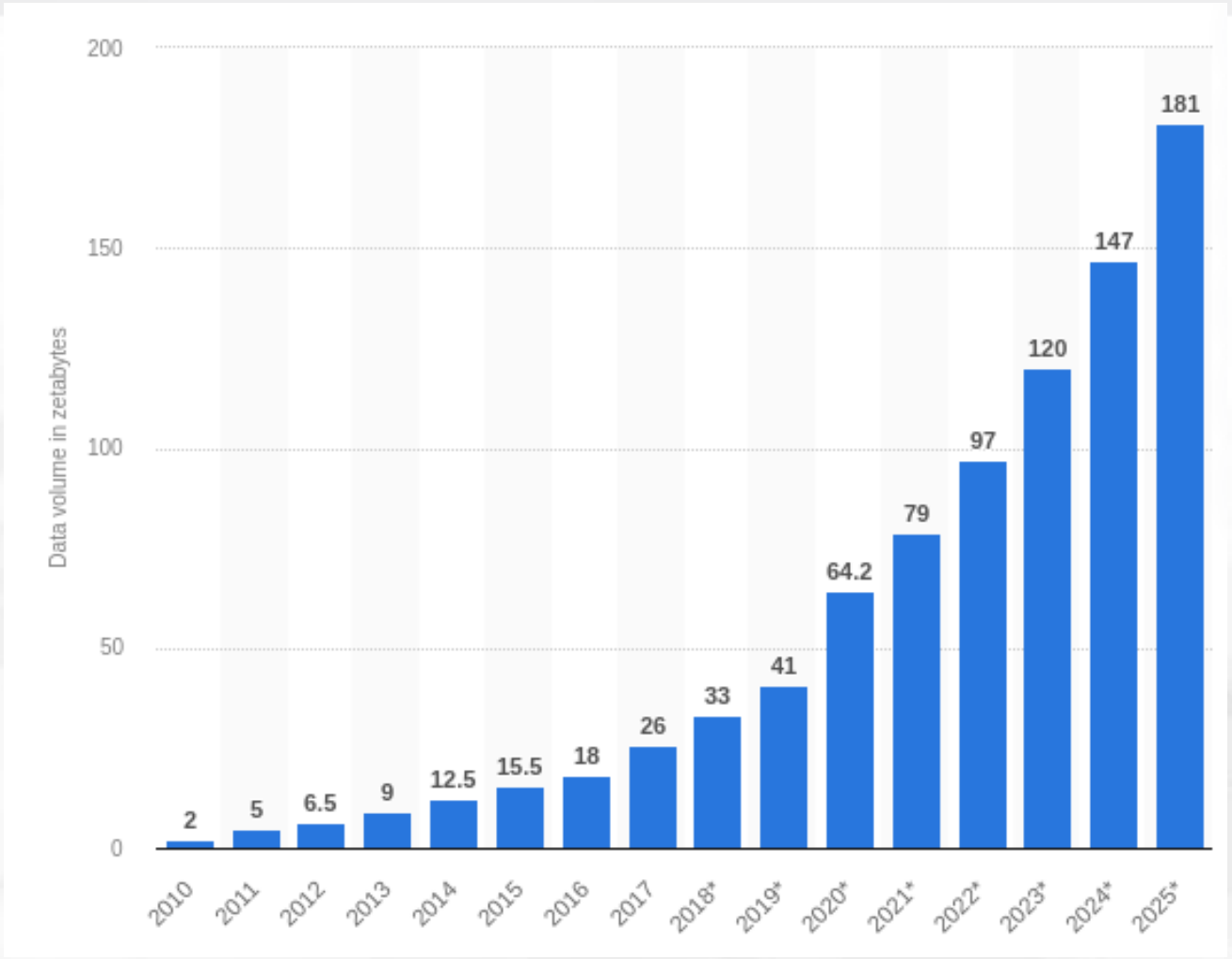


- Geleneksel yazılım araçları ve teknolojileri kullanarak analiz edilmesi ve yönetilmesi zor olan ya da mümkün olmayan muazzam büyüklük ve çeşitlilikteki dijital veriler **büyük veri** olarak adlandırılmaktadır.

# Veri Üretimi



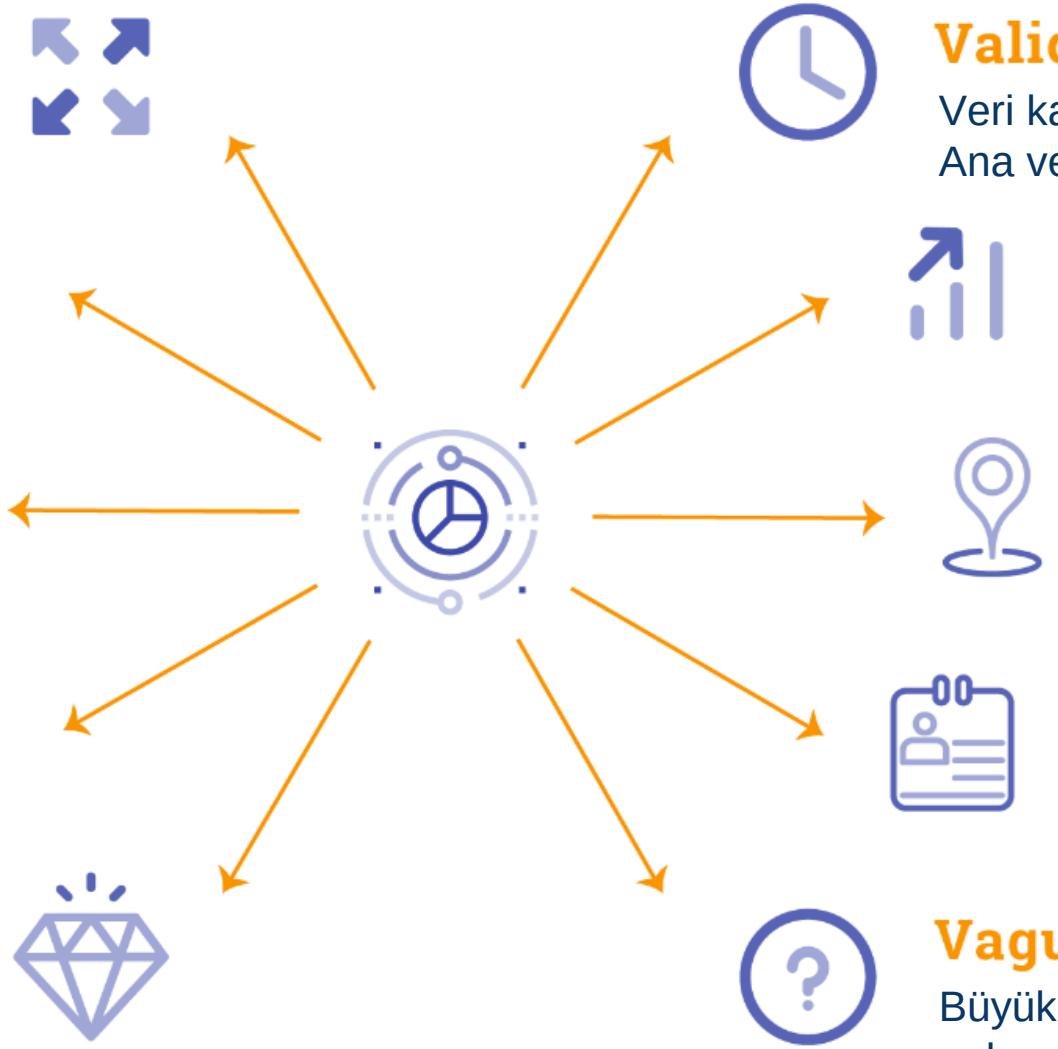



1 zettabyte = 1 trilyon gigabyte

# Büyük Veri: 10V





**Hacim - Volume**
Veri boyutu

**Hız - Velocity**
Verinin üretilme hızı

**Çeşitlilik - Variety**
Farklı veri türleri

**Doğruluk - Veracity**
Veri doğruluğu

**Değer - Value**
Faydalı veri

**Validity - Geçerlilik**
Veri kalitesi, yönetim,
Ana veri yönetimi

**Variability -Değişkenlik**
Veri kaynağında dinamik,
gelişen davranış

**Venue - Yer**
Birden çok platformdan
dağıtılılmış heterojen veri

**Vocabulary - Söz varlığı**
Veri yapısını tanımlayan veri
modelleri ve anlambilim

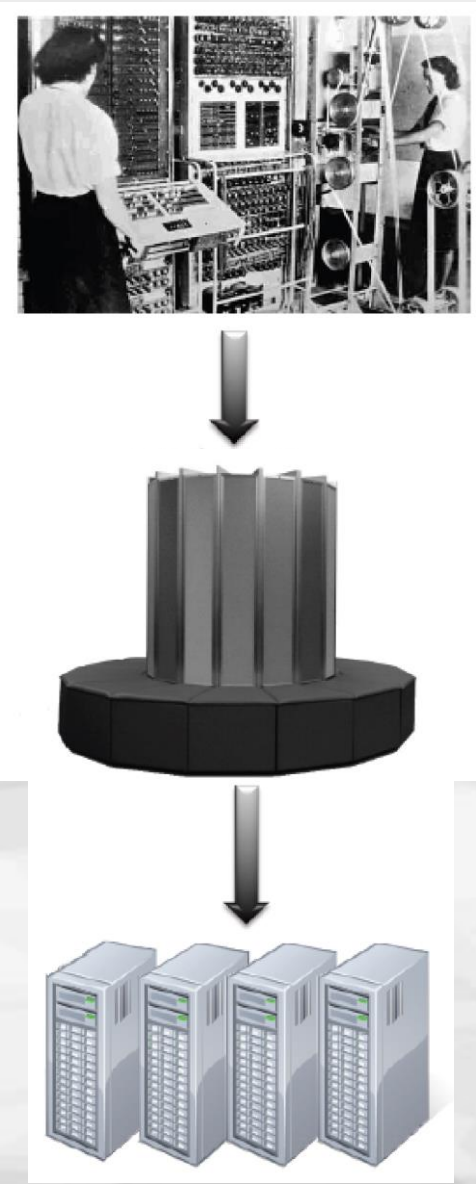
**Vagueness - Belirsizlik**
Büyük veri ve kullanılan araçların
anlamı üzerine kafa karışıklığı

**Kaynak:** www.xenonstack.com

# Hadoop Nasıl Ortaya Çıktı?



- Geleneksel hesaplama
  - Sınırlı işlemci, küçük ölçekte veri, karmaşık işlemler
- **İlk çözüm arayışı:** Büyük bilgisayarlar
  - İşlemci hızlarının artırılması
  - Bellek artırılması
  - Fakat ihtiyacıları karşılamada yetersizlik
- **Daha iyi bir çözüm arayışı:** Daha çok bilgisayar
  - Dağıtık sistemler
  - Bir iş için birden çok makine kullanma
  - Paralleştirme



# Hadoop Nasıl Ortaya Çıktı?

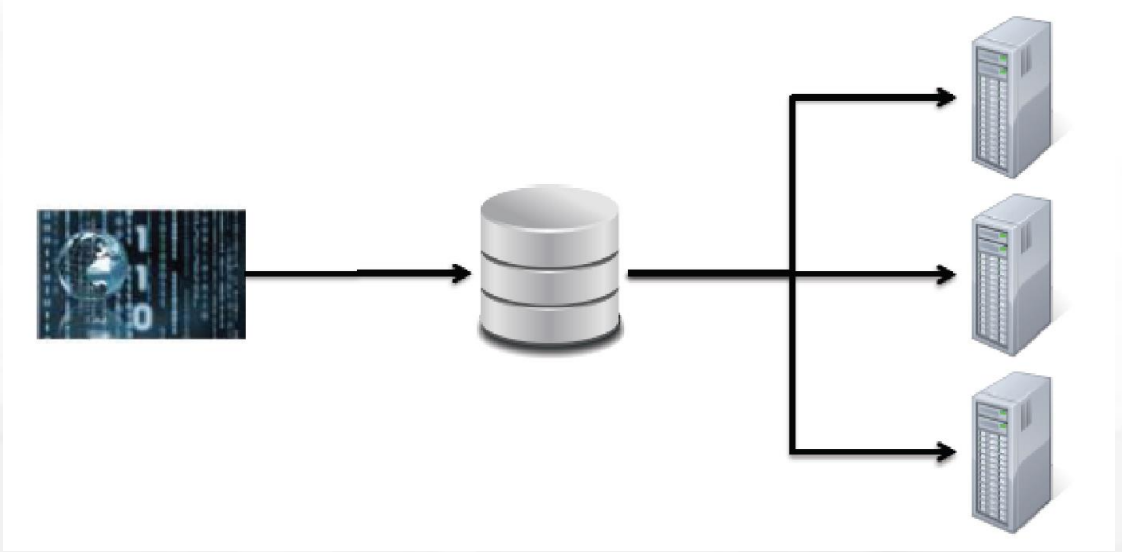


**Dağıtık sistemlerin zorlukları**

- Programlama karmaşıklığı
  - Veri ve işlemleri senkronize tutmak
- Sınırlı bant genişliği
- Parçasal hatalar (partial failures)

**Dağıtık sistemlerin dar boğazı**

- Veri, merkezi bir yerde depolanır.
- Veri, çalışma zamanında işlemcilere kopyalanır.
- Ancak, sınırlı veride sorun çıkmıyor.
- Modern sistemler artık daha fazla veriye sahip:
  - terabytes ve üzeri / günlük
- Yeni bir yaklaşıma ihtiyaç var



# Hadoop Kısa Tarihçesi



- Hadoop'un temelleri, Google'ın 1990 sonları 2000'lerin başındaki çalışmaları ile atılıyor.
- Google'ın sorunu:
  - Çok büyük miktardaki web içeriğinin indekslenmesi ve depolanması
- Google'ın çözümü:
  - 2003 yılında yayınlanan **GFS, Google File System** makalesi
  - 2004 yılında yayınlanan **MapReduce** makalesi
- Doug Cutting, Google'a ait yayınlardan hareketle Nutch projesini yeniden yazarak Hadoop'a giden yolu açmış oluyor.

# Hadoop Nedir?



## Hadoop

- Büyük Veri'yi depolayabilen, işleyebilen ve analiz edebilen bir platform
- Özellikleri :
  - Dağıtık
  - Ölçeklenebilir
  - Hata toleranslı
  - Açık kaynaklı

# Hadoop Nedir?



- Hadoop'un temel konseptleri:
  - **Dağıtık** biçimde veriyi depolaması ve **esnek** olabilmesi
  - Hesaplamayı depolamaya taşıyabilmesi (data locality)
  - Arka plandaki dağıtık sisteme ait detayları kendisinin halledebilmesi
  - Sistemdeki **arızalara karşı toleranslı** olabilmesi
- Hadoop etrafında araçların geliştirilmesiyle pek çok farklı konu için çözüm üretilebilmektedir. (**Büyük Veri Ekosistemi**)

| | |
|---|---|
| Extract Transform Load (ETL) | Tahminsel Analitik |
| İstatiksel Analiz | İş Zekası Ortamı |
| Makine Öğrenmesi | Veri Depolama |

# Hadoop Nedir?

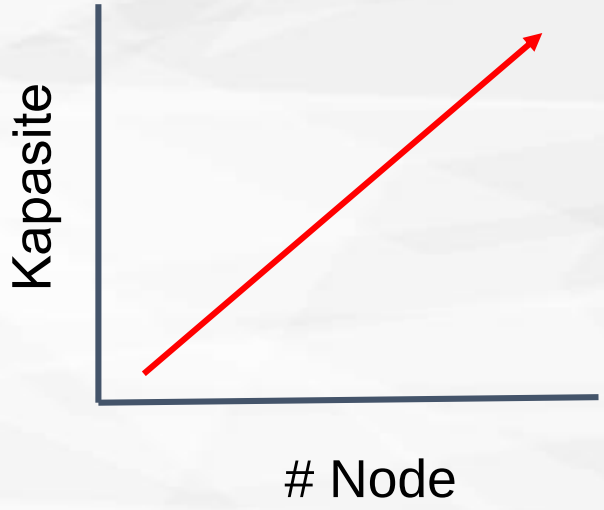


## Çekirdek Hadoop: Dosya Sistemi ve İşlem Platformu

- Hadoop Dağıtık Dosya Sistemi (HDFS)
  - Tüm veri formatlarını depolayabilir.
  - Veri bloklara ayrılıp, ardından pek çok kopya halinde tutuluyor.
- MapReduce
  - Veriyi dağıtık olarak işleyen platform
- YARN (Yet Another Resource Negotiator) (MapReduce v2)
  - Hadoop alt yapısındaki işlem kaynaklarının yönetimini sağlıyor.
  - İşlerin çizelgelerini düzenliyor.
  - İşlemlerin koordinasyonunu sağlıyor.

## Hadoop ölçeklenebilir

- Eklenen her bilgisayar ile kapasitesi orantılı olarak artar.
- Yükteki artışlara göre sistem performansı çok daha az düşer.
    - Sistem arızalarına karşı daha güçlü



# Hadoop Nedir?



## Hadoop sistem arızalarına karşı toleranslıdır

- Dağıtık sistemlerde tekil bilgasayar çökmeleri kaçınılmazdır.
- Peki çökme olduğu durumda ne olur?
    - Sistem çalışmasına devam eder
    - Kaynak yöneticisi, işleri diğer çalışan bilgisayarlara atar.
    - Veri kopyalama ile veri kaybı önlenir.
    - Tekrar ayağa kaldırılan bilgisayarlar otomatik olarak sisteme girer.

# Hadoop Nedir?



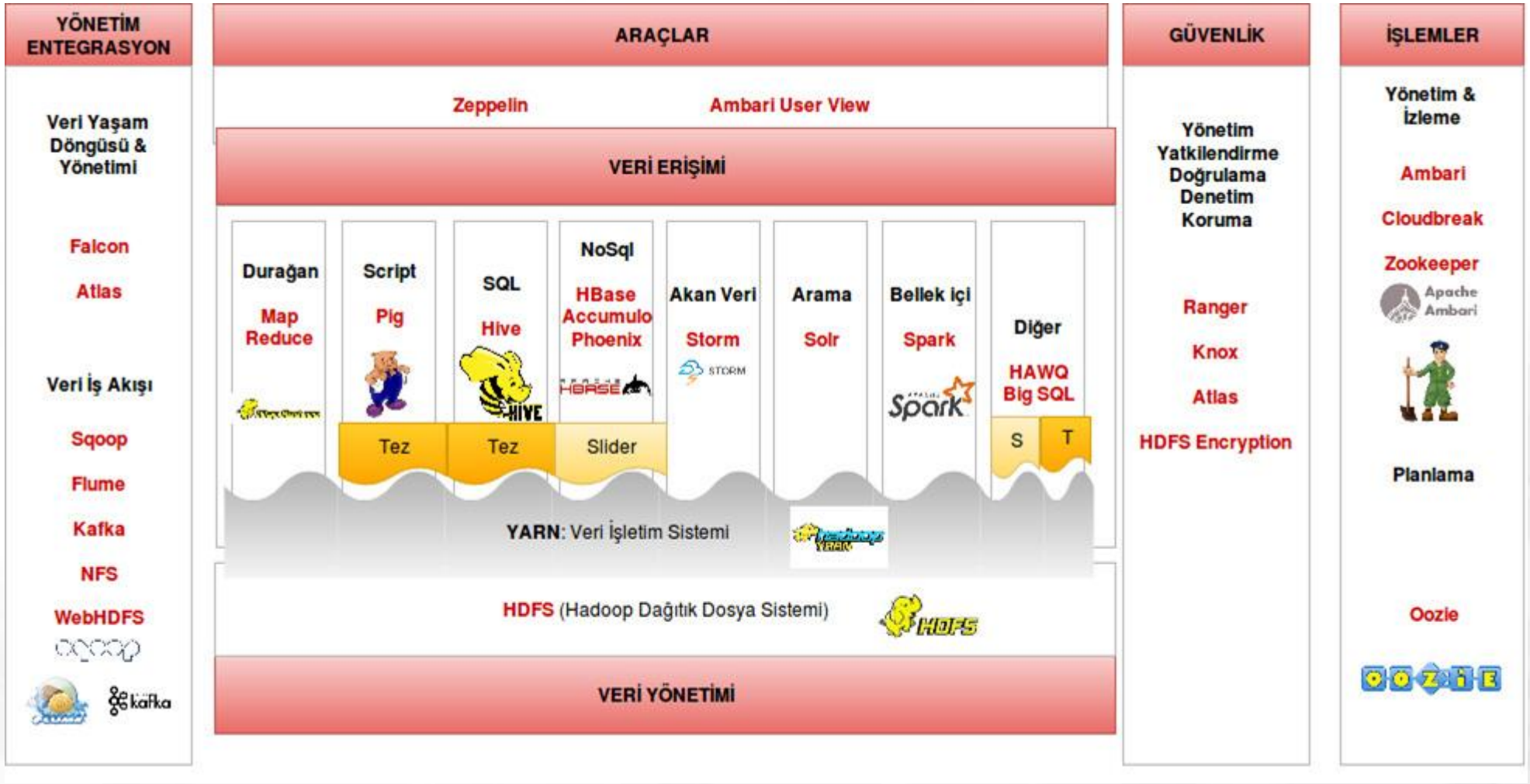
## Sadece Çekirdek Hadoop'tan mı meydana geliyor?

- Hadoop etrafında geliştirilen pek çok proje mevcutttur.
    - **Büyük Veri Ekosistemi**

- Tüm projeler açık kaynaklıdır.
    - **Apache Software Foundation**

# Büyük Veri Ekosistemi





# Büyük Veri Kullanım Alanları



## Finans
- Dolandırıcılık tespiti
- Risk yönetimi
- 360 derece müşteri görüşü

## Kamu Hizmetleri
- Hava analizi/tahminleme
- Akıllı şebeke yönetimi

## Lojistik
- Lojistik optimizasyonu
- Trafik sıkışlığı

## Bilgi İşlem
- Sistem log analizi
- Siber güvenlik
- Kaynak tahmini

## Sağlık
- Erken salgın alarmı
- Hastalık teşhisi
- Genom haritalama

## Perakende
- Kampanya verimliliği
- Hedef pazarlama
- Kaynak tahmini
- Marka algısı
- Müşteri deneyimi

## Telekomünikasyon
- Ağ izleme
- Telefon ve konum bazlı müşteri memnuniyeti

## İmalat
- İmalat verimliliği
- Talep tahmini
- Tedarik zinciri yönetimi

## Endüstri
- Müşteri tutma
- Müşteri kazanımı

# Büyük Veri Çözümlerine Neden İhtiyacımız Var?



- **Daha fazla veri** geliyor
  - Nesnelerin interneti (IoT:Internet of Things)
  - Sensör verileri
  - Akan veri
- Daha fazla veri beraberinde **daha büyük sorular** getiriyor
- Daha fazla veriyle bu sorulara **daha iyi cevap** verilebiliyor.
- Büyük veri çözümleri sayesinde daha **önce atılan veri**ler saklanabiliyor.
- Daha önce **anlamsız gelen veri**lerin içerisinde dolaşmak ve **yeni iş modelleri** geliştirmek mümkün.

## Hadoop Terminolojisi

- **Küme (cluster):** Birbiriyle çalışan bir grup bilgisayar
  - Veri depolama, veri işleme ve kaynak yönetimi

- **Düğüm (node):** Kümedeki her bilgisayar
  - **Master node:** İşin dağıtımını, kaynak yönetimini sağlar
    - Her zaman ayakta kalacak biçimde ayarlanmalı
  - **Worker node:** İşin kendisini yapar
    - Çökmesi kümeyi etkilemez

# Çekirdek Hadoop: HDFS



**Hadoop Dağıtık Dosya Sistemi Temelleri (Hadoop Distributed File System :HDFS)**

- HDFS, Hadoop'un **depolama** katmanıdır.
- Hadoop için optimize edilmiştir ve her türlü dosyayı depolayabilecek bir dosya sistemidir.
  - Veri bloklara ayrılır, bloklar tüm cluster üzerinde kopyalanarak dağıtılır.
- **NameNode ve DataNode, master-worker** mimarisinin üzerinde çalışır.
- Sıradan donanımlar üzerinde çalışarak, çok büyük dosyaları depolayabilir.
  - >100 MB, hatta petabytle'lar
- Veri yazma/okuma
  - Bir kere yaz, pek çok kez oku
  - Verinin tamamını okumak için geçen süre, verinin özellikle bir kısmını okumak için geçen süreden daha önemli

# Çekirdek Hadoop: MapReduce



## MapReduce Temelleri

- Hadoop kümesi üzerinde işlerin paralel biçimde dağıtılmasını, yönetilmesini ve işlenmesini sağlayan bir platformdur.
- Paralelleştirmeyle birlikte hesaplamayı verinin tutulduğu yere taşıyarak, bütün veri yerine verinin parçaları üzerinde işlemin yapılmasını sağlar.
  - **Veri işleme: Map ve Reduce Fazları**
- MapReduce bir dil değildir, bir programlama modelidir. Map ve Reduce fonksiyonları her dile uygulanabilir.
  - Java, C++, Python, Perl, Ruby, C vs.
- JobTracker ve TaskTracker, master-worker mimarisinin üzerinde çalışır.
- Sıradan donanımlar üzerinde çalışır.

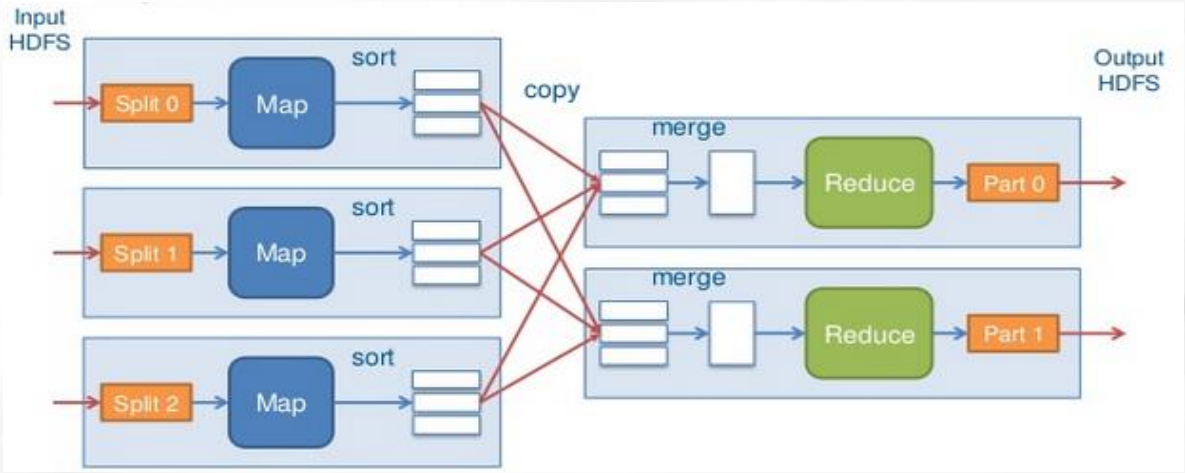
# Çekirdek Hadoop: MapReduce



**MapReduce Mimarisi**

**Map Fazı:** Her bir parça paralel olarak Map fonksiyonuna anahtar-değer şeklinde iletilir.

**Reduce Fazı:** Map fonksiyonundan çıkan değerler gruplanıp sıralandıktan sonra Reduce fonksiyonuna iletilir.



# Çekirdek Hadoop: YARN

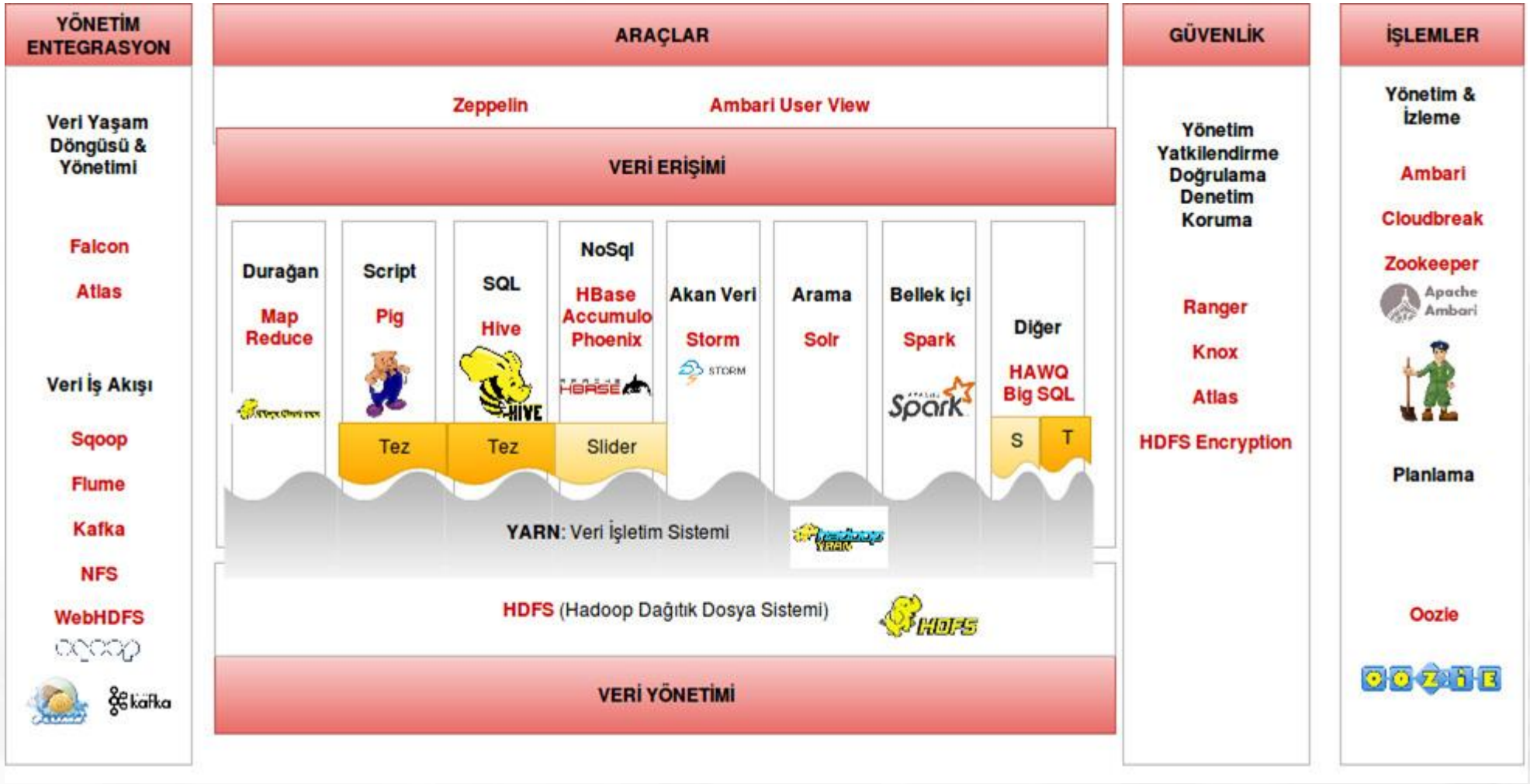


## YARN Temelleri

- Aynı zamanda MapReduce v2 olarak da adlandırılır.
- Hadoop temelinde tüm kümeye ait işlem gücü kaynağını sadece MapReduce işlem platformu ile yönetmektedir.
- Ancak, ilerleyen zamanlarda MapReduce dışında Spark, Storm vs. farklı platformlar ortaya çıkmaya devam etmektedir.
- Kaynakların farklı platformların isteğine göre cevap verebilecek **yönetimi, iş çizelgelerinin oluşturulması, izlenmesi** amacı ile geliştirilmiştir.

# Büyük Veri Ekosistemi



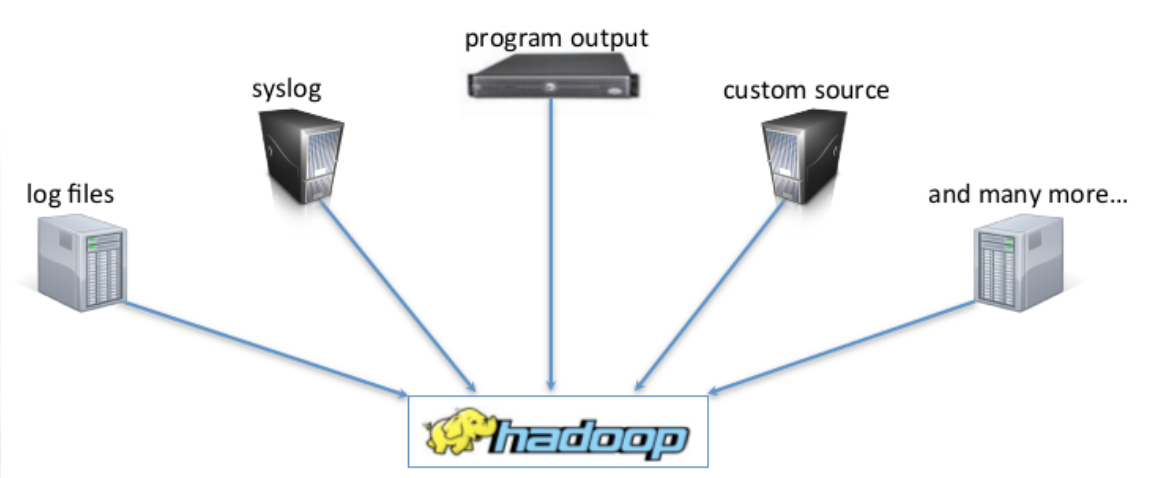


# Veri Entegresyonu: Flume



## Flume Temelleri

- Olay bazlı verilerin, özellikle akan verinin HDFS'e **aktarılmasını** sağlar.
  - log dosyaları, ağ verileri, email mesajları, GPS verisi, sosyal medya verisi, ...
- Verileri, üretildikleri formatta **toplayabilmekte ve depolanmasını** sağlayabilmektedir.
- Verileri, üretildikleri yerde **sıkıştırma, dönüştürme** vs. ön işlemler yapabilir.
- Parallel biçimde çalışarak ölçeklenebilmeyi sağlar.

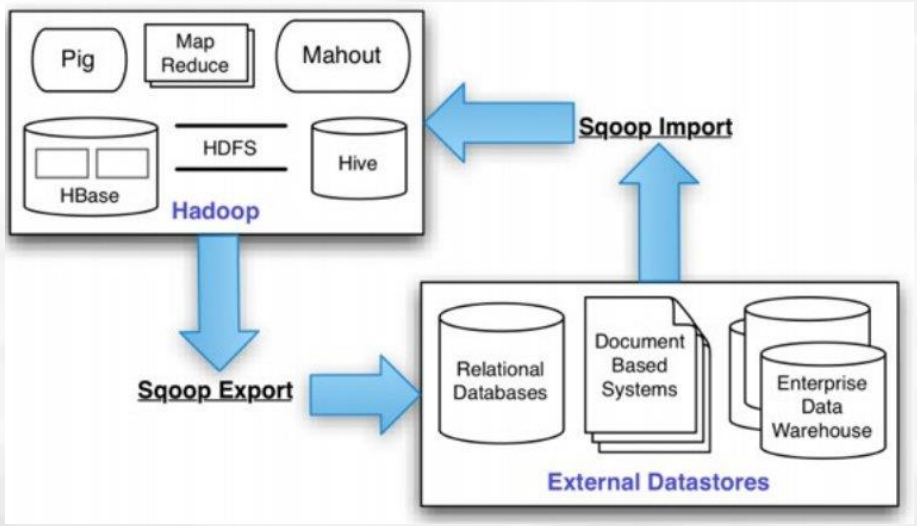



# Veri Entegresyonu: Sqoop



## Sqoop Temelleri

- HDFS ile ilişkisel veri tabanları arasında karşılıklı **veri aktarımını** sağlar.
  - Veritabanındaki tabloların tümü, bir kısmı
  - HDFS'deki verilerin veritabanına **tablo** olarak atılması
- JDBC kullanarak veritabanlarına bağlantıyı sağlar.
- JDBC yanı sıra belirli üreticilerin özel hazırladığı connector ile kendi sistemlerine bağlanabilir.

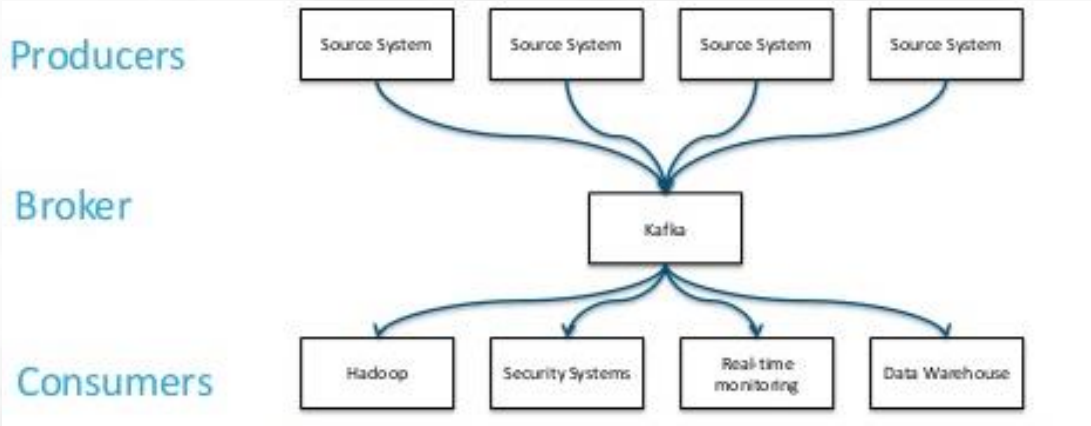


# Veri Entegrasyonu: Kafka



## Kafka Temelleri

- LinkedIn tarafından geliştirilen ve daha sonra açık kaynak Apache Projesi
- **Akan veriyi toplayarak, akan veri işleme** platformlarına veriyi yüksek bir çıkışla besleyebilen dağıtık mesajlaşma sistemidir.
  - Spark, Storm, Flink işleme platformlarına kolaylıkla entegre olabilmektedir.



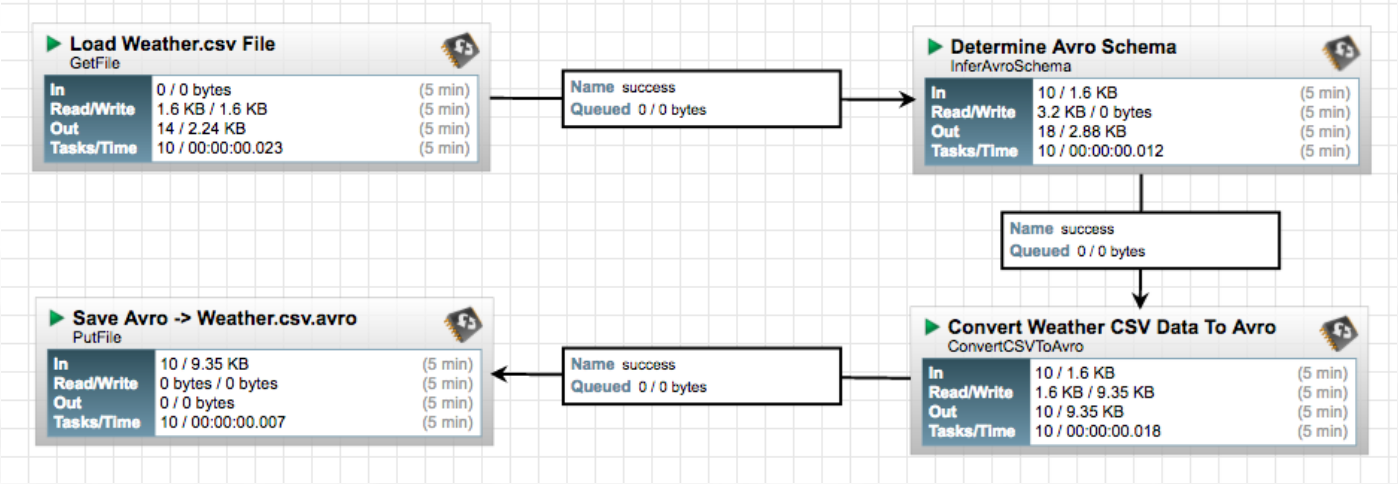

APACHE kafka

# Veri Entegrasyonu: Nifi



## NiFi Temelleri

- **Veri akış yönetimini** kapsamlı bir biçimde sağlayan bir platform.
  - Veri akış tasarımı, kontrolü, geri bildirimi ve izlenmesi
- Web tabanlı kullanıcı arayüzü ile kolaylıkla veri akışları oluşturulabiliyor.
  - Veri alma, gönderme, yönlendirme, dönüştürme, vs.

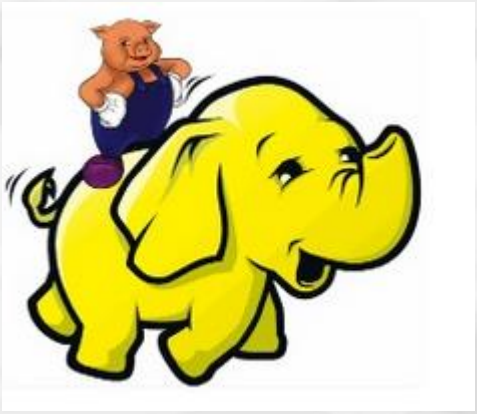



# Veri Analizi: Pig



## Pig Temelleri

- Hadoop'da MapReduce işleri oluşturmak için geliştirilen üst düzey platform.
- MapReduce yazmanın kolay alternatifi
  - PigLatin
- Hadoop üzerindeki **yapılandırılmamış** verinin analizini kolaylaştırır:
  - Basit dili ve yapısı sayesinde geliştirme süresini kısaltır.
  - Esnek veri modeli ve standart dosya formatlarına destek sağlar.
    - text, binary, sequence, json, vs.
  - SQL'den bildiğimiz bazı veri işleme ifadelerini kullanır.
    - group by, order by, vs.



# Veri Analizi: Hive



## Hive Temelleri

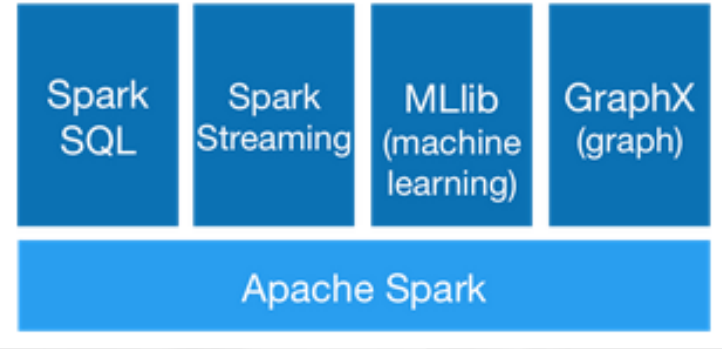
- Veri özetleme, sorgulama ve analiz işlemleri için Hadoop üzerinde geliştirilmiş bir Veri Ambarı (DataWarehouse)
- SQL benzeri bir dile sahiptir.
  - HiveQL
- Hadoop üzerindeki **yapılandırılmış** verinin yönetilmesini ve sorgulanmasını sağlar:
  - Veriler HDFS üzerinde saklanır.
  - Saklanan veriler, tablo yapısındadır.
  - Zengin veri tipleri sunar.
    - struct, array, map, vs
  - Farklı formatta tutulan verileri sorgulayabilir.
    - text, binary, sequence, vs
  - Ölçeklenebilir ve performanslıdır.

# Veri İşleme: Spark



## Spark Temelleri

- MapReduce işlem platformuna alternatif olarak geliştirilmiştir.
  - **Bellek içi hesaplama** yapma kabiliyeti sonucunda MapReduce göre ciddi hızlı sonuç üretebilmektedir.
- Genelleştirilmiş bir işlem platformudur.
  - **Akan veri** üzerinde çalışabilme
  - **Makine öğrenme** algoritma desteği
  - **SQL**, HiveQL analizleri yapabilme
  - **Grafik analizler** gerçekleştirebilme
- Kullanım kolaylığı önemli bir avantajıdır.
  - Java, Scala, Python, R dillerinde uygulama yazmaya imkan tanımaktadır.




APACHE Spark™

# Veri Depolama: HBase



**HBase Temelleri**

- HDFS üzerine kurulmuş bir **NoSQL veritabanı**dır.
- Google tarafından geliştirilen BigTable'ı temel almıştır.
- Hadoop üzerindeki verilerin düşük gecikmeli **yüksek performanslı** biçimde gerçek zamanlı yazma ve okuma işlemlerini yapabilmektedir.
  - Tablolara milyonlarca **ekleme, güncelleme** işlemlerini saniyeler içerisinde gerçekleştirir.
- Ölçeklenebilir dağıtık bir yapıya sahiptir.
  - Petabytes ve üzeri miktarlarda depolama yapabilir.

# Veri Arama: Solr



**Solr Temelleri**

- Apache Lucene üzerine kurulmuş açık kayak ileri düzey **arama platformu**
- Apache Lucene, açık kaynak Java arama/bilgi erişim kütüphanesi
- Solr belge tabanlı, yani belgeler arası arama yapar.
- Dağıtık indeksleme
- Hataya dayanıklılık
- Güçlü metin arama desteği
- İmla denetimi
- Veritabanı entegrasyonu
- Zengin doküman desteği

# Koordinasyon: ZooKeeper



**ZooKeeper Temelleri**

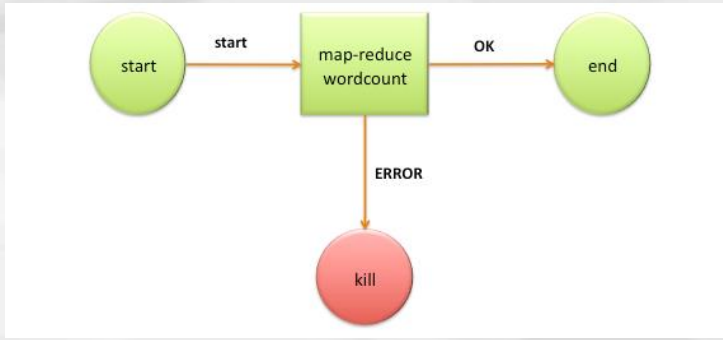
- Hadoop üzerinde dağıtık uygulama geliştirilmesini izin veren, dağıtık bir koordinasyon servisidir.
- Herkes işinin başında mı?
- Koordinasyon
- Dağıtık konfigürasyon
  - İstemci ve hizmet sağlayıcı listeleri
  - Hataya dayanıklılık
- Birlikte çalışabilirlik

# İş Akışı Zamanlayıcısı: Oozie



## Oozie Temelleri

- **Oozie**, Apache Hadoop işlerini yönetmek için bir **iş akışı zamanlayıcı sistemidir.**
- **Açık kaynak kodlu** ve ücretsiz
- **Periyodik işlemler** için uygun (dakikalık , saatlik , haftalık …)
- Birçok big data kütüphanesini destekler
  - Hadoop, Pig, Hive, Spark, Sqoop..
- Fork işlemini destekler. Bir MapReduce job'ı bittikten sonra aynı anda paralel bir şekilde devam edecek Pig ve Spark job'ı başlatabilir
- Hadoop dosya sistemi komutları kullanılabilir (hdfs dfs -copyFromLocal from to )
- Linux komutları çalıştırılabilir
- Java projesi çalıştırılabilir
- Kullanışlı bir arayüz sağlar
- Ölçeklenebilir, güvenilir ve genişletilebilir



# Veri Güvenliği: Ranger



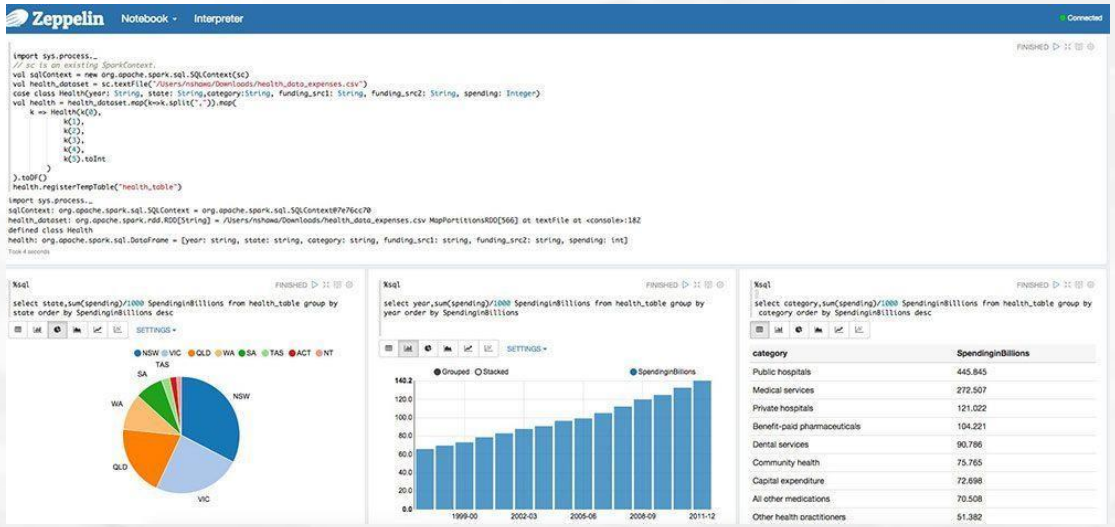
## Ranger Temelleri

- Hadoop üzerinde **güvenliğin** sağlanmasını kolaylaştırıyor:
  - Yetkilendirme (authorization)
  - Kimlik doğrulaması (authentication)
  - Denetleme (auditing)
  - Veri kriptolama (data encryption)
  - Güvenlik yönetimi (security administration)
- Hadoop ekosisteminde destek verdiği mevcut projeler:

# İnteraktif Veri Analizi: Zeppelin



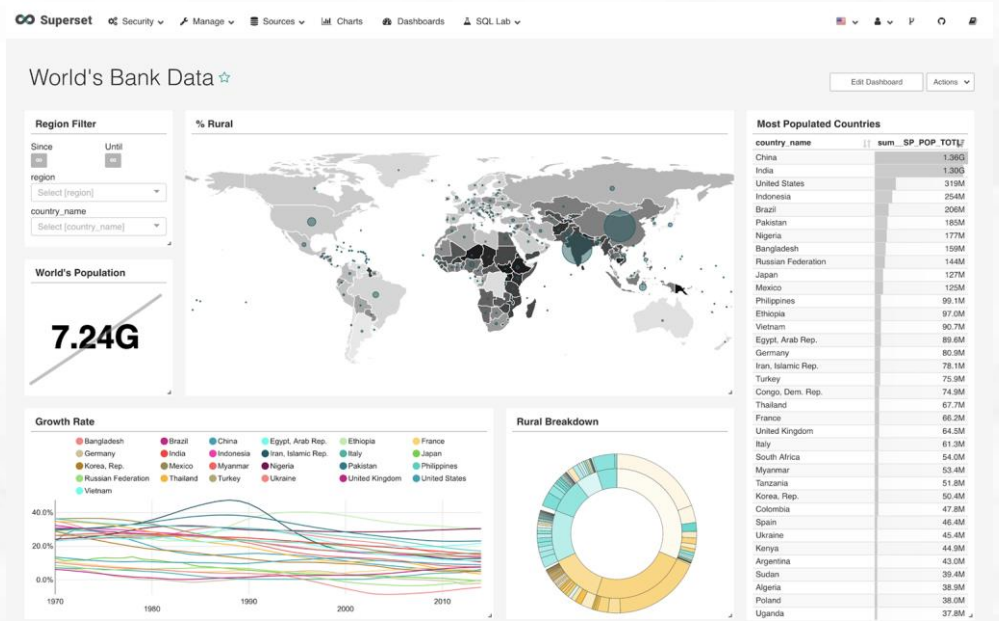
- Etkileşimli veri analizi sağlayan tamamen **açık bir web tabanlı notebook.**
- Veri mühendislerinin, veri analistlerinin ve veri bilimcilerin, veri **kodunun geliştirilmesi, organize** edilmesi, yürütülmesi ve paylaşılması ve sonuçların komut satırına atıfta bulunulmadan veya küme detaylarına ihtiyaç duymadan **görselleştirilmesi** yoluyla daha üretken olmasını sağlar.
- Apache Zeppelin, Spark için **veri araştırma, görselleştirme, paylaşım ve işbirliği** özelliklerini bir araya getiren yeni bir web tabanlı notebookdur.
- **Python, Scala, Hive, SparkSQL, shell ve markdown** gibi programlama dillerini desteklemektedir.

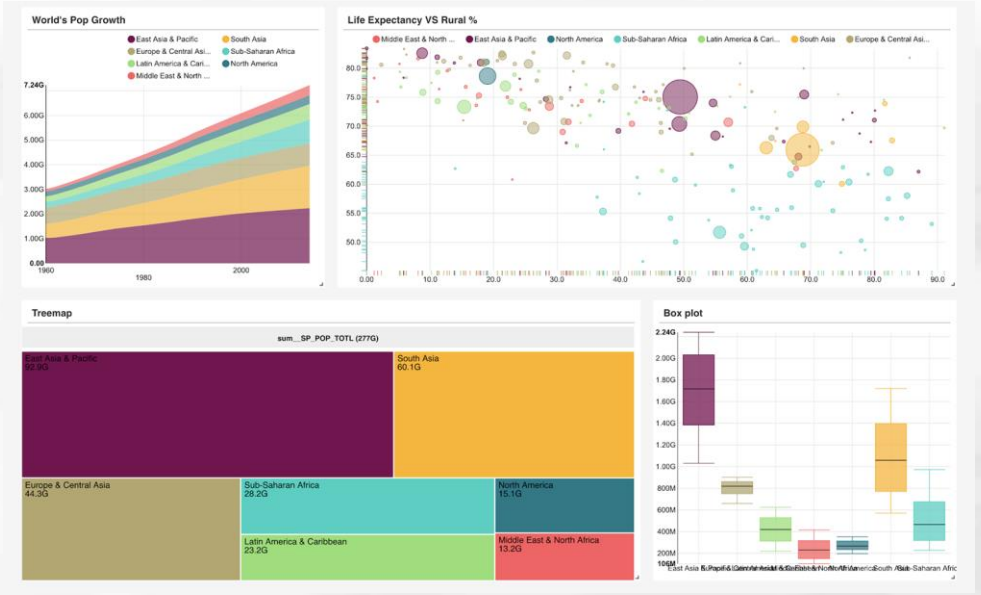



# Veri Görselleştirme: Apache Superset



- Web tabanlı **iş zekası aracı**
- Açık kaynaklı
- Etkileşimli **dashboard** oluşturma
- **Aggregations** desteği (Druid ile)
- Kylin, Presto, Hive, Impala, SparkSQL, MySQL, Postgres, Oracle, Redshift, SQL Server, Druid
- Time series, Bar chart, Bubble chart, Word cloud, World map, Histogram, ...
- Kimlik doğrulama desteği: Database, OpenID, LDAP, OAuth, REMOTE_USER - Flask AppBuilder)



# Teşekkür Ederiz..

www.b3lab.org

TÜBİTAK BİLGEM Gebze Yerleşkesi, Gebze, Kocaeli, TÜRKİYE, 41470
T: +90 262 675 23 92 E: b3lab.iletisim@tubitak.gov.tr