Büyük Veri & Alet Çantam


694,444
hours watched
3.8 MILLION
search queries
1 MILLION
logging in
18.1 MILLION
texts sent
2.1 MILLION
snaps created
4.5 MILLION
videos viewed
41.6 MILLION
messages sent
390,030
apps downloaded
60
SECONDS
$996,956
spent online
347,222
scrolling
4.86 MILLION
impressions served
87,500
people tweeting
2 MILLION
RTB queries
$1.05
$0.75
1.4 MILLION
swipes
1.8 MILLION
emails sent
41
music streaming
subscriptions

Spotify “Discover Weekly” Listesi

Discover Weekly

PAUSE

	TITLE	ARTIST	ALBUM	
♡	C Section	Dot Hacker	N°3	5 days ago
♡	Animal	Badflower	Temper	5 days ago
♡	Bottom of the Lake	The Builders and The Butchers	The Builders and the Butchers	5 days ago
♡	Tired Old Dog	The Devil and the Almighty Bl...	The Devil and the Almighty Bl...	5 days ago
♡	Impossible Dreams	Versus Them	Six A.M. Salvation	5 days ago
♡	The Rains Of Castomere - From The "Game Of Thron...	The National	Game Of Thrones: Season 2 (...	5 days ago
♡	Death in Greenpoint	Mishka Shubaly	When We Were Animals	5 days ago
♡	Sugar Boats	Modest Mouse	Strangers to Ourselves	5 days ago
♡	Love	Colour Haze	Colour Haze	5 days ago

dave horwitz
@Dave_Horwitz

It's scary how well @Spotify Discover Weekly playlists know me. Like former-lover-who-lived-through-a-near-death experience-with-me well.

♡ 278 10:09 PM - Oct 27, 2015

190 people are talking about this

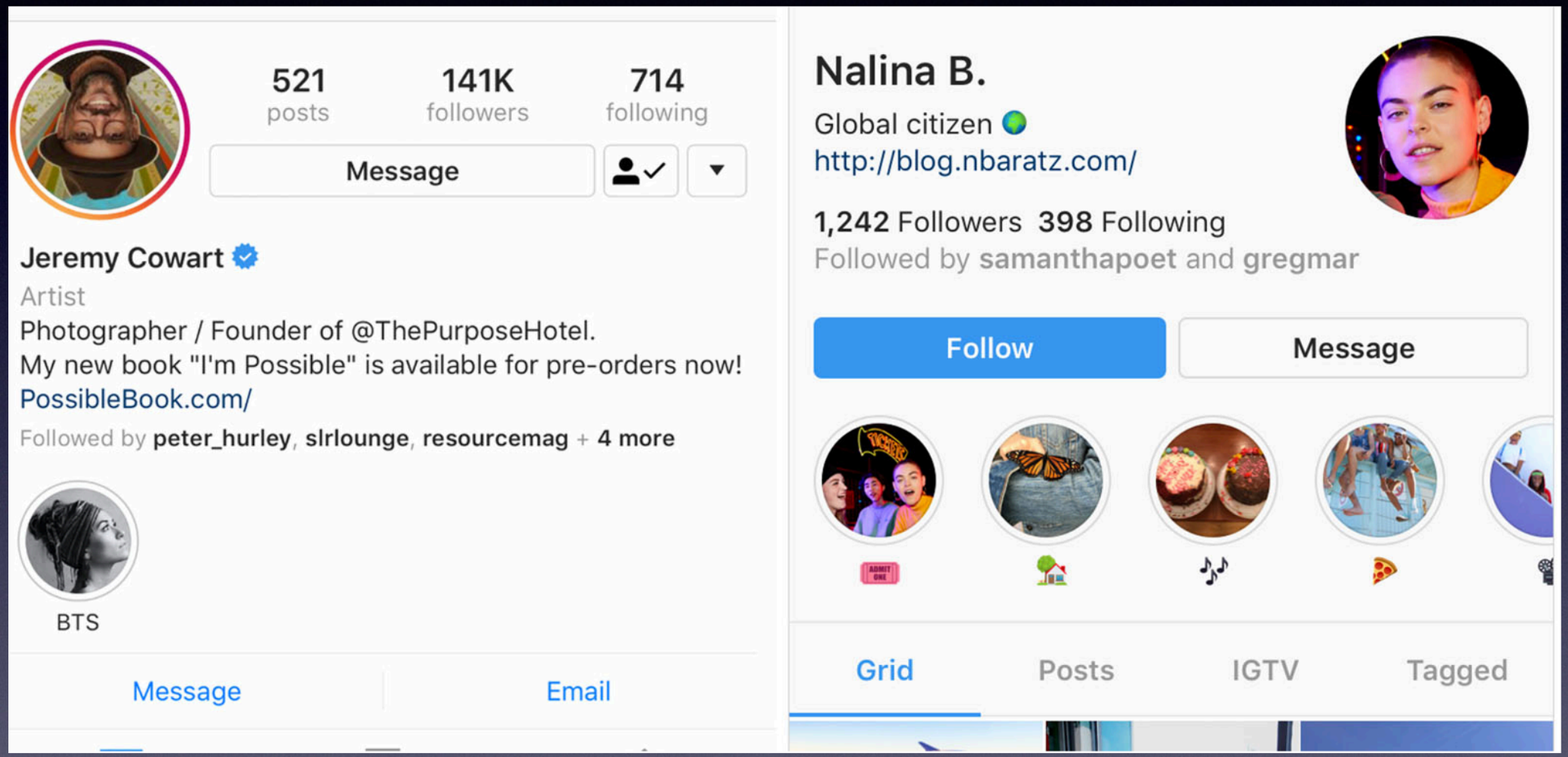
Amanda Whitbred
@amandawhitbred

At this point @Spotify's discover weekly knows me so well that if it proposed I'd say yes

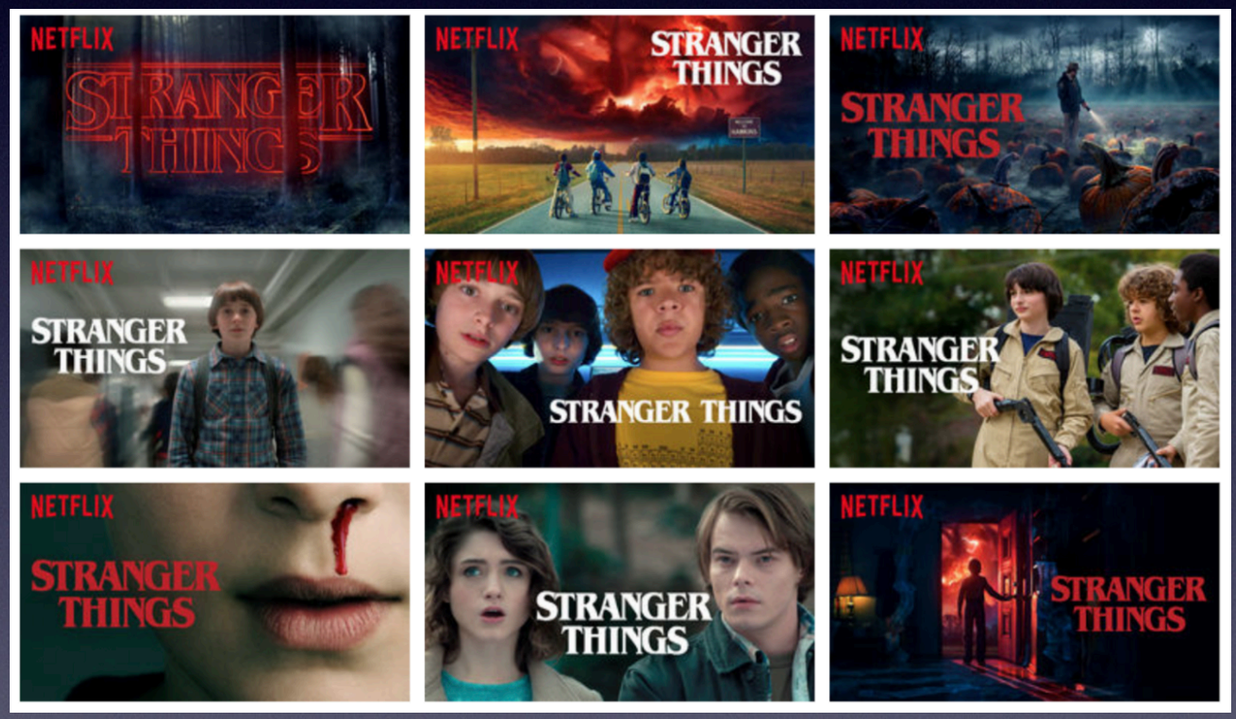
♡ 76 12:36 AM - Aug 19, 2016

25 people are talking about this

Instagram’ın aynı zamanda sunduğu farklı profil sayfaları

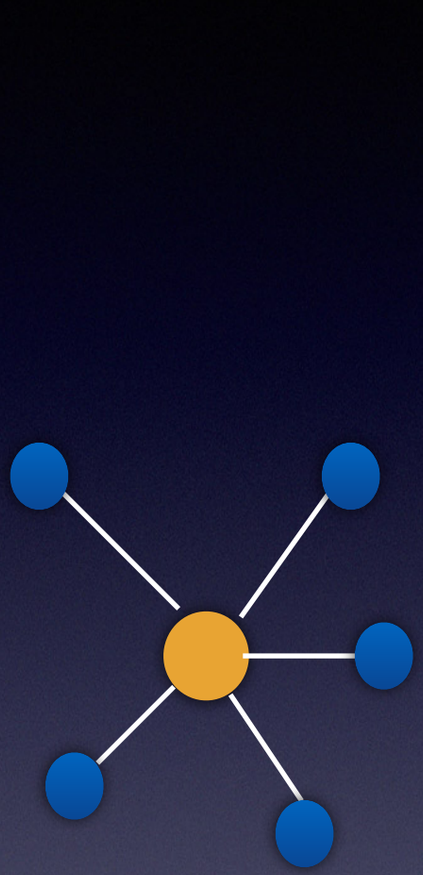


Netflix’in Kişileştirilmiş Özellikleri

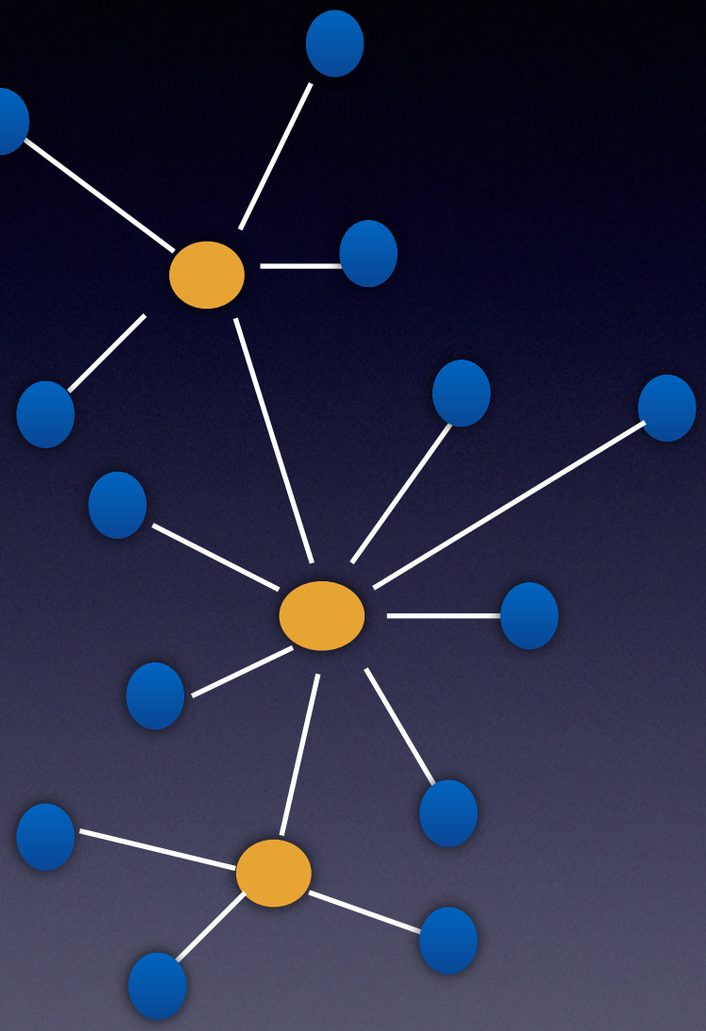


Alışveriş Sitelerinin Bizi Çevremizden iyi Tanıması

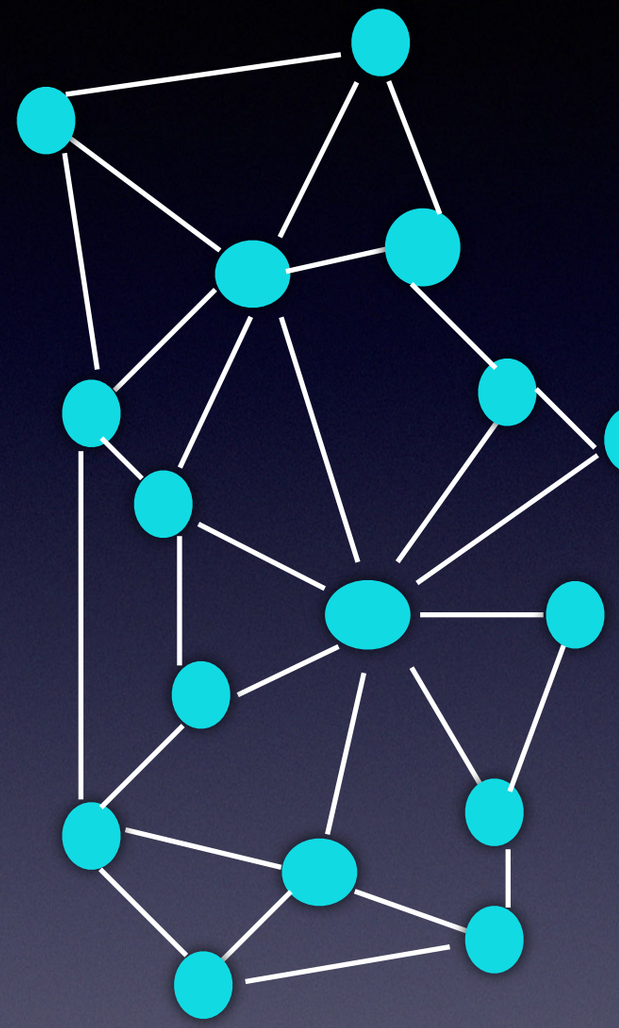
Veriyi Saklama ve Erişim



Centralized

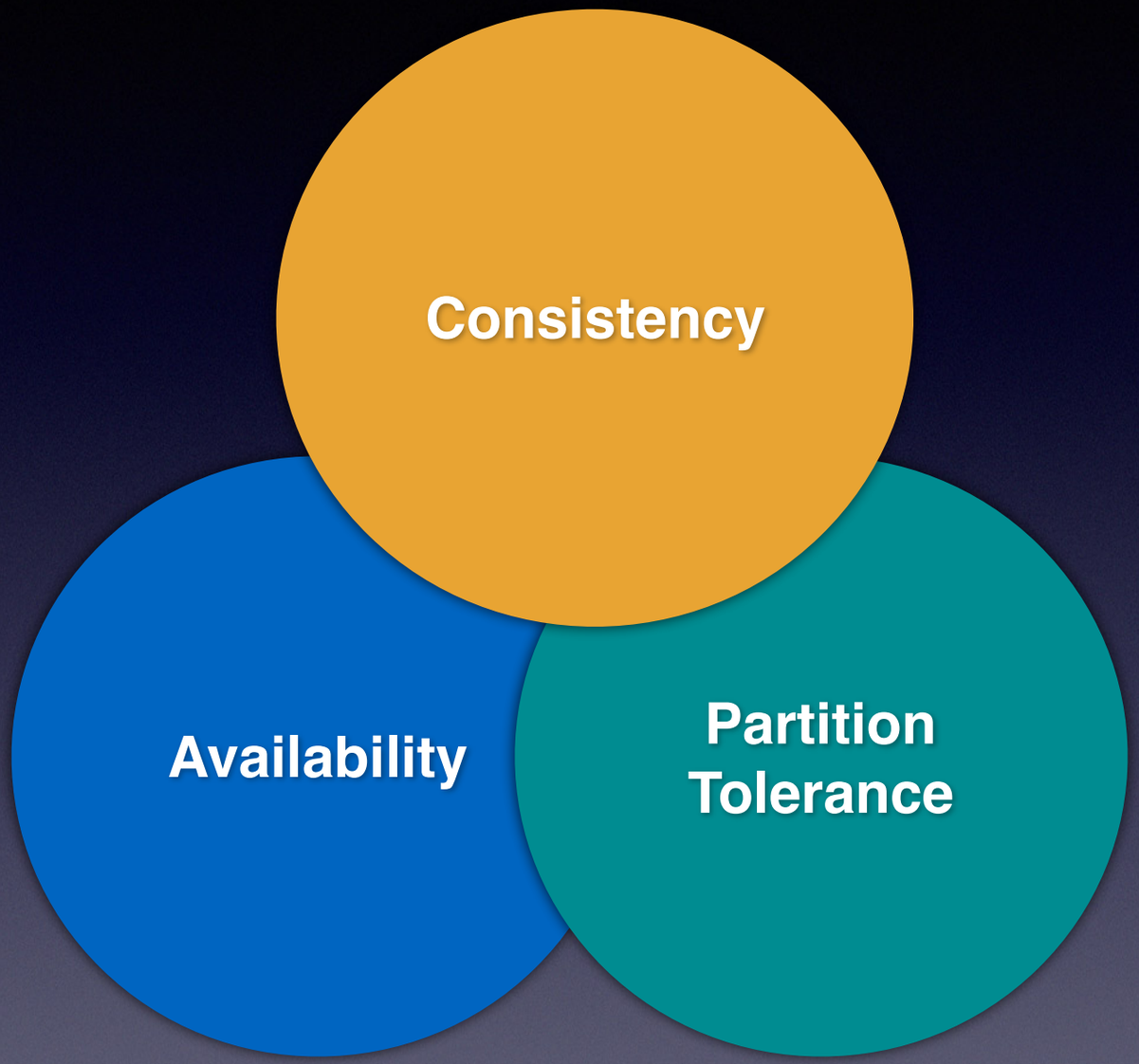


Decentralized



Distributed

CAP Teoremi



Big Data
5V

Volume

Variety

Veracity

Velocity

Value

Veri Modelleri

- Structured Data
- Semi-Structured Data
- Unstructured Data

Web / Mobile

Social Media

Web Logs

A/B Testing

IOT - Sensor

Smart Home

Self-driven Car

Web Scraping

Health Data

User History

Click Streaming

Behavioral Targeting

Boyut

Veri çeşitliliği ve karmaşıklığı

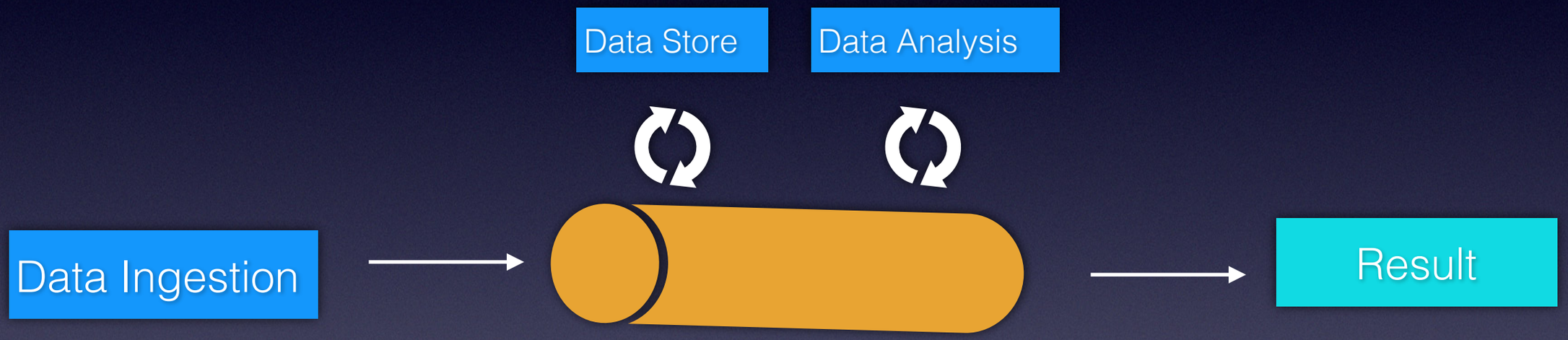
Aynı zamanda …

- Saklama maliyeti
- CPU maliyeti

- Network Erişimi

Big Data = İşlem + Etkileşim + Gözlem

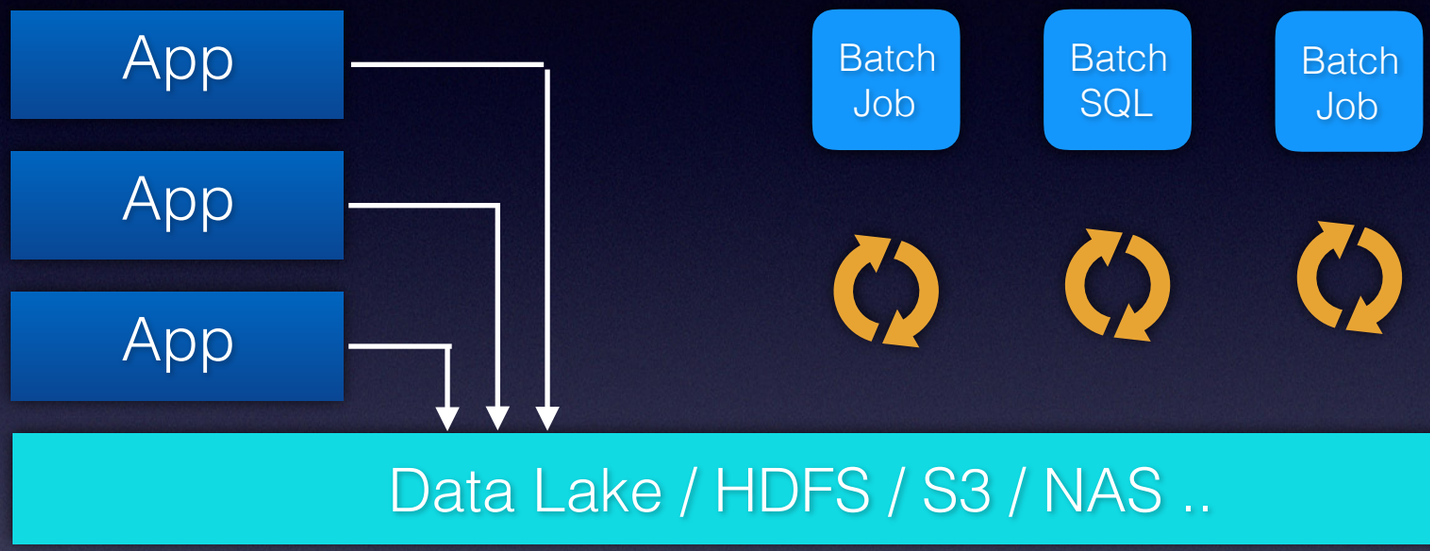
Data Pipeline



VERİ İŞLEME MODELLERİ

- Batch Processing
- Stream Processing / Real-time Processing

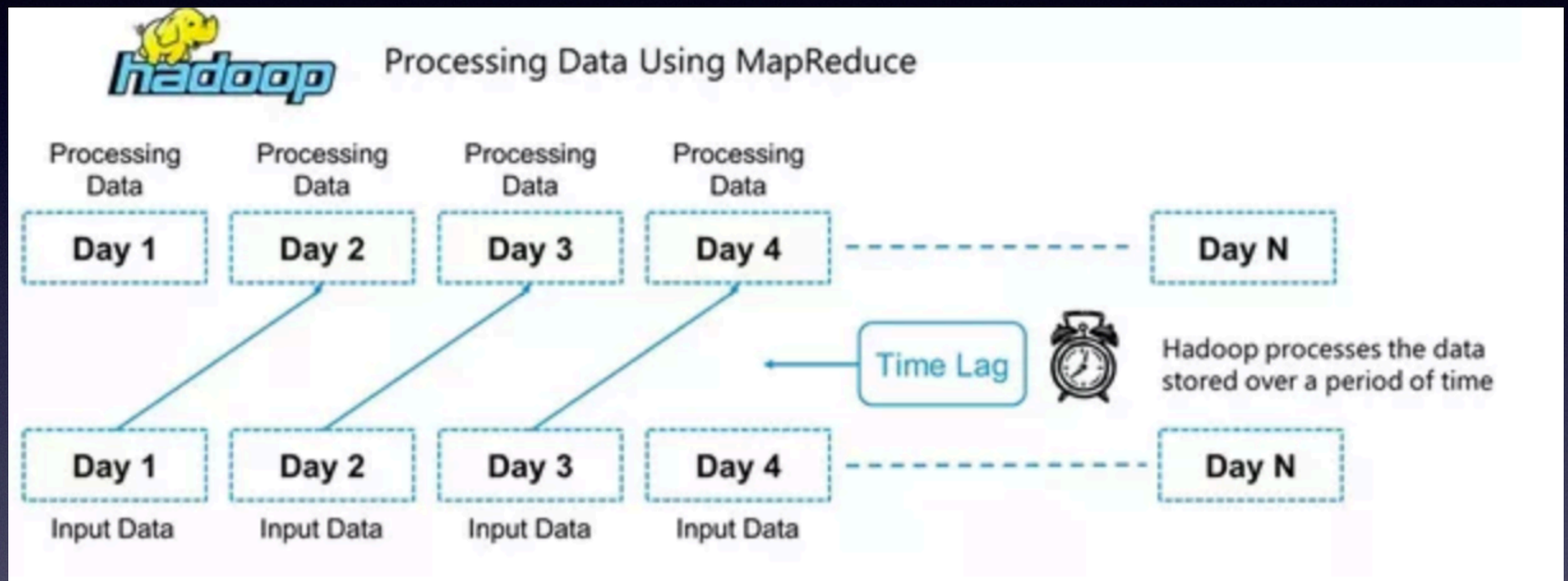
BATCH PROCESSING



Batch Processing

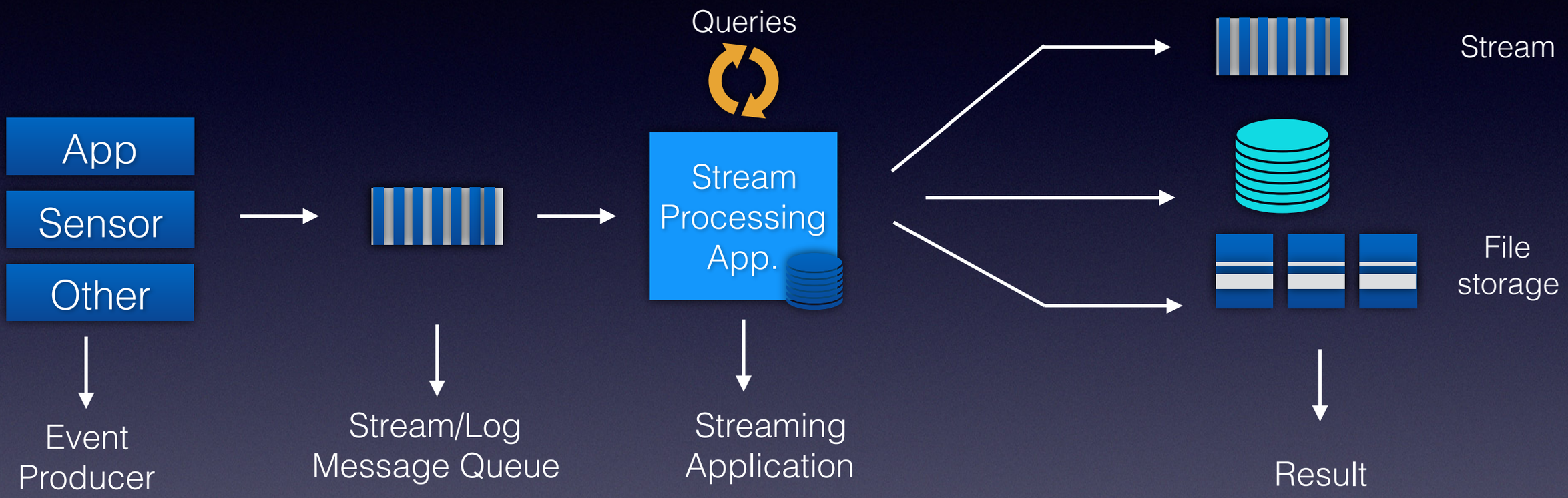
- Önceden depolanmış veriler üzerinde yapılan analizlerdir
- Veri limitlidir
- Belirtilmiş zaman aralığındaki analizlerde kullanılır.
- Anlık işlemeye ihtiyaç duyulmadığı durumlarda kullanılır.
- Bazı gerçek zamanlı teknolojilerde Micro-batch olarak da adlandırılır.

BATCH PROCESSING MAPREDUCE



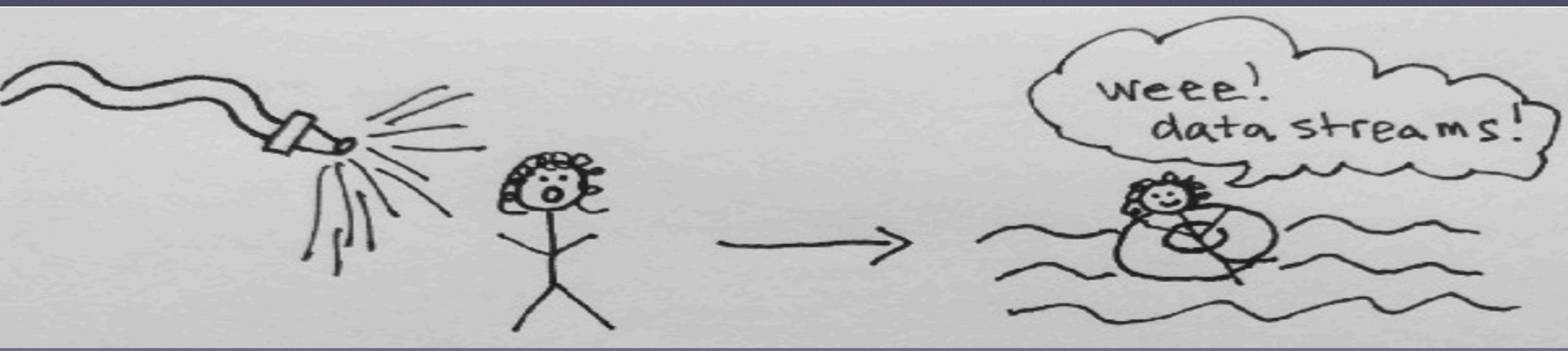
Processing Data Using MapReduce

STREAMING PROCESSING



STREAMING PROCESSING

- Verinin limiti yoktur.
- Veriyi elde ettiği anda çok kısa bir sürede kurgulanan şekilde işler.
- Hızlıdır.
- Paraleldir.



NEDEN GEREKLİ ?

- Bazı durumlarda verinin sonsuz olması
- Anlık takibini yapmak
- Birden fazla akış üzerinde veriyi eşzamanlı erişim
- Toplu şekilde işlemenin yüksek gecikmeyle sonucu vermesi ve depolama gereksinimi
- Veri depolanamayacak kadar büyük olabilir.

KULLANILAN DURUMLAR

- Sensör Verileri
- Web Site Aktiviteleri
- Finansal Veriler
- Zaman Serisi Verileri

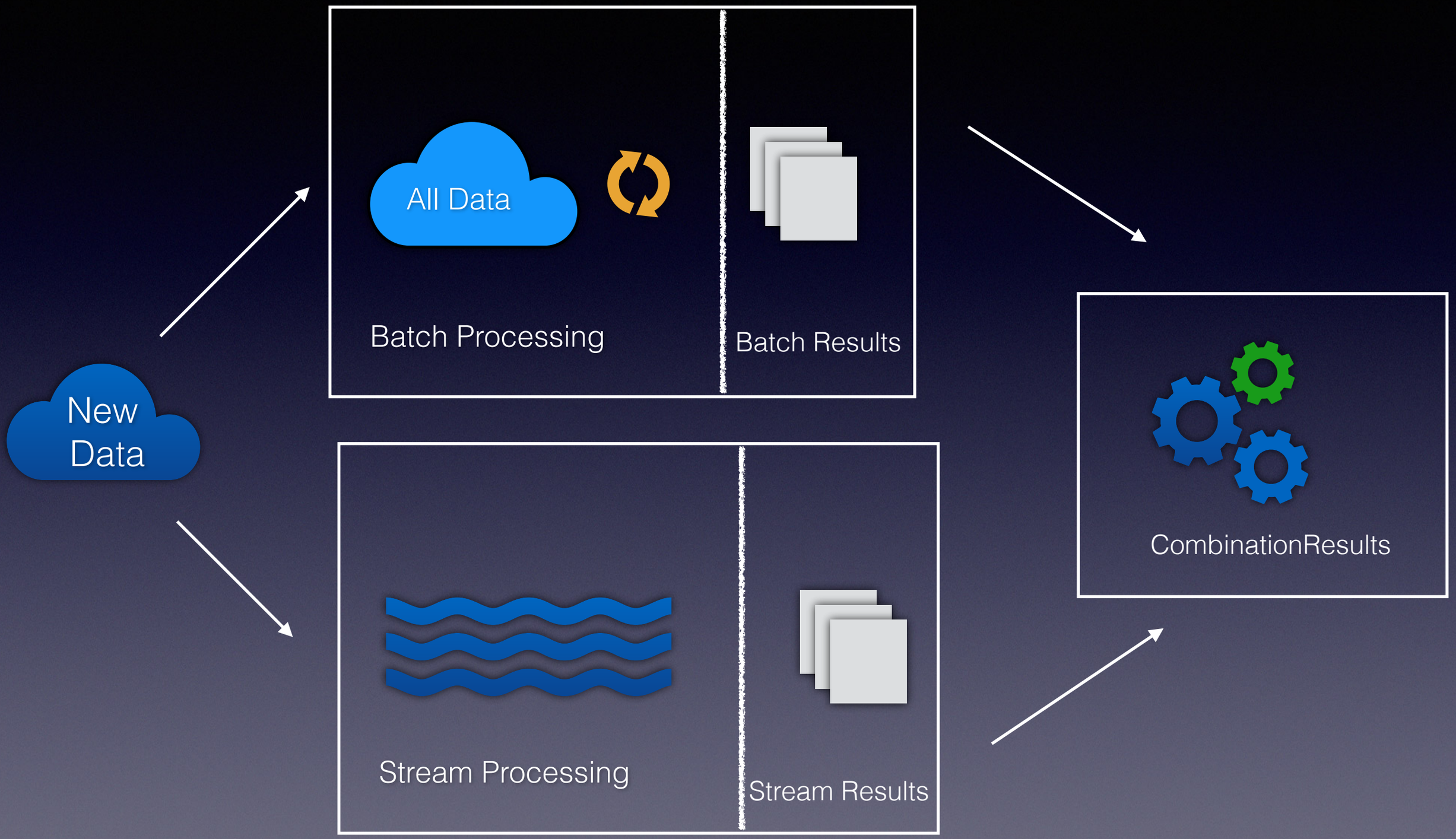
Sonuçlarında Neler Oluyor ?

- Uyarı Sistemleri
- Fraud Detection
- Akıllı Ev, Tarım sistemleri
- Sürücüsüz araçlar
- Makine öğrenmesine eğitim verisi

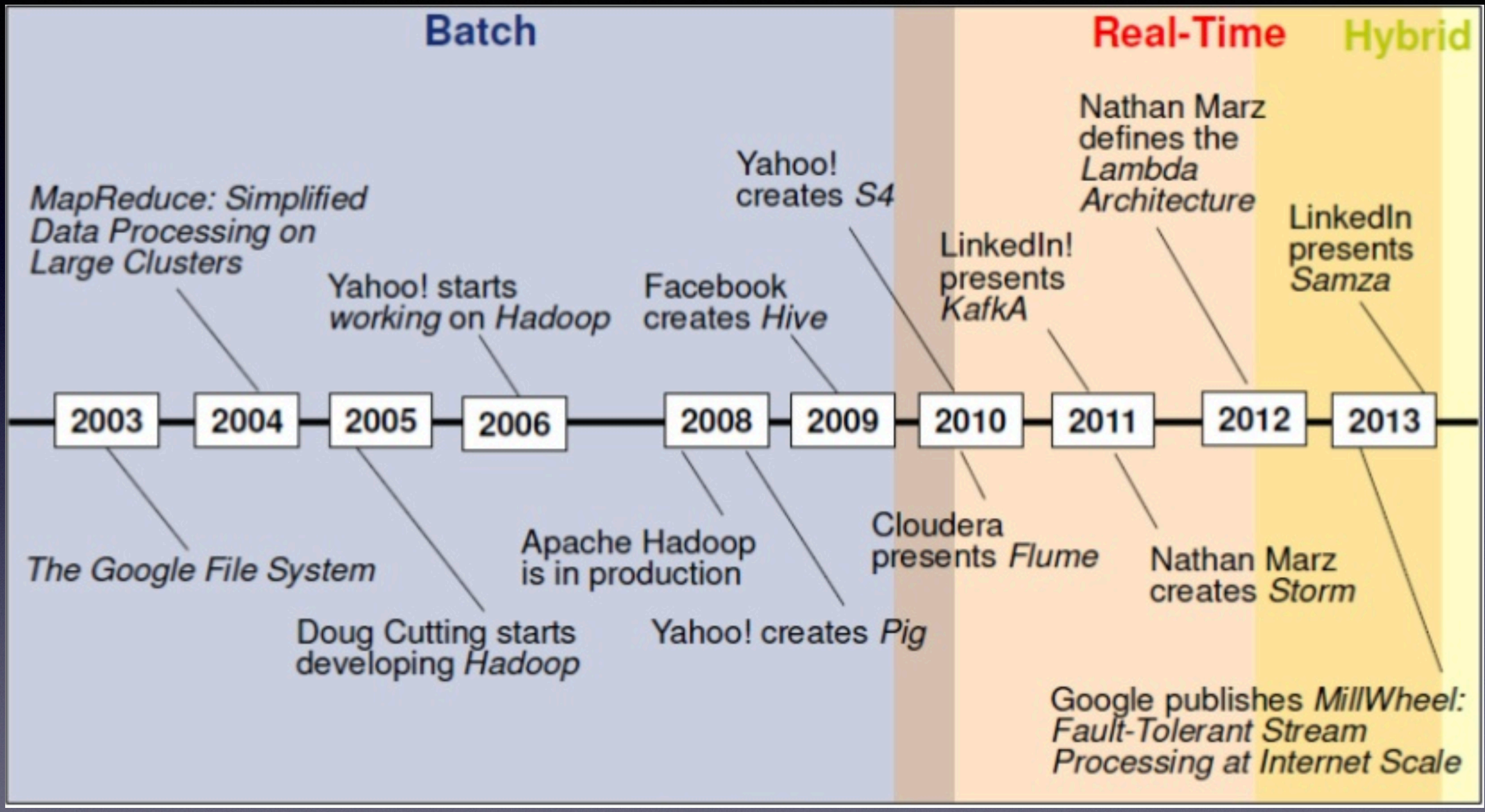
Veriyi İşleme Tarzı - Stream Processing

En fazla bir defa (At Most Once)	En az bir defa (At Least Once)	Sadece bir defa (Exactly Once)
Mesaj bir defa gönderilir	Mesaj bir veya daha fazla gönderilir, her defa işlenir	Mesaj bir veya daha fazla gönderilir ama bir defa işlenir
Mesaj karşılanabilir.	Mesajın tüketileceği garantidir.	Mesajın tüketileceği garantidir.
Veri çoklanmaz	Veri çoklanabilir	Veri çoklanmaz
Veri kaybı olabilir.	Veri kaybı olmaz	Veri kaybı olmaz

Hibrit Model



Source

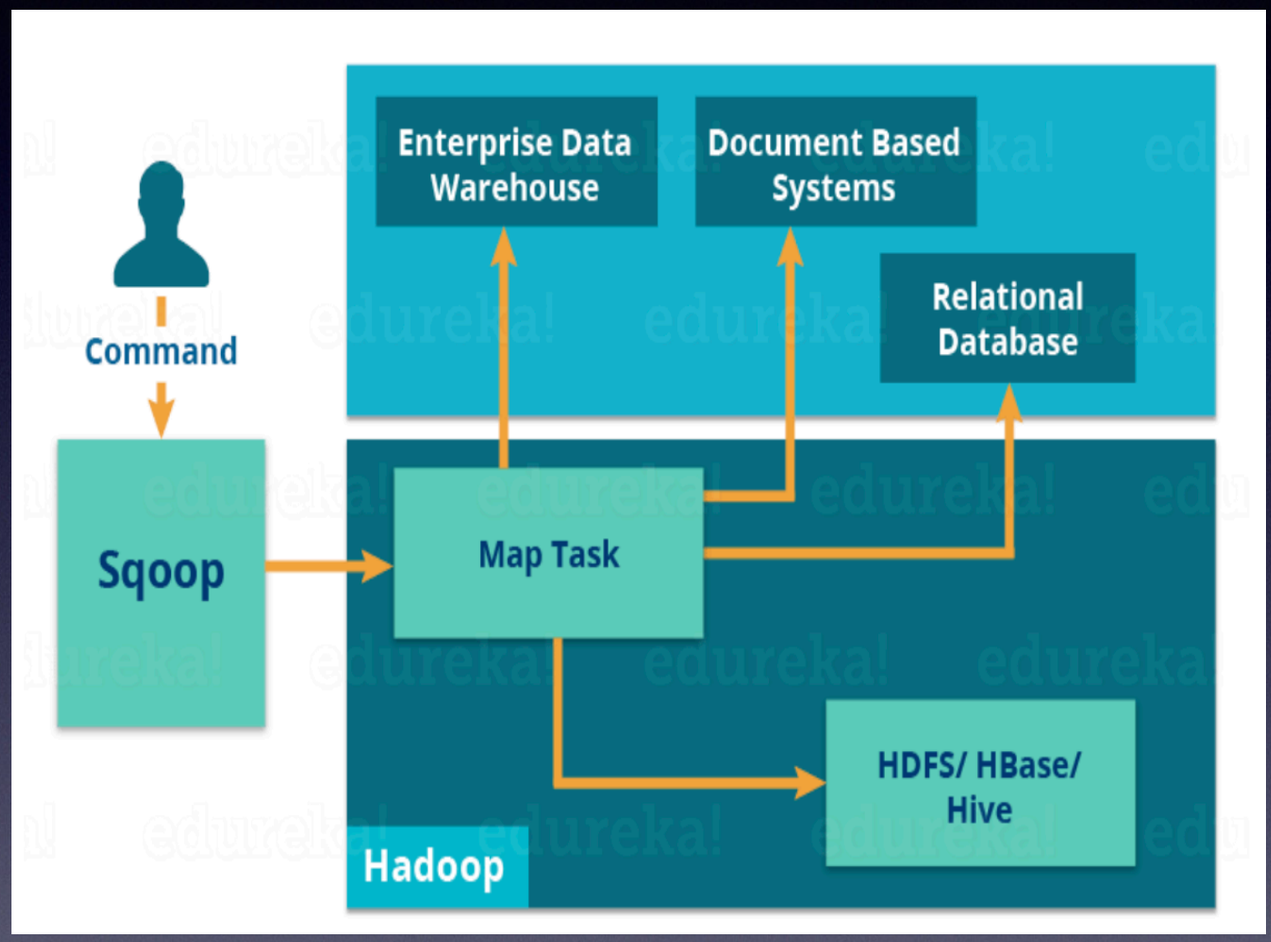

Batch
Real-Time
Hybrid
MapReduce: Simplified Data Processing on Large Clusters
Yahoo! starts working on Hadoop
Facebook creates Hive
Yahoo! creates S4
LinkedIn! presents KafkA
Nathan Marz defines the Lambda Architecture
LinkedIn presents Samza
2003
2004
2005
2006
2008
2009
2010
2011
2012
2013
The Google File System
Doug Cutting starts developing Hadoop
Apache Hadoop is in production
Yahoo! creates Pig
Cloudera presents Flume
Nathan Marz creates Storm
Google publishes MillWheel: Fault-Tolerant Stream Processing at Internet Scale

Büyük Veri Teknolojileri

Veri Toplama Teknolojileri

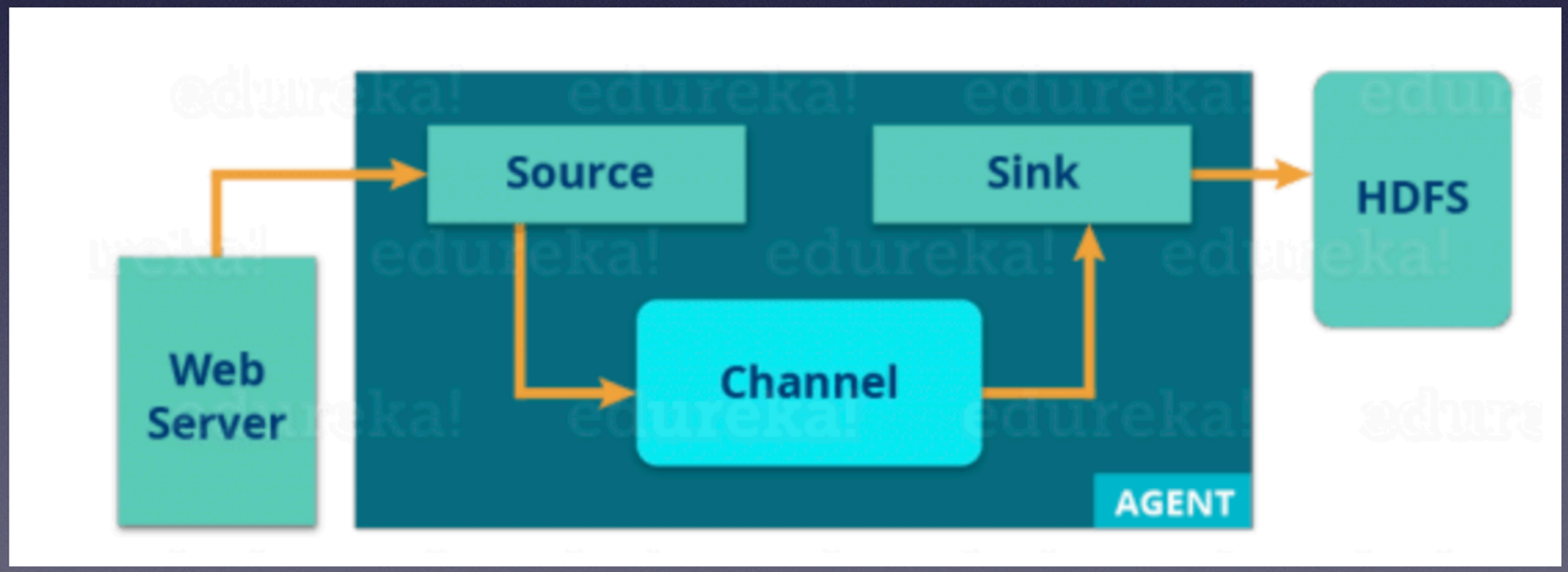
Apache Sqoop

- İlişkisel veritabanları ve Hadoop arasında veri aktarımı yapar
- Mapreduce Job'ları oluşturur.
- Yarn kullandığından paralel olarak import ve export işlemi yapabilir.
- Komut satırı ile istenilen boyuttaki veriyi istenilen yere aktarır.



Apache Flume

- Birden fazla kaynaktan veriyi toplayıp HDFS'e yazar
- Dağıtıktır.
- Channel bazlı transaction api'yı vardır. (Sender & receiver)



Apache Nifi

- Farklı kaynaklardan veri toplanmasına, işlenmesine ve farklı ortamlara aktarılmasını sağlar
- Web arayüzü sunar.
- Dağıtıktır.
- Kafka, HDFS, NoSQL gibi ortamlara veri akışını yönlendiren hazır işlemciler (processor) vardır.
- Kendi custom processor'lerimizi yazabiliriz.
- Kendi içersinde kuyruk yapısı vardır.

Fluentd

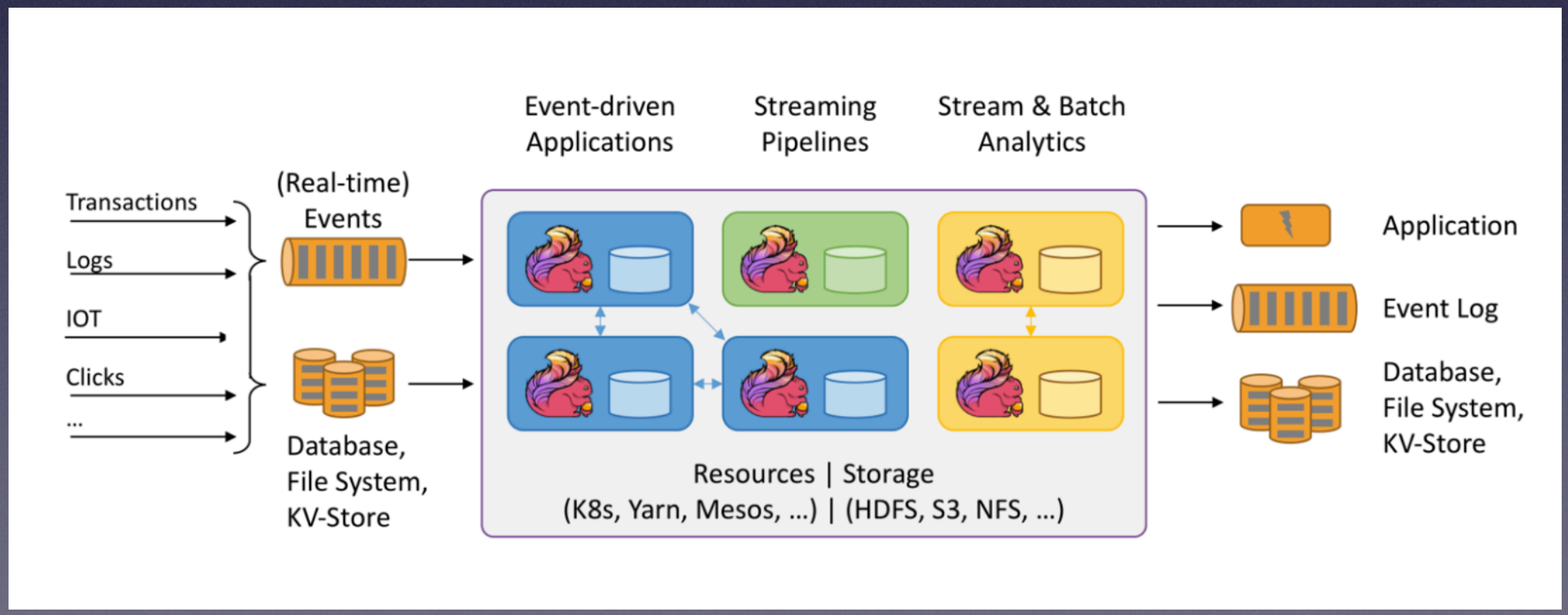
- Log yapılarını birleştirmek için tasarlanmıştır
- Mümkün olduğunca JSON veri formatında çalışır
- Çoklu kaynaklara rağmen log toplamayı, filtrelemeyi ve çıktı vermeyi kolaylaştırır.
- Kaynak tüketimi ve kurulumu oldukça kolaydır.

Realtime / Near-Realtime Streaming Teknolojileri

beam

Apache Flink

- Hadoop ile uyumlu veri akış platformudur.
- Batch ve stream programlar çalıştırabilir.
- FlinkML kütüphanesi ile makine öğrenmesi ortamı sunar.
- Stateful Streaming ile “exactly once” garantisi verir.



Source

Apache Beam

- Hem batch hem de streaming processing API'ıdır.
- Platform bağımsızdır.
- Oluşturulan data pipeline Apache flink, spark, Google dataflow gibi ortamlarda çalıştırılabilir.

Saklama Teknolojileri

NoSQL

Couchbase

mongoDB

elasticsearch

cassandra

Key/Value

redis

etcd

GraphSQL

neo4j

Time Series

influxdb

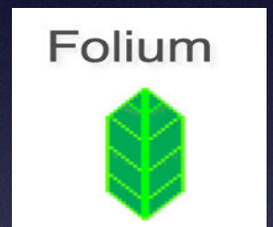
Prometheus

cloudera
IMPALA

APACHE
HBASE

HIVE

Görselleştirme Teknolojileri



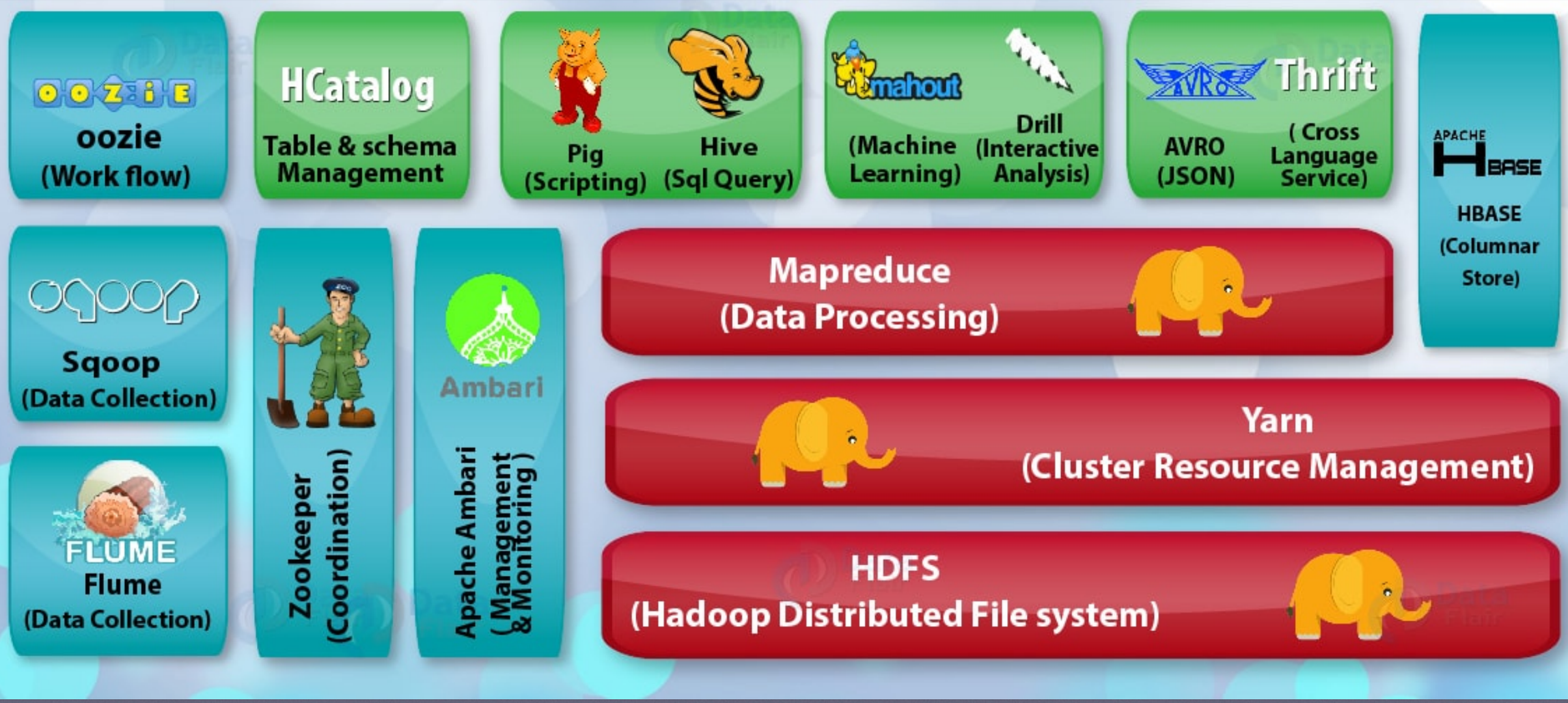
Hadoop Ekosistemi

Hadoop Nedir ?

- Framework
- Büyük veri üzerinde çalışan uygulamalara ortam sunar.
- Açık kaynak kodludur.
- Dağıtıktır.
- Ölçeklenebilir yapıdadır.

- İki ana katmanı vardır
    1. Dağıtık Dosya Sistemi (HDFS)
    2. Dağıtık veri işleme - MapReduce

Hadoop Ekosistemi



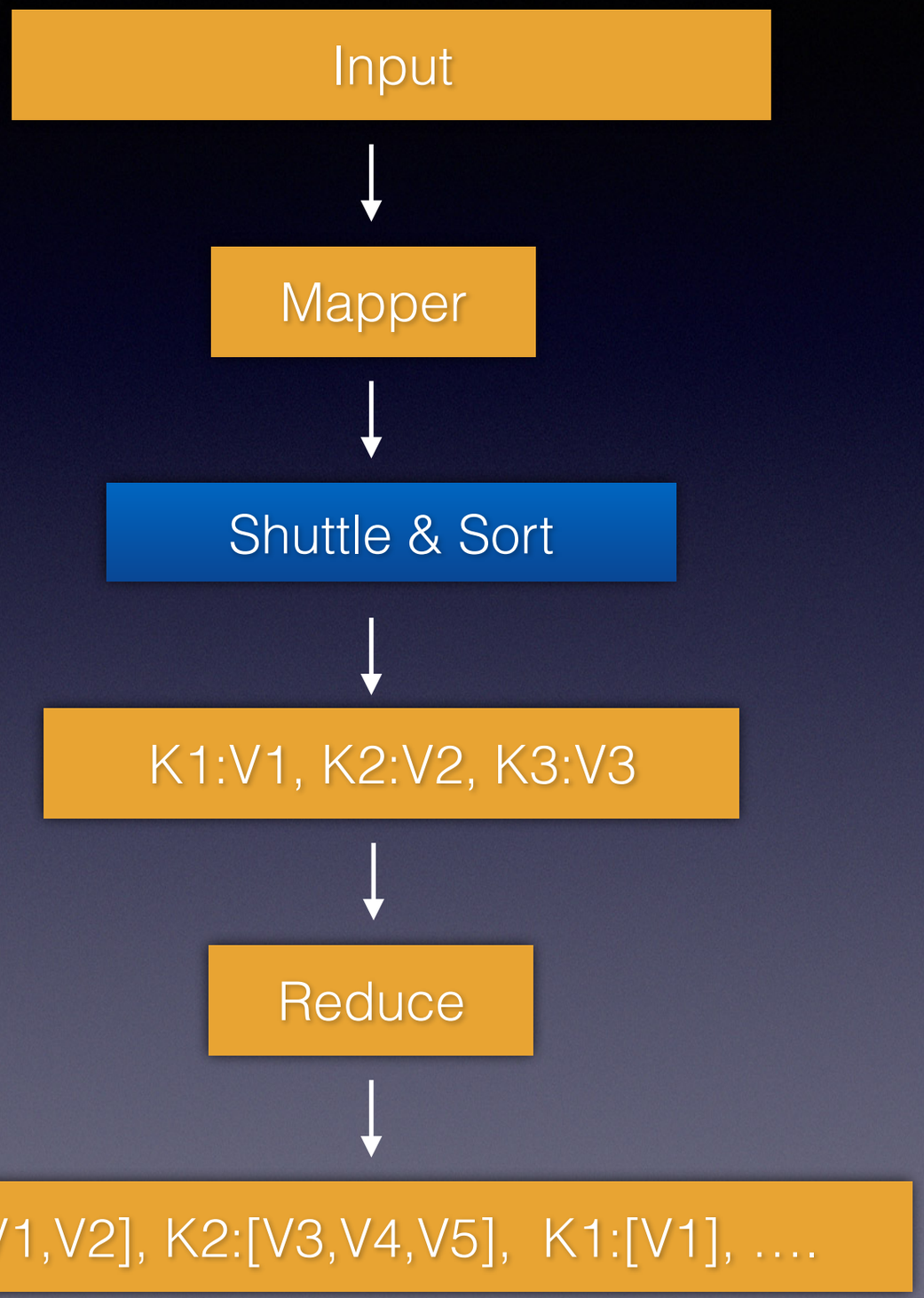
HADOOP YONETİM PLATROMLARI

	Apache Hadoop	Hortonworks	Cloudera	MapR
Open Source	Yes	Yes	Partially	No
Support	Community	Enterprise Support	Enterprise Support	Enterprise Suppo
Frontend	Apache Ambari	Apache Ambari	Cloudera Manager	MapR Control System
Price	Free	$$	$$	$$$
Focus	Open Source, reliable, scalable, distributed computing	Enterprise capabilities	Enterprise capabilities	Enterprise & Performance

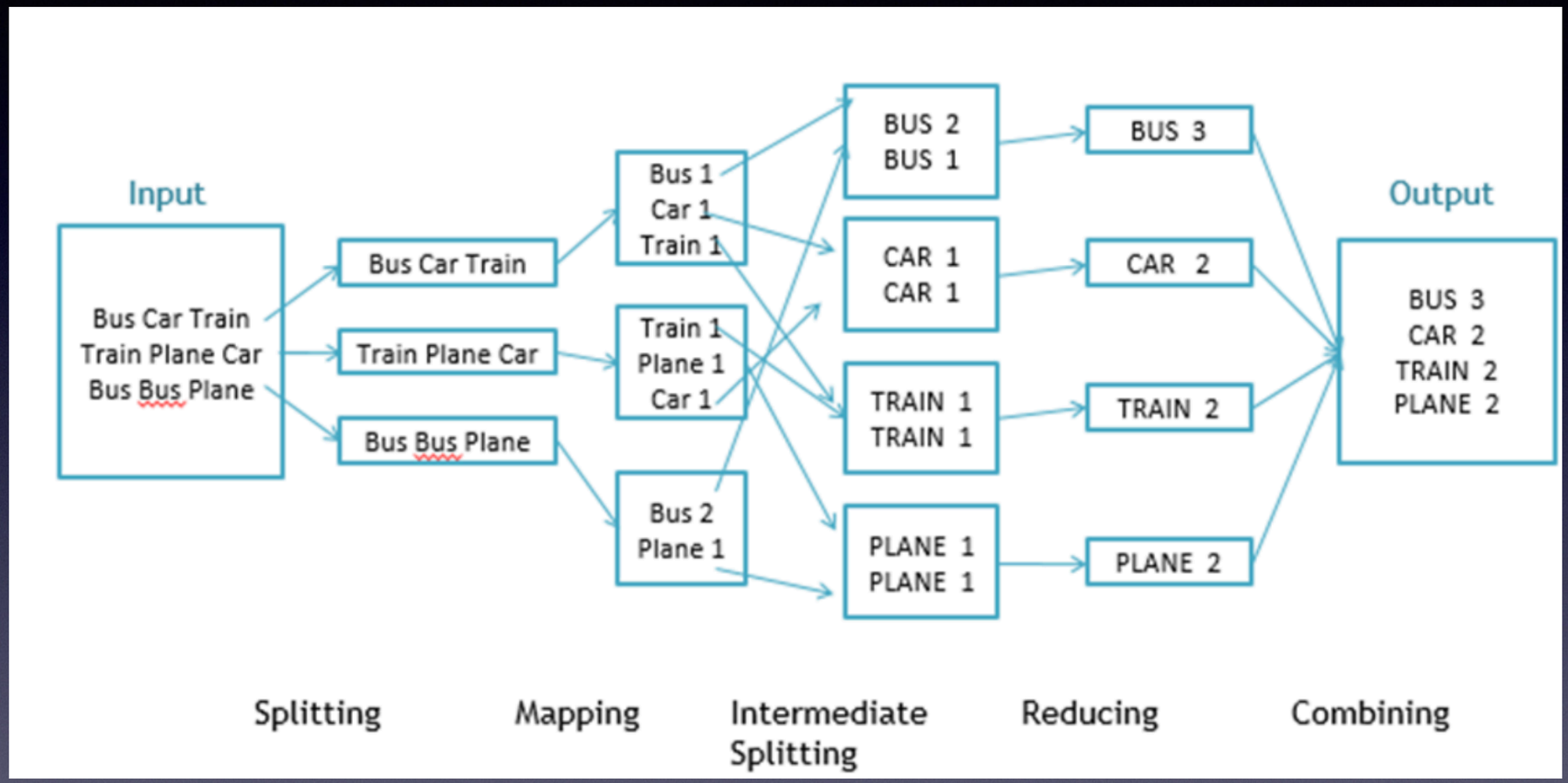
Mapreduce

- **Dağıtık** veri analiz framework'ü
- **Hadoop V1 ile hem kaynak yönetimini hem de veri analizi**
- Birçok dil ile geliştirme yapılır.
- **Pig** ile daha kolay geliştirme sunar.
- **Batch** işlemler gerçekleştirir.
- **Yüksek gecikme**

Mapreduce

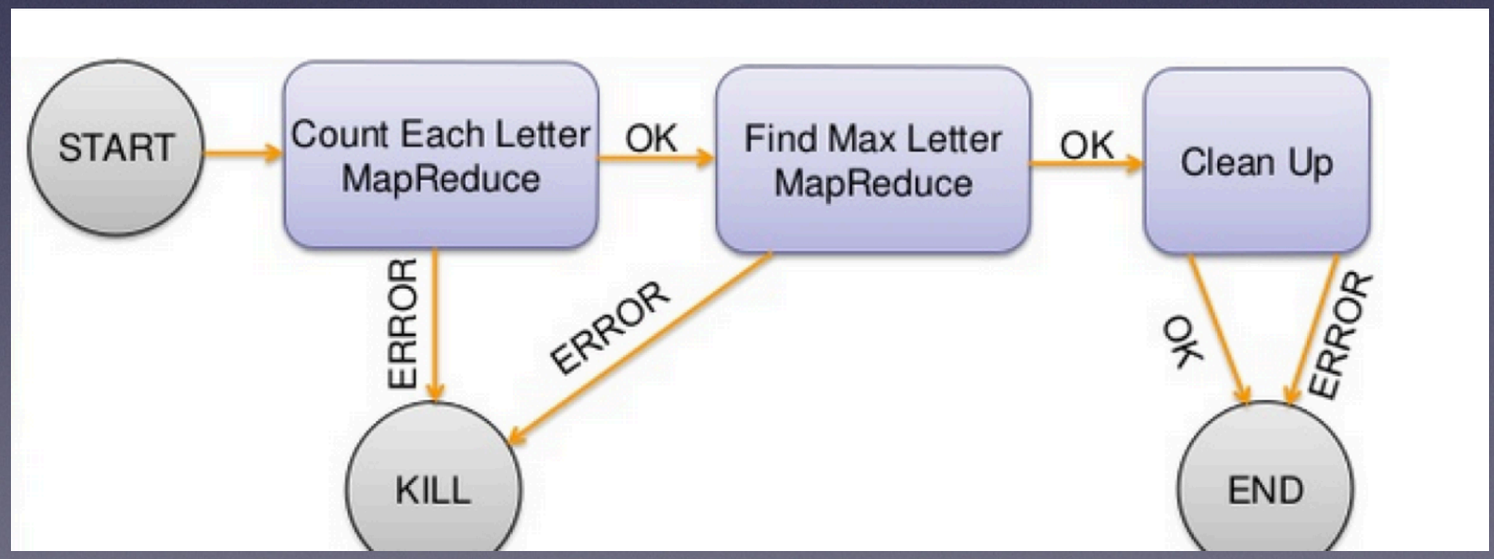

Input
Mapper
Shuttle & Sort
K1:V1, K2:V2, K3:V3
Reduce
K1:[V1,V2], K2:[V3,V4,V5], K1:[V1], ….

Mapreduce Word Count



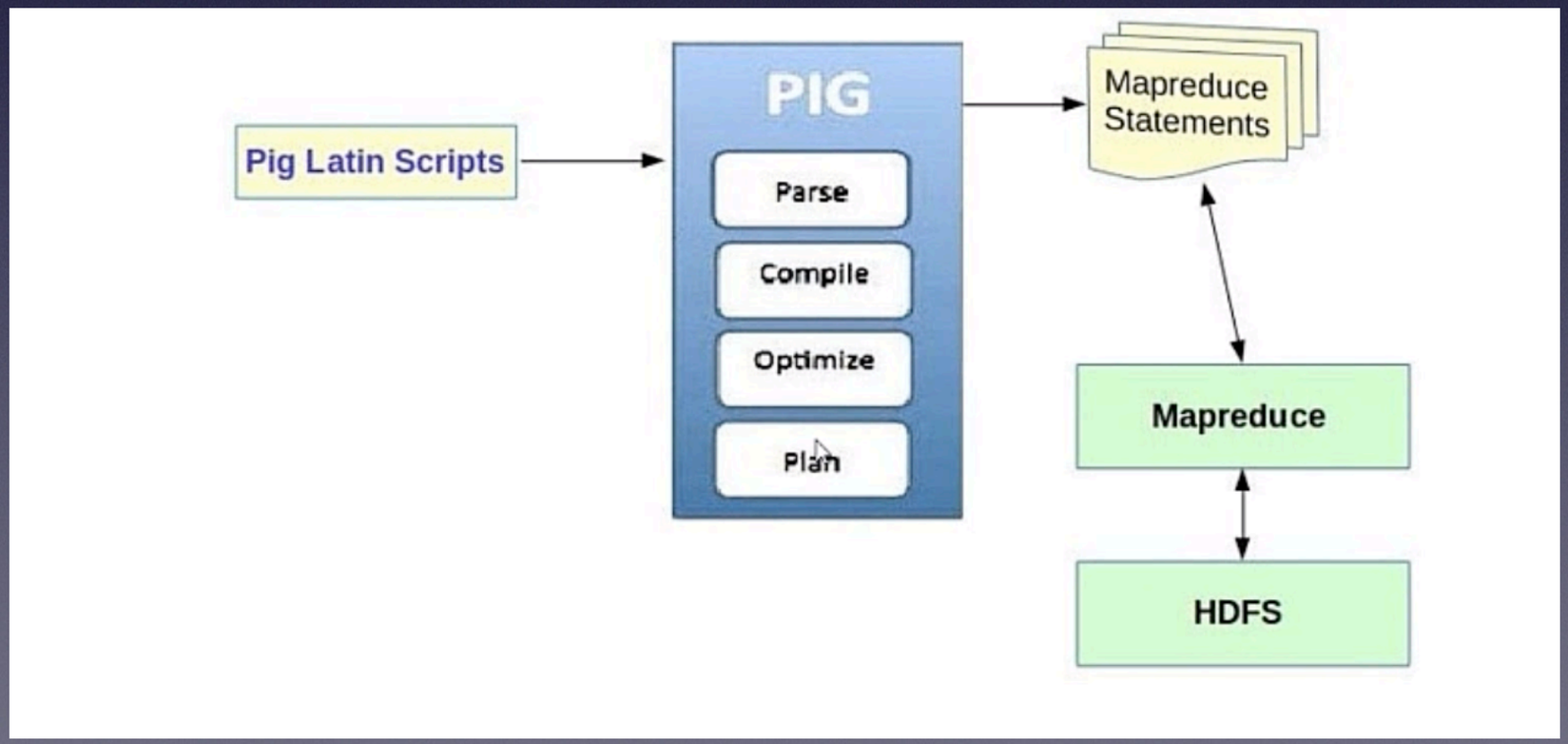
Oozie

- Workflow yönetimi & bunlarınların koordinasyonu
- Akış, “action node” lardan oluşur.
  - Start, End, Kill
- Akışları belli zamanlarda otomatik çalıştırır.
- Pig script, bash script, Hive Sorguları



Pig

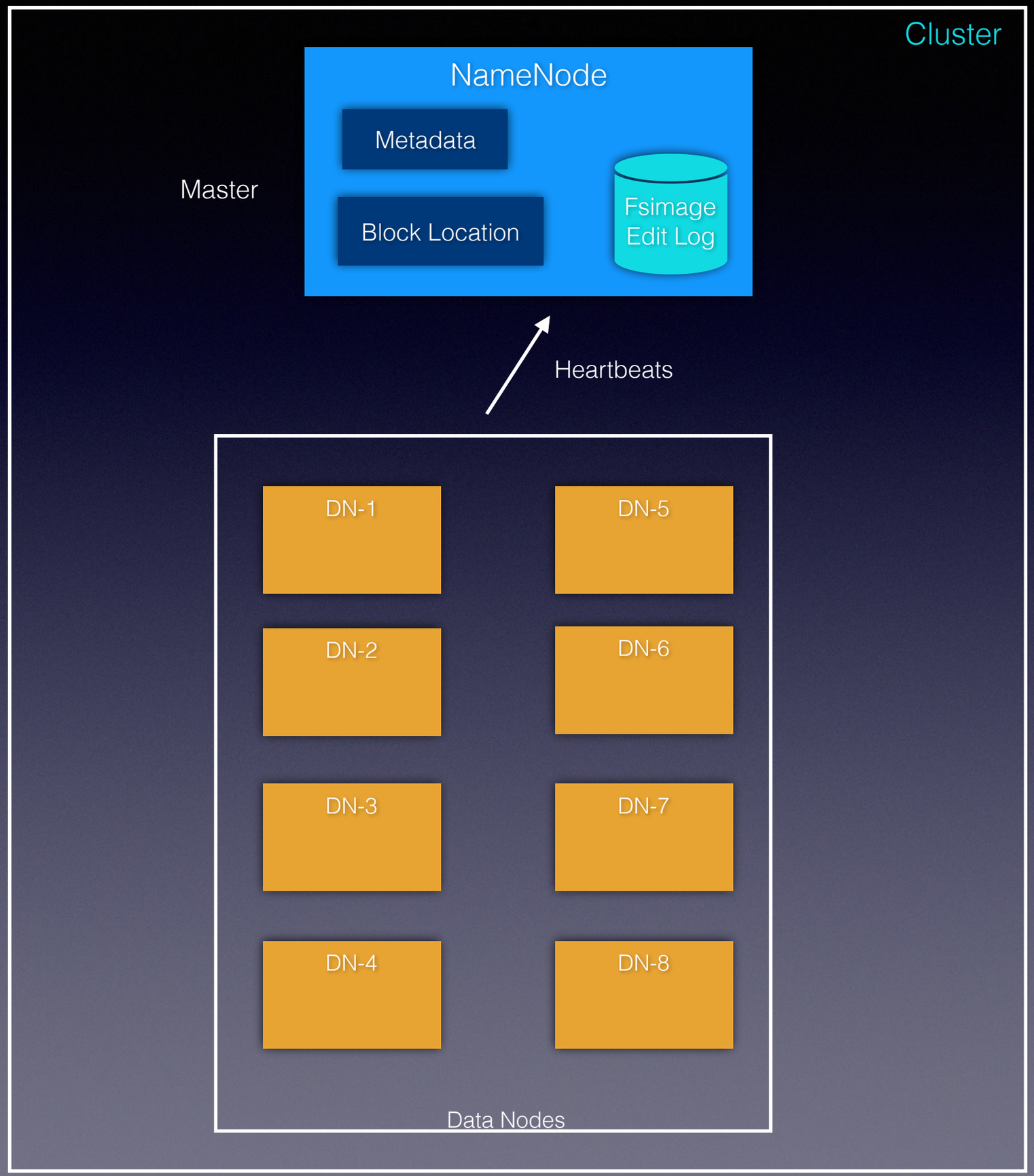
- Mapreduce ile veri analizi yapmamızı sağlar.
- Optimizasyon mekanizması vardır.
- Kendine özgü Pig Latin dili ile geliştirme yapılır.



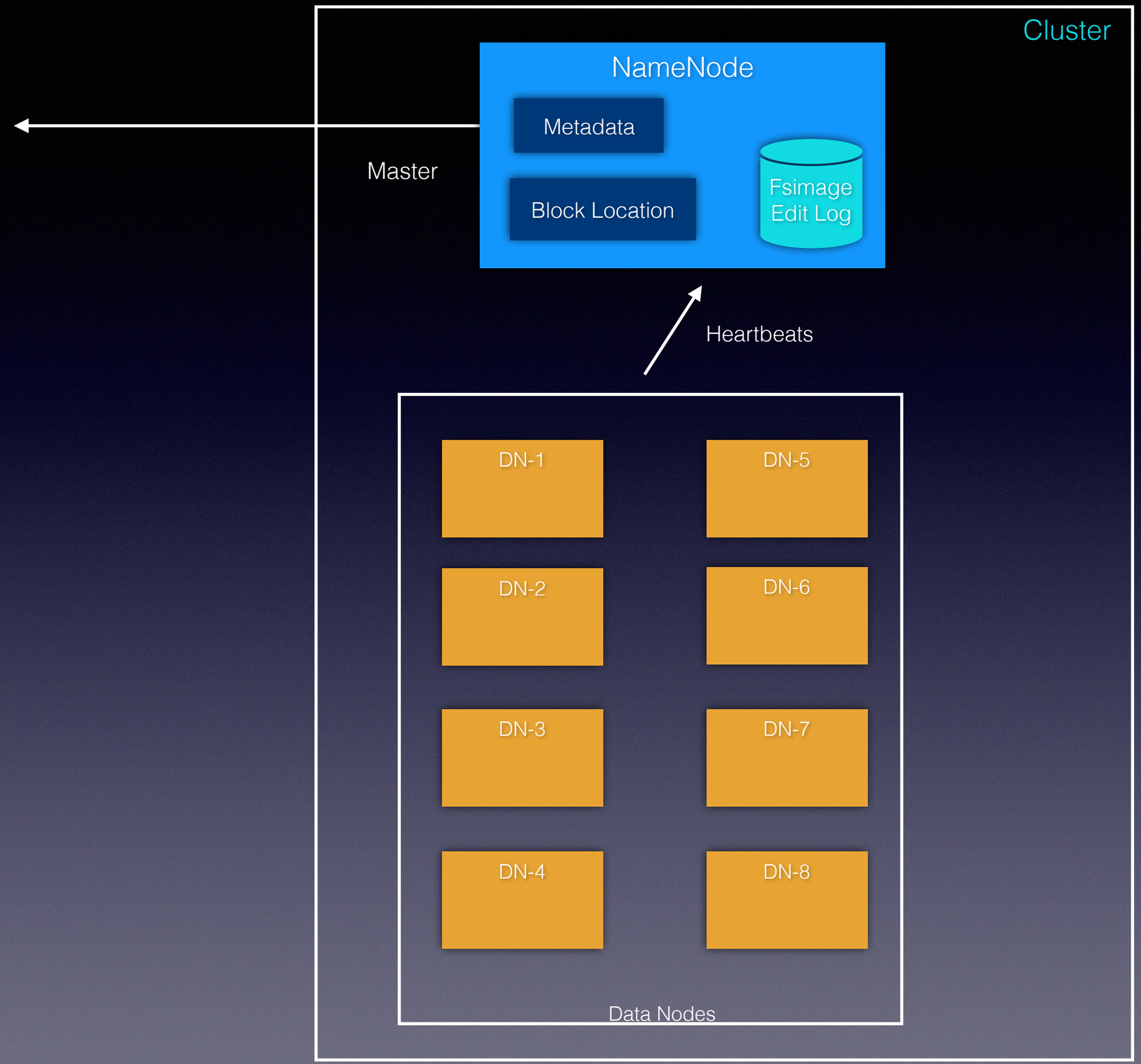
HDFS - Hadoop Distributed File System

- Veriyi **bloklar** halinde **dağıtık** olarak saklar.
- Master - Slave mimarisi ile çalışır.
  - Data Nodes
  - Name Nodes

HDFS Mimarisi



File name	log.txt
File size	300MB
No. of blocks	3
Block IDs	[#1, #2, #3]
User	hduser
Group	hadoop
Permission	-rw--r--r
Replication	3
Block size, etc	128MB, etc



HDFS Client

Access to file

Allow or Permission Denied

Master

Cluster

NameNode

Metadata

Block Location

Fsimage
Edit Log

Heartbeats

File name	log.txt
File size	300MB
No. of blocks	3
Block IDs	[#1, #2, #3]
User	hduser
Group	hadoop
Permission	-rw--r--r
Replication	3
Block size, etc	128MB, etc

DN-1

DN-5

DN-2

DN-6

DN-3

DN-7

DN-4

DN-8

Data Nodes

HDFS Client

Create file

Allow or Permission Denied

Cluster

Master

NameNode

Metadata

Block Location

Fsimage
Edit Log

Log.txt

Heartbeats

300MB

128MB

128MB

44MB

DN-1

DN-5

DN-2

DN-6

DN-3

DN-7

DN-4

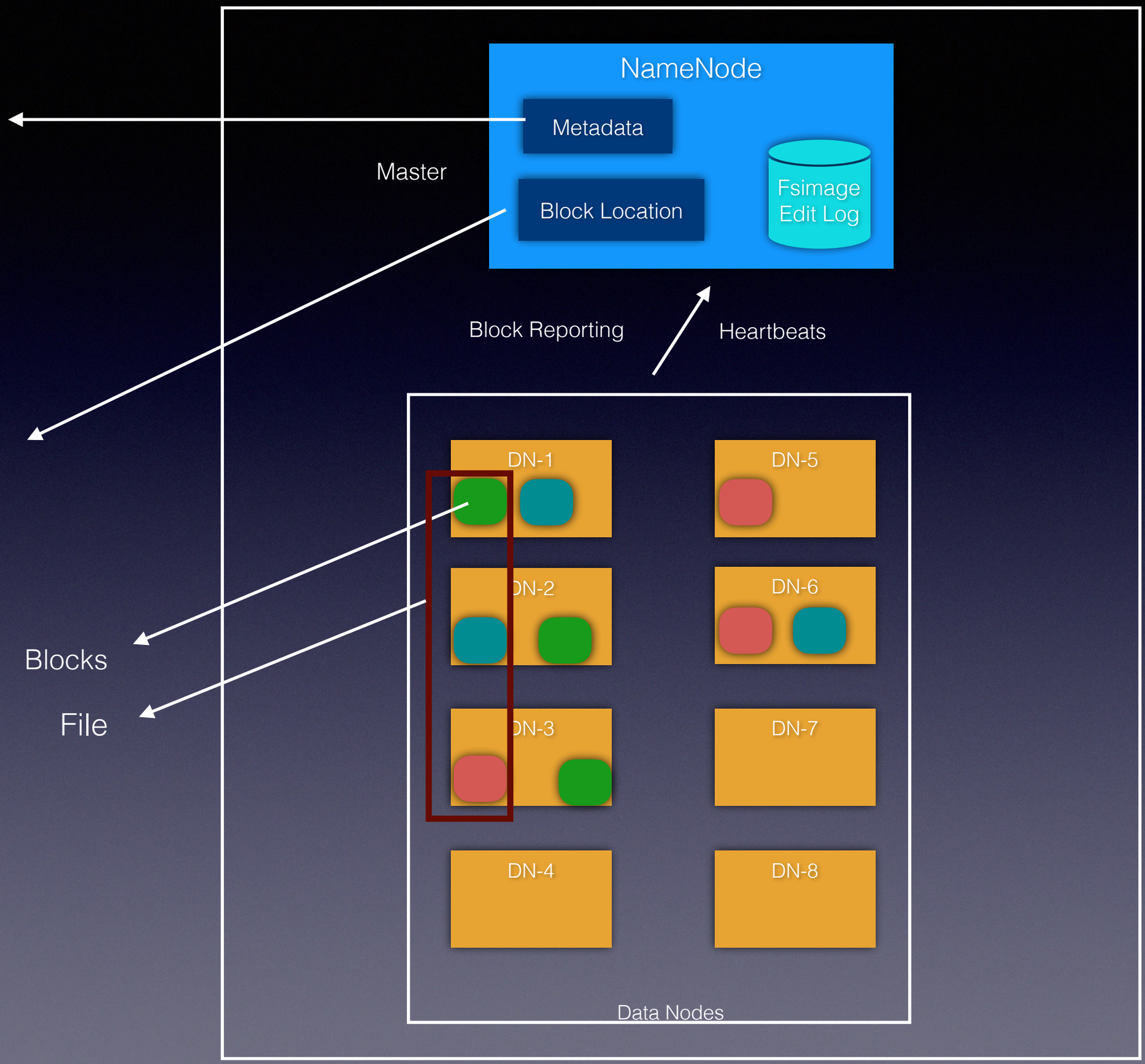
DN-8

Data Nodes

Replica Factor = 3

File name	log.txt
File size	300MB
No. of blocks	3
Block IDs	[#1, #2, #3]
User	hduser
Group	hadoop
Permission	-rw--r--r
Replication	3
Block size, etc	128MB, etc

Blok ID	Generation	Len
#1	T1	64MB
#2	T2	44MB
#3	T3	128MB

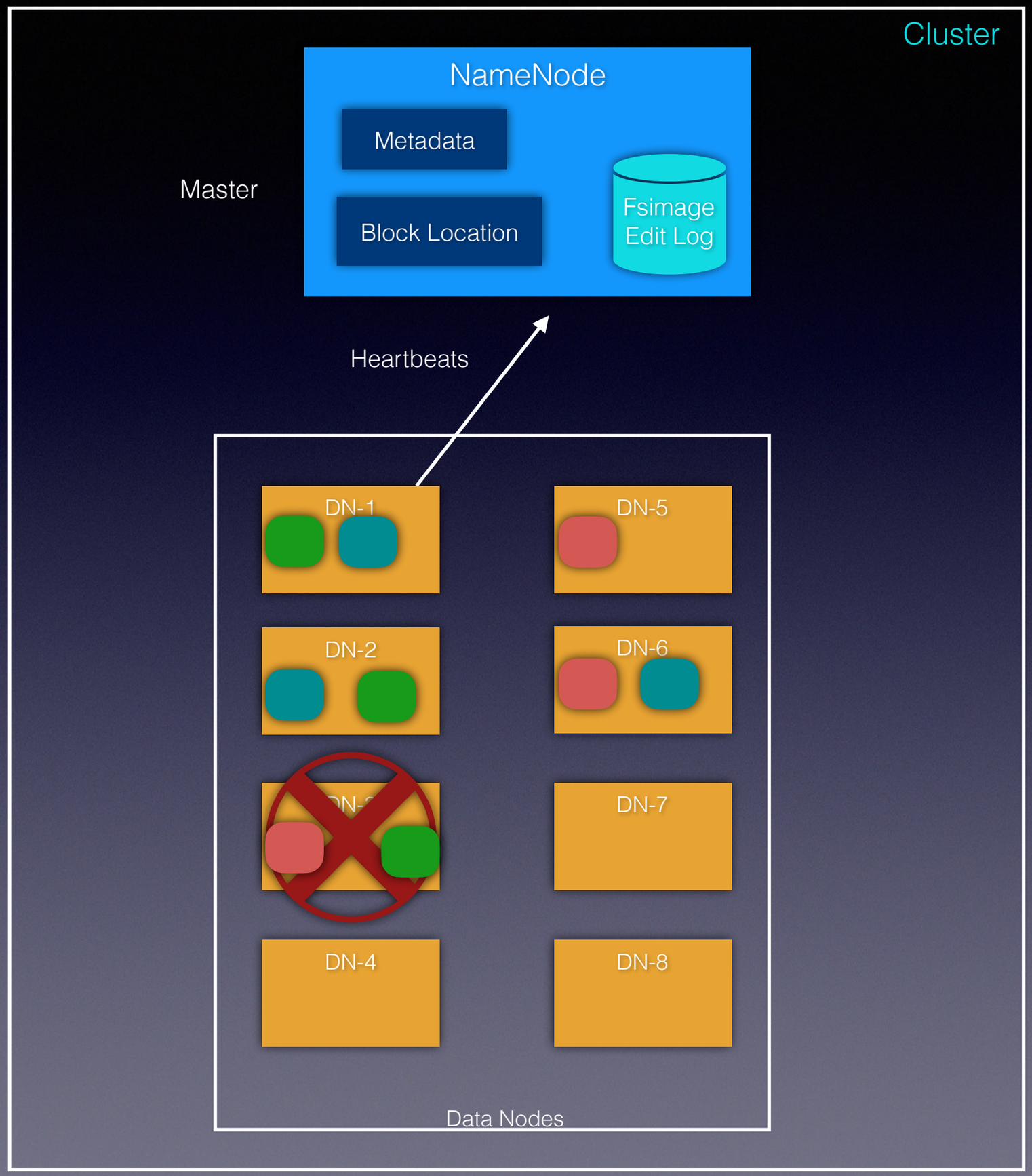


Heartbeats

Total Space
Used Space
Free Space
Data Transfer in Progress

LoadBalancing
Block Allocation

DN servis dışı kalırsa ?

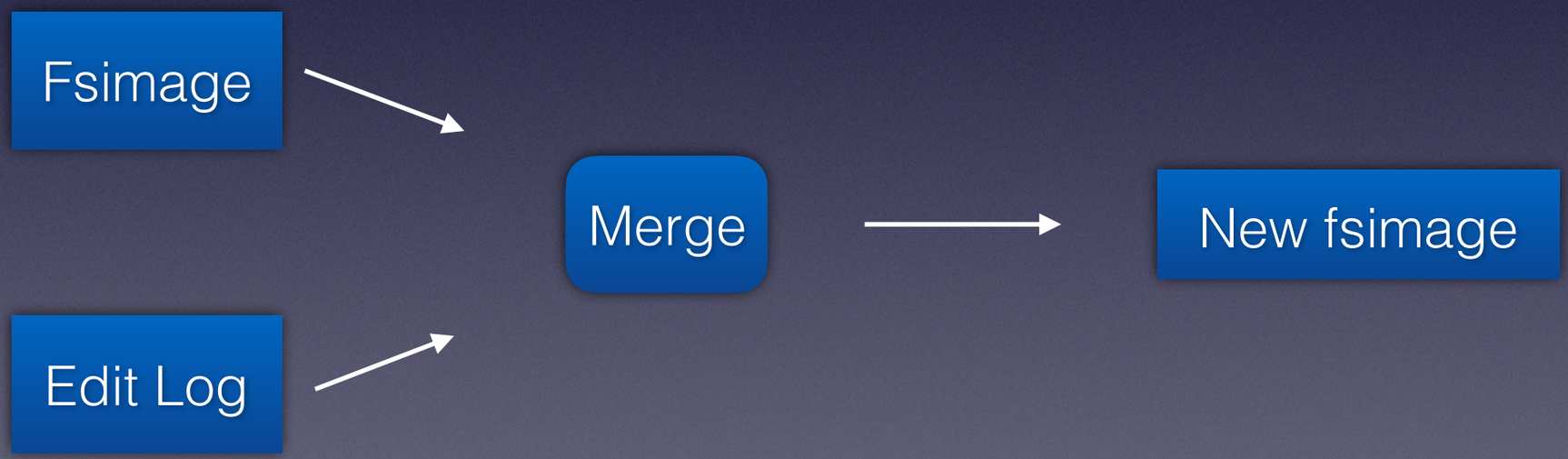


Replica Factor

- Replica faktör arttıkça;
  - Erişilebilirlik ve güvenilirlik artar.
  - Yazma ve network performansı azalır.

HDFS Check Pointing

- Secondary Name Node,
  - Name Node'un yerini **tutmaz.**
  - Aynı görevi yapmaz.



HDFS'i Kullanırken Nelere Dikkat Etmeliyiz

- Küçük boyutlu dosyalar yerine tek ve büyük boyutlu dosyalar halinde saklamaya
- Ham veriyi silmemeye
- Dosyaları kategorize edebileceğimiz şekilde tutmaya
    - /datawarehouse/tweets/2019-12-20

En önemlisi “MONITORING”

Nelerden Bahsettik

- Günlük hayatımızdaki sonuçlarından
- Büyük veri işleme modellerinden
- Teknolojilerinden

- Stream & Batch processing
- Hadoop Ekosistemi
- Kafka ve Kafka Streams

Büşra UMAN

Trendyol

Yazılım Geliştiricisi

Github: busrauman

E-mail: busrauman93@gmail.com