BTK AKADEMİ

Doç. Dr. Murat KARABATAK
Yazılım Müh. Öğretim Üyesi

Büyük Veri (Big Data)

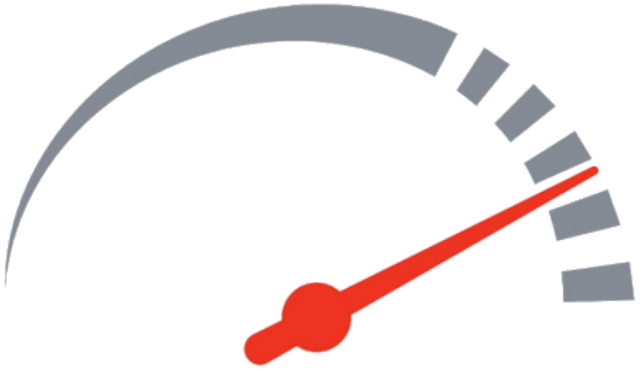
Büyük Veri (Big Data)

Gelişen teknolojiler ve artan veri kullanım oranı paralelinde daha fazla çeşitlilik içeren veri kümesidir.

Büyük Veri (Big Data)

Boyut

Hız



Çeşitlilik

Büyük Veri (Big Data)

- Sosyal medya hesapları,
- Arama motorları,
- Banka hesap hareketleri,
- Mailler ve kullanıcıların internet etkileşimleri,
- …

BIG DATA

Büyük Veri (Big Data)

- Birçok veri noktası karşılaştırılır, verilerin birbirleriyle olan ilişkileri ortaya çıkarılır ve bu ilişkiler öğrenmeyi, dolayısıyla daha akıllı kararlar alınmasını sağlar.

Büyük Verinin Kullanım Alanları

- Satıcılar,

- Şirketler,

- Araştırmacılar,

- Girişimciler.

Büyük Verinin Kullanım Alanları

- %50 daha fazla kazanç
- %41 daha etkili Pazar çalışması
- %37 reklam harcamasında azalma
- %37 sosyal medya kullanımında başarı

BTK AKADEMİ

Büyük Verinin Kullanım Alanları

- Sağlık hizmetlerinde hastalığın erken teşhisinde ve yeni ilaçların geliştirilmesinde kullanılabilir.

BTK AKADEMİ

Büyük Verinin Kullanım Alanları

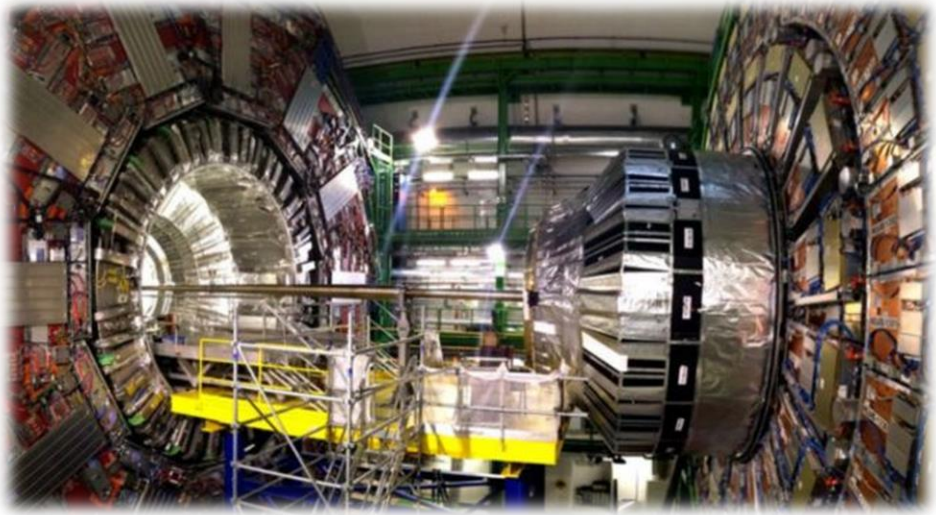
- Felaketleri tahmininde, veriler bir sonraki muhtemel deprem durumu tahmin etmek için analiz edilebilir.

Büyük Verinin Kullanım Alanları

- Güvenlik güçleri, mevcut kamu veri setlerine dayanan veri güdümlü stratejiler benimseyebilir.

Büyük Verinin Kullanım Alanları

Büyük Hadron Çarpıştırıcısı deneyi



- Saniyede 40 milyon kez veri,
- Yaklaşık 150 milyon sensor,
- Saniyede 600 milyon çarpışma.

BTK AKADEMİ

Büyük Verinin Kullanım Alanları

E-Ticaret Siteleri,

- Her gün milyonlarca arka uç işlemi,
- Yarım milyondan fazla satıcısından soruları,
- Dünyanın en büyük üç Linux veri tabanıdır.

BTK AKADEMİ

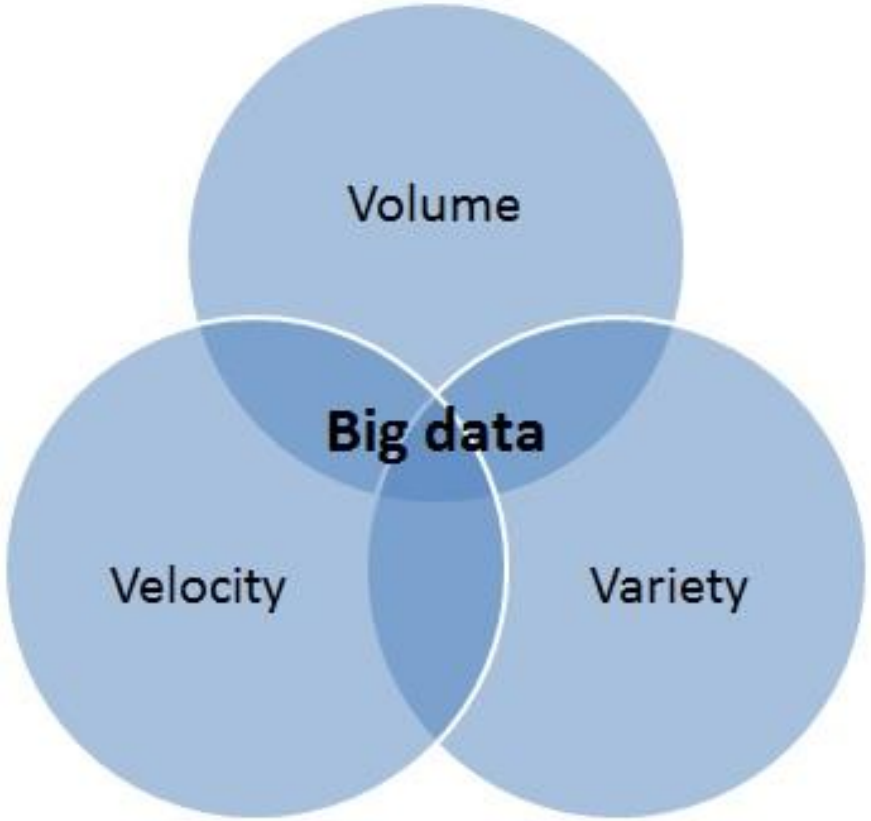
Büyük Verinin Kullanım Alanları

Sosyal Medya Siteleri,

- Her gün milyonlarca kullanıcı,
- Milyarlarca metin içerikli paylaşım,
- Milyarlarca resim ve video.

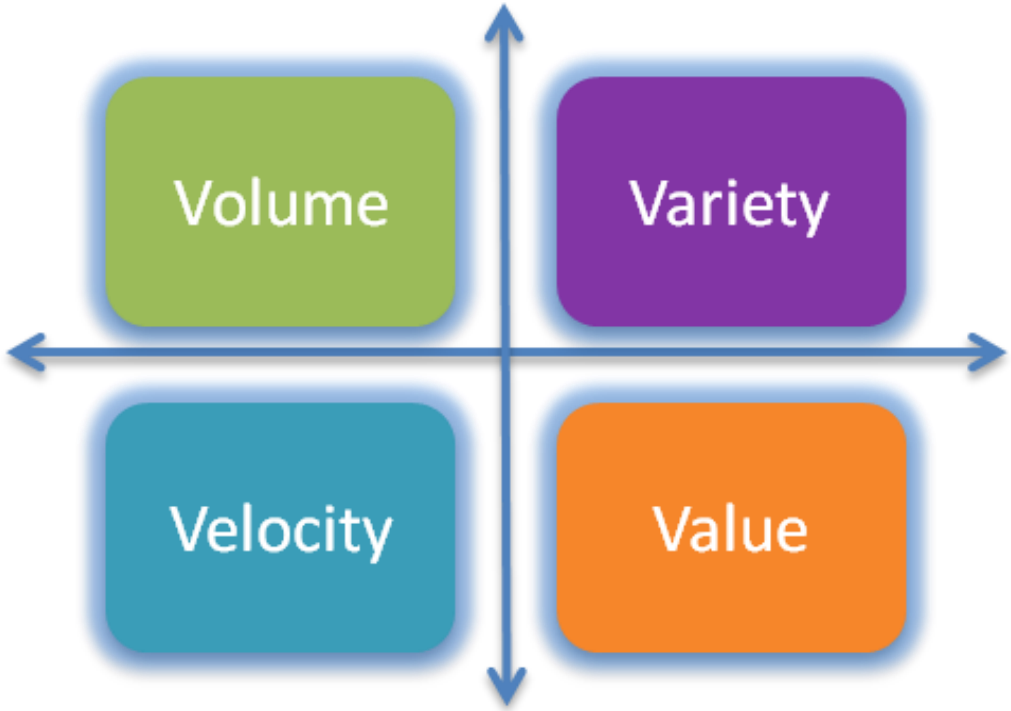
Büyük Veri Kavramı

- Büyük veri için genellikle 3V kavramı kullanılmaktadır.



Büyük Veri Kavramı

- Büyük veri için 4V kavramı



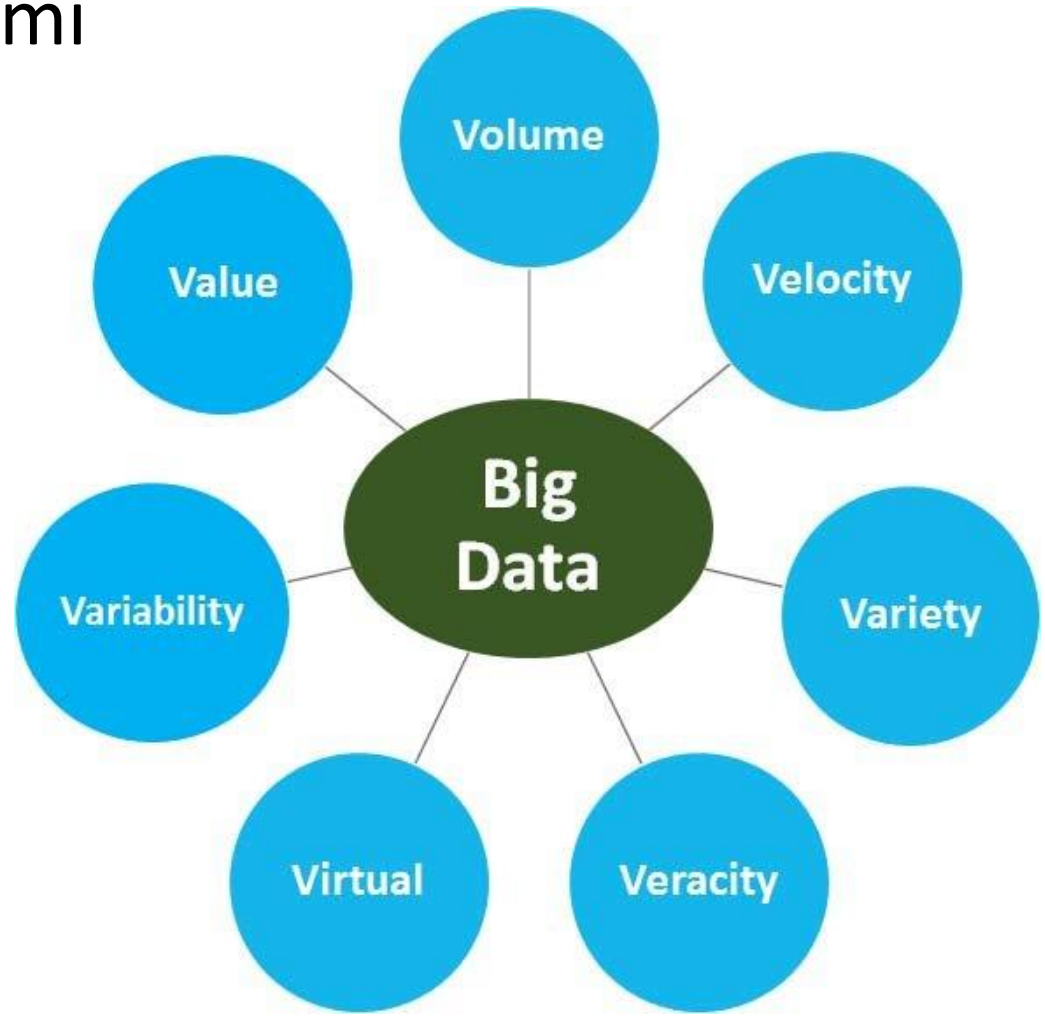
Büyük Veri Kavramı

- Büyük veri için 5V kavramı

BTK AKADEMİ

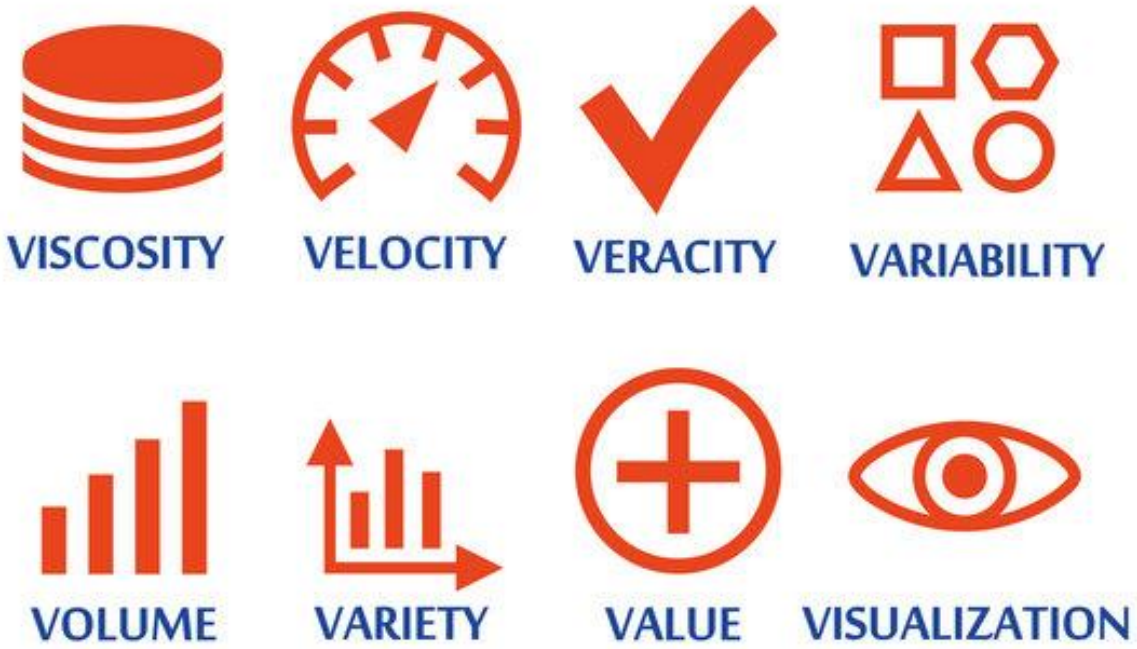
Büyük Veri Kavramı

- Büyük veri için 7V’ kavramı



Büyük Veri Kavramı

- Volume (Hacim)
- Variety (Çeşitlilik)
- Velocity (Hız)
- Value (Değer)
- Veracity (Doğruluk)
- Visualisation (Görselleştirme)
- Viscosity (Viskozite)
- Virality (Virallik)



BTK AKADEMİ

Volume (Hacim)

- Terabayttan petabaytlara kadar olan verinin boyutudur.

Variety (Çeşitlilik)

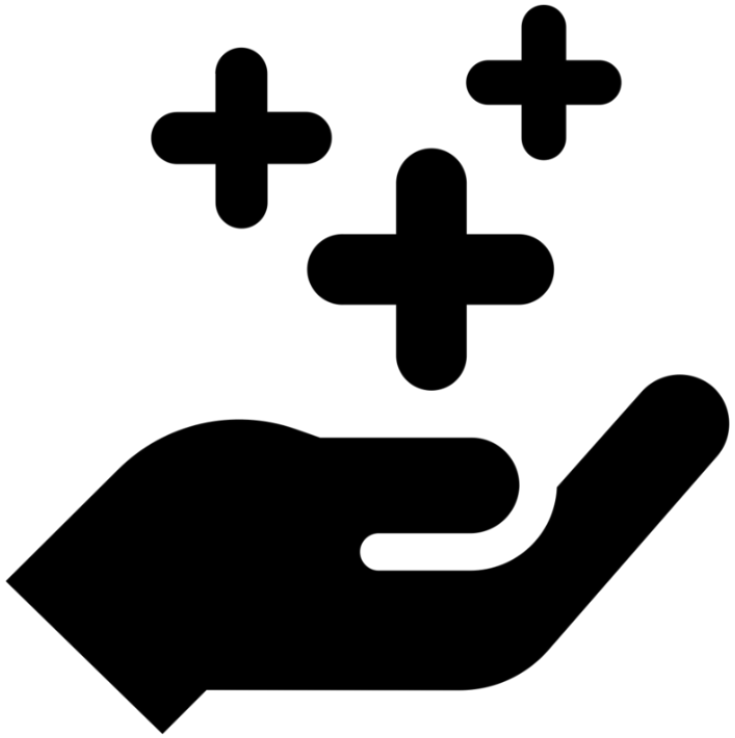
- Çeşitli kaynaklardan farklı verileri içerir. Örneğin sosyal medya etkileşimleri, finansal işlemler.

Velocity (Hız)

- İşletmelerin zaman verilerinin üretilmesinden, kullanıcılara uygulanabilecek zamana kadar Kaynak ve hedef arasındaki veri aktarım hızı

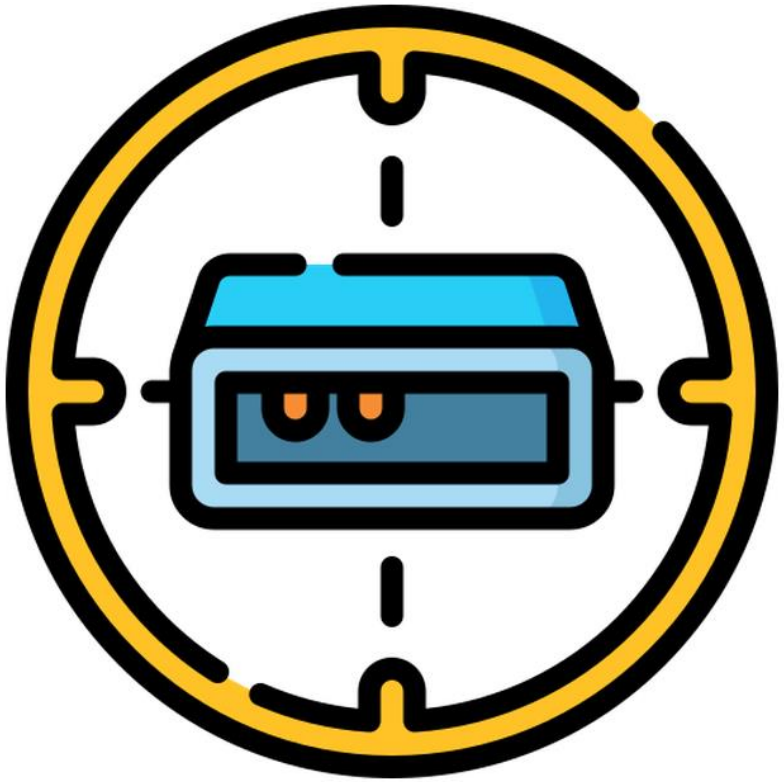
Value (Değer)

- Herhangi bir veri setinden elde edilecek olan ekonomik ve sosyal zenginliği ifade eder.



Veracity (Doğruluk)

- Verilerin eksiksizliğini ve doğruluğunu ifade eder.

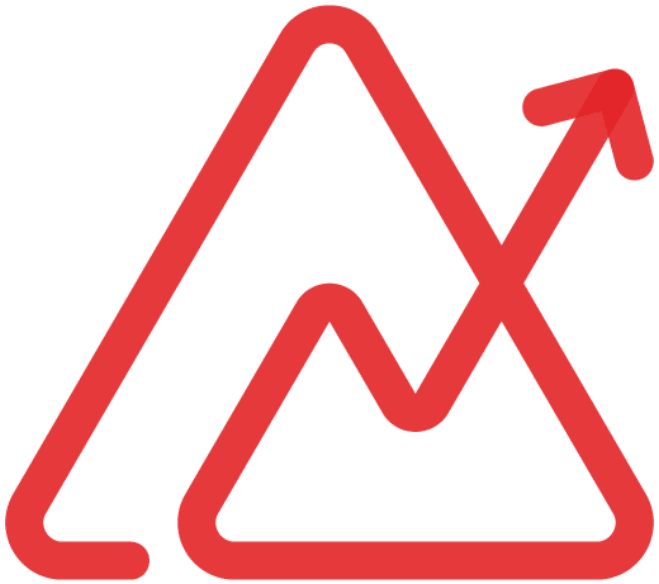


Visualisation (Görselleştirme)

- Görselleştirme sayesinde veriler, daha anlamlı ve etkili hale gelir. Karmaşık veriler daha ulaşılabilir, anlamlı ve kullanışlı olur.

Viscosity (Viskozite)

- Veri kaynaklarının çeşitliliği, veri akış hızı ve işlem karmaşıklıklarına karşı gösterilen direnci ifade etmektedir.



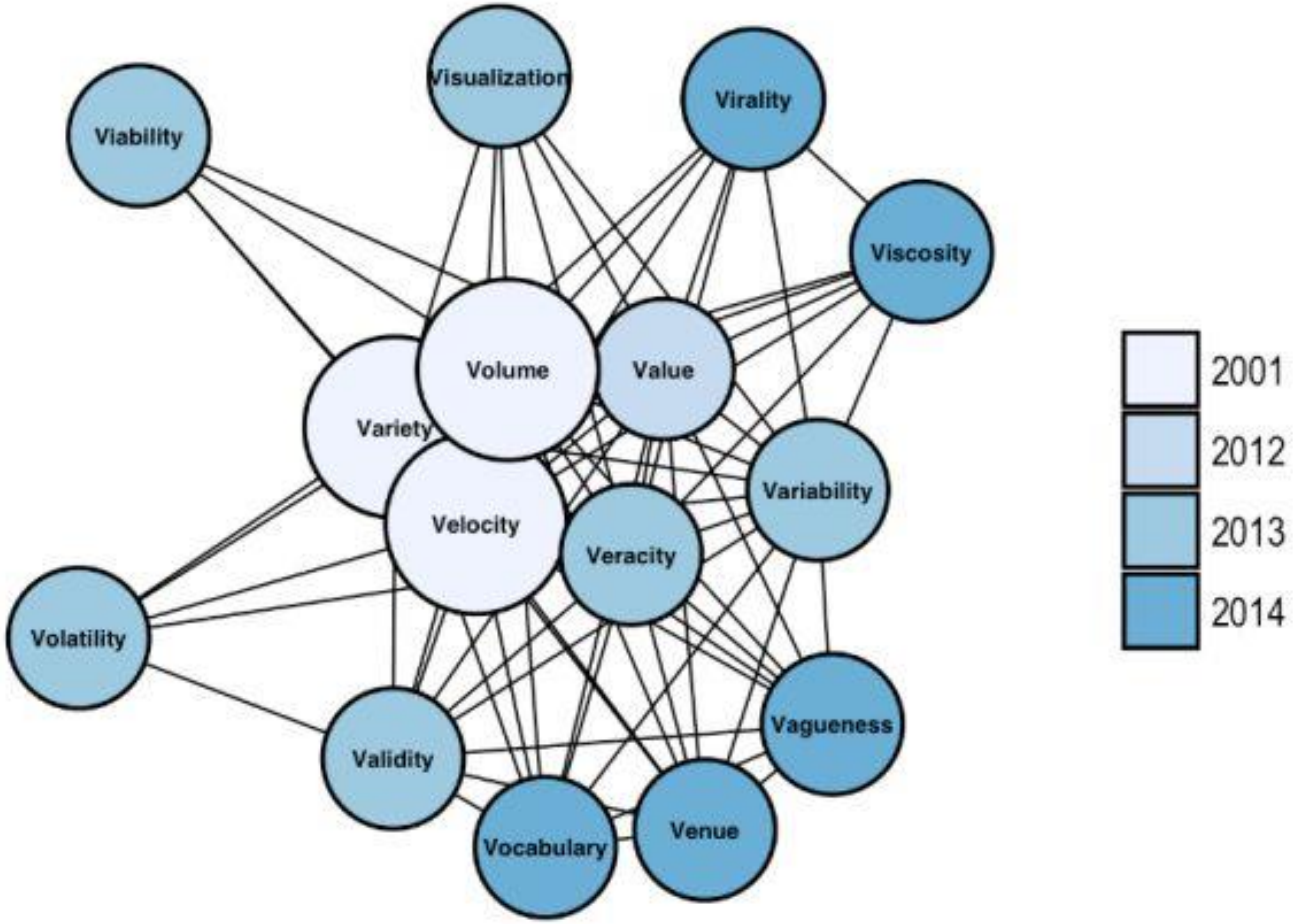
Virality (Virallik)

- Verilerin bir ağ üzerinden yayılabileceği hızı ölçer.
- Zaman, yayılma oranıyla birlikte belirleyici bir faktördür.

BTK AKADEMİ

Büyük Veri Kavramı

- Büyük veri için 15V kavramı



Büyük Veri Kavramı (42V)

Vagueness
Vane
Variety
Vastness
Veil
Verdict
Viability
Virtuosity
Vivify
Volatility
Vulpine
Validity
Vanilla
Varifocal
Vaticination
Velocity
Versed
Vibrant
Viscosity
Vocabulary
Volume
Visualization
Valor
Vantage
Varmint
Vault
Venue
Version Control
Victual
Visibility
Vogue
Voodoo
Value
Variability
Varnish
Veer
Veracity
Vet
Viral
Vexed
Voice
Voyage

BTK AKADEMİ

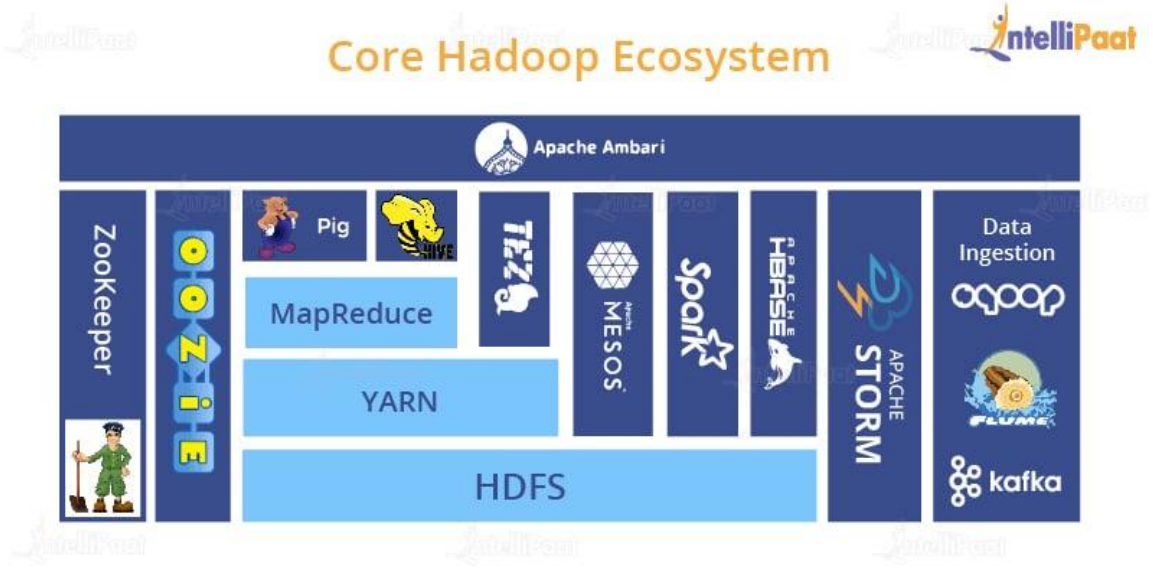
Büyük Veri Araçları

- Apache Hadoop
- Apache Spark
- Rapid Miner
- Cassandra
- Xplenty

Apache Hadoop

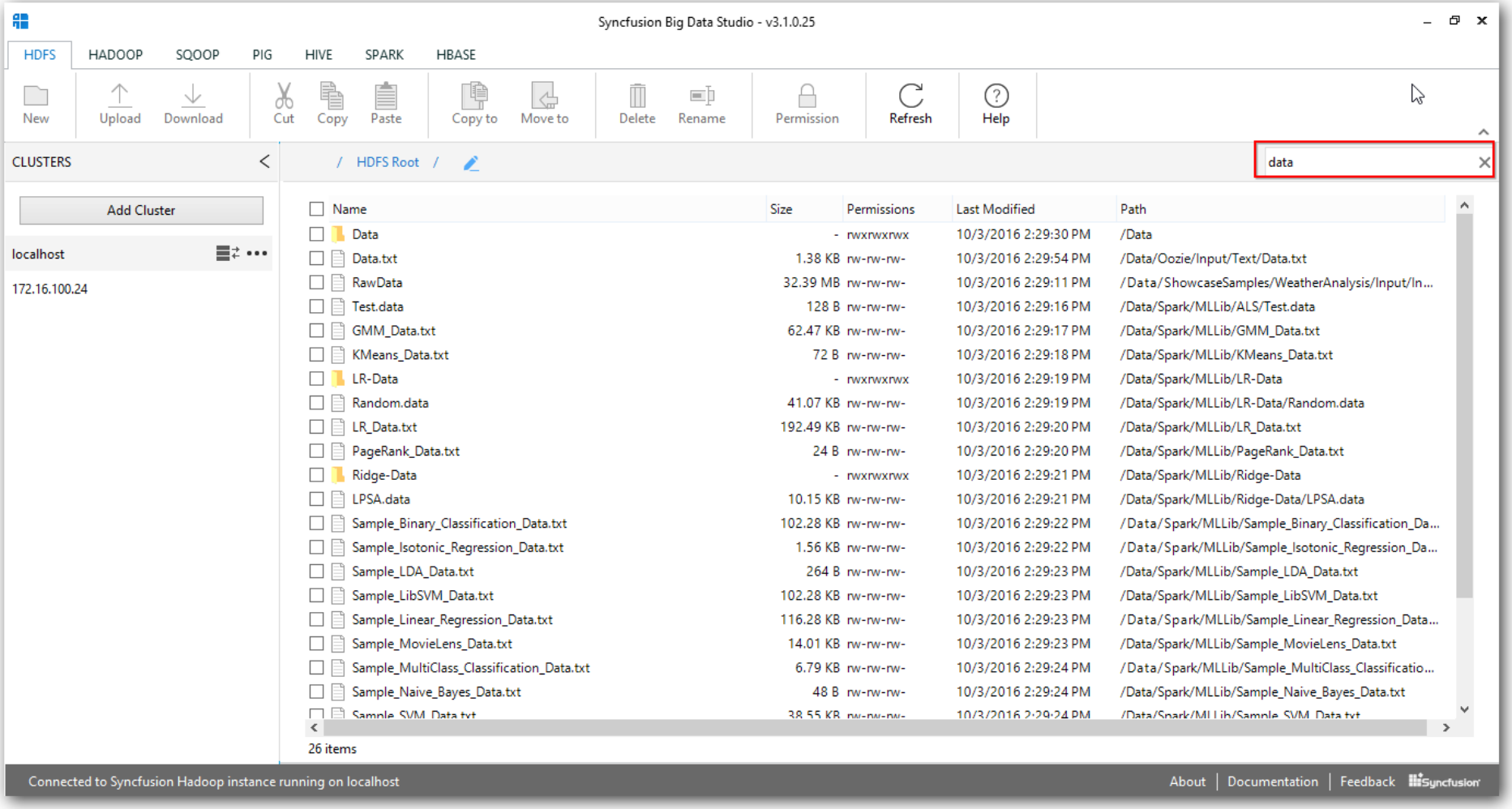
Açık kaynak kodlu, güvenilir, ölçeklenebilir paralel hesaplama yazılımı projesidir.

- Bulut Altyapısı
- Kütüphane Desteği
- MapReduce
- HDFS
- Verimli ve Esnek



BTK AKADEMİ

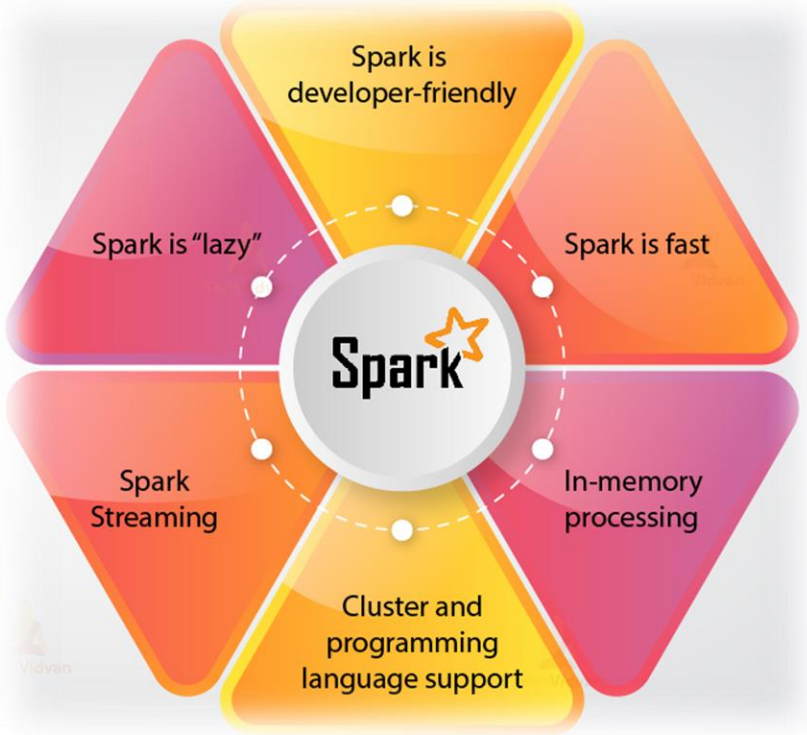
Apache Hadoop



Apache Spark

Apache Spark, hem gerçek zamanlı verileri hem de toplu verileri işleyen açık kaynaklı bir araçtır.

- Yüksek Akışlı İşlem
- Hızlı Grafik İşleme
- Bağımsız Küme Mod
- Birleşmiş Kütüphane
- Bulut yayımı



Apache Spark

Spark 2.0.2 | Jobs | Stages | Storage | Environment | Executors | SQL | Spark Pi application UI

Executors

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write
Active(2)	2	2.3 KB / 868.8 MB	0.0 B	15	15	0	4601	4616	7.6 m (12.7 s)	0.0 B	0.0 B	0.0 B
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B
Total(2)	2	2.3 KB / 868.8 MB	0.0 B	15	15	0	4601	4616	7.6 m (12.7 s)	0.0 B	0.0 B	0.0 B

Executors

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs	Thread Dump
0	xx.xx.xx.xx:1093	Active	1	1168.0 B / 434.4 MB	0.0 B	15	15	0	4601	4616	7.6 m (12.7 s)	0.0 B	0.0 B	0.0 B	stdout stderr	Thread Dump
driver	xx.xx.xx.xx:1089	Active	1	1168.0 B / 434.4 MB	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B		Thread Dump

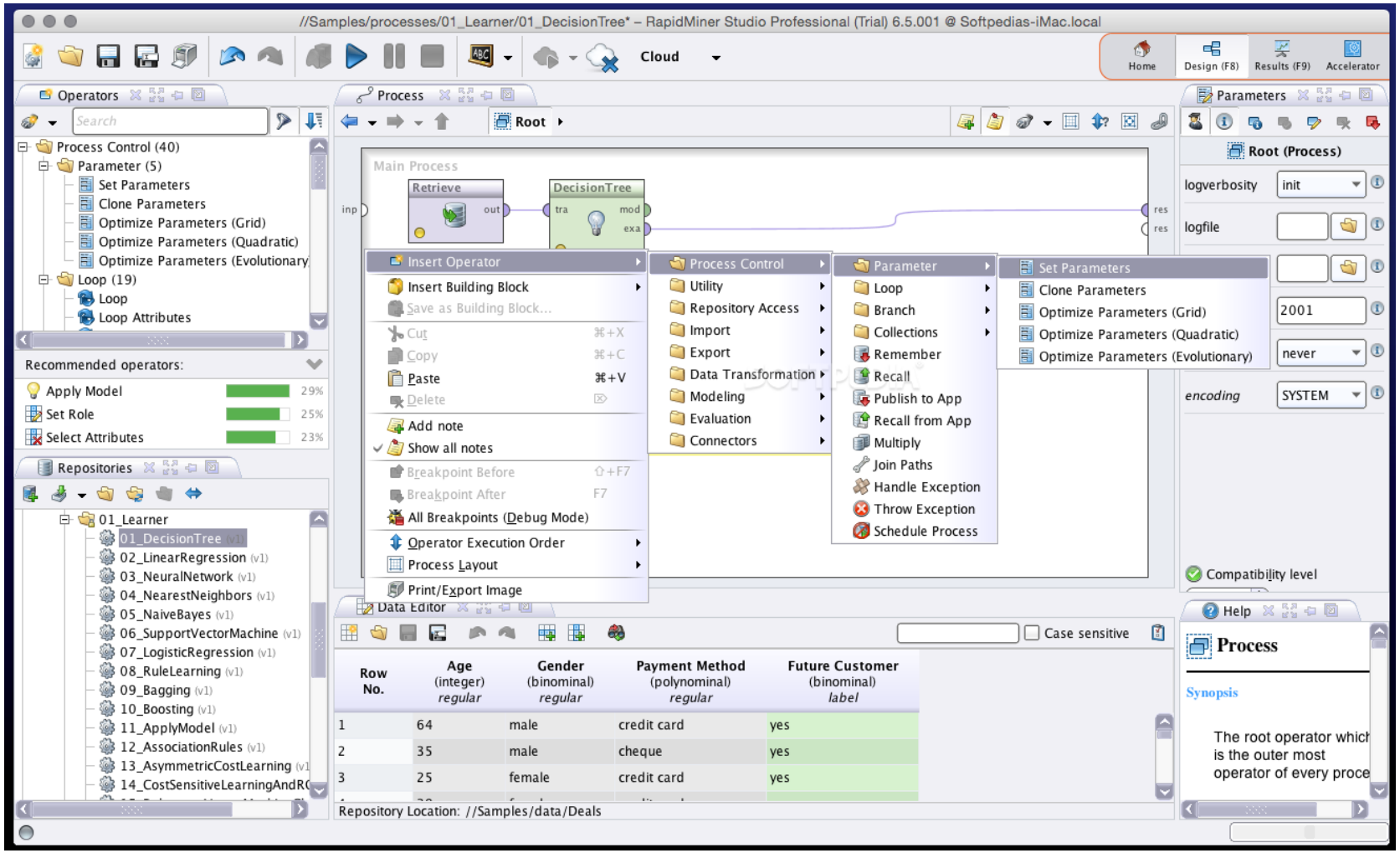
BTK AKADEMİ

Rapid Miner

Açık kaynaklı, çapraz platform büyük veri aracıdır. Tahmine dayalı analitik ve makine öğrenimi teknolojisi içerir.

- Bulut Entegrasyonu
- Şirket içi veri tabanları ile entegrasyon
- Tahmine dayalı modeller oluşturma
- İstemci-sunucu modelinin desteği

Rapid Miner

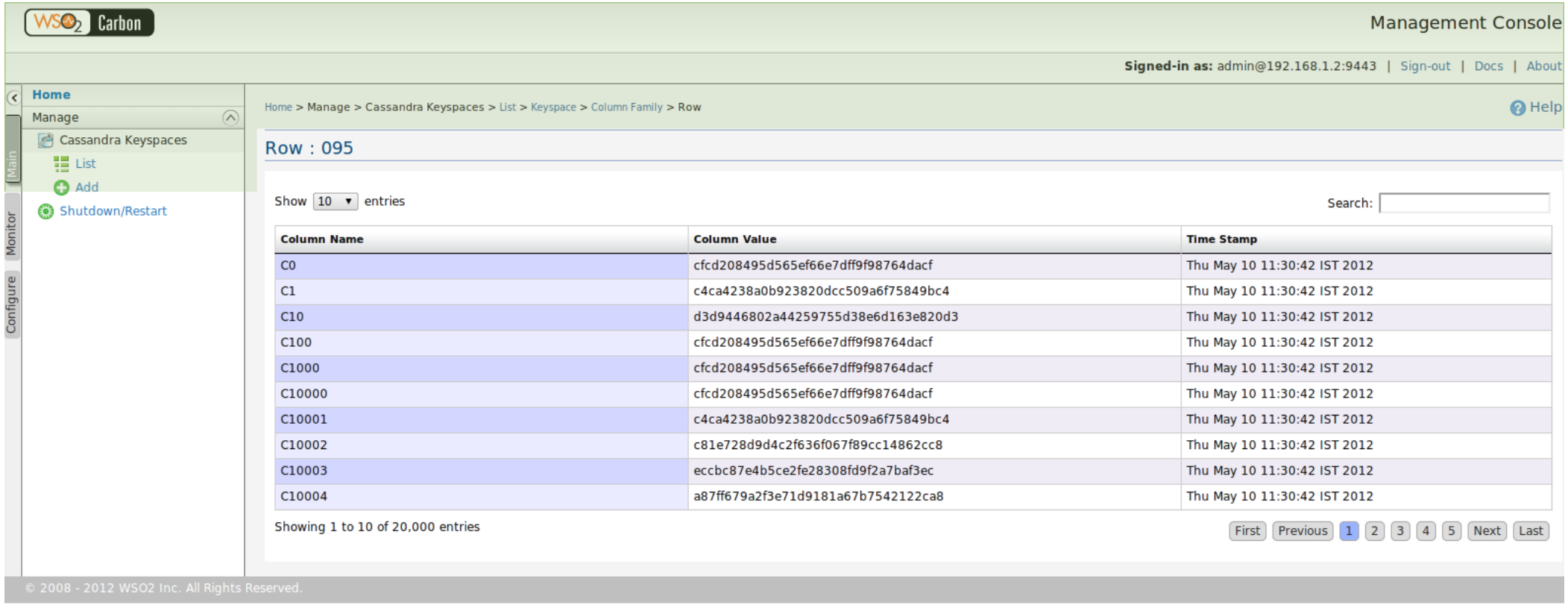

//Samples/processes/01_Learner/01_DecisionTree* – RapidMiner Studio Professional (Trial) 6.5.001 @ Softpedias-iMac.local
Cloud
Home
Design (F8)
Results (F9)
Accelerator
Operators
Process
Parameters
Root
Main Process
Retrieve
DecisionTree
Process Control (40)
Parameter (5)
Set Parameters
Clone Parameters
Optimize Parameters (Grid)
Optimize Parameters (Quadratic)
Optimize Parameters (Evolutionary)
Loop (19)
Loop
Loop Attributes
Recommended operators:
Apply Model 29%
Set Role 25%
Select Attributes 23%
Repositories
01_Learner
01_DecisionTree
02_LinearRegression
03_NeuralNetwork
04_NearestNeighbors
05_NaiveBayes
06_SupportVectorMachine
07_LogisticRegression
08_RuleLearning
09_Bagging
10_Boosting
11_ApplyModel
12_AssociationRules
13_AsymmetricCostLearning
14_CostSensitiveLearningAndR
Insert Operator
Insert Building Block
Save as Building Block...
Cut
Copy
Paste
Delete
Add note
Show all notes
Breakpoint Before
Breakpoint After
All Breakpoints (Debug Mode)
Operator Execution Order
Process Layout
Print/Export Image
Process Control
Utility
Repository Access
Import
Export
Data Transformation
Modeling
Evaluation
Connectors
Parameter
Loop
Branch
Collections
Remember
Recall
Publish to App
Recall from App
Multiply
Join Paths
Handle Exception
Throw Exception
Schedule Process
Set Parameters
Clone Parameters
Optimize Parameters (Grid)
Optimize Parameters (Quadratic)
Optimize Parameters (Evolutionary)
Root (Process)
logverbosity init
logfile
2001
never
encoding SYSTEM
Compatibility level
Data Editor
Case sensitive
Row No. | Age (integer) regular | Gender (binominal) regular | Payment Method (polynominal) regular | Future Customer (binominal) label
1 | 64 | male | credit card | yes
2 | 35 | male | cheque | yes
3 | 25 | female | credit card | yes
Help
Process
Synopsis
The root operator which is the outer most operator of every proce
Repository Location: //Samples/data/Deals

Cassandra

Apache Software Foundation tarafından 2008 yılında oluşturulmuş ücretsiz ve açık kaynaklı bir veritabanı yönetim aracıdır.

- Hızlı Veri İşleme,
- Bulut Kullanılabilirliği,
- Kolay veri dağılımı,
- Ölçeklenebilir performans,
- Basit ve Verimli.

Cassandra



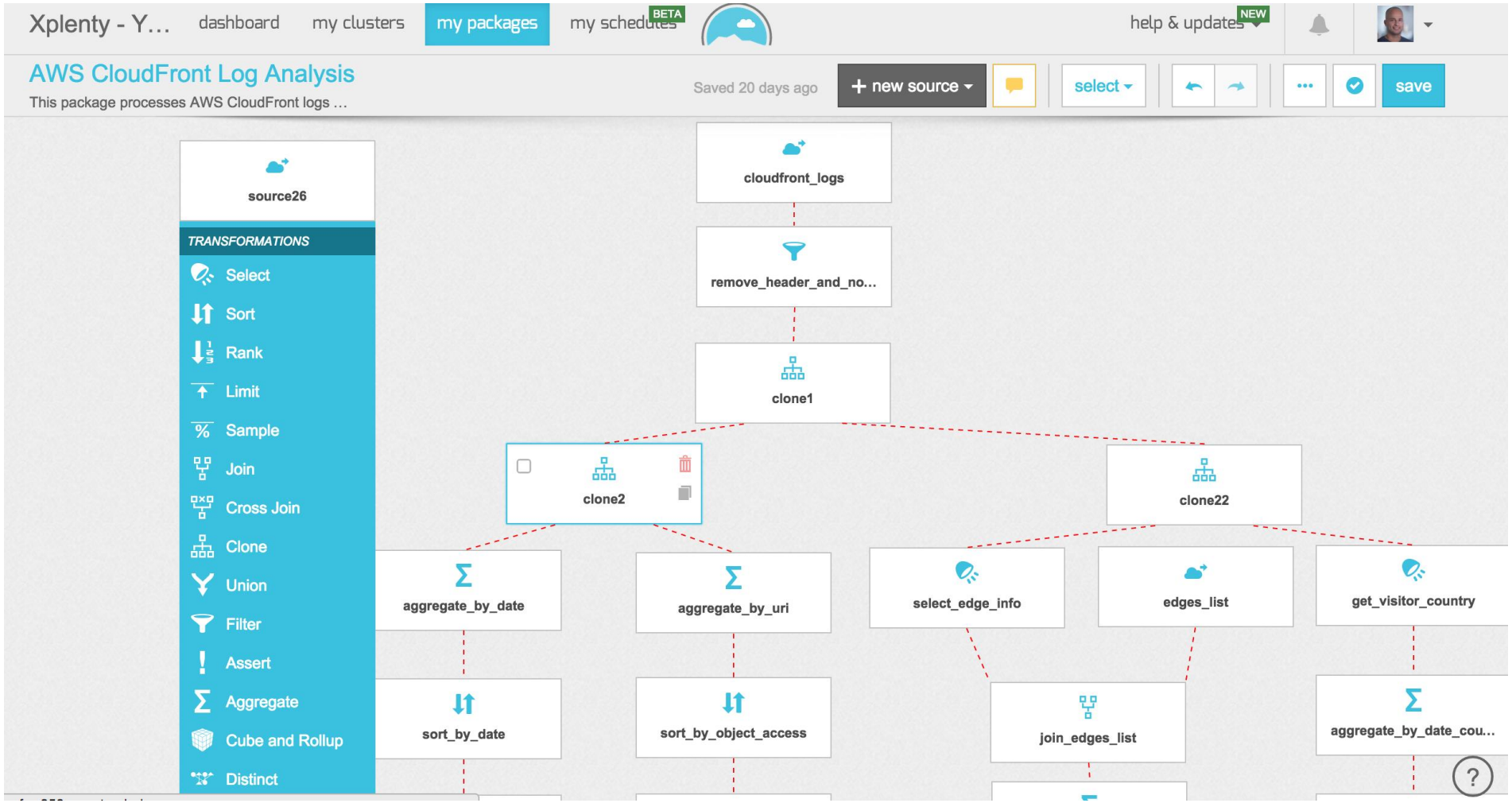
Column Name	Column Value	Time Stamp
C0	cfcd208495d565ef66e7dff9f98764dacf	Thu May 10 11:30:42 IST 2012
C1	c4ca4238a0b923820dcc509a6f75849bc4	Thu May 10 11:30:42 IST 2012
C10	d3d9446802a44259755d38e6d163e820d3	Thu May 10 11:30:42 IST 2012
C100	cfcd208495d565ef66e7dff9f98764dacf	Thu May 10 11:30:42 IST 2012
C1000	cfcd208495d565ef66e7dff9f98764dacf	Thu May 10 11:30:42 IST 2012
C10000	cfcd208495d565ef66e7dff9f98764dacf	Thu May 10 11:30:42 IST 2012
C10001	c4ca4238a0b923820dcc509a6f75849bc4	Thu May 10 11:30:42 IST 2012
C10002	c81e728d9d4c2f636f067f89cc14862cc8	Thu May 10 11:30:42 IST 2012
C10003	eccbc87e4b5ce2fe28308fd9f2a7baf3ec	Thu May 10 11:30:42 IST 2012
C10004	a87ff679a2f3e71d9181a67b7542122ca8	Thu May 10 11:30:42 IST 2012

Xplenty

Veri hazırlamak ve işlemek için kullanılan bulut tabanlı bir veri analiz bir platformdur.

- Donanım ve Yazılıma bağımlı değil,
- Esnek ve ölçeklenebilir,
- Karmaşık veriler hazırlanabilir,
- API bileşeni sunar.

Xplenty


Xplenty - Y...
dashboard
my clusters
my packages
my schedules
BETA
help & updates
NEW
AWS CloudFront Log Analysis
This package processes AWS CloudFront logs ...
Saved 20 days ago
+ new source
select
save
source26
TRANSFORMATIONS
Select
Sort
Rank
Limit
Sample
Join
Cross Join
Clone
Union
Filter
Assert
Aggregate
Cube and Rollup
Distinct
cloudfront_logs
remove_header_and_no...
clone1
clone2
clone22
aggregate_by_date
aggregate_by_uri
select_edge_info
edges_list
get_visitor_country
sort_by_date
sort_by_object_access
join_edges_list
aggregate_by_date_cou...

Diğer Büyük Veri Araçları

cloudera
Ask Bigger Questions

Open for Innovation
KNIME

mongoDB.

Datawrapper

lumify

Qubole
ELEMENTAL TO BIG DATA

talend

tableau

BTK AKADEMİ