



IBM Big Data Forum - Think Big

Büyük Veri Perdesini Aralıyoruz

Veri Ambarı Yetkinliklerinizi Büyük Veri ile Genişletin

Cüneyt Göksu, VBT

IBM Gold Consultant

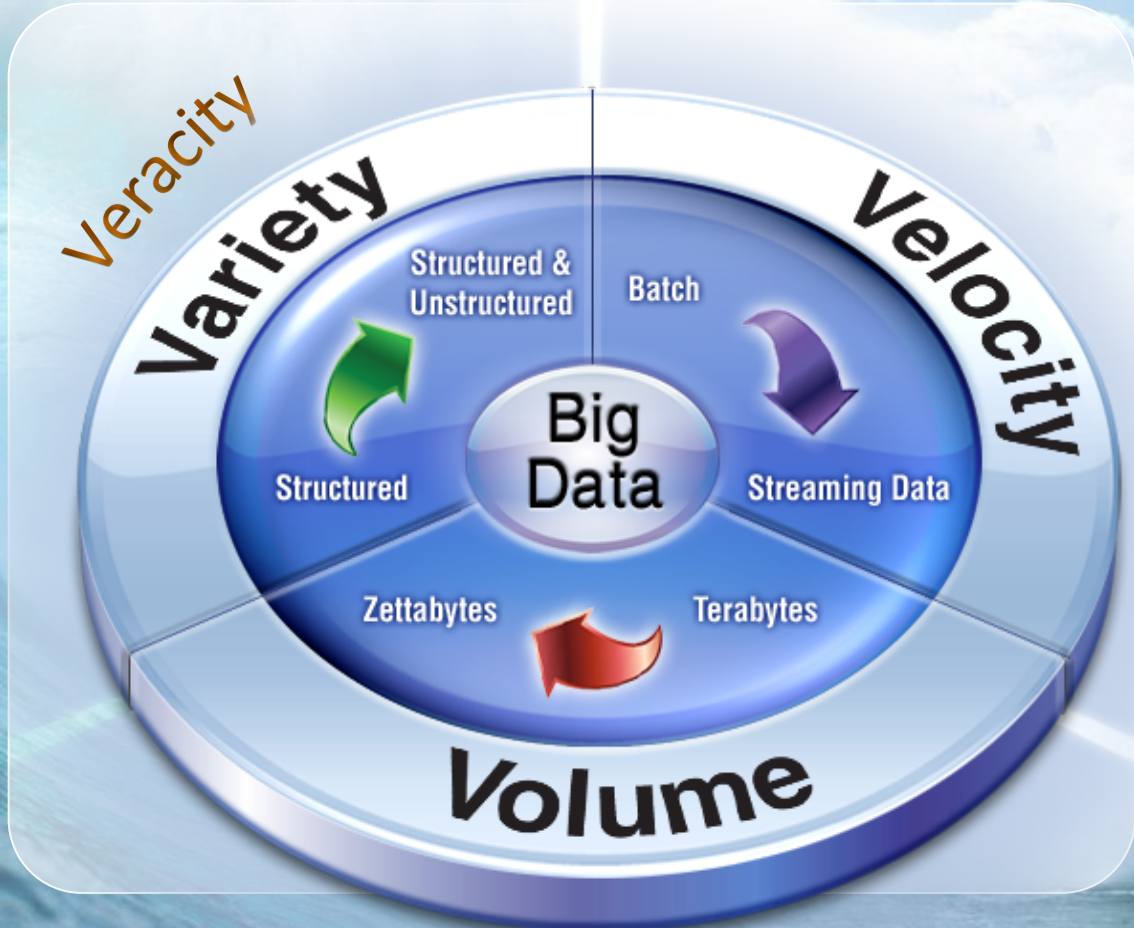
IBM Champion for Data Management

Ayhan Önder, IBM

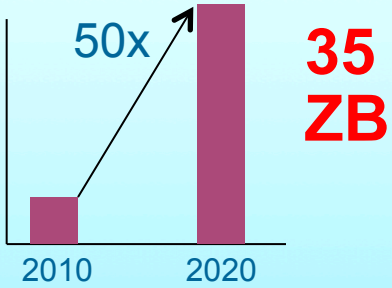
BigData Technical Specialist, ayhano@tr.ibm.com

Big Data

Daha önce analizi mümkün olmayan çok büyük, çok çeşitli ve akışkan veriler üzerinden bilgiye erişim.



Volume



Velocity



30 Billion
RFID
sensors and
counting

Variety



80% of the
world's data is
unstructured



Veracity

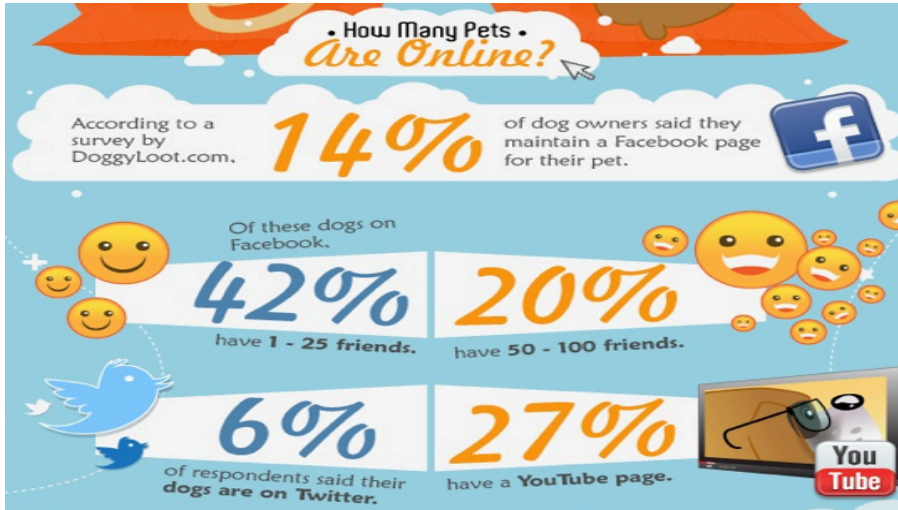
Yöneticilerin 1/3'ü karar vermek için kullandıkları verilerin doğruluğuna güvenmiyor



Value

$3 V + V = (V)alue$ (Değer)

"Makine ve sensörler tarafından üretilen veriler, sosyal medya tarafından üretilen verilerden en az 10 kat daha fazla...."

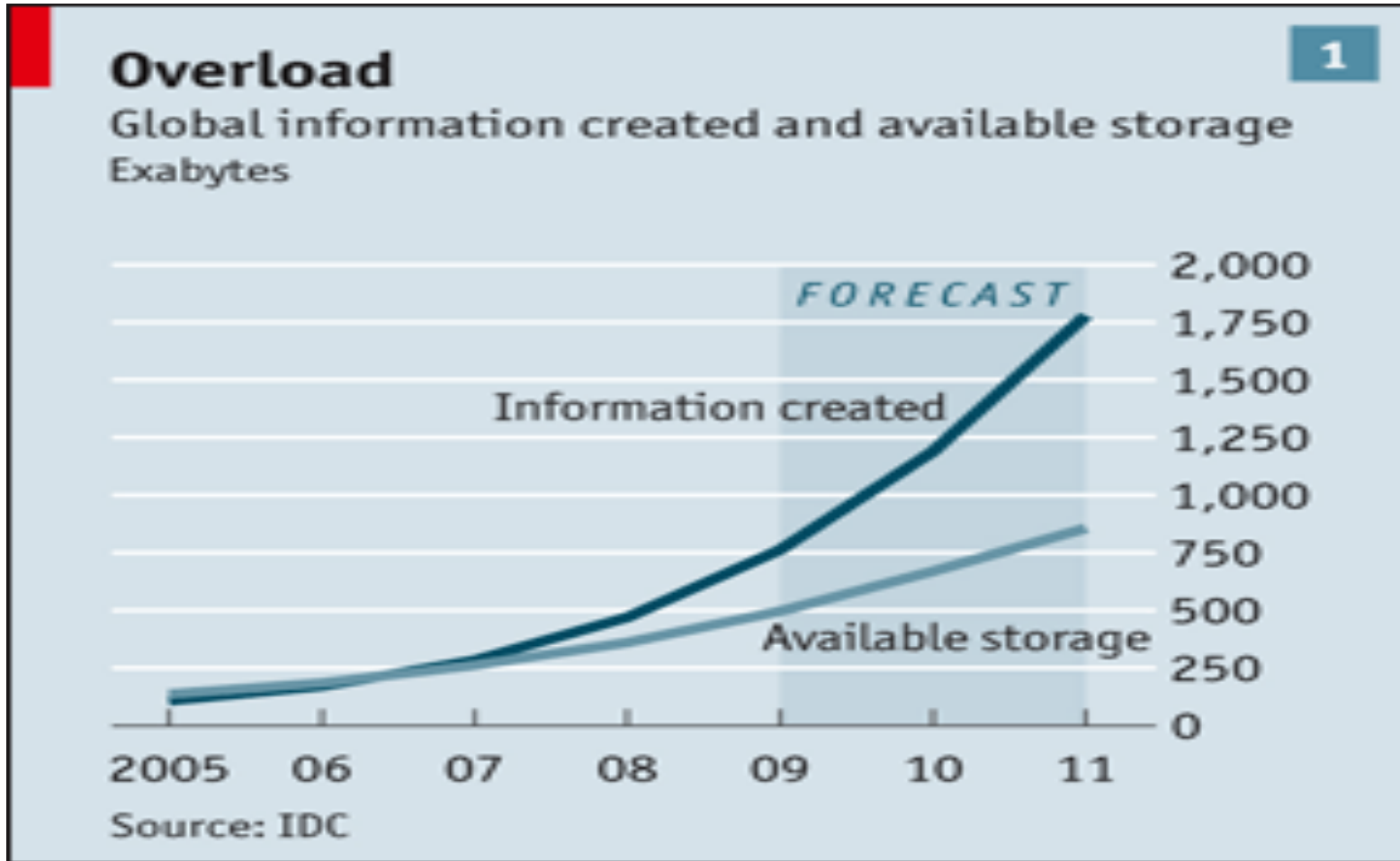


70 yeni domain kaydı
13000 iPhone uygulaması indirimi
700,000 Facebook güncellemesi
168 milyon e-mail

60 yeni blog
1500+ blog yazısı
98,000 yeni tweet
600+ yenivideo



Bu verileri saklayacak yer yok!



Houston... We have a problem!...

- Bir kurumun çözümlenebileceği verinin yüzdesi, o kuruma gelen verinin artış hızı ile orantılı olarak azalıyor.
- Başka bir deyişle, zaman geçtikçe, işimiz hakkında daha az bilgi sahibi oluyoruz.

Kurumların ÜRETİĞİ
erişilebilir veriler

Sinyaller
Ve Gürültü



Kurumların
İŞLEYEBİLDİĞİ veriler

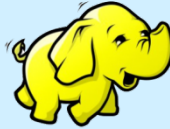
Büyük Veri Problemi Farklı Teknolojik Yetkinlikler Gerektiriyor

Federe veri kaynakları
üzerinde arama ve keşif



Federated Discovery and Navigation

Her çeşit büyük veriyi
saklama ve yönetme



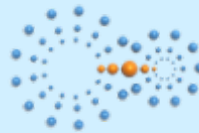
Hadoop File System
MapReduce

Biçimsel verilerin hızlı
analizi



Data Warehousing

Akışkan verilerin yönetimi



Stream Computing

Biçimsel olmayan verilerin
analizi



Text Analytics Engine

Veri kaynaklarının
entegrasyonu ve sahipliği



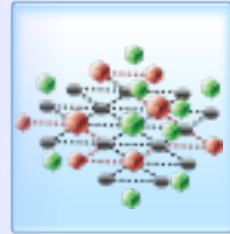
Integration, Data Quality, Security,
Lifecycle Management, MDM

Kullanım alanları sıklıkla farklı teknolojilerin harmanlanmasını gerektiriyor



Veri yükleme ve ön-işleme

Tüm verilerin saklanması, Yapısal olmayan verilerin ön-işlemeden geçirilerek, değerli bilgilerin ortaya çıkarılması



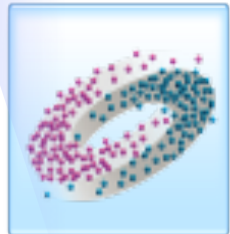
Yapısal ve yapısal olmayan veriler üzerinde derinlemesine analitik

Yapısal olmayan veriler üzerinde gelişkin text analitik fonksiyonları ve veri madenciliği modelleri kullanılarak mevcut veri ambarı analitik yetkinliklerinin geliştirilmesi



Akan veriler ile tarihsel verilerin birlikte kullanımı

Tarihsel verilerle oluşturulan modellerin, akan veriler üzerinde çalıştırılarak, gerçek zamanlı analizler yapılması ve aksiyon alınabilmesi



Keşifsel analizler için yapısal verilerin tekrar kullanımı

Deneysel çalışmalar, ad-hoc analiz ve görselleştirme

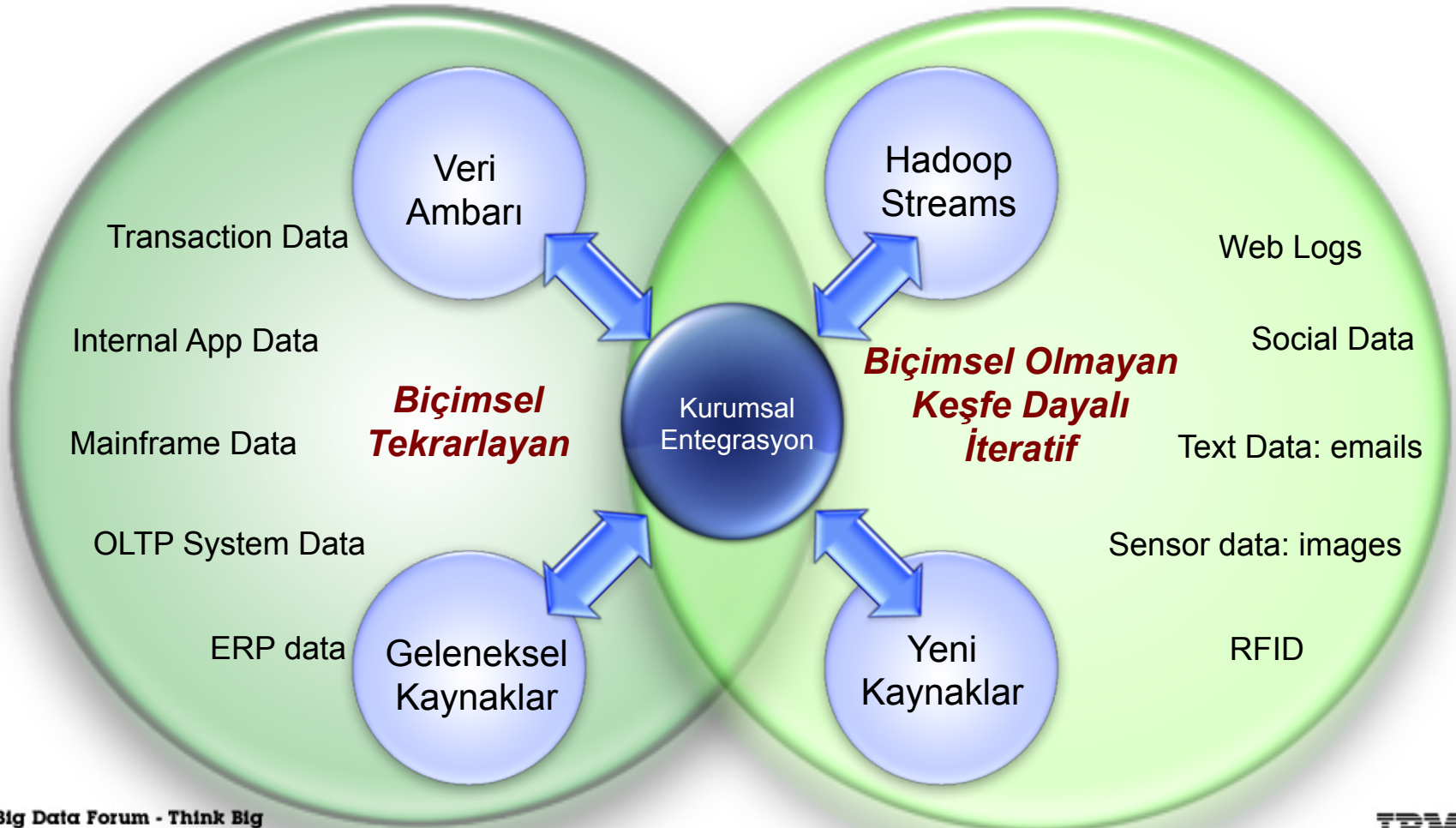
Farklı Uygulama Alanları için Birbirini Bütünleyen Yaklaşımlar

Geleneksel Metodlar

Biçimsel, Analitik, Tekrarlayan

Yeni Yaklaşımlar

Yaratıcı, bütünsel görüş, sezgisel



Geleneksel yaklaşım ile Big Data yaklaşımlarının sinerjisi

Geleneksel Yaklaşım

Biçimsel Veri ve Tekrarlayan Raporlar

Son kullanıcıların yapacağı analizler hemen hemen belirlidir



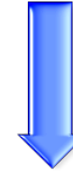
BT sorulacak soruları ve sorguları en hızlı yanıtlayacak platformu oluşturur



Big Data Yaklaşımı

İteratif ve Keşfe Yönelik

BT Yaratıcı keşflere yönelik bir platform ve veri zenginliği sunar



Son kullanıcılar neler keşfedilebilir araştırır, tarihsel veriye özgürce erişirler



Data Scientist!

CENTRAL INTELLIGENCE AGENCY

English
More>



THE WORK OF A NATION. THE CENTER OF INTELLIGENCE.

Career Opportunities

CIA Home > Careers and Internships > Career Opportunities > Science, Engineering & Technology Positions > Data Scientist



CIA Home

About CIA

▾ Careers and Internships

▾ Career Opportunities

View All Career Opportunities

Analytical Positions

Business, IT & Security Positions

Clandestine Service Positions

Language Positions

▾ Science, Engineering & Technology Positions

View Jobs

▶ Data Scientist

Data Scientist

Work Schedule: Full Time

Salary: \$51,418 – \$136,771

Location: Washington, DC metropolitan area

Do you have a passion for creating data-driven solutions to the world's most difficult problems? The CIA needs technically-savvy specialists to organize and interpret Big Data to inform US decision makers, drive successful operations, and shape CIA technology and resource investments. The CIA is looking for individuals from diverse educational backgrounds to fill the role of data scientist. If you have experience in data analytics, computer science, mathematics, statistics, economics, operations research, computational social science, quantitative finance, engineering or other data analysis fields, consider a career as a Data Scientist at CIA.

Hadoop Nedir?



Ölçeklenebilir

- Yeni nodelar canlı sisteme eklenebilir

Düşük Maliyetli

- MPP mimariyi düşük maliyetli sunucular ile sağlayabiliyor

Esnek

- Hadoop şema gerektirmez, her türlü veriyi kullanabilirsiniz

Hata Toleranslı

- MapReduce paradigması sayesinde

Hadoop hangi Big Data problemleri için düşünülebilir ?

**Büyük Hacimli Verilerin Analizi
ve Saklanması**



**Farklı tipteki verilerin bileşiminden
elde edilebilecek yeni bilgiler**



**Veri Hacmi Nedeniyle Pahalı Kalan
Mevcut Teknolojilerin Bütünlenmesi**



**Veri Kaynaklarının Keşfi
(Data Scientist)**



5 Temel Kullanım Alanı



Büyük Veri Arama

Karar alma sürecini geliştirmek için tüm büyük veri kaynaklarını aramak, veriyi anlamak, görselleştirmek



Gelişkin 360° Müşteri Görüntüsü

Yeni İç ve Dış veri kaynaklarıyla mevcut Müşteri Görüntüsünü genişletin



Yeni Güvenlik / İstihbarat Yetkinlikleri

Düşük risk, gerçek zamanlı dolandırıcılık izleme ve siber güvenlik algılama



Operasyon Analizi

Makinelerin ürettiği farklı çeşitlilikteki verilerin analiz edilerek iş sonuçlarının geliştirilmesi



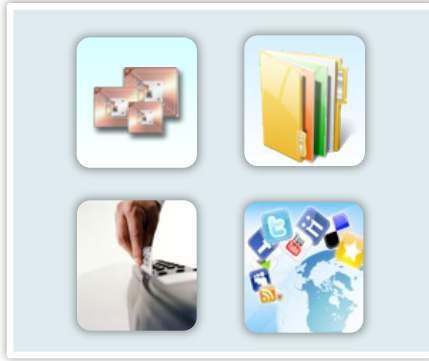
Veri Ambarı Eklentisi

Operasyonel verimliliği artırmak için büyük veri ve veri ambarı özelliklerini entegre etmek

Veri Ambarı Eklentisi olarak Büyük Veri

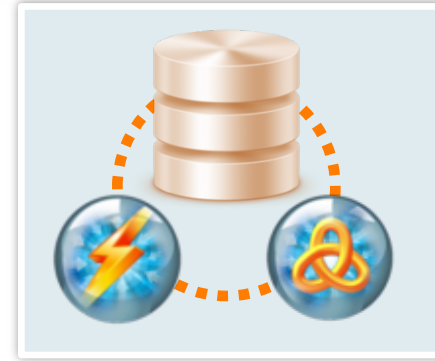


Operasyonel verimliliği artırmak için büyük veri ve veri ambarı özelliklerini entegre etmek



Farklı yapıdaki verilerden faydalanmak

- Yapısal, yapısal olmayan, ve akan verilerin analitik süreçlerde efektif kullanımı
- Düşük gecikme gereksinimleri
- Sorgulanabilir veri ortamı

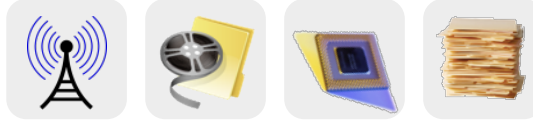


Veri Ambarının Genişletilmesi

- Nadir kullanılan verilerin taşınabileceği, veri saklama, lisans ve bakım maliyetleri için uygun çözüm
- Akan verilerin akıllıca işlenmesi sayesinde saklama maliyetlerinden kazanç
- Filtreden geçirilmiş veriler sayesinde daha hızlı veri ambarı performansı

Veri Ambarı Eklentisi : Farklı İhtiyaçlar

1 Ön-İşleme



Data Explorer

Streams
Gerçek zamanlı operasyonlar

BigInsights
Her tür veri için ara katman



Data Warehouse

2 Sorgulanabilir Arşiv

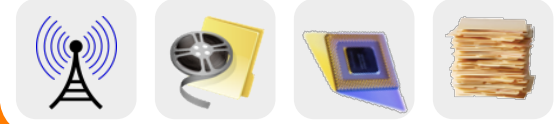
BigInsights
Information Integration

Data Explorer
Verinin aranıp görüntülenmesi



Data Warehouse

3 Keşifsel Analiz



Yapısal olmayan verilerle birlikte analiz



Streams
Analitiğin mikro saniyeler mertebesinde çalıştırılması

Data Warehouse

IBM Big Data Stratejisi: Analitiği Veriye yaklaştırın

Analitik Uygulamalar

BI / Raporlama | Analiz/ Görselleştirme | Fonksiyonel Uyg. | Endüstri Uyg. | Veri Madenciliği | İçerik Analitiği

IBM Big Data Platformu

Görselleştirme
ve Keşif

Uygulama
Geliştirme

Sistem
Yönetimi



Hızlandırıcılar

Hadoop
Sistemi



Akışkan
veri işleme



Veri
Ambarı



Veri Entegrasyonu ve Sahipliği

- Tipi, biçimi, boyutu ve akışkanlığı ne olursa olsun tüm verilerinizi entegre edilebilmesi ve yönetimi

- İleri düzey analitik fonksiyonların verinin doğal halinde uygulanması

- Elinizdeki tüm verileri görselleştirebilmesi

- Yeni nesil analitik uygulamalar geliştirebilmek için geliştirme araçları

- İşyükü optimizasyonu ve zamanlama

- Güvenlik ve Veri Sahipliği



BigInsights, Hadoop yetkinliklerini kurumunuza getiriyor

Analitik Uygulamalar

BI / Raporlama | Analiz/ Görselleştirme | Fonksiyonel Uyg. | Endüstri Uyg. | Veri Madenciliği | İçerik Analitiği

IBM Big Data Platformu

Görselleştirme ve Keşif

Uygulama Geliştirme

Sistem Yönetimi



Hızlandırıcılar

Hadoop Sistemi



Akışkan veri işleme



Veri Ambarı



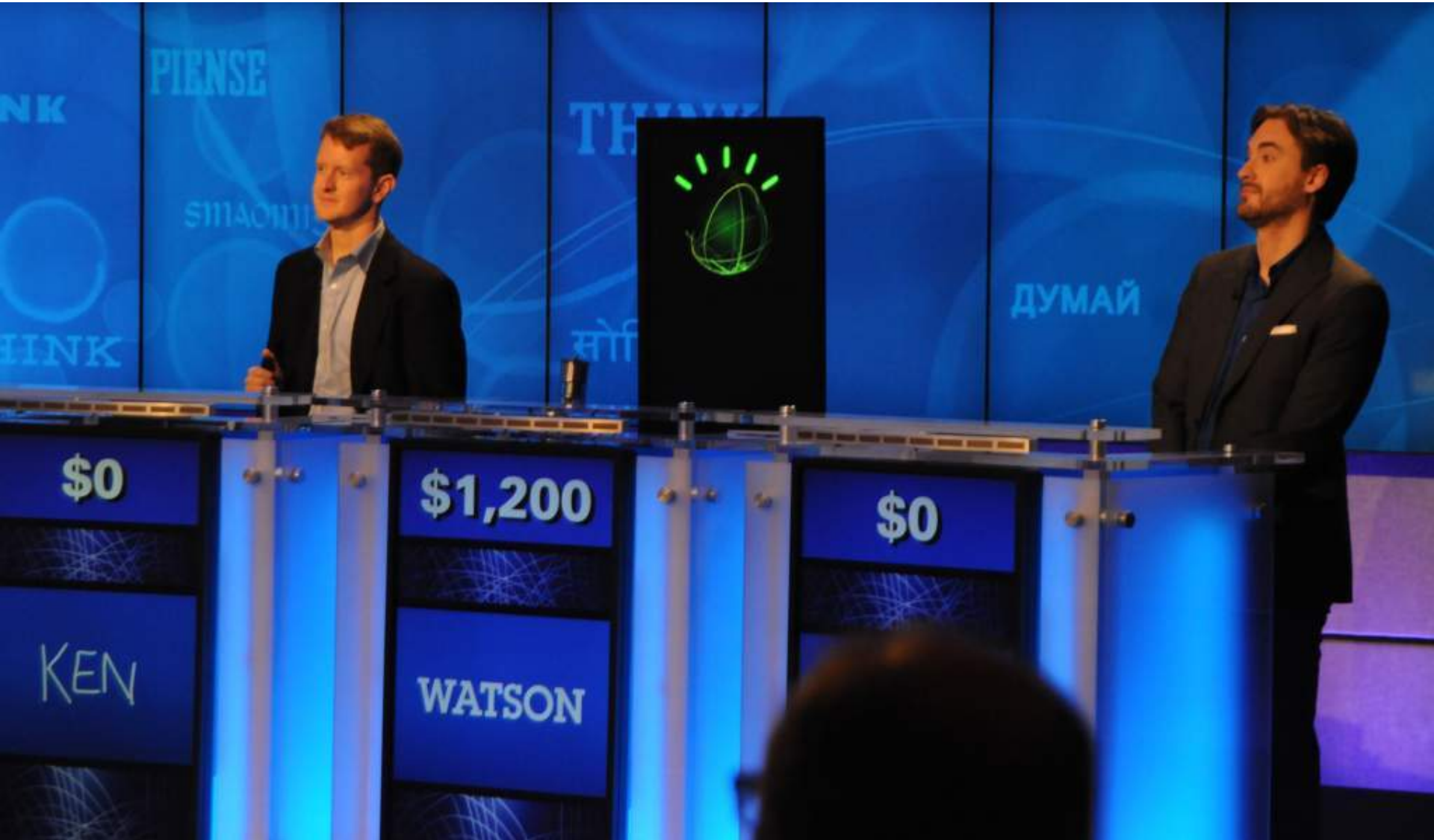
Veri Entegrasyonu ve Sahipliği

IBM InfoSphere BigInsights

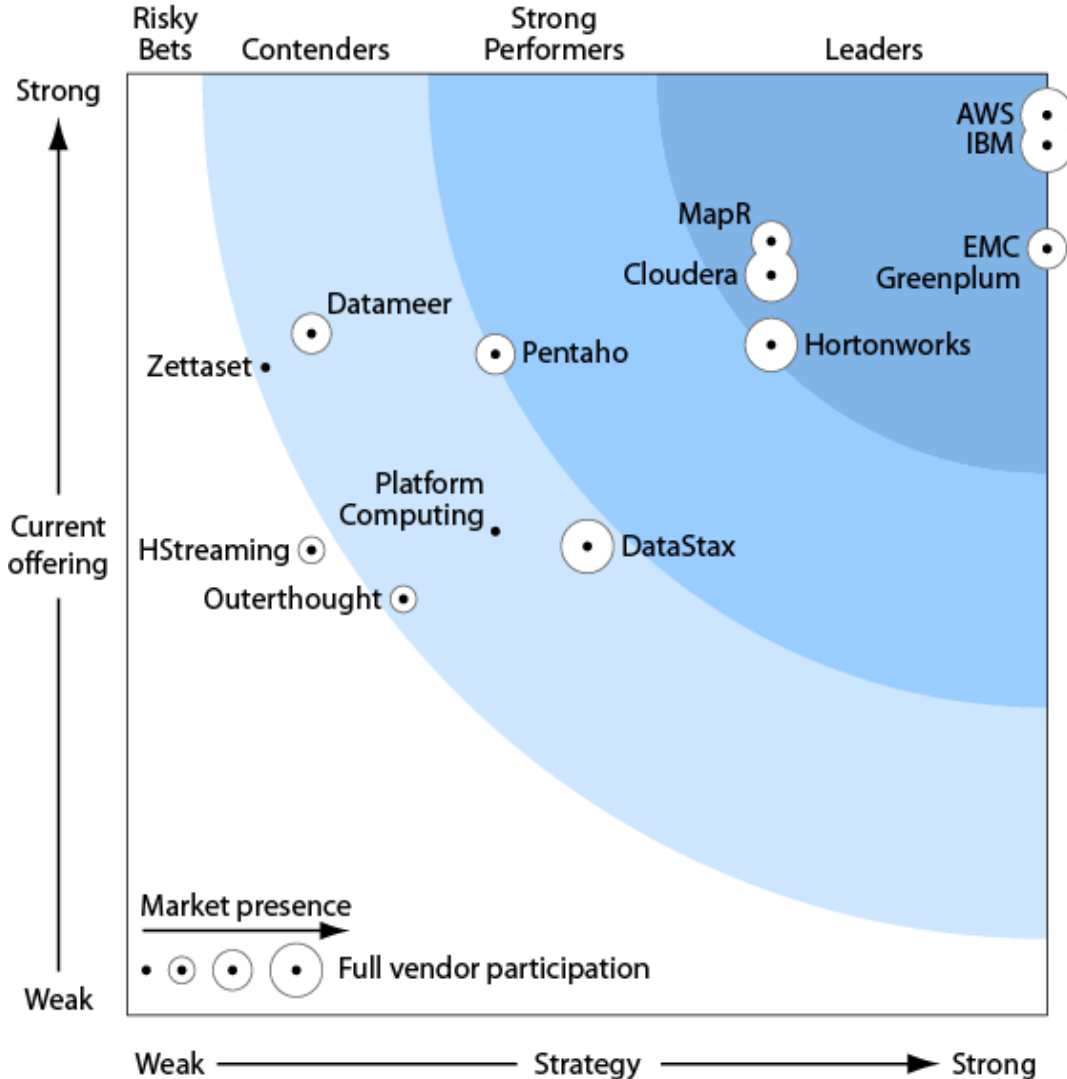
Açık kaynak kodlu Hadoop çözümünü tamamlıyor;

- Performans Optimizasyonları
- Geliştirme Araçları
- Kurumsal Entegrasyon Yetkinlikleri
- Analitik Hızlandırıcılar
- Hazır uygulamalar
- Görselleştirme
- Güvenlik
- Gelişkin Analitik Yetkinlikler

IBM Watson



Endüstrideki ilk Hadoop Sistemleri Değerlendirmesi



FORRESTER®

“IBM has the deepest Hadoop platform and application portfolio. IBM, an established EDW vendor, has its own Hadoop distribution; an extensive professional services force working on Hadoop projects; extensive R&D programs developing Hadoop technologies; connections to Hadoop from its EDW.”

–The Forrester Wave™: Enterprise Hadoop Solutions, 1Q12



Vestas optimizes capital investments based on 2.5 Petabytes of information

Need

- Model the weather to optimize placement of turbines, maximizing power generation and longevity

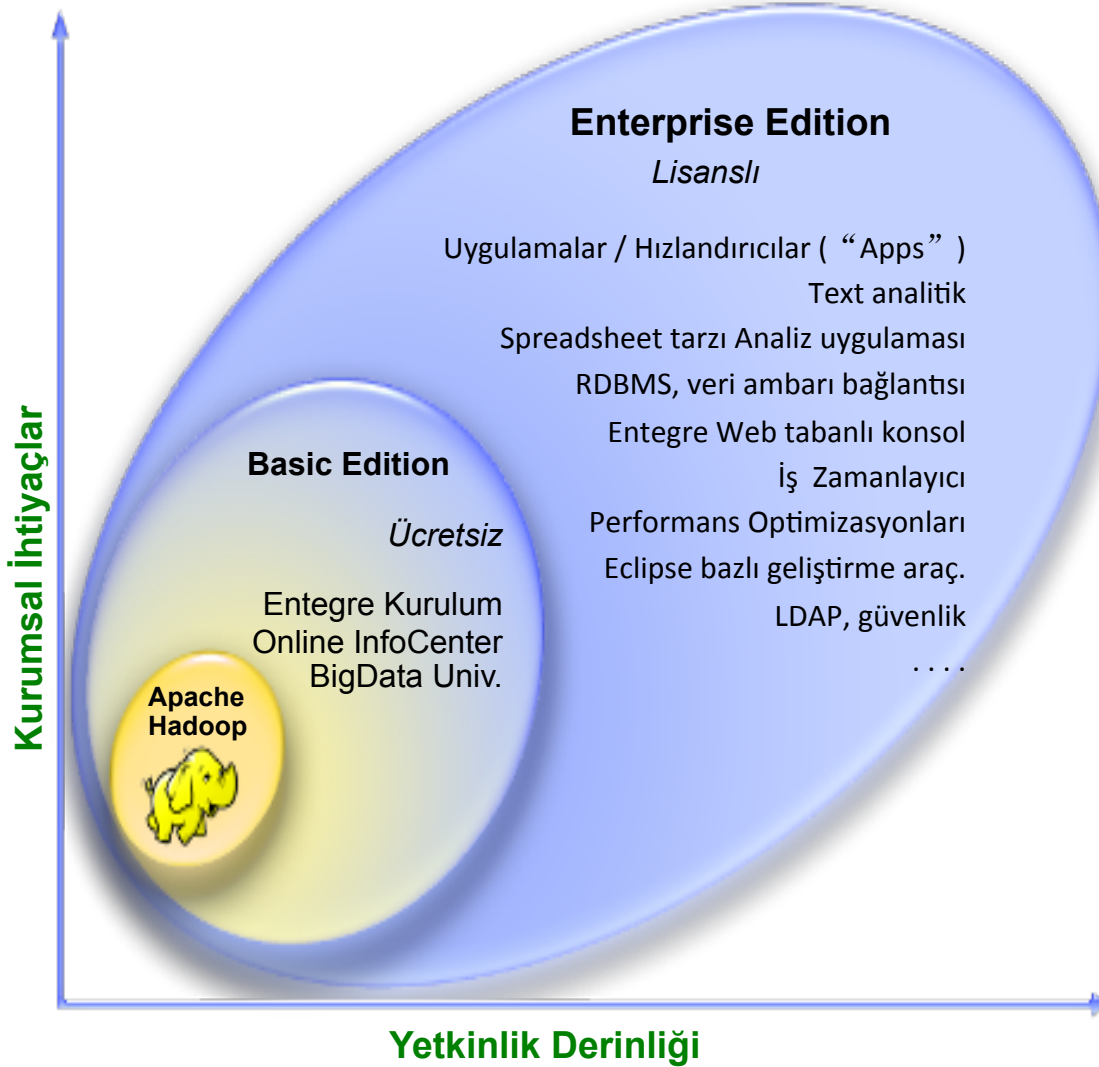
Benefits

- Reduce time required to identify placement of turbine from weeks to hours
- Reduces IT footprint and costs, and decreases energy consumption by 40 % -- while increasing computational power
- Incorporate 2.5 PB of structured and semi-structured information flows. Data volume expected to grow to 6 PB





IBM Infosphere BigInsights





Basic Edition

Open Source

IBM

Infrastructure

Integrated installer

ZooKeeper

Jaql

Pig

Oozie

HBase

Hive

Lucene

MapReduce

HDFS

Connectivity

JDBC

Flume



Enterprise Edition

Open Source

IBM

Optional IBM and partner offerings

Analytics and discovery

Text processing engine and library

BigSheets

“Apps”

DB export

Web Crawler

DB import

Boardreader

Hive query

Distrib file copy

Pig query

...

Jaql query

Administrative and development tools

Web console

- Monitor cluster health
- Add / remove nodes
- Start / stop services
- Inspect job status
- Inspect workflow status
- Deploy apps
- Launch apps / jobs
- Work with distrib file system
- Work with spreadsheet interface
- Support REST-based API
- ...

Eclipse plug-ins

- Text analytics
- MapReduce programming
- Jaql development
- Hive query development

Infrastructure

Integrated installer

Enhanced security

ZooKeeper

Jaql

Pig

Oozie

HBase

Hive

Text compression

BigIndex

Lucene

Adaptive MapReduce

MapReduce

Flexible scheduler

GPFS

HDFS

Connectivity

JDBC

Netezza

DB2

Streams

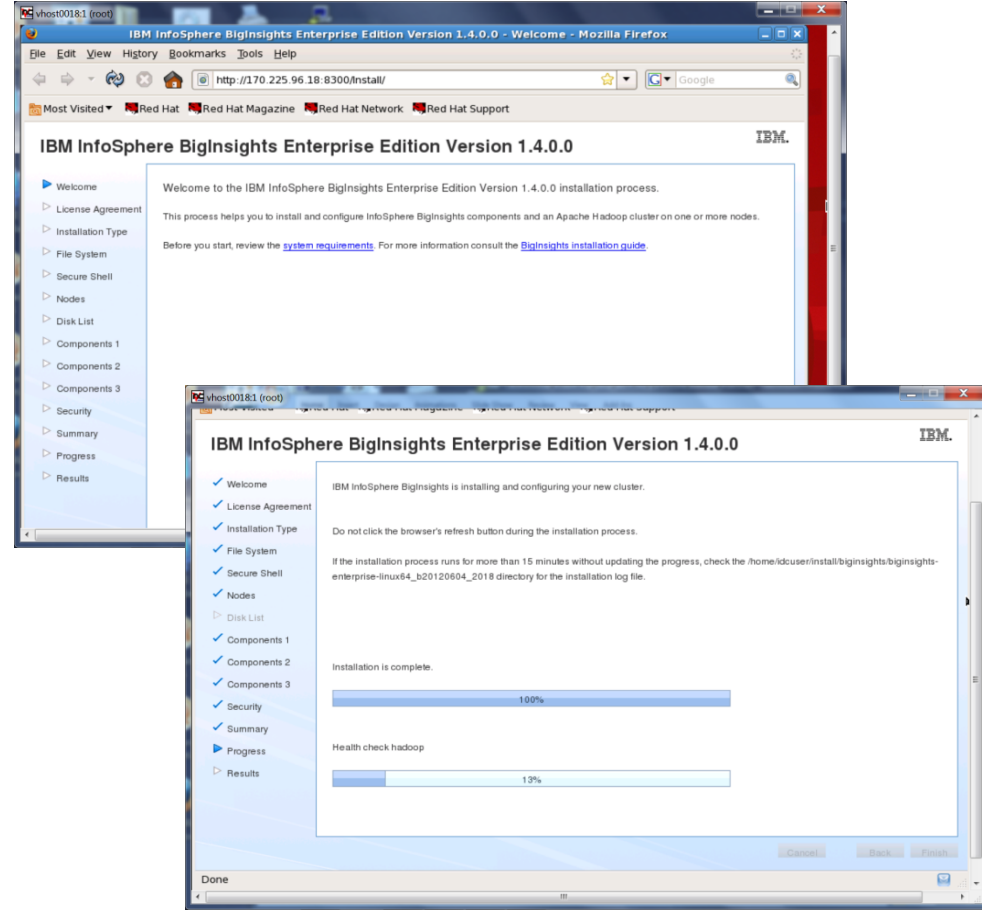
R

Flume

Streams*

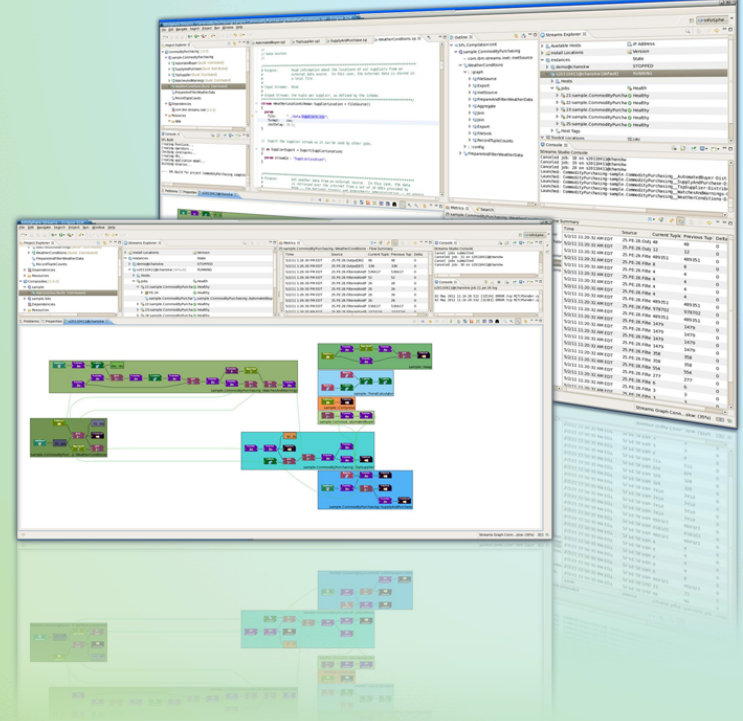
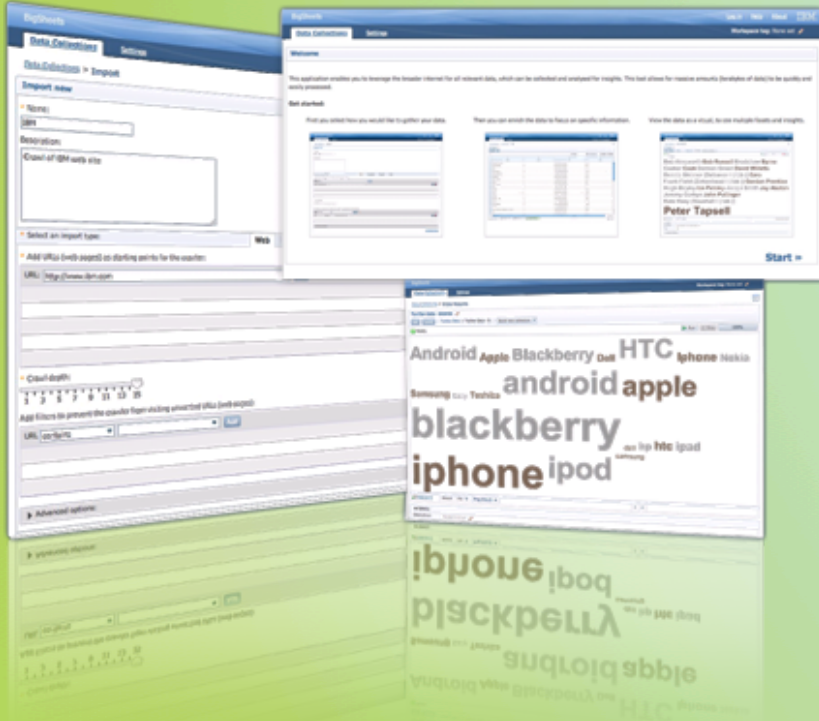
Web Arayüzlü Kurulum

- İster tek sunucu üzerine, ister çok sayıda sunucu üzerine son derece basit bir şekilde Hadoop Cluster kurulumu
- Tüm seçilen bileşenlerin otomatik kurulumu
- Kurulum sonrasında validasyon ve sağlık kontrolü; hem IBM ve hem açık kaynak kodlu programların tamamı için



Açık kaynak kodlu programların indirilip, konfigüre edilip, test edilmesine gerek kalmamakta, versiyon uyumsuzlukları gibi sorunlar çıkmamakta ve tüm yazılımsal bağımlılıklar otomatik olarak çözümlenmektedir.

Kullanıcı ve Geliştiriciler için Basit Arayüzler



Son Kullanıcı Görseleştirme

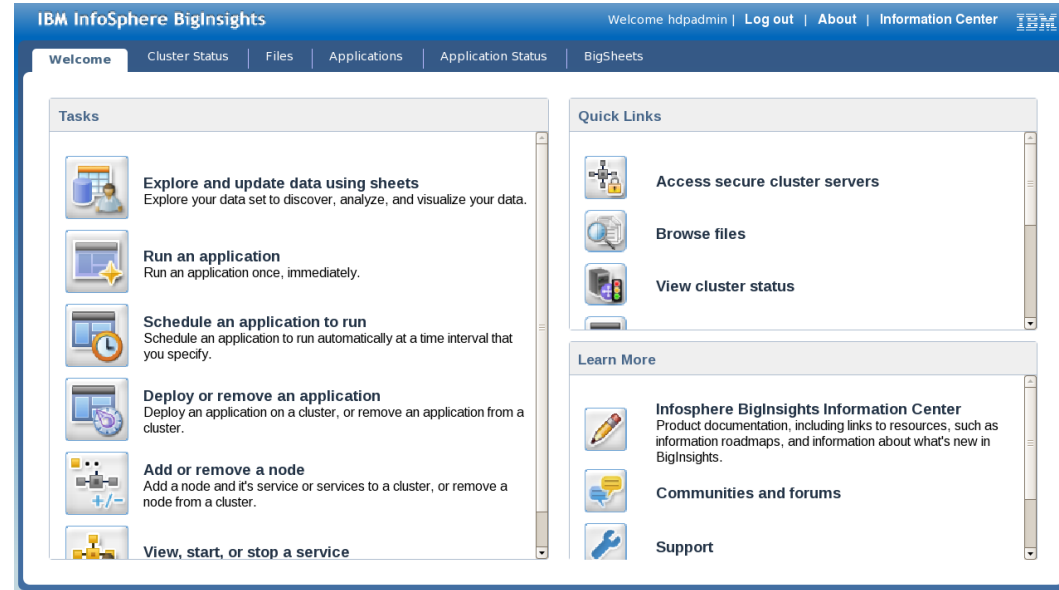
Veri keşfi, veri toplama (crawling), ve analitik

Geliştirme ve Yönetim Araçları

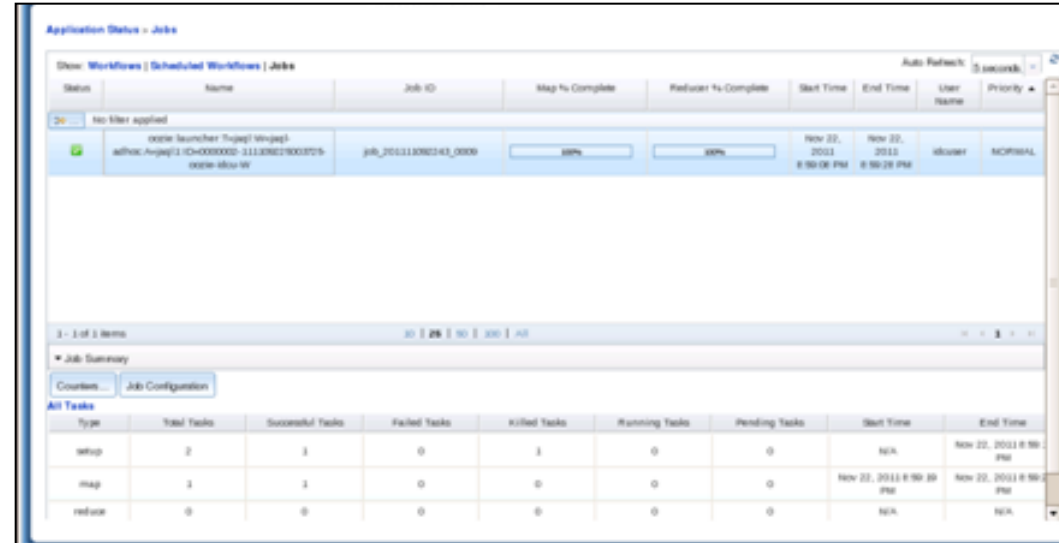
Aşına olduğunuz geliştirme araçları ve ortamı, test ve optimizasyon, sistem yönetimi

Web Arayüzlü Yönetim

- BigInsights Yönetimi
 - Sistem Sağlık Durumu
 - Yeni node ekleme / çıkarma
 - Servisleri başlatıp / durdurma
 - İş çalıştırma / durdurma
 - Dosya sistemini görüntüleme
 - Dosya ekleme/çıkarma



- Uygulamaları başlatma
 - Örn. BigSheets
 - Hazır uygulamaları çalıştırma (IBM kaynaklı veya geliştirdiğiniz uygulamalar)
- Uygulamaların yayınlanması
- Yardım sayfaları



Status	Name	Job ID	Map % Complete	Reducer % Complete	Start Time	End Time	User Name	Priority
Success	00000 Searcher [map] [map]	job_20111102141_0000	100%	100%	Nov 22, 2011 8:50:08 PM	Nov 22, 2011 8:50:28 PM	hduser	NORMAL

Type	Total Tasks	Successful Tasks	Failed Tasks	Killed Tasks	Warning Tasks	Pending Tasks	Start Time	End Time
map	2	2	0	0	0	0	N/A	Nov 22, 2011 8:50:08 PM
reduce	0	0	0	0	0	0	N/A	N/A
setup	2	2	0	0	0	0	Nov 22, 2011 8:50:08 PM	Nov 22, 2011 8:50:28 PM

Son Kullanıcı Deneyimi : Uygulamaların Çalıştırılması

IBM InfoSphere BigInsights

Welcome | Cluster Status | Files | **Applications** | Application Status | Sheets

About | Help IBM

Applications

- Ad hoc Hive query
- Ad hoc Jaql query
- Ad hoc Pig query
- Boardreader
- Database Export
- Database Import
- Distributed File Copy
- TeraGen-TeraSort
- Web Crawler
- Word Count

Execution Name:

▼ Parameters

Input path:

Output path:

▼ Advanced Settings

Update Sheets Collection

Schedule Job [View Schedule Configuration](#)

Start Date:

Frequency:

Until:

Application History

Status	Execution Name	Progress	Start Time	Elapsed Time	Output	Details
No filter applied						
0 items						

Web tabanlı uygulama erişim ara yüzü; HIVE sorgusu

The screenshot shows the 'Applications' tab in a web interface. The navigation bar includes 'Welcome', 'Cluster Status', 'Files', 'Applications', and 'Application Status'. The main area displays a grid of application tiles:

- BoardReader
- Pig sample
- Crawler
- Database
- Tera Gen-Sort
- Hive sample
- Word Count
- Database
- Simple Jaql Application
- Distributed Copy
- Hadoop Streaming Word Count Sample
- Ad h...

Name: Ad hoc Hive query

Description:
You can use the ad hoc Hive query to create your own customized...

Execution:
Execution Name:

Parameters:
Hive query:

```
CREATE EXTERNAL TABLE student (  
  NAME STRING,  
  AGE INT,  
  GPA FLOAT  
)  
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\n'  
STORED AS TEXTFILE  
LOCATION '/tmp/mystudents';
```

Advanced Settings

Application History

Status	Execution Name	Progress
No filter applied		
<input checked="" type="checkbox"/>	Default Execution	<input type="text" value="100%"/>

Son kullanıcılar da basitçe yeni uygulamalar geliştirebilir

IBM InfoSphere BigInsights

About | Information Center IBM

Welcome | Dashboard | Cluster Status | Files | **Applications** | Application Status | BigSheets

Execute Applications | Chain Applications

Applications

Boardreader Data Sample

JAQL Database Export

Data Subset Database Import

Distributed File Copy

Web Crawler Word Count

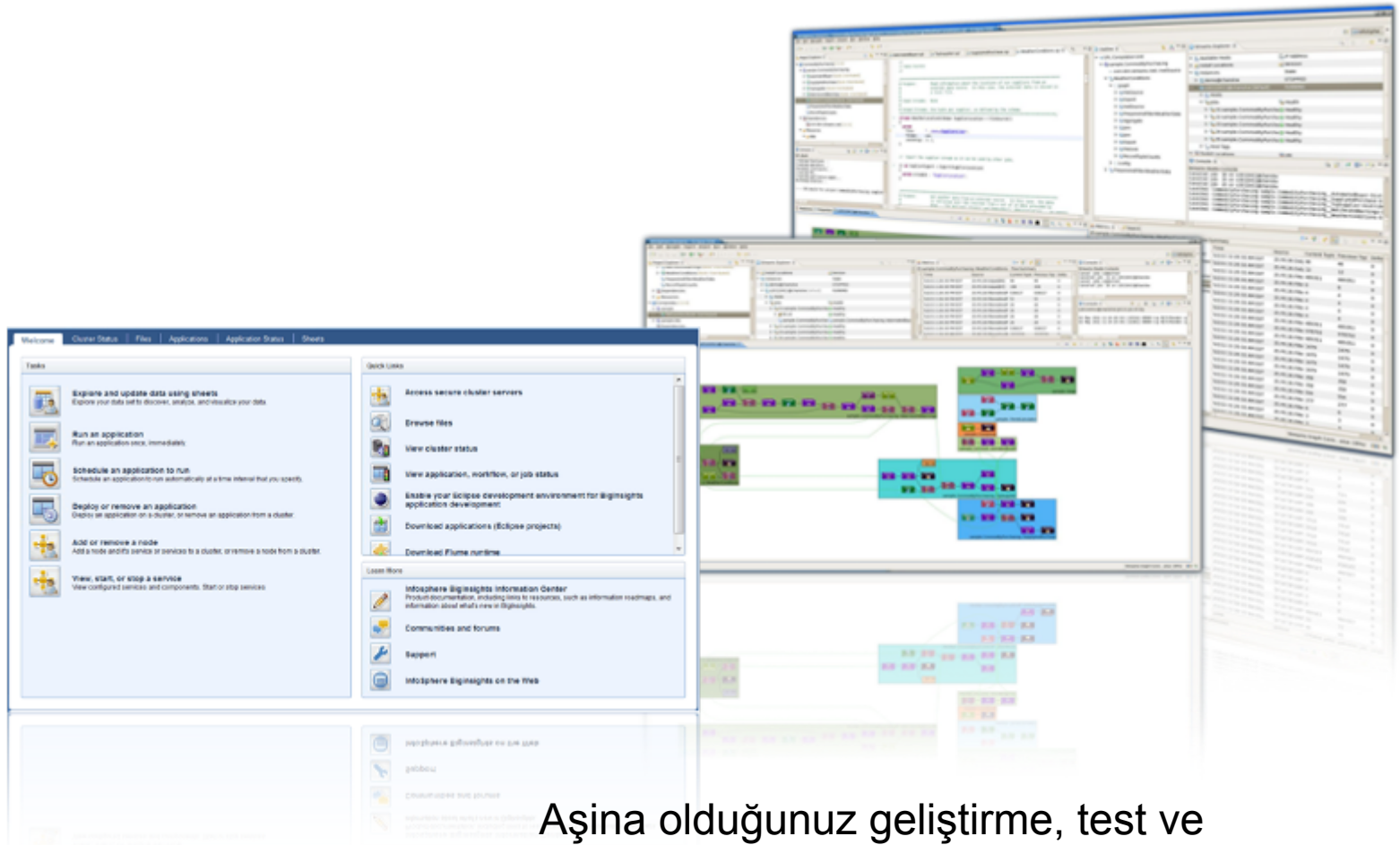
Application Name: Data Sampling Analysis

Application Description: Application transfers data from ftp to hdfs. Generates a sample of it and exports it into a database.

Distributed File Copy → Data Sample → Database Export → []

Next

Profesyonel Geliştirme ve Yönetim Araçları



Aşına olduğunuz geliştirme, test ve optimizasyon araçları

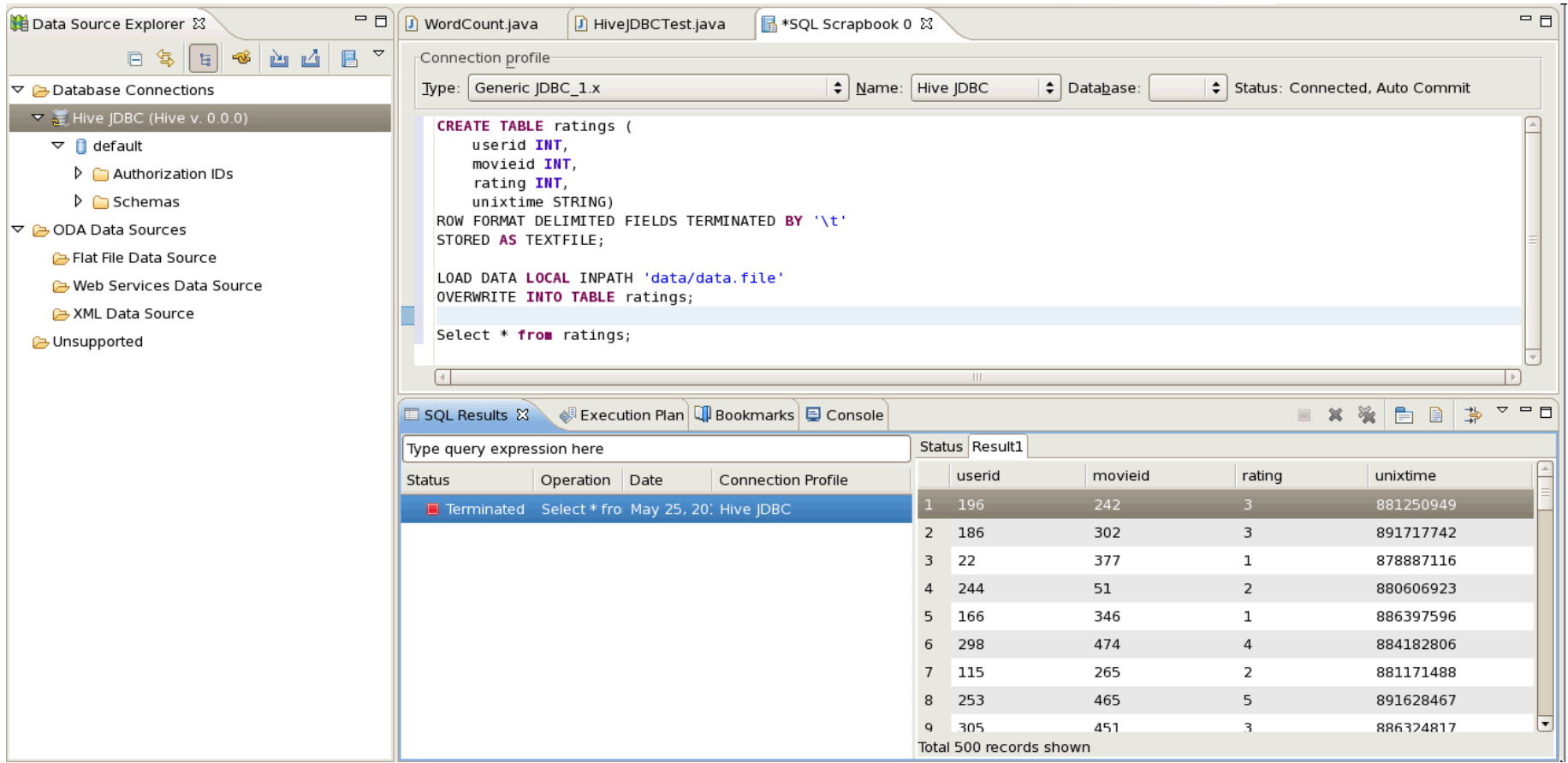
Big Data Programları geliştirme – Map Reduce örneği

Eclipse bazlı geliştirme araçları JAQL, Hive, Java MapReduce, Text Analytics

The screenshot displays the Eclipse IDE interface during the development of a Java MapReduce program. The 'New' wizard is open, showing the 'Wizards' list with 'Java MapReduce Program' selected. The 'Project Explorer' shows the project structure, including 'WordCount.java'. The 'Run As' context menu is open, with 'Java MapReduce' selected. The 'Console' window shows the output of the program execution, including the following log entries:

```
<terminated> New_configuration [Java MapReduce] /home/hadoop/java/ibm-java-i386-60/bin/javaw (Apr 25, 2012 3:49:17 PM)
12/04/25 15:49:21 INFO input.FileInputFormat: Total input paths to process : 2
12/04/25 15:49:22 INFO mapred.JobClient: Running job: job_local_0001
12/04/25 15:49:22 INFO util.ProcessTree: setsid exited with exit code 0
12/04/25 15:49:22 INFO mapred.Task: Using ResourceCalculatorPlugin : org.apache.hadoop.util.LinuxResourceCalcu
12/04/25 15:49:22 INFO mapred.MapTask: io.sort.mb = 100
12/04/25 15:49:25 INFO mapred.JobClient: map 0% reduce 0%
12/04/25 15:49:25 INFO mapred.MapTask: data buffer = 79691776/99614720
12/04/25 15:49:25 INFO mapred.MapTask: record buffer = 262144/327680
12/04/25 15:49:27 INFO mapred.MapTask: Starting flush of map output
12/04/25 15:49:30 INFO mapred.MapTask: Finished spill 0
12/04/25 15:49:30 INFO mapred.Task: Task:attempt_local_0001_m_000000_0 is done. And is in the process of commit
12/04/25 15:49:31 INFO mapred.LocalJobRunner:
12/04/25 15:49:31 INFO mapred.LocalJobRunner:
```


SQL Editör arayüzü ile Hive sorguları çalıştırma ve sonuçları görüntüleme



The screenshot shows a SQL IDE interface with the following components:

- Data Source Explorer:** Shows a tree view of database connections, including Hive JDBC (Hive v. 0.0.0) with sub-items for default, Authorization IDs, and Schemas.
- Connection Profile:** Shows a profile named 'Hive JDBC' with a status of 'Connected, Auto Commit'.
- SQL Editor:** Contains the following SQL code:

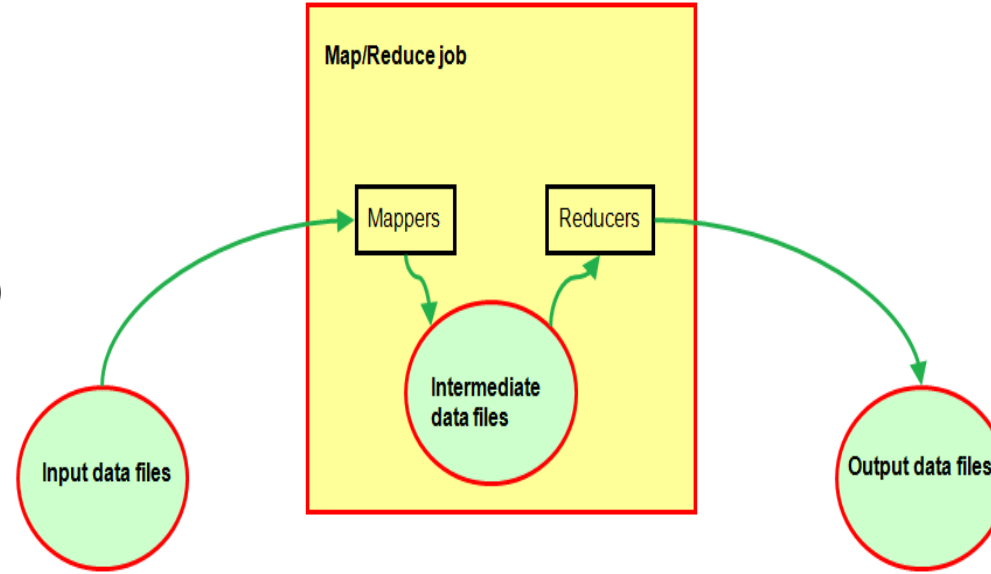
```
CREATE TABLE ratings (  
  userid INT,  
  movieid INT,  
  rating INT,  
  unixtime STRING)  
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'  
STORED AS TEXTFILE;  
  
LOAD DATA LOCAL INPATH 'data/data.file'  
OVERWRITE INTO TABLE ratings;  
  
Select * from ratings;
```
- SQL Results:** Shows the execution status as 'Terminated' and the results of the query. The results are displayed in a table with columns: Status, Operation, Date, Connection Profile, userid, movieid, rating, and unixtime.

Status	Operation	Date	Connection Profile	userid	movieid	rating	unixtime
Terminated	Select * fro	May 25, 20:	Hive JDBC	196	242	3	881250949
				186	302	3	891717742
				22	377	1	878887116
				244	51	2	880606923
				166	346	1	886397596
				298	474	4	884182806
				115	265	2	881171488
				253	465	5	891628467
				305	451	3	886324817

Total 500 records shown

Sıkıştırma

- Temelde LZO algoritması baz alınmıştır
- IBM'in kendi tasarımı
 - Izma ve bzip2'den çok daha hızlı
 - Makul sıkıştırma performansı (~60%)
- Faydaları :
 - Bütünüyle parçalanabilir
 - Index dosyasına ihtiyaç duymaz



	Size (Mbytes)	Comp . speed (sec)	Comp . memory used (MBytes)	Decomp. speed	Decomp. memory used (Mbytes)
uncompressed	96				
gzip	23	10	0.7	1.3	0.5
bzip2	19	22	8	5	4
lzo	36	1	1	0.6	0
lzm	18	63	14	3	1.8

Big Index - büyük boyutlu indeksleme

• Dağıtık İndeks ve Arama

– Parallel index

- İndeks operasyonları paralel çalışır, fakat bir indeks olarak saklanır

– Partitioned index

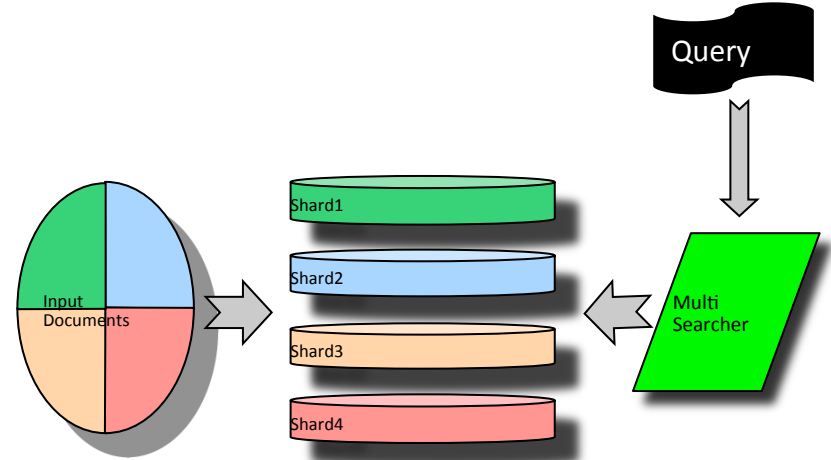
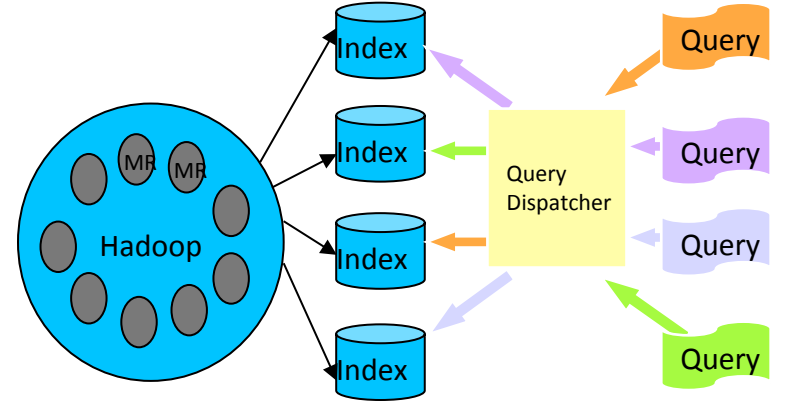
- İndeks farklı bölümlere bir anahtar bazında parçalanmıştır (örn. customer ID, tarih)
- Arama işlemi genellikle bu indeks bölümlerinden biri üzerinde koşturulur

– Distributed index

- Tek bir indeks fazlasıyla büyük olabilir
- Bu durumda indeks parçalara bölünüp, tek bir mantıksal indeks gibi davranır
- Her sorgu tüm parçalar üzerinden koşturulur

– Real-time index

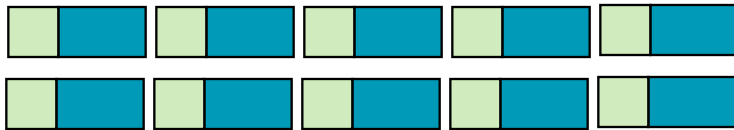
- Gerçek zamanlı kaynaklardan gelen veriler (örneğin twitter), gerçeğe yakın zamanlı olarak indekslenir



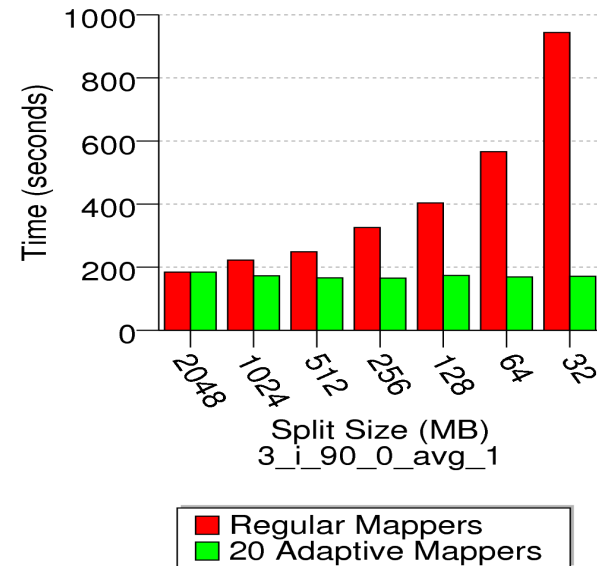
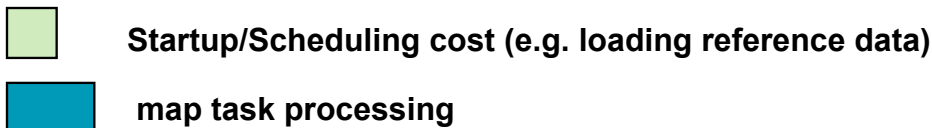
Adaptive MapReduce

- Geleneksel MAP işlemlerinin ayağa kalkması aşağıdaki durumlarda çok maliyetli olabilmektedir
 - Ufak dosyalar ve dosya parçacıkları üzerinde çalışıldığında
 - MAP işlemi basit hızlı işlemler gerçekleştirdiğinde
- Adaptive MR
 - Ufak dosyaların/parçaların dinamik olarak bütünleştirilerek az sayıda MAP işleminin koşturulmasını sağlar
- Sonuç:
 - Varsayılan (64MB) veya daha ufak bölümlenmeler de bile daha iyi performans

Traditional MR → n map tasks run consecutively on the same node/slot

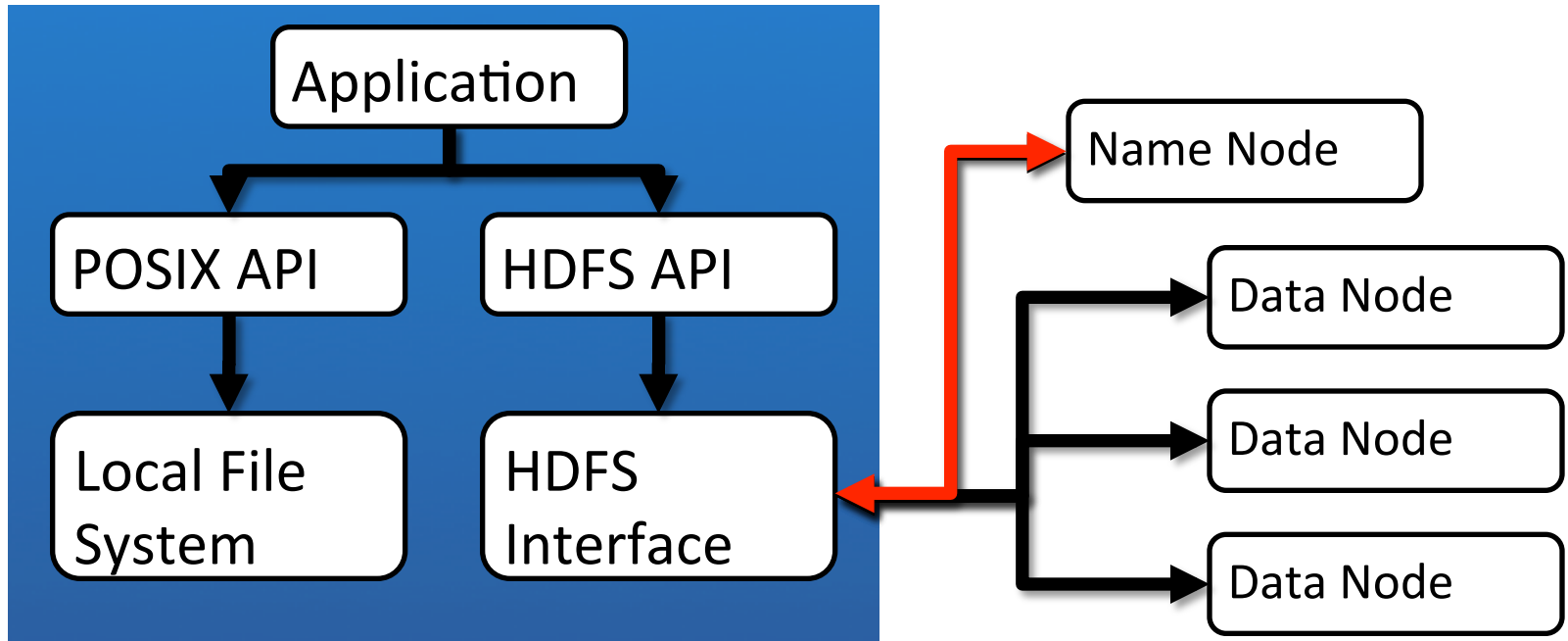


AdaptiveMR → One map task might process the seven splits



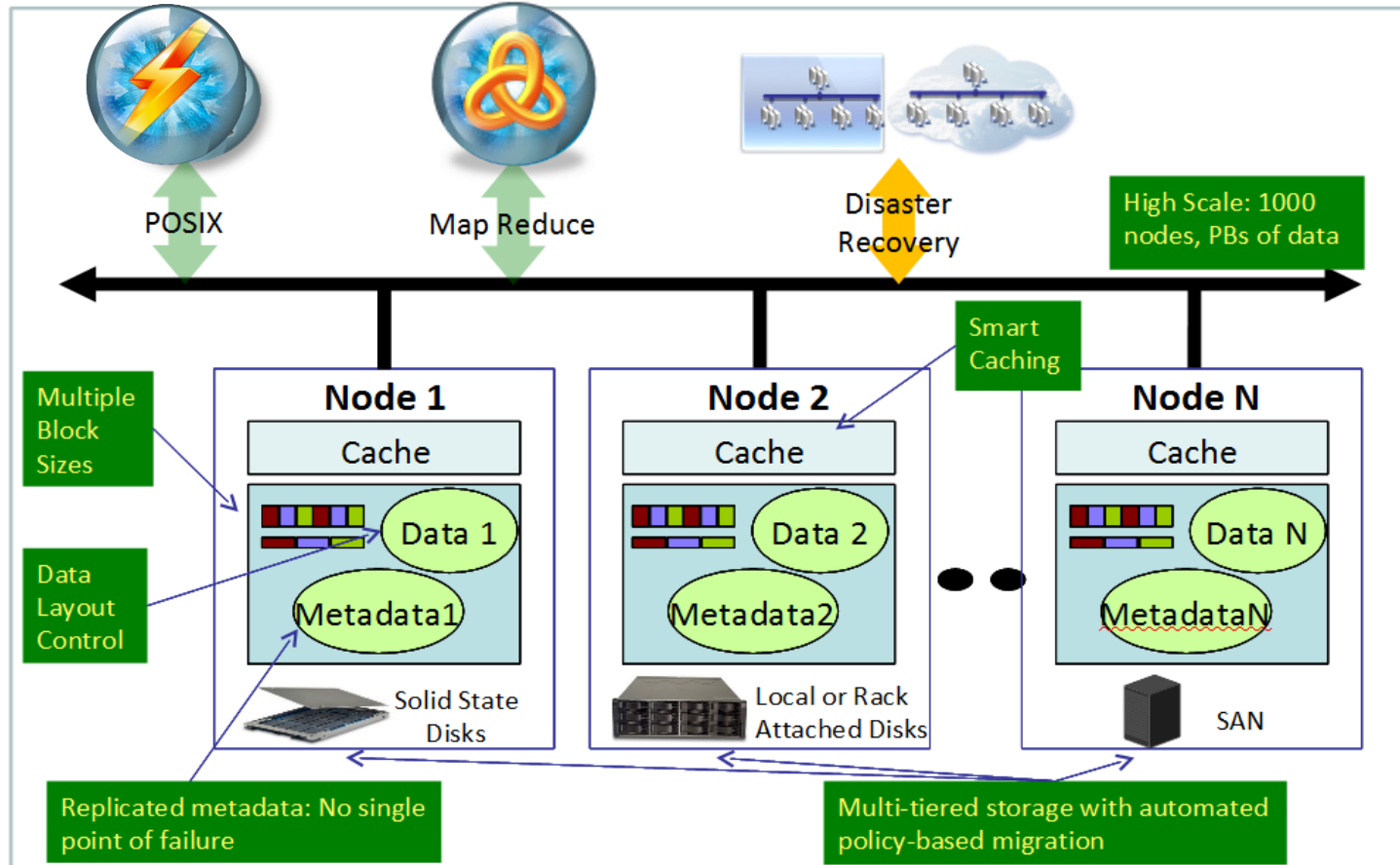
HDFS

- Hadoop Distributed File System
- Single Point of Failure



GPFS-FPO

- GPFS-FPO - Ölçeklenebilir, yüksek performanslı, ve güvenilir
- Hem MapReduce uygulamalarını hem de POSIX erişimini destekler
- Hem online hem de batch işlemleri destekler



BigInsights ve Veri Ambarı



Big Data
analitik
uygulamalar



BigInsights

facebook

twitter

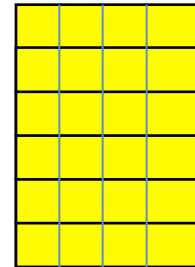
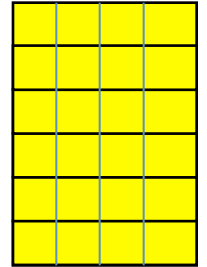
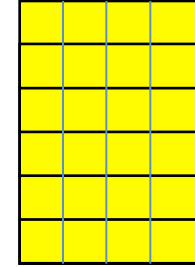


Filtreleme *Transformasyon* *Özetleme*



Geleneksel
Analitik

Veri Ambarı



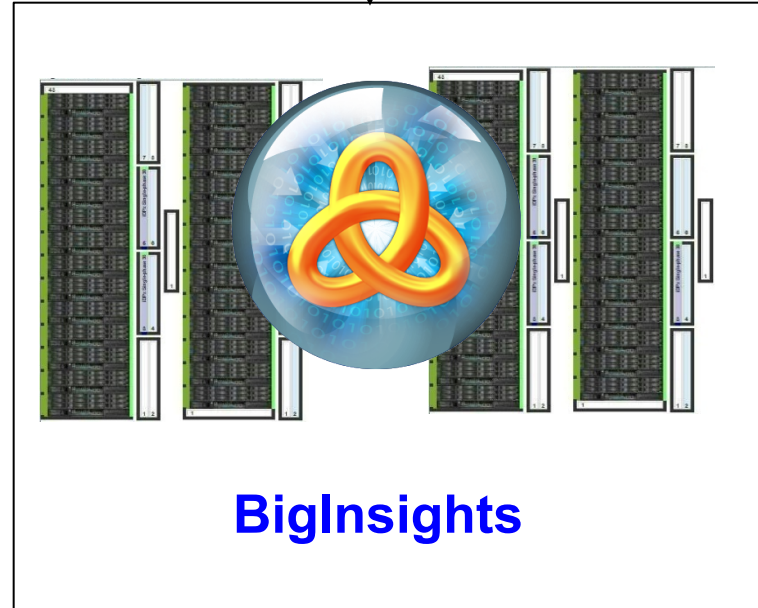
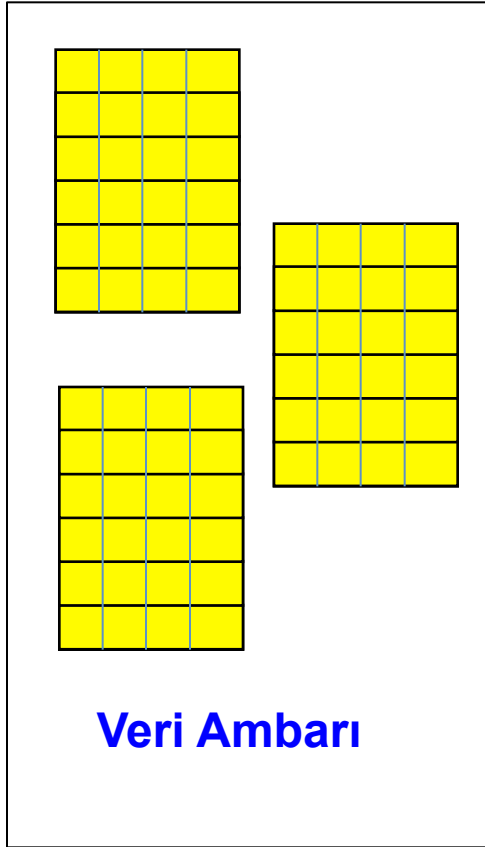
BigInsights ve Veri Ambarı



Geleneksel
Analitik



Big Data
Analitik



- *Sorgulanmaya hazır soğuk data, veri ambarı arşivi, yapısal olmayan, yarı yapısal veriler*

BigInsights Güvenlik Mimarisi



Kimlik Kontrol Ünitesi

Web Arayüzü

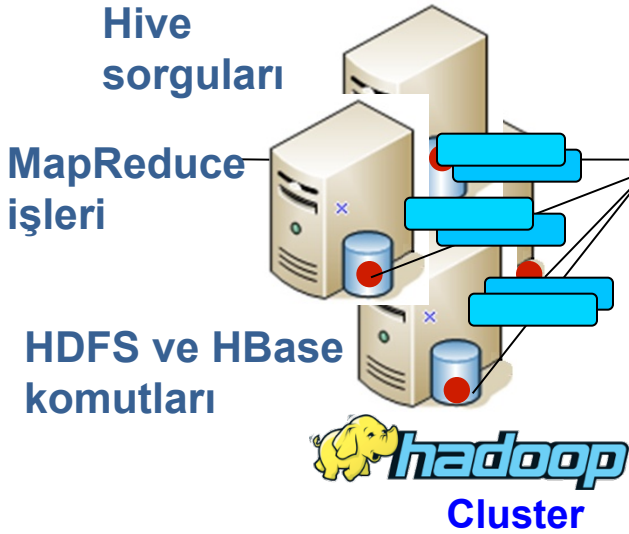


- Kullanıcı Kimlik Kontrolü
 - Dosya
 - LDAP
 - Pluggable Authentication Modules (PAM)
- Yetkilendirme (rol-bazlı)
 - System administrator
 - Data administrator
 - Application administrator
 - User

Guardium, BigInsights üzerindeki veri aktivitelerini izleyebilmektedir

İlşkili mesajlar collector'a gönderilir

InfoSphere Guardium S-TAP



InfoSphere Guardium collector



Access Rule Description	Client IP	Server IP	DB User Name	Full SQL String
Access Rule: sensitive files: Alert9.70.145.1189.70.145.113SVORUGA				__WGPB message {struct:1='getFileinfo',struct:2={struct:1='/user/svoruga/testme',struct:3='org.apache.hadoop.hdfs.protocol.ClientProtocol',varint:4=1}}
Access Rule: sensitive files: Alert9.70.145.1189.70.145.113SVORUGA				__WGPB message {struct:1='getListing',struct:2={struct:1='/user/svoruga/testme',struct:2='',varint:3=0},struct:3='org.apache.hadoop.hdfs.protocol.ClientProtocol',varint:4=1}}

InfoSphere Guardium raporlama ve alarm

- BigData sunucu kaynaklarına ve ağ trafiğine minimal etki
- Görevlerin ayrıştırılması – teftişe yönelik bilgiler ayrı bir güvenli sunucuda

Nasıl başlayabilirsiniz ?

- Cloud / Bulut Bilişim Sistemleri
 - RightScale, ya da Amazon, Rackspace, IBM Smart Enterprise Cloud, ya da özel bulut sistemleri üzerinde
 - Sadece kullandığınız kaynak kadar ödersiniz
- Sanal Sınıf
 - Ücretsiz eğitim : www.bigdatauniversity.com
- Kendi Sunucularınızda
 - Basic Edition'ı ücretsiz indirebilirsiniz
- Sınıf içi Eğitimler



The screenshot shows the BigDataUniversity website interface. At the top, there is a navigation bar with links for HOME, LEARN, DOWNLOAD, RESOURCES, JOBS, and LEARN Hadoop. A search bar is also present. The main content area features a video player on the left, a central text block with the heading "Why register?" and several bullet points, and a right sidebar with a "Study Made Easy!" section and "Student Testimonials". A prominent "sign up now" banner is overlaid on the right side of the page.

BigDataUniversity BETA
Learn from the industry's best

English

login sign up

HOME LEARN DOWNLOAD RESOURCES JOBS LEARN Hadoop

search courses

Why register?

- **Easy and Affordable**
Learning Hadoop and other Big Data technologies has never been more affordable! Many courses are FREE!
- **Latest industry trends**
Acquire valuable skills and get updated about industry's latest trends right here. Today!
- **Learn from the Experts!**
Big Data University offers education about Hadoop and other technologies by the industry's best!
- **Learn at your Own Pace!**
Find everything right here when you need it and from wherever you are.

sign up now
for a chance to receive:
FREE Books
SIGN UP

Study Made Easy!

Student Testimonials go to sign up

Balázs (USA)
The training material has short, easy to digest videos, with separately available transcripts allowing to execute the same commands after watching the videos. Each command line in the transcript can be cut and pasted making the examples easy to reproduce. There is very good support for Windows users both on 32bit and 64bit. Everything works right out of the box as described in the course materials. Online support on the Course Forums is excellent, most questions were answered before even I encountered the issue. The curriculum associates the chapters from the "Getting Started with"

about us | legal | contact | bug reports

Analitik Uygulamalar

BI / Raporlama | Analiz/ Görselleştirme | Fonksiyonel Uyg. | Endüstri Uyg. | Veri Madenciliği | İçerik Analitiği

IBM Big Data Platformu

Görselleştirme
ve Keşif

Uygulama
Geliştirme

Sistem
Yönetimi



Hızlandırıcılar

Hadoop
Sistemi



Akışkan
veri işleme



Veri
Ambarı



Veri Entegrasyonu ve Sahipliği

Operasyon Analizi: İhtiyaç



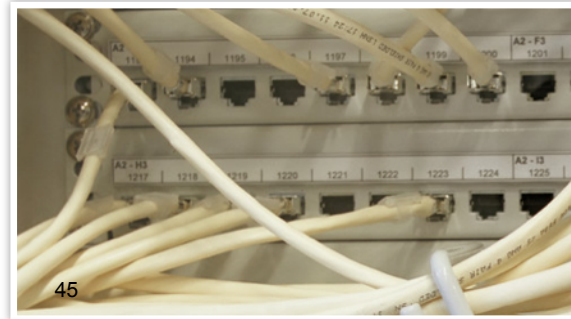
Gelişkin iş sonuçları için, çeşitli cihazların ürettiği verilerin analiz edilmesi

Machine Data Analytics Accelerator

Ne yapar ?

- Çok çeşitli log ve makine verisinin toplanıp, işlenmesi ve ilişkilendirmesini sağlar

1	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	Cisco - Num
2	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	AnyVendor
3	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	PNOC - Int
4	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	PNOC - Int
5	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	AnyVendor
6	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	Cisco - Num
7	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	AnyVendor
8	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	AnyVendor
9	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	PNOC - Int
10	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	PNOC - Int
11	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	AnyVendor
12	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	AnyVendor
13	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	Cisco - Int
14	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	Cisco - Num
15	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	PNOC - Int
16	11.1.2.4	Oct 21 06:33:45	hex('1043882D078A4D16') syslog	%L2-BDF	PNOC - Int



Makine Verilerinin Analizi Temelde Zorlu Sorunlar Barındırıyor

Veri Kaynakları ve Entegrasyon



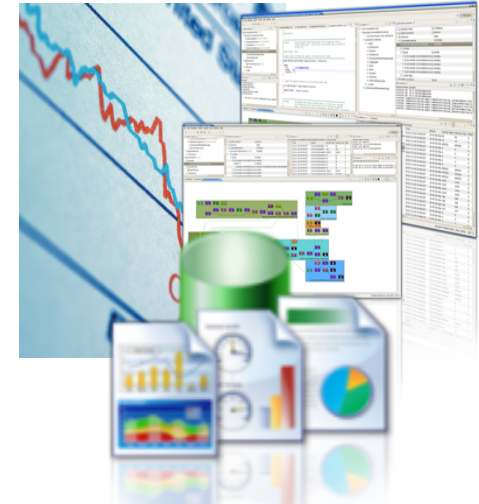
- Standart olmayan kompleks formatlar
- Yüksek hacim
- Kurumsal ve makine verisi karışımı
- Sürekli akan ve biriken veriler
- Birbiri arasında ilişkilendirmede tutarsızlıklar (zaaman damgası farklılıkları, IP adres formatları, zaman dilimi vs.)

Analitik



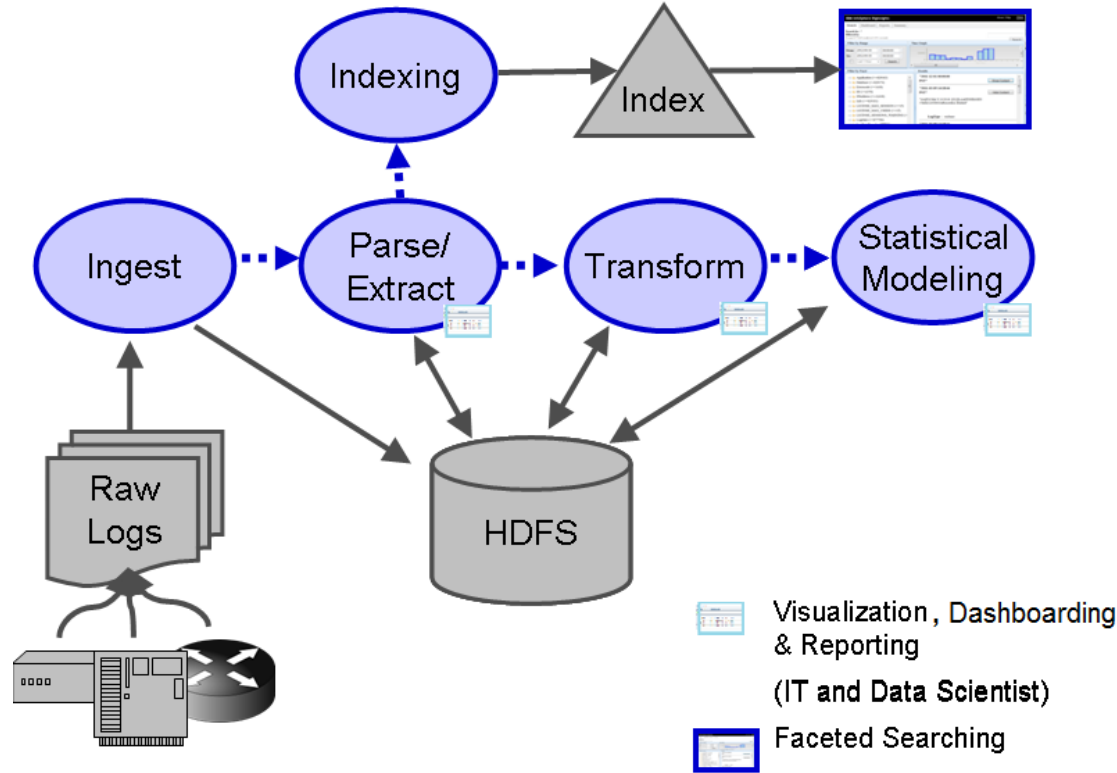
- Yüksek hacimli indeksleme
- Farklı veri setleri arasında korelasyon
- Farklı veri biçimleri için ileri düzey analitik

Görselleştirme / Aksiyon ve Çıktı



- Akan ve büyük hacimli veriler için yeni görselleştirme yetkinlikleri
- Gerçek zamanlı göstergeler
- Coğrafi zenginleştirme
- Büyük hacimli veriler üzerinde gezinti

Machine Data Accelerator – üst seviye akış



- Oturum bilgilerinin büyük veriler içinden ayıklanması
- Korelasyon
- Tekrarlanan serilerin analizi
- KPI ve olay ilişkileri
- Neden sonuç analizleri

BigInsights Avantajları

- Adaptive MapReduce
 - Dengeli performans iyileştirme
- Parçalanabilir LZ0-tabanlı Sıkıştırma
 - Daha hızlı sıkıştırma
 - IBM-LZO – Sadece IBM’de
- Güvenlik
 - Açık Port Sayısının azaltılması
 - SFTP ve LDAP desteği
- Big Index
 - IBM’in Lucene’e olan eklentisi
- Gelişkin Araçlar
 - Kurulum, Geliştirme & Yönetim
- “App Store”
 - Uygulama geliştirme ve yayınlama
 - 3. parti yazılımlar satın alma imkanı
- Big Sheets
 - Tablo görünümlü web bazlı arayüz
 - Veri toplama eklentileri
 - Görselleştirme eklentileri
- Text bazlı Analitik
- Hive SQL Editor, Eclipse
- Yüksek hızlı RDBMS erişimi
 - Netezza, DB2/InfoSphere Veri Ambarı
 - DataStage ile Parallel HDFS yazma/okuma
- Tek Hata Noktasının kaldırılması
 - GPFS-SNC