IBM Big Data Forum - Think Big

Büyük Veri Perdesini Aralıyoruz

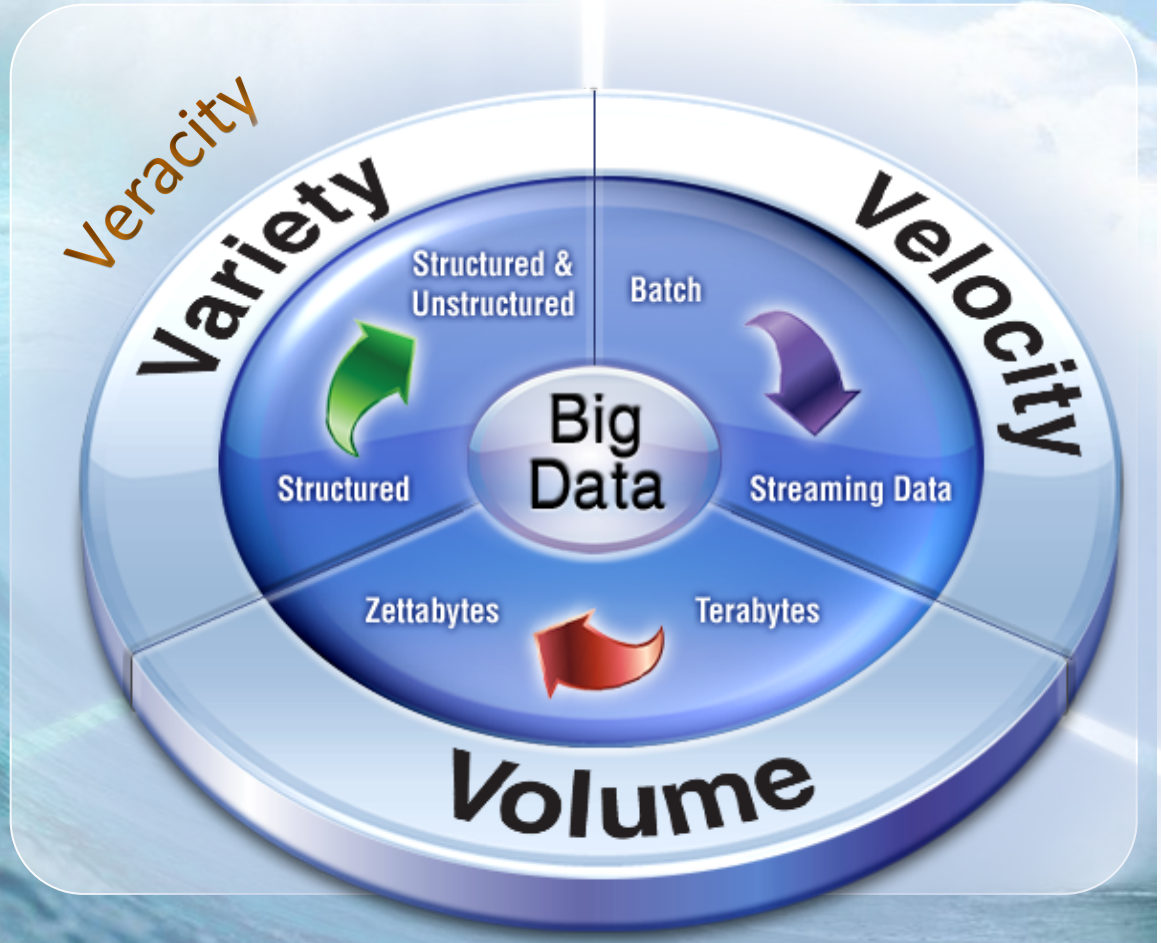
Veri Ambarı Yetkinliklerinizi Büyük Veri ile Genişletin

Cüneyt Göksu, VBT
IBM Gold Consultant
IBM Champion for Data Management

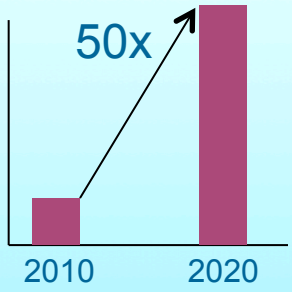
Ayhan Önder, IBM
BigData Technical Specialist, ayhano@tr.ibm.com

Big Data

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



Volume



35 ZB

Velocity

30 Billion RFID sensors and counting

Variety

80% of the worlds data is unstructured

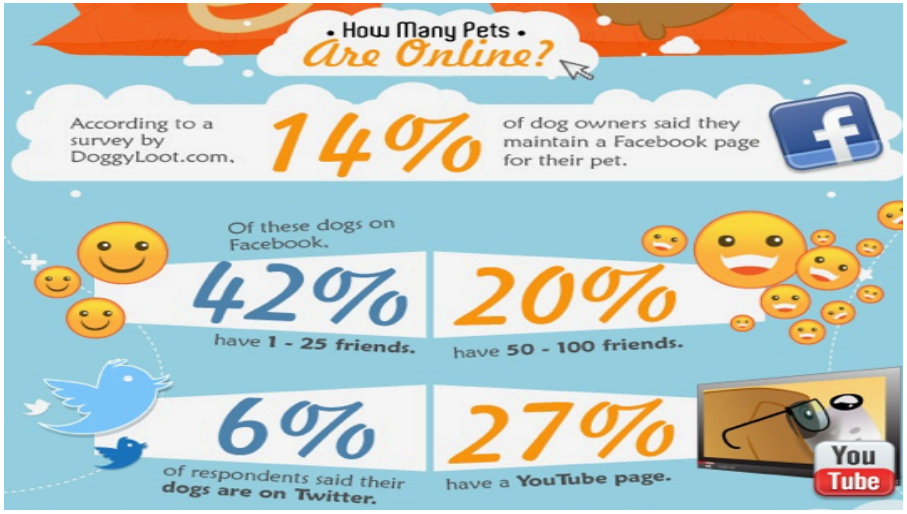
Veracity

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

Value

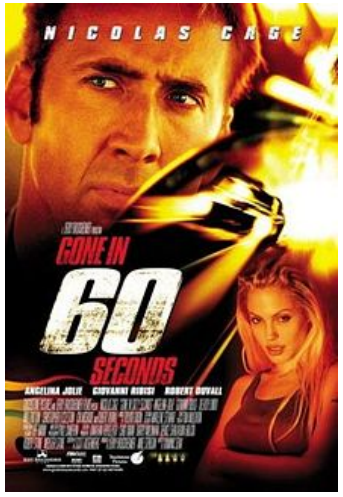
3 V + V = (V)alue (Değer)

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

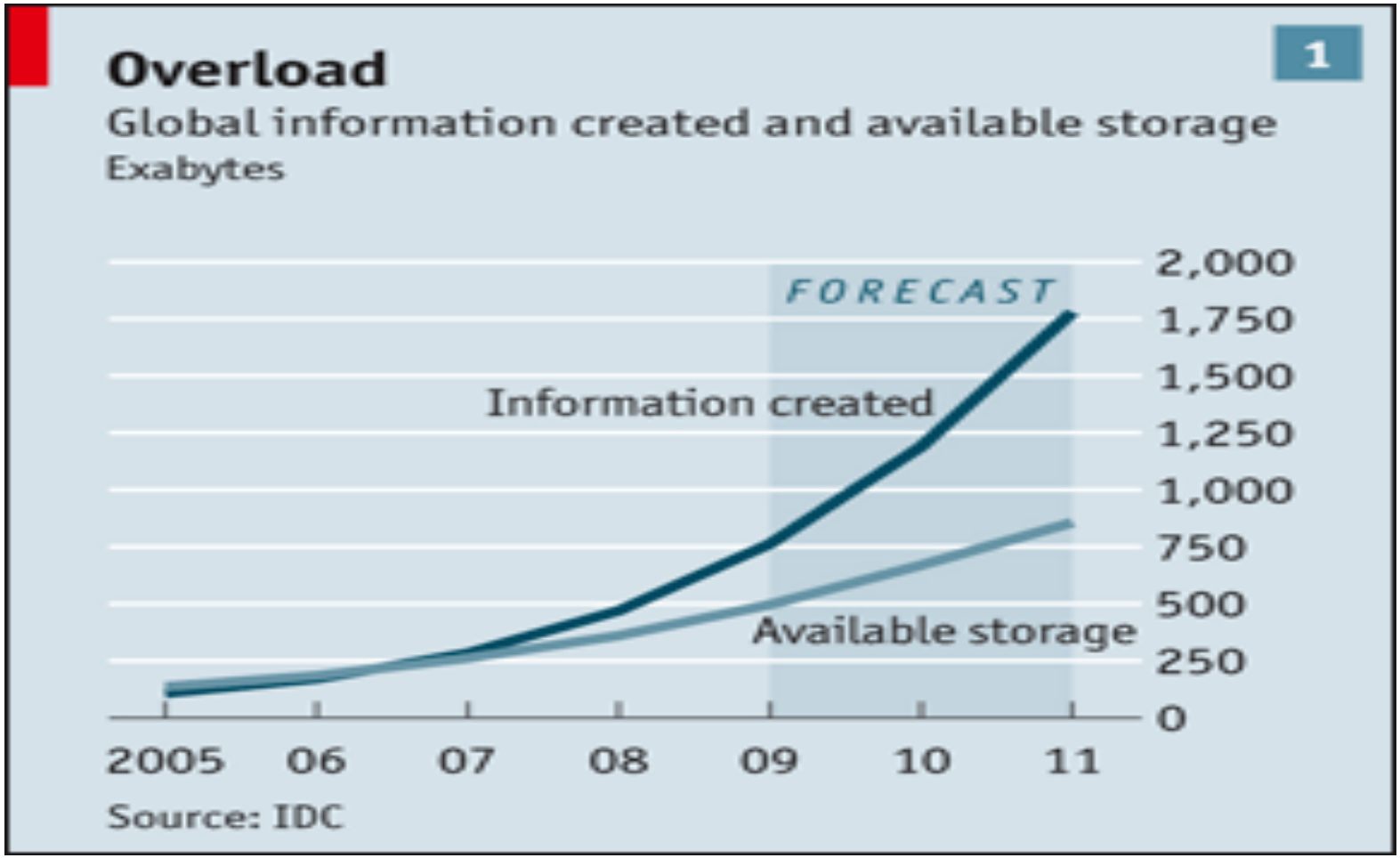


70 yeni domain kaydı
13000 iPhone uygulaması indirimi
700,000 Facebook güncellemesi
168 milyon e-mail

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



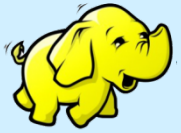
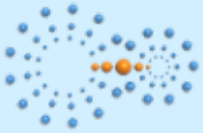
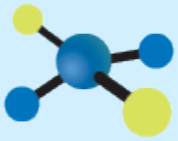
Bu verileri saklayacak yer yok!



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

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

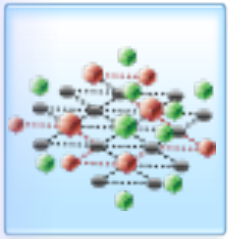
Büyük Veri Problemi Farklı Teknlojik Yetkinlikler Gerektiriyor

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-işlemden geçirilerek, değerli bilgilerin ortaya çıkarılması

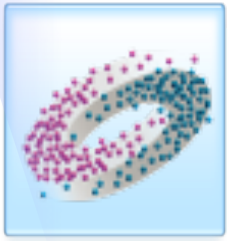


Yapısal ve yapısal olmayan veriler üzerinde derinlemesine analitik

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 anlizler yapılması ve aksiyon alınabilmesi



Keşifsel analizler için yapısal verilerin tekrar kullanımı

Deneysel çalışmalar, ad-hoc analiz ve görselleştirme

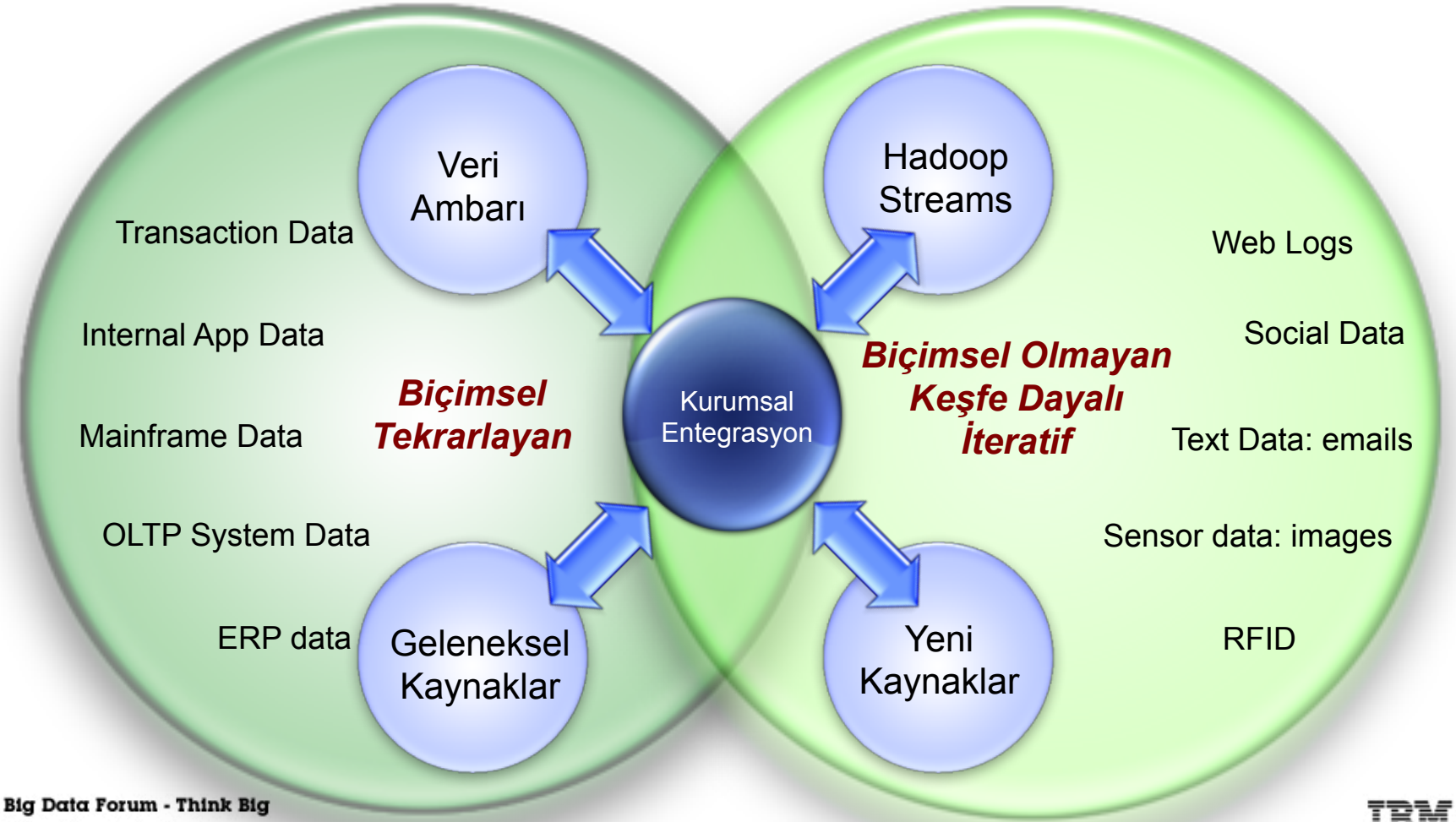
Farklı Uygulama Alanları için Birbirini Bütünleyen Yaklaşımlar

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 keşiflere yönelik bir platform ve veri zenginliği sunar

Son kullanıcılar
neler keşfedilebilir araştırır, tarihsel veriye özgürce erişirler

Data Scientist!

CENTRAL INTELLIGENCE AGENCY

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 360º 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

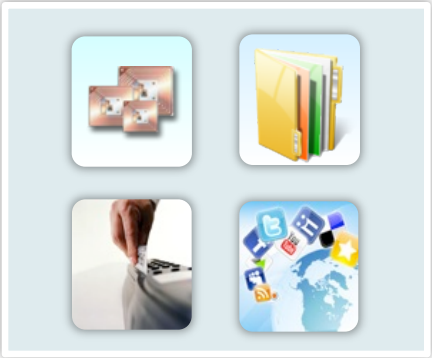
Makinelerin ürettiği farklı çeşitlilikteki verilerin analiz edilerek iş sonuçlarının geliştirilmesi

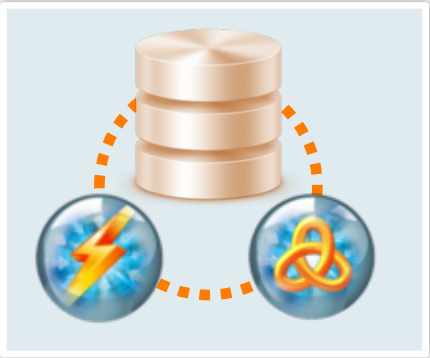
Veri Ambarı Eklentisi

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

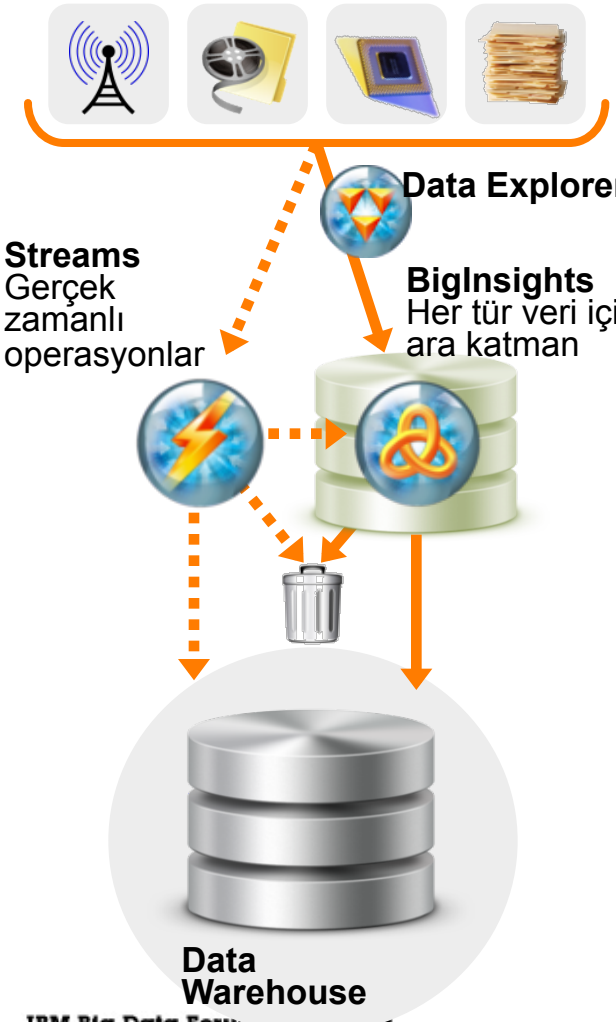
- Yapısal, yapısal olmayan, ve akan verilerin analitik süreçlerde efektif kullanımı
- Düşük gecikme gereksinimleri
- Sorgulanabilir veri ortamı

Veri Ambarının Genişletilmesi

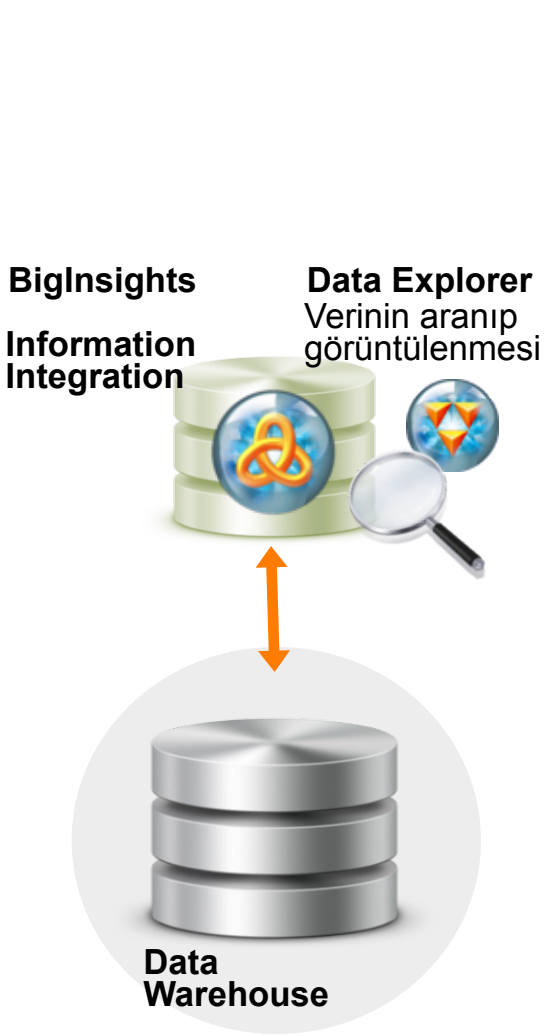
- Nadir kullanılan verilerin taşınabileceği, veri saklama, lisans ve bakım maliyetleri için uygun çözüm
- Akan verilerin akıllıca işlenmesi sayesinde saklama maliyetlerinden kazanç
- Filtreden geçirilmiş veriler sayesinde daha hızlı veri ambarı performansı

Veri Ambarı Eklentisi : Farklı İhtiyaçlar

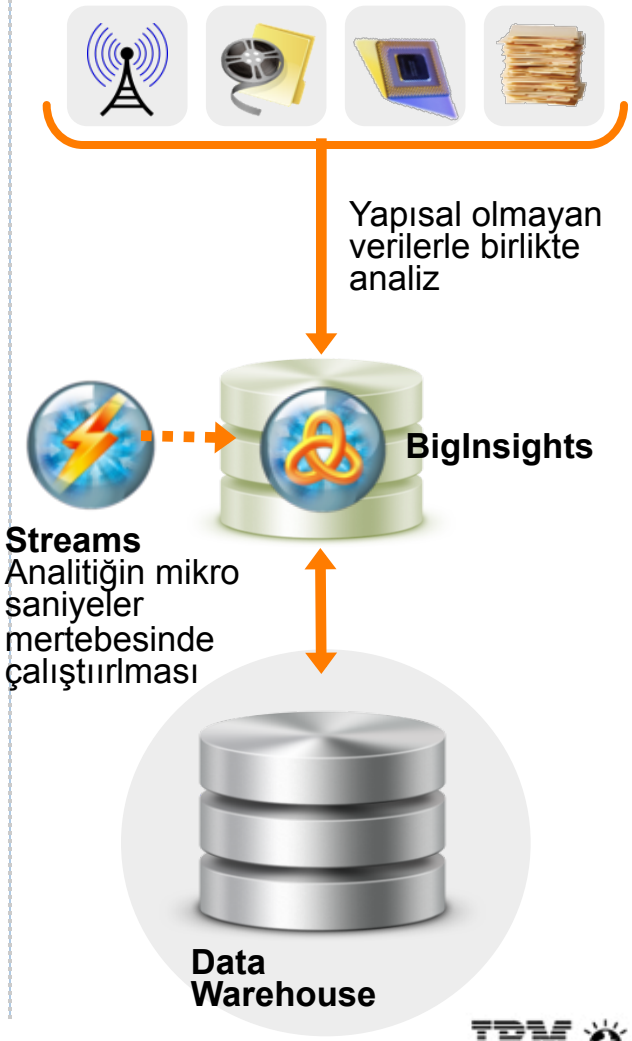
1 Ön-İşleme



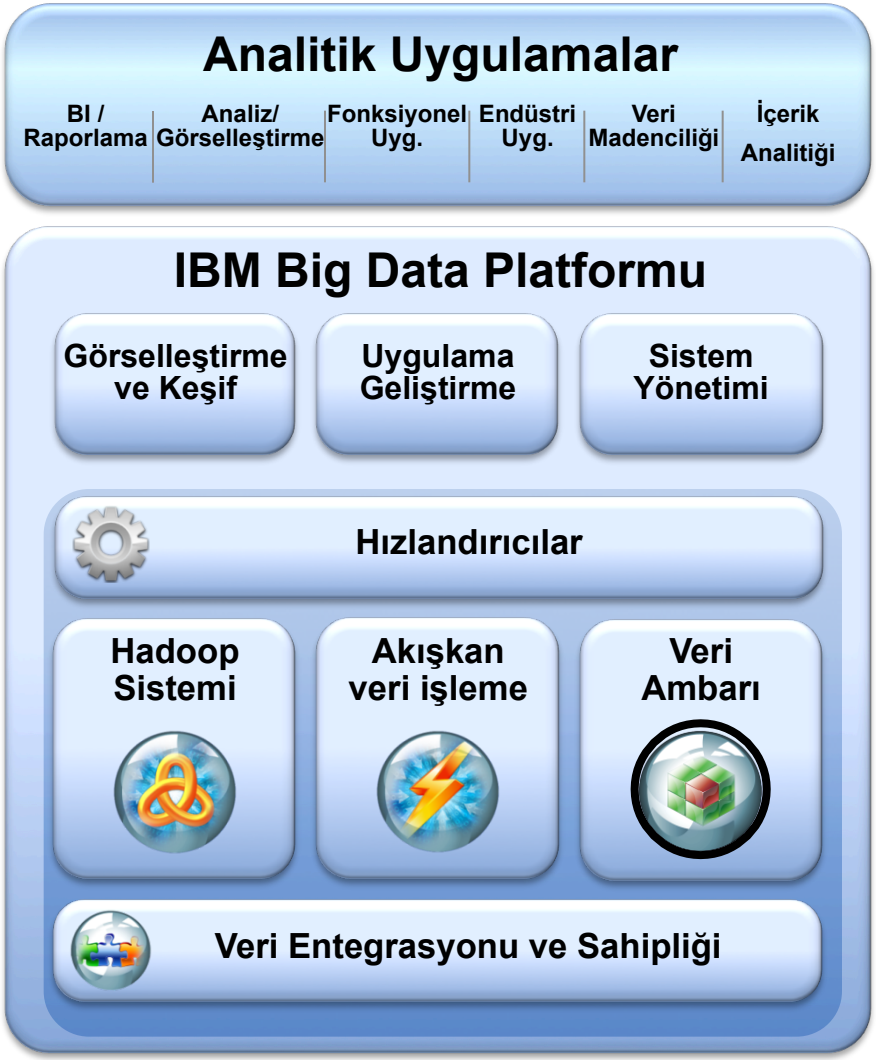
2 Sorgulanabilir Arşiv



3 Keşifsel Analiz

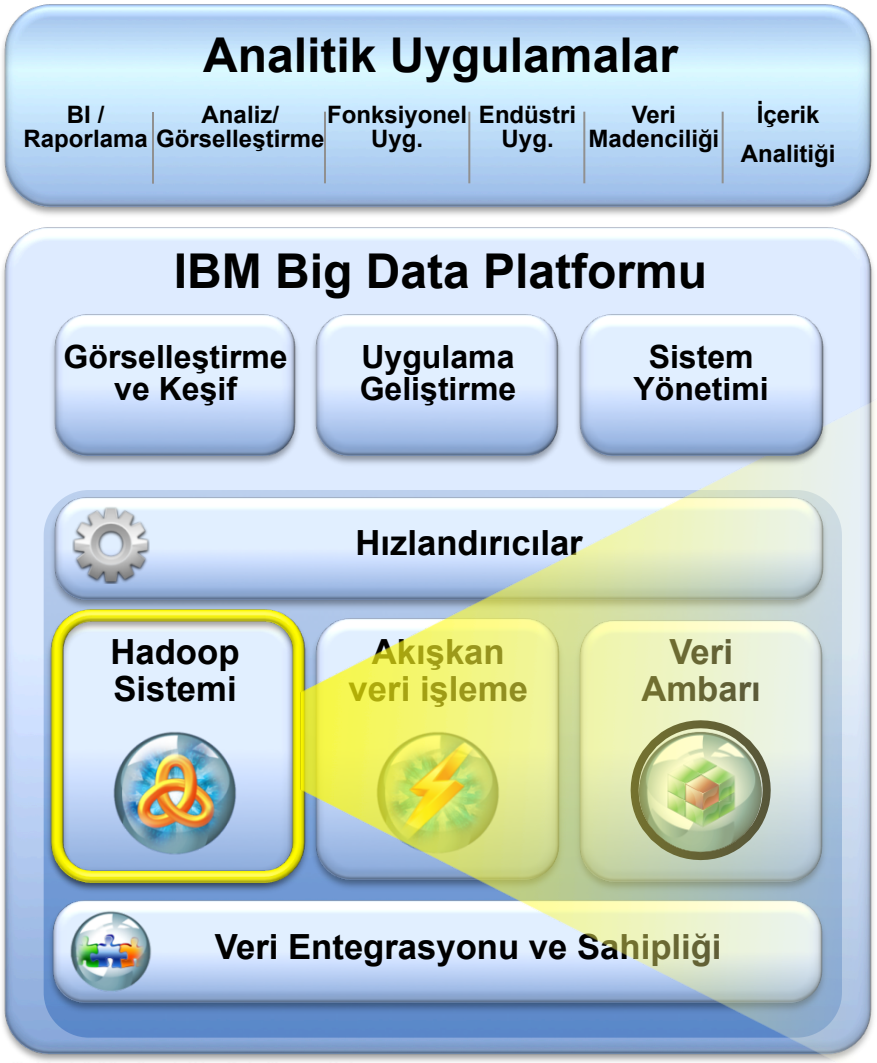


IBM Big Data Stratejisi: Analitiği Veriye yaklaştırın



- 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 görselleştirebimesi
- Yeni nesil analitik uygulamalar geliştirebilmek için geliştirme araçları
- İşyükü optimizasyonu ve zamanlama
- Güvenlik ve Veri Sahipliği

BigInsights, Hadoop yetkinliklerini kurumunuza getiriyor

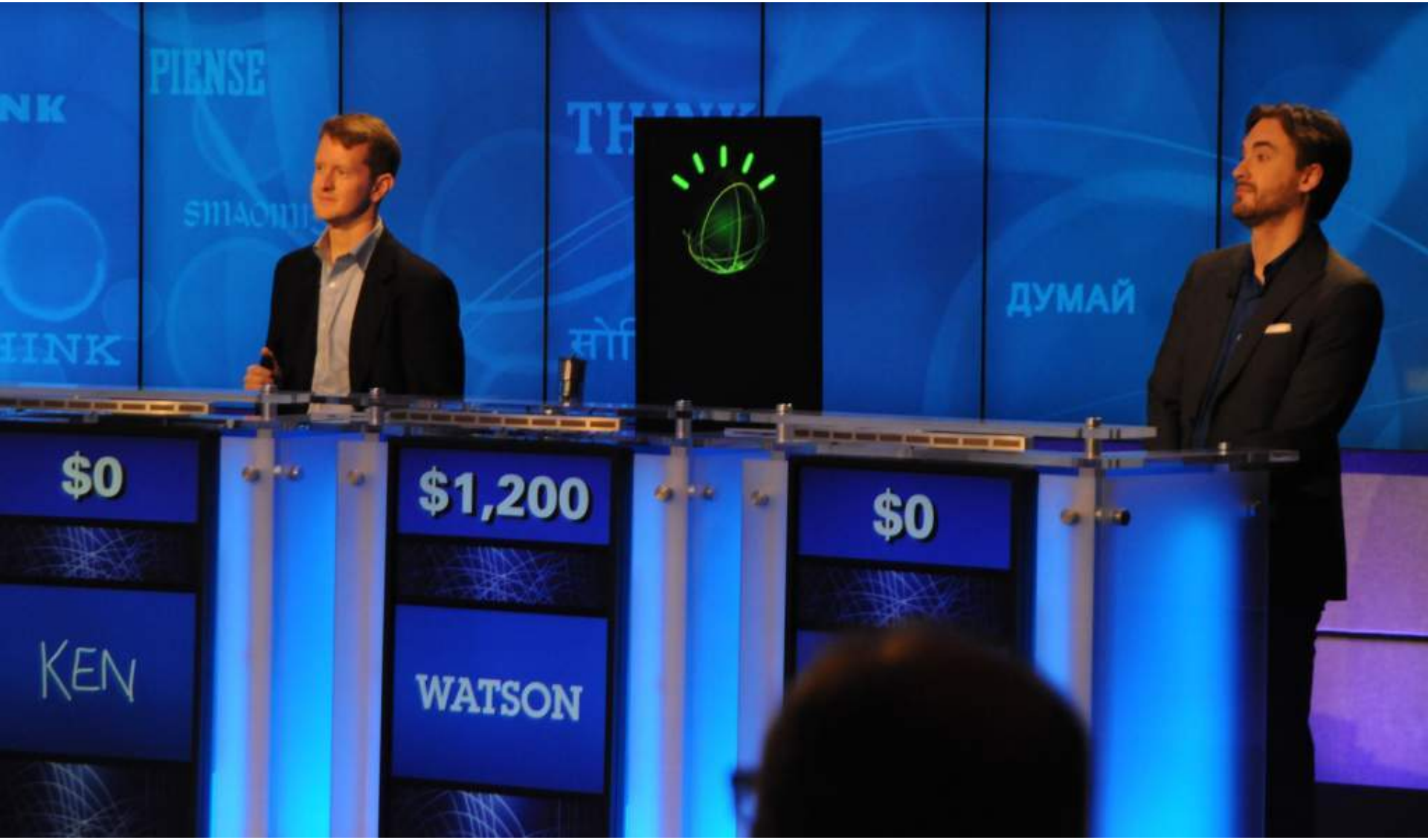


IBM InfoSphere BigInsights

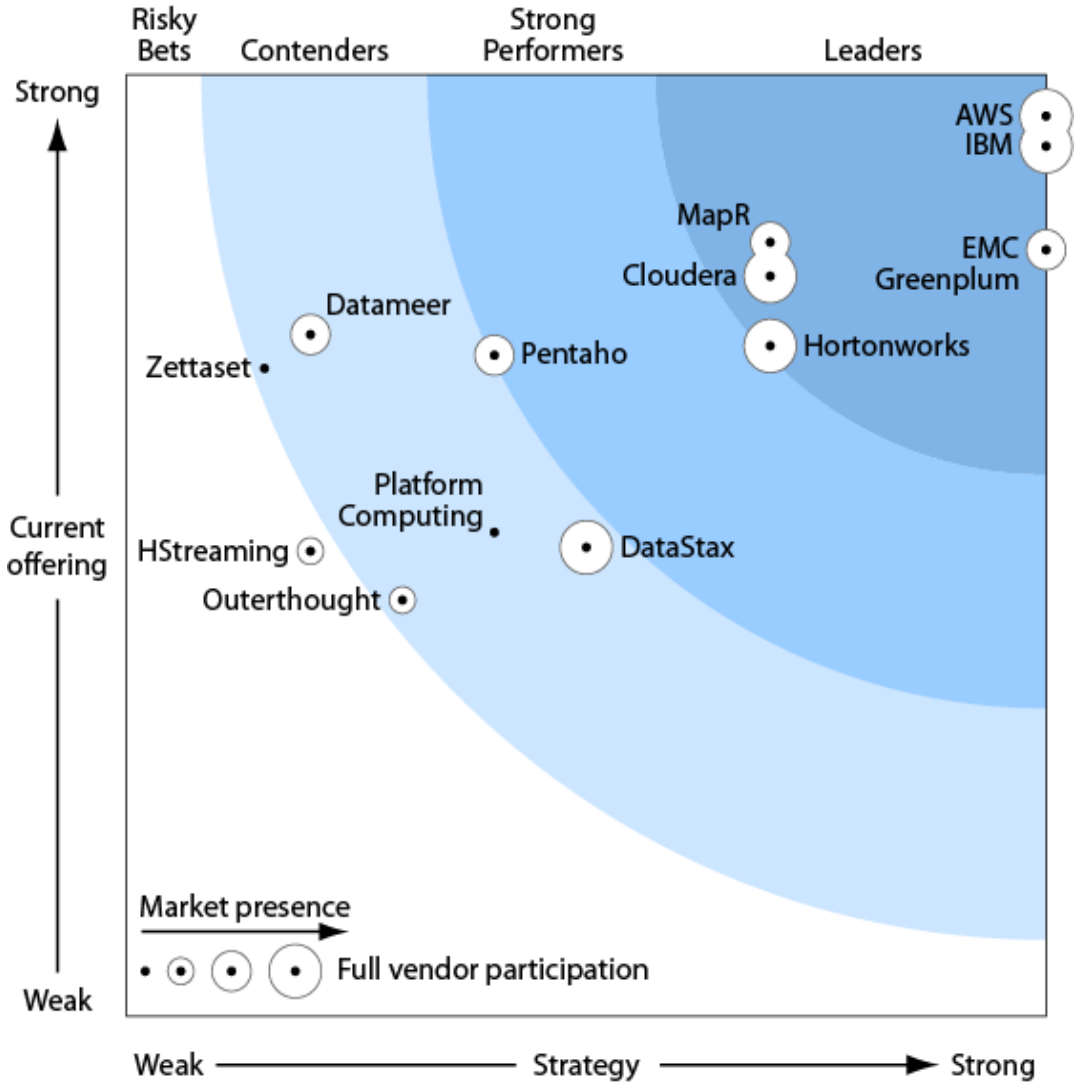
Açık kaynak kodlu Hadoop çözümünü tamamlıyor;

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

IBM Watson



Endüstideki 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

IBM

Vestas optimizes capital investments based on 2.5 Petabytes of information

Need

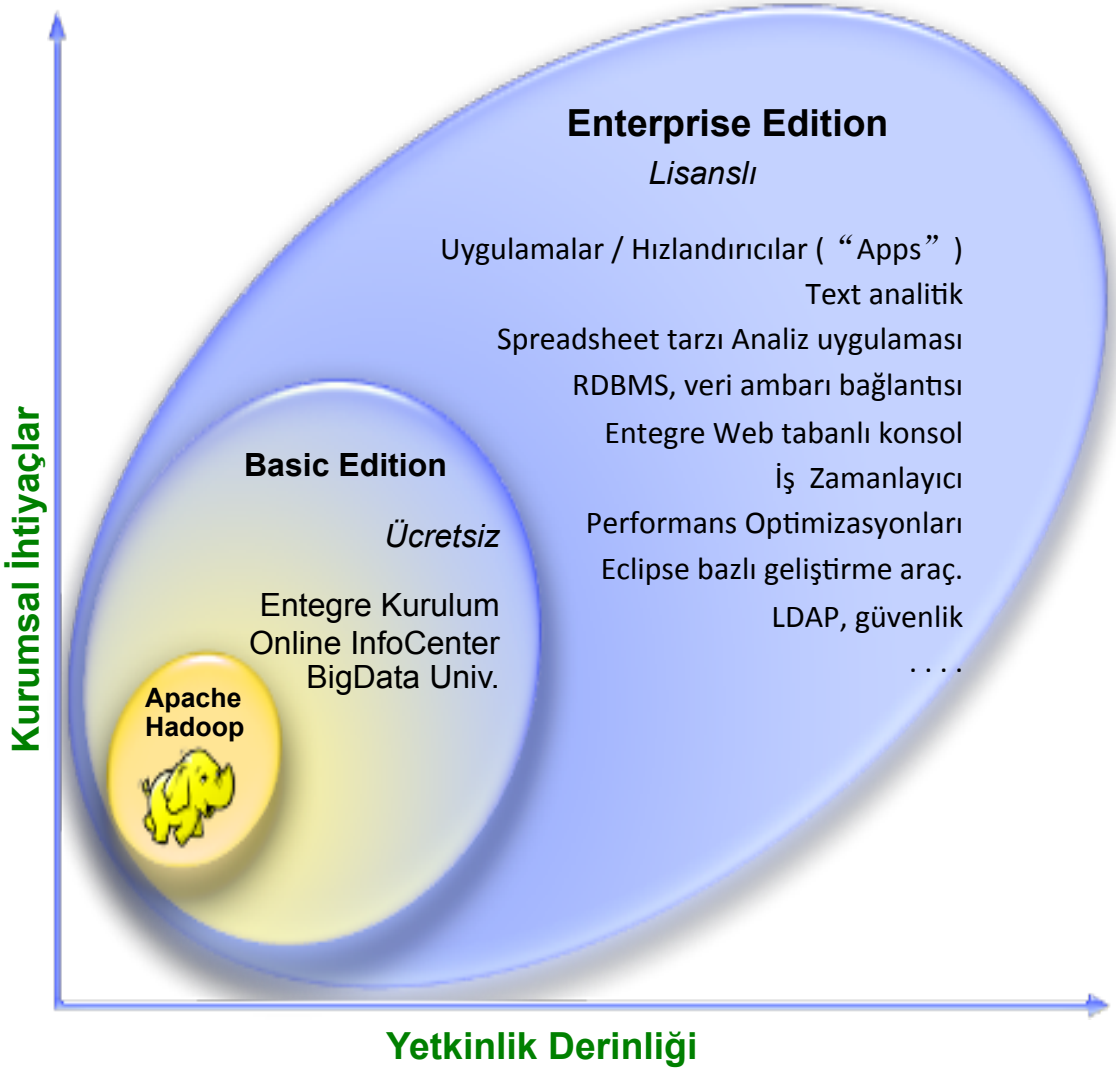
- Model the weather to optimize placement of turbines, maximizing power generation and longevity

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

Vestas®

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"

- Web Crawler
- Boardreader
- Distrib file copy
- . . .
- DB export
- DB import
- Hive query
- Pig query
- Jaql query

Infrastructure

- Integrated installer
- Enhanced security
- Text compression
- BigIndex
- Flexible scheduler
- ZooKeeper
- Oozie
- Lucene
- Jaql
- HBase
- Adaptive MapReduce
- GPFS
- Pig
- Hive
- MapReduce
- HDFS

Connectivity

- JDBC
- Netezza
- DB2
- Streams
- R
- Flume
- Streams*

Administrative and development tools

Web console

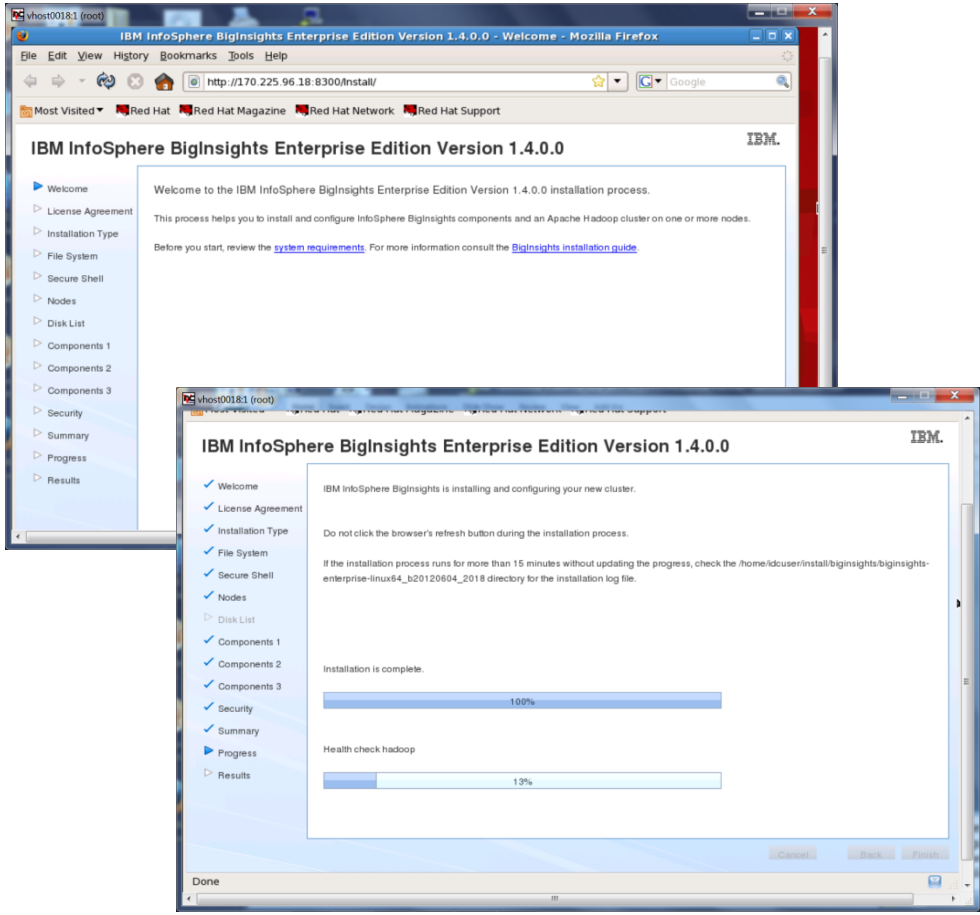
- Monitor cluster health
- Add / remove nodes
- Start / stop services
- Inspect job status
- Inspect workflow status
- Deploy apps
- Launch apps / jobs
- Work with distrib file system
- Work with spreadsheet interface
- Support REST-based API
- . . .

Eclipse plug-ins

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

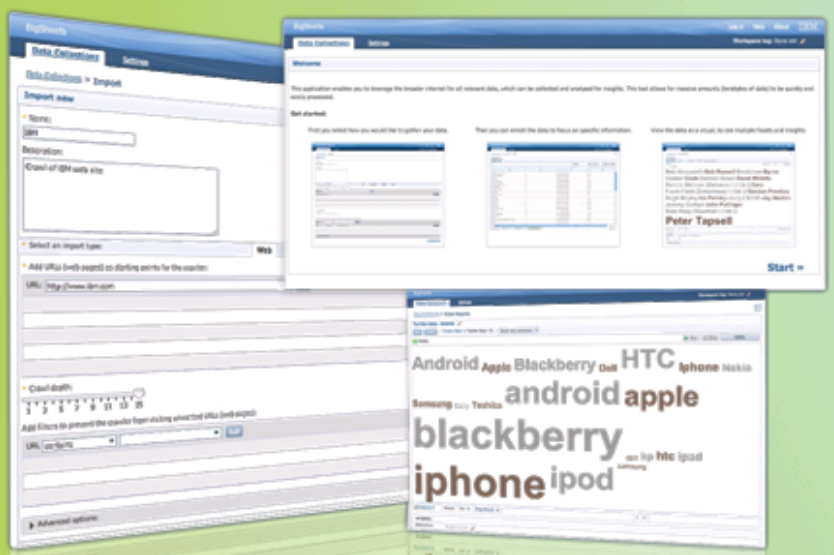
Web Arayüzlü Kurulum

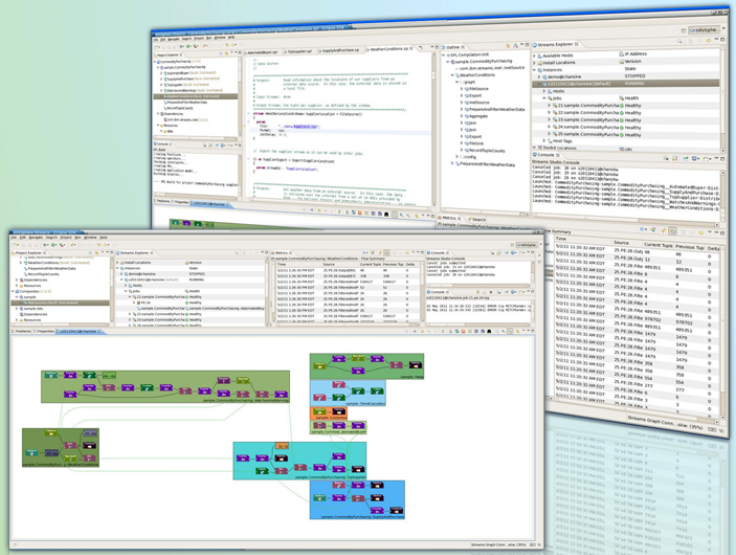
- İster tek sunucu üzerine, ister çok sayıda sunucu üzerine son derece basit bir şekilde Hadoop Cluster kurulumu
- Tüm seçilen bileşenlerin otomatik kurulumu
- Kurulum sonrasında validasyon ve sağlık kontrolü; hem IBM ve hem açık kaynak kodlu programların tamamı için



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

Kullanıcı ve Geliştiriciler için Basit Arayüzler





Son Kullanıcı Görselleştirme

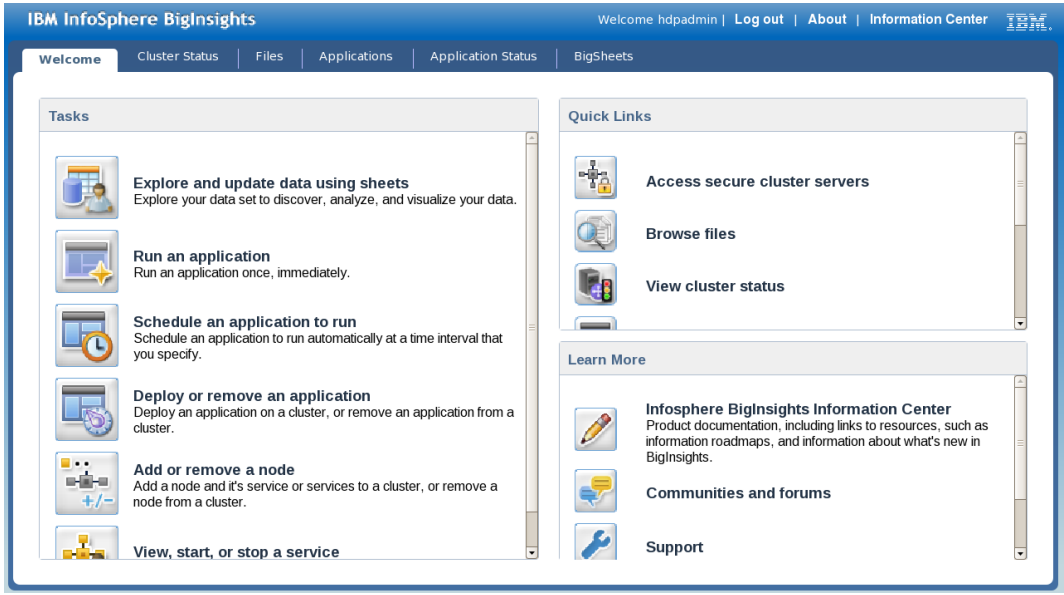
Veri keşfi, veri toplama (crawling), ve analitik

Geliştirme ve Yönetim Araçları

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

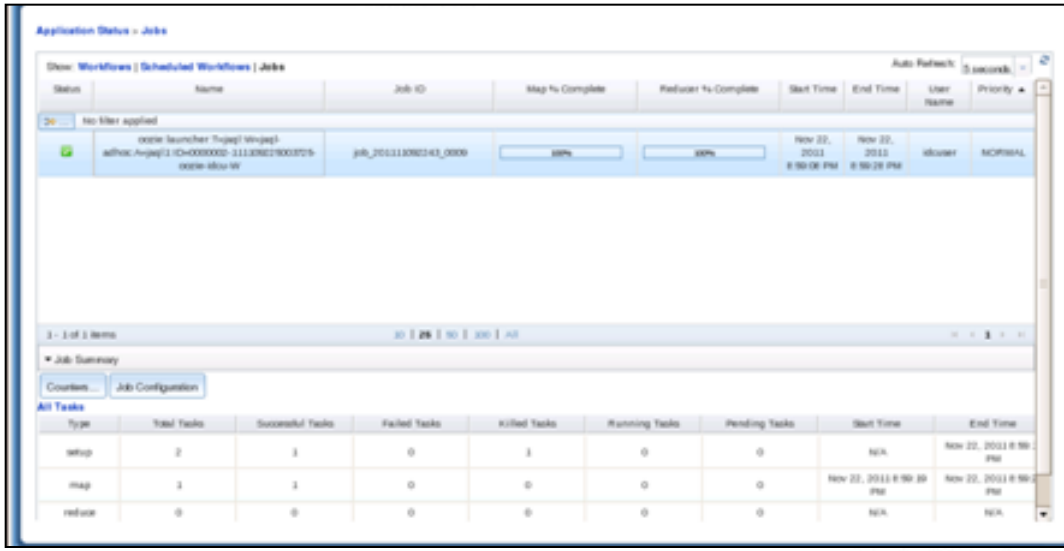
Web Arayüzlü Yönetim

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

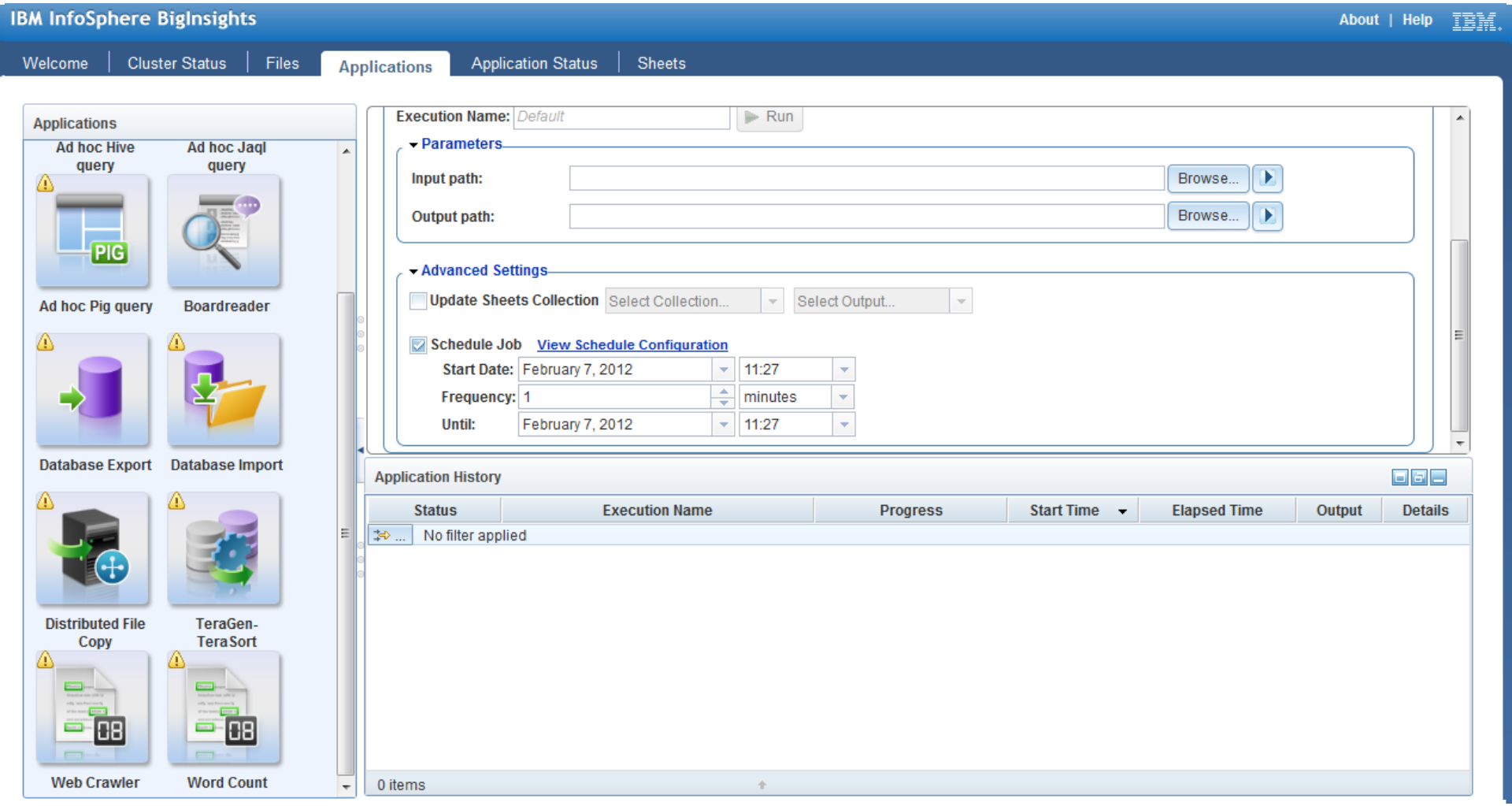


- Uygulamaları başlatma
  - Örn. BigSheets
  - Hazır uygulamaları çalıştırma

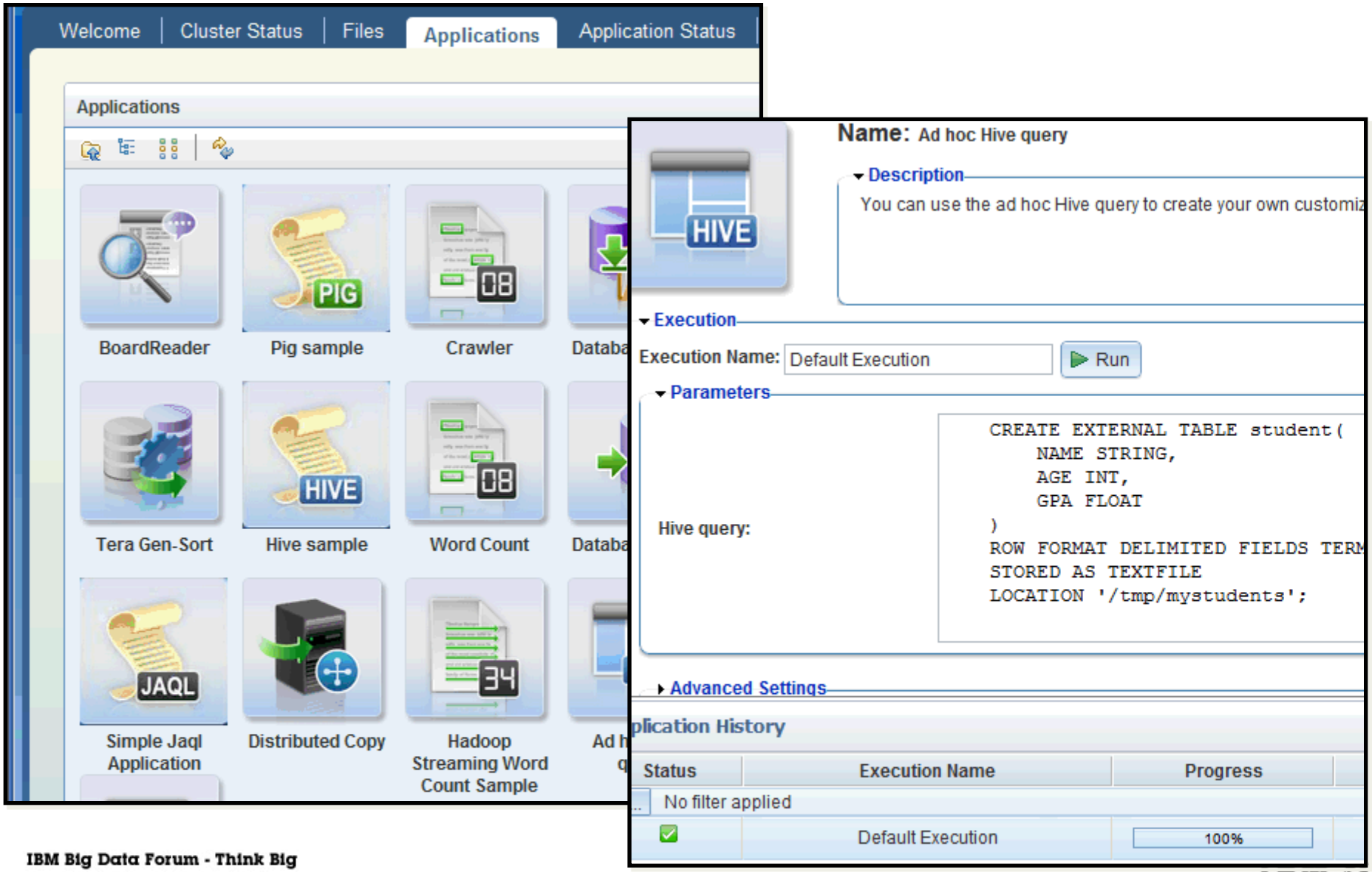
  (IBM kaynaklı veya geliştirdiğiniz uygulamalar)
- Uygulamaların yayınlanması
- Yardım sayfaları



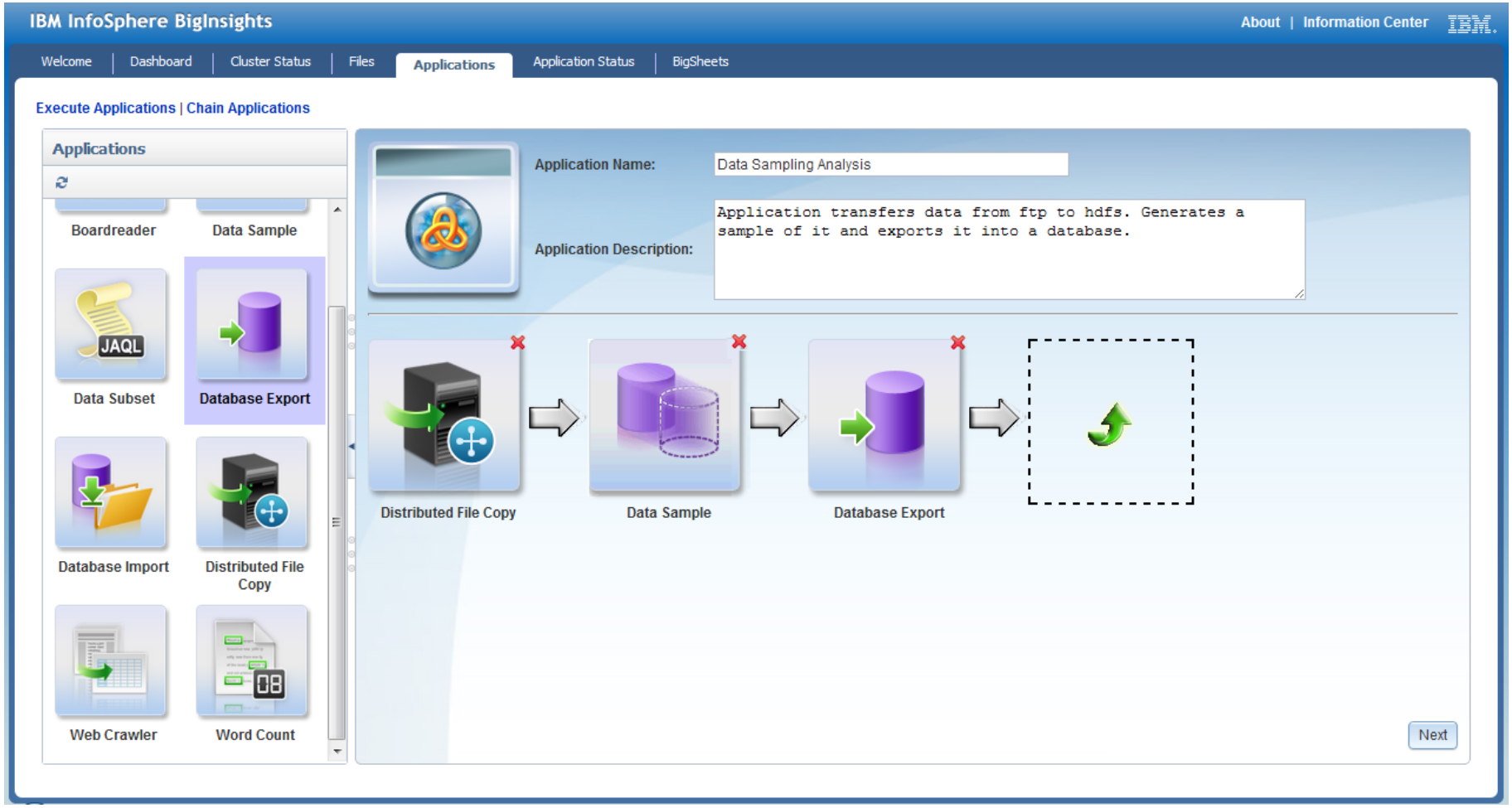
Son Kullanıcı Deneyimi : Uygulamaların Çalıştırılması



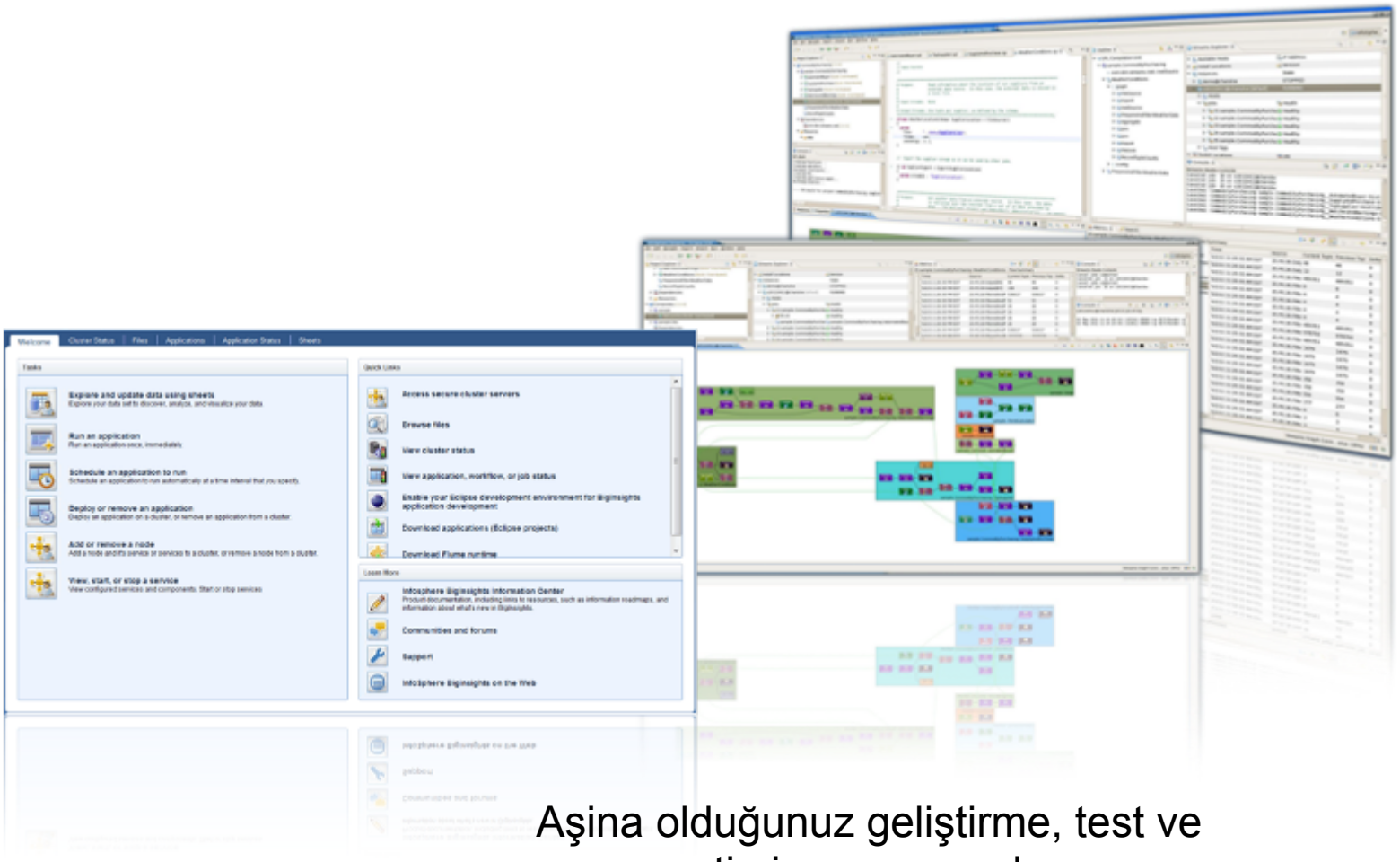
Web tabanlı uygulama erişim ara yüzü; HIVE sorgusu



Son kullanıcılar da basitçe yeni uygulamalar geliştirebilir

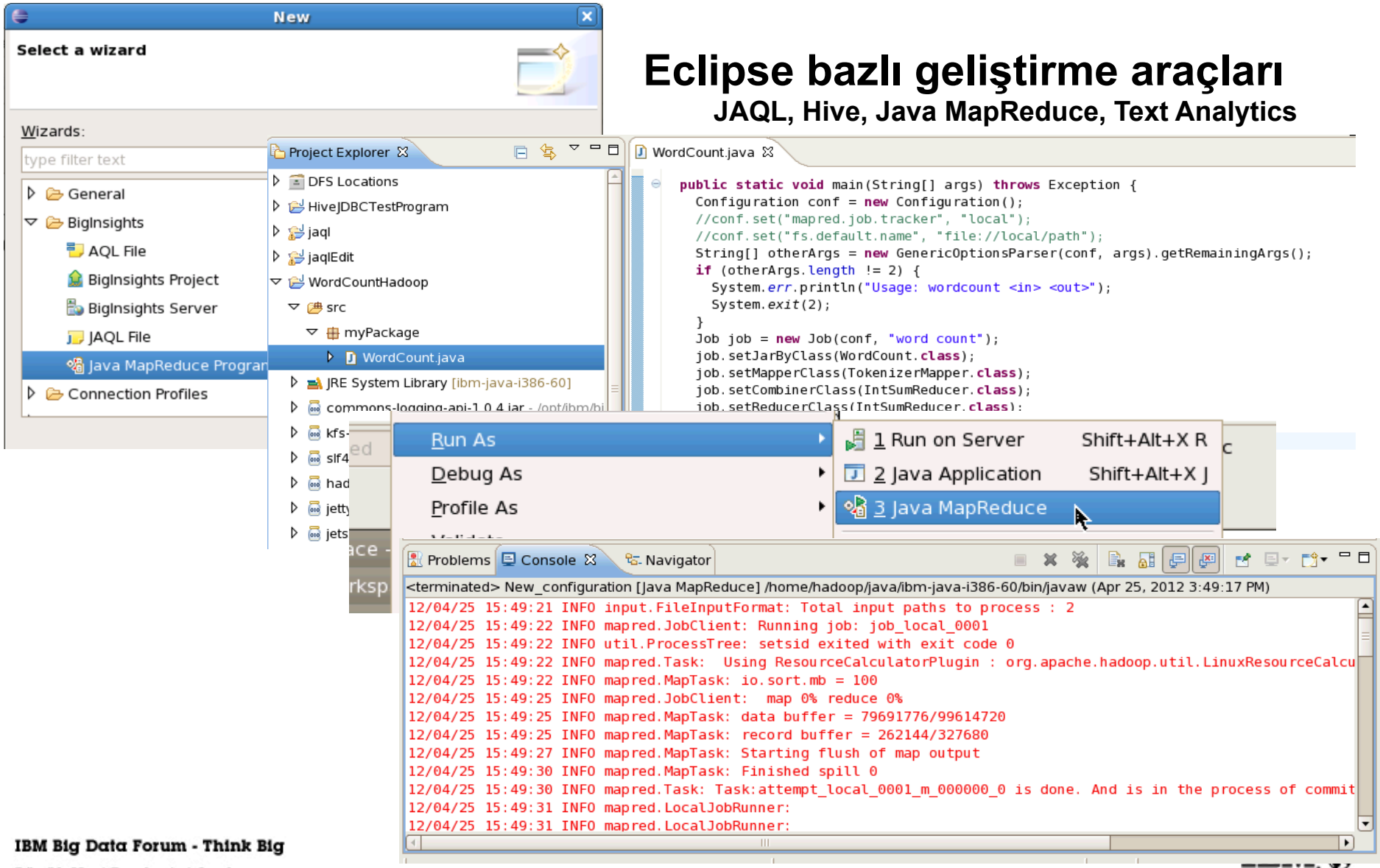


Profesyonel Geliştirme ve Yönetim Araçları



Aşina olduğunuz geliştirme, test ve optimizasyon araçları

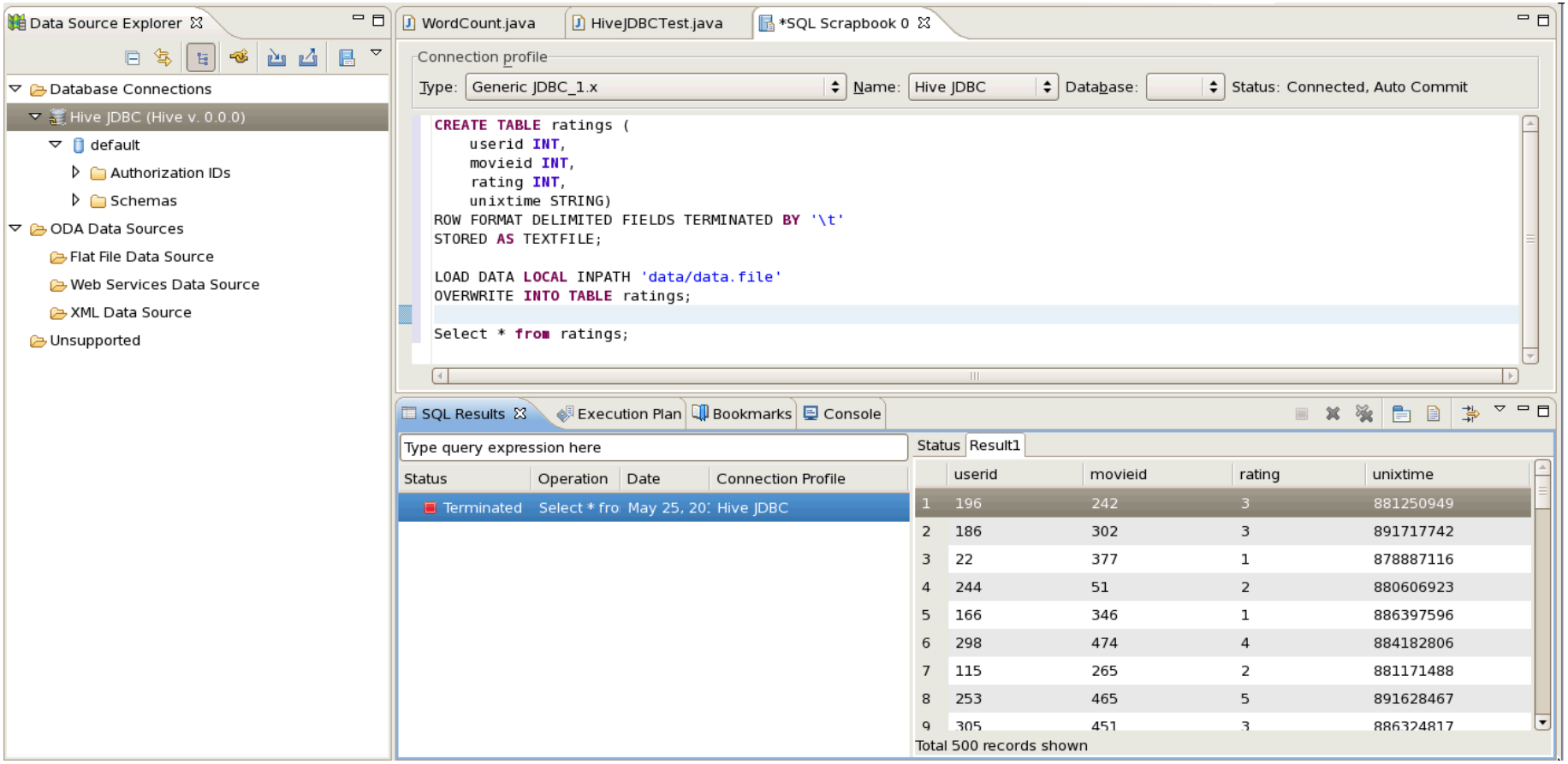
Big Data Programları geliştirme – Map Reduce örneği



Eclipse bazlı geliştirme araçları

JAQL, Hive, Java MapReduce, Text Analytics

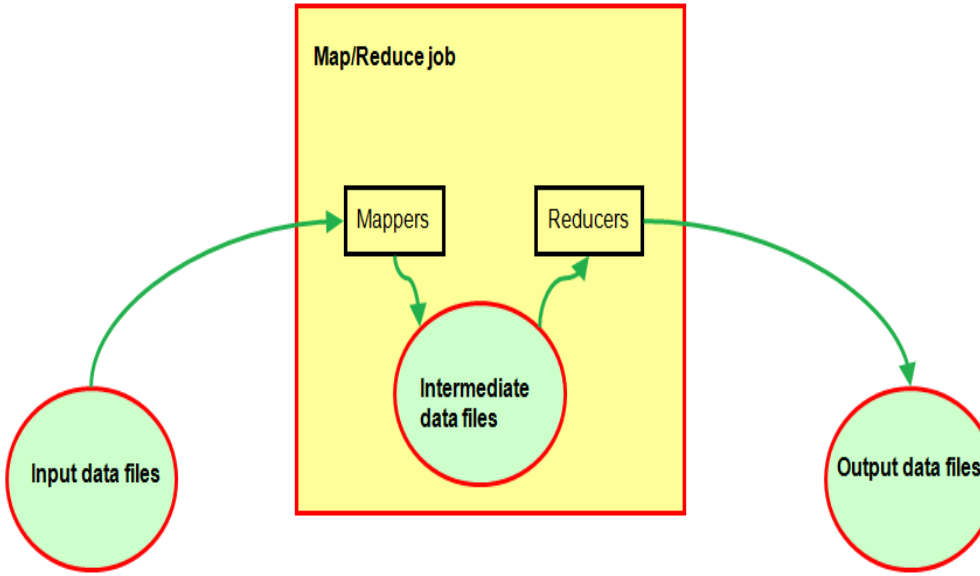
SQL Editör arayüzü ile Hive sorguları çalıştırma ve sonuçları görüntüleme



Sıkıştırma

- Temelde LZO algoritması baz alınmıştır
- IBM'in kendi tasarımı
  - lzma ve bzip2'den çok daha hızlı
  - Makul sıkıştırma performansı (~60%)

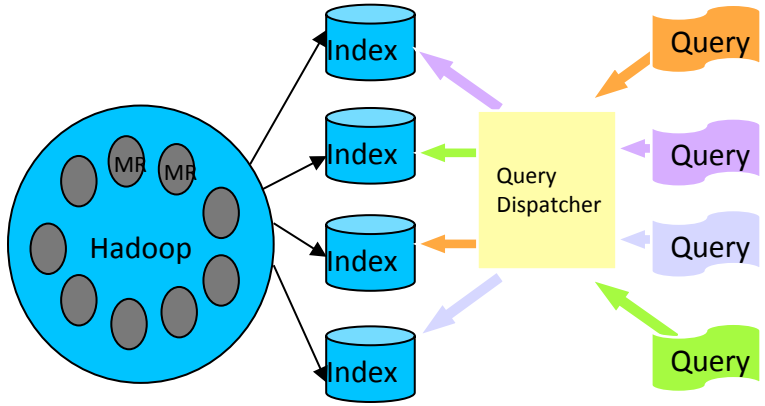
- Faydaları :
  - Bütünüyle parçalanalabilir
  - Index dosyasına ihtiyaç duymaz

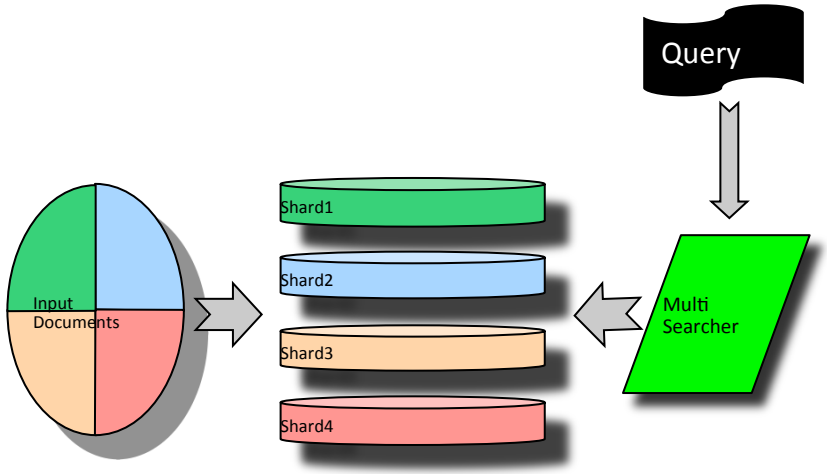


	Size (Mbytes)	Comp . speed (sec)	Comp . memory used (MBytes)	Decomp. speed	Decomp. memory used (Mbytes)
uncompressed	96				
gzip	23	10	0.7	1.3	0.5
bzip2	19	22	8	5	4
lzo	36	1	1	0.6	0
lzm	18	63	14	3	1.8

Big Index - büyük boyutlu indeksleme

- Dağıtık Indeks ve Arama
  - **Parallel index**
    - Indeks operasyonları paralel çalışır, fakat bir indeks olarak saklanır
  - **Partitioned index**
    - Indeks 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

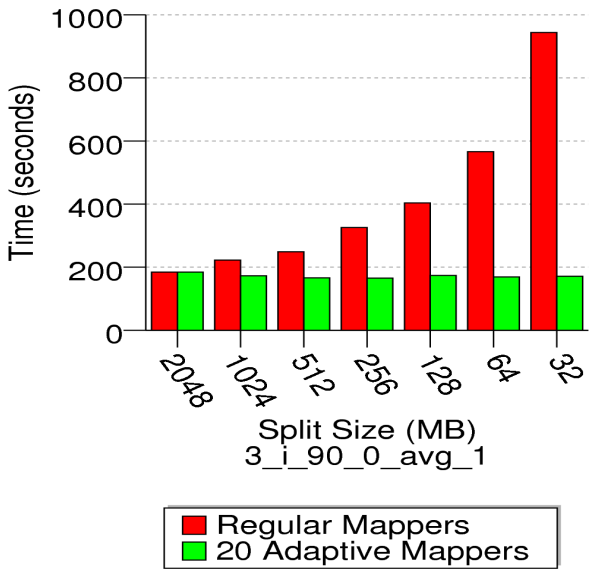
- **Geleneksel MAP işlemlerinin ayağa kalkması aşağıdaki durumlarda çok maliyetli olabilmektedir**
  - **Ufak dosyalar ve dosya parçacıkları üzerinde çalıştığında**
  - **MAP işlemi basit hızlı işlemler gerçekleştirdiğinde**
- **Adaptive MR**
  - **Ufak dosyaların/parçaların dinamik olarak bütünleştirilerek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

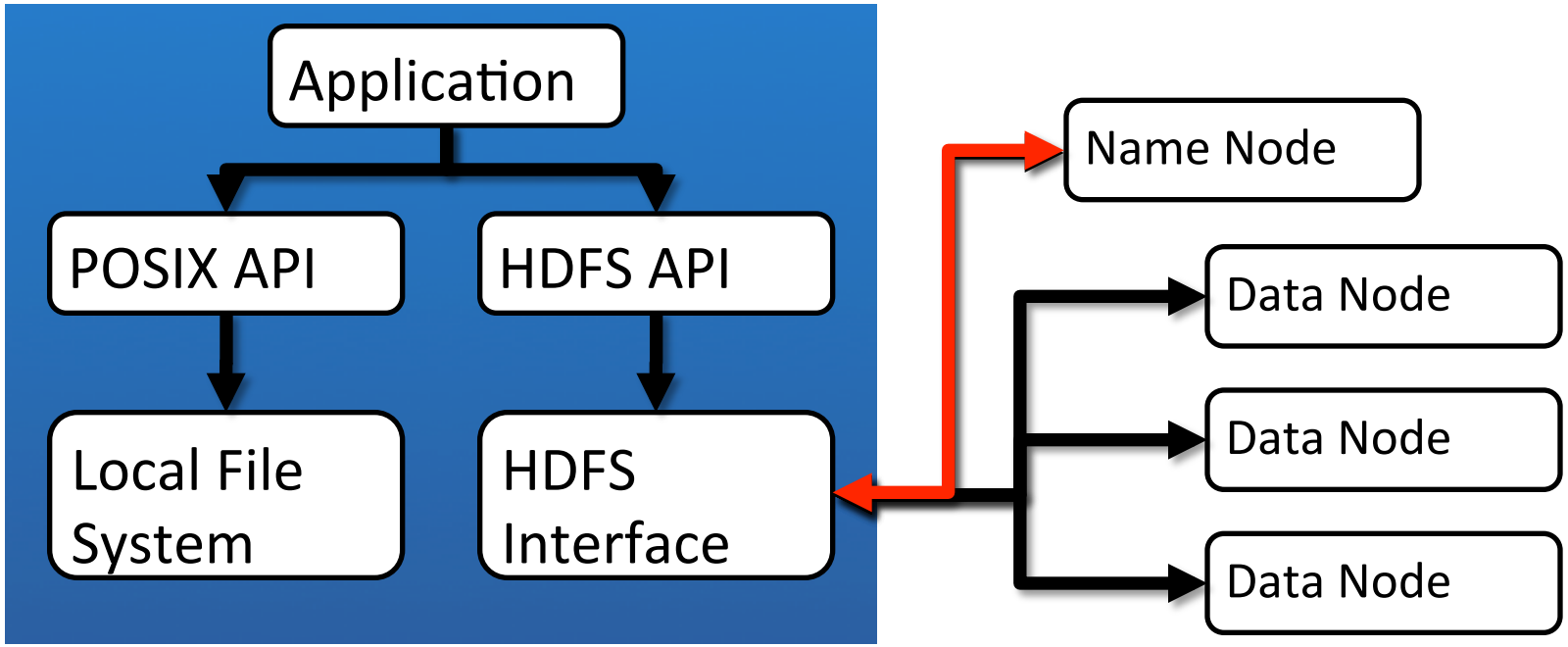
Startup/Scheduling cost (e.g. loading reference data)

map task processing



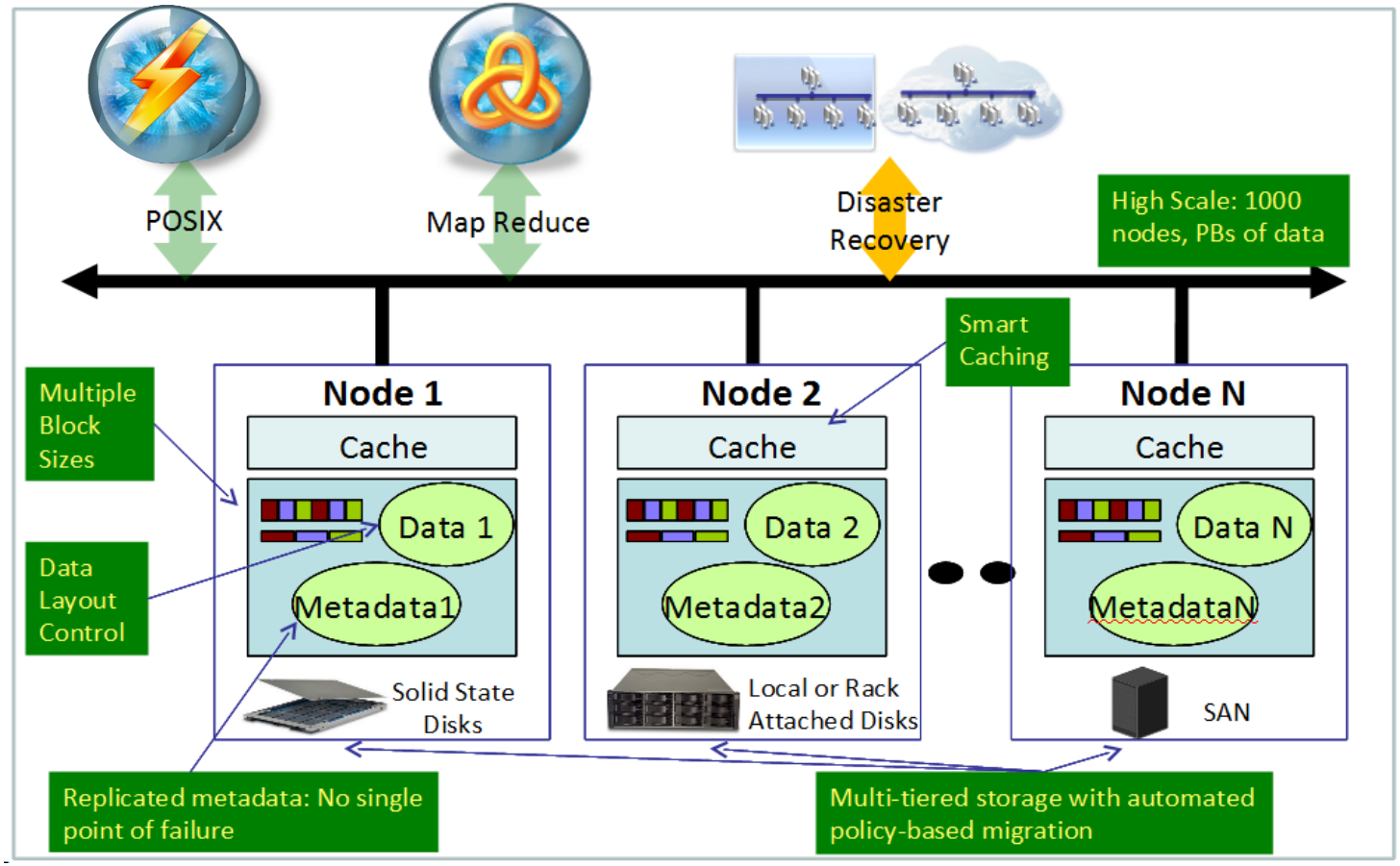
HDFS

- Hadoop Distributed File System
- Single Point of Failure



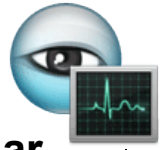
GPFS-FPO

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



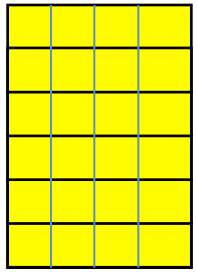
BigInsights ve Veri Ambarı

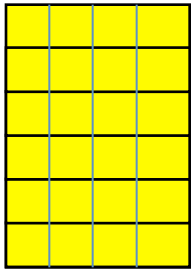
Big Data analitik uygulamalar

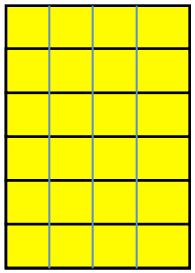


Geleneksel Analitik

Veri Ambarı



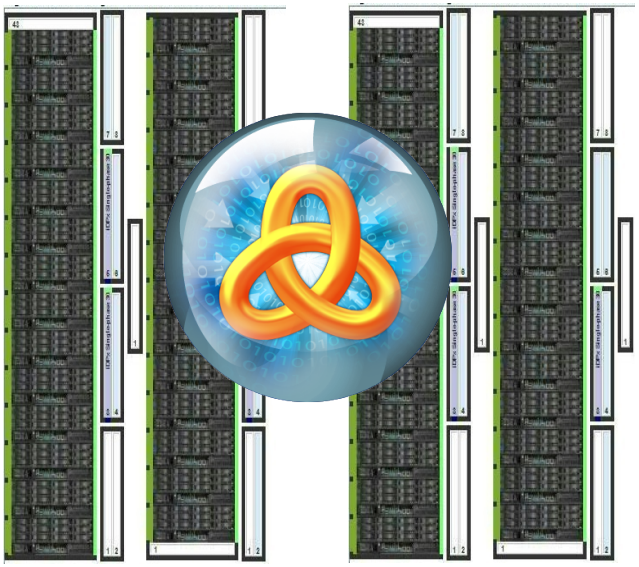




facebook

twitter

BigInsights



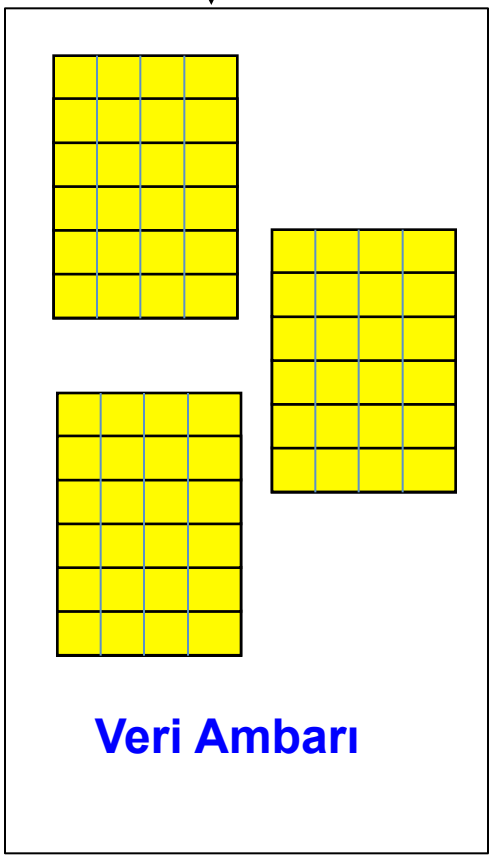
Filtreleme *Transformasyon* *Özetleme*

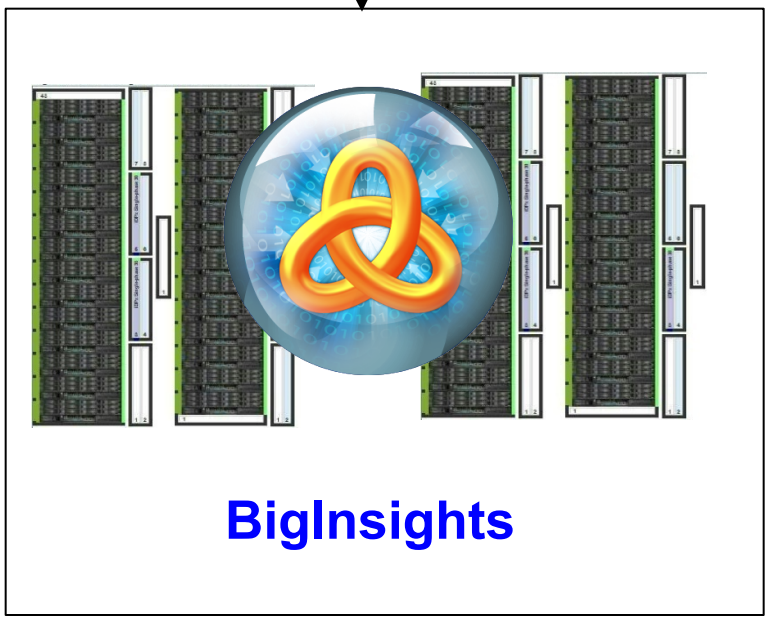
IBM

BigInsights ve Veri Ambarı

Geleneksel Analitik

Big Data Analitik





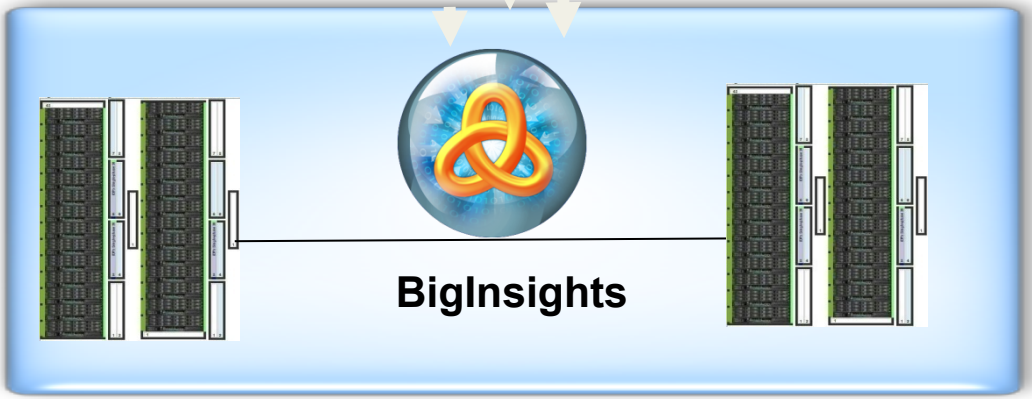
- ***Sorgulanmaya hazır soğuk data, veri ambarı arşivi, yapısal olmayan, yarı yapısal veriler***

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

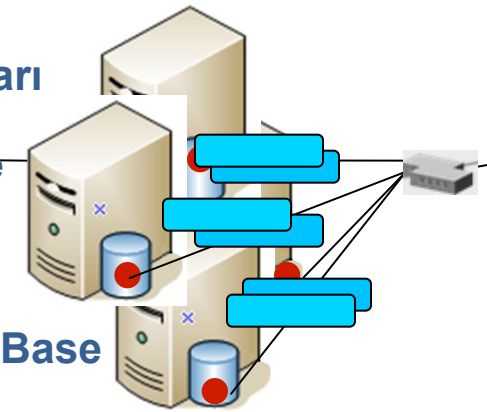
InfoSphere Guardium S-TAP

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

Hive sorguları

MapReduce işleri

HDFS ve HBase komutları



Cluster

InfoSphere Guardium collector

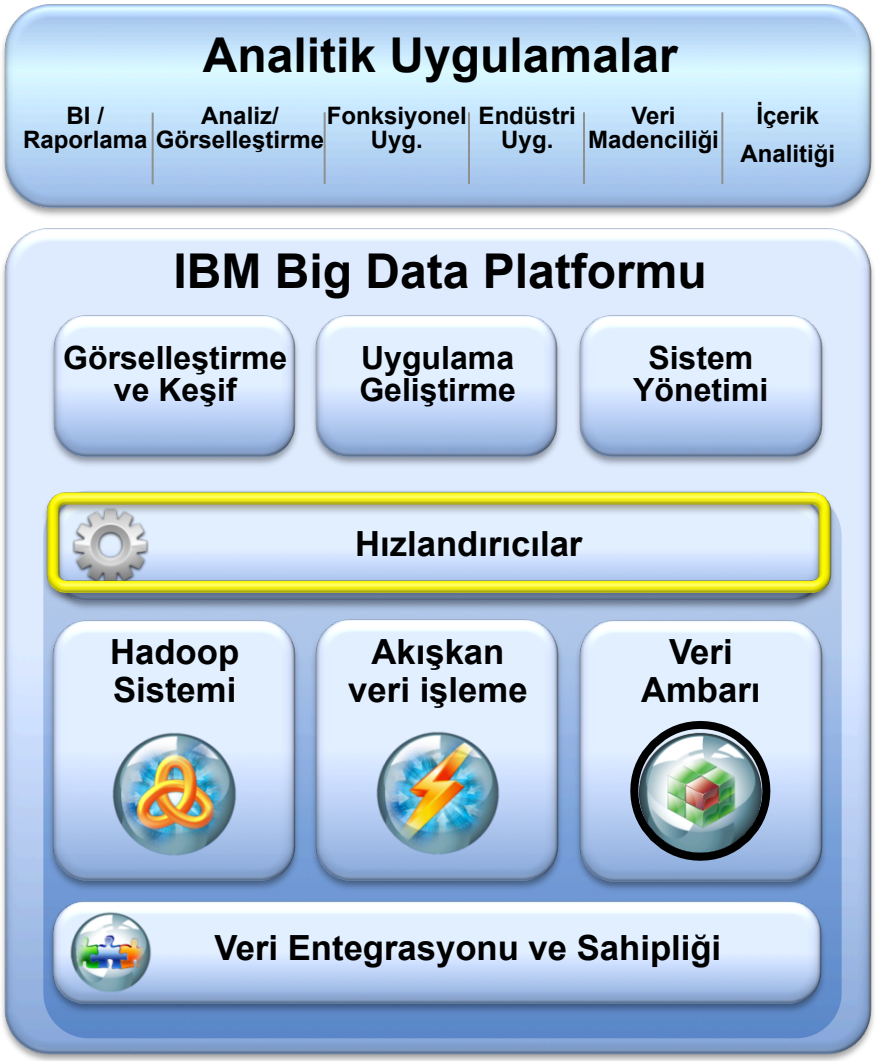
Kişisel veri erişimi !

Access Rule Description	Client IP	Server IP	DB User Name	Full SQL String
Access Rule: sensitive files: Alert	9.70.145.118	9.70.145.113	SVORUGA	__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: Alert	9.70.145.118	9.70.145.113	SVORUGA	__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

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 indirebilirisiniz
- Sınıf içi Eğitimler


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

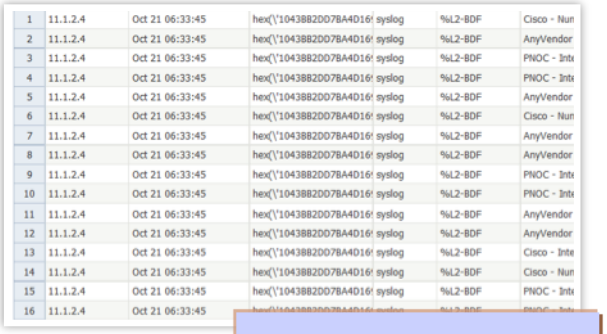
Operasyon Analizi: İhtiyaç

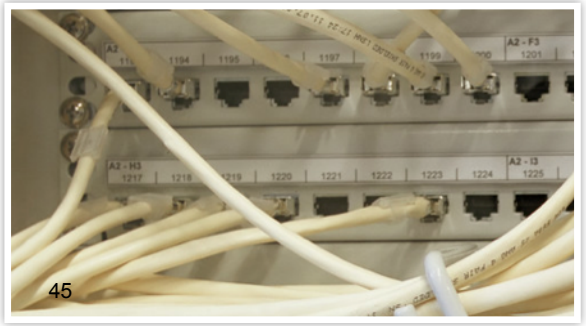
Gelişkin iş sonuçları için, çeşitli cihazların ürettiği verilerin analiz edilmesi

Machine Data Analytics Accelerator

Ne yapar ?

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



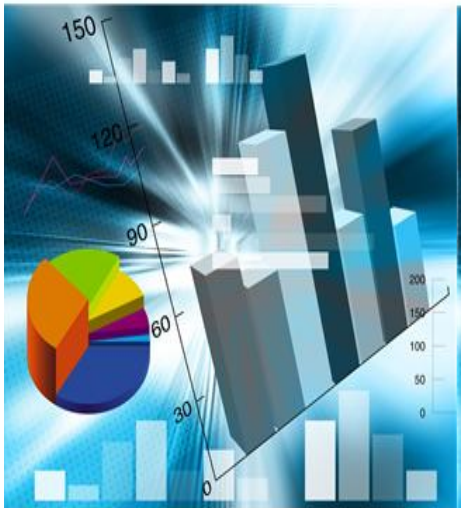


Makine Verilerinin Analizi
Temelde Zorlu Sorunlar Barındırıyor

Veri Kaynakları ve Entegrasyon

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

Analitik

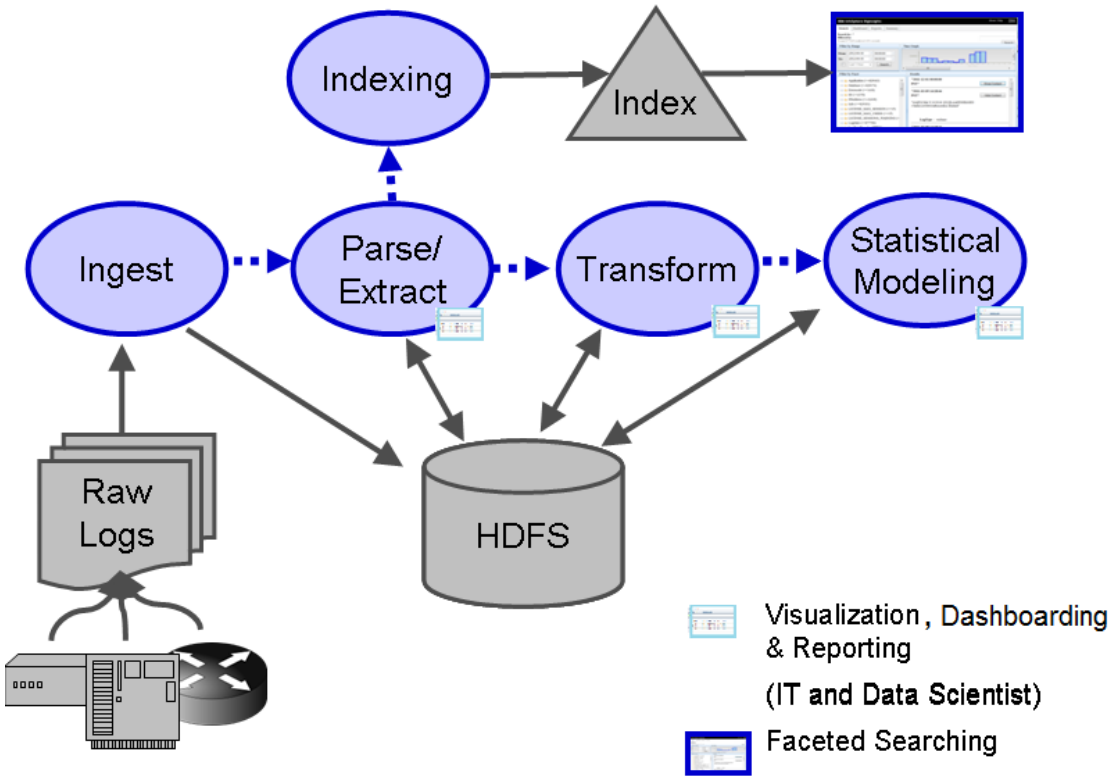


- Yüksek hacimli indeksleme
- Farklı veri setleri arasında korelasyon
- Farklı veri biçimleri için ileri düzey analitik

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 LZO-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