Information Management

BIG DATA

Luis Reina
Luis_reina@es.ibm.com

July 12, 2012

Un poco de Historia de los Datos…

OLTP

Bases de Datos Operacionales

1968

Base de datos

Jerárquicas

“IMS”

1970

Bases de datos

Relacionales

“System R”

OLAP

Data Warehousing

1983

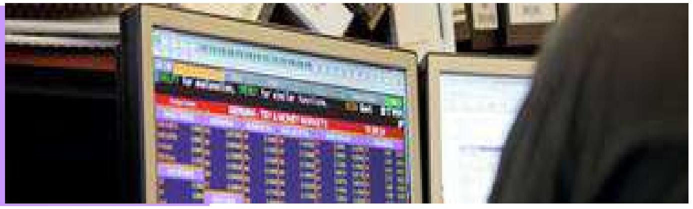
DB2 v1

Pero el mundo ha cambiado para ser más…

INSTRUMENTED

INTERCONNECTED

INTELLIGENT



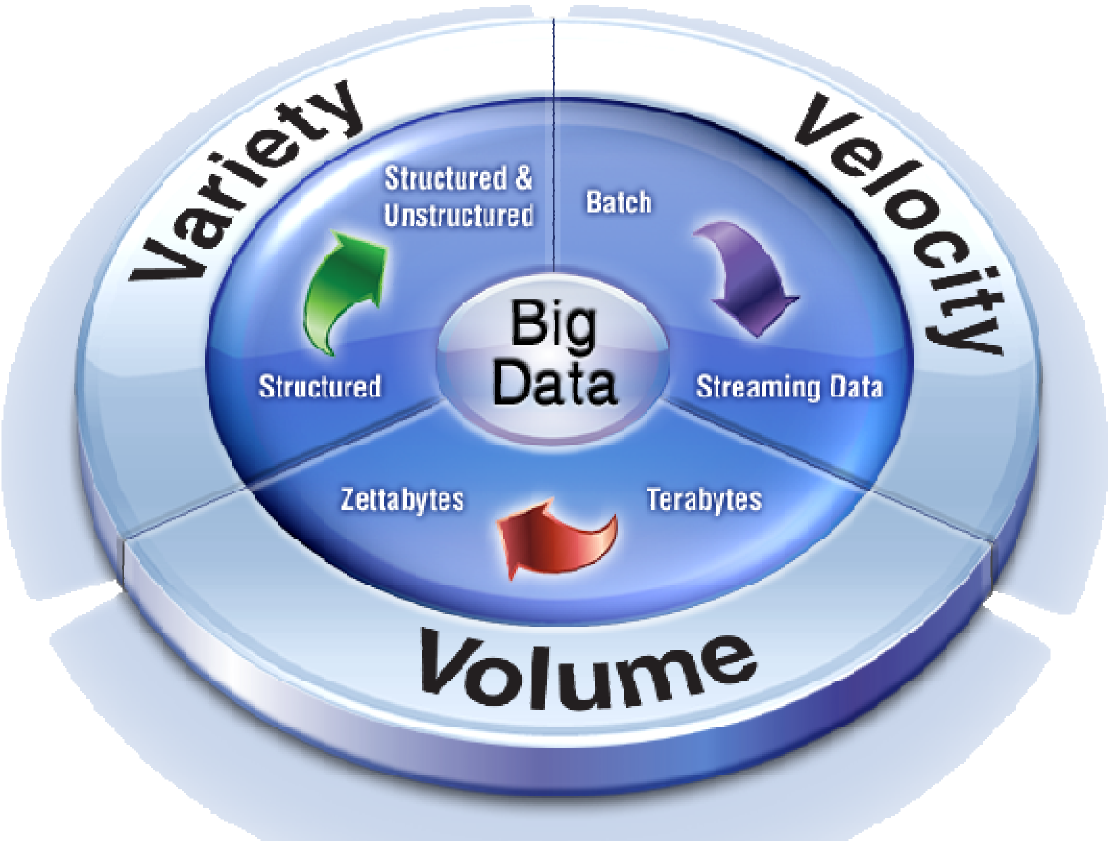
The resulting explosion of information creates a need for a new kind of intelligence

…to help build a Smarter Planet

In 2005 there were 1.3 billion RFID tags in circulation…

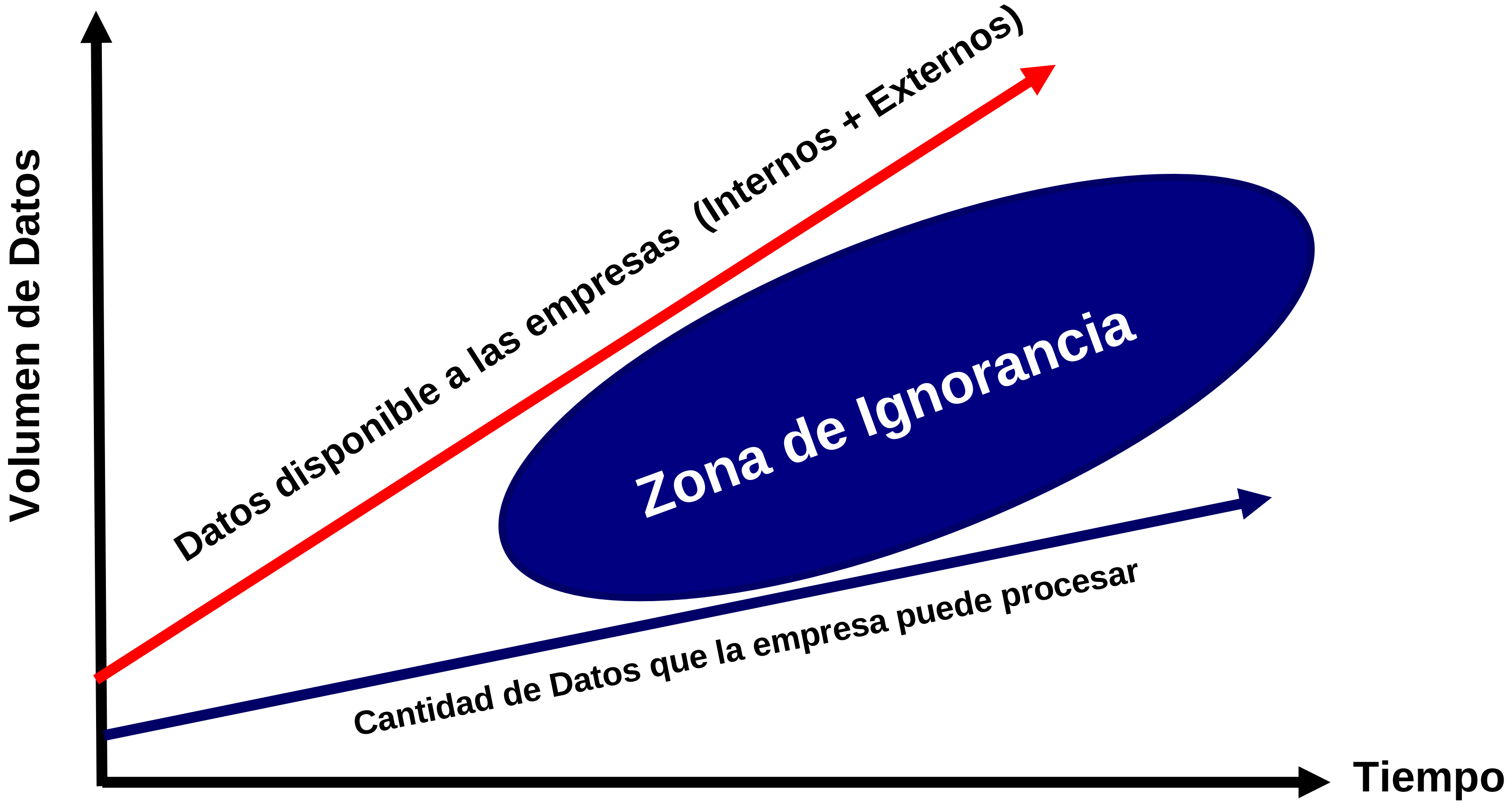
…by the end of 2011, this was about 30 billion and growing even faster

Modelo de las 3 Vs Describe la Situación actual de los Datos

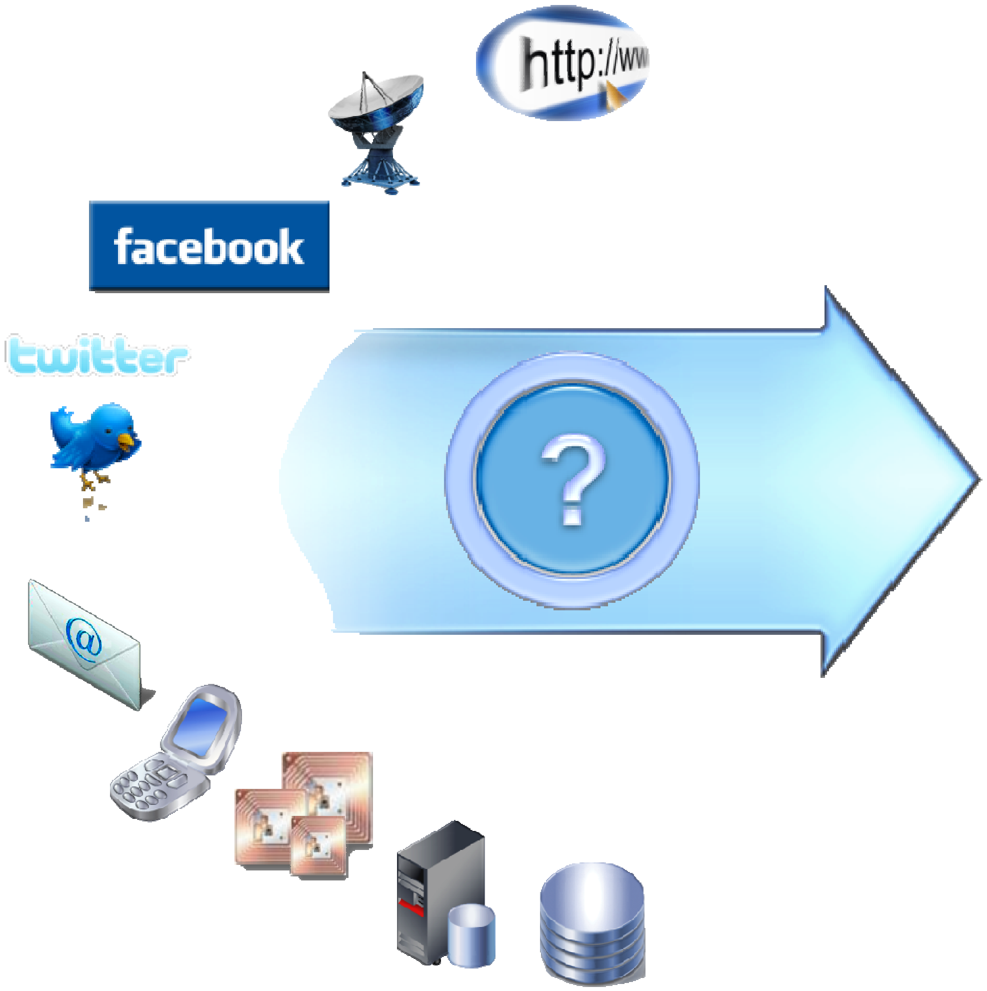


Se están añadiendo más Vs al modelo como Veracidad

Zona de Ignorancia Crece día a día


Volumen de Datos
Datos disponible a las empresas (Internos + Externos)
Zona de Ignorancia
Cantidad de Datos que la empresa puede procesar
Tiempo

¿Qué podría hacer si fuese capaz de Analizar estos datos?



Algunos Ejemplos:

Análisis de Sentimiento.

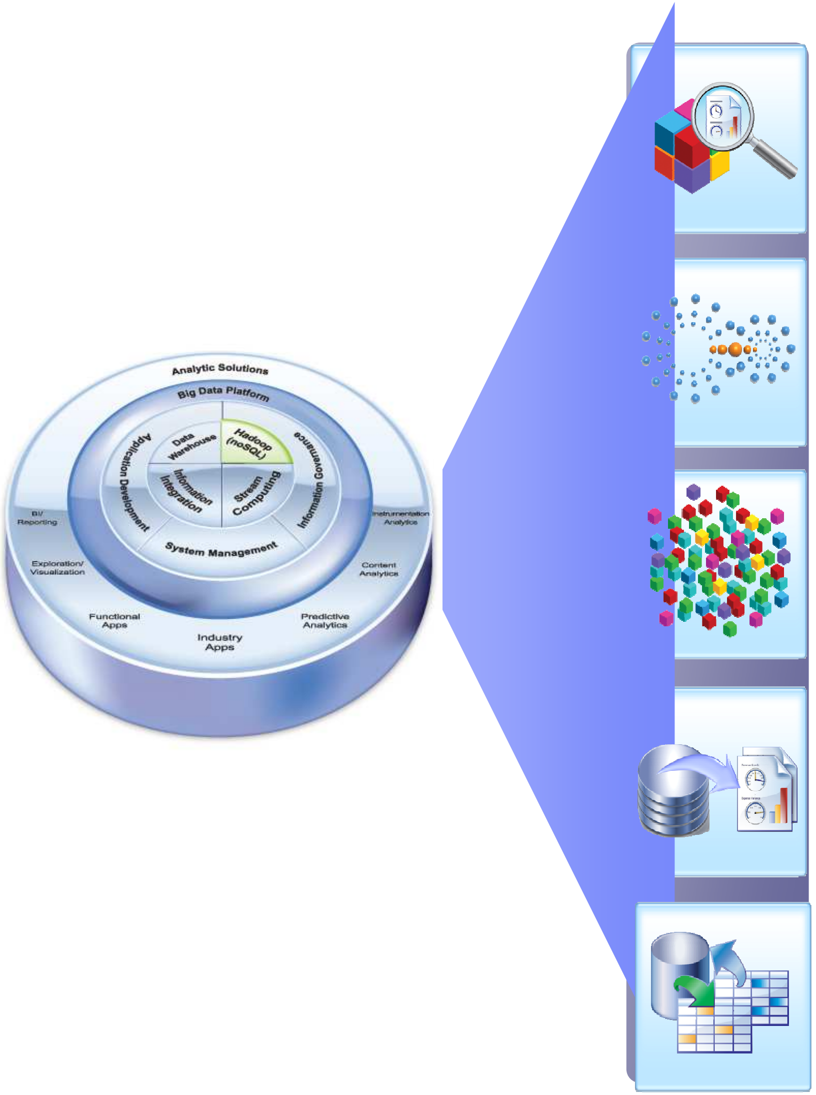
Tomar decisiones de riesgo basado en información transaccional en tiempo real.

Predecir patrones de tiempo para optimizar el uso de turbinas de viento.

Detectar a tiempo en pacientes de hospitales situaciones críticas.

Identificar criminales y amenazas desde información diversa como video, audio u otras fuentes.

¿Qué debe incluir una plataforma de Big Data?



Analyze a Variety of Information

Novel analytics on a broad set of mixed information that could not be analyzed before

Analyze Information in Motion

Streaming data analysis

Large volume data bursts & ad-hoc analysis

Analyze Extreme Volumes of Information

Cost-efficiently process and analyze petabytes of information

Manage & analyze high volumes of structured, relational data

Discover & Experiment

Ad-hoc analytics, data discovery & experimentation

Manage & Plan

Enforce data structure, integrity and control to ensure consistency for repeatable queries

Mezclando los enfoques Tradicionales y de Big Data

Traditional Approach

Structured & Repeatable Analysis

Business Users

Determine what question to ask

IT

Structures the data to answer that question

Monthly sales reports
Profitability analysis
Customer surveys

Big Data Approach

Iterative & Exploratory Analysis

IT

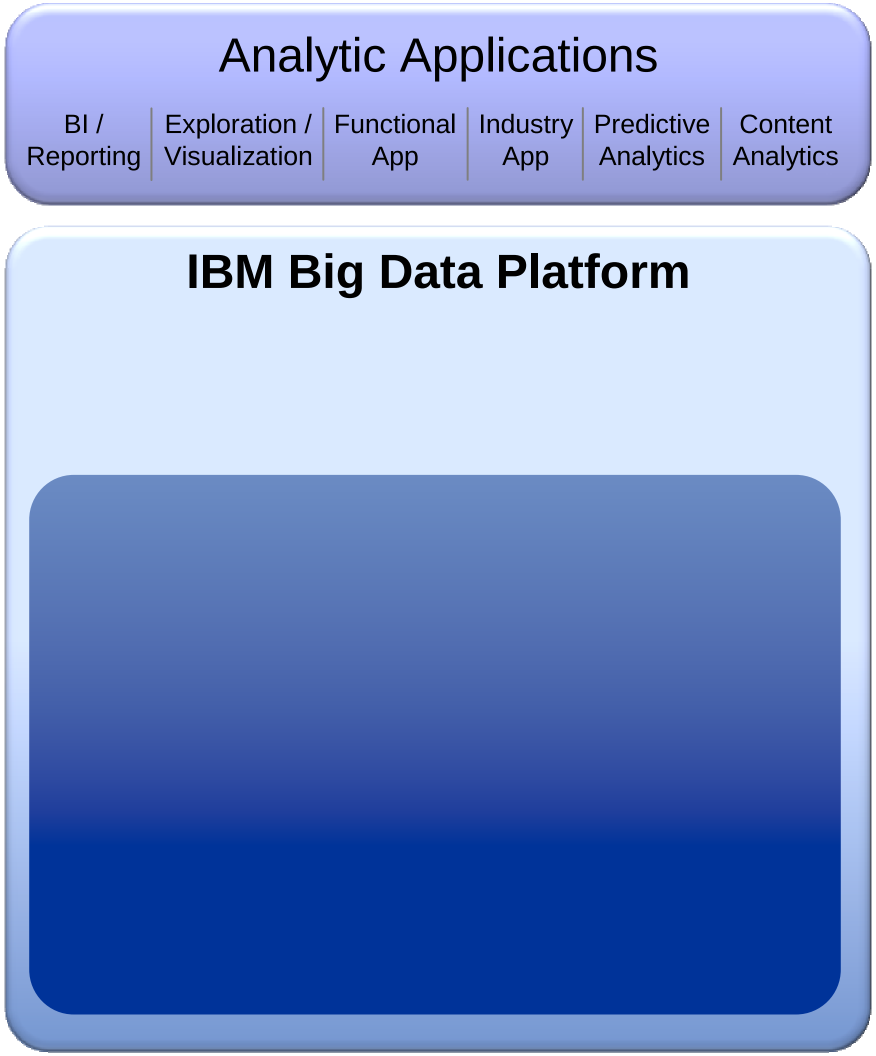
Delivers a platform to enable creative discovery

Business

Explores what questions could be asked

Brand sentiment
Product strategy
Maximum asset utilization

La Plataforma de IBM de Big Data

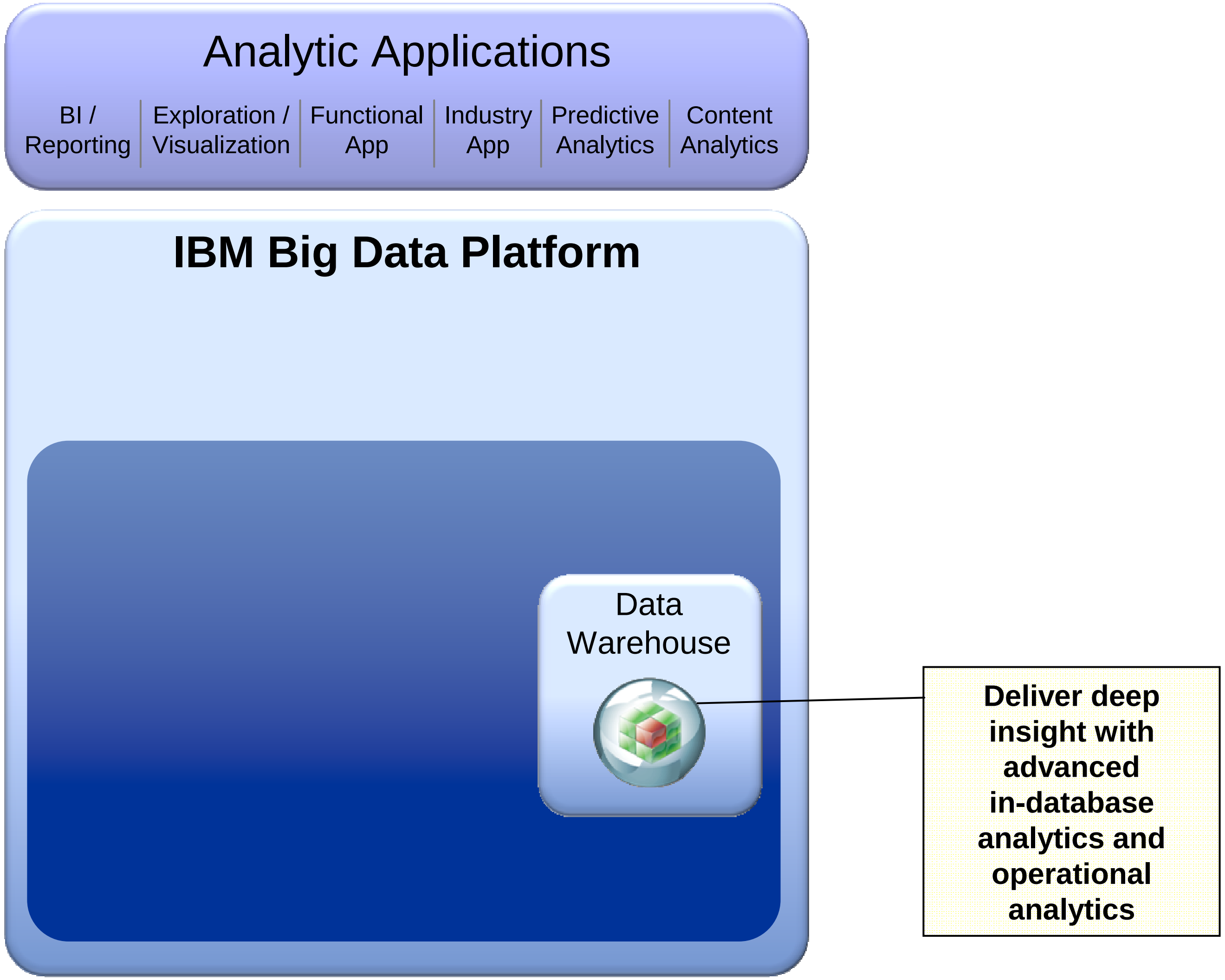


Analytic Applications

BI / Reporting | Exploration / Visualization | Functional App | Industry App | Predictive Analytics | Content Analytics

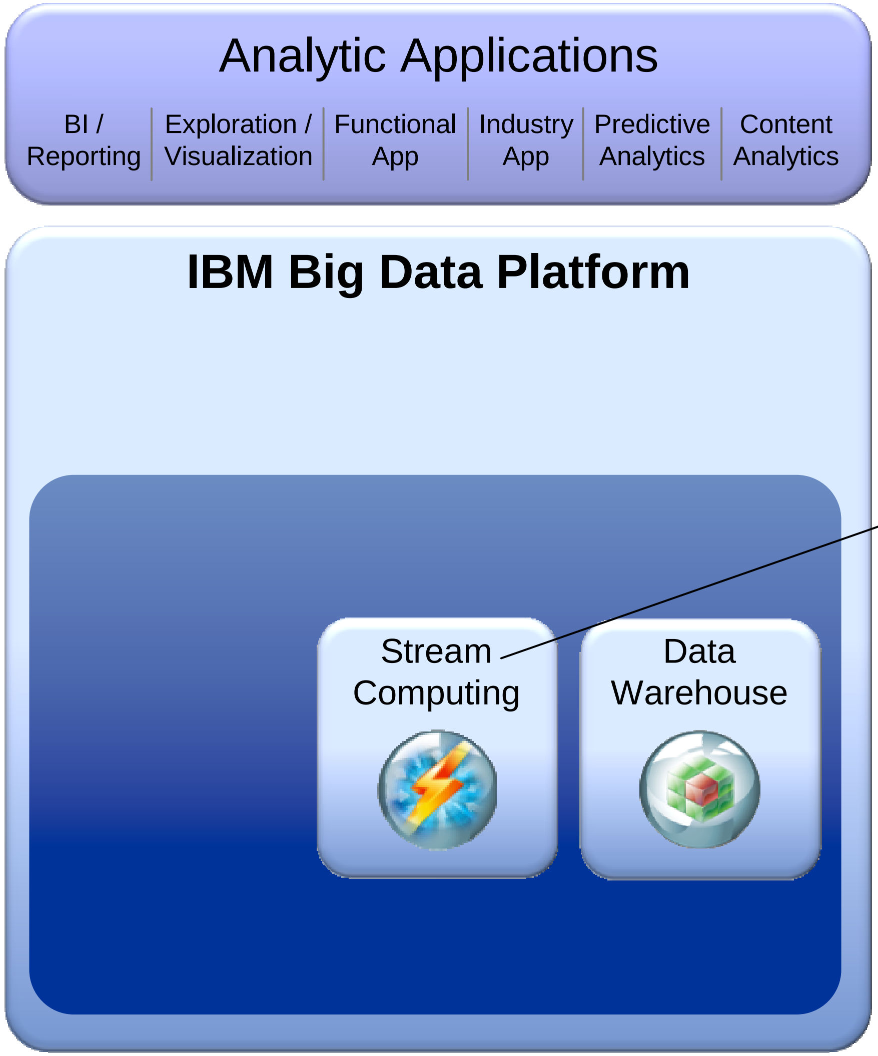
IBM Big Data Platform

La Plataforma de IBM de Big Data



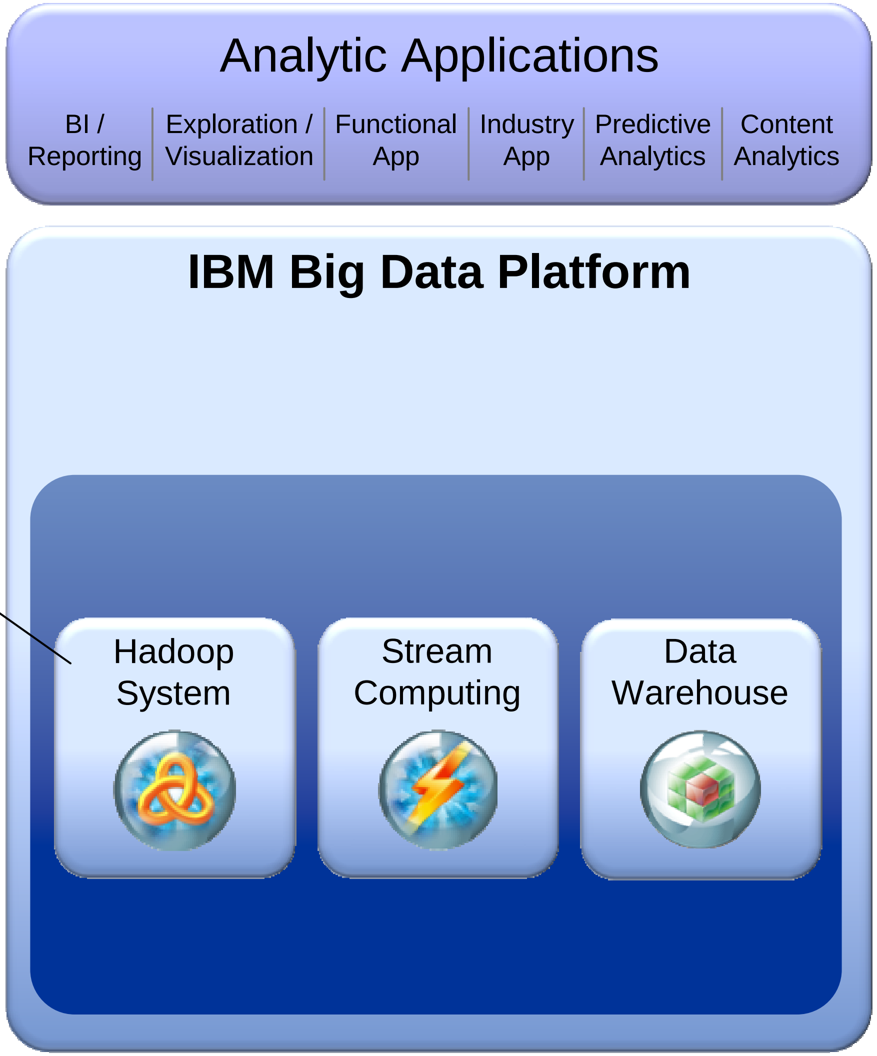
Deliver deep insight with advanced in-database analytics and operational analytics

La Plataforma de IBM de Big Data



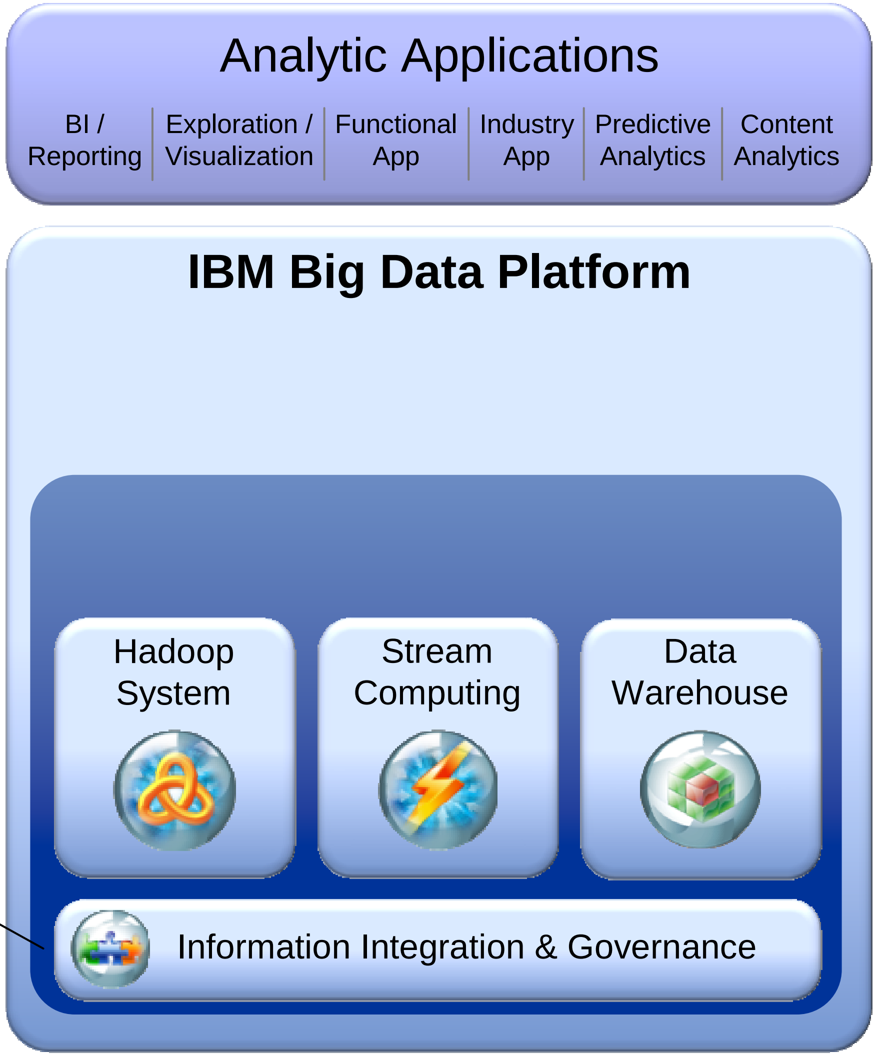
Analyze streaming data and large data bursts for real-time insights

La Plataforma de IBM de Big Data



Cost-effectively analyze petabytes of structured and unstructured information

La Plataforma de IBM de Big Data



Analytic Applications

BI / Reporting | Exploration / Visualization | Functional App | Industry App | Predictive Analytics | Content Analytics

IBM Big Data Platform

Hadoop System

Stream Computing

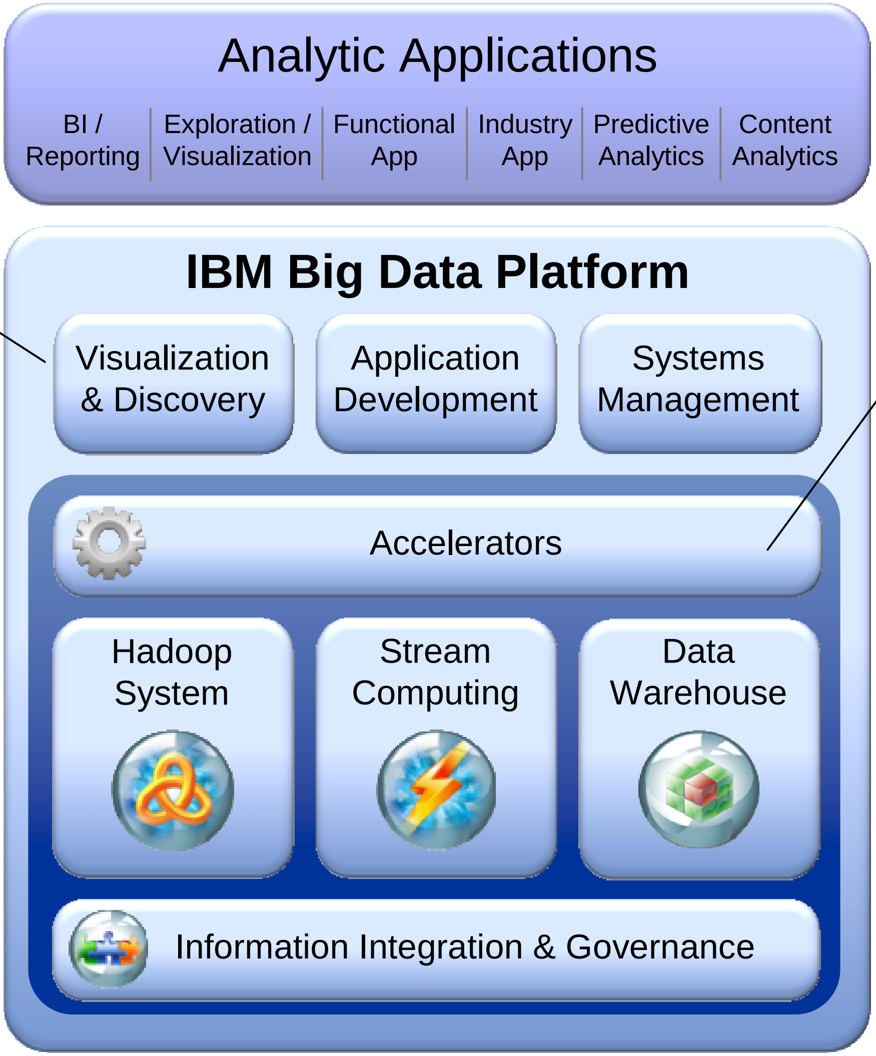
Data Warehouse

Information Integration & Governance

Govern data quality and manage information lifecycle

La Plataforma de IBM de Big Data

Gather, extract and explore data using spreadsheet metaphor

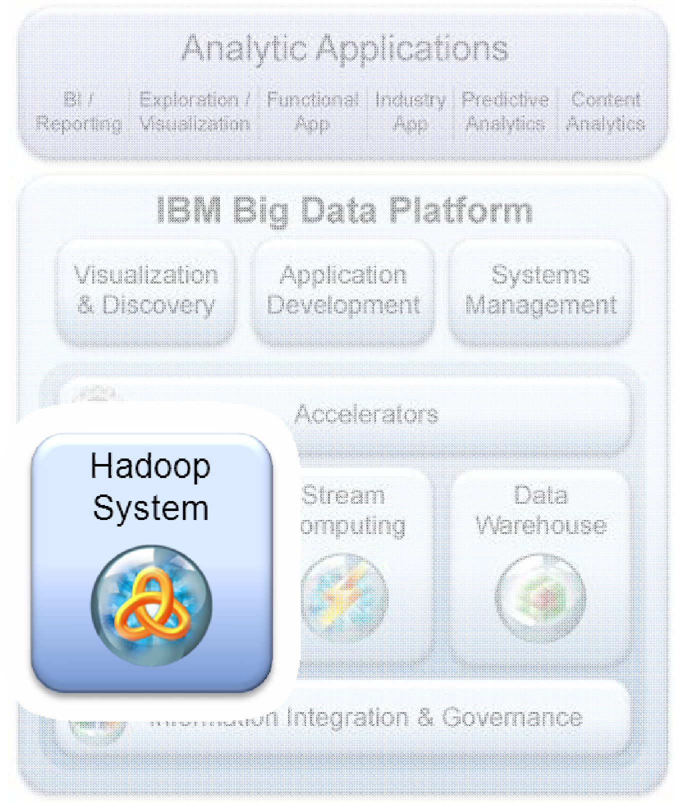


Speed time to value with analytic and application accelerators

La Plataforma de IBM de Big Data - Hadoop

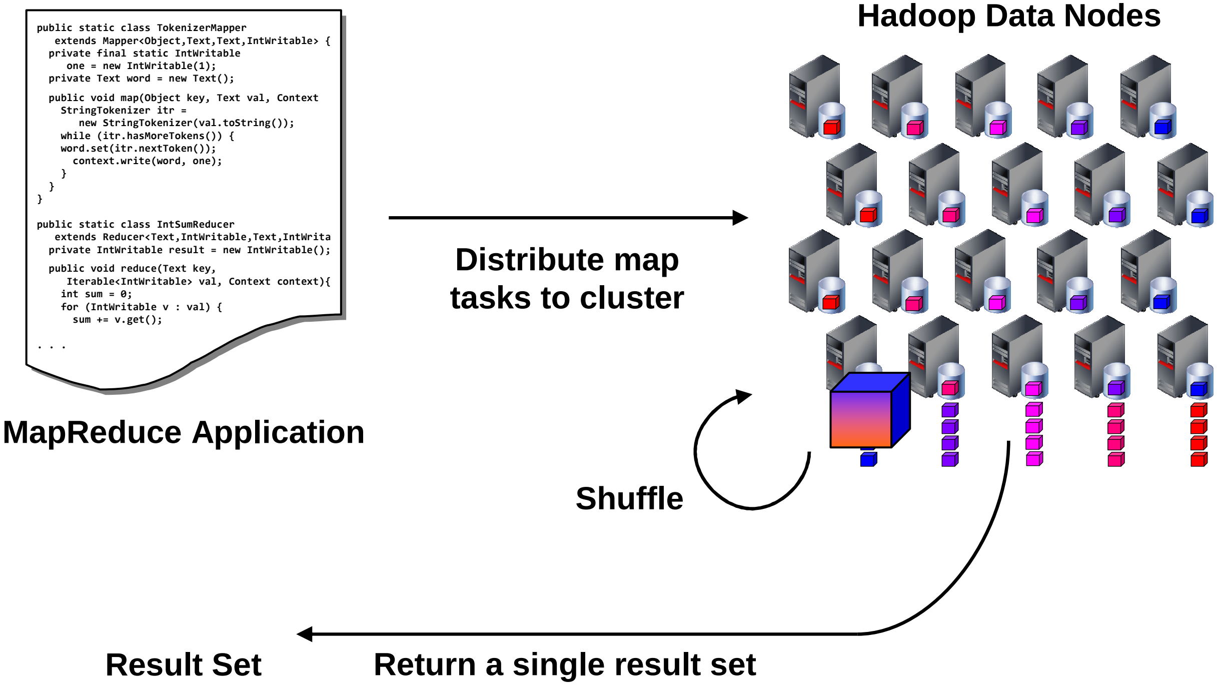
- **Manages a wide variety and huge volume of data**

- **Augments open source Hadoop with enterprise capabilities**
  - – Performance Optimization
  - – Development tooling
  - – Enterprise integration
  - – Analytic Accelerators
  - – Application and industry accelerators
  - – Visualization
  - – Security



Hadoop

- **Hadoop computation model**
  - Data stored in a distributed file system spanning many inexpensive computers
  - Bring function to the data
  - Distribute application to the compute resources where the data is stored
- **Scalable to thousands of nodes and petabytes of data**



```
public static class TokenizerMapper
   extends Mapper<Object,Text,Text,IntWritable> {
  private final static IntWritable
     one = new IntWritable(1);
  private Text word = new Text();

  public void map(Object key, Text val, Context
    StringTokenizer itr =
       new StringTokenizer(val.toString());
    while (itr.hasMoreTokens()) {
    word.set(itr.nextToken());
      context.write(word, one);
    }
  }
}

public static class IntSumReducer
   extends Reducer<Text,IntWritable,Text,IntWrita
  private IntWritable result = new IntWritable();

  public void reduce(Text key,
     Iterable<IntWritable> val, Context context){
    int sum = 0;
    for (IntWritable v : val) {
      sum += v.get();

. . .
```

1. **Map Phase**
   (break job into small parts)
2. **Shuffle**
   (transfer interim output for final processing)
3. **Reduce Phase**
   (boil all output down to a single result set)

IBM BigInsights= Hadoop Empresarial

hadoop + IBM Innovation

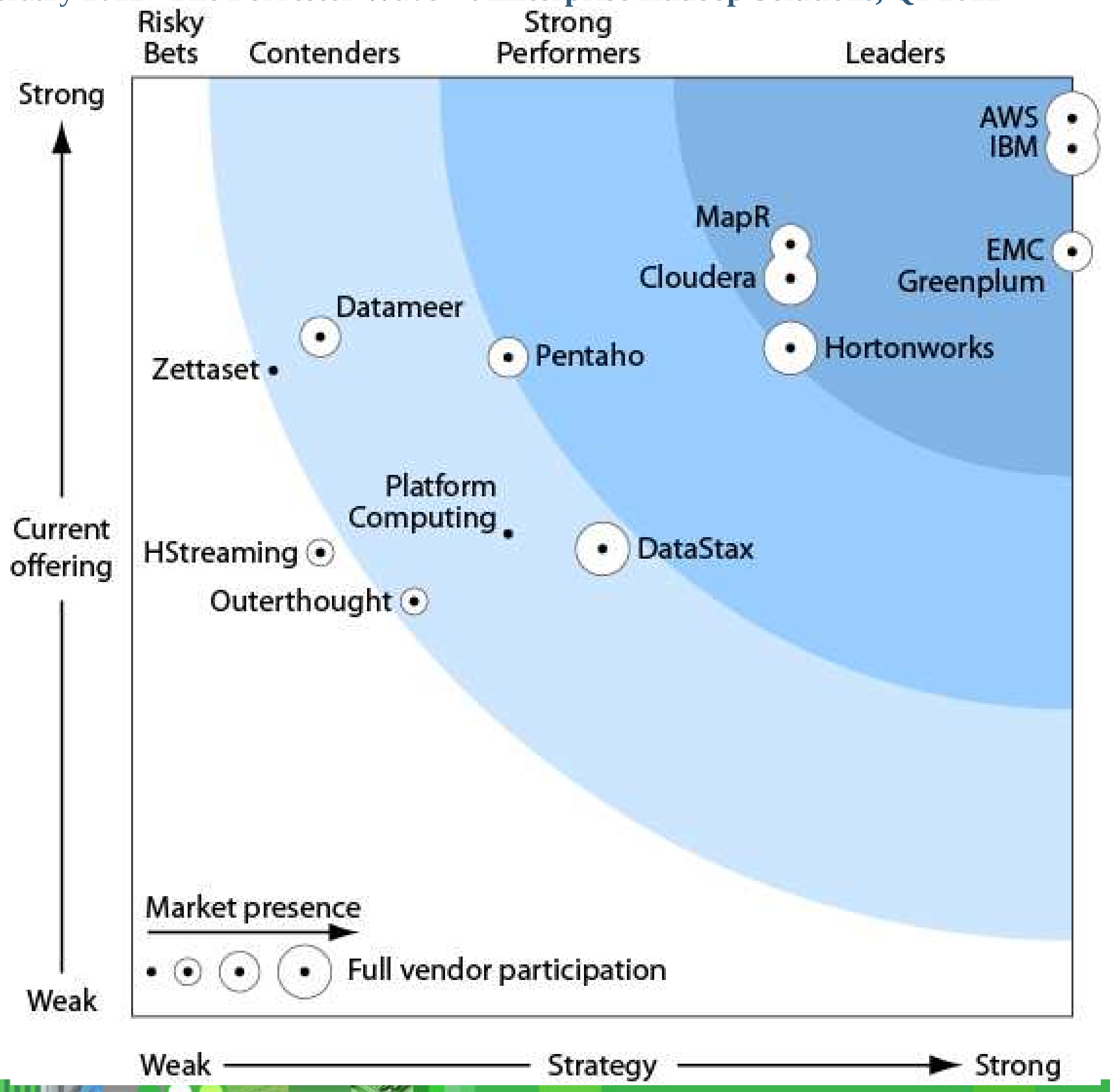
- **Scalable**
  - New nodes can be added on the fly.
- **Affordable**
  - Massively parallel computing on commodity servers
- **Flexible**
  - Hadoop is schema-less, and can absorb any type of data.
- **Fault Tolerant**
  - Through MapReduce software framework

- **Performance & reliability**
  - Adaptive MapReduce, Compression, BigIndex, Flexible Scheduler
- **Analytic Accelerators**
- **Productivity Accelerators**
  - Web-based UIs
  - Tools to leverage existing skills
  - End-user visualization
- **Enterprise Integration**
  - To extend & enrich your information supply chain.

Soluciones Hadoop Líderes

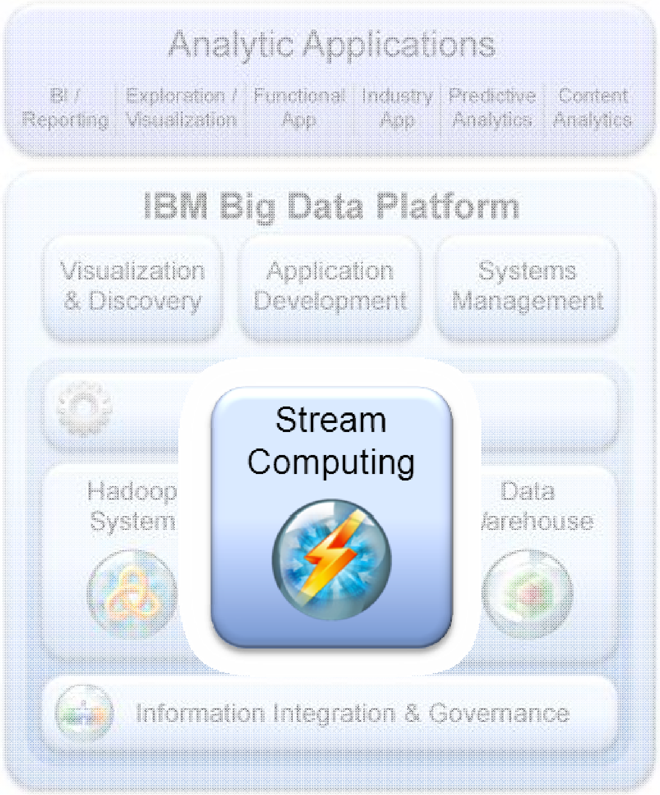
FORRESTER®

February 2012 **“The Forrester Wave™: Enterprise Hadoop Solutions, Q1 2012”**



La Plataforma de IBM de Big Data - *Stream Computing*

- **Built to analyze data in motion**
  - Multiple concurrent input streams
  - Massive scalability
- **Process and analyze a variety of data**
  - Structured, unstructured content, video, audio
  - Advanced analytic operators



Stream Computing: Nuevo Paradigma

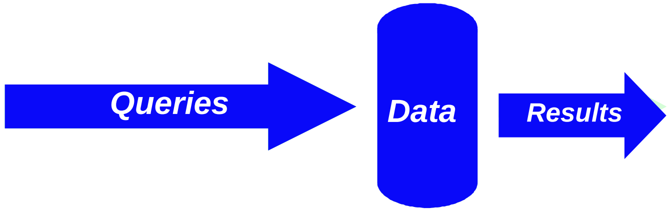
Traditional Computing

Historical fact finding with data-at-rest

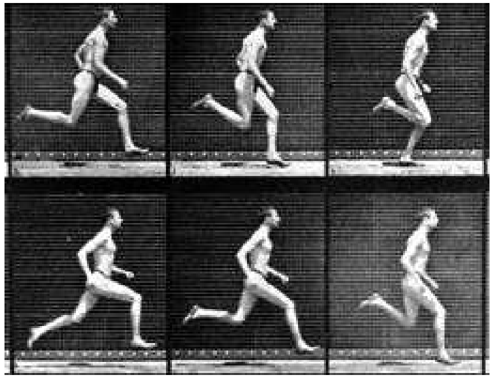
Batch paradigm, pull model

Query-driven: submits queries to static data

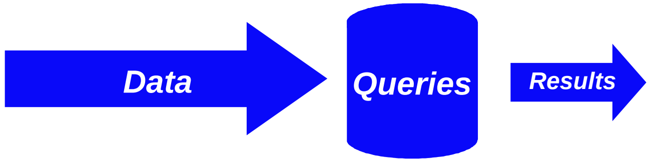
Relies on Databases, Data Warehouses



Stream Computing

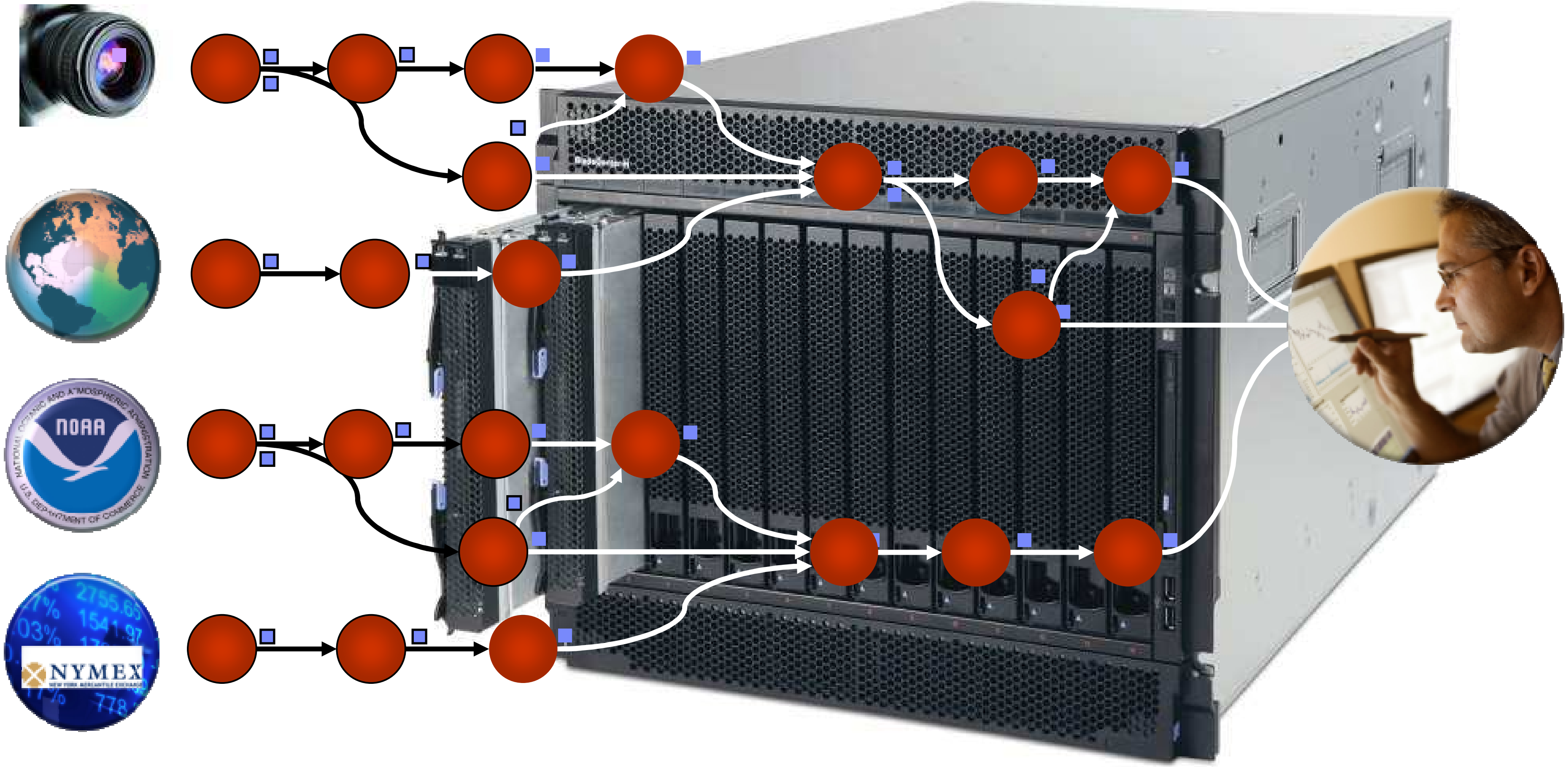


- **Real time analysis of data-in-motion**
- **Streaming data**
  Stream of structured or unstructured data-in-motion
- **Stream Computing**
  Analytic operations on streaming data in real-time



¿Qué es *Stream Computing?*

- Continuous Ingestion
- Continuous Complex Analysis in Microseconds


NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE
NYMEX

La Plataforma de IBM de Big Data – Aplicaciones Analíticas

Big Data Platform is designed for analytic application development and integration

BI/Reporting – Cognos BI, Attivio

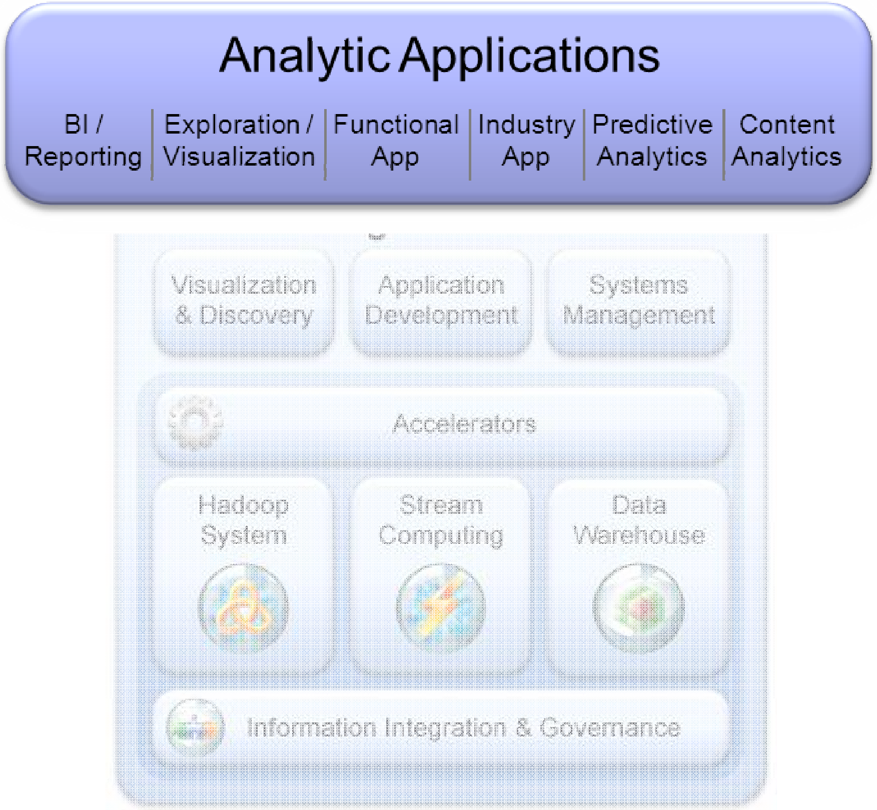
Predictive Analytics – SPSS, G2, SAS

Exploration/Visualization – BigSheets, Datameer

Content Analytics – IBM Content Analytics

Functional Applications – Algorithmics**,** Cognos Consumer Insights, Clickfox, i2, IBM GBS

Industry Applications – TerraEchos**,** Cisco, IBM GBS



¿Qué es *Text Analytics*?

- **High Performance and Scalable rule based Information Extraction Engine.**

- **Distill structured information from unstructured data**
  - Rich annotator library supports multiple languages

- **Provides sophisticated tooling to help build, test, and refine rules.**
  - Developer tools, an easy to use text analytics language, and a set of extractors for fast adoption.
  - Multilingual support, including support for DBCS languages.

- **Developed at IBM Research since 2004:** System T

- **Embedded in several IBM products**
  - Infosphere Warehouse
  - Infosphere Streams.
  - Lotus Notes
  - Cognos Consumer Insights

- **BigInsights is the first time IBM opens up the Text Analytics Engine technology for customization and development**

Text Analytic: Ejemplo Sencillo

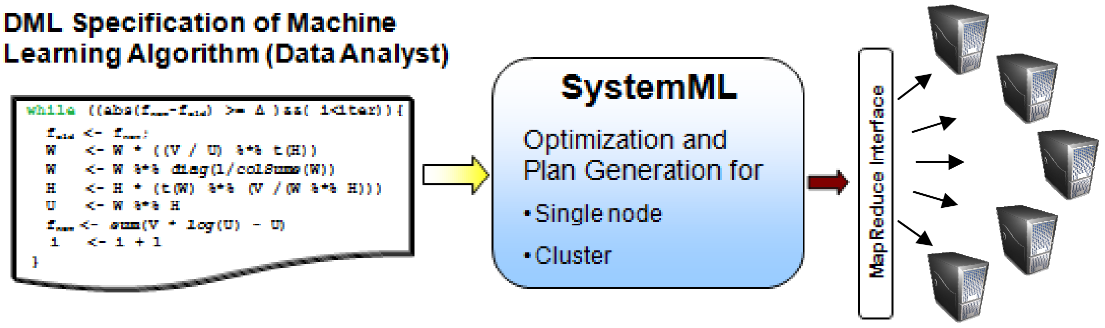
Football World Cup 2010, one team distinguished well from the rest winning the final. Early in the second half, Netherlands' striker, Arjen Robben, had a chance to score, but the awesome keeper for Spain, Iker Casillas made the save. Winner superiority was reflected when Winger Andres Iniesta scored for Spain for the win.

World Cup 2010 Highlights

Name	Position	Country
Arjen Robben	Striker	Netherlands
Iker Casillas	Keeper	Spain
Andres Iniesta	Winger	Spain

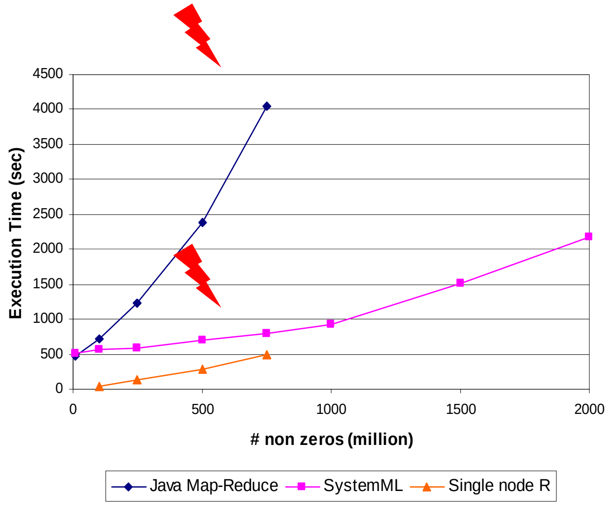
Análisis Estadístico y Predictivo

- Framework for machine learning (ML) implementations on Big Data
  - Large, sparse data sets, e.g. 5B non-zero values
  - Runs on large BigInsights clusters with 1000s of nodes
- Productivity
  - Build and enhance predictive models directly on Big Data
  - High-level language – Declarative Machine Learning Language (DML)
    - E.g. 1500 lines of Java code boils down to 15 lines of DML code
  - Parallel SPSS data mining algorithms implementable in DML
- Optimization
  - Compile algorithms into optimized parallel code
  - For different clusters
  - For different data characteristics
  - E.g. 1 hr. execution (hand-coded) down to 10 mins



DML Specification of Machine Learning Algorithm (Data Analyst)

```
while ((abs(f_new-f_old) >= Δ )&&( i<iter)){
   f_old <- f_new;
   W    <- W * ((V / U) %*% t(H))
   W    <- W %*% diag(1/colSums(W))
   H    <- H * (t(W) %*% (V /(W %*% H)))
   U    <- W %*% H
   f_new <- sum(V * log(U) - U)
   i    <- i + 1
}
```



#START013

Conectados con el progreso

IBM Software Summit
6 de noviembre de 2012
Palacio Municipal de Congresos de Madrid

▶ **Únete a la conversación en #Start013**

Encuentra todos los detalles en www.ibm.com/software/es/

@IBMSoftware_es   IBM Software España   Encuentro de Software   IBM España