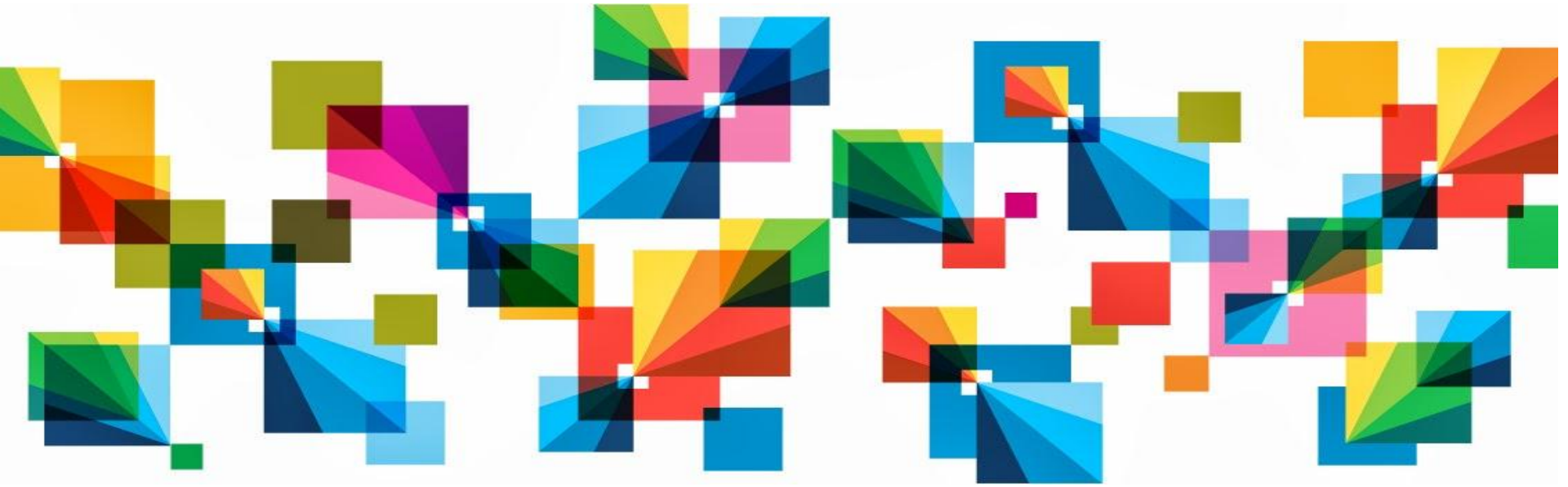
La Analítica de Datos no Estructurados

Análisis Big Data con Hadoop



Big Data es Más Que Solamente Hadoop

Que me podría decir de Big Data?

Quiero saber de Hadoop.

Director de Marketing

Big Data es mucho mas que Hadoop!

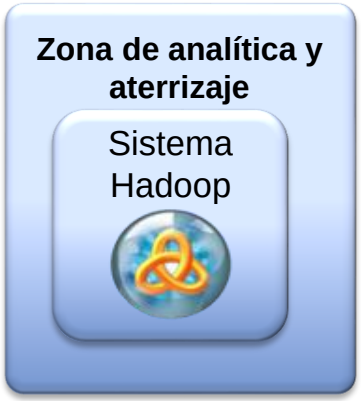
Nuestros competidores no pueden entregar todo el conjunto de casos de uso de Big Data.

IBM

Hay Dos Tipos Principales de Big Data

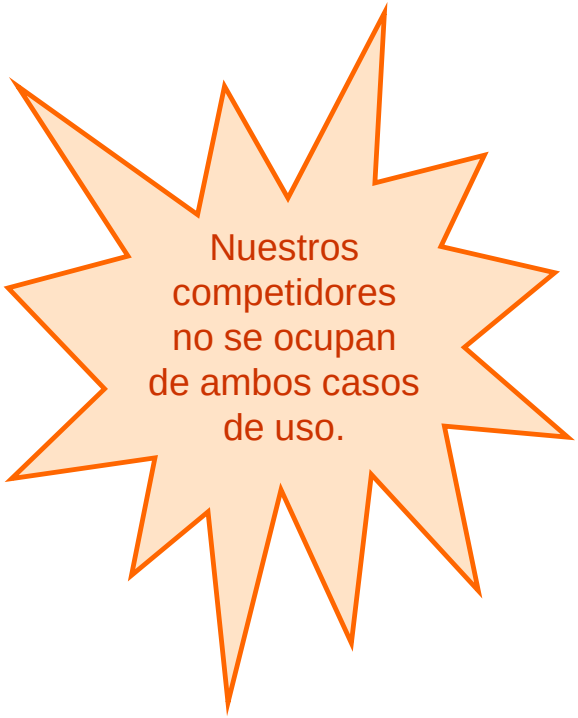
Datos en movimiento

- Datos no típicamente almacenados
- Alta velocidad
- Múltiple fuentes de datos
- Volúmenes enormes de datos no estructurados
- Latencia ultra baja requerida



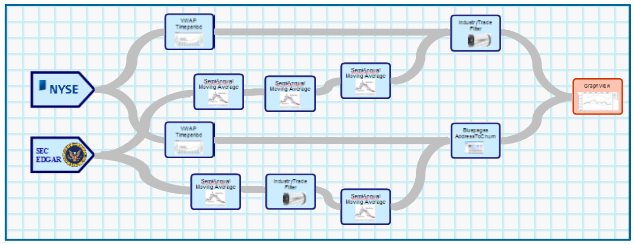
Datos en reposo

- Datos almacenados en disco
- Volúmenes enormes de datos no estructurados
- Sin esquemas predefinidas
- Demasiado grande para permitir que procesos o herramientas tradicionales ejecuten en manera oportuna.

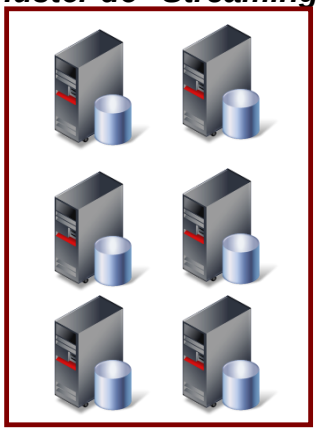


Nuevos Modelos de Programación y Hardware de Bajo Costo Resuelve Problemas de Big Data

Aplicación de “Data Streaming”

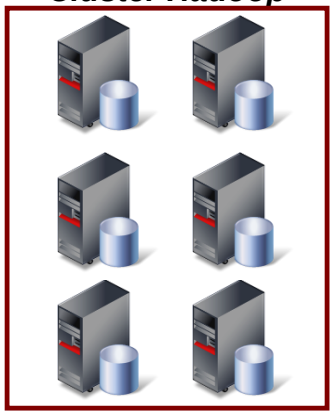


Clúster de “Streaming”



- Flujo de datos y aplicaciones Apache Hadoop
  - Sistema probado para el proceso de cantidades grandes de datos
  - Streaming data en movimiento, Hadoop para datos en reposo
  - Permite que las aplicaciones trabajen de forma transparente con grandes grupos de nodos en paralelo

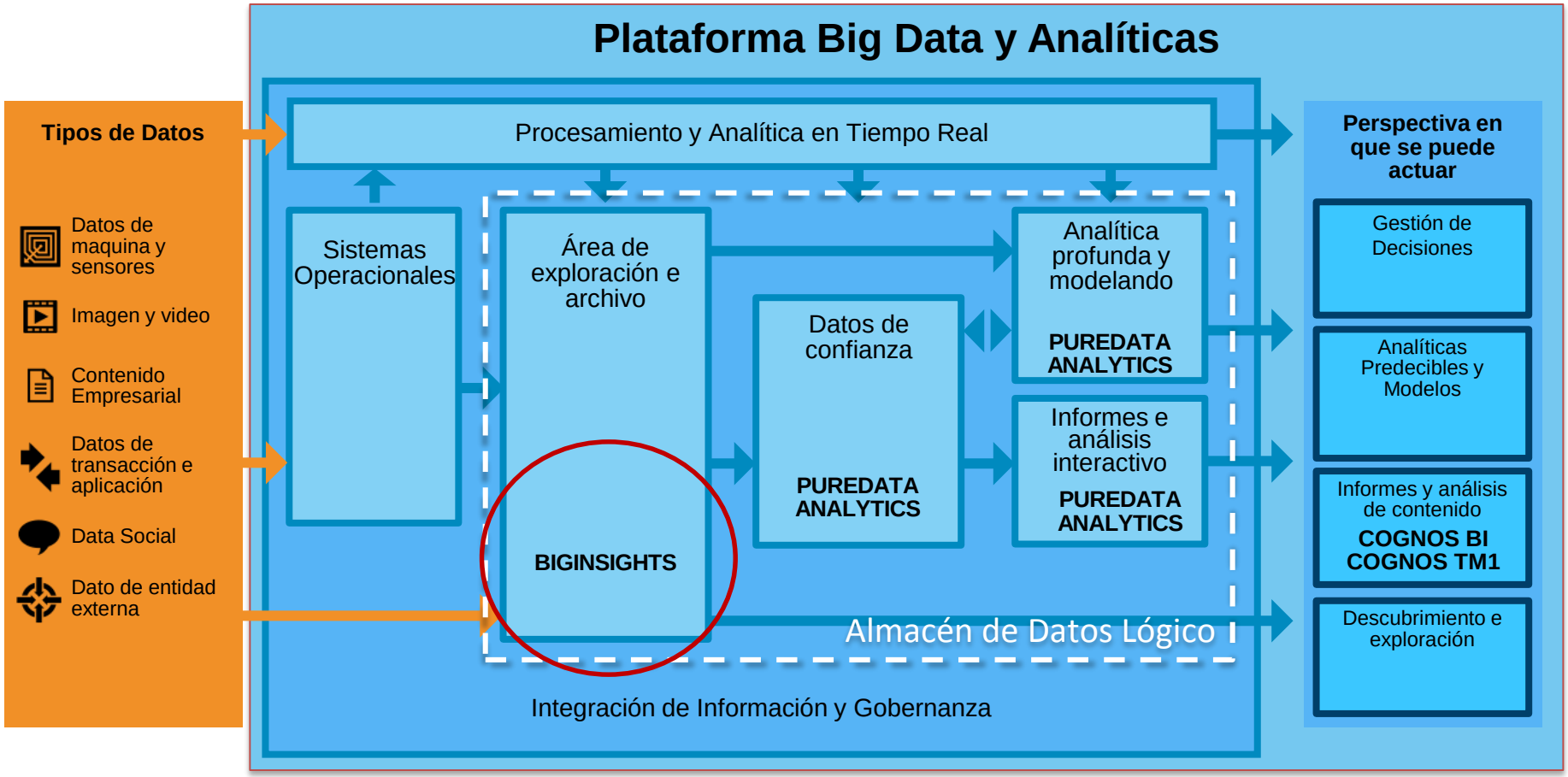
Clúster Hadoop



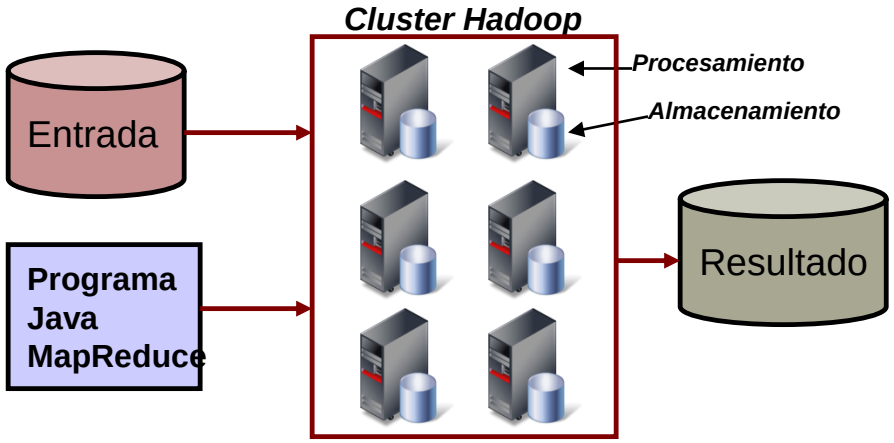
Ganando Valor de los Datos en Descanso

Fuente de Datos	Análisis	Valor al Negocio
Web Logs 	*Analizar el comportamiento del comprador en sitio internet de comercio*	*Maximizar las ventas en sitios a menor de e-comercio*
Medios Sociales 	*Analizar el sentimiento y la experiencia del comprador*	*Atraer y retener los clientes*
Datos Meteorológicos 	*Analizar grandes cantidades de datos meteorológicos históricos*	*Determinar la colocación optima de aerogeneradores*

Arquitectura de Siguiente Generación IBM – El Almacén de Datos Lógico

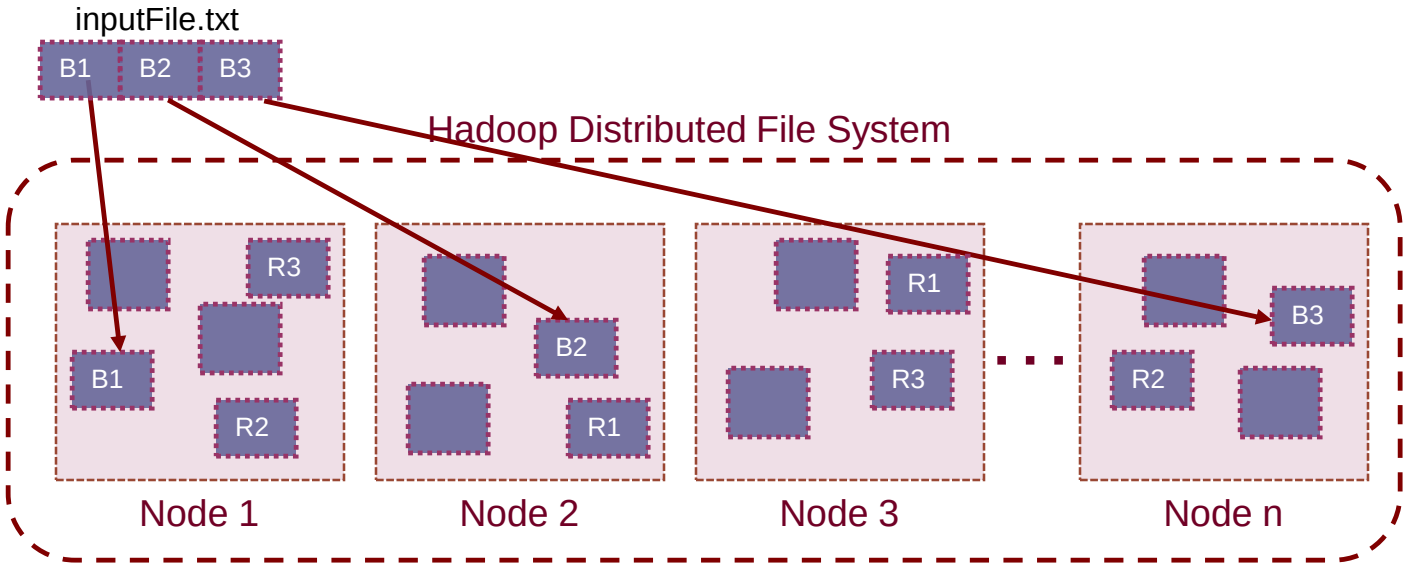


InfoSphere BigInsights Incluye Apache Hadoop Para Procesar los Datos en Reposo



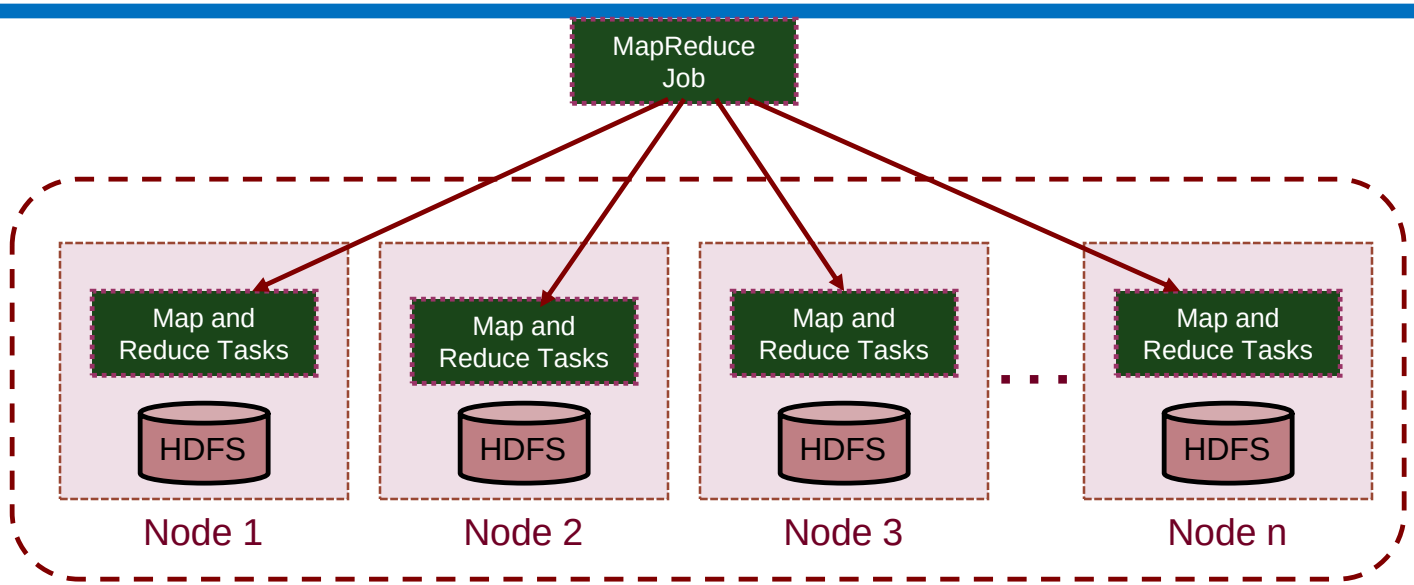
- Compuesto de un clúster de hardware de bajo costo
  - Los nodos tienen procesadores, memoria, y discos
- Sistema de archivos especializado – Hadoop Distributed File System (HDFS)
- Modelo de programación especial – MapReduce

El Sistema de Archivos Distribuidos Hadoop (HDFS) Distribuye los Datos en un Clúster Hadoop



- Un sistema de archivos distribuido que se extiende por todos los nodos de un clúster Hadoop
- A tiempo de usarse, los archivos son divididos en bloques y distribuidos entre los nodos de datos
- El sistema supone que los nodos fallaran
  - Logra fiabilidad mediante la replicación de datos a través de múltiples nodos
- Elásticamente escalable

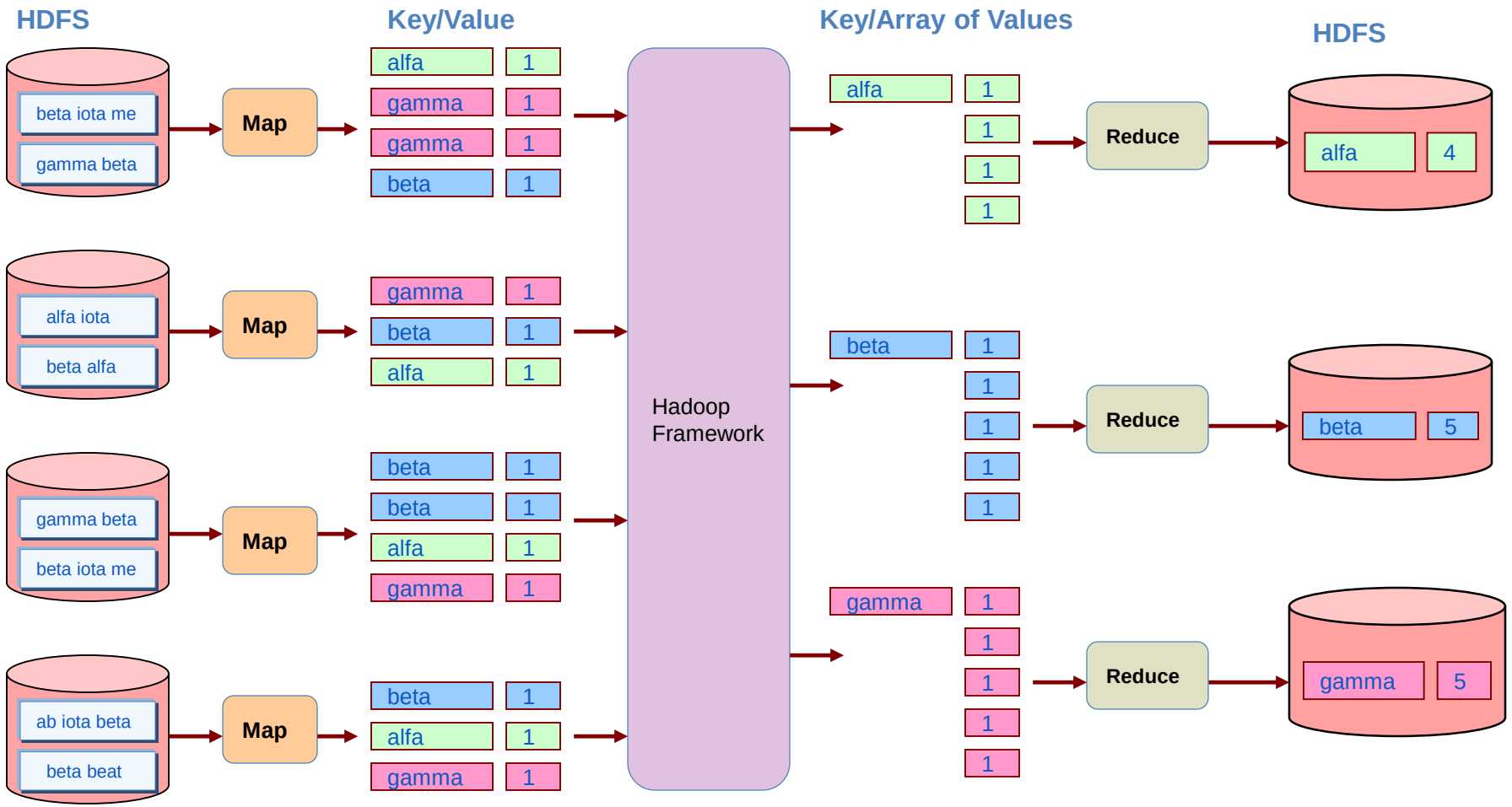
El MapReduce Framework Envía los Programas Hacia los Datos



- Tarea MapReduce se envía a cada nodo
- Las tareas Map y Reduce se ejecutan en paralelo a través de los nodos
- El sistema Hadoop hace mucho del “trabajo pesado”
  - Por ejemplo, mover datos entre las tareas Map y Reduce

Ejemplo MapReduce Simple de Contar Ocurrencias de Cadenas en Texto

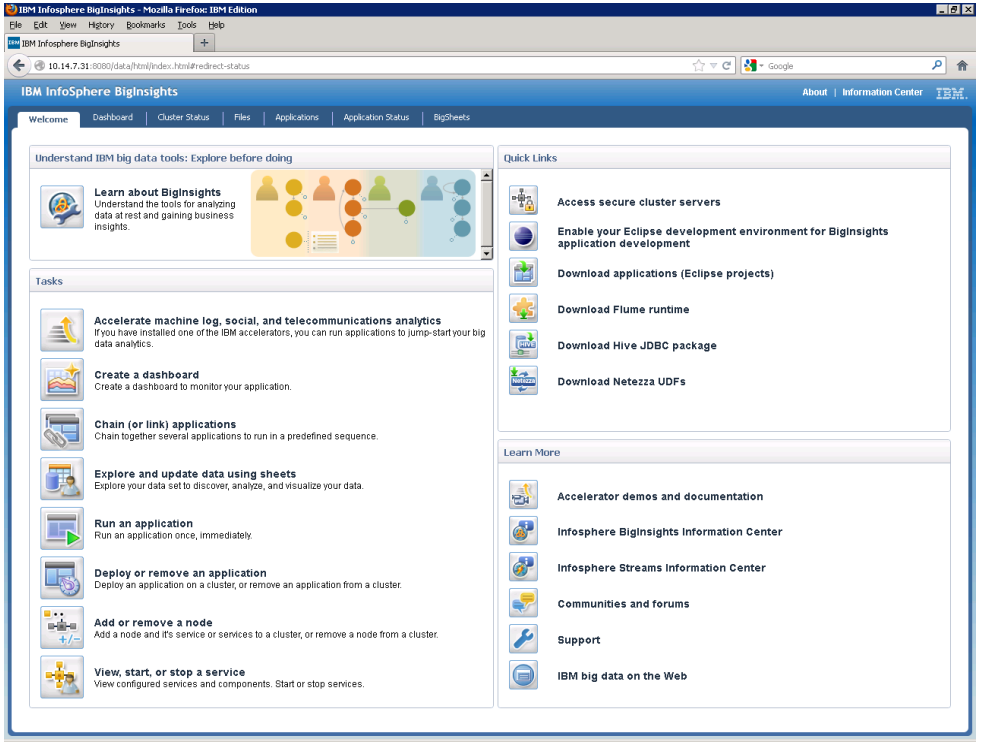
Objetivo: Contar el número de ocurrencias de texto "alfa," "beta," y "gamma" en un archivo.



BigInsights Lo Hace Fácil para Todos los Roles de Big Data

- Rol de administrador
  - Gerencia de clúster completo
    - Monitoreo / iniciar / detener componentes
    - Anadir / eliminar nodos
  - Tableros estilo portal
- Rol de desarrollador
  - Herramientas basadas en Eclipse
  - Acceso de leer/escribir a HDFS
  - Amplias vistas de trabajos y flujos de trabajos en el sistema
  - Centro de despliegue, lanzamiento, y programación
  - Aceleradores integrados
- Rol de usuario de negocio
  - No hay requerimiento de Java
  - Herramienta de Spreadsheet
  - Visualización

Consola InfoSphere BigInsights



Service Oriented Finance Quiere Analizar las Quejas de los Clientes

Necesitamos saber de que se están quejando los clientes.

Director de Marketing

Nosotros podemos ayudarle hacer esto con el análisis de el sentimiento del comprador usando BigInsights

IBM

El Análisis de Sentimientos – Un Desafío Big Data, Sino También una Oportunidad de Big Data

Grandes volúmenes de datos no estructurados

Tratando de determinar …

- Demanda de productos
- Aceptación del producto
- Amenazas competitivas
- Reputación de la marca
- Objetivos publicitarios

Encontrando los sentimientos de los clientes entre los datos de medios sociales

DEMO: Usando BigInsights para Analizar el Sentimiento Negativo en Twitter

Fuente
Twitter

Tema
Service Oriented
Finance

Gustos

- Me encanta la función de verificación de cheque
- Me gusta la habilidad de pagar facturas en línea
- Me gusta que cajeros automáticos se encuentran en toda la ciudad
- Me gustan los representantes de servicio

Disgustos

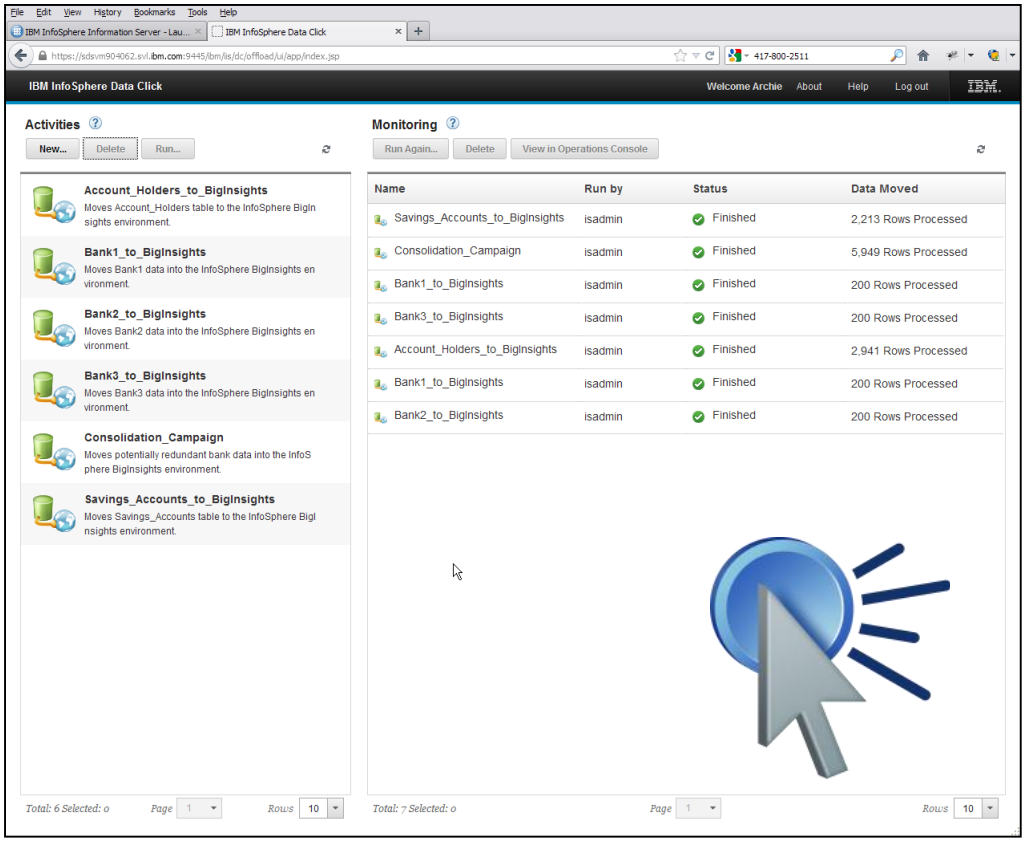
- No confió en la función de banca on-line
- No me gusta esperar en fila por mucho tiempo
- No me gustan las tarifas por uso de ATM
- Odio los cargos por sobregiro

BigInsights Tiene Habilidades que Otras Distribuciones de Hadoop no Tienen

- Dos motores de procesamiento poderosos
  - Procesamiento masivo paralelo de lotes con MapReduce
  - Motor SQL completamente ANSI compatible con Big SQL
- Desempeño y optimización
  - MapReduce adaptivo
  - Programador avanzado
  - BigIndex para indexación de gran escala
  - Compresión divisible, rápido
- Optim Development Studio
  - IDE para Java basado en Eclipse
- Integración Big Data
  - Information Server, InfoSphere Streams, Netezza, DB2
- Aceleradores analíticos
  - BigSheets spreadsheet y visualización
  - Datos de maquina
  - Medios de comunicación social
  - Análisis de texto avanzado
  - Lenguaje de consulta JAQL

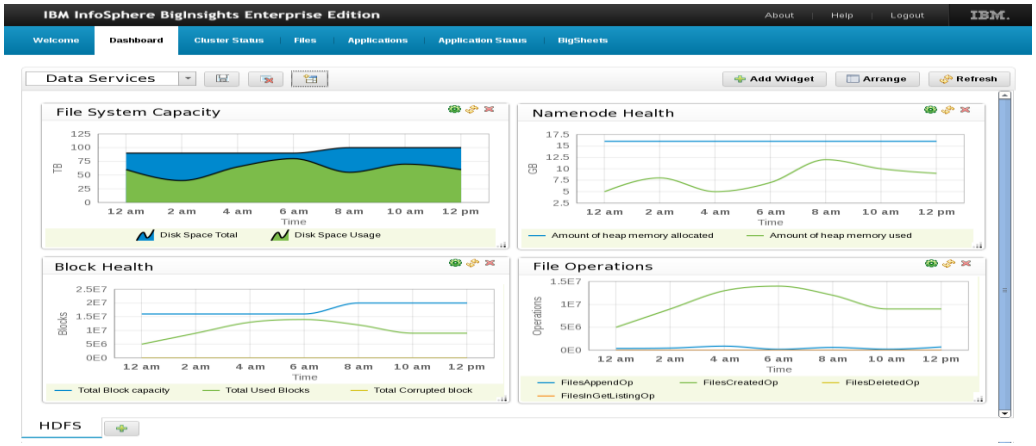
InfoSphere Data Click esta Liado con BigInsights

- Usuario técnico o de negocio
- Integración de plataformas relacionales y Hadoop
- Simple experiencia end-to-end
- Configuración basado en Web
- Soporta DB2, Netezza, Oracle, SQL Server, Teradata y otros
- Consulta tablas de datos descargadas a través de Big SQL



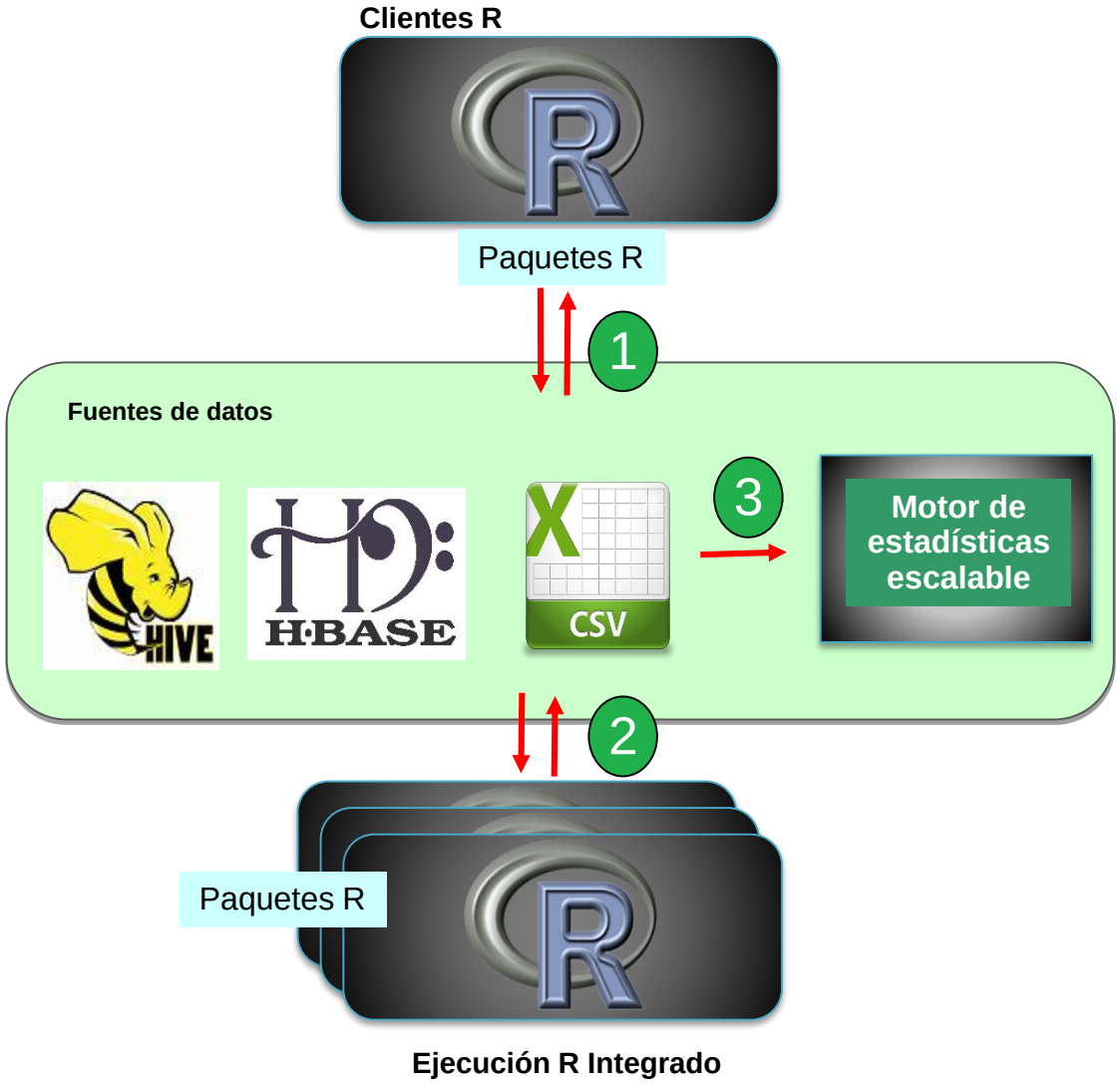
Visualice Resultados a Través de Tableros

- Tableros integrados para el monitoreo de la salud del sistema, estado de la aplicación, el sistema de archivos distribuido, etc.
- Fácil de personalizar (añadir, crear grupos, eliminar widgets)
  - Big Sheets colecciones y gráficos
  - Monitoreo de clúster / sistema
  - Monitoreo de HDFS
  - Métricas MapReduce
  - Gadgets sociales de código abierto
- Se pueden crear nuevos tableros personalizados!



Big R: La Integración Completa de R en IBM BigInsights

- Explorar, visualizar, transformar, y modelar Big Data utilizando la sintaxis familiar y paradigma R
- R escalable
  - Particiones de gran volúmenes de datos
  - Ejecución de clúster paralelo
  - Todo del entorno R
- Aprendizaje de maquina escalable

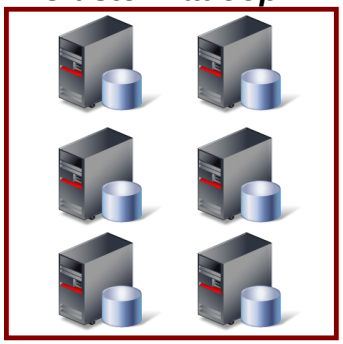


Demostrando el Valor de BigInsights con Experimentos de Investigación

Experiencia Simplificada

Hace cada parte del ciclo de vida de TI mas fácil

Clúster Hadoop

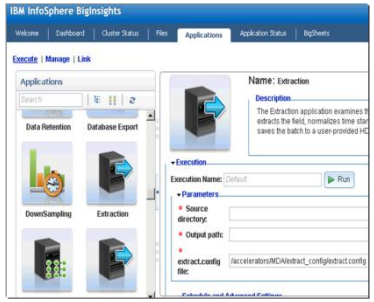


1

Levantando un Clúster Hadoop: BigInsights vs. Hacerlo usted mismo

Pericia Integrada

Captura y automatiza lo que los expertos hacen



2

Desarrollo de solución Hadoop: Machine Data Accelerator vs. Hacerlo usted mismo

Aceleradores BigInsights

- Aplicaciones aceleradores proporcionan un procesamiento de datos y la lógica de negocio adaptados a casos de uso específicos
  - Reducir la complejidad de la creación de aplicaciones Big Data
  - Mejorar el tiempo de valorar para implementaciones Big Data

- Machine Data Analytics Accelerator (MDA)
  - Analizar y extraer de gran variedad de datos de máquina
  - Búsqueda facetada para una fácil navegación
  - La visualización para facilitar el análisis de los datos

- Social Data Analytics Accelerator (SDA)
  - Analizar grandes volúmenes de diversos tipos de datos de los medios sociales con procesamiento en tiempo real

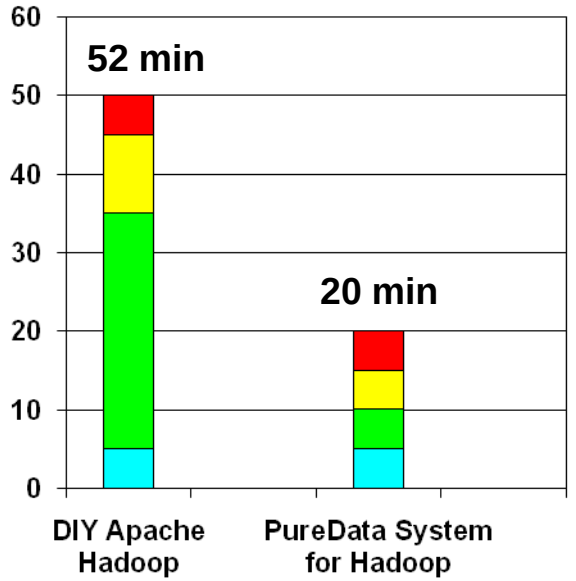
Aceleradores de Datos de Maquina

- Aplicaciones listos para usarse para crear soluciones de análisis de datos generados por maquina
  - Registros de sistema, e-mails, base de datos, y Web
- Despliegue, administrar, y ejecutar de la consola BigInsights
  - Importación, extracto, Índice, búsqueda, Unión, transformación, Análisis
- Reutilice o personalícelos para crear soluciones de análisis de Big Data rápidamente
  - BigInsights plugin de eclipse proporciona soporte de herramientas

Aceleradores de Datos de Maquina Significativamente Aumentan la Productividad del Desarrollador

Análisis de registros

Tarea	DIY	MDA
Configurar IDE	5 min	5 min
Desarrollar código	30 min	5 min
Empaquetar e implementar	10 min	5 min
Probar el código	7 min	5 min
Líneas de código	57	7



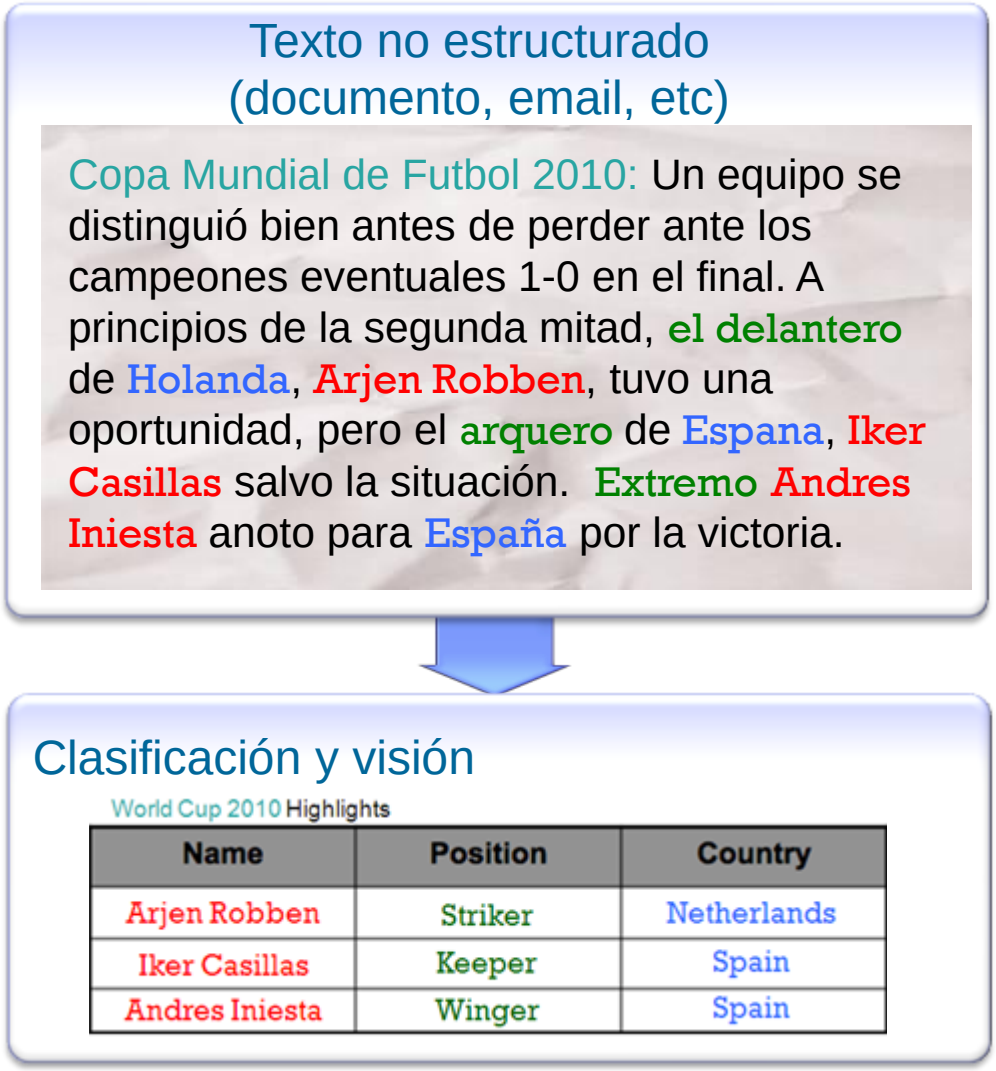
Machine Data Accelerator redujo el tiempo de desarrollo a la mitad para trabajos de análisis de archivos de registro

Requiere 8x menos código

The productivity gains depend on the scope of test case. More reuse of MDA modules will give bigger improvement

BigInsights Tiene Capacidad del Análisis de Texto Incorporado

- Destila información estructurada de texto no estructurado
  - El análisis de sentimiento
  - Comportamiento de consumidor
  - Actividad Ilegal o sospechosa
- Analiza texto y extrae información del email, blogs, redes sociales
- Incluye:
  - Lenguaje de consulta de texto - AQL
  - Herramientas de desarrollo extensas
  - Compilador y generador de perfiles
  - Varios modos de invocar y visualizar los resultados en BigInsights



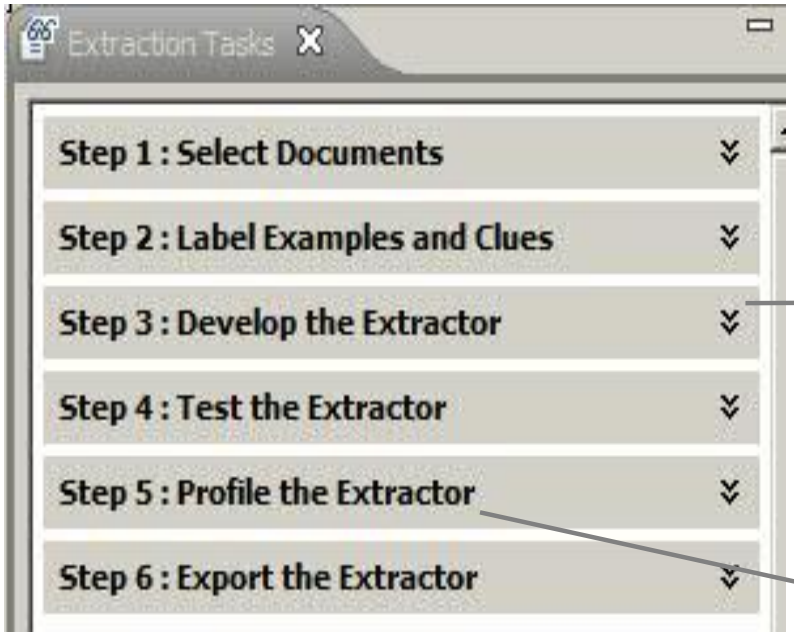
Annotated Query Language (AQL) Permite Extracción de Inteligencia del Texto no Estructurado

- Lenguaje declarativo similar a SQL
  - Muy fácil de leer y entender rápidamente
- Permite la definición de patrones, reglas y lógica relacional para extraer
- Permite extracción precisa de información
- Incluye extractores pre-construidos para; nombres, direcciones, y números de teléfono, etc.
- Soporte para ingles, español, francés, alemán, portugués, holandés, japonés, chino

Uso de la expresión regular para extraer números de teléfono

```
create view Phone as
extract regexes /\d{3} {-} \d{3} {-} \d{4}/ on D.text as num
from Document D;
```

Amplia Herramienta Basada en Eclipse Para Guiar a los Desarrolladores

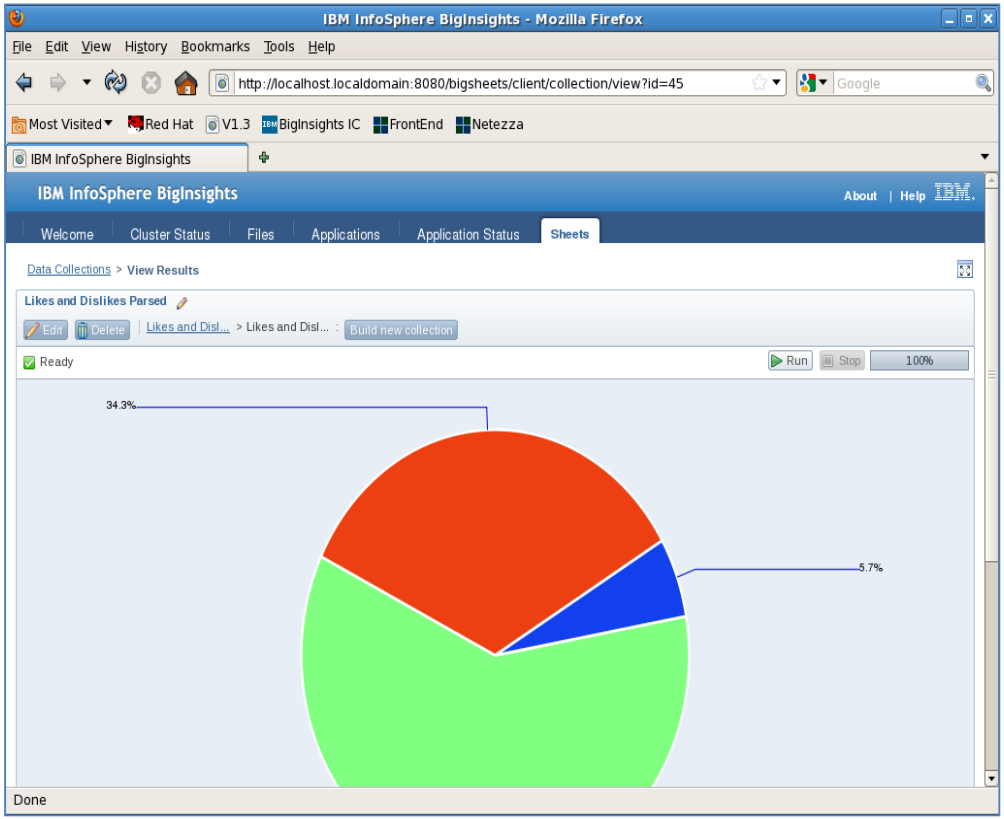


Guía de paso a paso

Cada paso expone una poderosa herramienta utilizada durante el desarrollo de los programas de análisis de texto

El runtime expone un perfilador para ayudar a entender el rendimiento de un extractor

Aplicaciones de Análisis de Texto se Integran con Otras Funciones de BigInsights



- Utilizar la aplicación de la analítica de texto como función en BigSheets
- Visualizar los resultados con BigSheets
- Ejecutar desde una pestana de la aplicación BigInsights

BigInsights Simplifica Big Data Para la Empresa

- Aceleradores integrados
- Herramientas de análisis de texto integradas
- Herramientas de visualización integradas
- Herramientas para integrar habilidades personalizadas
- Motor Big SQL
- Soporte por el lenguaje R
- Herramientas para todas las funciones
  - Desarrollador, administrador, usuario de negocio

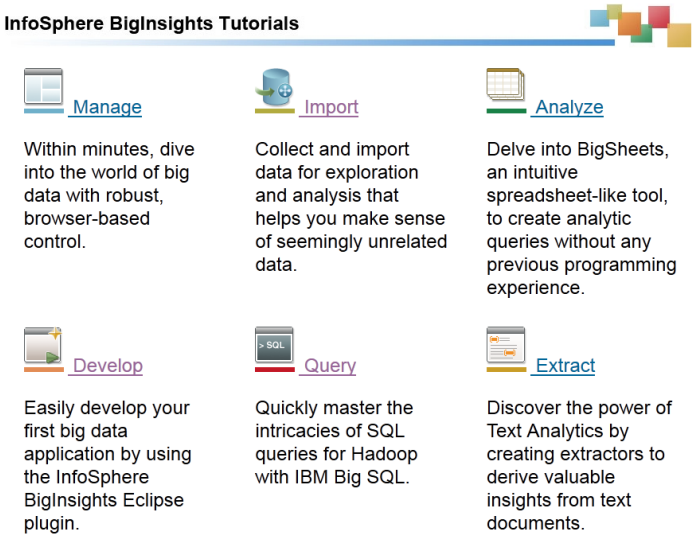
34X Mas Rápido
Levantando el Clúster

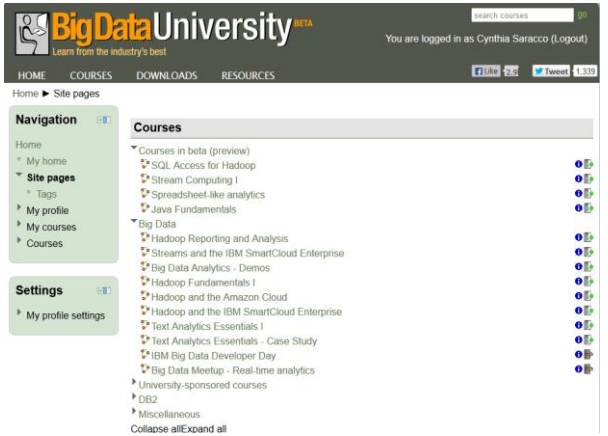
2X Mas Rápido
Creación de Aplicaciones Hadoop

1.7X Mas Rápido
Ejecutando el Benchmark Terasort

¿Le Gustaría Saber Más?

- **Descargue el Quick Start Edition**
  - ibm.com/infosphere/quickstart
- **Pruebe las tecnologias**
  - Siga tutoriales online
  - Inscríbase en clases on-line
  - Vea demostraciones en video, lea artículos, etc.
- **Links a todos recursos desde el BigInsights wiki**
  - Portal técnico en http://tinyurl.com/biginsights





Agenda de Hoy

Time	Topic
09:00 – 09:15 AM	Introducción: Lo Que la Analítica Big Data Puede Hacer Para Su Negocio
09:15 – 10:00 AM	Domine los Fundamentos: Analizando datos estructurados con sistemas PureData System for Analytics
10:00 – 10:15 AM	**Break**
10:15 – 11:00 AM	La Analítica de Datos no Estructurados: Análisis Big Data con Hadoop
11:00 – 11:45 AM	Amplía tu Estrategia de Análisis: Modernización de Almacén de Datos
11:45 – 12:00 PM	Resumen y Acción