Para: desarrolladores

Por: Raúl Bareño Gutiérrez

Empecemos desde ceros .. y unos..

bit es la unidad mínima de almacenamiento

1 bit = 0.0000000012 m = 12 atomos

posibles valores 0 1

En $1cm^2$ se pueden llegar a meter 1.5 Tb

$1cm^2$ = 12,000,000,000,000 bits

byte

1 byte = 8 bits

256 combinaciones

Es posible expresar cualquier carácter alfabético en un byte.

Hagan la prueba guardando un archivo de texto con una sola letra. Pesará 1 byte.

ASCII, UTF-8

Más allá del kilo, mega, giga, tera ..

Claro que hay más allá del Tera, y es importante que alguien que trabaje con Big Data no se asuste de escuchar unidades estratosféricas ...

kilo	$1x10^3$
Mega	$1x10^6$
giga	$1x10^9$
tera	$1x10^{12}$

¿Cuales siguen?

peta	$1x10^{15}$
exa	$1x10^{18}$
zetta	$1x10^{21}$
yotta	$1x10^{24}$
xonao hella	$1x10^{27}$
weka	$1x10^{30}$
vunda	$1x10^{33}$
uda	$1x10^{36}$
treda	$1x10^{39}$

y siguen ...

sorta	$1x10^{42}$
rinta	$1x10^{45}$
quexa	$1x10^{48}$
pepta	$1x10^{51}$
ocha	$1x10^{54}$
nena	$1x10^{57}$
minga	$1x10^{60}$
luma	$1x10^{63}$

continúan los numerales

Undecillion	$1x10^{66}$
Undecilliarde	$1x10^{69}$
Googol	$1x10^{100}$
Sexvigintillion	$1x10^{156}$
Zentillion	$1x10^{600}$
Googolplex	$1x10^{Googol}$
Googolplexplex	$1x10^{Googolplex}$
Googolplexplexplex	$1x10^{Googolplexplex}$

tabla

Unidades de información (del byte)			
Sistema Internacional (decimal)		**ISO/IEC 80000-13 (binario)**	
Múltiplo (símbolo)	**SI**	**Múltiplo (símbolo)**	**ISO/IEC**
kilobyte (kB)	10^3	kibibyte (KiB)	2^{10}
megabyte (MB)	10^6	mebibyte (MiB)	2^{20}
gigabyte (GB)	10^9	**gibibyte** (GiB)	2^{30}
terabyte (TB)	10^{12}	tebibyte (TiB)	2^{40}
petabyte (PB)	10^{15}	pebibyte (PiB)	2^{50}
exabyte (EB)	10^{18}	exbibyte (EiB)	2^{60}
zettabyte (ZB)	10^{21}	zebibyte (ZiB)	2^{70}
yottabyte (YB)	10^{24}	yobibyte (YiB)	2^{80}

cifras, cifras y más cifras

Cada 5 minutos se genera 1 exabyte de datos

= 1,000,000,000,000,000,000 bytes

y.... ¿Qué hacemos con tanta información?

Minarla ….

Así como los mineros encuentran piedras preciosas entre tanta tierra, en el caso de los datos es lo mismo, intentar encontrar patrones que ayuden a tomar decisiones es un arte….

y… ¿Qué se usa?

Las bases de datos tradicionales tienen algunos límites, por ejemplo:

MySQL soporta hasta 4GB en discos duros FAT o hasta 2TB en discos duros Windows NTFS, Linux ext3 y Mac HFS+

Por lo que se usan herramientas especiales.

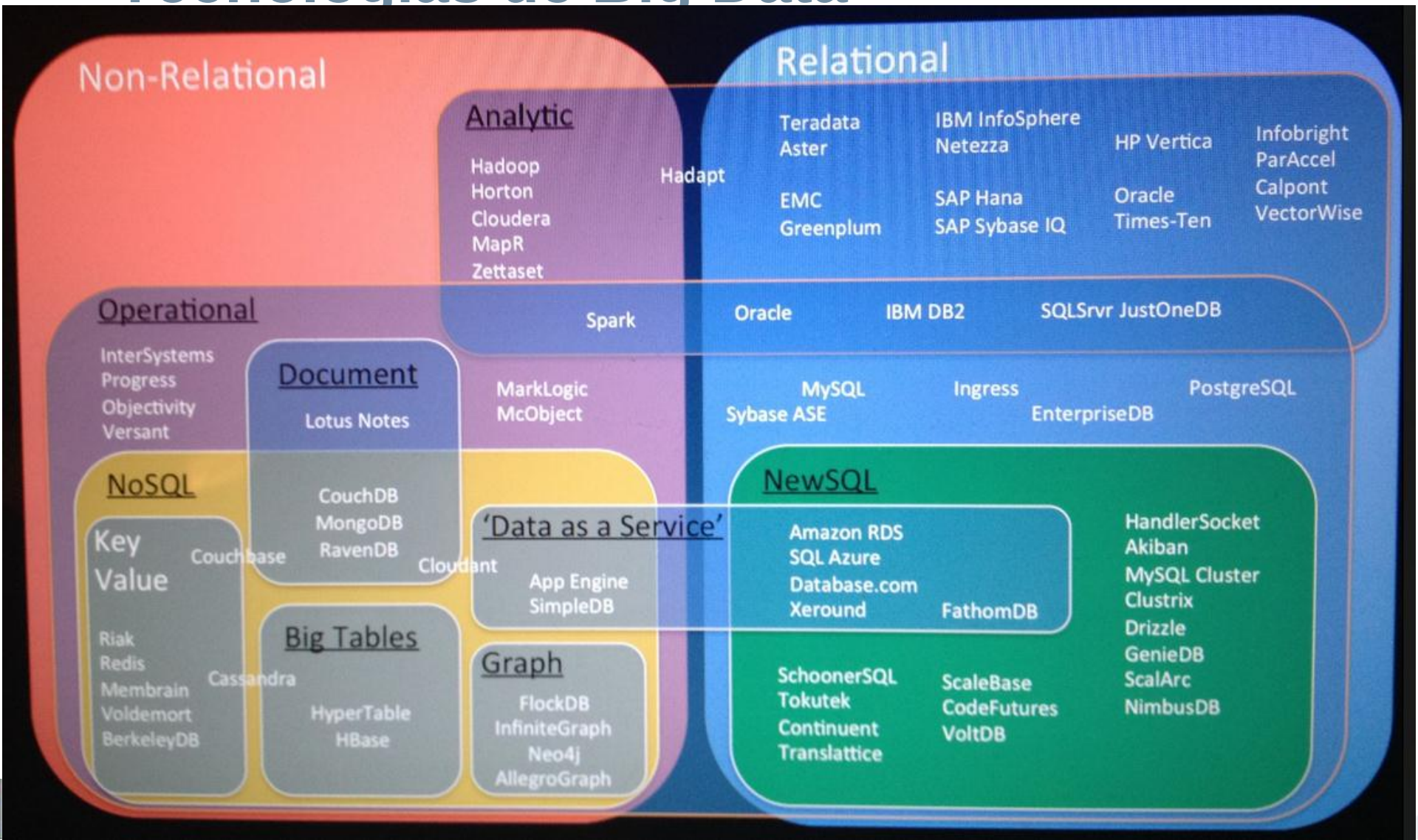
y… ¿Cuáles sí soportan?

Los más grandes del mercado son estos, pero existen alternativas Open Source como por ejemplo Hadoop o Cassandra.

Bases de datos que usan las grandes empresas

Facebook:	RocksDB
Amazon:	Dynamo
Google:	BigTable
Foursquare:	MongoDB
Twitter:	Cassandra
Wikipedia:	MariaDB
Ebay:	BerkeleyDB
Yahoo:	Oracle
Microsoft:	SQL Server

Tecnologías de Big Data



Bases de datos relacionales

Permiten establecer interconexiones (relaciones) entre los datos (que están guardados en tablas), y a través de dichas conexiones relacionar los datos de ambas tablas

Bases de datos no relacionales

Los datos almacenados no requieren estructuras fijas como tablas, no garantizan completamente ACID (atomicidad, coherencia, aislamiento y durabilidad), y habitualmente escalan bien horizontalmente.

NewSQL

Nace en el 2011 y trata de conseguir el mismo rendimiento escalable de sistemas no relacionales para el procesamiento de transacciones en línea y garantiza el ACID de un sistema de base de datos tradicional

NoSQL

No usan SQL como el principal lenguaje de consultas. Las principales compañías de Internet se dieron cuenta que el rendimiento era más importantes que cuidar la coherencia.

Key Value

Relacionan una llave con un valor, este es el principio fundamental que logra que consultas se ejecuten instantáneamente en bases de datos de muy muy alta escala. Buscar en bases de datos de quintillones de registros es instantáneo gracias a esto.

Funcionan mediante arreglos

En la programación hay una estructura de datos muy común que se llaman arreglos, que consiste en guardar varios valores en una variable y se mandan llamar por su posición.

Ejemplo de un arreglo

```
var arreglo = [];
arreglo[0] = “Hola”;
arreglo[1] = “Amigo”;
arreglo[2] = “Sentado”;
arreglo[3] = “Atrás”;
```

Extraigamos información

Si requerimos en valor de la casilla 1 hacemos lo siguiente:

arreglo[1]

Y nos devuelve “Amigo”

Pasa lo mismo con textos

Pero se le suelen llamar mapas (éste es la base de Big Data).

```
var mapa = {};
mapa[“que”] = “Adios”;
mapa[“mal”] = “Amigo”;
mapa[“ejem”] = “Dormido”;
mapa[“plo”] = “Adelante”;
```

Mandamos traer un registro

Si queremos traer el valor de la posición “ejem” entonces sería así:

mapa[“ejem”]

Y nos devolverá:

“Dormido”

¿Qué es lo que hace?

Inmediatamente ubica la posición de memoria ya sea dependiendo de la posición o trae el registro directamente de una tabla de Hash que genera internamente, lo que lo hace instantaneo y no necesita “buscar”, solo lo trae.

Podemos guardar una tabla

Ahora podemos hacer esto

```
var tabla = {}
tabla[“user1”] = {nombre:”Raul”,edad:48};
tabla[“user2”] = {nombre:”Robert”,edad:53};
tabla[“user3”] = {nombre:”Claudia”,edad:33};
tabla[“user4”] = {nombre:”Adriana”,edad:45};
```

Y mandarla traer

Si queremos traer la información de “user2” lo llamamos así

tabla[“user2”]

y nos trae sus datos

{nombre:”Robert”,edad:53}

O pedir algunos datos

O podemos pedir solo algún dato

tabla[“user2”][“nombre”]

Y regresa:

Robert

Así de sencillo funciona Big Data

De esta manera es como funciona el “Map - Reduce” que es la base de Big Data.

BigTable

Como mencionamos, las grandes empresas como Google necesitan velocidad en sus búsquedas y no podían perder tiempo buscando en miles de tablas, por lo que todo lo pusieron en una sola, con miles y miles de columnas, de ahí nació BigTable.

Esta super tabla no tiene porque estar en una sola computadora, puede estar distribuida en una granja de servidores.

Ejemplo

user_id	user_name	bank_name	bank_cash	group_name	school_name
dsf5ds6fds	Robert	GNB suda	1	teacher	UMB
sdfsdf7678	Arnolfo	Benemex	40	student	UIS
879s7dfsd	Juanelo	Banarte	20	student	UIS
hjgsdf653a	Petronilo	Banarte	100	student	UMB
6sdf58sd5f	Proculo	Scotte	250	staff	UMB
75dsfsdfs7	Agapito	Benemex	80	student	UIS
sdfsd76sf5	Panfilo	Scotte	133	student	UMB
dsf678sd6f	Raul	BBVA	244	teacher	UIS
usdf8sf6s8f	Anivdelarev	Scotte	412	staff	UIS
d6s6fs7df6	Delfino	Benemex	44	student	UMB
sdf6s78f6s	Ruperto	Scotte	2	student	UIS

Hagamos una consulta

Las búsquedas son por llave valor, ejemplo

{user_id:”dsf678sd6f”}

Y nos regresa:

[
{user_id:”dsf678sd6f”,user_name:”Raul”,bank_name:”BBVA”,bank_cash:”244”,
group_name:”teacher”,school_name:”UIS”}
]

Ahora por otro registro

Las búsquedas son por llave valor, ejemplo

{group_name:”teacher”}

Y nos regresa 2 registros:

[
{user_id:”dsf5ds6fds”,user_name:”Robert”,bank_name:”GNB
suda”,bank_cash:”1”,group_name:”teacher”,school_name:”UMB”},
{user_id:”dsf678sd6f”,user_name:”Raul”,bank_name:”BBVA”,bank_cash:”244”,group_name:”teacher”,sc
hool_name:”UIS”}
]

MongoDB

Cuenta con una versión estable desde el 2011. Es una de las principales plataformas usadas para Big Data debido a la escalabilidad, el uso de NoSQL y el eficiente uso de llave-valor. Es gratuita y de código abierto.

Funciona exactamente como el ejemplo

Hagamos el ejemplo en mongo

db.table.find({user_id:”dsf678sd6f”})

Y nos regresa exactamente lo mismo

[
{user_id:”dsf678sd6f”,user_name:”Raul”,bank_name:”BBVA”,bank_cash:”244”,
group_name:”teacher”,school_name:”UIS”}
]

Tutorial de MongoDB

http://docs.mongodb.org/manual/tutorial/getting-started/

Y de como probarlo

http://www.valleyprogramming.com/blog/big-data-datasets-large-examples-boulder-colorado-hadoop-mongodb

Prueba BigQuery de Google

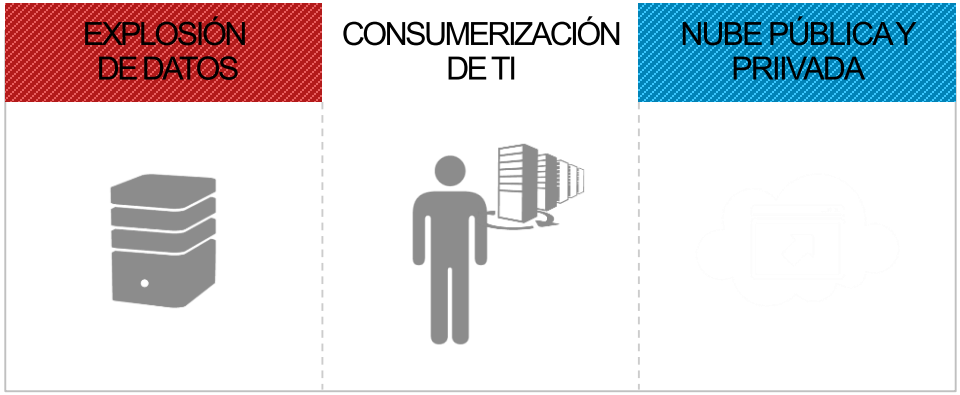
https://developers.google.com/bigquery/

¿Qué es Big Data?

"Volumen masivo de datos, tanto estructurados como no-estructurados, los cuales son demasiado grandes y difíciles de procesar con las bases de datos y el software tradicionales" (ONU, 2012)

LA REVOLUCIÓN DE LOS DATOS

Tendencias de la Industria



Los datos digitales crecerán **44x** próxima década

En 2016, servicios de nube pública tendrán 46% de crecimiento neto en gasto de TI

Introducción: La revolución de los datos

- Big Data, Data Science y lo que nos pueden proporcionar
- Actores de mercado en Big Data
- Utilidades de Big Data

Las 4 V's

Volumen

Velocidad

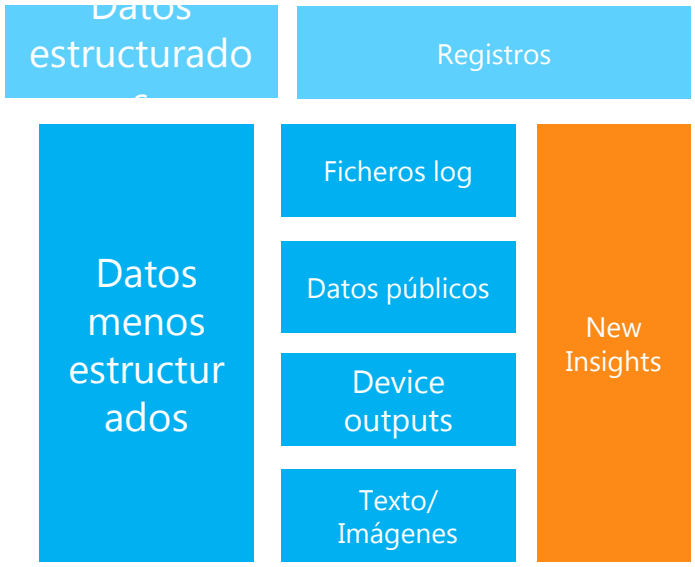
Variedad

Variabilidad

Ejemplos de Big Data

facebook	12 Tb día	21 Pb Hadoop cluster
bing	7 Pb mes	
twitter	1 Tb tweets/dia	7 Tb datos/dia
KLOUT	75 Millio n	4 Billion Graph edg/day
YAHOO!	14 Tb Hadoop cluster	

Entonces…¿cómo obtengo insights?



- Datos estructurados
  - Bases de Datos relacionales
  - Bases de Datos analíticas
- Datos menos estructurados
  - Intentar un ETL para transformarlo en relacional
    - Tiempo de desarrollo elevado
    - Son datos susceptibles a cambios de estructura
  - Archivados y Borrados
  - Acceso caro

¿Qué es Big Data para nosotros?

- Datos, datos, datos…..
- ¿Big? Hoy es Big, pero dentro de 2-3 años, ¿seguirá siendo big?
- Equipos y negocios “centrados en datos”
- ***Big Data: “ Barreras para que una organización o equipo puedan almacenar, procesar y acceder todos los datos que necesitan para operar con eficiencia, tomar decisiones, reducir riesgos, etc.”***

Vista completa de nuestros usuarios

Seguridad / inteligencia

Operaciones Analíticas (TI, dispositivos,…)

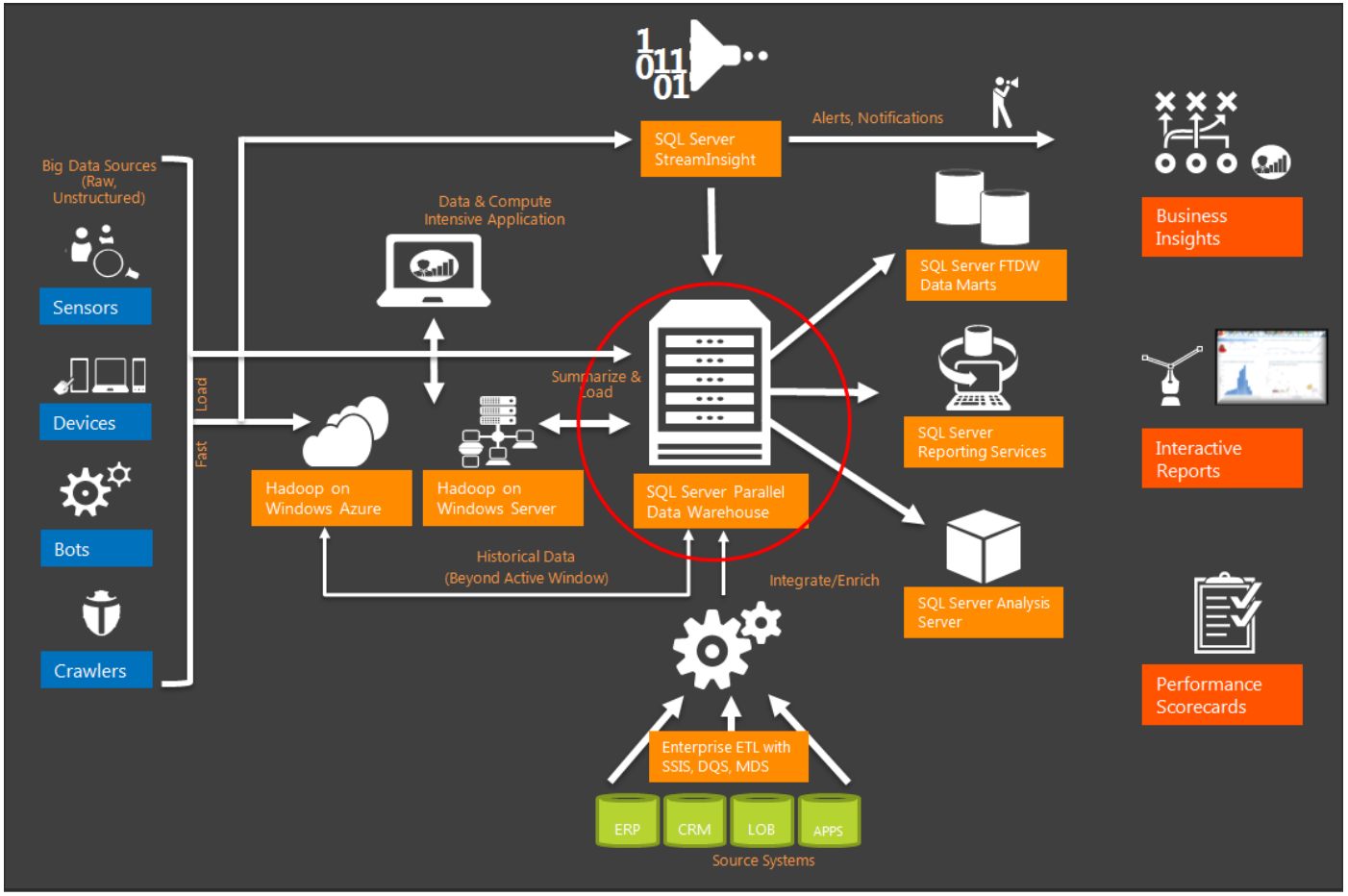
Utilidades de Big Data

MIDAMO


MIDAMO MANAGEMENT STUDIO
Temas
Configuración
Riesgo:
% Moderado:
Posts Moderados:
Posts totales:
Deportes 0,2 77,92% 8236,92 posts 10571 posts
Infantil 0,1 88,29% 2074,82 posts 2350 posts
Cultura 0,45 47,58% 2715,11 posts 5707 posts
Tecnología 0,6 30,4% 2706,51 posts 8903 posts
Política 0,25 71,06% 6939,25 posts 9765 posts
Total 0,32 63,05% 22672,61 posts 37296 posts
Aplicar cambios


SU NOMBRE FUE TAN COREADO COMO EL DE KAKÁ
Florentino sí que es un galáctico
Las más de 40.000 personas que se dieron cita en el coliseo de la Castellana tenían tantas ganas de ver a Kaká de blanco como de corear el nombre de la persona que ya trajo a Figo, Zidane, Ronaldo y Beckham .
Jose Antonio 12:30 29/7
A uno le cuesta imaginar cómo será la presentación de Cristiano Ronaldo, actual Balón de Oro y el traspaso más caro en la historia del fútbol, el próximo 6 de julio, pero desde ya les digo que el portugués tendrá muy difícil superar el espectáculo vivido la tarde de este martes en el Santiago Bernabéu.
Porque la presentación de hoy ha sido doble. La afición merengue ha tenido al fin la posibilidad de vitorear a su nuevo héroe y, de paso, al hombre que lo ha hecho posible, Florentino Pérez. La primera toma de contacto del presidente madridista con su afición en esta segunda etapa desmostró hasta qué punto era deseado su regreso.
Florentino camina por el escenario con la imagen de Raúl al fondo.
Comentarios:
Más comentarios: 85 ...
#85 javi86 Atención, última hora: el virus de Madriditis, según acaba de comentar la ministra de sanidad, alcanza el grado de Pandemia en Cataluña. Asimismo ha declarado que los principales afectados son tanto varones como mujeres, simpatizantes del Barsa. HALA
#84 madridl I belong to Jesús!! I belong to Kaká!! simplemente de pie, callado , sin decir nada, solo sonriendo..... es elegante. Solo podias ir a un sitio y has elegido el mejor. HALA MADRID!!!
Bienvenido jose233, añade comentari
Me parece genial que un equipo como el Madrid fiche a un jugador de estas características
Como Madridista que soy estoy orgulloso de ello, nos va a hacer muy grandes!!
HALA MADRID!!
Publicar
Normas:
Por favor, escribe correctamente, sin mayúsculas ni abreviaturas.
Recuerda que el tono del mensaje debe ser respetuoso. No se admitirán insultos ni faltas de respeto.

ESCENARIOS DE BUSINESS ANALYTICS


Big Data Sources (Raw, Unstructured)
Sensors
Devices
Bots
Crawlers
Load
Fast
SQL Server StreamInsight
Alerts, Notifications
Data & Compute Intensive Application
Summarize & Load
Hadoop on Windows Azure
Hadoop on Windows Server
SQL Server Parallel Data Warehouse
Historical Data (Beyond Active Window)
Integrate/Enrich
SQL Server FTDW Data Marts
SQL Server Reporting Services
SQL Server Analysis Server
Business Insights
Interactive Reports
Performance Scorecards
Enterprise ETL with SSIS, DQS, MDS
ERP
CRM
LOB
APPS
Source Systems

Escenarios de Business Analytics

Análisis sencillo de gran cantidad de datos no estructurados: Microsoft HDInsight

Análisis sencillo de datos en memoria: Microsoft StreamInsight

Análisis en profundidad: SQL Server y Self-Service BI

¿Qué es Hadoop?

- Open Source
- Plataforma de almacenamiento de datos y análisis para Big Data
- Optimizado para manejar
  - Datos masivos a través de paralelismo
  - Variedad de datos (Estructurados, No-estructurados, Menos estructurados)
  - Uso de hardware económico

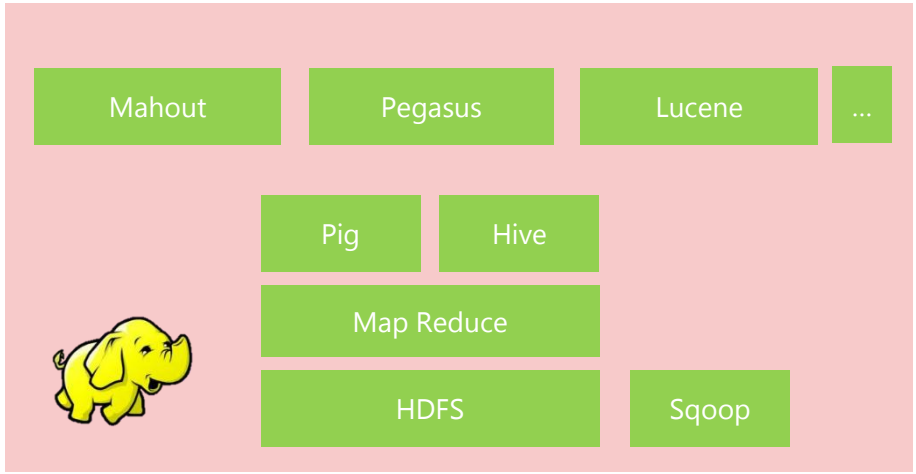
¿Qué es Hadoop?: Ventajas

- Escalable
- Escala linealmente en capacidad de almacenamiento y computación

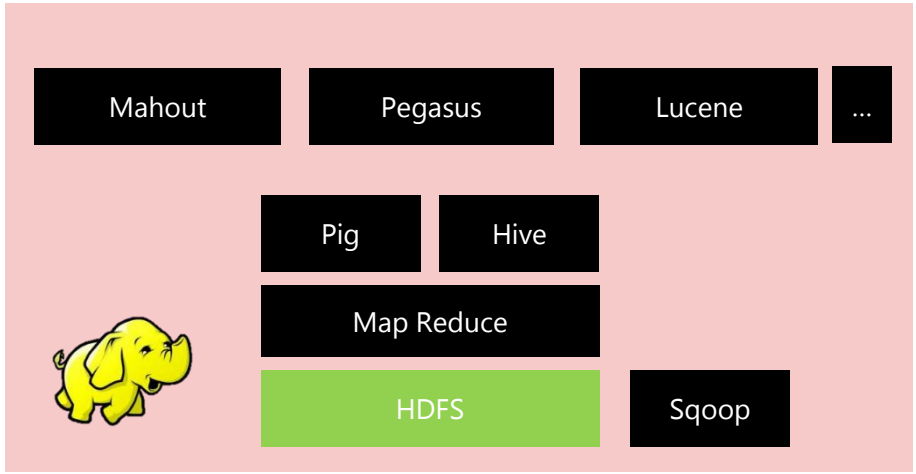
- Tolerante a Fallos
- Proporcionado por el Sistema de ficheros distribuido y el framework de lectura

- Procesamiento distribuido
- Sigue la estrategia de divide y vencerás

¿Qué es Hadoop?: Componentes



¿Qué es Hadoop?: Componentes



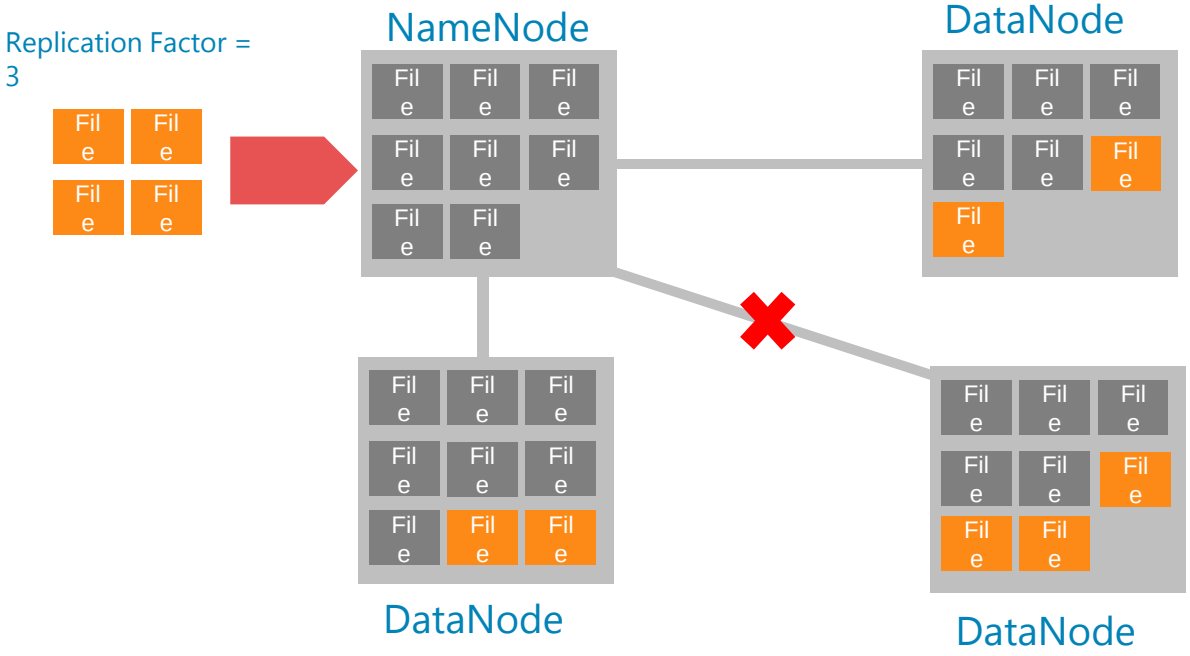
Hadoop Distributed File System (HDFS)

- Sistema de ficheros distribuido diseñado para grandes conjuntos de datos
- Fiable y con buen rendimiento
  - Alto rendimiento de acceso: Latencia de disco
  - Alto ancho de banda Almacenamiento Clustered auto-reparable
- Divide los datos entre los nodos en un Cluster
  - NameNode: Mantiene el mapeo de bloques de ficheros a nodos esclavos
  - DataNode: Almacena y sirve bloques de datos

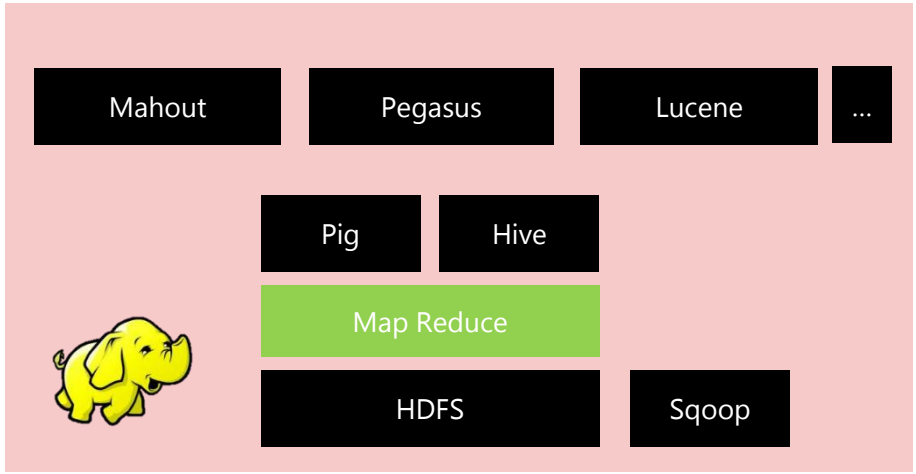
Hadoop Distributed File System (HDFS)

Block Size = 64 Mb

Replication Factor = 3



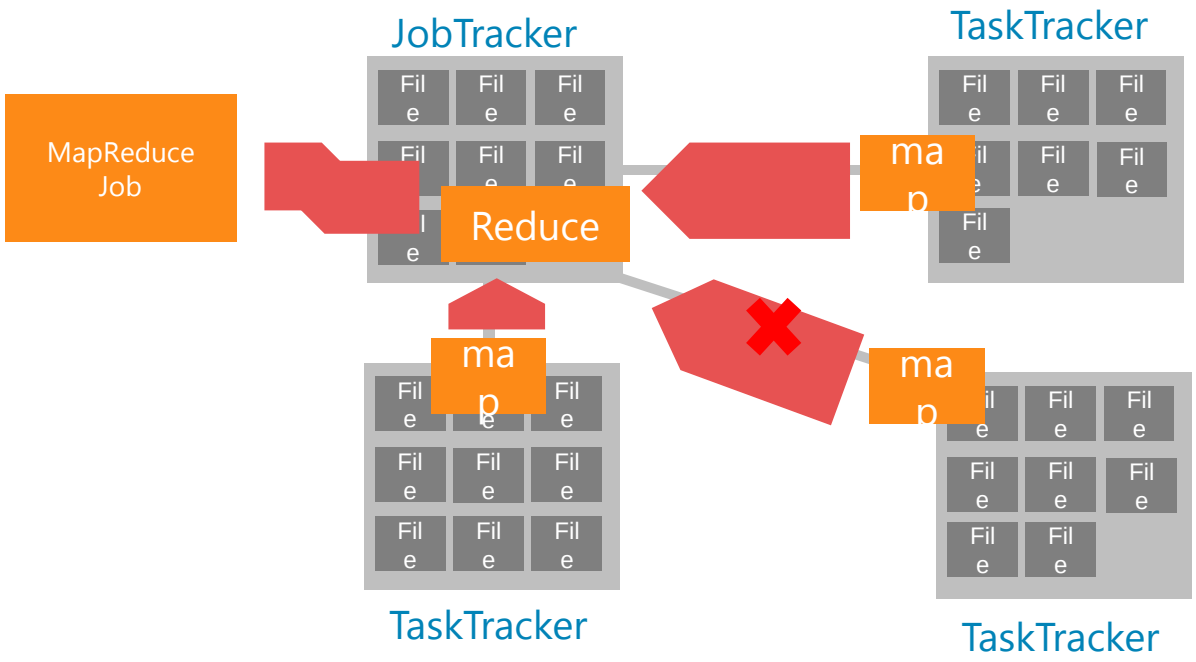
¿Qué es Hadoop?: Componentes



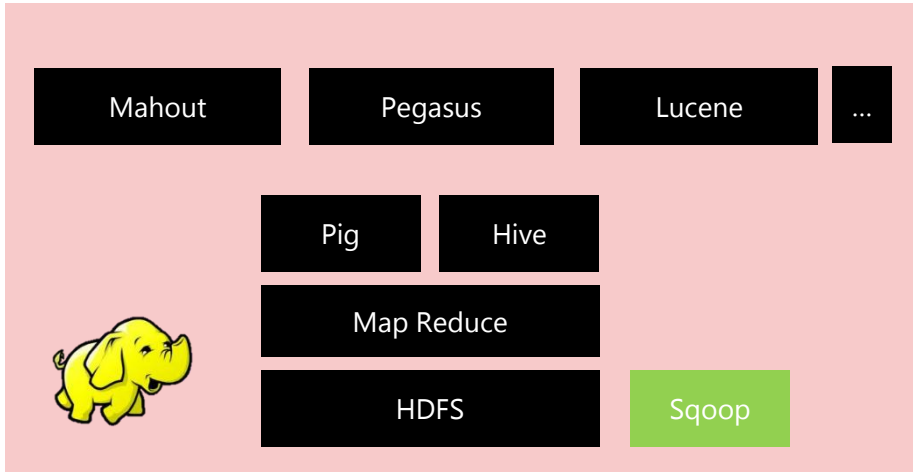
Map Reduce Framework

- Motor de planificación parar Procesamiento de carga distribuido
  - Pares Clave-Valor
  - Función Map
  - Función Reduce
- Lenguajes de Script : Java, python, Javascript…
- Saca provecho de la distribución de datos de HDFS
  - JobTracker: Planifica los trabajos entre los TaskTrackers
  - TaskTracker: unidades de trabajo

Map Reduce Framework



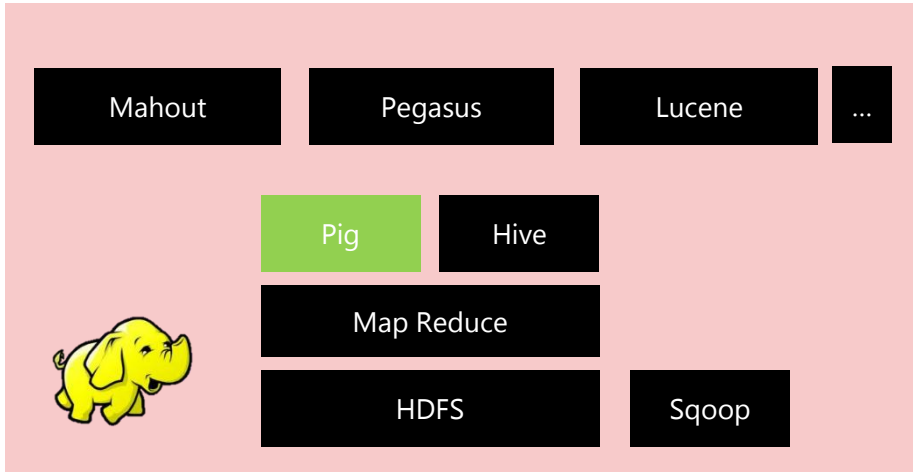
¿Qué es Hadoop?: Componentes



Sqoop

- Tecnología que sirve de interfaz entre HDFS y los Sistemas de información empresarial
- Orígenes de datos relacionales integrados
  - MySQL, Oracle, SQL Server …
- Importación / Exportación (Bidireccional)

¿Qué es Hadoop?: Componentes



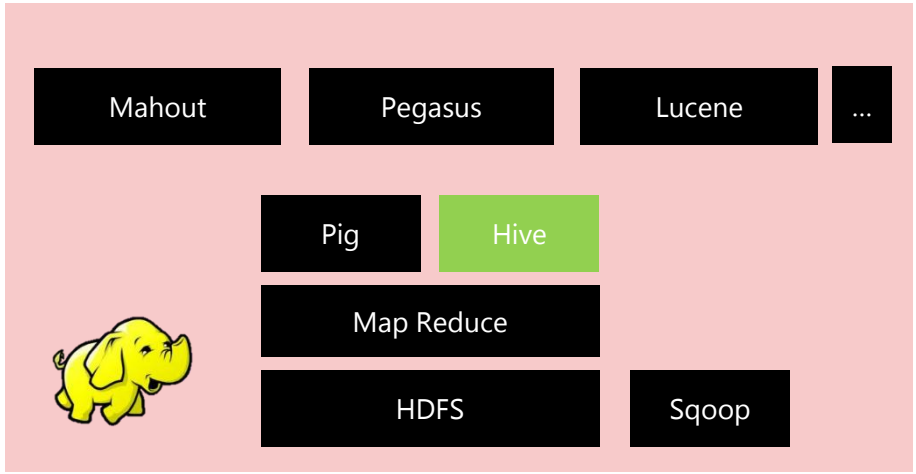
Pig

- Lenguaje de flujo de datos de alto nivel y framework de ejecución
- Lenguaje de consulta: PigLatin
  - Posibilidad de join de tablas

```
log  = LOAD 'excite-small.log' AS (user, time, query);
grpd = GROUP log BY user;
cntd = FOREACH grpd GENERATE group, COUNT(log);
DUMP cntd;
```

- Por detrás ejecuta trabajos MapReduce

¿Qué es Hadoop?: Componentes



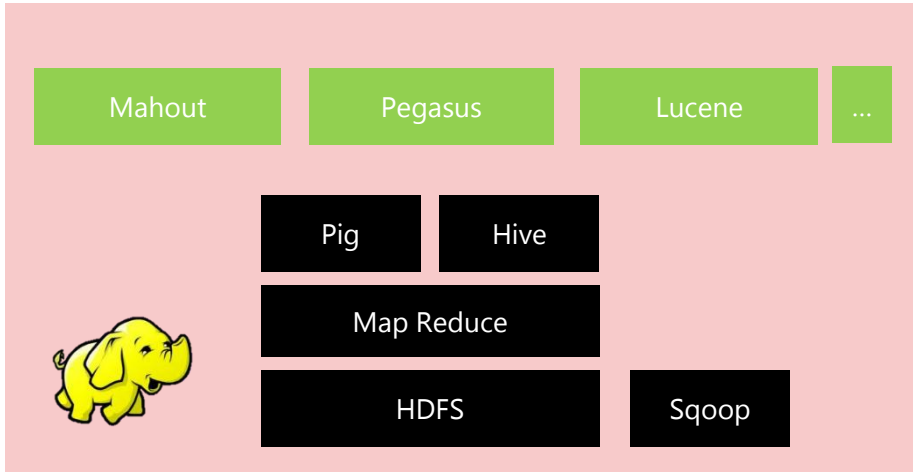
Hive

- Infraestructura Data Warehouse desde Hadoop
- Proporciona
  - Sumarización de Datos
  - Consultas Ad-hoc
- Lenguaje consulta estilo SQL: HiveQL

```
select regexp_replace(split(csuristem, "/")[1], "MainFeed.aspx", "Home"),
count(*)
from weblog_sample
group by regexp_replace(split(csuristem, "/")[1], "MainFeed.aspx", "Home")
```

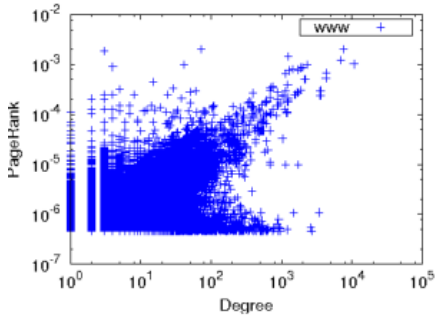
- Por detrás ejecuta trabajos MapReduce

¿Qué es Hadoop?: Componentes



Otros componentes: Hadoop Ecosystem

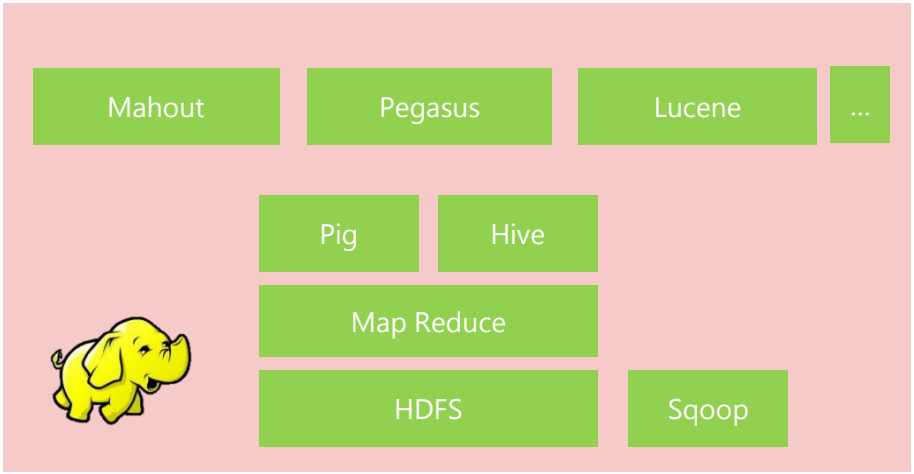
- Mahout
  - Minería de Datos y Machine Learning
- Pegasus
  - Page Rank y Graph Mining
  - Social Network Analysis



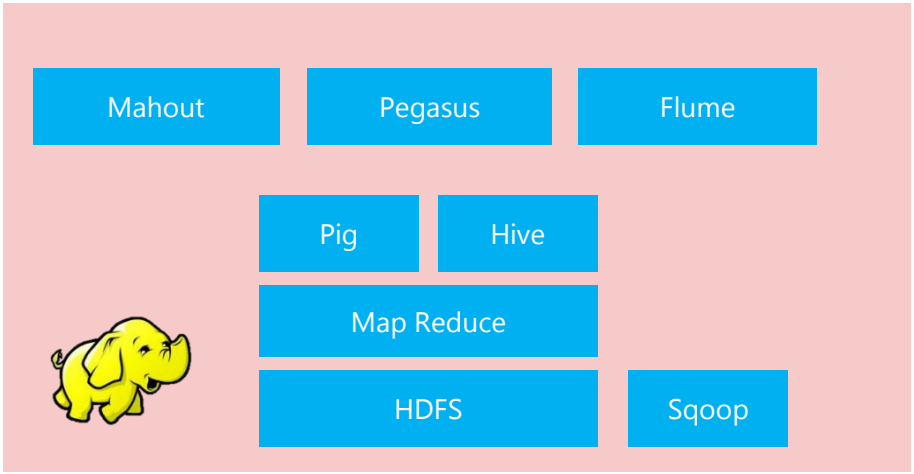
HDInsight

- Project Isotope
- Proporciona Apache Hadoop en
  - Windows Server
  - Windows Azure
- Active Directory & System Center

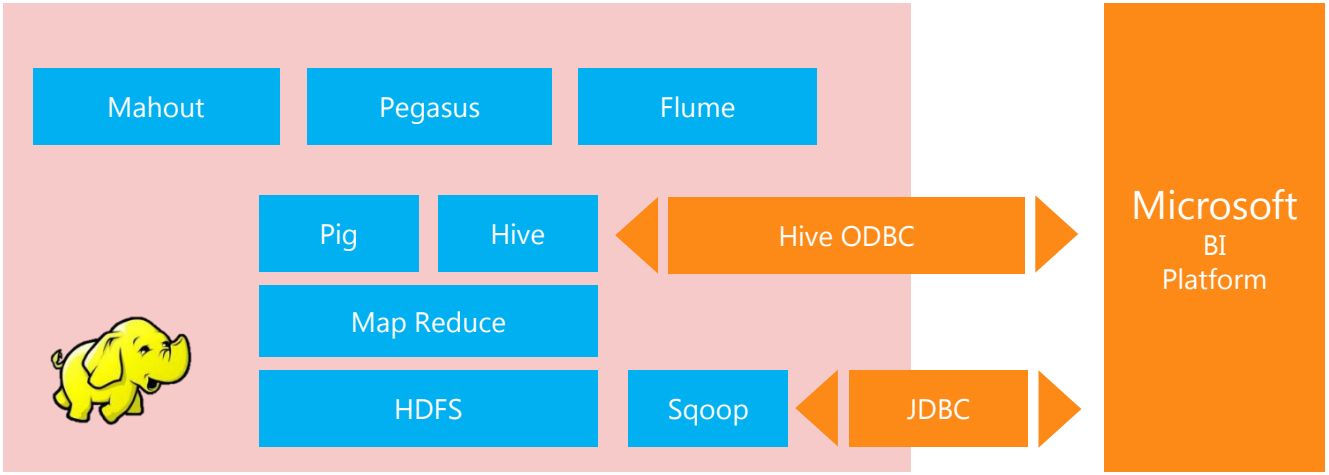
Hadoop: Componentes Originales



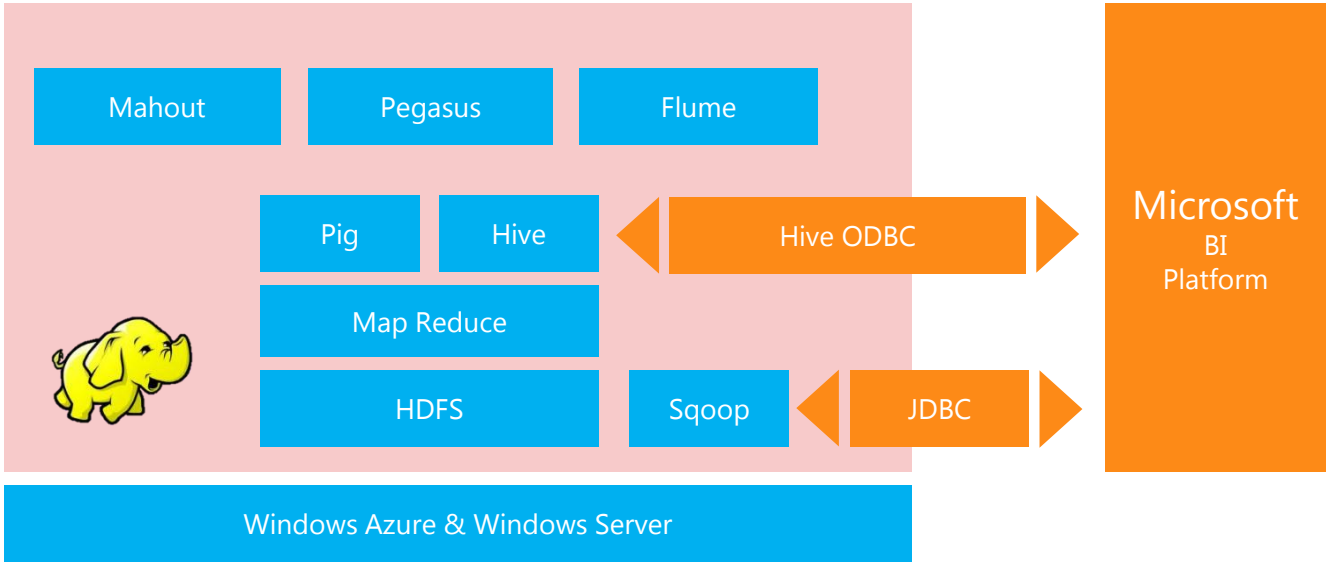
HDInsight



HDinsight


Mahout
Pegasus
Flume
Pig
Hive
Hive ODBC
Map Reduce
HDFS
Sqoop
JDBC
Microsoft
BI
Platform

HDInsight


Mahout
Pegasus
Flume
Pig
Hive
Hive ODBC
Map Reduce
HDFS
Sqoop
JDBC
Microsoft
BI
Platform
Windows Azure & Windows Server

Características HDInsight

- HDFS
- Basado en Windows
- Compatibilidad con Directorio Activo
- Almacenamiento compatible:
  - HDFS
  - Azure Blob Storage
  - Amazon S3

- MapReduce Framework
- Compatibilidad JavaScript
- Hadoop Streaming con compatibilidad F# y C#

Características HDInsight

- Hive
- Consola Interactiva
- Complemento Hive para Excel
- Hive ODBC Driver
- Potentes funciones regex
- Pig
- Consola Interactiva
- Sqoop
- Driver JDBC para SQL Server y SQL Server PDW

Cisco | Networking Academy®
Mind Wide Open™

MUCHAS GRACIAS

CONSTRUIMOS FUTURO