Base de Datos Analítica

Making sense of the Internet of Things.

Software Guru Virtual Conference
30 de abril de 2014
Landy Reyes
Edgar de los Santos

Agenda

- Antecedentes Base de Datos Analíticas
  El internet de las Cosas
- BD Tradicionales vs BD Columnares
- Arquitectura
- Mejores Prácticas y Optimización
- Requerimientos de Software y Hardware
- Infopliance
- Demo

Agenda

- Antecedentes Base de Datos Analíticas

    El internet de las Cosas

- BD Tradicionales vs BD Columnares
- Arquitectura
- Mejores Prácticas y Optimización
- Requerimientos de Software y Hardware
- Infopliance
- Demo

Antecedentes BD Analíticas

"The Internet of Things has the potential to change the world, just as the Internet did. Maybe even more so"

Kevin Ashton

Antecedentes BD Analíticas

Actualmente estamos rodeados de dispositivos que generan datos ...

- Tag del coche
- Pago con tarjeta de crédito
- Llamadas de celular
- Búsquedas en internet
- Sensores de movimiento
- Sensores de temperatura
- Logs de maquinas
- Sportbands

Antecedentes BD Analíticas

Para hacer sentido de los datos necesitas una BD Analítica...

INFOBRIGHT

- Es una base de datos analítica y columnar
- Diseñada para analizar grandes volúmenes de información
- Fácil de usar y administrar

Las bases de datos tradicionales ya no son suficientes para realizar análisis sobre el Internet de las cosas.

Agenda

- Antecedentes Base de Datos Analíticas
    El internet de las Cosas
- BD Tradicionales vs BD Columnares
- Arquitectura
- Mejores Prácticas y Optimización
- Requerimientos de Software y Hardware
- Infopliance
- Demo

BD Tradicionales vs BD Columnares

La diferencia entre una base de datos tradicional y una columnar es …

Employee_ID	Name	Dept	Salary
1	Joe Stevens	Sales	100,000
2	Chuck Berry	Operations	70,000
3	James Dean	Finance	80,000

Datos almacenados en **registros**

1	Joe Stevens	**Sales**	100,000
2	Chuck Berry	Operations	70,000
3	James Dean	Finance	80,000

Tradicional / OLTP

Guarda los valores de un registro como una sola entidad

Datos almacenados en **columnas**

1	Joe Stevens	**Sales**	100,000
2	Chuck Berry	**Operations**	70,000
3	James Dean	**Finance**	80,000

COLUMNAR

Almacena los datos de un registro en columnas

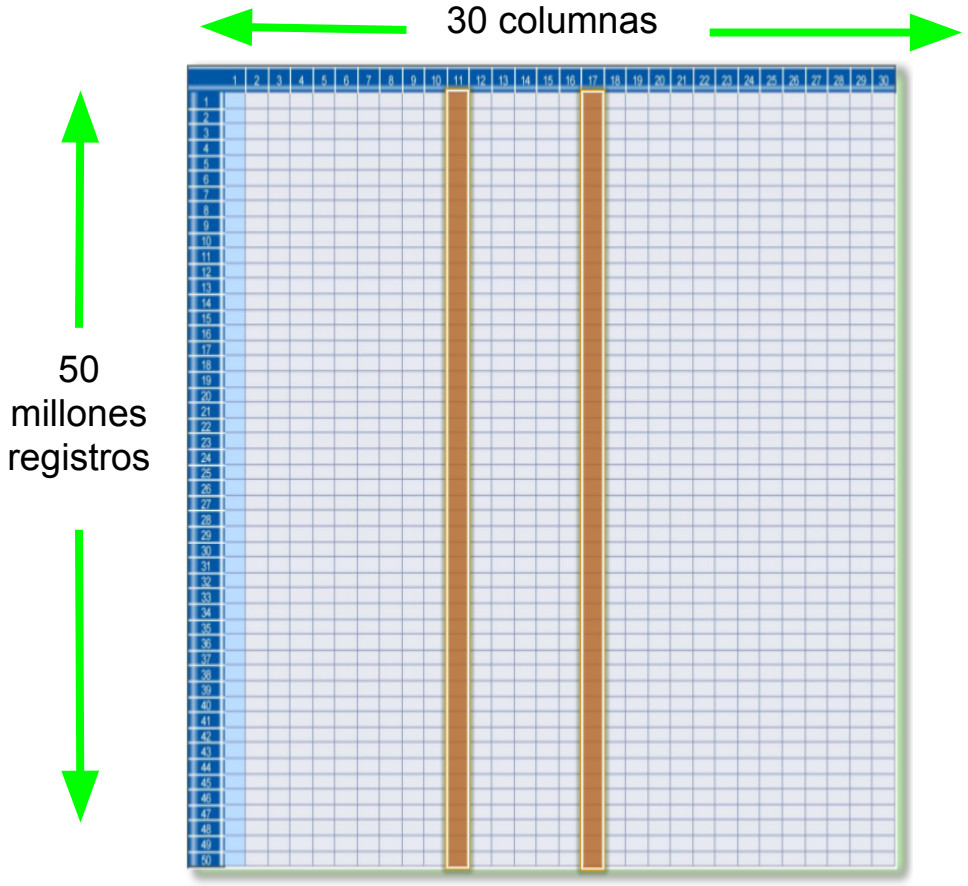
BD Tradicionales vs BD Columnares

Si realizamos una consulta a una tabla con 50 millones de registros ...

Consulta

- 50 días de datos
- 1 millón de registros por día

Select Columna11,
Columna17
Where Columna17
sea la 3a semana
(día 15 – 21)



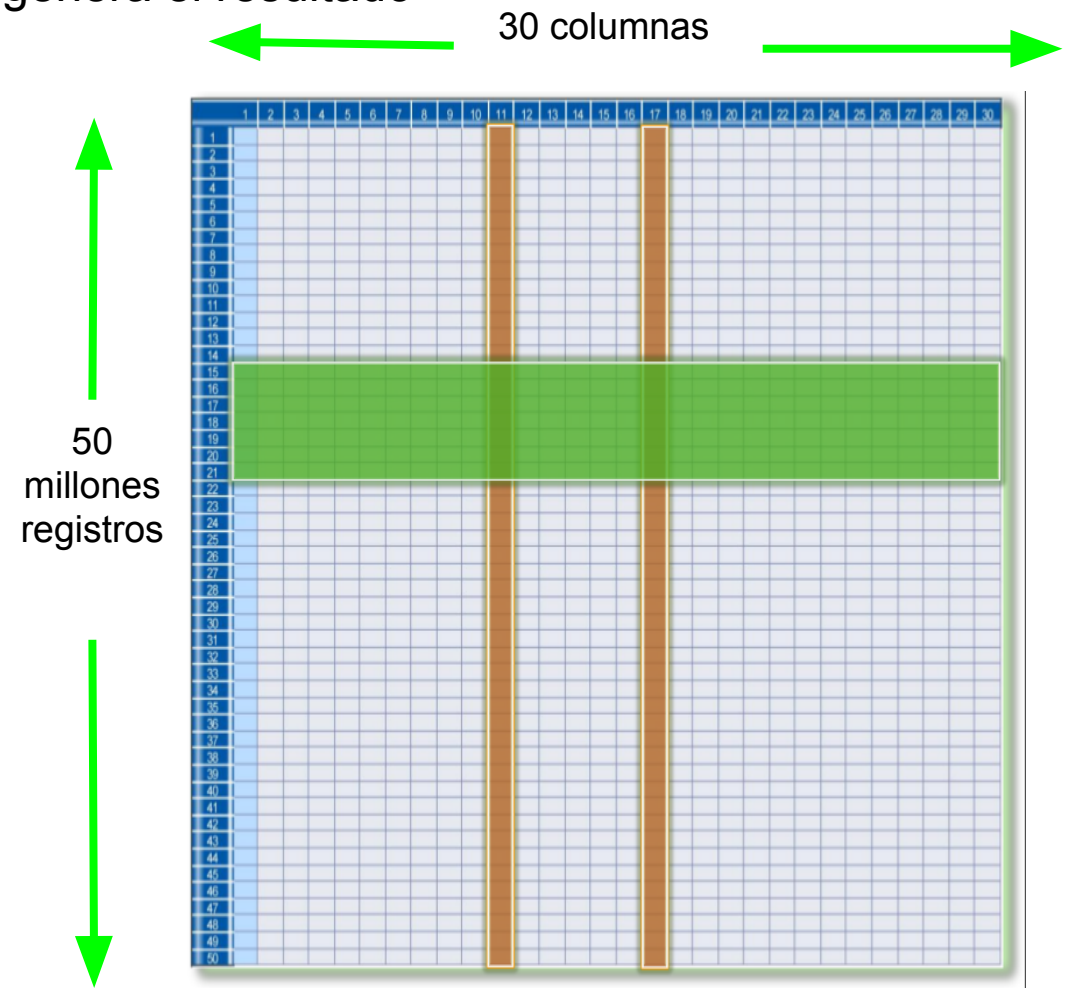
BD Tradicionales vs BD Columnares

La base de datos transaccional genera el resultado

Resultado

BD OLTP - Transaccional

- Elimina 43 días
- Recupera 7 millones de registros
- **Regresa 210 millones de datos**



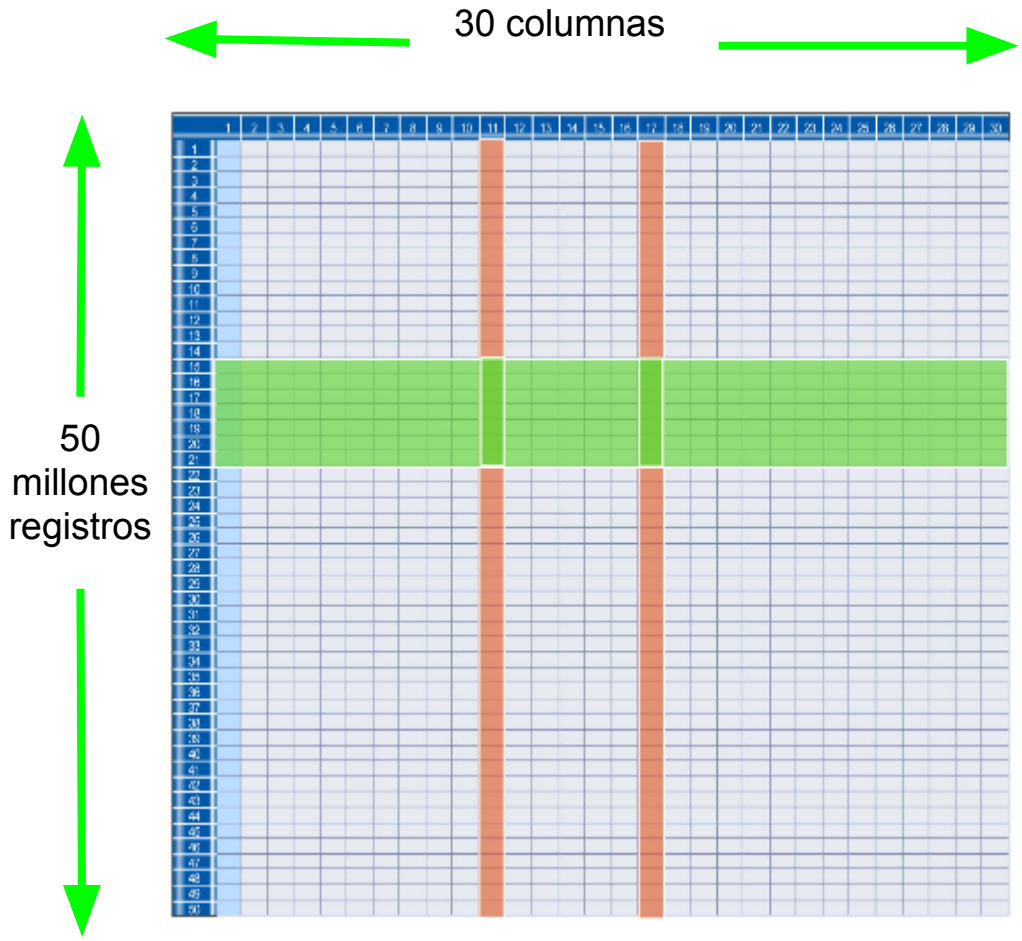
BD Tradicionales vs BD Columnares

La base de datos Columnar genera el resultado con 93% menos de datos

Resultado

BD Columnar

- Elimina 43 días
- Elimina 28 de las 30 columnas
- **14 millones de datos recuperados: 93% menos de datos**

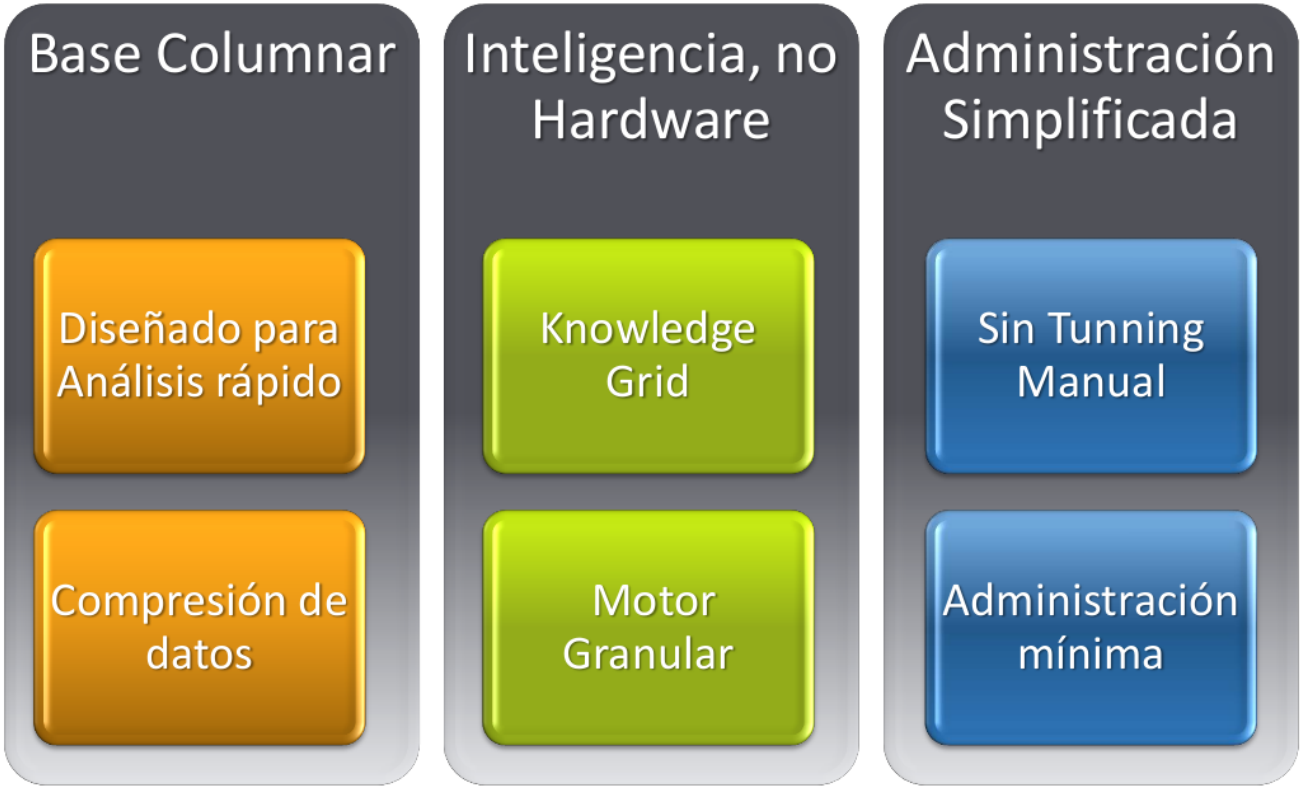


Agenda

- Antecedentes Base de Datos Analíticas
  El internet de las Cosas
- BD Tradicionales vs BD Columnares
- Arquitectura
- Mejores Prácticas y Optimización
- Requerimientos de Software y Hardware
- Infopliance
- Demo

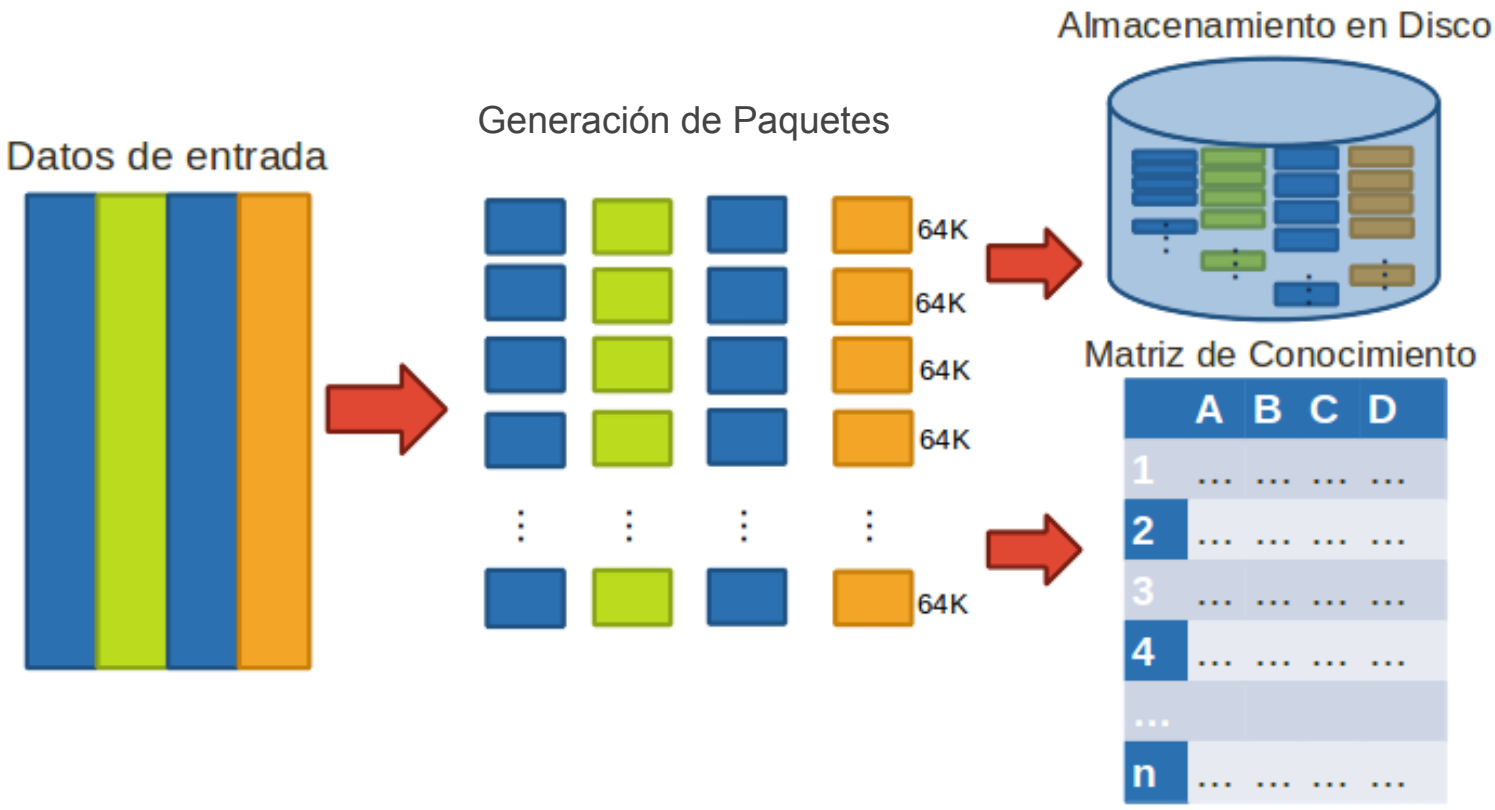
Distintivo de Infobright

Inteligencia, no Hardware, los principios de Infobright ...



Paquetes de Datos y Compresión

La inteligencia de infobright inicia en la carga y organización de los datos ...

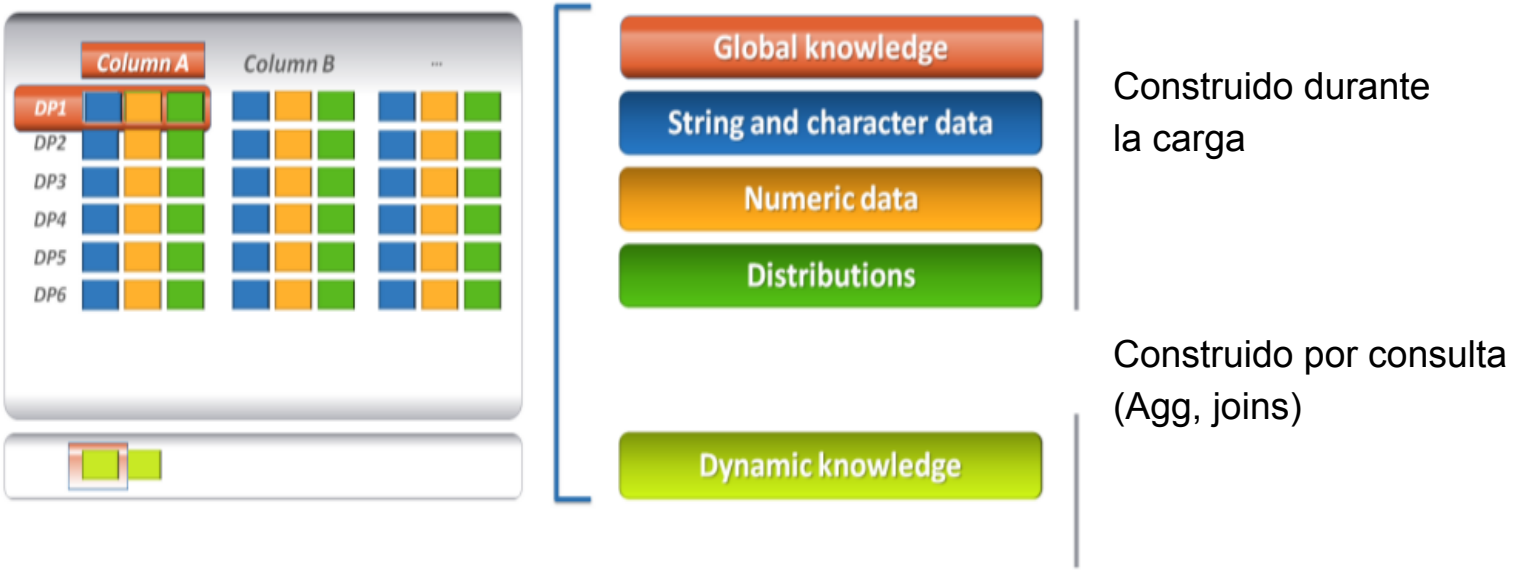


Matriz y Nodos de Conocimiento

La inteligencia de infobright continúa con la generación del conocimiento ...

Matriz de Conocimiento
Knowledge Grid
Información acerca de los datos

Nodos de Conocimiento
Knowledge Nodes
Construido para cada paquete (64 kB)



Esta capa de metadata = 1% del volumen total después de la compresión

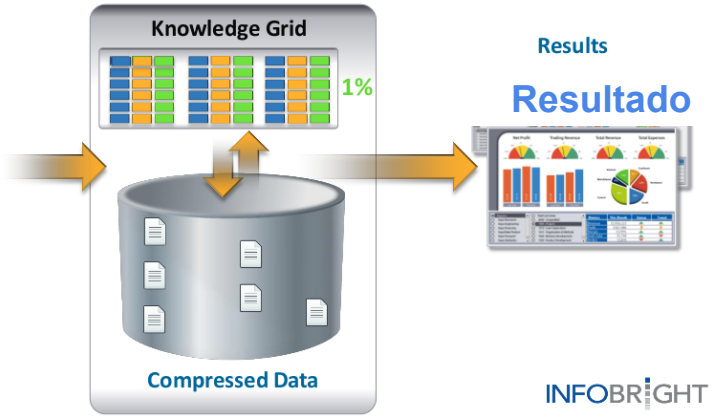
Motor Granular

Para responder una consulta, Infobright utiliza su motor granular alimentado de la matriz de conocimiento:

1. Se recibe una consulta
2. El motor pregunta iterativamente al Knowledge Grid
3. Con cada iteración se eliminan Data Packs
4. Solo los Data Packs necesarios son descomprimidos

Consulta

¿Cómo están mis ventas este año?



Motor Granular - Ejemplo

Solución de consulta usando la matriz de Conocimiento

```
SELECT count(*)
FROM employees
WHERE salary > 50,000
AND age < 35
AND job = ‘Accounting’
AND city =’TORONTO’;
```

Cargadores

Infobright Loader

- Cargador rápido.
- Manejo simple de errores
- Estricto format de datos.
- Soporta archivos delimitados y binarios

MySQL Loader

- Más lento que Infobright loader.
- Mejor manejo de errores.
- Amplio soporte a archivos de texto, inclusive de tamaño delimitado.

INSERT

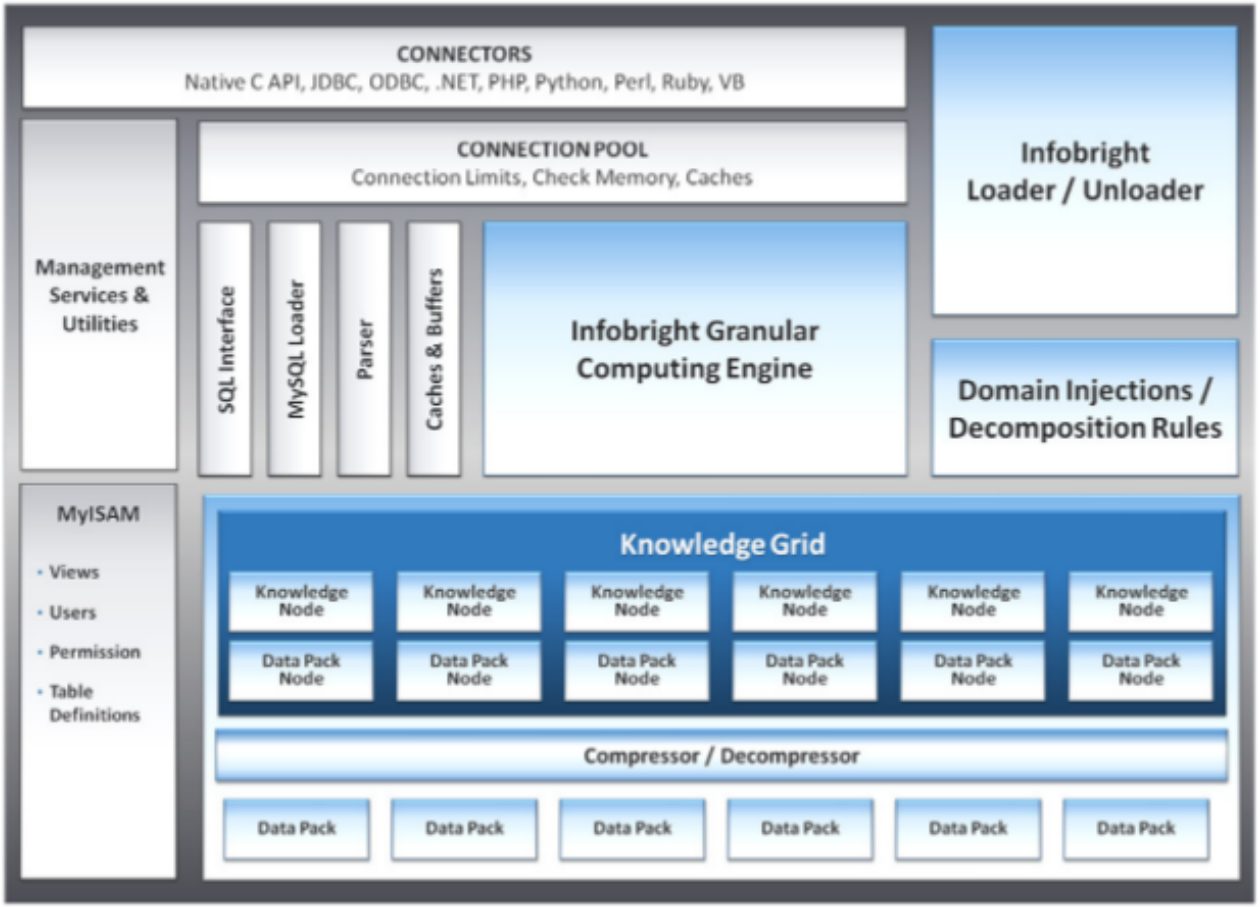
- Soportado por todas las herramientas de ETL.
- Puede ser muy lento dependiendo del bloque de datos a insertar.

Distributed Load Processor (DLP)

- Add-ON a Infobright Enterprise
- Procesos de carga remota
- Permite escalabilidad al correr en múltiples DLP concurrentemente
- Conectividad on clusters de Hadoop

Infobright + MySQL

Infobright esta construido dentro de la arquitectura de MySQL



Agenda

- Antecedentes Base de Datos Analíticas
  El internet de las Cosas
- BD Tradicionales vs BD Columnares
- Arquitectura
- Mejores Prácticas y Optimización
- Requerimientos de Software y Hardware
- Infopliance
- Demo

Optimizador de Consultas

Recomendaciones aprovechar al máximo el potencial de la base de datos

Tipos de Datos

Enteros, mejor rendimiento para:

- Uniones (joins)
- Llaves sustitutas (surrogate keys)
- Opción 'lookup' para búsquedas

Caracteres, mejores prácticas:

- Sub-selects con llaves sustitutas
- Columnas Checksum para cadenas largas

```
Create Table Customer(
Customer_Key varchar(10),
Customer_Name varchar(50),
Customer_Address  varchar(300),
Category varchar(10));
```

```
Create Table Customer(
Customer_Key integer,
Customer_Name varchar(50),
Customer_Address  varchar(300),
Category varchar(10) comment 'lookup',
Customer_Name_MD5 bigint,
Customer_Address_MD5 bigint);
```

Construcción de Consultas

Para que la respuesta en las consultas en Infobright incrementen su rendimiento, solo hay que cambiar la lógica de los cruces en tablas...

SQL Original

```
select sum(dlr_trans_amt), a.msa_id
from fact_sales a, dim_dates b, dim_msa c
where a.trans_date=b.trans_date and a.msa_id=c.msa_id
 and b.trans_year=2006 and b.trans_month='MARCH'
and c.msa_name in
('BIRMINGHAMHOOVER', 'NAPLESMARCO ISLAND', 'CHAMPAIGNURBANA')
group by a.msa_id;

3 rows in set (3 min 11.65 sec)
```

Infobright – Alto rendimiento

```
select sum(dlr_trans_amt), msa_id
from fact_sales a
where
trans_date in (select trans_date from dim_dates b where b.trans_year=2006 and
b.trans_month='MARCH')
and
msa_id in (select msa_id from dim_msa where msa_name in
('BIRMINGHAMHOOVER', 'NAPLESMARCO ISLAND', 'CHAMPAIGNURBANA')
group by msa_id;

3 rows in set (21.28 sec)
```

Agenda

- Antecedentes Base de Datos Analíticas
  El internet de las Cosas
- BD Tradicionales vs BD Columnares
- Arquitectura
- Mejores Prácticas y Optimización
- Requerimientos de Software y Hardware
- Infopliance
- Demo

Plataformas y Recursos

Plataformas Soportadas - 64 bits

- Windows Server 2003, 2008
- Solaris 10
- Red Hat Enterprise Linux 5.4, 6.2
- Debian 6
- CentOS 5.4, 6.2
- Novell SUSE Linux Enterprise 10
- Novell SUSE Linux Enterprise 11

Recursos Necesarios

Procesador:

- 2.0 GHz o mayor
- Dual o quad core

CPU y Memoria

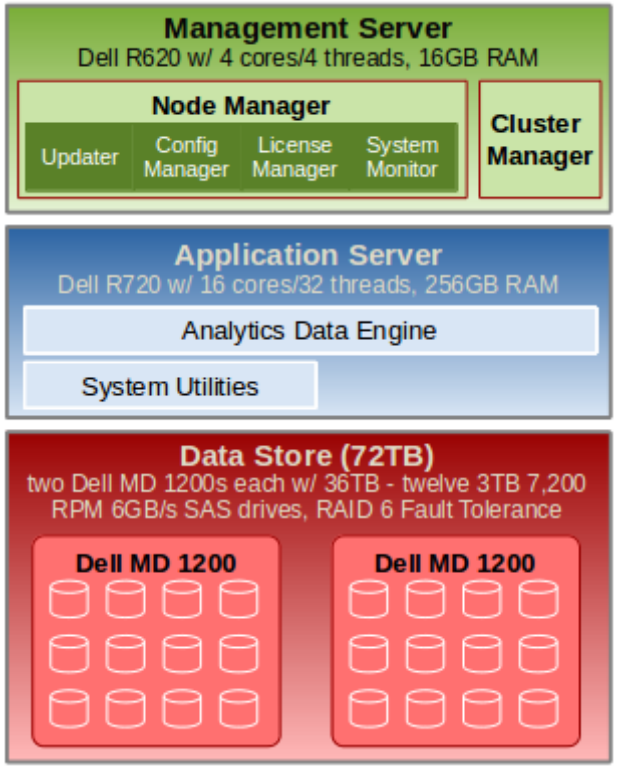
- Depende de los requerimientos de rendimiento
- Recomendación:
- 1 core y 2GB of memoria para consultas simples
- 2 cores y 4 GB of memoria para consultas complejas
- El cargador es es extra

Agenda

- Antecedentes Base de Datos Analíticas
  El internet de las Cosas
- BD Tradicionales vs BD Columnares
- Arquitectura
- Mejores Prácticas y Optimización
- Requerimientos de Software y Hardware
- Infopliance
- Demo

Infopliance

Appliance de propósito específico con hardware y software integrado para volúmenes mayores a 10 TB



Hardware

- Commodity Hardware
- Plug & Play

Software

- Infopliance Manager
- Infopliance Analytics Data Enginer
- Infopliance Data Processors

Agenda

- Antecedentes Base de Datos Analíticas
  El internet de las Cosas
- BD Tradicionales vs BD Columnares
- Arquitectura
- Mejores Prácticas y Optimización
- Requerimientos de Software y Hardware
- Infopliance
- Demo

Demo

Demo de una empresa retail con un año de historia y datos pos enriquecidos por la geografía

- Modelo Normal (Estrella)
  - Tablas Ventas con 1.3 billones de registros
  - Tabla Clientes (Consumo y Pymes) con 3.2 millones registros
  - Catálogo de Geografía
  - Creación de tablas: DDL
  - Motor de Infobright
  - Optimización para búsquedas por columnas de texto
  - Compresión de datos
- Modelo Optimizado para Infobright
  - Tabla plana con Ventas + Clientes + Geografía 1.3 billones registros
  - Compresión de datos
- Consultas