Conceptos de Hadoop

Rogelio Ferreira Escutia

Conceptos

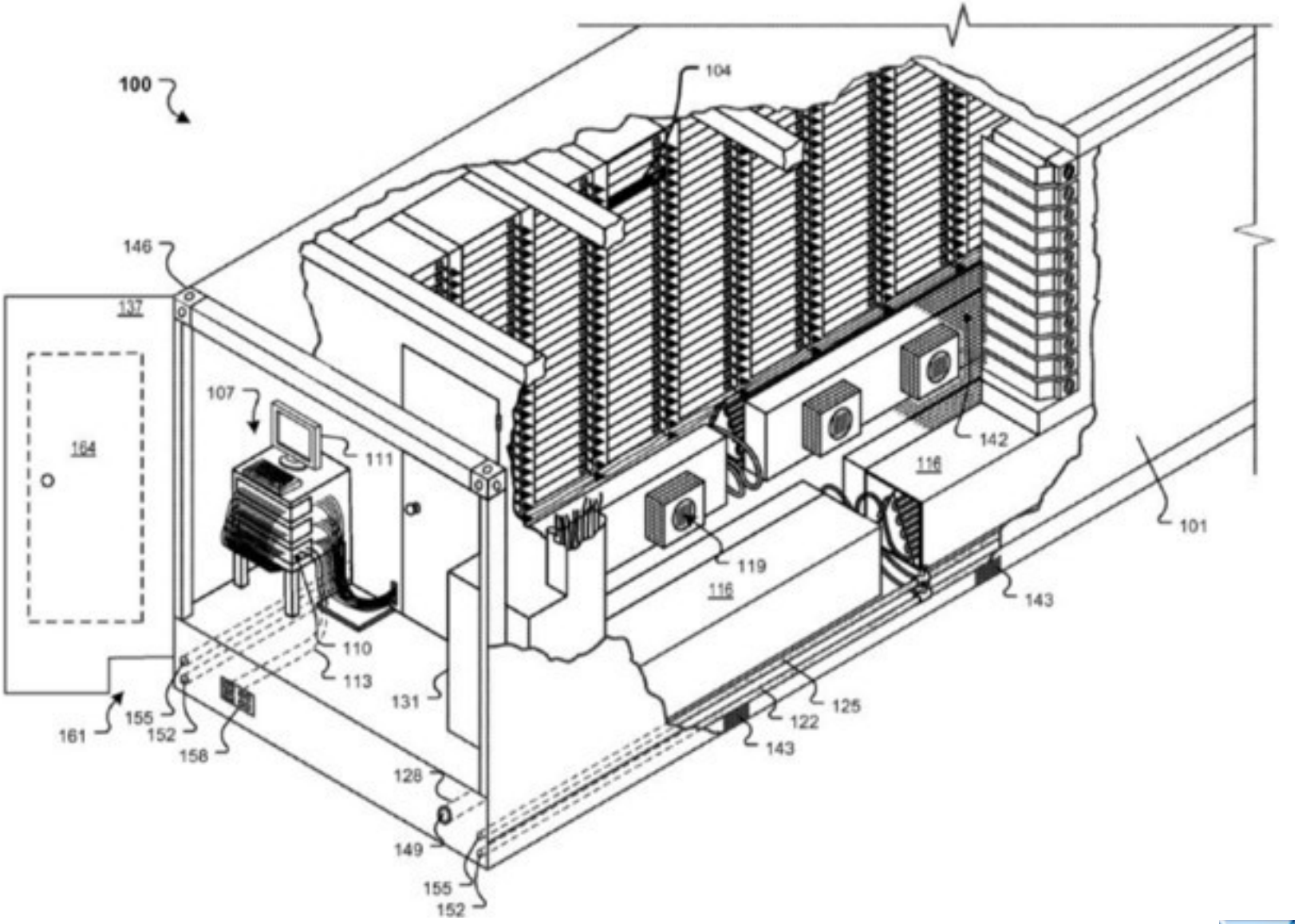
Google Modular Data Center

- **Desde que apareció Google en 1996, la cantidad de datos que debía almacenar para su algoritmo PageRank de búsqueda y procesamiento de datos fué creciendo.**

- **De lo anterior, Google prefirió diseñar y construir su propia tecnología en vez de comprarla y de ahí surge su proyecto "Google Modular Data Center" para empezar a estandarizar y adecuar sus centros de datos.**

- **Cada Módulo incluye su propia fuente de poder, aire acondicionado y soporte para 1000 servidores modulares con Linux.**

Google Modular Data Center

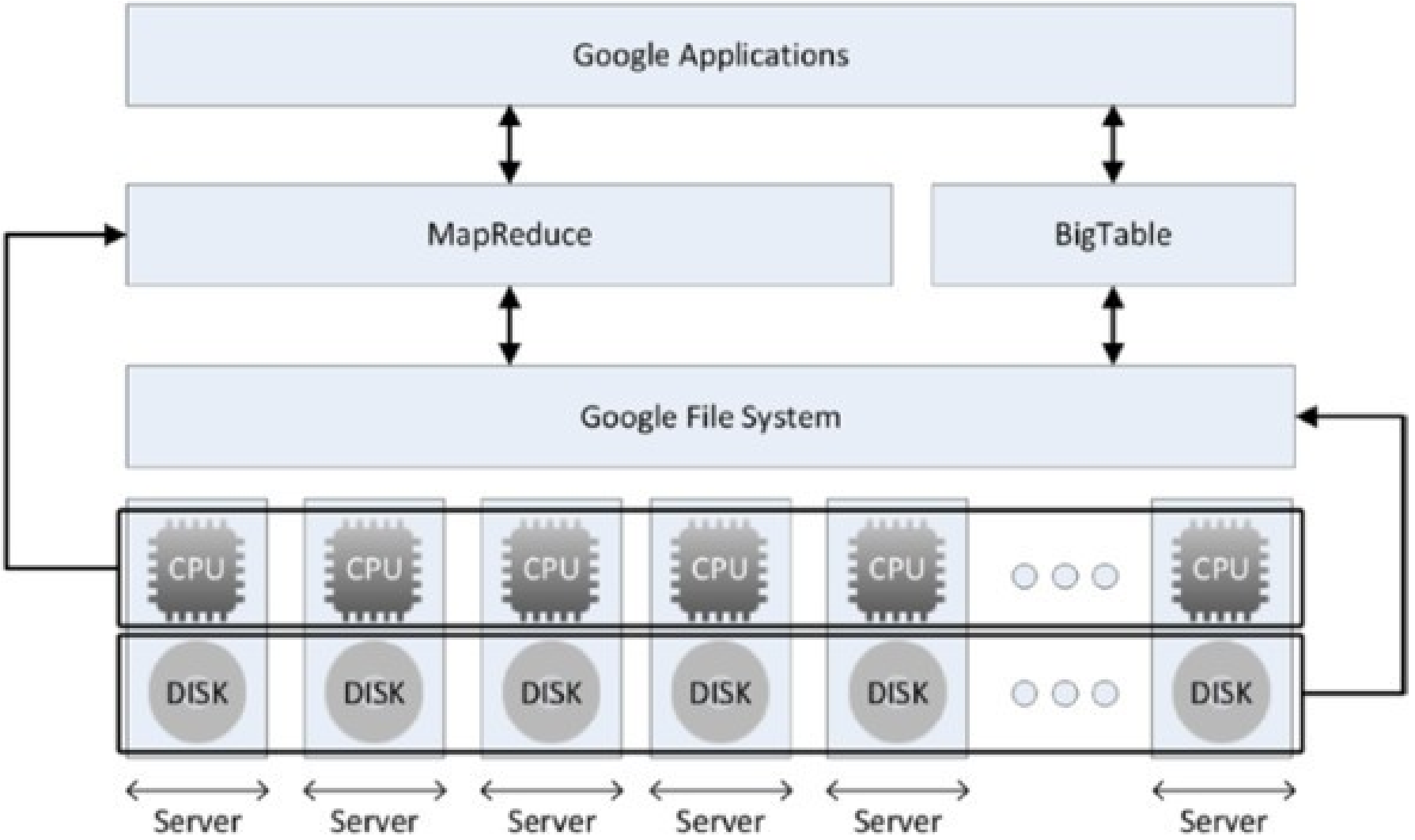


"Next Generation Databases", Guy Harrison, Apress Media 2015

Software de Google

- **Google desarrolló el siguiente software para sus propios centros de datos:**
  - **Google File System(GFS): un sistema de archivos distribuido.**
  - **MapReduce: un framework de procesamiento distribuido para paralelización de algoritmos.**
  - **BigTable: un sistema norelacional de base de datos que utiliza GFS para el almacenamiento.**

Software de Google

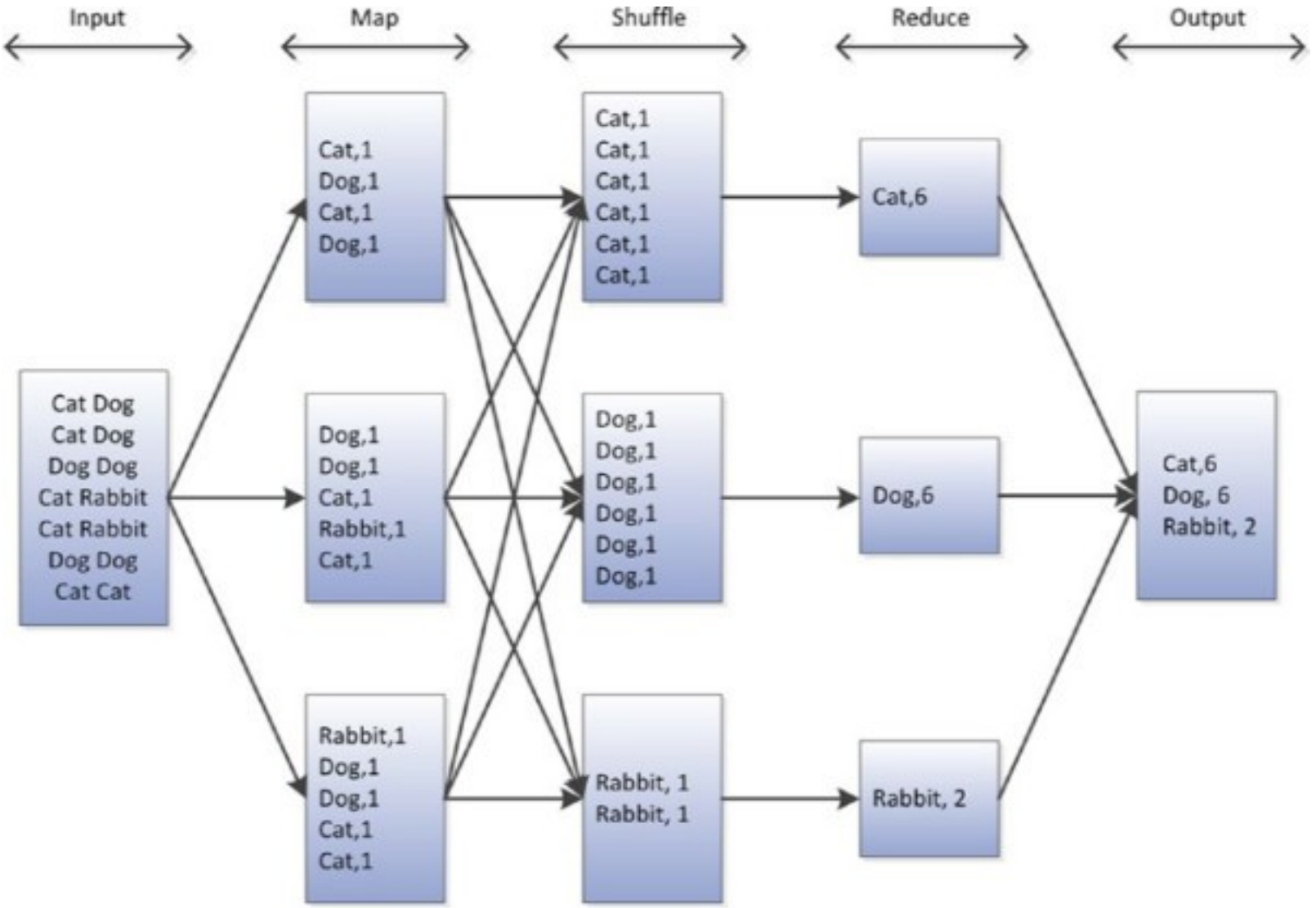


"Next Generation Databases", Guy Harrison, Apress Media 2015

MapReduce

- **MapReduce es un modelo de programación paralela para propósitos generales que se divide en 2 fases:**
  - **Mapping: Se dividen las peticiones para ser procesadas por hilos en diferentes computadoras.**
  - **Reduce: Se combinan las salidas del mapping para obtener una salida final.**

MapReduce

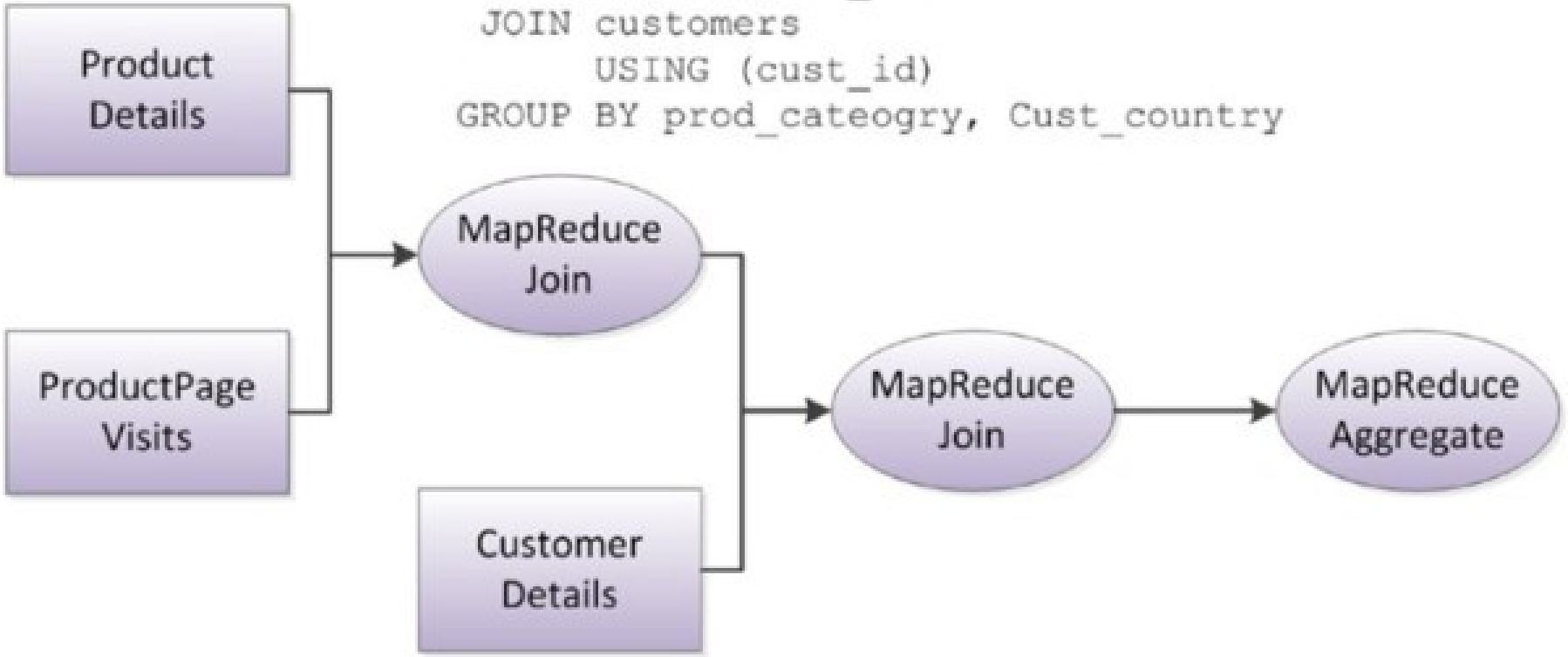


"Next Generation Databases", Guy Harrison, Apress Media 2015

MapReduce

Equivalent of:

```
SELECT prod_category, Cust_country,
       SUM(visits)
  FROM products
  JOIN product_page_visits
       USING (prod_id)
  JOIN customers
       USING (cust_id)
 GROUP BY prod_cateogry, Cust_country
```



Primeras Implementaciones

Yahoo (2008)

- **Yahoo anunció en 2008 un cluster usando Hadoop con 5 petabytes de almacenamiento y mas de 10,000 núcleos.**

- **Este sistema fué utilizado para crear el índice y las búsquedas del motor de Yahoo.**

Facebook (2008)

- **Facebook empezó a trabajar con Hadoop en el 2007, y en 2008 ya tenía en funcionamiento 2,500 CPUs.**

- **En el 2012 el cluster Hadoop de Facebook ya superaba los 100 petabytes en almacenamiento.**

Rogelio Ferreira Escutia

Instituto Tecnológico de Morelia
Departamento de Sistemas y Computación

Correo: ***rogelio@itmorelia.edu.mx***
rogeplus@gmail.com

Página Web: ***http://sagitario.itmorelia.edu.mx/~rogelio/***
http://www.xumarhu.net/

Twitter: ***http://twitter.com/rogeplus***
Facebook: ***http://www.facebook.com/groups/xumarhu.net/***