

Capítulo 4: Hadoop

Capítulo 5: Arquitectura de Hadoop

Capítulo 6: Componentes de Hadoop



División de Alta Tecnología

5

Arquitectura de Hadoop

La información es la gasolina del siglo XXI, y la analítica de datos el motor de combustión.
Peter Sondergaard.

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.




Objetivos

Al finalizar el capítulo, el alumno logrará:

- Desplegar la arquitectura distribuida de Hadoop.
- Implementar un Clúster Hadoop.
- Manejar el sistema de archivos HDFS.

5 - 2

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Agenda

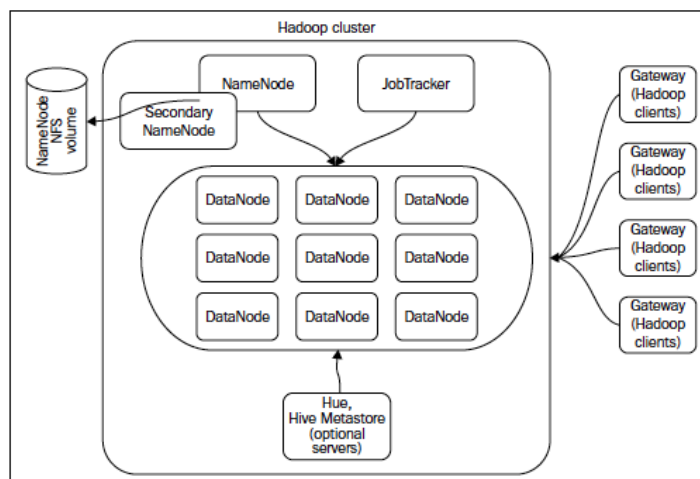
- Componentes de Hadoop
- Implementación de un Cluster Hadoop
- HDFS

5 - 3

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop

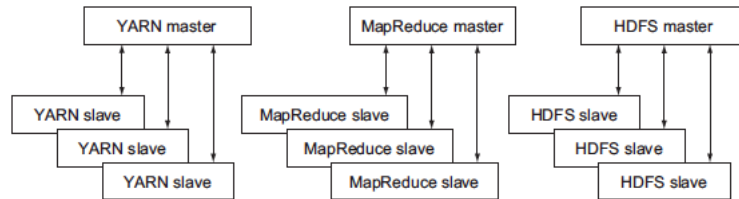


5 - 4

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop



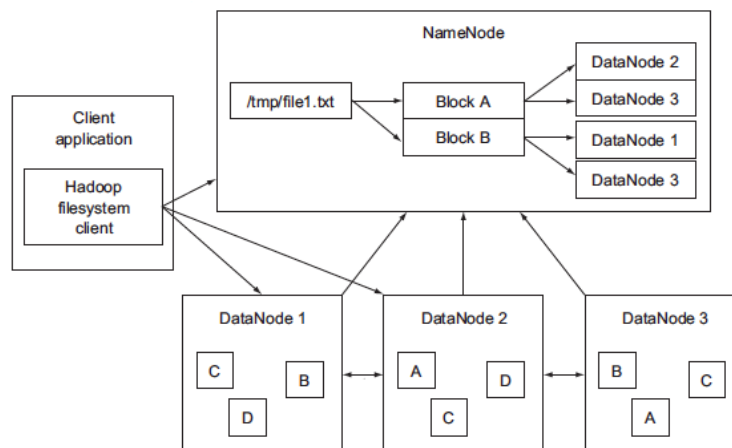
- HDFS o Hadoop Distributed File System, es un sistemas de almacenamiento distribuido para almacenamiento de información estructurada o no estructurada.
- MapReduce es un motor de procesamiento batch, que en la versión 2.0 de Hadoop, se implementa con YARN.
- YARN ó Yet Another Resource Negotiator, incluido en la versión 2.0 de Hadoop, es un administrador de recursos mediante la calendarización de actividades.

5 - 5

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop



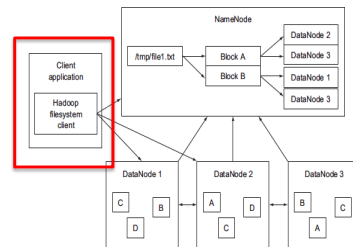
5 - 6

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Aplicación Cliente

- Las aplicaciones clientes acceden al catálogo del NameNode, para identificar los elementos que son parte de la consulta.
- Una vez identificado los datos propiamente tal, en la ubicación del DataNode, el cliente accede a los datos en forma directa.



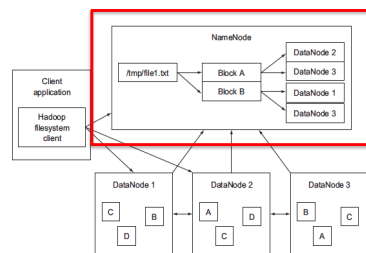
5 - 7

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Name Node

- Es el componente que contiene un catálogo de todos los DataNode que son parte de la infraestructura.
- En ésta capa se almacena la metadata que permite la ubicación de los datos que se ubican en los DataNodes.



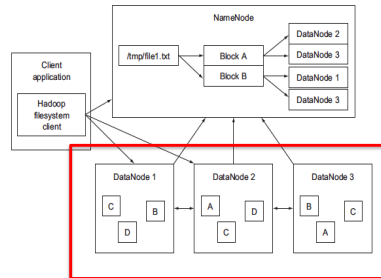
5 - 8

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Data Node

- Son los fragmentos de información que se almacenan en la infraestructura.
- Se genera redundancia de información mediante los servicios de sincronización interna.

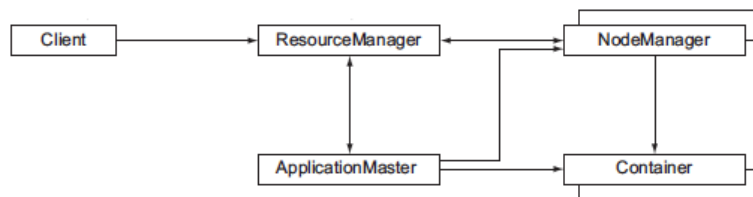


5 - 9

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Map Reduce – Flujo de Trabajo



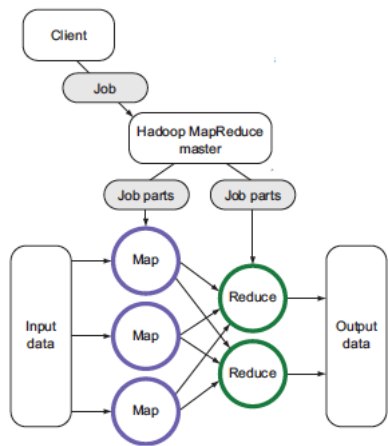
MapReduce es un marco de software para escribir fácilmente aplicaciones que procesan grandes cantidades de datos (conjuntos de datos de varios terabytes) en paralelo en clústeres grandes (miles de nodos) de hardware básico de una manera confiable y tolerante a fallas.

5 - 10

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Modelo Map Reduce

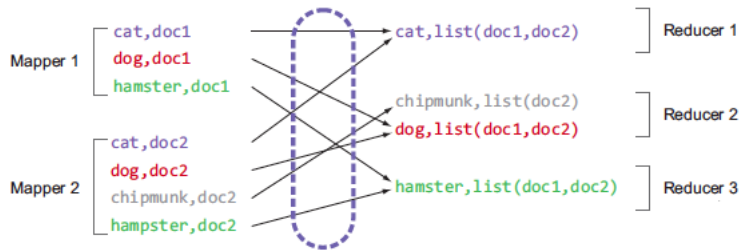


5 - 11

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Modelo Map Reduce

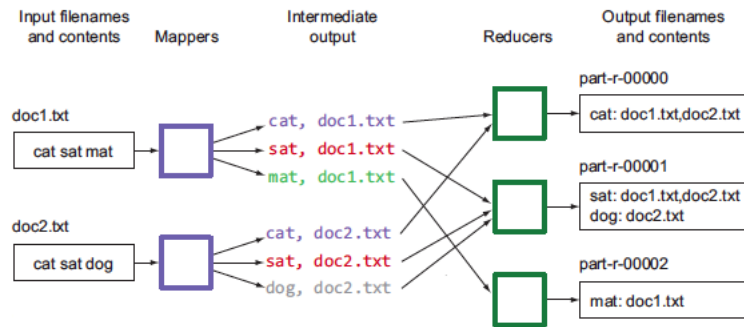


5 - 12

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Modelo Map Reduce

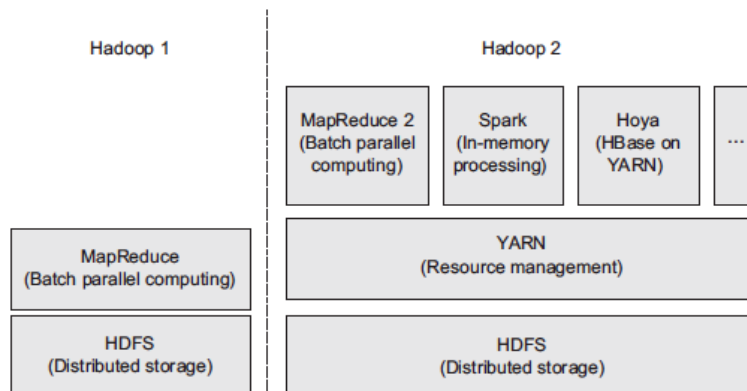


5 - 13

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Yarn en Hadoop 2

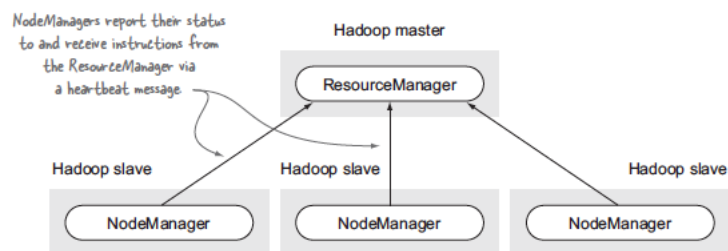


5 - 14

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Yarn – Resource Manager y Node Manager

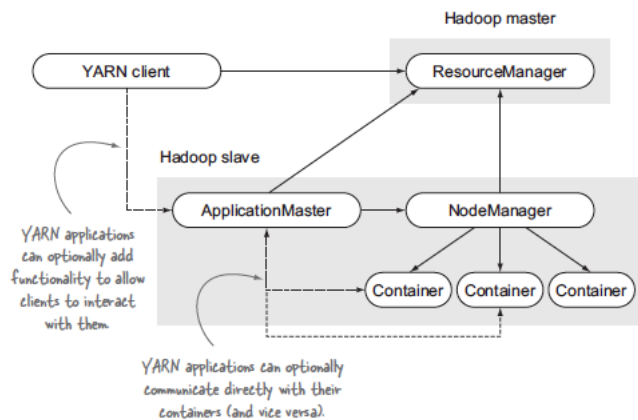


5 - 15

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Yarn – Aplicaciones Cliente

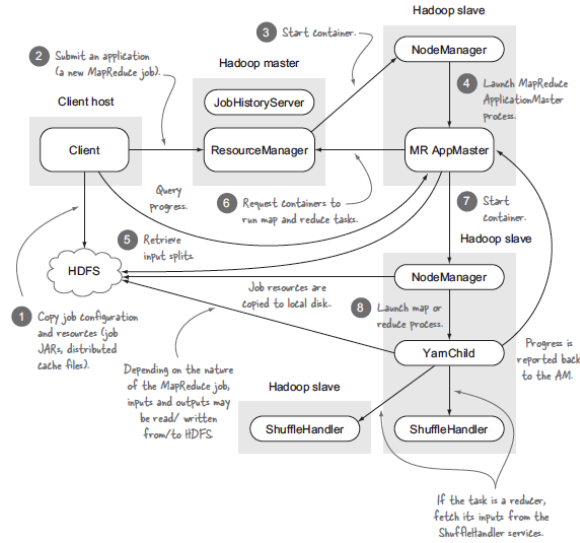


5 - 16

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Componentes de Hadoop Yarn – Flujo de Trabajo

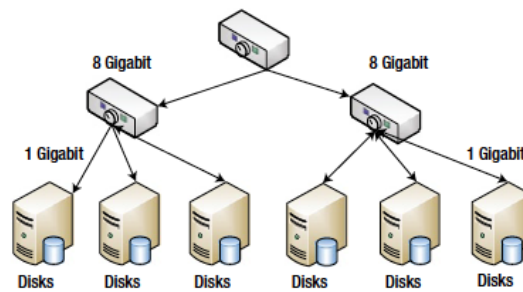


5 - 17

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Implementación de un cluster Hadoop



- Infraestructura de comunicaciones de alta velocidad.
- Granjas de servidores distribuidos.
- Tolerancia a fallos.

5 - 18

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



HDFS

- HDFS, es un sistema de almacenamiento tolerante a fallos que puede almacenar gran cantidad de datos, escalar de forma incremental y sobrevivir a fallos de hardware sin perder datos.
- HDFS gestionar el almacenamiento en el cluster, dividiendo los ficheros en bloques y almacenando copias duplicadas a través de los nodos.
- Por defecto se replican en 3 nodos distintos.

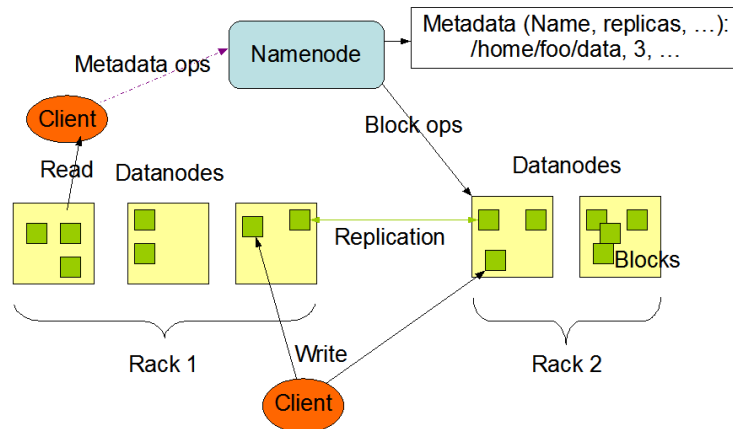
5 - 19

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



HDFS

HDFS Architecture



5 - 20

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



HDFS

¿Como trabaja?

- Nombres del sistema de archivos
- Replicación de datos
- La persistencia de los metadatos del sistema de archivos
- Los protocolos de comunicación
- Robustez
- Organización de datos
- Accesibilidad
- Recuperación de archivos

5 - 21

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Tarea Nº 5

Al finalizar la tarea, el alumno logrará:

- Aprender las características de YARN y MAP REDUCE.



5 - 22

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Ejercicio N° 5.1: Trabajando con HDFS

Al finalizar la tarea, el alumno logrará:

- Aprender a trabajar con los diversos comandos HDFS

5 - 23

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Ejercicio N° 5.2: Ejecutar procesos en Map Reduce

Al finalizar la tarea, el alumno logrará:

- Aprender como trabajan los procesos de Map Reduce

5 - 24

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Lecturas adicionales

Se sugiere revisar los siguientes enlaces para profundizar en los conceptos tratados en el presente capítulo:

- [Cluster Hadoop](#)

5 - 25

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Resumen

En este capítulo, hemos aprendido como es la arquitectura distribuida de Hadoop y como implementar un Cluster Hadoop.

Además, hemos hecho una breve revisión de como administrar el sistema de archivos HDFS.

5 - 26

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.

