


Capítulo 4: Hadoop

Capítulo 5: Arquitectura de Hadoop

Capítulo 6: Componentes de Hadoop




DAT
División de Alta Tecnología

4

Hadoop

Big Data

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.




Objetivos

Al finalizar el capítulo, el alumno logrará:

- Diseñar una arquitectura Hadoop.

4 - 2

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Agenda

- Introducción al Hadoop
- Ecosistema de Hadoop
- Big Data y el Cloud

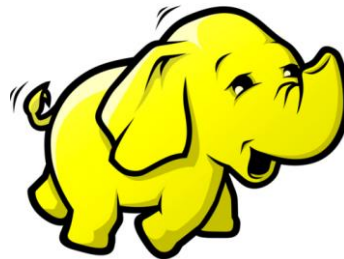
4 - 3

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Introducción a Hadoop

- Hadoop = Big Data
- Entorno Distribuido (Datos y Procesos)
- Es escalable de forma horizontal con commodity hardware.

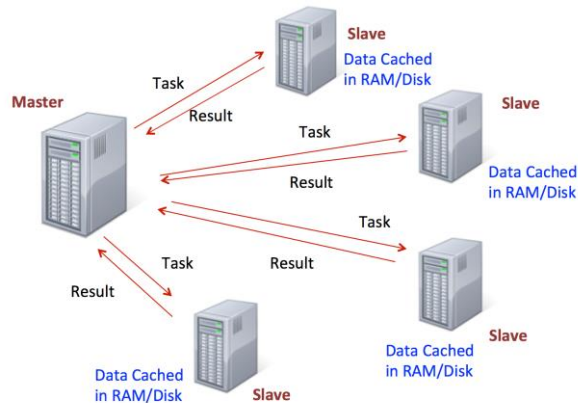


4 - 4

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Introducción a Hadoop



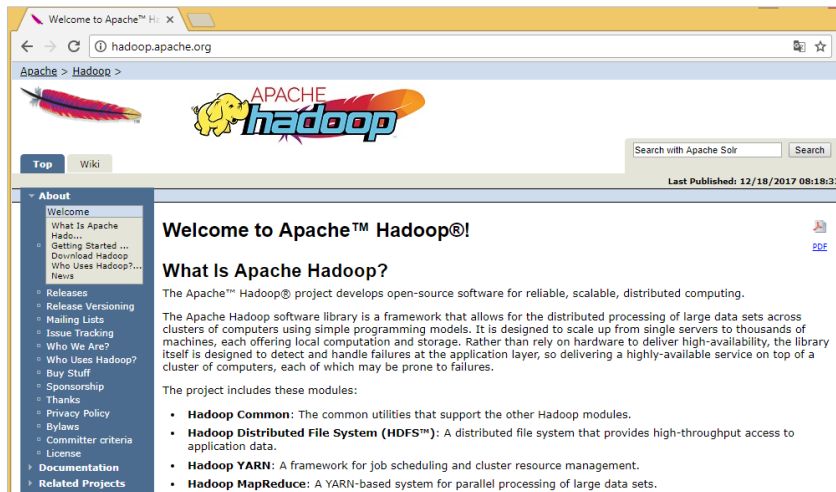
- Trabaja con procesamiento en paralelo a través de nodos de datos en un sistema de ficheros distribuidos.

4 - 5

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Introducción a Hadoop

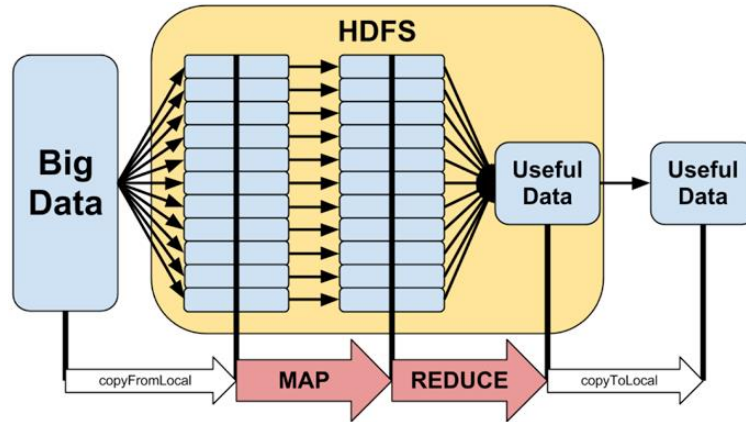


4 - 6

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Introducción a Hadoop



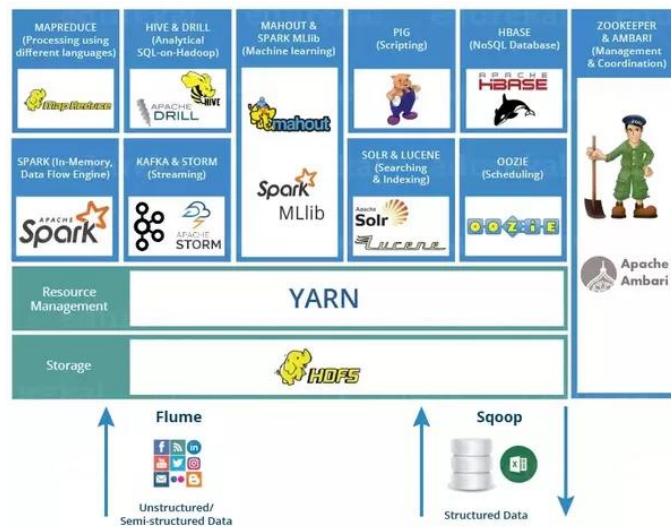
Datos (HDFS) y Procesamiento (MAP / REDUCE)

4 - 7

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Ecosistema de Hadoop



4 - 8

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Ecosistema de Hadoop

- Ambari: una herramienta basada en web para aprovisionar, administrar y monitorear clústeres de Apache Hadoop.
- HBase : Una base de datos orientada a valores/claves que se ejecuta sobre HDFS
- Hive : sistema de funciones que soportan agregación de datos y consultas ad hoc sobre MapReduce
- Pig: Lenguaje de alto nivel para gestionar flujos de datos y ejecución de aplicaciones sobre Hadoop
- Mahout: entorno de machine learning implementado en hadoop
- Zookeeper : servicio centralizado para mantener información de configuración, gestión de nombre, y para facilitar la sincronización de servicios.
- Sqoop : Herramienta diseñada para transferir datos masivos desde Hadoop a otros entornos como Bases de datos relacionales.

4 - 9

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Hive

El software de almacenamiento de datos Apache Hive TM facilita la lectura, escritura y administración de grandes conjuntos de datos que residen en el almacenamiento distribuido y se consultan mediante la sintaxis SQL.

- Herramientas para permitir el acceso fácil a los datos a través de SQL, lo que permite tareas de almacenamiento de datos como extraer / transformar / cargar (ETL), informes y análisis de datos.
- Un mecanismo para imponer estructura en una variedad de formatos de datos
- Acceso a archivos almacenados directamente en Apache HDFS TM o en otros sistemas de almacenamiento de datos como Apache HBase TM

<http://hive.apache.org/>



4 - 10

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Pig

Apache Pig es una plataforma para analizar grandes conjuntos de datos que consiste en un lenguaje de alto nivel para expresar programas de análisis de datos, junto con infraestructura para evaluar estos programas. La propiedad principal de los programas de Pig es que su estructura es susceptible de una paralelización sustancial, lo que a su vez les permite manejar conjuntos de datos muy grandes.

- Genera comando MapReduce

<http://pig.apache.org/>



4 - 11

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache sqoop

Apache Sqoop es una herramienta diseñada para transferir datos de manera eficiente entre fuentes de datos estructurados, semiestructurados y no estructurados. Las bases de datos relacionales son ejemplos de fuentes de datos estructurados con un esquema bien definido para los datos que almacenan. Cassandra, Hbase son ejemplos de fuentes de datos semiestructuradas y HDFS es un ejemplo de fuente de datos no estructurados que Sqoop puede admitir.

<http://sqoop.apache.org/>



4 - 12

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Flume

Flume es un servicio distribuido, confiable y disponible para recopilar, agregar y mover grandes cantidades de datos de registro de manera eficiente. Tiene una arquitectura simple y flexible basada en flujos de datos de transmisión. Es robusto y tolerante a fallas con mecanismos de confiabilidad ajustables y muchos mecanismos de conmutación por error y recuperación. Utiliza un modelo de datos extensible simple que permite la aplicación analítica en línea.

Útil para cargar y mover en Hadoop información de tipo textos como ficheros de logs, paquetes de twitter, etc.

<http://flume.apache.org/>



4 - 13

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Zookeeper

ZooKeeper es un servicio centralizado para mantener la información de configuración, nombrar, proporcionar sincronización distribuida y proporcionar servicios grupales. Todos estos tipos de servicios son utilizados de una forma u otra por aplicaciones distribuidas.

Elimina la complejidad de la gestión distribuido de la plataforma

<http://zookeeper.apache.org/>



4 - 14

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Spark

Apache Spark es un sistema de computación en clúster rápido y de uso general. Proporciona API de alto nivel en Java, Scala, Python y R, y un motor optimizado que admite gráficos de ejecución general. También es compatible con un amplio conjunto de herramientas de alto nivel que incluyen Spark SQL para SQL y procesamiento de datos estructurados, MLlib para Machine Learning, GraphX para procesamiento de gráficos y Spark Streaming.

Velocidad: Ejecute programas hasta 100 veces más rápido que Hadoop MapReduce en la memoria, o 10 veces más rápido en el disco.

<http://spark.apache.org/>



4 - 15

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Big Data y el Cloud

¿Por qué ir al Cloud?



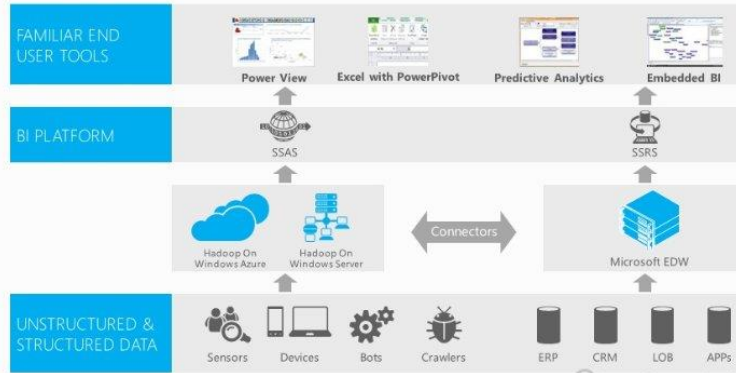
4 - 16

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Big Data y el Cloud

Microsoft Big Data Solution



<https://www.microsoft.com/es-es/sql-server/big-data>

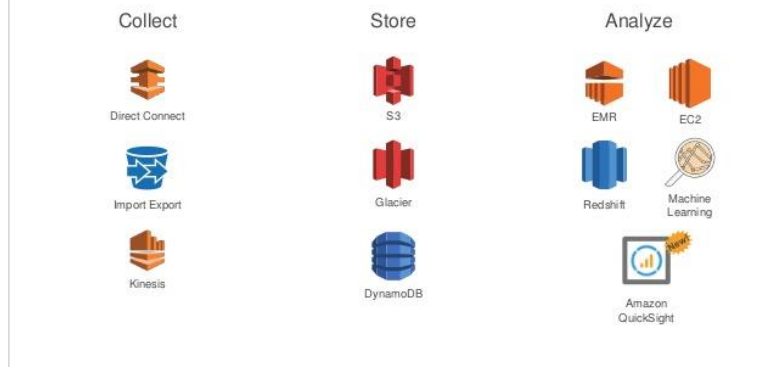
4 - 17

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Big Data y el Cloud

AWS Big Data Platform



<https://aws.amazon.com/es/big-data/>

4 - 18

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Laboratorio N° 4.1: Revisión de una arquitectura Hadoop

Al finalizar la tarea, el alumno logrará:

- Aprender como trabaja un cluster de Hadoop

4 - 19

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Tarea N° 4: Curso Hadoop

Al finalizar la tarea, el alumno logrará:

- Aprenda los conceptos básicos de Apache Hadoop, un marco de programación libre, de código abierto basado en Java.

4 - 20

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:

¿Qué significa Hadoop en el mundo del Big Data?

4 - 21

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Resumen

En este capítulo, hemos aprendido la arquitectura Hadoop, el ecosistema Hadoop y las diversas alternativas de Big Data en el Cloud.

4 - 22

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.

