

ECOSISTEMA HADOOP

Elementos para ciencia de datos:

1

HDFS: El Hadoop Distributed File System (HDFS) es un sistema de archivos distribuido que proporciona la capacidad de almacenar grandes volúmenes de datos en múltiples nodos en un cluster.

2

MapReduce: Es el motor de procesamiento de Hadoop que permite a los usuarios procesar y analizar grandes cantidades de datos de forma distribuida.

3

Hive: Es un motor de consulta de datos en Hadoop que proporciona una interfaz de consulta SQL para acceder a los datos almacenados en HDFS.

4

YARN: Es un componente de gestión de recursos en el ecosistema de Hadoop que permite ejecutar aplicaciones y tareas de procesamiento de datos de manera eficiente y equilibrada en el cluster.

5

Pig: es un lenguaje de programación de alto nivel que permite a los usuarios procesar grandes cantidades de datos usando una sintaxis sencilla y natural.

6

HBase: es una base de datos NoSQL distribuida que se ejecuta en Hadoop y proporciona acceso en tiempo real a grandes cantidades de datos estructurados y no estructurados.

7

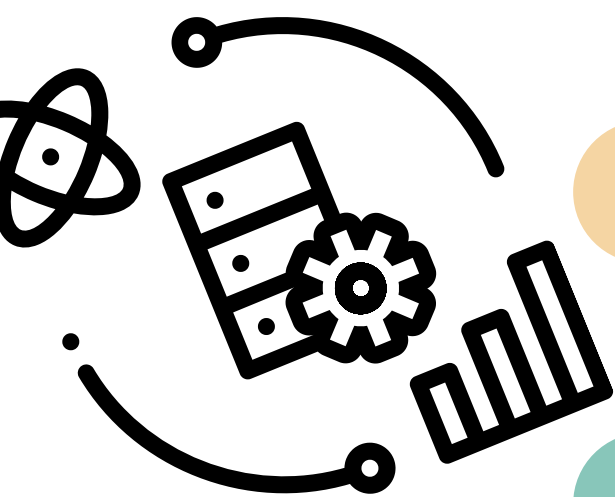
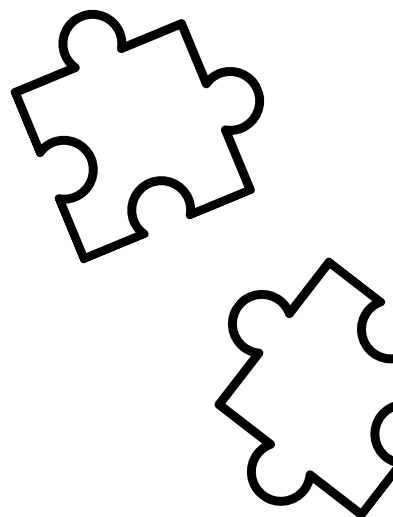
Flume: es una herramienta de ingesta de datos en tiempo real que permite a los usuarios capturar, agregar y transferir grandes cantidades de datos en tiempo real desde diferentes fuentes a HDFS.

8

Sqoop: es una herramienta de transferencia de datos que permite a los usuarios importar y exportar grandes cantidades de datos entre Hadoop y otras fuentes de datos externas, como bases de datos relacionales.

9

Kafka: es un sistema de cola de mensajes distribuido que permite a los usuarios procesar grandes cantidades de datos en tiempo real y realizar la ingesta de datos en Hadoop.



10

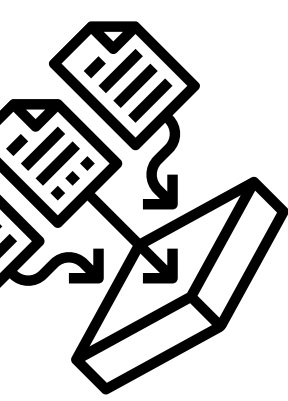
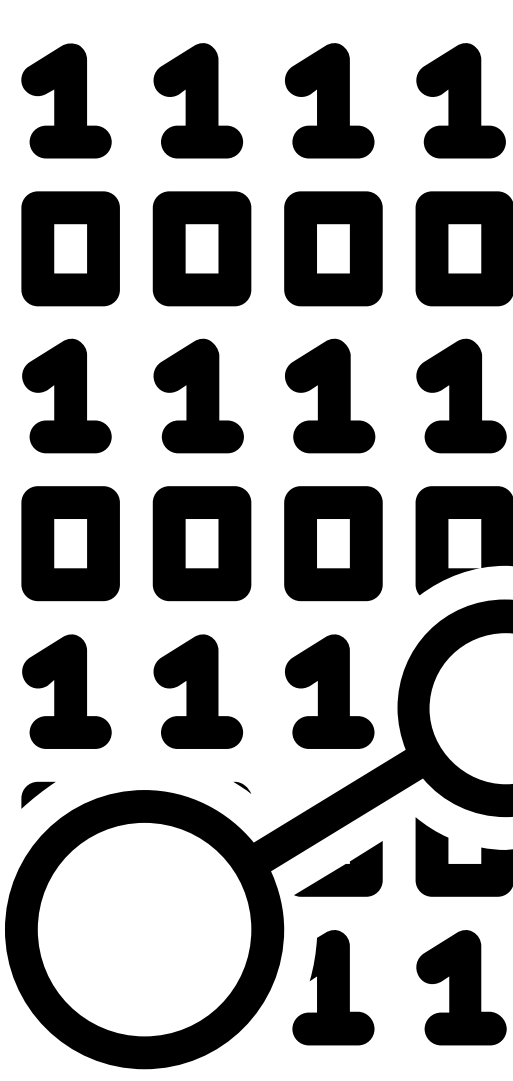
Mahout: es una biblioteca de aprendizaje automático de código abierto diseñada específicamente para ejecutarse en Hadoop y proporciona algoritmos de aprendizaje automático escalables para grandes cantidades de datos.

11

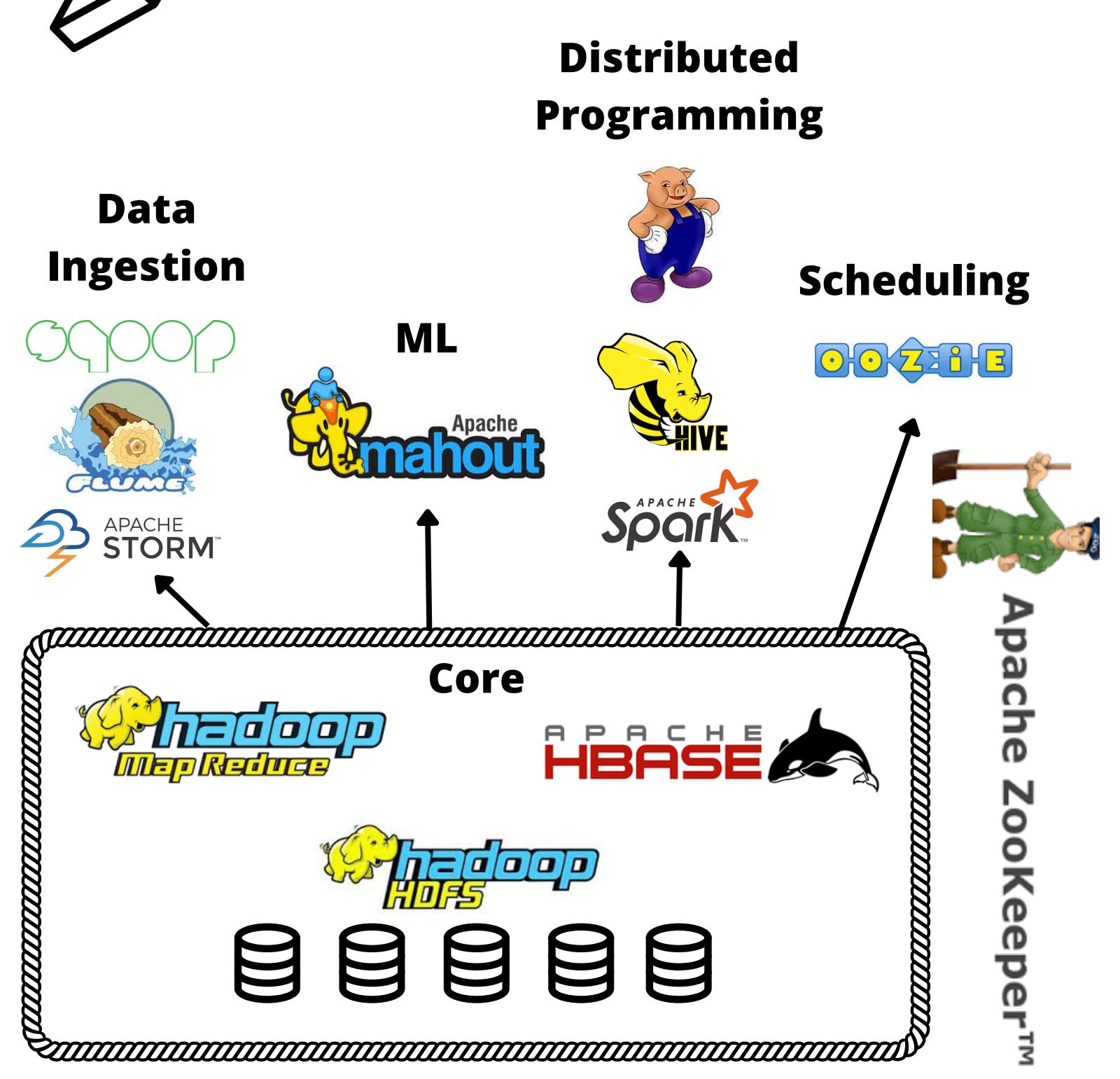
Oozie: es una plataforma de planificación de trabajos distribuidos diseñada específicamente para ejecutar trabajos de Hadoop de manera eficiente y escalable.

12

ZooKeeper: es un servicio de gestión de configuración distribuido que se utiliza en conjunto con Hadoop para proporcionar una gestión de configuración consistente y fiable en un entorno distribuido.



Interacciones entre componentes:



Integrantes

- Juan Carlos Alvarado Carricarte, A01793486
- Bryan Rodolfo Alvarado Cruz, A01793670
- Eduardo Gabriel Arévalo Aguilar, A01793897

Referencias

- Achari, S. (2015). *Hadoop Essentials: Delve Into the Key Concepts of Hadoop and Get a Thorough Understanding of the Hadoop Ecosystem*. Packt Publishing

