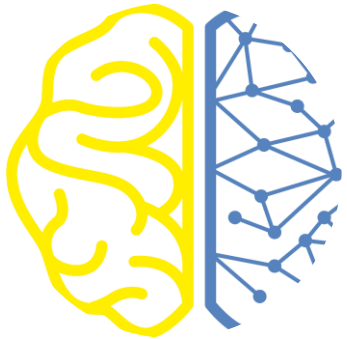
Big Data con Hadoop



y otras tecnologias para procesar grandes datos – Paso a Paso

Por: Eric Alexander

https://datadosis.com

Que es Hadoop

Significado

una **plataforma de software** de código abierto para el **almacenamiento distribuido** y el **procesamiento distribuido** de **conjuntos de datos muy grandes en clusters** de computadoras

construidos a partir de hardware de productos básicos - Hortonworks

DataDosis

Porque Hadoop?

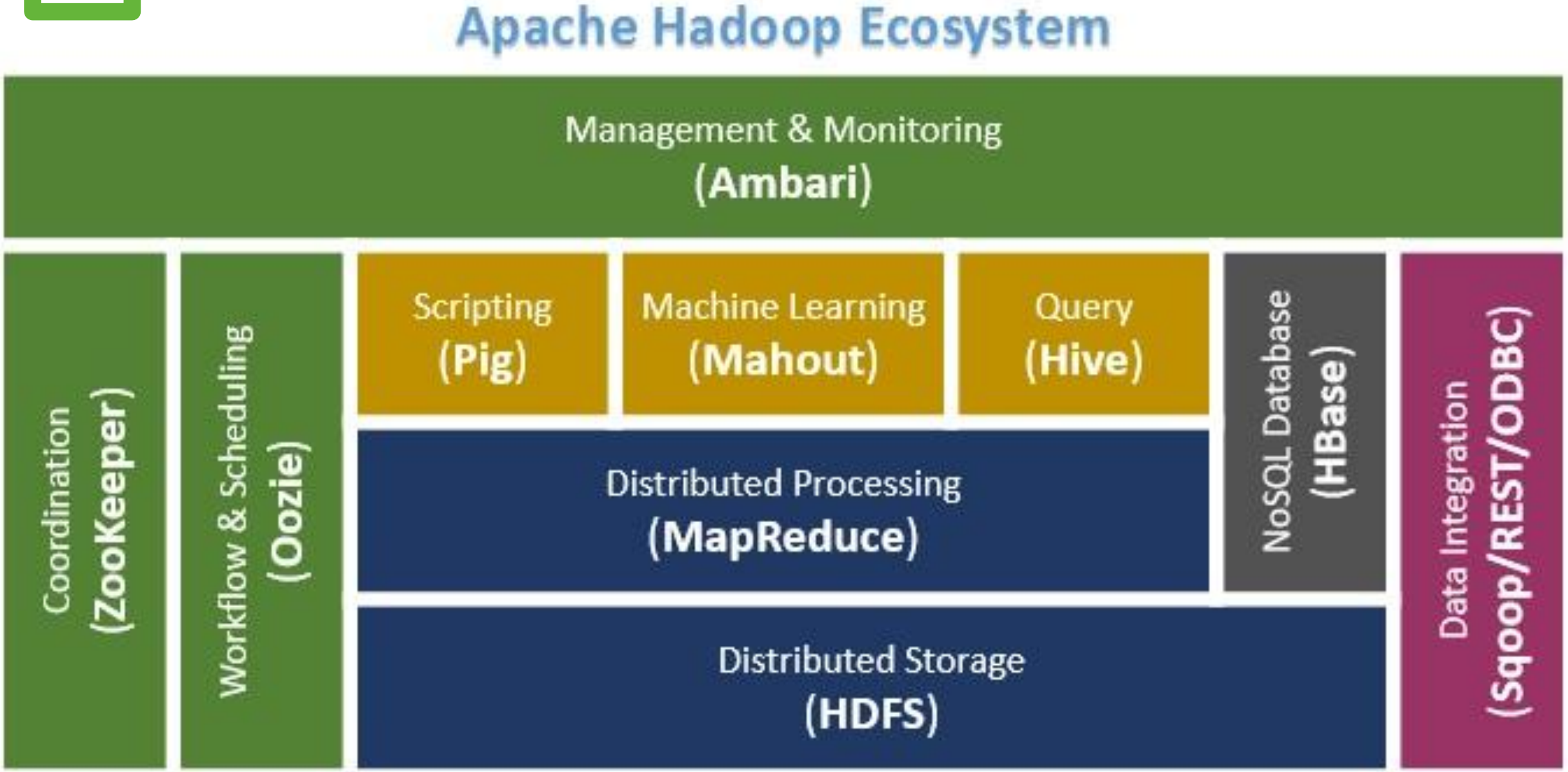
- Procesa Terabytes al Dia!
- Escalado Vertical no es suficiente:
  - tiempos de búsqueda del disco
  - fallas de hardware
  - tiempos de procesamiento
- Escalado Horizontal es lineal
- Hadoop: hace mas que proceso por lotes

El Mundo de Hadoop

- Motores de Query
- Ecosistema de Hadoop
- Almacenamiento externo de datos

DataDosis

Ecosistema de Hadoop



HDFS (Almacenamiento Distribuido)

Hadoop Distributed File System

- Permite distribuir los datos a travez de un cluster de computadoras
- Hace que todo se vea como una sola computadora gigante
- Mantiene Copias Redundantes de los Datos
  - Si una pc se rompe, el Sistema sigue corriendo sin problemas

DataDosis

YARN

Yet Another Resource Negotiator

- Aqui empieza el proceso de datos
- Es lo que maneja los recursos de tu cluster
- Decide cuanta memoria alocar en los procesos
- Que procesos deben corer en que orden
- Que nodos estan disponibles
- Es el *Corazon del cluster*

MapReduce

Modelo de programacion

- permite procesar los datos en un cluster
- Consiste de:
  - Mappers
  - Reducers

  Son distintos libretos que escribes, o funciones
- **Mappers**: permiten transformar datos en paralelo a través de un cluster, de manera eficiente
- **Reducers**: Agrega todos esos datos y los junta

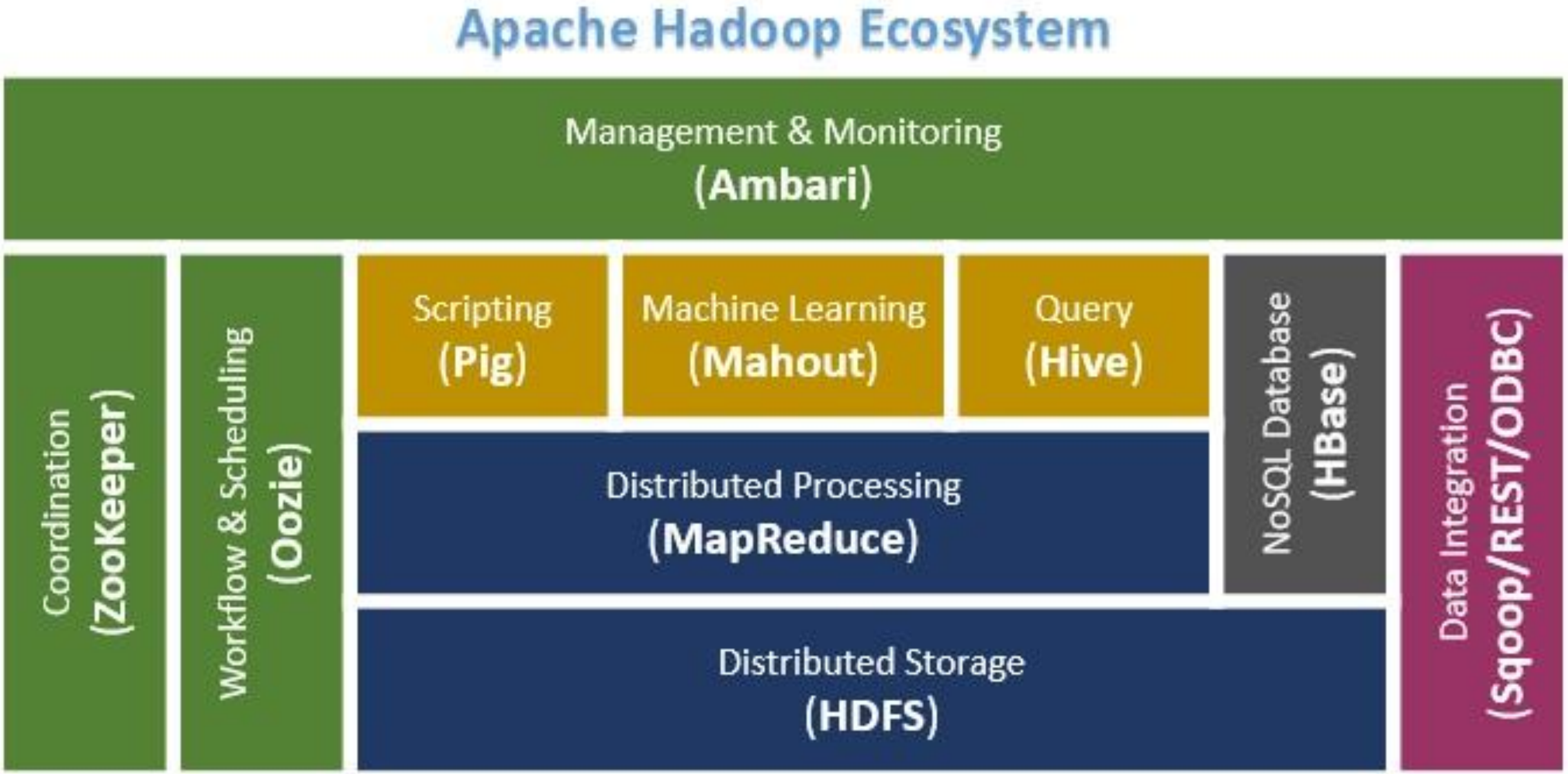
DataDosis

Pig

Funciona Encima de MapReduce

- Si no quieres escribir código java o Python puedes usar Pig
- Utiliza escritura estilo SQL

Ecosistema de Hadoop: **HIVE**

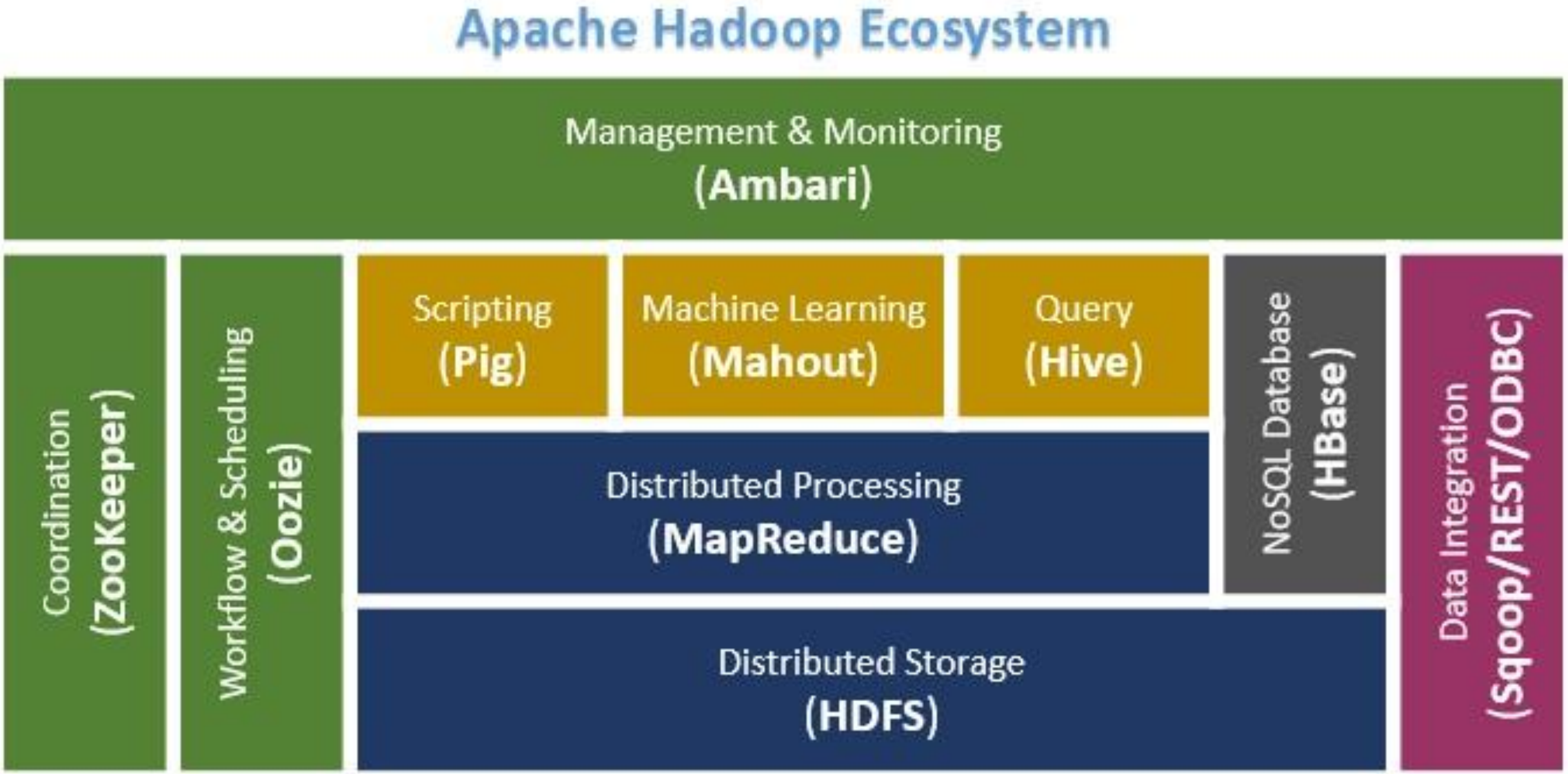


Hive

Resuelve Problema Similar a Pig

- Se ve como una base de datos SQL
- Toma queries SQL y hace que los datos distribuidos en tu sistema se vean como una base de datos SQL
- *Si te gusta SQL*, Hive es para ti!

Ecosistema de Hadoop: **Apache Ambari**



Apache Ambari

Monitorea el Sistema Completo

- Visualizar el cluster
- El ojo de todo, nos permite ver que esta corriendo
- Que sistemas usan cuantos recursos
- Permite ejecutar Queries con HIVE o pig

Mesos

Alternativa a YARN

- Resuelve los mismos problemas de otra manera

DataDosis

Spark

Esta al mismo nivel que MapReduce

- Encima de YARN o Mesos
- Permite correr Queries en los datos
- Requiere saber programacion
- Escribes programas con Scala o Python
- Es muy **rapido** y **bastante utilizado**
- Machine Learning
- Data Streaming

Tez

Similar a Spark

- Permite Ejecutar planes Optimos para correr Queries
- Corre en conjunto con HIVE para acelerarlo
  - HIVE puede ir encima de MapReduce o encima de Tez
- Es un medio para hacer queries y obtener respuestas rapidas del cluster

Apache HBase

Expone Datos a Plataformas Transaccionales

- Base de datos NoSQL
- Permite exponer nuestros datos procesados por MapReduce a los sistemas externos que los necesitan

Apache Storm

Transmicion de Datos (Data Streaming)

- Datos de Web
- Datos de Sensores
- ***Spark Streaming*** resuelve en mismo problema
- Procesa Datos Rápidamente en Tiempo Real

Oozie

Programar Trabajos en un Cluster

- Si tienes un cluster con muchos pasos
  - Oozie se encarga de poder hacer la programacion de los pasos facilmente
- ***Ejemplo***: Cargar datos a Hive, transformarlos en Pig, Pasar los resultados por Spark, mostrar todo en Hbase
  - Cada paso lo puedes programar desde Oozie

Zookeeper

Coordiandor de cluster

- Coordina los nodos
- Observa estados compartidos en el cluster
- Varias de las Apps que vimos requieren de Zookeper

Ingestion de Datos

Como asimilas la informacion

- ***SQOOP***: permite convertir Base de datos en base de datos Relacional
  - Conector entre Hadoop y base de datos relacionales
- ***FLUME***: transforma web logs en informacion para Hadoop
  - Informacion de servidores web Hadoop
- ***KAFKA***: recolecta datos de las computadoras, servidores, sensors y envia los datos a Hadoop

Almacenamiento Externo de Datos

- MySQL
  - Puedes exportar los datos a MySQL
  - Spark permite guardar en JDBC o ODBC y puedes guardarr los datos en mySQL para acceder a ellos en otro momento
- Cassandra
- mongoDB
  - Ambos Cassandra Y MongoDB son Buenos para exponer los datos a un API que requiera de estos

DataDosis

Motores de Query

Permite entrar Queries Interactivamente

Apache Drill

- Escribir queries SQL que pueden correr en bases de datos NoSQL
- Puedes hablar con Hbase, Cassandra y MongoDB

Hue

- Otra manera de crear queries interactivamente
- Puede tomar el rol de Apache Ambari y permite visualizar todo el cluster

Apache Phoenix

Similar a Apache Drill

- Tambien hace que una NoSQL se vea como una bd Relacional

DataDosis

Presto

Ejecutar Queries a Travez de Todo el Cluster

DataDosis

Apache Zeppelin

UI estilo Notebook que permite interactuar con el Cluster

DataDosis