# EJERCICIOS

## ECOSISTEMA HADOOP

### EJERCICIO 1

Rellena la siguiente tabla:

|  |  |  |
| --- | --- | --- |
| **NOMBRE** | **DESCRIPCIÓN** | **LOGOTIPO** |
| **Apache Hive** | Apache Hive es un sistema de almacenamiento de datos distribuido y tolerante a fallos que permite realizar análisis a gran escala. Hive Metastore (HMS) proporciona un repositorio central de metadatos que se pueden analizar fácilmente para tomar decisiones informadas basadas en datos y, por lo tanto, es un componente fundamental de muchas arquitecturas de lagos de datos. Hive está construido sobre Apache Hadoop y admite el almacenamiento en S3, adls, gs, etc. a través de hdfs. Hive permite a los usuarios leer, escribir y administrar petabytes de datos mediante SQL. |  |
| **Apache Pig** | Apache Pig es una plataforma para analizar grandes conjuntos de datos que consta de un lenguaje de alto nivel para expresar programas de análisis de datos, junto con una infraestructura para evaluar dichos programas. La característica más destacada de los programas Pig es que su estructura es susceptible de una paralelización sustancial, lo que a su vez les permite manejar conjuntos de datos muy grandes | Apache Pig - Wikipedia |
| **Apache HBase** | Apache HBase es la base de datos de Hadoop , un almacén de big data distribuido y escalable. Utilice Apache HBase cuando necesite acceso aleatorio y en tiempo real de lectura/escritura a su Big Data. El objetivo de este proyecto es alojar tablas muy grandes (miles de millones de filas por millones de columnas) sobre clústeres de hardware de consumo. | Apache HBase – Big Data Dummy |
| **Apache Flume** | Flume es un servicio distribuido, confiable y disponible para recopilar, agregar y mover de manera eficiente grandes cantidades de datos de registro. Tiene una arquitectura simple y flexible basada en flujos de datos en tiempo real. Es robusto y tolerante a fallas con mecanismos de confiabilidad ajustables y muchos mecanismos de recuperación y conmutación por error. Utiliza un modelo de datos simple y extensible que permite la aplicación analítica en línea. | Get to Know Apache Flume from Scratch! - Analytics Vidhya |
| **Apache Sqoop** | Apache Sqoop(TM) es una herramienta diseñada para transferir eficientemente datos masivos entre Apache Hadoop y almacenes de datos estructurados, como bases de datos relacionales. |  |
| **Apache Oozie** | Oozie es un sistema de programación de flujo de trabajo para administrar trabajos de Apache Hadoop.  Los trabajos de flujo de trabajo de Oozie son gráficos acíclicos dirigidos (DAG) de acciones.  Los trabajos de Coordinador de Oozie son trabajos recurrentes de Flujo de trabajo de Oozie que se activan por tiempo (frecuencia) y disponibilidad de datos. |  |
| **Apache ZooKeeper** | ZooKeeper es un servicio centralizado para mantener la información de configuración, asignar nombres, proporcionar sincronización distribuida y proporcionar servicios de grupo. Todos estos tipos de servicios son utilizados de una forma u otra por aplicaciones distribuidas. Cada vez que se implementan, hay mucho trabajo que se dedica a corregir los errores y las condiciones de carrera que son inevitables. Debido a la dificultad de implementar este tipo de servicios, las aplicaciones inicialmente suelen escatimar en ellos, lo que los hace frágiles en presencia de cambios y difíciles de administrar. | Apache ZooKeeper - Wikipedia, la enciclopedia libre |
| **Apache Storm** | Apache Storm se integra con cualquier sistema de colas y cualquier sistema de base de datos. La abstracción spout de Apache Storm facilita la integración de un nuevo sistema de colas.  De la misma manera, la integración de Apache Storm con sistemas de bases de datos es sencilla. Simplemente abra una conexión a su base de datos y lea/escriba como lo haría normalmente. Apache Storm se encargará de la paralelización, la partición y la reintención de errores cuando sea necesario. | Apache Storm - Wikipedia |
| **Apache Spark** | Apache Spark ™ es un motor multilenguaje para ejecutar ingeniería de datos, ciencia de datos y aprendizaje automático en máquinas de un solo nodo o clústeres. | Apache Spark - Wikipedia |
| **Apache Kafka** | Apache Kafka es una plataforma de transmisión de eventos distribuida de código abierto utilizada por miles de empresas para canalizaciones de datos de alto rendimiento, análisis de transmisión, integración de datos y aplicaciones de misión crítica. | Apache Kafka |
| **Apache Atlas** | Atlas es un conjunto escalable y extensible de servicios básicos de gobernanza que permiten a las empresas cumplir de manera eficaz y eficiente sus requisitos de cumplimiento dentro de Hadoop y permite la integración con todo el ecosistema de datos empresariales.  Apache Atlas proporciona capacidades abiertas de gestión y gobernanza de metadatos para que las organizaciones creen un catálogo de sus activos de datos, clasifiquen y gobiernen estos activos y brinden capacidades de colaboración en torno a estos activos de datos para científicos de datos, analistas y el equipo de gobernanza de dato | Apache Atlas – Data Governance and Metadata framework for Hadoop |
| **Apache Mahout** | Apache Mahout(TM) es un marco de álgebra lineal distribuida y un DSL de Scala con capacidad de expresión matemática diseñado para permitir que los matemáticos, estadísticos y científicos de datos implementen rápidamente sus propios algoritmos . Apache Spark es el back-end distribuido listo para usar recomendado, o se puede extender a otros back-end distribuidos. | Apache Mahout |
| **Apache Phoenix** | Apache Phoenix permite OLTP y análisis operativo en Hadoop para aplicaciones de baja latencia al combinar lo mejor de ambos mundos:  el poder de las API estándar de SQL y JDBC con capacidades completas de transacciones ACID y  La flexibilidad de las capacidades de lectura de esquemas con enlaces tardíos del mundo NoSQL al aprovechar HBase como su almacenamiento de respaldo | Overview | Apache Phoenix |
| **Apache Impala** | Impala eleva el nivel de rendimiento de las consultas SQL en Apache Hadoop y, al mismo tiempo, mantiene una experiencia de usuario familiar. | Apache Hive vs Apache Impala: principales diferencias - Geekflare |
| **Apache Nifi** | En pocas palabras, NiFi se creó para automatizar el flujo de datos entre sistemas. Si bien el término "flujo de datos" se utiliza en una variedad de contextos, lo usamos aquí para referirnos al flujo automatizado y administrado de información entre sistemas. Este problema ha existido desde que las empresas tenían más de un sistema, donde algunos de los sistemas creaban datos y otros los consumían. | Apache Nifi. Flujo de datos y ETL/ELT en Streaming. - Inteligencia  Artificial y Big Data |
| **Apache Ambari** | El proyecto Apache Ambari tiene como objetivo simplificar la gestión de Hadoop mediante el desarrollo de software para el aprovisionamiento, la gestión y la supervisión de clústeres Apache Hadoop. Ambari ofrece una interfaz de usuario web de gestión de Hadoop intuitiva y fácil de usar, respaldada por sus API RESTful. | Ambari - |
| **Mapreduce** | MapReduce es un modelo de programación que se ejecuta sobre Hadoop, un motor de análisis de datos ampliamente utilizado para Big Data. Programa aplicaciones que se ejecutan en paralelo para procesar grandes volúmenes de datos almacenados en clústeres. | Hadoop Ecosystem | Hadoop Tools for Crunching Big Data | Edureka |