Sección #5 BigData, Hadoop y Hive

1. **Las 5 Vs del Big Data:**
   - ***Volumen***: Esto se refiere a la enorme cantidad de datos que se generan cada día. Es algo así como la cantidad de información que podríamos tener en miles de discos duros llenos. Un ejemplo es todo lo que se publica en redes sociales en un solo día, como fotos, videos y textos. Si intentáramos guardar todo eso en una base de datos como MySQL, tendríamos problemas, porque MySQL no está hecho para manejar tanta información sin volverse lento o hasta colapsar.
   - ***Velocidad***: Aquí hablamos de la rapidez con la que los datos son generados y también cómo deben ser procesados casi en tiempo real. Un buen ejemplo es el procesamiento de transacciones bancarias o las actualizaciones en redes sociales, donde cada segundo se están registrando miles de acciones. En MySQL, manejar ese tipo de velocidad es complicado porque no es tan rápido y eficiente para procesar información en tiempo real a gran escala.
   - ***Variedad***: Los datos no vienen todos en el mismo formato, pueden ser de diferentes tipos, como texto, imágenes, videos o incluso sonidos. Por ejemplo, pensemos en los datos de una tienda online: hay texto (como descripciones de productos), fotos, comentarios, etc. MySQL se especializa en datos estructurados (tablas y filas), y manejar diferentes tipos de datos a la vez puede volverse un lío y afectar el rendimiento.
   - ***Veracidad***: Aquí nos referimos a la calidad o precisión de los datos. No todos los datos son confiables, pueden contener errores o ser inconsistentes. Por ejemplo, en un sistema de sensores, algunos podrían enviar datos incorrectos. MySQL no tiene herramientas avanzadas para lidiar con datos que pueden no ser precisos o para hacer verificaciones de calidad, lo que hace más complicado filtrar los datos confiables de los que no lo son.
   - ***Valor***: Esta es la capacidad de extraer algo útil de los datos, encontrar patrones que realmente aporten. Por ejemplo, en una tienda online, analizar los datos para saber qué productos interesan más a los clientes. MySQL se queda corto cuando se trata de extraer valor de enormes cantidades de datos y suele necesitar herramientas adicionales, como sistemas de análisis o de minería de datos.

2. **Distribuciones de Hadoop:**
   - ***Cloudera***: Es una de las distribuciones más populares de Hadoop. Viene con módulos como HDFS (Hadoop Distributed File System) que sirve para almacenar datos distribuidos en varios nodos y MapReduce, que es un modelo de procesamiento de datos para hacer cálculos en paralelo. Cloudera también incluye herramientas de administración que hacen más fácil controlar y monitorear los clústeres, es decir, los grupos de computadoras que trabajan juntos en el procesamiento de datos.
   - ***Hortonworks:*** Otra distribución conocida que también incluye HDFS y YARN (Yet Another Resource Negotiator). YARN se encarga de gestionar los recursos del sistema y asignarlos a diferentes tareas. Hortonworks es ideal para empresas que necesitan escalar su capacidad de procesamiento de datos y ofrece integración con otras herramientas de Big Data como Apache Spark y Apache Hive.

3. **Modos de ejecución de Hadoop:**
   - ***Modo Local***: Todo se ejecuta en una sola máquina, sin distribución. Es muy útil para hacer pruebas o experimentar porque no necesitamos un clúster completo. Aquí es donde solemos comenzar a hacer pruebas antes de pasar a algo más grande.
   - ***Modo Pseudo-distribuido***: En este modo, cada componente de Hadoop corre en un solo nodo, pero imita el comportamiento de un clúster real. Ideal para desarrollo y pruebas en un entorno que se asemeja más al de producción.
   - ***Modo Distribuido***: Aquí es donde Hadoop realmente brilla. Se ejecuta en un clúster de múltiples nodos, ideal para trabajar con grandes volúmenes de datos en producción. Es lo que usan las empresas para análisis de datos a gran escala.
4. **Roles en YARN:**
   - ***ResourceManager (RM)***: Es como el jefe o administrador de recursos. Su trabajo es asignar los recursos del sistema (como la memoria y el procesador) a las tareas que lo necesiten y asegurarse de que todo esté funcionando correctamente.
   - ***NodeManager (NM)***: Este se encarga de ejecutar las tareas que le asigna el ResourceManager en su propio nodo. Cada nodo en el clúster tiene un NodeManager que realiza el trabajo de procesar los datos y reporta el progreso al ResourceManager.
5. **Diferencias entre MySQL y HiveSQL:**
   - ***MySQL***: Es una base de datos relacional diseñada para datos estructurados y transacciones rápidas. Funciona genial para bases de datos pequeñas o medianas, y para consultas rápidas de datos estructurados. No está hecha para manejar petabytes de datos o para hacer consultas en paralelo a gran escala.
   - ***HiveSQL***: Es más una interfaz para trabajar con datos en Hadoop y usa un lenguaje similar a SQL para hacer consultas sobre datos almacenados en HDFS. Está optimizado para Big Data y permite consultas en clústeres distribuidos. Es más lento que MySQL en consultas pequeñas, pero mucho más eficiente cuando se trata de grandes volúmenes de datos, porque está diseñado para hacer análisis y no para transacciones en tiempo real.