Tipo : Lectura
Capítulo : Administración de Hadoop

I. OBJETIVO

Ampliar sus conocimientos sobre cómo utilizar una distribución de Hadoop

II. LECTURAS COMPLEMENTARIAS

Distribuciones Hadoop: ¿Cómo gestionar tu clúster Hadoop?

Es una realidad que los cinco gigantes del big data (Amazon, Apple, Google, Microsoft y Facebook) cuentan el número de servidores de sus clústeres para el almacenamiento y análisis de datos por miles e incluso millones. Aunque la gestión de estos clústeres no es una tarea sencilla, poseen recursos y herramientas desarrolladas por ellos mismos que solucionan estas tareas.

Instalar algunas herramientas del ecosistema Hadoop en un único servidor para realizar pruebas es relativamente sencillo y existen manuales que lo explican perfectamente. Pero cuando se trata de administrar un clúster la tarea se vuelve más complicada e implica la intervención de varios agentes: analistas de requisitos que evalúan los requisitos de rendimiento, arquitectos de sistemas que evalúan las configuraciones de hardware, ingenieros de sistemas que instalan y configuran el ecosistema de Hadoop…

Existen diferentes herramientas o distribuciones que nos permiten administrar nuestro clúster de manera sencilla. Aunque existen varias en el mercado como Cloudera, EMC, Hortonworks, IBM Big Blue, Intel Distribution o MapR. En este artículo nos vamos a centrar en las dos principales: Cloudera y Hortonworks.

Cloudera

Fundada en 2008 por tres ingenieros de Google, Yahoo y Facebook más un antiguo ejecutivo de Oracle. Cloudera Inc. proporciona software, soporte, servicios y formación basados en Apache Hadoop. Cloudera Distribution for Hadoop (CDH) es la distribución líder y más conocida para el ecosistema Hadoop. Ha sido la primera en lanzar una distribución comercial y posee un gran número de clientes importantes.
CDH posee una consola de administración (Cloudera Manager) fácil de usar que muestra la información de manera clara. También provee la Cloudera Management Suite que posee un asistente para el despliegue, un panel de gestión y un módulo de gestión de recursos con el fin de simplificar la planificación y expansión.
En gran parte CDH es open source y solo algunos componentes son propietarios. Esto beneficia a los usuarios que buscan minimizar los riesgos de la dependencia obligada con el proveedor de servicios (vendor lock-in).

Por otro lado, **Cloudera ofrece servicios de consultoría que ayudan a las organizaciones a integrar las tecnologías Hadoop con su estrategia** de gestión de datos.
En su cartera de clientes están presentes firmas como Cisco, Siemens, Samsung, MasterCard o SanDisk.

Hortonworks
Fue fundada en 2011 por ingenieros de Yahoo. Hortonworks Hadoop distribution (HDP) es la única distribución completamente libre y puede ser descargada e integrada fácilmente. **HDP está centrada en mejorar la usabilidad del ecosistema Hadoop**. Su instalación es completamente gratuita, existiendo una versión de pago con la única diferencia de ofrecer soporte técnico.

Es la única distribución que soporta Windows. Los usuarios pueden desplegar en un clúster hadoop basado en Windows en Azure a través del servicio HDInsight.
Además, Hortonworks es el principal contribuyente del proyecto Ambari, perteneciente al ecosistema Hadoop cuyo objetivo es proporcionar herramientas de administración y monitorización.

Al ser completamente libre **su integración con nuevas actualizaciones de las distintas herramientas se realiza de manera fácil y rápida**. Por otro lado, también evita el vendor lock-in.
Entre sus principales clientes se encuentra Ebay, Bloomberg, Spotify y Samsung Electronics.

Cloudera & Hortonworks (Similitudes)

- Están enfocadas a Hadoop y ofrecen distribuciones preparadas para el entorno empresarial.
- Tienen comunidades que participan activamente.
- Tiene arquitectura master-slave.

Cloudera vs. Hortonworks (Diferencias)

- La estrategia de largo plazo de Cloudera es convertirse en una data hub empresarial, disminuyendo la necesidad de data warehouse. En cambio, Hortonworks continuará siendo un proveedor de distribuciones Hadoop.
- Cloudera puede ser ejecutado en servidores Windows. Hortonworks está disponible como un componente nativo en servidores Windows.
- Cloudera tiene componentes propietarios mientras Hortonworks no posee software propietario.
- Cloudera tiene una versión de prueba de 60 días mientras Hortonworks es completamente libre.

¿Qué hay que tener en cuenta antes de seleccionar una distribución?

Cada distribución posee una serie de características y herramientas que se deben analizar de manera previa a su selección:

- ¿Cuáles son tus objetivos? ¿Qué resultado estás esperando obtener? ¿Qué problema quiere resolver?
- ¿Cómo puede ayudar Hadoop a cumplir esos objetivos? ¿La infraestructura que vas a seleccionar es lo suficientemente flexible para tus necesidades?
- ¿Qué tipo de datos se analizarán? ¿Existen cuestiones legales sobre los datos (LOPD)? ¿La distribución es compatible con las políticas de protección de datos de tu compañia?
- ¿Componentes propietarios o código abierto?
- ¿Qué herramientas de administración necesitas?
- ¿La distribución que selecciones cumplirá tus necesidades en un futuro?
- ¿La distribución te ata a un proveedor de servicios cloud (vendor lock-in)?
- ¿Te proporcionan soporte técnico?

Conclusión

Aunque Cloudera fue la primera compañía en crear distribuciones de Hadoop y actualmente es número 1 en cuanto a distribuciones Hadoop, mientras que Hortonworks ha experimentado una gran evolución y expansión.

Por lo tanto, si se desea seleccionar una distribución se hace necesario responder a las preguntas del apartado anterior.

No existe un proveedor claramente ganador.

Bibliografía:

- "Distribuciones Hadoop: ¿Cómo gestionar tu clúster Hadoop?"
  http://www.agiliacenter.com/distribuciones-hadoop-gestionar-cluster-hadoop/
  Consulta: 01 de diciembre del 2017