* + 1. Contenido Entrenamiento Plataforma BigData

A continuación, se listan los diferentes temas de capacitación:

Información acerca de Hadoop

Este es un curso introductorio sobre Hadoop y su ecosistema. La tecnología Hadoop se utiliza para construir complejas y poderosas aplicaciones de procesamiento de datos. Este curso está diseñado para administradores y analistas que son nuevos en Big Data y Hadoop y quieren entender cómo las organizaciones están usando Hadoop para maximizar el valor de los datos.

Hadoop para Desarrolladores

Este curso de capacitación es para desarrolladores que quieren aprender a usar Hadoop para crear poderosas aplicaciones de procesamiento de datos. Al final de la capacitación, los participantes aprenderán el proceso de tomar un proyecto Hadoop desde la concepción hasta su finalización, las mejores prácticas para usar Hadoop más eficazmente para proporcionar soluciones, cómo escribir un programa MapReduce usando Hadoop API, utilizando HDFS para la carga y procesamiento efectivos De datos y cómo funcionan HDFS, MapReduce y YARN.

Además, el curso también cubre los temas a considerar cuando se desarrollan los trabajos de MapReduce, cómo implementar algoritmos comunes en Hadoop, cómo depurar, monitorear y optimizar las soluciones de Hadoop, cómo aprovechar otros proyectos como Hive, Pig, HBase y advanced Hadoop API temas requeridos para el análisis de datos del mundo real.

Información acerca de Bedrock

Este curso se centra en la plataforma de gestión de datos Bedrock, que es una solución integral de gestión de datos basada en Hadoop. Esta plataforma le permite visualizar, analizar, organizar y recopilar datos para crear e implementar aplicaciones de producción de Hadoop. La plataforma incorpora los principios fundamentales de la gestión de datos.

Bedrock para Desarrolladores

Este curso se centra en la plataforma de gestión de datos Bedrock, que es una solución completa de gestión de datos basada en Hadoop. Esta plataforma le permite visualizar, analizar, organizar y recopilar datos para crear e implementar aplicaciones de producción de Hadoop. La plataforma incorpora los principios fundamentales de la gestión de datos.

Este curso puede ser personalizado para cubrir las herramientas y tecnología de desarrollo / integración de datos

Administración de Apache Hadoop

Este entrenamiento combina la experiencia, la práctica y la lección para enseñar a los estudiantes cómo configurar y administrar los clústeres que ejecutan la distribución de Apache Hadoop. Este curso abarca Hadoop Fundamentos y Arquitectura, Hadoop Ecosystems Overview, Requisitos de hardware y software para Apache Hadoop, Información general sobre la instalación y la configuración, Implementación de un clúster Apache Hadoop, Implementación de los servicios del ecosistema Hadoop, Configuración de la seguridad mediante Kerberos, Hadoop, Manage and Monitor Apache Hadoop Cluster, Solución de problemas del clúster Apache Hadoop, Nagios y Ganglia.

Operación y Apoyo de Data Lake

Este entrenamiento personalizado enseña a los estudiantes todos los aspectos operacionales de un lago de datos saludable - desde la hidratación hasta la organización, preparación y extracción de datos. Se consideran en detalle varios aspectos de la seguridad desde la infraestructura hasta el acceso a los datos y los controles basados ​​en roles. La gestión de datos de Lake utilizando Bedrock se describe en detalle.

Migración de Datos

Este módulo de capacitación personalizado enseña a los estudiantes los principios de migración de datos de los sistemas de bases de datos relacionales tradicionales a los almacenes de datos basados ​​en datos de Big Data Hadoop. Cubre los fundamentos, comparando y contrastando los dos sistemas, y guía a los estudiantes a través de la consideración práctica de la migración tanto de datos históricos como de tuberías de datos. Las herramientas y tecnologías de uso común se introducen junto con los patrones arquitectónicos para proyectos de migración exitosos.

**Capacitación HADOOP – DAY 1**

1. Computation
2. Flexibility: Storage first, then the porpuse
3. Fault torerance
4. Scalability
5. Low Cost

SSSD Service to use to login to machine. Con el usuario y clave del directorio active

Cluster in Colombia es Cloudera (Hadoop)

Data hub: 3 capas:

* Storage (Filesystem and Online NoSQL)
* Manage the resources
* Data Management (Analytic SQL (impala), Machine learning (spark), batch processing (mapreduce), etc…)

**HDFS Architecture:**

1. **Namenode:** Manage the whole filesystem.
   * FS Image (Metadata of the file) 🡪 Filename, replication factor, modificatons (change the file or change the permisions of the file), access time (how long you have access to the file), permisions
   * Edit logs 🡪Cualquier cambio en el file queda capturado en el edit logs. Cambios en los permisos del archivo, quien lo modificó, quien accedió al archivo, etc…
2. **Secondary Namenode:** Made Checkpointing: Combine Fs image with edit logs y lo envía de vuelta al namenode. Checkponting is every hour para que el namenode y secondary esten sincronizados.
3. **Datanode:** Son los datos. Cualquier dato que alimenta el HDFS está en el datanode. El Datanote se alimenta por bloques: 1 bloque =64MB.

La información se replica en 3 datanodes siempre. En este caso 1 bloque por cada nodo esclavo. Total 3 bloques.

Cada bloque tiene que estar replicado 3 veces

Datanode send a heartbeat to the namenode node every 10 min. Si uno de los datanode no envia el heart beat y ya no hay otro nodo para replicar, se genera un warning.

**YARN Architecture (is call Map Reduce 2)**

Yet Another Resource Negociator

Comparando con el Mapreduce1 las tareas a, b y c del JobTracker se separan en tres diferentes componentes: REsource Manager, Application Master and Timeline Server). Mapreduce 2 puede correr hasta en 10.000 nodos

1. **Resource Manager 🡪** hace lo que hace el “Schedule the job” en Mapreduce1
2. **Node Manager:** Report all the task to the resource manager.

Aplication Master is inside of the Node manager

1. **Aplication Master 🡪** hace lo que hace el “assign task” to the node manager en Mapreduce1
2. **Time Line Server 🡪** hace lo que hace el “monitor progres to task” en Mapreduce1
3. **Container based model**

* Mapper: 1)Reads the blocks (data), 2)organaiced and 3)output to the reducer
* Reducer:

La asignación de los recursos se hace por default para el mapper y el reducer.

Overview Map Reduce 1 (requerido para entender YARN Achitecture):

MapsReduce: is an application, it runs jobs at the storage system. It contains:

1. **Job Tracker:** run in the master nodes (namenode ans secondary)
   1. Schedule the job
   2. Assign Task to the Task Tracker
   3. Monitor the progress of task
2. **Task Tracker:** The Task Tracker components are the dataNodes

The task tracker reports to the Job Tracker. The task tracker debe ser instalado en en los datanodes. Este leer los datos , corre los queries, y reporta la salida al Job Tracker

1. **Slots:** Para cada cosa que quiero ejecutar se requieren los resources. Se requiere memoria.

En mapreduce 1 se define manualmente cuales son los recursos que se requiren para correr las tareas de mapper y reducer. Los recursos se deben asignar para cada uno ejemplo 44M de memoria para el mapper y 44 para el reducer

Mapreduce1 solo puede correr hasta en 4000 nodos

MapeReduce Ejemplo:

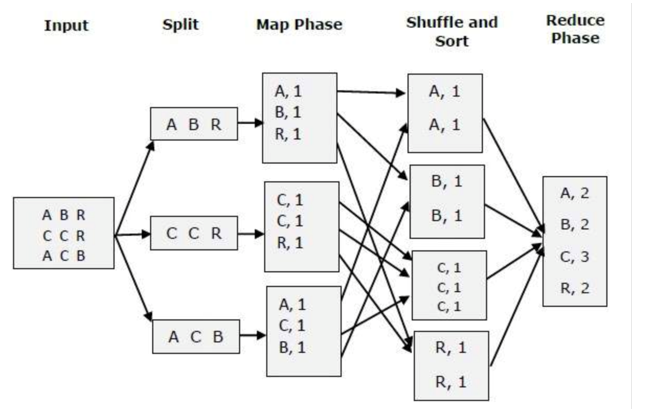
Se requiere el conteo de cantidad de letras que hay en un libro (ej datalake con las letras de un libro).

Se debe correr dos funciones: El MAPPER y el REDUCE

El mapper lee los datos, luego los organiza y luego la salida al reduce.

El Split y el Shuffle and sort no lo hacemos nosotros.

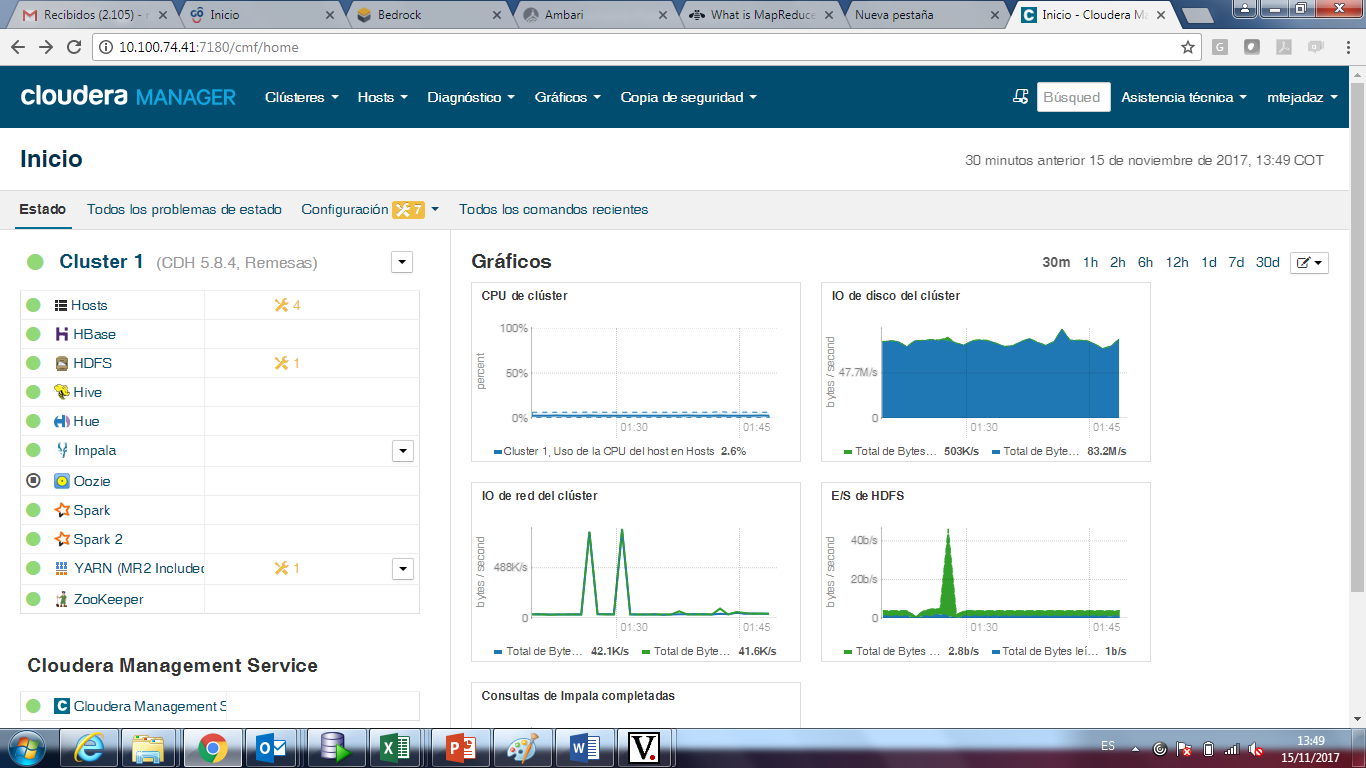
**Cloudera Manager Architecture:**



Otro ejemplo puede ser con temperaturas, por ejemplo se tiene en el datalake las temperaturas de todos los días desde 1970, para encontrar cuál es fue el día más caliente.

En UNE se tiene HIVE para mapreduce. Este hace lo hace internamente, es como un editro de SQL, es lo que se etá utiliznado en los modelos que tenemos migrados ahora en UNE.

**Cloudera Manager Architecture:**



HDFS:



YARN



**Security - Kerberos**

Se monta kerberos para poder tener la administración de los usuarios que puedan acceder a cierta información o no. Que puedan o no modificar o borrar.

Existen dos tipos de usuarios:

* System User
* Human User: Se autentica por SSH con un username y un login

En Kerberos el user name se llama: principal name

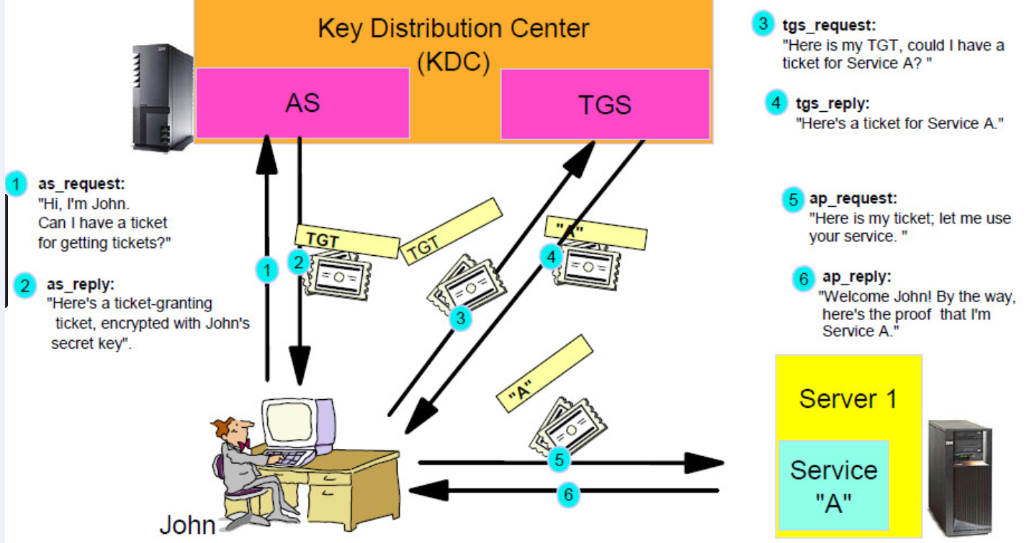
**Kerberos Architecture**

1. Authetication server
2. Ticket Granting Server
3. Service request

Key Distributios Server (KDC) es la BD donde están los usuarios que pueden acceder a Kerberos.

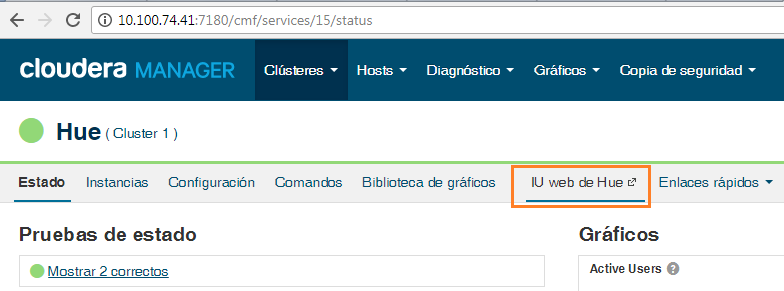
Tiene dos partes: AS (Authentication Server) and TGS (Tiquet granting Service)

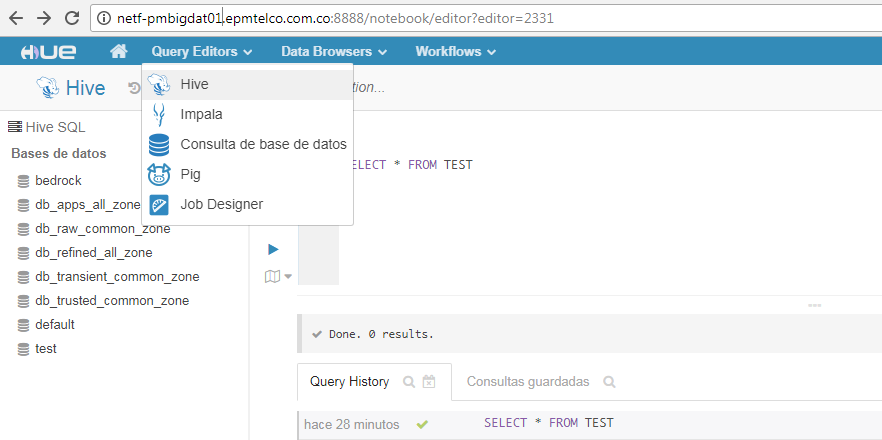
Kerberos se puede instalar en cualquier nodo, pero lo recommendable es instalarlo en los master nodes que son los que tiene más memoria



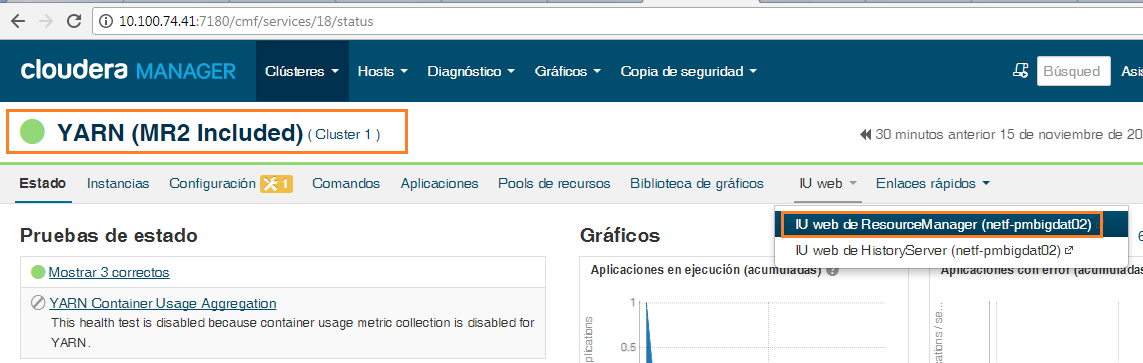
**HUE:**

****

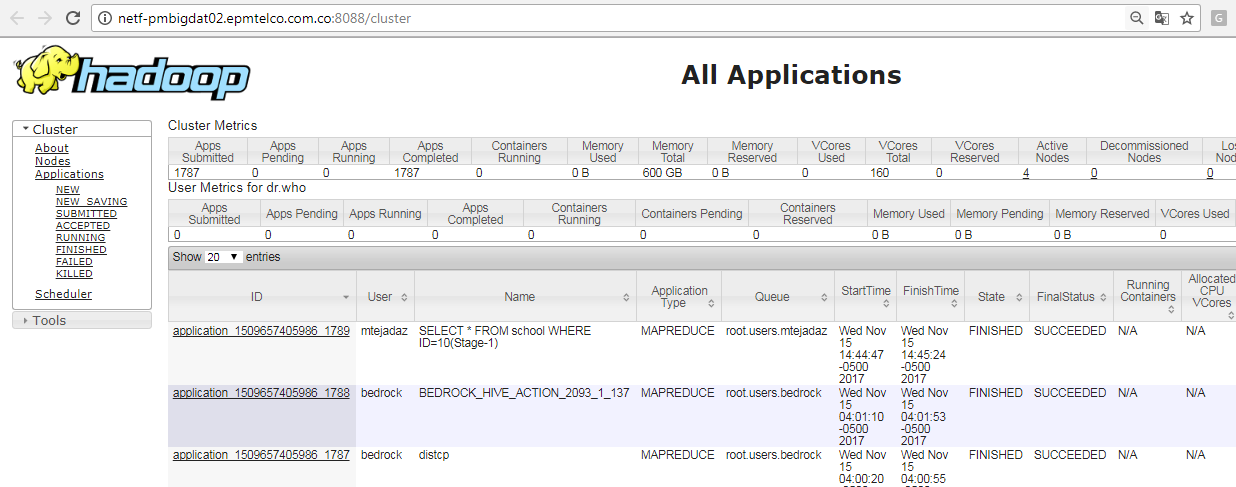
****



Para ver el histórico de los Jobs, en la opción YARN, IU Web:



Salen todos los Jobs, estados, etc…



**CAPACITACIÓN BIG DATA - DAY 2**