Descripción

Esta práctica se puntuará. La nota de la misma formará parte de la nota final de la asignatura. El alumno deberá escribir la configuración de un flujo, con uno o más agentes Flume en una o más capas, que de respuesta a lo solicitado en el enunciado.

Serán necesarios los conceptos mostrados en la sesión teórica para completar esta actividad.

Tiempo

La actividad deberá ser entregada en el plazo de un mes desde su publicación.

Enunciado

La empresa Floristerías Reunidas (FR) tiene 50 provinciales sedes en España. Cada una de estas sedes tiene un data-center que da servicio a cientos de floristerías asociadas. Cada una de las floristerías accede a los sistemas de su sede mediante una aplicación web que se ejecuta en el data-center de la sede. En esta aplicación se registra cada venta y cada pedido que se realiza.

Esta información es registrada en un log de aplicación presente en cada uno de los servidores web (varios por data-center, en máquinas diferentes). Este archivo de log se encuentra en la ruta /var/log/operaciones.log.

El log se rota diariamente hacia una carpeta /var/operaciones con el nombre operaciones-<fecha>.log donde se almacenan para su análisis.

Las operaciones se almacenan en el fichero de log con el formato:

[TIMESTAMP] [TIPO_OPERACION] [DATOS_OPERACION]

El timestamp es el epoch en tiempo universal y existen más de veinte tipos de operación diferentes. Entre ellos se encuentran los tipos 'VENTA' y 'PEDIDO'. Los datos de las asociados a una operacióno son diferentes según el tipo del que se trate.

El tamaño máximo de cada una de estas entradas de log en 300 bytes. El número de eventos por minuto en cada sede es de media 60.

Todas las sedes pueden acceder por red a dos máquinas 'fr-central-1' y 'fr-central-2' que han sido instaladas para recolectar la información hacia un HDFS que solo es accesible desde la sede central. Los puertos 10000 a 10100 están abiertos para que las aplicaciones puedan usarlos.

FR quiere almacenar toda la información relativa a ventas y pedidos en este HDFS para su posterior análisis. La URI del HDFS es hdfs://fr-hdfs:9000.

Definir un flujo que recolecte esta información y la deje en el HDFS separada primero por tipo de operación y luego por fecha del evento.

El sistema deberá tener en cuenta que el sistema HDFS puede estar parado durante un periodo máximo de tres horas. Se supondrá que el disco en las máquinas colectoras es suficiente para la configuración que diseñe el alumno.

El sistema tiene que implementar un mecanismo de failover en la comunicación desde las sedes a las máquinas recolectoras.