Sistema de Agrupamiento y Búsqueda de Contenidos de la Blogosfera de la ESPOL, Utilizando Hadoop como Plataforma de Procesamiento Masivo y Escalable de Datos.

Allan Avendaño

Agenda

- Introducción
- Paradigma Map/Reduce
- Amazon Web Services
- Information Retrieval
- Diseño
- Implementación
- Pruebas

INTRODUCCIÓN

Estado Actual

- **ESPOL** se ha planteado el objetivo de **mejorar su posicionamiento** en la lista de **sitios Web** de universidades latinoamericanas
- Se crea la blogosfera politécnica
  - Autores: estudiantes, profesores y personal afín
  - Propósito: Diversificación de contenidos
  - Aún no existen datos que permitan **determinar las características** de la **comunidad**
- Solución Actual: **Directorio de Blogs**
  - **No** permite realizar **búsquedas**
- Solución Propuesta: **Sistema de Búsqueda de Contenidos en la Blogosfera Politécnica**
  - Y un motor de recomendación de entradas

INTRODUCCIÓN

Motivación

- **Consolidar la comunidad** de autores de blogs de la ESPOL

- **Proporcionar** de una visión general de lo que **se escribe en la blogosfera politécnica** a los lectores y miembros de la comunidad

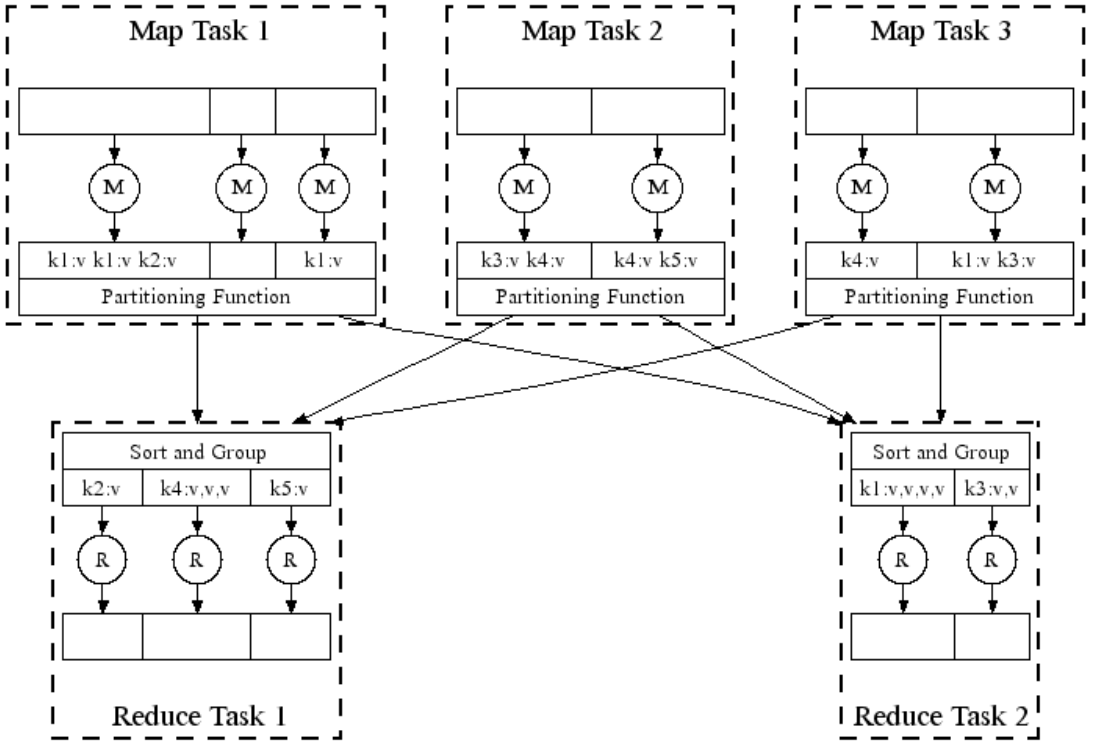
INTRODUCCIÓN

Objetivos

- **Implementar un módulo de agrupamiento** de contenidos de la blogosfera de la ESPOL
  - Usando Hadoop como plataforma de procesamiento masivo de datos
- **Implementar un módulo de búsqueda** de entradas publicadas en los blogs
- Implementar una **interfaz Web** para el sistema de búsqueda y agrupamiento de contenidos desarrollado

Paradigma Map/Reduce

- **Modelo de programación** desarrollado por **Google** para resolver **tareas de procesamiento** de datos a gran escala
  - Inspirado en las operaciones de lenguajes funcionales



Hadoop

- Plataforma de procesamiento distribuido y masivo de datos de la ASF
  - Sistema de archivos distribuido → HDFS
- Esconde las características complejas
  - Paralelización de tareas
  - Control de trabajos
  - Tolerancia a fallos
  - Escalabilidad
  - Detección de errores
  - Sincronización

Amazon Web Services

- **Computación en la nube**
- Uso de recursos **bajo demanda**
- Algunos servicios útiles:
  - Elastic Computing Cloud (EC2)
    - Permite, entre otras cosas, levantar clústeres con Hadoop instalado
  - Simple Storage Service (S3)
    - Almacenamiento escalable de datos
  - Elastic MapReduce (EMR)
    - Automatiza tarea de levantar un clúster Hadoop sobre EC2

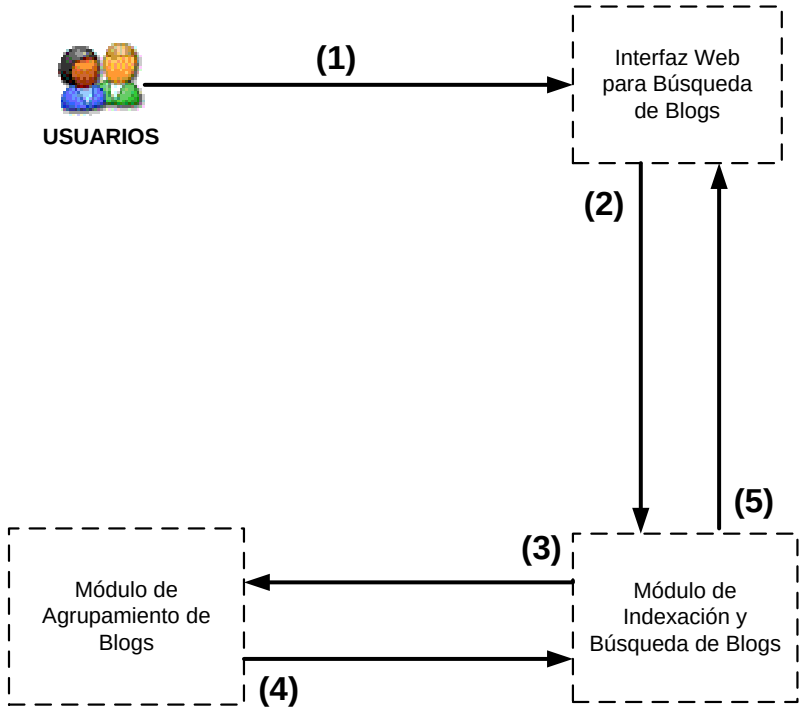
Information Retrieval

- **Representar, almacenar, organizar y acceder** a **documentos** relevantes tomados a partir de una colección de documentos **sin estructurar** (generalmente en lenguaje natural), con el objetivo de satisfacer las **necesidades de los usuarios**

- Técnicas de Extracción de Información:
  - Agrupamiento (Clustering).
  - Clasificación.

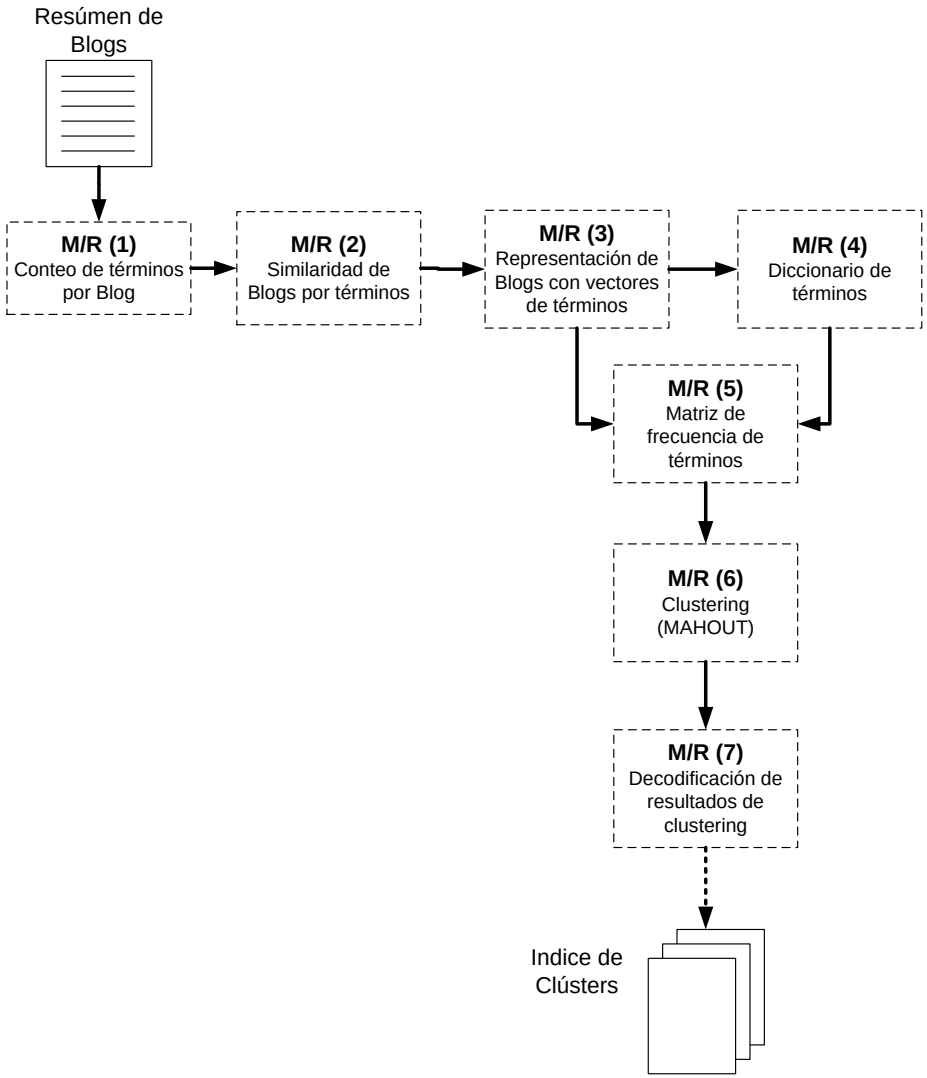
Diseño (I)

- Los usuarios ingresan los **términos de búsqueda** desde la interfaz Web del sistema **(1)**
- Los términos son enviados al **módulo de indexación y búsqueda** para obtener los blogs relevantes a los términos **(2)**
- A partir de las entradas resultantes de la búsqueda **(3)** se obtienen los **blogs relacionados (4)**
- Finalmente, el resultado es **procesado y visualizado** en la interfaz Web **(5)**



Diseño (II)

- Se implementan técnicas de extracción de información
- Siete procesos Map/Reduce



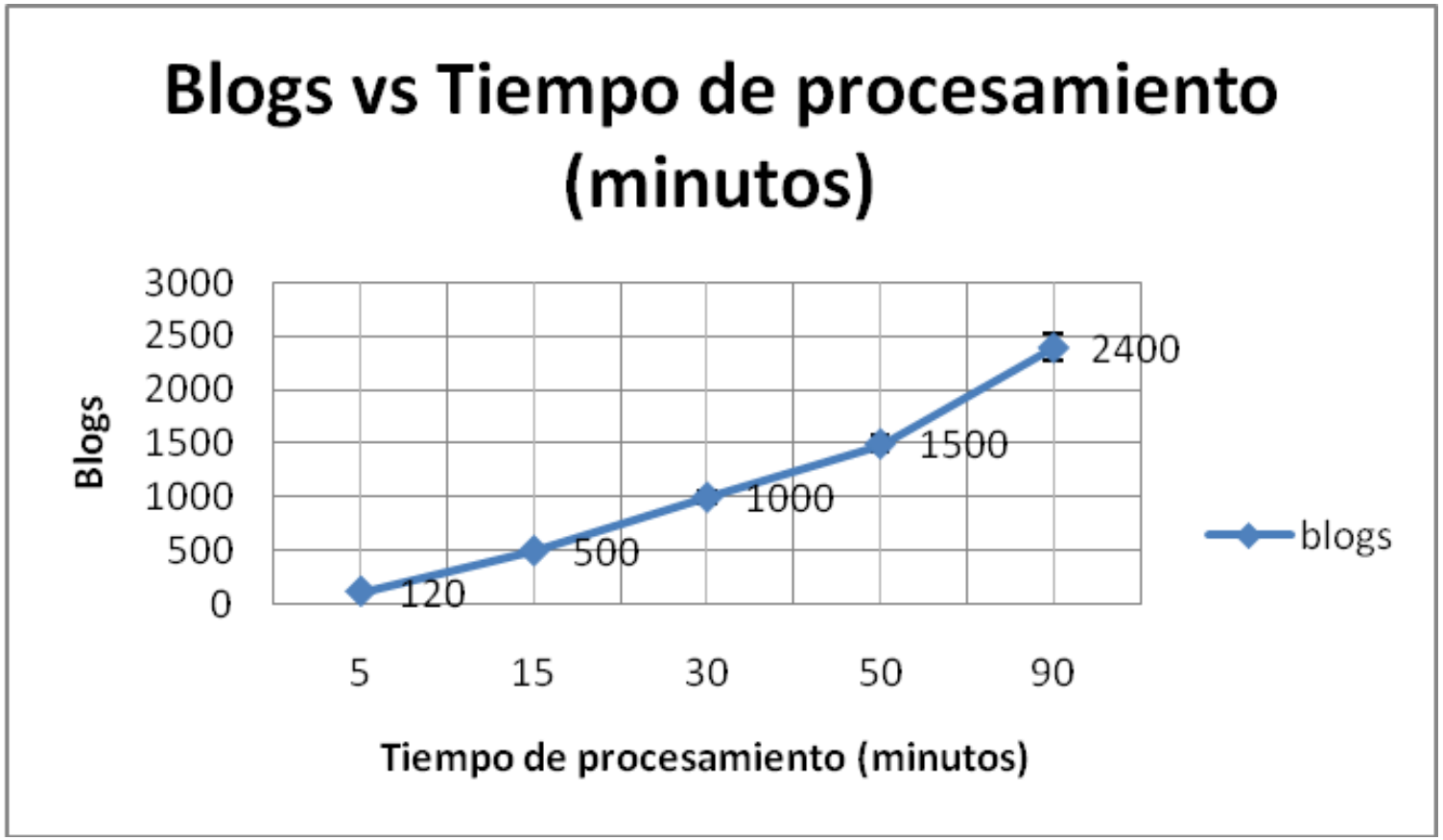
Implementación

- El sistema está compuesto por dos módulos:
  - Agrupamiento de contenidos
  - Indexación y búsqueda de entradas
- Datos a procesar:
  - Entradas de la blogosfera politécnica
  - Extraídas durante los meses de Febrero y Marzo del 2009
  - Gran cantidad de datos a analizar

Pruebas

- Se evaluó el rendimiento de la fase medular del sistema:
  - Agrupamiento de entradas.
- Plataforma de evaluación:
  - Recursos de Amazon Web Services
  - Diez nodos con Hadoop instalado
  - Pequeños grupos de entradas

Resultados


Blogs vs Tiempo de procesamiento (minutos)
Blogs
3000
2500
2000
1500
1000
500
0
120
500
1000
1500
2400
5
15
30
50
90
Tiempo de procesamiento (minutos)
blogs

Conclusiones

- **Tiempo de procesamiento** directamente relacionado al **número de entradas** que componen cada grupo
  - Muestra la escalabilidad lineal de Hadoop.
- Resulta **conveniente implementar** técnicas de **extracción de información** en un ambiente distribuido.
- Servicio **EC2** permite utilizar clústeres a bajo costo en lugar de adquirirlos.

Recomendaciones

- **Agrupamiento de entradas** relacionadas debe continuar realizándose como un proceso en **segundo plano.**
- Previo a la **puesta en producción**, es conveniente **diseñar un plan de extracción de entradas** de la blogosfera politécnica.
- Estudiar el grado de afectación del sistema en la comunidad de autores de blogs.

GRACIAS POR SU ATENCIÓN