Informe de materia de graduación

Evaluación de MapReduce, Pig y Hive, sobre la plataforma Hadoop

Franklin Parrales Bravo

Marco Calle Jaramillo

Jueves, 06 de Mayo de 2010

Contenido

- Herramientas
- Motivación
- Servicios y herramientas usadas
- Esquema
- Resultados
- Conclusiones y Recomendaciones

Herramientas

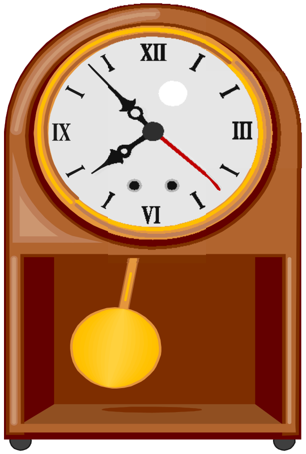
- Hive
- Pig
- Java nativo(Hadoop)

Contenido

- Herramientas
- Motivación
- Servicios y herramientas usadas
- Esquema
- Resultados
- Conclusiones y Recomendaciones

Motivación

¿Cuál de las herramientas anteriormente mencionadas es más adecuada para el procesamiento masivo de datos?



Contenido

- Herramientas
- Motivación
- Servicios y herramientas usadas
- Esquema
- Resultados
- Conclusiones y Recomendaciones

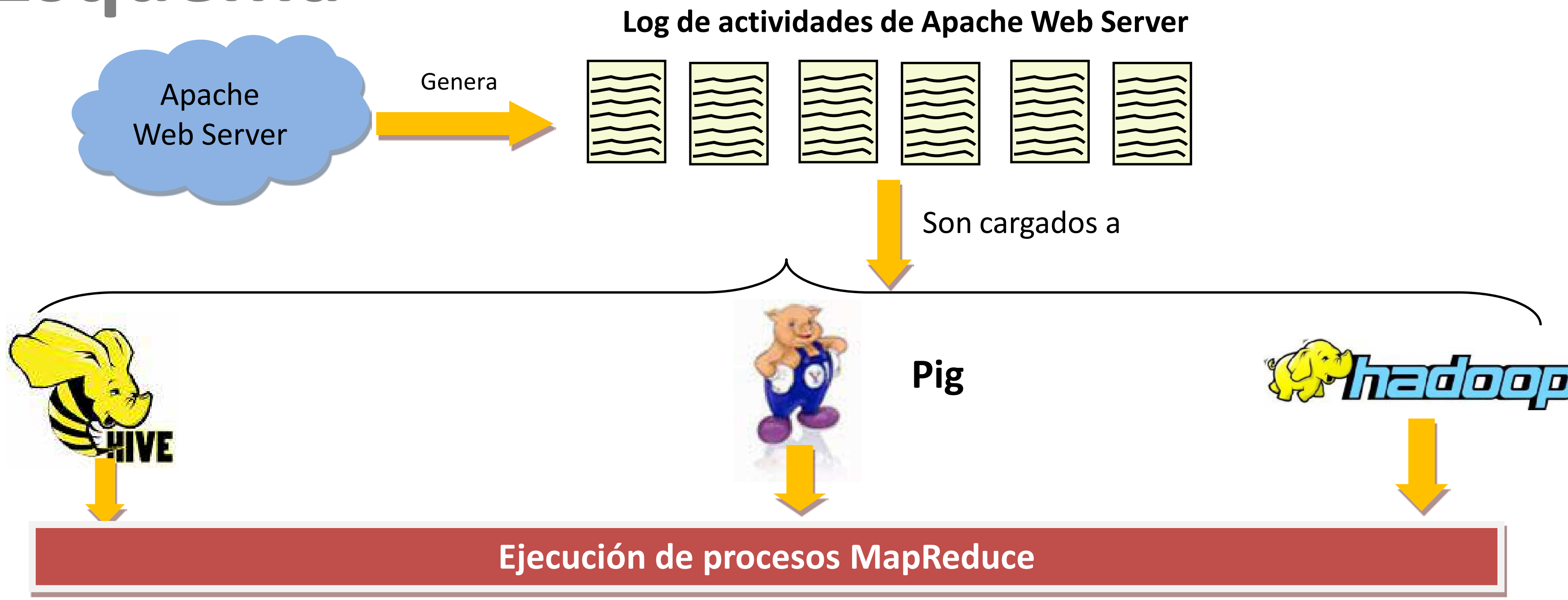
Servicios y herramientas usadas

- Hadoop 0.18, Pig 0.5, Hive 0.4.0.
- Imagen Fedora de Cloudera que nos provee Amazon Machine Image (AMI).
- Elastic Computing Cloud (EC2) y Simple Storage Service (S3) de AWS (Amazon Web Services)

Contenido

- Herramientas
- Motivación
- Servicios y herramientas usadas
- Esquema
- Resultados
- Conclusiones y Recomendaciones

Esquema



Obtención de reportes y evaluación de tiempo de ejecución

1. Conteo de direcciones IP
2. Identificar fechas de eventos de advertencia o error
3. Identificación de la página que más veces no ha sido hallada, y la hora a la que más veces no se la halló

Contenido

- Herramientas
- Motivación
- Servicios y herramientas usadas
- Esquema
- Resultados
- Conclusiones y Recomendaciones

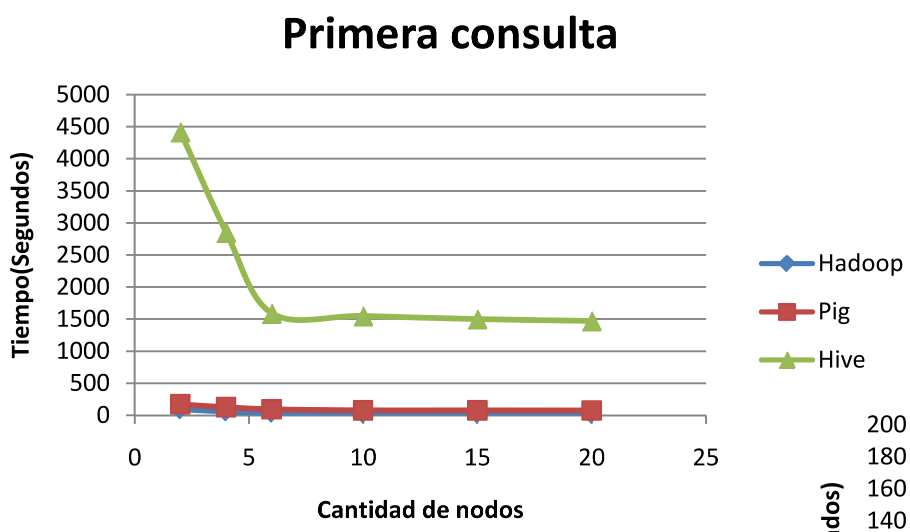
Comparación de las tres herramientas por cantidad de nodos

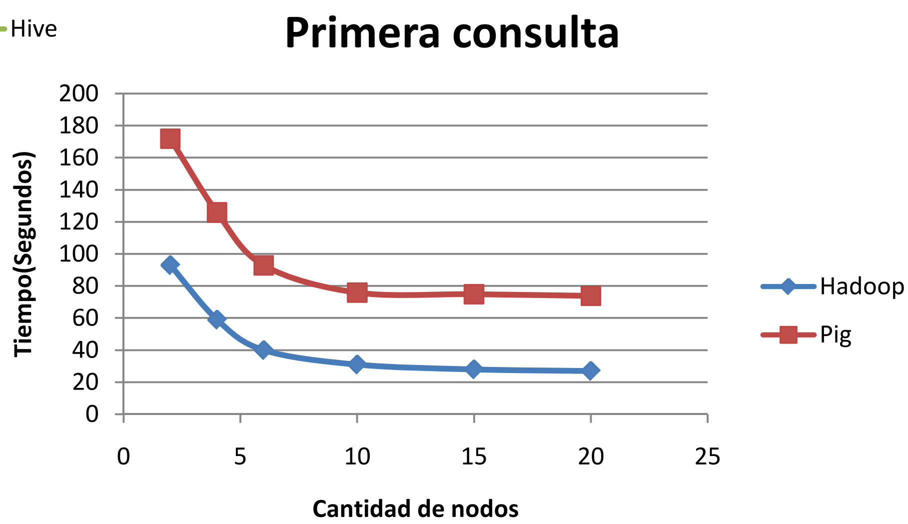
Nodos	2	4	6	10	15	20
Consulta 1	93	59	40	31	28	27
Consulta 2	72	51	36	27	24	21
Consulta 3	71	51	37	28	24	20

Nodos	2	4	6	10	15	20
Consulta 1	172	126	93	76	75	74
Consulta 2	133	74	63	53	52	69
Consulta 3	279	201	171	154	154	142

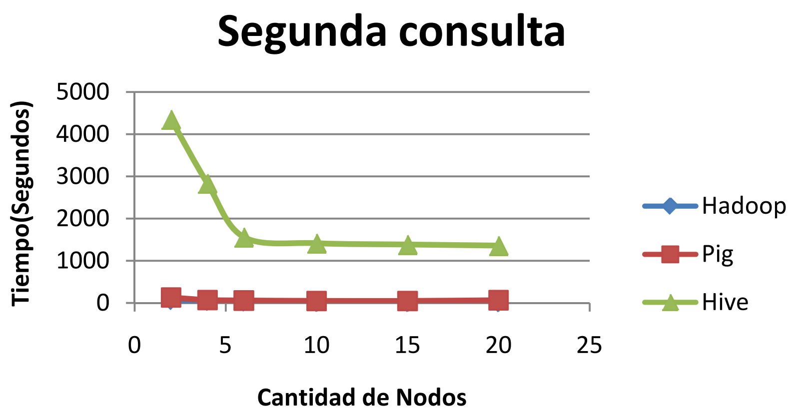
Nodos	2	4	6	10	15	20
Consulta 1	4414.7513	2852.0139	1588.4087	1548.1025	1499.1968	1470.5606
Consulta 2	4352.8812	2834.0337	1561.4006	1415.2823	1384.61	1359.4257
Consulta 3	8898.2346	5776.0148	3117.9123	3087.789	2889.1818	2880.6259

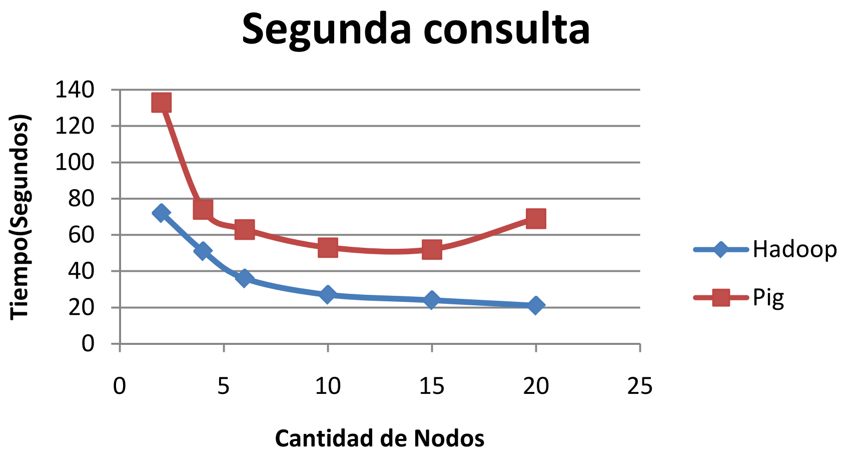
Comparación de herramientas en la primera consulta



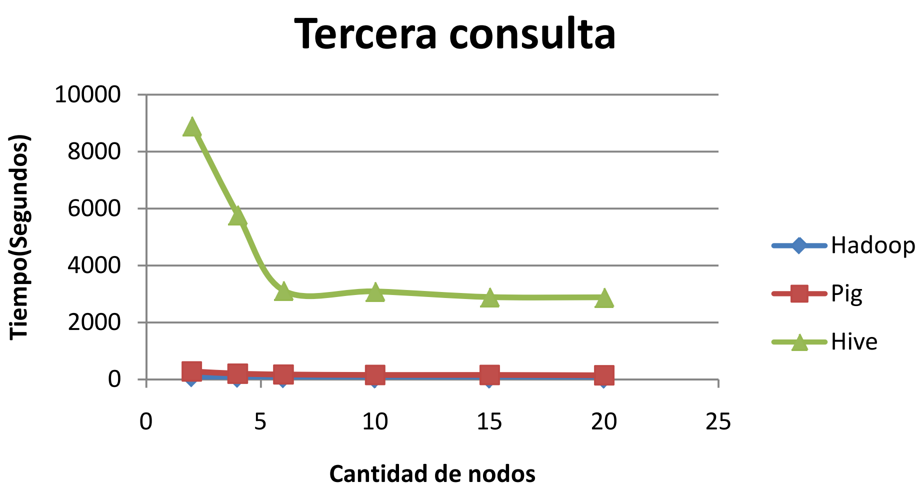


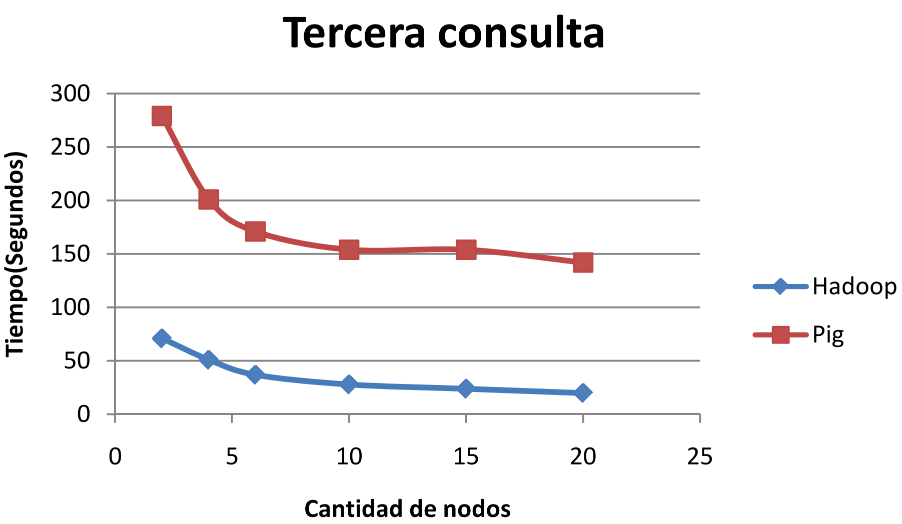
Comparación de herramientas en la segunda consulta





Comparación de herramientas en la tercera consulta





Comparación de las tres herramientas por cantidad de nodos

Hadoop	Correlación	Significancia estadística
Consulta 1	-0.686623	1.39E-06
Consulta 2	-0.7078144	2.55E-07
Consulta 3	-0.7147704	1.41E-07

Pig	Correlación	Significancia estadística
Consulta 1	-0.7864096	9.63E-14
Consulta 2	-0.6801039	2.28E-06
Consulta 3	-0.7843727	1.23E-13

Hive	Correlación	Significancia estadística
Consulta 1	-0.7225381	7.16E-11
Consulta 2	-0.7378843	1.74E-08
Consulta 3	-0.7274289	4.61E-08

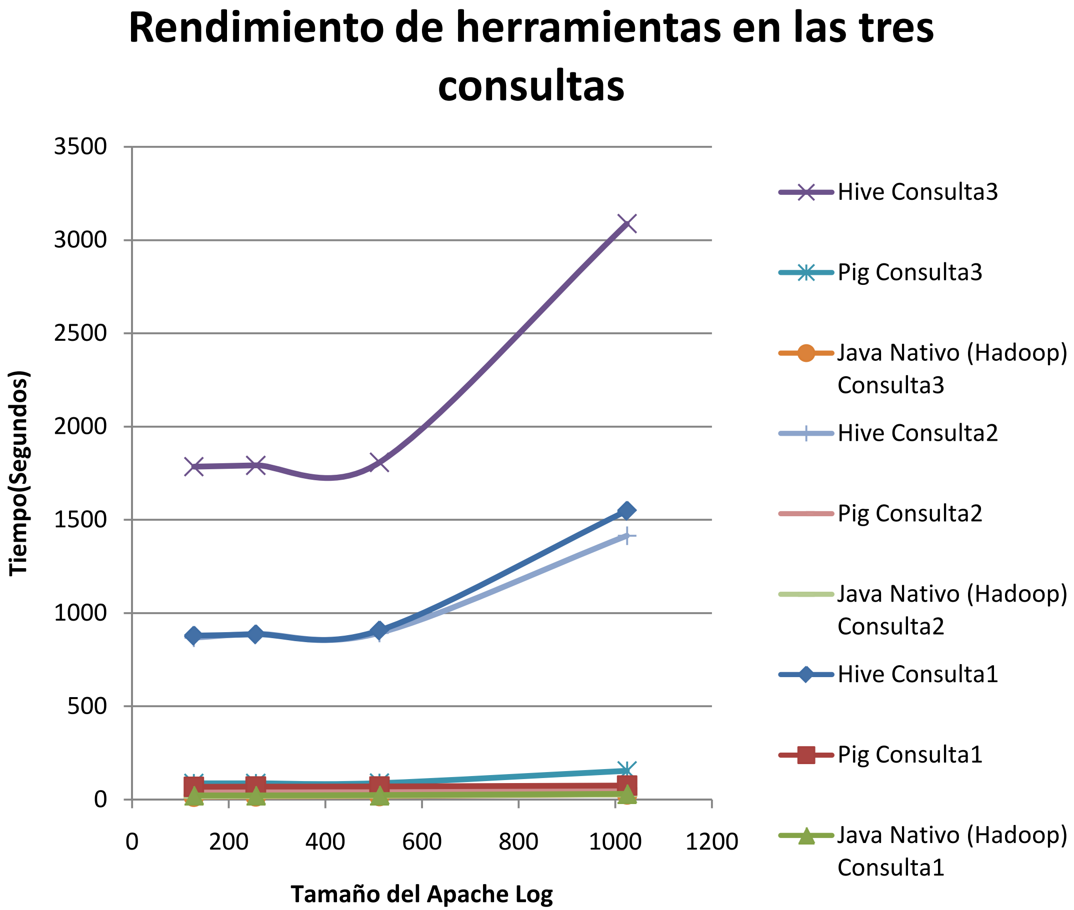
Comparación de las tres herramientas por tamaño del Apache Log

Tamaño del Log (MB)	128	256	512	1024
Hive	879	886	907	1548
Pig	69	70	71	76
Java Nativo (Hadoop)	22	22	24	31

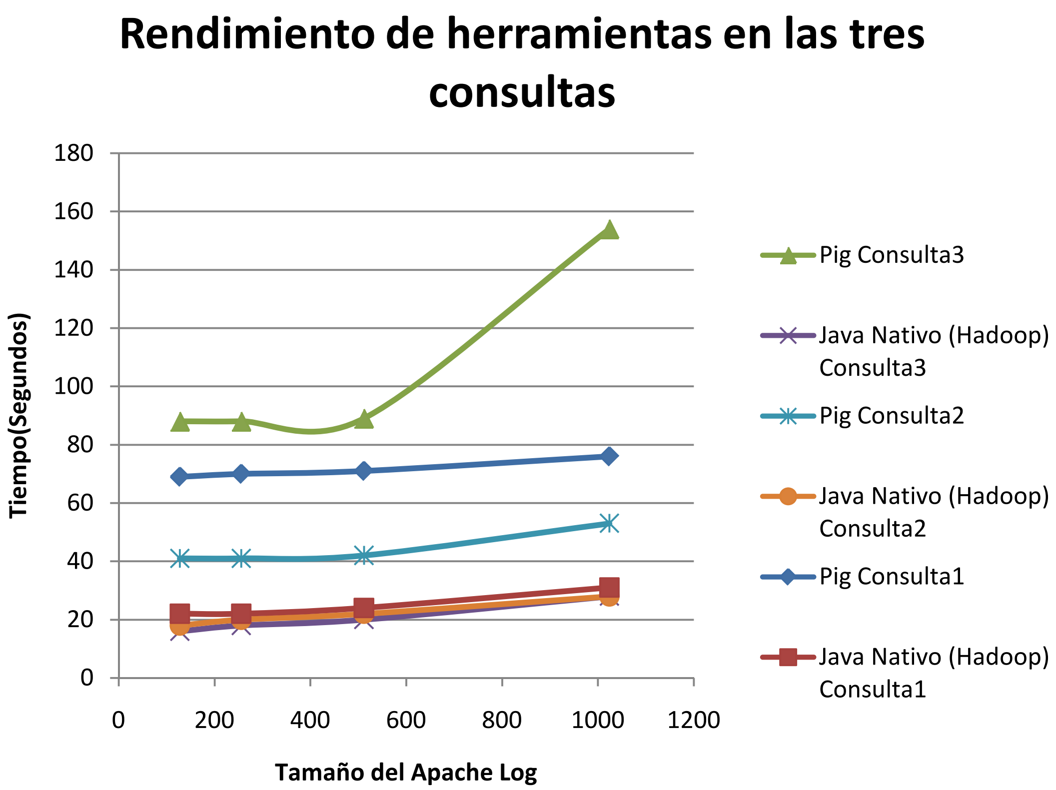
Tamaño del Log (MB)	128	256	512	1024
Hive	868	888	894	1415
Pig	41	41	42	53
Java Nativo (Hadoop)	18	20	22	28

Tamaño del Log (MB)	128	256	512	1024
Hive	1785	1792	1808	3087
Pig	88	88	89	154
Java Nativo (Hadoop)	16	18	20	28

Comparación de las tres herramientas por tamaño del Apache Log



Comparación de las tres herramientas por tamaño del Apache Log



Comparación de las tres herramientas por tamaño del Apache Log

Hadoop	Correlación	Significancia estadística
Consulta 1	0.6093751	0.0003512
Consulta 2	0.843469	4.85E-06
Consulta 3	0.7084398	1.18E-02

Pig	Correlación	Significancia estadística
Consulta 1	0.2567449	0.1708
Consulta 2	0.4447525	1.38E-02
Consulta 3	0.1553797	4.12E-01

Hive	Correlación	Significancia estadística
Consulta 1	0.8654805	6.71E-07
Consulta 2	0.78687	2.51E-04
Consulta 3	0.8690335	4.72E-07

Contenido

- Herramientas
- Motivación
- Servicios y herramientas usadas
- Esquema
- Resultados
- Conclusiones y Recomendaciones

Conclusiones

- Pig es fácil aprender y es rápido.
- Hive usa sentencias parecidas a SQL pero se necesita de mas nodos para que tome menor tiempo.
- Java nativo de Hadoop permite maximizar el excelente uso de recursos, para obtener el resultado más óptimo, pero sacrifica:
  - Facilidad en escritura de código
  - Tiempo empleado en la implementación de la solución

Conclusiones

- Es importante hacer notar que la consulta 3, es mucho más lento en Hive y Pig que las otras consultas, pero no así en Java nativo, ya que en este caso la consulta 1 es ligeramente más lenta que las otras.

Recomendaciones

- A menos que sea de vital importancia el tiempo, es mejor elegir a Pig por
  - Facilidad de código
  - No demora mucho
- Tareas sobre logs del orden de Gigabytes, usar solamente diez nodos dependiendo de la tarea ya que sería un desperdicio de recursos.

Informe de materia de graduación

Evaluación de MapReduce, Pig y Hive, sobre la plataforma Hadoop

Franklin Parrales Bravo

Marco Calle Jaramillo

Jueves, 06 de Mayo de 2010

ANEXOS

Evaluación de MapReduce, Pig y Hive, sobre la plataforma Hadoop

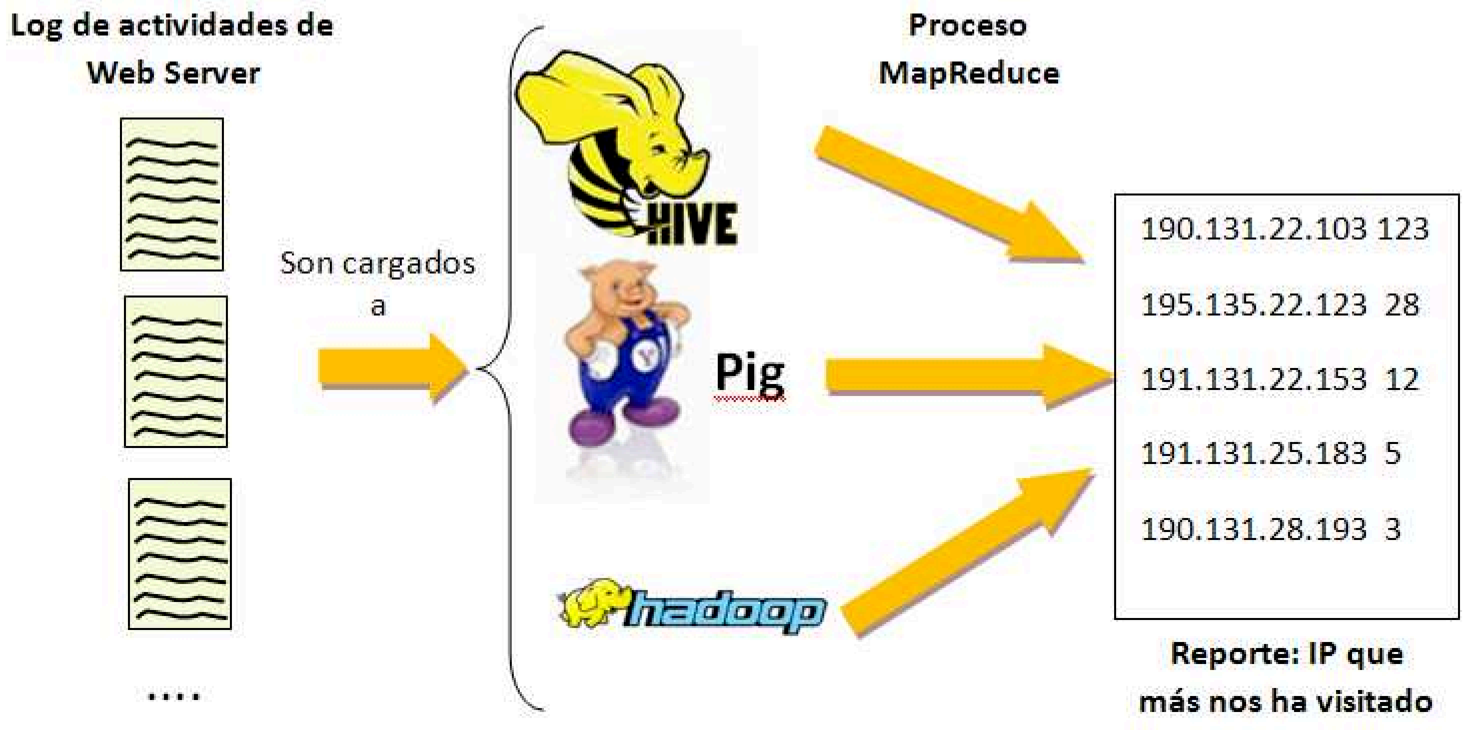
Franklin Parrales Bravo

Marco Calle Jaramillo

Jueves, 06 de Mayo de 2010

Información de la IP que más ha visitado el servidor

Dirección ip	Número de veces
190.131.22.103	123



Información de la hora a la que se han producido la mayor cantidad de errores en el servidor

Hora	Número de veces
08	523

Reporte de la primera parte de la consulta

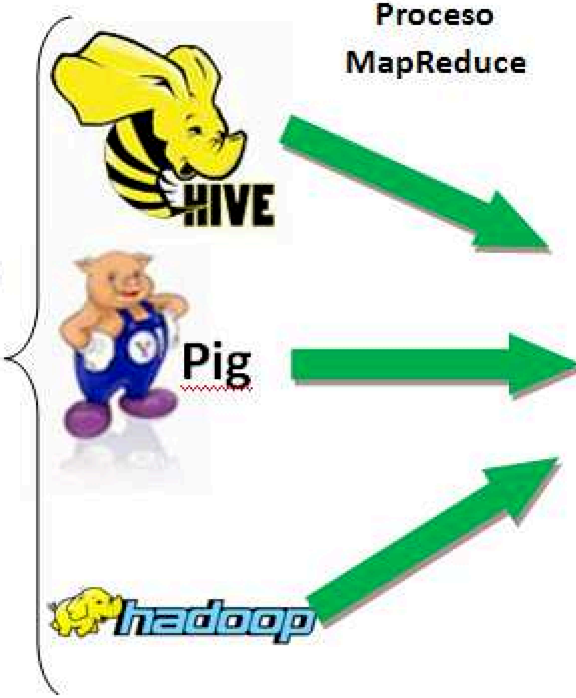
06/Dec/2009:04:13:07 -0500

06/Dec/2009:04:23:17 -0500

06/Dec/2009:04:43:57 -0500

07/Dec/2009:05:56:18 -0500

Es cargado a



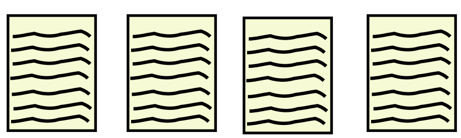
Proceso MapReduce

20	2468
08	3576
16	4464
15	4698
17	4704
13	5031
14	5811
12	6136
09	6543
10	6847
11	7349

Reporte: horas-errores

Información de la página o recurso que más veces ha producido error y a qué hora lo ha generado más veces

Página	Número de veces
/templates/fiec_inicio_template/css/template_css.css	2776

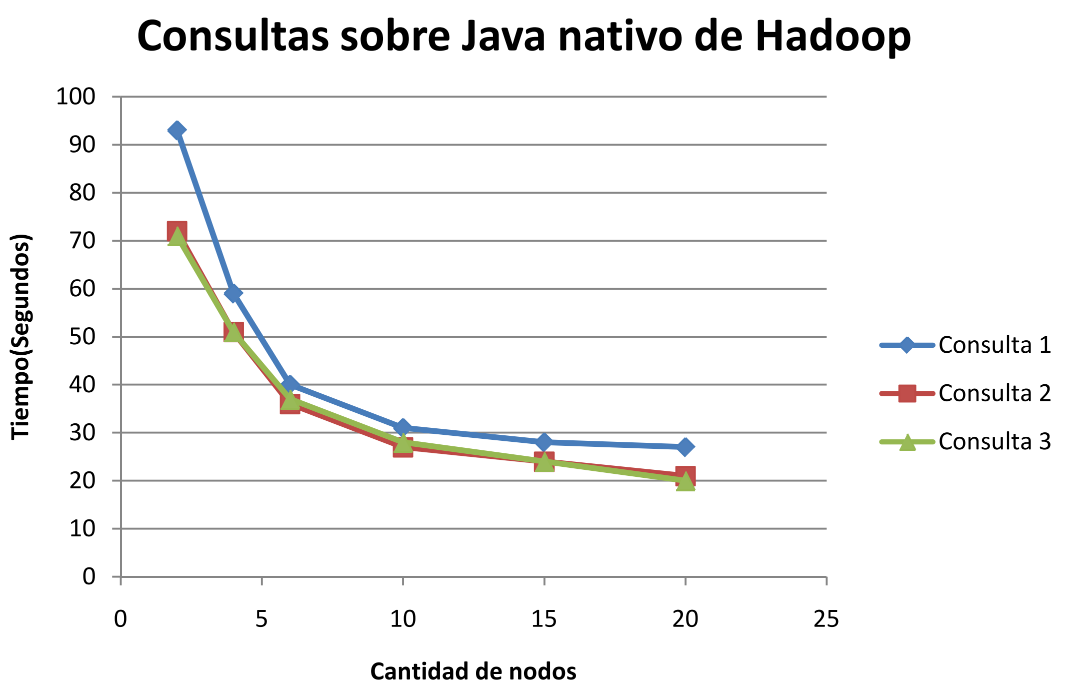


Hora	Número de veces
08	523

Volvemos a consultar los Logs de actividades de Apache Web Server por el recurso

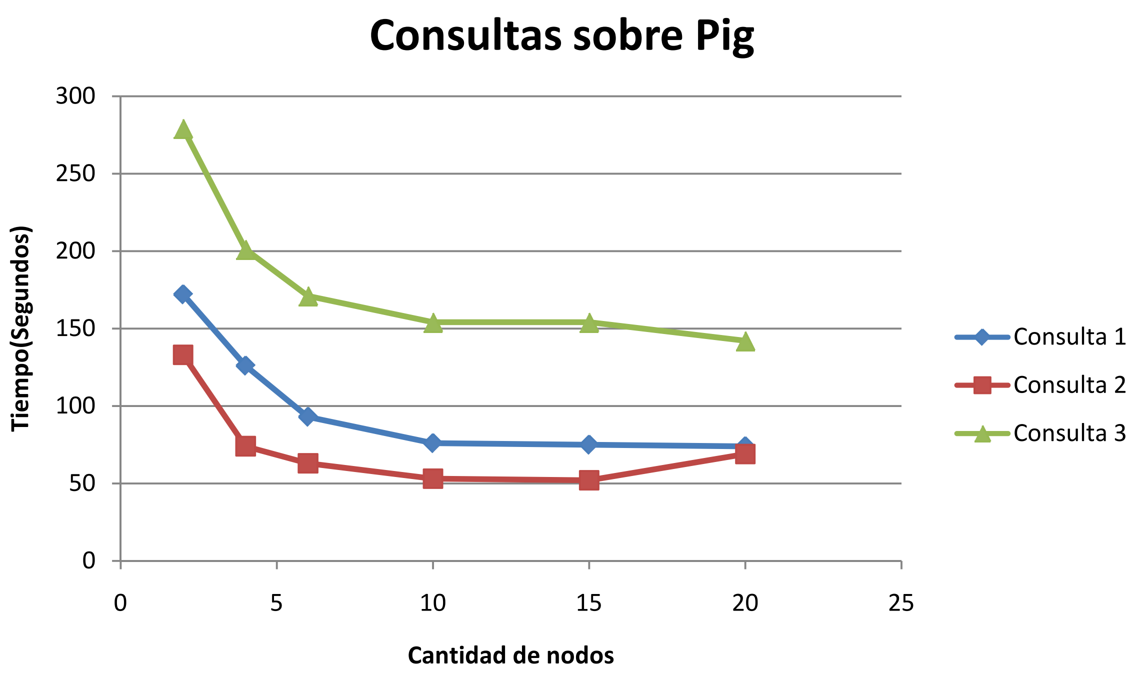
Consultas sobre Java nativo en Hadoop

Nodos	2	4	6	10	15	20
Consulta 1	93	59	40	31	28	27
Consulta 2	72	51	36	27	24	21
Consulta 3	71	51	37	28	24	20



Consultas sobre Pig

Nodos	2	4	6	10	15	20
Consulta 1	172	126	93	76	75	74
Consulta 2	133	74	63	53	52	69
Consulta 3	279	201	171	154	154	142



Consultas sobre Hive

Nodos	2	4	6	10	15	20
Consulta 1	4414.7513	2852.0139	1588.4087	1548.1025	1499.1968	1470.5606
Consulta 2	4352.8812	2834.0337	1561.4006	1415.2823	1384.61	1359.4257
Consulta 3	8898.2346	5776.0148	3117.9123	3087.789	2889.1818	2880.6259