Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

PROYECTO DE ANALÍTICA Y BIG DATA DE LOGS DE SHAREPOINT

Entidad

Departamento Nacional de Planeación

- Dirección de Desarrollo Digital
- Oficina de Tecnologías y Sistemas de Información

Sector	Lenguaje	Fuente de datos
Planeación	R	Logs de SharePoint

Introducción

Se consolidó una base de datos de logs de los portales y sub-portales, con el fin de establecer patrones de comportamiento de las variables analizadas en la base de datos consolidada, para clasificarlos mediante métodos estadísticos y así ddeterminar las causas de las caídas de conectividad en los portales del DNP. Se caracteriza los intervalos (por minuto) que presentan caídas en los portales del DNP.

Objetivo general

Identificar el comportamiento del acceso a los portales y subportales alojados en SharePoint

Metodología

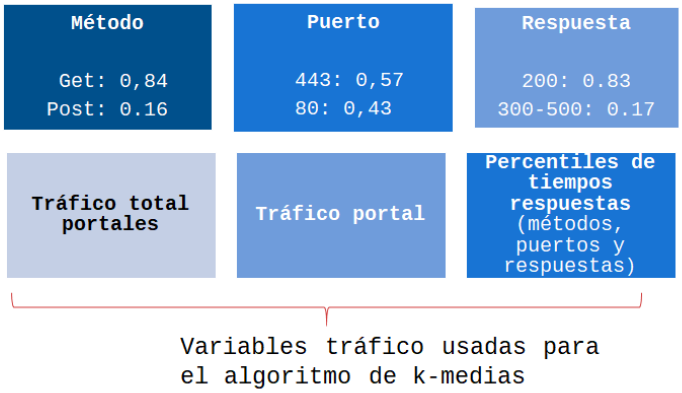
El proceso para desarrollar este análisis implico los siguientes pasos:

- Recolección de los logs por medio de un agente para unificar y consolidar la base de datos.
- Vectorización de los registros (líneas de Logs) indexados.
- Estadística descriptiva y agrupamiento (**cluster**) para descubrir los patrones de comportamiento mediante los componentes de analítica en las variables.
- Visualización de resultados de los patrones de comportamiento.

Se consideran en total 18 portales del DNP (Algunos destacados son el portal DNP, SISBEN, Intranet, entre otros) en los cuales se analiza cada minuto de tráfico. Si se presenta ausencia de tráfico en los siguientes dos minutos se considera una alerta (Variable Y en la ilustración). Esta variable se modela en términos de la información disponible en cada minuto que se recolecta a partir de una información extraída cada milisegundo. Las variables de análisis son el método de conexión (GET, POST), los puertos utilizados (443, 80), el tipo de respuesta (200: petición exitosa, 300: redirección, 400: error del cliente, 500: problema del servidor), el tráfico total en el portal del intervalo, el tráfico total en todos los portales en el intervalo y los tiempos de respuesta.

Rta	Y
Si	1
No	0
No	0
Si	0
Si	0
No	0
Si	1
No	0
No	0

Variable objetivo

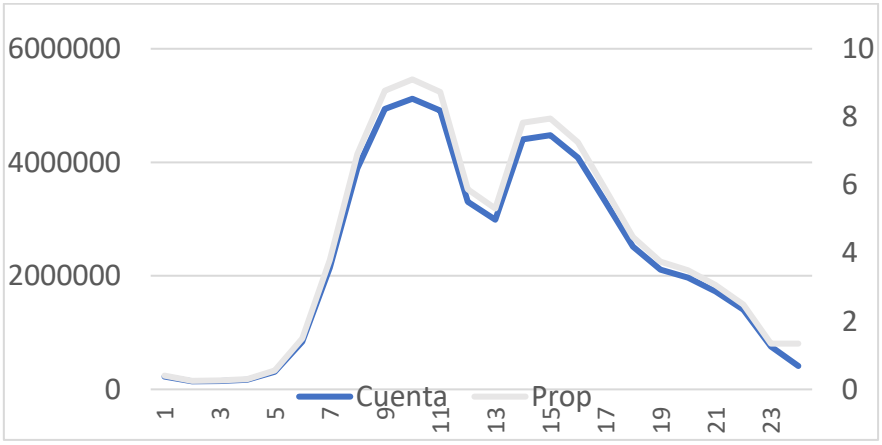


Para llevar a cabo el análisis se realizó el siguiente proceso:

- Lectura y depuración de los logs
- Análisis exploratorio de datos
- Reducción de la dimensionalidad de la información disponible
- Clustering: con el objetivo de identificar grupos de intervalos de tiempo sin tráfico que se debe a problemas técnicos y no a auscencia de tráfico,
- Caracterización

Resultados

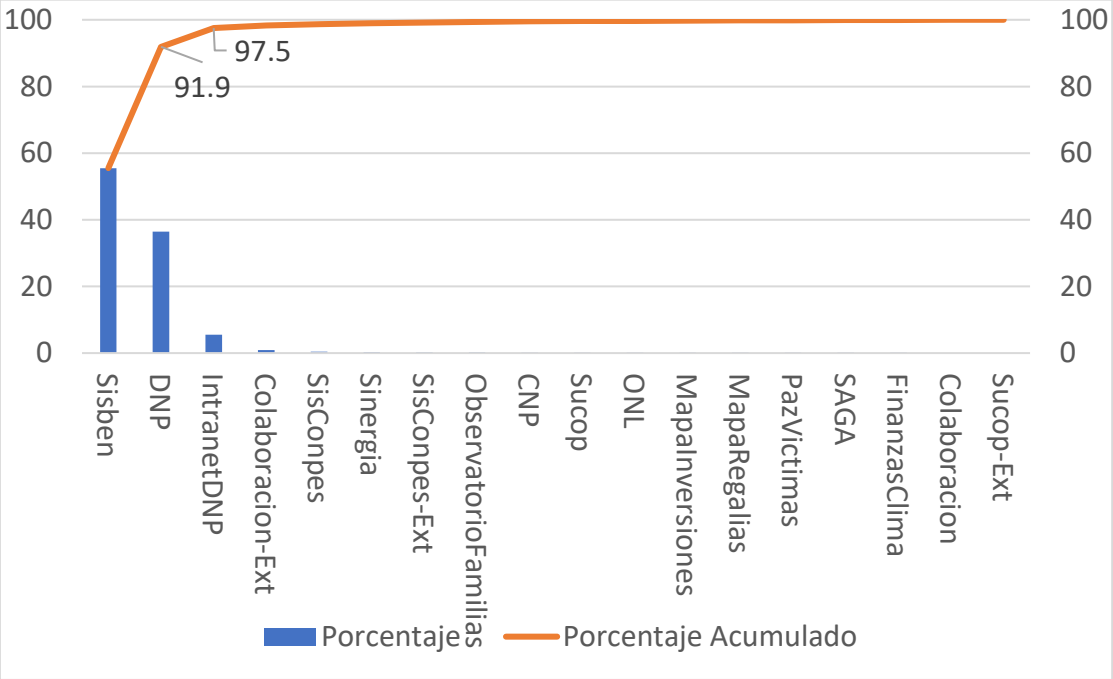
Del análisis descriptivo se destaca el comportamiento del tráfico en los portales por hora:



El tráfico se concentra entre las 7 am y las 7 pm. Los portales tienen el siguiente comportamiento de tráfico:

Portal	Total respuestas	Porcentaje	Porcentaje acumulado
Sisben	31254604	55,5	55,5
DNP	20481974	36,4	91,9
IntranetDNP	3114663	5,5	97,5
Colaboracion-Ext	492592	0,9	98,3
SisConpes	208488	0,4	98,7
Sinergia	158459	0,3	99,0
SisConpes-Ext	113478	0,2	99,2
ObservatorioFamilias	92848	0,2	99,4
CNP	53650	0,1	99,5
Sucop	53361	0,1	99,6
ONL	46586	0,1	99,6
MapaInversiones	42521	0,1	99,7
MapaRegalias	35189	0,1	99,8
PazVictimas	33581	0,1	99,8
SAGA	33581	0,1	99,9
FinanzasClima	32541	0,1	99,9
Colaboracion	18963	0,0	100,0
Sucop-Ext	9806	0,0	100,0

Se destacan los portales del SISBEN, DNP y la Intranet, que concentran cerca del 97,5% de las respuestas.



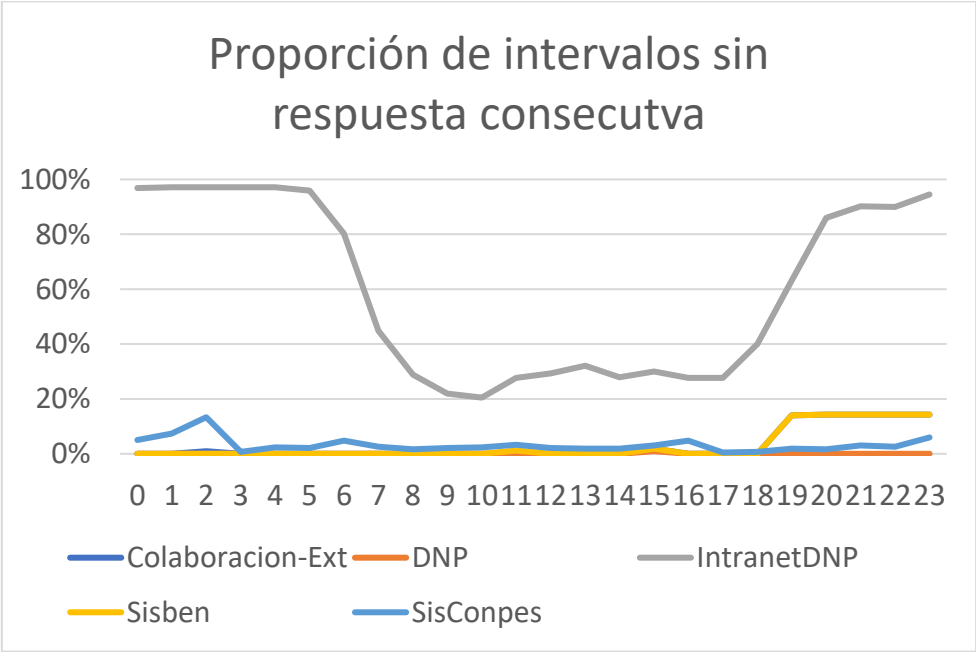
El comportamiento de las caídas consecutivas es el siguiente:

Caídas Sucesivas	Frecuencia	Porcentaje (%)
0	70019	71,2
1	25159	25,6
2	1112	1,1
3	193	0,2
4	122	0,1
5	89	0,1
6	72	0,1
7	56	0,1
8	47	0,0
9	45	0,0
10	39	0,0

Se observa que las caídas de dos minutos o más son una fracción muy pequeña en el tráfico de los portales. Si se realiza este análisis por portal se observa que las caídas se concentran que hay una proporción de caídas en la intranet y en el portal de colaboración (8,8% y 5,9 % respectivamente), se presentan en contraste muy pocas caídas en los portales del DNP y el SISBEN.

Portal	Caídas seguidas Portal	Respuestas Portal	% Caídas por portal
IntranetDNP	307	3483	8,8 (307 / 3483)
Colaboracion	307	5161	5,9
PazVictimas	280	5439	5,1
SAGA	280	5439	5,1
SisConpes	283	5550	5,1
Sucop	277	5648	4,9
SisConpes-Ext	238	6257	3,8
FinanzasClima	169	6359	2,7
ONL	136	6772	2,0
MapaInversiones	126	6803	1,9
Sucop-Ext	78	5964	1,3
MapaRegalias	81	7543	1,1
CNP	42	7231	0,6
Sinergia	41	7352	0,6
ObservatorioFamilias	22	8019	0,3
Colaboracion-Ext	6	9529	0,1
Sisben	4	9753	0,0
DNP	1	10060	0,0

Se logra también identificar que grande parte de las caídas consecutivas se puede confundir por ausencia de tráfico en las franjas por entre 7 de la noche y 7 de la mañana:



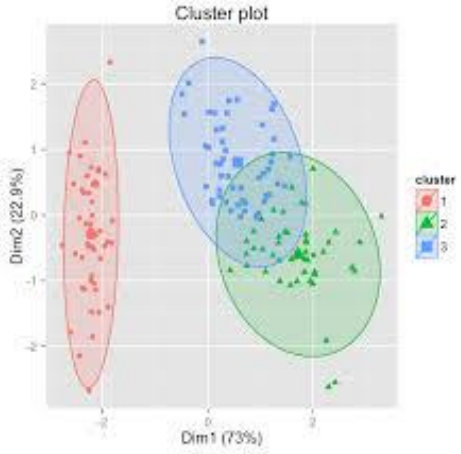
Por otro lado 15 de los 18 portales presentan muy bajo tráfico por minuto, en contraste los portales SISBEN, DNP, Intranet presentan un alto tráfico y las no respuestas que se presenta en 15 de los 18 portales en los próximos 2 minutos se debe con una alta probabilidad a un bajo tráfico. Se presenta también un número considerable de caídas en la Intranet.

Portal	Mediana Tráfico (minuto)	Número de intervalos sin Rta. en siguientes 2 min. (Cluster)	Número de registros sin Rta. en siguientes 2 (no vinculado al cluster)	Número de registros con respuestas en los siguientes 2 minutos
Sisben	1547,4	1	3	5379
DNP	856,4	0	1	5454
IntranetDNP	231	60	117	3125
SisConpes	1	0	111	3076
Colaboracion-Ext	12	0	2	5385
Sinergia	2	0	11	4201
SisConpes-Ext	1	0	72	3758
ObservatorioFamilias	2	0	13	4471
Sucop	2	0	109	3035
CNP	2	0	17	4007
PazVictimas	2	0	114	2954
SAGA	2	0	114	2954
ONL	2	0	44	3768
MapaInversiones	2	0	52	3723
FinanzasClima	2	0	38	4065
MapaRegalias	1	0	74	3458
Colaboracion	2	2	128	2850
Sucop-Ext	1	0	33	3240

Clustering

Para identificar esos intervalos críticos primero se llevó a cabo 100 ejecuciones del algoritmo de clasificación no supervisado *k-medias.* Posteriormente se seleccionan en cada ejecución los intervalos pertenecientes a los clusters que contienen intervalos con una alta proporción de caídas sucesivas (>10%, en contraste con 1,6% de caídas en todos los intervalos). Se conforman tres grupos de intervalos tras 100 corridas del algoritmo *k-medias*:

a. Grupo de Intervalos sin caídas. (Azul)
b. Grupo Objetivo: grupo de intervalos con caídas que en más del 75% de las veces pertenece a clusters asociados a una alta proporción de caídas (Rojo).
c. Grupo de intervalos con caídas que en menos del 75% de las veces pertenece a clusters asociados a una alta proporción de caídas. (Verde)

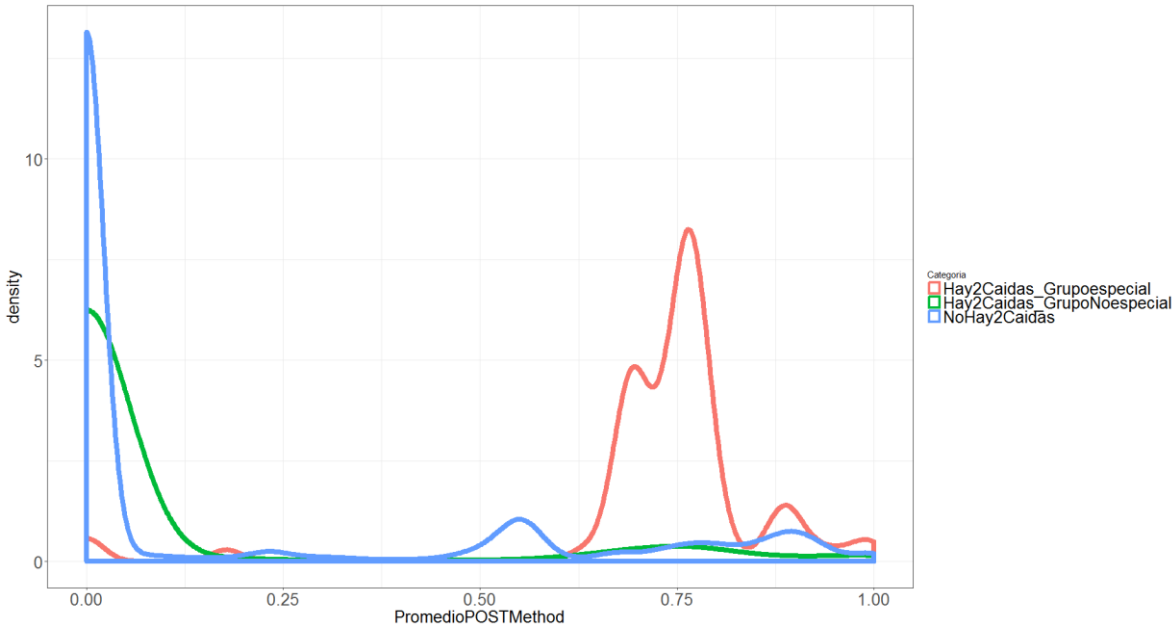


Se observa en el siguiente gráfico que los intervalos de minutos previos a la caída que no presentan tráfico y que pertenecen al *cluster* donde se concentran los problemas técncios (rojo) presenta un tráfico muy similar a los intervalos sin caídas (Azul), los intervalos sin respuesta que no están en el clúster tienen en los minutos previos muy poco tráfico de los 18 portales lo que hce pensar que estos intervalos son la gran mayoría de intervalos sin tráfico.

Los intervalos de minutos previos a la caída que no presentan respuesta que pertenecen al *cluster* especial (rojo) donde se presentan los problemas técnicos presenta en su respectivo portal un tráfico muy similar a los intervalos sin caídas en el tráfico (verde), en contraste, los intervalos sin respuesta que no están en el clúster tienen previamente muy poco tráfico en el portal (estos son la gran mayoría de intervalos sin tráfico)

Al caracterizar los cluster por método se observa:

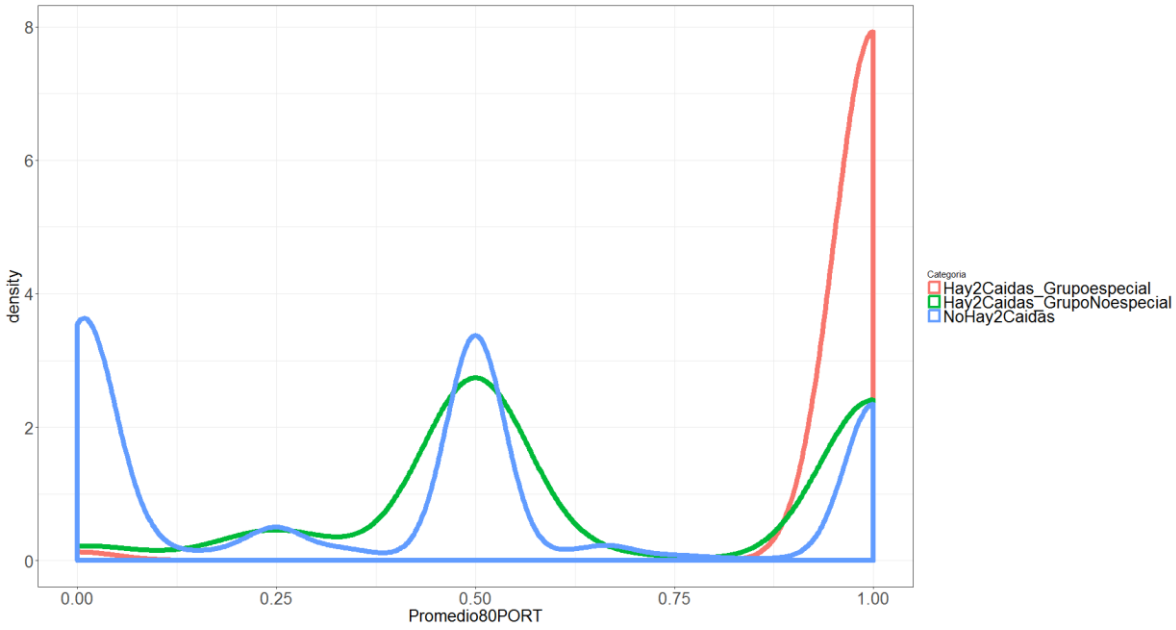
- En los intervalos sin tráfico pertenecientes al *cluster* (asociado a una alta proporción de caídas) el método predominante es **POST** (rojo)
- En contraste, en los intervalos con tráfico (azul) y sin tráfico no perteneciente al *cluster* (verde) predomina el método GET
- Las distribuciones correspondientes a los intervalos sin respuestas (azul) y a los intervalos sin respuesta no pertenecientes al *cluster (verde)* son muy similares



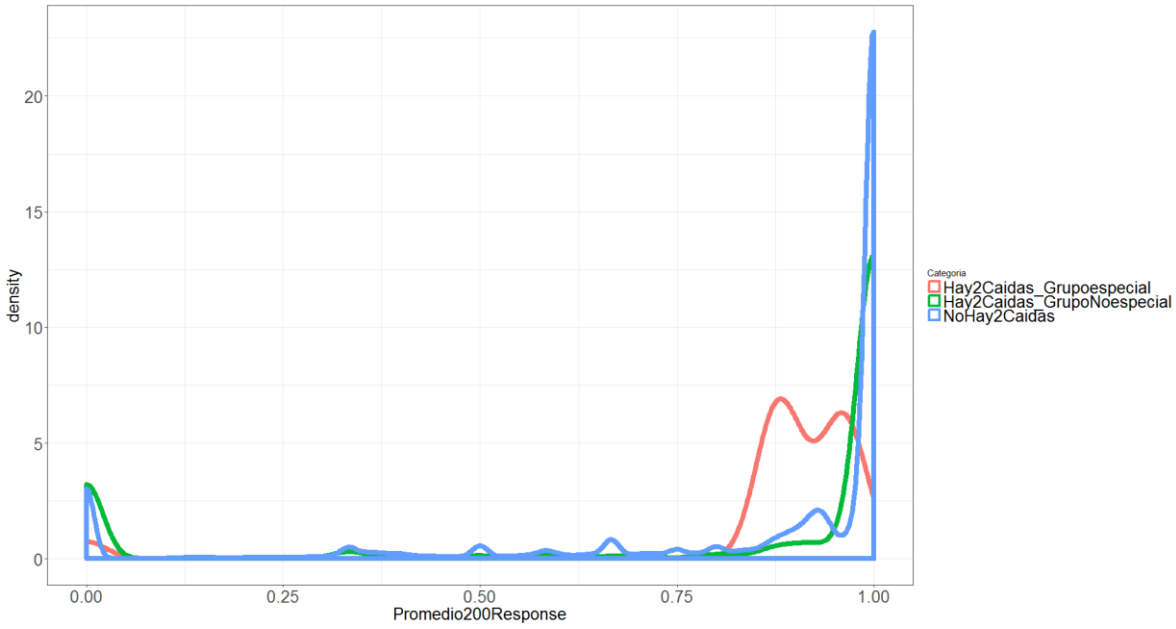
- En los intervalos sin tráfico pertenecientes al *cluster* (asociado a una alta proporción de caídas) el puerto predominante es el 80 (rojo).

- En contraste, en los intervalos con tráfico (azul) y sin tráfico no perteneciente al *cluster* (verde) es balanceado el puerto 80 y el puerto 443.

- Las distribuciones correspondientes a los intervalos sin respuestas (azul) y a los intervalos sin respuesta no pertenecientes al *cluster*



En los intervalos asociados a caídas se presenta una menor respuesta en la respuesta 200 (en detrimento de las respuestas 400 y 500) que en los intervalos que no presentan caídas o presentan no respuestas no asociadas al cluster.



Conclusiones

1. Se identificaron las caídas del sistema que se debieron a problemas técnicos y no a ausencia de tráfico.
2. Las caídas están asociadas a un tráfico previo altamente concentrado en el puerto 80 y en el método Post.
3. Las no respuestas que se dan en un intervalo obedecen en la gran mayoría de los casos a un bajo tráfico en los portales y no a caídas propiamente dichas.
4. Las caídas están mayormente asociadas a las respuestas de tipo 400 y 500.