Preprocesamiento de datos climáticos para el Extractor Rápido de Información Climatológica IV

Yolanda Solís Alvarado, Elder Jaimes Leyva, Javier García Hernández y Jaqueline Lafragua Contreras

Introducción

Durante este 2016 se ha venido desarrollando la cuarta entrega del *Extractor Rápido de Información Climatológica* (ERIC), la cual se divide en dos ediciones: la de escritorio (***ERIC IV versión 4.1***) y la web (***SIG ERIC versión 1.0***). La primera tiene como objetivo agilizar las consultas de manera dinámica mediante un constructor de consultas con manejo de operadores lógicos y operaciones aritméticas básicas, así como la sobreposición de capas cartográficas de contexto como son estados, cuencas, etcétera para la consulta espacial de información climatológica.

La segunda tiene como objetivo permitir al público, en general, tanto la descarga de los registros históricos como la ubicación de las estaciones climatológicas sobre un mapa de *Google Maps*, así como también modificar atributos de las estaciones *online* (por medio de un usuario con permisos). Es decir, hacer posible descargar información de una o varias estaciones climatológicas y descargar capas geográficas en formato *ShapeFile*. Los usuarios que cuenten con usuario y contraseña, además, podrán dar mantenimiento al catálogo de estaciones para corregir cualquiera de sus atributos, pero principalmente las coordenadas para ubicar las estaciones sobre el mapa. Este desarrollo también es conocido como ***ERIC IV para web*** o ***SIG ERIC***.

Para ello, se consideró la base de datos de CLICOM completa y el listado de estaciones que forman la Red Climática de Referencia Nacional (RCRN), misma que es básicamente un subconjunto de CLICOM, más unas cuantas fuentes más. Se analizaron en cuanto a tamaño, formato, etcétera, para determinar la pertinencia de uso de manejador de base de datos o manejador de archivos. La RCRN fue definida a través del Programa Nacional contra Contingencias de Sequía (PRONACOSE) en 2015, y las estaciones que la forman cuentan con registros climatológicos que cumplen con homogeneización y validación de datos estimados.

El primer formato considerado fue *Microsoft Access* debido a la popularidad del mismo y así la mayoría de usuarios podría manipular los datos. Además, al desarrollar rutinas para acceder a la base de datos, se consideró elaborar las llamadas consultas complejas, consultas construidas por el usuario bajo el uso de operadores básicos (+, -, /, *) y lógicos (*and*, *or*, etcétera) en la versión de escritorio. Sin embargo, dado que el universo de datos más grande es de 5 400 estaciones, las consultas complejas resultaban de hasta diez minutos y, en

equipos con menores recursos, llegaron a quedar bloqueadas. Bajo este escenario, se optó por convertir el banco de datos a formato binario en una estructura de datos definida por el equipo de desarrollo y manipulada desde programación C# (Microsoft, 2016) con librerías (COM) diseñadas exprofeso para manipular de manera eficiente este formato.

Derivado de las pruebas de desempeño, se concluyó dejar el formato binario así probado para la versión escritorio y el formato *Microsoft Access* para la versión web, donde únicamente se considera la descarga de archivos, uno por estación, incluyendo todas sus variables medidas y con el beneficio de tener la ubicación sobre servicios de mapa *Google Maps* (Google, 2016) y las capas de los puntos de las estaciones, una para CLICOM y otra para RCRN. La implementación de la interfaz que incluya los beneficios de este análisis está en proceso de desarrollo y validación. Se espera sea liberada durante 2017.

Metodología

La base de datos de origen para la presente versión es el CLICOM y el *Sistema de Información Hidroclimatológica* (SIH), ambos de la Gerencia de Aguas Superficiales e Ingeniería de Ríos (GASIR) de la Comisión Nacional del Agua (CONAGUA). Para las consultas, se dividió la información en dos fuentes de datos: el CLICOM y la Red Climática de Referencia Nacional.

La RCRN es básicamente un subconjunto del CLICOM, más unas fuentes más. La RCRN fue definida a través del PRONACOSE en 2015, y las estaciones que la conforman cuentan con registros climáticos que cumplen con homogeneización y validación de datos estimados. En Lobato (2015), se describe el procedimiento aplicado a cada estación climática para el control de calidad y, con ello, determinar las 1 099 estaciones que resultaron del análisis correspondiente. De este modo, el ***ERIC IV versión 4.1*** para escritorio presenta dos fuentes de datos a consultar: CLICOM (consta de 5 492 estaciones) y RCRN (consta de 1 099 estaciones).

En la figura 1 se observa un esquema del listado de las estaciones de la RCRN, y en la figura 2 un esquema de los conjuntos de estaciones tomados de cada fuente: CLICOM y SIH. Se analizaron ambos conjuntos de estaciones en cuanto a tamaño, formato, velocidad de lectura de datos, velocidad de despliegue de consulta, etcétera, para determinar la pertinencia de uso de un determinado manejador de base de datos, o bien, generar únicamente un manejador de archivos.

Se observó que la estructura de los datos viene en formatos diarios y mensuales con clave de estaciones, claves de variables, año, mes y el día (en caso de datos diarios). Al existir más de un formato de datos fuente, tanto diarios y mensuales de CLICOM y SIH, se tuvo la necesidad de unificar los formatos para que el programa/interfaz prototipo que los use pueda consultar los datos de una forma rápida, utilizando los mismos métodos/rutinas/funciones de extracción de datos.

No.	Fuente	Estación	Latitud	Longitud
1	SPI	1004	22.01	-102.2
2	SPI	1005	21.78	-102.37
3	SPI	1008	22.2	-102.14
4	SPI	1012	21.79	-102.8
5	SPI	1015	21.92	-101.97
6	SPI	1017	22.23	-102.44
7	SPI	1018	22.14	-102.42
8	SPI	1019	22.13	-102.36
9	SPI	1020	22	-102.67
10	SPI	1022	21.74	-102.18
11	SPI	1028	22.14	-102.07
960	Méndez, 2010	29015	19.55	-98.57
961	Méndez, 2010	30022	18.42	-95.11
962	Méndez, 2010	30031	18.37	-95.8
1055	2003	32027	23.1	-104.13
1056	IMTA, 2015	2063	30.4	-115.89
1057	IMTA, 2015	3006	26.38	-112.51
1058	IMTA, 2015	3009	24.84	-111.13
1059	IMTA, 2015	3022	25.05	-111.12
1060	IMTA, 2015	3065	25.09	-111.33
1061	IMTA, 2015	7002	16.38	-92.24
1062	IMTA, 2015	7104	16.31	-91.97
1063	IMTA, 2015	7129	15.7	-93.21
1064	IMTA, 2015	7191	14.92	-92.26
1065	IMTA, 2015	13079	20.07	-98.65

No.	Fuente	Estación	Latitud	Longitud
38	SPI	5026	25.54	-103.47
39	SPI	5029	25.07	-103.28
40	SPI	5030	27.52	-100.62
41	SPI	5031	27.42	-101
42	SPI	5033	27.85	-101.12
43	SPI	5035	25.27	-100.58
44	SPI	5036	25.76	-103
45	SPI	5039	27.28	-103.7
46	SPI	5044	26.99	-102.07
47	SPI	5048	25.43	-101
48	SPI	5086	29.04	-100.96
997	Pieyns, 2003	2060	30.88	-115.63
998	Pieyns, 2003	3002	27.69	-114.9
999	Pieyns, 2003	3031	24.81	-110.98
1090	2015	24009	21.37	-98.66
1091	IMTA, 2015	24013	22.25	-99.67
1092	IMTA, 2015	24016	22.49	-99.43
1093	IMTA, 2015	24028	22.25	-99.12
1094	IMTA, 2015	24092	21.23	-98.89
1095	IMTA, 2015	25009	26.07	-108.78
1096	IMTA, 2015	25033	25.09	-107.39
1097	IMTA, 2015	25045	23.97	-106.61
1098	IMTA, 2015	25049	22.53	-105.45
1099	IMTA, 2015	25115	25.47	-108.08

Figura 1: Red Climática de Referencia Nacional. Fuente: Elaborada a partir de Lobato, 2015.

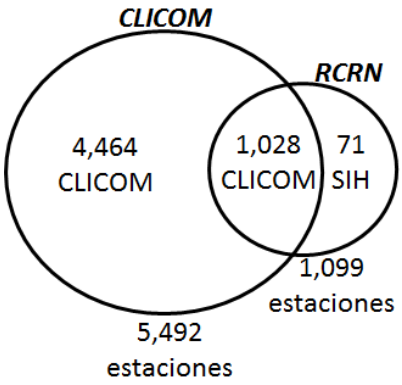

Figura 2: Fuentes que consulta el ERIC IV: CLICOM y RCRN.

El primer formato considerado fue *Microsoft Access* (MDB), debido a la popularidad del mismo y en por lo cual la mayoría de usuarios podría manipular los datos. Sin embargo, se descartó su el uso para la aplicación de consultas complejas con base en dos criterios: su lentitud de acceso y recuperación de datos cuando se extrae para grupos superiores a cincuenta estaciones, por lo que se creó un método/rutina de consulta parecido a una base de datos relacional. Con base en estos criterios, se consideró que el formato MDB es una buena opción de formato de almacenamiento para la versión web.

El formato que se utilizó se dividió en dos subformatos: el primero, un archivo tipo binario por cada variable utilizada donde se guardan bloques de datos por año de tipo numérico con punto flotante, y un segundo archivo de metadato e índices de formato *MDB Access*, en el que se puede consultar sólo la estación, año y las direcciones de índices que apuntan al archivo binario. Este método de acceso a datos se utiliza en bases de datos relacionales para mejorar la velocidad de respuesta de consulta de datos.

La estructura de los archivos binarios aquí diseñada, consiste en una lista incrementable de datos con cortes o separaciones de datos por bloques anuales. Dependiendo la posición en el bloque de datos, corresponde el dato de un día (datos diarios) o al dato de un mes (datos mensuales).

Los datos diarios se encuentran en una estructura uniforme de 31 x 12 (31 columnas y 12 renglones), tal como se observa en la figura 3, incluyendo los datos nulos que se representan con -99999. El bloque de datos cuenta con un índice que apunta al dato inicial de año y otro al dato final, con lo cual, al querer consultar un dato de un año en particular, se obtiene la dirección a la que apunta el año en el archivo binario, y así se localiza el dato preciso.

22.5	22.4	23.8	23	23	22.4	22	17	19.4	22	23.4	22	22.9	23.2	23.4	18.4	22.6	24.4	23	16	20.3	15.4	13.9	18.6	21.4	23	23.4	23.4	25.4	25.3	26.2
23.4	20.4	20	20.2	20.3	19.4	22	21.2	22	19.4	22.4	24.8	27	26.4	23.4	24.8	24.6	25.7	26.4	25.7	24.2	24.4	25.7	25	23.4	26.6	26.6	27	-99999	-99999	-99999
26	24	22.2	22	22.9	25.6	29	25.6	27.2	26.4	26.2	25.4	27.6	24.4	24.4	25.6	26	26	26.4	26.9	25	26.4	24.8	24	24.4	22	24.7	27.1	27.6	25.8	26.4
23.2	25.4	25	25.8	22.4	26.7	28.4	28	29	28.6	28.9	30	32	32	31.8	31	30.1	31.4	32.5	33.5	33.8	32.8	34	33.3	33.4	34.6	35.3	23.4	32.3	32.3	-99999
32	32	31.7	32.8	32.8	34	34	34	31	30.4	32.8	34.2	35	35.4	35	34.2	33	31.4	32.4	32	32.5	29.5	31	31.3	33	34.2	33.2	31.2	32	30	30
32	32.2	31.4	28.6	30.6	31.6	30	29.6	29.6	28.3	28.6	29	29.6	25.6	26.6	24.4	25	24.8	22.4	27	29	25.3	27.6	28.3	30	29.8	29	24	24	26	-99999
23	25.4	26	25.4	24	23.8	24.5	24.4	27.2	27.8	28	28.2	26.4	28.2	29.9	27.6	27.6	29.4	30.2	28	28.6	28.4	23	27	28.2	26.2	26.4	24.2	24.4	28.6	25.4
22.7	23.6	26.8	27.7	29	30	29.1	27	26.9	24	26	24	27.2	27.1	27.5	27.2	26.2	27.9	27.8	23.6	23.8	28	30.1	30.2	29.3	29.4	26.4	29	30.2	30	31.4
30.4	31.4	25.8	27	26	28.4	28	28.7	29	29	29.4	29.9	28.6	28.6	25.4	26.4	24.4	27.3	24.4	25	27.4	28.4	27	27.3	27.6	25	23.4	26.4	26.6	29	-99999
29	27.9	25	24.2	25	25.4	29	28.4	25	27	27.9	27.4	27.7	22.1	22.4	23.9	27.4	28.6	28.2	26.6	27.4	28.6	29.8	29	28	23.8	25	25.8	27	28.8	29
17.9	17.4	18.3	27	26.4	21.4	20	21	24.5	25.5	26.6	26.4	25.4	26	26.7	27.8	23.7	25.6	26	26.4	26	23.4	25	26.4	25.4	25	26.5	26.4	26.9	25.4	-99999
24.2	24	23	23	23.7	23	23	22	22.5	24	23.4	21	23.6	23.4	21	21.6	22.2	23	23	23	23	22	23	16.8	21.6	21.6	21	22	23	23.6	22

Figura 3: Estructura de datos en archivo binario; representa un año de una variable.

El metadato o tabla de índices de cada bloque de datos por año se encuentra en formato MDB, para ser extraídos mediante una consulta SQL (*Structured Query Language*). Cada registro en la tabla de datos es un índice hacia los datos del archivo binario; es decir, hacia el bloque anual. En la figura 4 se observa un ejemplo de registros de estaciones, años e índices que apuntan a un archivo binario de datos.

Para realizar una consulta, es necesaria la manipulación de estos archivos, tanto el archivo binario como el MDB. Con tal fin, se crearon clases en lenguaje C#. Se construyó, también, un método/rutina llamado *consulta_compleja*, el cual concatena las estaciones, variables y operadores, y luego formar un *query* y mostrarlo en una caja texto al usuario. Por último, valida que sea una expresión válida y envía el *query* al método *consulta_MDB*, obteniendo una tabla de resultados.

Metadata

Clave	Anio	Indice_Inicial	Indice_Final
1001	1961	0	1488
1001	1962	1488	2976
1001	1963	2976	4464
1001	1964	4464	5952
1001	1965	5952	7440
1001	1966	7440	8928
1001	1967	8928	10416
1001	1968	10416	11904
1001	1969	11904	13392
1001	1970	13392	14880
1001	1971	14880	16368
1001	1972	16368	17856
1001	1973	17856	19344
1001	1974	19344	20832
1001	1975	20832	22320
1001	1976	22320	23808
1001	1977	23808	25296

Figura 4: Registros de estaciones, años e índices que apuntan al archivo binario de datos.

Cada dato en el archivo binario es de tipo *float*, y ocupa 4 *bytes* de espacio en disco. Cada año de datos ocupa 1 488 *bytes*, lo que equivale a un índice de datos.

Si se leen los archivos binarios, cada 1 488 *bytes* es un bloque de datos anual correspondiente a una estación de la variable que contiene el archivo. Este proceso brinda velocidad a la extracción de información.

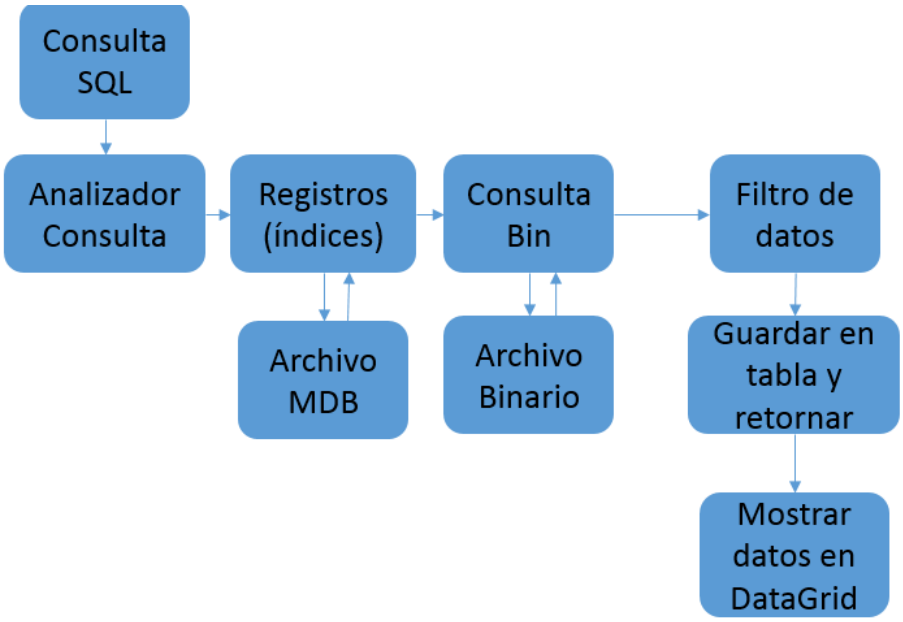

Figura 5: Proceso de consulta de datos.

Los archivos están diseñados para ser actualizados por periodos de datos, dependiendo del último dato agregado, sumando bloques de datos por año al archivo y así sucesivamente, incorporando los índices necesarios a la estación y año.

Por otro lado, el prototipo con el esquema binario descrito es compatible con las herramientas tecnológicas de una interfaz ágil, en el entorno de desarrollo integrado por *Microsoft Visual Studio* 2015, en complemento con el componente *MapWinGIS.ocx* que proporciona la funcionalidad SIG, así como cartografía a cualquier aplicación basada en *Windows Forms*.

Conclusiones

Se analizaron los registros de las 5 492 estaciones climatológicas del CLICOM, en cuanto a velocidad de acceso al disco duro, velocidad de lectura de archivos en diferentes formatos como son *Microsoft Access MDB*, plano de formato CSV, plano *ASCII*, plano .dat y en formato binario con estructura de diseño propio. Todo ello con el fin de que el prototipo del *ERIC IV*, en PC y web, funcione adecuadamente tanto en tiempos de respuesta como en productos que genere, ya sean consultas en pantalla, mapas, gráficas o reportes.

Se concluyó dejar el manejo de formato binario combinado con formato *Access* (sólo manejo de índices que apuntan a bloques del archivo binario) para la versión escritorio, y el formato *Microsoft Access* para la versión web, en la cual únicamente se considera la descarga de archivos, uno por estación, incluyendo todas sus variables medidas, y con el beneficio de tener la ubicación sobre servicios de mapa *Google Maps* (Google, 2016) y las capas de los puntos de las estaciones, una para CLICOM y otra para la Red Climática de Referencia Nacional.

La implementación de la interfaz que incluya los beneficios del presenta análisis está en proceso de desarrollo y se liberará durante 2017.

Referencias

Google, *Google Developers* (Googly Maps API v3), 2016, en: https://developers.google.com/maps/documentation/javascript/reference.

Lobato, S. R., *Propuesta de ampliación de estaciones climatológicas e incorporación de nuevas variables para mejorar la estimación del monitor de sequías*, proyecto de coordinación, seguimiento, supervisión, integración y análisis del Programa de Medidas para Prevenir y Enfrentar la Sequía (etapa 3 de 3), 2015.

Microsoft, *Microsoft Developer Network (Asp .Net),* 2016, en: https://msdn.microsoft.com/en-us/library/System.Web.UI(v=vs.110).aspx.