Análisis de requerimientos

La principal necesidad de la Organización Mundial de la Salud reside en disponer de información respecto a la vacunación de diversas enfermedades en todos los países de la organización. Dicha información permite realizar un seguimiento de los objetivos marcados, así como utilizarla para analizar la situación y tomar medidas al respecto con el objetivo de salvar el mayor número de vidas posibles.

Para desarrollar un proyecto según el contexto planteado, previamente se has de determinar los requisitos:

1. Conocer el grado de consecución de los objetivos en el compromiso definido en el GVAP.
2. Conocer la cobertura de vacunación por parte de los países de la OMS.
3. Conocer la cobertura sobre las vacunas DTP.
4. Conocer el porcentaje da vacunación del sarampión en las 5 regiones de la OMS con el objetivo de comprobar el cumplimiento del objetivo 2 de las metas impuestas para 2015 y 2020.
5. Conocer la evolución histórica para la vacunación desde que se tienen datos, así como mostrar dicha evolución de manera visual y entendible.
6. Facilitar el acceso a de acorde a las necesidades de los usuarios.
7. Los requisitos deben poder observarse desde diferentes perspectivas:
   1. Región
   2. País.
   3. Enfermedad.
   4. Porcentaje de vacunación.
   5. Grupos de vacunas.

Teniendo en cuenta estos requisitos el sistema deberá ser capaz de responder a las siguientes preguntas:

* Cobertura de inmunización que alcanzan los países miembros a lo largo del tiempo.
* Países que han logrado cumplir el objetivo 1 de 2015.
* Regiones que han cumplido el objetivo 2 de 2015.
* Evolución histórica de cualquier vacuna o grupo de vacunas.
* Ranking de países por grado de vacunación de cualquier enfermedad.
* Top países con cobertura contra el sarampión.
* Top países con cobertura contra las enfermedades DTP.

ANÁLISIS DE FUENTES DE DATOS

**coverage\_estimates\_series.xls:** series cronológicas de coberturas de inmunización por región de la OMS, Estados miembros y vacunas.

* Formato: CSV
* Primera línea con etiquetas de los campos.
* Separador de campos: Punto y coma (;)
* Campos de Texto: Entre comillas (“)

|  |  |  |
| --- | --- | --- |
| Nombre de etiqueta | Tipo | Ejemplo |
| WHO\_REGION | Texto | EMR |
| ISO\_code | Texto | AFG |
| Cname | Texto | Afghanistan |
| Continent | Texto | “Asia” |
| Vaccine | Texto | “BGC” |
| Year | Numérico | 1980 |
| Percent\_coverage | Numérico | 33 |
| Asterisc | Texto | \* |

Datos relevantes:

1. No todos los países informan acerca de todas las vacunas.
2. El carácter asterisco (“\*”) en el campo Asterisc significa una cobertura del 100%. Lo más común es que esté en blanco.
3. No hay información de los países en los mismos años.

Total de registros: 56.646

**wuenic2018rev\_data\_2019-11-16.csv:** datos de la encuesta de cobertura de inmunización, número de niños vacunados y población objetivo por región de la OMS, Estados miembros y vacunas. Años 1997-2018.

* Formato: CSV
* Primera línea con etiquetas de los campos.
* Separador de campos: Punto y coma (;)
* Campos de Texto: Entre comillas (“)

|  |  |  |
| --- | --- | --- |
| Nombre de etiqueta | Tipo | Ejemplo |
| Group | Texto | “WHO Regions” |
| Subgroup | Texto | “EMR” |
| Name | Texto | “Afghanistan” |
| Year | Numérico | 2018 |
| Vaccine | Texto | “BGC” |
| Coverage | Numérico | 73 |
| Vaccinate | Numérico | 941,000 |
| Target | Numérico | 1,207,000 |
| Source | Texto | WHO/UNICEF estimates of national immunization coverage, 2018 revision |

Datos relevantes:

1. No todos los países informan acerca de todas las vacunas.
2. No hay información de algunos países en todos los años.
3. No todos los países tienen el mismo número de registros-

Total de registros: 35365

**Ficheros.xml:**  datos sobre casos de sarampión (measles) de los países miembros de la región de Europa. Años 2008-2018:

1. Sarampion\_5001.xml: número de casos detectados.
2. Sarampion\_5002.xml: número de muertes.
3. Sarampion\_5003.xml: número de hospitalizaciones
4. Sarampion\_5005.xml: número de casos confirmados en laboratorio.

* Formato: xml
* Etiquetas de los campos. <Etiqueta> </ Etiqueta >
* Separador de campos: <RegistroS> </RegistroS>

**Estructura común:**

|  |  |  |
| --- | --- | --- |
| Nombre de etiqueta | Tipo | Ejemplo |
| cod | Numérico | 2 |
| nombre | Texto | Albania |
| a2008 | Numérico | 12 |
| a2009 | Numérico | 1367 |
| a2010 | Numérico | 4 |
| a2011 | Numérico | 7 |
| a2012 | Numérico | 12 |
| a2013 | Numérico | 27 |
| a2014 | Numérico | 1 |
| a2005 | Numérico | 154 |
| a2016 | Numérico | 13 |
| a2017 | Numérico | 38 |
| a2018 | Numérico | 20 |

Datos relevantes:

1. Es frecuente que no haya datos de los países en algunos años. Algunos incluso están vacíos, como es el caso de Mónaco.
2. Todos los registros tienen el campo con y nombre con datos.
3. Los países de todos los ficheros son los mismos y están en el mismo orden.
4. En el fichero de Sarampion\_5002.xml la mayoría de los registros están vacíos

Total de registros: 220 en cada fichero. En total: 880.

**REGION.json:** contiene información de las regiones de la OMS.

* Formato: json.

|  |  |  |
| --- | --- | --- |
| Nombre de etiqueta | Tipo | Ejemplo |
| Código de Región | Texto | “AMR” |
| Nombre | Texto | “Americas” |

Total de registros: 58.

**COUNRTY.json:** contiene información acerca de los países miembros de la OMS.

* Formato: json.

|  |  |  |
| --- | --- | --- |
| Nombre de etiqueta | Tipo | Ejemplo |
| Código de país | Texto | “BEL” |
| Nombre de país | Texto | “Belgium” |
| Nombre de atributo | Texto | “WORLD\_BANK\_INCOME\_GROUP\_RELEASE\_DATE” |
| Atributo | Texto | “2017” |

Total de registros: 247.

Datos relevantes:

1. El número de atributos por país no es el mismo. Por ejemplo, Bélgica tiene 23 y Gibraltar solo 6.
2. Algunos países como china tienen varios registros, asociados a algunas de sus regiones/provincias.

**MORTCAUSE.json:** contiene información acerca de los códigos de las enfermedades.

* Formato: json.

|  |  |  |
| --- | --- | --- |
| Nombre de etiqueta | Tipo | Ejemplo |
| Código de enfermedad | Texto | “0040” |
| Nombre de enfermedad | Texto | “Tetanus” |

Total de registros: 78.

**Estimación de volumetría:**

|  |  |  |
| --- | --- | --- |
| Fuente de datos | Valores por almacenar | Total de registros |
| coverage\_estimates\_series   * 1 fichero anual * 194 países miembros (6 regiones OMS) * 45 vacunas | Años: 53 datos | 1 fichero  × 194 países  × 45 vacunas  × 53 datos =  462.690 |
| wuenic2018rev\_data\_2019-11-16   * 1 fichero anual * 194 países miembros (6 regiones OMS) * 45 vacunas | Años: 22 datos | 1 fichero  × 194 países  × 45 vacunas  × 53 datos =  192.060 |
| Sarampion\_5001.xml |  | 220 |
| Sarampion\_5002.xml |  | 220 |
| Sarampion\_5003.xml |  | 220 |
| Sarampion\_5005.xml |  | 220 |
| REGION.json |  | 58 |
| COUNTRY.json |  | 247 |
| MORTCAUSE.json |  | 78 |
|  | **TOTAL** | 656013 |

Análisis funcional.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Requerimiento | Prioridad | Exigible/deseable |
| 1 | Se extraerá de manera adecuada la información de las fuentes de datos (se considerará solo la información relevante). | 1 | E |
| 2 | Se creará un almacén de datos. | 1 | E |
| 3 | Se cargará en el sistema toda la información proporcionada | 1 | E |
| 4 | Se creará un modelo multidimensional OLAP para las consultas de todos los usuarios, será posible elegir dimensiones como la región, país, vacuna, año… | 1 | E |
| 5 | Se garantizará el acceso a todos los usuarios mencionados en el apartado de los requisitos, siempre con los permisos adecuados. | 1 | E |
| 6 | Se crearán los informes estadísticos para analizar la evolución de la vacunación de cualquier enfermedad. | 2 | E |
| 7 | Se habilitarán gráficas que muestren de manera intuitiva los datos a los usuarios. | 3 | D |
| 8 | Se redactará un manual de carga de datos incremental | 3 | D |
| 9 | Serán creadas una serie de alertas en el sistema warehouse para determinadas situaciones en las que peligren los objetivos de la OMS: | 3 | D |

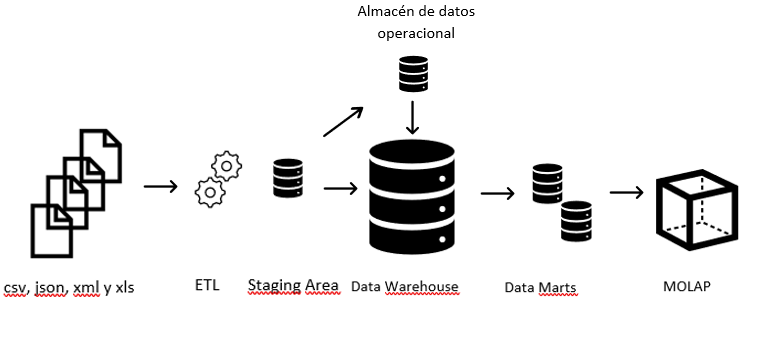
A la hora de definir la arquitectura funcional debemos tener en consideración los siguientes elementos:

1. Los datos proporcionados son ficheros csv, json, xls y xml en los que se proporcionan tanto metadatos como información histórica sobre el grado de vacunación de los países de la OMS. También poseemos una extensa información sobre la evolución de la enfermedad del sarampión en dichos países desde 2008 a 2018.

Tanto la variedad de archivos como la información que proporcionan determinan los elementos restantes de la arquitectura funcional de la FIC a desarrollar.

1. La arquitectura queda compuesta por:
   1. Staging Area: a pesar de ser opcional, el staging área proporciona una ventaja relevante en el caso de cargas de datos heterogéneas. Este elemento permitirá reducir el impacto en las cargas de los modelos ETL homogeneizándolos.
   2. Procesos ETL: serán indispensables para la carga de los datos, actuarán en sintonía con el staging area.
   3. Almacén de datos o Data Warehouse: dado que nos encontramos ante información no volátil e historificada, el almacén de datos es un elemento indispensable que actuará como núcleo de la FIC.
   4. Almacén departamental o Data Mart: con el fin de aligerar la carga sobre el almacén de datos y de aligerar el tiempo de respuesta del mismo se implementarán almacenes de datos departamentales de acuerdo a las características de los usuarios.
   5. Almacén de datos operacional: servirá como medio a los analistas para realizar informes estadísticos.
   6. MOLAP: a partir de la información de los datamart se crearán cubos multidimensionales para realizar las consultas.

Representación gráfica:



Cabe destacar que tanto el almacén de datos operacional como la staging área son elemento que no se podrían extraer de la arquitectura sin resultar fatídico para el funcionamiento de la FIC, no obstante, son útiles y mejoran notablemente su rendimiento.

IV. Diseño del modelo conceptual, lógico y físico del almacén de datos

Para el correcto desarrollo del almacén de datos es preciso definir los hechos (facts), las dimensiones de análisis (dimensions), las métricas y los atributos que nos permitan dar respuesta a las preguntas que se han definido en el análisis de requerimientos.

**Diseño conceptual**

Del análisis de la fuente de datos coverage\_estimates\_series, que contiene datos sobre inmunización de cada año, se determina que uno de los hechos que hay que analizar es la cobertura de inmunización.

Teniendo en cuenta los requerimientos identificados, se analizará el hecho y la cobertura de inmunización para resolver la necesidad de los usuarios de analizar la cobertura de inmunización a lo largo del tiempo.

El análisis de la cobertura de inmunización determina el diseño de la siguiente tabla de hechos:

|  |  |
| --- | --- |
| Tabla de hechos | Descripción |
| fact\_cobertura | Datos de cobertura de inmunización |

Una de las métricas de la tabla de hechos fact\_cobertura es la que se incluye en la siguiente tabla:

|  |  |
| --- | --- |
| Métricas | Descripción |
| Cobertura | Cobertura de inmunización |

Las métricas que se identifiquen para la tabla de hechos fact\_cobertura podrán ser analizadas desde las diferentes perspectivas por medio de las siguientes dimensiones:

|  |  |
| --- | --- |
| Dimensiones | Descripción |
| Geografía | Distribución geográfica de la OMS a la que corresponde la información de la cobertura de inmunización. |
| Vacunas | Vacunas incluidas en el calendario de vacunación. |
| Temporalidad | Periodo temporal de vacunación. |

A partir de las dimensiones y de la tabla de hechos identificados se construye el modelo conceptual, en el que tanto las dimensiones como los hechos son entidades independientes que forman parte de nuestro modelo de estrella.

El diseño conceptual con forma de estrella para la tabla de hechos y las dimensiones identificadas es:

DIM\_VACUNAS

FACT\_COBERTURA

DIM\_GEOGRAFÍA

DIM\_TEMPORALIDAD

Del archivo coverage\_estimates\_series también se determina que otro hecho a analizar es la comparativa entre países y regiones de la cobertura de vacunación. La métrica de este hecho es la misma que en el caso de fact\_cobertura.

|  |  |
| --- | --- |
| Tabla de hechos | Descripción |
| fact\_cobertura\_comparativa\_país | Datos de cobertura de inmunización por país |

|  |  |
| --- | --- |
| Dimensiones | Descripción |
| País | Información de la cobertura de inmunización por país. |
| Vacunas | Vacunas incluidas en el calendario de vacunación. |
| Temporalidad | Periodo temporal de vacunación. |

FACT\_COBERTURA\_

PAÍS

DIM\_TEMPORALIDAD

DIM\_PAÍS

DIM\_VACUNAS

|  |  |
| --- | --- |
| Tabla de hechos | Descripción |
| fact\_cobertura\_comparativa\_región | Datos de cobertura de inmunización por región |

|  |  |
| --- | --- |
| Dimensiones | Descripción |
| Región | Información de la cobertura de inmunización por región. |
| Vacunas | Vacunas incluidas en el calendario de vacunación. |
| Temporalidad | Periodo temporal de vacunación. |

DIM\_VACUNAS

DIM\_PAÍS

DIM\_TEMPORALIDAD

FACT\_COBERTURA\_

REGIÓN

Los archivos sarampión\_500\*.xml nos permiten realizar un estudio más detallado de la enfermedad en concreto, con ellos podemos determinar el siguiente hecho.

|  |  |
| --- | --- |
| Tabla de hechos | Descripción |
| fact\_cobertura\_sarampión | Datos de casos detectados de sarampión, muertes confirmadas, casos de hospitalizaciones y casos confirmados en laboratorio. |

|  |  |
| --- | --- |
| Métricas | Descripción |
| Casos | Casos detectados de sarampión. |
| Muertes | Número de muertes. |
| Hospitalizaciones | Número de hospitalizaciones. |
| Casos laboratorio | Número de casos confirmados en laboratorio. |

Todos los hechos tienen las mismas dimensiones:

|  |  |
| --- | --- |
| Dimensiones | Descripción |
| País. | Información del hecho por país. |
| Año. | Año que se desea observar del hecho. |

Los diseños conceptuales con forma de estrella para las tablas de hechos y las dimensiones identificadas son:

DIM\_AÑO

FACT\_COBERTURA\_

SARAMPIÓN

DIM\_PAÍS

Del análisis de la fuente de datos wuenic2018rev\_data\_2019-11-16.csv, que contiene datos sobre inmunización en la población infantil y objetivo, se determina el hecho de cobertura de vacunación.

|  |  |
| --- | --- |
| Tabla de hechos | Descripción |
| fact\_cobertura\_infantil\_objetivo | Datos de cobertura de inmunización en la población infantil y objetivo. |

|  |  |
| --- | --- |
| Métricas | Descripción |
| Vacunados | Números de vacunados. |
| Target | Número de población objetivo a vacunar. |

Las métricas que se identifiquen para la tabla de hechos fact\_cobertura\_infantil\_objetivo podrán ser analizadas desde las diferentes perspectivas por medio de las siguientes dimensiones.

|  |  |
| --- | --- |
| Dimensiones | Descripción |
| País. | Información del hecho por país. |
| Año. | Año que se desea observar del hecho. |

El diseño conceptual con forma de estrella para la tabla de hechos y las dimensiones identificadas es:

DIM\_PAÍS

FACT\_COBERTURA\_

INFANTIL\_OBJETIVO

DIM\_AÑO

**Diseño lógico**

|  |  |
| --- | --- |
| **Tabla de Hechos** | **Métricas** |
| FACT\_COBERTURA | DATO\_COBERTURA |
| FACT\_COBERTURA\_PAÍS | DATO\_COBERTURA\_PAÍS |
| FACT\_COBERTURA\_REGÍON | DATO\_COBERTURA\_REGIÓN |
| FACT\_COBERTURA\_SARAMPION | SARAMPION\_CASOS |
|  | SARAMPION\_MUERTES |
|  | SARAMPION\_HOSP |
|  | SARAMPION\_CASOS\_LAB |
| FACT\_COBERTURA\_INFANTIL\_OBJETIVO | DATO\_ VACUNAS |
|  | DATO\_TARGET |

En la siguiente tabla, se muestran los atributos descriptores con las referencias a sus dimensiones de la tabla de hechos FACT\_COBERTURA:

|  |  |
| --- | --- |
| **Dimensiones** | **Atributos descriptores** |
| Geografía | COD\_GEO |
| Vacunas | COD\_VAC |
| Temporalidad | COD\_TEMP |

El diseño lógico propuesto para este hecho es:

DATO\_COBERTURA

COD\_GEO

COD\_TEMP

COD\_VAC

DIM\_GEOGRAFÍA

FACT\_COBERTURA

DIM\_VACUNAS

DIM\_TEMPORALIDAD

En la siguiente tabla, se muestran los atributos descriptores con las referencias a sus dimensiones de la tabla de hechos FACT\_COBERTURA\_PAÍS:

|  |  |
| --- | --- |
| **Dimensiones** | **Atributos descriptores** |
| País | COD\_PAIS |
| Vacunas | COD\_VAC |
| Temporalidad | COD\_TEMP |

El diseño lógico para este modelo es:

DIM\_PAIS

DIM\_VACUNAS

FACT\_COBERTURA\_PAIS

COD\_VAC

COD\_TEMP

DIM\_TEMPORALIDAD

DATO\_COBERTURA\_PAIS

COD\_PAIS

En la siguiente tabla, se muestran los atributos descriptores con las referencias a sus dimensiones de la tabla de hechos FACT\_COBERTURA\_REGION:

|  |  |
| --- | --- |
| **Dimensiones** | **Atributos descriptores** |
| Región | COD\_REGION |
| Vacunas | COD\_VAC |
| Temporalidad | COD\_TEMP |

El diseño lógico para este modelo es:

FACT\_COBERTURA\_REGION

COD\_VAC

COD\_REGION

COD\_TEMP

DIM\_TEMPORALIDAD

DIM\_REGION

DIM\_VACUNAS

DATO\_COBERTURA\_REGION

En la siguiente tabla, se muestran los atributos descriptores con las referencias a sus dimensiones de la tabla de hechos FACT\_COBERTURA\_SARAMPIÓN:

|  |  |
| --- | --- |
| **Dimensiones** | **Atributos descriptores** |
| País | COD\_PAIS |
| Año | COD\_AÑO |

SARAMPION\_CASOS, SARAMPION\_MUERTES, SARAMPION\_HOSP, SARAMPION\_CASOS\_HOSP

COD\_PAIS

COD\_AÑO

DIM\_PAÍS

FACT\_COBERTURA\_

SARAMPIÓN

DIM\_AÑO

En la siguiente tabla, se muestran los atributos descriptores con las referencias a sus dimensiones de la tabla de hechos FACT\_COBERTURA\_INFANTIL\_OBJETIVO:

|  |  |
| --- | --- |
| **Dimensiones** | **Atributos descriptores** |
| País | COD\_PAIS |
| Año | COD\_AÑO |

DATOS\_VACUNAS, DATO\_TARGET

COD\_AÑO

COD\_PAIS

DIM\_PAÍS

FACT\_COBERTURA\_

INFANTIL\_OBJETIVO

DIM\_AÑO