**Requerimentos para el proyecto Fondef IDeA I+D ID19I10198:** “*Hacia la modernización del Estado: Modelado analítico de capacidades estatales basado en datos de dotación de personal*”**.**

**Introducción**

El proyecto Fondef ID19I10198 utiliza datos de dotación de personal del Estado Central[[1]](#footnote-1) desde el 2006 hasta la fecha, descargados desde los distintos Portales de Transparencia. Estos datos son procesados, limpiados y validados, al mismo tiempo que se procede a desambiguar las identidades del personal con alta precisión. Posteriormente, los datos son tratados en sus distintas dimensiones mediante indicadores/métricas, son mostrados con ayuda de tableros interactivos y visualizaciones, y son usados para construir modelos que permiten analizar/describir la evolución y composición del personal del Estado de Chile.

Los tableros construyen una capa interactiva que proporciona la posibilidad de explorar los datos de dotación de personal en las distintas dependencias de los Ministerios, teniendo en cuenta las dimensiones de Dotación, Remuneración, Longevidad entre otras y utilizando además distintos contextos caracterizados por Profesión, Estamento, Grado, Sexo, Región, etc. Numerosas preguntas pueden responderse con estos datos. A modo de ejemplo: ¿Cuál es la composición de cada servicio de acuerdo al género, grado y régimen de contratación?, ¿Cuántos funcionarios trabajan en el Estado, en sus distintas dependencias y en regiones?, ¿Cuál es la proporción de profesionales contratados en cada servicio?, ¿Hay diferencias relevantes en la permanencia de los funcionarios entre los distintos servicios?, ¿Cuánto difieren los servicios en sus niveles salariales?, ¿Cómo han cambiado los salarios a lo largo del tiempo y también en regiones?, entre otros muchos temas.

Las ventajas que ofrecen los productos finales asociados a este proyecto respecto a otras fuentes de datos son múltiples, pero fundamentalmente destacan el alto grado de resolución, disponibilidad y alcance de la data[[2]](#footnote-2).

La sostenibilidad, escalabilidad y operatividad de este proyecto radican fundamentalmente en asegurar un conjunto de actividades y recursos que den garantías para que estas ventajas continúen siendo viables.

**Las actividades y recursos/servicios requeridos.**

De manera muy resumida, las actividades del proyecto están relacionadas con los siguientes puntos:

1. Descarga/ingesta automática de los datos e integración (también automática) a un Data Lake.
2. Procesamiento/limpieza automatizada de los datos.
3. La construcción de tablas con indicadores y métricas a partir de la data cruda.
4. El suministro de datos a los tableros y visualizaciones.
5. La actualización periódica y automatizada de la data.
6. El modelado basado en datos.

Para realizar estas actividades, este proyecto utiliza un conjunto de servicios y herramientas tecnológicas contenidas en la nube Amazon Web Services[[3]](#footnote-3) (AWS). Estas herramientas permiten almacenar los datos en forma de *Data Lake*, para posteriormente interaccionar con ellos mediante varias vías. La arquitectura de nube permite operar sin interrupción y, a diferencia de un servicio on-premise, no requiere mantenimiento o actualización del software. Además, provee de recursos y servicios ad-hoc para generar reportes y visualizaciones basados en los datos de manera automática. Dichos recursos son usados en los tableros, en el sitio Web del proyecto[[4]](#footnote-4) y en los modelos. Es por ello que, para la sostenibilidad/continuidad del proyecto es fundamental garantizar la operación y gestión de la cuenta de AWS (lo que se hace actualmente dentro del proyecto). Para este fin, se requiere como mínimo:

1. Contar con personal técnico adecuado.
2. Destinar los recursos económicos necesarios para el mantenimiento de la cuenta AWS.
3. Incluir las actividades necesarias, dentro de los procesos internos de nuestra contraparte ( que también recibirá el proyecto), que permitan operar en la nube para garantizar la actualización periódica de los productos/resultados del proyecto.

Por otro lado, los tableros interactivos con visualización de indicadores basados en datos y elaborados en el proyecto están creados con Microsoft Power BI[[5]](#footnote-5). Estos tableros se “alimentan” directamente de las tablas generadas en la nube a partir de los datos que son capturados y procesados con distintos servicios de AWS (estos servicios se detallan más adelante). Para conservar actualizados los tableros se requiere mantener una cuenta profesional de Power BI, como parte de las exigencias para garantizar la sustentabilidad del proyecto.

**El perfil y requerimientos del personal técnico que recibe el proyecto.**

Subrayamos aquí, en base a las actividades que deberán realizarse para la continuidad del proyecto, el perfil del personal técnico que lo recibirá. El equipo técnico deberá contar con al menos un profesional con conocimientos de *cloud computing*, en particular de AWS (por ejemplo: profesional del área informática, o sistemas de información, o analista programador u otro relacionado). Se destaca AWS, debido a que la migración del proyecto a una plataforma *cloud* diferente puede resultar complejo. Entre los servicios de AWS que deberá manejar el equipo técnico están: *S3*, *GLUE*, *Athena*, *Sagemaker*, *EC2*, *IAM* y *Billing*.

Adicionalmente, los profesionales a cargo de recibir y darle continuidad al proyecto deberán manejar los lenguajes de programación utilizados: SQL, Python, R y Spark. Tendrán que estar familiarizados con los procesos de limpieza, estructuración e ingesta de datos, conocer las librerías básicas asociadas al manejo de datos y de expresiones regulares. Deberán también manejar Docker y Power BI. Por último, se requiere estar capacitado para la administración de usuarios, permisos, roles y políticas en la nube mediante el servicio AWS - IAM. Esto es importante para controlar el acceso a los datos y también a las máquinas virtuales generadas para el trabajo con los procesos de carga y la comunicación entre los distintos servicios de AWS que utiliza el proyecto.

**Los procesos y pipeline que deben asegurarse.**

Si bien la mayoría de las operaciones con los datos actualmente ya están automatizadas (dentro de un pipeline) en el proyecto, se requiere una intervención semi-manual en algunas etapas del flujo de trabajo. En particular, en la actualización y carga de los datos que provienen de distintas fuentes. Estas fuentes son independientes del proyecto y pueden generar algunas modificaciones propias.

La fuente principal de los datos utilizados en el proyecto es Transparencia Activa del Consejo para la Transparencia (CPLT). Por ley 20285 de transparencia, los datos de dotación de personal estarán disponibles. Sin embargo, y no obstante que con el CPTL hemos establecido algunos acuerdos para la descarga y disponibilidad de los datos con ciertos formatos y procedimientos previamente definidos, no existe en la actualidad un proyecto formalizado en este sentido y puede que eventualmente en el futuro se realicen cambios que no hayan sido considerados en el proyecto hoy. Por esta razón, y a menos que en el futuro cercano se establezcan procedimientos modernos y bien definidos para disponibilizar los datos de Transparencia Activa, se requerirá inevitablemente de intervención manual como una tarea periódica. Eso sí, este proyecto garantiza que estas intervenciones estén optimizadas y sean las mínimas posibles. De hecho, si no hay modificaciones en la entrada de los datos entonces, las tareas de limpieza, validación, construcción de indicadores y generación de reportes en tableros y sitio Web (actualmente diseñadas/implementadas) se ejecutarán con una mínima supervisión. Si hay modificaciones en el origen (ej. cambios en: nombre de las variables, enlaces de descarga, nueva parcelación de los datos, etc.) entonces, las lógicas de captura y procesamiento de datos deberán adaptarse directamente en la nube.

Al día de hoy, el CPTL nos proporciona un listado de enlaces de descarga con una nomenclatura definida previamente. Asociados a estos enlaces están los datos parcelados por Ministerio/Servicio y por Régimen de contratación (Contrata, Honorarios, Planta y Código del trabajo). En base a esto, hemos generado códigos que permiten la descarga de los datos, el filtrado por el año actual y el almacenamiento en AWS-S3. Estos códigos están alojados en un *Docker* que corre todas las semanas, debido a que el CPTL actualiza la data semanalmente. De manera secuencial, se ejecutan otros procesos que corren también una vez a la semana para dar estructura a la data, realizar la limpieza requerida (en particular de las fechas y las remuneraciones) y asegurar la ingesta de los datos en una base de datos que denominamos “*struct*”. Desde aquí, se le otorga/desambigua la identificación a cada funcionario. Posteriormente, se construyen las tablas con indicadores y se dejan disponibles para su uso en los tableros de Power BI. Esto es a grandes rasgos el pipeline que hay que mantener para que los resultados del proyecto se actualicen con nueva data.

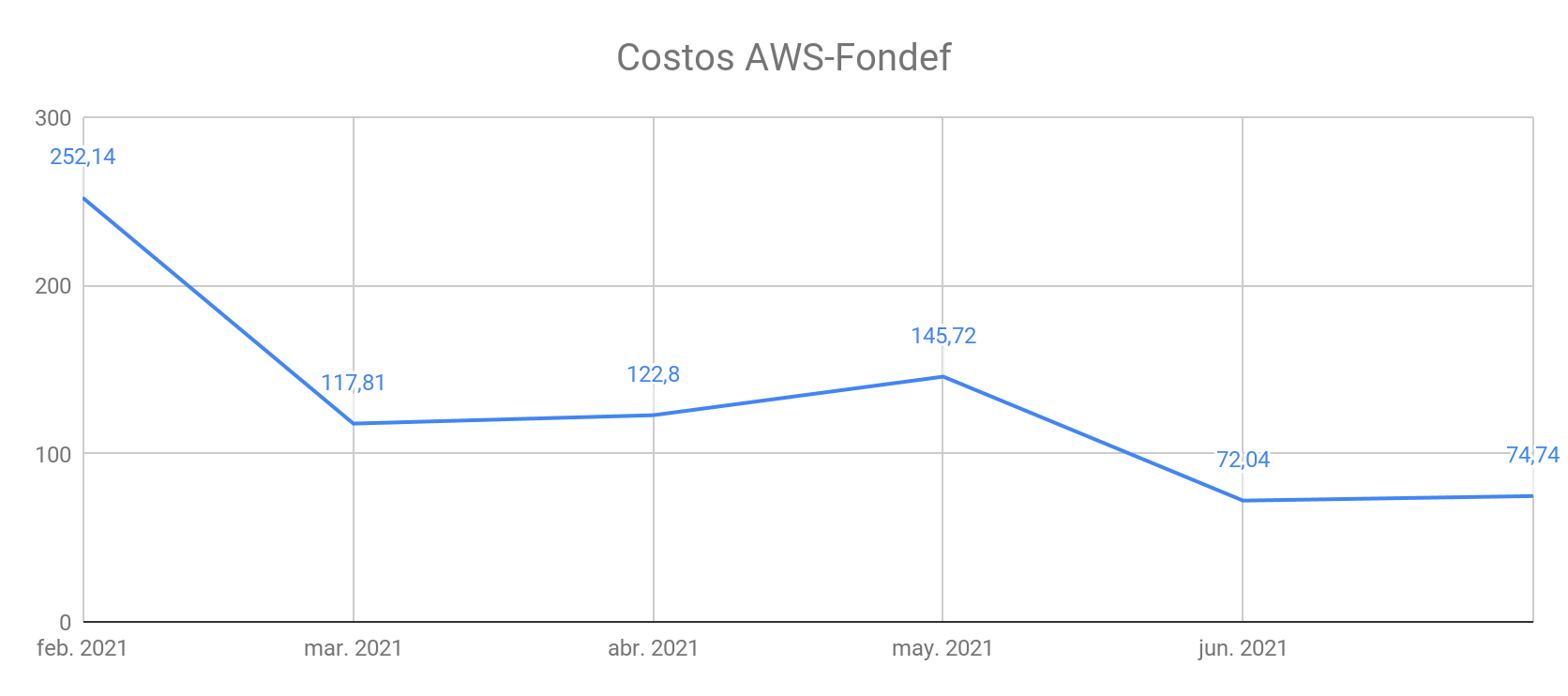
**Los elementos mínimos para la continuación del proyecto y los costos asociados.**

En resumen, los elementos mínimos para la sostenibilidad del proyecto son:

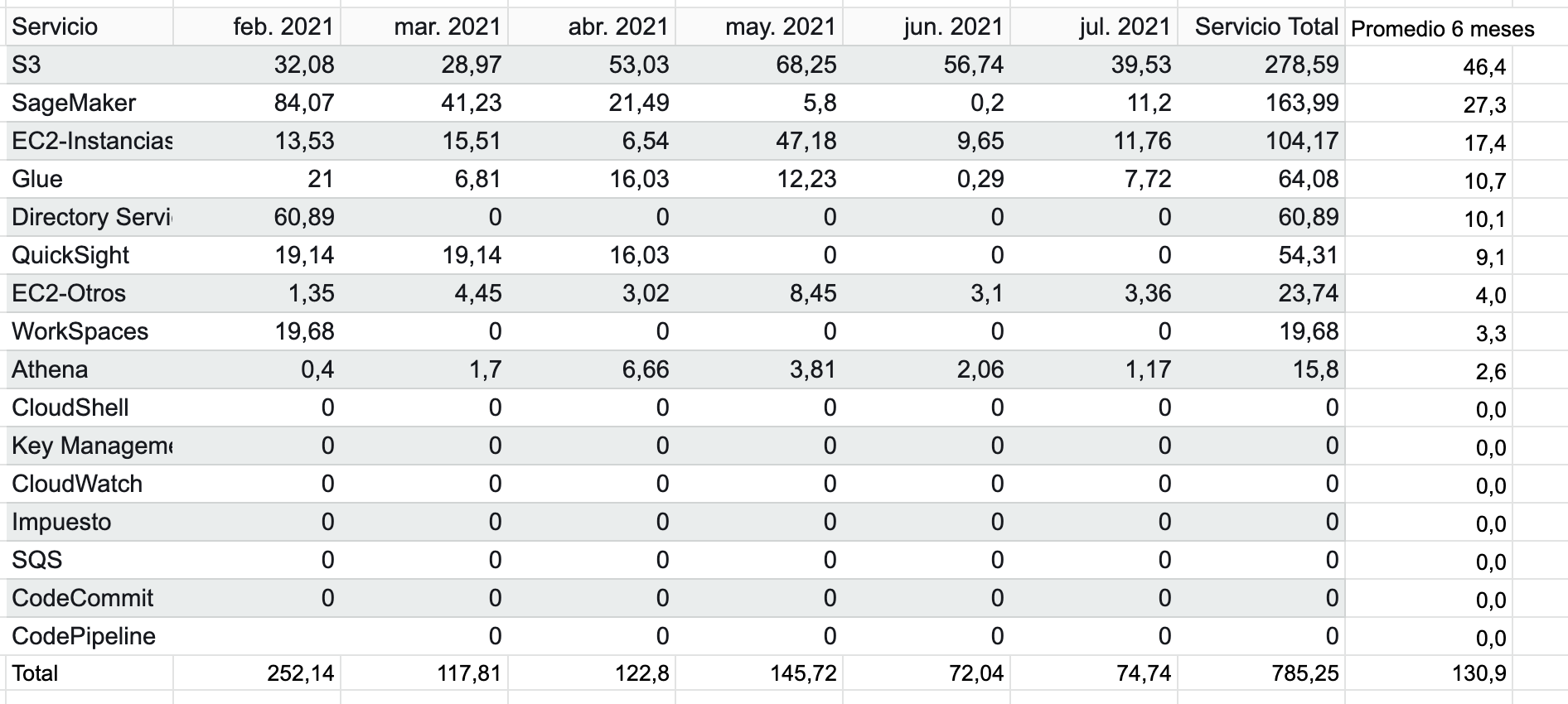
1. Contar con personal técnico adecuado en la Secretaría de Modernización del Estado (lo que estaría siendo garantizado con el actual equipo técnico y la capacitación que recibe de parte del equipo del proyecto)
2. Destinar los recursos necesarios para el mantenimiento de la cuenta de AWS.
3. Destinar los recursos necesarios para el mantenimiento de la cuenta profesional de Power BI.
4. Incluir dentro de los procesos internos las actividades para mantenimiento del proyecto.

Para dar más claridad en los recursos comprometidos en los puntos 2) y 3) indicamos con mayores detalles los costos y servicios requeridos.

Los costos asociados a servicios de AWS dependen de cuánto uso se les de (en acuerdo al modelo de negocios que plantea la nube). Si bien en los primeros tiempos de ejecución del proyecto, los costos fueron más elevados, debido a la intensidad del trabajo,  en este último tiempo las operaciones ya están en régimen y los costos son más bajos. El gasto promedio mensual de los últimos 6 meses es aproximadamente de USD$130 (en contraste con los USD$215 del primer año de ejecución del proyecto). Como ilustración, se recoge de manera gráfica (figura 1)  el gasto incurrido en los últimos 6 meses por uso de AWS y se da una tabla de costos diferenciados por servicio (tabla 1).



**Figura 1:** Costos de uso de AWS en los últimos 6 meses a modo de referencia



**Tabla 1:** Costos asociados a los distintos servicios empleados de AWS.

1. Esto incluye 25 Ministerios, 374 servicios y alrededor de 600.000 funcionarios/año. [↑](#footnote-ref-1)
2. Actualmente existen otras fuente de datos (no muchas más), que tienen algunas desventajas. En el caso de la data del Artículo 70 (Ley de Reajuste), si bien la data puede estar desagregada en la fuente, los datos públicos están agregados. En el caso de la data del Balance de Gestión Integral (BGI), manejada por DIPRES, se va mucho más allá de los datos de personal, pero los datos públicos son agregados o no están disponibles. SIAPER, por otra parte, tiene datos con un buen identificador (RUT), pero los datos no son públicos. [↑](#footnote-ref-2)
3. https://aws.amazon.com/es/ [↑](#footnote-ref-3)
4. Al momento de generar este informe el sitio Web oficial del proyecto aún está en construcción, aunque en su etapa final. Una muestra del sitio puede verse en: https://fondef.vercel.app/min-hacienda/persona-en-el-estado/fondef [↑](#footnote-ref-4)
5. https://powerbi.microsoft.com/es-es/ [↑](#footnote-ref-5)