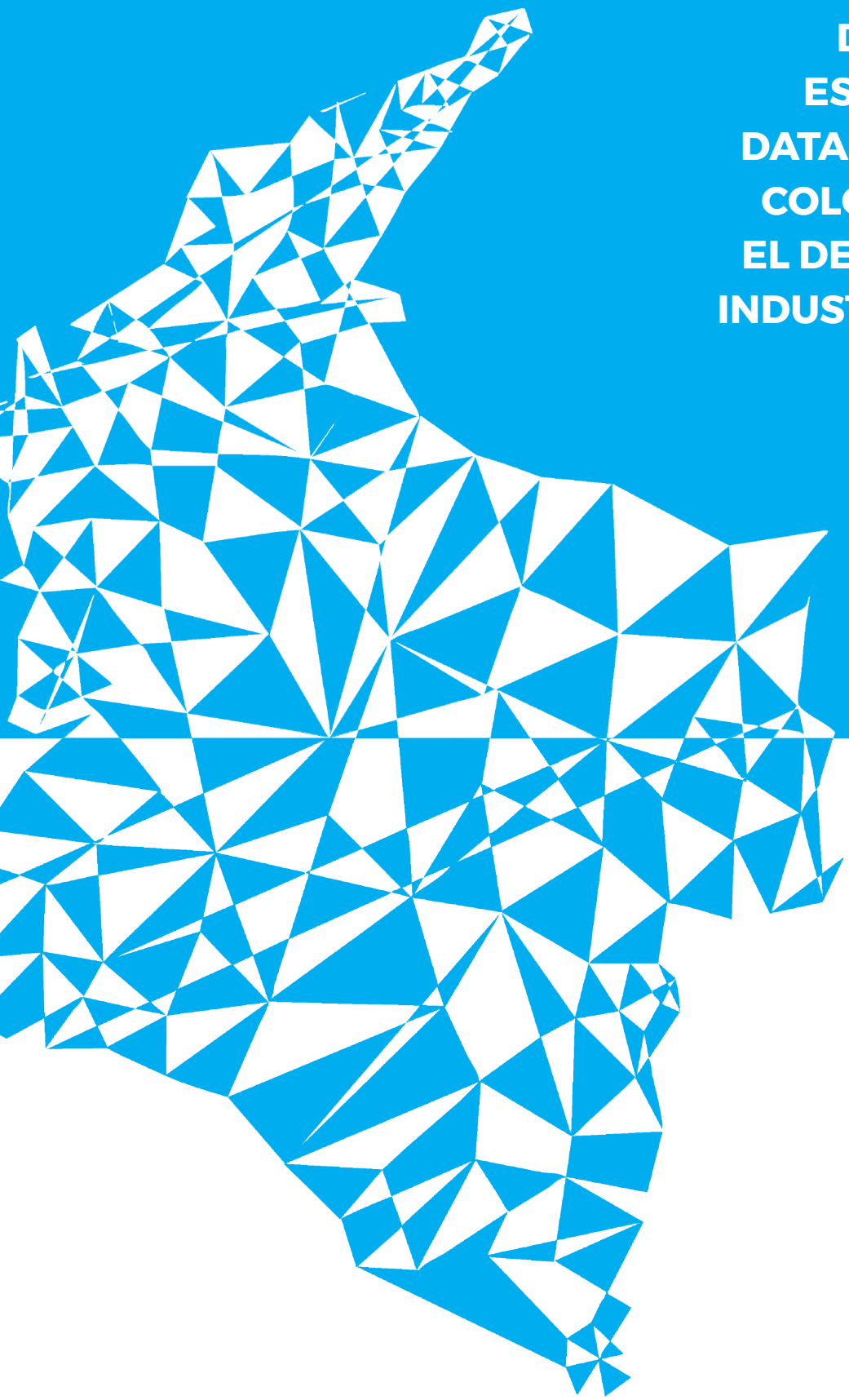


DEFINICIÓN DE LA ESTRATEGIA DE BIG DATA PARA EL ESTADO COLOMBIANO Y PARA EL DESARROLLO DE LA INDUSTRIA DE BIG DATA EN COLOMBIA

MODELO DE PROYECTOS DE BIG DATA: MANUAL Y RECOMENDACIONES

Marzo 2020

9

DATA-POP ALLIANCE

mit media lab

DNP Departamento Nacional de Planeación

iNNpulsa

Producido por un equipo compuesto por:

Maria Antonia Bravo, Andres Lozano, Valentina Casasbuenas, Emmanuel Letouzé

Coordinado por Emmanuel Letouzé

Con insumos de David Shrier

Bajo la supervisión general de: Alex Pentland

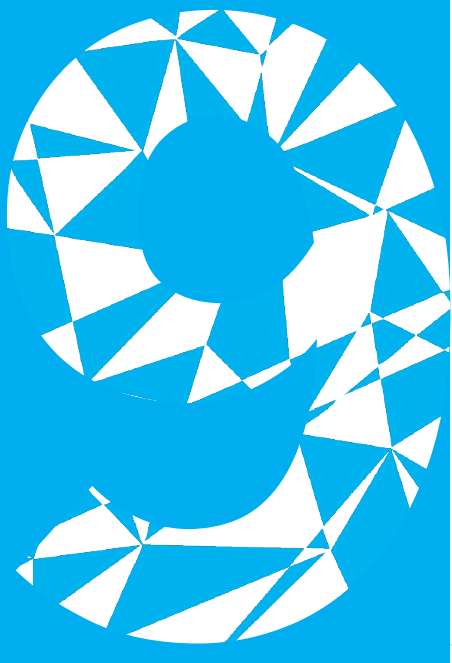
Diagramación editorial por Paola Caile

Asistencia de investigación y edición por Mariana Rodríguez

Marzo de 2020

Versión revisada y ajustada

MODELO DE PROYECTOS DE BIG DATA: MANUAL Y RECOMENDACIONES



ÍNDICE

MÁS Y MEJORES DATOS - ¿MEJORES POLÍTICAS PÚBLICAS?

Este documento, incluye recomendaciones y lineamientos de cómo el Gobierno puede adelantar proyectos de analítica de datos, detallando un modelo de Big Data para quienes buscan resolver, analizar o quizá, entender con más granularidad problemáticas de la gestión pública por medio de los datos.

Los datos como fuente de valor agregado para la toma de decisiones se ha convertido cada vez más en un hecho probado. Desde crear modelos de analítica para predecir la tasa de contagio de una enfermedad, o caracterizar individuos según su riesgo crediticio a futuro, los datos - y las metodologías y herramientas necesarias para darle sentido a estos - cada vez se vuelven más importantes para los gobiernos. En el marco de las discusiones sobre la Cuarta Revolución Industrial y de cómo aprovechar las nuevas capacidades para la explotación de datos, emergen también discursos alarmistas y optimistas de lo que se puede hacer y no con las tecnologías. Se discute quien se aprovecha en mayor medida de estos beneficios (y por qué), y quienes se dejan atrás, sí de verdad hay valor agregado o si todo es una discusión sin fundamento. Es cierto que aun cuando existen límites de lo que se puede hacer con los datos, la generación de nuevos tipos de conocimientos a través de estos no es solo posible, sino necesario.

Mientras en el sector privado diferentes empresas y corporaciones utilizan de manera incremental estas innovaciones en su trabajo, la impresión sobre el sector público es de rezago ante la innovación. Aun cuando este no es el caso, es necesario crear lineamientos y orientar a las entidades públicas hacia una integración transversal de la revolución digital, no sólo limitándose a pensar en herramientas y metodologías, pero entendiendo esta como una mentalidad. No bastará solo con implementar procesos aislados en varias entidades; es necesario impulsar un cambio de paradigma, donde la tecnología aporte a todos los pasos que requiere la toma de decisiones.

Al lograr acceder a distintas fuentes, combinar diversos conjuntos de datos y compartir información, cada entidad tendrá la posibilidad de obtener información que complemente y genere valor a sus propias funciones. Al mejorar los insumos para tomar decisiones, el conocimiento resultante sobre la gestión y función pública — por ejemplo, tener a disposición información actualizada o más granular sobre qué cambios realizar y qué continuar — permitirá obtener un conocimiento más preciso, que traerá consigo decisiones informadas y proyectos cuyo impacto se verá potencializado.

Por lo tanto, el objetivo general de este documento es poder guiar al lector, de una manera clara, a través de las distintas secciones y variables que deberá tener en cuenta para formular un proyecto de gestión pública con datos que impacte de manera positiva la política pública del país.

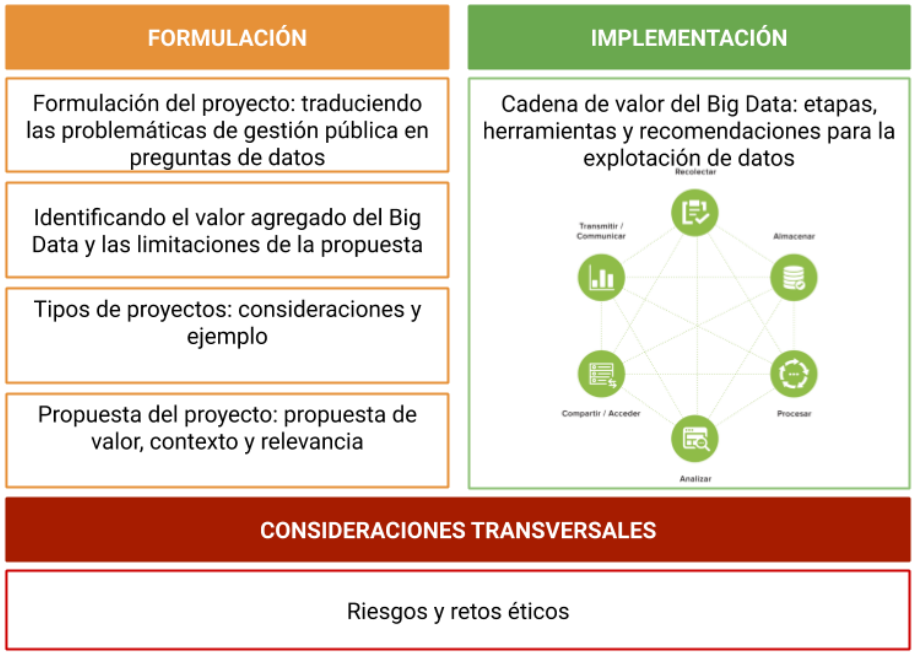
MANUAL PARA LA FORMULACIÓN DE PROYECTOS DE BIG DATA

Este manual busca orientar a quienes diseñan proyectos de gestión pública hacia el uso de Big Data y fuentes de información no-tradicionales, brindando una introducción a los conceptos necesarios, e incluyendo un modelo que abarca las consideraciones requeridas para formular dichos proyectos. De manera complementaria, el documento identifica potenciales casos de uso de dicho modelo, así como resúmenes breves de los hallazgos generados a partir de los pilotos de analítica adelantados en el marco de esta consultoría. Por medio de distintas recomendaciones, se presenta un modelo de proyectos de datos estandarizado que sirva como insumo para que funcionarios del Estado puedan hacer uso del Big Data para la toma de decisiones.

Teniendo en cuenta que la mayoría de profesionales involucrados con la formulación e implementación de política pública no tienen una formación profesional relacionada con la explotación de datos, el presente documento pretende entonces cubrir de una manera sencilla y entendible para todo tipo de público, los fundamentos que involucra la formulación de un proyecto de Big Data, incluyendo recomendaciones y sugiriendo preguntas clave que facilite el proceso para el lector.

Este manual va dirigido a una audiencia profesional vinculada a la gestión y creación de políticas públicas, más no necesariamente para expertos en programación y/o sistemas de información. Para el uso y entendimiento de este documento, no es requerido contar con experiencia en computación, estadística, gestión y/o análisis de datos.

Este documento se encuentra dividido en cinco secciones. Ya que parte de la audiencia esperada para el manual no cuenta con formación ni experiencia técnica en datos, la primera sección aclara los conceptos y contextos generales que se consideran necesarios para interpretar este manual y formular un proyecto de esta naturaleza. Además,se explica brevemente qué es el Big Data y cómo pensar en este como un ecosistema útil para la formulación de proyectos. Adicionalmente, se introduce la Inteligencia Artificial (AI) y los diferentes tipos de esta.



Siguiendo la misma estructura de la imagen anterior que resume las fases del proyecto, la segunda sección cubre la estructura del modelo de formulación de proyectos de Big Data, incluyendo tres fases principales: 1) el proceso de formulación del proyecto, 2) las herramientas y métodos para la implementación de este, y 3) las consideraciones transversales a tener en cuenta en el diseño de un proyecto de Big Data, incluyendo riesgos y retos éticos. A través de estas, el lector encontrará diferentes plantillas que guiarán el proceso de formulación del proyecto, desde la selección de la problemática objetivo del proyecto, los datos y aliados necesarios para su ejecución, el tipo de proyecto a desarrollar y consideraciones éticas a tener cuenta, entre otras.

¿Cómo utilizar las plantillas de este documento?

Este manual, incluye diferentes plantillas que buscan orientar hacia el uso de Big Data a quienes diseñan proyectos de gestión pública. A modo de ejemplo, a lo largo del manual el lector encontrará estas plantillas diligenciadas con un caso de uso puntual, pero estas pueden encontrarse en blanco en el anexo para ser replicadas con facilidad. En itálica se encuentran las instrucciones o consideraciones que se deben tener en cuenta para cada una de estas áreas. Aunque se sugiere diligenciar estas en el orden en la que se encuentran en el documento (de la plantilla I, a la IV) las instrucciones o la información relevante a estas no está necesariamente en orden cronológico en el manual.

¿De dónde provienen las recomendaciones expuestas en el manual?

A lo largo del manual, el lector podrá encontrar recomendaciones atinadas a cada una de las fases del diseño del proyecto. Estas recomendaciones, y en general el contenido de este manual provienen del formato de proyectos de Data-Pop Alliance, construido en conjunto con el United Nations Staff Systems College (UNSSC) y el MIT, y actualizado según guías y recomendaciones de fuentes tales como "A Guide to Data Innovation for Development: From Idea to Proof-of-Concept" del PNUD y UN Global Pulse, y "Data Collaboration for the Common Good" del World Economic Forum[1] . La información recogida en estas fuentes, ha sido adaptada al contexto de la gestión pública, considerando también las lecciones aprendidas de los proyectos pilotos de analítica de datos implementados en el marco de esta consultoría. Cabe resaltar que este manual fue plenamente divulgado en los talleres realizados del 21-25 de Octubre del 2019 en Bogotá, Colombia. Para más detalles sobre el contenido y el material empleado en estos, haga click aquí.

La cuarta sección del documento aborda recomendaciones de política pública que pueden facilitar la implementación de proyectos de analítica desde el gobierno. Estas recomendaciones hacen un llamado a integrar formatos y procesos que habiliten y expediten la explotación de datos, brindando especial atención a los aprendizajes producto de los pilotos de analítica adelantados en el marco de esta consultoría. La quinta sección, detalla usos del modelo propuesto en el manual, describiendo dos casos de uso potenciales y diligenciados según las plantillas de formulación del proyecto. Por último, la sexta sección resume brevemente los hallazgos generados a partir de dos proyectos de datos realizados en el marco de esta consultoría.

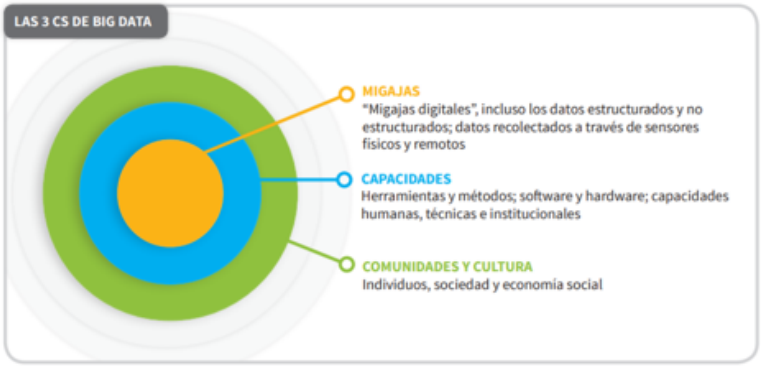
SECCIÓN I.CONTEXTOS Y CONCEPTOS

1. INTRODUCCIÓN AL BIG DATA

Comúnmente, el Big Data se ha traducido al español como 'datos masivos', una definición que enfatiza el tamaño de los datos como su característica principal. En términos críticos, ni los datos masivos ni el Big Data se deben reducir a grandes conjuntos de datos. En lugar de centrarse en la cantidad de materia prima, es mucho más útil pensar en términos cualitativos acerca de la naturaleza de esta materia prima, el ecosistema que la rodea y el fenómeno socio-tecnológico que alimenta al mismo.

Como se explica a continuación, es más útil entender el Big Data como un ecosistema, donde los "datos masivos" son definitivamente una parte relevante de este ecosistema, pero donde la definición de este no se limita únicamente al tamaño de los datos. De manera que incluso un *pequeño conjunto* de 'big data' es relevante a este concepto, simplemente por el hecho de que estos pudieron ser recolectados de manera pasiva o de procesos de datos controlados.

Figura 1. Las 3Cs del Big Data



[1] UN Global Pulse, "A Guide to Data Innovation for Development. From Idea to Proof-Of-Concept", diciembre de 2016; World Economic Forum, "Data Collaboration for the Common Good: Enabling Trust and Innovation Through Public-Private Partnerships", el 20 de mayo de 2019.

Big Data is not just big data

Para definir el ecosistema de Big Data, es útil guiarse por el marco de las 3Cs, concepto desarrollado por Data-Pop Alliance. Esta conceptualización de Big Data como un ecosistema delinea las partes que componen este - "Crumbs" o *migajas*, "Capacities" o *capacidades* y "Communities" o *comunidades*.

Crumbs o Migajas

Las migajas digitales se pueden pensar como las migajas que va dejando alguien mientras se mueve en el mundo digital. En otras palabras, se refiere principalmente a los datos emitidos y recolectados de forma pasiva como un subproducto de la interacción entre las personas con dispositivos digitales. En el centro de nuestras sociedades 'informáticas' está la producción de este tipo de datos masivos, producto también de conexiones entre diferentes plataformas, redes sociales y máquinas. Esta es una característica clave, ya que son estas migajas las que generan un cambio fundamental a nivel cualitativo y cuantitativo, dándole al Big Data un carácter político.

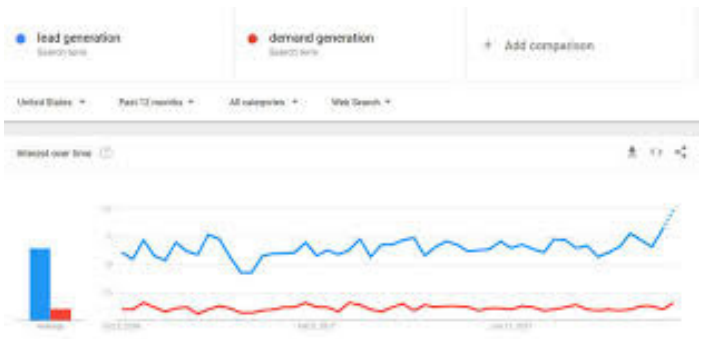
Dentro de las 'migajas' existen varios tipos tipos de datos. El primer tipo de "migajas" se llaman los "datos de escape" que son los resultados de las interacciones de los usuarios con dispositivos digitales y servicios. Se originan por ejemplo del uso de teléfonos celulares (registros de llamadas), tarjetas de crédito (transacciones), de medios de transporte (registros de metro o autobús o un sensor electrónico de cobro de peaje, por ejemplo), redes sociales y motores de búsqueda.

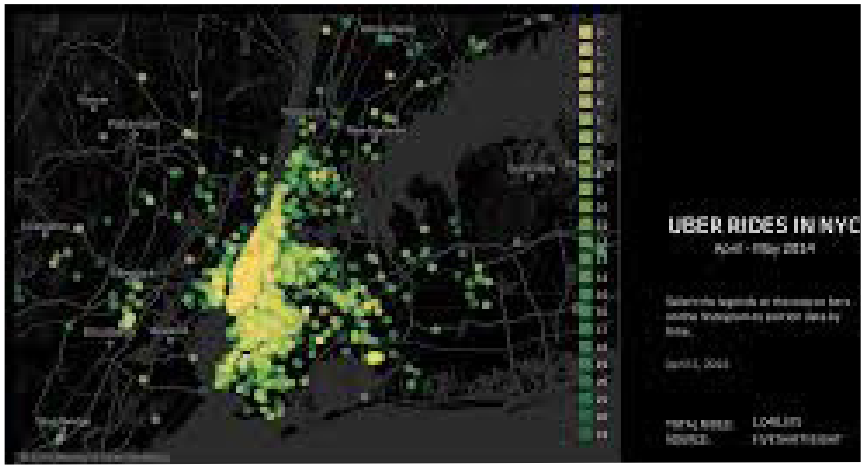
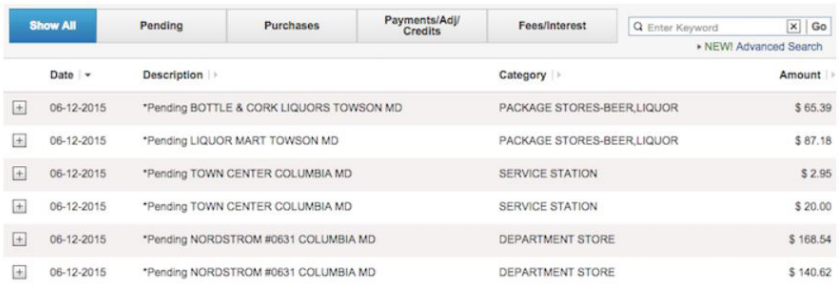
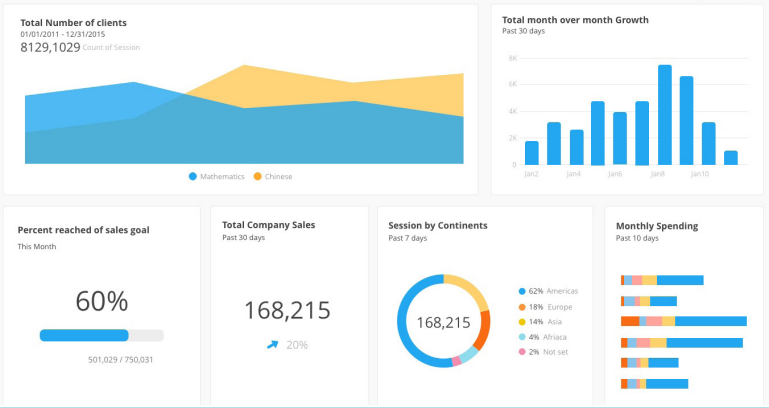
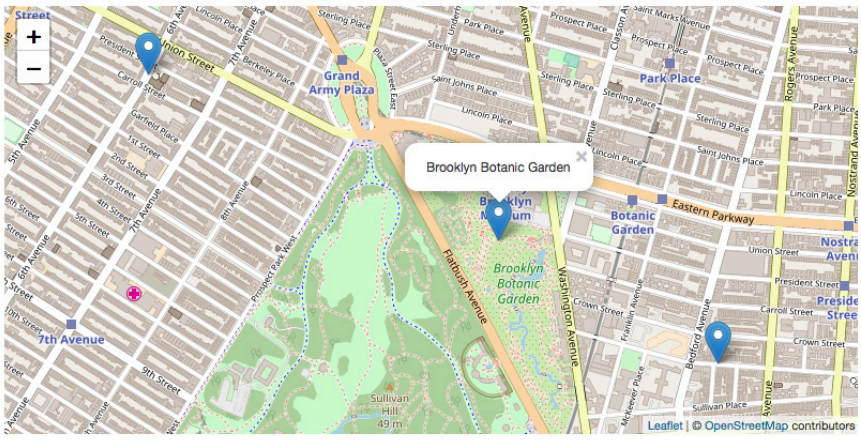
El segundo tipo de "migajas" en nuestra taxonomía incluye videos, documentos, imágenes, publicaciones en blogs y otros contenidos en redes sociales. Estos son datos no estructurados que son más difíciles de analizar de manera automatizada (ya que no se encuentran en la presentación tradicional de filas y columnas), y también están más expuestos a las decisiones editoriales de sus autores; pueden ser, como Alex Pentland lo dijo, "editados de acuerdo con las normas del día de la red social que le corresponda"[2] Esto puede limitar su potencial para la investigación o llevar a conclusiones equivocadas si se toma por ejemplo su valor nominal. Sin embargo, este tipo de datos se ha vuelto clave para el estudio de los sentimientos, deseos, percepciones, creencias, etc. Es importante señalar que estos datos requieren ser convertidos en datos estructurados para poder ser analizados automáticamente - por ejemplo, el número de píxeles en una foto, o el número de veces que una palabra clave aparece en un libro. Este proceso es aquel sobre el que se hacen preguntas y sobre los datos iniciales que reflejan las intenciones humanas de la misma forma que los CDRs reflejan lo que los operadores de telecomunicaciones u otros usuarios finales necesitan y quieren saber. Una consecuencia interesante es a lo que los datos no estructurados se prestan para muchos tipos de preguntas de investigación, incluyendo la posibilidad de desafiar los tipos de datos estructurados que se crean a partir de ellos.

El tercer tipo de migajas es recolectado por sensores digitales, ya sean físicos o remotos. Los primeros son dispositivos físicamente instalados para capturar información sobre acciones humanas - como medidores eléctricos, o, cada vez más, cierto tipo de dispositivos portátiles. Una vez más, los datos que se toman a partir de estos dispositivos se recogen para un propósito específico, pero pueden terminar siendo utilizados también para fines distintos a los que fueron recogidos inicialmente

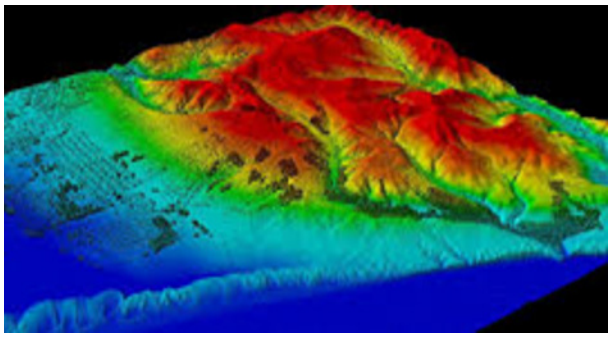
No existe una dicotomía bien definida, una definición universal o estándar, acerca de lo que forma parte de los datos masivos y lo que no. Es principalmente una cuestión de definirlo en un contexto específico - así como se muestra a continuación.

Migajas - Ejemplos de Tipos de Datos

<table>
<tr><td rowspan="2">Basadas en móviles (teléfonos celulares)</td><td>Datos de GPS móvil</td><td rowspan="2"><table>
<tr><th>CALLER ID</th><th>CALLER CELL TOWER LOCATION</th><th>RECIPIENT PHONE NUMBER</th><th>RECIPIENT CELL TOWER LOCATION</th><th>CALL TIME</th><th>CALL DURATION</th></tr>
<tr><td>X76VG588RLPQ</td><td>2°24' 22.14", 35°49' 56.54"</td><td>A81UTC93KK52</td><td>3°26' 30.47", 31°12' 18.01"</td><td>2013-11-07T15:15:00</td><td>01:12:02</td></tr>
</table></td></tr>
<tr><td>Registro de detalles de llamadas</td></tr>
<tr><td rowspan="3">Trazas de navegación en línea</td><td>Cookies de un navegador</td><td rowspan="3"> 
</td></tr>
<tr><td>Direcciones IP</td></tr>
<tr><td>Historial de búsqueda</td></tr>
</table>

Transporte	Seguimiento de flota (GPS, taxi, buses, etc)	
	Rideshare (desplazamiento compartido) (Uber, Cabify, etc)	
Financiero	Mercado virtual (Airbnb, etc)	
	Transacción crédito/ débito	
	Programas de fidelidad y membresías de recompensas	
Redes sociales	Contenido de publicaciones	
	Gráficos sociales basados en creaciones de usuarios	
	Metadata de publicaciones	
Datos de Crowd-sourcing	Llamadas en redes sociales usando hashtags	
	Contenido estructurado en línea	
	Contenido no estructurado	

[2] Alex Pentland, "Reinventing society in the wake of big data," Edge, 2012.

Sensores Físicos	Contadores inteligentes	
	Sismómetros oficiales	
	Rastreadores de velocidad/peso	
Sensores Remotos	Vehículos aéreos no tripulados (drones)	
	Imágenes satelitales (NASA, TRMM, LANDSAT)	

Capacidades

Las capacidades son entendidas como el conjunto de herramientas y métodos, hardware y software, know-how y habilidades necesarias para procesar y analizar este nuevo tipo de datos, incluyendo técnicas de visualización, aprendizaje estadístico automatizado (machine learning), algoritmos, etc. Estas actúan como puente entre las migajas y las comunidades, para analizar los usos innovadores y alternativos de los datos. Además, influyen en la forma en que las comunidades de Big Data interactúan y movilizan en torno a nuevos conocimientos y metodologías.

Estas capacidades incluyen computadoras potentes, infraestructuras computacionales, centros de datos, así como técnicas de visualización, familias de algoritmos, aprendizaje de máquinas (machine-learning) y técnicas de aprendizaje profundo (deep-learning) que tienen la capacidad para buscar e identificar patrones y tendencias en grandes cantidades de datos complejos, entre otros.

Comunidades

La tercera y última C son las 'comunidades' o en otras palabras, las partes involucradas en la generación, gobierno y uso de datos, incluidos los productores de datos, usuarios finales, autoridades, la sociedad civil, expertos, defensores de la privacidad y comunidades de hackers cívicos, así como cualquier persona representada en un conjunto de datos. En otras palabras, las comunidades son la superestructura de Big Data, incluyendo los actores involucrados en el ecosistema (desde ciudadanos hasta generadores de datos, analistas y usuarios finales) y las instituciones socio-políticas (leyes, cuadros político-institucionales).

La economía política de Big Data no puede ser capturada adecuadamente sin reconocer y entender las relaciones e interacciones que se desarrollan entre sus diversos actores y los incentivos subyacentes que los conectan. Los objetivos, habilidades y limitaciones de estos individuos y grupos deben ser tomados en cuenta para entender lo que es Big Data, lo que se puede hacer con Big Data y cómo podría moldearse para producir resultados positivos.

Entre los actores de estas comunidades se encuentran:

- Academia e investigación
- Sociedad civil/ONGS
- Instituciones multilaterales
- Gobiernos/Institutos nacionales de estadística
- Sector privado (E.G. medios de comunicación, financiero, etc)

2. INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

El concepto de Inteligencia Artificial se le acredita a Alan Turing, científico que en 1950 se preguntó si las máquinas eran capaces de pensar al mismo nivel que un ser humano. De este cuestionamiento, el cual busca responder a esta pregunta de manera afirmativa, nació el reconocido "*Turing test*" — si una máquina puede hacer que los humanos piensen que la máquina es humana, entonces esta es humana también. Años después, siguiendo a Turing, fue John McCarthy quien utilizó el término de Inteligencia Artificial para hacer referencia a aquellas máquinas que pudieran pensar de manera autónoma. De aquí que la Inteligencia Artificial se pueda describir como la ciencia de entrenar máquinas o programas que aprenden a realizar tareas humanas o que tradicionalmente requieren tipos de inteligencia humana[3](West, D. 2018 & SAS. s.f.)[4].

Actualmente, se afirma que existen tres tipos de Inteligencia Artificial. La primera es la denominada "**inteligencia artificial débil**", que permite que los computadores hagan una serie de tareas específicas, bajo una serie de restricciones y limitaciones. Un ejemplo de esta son las máquinas que pueden jugar ajedrez o póker. Aquí, una persona introduce manualmente todas las reglas del juego y en cada jugada la máquina usa las reglas para calcular la movida que tiene mayor probabilidad de éxito; sin hacer uso de datos comportamentales de los contrincantes o de datos históricos de jugadas.

La segunda es la "**inteligencia artificial general**" o "**inteligencia artificial fuerte**", la cual se refiere a la habilidad de una máquina de aplicar su "inteligencia" a diferentes problemáticas; esto supondría realizar las mismas tareas que un ser humano autónomamente, lo cual estamos lejos de lograr (*Builtin*. s.f.)[5]. Por encima de esta, estaría la "**superinteligencia artificial**", que es cuando una máquina sobrepasa las habilidades de las mentes más brillantes en el mundo. Esta última, por ahora no está en ningún futuro cercano.

Existen varios campos de aplicación de la inteligencia artificial (Kumar, C. 2018)[6]:

- Aprendizaje de máquinas (machine learning o ML): aunque frecuentemente se usa como sinónimo de inteligencia artificial, el aprendizaje de máquinas representa solo una parte de la IA. En el ML, se le permite a la máquina crear sus propias reglas para analizar los datos de entrada y generar la información de salida. Algunas de las formas de aprendizaje de máquinas más comunes son[7]:
  - Aprendizaje supervisado: son modelos en los que se busca aproximar la función matemática que explica la relación entre un grupo de variables de entrada y una o más variables de salida. Por medio de un proceso iterativo de prueba y error, el algoritmo actualiza los valores de los parámetros de cada función, hasta que el conjunto de funciones permite la predicción de los valores en la variable de salida a partir de los datos de entrada. Es supervisado en el sentido en el que tener los datos de salida le da un referente al modelo para calcular sus errores de predicción y así mejorar en cada iteración la capacidad predictiva del modelo. Un uso de estos modelos es la detección de patologías a partir de imágenes médicas, como radiografías. Al tener información sobre qué imágenes muestran la existencia de una patología y qué imágenes no, el modelo puede aprender las características de la patología presentes en la imagen y ayudar en la predicción de las mismas en otros pacientes.
  - Aprendizaje no supervisado: son modelos en los que se busca encontrar los patrones presentes en un grupo de datos. Se busca encontrar una función que encuentre distancias, similitudes y diferencias entre los datos de entrada para agruparlos y diferenciarlos de otros según un gran número de características. Encontrar patrones en la información resulta útil en casos como la segmentación de mercados para los negocios a partir de datos demográficos y de actividad en línea. Es de notar que en el caso de segmentación de mercados se quiere encontrar grupos de personas para dirigirles un producto que responda a sus gustos, a través de publicidad. En este caso no se necesitan datos de salida y hace de este un problema de aprendizaje no supervisado.
- Procesamiento de lenguaje natural: corresponde a la manipulación automática del lenguaje natural por medio de un software. Es decir, se procesa esta habla, ya sea verbal o textual, por medio de un software que lo identifica y puede determinar diversas cosas sobre él, como por ejemplo el lenguaje vulgar o fuerte.

---

[3] En general, se reconoce la inteligencia humana como una serie de procesos y habilidades - procesos cognitivos que no son aislados

[4] West, D. 2018. What is artificial intelligence? Brookings; SAS. SAS Insights: Analytic insights. (s.f.).

[5] Builtin. Artificial Intelligence. s.f. https://builtin.com/artificial-intelligence

[6] Kumar, C. 2018. "Artificial Intelligence: Definition, Types, Examples, Technologies". Medium.

- Robótica: es el área de la ingeniera encargada de diseñar y manufacturar robots, que son utilizados para ciertas tareas.
- Visión de computador: es el campo en el cual las máquinas tienen la habilidad de "ver", por medio de la captura y análisis de información recibida a través de cámaras, conversión de análogo a digital y procesamiento de señal digital. Aunque la visión de computador se desarrolló de manera separada al aprendizaje de máquinas, la aplicación de este último al análisis de imágenes y video permitió aumentar la cantidad de datos que se procesan y la complejidad de las tareas que se realizan.

Independiente de su complejidad, la inteligencia artificial debe ser vista como una tecnología de utilidad general, lo que implica entender esta como una tecnología capaz de inducir cambios considerables y a largo plazo, tales como los precipitados por el ferrocarril, el automóvil y el internet (Brynjolfsson, Rock, y Syverson 2017; Biagi 2013)[8]. Particularmente, se espera que influya de manera significativa sobre el crecimiento de la economía, pues endógenamente llevará a innovaciones en productos, procesos, y generalmente en organizaciones en sectores que invierten en tecnología.

SECCIÓN 2. PROCESO DE FORMULACIÓN E IMPLEMENTACIÓN DE PROYECTOS DE BIG DATA PARA EL GOBIERNO

El desarrollo de proyectos de Big Data para la administración pública requiere de un riguroso planteamiento de proyecto, logrando el entendimiento de una problemática social por medio de preguntas que los datos pueden responder. Esta sección del manual divide las etapas de formulación de un proyecto de Big Data y en cada una de las subdivisiones se hacen las preguntas clave que permiten entender la problemática y crear un plan de acción; brindando la flexibilidad necesaria para permitir los procesos iterativos inherentes a este tipo de proyectos.

El primer punto de esta sección se enfoca en el planteamiento y formulación del proyecto. Empezando con la formulación de una problemática de la gestión pública en una pregunta que se pueda responder a través del análisis de datos, se tocan temas fundamentales que permitan al lector encuadrar el proyecto y finalmente proponer un tipo de proyecto teniendo en cuenta los riesgos y desafíos que este supone.

La segunda parte de esta sección, recoge las herramientas necesarias para la implementación de un proyecto de esta naturaleza en cuanto a la recolección, almacenamiento, procesamiento, análisis, transferencia y comunicación de los datos,,caracterizando algunos prototipos de proyectos posibles.

Por último, la tercera parte de esta sección recoge consideraciones transversales a ser tenidas en cuenta durante todo el ciclo de proyecto, como lo son los retos y riesgos éticos a lo largo de un proyecto.

Para complementar esta sección y con el objetivo de ejemplificar y brindar casos de uso del manual, dos ejemplos de Big Data útiles para la administración pública se encuentran en la Sección 4 del documento.

Sección 2.1: Planteamiento del proyecto

1. FORMULACIÓN DE PROYECTOS DE BIG DATA: TRADUCIENDO LAS PROBLEMÁTICA DE GESTIÓN PÚBLICA EN PREGUNTAS DE DATOS

El primer objetivo para el planteamiento del proyecto es el de traducir una problemática de la gestión pública en una pregunta que se pueda responder a través del análisis de datos. Es importante que desde el comienzo del proyecto se busque tener un balance entre la problemática a tratar y lo que de forma realista se puede responder con datos, para delimitar el tema y conceptualizar un proyecto que a *priori* es realizable en cuestiones de alcance y acceso a los datos.

A continuación se presenta la plantilla que guía el planteamiento del proyecto, acompañada de las instrucciones que permiten diligenciarla. Posteriormente, se ilustra el proceso de formulación de un caso de uso específico diligenciado la diferentes parte de la plantilla.

---

[7] Maheshwari, V. 2018. "Supervised and Unsupervised Learning". Medium.

[8] Brynjolfsson, Erik, Daniel Rock, y Chad Syverson. 2017. “Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics”. w24001. Cambridge, MA: National Bureau of Economic Research; Biagi, Federico. 2013. ICT and Productivity: A Review of the Literature. Luxembourg: Institute for Prospective Technological Studies.

1. Para empezar, elija el área de impacto en la que se va a enfocar su proyecto, para luego definir la(s) pregunta(s) clave(s) que quisiera poder contestar al llevarlo a cabo. Entre más específica sea esta área, más fácil será plantear el proyecto.
2. Determine cuales son las preguntas claves que le interesan sobre dicha temática. Estas preguntas guían al equipo de formulación de proyectos hacia la caracterización de los desafíos importantes que se presentan en el área que se desea impactar. Para ello, se hacen preguntas generales como *¿cuáles son los desafíos más importantes en dicha área?, ¿por qué no se han resuelto satisfactoriamente hasta ahora?, ¿qué información se necesita para afrontar aquellos desafíos?, ¿qué información es importante en el contexto social, económico y político en el que se desarrolla la problemática?* Tenga en cuenta que en esta etapa es ideal empezar a pensar qué información necesitará para responder estas preguntas claves. En vez de enfocarse en la oferta de información y construir su proyecto, se recomienda hacerlo desde una perspectiva de la demanda - *¿que información necesito?¿puedo acceder a esta?*
3. Teniendo en cuenta estas preguntas, reflexione acerca de las migajas (fuentes de datos), capacidades y comunidades que van a ser necesarias para responderlas. En cuanto a las migajas, piense: *¿qué datos necesita?, ¿están estos disponibles?, ¿debo poner en marcha procesos de recolección de datos?, ¿estos datos reflejan alguna dimensión del área de impacto?* Para capacidades, piense: *¿qué métodos y herramientas son necesarias para almacenar, analizar y/o procesar esta información?, ¿necesito conocimiento externo a mi entidad para la explotación de estos datos?.* Por último, piense: *¿cuáles son las organizaciones clave con las que va a trabajar?, ¿quiénes tienen los datos necesarios?, ¿quiénes son los sujetos de datos?, ¿qué comunidades se pueden ver afectadas o están involucradas en el proyecto?, ¿qué actores deben estar involucrados en este proyecto?* En lo posible, formule una hipótesis - *¿la migajas, capacidades y comunidades que incluyo me ayudaran a generar valor sobre el área de impacto pensado?*
4. Tras un análisis de la pregunta clave, el área de impacto y las 3Cs, plantee el problema que intenta resolver por medio de este proyecto.El objetivo de esta etapa es encontrar un balance entre la necesidad de la gestión pública (primera etapa) y el alcance de los datos (segunda etapa). Mientras la pregunta clave sobre los desafíos a afrontar guía el planteamiento del proyecto, cada una de las 3Cs delimita el alcance de los datos para responder dicho desafío. Las migajas delimitan la problemática en cuanto a quiénes y qué puede ser estudiado. Las capacidades, particularmente los métodos, enmarcan el tipo de respuesta que se busca, por ejemplo si el problema requiere de un análisis comparativo, el problema se puede plantear hacia las diferencias entre grupos (*¿cuáles son los diferentes grupos que presentan el comportamiento de interés?*) o descriptivo de una población (*¿qué características tienen las personas que presentan cierto comportamiento?*). El análisis puede ser correlacional (*¿qué características se relacionan con el comportamiento?*), causal (*¿cuál es el efecto de un programa en el comportamiento del grupo de interés?*) o predictivo (*¿cuál será el comportamiento del grupo de interés en un futuro?*). Finalmente, las comunidades brindan la información sobre los actores con los que es importante contar para desarrollar el proyecto. Ya que las fuentes de migajas pueden ser variadas, la identificación del proveedor de datos ayuda a definir características del proyecto, como la temporalidad del estudio (*¿desde cuándo tiene este proveedor los datos del grupo de interés?*).
5. Tomando como base la problemática planteada, diligencie cada una de las características clave del proyecto. En este punto, tenga en cuenta que su proyecto debe ser factible - verifique el alcance y el acceso a los datos y no olvide delimitar el proyecto en el tiempo y a nivel geográfico.
6. Reúnase con su equipo de trabajo, revise las características clave del proyecto y elija las características que usted determina son prioritarias de su proyecto. Ya que durante la implementación del proyecto, varias de estas características estarán sujetas a cambios, se realiza una priorización de aquellas consideradas claves para que el proyecto responda a la pregunta problema. Por ejemplo, en algunos proyectos la problemática existe en una amplia región geográfica, lo cual permite tener flexibilidad sobre el lugar en el que se efectúa el estudio y permite darle prioridad al lugar en el que los datos estén disponibles. En otro caso, la geografía puede ser tan importante que sería preferible cambiar el tipo de datos o su proveedor para responder a la pregunta problema. Elija las tres primeras prioridades y listeles al final de la plantilla. Estas serán fundamentales al momento de elegir el prototipo de proyecto que se llevará a cabo.

Plantilla I: Volviendo problemáticas de gestión pública en preguntas de datos		
1. Área de Impacto	2. Preguntas Clave:	
3. Las 3C		
Migajas	*Capacidades*	*Comunidades*
4. Planteamiento del problema		

Plantilla I: Volviendo problemáticas de gestión pública en preguntas de datos		
5. Características Clave del Proyecto		
Meta estratégica	*Beneficiarios*	*Acceso a migajas*
Objetivos operacionales específicos	*Gobernanza y alianzas*	*Capacidades y herramientas*
Resultados o productos esperados	*Plazo y alcance geográfico*	*Otros*
6. Priorización de características del proyecto		
Prioridad 1		
Prioridad 2		
Prioridad 3		

Ejemplo: Desarrollo de un caso de uso - Más Familias en Acción

Una entidad del gobierno está interesada en entender la efectividad de uno de sus programa para el alivio de la pobreza de Transferencias Monetarias Condicionadas. Para este ejemplo, se toma el programa colombiano, Más Familias en Acción (MFA), el cual brinda transferencias monetarias a las familias vulnerables cuyos hijos cumplen con un requisito de asistencia escolar y/o a controles de salud. En este caso, el interés de la entidad está en entender en qué se está usando el dinero de las transferencias. El área de impacto específica del ejemplo es la *efectividad de transferencias monetarias condicionadas*.

Entre las preguntas preliminares que pueden ser de interés se encuentran las siguientes:

Elegibilidad	¿Las transferencias son recibidas por el grupo poblacional objetivo?
Acceso	¿Los mecanismos para realizar las transferencias permiten que las personas las reciban?
Uso	Una vez se reciben las transferencias, ¿en qué se usan?
Impacto	¿El uso de las transferencias ayuda a aliviar algún componente de la pobreza?

Para formular dicho proyecto, se decide tomar el uso de las transferencias como objeto de estudio. Las selección de esa pregunta se hace pensando principalmente en la limitación que tiene el Departamento de Prosperidad Social en conocer el uso que se le da a este dinero tras ser entregado, así como la posibilidad de utilizar datos de entidades bancarias para conocer este uso, dado que las transferencias a los beneficiarios se realizan a entidades bancarias para retiros en persona o uso a través de tarjetas débito. Normalmente, dicha información se limita a lo que es recolectado a través de encuestas, que en muchas ocasiones incurre en altos costos de despliegue y baja participación.

Para el estudio de MFA sobre el uso de las transferencias, las migajas estarían representadas por los datos socioeconómicos de los bene beneficiarios y la información bancaria o financiera de la entrega y uso de subsidios.

Se puede pensar en datos del sector privado para obtener información sobre el consumo de bienes y servicios (locales comerciales) o el consumo en tarjetas bancarias (Para este ejemplo, se asume que es posible obtener datos tanto de la entidad bancaria como de MFA, de lo contrario, en esta etapa se replantearía el proyecto). Posiblemente, se puede pensar en mapear el tipo de lugares donde se están haciendo transacciones con el dinero de las transferencias, o alternativamente, se puede mapear temporalmente los hábitos de gasto de esta transferencia en un mes (por ejemplo, si solo ocurre en los primeros tres días tras recibir los beneficios). En cuanto a las capacidades, ya que se requiere una caracterización de grupos que pueden tener diferentes estilos de consumo, se pueden usar métodos de aprendizaje de máquinas no supervisado. En cuanto al almacenamiento y procesamiento, es importante consultar sobre el tamaño de la base de datos a la que se quiere acceder para decidir si es más rentable usar los equipos que la organización que ejecuta el proyecto tiene, o el almacenamiento en la nube.

Las comunidades estarían representadas por miembros del Gobierno, particularmente del Departamento de Prosperidad Social, quienes administran MFA. El sector privado, como los bancos o locales comerciales que tienen

información sobre el uso del dinero y finalmente, los beneficiarios del programa.

Tomando estas consideraciones en cuenta, la pregunta se centra entonces en un análisis descriptivo del tipo de consumo que los beneficiarios de MFA le dan a las transferencias monetarias condicionadas a través de transacciones bancarias.

La meta estratégica del proyecto es identificar el tipo de consumo que los beneficiarios de MFA le dan a las transferencias monetarias. Lo cual requiere, a nivel operacional, que se asegure el acceso a los datos tanto de los beneficiarios de MFA como de sus transacciones bancarias. Esto se debe hacer respetando la privacidad de los beneficiarios, por lo cual, entre otras medidas de privacidad se propone la anonimización de ambas bases de datos (el detalle de las medidas para proteger la privacidad se encuentra en la sección de ética, más adelante en el documento). El resultado del proyecto, debe estar alineado con la meta estratégica del mismo. Por lo cual, se espera que con el proyecto se logre clasificar los tipos de consumo de transferencias que se encuentran en los beneficiarios del programa. En un principio, el beneficiario del proyecto sería la entidad a cargo de MFA, el Departamento de Prosperidad Social (DPS) de Colombia, ya que esta información podría ayudar a entender si los beneficiarios están logrando comprar alimentos con las transferencias. A su vez, se podría ver si en tiempos de crisis, se afectan su capacidad a conseguirlos. En segundo lugar, con esta información el DPS podría promover medidas para apoyar a las familias cuyos hábitos de consumo de vean afectados en determinado momento.

El alcance geográfico del programa de MFA es nacional, por lo cual se podrían elegir diferentes áreas del país. Sin embargo, es necesario que los beneficiarios del programa tengan una cuenta de banco para poder analizar sus transacciones bancarias. Para aumentar la probabilidad de tener beneficiarios con cuentas bancarias, se asume que hay una mayor probabilidad de encontrar una población bancarizada en áreas con mayor urbanización, por lo cual Bogotá podría ser una buena opción. Posteriormente, se corrobora esta información con estudios de inclusión financiera. En este caso, el reporte de inclusión financiera del 2018, afirma que Bogotá tiene la mayor tasa de inclusión financiera (98%) del país[9].

En cuanto al plazo del proyecto, es importante estimar que si los datos no están disponibles y se requiere de diferentes entidades para obtenerlos (en este caso el DPS y una entidad bancaria), es probable que una gran parte del proyecto se vaya en la consecución y estructuración de la base de datos. Para tener un margen prudente, se planean 6 meses para la consecución de los datos y 6 meses para el desarrollo del análisis, validación de los resultados y reporte de los mismos.

En gobernanza y alianzas de datos, se recomienda incluir en un órgano de gobernanza a las comunidades identificadas en el proyecto. Esto con el fin de tener un grupo de personas que puedan guiar las decisiones del proyecto y que tengan un buen entendimiento de su contexto. En este caso un grupo con el el DPS, el banco y los beneficiarios de MFA. La conformación de dicho grupo, puede ayudar la coordinación de acceso a las migajas, a quienes se les solicita un permiso para acceder a los datos[10].

2. IDENTIFICANDO EL VALOR AGREGADO DEL BIG DATA Y LAS LIMITACIONES DE LA PROPUESTA

Las siguiente etapa corresponde a la identificación del potencial valor agregado generado a partir de la explotación de datos y del uso de metodologías y herramientas del Big Data y la Inteligencia Artificial. Esta considera las limitaciones de la propuesta con el fin de generar claridad sobre el alcance del proyecto, sus usos potenciales y facilitar la comunicación con otros actores sobre los beneficios de implementar el proyecto de datos. Esta fase invita a replantearse la pregunta inicial, y si es necesario, reformular esta tras considerar el impacto previsto y potenciales limitaciones. Para esto, considere las siguientes instrucciones:

1. Comience diligenciado el planteamiento del problema, según definió en la primera plantilla. Reformule su planteamiento si lo considera necesario tras revisar las prioridades y características del proyecto.
2. Defina el impacto previsto que tendrá el proyecto, es decir, considere la utilidad que generaría darle respuesta a su pregunta, así como el valor agregado que espera se genere a través de su proyecto (*¿para quien se genera esta valor?*)
3. Determine la utilidad que tiene utilizar metodologías y herramientas de Big Data o Inteligencia Artificial para su proyecto. *¿Por qué es mejor responder a esta pregunta con estas herramientas y no con herramientas tradicionales? ¿Utilizar estas herramientas conlleva a un mejor uso de mis recursos? ¿Que tipo de resultados estoy buscando, y definitivamente, estas metodologías me permitirían esbozar este tipo de resultados?*

[9] Super Intendencia Financiera, 2018. “Reporte de Inclusión Financiera”.

[10] En la sección 4, ‘uso del modelo’ podrá encontrar este ejemplo completamente diligenciado.

4. Considere el contexto y las limitaciones a las que se enfrenta el proyecto. Dentro de estas se pueden encontrar desafíos en cuanto a la obtención de los datos, las necesidades tecnológicas para analizarlos, el manejo de la ética y de la privacidad en el proyecto.
5. Por último, habiendo realizado los pasos anteriores y en caso de considerarse pertinente, se evalúa si es necesario **reformular la pregunta inicial** para tener en cuenta el impacto, la utilidad del Big Data o IA y las limitaciones del proyecto.

Plantilla II: Planteamiento del problema	
¿Qué problema está tratando de resolver utilizando Big Data o Inteligencia Artificial?	
1. *Enmarque el problema en una pregunta de datos*	*2. ¿Cuál es el impacto previsto con su proyecto?*
3. ¿Cómo puede ser el Big Data o la Inteligencia Artificial útil para la resolución de este problema?	*4. ¿Cuáles son las limitaciones y el contexto al que se enfrenta este proyecto?*
5. Reformule su pregunta	

Ejemplo: Desarrollo de un caso de uso - Más Familias en Acción

Según se determinó en el cuadro de arriba, el proyecto se enfocará en la efectividad de programas de transferencias monetarias condicionadas. Para ello, se requiere tener acceso a los datos de consumo de productos a través del uso de tarjetas bancarias.

Para analizar el impacto esperado, es pertinente considerar cuál sería el resultado de responder a la pregunta del proyecto. Por una parte, es de esperar que los beneficiarios de estas transferencia le den diferentes usos a este beneficio, por lo cual pueden existir diferentes patrones en el uso de estos de acuerdo a ciertas de las características de los beneficiarios. El beneficio que se generaría al entender los tipos de consumo y las características de las personas en cada patrón de consumo, determina entonces el impacto del proyecto. En este caso, se puede pensar que el entendimiento de dichos patrones le permitiría al DPS focalizar esfuerzos para corregir conductas de consumo. Esto puede ser particularmente relevante en épocas de crisis, en las cuáles los patrones de consumo pueden verse afectados en las poblaciones vulnerables.

Por otra parte, los beneficios de usar Big Data e Inteligencia Artificial en este contexto se relacionan con la frecuencia y rapidez con la que esta información se puede obtener. Las metodologías tradicionales para obtenerla suele incluir encuestas, que son costosas y requieren de periodos de tiempo largos para su implementación. En este caso, las transacciones bancarias se actualizan casi en tiempo real, lo cual disminuiría costos en la recolección de información. A su vez, la Al permitiría crear grupos de consumo teniendo en cuenta diferentes características de los usuarios de forma no supervisada, creando categorías que pueden no ser evidentes para quienes gestionan el programa.

En cuanto a las limitaciones del proyecto, se puede pensar en las limitaciones de las fuentes de datos. Por ejemplo, analizando cómo se registran los datos de transacciones bancarias, es posible evidenciar que las compras con débito se efectúan con el dinero total disponible en la cuenta. Es decir, para una persona que tiene más de una fuente de ingreso, no es posible diferenciar si una compra se realizó con los fondos de una u otra fuente. En consecuencia, los datos no permitirían evidenciar el uso exacto que los individuos dan a la transferencia recibida.

Por otra parte, dado que el banco permite retirar el dinero y utilizarlo en efectivo, este análisis no sería representativo de los usuarios que pagan en efectivo, o que no tienen cuenta bancaria. Una potencial alternativa para observar las

diferencias en los patrones de consumo sería comparando la diferencia entre el patrón de consumo de beneficiarios y no beneficiarios del programa que tienen tarjeta débito.

Para proteger la privacidad de las personas se puede considerar aplicar el mismo algoritmo de anonimización a datos que contengan la cédula de los beneficiarios del programa MFA en los datos del gobierno y del banco, para luego unir las bases de datos con los mismos números pero sin nombres o datos que permitan la identificación de las personas. Otra consideración importante es que los datos obtenidos a través de los bancos no son representativos de la población no bancarizada. Es decir, que la interpretación de los resultados debe hacerse reconociendo que los resultados de la misma pueden no aplicar a un segmento de la población.

Finalmente se replantea la pregunta del proyecto: ¿Existen diferencias en los patrones de consumo de beneficiarios y no beneficiarios de transferencias monetarias condicionadas medibles a través de transacciones bancarias?

3. TIPOS DE PROYECTOS DE BIG DATA

Para garantizar el éxito de un proyecto de datos, hace falta pensar también cuál es la mejor modalidad para este proyecto, pensando tanto en las alianzas y la gobernanza necesarias, así como la audiencia que utilizará los *insights* generados a partir de la explotación de datos. Pensar únicamente en proyectos de investigación es ignorar otro tipo de proyectos con enfoques multidisciplinarios o colaborativos que tienen la posibilidad de construir capacidades y comunidades alrededor de los datos. La decisión sobre qué tipo de proyecto adelantar dependerá en gran parte en los recursos necesarios para adelantar estos así como la problemática u objetivo en cuestión (*pensar en la información de la Plantilla I:Volviendo problemáticas de gestión pública en preguntas de datos*)

Para este punto, se sugiere al lector **retomar las tres prioridades expuestas** en la Plantilla I, diligenciada en el paso número uno de la sección 2.1 (ver página 21). Al elegir el prototipo de proyecto (distintas opciones listadas a continuación), tenga en consideración las siguientes instrucciones:

1. Lea los prototipos de proyecto a continuación, y pregúntese qué tipo de proyecto considera se podría llevar a cabo para cumplir con las prioridades delineadas anteriormente. Deseche aquellos en donde ninguna de sus tres prioridades vaya a ser cumplida. Paso siguiente, dentro de los posibles prototipos elija aquel en el mayor número de prioridades se cumplan, teniendo en cuenta variables clave como el tiempo, los recursos y la problemática que se busca resolver.

<table>
<tr><th colspan="3">Herramientas de Ciencia Ciudadana</th></tr>
<tr><td colspan="3">Este tipo de proyectos se enfocan en recolectar data y extraer conocimiento de una red de expertos y no expertos interesados en un tema. Principalmente estas herramientas se construyen para recolectar datos que son crowdsourced sobre un tema especifico, por ejemplo, permitir a usuarios de una aplicación notificar sobre una violación en los derechos humanos y georeferenciar esta denuncia.</td></tr>
<tr><td colspan="3">Recursos</td></tr>
<tr><td>Técnicos: plataforma/panel de control, diseño e interfaz centrados en el usuario; mantenimiento técnico</td><td>Humanos: especialistas calificados, diseñadores de encuestas, organizadores comunitarios, desarrolladores, científicos de datos, especialistas en ética y gobernanza de los datos, especialistas en comunicaciones</td><td>Data: datos procedentes de crowdsourcing, datos abiertos</td></tr>
<tr><td colspan="3">Diseño y Consideraciones:</td></tr>
<tr><td colspan="3">• Calidad de los datos recopilados puede variar al ser datos que una red de no-expertos voluntariamente entrega al sistema
• Protección de privacidad y consentimiento informado
• Riesgos de privacidad, al permitir que cualquier usuario del sistema pueda entrega información acerca otros
• Representación y participación, incluyendo que solo quienes tienen acceso a estos sistemas tienen la posibilidad de utilizarlos</td></tr>
<tr><td colspan="3">Ejemplo</td></tr>
<tr><td colspan="3">‘Nuestras aves, nuestro café’, Jardín, Antioquia.
Junto con los caficultores se recolectan datos para hacer un registro formal de todas aves que ellos conocen que habitan sus cafetales. Estos datos benefician a la comunidad tanto para educar a los niños como para formar guías turísticos. Por el otro lado, Cornell, encargado de la investigación, usa esta información para investigar incentivos para prácticas productivas sostenibles para el café generan un beneficio para la biodiversidad de este paisaje natural[11].</td></tr>
</table>

Retos de Datos		
Poner a disposición de una red dispersa y diversa de expertos académicos, profesionales y no profesionales, conjuntos de datos específicos para que realicen interpretaciones y generen conocimiento a partir de la explotación de datos.		
Recursos		
Técnicos: Conjuntos de datos API/ batch	**Humanos:** Científicos de datos, investigadores de ciencias sociales, especialistas en ética y gobernanza de los datos, desarrollo de asociaciones, participación de la comunidad	**Data:** Datos gubernamentales abiertos, datos del sector privado
Diseño y Consideraciones:		
• Acceso a los datos es una de sus mayores limitaciones - en varios casos se han utilizado datos de celulares, lo que hace el acceso un proceso complejo • Implementación de alianzas público-privadas se hacen necesarias, ya que usualmente se necesitan varios tipos de conocimiento, incluyendo información sobre los datos y la infraestructura. Muchas veces son realizados entre conjuntos de diferentes organizaciones, incluyendo universidades y centros de investigación • Tener en cuenta si se tiene el consentimiento informado de los individuos cuyos datos se va a utilizar, o si el uso de los datos está reglamentado por las políticas de privacidad y acceso por parte de quien tiene custodia de los datos		
Ejemplo		
Data4Refugees JEste reto de datos co-organizado por Türk Telekom, Bogazici University and Tübitak in collaboration with Fondazione Bruno Kessler (FBK), MIT Media Lab, Data-Pop Alliance, UNHCR, IOM e UNICEF puso a disposición CDRs anonimizados a un grupo de investigadores con el objetivo de producir nuevos conocimiento relacionados los refugiados Sirios en Turquía, específicamente relacionados con su seguridad, educación, salud desempleo e integración.		

Productos Inteligentes		
Proporcionar productos y conocimientos basados en conjuntos de datos cerrados para el análisis y la toma de decisiones		
Recursos		
Técnicos: Plataforma/panel de control, diseño e interfaz centrados en el usuario; mantenimiento técnico	**Humanos:** Científicos de datos, investigadores de ciencias sociales, especialistas en ética y gobernanza de los datos, profesionales en desarrollo de alianzas, participación de la comunidad	**Data:** Datos de propiedad/cerrados, análisis y visualización de datos
Diseño y Consideraciones:		
• Requiere de una interpretación de los datos hecha previa a la producción de indicadores, lo que puede conllevar a diferentes sesgos • Al ser un tipo de proyecto enfocado en agregaciones e indicadores precomputados se puede complicar el acceso a datos muy granulares • Se hace esencial implementar mecanismos de transparencia y rendición de cuentas		
Ejemplo		
SmartSteps de Telefonica Es una plataforma que permite a Telefónica extraer *insights* de patrones globales, filtrados por diferentes tipos de personas para ayudar a organizaciones afianzar sus propuestas comerciales. Es un ejemplo donde datos de empresas públicas se pueden utilizar en proyectos, sin embargo estos solo están disponibles a un costo.		

[11] https://www.javeriana.edu.co/pesquisa/tag/ciencia-ciudadana/

Proyectos de Investigación		
Llevar a cabo investigación basada en datos para obtener conocimiento, a través de la creación de nuevas metodologías y productos basados en datos		
Recursos		
Técnicos: Conjuntos de datos, almacenamiento y procesamiento seguros de datos	**Humanos:** : Científicos de datos, investigadores de ciencias sociales, especialistas en ética y gobernanza de los datos, profesionales en desarrollo de Alianzas, participación de la comunidad	**Data:** Datos gubernamentales abiertos, acceso controlado a datos privados, teledetección, datos de fuentes tradicionales (incluyendo sistemas, registros, datos de encuestas, etc.)
Diseño y Consideraciones:		
• Las iniciativas de investigación requieren de estudio de campo y de mercado (en otras palabras si ya existe dicha iniciativa) para evitar duplicidad de esfuerzos • Estos estudios deben velar por la validez, representatividad y parcialidad de sus resultados • En varios casos requieren de la justificación del beneficio público de llevar a cabo estas investigaciones		
Ejemplo		
Tackling Climate Change with Machine Learning *Abordando el cambio climático con aprendizaje de máquinas* El aprendizaje de máquinas, o *machine learning*, es una herramienta útil para pensar en soluciones y aplicaciones para combatir el cambio climático. Este proyectos de investigación es un intento de mapear todos los usos actuales de esta tecnología para cuestiones de cambio climático; los autores proponen, que al analizar los datos recogidos sobre el tema, el aprendizaje de máquinas puede mejorar la ingeniería de vehículos, habilitar infraestructura inteligente y proporcionar información relevante para las políticas, entre otros. Un ejemplo de los casos de usos estudiado, es la posibilidad de analizar los patrones de movilidad de una ciudad, para crear modelos de demanda, y a partir de ellos planificar una nueva infraestructura para reducir la duración de los viajes e influir en la elección de modos de transporte, para de esta forma reducir los gases de efecto invernadero[12].		

Coaliciones y colaboraciones		
Fomentar la colaboración de diversas partes interesadas en torno a aplicaciones y políticas de datos		
Recursos		
Técnicos: Plataforma de comunicaciones / panel de control, mantenimiento técnico	**Humanos:** : Científicos de datos, investigadores de ciencias sociales, especialistas en ética y gobernanza de los datos, desarrollo de alianzas, participación de la comunidad	**Data:** Dependerán del sector
Diseño y Consideraciones:		
• Importante evaluar quienes son parte de las alianzas y los incentivos que cada uno tiene en el proyecto • Considerar las diferentes metas a corto y largo plazo para crear una coalición que funcione a lo largo del tiempo		
Ejemplo		
CGIAR CGIAR es una asociación de investigación global, que busca un futuro alimentario seguro, con el fin de reducir la pobreza, mejorar la seguridad alimentaria y nutricional y mejorar los recursos naturales. Con este fin en mente, la asociación creó una plataforma, con duración de 5 años (2017-2021), para el uso de Big Data en la agricultura. Su objetivo principal es el de aprovechar las capacidades de big data para acelerar y mejorar el impacto de la investigación agrícola internacional. Por lo tanto, por medio de esta se busca abrir y compartir datos agrícolas, para así convocar a socios para desarrollar ideas y proyectos innovadores[13].		

[12] https://arxiv.org/abs/1906.05433
[13] https://bigdata.cgiar.org/

Herramientas de código abierto		
Generar conocimiento e impulsar modelos colaborativos para el uso y análisis de los datos desarrollar productos basados en estos		
Recursos		
Técnicos: Dependen del contexto	**Humanos:** : Científicos de datos, investigadores de ciencias sociales, especialistas en ética y gobernanza de los datos, desarrollo de alianzas, participación de la comunidad	**Data:** Dependerán del sector
Diseño y Consideraciones:		
• Este tipo de proyectos generalmente son de aplicabilidad en varios contextos, dado la replicabilidad del código y potenciales casos de uso que emergen a partir de estos • Importante considerar la no duplicación de esfuerzos existentes • Contribuye al fortalecimiento de la comunidad de investigación y fomenta la creación de alianzas		
Ejemplo		
Software Público de Colombia El Software Público de Colombia es una iniciativa, impulsada por el Gobierno Nacional de Colombia, que pretende fomentar la transferencia tecnológica. En esta, las personas y/o entidades pueden publicar, compartir, mejorar y reutilizar aplicaciones desarrolladas con recursos del estado, cuyas licencias son de código abierto. Lo que pretende la plataforma es, tanto racionalizar la inversión, como sacar provecho de los desarrollos ya existentes, para así generar iniciativas de innovación pública, a través de la co creación entre las entidades públicas y privadas y ciudadanos y comunidades[14].		

Plataformas		
Proporcionar herramientas digitales fáciles de usar para recopilar, usar y analizar datos derivados de diversas fuentes		
Recursos		
Técnicos: Plataforma / panel de control, mantenimiento técnico	**Humanos:** : Científicos de datos, investigadores de ciencias sociales, especialistas en ética y gobernanza de los datos, desarrollo de asociaciones, participación de la comunidad	**Data:** Datos de encuestas, datos abiertos, datos procedentes de crowdsourcing
Diseño y Consideraciones:		
El diseño e interfaz de estas plataformas debe estar centrada en el usuario, así como consideraciones relacionadas a la accesibilidad y facilidad de uso de estas.		
Ejemplo		
ESRI y ArcGIS ArcGis es un sistema en el cual se puede recopilar, organizar, administrar, analizar, compartir y distribuir información geográfica. La plataforma ESRI, permite crear y utilizar sistemas de información geográfica (SIG) y poner dicho conocimiento al servicio y acceso de cualquier usuario. De esta forma, se fomenta la creación de información colaborativa - en este caso, la plataforma creada lleva registro de los casos de coronavirus en Colombia a través de un dashboard interactivo. Así mismo, se puede acceder a él por medio de la web, dispositivos móviles o computadores[15].		

Ejemplo: Desarrollo de un caso de uso - Más Familias en Acción

Según se determinó en la primera plantilla, el proyecto se enfocará en la efectividad de programas de transferencias monetarias condicionadas. Para ello, se requiere tener acceso a los datos de consumo de productos a través del uso de tarjetas bancarias. Al observar los tipos de proyecto propuestos en esta sección, se encuentra que el proyecto se ajusta a las características de un proyecto de investigación.

[14] https://www.softwarepublicocolombia.gov.co/es/content/herramientas
[15] https://esri.co/covid-19/

Esto se debe a que el proyecto parte del supuesto de la falta de información sobre los patrones de consumo de los beneficiarios de MFA. Para este proyecto no se propone un producto inteligente puesto se considera mejor evaluar la utilidad y escalabilidad de la propuesta, antes de transformarla en un producto. Esto se logra con mayor facilidad con el formato de proyecto de investigación. Sin embargo, es aconsejable que durante el proyecto de investigación se considere la transformación del estudio en un producto replicable.

En el contexto de este proyecto, que involucra el tratamiento de datos sensibles de poblaciones vulnerables provenientes del gobierno y de entidades bancarias, es importante que el equipo que ejecute el proyecto sea multidisciplinario. Es decir, que cuente con personas competentes en ingeniería de datos, para crear la estructura necesaria para unir las bases de datos; ciencia de datos y estadística, para efectuar el análisis de los mismos; ciencias sociales, para asegurar que el análisis e interpretación de los datos sean entendidos correctamente dentro del contexto de poblaciones vulnerables y de alivio de pobreza; ética y gobernanza de datos, con el fin de evitar que el proyecto ponga en riesgo la privacidad de la información de estas personas.

Adicionalmente, durante el desarrollo del proyecto es importante incluir a la comunidad de estudio. Esto con el fin de incorporar la voz de los beneficiarios del programa en la interpretación de sus datos. Dependiendo de la duración del proyecto y los fondos disponibles, se puede incluir a la comunidad a través de un panel para la interpretación de resultados, con representantes en la comunidad de estudio. También, se puede pensar en la realización de estudios cualitativos para informar los hallazgos del estudio cuantitativo.

4. PROPUESTA DE PROYECTO

Finalmente, y para finalizar la formulación de su proyecto, es clave ahondar en el potencial valor agregado de este, pensando en cómo es una innovación técnica relevante, así como determinar retos a priori y cómo sobrellevarlos. Este proceso busca consolidar la propuesta de valor del proyecto, guiando al lector a pensar de manera macro, si su proyecto puede ser una iniciativa factible para su organización. La plantilla abajo se enfoca en detallar la relevancia de su proyecto en el contexto institucional en el que se ubica, anticipándose a los riesgos y desafíos que podrá enfrentar el desarrollo del proyecto, y que deben ser resueltos. Para esta parte, es fundamental tener en cuenta los desafíos que podrían presentarse con respecto a la capacidad humana y colaboraciones, las herramientas y capacidades técnicas y los aspectos de gobernanza, legales y éticos de una potencial implementación del proyecto.

Para diligenciar la siguiente plantilla, responda a cada una de las preguntas, teniendo en cuenta el contexto institucional en el que se ubica. Sintetice el valor agregado y la innovación de la propuesta, haciendo una breve mención a los riesgos y desafíos y los mecanismos de mitigación que se llevarán a cabo con el proyecto.

Plantilla 3: Propuesta de proyecto	
Determine ¿quién es su audiencia y cuál es la mejor manera de llegar a ellos?	
1. Contexto y relevancia	2. Innovación
¿Cuál es la relevancia de este proyecto a nivel local, nacional o global? *¿Dónde se va a implementar el proyecto y porqué (enfoque geográfico)?* *¿Quiénes son los potenciales beneficiarios (población objetivo) de este proyecto?* *¿Cuáles son los objetivos operacionales específicos del proyecto, y porqué son relevantes?*	*¿Cuál es la innovación técnica más relevante para su proyecto y porque?* *¿Qué valor agregado aporta su propuesta para mejorar la forma en la que actualmente se están haciendo las cosas?.* *¿De qué manera presenta su proyecto una innovación en términos de metodología? De datos? De resultados y productos finales?*
3. Retos y cómo sobrellevarlos	4. Propuesta de valor
¿Cuáles son los riesgos y desafíos que puede enfrentarse durante el desarrollo e implementación del proyecto? *¿Algún reto específico relacionado a los datos o tecnologías que está proponiendo utilizar?* *¿Cuales son las soluciones previstas para abordarlo?*	*¿Cómo espera que las partes interesadas adopten su proyecto?* *¿A corto plazo, cuáles los próximos pasos para el desarrollo e implementación del proyecto?* *¿Cuál es la visión a largo plazo del proyecto?* *¿Cómo encaja este proyecto entre las prioridades del gobierno nacional? Del gobierno local?*

Ejemplo: Desarrollo de un caso de uso - Más Familias en Acción

La audiencia del proyecto está compuesta por la entidad a cargo de Más Familias en Acción, el Departamento de Prosperidad Social, la entidad que brinda los datos bancarios y los beneficiarios del programa.

Teniendo en cuenta que el proyecto surge dentro de Prosperidad Social, se piensa en la mejor forma de acercar al DPS con la entidad bancaria. Esto se puede lograr a través del establecimiento de un acuerdo de confidencialidad para que se puedan compartir los datos entre las entidades. Este proceso se puede agilizar si los expertos en datos proporcionan un protocolo y código para la anonimización de los datos de ambas entidades, lo cual aligera la carga legal para el tratamiento de datos sensibles. De la misma forma, una buena forma de llegarle a los beneficiarios del programa es a través de un convenio interno con Prosperidad Social, realizando un acuerdo en donde se pauten las medidas de confidencialidad de la información, anonimidad de los informantes para proteger su seguridad durante y después de la participación en el proyecto.

Este proyecto puede ser altamente relevante para el Gobierno, ya que Familias en Acción tiene cobertura nacional y usa recursos públicos. Entender los usos que las personas le dan a las transferencias es de interés para el Gobierno de Colombia, puesto que permite entender el consumo que se promueve a través de los fondos para programas sociales.

El alcance geográfico del proyecto se delimita a Bogotá, puesto que una de las prioridades de la propuesta es la realización del cruce de datos entre entidades bancarias y del Gobierno. Ya que para tener información sobre las transacciones bancarias, es necesario que los beneficiarios del programa tengan una cuenta de banco, se asume que hay una mayor probabilidad de encontrar una población bancarizada en áreas con mayor urbanización, por lo cual Bogotá podría ser una buena opción[16].

En el proyecto se han identificado elementos innovadores, como el uso de nuevas fuentes de datos para resolver interrogantes sobre el consumo de las personas. El carácter innovador de usar los datos bancarios se basa en que la información sobre las transacciones bancarias se actualiza de forma casi simultánea; mientras que la información sobre el consumo recogida a través de encuestas resulta costosas y se realiza con menor frecuencia. Es decir, que el proyecto tiene un potencial importante para que el gobierno tenga acceso a más información, más reciente, a un menor costo. En general, usar aprendizaje de máquinas no supervisado permite procesar una mayor cantidad de información que con los métodos estadísticos tradicionales, integrando más variables y estableciendo patrones de consumo combinando toda la información.

El reto más importante para el proyecto es lograr la autorización para acceder a datos bancarios. Por ello, se propone trabajar un convenio donde ambas organizaciones anonimicen los datos con un método determinístico, el cual permite que un mismo número de cédula tenga el mismo número anónimo. Este proceso facilita la unión de las bases de datos. Igualmente, se pueden realizar transformaciones aleatorias a los datos de las personas sin afectar sus valores promedio (agregar ruido), para evitar identificar patrones que llevaran a la reidentificación de las personas[17].

Otra medida puede ser la agregación de los datos a nivel de barrio. De esta manera se podrían comparar los hábitos de consumo promedio de las personas, con los indicadores socioeconómicos promedio en cada barrio. La agregación evita la identificación de cualquier individuo, asegurando la protección de la privacidad por defecto.

A largo plazo, se busca que el DPS tenga acceso a una herramienta que en caso de una crisis permita ver los cambios en las tendencias de consumo de sus beneficiarios. De esta manera, se pueden idear mecanismos para ayudar a promover hábitos de consumo saludables. Por ejemplo, si en una crisis se detecta una baja en la compra de alimentos, se puedan hacer campañas para soportar la nutrición de la familia y especialmente la infantil, que es fundamental para Más Familias en Acción.

[16] Dicha información se corrobora con estudios de inclusión financiera. Como el de reporte de inclusión financiera del 2018, donde se afirma que Bogotá tiene la mayor tasa de inclusión financiera (98%) del país. Superintendencia Financiera, 2018. “Reporte de Inclusión Financiera”.

[17] Para mayor información ver el código abierto desarrollado por el Banco Mundial y otros, en Statistical Disclosure Control for Microdata, accesible en el siguiente link.

Recomendaciones Transversales: Formulación del proyecto

- En esta etapa es importante asegurar que el proyecto de analítica planteado, responda correctamente a las problemáticas sociales identificadas y que los objetivos formulados estén al alcance de lo que se puede lograr con los datos. Para hacer esta determinación piense qué información se comunica a través de los datos, y si esa información conlleva a producir *insights* sobre su área de impacto (*es posible utilizar datos de x para determinar y?*). De este balance depende la efectividad del proyecto y el manejo de las expectativas. Aclarar estas últimas es imprescindible, pues serán las encargadas de definir, al finalizar el proyecto, si se cumplió o no con lo que se esperaba.
- Sí su proyecto requiere la contratación de terceros para la ejecución de este, se recomienda que los términos de referencia o lineamientos que guíen el proyecto no estén formulados sólo para la consecución de un objetivo concreto, pero que consideren también como los modelos, algoritmos y códigos desarrollados por expertos podrían ser reutilizados por la entidad quien adelanta el proyecto. Es relevante desarrollar proyectos replicables, escalables que puedan contribuir a la continuidad de análisis y construcción de habilidades. En lo posible, incluir en los entregables de estos proyectos procedimientos que construyan habilidades dentro de las organizaciones relevantes, incluyendo gestión del conocimiento y documental. Así mismo, priorizar entregables que sean de fácil uso para grupos no-técnicos. Por ejemplo, el uso de tableros de control o balanced scorecards facilitan el acceso a los resultados del ejercicio analítico para la toma de decisiones.
- Al escoger los datos que utilizará en su proyecto, es necesario entender qué tipo de datos son y quién tiene acceso a dicho tipo de datos, además de revisar si el uso previsto que le dará a los datos cuenta con los permisos o restricciones legales para el uso de estos (son datos públicos? Están abiertamente disponibles?). Por ejemplo, saber para qué fueron creados y si se pueden o no usar para algo diferente a este objetivo, si las personas quienes los suministraron están de acuerdo con su uso, si deben ser anonimizados, entre otros.
- Es fundamental conformar equipos multidisciplinares con capital humano experto tanto en las preguntas de negocio objetivo del proyecto como en datos para el uso y aprovechamiento de los modelos de análitica.
- Asegúrese de que su proyecto mantenga siempre un enfoque de derechos, teniendo como prioridad el respeto de los derechos de la población que está siendo analizada.

Sección 2.2: Consideraciones transversales al proyecto

1. RIESGOS Y RETOS EN PROYECTOS DE BIG DATA

Es imprescindible que todo proyecto de datos, especialmente de Big Data y/o IA, incluya una etapa de revisión ética. En esta, se deben analizar los posibles desafíos del uso de fuentes no tradicionales como lo son la representatividad y la participación, y se deben establecer soluciones o estrategias de mitigación para los mismos, priorizando la protección de los derechos de los sujetos de datos. Esta etapa cobra aún más importancia cuando se tiene en cuenta que usualmente, los datos de Big Data recogen información sobre comportamientos humanos de manera pasiva, a través del uso que usuarios le dan a diferentes dispositivos digitales. Por lo tanto, se debe garantizar el respeto a la privacidad y al consentimiento informado de los individuos cuyos datos estén siendo usados en la investigación.

Además del consentimiento informado, algunas formas de respetar la privacidad incluyen la anonimización de los individuos o la agregación de los datos a un nivel en el que ningún individuo pueda ser identificado (datos a nivel de ciudades por ejemplo). En cuanto a la representatividad, es fundamental reconocer los sesgos y si es posible disminuirlos por medio de otras fuentes de información. En el siguiente recuadro, se guía a los usuarios del manual a través de diferentes factores clave de la ética y la participación a tener en cuenta durante todo el ciclo del proyecto.

Para llenar la siguiente plantilla, pregúntese lo siguiente:

Privacidad y consideraciones jurídicas

¿La información recopilada expone a los usuarios (u otros) a riesgos si esta fuese vista por personas no autorizadas?

¿Cuáles son las medidas preventivas para evitar el acceso no autorizado durante el intercambio de datos confidenciales/ protegidos, a fin de proteger la privacidad individual y colectiva?

¿Qué marcos normativos regulan el uso de datos individuales y colectivos de este proyecto?

Protección de datos y consentimiento informado

¿Los usuarios conocen cuándo y por qué sus datos serán recopilados y cómo se van a emplear? ¿Cuáles son las consecuencias para el usuario si niega el consentimiento? Para el proyecto? ¿Hay alternativas de recolección de datos si usuarios optan por no participar?

Si se utilizan conjuntos de datos previamente recopilado, ¿cumple el uso actual de dichos datos a los principios y usos acordado en principio por el usuario? ¿se ha comunicado con los sujetos de datos los usos alternativos o futuros de los datos?

¿Cómo asegura usted la protección y el consentimiento a datos pasivos generados por personas en espacios públicos (por ejemplo, cámaras de tráfico, datos de tránsito, etc)?

Representatividad y participación

Sesgo de la selección:

¿de quién se incluyen los datos? ¿de quién se excluyen los datos? ¿es consciente de algún sesgo en sus datas? ¿el proyecto considera el sesgo de la fuente de datos y/o intenta evaluar y comunicar dicho sesgo?

Coherencia/calidad de los datos:

¿tiene usted información de todos los usuarios de la misma forma?

Participación de la comunidad: ¿los beneficiarios a quienes va destinado el proyecto están involucrados en el proceso de observaciones/decisión?

- ¿Existe el riesgo de que el análisis produzca resultados (consecuencias, recomendaciones, precios, etc.) que afecten desproporcionadamente a ciertos individuos o grupos (por ejemplo, discriminación algorítmica[18])?

Gobernanza de datos

¿Se están integrando los datos con otros conjuntos de datos? ¿Cómo se evalúan los nuevos riesgos creados al integrar múltiples conjuntos de datos?

¿Qué marcos guían las decisiones sobre las aplicaciones y repercusiones de la recopilación,

almacenamiento, análisis y acceso de los datos?

¿Qué marcos de gobernanza de los datos y del proyecto puede poner en lugar para mitigar riesgos éticos?

Transparencia

¿Está claro para los usuarios cuál es el propósito y uso para el que están aportando sus datos?

¿Los metadatos sobre las metodologías, el consentimiento del usuario y los usos aceptables

están siendo registrados junto con el conjunto de datos en sí?

¿Las metodologías que piensa utilizar son replicables/ de código abierto? ¿O hay un mecanismo para la revisión/aprobación/ validación de la comunidad de estas metodologías?

¿Alguno de los datos o resultados está sujeto a las leyes de libertad/protección de la información? ¿Qué riesgos podría crear esto?

¿Cuáles son los mecanismos de confianza para validar y revisar las decisiones que sustentan el marco de gobernanza del proyecto?

Plantilla 4: Ética y Participación: Retos y Soluciones
Consideraciones legales y de privacidad -
Protección de datos y consentimiento informado -

[18] La discriminación algorítmica hace referencia a las ocasiones en que los algoritmos presentan sesgos en el momento de tomar decisiones, las cuales están basadas en algoritmos supuestamente neutros. Estos sesgos se han encontrado, por ejemplo, en la publicidad en línea, las acciones de reclutamiento y las tarifas de los servicios. Todavía es un reto tener certeza de que los algoritmos utilizados en la inteligencia artificial serán justos con las personas, sobre la decisiones que toman. (Byrnes, N. 2016. MIT Technology Review. El día que los algoritmos empezaron a discriminar a la gente sin querer.)

Representatividad y participación -
Gobernanza responsable -
Transparencia -

Realizar el cruce de datos de consumo en tarjetas de banco y los de Más Familias en Acción es una operación sensible que podría infringir la privacidad de las personas. Como estrategia para asegurar la protección de los datos, se propone crear un convenio en el que ambas organizaciones establecen un protocolo de anonimización y modificación de datos (añadiendo ruido aleatorio).

En cuanto al consentimiento informado, es necesario verificar los permisos que dan los beneficiarios de MFA para la utilización de sus datos, al igual que la autorización que se le da a los bancos. De no ser posible hacer el uso de los datos anonimizados, se puede seleccionar una muestra de beneficiarios y solicitarles autorización expresa para usar sus datos en el estudio.

La representatividad y participación se tuvo en cuenta al incluir entrevistas y/o un panel con los beneficiarios del programa para discutir los hallazgos del proyecto. Además, para promover la gobernanza responsable del proyecto, se pueden incluir en dicho panel representantes del Departamento de Prosperidad Social y del banco. El objetivo de este panel sería el de supervisar las decisiones éticas, de uso de datos, análisis y divulgación de resultados.

Una consideración de representatividad adicional es recalcar que el proyecto, por sus migajas, excluye aquellos individuos que no están bancarizados. Es decir, que sus resultados no se pueden generalizar a la población no bancarizada. Para crear transparencia en el proyecto, toda divulgación de resultados debe incluir la sección de limitaciones. Adicionalmente, se debe dejar un repositorio con el código para que el DPS pueda replicar el estudio.

RECOMENDACIONES TRANSVERSALES: ÉTICA Y PARTICIPACIÓN

Consideraciones Legales y de Privacidad

- La protección a la privacidad de las personas y grupos debe asegurarse en todas las etapas del ciclo de vida de los datos, y del proyecto, a través de diferentes mecanismos. Esto se puede lograr a través de la agregación de información, la anonimización de los datos, y a través de diferentes estándares de procesamiento de datos - en última esto dependerá del tipo de datos y el objeto del proyecto.
- Es necesario encontrar, seleccionar, entender y seguir los lineamientos y marcos legales aplicables y relacionados al uso de los datos. Este proceso puede ser particularmente complejo cuando los datos a utilizar en un proyecto provienen de múltiples fuentes o colaboradores, y están sujetos a diferentes lineamientos. Sin embargo, es imprescindible reconciliar los diferentes marcos normativos y regulaciones aplicables a estas fuentes. Esto también puede ocurrir cuando se extraen datos provenientes de varios países u organizaciones con marcos normativos diferentes aplicables a la explotación de datos (por ejemplo, datos sujetos al GDPR).
- Considerar riesgos imprevistos en materia legal y de privacidad que puedan surgir al combinar diferentes tipo de datos - por ejemplo, pensar si la unificación de dos bases de datos permite la reidentificación de datos previamente anonimizados.

Retos en Representatividad y Participación

- Identificar y tomar en cuenta los posibles sesgos – tanto en recopilación, análisis y comunicación de resultados — y verificar y balancear la utilidad de ciertas herramientas (por ejemplo, parcialidades latentes o ignoradas).
- Tener en cuenta que ciertas fuentes de datos pueden ser representativas únicamente de una parte de la población. Por ejemplo, los estudios que traten con datos de redes sociales deben considerar que en países o ciudades de ingresos medios y altos, la penetración y uso de redes sociales es mayor que en países o ciudades de ingreso bajos, donde el total de la población capturada en la fuente de datos puede que no sea representativa de toda la población.
- En lo posible, asegurar que su proyecto incluya en todo el ciclo de vida del proyecto la participación de los grupos de interés estudiados. Busque generar acceso a los resultados de su proyecto a comunidades objeto del estudio.

[19] Dittrich, D. and Kenneally, E. 2012. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. United States: The Department of Homeland Security; European Commission. 2019. Ethics guidelines for trustworthy AI. European Union.

Retos en la Protección de Datos y el Consentimiento

- Definir y asegurar la protección de los datos en torno a la recopilación, almacenamiento, y uso de los datos a través del ciclo de vida del proyecto. Esto requiere cumplir con normativas de protección de datos, incluyendo revisión de consentimiento informado y permisibilidad de utilizar los datos para el objeto del proyecto.

Riesgos en la Gobernanza Responsable de los Datos

- Definir mecanismos de gobernanza y supervisión para el proyecto, siguiendo los principios de '*do no harm*'. Revisar diferentes marcos de gobernanza incluyendo el '*Menlo Report*', la guía ética de la Unión Europea para una Inteligencia Artificial confiable, o los estándares de organizaciones tales como el IEEE[19].
- Pensar si se hace necesario crear un organismo de supervisión y gobernanza ética del proyecto que se asegure del seguimiento de estándares, éticos y legales para el uso de datos durante todo el ciclo de vida del proyecto.

Retos en Transparencia

- Equilibrar las motivaciones complementarias u opuestas de los generadores de datos, usuarios, y científicos de datos involucrados en el proyecto
- Establecer y hacer cumplir una gobernanza del proyecto transparente con procesos, información y componentes abiertos en la mayor medida posible.

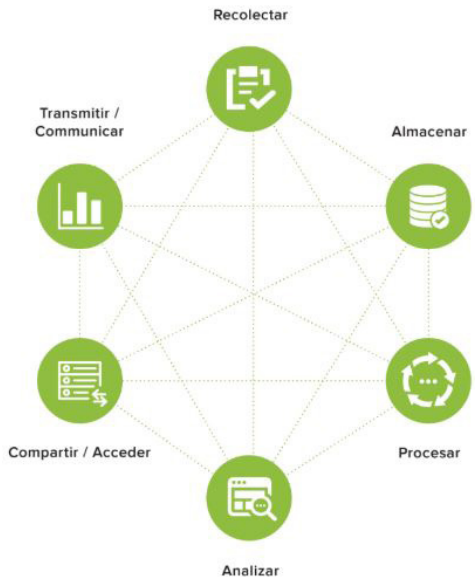
Sección 2.3: Implementación del proyecto

1. CADENA DE VALOR DEL BIG DATA: ETAPAS, HERRAMIENTAS Y RECOMENDACIONES PARA LA IMPLEMENTACIÓN DE PROYECTOS

Al diseñar e implementar un proyecto de Big Data, se hace esencial pensar en la cadena de valor del Big Data como las diferentes etapas que se deben realizar, y al mismo tiempo debidamente planear, para transformar los datos crudos en información accionable. No se trata de sólo priorizar la explotación de datos o la Inteligencia Artificial por su carácter innovador, más bien, se trata de evaluar los beneficios de utilizar estos, de pensar en cómo y qué tecnologías pueden ayudar, de seleccionar qué datos pueden ser útiles para ciertos temas, y entender cómo se puede crear un marco robusto para utilizar la evidencia producida en un proyecto de datos.

La idea es lograr gestionar y coordinar los datos utilizados en el marco de un proyecto desde el punto de generación u obtención de los datos, hasta el momento en donde estos son utilizados para el objeto del proyecto. Entre otros, esto implica pensar en el acceso a los datos - y qué herramientas se utilizarán - o en otros casos, también implica crear los procesos de recolección de estos. El siguiente diagrama representa estas etapas de uso de datos y proporciona ejemplos clave de las herramientas que se relacionan con cada una de estas. Las etapas no necesitan darse en un orden específico, y pueden repetirse, omitirse u ocurrir simultáneamente. De hecho, gran parte del diseño de proyectos implica elegir la secuencia y combinación apropiada de etapas para ejecutar un proyecto viable que atienda los requerimiento éticos, busque propender un impacto positivo, y sea escalable a futuro. Esto, teniendo en cuenta las limitaciones, necesidades y capacidades de los diversos interesados y los beneficiarios[20].

Figura 2. Cadena de valor del Big Data



La cadena de valor ilustrada arriba delinea las diferentes actividades, procesos e interacciones que deben ocurrir y deben

[20] Abou Zakaria Faroukhi et al., "Big data monetization throughout Big Data Value Chain: a comprehensive review", Journal of Big Data 7, núm. 1 (el 8 de enero de 2020): 3, https://doi.org/10.1186/s40537-019-0281-5.

[21] H. Gilbert Miller y Peter Mork, "From Data to Decisions: A Value Chain for Big Data", IT Professional, 2013, https://doi.org/10.1109/MITP.2013.11.

de organizarse en el planteamiento estratégico de un proyecto. Esto implica aislar y entender los procesos que deben ocurrir y entender el impacto de estos sobre otros eslabones de la cadena para así crear una serie de procesos iterativos y bien organizados para extraer el valor de los datos[21].

Este proceso se concentra principalmente en seis etapas[22].

Recolectar

Quizá la etapa más importante, pues involucra asegurar el acceso a los datos en crudo. Alternativamente, esta etapa también puede incluir poner en marcha los diferentes procesos para la generación de los datos.

Herramientas para recolectar datos de personas (crowdsourcing)	
U Report VOICE MATTERS	Sistema de recolección de datos creado por UNICEF para mejorar la participación ciudadana a través de encuestas y alertas enviadas por SMS
 OpenStreetMap	Proyecto colaborativo de datos geográficos para recolectar datos de manera manual, a través de GPS o fotografía aérea
 Ushahidi	Herramienta gratis para recolectar y gestionar datos utilizando dispositivos móviles.
 KoBoToolbox	Software de código libre que permite que observadores neutros envíen reportes utilizando sus celulares o a través de internet. Se ha utilizado en el caso de respuestas a emergencias y violaciones de derechos humanos, entre otras.
Herramientas para recolectar datos de dispositivos	
bandicoot	Una herramienta de código libre de Python que permite extraer atributos de metadatos estándares de CDRs[23].
funf Open Sensing Framework	Un marco de procesamiento y recolección de datos para dispositivos móviles. El objetivo de este es proporcionar un conjunto de funcionalidad de código abierto que permite la recopilación, carga y configuración de una amplia gama de datos accesibles a través de teléfonos móviles. [24]

[22] Información para esta sección se enfoca en el Toolkit de Data-Pop Alliance (2017), en "Oportunidades y requerimientos para aprovechar el uso de Big Data para las estadísticas oficiales y los ODS en América Latina" de Data-Pop Alliance (2016) y Faroukhi et al., 2020.

[23] Yves-Alexandre de Montjoye, Luc Rocher, y Alex Sandy Pentland, "bandicoot: a Python Toolbox for Mobile Phone Metadata", Journal of Machine Learning Research 17, núm. 175 (2016): 1–5.

[24] Para más información, ver https://www.funf.org/about.html

Compartir y acceder

Involucra los procesos de hacer que la información almacenada se haga disponible para otros, u obtener datos compartidos por otros. Entre otros, incluye determinar acceso a los datos para sistemas compartidos, incluyendo también restricciones de seguridad y privacidad del uso de los datos.

Herramientas para compartir datos y código	
GitHub	Plataforma que ofrece un sistema de control de versiones, principalmente utilizado en la creación de código fuente para programación.
	Sistema que alimenta a una base de datos con los tweets publicados en Twitter casi a tiempo real, filtrando por criterios predeterminados.
Herramientas para compartir datos confidenciales	
OPAL open algorithms for better decisions	El proyecto opal es una innovación socio-tecnológica para compartir datos de empresas privadas detrás de sus firewalls de una manera segura y participativa[25].
OPENPDS/SA Personal Data with Privacy	El Personal Data Store (PDS) está propuesto como un repositorio centralizado de los datos de un usuario móvil, que permite a éste ver los datos recolectados sobre su comportamiento en línea y conocer los usos que se le están dando a estos.

Almacenar

Incluye no sólo el almacenamiento, sino la preparación y gestión de la arquitectura sobre la cual se albergan los datos. Entre estas, se debe considerar la gestión de grandes volúmenes de datos, y la manera de alimentar los sistemas donde se almacenarán los datos. La transferencia de los datos a un sistema de almacenamiento se pueden hacer en una sola entrega o *batch*, en varias entregas pequeñas (*micro-batch*) o en *stream*, lo que quiere decir que diferentes fuentes de datos alimentan el sistema en tiempo real. Además, el importante tener en cuenta la manera en la que se realizará la transferencia, incluyendo la infraestructura de red utilizada para transferir los datos.

Herramientas para almacenar Grandes volúmenes de datos	
hadoop	Una colección de utilidades de software de código libre que utiliza una red de computadores para procesar grandes cantidades de datos. Este proporciona un framework de almacenamiento y procesamiento distribuido.
MySQL®	Un sistema de gestión de bases de datos relacionales que utiliza el *Structured Query Language.*

	Servicio ofrecido por Amazon que ofrece almacenamiento en la nube de objetos a través de una interfaz en la web. Sus servicios son escalables y globales.
	Sistema de almacenamiento de archivos de Google que permite a los usuarios sincronizar servidores a través de diferentes equipos.
Datos confidenciales de forma segura	
	Una cadena de bloques (*blockchain*) es una estructura de datos en la que la información contenida se agrupa en conjuntos (de bloques) los que se les añade metadatos relativas a otro bloque de la cadena anterior en una línea temporal, de manera que gracias a técnicas criptográficas, la información contenida en un bloque solo puede ser repudiada o editada modificando todos los bloques posteriores.

Procesar

Esta etapa incluye tanto el procesamiento de los datos como el preprocesamiento de estos. Para el preprocesamiento de los datos, algunos de los procesos que pueden incluirse son la limpieza de datos, procesos de validación y transformación, reducción de dimensionalidad e integración de diferentes fuentes para almacenar, posiblemente en un data lake[26]. Limpiar los datos consiste en remover valores faltantes, o completar éstos con valores agregados, promedios o el valor más probable para el atributo. Reducción de los datos consiste en asumir que los datos de una fuente siguen un modelo específico, para luego estimar y almacenar sólo los parámetros de este modelo (puede ser útil para reducir el volumen de los datos). La transformación de los datos puede incluir convertir los datos en bruto a diferentes formatos, incluyendo la integración de diferentes fuentes.

Herramientas para ejecutar instrucciones (en máquinas)	
	Colección de servicios de computación en la nube pública ofrecidas a través de Amazon.com
	Protocolo de código abierto descentralizado que permite realizar computaciones sobre datos encriptados
	Servicio web que permite crear y programar acciones para automatizar diferentes tareas y acciones

[25] https://www.opalproject.org/about-opal

[26] Un data lake es un repositorio de almacenamiento centralizado que almacena una gran cantidad de datos en bruto (estructurados y no estructurados) a cualquier escala. Este, permite importar datos en tiempo real, sin importar su cantidad. Así mismo, permite que se pueda acceder a los datos en las herramientas y marcos analíticos de preferencia, sin que sea necesario tener que mover los datos a un sistema de análisis separado para su ejecución. ("Data Lake: Definición, Conceptos Clave y Mejores Prácticas" s/f; What is data lake? AWS. s.f.).

Herramientas para darle sentido a datos sin procesar	
OpenRefine	Aplicación de desktop de código libre que permite limpiar y transformar datos a otros formatos, en otra palabra *data wrangling*
data BASIC.io	Suite de herramientas de uso fácil para principiantes sin experiencia datos.

Analizar

Esta etapa involucra analizar datos pre procesados y almacenados para encontrar correlaciones, identificar patrones y crear *insights* accionables. Entre las funciones analiticas de los modelos de big data se encuentra la función descriptiva, predictiva (para hacer inferencias acerca de las condiciones actuales y predicciones sobre eventos futuros), prescriptiva (para hacer inferencias causales) y discursiva o diagnóstica (para identificar las causas que conllevan a un resultado).

Herramientas para programación y análisis estadístico	
MATLAB	Sistema de cómputo número con su propio sistema de programación, que permite la manipulación de matrices, la representación de datos y funciones, la implementación de algoritmos, la creación de interfaces de usuario (GUI) y la comunicación con programas en otros lenguajes y con otros dispositivos.
python	Lenguaje de programación interpretado (i.e permite la legibilidad de su código)
R	Lenguaje y entorno de programación con enfoque en análisis estadístico y visualizaciones.
pandas	Biblioteca de software para manipulación y análisis de datos para el lenguaje de programación Python
Herramientas para darle sentido a datos sin procesar	
ArcGIS ESRI	Software de sistema de información geográfica que agrupa varias aplicaciones para la captura, edición, análisis, tratamiento, diseño, publicación e impresión de información geográfica
APACHE Spark	Framework de computación en clúster open-source de Apache
scikit learn	Biblioteca de aprendizaje automático de software libre en lenguaje Python
Natural Language ToolKit python	Conjunto de bibliotecas o programas para el procesamiento de lenguaje natural

Transmitir y comunicar

Esta etapa involucra encontrar maneras para comunicar la información e insights obtenidos del análisis de los datos en una manera comprensible y convincente. Para esto es esencial la visualización de los datos, incluyendo los mapas, modelos 3D y gráficos. También involucra transmitir esta información a las partes interesadas.

Herramientas para compartir ideas y transferir datos	
	Las siglas CSV hacen referencia a comma-separated-values, dado que denominan a los tipos de documento, de formato abierto sencillo, que representan los datos en forma de tabla; donde las columnas se separan por comas y las filas por saltos de línea.
	Las siglas JSON, corresponden en inglés a Javascript Object Notation y hace referencia a un formato sencillo de intercambio de datos. Está constituido por dos estructuras: colección de pares de nombre/valor y lista ordenada de valores.
	El Advanced Encryption Standard (AES) es un esquema o algoritmo de cifrado simétrico.
Herramientas para comunicar información	
	Shiny es un paquete de R que facilita la creación de aplicaciones web interactivas, directamente desde R.Shiny le permite a los usuarios interactuar con sus datos, sin tener que manipular el código.
	Es una plataforma de visualización que permite crear gráficas y tablas interactivas, en forma de paneles y hojas de trabajo. Es principalmente utilizado para inteligencia de los negocios, pero no se limita a esto.
	Data Driven Documents (D3) es una biblioteca de JavaScript para manipular documentos basados en datos. Con estos se crean infogramas, utilizando HTML, SVG y CSS.

RECOMENDACIONES TRANSVERSALES: CADENA DE VALOR DEL BIG DATA

- Priorizar la creación de un inventario de fuentes de datos disponible para el proyecto, incluyendo la metadata de éste. Tener claro la calidad de las fuentes de datos en términos de completitud, validez, consistencia, timeliness y precisión antes de compartir.
- En lo posible, compartir información sobre la organización interna de los datos (arquitecturas de entidades dueñas de los datos). A futuro, esto puede permitir una integración más fluida para quienes estén procesándolos.
- Considerar si los insights generados a partir de la explotación de datos se compartirán internamente únicamente o si también serán compartidos con terceros. Esto definirá en gran parte las herramientas que se debe utilizar para transmitirlos, por ejemplo, si se debe tener en cuenta la creación de una plataforma de datos abiertos o un dashboard de uso compartido.
- Cuando desee compartir datos a través de un panel de visualización interactivo, debe asegurarse de que la organización que albergue los datos a compartir cuente con los requerimiento y capacidades computacionales para gestionar este tipo de aplicación. Especialmente cuando los datos subyacentes a la visualización se actualizan con frecuencia, y el usuario está limitado por las funciones de análisis que la organización haya configurado y puesto a disposición.
- En lo posible para su proyecto, intente priorizar sistemas de almacenamiento flexibles, pues las soluciones de almacenamiento pueden incidir en la facilidad de escalar y en el performance del proyecto de datos. Las soluciones de almacenamiento en la nube son una buena solución cuando servidores propios no son los suficientemente potentes para realizar el análisis pensado. Esto permite además, alquilar recursos de almacenamiento a medida que las

necesidades de almacenamiento, procesamiento o análisis aumentan o disminuyen.

- En lo posible, para garantizar la interpretabilidad de los modelos utilizados en su proyecto, se recomienda reducir en la mayor medida posible el uso de algoritmos *blackbox*[27]. Esto permite que la interpretación de los resultados sea más transparente, y también puede contribuir a reducir los sesgos en los modelos, y por consiguiente en los resultados.
- Si piensa tercerizar los servicios de procesamiento y almacenamiento de datos, tener en cuenta las implicaciones que esto podría traer a sus estándares de seguridad y calidad. Si está tratando con datos sensibles, priorice almacenamiento y análisis en servidores locales.
- Utilizar soluciones de transferencia de datos con mayores estándares de seguridad que repositorios o softwares de acceso libre, como por ejemplo Google Drive, DropBox o SharePoint. Para datos y transferencia de estos realizados desde el gobierno, la recomendación principal es utilizar el SandBox de Big Data disponible para realizar transferencias de grandes volúmenes de datos entre diferentes entidades; esta no requiere esfuerzos de implementación y la confianza de los sistemas es alta.
- Si su proyecto depende del procesamiento de datos en tiempo real (por ejemplo, datos meteorológicos, o de calidad del aire) es necesario revisar si es factible procesar automáticamente estos datos para extraer métricas clave y/o buscar patrones específicos que ayuden a determinar si vale la pena conservar la totalidad de datos en crudo, dado el volumen de los datos y las necesidades de almacenamiento. Si decide no almacenar la totalidad de los datos en brutos, y decide almacenar indicadores agregados, debe tener en cuenta que es difícil regresar y corregir los cálculos hechos sobre estos datos, hacer cambios en el proceso o en los algoritmos utilizados, por lo cual se hace esencial permitir que otros verifiquen y validen el proceso utilizado para agregar los datos.
- Si su proyecto requiere analizar datos detrás de un firewall (por ejemplo, si necesita calcular indicadores agregados de movilidad a través de datos de redes celulares), asegúrese de que la organización que aloja los datos tenga las capacidades informáticas (capacidad de procesamiento y hardware adecuado) como la experiencia suficientes para llevar a cabo el análisis de una manera eficaz. Esto dependerá del tipo y volumen de los datos. Esto puede ocurre cuando no se permite que los datos en bruto abandonen los servidores seguros en el que están alojados, por lo que, si alguien quiere explorar los datos o ejecutar un modelo con estos datos, debe hacerlo a través de un firewall - usualmente esto se limita a proyectos donde el intercambio de los datos está limitado por consideraciones éticas, legales y estratégicas. Analizar datos detrás de un firewall es más complejo que otros métodos, pero en estos casos puede ser el más adecuado.
- Si su proyecto requiere de herramientas de crowdsourcing para recolectar datos, considere si la información que pretende obtener puede ser reportada de manera precisa y al nivel de granularidad que necesita por su audiencia. Hay momentos en los que tiene sentido aprovechar a personas no expertas para reunir información y conocimientos, pero los desafíos de tratar con datos de fuentes varias y dispares a veces pueden superar los beneficios.
- Si su proyecto requiere crear portales o aplicaciones para agregar datos - por ejemplo, para recopilar y almacenar diferentes conjuntos de datos y ponerlos a disposición de usuarios (i.e, repositorios de datos abiertos), tenga en cuenta que no todos los datos pueden ser compartidos de forma segura de esta manera, y que esta modalidad de proyecto pone mucha responsabilidad en el proveedor de la plataformas al tener que monitorear la actividad del conjunto de datos y mantener el almacenamiento de los mismos.

[27] Se refiere a los sistemas en los cuales sólo es posible observar los inputs y outputs del proceso, más no el trabajo interno, o proceso por el cual dichos inputs pasan. Esto sucede con algoritmos de aprendizaje automático que internalizan datos de formas que no son fácilmente comprendidas por los humanos, lo que genera poca transparencia en los procesos, ya sea por su complejidad (profunda red neuronal trabajando de manera difusa) o por su dimensionalidad (algoritmo basado en relaciones geométricas, no visualizables por el hombre). (Card, D. 2017. Towards data science. The “black box” metaphor in machine learning. & Bathaee, Y. 2018. Harvard Journal of Law & Technology. The Artificial Intelligence Black Box And The Failure Of Intent And Causation.)

SECCIÓN 3. USO DEL MODELO

1. CASO 1: USO DE TRANSFERENCIAS MONETARIAS CONDICIONADAS

<table>
<tr><th colspan="4">Plantilla I: Volviendo problemáticas de gestión pública en preguntas de datos</th></tr>
<tr><td>1. Área de Impacto</td><td colspan="3">2. Preguntas Clave:</td></tr>
<tr><td>Efectividad de transferencias monetarias condicionadas</td><td colspan="3">Una vez que se reciben las transferencias monetarias, ¿en qué se usan?</td></tr>
<tr><td colspan="4">3. Las 3C</td></tr>
<tr><td>Migajas

Datos de Más Familias en Acción de Prosperidad Social; transaccones de tarjetas bancarias o digitales de davivienda/daviplata.</td><td colspan="2">Capacidades

Almacenamiento y procesamiento en la nube; métodos de ML no supervisados para agrupar tipos de transacciones según grupos de personas; estadística tradicional para comparar grupo de tratamiento (beneficiario de Más Familias en Acción) y grupo de control (no beneficiario).</td><td>Comunidades

Departamento de Prosperidad Social; entidades bancarias relacionadas con Más Familias en Acción (Davivienda); Beneficiarias y beneficiarios de Más Familias en Acción.</td></tr>
<tr><td colspan="4">4. Planteamiento del problema</td></tr>
<tr><td colspan="4">¿Cuál es el uso que los beneficiarios de Más Familias en Acción le dan a las transferencias monetarias condicionadas que se reciben a través de transacciones bancarias?</td></tr>
<tr><td colspan="4">5. Características Clave del Proyecto</td></tr>
<tr><td>Meta estratégica

Identificar los tipos de consumo de transferencias monetarias presentes en los beneficiarios de Familias en Acción.</td><td>Beneficiarios

El Departamento de Prosperidad Social, posteriormente los beneficiarios de Más Familias en Acción.</td><td colspan="2">Acceso a migajas

Permiso de Prosperidad Social y de Entidad Bancaria para acceder a datos de beneficios y consumo.</td></tr>
<tr><td>Objetivos operacionales específicos

Coordinar un cruce de los datos de ambas plataformas.</td><td>Gobernanza y alianzas

Gobierno (DPS), Sector Privado (Banco) y sociedad civil (beneficiarios).</td><td colspan="2">Capacidades y herramientas

Acceso a la nube, Stata, Python, Expertos en ciencias sociales y ciencia de datos.</td></tr>
<tr><td>Resultados o productos esperados

Clasificación de tipos de consumo de beneficiarios y no beneficiarios del programa.</td><td>Plazo y alcance geográfico

1 año, teniendo en cuenta el tiempo de recibir datos. Alcance georgáfico mínimo 1 ciudad, máximo áreas urbanas.</td><td colspan="2">Otros</td></tr>
<tr><td colspan="4">6. Priorización de características del proyecto</td></tr>
<tr><td>Prioridad 1 - Meta estratégica</td><td colspan="3">Identificar los tipos de consumo de transferencias monetarias presentes en los beneficiarios de Familias en Acción.</td></tr>
<tr><td>Prioridad 2 - Acceso a migajas</td><td colspan="3">Permiso de Prosperidad Social y de Entidad Bancaria para acceder a datos de beneficios y consumo.</td></tr>
<tr><td>Prioridad 3 - Objetivos operacionales específicos</td><td colspan="3">Coordinar un cruce ético de los datos de ambas plataformas.</td></tr>
</table>

Plantilla II: Planteamiento del problema	
¿Qué problema está tratando de resolver utilizando Big Data o Inteligencia Artificial? Entender el uso que los beneficiarios de Más Familias en Acción dan a las transferencias monetarias condicionadas.	
1. *Enmarque el problema en una pregunta de datos* ¿Qué tan efectivo es el uso que le dan los beneficiarios del programa Más Familias en Acción a las transferencias monetarias que reciben?	*2. ¿Cuál es el impacto previsto con su proyecto?* Permitir que el DPS tenga información sobre la efectividad del consumo del dinero que reciben los beneficiarios de las transferencias monetarias condicionadas, con el fin de promover buenas prácticas de uso de los recursos.
3. ¿Cómo puede ser el Big Data o la Inteligencia Artificial útil para la resolución de este problema? Las transferencias bancarias permiten evidenciar los patrones de consumo en tiempo real y a un menor costo que la colecta de información con encuestas. Los modelos de aprendizaje no supervisado permiten crear grupos en grandes volúmenes de datos.	*4. ¿Cuáles son las limitaciones y el contexto al que se enfrenta este proyecto?* Las transacciones bancarias no permiten identificar el uso que se le da a las transferencias monetarias.
5. Reformule su pregunta	
El análisis de las limitaciones permite darse cuenta que el proyecto se debe enfocar en las diferencias de patrones de consumo entre beneficiarios y no beneficiarios del programa. Por esta razón se reformula la pregunta de la siguiente manera: **¿Existen diferencias en los patrones de consumo de beneficiarios y no beneficiarios de transferencias monetarias condicionadas medibles a través de transacciones bancarias?**	

Plantilla 3: Propuesta de proyecto	
Determine ¿quién es su audiencia y cuál es la mejor manera de llegar a ellos? - Audiencia: el Departamento de Prosperidad Social, la entidad que brinda los datos bancarios y los beneficiarios del programa. - Forma de acercar a las entidades: creación de un acuerdo de confidencialidad compartir los datos entre las entidades. - Forma de acercarse a los beneficiarios: acuerdo con Prosperidad Social con pautas de confidencialidad de la información, anonimidad de los informantes, para proteger su seguridad durante y después de su participación en el proyecto.	
1. Contexto y relevancia	2. Innovación
¿Cuál es la relevancia de este proyecto a nivel local, nacional o global? Familias en Acción tiene cobertura nacional con recursos públicos, entender el tipo que consumo que se promueve a través del programa es importante para el país. *¿Dónde se va a implementar el proyecto y porqué (enfoque geográfico)?* Bogotá. *¿Quiénes son los potenciales beneficiarios (población objetivo) de este proyecto?* Prosperidad Social y los beneficiarios de Más Familias en Acción. *¿Cuáles son los objetivos operacionales específicos del proyecto, y porqué son relevantes?* Lograr la obtención y unión de forma ética de los datos de transacciones bancarias con datos de Más Familias en Acción.	*¿Cuál es la innovación técnica más relevante para su proyecto y porque?* Nuevas fuentes de datos para resolver interrogantes sobre el consumo de las personas. *¿Qué valor agregado aporta su propuesta para mejorar la forma en la que actualmente se están haciendo las cosas?* La actualización en tiempo real de las transacciones bancarias. *¿De qué manera presenta su proyecto una innovación en términos de metodología? De datos? De resultados y productos finales?* - Metodología: el uso de modelos no supervisados de aprendizaje de máquinas para beneficio social. - Datos: unir datos del sector privado con los del sector público. - Resultados: identificar hábitos de consumo que promueven a través del programa de Más Familias en Acción.

3. Retos y cómo sobrellevarlos	4. Propuesta de valor
¿Cuáles son los riesgos y desafíos que puede enfrentarse durante el desarrollo e implementación del proyecto? ¿Algún reto específico relacionado a los datos o tecnologías que está proponiendo utilizar? No poder vincular los datos de los beneficiarios con los bancarios, por respeto a privacidad. *¿Cuales son las soluciones previstas para abordarlo?* respeto Anonimización de datos de las personas o agregación de los datos de MFA y bancarios a nivel de barrio.	*¿Cómo espera que las partes interesadas adopten su proyecto?* Como una herramienta que le permite al DPS para ver las tendencias de consumo y ayudar a la promoción y mantenimiento de hábitos de consumo saludables, incluyendo la nutrición infantil. *¿A corto plazo, cuáles los próximos pasos para el desarrollo e implementación del proyecto?* La consecución de los datos. El proyecto tiene como etapa recolección, almacenamiento, análisis (para limpieza de datos), procesamiento, análisis (modelos no supervisados y de estadística tradicional)y transmisión. *¿Cuál es la visión a largo plazo del proyecto?* Tener herramientas que permitan incentivar y apoyar a las familias para que promuevan una mejor nutrición en la familia. *¿Cómo encaja este proyecto entre las prioridades del gobierno nacional? Del gobierno local?* La nutrición infantil es parte fundamental del programa de Más Familias en Acción, el cual es política de estado con cobertura nacional.

Plantilla 4: Ética y Participación: Retos y Soluciones
Consideraciones legales y de privacidad - cruce de datos entre MFA y el banco, a través de anonimización de bases de datos, adición de ruido a los indicadores y posibilidad de agregar datos a nivel de barrio.
Protección de datos y consentimiento informado - se verifica el consentimiento de los dueños de los datos para el uso de estos datos en estudios. De lo contrario, se selecciona una muestra de beneficiarios a los que se les solicita autorización y se realiza el estudio con ellos. La protección de datos se garantiza al usar plataformas de manejo de datos seguras o soluciones on-premise, si se deseara que el proyecto se hiciera en las oficinas del DPS o del banco.
Representatividad y participación - se propone una etapa en la que se dialoga con beneficiarios y no beneficiarios para entender sus patrones de consumo y que permitan ajustar el análisis.
Gobernanza responsable - se crea un consejo de ética con los portadores de interés del proyecto (DPS, Banco y Beneficiarios) para guiar las decisiones del proyecto.
Transparencia - se especifican todas las limitaciones del análisis en las presentaciones sobre el proyecto y en el resultado escrito de la investigación. De la misma forma, se publica el código usado para analizar los datos, de tal forma que el estudio sea replicable.

2. CASO 2: MONITOR DE VIOLENCIA DE GÉNERO EN EL PAÍS

<table>
<tr><th colspan="3">Plantilla I: Volviendo problemáticas de gestión pública en preguntas de datos</th></tr>
<tr><td>1. Área de Impacto</td><td colspan="2">2. Preguntas Clave:</td></tr>
<tr><td>Monitorear la violencia de género en el país</td><td colspan="2">¿Quiénes son los beneficiarios?
Tomadores de decisiones, víctimas y potenciales víctimas de violencia de género.

¿En qué se espera que se usen los beneficios?
Entender con mayor granularidad y precisión el contexto en el que ocurren estos delitos, y la magnitud con la que ocurren.

¿En qué se usan los beneficios?
Para la creación de medidas preventivas la violencia de género.</td></tr>
<tr><td colspan="3">3. Las 3C</td></tr>
<tr><td>Migajas

-Estadísticas Violencia de Género 2018 - Ministerio de Salud
-Datos de denuncias y delitos de la Policía Nacional
-Twitter, Facebook
-Revistas y periódicos del país
-Crowdsourced data</td><td>Capacidades

-Diseño UX y UI
-Minería de datos de fuentes online
-Desagregación y unificación de bases de datos
-Machine learning e inteligencia artificial</td><td>Comunidades

-Academia e instituciones de investigación (equivalente de Geochicas en Colombia)
-Ministerio de Salud
-Polícia Nacional</td></tr>
<tr><td colspan="3">4. Planteamiento del problema</td></tr>
<tr><td colspan="3">La ausencia de una base de datos o plataforma que monitoree la violencia de género en Colombia, a parte de las fuentes oficiales, limita conocer la verdadera magnitud de los delitos de género en el país. Hasta el momento, no existe una base de datos centralizada, que además esté acompañada de visualizaciones, gráficas y análisis sobre la información de delitos de género. La ausencia de este tipo de iniciativas limita el entendimiento de la magnitud y patrones en este tipo de violencia.</td></tr>
<tr><td colspan="3">5. Características Clave del Proyecto</td></tr>
<tr><td>Meta estratégica

Crear una plataforma de visualización que reúna toda la información del país en términos de violencia de género, y que se actualice a medida que se hacen disponibles nuevos datos.</td><td>Beneficiarios

Tomadores de decisiones, víctimas y potenciales víctimas de violencia de género.</td><td>Acceso a migajas

Se realizará a través de técnicas de minería de texto, lo que permite recolectar información de varias fuentes de datos. Así mismo, se habilitará una manera de alimentar este sistema con datos reportados por víctimas.</td></tr>
<tr><td>Objetivos operacionales específicos

1) Crear una base de datos centralizada con toda la información disponible, 2) alimentar esta base de datos con información minada de diferentes fuentes, 3) crear plataforma de visualización interactiva.</td><td>Gobernanza y alianzas

Se hace necesario contar el apoyo de entidades quienes producen este tipo de información (Ministerio de Salud, por ejemplo), así como de activistas cuyas redes puedan dar visibilidad a la plataforma.</td><td>Capacidades y herramientas

Fundamental para el proyecto es lograr recolectar e integrar diferentes fuentes de datos para crear una base de datos con información similar.</td></tr>
<tr><td>Resultados o productos esperados

Una plataforma funcional que se actualiza a través de procesos automáticos de integración de fuentes.</td><td>Plazo y alcance geográfico

Seis meses de preparación, y actualización y gestión constante.</td><td>Otros</td></tr>
</table>

6. Priorización de características del proyecto	
Prioridad 1 - Meta estratégica	Crear una plataforma funcional que se actualiza a través de procesos automáticos de integración de fuentes.
Prioridad 2 - Acceso a migajas	Fundamental para el proyecto es lograr recolectar e integrar diferentes fuentes de datos para crear una base de datos con información similar, i.e integración de bases de datos y minería de datos.
Prioridad 3 - Objetivos operacionales específicos	Se hace necesario contar el apoyo de entidades quienes producen este tipo de información (Ministerio de Salud, por ejemplo), así como de activistas cuyas redes puedan dar visibilidad a la plataforma.

Plantilla II: Planteamiento del problema

¿Qué problema está tratando de resolver utilizando Big Data o Inteligencia Artificial?

Ausencia de información georeferenciada y detallada sobre delitos sexuales y de género, incluyendo feminicidios, en el país.

1. *Enmarque el problema en una pregunta de datos*	*2. ¿Cuál es el impacto previsto con su proyecto?*
¿Qué nuevos conocimientos puedo adquirir al integrar datos georeferenciados y minados con registros administrativos y datos de entidades públicas sobre la prevalencia y patrones de los delitos sexuales y de género en el país?	Crear una plataforma para visualizar y monitorear los crímenes de género en el país, que le permita al gobierno reunir diferentes fuentes de datos confiables de manera sistemática, para 1) concientizar a la ciudadanía y a los tomadores de decisiones sobre la magnitud del problema, y 2) para entender mejor los contextos donde este tipo de crímenes ocurren en busca de crear medidas preventivas.
3. ¿Cómo puede ser el Big Data o la Inteligencia Artificial útil para la resolución de este problema?	*4. ¿Cuáles son las limitaciones y el contexto al que se enfrenta este proyecto?*
Por una parte, se puede utilizar herramientas de minería de datos para extraer información de diferentes fuentes en línea, por ejemplo de redes sociales o de periódicos en línea, buscando complementar la información recolectada por el estado. Por otro lado, y quizá en un segunda etapa del proyecto, se puede aplicar algoritmos de aprendizaje de máquinas para entender la incidencia de diferentes factores sobre un crimen e identificar marcadores de que un crimen pueda ocurrir.	Dado la sensibilidad del tema, debe establecerse un marco ético para el manejo de los datos en el proyecto. Por otra parte, la minería de datos puede conllevar a datos poco confiables (dado su naturaleza voluntaria) y puede suponer riesgos éticos de acceso a la información también. Por otro lado, mantener una plataforma de visualización actualizada con datos nuevos constantemente es un reto operacional, dado los recursos necesarios para mantener dicha plataforma.

5. Reformule su pregunta

¿Cómo pueden integrarse fuentes de datos de registros administrativos, crowdsourced y datos minados acerca delitos sexuales y de género para crear una plataforma de visualización que permita monitorear y entender a más detalle la violencia de género?

Plantilla 3: Propuesta de proyecto	
Determine ¿quién es su audiencia y cuál es la mejor manera de llegar a ellos? Tomadores de políticas públicas en temas de violencia de género; activistas y organizaciones que abogan en contra de la violencia de género.	
1. Contexto y relevancia	2. Innovación
Este proyecto es relevante en general dadas las alarmantes cifras de violencia de género en Colombia. En particular es un esfuerzo para visualizar los feminicidios y otros crímenes puntualmente en contra de la mujer; estos crímenes a veces se agrupan en otras categorías, lo que hace difícil llevar una cuenta exacta de los casos en el país, y por ende conlleva a no priorizar medidas preventivas contra este tipo de crimen. Al ser una plataforma de visualización, este proyecto se enfocara en todo el país.	Este proyecto propone la agregación de diferentes fuentes de datos de entidades públicas y privadas, así como fuentes de redes sociales y auto-declaradas. Encontrar una manera de integrar estas fuentes es un reto, así como encontrar una manera de georeferenciar todos los crímenes o delitos. El valor agregado es no solo crear una base de datos pero crear visualizaciones fáciles de digerir y extraer patrones de los datos que puedan informar sobre los factores, variables o condiciones que más incide, o que tienen alguna correlación con la ocurrencia de un crimen.
3. Retos y cómo sobrellevarlos	4. Propuesta de valor
Uno de los riesgos principales es no representar de manera adecuada cuando verdaderamente se trata de un crimen de violencia de género, versus otro tipos de crímenes. Para sobrellevar esta limitación, se hace necesario contar con expertos en el tema, y quizá con etiquetas diferentes que reflejen la certeza que se tiene sobre un caso. Otro reto es crear una visualización que se actualice a medida que se van creando más datos, ya que este presenta un reto en la arquitectura de la visualización y en la gestión de esta plataforma. Para abordarlo, es necesario contar con una persona dedicada a monitorear y actualizar la visualización, tomando en cuenta las nuevas fuentes de datos que pueden ser creadas a través del tiempo.	A nuestro conocimiento, no existe una fuente centralizada de datos sobre este tipo de delito que contenga información fácil de digerir. A largo plazo, se podría crear una red internacional de gobiernos que comparten los datos acerca este tipo de crímenes que permite alimentar a los modelos de más datos.

Plantilla 4: Ética y Participación: Retos y Soluciones
Consideraciones legales y de privacidad - los datos deben ir propiamente anonimizados y con 'ruido' en su georeferenciación para no permitir la identificación de estos. Además deben considerar medidas para prevenir la revictimización de las personas cuya información se encuentra en la plataforma de visualización.
Protección de datos y consentimiento informado - asegurarse de que la minería de texto no está violando las diferentes cláusulas de protección de datos de las fuentes utilizadas.
Representatividad y participación -
Gobernanza responsable - contar con amplia representación de organizaciones en la gobernanza del proyecto.
Transparencia - crear etiquetas que comuniquen la certeza con la que se puede asumir que un caso es de hecho un caso de violencia de género (varía dependiendo de lo que reportan las fuentes).

SECCIÓN 4. RECOMENDACIONES SOBRE CÓMO EL ESTADO PUEDE IMPLEMENTAR PROYECTOS DE ANALÍTICA PARA ESTIMULAR EL SECTOR NACIENTE DE BIG DATA

A continuación se detallan recomendaciones específicas, que buscan facilitar los cuellos de botella actuales en la formulación e implementación de proyectos de datos desde el Estado. Estas recomendaciones son pensadas como insumos de varias políticas públicas, lineamientos o guías cuya función delimite o facilite la explotación de datos desde el gobierno. Estas se desarrollan, principalmente, de las recomendaciones y aprendizajes, producto de los pilotos de supervivencia empresarial y detección de fraude en el SISBEN (ver documentos 7 y 8 de esta consultoría). Cabe recalcar que a lo largo de esta consultoría se han identificado recomendaciones adicionales que pueden encaminar al gobierno hacia la transversalización de la explotación de datos en sus funciones, incluyendo diferentes lineamientos y estándares que también aplican para la formulación e implementación de proyectos.

1. Promover el afianzamiento y la consolidación de las infraestructuras físicas adecuadas para adelantar proyectos de Big Data e Inteligencia Artificial en entidades gubernamentales. En los proyectos pilotos, adelantados en el marco de este contrato, se pudo determinar que los recursos de procesamiento y almacenamiento del Estado resultan insuficientes para la preparación de los datos a ser entregados a consultores en proyectos de análitica. Esto presenta también, una barrera para la validación, de parte del gobierno, de los resultados obtenidos en estos proyectos. La ausencia de capacidad de procesamiento redunda en conjuntos de datos incompletos, demoras prolongadas de integración y anonimización de datos, y hace necesario ejecutar procesos adicionales de preparación de los datos. Sobre todo, esto también involucra una incapacidad en la replicación de los resultados o un posterior uso de los modelos creados en el marco de los proyectos.
2. Contar con soluciones de almacenamiento flexibles que se adapten a las necesidades específicas de la entidad que las quiere contratar; si el departamento o entidad objeto requiere un procesamiento rápido y constante durante todo el año, puede ser aconsejable tener un sistema de procesamiento físico; por el contrario, a niveles más bajos de procesamiento, puede ser más aconsejable el uso de la nube.
3. Es recomendable reforzar la implementación de guías gubernamentales en temas de interoperabilidad, anonimización y calidad de datos. Con dichas guías será posible la conexión o comunicación entre los diversos sistemas y productos del Gobierno, sin retrasos o procesos administrativos adicionales. El cruce automático de datos reducirá la carga de preparación de conjuntos de datos para potenciales proyectos. Priorizar estos lineamientos tienen la posibilidad de incrementar la homogeneidad de las fuentes de datos del Estado, lo cual reducirá el trabajo de la edición y reingeniería de los datos, facilitando así su interpretación e identificación. Adicionalmente, permitiría una consistencia en la composición de variables a través de una base de datos, para que de esta forma, al cruzar bases o fuentes de datos que quieren estudiar lo mismo, se asegure que cada variable está considerando los mismos factores.
    - Hacer una verificación periódica de la calidad de los datos que se están utilizando en el proyecto, y en general, de las fuentes del Estado. Esto permite tener un registro actualizado de los datos albergados, incluyendo la cobertura y confiabilidad de estos. En la estrategia propuesta en esta consultoría, dicha tarea es responsabilidad de los CIOs u Oficinas de TI de las entidades, quienes deben ser responsables de generar reportes tanto del software y hardware, como de los conjuntos de datos de la entidad.
4. Considerar implementar reglamentación que haga obligatorio y estandarice la creación de diccionarios de datos para todas las fuentes de datos del Estado. El tener un diccionario exhaustivo y actualizado facilita la integración de diferentes fuentes, así como el entendimiento de los datos en general. Priorizar la creación de estos diccionarios va de la mano de las necesidades generales de gestión de datos maestros del país, entre las que se encuentra la necesidad de definir estándares transversales[28].
5. Considerar cómo permitir que los estándares para la recepción de datos, en el marco de un proyecto de datos, se realice en conjunto con quienes llevan a cabo el procesamiento y análisis de estos. Esto involucraría una modificación a los términos de referencia utilizados actualmente para contratistas, buscando permitir que decisiones sobre el proyecto se puedan realizar previo al inicio formal de este.

[28] En los documentos anteriores de esta consultoría, se identifica que los principios contemplados en la International Open Data Charter son un buen punto de partida (incluye apertura for defecto, lineamientos para hacer datos accesible y utilizables, entre otros). Para más información ver “Propuesta de plan de implementación de la Estrategia de Big Data para el Estado”, página 12.

6. Al otorgar contratos para la realización de proyectos de Big Data, es fundamental buscar contratistas con un balance entre experiencia y competencias técnicas, en el área de negocio o tema a tratar. Además, en la creación de estos proyectos, es importante asegurar el acompañamiento a dichos contratistas, tanto para resolver dudas sobre los datos estudiados, como para garantizar la transferencia de conocimiento a entidades del sector público. En el caso de los pilotos de analítica, la mediación de iNNPulsa fue clave para que ambas empresas pudieran contactar a las entidades dueñas de los datos para solucionar inconvenientes que surgieron en el proceso de los pilotos y para aterrizar y verificar el análisis de la problemática social en cuestión.
7. Al otorgar contratos para la realización de proyectos de Big Data, es fundamental buscar contratistas con un balance entre expe Crear e impulsar una red de transformación digital, como el punto de partida para profesionales de Big Data y tecnologías de la Cuarta Revolución Industrial. Esta red de profesionales será organizada a través de tertulias trimestrales en temáticas de formación relevantes a la explotación de datos, con participación de profesionales del sector público y privado. Esta red puede ser parte de la gobernanza, verificación y validación de proyectos de analítica adelantados desde el gobierno.

SECCIÓN 5. HALLAZGOS GENERADOS A PARTIR DE PILOTOS

En el marco del contrato suscrito entre iNNpulsa Colombia y el Massachusetts Institute of Technology (MIT) para el diseño de una Estrategia de Big Data para Colombia, se llevaron a cabo dos proyectos pilotos de analítica de datos. Cada uno estaba encaminado a resolver una problemática pública concreta usando Big Data, con el objetivo de estimular la participación del sector privado y de fortalecer esta industria en el país. Esta sección describe los resultados generados a partir de los dos pilotos, brevemente introduciendo el contexto y metodología utilizadas, e incluye los hallazgos principales de estos dos proyectos. Para más información y detalle sobre el proceso efectuado y las metodologías utilizadas, ver aquí y aquí .

1. PROYECTO PILOTO DE ANALÍTICA DE DATOS: SISBEN

El SISBEN (Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales) es el mecanismo usado por el Gobierno de Colombia para focalizar recursos de inversión pública social hacia la población más vulnerable del país. El SISBEN usa encuestas socioeconómicas para estimar el nivel de vulnerabilidad de los hogares encuestados. Con esta información, diversos programas sociales del gobierno definen puntos de corte para determinar el nivel de vulnerabilidad mínimo para poder entrar a sus programas; de tal forma que se le garantice el acceso a las poblaciones más vulnerables. Sin embargo, la naturaleza declarativa de las encuestas socioeconómicas y el alto costo de mantener la base de datos actualizada dificultan la estimación de la condición socioeconómica real de las personas. En consecuencia, se crean anomalías en la base de datos, es decir, encuestas que no reflejan la situación socioeconómica de los individuos. Las anomalías tienen importantes repercusiones económicas y sociales ya que estas crean desviaciones en la inversión social, evitando que esta llegue a las personas con más necesidades socioeconómicas en el país.

En este contexto, la empresa Data Innova SAS fue contratada por iNNPulsa con dos objetivos: primero, ayudar en la detección de las anomalías en la base de datos del SISBEN y segundo, encontrar las características socioeconómicas y los programas sociales que más se relacionan con la movilidad social. Para este fin, se utilizaron las bases de datos del SISBEN III, y de Programas Sociales. Además se realizó una prueba de concepto donde se validaron los datos del SISBEN con las bases de datos de la UGPP, RUNT, RUES y ADRES. Los modelos de aprendizaje de máquinas usados fueron: Gradient Boosting Machine (GBM), Isolation Forest, Long-Short Term Memory (LSTM) y Long-Short Term Memory con Autoencoder.

Algunos de los resultados del proyecto piloto son los siguientes:

- Las mejores predicciones de anomalías se alcanzaron con el GBM y el LSTM[29].
- De una muestra de 7.500 casos validados en el RUNT, 88.6% de las personas reportaron no tener vehículo en el SISBEN pero tenían vehículo en el RUNT.
- Se detectó que el puntaje del SISBEN tiene una alta variación cuando las personas reportan un cambio en la tenencia vehículo, o un cambio en la manera en la que se cocina en el hogar. En el caso del cambio en la tenencia de vehículo se encontró una gran cantidad de anomalías. Se le recomienda al DNP incluir estas variables dentro de las que podrían generar alertas de anomalías.

[29] Área bajo la curva: GBM .91, LSTM .92. En 4 mil personas, se detectan: GBM 71%, LSTM 72% de anomalías.

- Las variables que más se relacionan con aumentos en el puntaje en el SISBEN son las de Educación y Salud (Paso de primaria a secundaria y de régimen subsidiado al contributivo de salud).
- La mayor cantidad de aumentos en el puntaje del SISBEN se relaciona con la participación en los programas de Mujeres Ahorradoras en Acción, Incentivo para la Capacitación al Empleo y Más Familias en Acción.
- La base de datos del DNP no tiene una estructura de panel, la cual se creó para este proyecto y permitió encontrar más de 3.800 duplicados no reportados en el SISBEN, además de permitir el análisis de movilidad social y anomalías en el tiempo.

2. PROYECTO PILOTO DE ANÁLITICA DE DATOS: SUPERVIVENCIA EMPRESARIAL

Con el objetivo de crear y ejecutar un modelo de analítica de datos predictivo y diagnóstico de los factores determinantes de la supervivencia y exportación de las empresas colombianas, se realizó la contratación de la empresa EyS Soluciones Empresariales para adelantar un proyecto piloto de analitica de datos. Este proyecto consistió de dos objetivos específicos, el primero enfocado en identificar la probabilidad de que una empresa sobreviva en el tiempo; y el segundo, fue identificar la probabilidad de que una empresa comience a exportar de acuerdo a ciertas características, variables y factores. Con el fin de lograr los objetivos mencionados, se utilizaron diferentes bases datos, las cuales fueron insumo para los análisis posteriores. Entre estas se incluye información del Registro Único Empresarial y Social (RUES)*, Importaciones DANE-DIAN*, Exportaciones DANE-DIAN* y Sistema de Matrículas Estudiantil (SIMAT) del Ministerio de Educación.

Para este análisis, el piloto hizo uso principalmente de Survival Analytics, familia de métodos estadísticos y de aprendizaje de máquinas para estimar la duración y analizar el tiempo de eventos. Para complementar estos métodos, se utilizaron también otros modelos estadísticos paramétricos y no-paramétricos, y de aprendizaje de máquinas, tales como Random Forests y Random Survival Forests. Entender con mayor precisión la incidencia que factores sociales, económicos, y empresariales tienen sobre la supervivencia y la vocación exportadora de empresas colombianas, es clave para generar políticas públicas informadas que reflejen la realidad contextual y cambiante del país.

Entre los hallazgos generados a partir de este proyecto, se encuentran los siguientes:

- Cuando se observa la probabilidad de supervivencia de todas las empresas colombianas, sin importar su ubicación o actividad, se encuentra que esta decrece con los años. En promedio un 77.6% de las empresas pasan del primer año, mientras que solo el 40.5% de las empresas superan el décimo año.
- Al comparar la supervivencia de las empresas constituidas como persona natural o persona jurídica, se encuentra que las últimas tienen una mayor probabilidad de supervivencia. En el primer año, la supervivencia de las empresas como persona jurídica es 6 puntos porcentuales mayor que las de empresas como persona natural. En el décimo año, esta diferencia se aumenta a 7 puntos porcentuales.
- Cuando se compara la supervivencia empresarial según la ubicación geográfica de las empresas en el primer, quinto y décimo año, se encuentra que la supervivencia de las empresas ubicadas en Medellín, Barranquilla y Bucaramanga se mantienen por lo menos 20 puntos porcentuales por encima del promedio nacional. En Cali, a través del tiempo, la supervivencia se encuentra consistentemente abajo del promedio nacional, por aproximadamente 9 puntos porcentuales.
- Según los modelos, entre los factores que tienen una incidencia inversamente proporcional al riesgo de desaparición de una empresa se incluye el tamaño de la empresas medido en activos - esta tienen un afectación inversa del 6%, queriendo decir que un aumento en los activos de una empresa contribuye a reducir las tasas de riesgo de aumento en un 6%. Adicionalmente, las empresas tienden a durar más en el tiempo cuando tienen más establecimientos.
- Según los cálculos de los modelos desplegados en este proyecto, se estima que 4 de cada 1000 empresas colombianas realizarán su primera exportación al término del primer año. Esta figura se incrementa a 2 de cada 164 empresas al quinto año, y 1 de cada 65 empresas después de 10 años.