Contexto del proyecto

El Ministerio de Agricultura en Colombia tiene el programa "**Coseche y venda a la fija**" como estrategia de comercialización agropecuaria (Agricultura por contrato).

A grandes rasgos el programa busca que se pueda establecer una relación de confianza entre el productor (agricultor colombiano) y el comercializador donde se pueda asegurar que el agricultor pueda recibir un abono antes de recolectar su cosecha y así el comercializador pueda asegurar su pedido. Reduciendo la brecha de los intermediarios.

El principal objetivo de este programa es:

- Reducir la incertidumbre y riesgos de comercialización.
- Fortalecer las relaciones entre Agricultor y comercializador
- Asegurar la comercialización de los cultivos.
- Evitar el desperdicio de los productos.

Indicadores de rendimiento

N	INDICADOR	TIPO	DESCRIPCIÓN
1	Rendimiento de Cosecha	numérico (#)	Cantidad de toneladas producidas por hectáreas cosechadas.
2	Cosechado vs lo sembrado	porcentaje (%)	Efectividad entre las hectáreas sembradas y las hectáreas cosechadas.
3	Provisión	Boleano	Corresponde a los productos de ciclo transitorio y anual el cual indica los periodos del año en qué un producto es cultivado o no
4	Abastecimiento	porcentaje (%)	Producción de cada uno de los productos por departamento.
5	Productividad por ciclos	porcentaje (%)	Producción por ciclos cada año.
6	Proporción por cultivo	porcentaje (%)	Producción por cultivo y hectárea
7	Área sembrada	numérico (#)	Hectáreas sembradas por grupo de cultivos

Entendimiento de la Data

La data de este proyecto es pública en la página www.datos.gov.co publicada por el Ministerio de Agricultura de Colombia y la comprensión y abordaje que se realizará se toma como fuente las necesidades obtenidas en el documento "PLAN ESTRATEGICO SECTORIAL 2019-2020 - CAMPO CON PROGRESO: TRANSFORMACIÓN PRODUCTIVA, COMPETITIVIDAD, Y DESARROLLO RURAL"

La fuente de la data es: https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Evaluaciones-Agropecuarias-Municipales-EVA/2pnw-mmge

Perfilado de la Data

La data esta distribuida en 17 columnas con 206.068 registros, los atributos se dividen entre numéricos, textos y fechas de la siguiente forma:

Nombre de Columna	Descripción	Tipo
COD_DEP	Código del departamento, según lo establecido por el DANE	Número
DEPARTAMENTO	Departamento Colombiano	Texto
COD_MUN	Código del municipio, según lo establecido por el DANE	Número
MUNICIPIO	Municipio Colombiano	Texto
GRUPO_CULTIVO	Categoria del cultivo	Texto
SUBGRUPO_CULTIVO	Tipo de cultivo según categoria	Texto
CULTIVO	Nombre del cultivo	Texto
DESAGREGACION_REGION_SISTEMA_PRODUCTIVO	Nombre generico del cultivo	Texto
ANNO	Año de producción	Número
PERIODO	Periodo medico	Texto
AREA_SEMBRADA(HA)	Area sembrada en hectáreas	Número
AREA_COSECHDA(HA)	Area cosechada en hectáreas	Número
PRODUCCCION(T)	Tiempo de producción	Número
RENDIMIENTO(T/HA)	Rendimiento de la cosecha	Número
ESTADO_FISICO_PRODUCCION	Estado del producto	Texto
NOMBRE_CIENTIFICO	Nombre cientifico del cultivo	Texto
CICLO_CULTIVO	Ciclo del cultivo en el país	Texto

Metadatos

El contenido de cada una de las variables esta dado por:

COD_DEP: Equivale a los IDs de los 32 departamentos que conforman la división política de Colombia, esta identificación concierne a la nomenclatura asignada por el DNE.

DEPARTAMENTO: Corresponde al nombre de cada uno de los departamentos.

COD_MUN: Contiene 1105 registros únicos equivalentes a los IDs de los municipios, esta nomenclatura esta dada por el DANE.

MUNICIPIO: Corresponde al nombre de los municipios. Valores perdidos 1.

GRUPO_CULTIVO: Atributo correspondiente a los grupos de cultivos y esta conformado por 13 categorías.

- Hortalizas
- Plantas aromáticas, condimentarías y medicinales
- Tubérculos y plátanos
- Frutales
- Oleaginosas
- Leguminosas
- Fibras
- Flores y follajes
- Cereales
- Otros permanentes

- Forestales
- Hongos
- Otros transitorios

SUBGRUPO_CULTIVO: Variable desagregada de la variable **GRUPO_CULTIVO** contiene 120 categorías.

CULTIVO: Atributo desagregado de la variable **SUBGRUPO_CULTIVO** y contiene 223 categorías.

DESAGREGACION_REGION_SISTEMA_PRODUCTIVO: Esta conformada por 271 categorías.

ANNO: Esta conformado por cada una de las categorías de años que hacen parte de la base de datos desde 2006 a 2018, 13 categorías.

PERIODO: El atributo periodo esta dado por 12 categorías de los años 2007 - 2018, adicional a estás cada año esta dividido en dos categorías equivalentes a la letra **A** para el periodo Enero Junio y B Julio Diciembre, para un total de 37 categorías.

AREA_SEMBRADA(HA): Corresponde al total de hectáreas sembradas.

AREA_COSECHDA(HA): Corresponde al total de hectáreas cosechadas derivado del total de hectáreas sembradas.

PRODUCCION(T): Atributo que hacer referencia al número de toneladas producidas.

RENDIMIENTO(T/HA): Variable derivada de la división entre las toneladas (T) de producción y el área cosechada en hectáreas (HA)**, algunos registros muestran diferencias en esta operación.** Valores nulos 3433

ESTADO_FISICO_PRODUCCION: Variable que clasifica los registros acordes al estado de cada uno de los productos y esta conformada en 23 categorías.

NOMBRE_CIENTIFICO: Categorizado en 214 nombre científicos diferentes. Valores perdidos 2853.

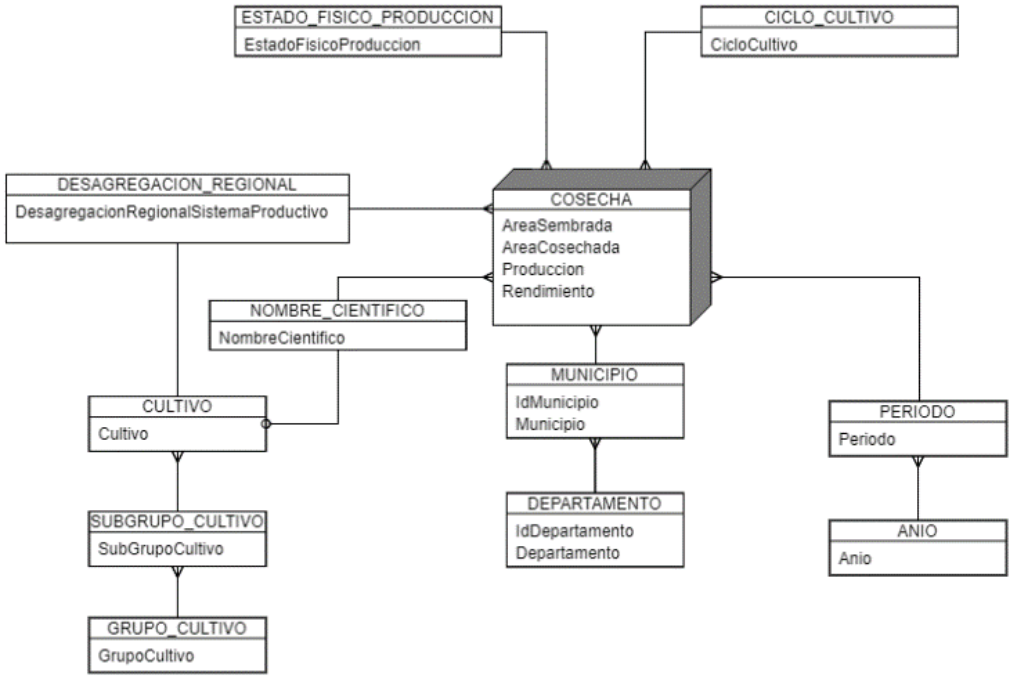
CICLO_CULTIVO: Se divide en los ciclos de los cultivos

- Transitorio
- Anual
- Permanente

Cultivos transitorios o anual: Cultivos cuyo ciclo de crecimiento es, en general, menor de un año y tienen como característica fundamental que después de la cosecha deben volver a sembrarse para seguir produciendo.

Cultivos permanentes: Cultivos que después de plantados llegan en un tiempo relativamente largo a la edad productiva, dan muchas cosechas y terminada su recolección no se los debe plantar de nuevo.

Diagrama Entidad Relación



En el proceso de limpieza de datos se hace ajustes:

- En la dimensión NOMBRE_CIENTIFICO
- Se hacen equivalencias y transformaciones frente a los departamentos y municipios utilizando los códigos del DANE
- Se hacen equivalencias en la dimensión COSECHA
    - AREA_SEMBRADA(HA) → AreaSembrada
    - AREA_COSECHDA(HA) → AreaCosechada
    - PRODUCCION(T) → Producción
    - RENDIMIENTO(T/HA) → Rendimiento

Tabla de hechos o Fact_prod

Se crea una tabla de hechos teniendo en cuenta las siguientes condiciones:

- Tabla principal del modelo dimensional
- Contienen campos claves que se unen a las tablas de dimensión
- Contiene métricas o también llamadas medidas y es aquello que queremos medir o analizar. Generalmente son valores numéricos que se suelen agregar
- Evitan la redundancia de atributos por estar estos en las tablas de dimensiones

Fact_prod	Tipo Dato	Descripción
ID_FACT	Entero	Identificador único para la tabla de hechos
ID_MUNICIPIO	Entero	Identificador único para la dimeción de ubicación geográfica
ID_CULTIVO	Entero	Identificador único para la dimención de cultivo
ID_TIEMPO	Entero	Identificador único para la dimención de tiempo
ID_CICLO	Entero	Identificador único para la dimención de ciclo
ID_ESTADO	Entero	Identificador único para la dimención de estado
AREA_SEMBRADA_HA	Entero	Medida en hectareas de siembra
AREA_COSECHADA_HA	Entero	Medida en hectareas de cosecha
PROCUCCION_T	Entero	Total toneladas cosechadas
RENDIMINETO_T_HA	Decimal	Medida equivalente a las T/HA

Descripción tabla de hechos

Identificar las dimensiones teniendo en cuenta las siguientes condiciones:

- Tablas desnormalizadas
- Se unen a las tablas de hechos a través de un campo clave
- Los atributos de la tabla de dimensión ofrecen información característica de las tablas de hechos
- Las dimensiones pueden contener una o varias relaciones jerárquicas

Nombre tablas de dimisiones: **Dim_ciclo, Dim_Time, Dim_Geografïa, Dim_Estado, Dim_Cultivo, Dim_Sis_Produccion, Dim_Científica**

PROCESO ETL

El proceso ETL se realiza utilizando la herramienta PowerBI, previo desarrollo del modelo ER, los datos se almacenaron en un archivo de Excel dividido por hojas de calculo.

Posterior al cargue (E) de los datos se realizaron una serie de trasformaciones(T) como: estandarizaciones, limpieza de datos y creación de nuevas variables útiles y necesarias para la visualización, al igual se tuvo en cuenta los hallazgos detectados durante el perfilamiento de los datos.

Limpieza de datos

a) Se agrego el nombre del municipio faltante basados en el ID del municipio.
b) Al verificar los valores perdidos del atributo **RENDIMIENTO_T_HA** (3433) encontramos que (2354) registros la cosecha es 0 y presentaban rendimiento lo cual no tiene sentido de negocio, por lo tanto, se recreo la variable basados en la formula T/HA_Cosechadas, para el caso que el valor de hectáreas es 0 la formula no aplica por lo tanto se remplazo con un cero.
c) Para los registros de nombres científicos faltantes se creo un registro que los identifica con “SIN_NOMBRE” dadas las características son productos varios es diferentes categorías.

Estandarización

a) Se ajustaron los nombres de las variables a mayúsculas.
b) Se asigno el tipo de datos entero para todas las variables de IDs.

c) Se asignaron los tipos de datos acorde al diseño de la tabla de hecho y las dimensiones.
d) Se acotaron los dígitos de rendimiento a dos.
e) Se eliminaron los resúmenes de las variables numéricas.
f) Agregaron las categorías dependiendo del tipo del atributo.

Variables nuevas

a) Se creo la variable % **RENDIMIENTO_POR_AREA** la cual equivale al porcentaje de efectividad entre las hectáreas sembradas y las hectáreas cosechadas.
b) Se creo la variable **PAIS y CO_DEPTO** las cuales contiene información geografía adicional, esto con el fin de generar mapas en PBI, de la misma forma configurar.
c) Se creo la variable **PER2** esta variable nos permite entender los ciclos de producción en el tiempo.

Frecuencia de actualización

Acorde a la dinámica del negocio la actualización día vencido se ajusta a las necesidades e indicadores requeridos.

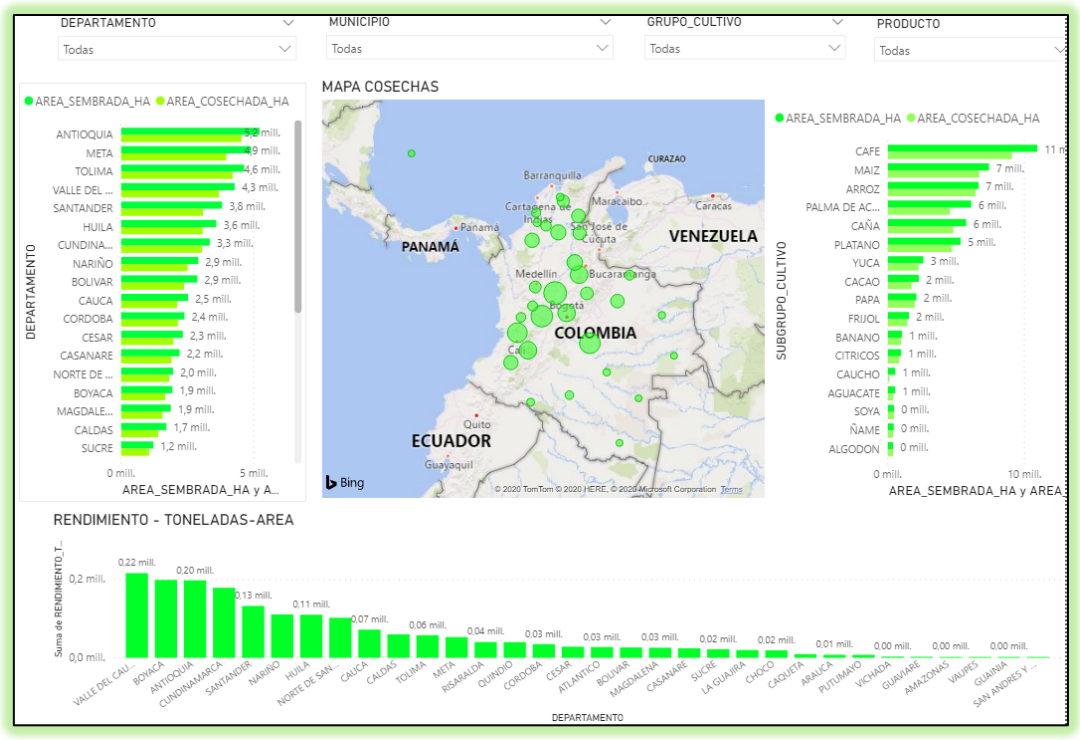
Visualización

Para la visualización de la data se utilizo un dashboard de power BI con 3 tableros principales:

- Productividad
- Tipos de Cultivos
- Períodos

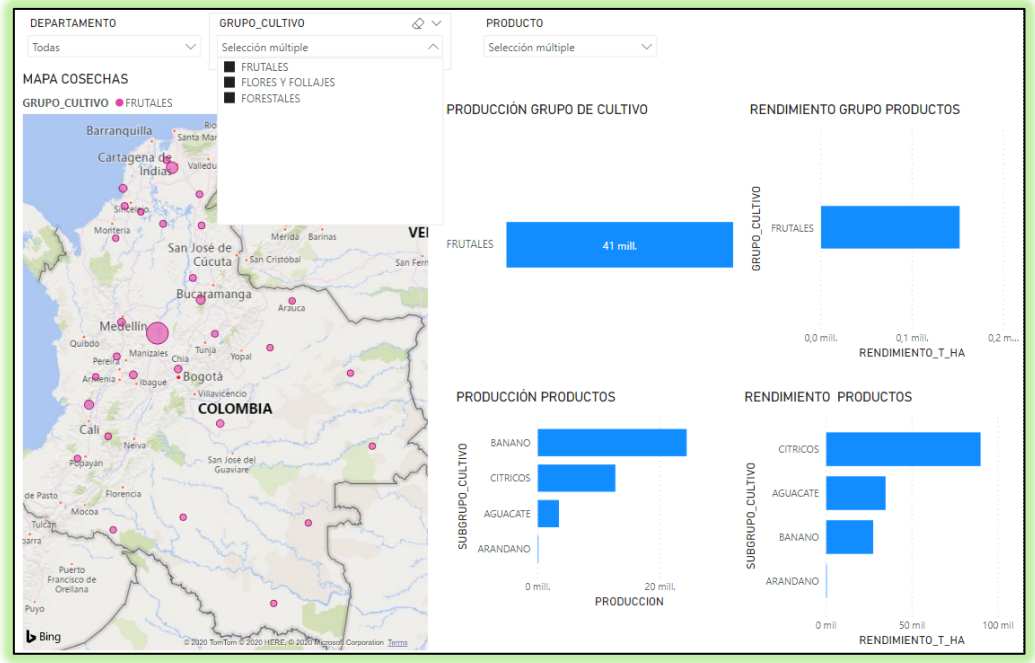
Así como se creo un home con 3 botones que redirigen a cada tablero.

En el tablero de productividad se visualizan indicadores de áreas sembradas vs áreas cosechadas a nivel de localización y por producto, se puede filtrar por departamento, Municipio, grupo de Cultivo y producto observando el resultado en las diferentes graficas y puntos de localización del mapa.



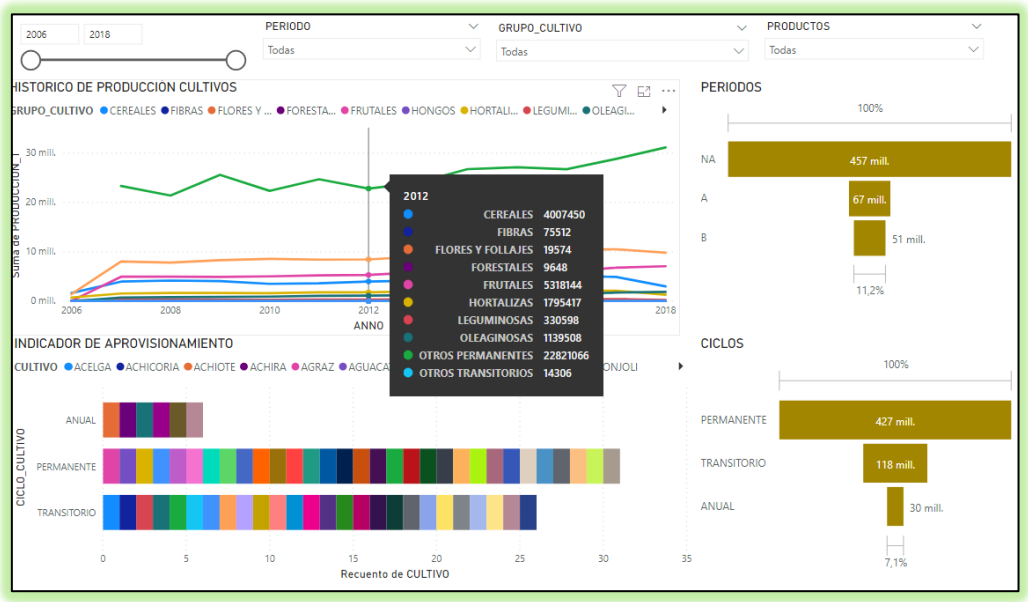
Dashboard de Productividad

En el tablero de Tipos de Cultivos están los indicadores que muestran estadísticas acerca del grupo de cultivo, productos y localización, se visualizan la producción y rendimiento de estos.



Dashboard de Productos

En el tablero de Períodos se hizo énfasis en cuanto al tiempo de los cultivos, periodos y ciclos de cultivos de los diferentes productos analizando cuales son anuales, permanentes y transitorios y de acuerdo con esto analizar la producción. Se puede filtrar por un rango de fecha, periodos, grupo de cultivo y productos.



Dashboard de Periodos

Datos de contacto:

Sandra Delgado - Analista de Datos

- https://www.linkedin.com/in/sandra-delgado-gomez
- https://github.com/salidego
- https://medium.com/sandra-delgado