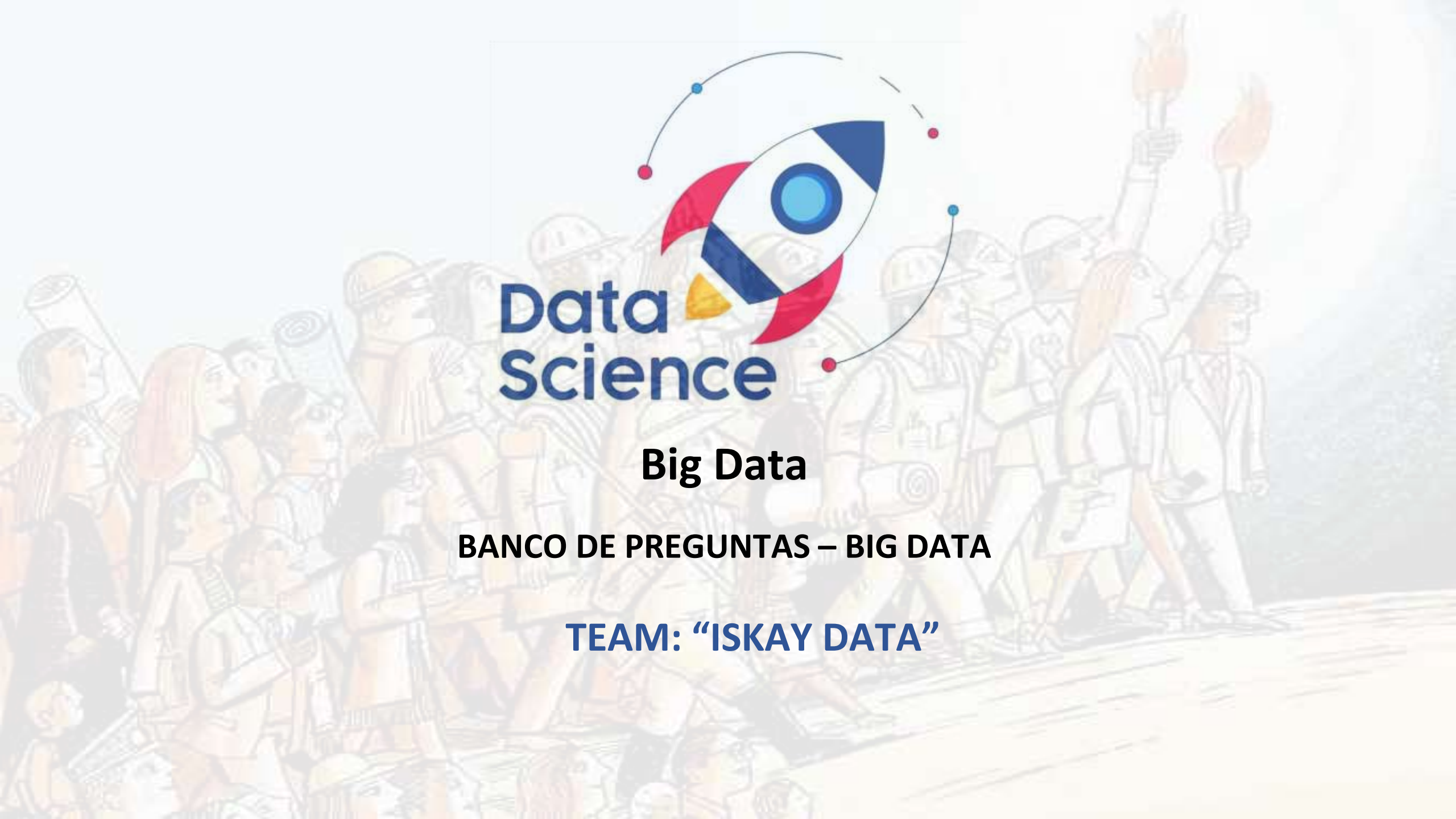


Big Data

BANCO DE PREGUNTAS – BIG DATA

TEAM: “ISKAY DATA”

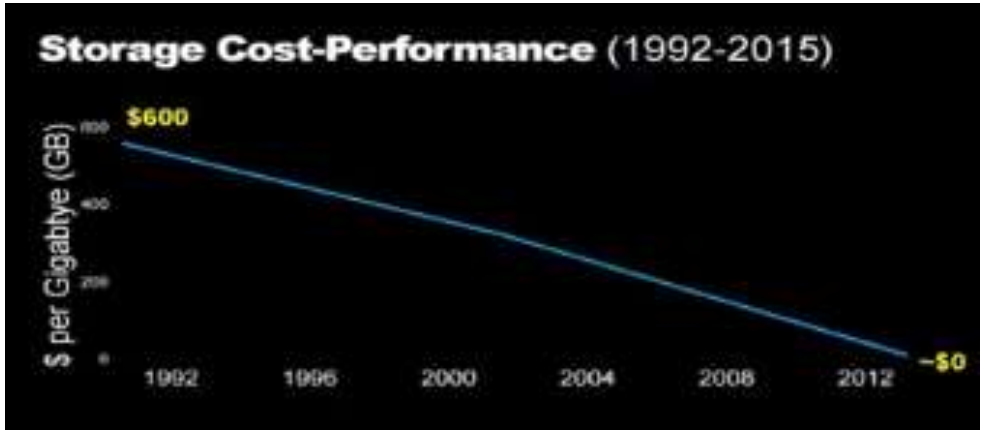
PREGUNTA 1

¿A qué se debe el crecimiento de Big Data?

a) Bajo costo en Procesamiento
b) Bajo costo en Almacenamiento
c) Bajo costo para trabajar en la cloud.
d) Todas las anteriores

RESPUESTA (D)

La principal causa que haya nacido Big Data como mercado es la evolución tecnológica, es decir, la caída de los costos de procesamiento, los costos de almacenamiento de datos y el bajo costo de poder trabajar en la cloud.



PREGUNTA 2

¿Cuál será el principal rol del área de TI con la llegada del Big Data?

a) Gestionar el hardware y software.
b) Lograr que los datos estén disponibles para la organización.
c) Capacitar a los usuarios con las aplicaciones.

RESPUESTA (B)

En un entorno Big Data, el área de TI ya no se encargará sólo de gestionar los temas de hardware y software de la empresa, su rol más importe será el de disponibilizar la data en la organización.

PREGUNTA 3

Cuáles son las funciones de un Big Data Engineer.

a) Definir las arquitecturas Big Data, si se trabajará en la nube o no.
b) Desarrollar modelos predictivos que ayuden en la toma de decisiones.
c) Desarrollar ETLs y mantener la operativa de los softwares que procesan datos a grandes escalas.

RESPUESTA (C)

Las funciones de un Data Engineer son las de desarrollar ETLs para la ingesta de datos y mantener la continuidad operativa de los distintos componentes Big Data que procesan un gran volumen de información.

PREGUNTA 4

Cuales son las principales V’s en Big Data

a) Veracidad, valor, volumen.
b) Volumen, velocidad, variedad.
c) Variedad, valor, viabilidad.

RESPUESTA (B)

Las características principales de los datos en un entorno Big Data son el volumen, la velocidad y la variabilidad.

PREGUNTA 5

Cuáles son las funciones de un Data Scientist.

a) Definir la arquitectura Big Data y desarrollar ETLs.
b) Crear modelos óptimos que perduren y ayuden en la toma de decisiones.
c) Generar datos y cargarlos al ecosistema Big Data.

RESPUESTA (B)

Las funciones de un Data Scientist es la de crear modelos analíticos que perduren, es decir que sean estables y que ayuden en la toma de decisiones .

PREGUNTA 6

En la fase de comprensión de Datos cuales son los dos retos principales

a) Diferenciar los datos estructurados y los semi estructurados.
b) Identificar las fuentes de información asociadas al problema y relacionar los conceptos.
c) Identificar a los actores en la recopilación de datos y darles la prioridad.

RESPUESTA (B)

En la fase de comprensión es importante identificar las fuentes de información que están asociadas a nuestra necesidad o problema y también es importante comprender los conceptos de los datos; es decir, que usuario lo genera, donde y de que forma.

PREGUNTA 7

Cuáles son las principales ventajas de los Sistemas Distribuidos.

a) Seguridad.
b) Tolerancia a fallos.
c) Alta Disponibilidad.
d) Todas las anteriores

RESPUESTA (D)

Las principales ventajas de un sistema distribuido son la seguridad, tolerancia a fallos y alta disponibilidad .

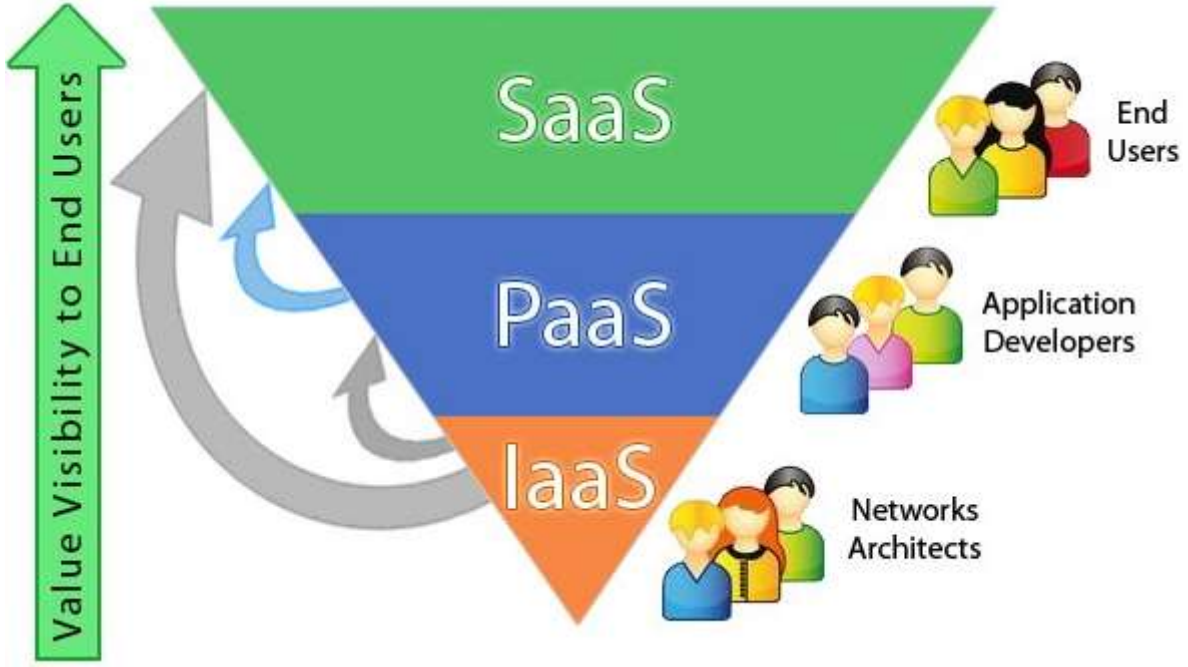
PREGUNTA 8

¿Cuáles son los servicios de la nube orientados para desarrolladores?

a) Iaas (infraestructura).
b) Saas (software)
c) Paas (plataforma).

RESPUESTA (B)

Los servicios en la nube orientados para desarrolladores son las Paas (plataforma como servicio).



PREGUNTA 9

Qué componentes Big Data nos ayudan en la ingesta.

a) Flume.
b) Hive
c) Sqoop.
d) a y c

RESPUESTA (D)

Algunos de los componentes Big Data que se utilizan para la ingesta de data son Flume (carga de archivo de distintas fuentes) y Sqoop (carga de archivos estructurados).

PREGUNTA 10

Respecto a las zonas del Data Lake, como se llama la zona donde los datos llegan en crudo.

a) Staging
b) Gold
c) Landing

RESPUESTA (C)

La zona Landing de Big Data es donde llega la data cruda, también se le conoce como RDV (Raw Data Voult).

PREGUNTA 11

¿Qué significa la modelización en el proceso de Big Data?

RESPUESTA

Luego de obtener la tabla de modelado,en la etapa de fusión con la construcción de variables derivadas,
Debemos conocer y elegir la técnicas a utilizar para la construcción del modelo,**supervisado o no superivsado.**
Los modelos se construyen aplicando el método científico sobre los datos.
Escoger el diseño de técnica de modelado
Utilizar métricas de evaluación (AIC, BIC, RMSE, KS) para los modelos.
Dividir los datos para mejorar la capacitdad analitica y evitar el overfitting.
Estimar los parámetros de los datos y seleccionar el mejor modelo con conjunto de evaluación.
Finalmente la capacidad analitica se calcula en el conjunto de test.

PREGUNTA 12

¿Cómo relacionar los datos de manera funcional?

RESPUESTA

Buscar identificadores para agrupar información y encontrar un identificador global o crear una regla para relacionar los datos.

De manera conceptual buscar como los conceptos se relacionan entre los datos (personas y productos relacionado mediante contratos)

No es recomendable focalizarse en las datos disponibles, ya que se debe de utilizar solo los datos necesarios.

PREGUNTA 13

¿Cual es la fase para desplegar en la plataforma de explotacion el modelo construido?.Defina sus etapas.

RESPUESTA

La fase se llama Despliegue y esta constituido por 3 etapas:

-Integración en la arquitectura: Tomar el modelo obtenido e introducir en el modelo de explotación de la organizacion

-Planificación temporal :Ejecutar el modelos cuando se tengan los datos disponibles y su captura.

-Integración con aplicaciones: Plantear como integral el modelo en la aplicación usando outputs , CRM , uso de apis o dentro de otra aplicación.

PREGUNTA 14

¿Que objetivo y técnicas de Big Data se utilizaría para la detección de fraudes en una organización del sector de seguros?

RESPUESTA

Para detectar los futuros fraudes de contratación a ocurrir en la empresa de seguros, se tendría , que identificar como prioridad relaciones ocultas entre titulares y tomadores.

Las técnicas aconsejables a utilizar serian una teoría de grafos para conocer las relaciones y analizarlos en un mapa.
Para el análisis de la influencia se utilizará un modelo supervisado para los coring de relación

PREGUNTA 15

¿Cual es el concepto de HDFS en el software libre Hadoop?

RESPUESTA

“Hadoop Distributed File System”, es un sistema de ficheros creado por el software hadoop, el cual lo subdivide en pequeños ficheros llamados **chunks.**
Los chunks se distribuyen en distintas máquinas de un sistema distribuido usando un concepto de replicado para evitar la accesibilidad del dato.
Cada chunks contiene información de control o datos de usuario.

Estos chunks se encuentran en el name node de la maquina con los metadatos de los datos, y estos se encuentran almacenados en los data nodes

Caso de Uso

Ingesta de datos de la Municipalidad de Miraflores (datos abiertos) en un ecosistema Big Data

Fuentes de Origen

DATA ESTRUCTURADA .CSV

Se encontró información de la Municipalidad de Miraflores las cuales fueron obtenidas en formato CSV.

Nombre	Clase
Autos_Mal_Estacionados.csv	Documento CSV
Estado_de_Infracciones.csv	Documento CSV
Fotopapeletas.csv	Documento CSV
Infracciones_Agosto2018.csv	Documento CSV
Infracciones_Empresas.csv	Documento CSV
Infracciones_Julio2018.csv	Documento CSV
Relacion_Empresa.csv	Documento CSV
Ruidos_ABRIL2018.csv	Documento CSV
Tipo_Infraccion.csv	Documento CSV

	A	B	C	D	E	F	G	H
1	ANIO,ID_ESTADO_INFRAC,ESTADO,NRO. RSAD,NOT RSAD,DETALLE INFRACCIÃ“N,LU							
2	2018,2,ORDINARIA,2017-8645,20180110,Por ocasionar ruidos molestos y constantes							
3	2018,2,ORDINARIA,2017-8700,20180110,Por ocasionar ruidos molestos y constantes							
4	2018,2,ORDINARIA,2018-0546,20180209,Por ocasionar ruidos molestos y constantes							
5	2018,2,ORDINARIA,2018-0594,20180214,Por ocasionar ruidos molestos y constantes							
6	2018,2,ORDINARIA,2018-0834,20180226,Por ocasionar ruidos molestos y constantes							
7	2018,2,ORDINARIA,2018-1441,20180402,Por ocasionar ruidos molestos y constantes							
8	2018,2,ORDINARIA,2018-1490,20180404,Por ocasionar ruidos molestos y constantes							
9	2018,2,ORDINARIA,2018-0116,20180112,"Por ocasionar ruidos molestos o persistent							
10	2018,2,ORDINARIA,2018-1163,20180314,Por ocasionar ruidos molestos y constantes							
11	2018,2,ORDINARIA,2018-1162,20180321,Por ocasionar ruidos molestos y constantes							
12	2018,2,ORDINARIA,2018-1159,20180321,Por ocasionar ruidos molestos y constantes							
13	2018,2,ORDINARIA,2018-1435,0,Por ocasionar ruidos molestos y constantes proveni							
14	2018,2,ORDINARIA,2018-1366,20180326,Por ocasionar ruidos molestos y constantes							
15	2018,2,ORDINARIA,2018-1361,20180405,Por ocasionar ruidos molestos y constantes							
16	2018,2,ORDINARIA,2018-1165,20180322,Por ocasionar ruidos molestos y constantes							
17	2018,2,ORDINARIA,2018-1164,20180321,Por ocasionar ruidos molestos y constantes							
18	2018,2,ORDINARIA,2018-1363,20180327,Por ocasionar ruidos molestos y constantes							
19	2018,2,ORDINARIA,2018-1365,20180328,Por ocasionar ruidos molestos y constantes							
20	2018,2,ORDINARIA,2018-1350,20180327,Por ocasionar ruidos molestos y constantes							
21	2018,2,ORDINARIA,2018-1158,20180321,Por ocasionar ruidos molestos y constantes							

Ingesta de Datos

Creación de directorios en HDFS

En esta etapa, se realiza la configuración de directorios en el sistema de archivos de Hadoop (HDFS)

```
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -mkdir /datalake/landing/proyecto/portalweb/fiscalizacion
18/08/25 20:41:37 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -mkdir /datalake/landing/proyecto/portalweb/fiscalizacion/ruido
18/08/25 20:41:43 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -mkdir /datalake/landing/proyecto/portalweb/fiscalizacion/estinfraccion
18/08/25 20:42:07 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -mkdir /datalake/landing/proyecto/portalweb/fiscalizacion/automalestacionado
18/08/25 20:42:53 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
```

```
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -ls /datalake/landing/proyecto/portalweb/fiscalizacion
18/08/25 20:43:39 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
Found 3 items
drwxr-xr-x   - ahuamani hadoop          0 2018-08-25 20:42 /datalake/landing/proyecto/portalweb/fiscalizacion/automalestacionado
drwxr-xr-x   - ahuamani hadoop          0 2018-08-25 20:42 /datalake/landing/proyecto/portalweb/fiscalizacion/estinfraccion
drwxr-xr-x   - ahuamani hadoop          0 2018-08-25 20:41 /datalake/landing/proyecto/portalweb/fiscalizacion/ruido
```

Ingesta de Datos

Carga de archivos CSV a Hadoop.

Realizamos la ingesta de data (archivos CSV, JSON) al sistema de archivos HDFS.

Esta actividad se realiza con el comando:

$> hdfs dfs -put [*Ruta de archivo origen*] [*Directorio destino*]

```
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -put "Relacion_Empresa.csv" /datalake/landing/proyecto/portalweb/transito/emptransporte
18/08/25 21:22:14 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -put "Fotopapeletas.csv" /datalake/landing/proyecto/portalweb/transito/fotopapeleta
18/08/25 21:23:45 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -put "Infracciones_Julio2018.csv" /datalake/landing/proyecto/portalweb/transito/infraccionubicacion
18/08/25 21:26:39 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -put "Infracciones_Agosto2018.csv" /datalake/landing/proyecto/portalweb/transito/infraccionubicacion
18/08/25 21:26:54 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -put "Infracciones_Empresas.csv" /datalake/landing/proyecto/portalweb/transito/rankempinfr
18/08/25 21:28:25 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
ahuamani@clustercandidatos-w-0:~$ hdfs dfs -put "Tipo_Infraccion.csv" /datalake/landing/proyecto/portalweb/transito/tipoinfraccion
18/08/25 21:29:03 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.7-hadoop2
```

Creación de Tablas

Se seleccionó Hive como almacenamiento de datos ya que cumple con las siguientes características:

- Hive es totalmente compatible con Haddop y MapReduce.
- Nos permite crear tablas externas a partir de carpetas del sistema HDFS.
- Consultas al estilo SQL.

Como punto de partida en HIVE, se procede a crear la Base de Datos:

```
-- Creación de Base de Datos
-- ================================================
CREATE DATABASE candidatura;
```

Creación de Tablas

Una vez creada la Base de Datos, se procede a crear las tablas. Estas serán creadas con el tipo EXTERNAL:

Las tablas a crear son:

- EMP_TRANSPORTE
- FOTOPAPELETA
- INFRACCION_UBICACION
- RANK_EMP
- TIPO_INFRACCION
- AUTOMALESTACIONADO
- ESTINFRACCION
- RUIDO

Ejemplo:

```
-- Tabla: emp_transporte
CREATE EXTERNAL TABLE IF NOT EXISTS candidatura.emp_transporte(
CODIGO STRING COMMENT 'Código de empresa',
EMPRESA_TRANSPORTE STRING COMMENT 'Nombre de empresa'
)
COMMENT 'tabla de nombre de empresas'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS TEXTFILE
LOCATION '/datalake/landing/proyecto/portalweb/transito/emptransporte'
tblproperties("skip.header.line.count" = "1");
```

Creación de Tablas

Procedemos a consultar algunas de las tablas para validar que fueron generadas correctamente y que contienen la información correspondiente:

```
hive> select * from candidatura.fotopapeleta limit 10;
OK
2016    1       Av. 28 de Julio 9       Septiembre      2
2016    2       Ca. Colon       9       Septiembre      32
2016    5       Av. La Paz      9       Septiembre      ---
2016    7       Mlcon. De La Reserva    9       Septiembre      116
2016    9       Ca. San Martin  9       Septiembre      61
2016    11      Av. Del Ejercito        9       Septiembre      ---
2016    12      Ca. Grimaldo del Solar  9       Septiembre      3
2016    14      Ca. Alcanfores  9       Septiembre      11
2016    15      Ca. Enrique Palacios    9       Septiembre      1
2016    16      Ca. Schell      9       Septiembre      1
Time taken: 0.107 seconds, Fetched: 10 row(s)
```