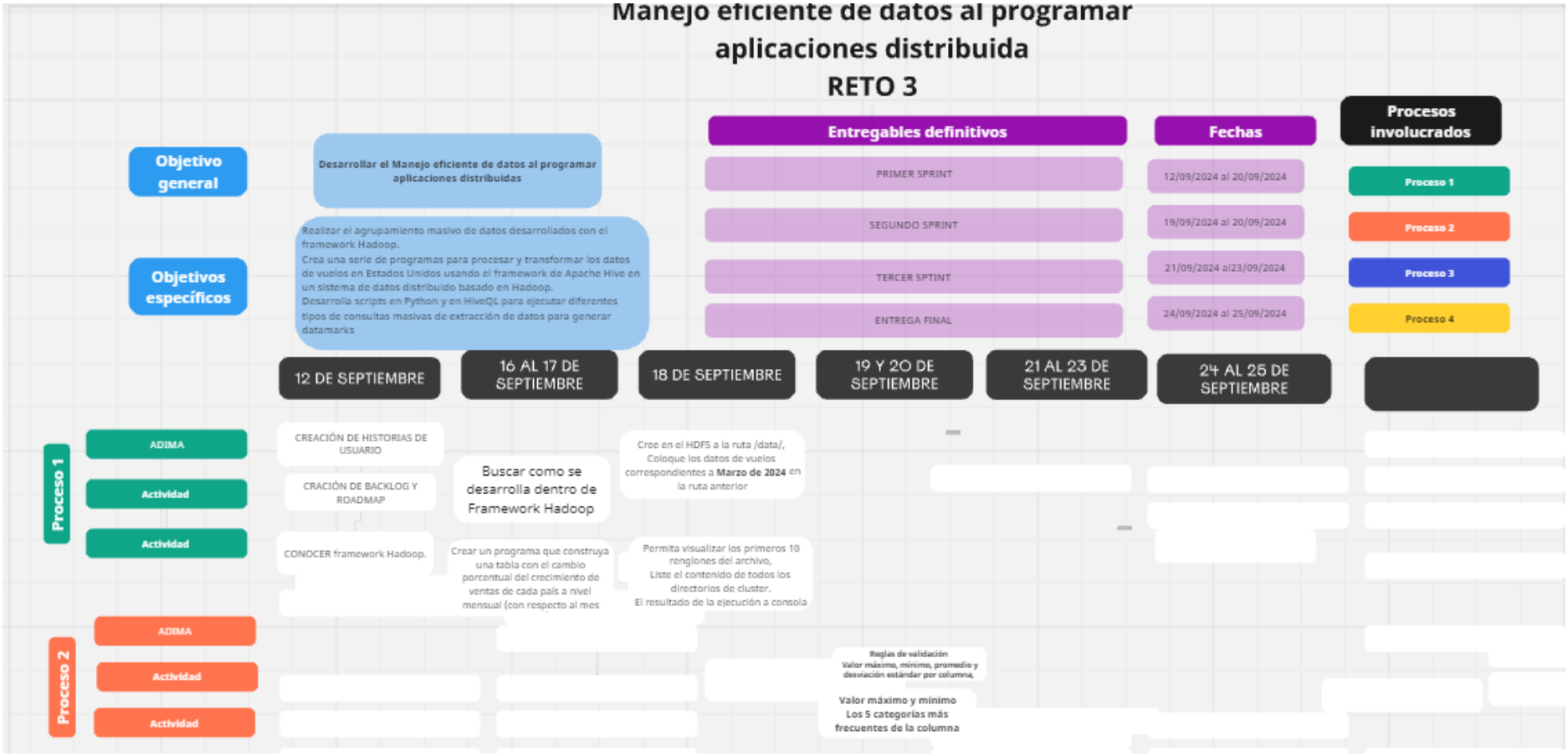
Manejo eficiente de datos al programar aplicaciones distribuida
RETO 3
Objetivo general
Desarrollar el Manejo eficiente de datos al programar aplicaciones distribuidas
Objetivos específicos
Realizar el agrupamiento masivo de datos desarrollados con el framework Hadoop.
Crea una serie de programas para procesar y transformar los datos de vuelos en Estados Unidos usando el framework de Apache Hive en un sistema de datos distribuido basado en Hadoop.
Desarrolla scripts en Python y en HiveQL para ejecutar diferentes tipos de consultas masivas de extracción de datos para generar datamarks
Entregables definitivos
PRIMER SPRINT
SEGUNDO SPRINT
TERCER SPTINT
ENTREGA FINAL
Fechas
12/09/2024 al 20/09/2024
19/09/2024 al 20/09/2024
21/09/2024 al23/09/2024
24/09/2024 al 25/09/2024
Procesos involucrados
Proceso 1
Proceso 2
Proceso 3
Proceso 4
12 DE SEPTIEMBRE
16 AL 17 DE SEPTIEMBRE
18 DE SEPTIEMBRE
19 Y 20 DE SEPTIEMBRE
21 AL 23 DE SEPTIEMBRE
24 AL 25 DE SEPTIEMBRE
Proceso 1
ADIMA
Actividad
Actividad
CREACIÓN DE HISTORIAS DE USUARIO
CRACIÓN DE BACKLOG Y ROADMAP
CONOCER framework Hadoop.
Buscar como se desarrolla dentro de Framework Hadoop
Crear un programa que construya una tabla con el cambio porcentual del crecimiento de ventas de cada país a nivel mensual (con respecto al mes
Cree en el HDFS a la ruta /data/, Coloque los datos de vuelos correspondientes a Marzo de 2024 en la ruta anterior
Permita visualizar los primeros 10 renglones del archivo,
Liste el contenido de todos los directorios de cluster.
El resultado de la ejecución a consola
Proceso 2
ADIMA
Actividad
Actividad
Reglas de validación
Valor máximo, mínimo, promedio y desviación estándar por columna,
Valor máximo y mínimo
Los 5 categorías más frecuentes de la columna