Presentación Ejecutiva: Proyecto Final Gran Escala

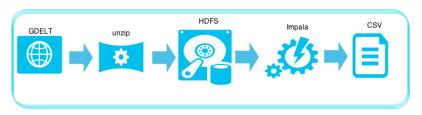
Luis Roman, Alfonso Kim, Fernanda Mora, Andrea García y Sergio Haro

5 de junio de 2015

Implementación de una arquitectura de Gran Escala usando distintos frameworks y lenguajes

El objetivo de este proyecto es practicar las herramientas vistas en clase tales como: Sqoop, Flume, Hive o Impala, Luigi y Pig o PySpark para un flujo significativo de datos.

El flujo que Luigi que se implementó es el siguiente



Arquitectura

El flujo de datos debe estar orientado a procesos y no tareas fragmentadas. El proceso de orquestación es vital para asegurar la resiliencia del flujo de datos.

En cuanto a la ingesta de datos preferimos Flume a Kite ya que Flume te permite procesar y filtrar directamente sin pasar por tantos procesos como en Kite (para leer de diferentes fuentes y esquemas, procesarlos y pasarlos a diferentes destinos etc..). Sqoop nos gustó por qué nos permite manejar los datos desde y hacia una base relacional. a pesar de ser menos potente que Flume (por qué solo es para BD relacionales) nos gustó por su facilidad de uso (se conecta a Postgres y a otras bases de datos).

Extras

Para mayor información dirigirse a la presetacion extendida y a la documentación del trabajo https://github.com/AndyGT/big-data/tree/master/grupos/LAFAS/Proy%20final_v4

"proyecto_final.pdf"

"presentacion_extendida.html"

