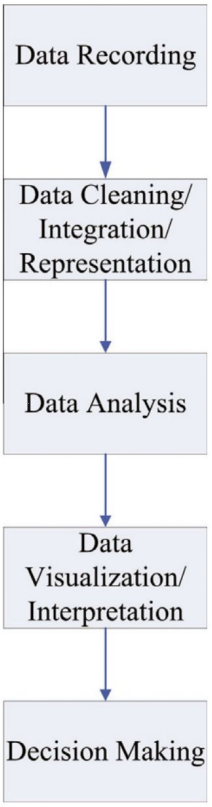
Trabajo Final Sistemas de Bases de Datos Masivos

1. Introducción

En este trabajo usted demostrará y aplicará los conocimientos adquiridos en el curso de Sistemas de bases de datos masivos. Durante este trabajo usted entenderá las distintas fases del procesamiento de grandes conjuntos de datos como se muestra en la siguiente gráfica:

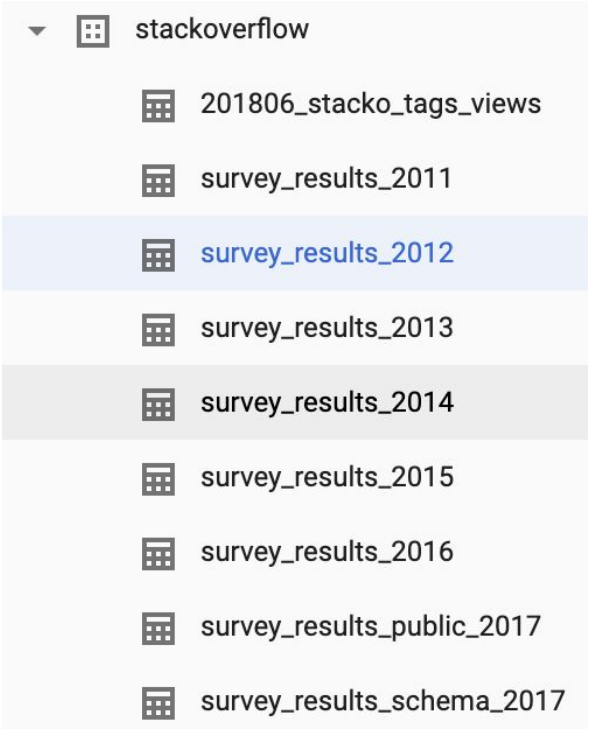


2. Motivación

Analizar tendencias dentro del mundo de la tecnología se ha convertido en uno de los más grandes retos de la industria, prever ¿Cuál será el siguiente y más éxitoso lenguaje de programación? Ó ¿Cuál sería el salario ideal para un desarrollador por tecnología? Se convertirá en una tarea crítica durante los próximos años, Stack Overflow es un sitio web ampliamente utilizado por la comunidad de desarrolladores de software, en la cual otros desarrolladores pueden encontrar soluciones a problemas de programación en diferentes lenguajes. Este sitio realiza desde el 2011 una encuesta a sus usuarios para observar y analizar tendencias en la industria de la tecnología y el software: https://insights.stackoverflow.com/survey/2019 La más reciente fue durante el mismo 2019.

3. Enunciado

Las encuestas han sido cuidadosamente almacenadas y expuestas en directorios de datos abiertos. Para nuestro propósito utilizaremos 7 conjuntos de datos masivos completamente libres y accesibles a través de BigQuery:



Considerando los siguientes conjuntos de datos:

Fh-bigquery.stackoverflow.survey_results_2011
Fh-bigquery.stackoverflow.survey_results_2012
Fh-bigquery.stackoverflow.survey_results_2013
Fh-bigquery.stackoverflow.survey_results_2014
Fh-bigquery.stackoverflow.survey_results_2015
Fh-bigquery.stackoverflow.survey_results_2016
Fh-bigquery.stackoverflow.survey_results_public_2017

I. Haga una limpieza de los datos identificando los campos comunes entre los conjuntos de datos. (Adjuntar Rutina SQL o Python)

II. Genere una Bodega de datos que permita responder al menos 15 preguntas elaboradas por el equipo. (Adjunte las 15 preguntas con las 15 consultas SQL que respondan a las preguntas y la respectiva gráfica)

III. Adjunte sobre la bodega de datos los modelos de Estrella, Cubo y Malinowski.

IV. Aplique MapReduce para llevar esos datos a una base de datos NoSQL que permita agrupar el número de programadores por lenguaje de programación y las respuestas por pregunta de los 5 conjuntos en paralelo* (Mediante Map Reduce).
(Adjunte la rutina en Python, SQL y/o Javascript que le permitió obtener la base de datos SQL en paralelo, deberá adjuntar evidencia del procesamiento en paralelo y hará parte de la exposición la demostración de la rutina).

V. Genere un conjunto de datos nuevos a partir de CSV con al menos 50 mil registros y prediga cuáles serían las respuestas de una nueva encuesta en el 2018. (Adjunte el análisis y el algoritmo que le permitió generar los registros manteniendo la línea de tendencia con base en las encuestas anteriores).

VI. Genere una rutina** que permita ingresar un registro en cualquiera de las encuestas y que actualice todos los indicadores hacia delante.

VII. Genere una rutina que actualice un indicador y “modifique” los “datos operativos” y la bodega de datos de manera coherente para corresponderse con el indicador.

Nota: Para los siete (7) Entregables anteriores elabore un informe de lectura en formato MarkDown* que explique y reúna claramente lo solicitado en los puntos anteriores debe ser enviado <u>antes</u> del viernes 22 de Junio a la 1:59PM.**

* El objetivo de MapReduce es paralelizar y resumir grandes conjuntos de datos
** Porción de código en cualquier lenguaje.
*** https://es.wikipedia.org/wiki/Markdown y https://stackedit.io