Descripción Práctica 3
Persistencia, Buscadores y Arquitectura

Jenaro Gallego

Práctica **P3**

- Objetivo de la práctica:
  - Utilizar distintas opciones en el formato de almacenamiento
  - Utilizar distintas tecnologías en el entorno Hadoop y fuera del entorno hadoop que permitan realizar una analítica de los datos
  - Explotar la información utilizando buscadores y su ecosistema
  - Diseñar esquemas mostrando los componentes utilizados en cada paso

- Haced backups continuamente de vuestro trabajo
- Apoyaros en Python/notebooks en la medida en la que sea necesario
- Utilizad scripts para poder repetir lo hecho hasta el momento e incluidlos en el entregable en la medida de lo posible

Índice del Entregable

- **Introducción**
  - Descripción y/o Descubrimiento del juego de datos a utilizar (opendata o cualquier otro juego de datos disponible)
  - Hipótesis a comprobar
  - Pasos a seguir
- **Por cada paso seguido o componente utilizado**
  - Arquitectura/Visión general de componentes utilizados
  - Desarrollo realizado:

    Analítica o búsqueda o visualización realizadas con el objetivo de comprobar la hipótesis
  - Resultados conseguidos
- **Conclusiones**

Secuencia

- Seleccionad un juego de datos de Open Data
- Familiarizaros con los datos
- Seleccionad una hipótesis que pueda ser aplicable
- Tecnologías a utilizar:
  - Parquet, ORC como formato para facilitar las consultas analíticas
  - Hive y Hbase como bases de datos sobre las que realizar la analítica utilizando distintos formatos de persistencia (Parquet/ORC) e interoperatividad entre componentes (Hive y Hbase)
  - ElasticSearch/SolR como entornos de búsqueda y descubrimiento
  - Kibana para visualización y descubrimiento
- Revisamos resultados
- Confirmamos/rechazamos la hipótesis

Valoración de la práctica

Tecnología	Valoración	Alternativas
Descubrimiento (Buscadores)	Mínima	Query utilizando buscador
	Media/Alta	Si los datos incluyen series temporales, cargarlas en Kibana y mostrar gráficos temporales
	Media/Alta	Visualización de datos con opciones de Kibana
Persistencia (Hadoop)	Mínima	Hdfs, hive
	Media	Incorporando formato de almacenamiento distinto a CSV (y utilizándolo): Parquet, OCR
	Alta	Carga en Hbase (utilizando compatibilidad con Hive)
Analítica (Hadoop)	Mínima	SQL sobre HIVE
	Media / Alta	SQLs complejos, joins

Anexos del entregable

- **Código:** scripts/programas utilizados en cada paso

- **Juego de datos original** (enlace para poder acceder a él o copia si es factible)

Por ejemplo

- Cojo los datos de covid 19 de la Comunidad de Madrid: https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos
- Reviso qué información incluye los juegos de datos
- Hipótesis: ¿cómo influyeron las vacaciones o las fiestas?
- Cargo los datos en ElasticSearch, configuro series temporales
- Visualizo en Kibana (sin interpretar)
- Cargo los datos en Hive y veo con qué tipo de consultas puedo confirmar o rechazar la hipótesis.
- Reviso las conclusiones obtenidas con la información obtenido a través de los buscadores

Otros juegos de datos

- Son válidos cualquier otro juego de datos que deseéis y con los que estéis familiarizados: deportes, vuelos, naturaleza, sanidad, películas, música, etc.
- Algunas otras fuentes:
  - https://www.kaggle.com
  - https://www.data.world
  - https://www.grouplens.org