UNIVERSIDAD EAFIT

ST0263: Tópicos Especiales en Telemática, 2024-2

Trabajo 3 – Automatización del proceso de Captura, Ingesta, Procesamiento y Salida de datos accionables para realizar la gestión de datos de Covid en Colombia (Arquitectura Batch para big data)

Fecha de entrega: 24 de noviembre de 2024.

Descripción:

Durante los diferentes temas vistos en la unidad 3, se ha podido evidenciar los retos que puede conllevar ejecutar cada una de las etapas del ciclo de vida de un proceso analítico conformado por la Captura de datos en la Fuente, la Ingesta, el Almacenamiento, el Procesamiento y los resultados.

Muchos de estos procesos, los realizamos durante el curso con procesos manuales, pocos datos y demos muy básicos que nos alejan de la realidad de los procesos de ingeniería de datos reales en las empresas.

Para esto, realizaremos un proyecto más cercano a un prototipo real de un caso de ingeniería de datos big data.

Como fuente de datos, utilizaremos los datos provistos en línea por el Ministerio de Salud para datos Covid, el acceso se puede hacer por archivos o por APIs (explorar ambos casos). Además, contaremos con una base de datos relacional real, tipo MySQL o Postgres, que contendrá datos simulados que se puedan requerir para completar el análisis de datos de Covid en Colombia. Esto nos permitirá al menos experimentar con 2 fuentes reales en las empresas (archivos en URLs o APIs y acceso a bases de datos).

A nivel de ingesta de datos, debemos crear algún proceso automático, que nos permita descargar los datos de covid por Archivo y API, y almacenarlos en un Bucket S3 en la zona Raw. También debemos crear un proceso automático, para la extracción de los datos de la base de datos relacional mysql o postgress y almacenar los datos en un bucket S3 en la zona raw. Analice la posibilidad de utilizar Hadoop Sqoop de EMR para realizar la ingesta de datos desde una BD-SQL hacia S3. El producto natural para este proceso es DMS pero no está habilitado en AWS Academy. Explore posibilidades.

A nivel de procesamiento, utilizaremos un clúster de EMR con procesamiento en Spark para realizar 2 tareas:

- Automatizar procesos ETL mediante Steps en un clúster EMR con Spark. Los requerimientos de estos procesos ETL será realizar procesos de preparación de

datos y unión de datos de covid provenientes del Ministerio de Salud con Datos de la base de datos relacional. El resultado de los procesos ETL en Spark deberán ser almacenados en un bucket S3 en la zona Trusted. Tanto el proceso de creación del clúster EMR como el procesamiento ETL deberán ser automáticos.

- Automatizar procesos de análisis, analítica o aprendizaje de máquina sobre los datos preparados en la zona Trusted. En principio, aplicar los mismos procesos analíticos realizados en el lab3-3 con mejoras combinando los datos de la base de datos relacional. **DE MANERA OPCIONAL PUEDE implementar al menos un modelo de aprendizaje de máquina supervisado o no supervisado utilizando dataframes pipelines con SparkML**. Los procesos de analítica de datos descriptivos deben ser implementados con dataframes pipelines y con SparkSQL. El resultado del análisis de datos, deben ser enviados a un bucket S3 en la zona Refined. Estos resultados deben poder ser consultados de diferentes formas: Athena y API Gateways. Implementar ambos.
  Recuerde que todos los procesos deben ser automáticos y no manuales.

A nivel de aplicación, los resultados del proceso analítico deben poder ser consultados por Athena o vía una API, realizar el prototipo que corresponda para demostrar estas aplicaciones.

Algunos otros requerimientos y recomendaciones:

- La captura e ingesta de datos hacia S3 deberá ser automática sin intervención humana.
- La creación del clúster, las pruebas y el desarrollo las puede hacer manualmente, pero cuando ya esté listo el proyecto3 tanto la creación del clúster, la realización del procesamiento ETL y analítica, y los resultados programados en los 'Steps' del clúster deben ser automáticos sin intervención humana.
- La primera opción de implementación es con AWS Academy, aunque todos conocemos las limitaciones que puede tener para utilizar algunos servicios. Deberá buscar alternativas en AWS Academy sino cuenta con un servicio, sea creativo. SIN EMBARGO, si se siente con confianza y dominio de las mismas tecnologías que utilizaremos en AWS pero en otra nube como GCP o Azure, se invita y promueve su realización en estas otras nubes para implementar este mismo proyecto3. Se cuenta con créditos oficiales para otorgar en GCP y Azure. La ventaja de estas últimas 2, es que podemos utilizar todos los servicios sin limitación de permisos, pero si de costo (Máx 50 USD).

A nivel de entregables:

- Repositorio github del proyecto3 donde estén todos los scripts, programas, instrucciones, documentación, etc, para replicar el proyecto.
- Un documento readme.md en el repositorio github del proyecto3.
- Una videosustentación del proyecto3
- Presentación y sustentación del proyecto3, activando todos los recursos, mostrando la automatización y exponiendo las diferentes etapas. Esta sustentación se realizará el día 25 de noviembre, con turnos de 30 mins. Se enviará agenda con la citación. Debe ser presencial.

Referencias:

- Video y github de un caso sencillo de ejecución de Spark en EMR con Steps.
    - https://youtu.be/ZFns7fvBCH4?si=hu5Y34JDB9yY7bsd
    - https://github.com/airscholar/EMR-for-data-engineers/tree/main

Fecha de entrega máxima: 24 de noviembre de 2024 23:59 por buzón de Interactiva virtual