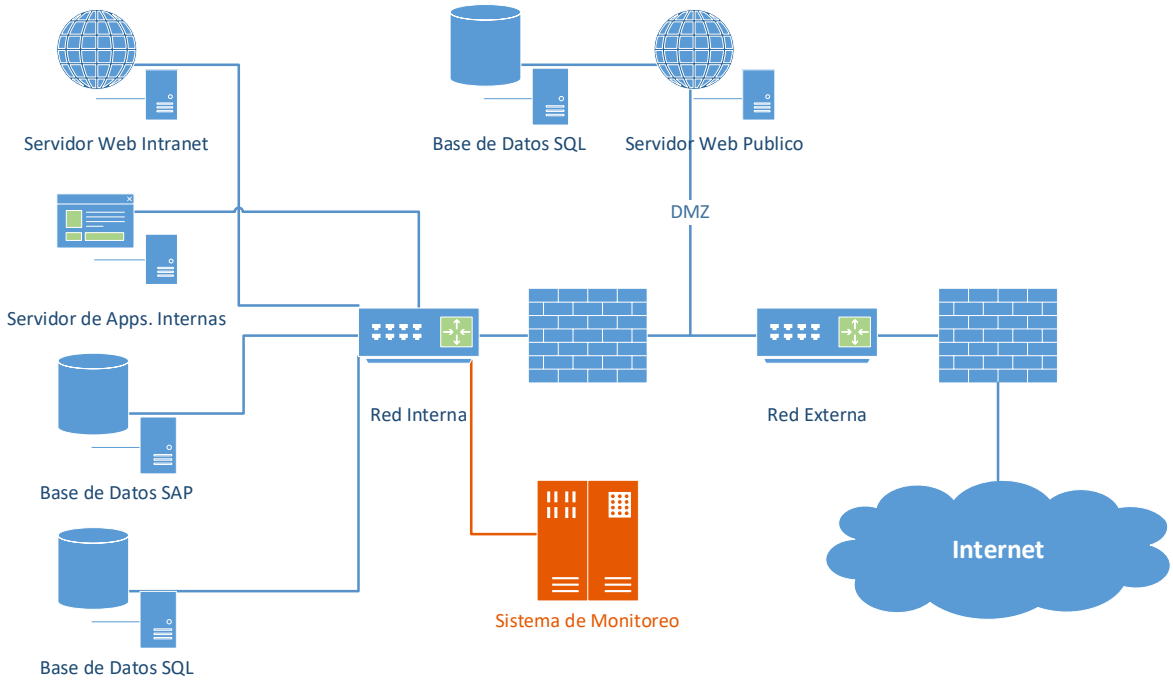
Evaluación Módulo 6 – Big Data

La empresa ACME S.A. se ve enfrentada a la necesidad de centralizar el monitoreo de diversos sistemas dentro de la arquitectura empresarial:



En dicho Sistema de Monitoreo se deben recoger los logs de los servidores:

- Servidor Web Intranet
- Servidor de Apps. Internas
- Base de Datos SAP
- Base de Datos SQL
- DMZ:
  - Servidor Web Publico
  - Base de Datos SQL

1. Pasando por alto las consideraciones de arquitectura, seguridad y otros supuestos, **diagramar funcionalmente los aspectos requeridos para lograr el monitoreo inicialmente planteado. Vale decir:**
   - ¿Cómo recolectamos los logs (archivos planos, recolección directa, en tiempo real, etc.)? Justifique su elección
   - ¿Dónde almacenamos los datos (Hadoop HDFS, almacenes temporales, Data Lakes, etc.)? ¿Es necesario utilizar almacenes intermedios? Justifique su elección
   - ¿Qué sistemas utilizamos para procesar la información (Hive, Spark, HBase, etc.)? Justifique su elección
   - ¿Cómo consumimos la información (Reportería, Machine Learning, modelos estocásticos, etc.)? Justifique su elección

Para responder lo anterior puede utilizar cualquier supuesto que estime conveniente. Ver Anexo y fuente del Paper. Puede ayudar a responder de mejor manera

2. Para realizar la prueba de concepto del sistema antes planteado, se tiene una muestra de los Logs del Sistema Operativo del Servidor de Apps (*EventosSistema.txt* → Separado por TABS). Con esta muestra, debe realizar los siguientes ejercicios:

- **I.** Cargar los datos en el HDFS Usando Flume
- **II.** Cargar los datos en HIVE
- **III.** Conteo de Palabras en HIVE
- **IV.** Generar una Tabla de Eventos. Vale decir, una tabla que contenga una lista con los distintos tipos de eventos, mostrando ***Id. del evento, Origen, Nivel y Conteo (Cantidad de Eventos por ID).***

  Por ejemplo (***Valores de ejemplo, no representan la solución final***):

Id. del Evento	**Origen**	**Nivel**	**Conteo**
10016	Microsoft-Windows-DistributedCOM	Advertencia	1345
16	Microsoft-Windows-Kernel-General	Información	986
234	Microsoft-Windows-Hyper-V-VmSwitch	Información	32

- **V.** **(BONUS)** Almacenar la tabla creada en una tabla de HBase
- **VI.** **(BONUS)** Crear una pequeño reporte (2 – 4 objetos visuales) con los datos extraídos.

3. Responda:

- **I.** Explique el término "Big Data" y cuáles son las cinco "V" del mismo.
- **II.** ¿Cuáles son las diferencias entre un sistema de archivos normal y HDFS?
- **III.** ¿Cuáles son las diferencias entre lo que llamamos Hadoop 1.0, Hadoop 2.0 y Hadoop 3.0?
- **IV.** ¿Por qué HDFS es tolerante a fallos?
- **V.** Explique la arquitectura de HDFS.
- **VI.** ¿Qué problema soluciona YARN dentro del ecosistema de Hadoop?
- **VII.** ¿Cuáles son los componentes clave de HBase?
- **VIII.** Explique qué es una *Row Key* y las *Column Families* en HBase.

Instrucciones de Entrega

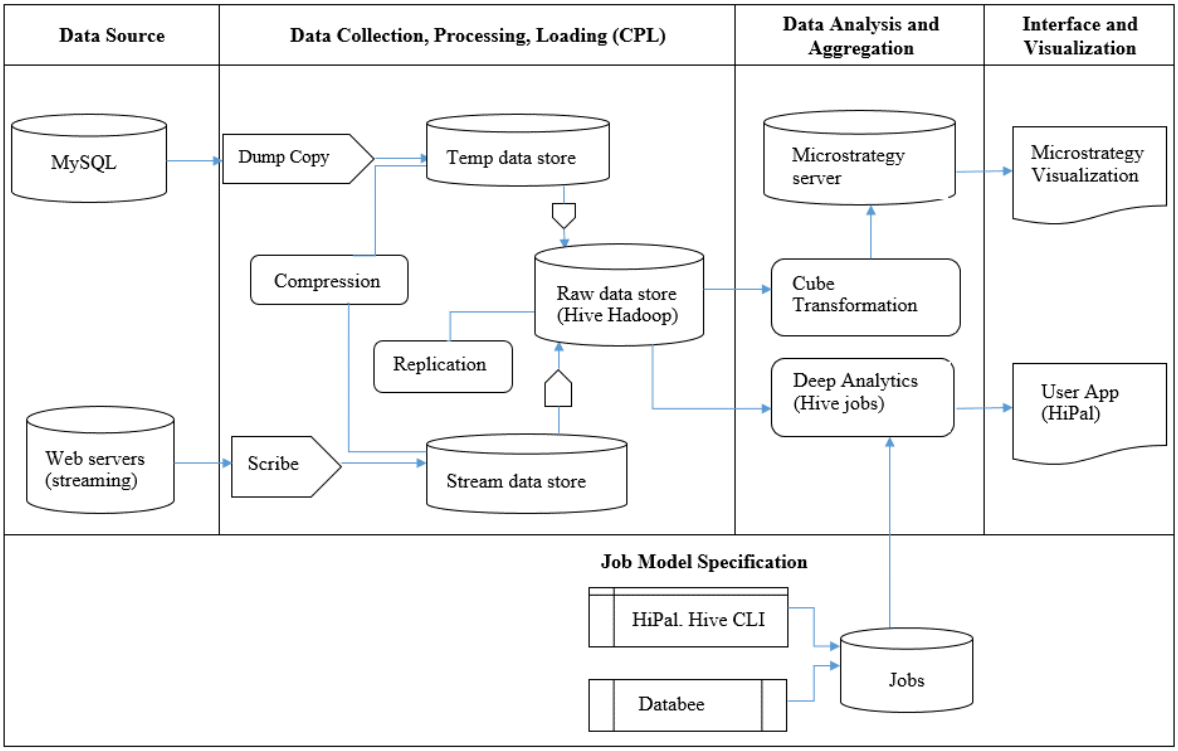
Para entregar la evaluación, subir a su Carpeta Personal de Sharepoint el o los archivos creados. No se exige formalidad, por lo que es válido cualquier formato:

- PDF
- Word
- Imágenes
- Scan de Cuaderno
- Etc.

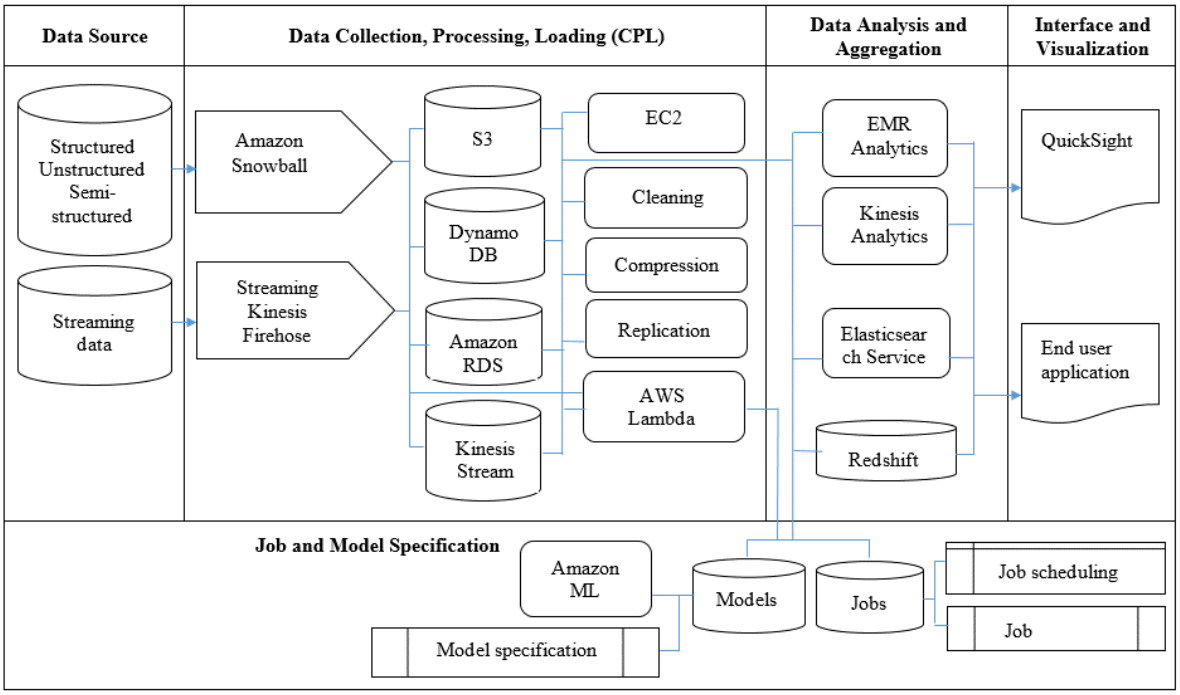
El plazo máximo de entrega es hasta el día **11 de mayo hasta las 23:59.**

Ante cualquier situación, favor de comunicarse con los relatores y/o la coordinación a cargo del curso.

Anexo – Arquitecturas de Referencia – Facebook y Amazon



Arquitectura Big Data Facebook



Arquitectura Big Data Amazon

Fuente: *A Reference Architecture for Big Data Systems*, Go Muan Sang, Lai Xu, Paul de Vrieze Faculty of Science and Technology, Bournemouth University
https://pdfs.semanticscholar.org/49d8/61279e15d36c4af1448dcb550976ced62cf7.pdf