UNIVERSIDAD TECNICA FEDERICO SANTA MARIA | DEPARTAMENTO DE INFORMÁTICA

Tarea de Laboratorio 1: Cassandra

INF-325 Bases de Datos Avanzadas

Profesor: Mauricio Figueroa Colarte

08 de octubre de 2021

1 CONTEXTO

Se requiere diseñar, poblar y consultar una base de datos en Cassandra que almacene las postulaciones a **Universia**, para lo cual se dispone del dataset llamado ***postulaciones.xlsx***. Esta fuente de datos es un Excel que recopila registros de postulaciones y matriculas efectivas para los períodos 2015, 2016 y 2017, con información consolidada desde el DEMRE[1] quien aporta información de las postulaciones, datos demográficos, geográficos, académicos, preferencias, becas y gratuidad.

Los datos fueron consolidados y depurados para hacer gestión a través de análisis OLAP, presentando 16 campos denominados como: CEDULA, PERIODO, SEXO, PREFERENCIA, CARRERA, MATRICULADO, FACULTAD, PUNTAJE, GRUPO_DEPEN, REGION, LATITUD, LONGITUD, PTJE_NEM, PSU_PROMLM, PACE, GRATUIDAD (ver diccionario de datos en tabla 1).

Tabla 1: Diccionario de Datos

Campos Categóricos	Campos Numéricos
CEDULA (RUT identificador del postulante) PERIODO (2015, 2016, 2017) SEXO (MASCULINO, FEMENINO) PREFERENCIA (1,2,3,4,5,6,7,8,9,10) CARRERA (Lista de carreras de la UCM) ESTADO (MATRICULADO, NO MATRICULADO) FACULTAD (Facultades de la UCM) GRUPO_DEPEN (MUNICIPAL, PARTICULAR SUBVENCIONADO, PARTICULAR PAGADO) REGION (Nombres de las regiones de Chile) PACE (PACE o Blanco) GRATUIDAD (SI, NO)	PUNTAJE (Puntaje Ponderado PSU) LATITUD (Latitud de la región) LONGITUD (Longitud de la región) PTJE_NEM (Puntaje Enseñanza Media) PSU_PROMLM (Puntaje Promedio Lenguaje Matemáticas)

2 REQUISITOS

1. Implementar la arquitectura de Cluster con un Data Center Cassandra (Simple Strategy) con 3 nodos operativos sobre Docker y con Factor de Replicación 3. Un nodo deberá estar operativo en el computador local del estudiante, los otros dos nodos deberán estar operativos en máquinas virtuales de AWS, uno por instancia. Para esto último se encuentra habilitado el *Classroom* en *Amazon Educate* para la asignatura.

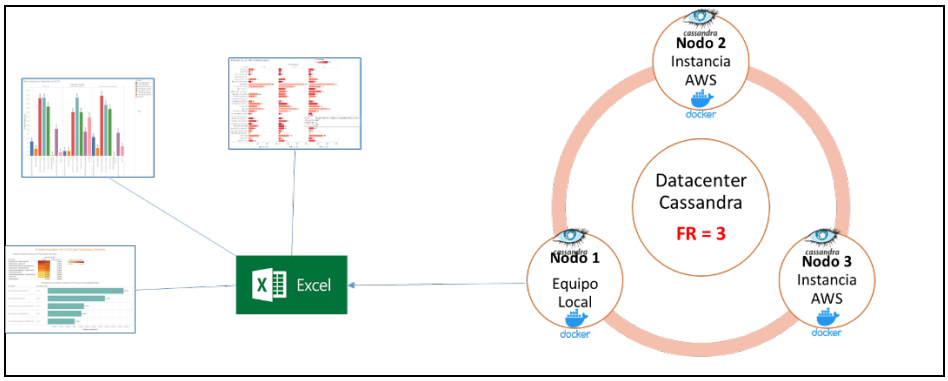

Figura 1: Arquitectura General

[1] http://www.demre.cl

2. En base al dataset ***postulaciones.xlsx*** entregado y las consultas requeridas para resolver las preguntas del negocio, implementar el Diseño físico de la base de datos en CQL, explicando razonadamente el motivo por el que se diseña de la forma propuesta. En este sentido, el modelado de los datos debe permitir cumplir con los siguientes objetivos de la manera más equilibrada posible:

   Regla 1: Distribuir los datos por todo el clúster:

   - Es deseable que cada nodo del clúster tenga un volumen de datos similar (equilibrio).
   - Como las filas se distribuyen en base a la *partition key* es conveniente escoger una clave primaria adecuada para la aplicación que se trate.

   Regla 2: Minimizar el número de particiones a leer:

   - Las particiones son grupos de filas que comparten la misma partition key.
   - Cuantas menos particiones tengan que ser leídas, más rápida será la lectura.

3. La Base de Datos Cassandra debe permitir realizar las siguientes consultas más frecuentes solicitadas por el negocio, **utilizando CQL**, sobre la base de datos diseñada:
   a. Devolver todos los postulantes matriculados en la carrera de medicina ordenados por periodo.
   b. Devolver todos los postulantes matriculados provenientes de la región del Maule en la carrera Ingeniería Civil Informática ordenados por periodo.
   c. Devolver todos los postulantes matriculados en la facultad de Ciencias de la Salud ordenado por puntaje PSU.

4. Una vez implementadas las consultas más frecuentes solicitadas por el negocio **utilizando CQL** extraiga el resultado cada una de ellas utilizando una conexión desde Excel, de tal forma que los datos se puedan visualizar en planillas y Ud. pueda generar alguna tabla dinámica o gráficos que permitan presentar los datos de una manera resumida (libre).

5. Una vez realizados los puntos anteriores, se pide mostrar evidencia objetiva que permita demostrar la **consistencia** de los datos en los 3 nodos (Método *Three*) y también demostrar la alta **disponibilidad**, por ejemplo, bajando el nodo local para que responda alguno de los nodos replicados sin que el usuario se percate del problema.