# **GEODB**

## RETO 4.

En primer lugar, tendremos en cuenta una de las premisas que aparecen expuestas en párrafos anteriores:

*“2TICKET no guarda ninguna información sobre el precio de las entradas, aforo de los recintos, dni del comprador, etc.”*

Por tanto, no guardamos ninguna información relacionada con un usuario particular. Si esto no fuera así, el escenario cambiaría de una manera muy importante. Pero nos atenemos a esta directiva de nuestra compañía, 2TICKET.

Teniendo en cuenta la volumetría de eventos y la velocidad con la que el sistema recibe interacciones (nuevos ententos, nuevos tickets, nuevas peticiones de usuarios…) una arquitectura Big Data encajará bastante bien. A nivel de definición de esta arquitectura, no hay una opción posible (¡afortunadamente!). Nosotros podemos optar por una configuración “clásica” basada en arquitectura de procesos Spark, Hive, herramientas habituales del stack Spark (Yarn…) y apoyándonos en la nube (optamos por AWS porque la conocemos más y porque aunque no sea la más barata es muy robusta y confiable; dispondremos de servicios típicos como S3, EC2, EMR, Data Pipelines, Lambdas, Redshift, herramientas de auto-scaling…)

Debido a la alta frecuencia y continuidad de nuestros inputs, optaríamos por frameworks de Streaming para Big Data (Apache Streaming, Flink…)

Respecto a las mejoras de la aplicación:

No disponemos de información relativa a los usuarios particulares. Pero sí disponemos de información relativa a ciertas “dimensiones” sobre las que podemos trabajar: dimensión “Tiempo”, dimensión “Geografía”, dimensión “Tipos/descripciones de eventos”, y poder usar sobre ellas métricas de Nº eventos, de Nº de entradas vendidas…

En cada una de estas dimensiones, y gracias a nuestros millones de datos para cada una de ellas, podremos realizar estudios, informes, dashboards, mapas de calor…y analizar tendencias, pudiendo de este modo obtener información útil tanto para nuestra compañía como para el usuario final.

Un ejemplo de esto serían los mapas de calor:

* Podemos elaborar para nuestra aplicación visualizaciones que nos indiquen “sitios calientes” a nivel geográfico, en los que los eventos sean multitudinarios, rompedores a nivel asistencia. O, para los que no gustan de las multitudes excesivas, localizar “sitios fríos” donde poder disfrutar de eventos en un ambiente tranquilo y sosegado.

También, a nivel geográfico se pueden desarrollar mapas según la densidad de eventos (no es lo mismo Nueva York que Tomelloso). Y poder cruzar todas las posibilidades (multidimensión).

* Al igual que a nivel geográfico, podemos elaborar mapas de calor para la dimensión temporal: “fechas calientes” (p.ej. Navidades) y “fechas frías” (p.ej. en estos momentos de confinamiento mundial el mapa mostraría un color gélido, helado, nada de actividad).

Se puede desarrollar también a nivel temporal para el nº de asistentes (nº entradas vendidas), y poder combinar las posibilidades al igual que antes.

* Lo mismo podemos hacer con la dimensión “Tipología de eventos”.

Otro ejemplo sería obtener las tendencias, aplicadas a cada dimensión: poder analizar la evolución temporal de todos estos datos, apoyándonos en estas dimensiones que hemos mencionado antes. De este modo podríamos analizar la evolución que han tenido/están teniendo ciertas localizaciones, ciertas tipologías de eventos…

Este punto entroncaría con el siguiente (los modelos predictivos).

## RETO 5.

En este epígrafe, queremos realizar modelos que permitan realizar predicciones de compras de tickets. Esto lo podríamos realizar basándonos en los datos anteriores y en los escenarios mencionados antes (seguimos suponiendo que NO almacenamos datos concretos de usuarios particulares).

En el punto anterior llegamos a realizar análisis de tendencias (datos pasados y presentes). Ahora deberemos construir un modelo que nos permita “trabajar” con datos futuros. Para ello, desarrollaremos un modelo predictivo, que usará los datos a disposición de nuestra compañía, y con la posibilidad de añadir datos procedentes de partners:

* Deberemos determinar qué tipo de modelo y de algoritmos utilizamos: intuimos que deberíamos apostar por Aprendizaje Supervisado (¿o Reforzado?). Y después habría que determinar si apostar por redes neuronales, regresiones…
* A nivel de datos, dispondremos de nuestra pequeña arquitectura Big Data para ingestar nuestros modelos e ir afinándolos.
* Partners: si son de nuestro sector podremos integrar sus datos de una manera muy directa; si son de otros sectores (restauración, hostelería…) podremos integrarlos creando nuevas dimensiones y enriqueciendo nuestro modelo de datos, y con ello enriqueciendo nuestro modelo predictivo.
* Existe también la posibilidad de aumentar el ámbito de nuestro modelo de datos (y por tanto de nuestro modelo predictivo) incorporando información proveniente de otras fuentes y que nos de nuevos puntos de vista y nos permita descubrir posibles causalidades en el acto de compra de tickets. Podríamos considerar, por tanto, incorporar:
  + Información climática (¿afecta el clima a la venta de tickets? Por supuesto que sí)
  + Información de Empleo (datos de la EPA, etc). Si uno es desempleado, tiene menos fuentes de ingresos y menos posibilidades de comprar entradas para ocio.
  + Información de rentas (u otro tipo de indicador) por ayuntamientos, barrios (si es posible).
  + Información de pirámide poblacional por edad, a nivel de territorio también (ayuntamientos, provincias…).
  + Etc

Todo ello nos aportaría nuevas dimensiones en nuestro modelo, y podríamos cruzar los datos “básicos” de 2Ticket con estos otros, para buscar relaciones causa-efecto a nivel edad, renta, empleo, clima… Todo ello haría que nuestro modelo predictivo fuera más …complejo, lo primero, pero también más potente, si conseguimos afinarlo lo suficiente.