Arquitectura para Bodegas López S.L

1.- Objetivo

En este proyecto se quiere diseñar la arquitectura para centralizar los datos de la bodega mediante un Data Lake. Para ello, la bodega cuenta con datos recogidos de diversas fuentes que son: ERP, CRM, Sensores, Campañas de marketing, e información de las redes sociales, que deben ser almacenados y explotados al máximo. Además, la bodega pide el uso de la aplicación meteorológica del AEMET, que produce los datos a tiempo real de las temperaturas, lluvias, entre otros.

Esta arquitectura a diseñar debe de satisfacer una serie de requisitos que son:

- Creación del Data Lake con los datos centralizados
- Diseñar un Data Warehouse
- Aplicación de modelos matemáticos para hacer predicciones
- Creación de visualizaciones y dashboards
- Control de acceso para los usuarios
- Implementación de una API que podrá ser explotada de manera segura por una aplicación de desarrollo

2.- Solución

En cuanto a la solución propuesta, se ha tenido en cuenta los cinco pilares de AWS, seguridad, confiabilidad, rendimiento, optimización de los costes y excelencia operacional.

Un aspecto importante para la elección de la región donde se va a trabajar es que la empresa Bodegas López S.L sigue el Reglamento General de Protección de Datos (RGPD), por lo tanto, la región a elegir además de tener todos los servicios que se van a utilizar debe cumplir esta regulación. La elección final teniendo en cuenta la localización, los servicios necesarios y las regulaciones, ha sido la región *eu-west-1*, correspondiente a Irlanda (Europa).

Inicialmente, se conocen las fuentes de datos de las cuales se saca la información:

1. Enterpraise Resource Planning (ERP): Es una base de datos relacional en la cual la empresa almacena los pedidos, las facturas, los gastos, y los clientes. Esta fuente, está almacenada en Cloud, concretamente, en *Amazon RDS*.
2. Customer Relationship Management (CRM): Base de datos relacional que almacena información tanto de los clientes como la interacción con ellos. Está almacenado en *Amazon RDS*.

3. Campañas de marketing: La información que se puede sacar de las campañas de marketing como el e-mail marketing, es una base de datos no relacional que está almacenada en *Dynamo DB*.
4. Sensores de la bodega: En el último año la empresa hizo una gran inversión en sensorizar la bodega para conseguir más información de cada uno de los procesos como la maceración del azúcar. Estos sensores instalados captan información de la temperatura, humedad, presión atmosférica, acidez de la tierra. Los datos se van a proteger y procesar mediante *AWS IOT*. Esto también hace posible una actuación sobre los mismos y habilita a la futura API a interactuar con los dispositivos.
5. Agencia Estatal de Meteorología (AEMET): Mediante la API que ofrece, se utilizan datos ambientales como la temperatura o las lluvias.
6. Redes sociales: Mediante la API de las distintas redes sociales como Instagram o Twitter, se extraen datos del perfil de la empresa.

Una vez estudiadas las fuentes de datos y donde están situadas, las vamos a centralizar en un Data Lake con el servicio de almacenaje *S3* de AWS. Dado lo cual, algunas de ellas van a sufrir cambios. Haciendo uso del servicio *Data Pipeline* se hacen los procesos de transformación ETL y movimiento de los datos del ERP, que está en *Amazon RDS*, CRM que está en *Amazon RDS* y Marketing, que está en *Dynamo DB*, a *S3*. En relación con las demás fuentes de datos, los sensores mediante la plataforma *AWS IOT* conectan la información a *S3.* Sin embargo, las APIs de la Agencia Estatal de Meteorología y de las Redes sociales, mediante un disparador o reloj y una función *Lambda* recogen la información automáticamente y la vuelcan en *S3.* Con toda la información centralizada el Data Lake en *S3*, para conseguir una optimización mayor de costos se utiliza el servicio *Glacier S3* que tiene un coste de almacenaje menor que *S3* para almacenar los históricos o datos que no se consultan pero que hay que guardar. Además de este almacenaje, se almacenan las tablas que se van a explotar en *Dynamo DB* reduciendo así los costes en producción de la arquitectura.

Después de crear el Data Lake, utilizando el servicio *Amazon RedShift* se crea un datawarehouse que contiene información de distintas fuentes de datos, quedando las siguientes dimensiones: Pedidos, Clientes, Gastos, Facturas, Campañas y Redes Sociales

En paralelo, partiendo de la información de S3, se pretende la aplicación de un modelo matemático a un determinado datasheet para el pronóstico de ventas. Con el fin de conseguir dicho datasheet, se utiliza el servicio *AWS Glue,* que va a tratar la información, transformarla para dejar únicamente la cosecha, temperatura media por cosecha, acidez media por cosecha, presión media por cosecha, humedad media por cosecha y ventas que es la variable a predecir. Tras esto, mediante *Amazon Sagemaker* se utilizan modelos matemáticos para la predicción.

Por otro lado, en la visualización el servicio que se utiliza es *Amazon Quicksight* con el cual se crean los distintos dashboards que van a permitir una rápida visualización de los datos. Este servicio se nutre de las nuevas tablas que están en *Dynamo DB*, que son las que se quieren explotar.

Otro de los requisitos iniciales de Bodegas López S.L cuando nos contrató para hacer el proyecto es la implementación de una API que será explotada por una aplicación de desarrollo, lo que se implementa mediante el servicio *Amazon API Gateway*, que va a permitir a los desarrolladores la creación, mantenimiento y monitoreo entre otras cosas de la API a cualquier escala. Con el fin de tener un control total con la arquitectura implementada, también se utiliza el servicio *Cloudwatch* para responder eficazmente a cambios de rendimiento, posibles datos erróneos, poner alarmas y utilizar como disparador de acciones automáticas como en el caso de la lectura de datos de las APIs.

Finalmente, el control de acceso a los usuarios se realiza mediante el servicio *Cognito*, que proporciona la autenticación y autorización para la administración de los usuarios mediante inicios de sesión. Las identificaciones *IAM* con permisos específicos se incluyen por si en algún caso se requieren permisos específicos para algún usuario.

En la siguiente página se encuentra el diseño con los servicios utilizados y las conexiones entre los mismos.

Diseño de la arquitectura