
This is the **published version** of the bachelor thesis:

Deu Marín, Arnau; Franco Puntos, Daniel, dir. Analytics as a Service sobre la plataforma d'Amazon Web Services. 2023. (Enginyeria de Dades)

This version is available at <https://ddd.uab.cat/record/281542>

under the terms of the  license

Analytics as a Service sobre la plataforma d'Amazon Web Services

Arnau Deu Marín

7 de juliol del 2023

Resum— El tractament de les dades és una activitat que està en augment dia rere dia, i amb mot més protagonisme dins del món de l'empresa. El tractament d'aquestes ofereix opcions d'optimització del teu negoci, noves oportunitats comercials i comparació contra la competència d'entre moltes. Aquest projecte ofereix la capacitat d'obtenir aquest tractament a empreses que no disposen d'una gran massa crítica, substituint la necessitat de l'adquisició del software i hardware adient, i la pròpia captació de personal qualificat.

El projecte es nodreix dels serveis oferts per Amazon Web Service per, no tant sols un servei que compleix el que s'ha comentat anteriorment, sinó també oferir un servei amb el mínim cost possible. A més a més, aquest treball conta amb un cas d'us real, on una empresa a facilitat les seves dades.

Paraules clau—Amazon Web Service, Cloud, ETL, software com a servei, sense servidor, tractament de dades, AWS Glue, data-lake sense servidor

Abstract—The treatment of data is an activity that is increasing day by day, and with more prominence in the business world. The processing of this data offers optimization options for your business, new commercial opportunities, and comparison against the competition, among many others. This project provides the capability to obtain this treatment for companies that do not have a large critical mass, replacing the need for acquiring appropriate software and hardware, as well as hiring qualified personnel.

The project leverages the services offered by Amazon Web Services to not only fulfill the aforementioned requirements but also provide a service at the minimum possible cost. Furthermore, this work includes a real use case where a company has provided its data.

Index Terms—Amazon Web Service, Cloud, ETL, software as a service, serverless, data treatment, AWS Glue, serverless datalake

1 INTRODUCCIÓ - CONTEXT DEL TREBALL

A VUI en dia el control de les dades i el seu tractament ha agafat un protagonisme sorprenent i segueix creixen a mesura que el temps passa.

Gran part dels negocis, a causa del gran desconeixement sobre les eines que existeixen actualment i els elevats preus que algunes d'aquestes imposen, mantenen un registre amb eines senzilles de les dades que generen. Són pocs els negocis que s'atreveixen a experimentar amb eines d'intel·ligència artificial i processos de tractament de dades més complexos.

Per aquest motiu, en aquest projecte es proposa el disseny d'un sistema replicable a tot negoci, capaç de suportar càrregues de dades periòdiques i tractar aquesta in-

formació, per ajudar a cada un d'aquests a evolucionar i ser més competitius dins del mercat.

En aquesta situació es troba una empresa que centre el seu negoci en la venda de piscines d'hidromassatge. La corporació en qüestió, veu vital l'anàlisi de les seves dades per millorar el seu negoci i lluitar contra la competència. Els alts preus dels softwares més desenvolupats i dels recursos humans necessaris els priven de fer tractaments de dades més complexos.

Al llarg del projecte, es farà ús d'aquest cas per posar a prova el funcionament i la viabilitat del sistema.

En aquest document es pot trobar una proposta desenvolupar aquest sistema, l'arquitectura dissenyada, i els resultats d'aquesta implementació.

-
- E-mail de contacte: arnau10marin@gmail.com
 - Treball tutoritzat per: Daniel Franco Puentes (Àrea d'Arquitectura i Tecnologia de computadors)
 - Curs 2022/23

2 OBJECTIUS

Havent analitzat la situació plantejada, és vital per a l'èxit del projecte que el sistema plantejar respongui a certes característiques:

- Sistema *serverless* i de baix cost.
- Sistema amb alta capacitat d'escalabilitat.
- Sistema amb processos automatitzats
- Sistema fàcilment replicable i amb viabilitat de ser creat a través de línies de codi.

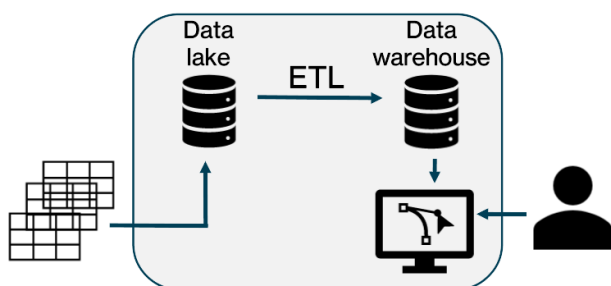


Fig. 1: Proposta de solució

Per assolir tots els objectius esmentats es planteja dissenyar un sistema com es presenta en la figura 1. Concretament, es vol que el sistema consti de 3 grans parts:

- Un sistema d'emmagatzematge òptim per abaratir el cost de les consultes d'informació.
- Una ETL robusta per requerir del mínim manteniment possible.
- Una interfície capaç de facilitar l'anàlisi de les dades per poder visualitzar la informació pertinent.

El sistema dependrà del format de les dades de la corporació, les necessitats i preocupacions que vulguin conèixer i resoldre, i la visualització de les dades.

3 CAS D'ÚS

Per demostrar el correcte funcionament del sistema a desenvolupar s'ha aconseguit plantejar un cas d'ús real, utilitzant les necessitats i dades d'una empresa. Aquesta, és una companyia dedicada al disseny, fabricació i comercialització de productes i solucions per a les piscines.

La venda de spas, el seu producte principal, es produeix a través del contacte amb un comercial, i aquest contacte es pot aconseguir a través de la web i de les diferents botigues que tenen.

Actualment disposen d'una base de dades dins d'un ser-

vidor remot, però no tenen cap persona encarregada al tractament, processament i anàlisi d'aquestes. Utilitzen eines de Microsoft Excel, com Power Pivot per realitzar la visualització de la informació.

S'ha pactat amb l'empresa el consentiment sobre les seves dades per realitzar aquesta prova a canvi d'intentar respondre les següents preguntes que ens han plantejat:

- Quines franges de preu no es cobreixen?
- Quina relació presenten les visites rebudes online amb les ventes realitzades?

Per introduir les dades de la seva base de dades dins del sistema, l'empresa facilitarà arxius *csv* representant cada un una taula de la base de dades.

4 ESTAT DE L'ART

Avui en dia, ja son diverses les empreses que ofereixen serveis informàtics al núvol, que permeten desenvolupar un sistema capaç de complir els requisits anteriors. Aquests proveïdors de serveis, permeten muntar aquests sistemes només havent de pagar per allò que utilitzes i sense preocupacions del manteniment del material utilitzat.

Cada vegada més negocis opten per utilitzar els serveis d'aquests proveïdors, degut al seu baix cost i la facilitat d'utilitzar aquests serveis [1]. Les grans avantatges d'aquests serveis no només es basen en aquestes idees, sinó que també ofereixen una gran facilitat per l'escalabilitat els seus sistemes, igual que la capacitat de fer-ne ús sense alts coneixements tècnics [9].

Actualment, les dos grans empreses que ofereixen més serveis i que estan barallant-se pel mercat son Azure i AWS.

Alguns dels serveis més utilitzats d'AWS son:

- Amazon S3: És un servei d'emmagatzematge d'objectes en línia que permet als usuaris emmagatzemar i recuperar dades. Altament escalable i amb capacitat de replicació de dades.
- Amazon SQS: És un servei de cues que permet la comunicació entre diferents serveis d'AWS.
- Amazon Lambda: És un servei de computació sense servidor que permet als usuaris executar codi sense la necessitat de gestionar i escalar la infraestructura de servidor.
- Amazon CloudFormation: És una servei que es basa en el control de la infraestructura sobre codi. Aquesta permet el aprovisionament dels recursos d'AWS a través d'arxius de codi.

5 METODOLOGIA I PLANIFICACIÓ

La metodologia que es vol dur a terme durant aquest projecte, s’adapta al contacte i disponibilitat de l’empresa col·laboradora. Per aquest motiu es divideix la implemen-tació en diferents blocs que poden ser desenvolupats simultàniament, per tal de no encallar-se en una única tasca.

La metodologia Kanban [2], basada en fases de treball, encaixa perfectament amb la implementació descrita, ja que permet marcar diferents blocs de treball amb múlti-ples feines. A la taula 1 s’exhibeix l’estructura plantejada.

TAULA 1: PLANIFICACIÓ DEL PROJECTE EN FASES

Fase 1	- Proposta del projecte - Investigació d’eines per desenvolupar el sistema i proposta - Primer contacte amb empresa col·laboradora i coneixement de les seves necessitats.
Fase 2	- Definició d’arquitectura bàsica - Obtenció de les dades necessàries per respondre les necessitats de l’empresa
Fase 3	- Definició específica del <i>workflow</i> de les dades. - Desplegament del servei d’emmagatzematge - Desplegament dels serveis per realit-zar la ETL
Fase 4	- Desplegament del servei de visualit-zació de les dades. - Carrega de dades dins del sistema.
Fase 5 (iterativa)	- Mostar de resultats a l’empresa col·laboradora. - Millora dels serveis per realitzar la ETL.

Per el control de la planificació s’ha fet ús de l’eina Mi-crosoft OneNote.

Aquesta eina ha permès crear taulells personalitzats, concretament descrivint una metodologia Kanban, per fer el seguiment de les tasques. A més, aquesta eina permet la creació de pàgines buides per emplenar-les amb esbor-rans o conceptes importants a tenir en compte del pro-jecte.

El disseny de la planificació s’ha basat en conèixer aquells treballs que suposen una coll d’ampolla, fent que aquests estiguin programes dins de la mateixa fase que treballs que no són influenciats per aquestes tasques.

Per dur a terme el desenvolupament del projecte s’ha

utilitzat Google Meets i Microsoft Teams.

6 PROPOSTA

Es proposa que el sistema descrit sigui desenvolupat sobre les diferents eines dins d’AWS [8], ja que a part de la gran quantitat de serveis ofertes, també faciliten la integració entre elles.

Es muntarà i definiran els diferents mòduls del sis-tema utilitzant:

- AWS S3 com a principal sistema d’emmagatzematge dels arxius i de les bases de dades.
- AWS Glue com a eina per realitzar la definició de les bases de dades, la transformació, i càrrega de dades. Concretament utilitzant els serveis de *Crawler*, *Trigger* i *Job*.
- AWS QuickSight com a eina per realitzar la visu-alització dels resultats.
- AWS SQS com a eina per activar nous treballs a través d’esdeveniments del propi sistema.
- AWS CloudFormation com a eina per definir tots els serveis anteriors i permetre la seva implemen-tació ràpidament i amb cost marginal baix.

Es vol que els sistema consti de diferents blocs de ser-veis:

- Bloc de càrrega de dades.
- Bloc de transformació de dades.
- Bloc d’anàlisi de dades.

Tenint en compte el volum de dades i els serveix que s’utilitzaran, es fa una predicció del cost del sistema uti-litzant el servei de càlcul de costos d’AWS.

TAULA 2: COSTOS SERVEIS D’AMAZON WEB SERVICES

AWS S3	0,46 €/mes 20 GB	1,05 €/mes 50 GB
AWS Glue Crawler	0.02 €/csv rastrejat	
AWS Glue Job	0,75 €/mes 10’	1,50 €/mes 20’
AWS QuickSight	~ 30 €/mes 1 a 6 lectors	

En la taula 2 es presenten els costos d’ús dels serveis per mes [10]. Mirant els imports es pot veure com els serveis d’emmagatzematge i processament de les dades per una empresa que no gestioni grans quantitats d’informació, suposen un cost inferior als dos euros.

7 ARQUITECTURA

Seguint la proposta plantejada, es proposa l'arquitectura de la figura 2. En aquesta imatge es mostren 4 mòduls específics.

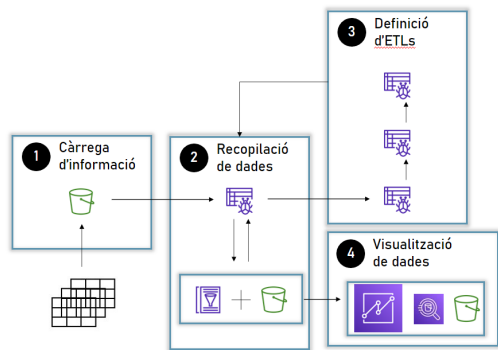


Fig. 2: Proposta de solució

Mòdul de càrrega d'informació: En aquesta fase, el sistema ha de ser capaç de carregar informació en format *csv*, i identifica els arxius carregats per donar la ordre d'executar la següent fase automàticament.

Mòdul de recopilació de dades: En aquesta fase, el sistema ha de ser capaç de detectar l'estructura dels diferents documents i emmagatzemar-la, creant una base de dades.

Mòdul de definició d'ETLs: Aquesta fase s'encarrega de realitzar els diferents processos de transformació, i creació de noves dades sobre la informació carregada al sistema en les fases anteriors. A més, la informació ja tractada ha d'emmagatzemar-se dins d'una nova base de dades.

Mòdul de visualització de dades: L'últim mòdul del sistema ha de permetre l'accés als diferents usuaris i ha de mostrar les dades carregades dins de la nova base de dades on està la informació ja tractada.

Cal tenir en compte, que com es tracte d'un sistema automatitzat, els mòduls han de tenir activadors que siguin capaçs d'activar els treballs dels següents mòduls.

8 CONFIGURACIÓ

Seguint els objectius i l'arquitectura dissenyada s'ha desenvolupat el projecte, desglossant-lo en una etapa per cada mòdul.

S'ha volgut incloure una etapa extra per la integració de tots els mòduls sobre CloudFormation.

D'aquesta manera s'aconsegueix una aproximació individualitzada al problema.

- Mòdul de càrrega d'informació.

- Mòdul de recopilació de dades
- Mòdul de definició d'ETLs
- Mòdul de visualització de dades.



Fig. 3: Llegenda serveis AWS

En la figura 3 es mostra la llegenda dels serveis d'AWS que es troben representats en la figura 2 i les que es troben més endavant [6].

8.1 Mòdul de càrrega d'informació

Per el desenvolupament d'aquest primer mòdul s'havia de buscar un servei capaç de suportar càrregues d'arxius i activar altres serveis quan algun d'aquests arxius fos carregat.

S'ha seleccionat el servei de contenidors de S3, on classificat per carpetes, es carreguen els diferents documents *csv* de les diferents taules de la base de dades de la corporació.

L'ús dels contenidors de S3 [5] aporta una sèrie de beneficis sobre el problema plantejat que els altres serveis no proporcionen.

S3 es un sistema sense servidor, on només pagues per les quantitat de dades emmagatzemades. A més, és fàcilment escalable i configurable per activar altres serveis quan qualsevol arxiu és carregat [7].

Altres serveis descartats per realitzar aquesta funció són els serveis d'Amazon Redshift i RDS. Encara que estan més orientats al tractament de dades i també son escalables, pagués per hores i la quantitat de dades emmagatzemades, fent que un sistema que només requereix de consultes periòdiques i no constants, el preu sigui més elevat.

8.2 Mòdul de recopilació de dades

Aquest segon mòdul havia de tenir la capacitat de extraure l'estructura de les diferents taules de la base de dades

carregada anteriorment per definir el seu catàleg de dades, igual que ser un servei que fos activable mitjançant notifikacions d'altres serveis. El coneixement d'aquesta estructura és vital per poder definir la base de dades sobre els contenidors de S3.

El servei seleccionat és el Crawler de Glue, que te la capacitat de determinar el Glue Catalog i així poder utilitzar el contenidor S3 com una base de dades [3].

Principalment es fa ús d'aquest servei perquè permet una fàcil integració amb S3, té un baix cost marginal, i permet ser activat sense la necessitat d'interacció humana. A part dels beneficis anteriors, aquesta eina també ofereix una alta flexibilitat, poder tractar arxius de diferent format, i actualitzar els diferents canvis sobre l'esquema original les modificacions de dades que apareguin.

Una altra possible solució per definir aquest catàleg, que no es contempla utilitzar, és la implementació manual on l'únic benefici és el cost nul, però on la definició de l'esquema passa a ser manual i es perd tota capacitat d'automatització del mòdul.

Aquest mòdul acaba amb la implementació del servei de Glue Trigger, que és un servei que permet la concatenació dels diferents serveis de Glue, on en aquest cas activarà els treballs del mòdul de definició d'ETLs.

TAULA 3: Comparació Lambda Vs. Glue Job

Característiques	AWS Lambda	AWS Glue Job
Funció principal	Execució de codi	Transformació de dades [11]
Escalabilitat	Automàtica	
Llenguatges permesos	Múltiples	Python y Scala
Temps màxim d'execució	15 minuts	Indefinit
Càlcul del Cost	Per temps i memòria utilitzada	Per temps i capacitat de processament
Integració amb AWS Glue	Pot integrar-se	Dissenyat per integrar-se amb Glue

8.3 Mòdul de definició d'ETLs

El tercer mòdul havia de permetre la lectura, transforma-

ció i càrrega de dades a través d'un activador.

Dos serveis s'aproximen a la solució buscada, en la taula 3 es mostra la comparació entre AWS Lambda i AWS Glue Job.

Finalment s'ha decidit fer ús d'AWS Glue Job, ja que és un sistema orientat al tractament de dades i que no té cap limitació de temps d'execució.

8.4 Mòdul de visualització de dades

La última part del sistema la defineix una eina que sigui capaç de llegir les dades i mostrar-les, per oferir així el seu anàlisis.

Serveis dins de l'entorn d'AWS que compleixin aquesta funció només n'hi ha 1, que s'anomena QuickSight [4]. Aquest funciona com una eina de reportatge de dades en forma de taulells interactius, que una de les seves propietats principals és que permet la ingesta de dades a través de consultes del servei Athena, que permet la lectura d'informació dins de contenidors S3.

Una altre opció viable era la descarrega dels arxius ja processats en el tercer mòdul, i fer ús d'altres eines de visualització, fora de l'entorn d'AWS.

El servei seleccionat ha estat el de QuickSight, ja que s'adapta a la condició de que el sistema estigui el màxim automatitzat possible, permeten fer programacions d'actualitzacions de les consultes dels informes creats.

8.5 CloudFormation i creació de Stacks

Un dels objectius del projecte és crear un sistema que sigui fàcilment replicable. AWS disposa d'un servei que permet implementar els seus serveis a través de la seva definició en línies de codi, fent així que amb poques modificacions es pugui desplegar el mateix servei per un cas de negoci diferent. Aquest s'anomena CloudFormation

La eina en qüestió opera de la mà amb AWS Cloud 9, un entorn de desenvolupament integrat que permet la execució i depuració de codi. Fent que si s'executa un codi de CloudFormation, es desplegui un paquet amb tots els serveis desenvolupats dins de l'arxiu. Aquesta agrupació de serveis s'anomena Stack [12].

Indicar també que el servei de CloudFormation es gratuït, que el cost de la acció recau sobre AWS Cloud 9, ja que opera sobre una instància EC2

També cal indicar que no tots els serveis s'han pogut implementar amb aquesta eina. Concretament, l'activador per cridar al serveis del mòdul de recopilació de dades han hagut d'estar definits manualment, ja que Clou-

dFormation actualment no soporta el trigger de S3 sobre Crawlers.

Per la resta de serveis, tots han estat definits i creats dins del Stack.

8.6 Encaix dels mòduls

L'encaix de tots aquests mòduls i els passos que el sistema prendrà venen definits per activadors automàtics i activacions manuals.

Inicialment el sistema no desenvoluparà cap servei fins que un arxiu *csv* sigui carregat dins del contenidor d'arxius pertinent. Cada cop que un arxiu es carregat o modificat dins d'aquest contenidor, una alerta s'enviarà desde d'aquest fins al Crawler de Glue pertinent.

Un cop l'alerta es processada el Crawler de Glue s'activarà, generant o modificant les taules de dins de la base de dades associada al contenidor. Un cop la acció finalitza es donarà la ordre d'executar un Job de Glue a través d'un activador de Glue.

Aquest Job de Glue aplicarà diverses transformacions sobre les taules de la base de dades ja creada, i les guardarà dins d'arxius de format optimitzat (Parquet) sobre un altre contenidor diferent.

Quan acabi les transformacions, el Job de Glue pot cridar a través d'activadors a altres Jobs, o a un Crawler, diferent de l'anterior, encarregat de crear les taules dels arxius en format parquet dins d'una base de dades diferent.

Finalment serà l'usuari qui hagi d'entrar dins de QuickSight i actualitzar les visualitzacions per poder veure les noves dades carregades dins de les diferents representacions prèviament definides.

Mirant l'estructura des de lluny, es pot comparar el primer contenidor de CSV com un Data-Lake, i el contenidor que conté arxius en format Parquet com un Data-Warehouse.

L'explicació comentada de l'encaix planteja el concepte més simple aplicable sobre el sistema desenvolupat. La quantitat de bases de dades, conjuntament amb els seus contenidors, Crawlers i Jobs pot augmentar a les necessitats de l'empresa.

Poden arribar a no tenir necessitat d'aplicar cap transformació sobre les dades carregades, si aquestes ja són prou vàlides per realitzar les visualitzacions pertinents.

9 RESULTATS

A continuació s'exposa el comportament de tot el sistema implementat. Per mostrar com es comporta el sistema sobre un cas real, es mostrarà:

- Com es comporta cada mòdul dins del cas d'ús plantejat a l'informe
- Un exemple de cas de negoci aplicat sobre el cas d'ús plantejat a l'informe.
- Extraccions d'informació per respondre les preguntes del cas d'ús plantejat

9.1 Mòdul de càrrega d'informació

Adaptant el servei seleccionat per aquest mòdul del sistema, sobre el cas d'ús, s'ha acabat tenint 3 contenidors d'informació:

- Csv-collector: Un contenidor per emmagatzemar tots els arxius *csvs* a carregar.
- Parquet-lake: Un contenidor per emmagatzemar tots els arxius parquet de les transformacions realitzades
- Py-Etl: Un contenidor per emmagatzemar els arxius de Python dels diferents Jobs de Glue.

9.2 Mòdul de recopilació de dades

L'encaix d'aquest segon mòdul sobre el cas plantejat, ha requerit de la implementació de dos bases de dades. Una per definir el Data-Lake de *csvs*, i una altre per definir l'únic Data-Warehouse. Cada una d'aquestes ha estat acompanyada del seu propi Crawler de Glue.

En aquest cas, els dos Crawlers han estat definits per ser capaços d'actualitzar les taules a mesura que les dades carregades varien, igual que la capacitat de generar particions amb aquestes per tal de fer més òptim la extracció d'informació i les seves consultes.

9.3 Mòdul de definició d'ETLs

Per intentar donar resposta a les preguntes realitzades per l'empresa que ha facilitat les seves dades, només ha estat necessari la definició d'un Job de Glue.

Concretament, el Glue Job funciona a través de l'extracció de dades utilitzant la base de dades creada amb Glue, sobre un arxiu escrit amb Pyspark.

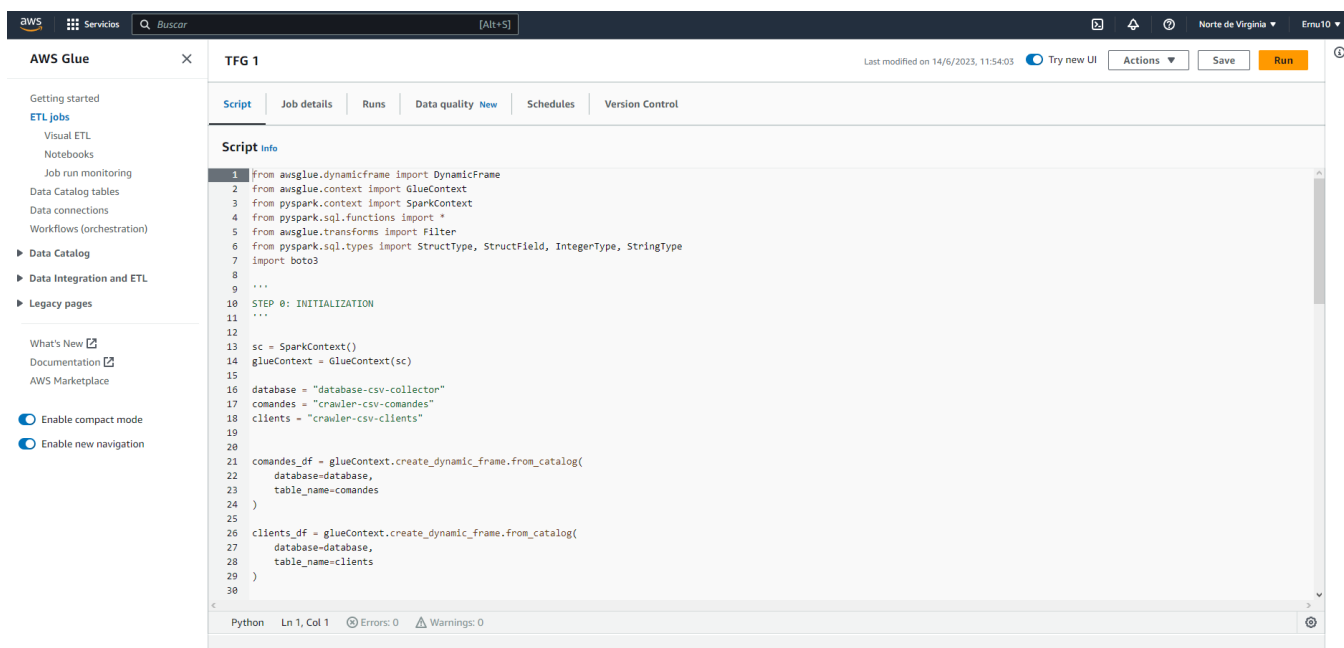


Figura 4: Extracció codi Pyspark Glue Job

9.3 Mòdul de definició d'ETLs

Per intentar donar resposta a les preguntes realitzades per l'empresa que ha facilitat les seves dades, només ha estat necessari la definició d'un Job de Glue.

Concretament, el Glue Job funciona a través de l'extracció de dades utilitzant la base de dades creada amb Glue, sobre un arxiu escrit amb Pyspark.

Inicialment, aquestes son extreptes i guardades dins d'un dynamic frame Seguidament, es filtra per les columnes i files desitjades, i s'aplica un *mapping* per crear una nova columna. Finalment es guarda aquest dynamic frame processat dins del contenidor seleccionat en format parquet. En la figura 4 es mostra part d'aquest codi implementat sobre la eina Job de Glue.

Resumint, es seleccionen les dades de la base de dades seleccionada, s'aplica una neteja i transformació, i finalment es guarden en format optimitzat.

9.4 Mòdul de visualització de dades

L'últim pas consisteix en mostrar les visualitzacions de les dades transformades. Per realitzar aquesta tasca s'ha fet ús dels serveis bàsics d'AWS QuickSight, definint només un lector i un editor sobre l'eina.

Aquest servei és el més car, fent que el cost total augmenti molt si es defineixen més editors i lectors. A part, es fa servir la ingesta de dades a través de SPICE [13], un

sistema més optimitzat i amb l'enviament d'un correu automàtic si la carrega d'aquestes dades falla, ja sigui activada manualment o programada.

9.5 Cas de negoci

Per la creació d'un cas de negoci coherent sobre la solució implementada, s'ha utilitzat els costos mostrats en les diferents execucions ja realitzades i s'han desglossat segons servei i si aquests son costos fixes o variables. Aquests costos s'han extret de l'explorador de costos del que disposa AWS [14].

La taula 4 mostra els costos dels serveis utilitzats de forma aproximada. Per entendre la taula es important conèixer que tots els valors representen dòlars americans. A part, aquests costos son els aconseguits en les diferents execucions realitzades en el moment de definició del projecte. Aquest import pot variar si la regió on els diferents serveis han estat implementats i si la seva configuració varia.

Fent una ullada sobre els valors mostrats dins d'aquesta, es pot veure com els cost total dels serveis es molt assequible. Inclús si es decidís fer servir un altre servei de visualització i reportatge de dades, els costos disminuirien en picat.

En total surt que el cost de mantenir aquest sistema anualment no supera els 330\$. Si s'arribés a fer servir una eina de visualització gratuïta, el cost anual rondaria sobre els 10\$.

TAULA 4: Cas de Negoci

SERVEI	UNITAT MESURA	ÚS MENSUAL	COSTS			
			BASE	INCREMENTAL	MENSUAL	ANUAL
S3	GB	20	0	0.023	0.46	5.52
SQS	-	0	0	0	0	0
Glue Crawler	DPU-Hour	0.2	0	0.44	0.088	1.056
Glue Catalogue	-	0	0	0	0	0
Glue Database	-	0	0	0	0	0
Glue Job	Minuts	3	0	0.09	0.27	3.24
Athena	Terabyte	0.002	0	5	0.01	0.12
Quicksight	-	0	26	0	26	312
TOTAL					26.828	321.936

Els costos consideren una càrrega d’informació i execució del sistema mensual. Per calcular el cost amb més execucions, s’haurà de multiplicar els costos mensuals i finals pel número d’execucions que es volen realitzar.

9.5 Respostes Preguntes cas d’us

Per tal de respondre a les preguntes plantejades per la empresa que ha cedit les dades, s’ha realitzat les visualitzacions pertinents després d’obtenir les dades transformades.

S’ha dedicat un informe sencer de QuickSight per fer l’anàlisi de quins imports de venda no estan coberts. Aquest informe conté una gràfica combinada que mostra els articles per franja de preu en columnes i la quantitat de unitats repartides amb una línia.

Una altre informe ha estat dedicat a mostrar les comparacions entre visites de la pàgina web i ventes realitzades. Per mostrar-la s’ha definit un gràfic de línies mostrant la quantitat de visites i comandes per cada mes de l’any.

Tots els estudis fets només mostren les dades, en cap moment s’extreuen conclusions. L’anàlisi d’aquestes i la conclusió final es deixa a càrrec de l’empresa que forma part dins del cas d’ús.

10 CONCLUSIONS

Només començar a desenvolupar el projecte, ja em vaig anar donant compte de la gran dimensió dels serveis Cloud, i el creixement continu que tenen.

Centrant-nos en els objectius plantejats a l’inici, es pot veure com aquest han estat assolits al llarg del projecte. Actualment, els serveis implementats compleixen que tenen una alta escalabilitat, fent que si un dia haguessin de suportar grans volums de dades, no es necessites cap

canvi. També s’ha aconseguit que tots aquests sistemes siguin sense servidor, fent que el cost de les execucions sigui mínim, sense haver de pagar tot el temps que els serveis no estan executant-se. Gràcies a la gran varietat d’eines d’AWS, també s’ha aconseguit fer un sistema automatitzat, que no requereix d’interacció humana per executar tot el procés.

Com a últim objectiu es tenia que el sistema es pogués replicar. Això s’ha aconseguit gracies a haver definit totes les instancies dins de CloudFormation, fent que si es volgués crear la mateixa arquitectura sobre un altre negoci, només s’haguessin de definir els arxius de Pyspark i les visualitzacions de Quicksight pertinents.

Tot i havent complert els objectius el sistema té una mancança. En cas de que l’usuari carregués un arxiu incorrecte, el sistema reportaria resultats incorrectes i segons com sigui l’arxiu carregat, la transformació de dades a través de Pyspark donaria error. Tot això es pot solucionar a través de la incorporació d’un nou mòdul que s’encarregués de comprovar la naturalesa de l’arxiu carregat abans de l’execució del Glue Crawler.

A part, el mòdul de definició d’ETLs es molt customitzable, poguent suportar la implementació de models d’aprenentatge profund i filtres d’alta complexitat, ja sigui per obtenir indicadors, una previsió de ventes o conjunts de metadades.

11 AGRAÏMENTS

Inicialment, agrair als tutors del projecte, tant al Daniel Franco com a l’Antoni Espinosa per l’ajut en poder portar la idea inicial a una realitat.

En especial menció, agrair a Artur Deu, el meu avi. Una persona que ha estat present des de l’inici del projecte, i tot i que mai no ha entès del que tracta, m’ha ajudat en tot el que he demanat i més.

BIBLIOGRAFIA

- [1] Workshop: Implementando un ETL Serverless en AWS [Consultat: 5 de febrer de 2023]. Disponible a internet: <https://www.youtube.com/watch?v=7RYketNCEGw>
- [2] What Is Kanban In Project Manager? (2021) [Consultat: 20 de març 2023]. Disponible a internet: <https://www.youtube.com/watch?v=xjcvHJZDlNY>
- [3] AWS Glue – Developer Guide [Consultat: 22 de maig de 2023]. Disponible a internet: <https://docs.aws.amazon.com/pdfs/glue/latest/dg/glue-dg.pdf>
- [4] AWS QuickSight – Developer Guide [Consultat 24 de maig de 2023]. Disponible a internet: <https://docs.aws.amazon.com/pdfs/quicksight/latest/develo perguide/amazon-quicksight-dg.pdf>
- [5] Amazon Simple Storage Service – User Guide [Consultat: 26 d'abril de 2023]. Disponible a internet: <https://docs.aws.amazon.com/pdfs/AmazonS3/latest/userguide/s3-userguide.pdf>
- [6] AWS Architecture Icons [Consultat: 20 de març de 2023]. Disponible a internet: <https://aws.amazon.com/es/architecture/icons/>
- [7] Amazon Simple Queue Service – Developer Guide [Consultat: 26 d'abril de 2023]. Disponible a internet: <https://docs.aws.amazon.com/pdfs/AWSSimpleQueueService/latest/SQSDeveloperGuide/sqs-dg.pdf>
- [8] AWS Data Engineer Project | AWS Glue | ETL in AWS [Consultat: 15 de març de 2023]. Disponible a internet: <https://www.youtube.com/watch?v=QPWzrllvnag>
- [9] Build and automate Serverless using AWS Glue, Lambda, Cloudwatch [Consultat: 5 de febrer de 2023]. Disponible a internet: <https://www.youtube.com/watch?v=3f7UY5R9O9U>
- [10] AWS Pricing Calculator [Consultat: 15 de març de 2023]. Disponible a internet: <https://calculator.aws/#/>
- [11] DynamicFrame Class [Consultat: 09 de juny de 2023]. Disponible a internet: <https://docs.aws.amazon.com/glue/latest/dg/aws-glue-api-crawler-pyspark-extensions-dynamic-frame.html>
- [12] Working with stacks [Consultat: 11 de juny de 2023]. Disponible a internet: <https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/stacks.html>
- [13] Importing data into SPICE [Consultat: 11 de juny de 2023]. Disponible a internet: <https://docs.aws.amazon.com/quicksight/latest/user/spice.html>
- [14] Facturas AWS [Consultat: 16 de juny de 2023]. Disponible a internet: <https://us-east-1.console.aws.amazon.com/billing/home#/bills>