Anàlisi de dades de consum en una empresa de touroperació.

Gregorio Cobacho Navarro

Grau d'Enginyeria Informàtica

TFG – Business Intelligence

Humberto Andrés Sanz

17/06/2019

FITXA DEL TREBALL FINAL

Títol del treball:	Anàlisi de dades de consum en una empresa de touroperació.
Nom de l'autor:	Gregorio Cobacho Navarro
Nom del consultor:	Humberto Andrés Sanz
Data de lliurament (mm/aaaa):	06/2019
Àrea del Treball Final:	Business Intelligence
Titulació:	Grau d'Enginyeria Informàtica

Resum del Treball (màxim 250 paraules):

Implementar sistema business intelligence en una empresa dedicada a la touroperació per tal d'aprofitar la gran quantitat d'informació que genera en processar les diferents transaccions que conformen un flux complet de procés de compra d'una reserva d'hotel.

Prèviament al Desenvolupament del projecte, es realitza una introducció a les tecnologies Business Intelligence per tal de presentar al lector aquestes tecnologies i que serveixi com a base per a la comprensió de la resta del projecte.
Un cop assentades les bases, es seleccionen les eines a utilitzar preferiblement cloud ja que l'empresa té el seu sistema implementat en base a aquesta tecnologia, en concret Amazon Web Services (AWS).
Posteriorment es realitzarà una captura de dades de mostra per tal de poder realitzar el processament de les dades capturades.
A la part final del projecte es generaran una sèrie d'informes de mostra orientats a ser una eina de consulta a l'hora de prendre decisions de negoci.

Abstract (in English, 250 words or less):

Implement business intelligence system in a company dedicated to the touroperation in order to take advantage of the large amount of information generated by processing the different transactions that make up a complete flow of the purchase process of a hotel reservation.

Prior to the development of the project, an introduction to Business Intelligence technologies is made in order to present the reader with these technologies and to serve as a basis for understanding the rest of the project.
Once the bases are established, the tools to be used are preferably selected, as the company has its system implemented based on this technology, specifically Amazon Web Services (AWS).
Subsequently, a capture of sample data will be performed in order to be able to process the captured data.
In the final part of the project, a series of sample reports aimed at being a consulting tool when making business decisions will be generated.

Paraules clau (entre 4 i 8):

AWS, Business, Intelligence, QuickSight, touroperador

Índex

Llista de figures

1. Introducció

1.1 Context i justificació del Treball

Molts operadors turístics ofereixen els seus productes en línia mitjançant web services que són consumits per metacercadors o agències online. Aquests sistemes produeixen una quantitat ingent d'informació constant.

Tenint en compte que un sistema Business Intelligence dota les empreses d'un sistema que transforma la informació generada en coneixement del seu propi negoci per tenir avantatges competitius, la informació generada degudament analitzada podria donar informació molt valuosa a l'empresa pel que fa a l'hàbit de consulta i consum real dels clients finals. A més, també pot obtenir informació determinant per al correcte dimensionament dels seus propis sistemes optimitzant els costos.
En aquest sentit, faré servir com a base una empresa fictícia que tingui allotjats els seus sistemes en cloud, (AWS).

Treballar en aquest context ens assegura un major dinamisme a l'hora d'ampliar o reduir la capacitat de procés dels sistemes segons la informació obtinguda. Fins i tot es podrien determinar necessitats de còmput a futur per tal de poder realitzar reserves d'instàncies a un preu competitiu (avantatge competitiu a través del cost de processos).

1.2 Objectius del Treball

En aquest TFG l'objectiu serà:

Analitzar els elements idonis dels oferts per Amazon AWS per implementar un sistema BI capaç de transformar la informació generada pels serveis que gestionen les transaccions d'operació de negoci de l'empresa en informes capaços de respondre a les preguntes que es realitzen des de l'àrea comercial pel que fa a hàbits de consum i l'àrea de sistemes pel que fa a la càrrega de treball d'aquests sistemes.

Anàlisi de transaccions executades al web service per tal d'analitzar les peticions de disponibilitat, bloqueig de reserva i tancament de reserva final. S'intentarà determinar a quines hores es produeixen un major nombre de peticions de disponibilitat i la ràtio de tancaments de reserves. A més, s'intentarà determinar per segments els hàbits de consum dels clients finals.

Concretament el sistema hauria d'oferir informes que responguin a les següents preguntes:

- A quina hora es realitzen més peticions de disponibilitat de serveis de vacances?
- Que zona geogràfica és la més demandada?
- Quina zona geogràfica és la menys demandada?
- Quin tipus de distribució és la més demandada (1 pax, 2 pax, amb nens ...)?
- ¿Quin ha estat el temps mitjà de resposta segons l'hora del dia?
- ¿Quin ha estat el ràtio de conversió entre disponibilitat i bloqueig de reserva?
- ¿Quin ha estat el ràtio de conversió entre bloqueig de reserva i confirmació de reserva?
- ¿Quin ha estat el ràtio d'error en disponibilitat, bloqueig i tancament?

Les preguntes anteriors haurien de poder respondre a nivell general i també a nivell de client específic, per tenir informació comercial sobre quins clients potenciar.

1.3 Enfocament i mètode seguit

S'enfocarà el treball en base a l'anàlisi d'informació d'una empresa "fictícia" encara que la informació que es manejarà serà obtinguda d'un entorn real de negoci per tal de que la implementació final pugui ser extrapolada a qualsevol empresa del mateix tipus.

Es realitzaran dues fases molt diferenciades, una d'estudi i una altra d'implementació:

- Determinar eines a utilitzar del catàleg ofert per la companyia proveïdora de serveis cloud.
- Utilitzar les eines seleccionades per a:
    - Recopilar informació de mostra d'una empresa operador turístic per tal d'utilitzar-la com a base de la implementació.
    - Analitzar la informació prèviament recopilada per tal de definir un model de persistència apte per al seu posterior processament.
    - Creació dels informes necessaris.

La fase més llarga serà la d'implementació ja que no es pretén aprofundir massa en l'anàlisi de les diferents propostes del proveïdor de servei (cloud), sinó en la implementació un cop decidida l'arquitectura o elements idonis. Això és degut al fet que els proveïdors d'aquests sistemes cloud són molt dinàmics i van canviant de forma molt ràpida els productes oferts, en pocs mesos el portfoli de solucions es pot veure incrementat per solucions que cada vegada són més ràpides i barates.

1.4 Planificació del Treball

Per al seguiment de la temporalització de les tasques s'utilitzarà MS Project, a continuació es mostren els diferents tasques amb la seva respectiva previsió de temps d'esforç, a més de les diferents fites (lliuraments) que coincideixen amb la planificació proposada per a aquest semestre per la UOC .

Nombre de tarea	Duración	Comienzo	Fin
TFG	**81 días**	**lun 25/02/19**	**lun 17/06/19**
PAC1	**11 días**	**lun 25/02/19**	**lun 11/03/19**
Contextualitzar i justificar el projecte.	2 días	lun 25/02/19	mar 26/02/19
Definir els objectius	3 días	mié 27/02/19	vie 01/03/19
Definir enfocament i mètode de treball	3 días	lun 04/03/19	mié 06/03/19
Definir productes a obtenir	2 días	jue 07/03/19	vie 08/03/19
Crear document pla de treball.	2 días	sáb 09/03/19	dom 10/03/19
Lliurament Pla de Treball	0 días	lun 11/03/19	lun 11/03/19
PAC2	**30 días**	**mar 12/03/19**	**lun 22/04/19**
Revisió conceptes bàsics arquitectura Business Intelligence	5 días	mar 12/03/19	lun 18/03/19
Selecció eines Business Intelligence	5 días	mar 19/03/19	lun 25/03/19
Captura dades de mostra	6 días	mar 26/03/19	mar 02/04/19
Processament de dades i guardat	6 días	mié 03/04/19	mié 10/04/19
Redacció document PAC2, característiques tècniques, problemes detectats, estat de les dades obtingudes	5 días	jue 11/04/19	mié 17/04/19
Lliurament	0 días	lun 22/04/19	lun 22/04/19

PAC3	**20 días**	**mar 23/04/19**	**lun 20/05/19**
Definir els diversos informes que conformaran l'anàlisi de dades.	3 días	mar 23/04/19	jue 25/04/19
Implementar lectura de dades per a l'obtenció dels informes.	6 días	vie 26/04/19	vie 03/05/19
Creació i refinament informes.	7 días	sáb 04/05/19	lun 13/05/19
Redacción document PC3 resultat anàlisi de les dades a través de diversos informes, característicasc tècnicas, problemas detectados.	5 días	mar 14/05/19	dom 19/05/19
Lliurament	0 días	lun 20/05/19	lun 20/05/19
Lliurament memoria/producte final	**20 días**	**mar 21/05/19**	**lun 17/06/19**
Redacció document memòria TFG (PAC1/PAC2/PAC3)	10 días	mar 21/05/19	dom 02/06/19
Redacció autoinforme de competències transversals	3 días	lun 03/06/19	mié 05/06/19
Lliurament autoinforme competències transversals	0 días	jue 06/06/19	jue 06/06/19
Gravació video presentació	6 días	vie 07/06/19	vie 14/06/19
Lliurament video presentació	0 días	sáb 15/06/19	sáb 15/06/19
Lliurament document memòria TFG	0 días	lun 17/06/19	lun 17/06/19

1.5 Breu sumari de productes obtinguts

- Conjunt de dades en format csv que es poden descarregar si es desitja per temps limitat des del següent enllaç: https://drive.google.com/file/d/10vJysjt4ZjPdjYLCfwZNsaAOOUGcWdMV/view?usp=sharing
- Arxiu manifest per carregar les dades de S3 que es pot descarregar si es desitja per temps limitat des del següent enllaç: https://drive.google.com/file/d/1xEUr_WDCv85dMKrL_hmowaVIwyCDuxLE/view?usp=sharing
- Conjunt d'informes, aquests informes estan embeguts en la pròpia memòria (capítol 5.6.1)
- Documento de memoria.
- Video presentación.

1.6 Breu descripció dels altres capítols de la memòria

Capítol 2. Revisió conceptes bàsic arquitectura Bussines Intelligence.
En aquest capítol es fa un ràpid recorregut de diferents conceptes relacionats amb el Big Data. També s'intenta definir el concepte de Big Data però des d'una perspectiva i enfocament pràctic per al tipus de negoci al qual està destinat aquest projecte.

Capítol 3. Ecosistemes.
En aquest capítol es defineix el concepte d'ecosistema i se centra en ecosistemes Cloud, en concret en l'ecosistema AWS exposant les ventenjas

d'adoptar aquest tipus de solucions tenint en compte que l'empresa en la qual se centra l'estudi ja té els seus serveis en Cloud .

Capítol 4. Selecció eines Business Intelligence.
En aquest capítol es realitza la selecció d'eines de l'Ecosistema AWS que s'utilitzaran en el projecte explicant els motius del seu ús i avantatges.

Capítol 5. Implementació
En aquest capítol es detalla la implementació del sistema BI, explicant detalladament el procés.

S'explica la fase de captura de dades i la extructura d'aquestes dades. Posteriorment s'explica el processament que segueixen aquestes dades i la seva posterior guardat. Finalment es detallen els informes que es defineixen per repondre a una sèrie de qüestions de negoci típiques i s'explica el procés que s'ha seguit per a la consecució d'aquests informes.

2. Revisió conceptes bàsics arquitectura Business Intelligence.

A manera d'introducció, definirem molt breument el que s'entén per big data, quan és necessari i les tecnologies que el conformen.

Hem de tenir en compte que l'objectiu d'aquest projecte no és la divulgació o presentació en profunditat de tecnologies big data. L'audiència a la qual està destinada haurà de tenir uns mínims coneixements en la matèria ja que el projecte té un enfocament pràctic per a un tipus de negoci molt concret.

De totes maneres es realitza un repàs ràpid de multitud de conceptes.

2.1 Què és big data?

La suma d'estratègies, tecnologies i sistemes destinades a l'emmagatzematge, processament, anàlisi i visualització de dades.

Quan és necessari el big data?

El big data és necessari en els següents escenaris o cas d'ús:

1. **Presa de decisions.**

   Les organitzacions amplien la capacitat de decisió tradicional alimentant-se o combinant els sistemes d'intel·ligència i el magatzem de dades corporatives amb els grans repositoris de dades.
2. **Operacions i intel·ligència operativa.**

   Les organitzacions apliquen tècniques de Big data en el camp de les operacions i la intel·ligència operativa per detectar patrons en temps real.

3. **Validació d'hipòtesis i resolució de problemes.**

   Aquest escenari consisteix a buscar solucions per a problemes de negoci que no han estat anteriorment tractats en l'organització.

   En aquest escenari no hi ha preguntes predefinides. S'intenta conèixer què ha passat, quins factors són els més rellevants i el perquè. Es basa en crear hipòtesis que posteriorment es validaran a través de diferents tècniques.

4. **Productes i serveis de dades.**

   En aquest escenario la dada és la peça angular per millorar l'experiència d'ús del producte o servei, així com per al disseny i desplegament d'aquest per tal de generar valor tant per al client com per l'organització.

5. **Comerç de dades.**

   En aquest escenari la dada es prepara per a la venda a tercers i és comercialitzat en brut o en forma de coneixement. Per a això es poden utilitzar diversos processos com agregació, transformació i distribució.

   En cas d'informació confidencial, s'aplicarà un procés d'emmascarament per tal de transformar aquesta informació perquè finalment contingui dades anònims.

2.2 Tecnologías de big data.

- Tecnologies de processament per lots o batch processing:

  Permeten resoldre problemes vinculats amb el volum de la dada.

- Tecnologies de processament en flux o (streaming processing):

  Permeten resoldre problemes vinculats amb la velocitat de les dades.

- NoSQL:

  Permeten resoldre problemes relacionats amb la varietat de dades.

Aquestes tecnologies cobreixen les necessitats envers la dada següents:

1. **Emmagatzematge**.

   Emmagatzemar la dada segons la necessitat de negoci, per a això els sistemes d'emmagatzematge s'han vist obligats a evolucionar per poder ser escalable, compatible amb dades estructurades i sense necessitat de high performance computing (HPC) per a la seva execució.

2. **Processament**.

   Capturar, transformar i moure les dades segons la necessitat de negoci, el processament en Big data es basa en l'extracte, la càrrega, la transformació (ETL).

   En altres paraules, s'intenta guardar les dades en brut i després es processa per complir amb el Data model dissenyat.

   En el context de les dades massives hi ha diferents tipus de processament:

   a) Processament per lots o en mode batch.

      Les dades es processen fora de línia i la seva latència pot ser hores. Per fer-ho, les dades s'emmagatzemen prèviament i es processen. Apache MapReduce i Spark permeten aquest tipus de processament.

   b) Processament en temps real:

      Les dades es processen en mode online i la seva latència és inferior a un minut. Per fer-ho, les dades es processen en memòria en el moment de la seva captura abans d'emmagatzemar-la.

      Hi ha dos tipus:

      Processament en stream on les dades venen en forma ininterrompuda i processament d'interval on les dades venen damunt i fora. Apache Storm, Apache flink i Spark permeten aquest tipus de processament.

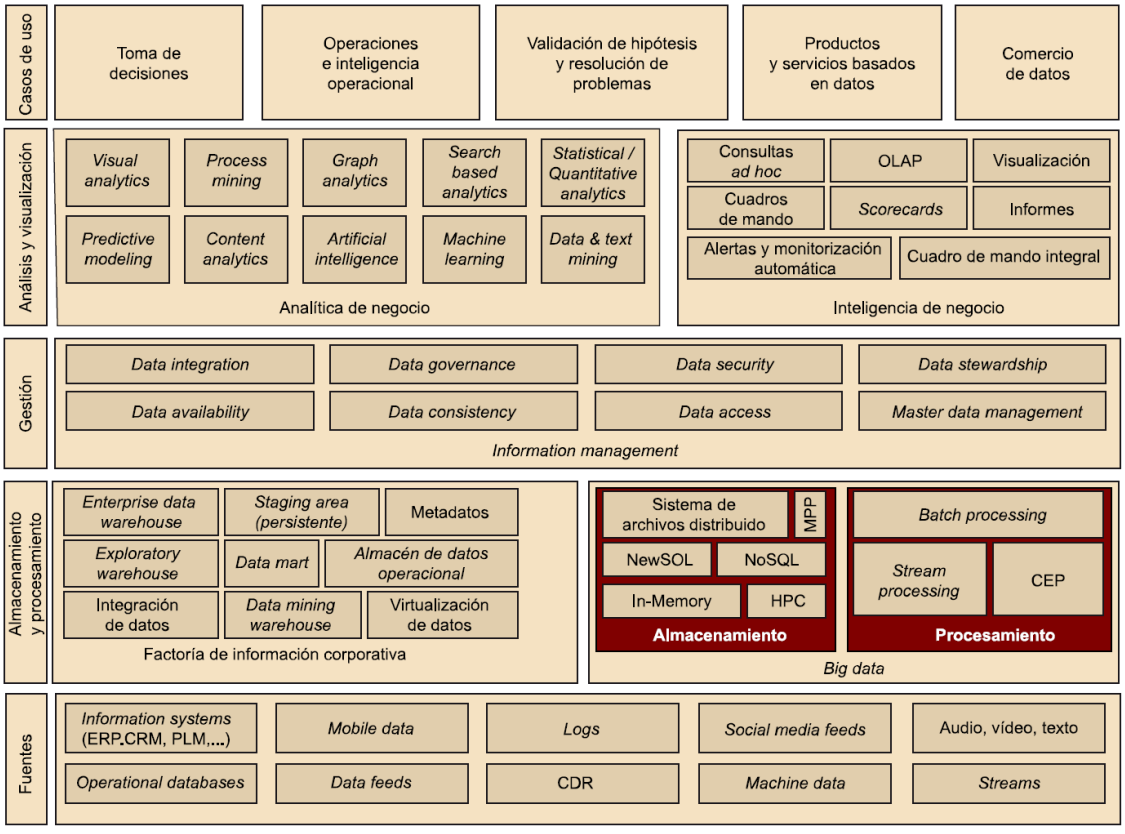
3. **Anàlisi.**

   Generar valor per a l'empresa a partir de les dades, per a això l'anàlisi es concentra principalment en dues àrees: Business Intelligence (rendiment passat) i Business Analytics (futura actuació) centrada en conèixer patrons ocults en les dades.

4. **Visualització**.

   Presentació dels resultats segons la necessitat del negoci, per a això hi ha dues disciplines principals: la visualització de dades (data Visualització) i data storytelling (històries basades en dades).

Arquitectura big data.



Fuente: Josep Curto.

Il·lustració 1. Arquitectura big data.

Les tecnologies Big data, bàsicament, estenen l'arquitectura tradicional de dades augmentant la capacitat de generar valor a partir d'aquestes dades. Aquesta extensió es pot realitzar a través d'una plataforma de negocis de dades massives o a través de la integració o desenvolupament d'una sèrie de components.

Com s'ha descrit anteriorment, hi ha dos tipus de solucions:

- **Solucions Enterprise-ready**: Plataformes estàndards que permeten donar resposta a casos específics.
- **Solucions ad hoc**: Plataformes independents basades en components que cal integrar internament.

Per a la creació d'un producte o servei de dades tenim els següents components:

- **Ingestió de dades (captura de dades)**: Per exemple: Kafka, Logstash, RabbitMq, Fluentd, Chuckwa y AWS Kinesis (Ecosistema Amazon).
- **Format de dades**: Transformació de dades a un format òptim per a l' emmagatzematge en brut, per exemple: Avro, ProtoBuf, Thrift i Parquet.
- **Sistema de fitxers distribuït**: Emmagatzematge de dades en brut, componentes de exemple: HDFS, AWS S3, Microsoft Azure, Alluxio i Ceph.
- **Procesament batch como per exemple**: Spark, Hadoop MapReduce, AWS EMR, Flink i Tez.

  Podem distingir els següents components:
  - MR alt nivell: MapReduce mitjaçant scripting, com per exemple: Pig, Cascading, Hadoop Streaming i Cascalog.
  - ML batch: Machine learning, com per exemple: Mahout, Spark Mlib, FlinkML y H_2O.
  - Graph batch: Anàlisi de grafs, com per exemple: GraphLab, Giraph, Spark GraphX i Hama.

- SQL batch: Ús de SQL, com per exemple: Hive, Presto, Drill, Hue e Impala.

- **Processament en streaming com per exemple**: Storm, Spark Streaming, AWS Lambda, Samza, Akka y Flink.

  Podem distingir els següents components:

  - ML Streaming: Machine learning com por exemple: Spark Mlib y SAMOA.
  - Graph streaming: Anàlisi de grafs com per exemple: X-stream/Chaos.
  - SQL streaming: Ús de SQL com per exemple:Spark SQL i Flink Table API.
  - CEP Streaming: Flink CEP.

- **Emmagatzematge de dades**: Existeixen bases de dades específiques: analítiques, gràfiques, transaccionals, Geospacial, ambients crítics, sèries de temps, cerca i memòria cau.

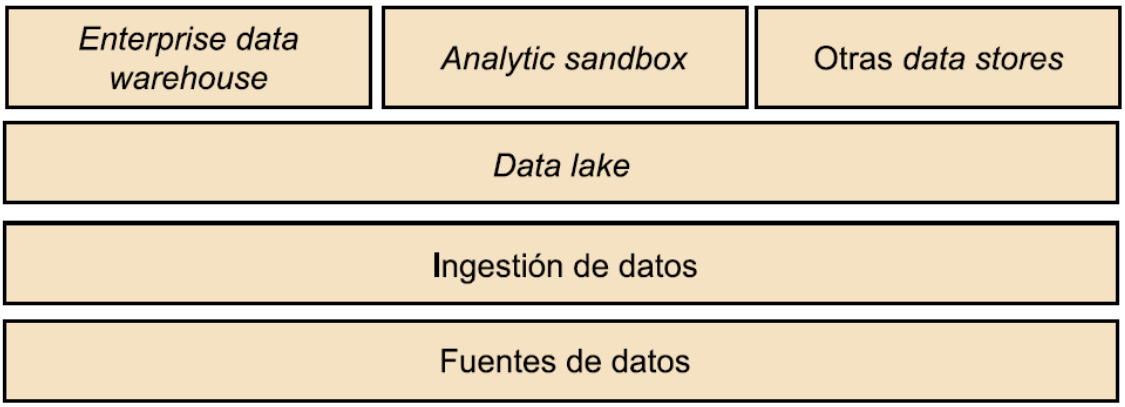
  - Analítica: execució naTiva d'algoritmes i tècniques analítiques com per exemple: AWS Redshift, Teradata i Vertica.
  - Grafs: Guarda informació en forma de graf i habilita la seva anàlisi com per exemple: Neo4j, OrientDB i ArangoDB.
  - Transaccionals: Optimizada per a transaccions que suporten ACID com ara: MySql, Oracle, Microsoft SQLServer i PostgreSQL.
  - Geoespacial: Optimitzades para a dades geolocalitzades com per exemple PostGIS i Elasticsearch.
  - Documents: Optimitzades per documents com per exemple: MongoDB i CouchDB.
  - Entorns crítics: Optimitzades para entorns crítics com per exemple: Cassandra, Riak i AWS DynamoDB.
  - Series temporals: Anàlisis de dades com una sèrie temporal com per exemple: InfluxDB, Casandra i Druid.
  - Cerca: Optimitzada per a la cerca de informació. ElasticSearch, Solr i MongoDB.

    - Caché: Us in-memory com a tècnica para accelerar consultes freqüents com per exemple: Redis, Memcache i Hazelcast.
- **Framework de desenvolupament web**: Desenvolupament d'un entorn web per exemple: Ruby on Rails, Node, js, Django, AngularJS i Flask.
- **Visualització de dades com per exemple**: D3, js, TAbleau, Qliksense, Leaflet, Highcharts i Kibana.
- **Gestió**: Utilitzats en producció com poden ser la gestió de clústers, monitorització / programació i seguretat.
    - Gestió de clústers (i recursos): Gestió eficient de clústers com ara: Docker, Zookeeper, YARN, Mesos, REEF i Helix.
    - Monitorització / programació: Monitorització de diferents components i programació de tasques com ara: Luigi, Airflox, Nagios, Graphite i Azkaban.Seguridad: Seguridad del dato como por ejemplo: Sentry, REcordService y Knox.

Finalment, cal esmentar un nou enfocament de l'arquitectura anomenat data Lake. Data lake és el repositori d'informació d'una organització en el qual s'inclouen dades estructures i no estructurats en una mateixa taula. En aquest enfocament s'estableix un procés sistemàtic d'explotació de la informació.

El data lake no exclou al data warehouse, sinó que el complementa creant una arquitectura més complexa en la qual es poden combinar tots dos enfocaments.

Cal comentar que data lake no suporta escenaris de processament en streaming. Per tant, podríem concloure que data lake és la font de dades de data warehouse, Analytic sandbox altres data stores existents, és a dir un magatzem de dades operacional persistent. A continuació es mostra una arquitectura genèrica de data lake:



Il·lustració 2. Data Lake.

3. Ecosistemes

Podríem definir l'ecosistema de programari com l'espai de treball en el qual s'agrupen un conjunt d'eines i bones pràctiques que permeten a un equip seguir una metodologia de treball.

Per tant, segons la definició anterior podríem concloure que un ecosistema en l'entorn del Big data és la implementació d'una certa arquitectura definida per a tal fi com el processament de bach, streaming o ambdós.

Hi proveïdors de big data en modalitat cloud computing que faciliten la implementació de solucions big data sobretot si el programari de l'organització està desplegat en aquestes infraestructures o plataformes com poden ser IaaS (infraestructura com a servei), Paas (plataforma com a servei) o Saas (programari com a servei), entre les quals destaquen Google, Microsoft o Amazon.

En aquest projecte ens centrarem en l'ecosistema Amazon AWS, la raó principal és que el programari de l'organització que implementarà la solució de BI es desplegarà en aquesta plataforma, això presentarà diversos avantatges:

- **Reducció del temps de desplegament**.
  Els components i les solucions de AWS estan dissenyats per funcionar òptimament en els seus propis serveis, de manera que el programari desplegat en aquesta plataforma serà fàcilment adaptable als serveis específics de AWS per a Big data.

- **Rendiment**.
  Com s'ha comentat anteriorment, les solucions ofertes per AWS estan

optimitzades per al seu ús en els seus serveis, si s'han seguit bones pràctiques en el disseny de la principal arquitectura de programari de l'organització, el rendiment dels grans components Les dades seran òptimes, ja que generen el menor trànsit de xarxa possible i el millor ús dels recursos de processament disponibles en tot moment mitjançant tècniques d'escala automàtica per exemple.

- **Costos**:
  Inicialment es pot pensar que el cost de les solucions cloud és elevat enfront solucions de codi obert, però per arribar a una conclusió clara cal realitzar-se diverses preguntes:

  - **La plataforma de codi obert cobreix totes les necessitats de l'organització?**
    Si la plataforma de codi obert no s'adapta a les necessitats de l'organització i requereix certa adaptació, el cost de l'adaptació pot superar el cost del servei durant anys.
  - **Quin és el cost d'executar la plataforma de codi obert en els seus propis servidors o Hosting extern?**
    Com a regla general el cost dels grans serveis de dades en mode núvol inclou allotjament, ja que s'ofereixen majoritàriament com a serveis (no es requereix cap servidor específic).
  - **Quin és el cost del sistema de manteniment i còpia de seguretat de la plataforma de codi obert?**
    Com a norma general, els contractes SLA d'empreses del núvol asseguren una alta disponibilitat de garanties sense simplement manteniment i gestió de les còpies de seguretat per l'usuari (encara que és aconsellable en tot cas per definir una rutina a aquest efecte).

3.1 Ecosistema Amazon AWS

Com es va comentar en l'apartat anterior el programari de l'organització està allotjat a Amazon AWS per la qual cosa la solució escollida estarà enquadrada en l'ecosistema Amazon AWS per temps d'implementació, cost i simplicitat de manteniment en termes generals. No obstant això, en el propi ecosistema AWS podem trobar múltiples solucions que es poden adaptar en major o menor mesura a la nostra solució.

En particular, l'ecosistema AWS s'estructura en tres grans blocs:

- **Captura de dades**: procés de ETL en format batch i streaming.
- **Emmagatzematge de dades**: Emmagatzematge de dades en tot tipus d' opcions (NoSQL, sistema distribuïts, relacionals).
- **Anàlisis de dades**: Anàlisis de dades incloent Business Intelligence (Quicksight) y machine learning.

4. Selecció eines Business Intelligence.

4.1 Captura de dades.

En l'ecosistema Amazon AWS hi ha diverses solucions com ara: AWS Direct Connect, Kinesis, Snowball.AWS.

Breu descripció:

- **Direct Connect**: conectividad privada entre AWS y entorno local.
- **Kinesis**: recull, processa i analitza dades i vídeos en temps real.

- **Snowball**: Servei de transport de dades a escala de petabyte mitjançant dispositiu físic. Utilitza dispositius dissenyats per a ser segurs i transferir grans volums de dades al núvol AWS.

Per al nostre projecte no utilitzarem cap de les solucions proposades per diversos motius:

- **Direct Connect**: no és necessari perquè els servidors estan allotjats a la mateixa AWS, per tant ja es beneficien de la connectivitat privada entre AWS.
- Kinesis: No és necessari, perquè els informes que es generaran no requereixen un processament en temps real, el processament per lots serà suficient.
- Snowball: No és necessari com Direct Connect, els servidors estan allotjats en el mateix AWS.

A més, la captura de dades es fa pel propi sistema a través de les instàncies (EC2) que s'encarregaran de guardar directament els arxius de log al servei d'emmagatzematge. En aquestes instàncies es troben allotjats els components que porcesan les peticions i la quantitat d'informació que guarda aquests components és configurable a través dels arxius de configuració, d'aquesta manera en moments crítics de càrrega es poden desactivar certs logs per tal d'augmentar la eficiència. El sistema de logs es gestiona amb llibreries log4net per als components escrits en .net i log4j per als components escrits en Java.

4.2 Emmagatzematge de dades.

En l'ecosistema Amazon AWS hi ha diverses solucions com:
S3, EFS, FSx, S3 Glacier, Storage Gateway, AWS Backup, RDS, DynamoDB, Neptune, Amazon Redshift, Amazon DocumentDB.

Breu descripció:

- **S3**: servei d'emmagatzematge d'objectes que ofereix escalabilitat, disponibilitat de dades, seguretat i alt rendiment. Els objectes es magatzem en buckets, cada objecte es compon d'un arxiu i de forma opcional, de qualsevol metadada que descrigui aquest arxiu.
- **S3 Glacier**: Similar a S3 però centrat en arxivar dades i fer reserves a llarg termini i baix cost.
- **RDS**: Base de dades relacional servei disponible per a múltiples motors: Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle Database i SQL Server.
- **DynamoDB**: Base de dades NoSQL claus-valor i documents que ofereixen el rendiment d'un milisegons d'un sol dígit a qualsevol escala.
- **Neptune**: Servei de bases de dades gràfiques d'alta disponibilitat, recuperació en un moment donat, backups continus a Amazon S3, i replicació entre zones de disponibilitat.
- **Redshift**: Servei d'emmagatzematge de dades àgil i escalable que permet analitzar totes les dades del llac de magatzem o dades simplement.
- **DocumentDB**: Base de dades compatible amb MondoDB per emmagatzemar, recuperar i gestionar dades semi-estructurades.

El nostre sistema genera una gran quantitat d'informació a través dels logs de procés,

en aquests fitxers de log es pot trobar la informació sobre les consultes que els clients han fet.

Perquè el nostre sistema genera un Gran Quantitat d'objectes de tipus arxiu, en principi la tecnologia que més ens encaixa per guardar les dades inicialment és S3 perquè treballa nativament amb objectes de tipus file, cal recordar que és justament el tipus d'objectes que anem a guardar, arxius csv.

En concret **farem servir Amazon S3 Estàndard** per les següents raons i característiques:

Característiques principals Amazon S3.

- **Baixa latència i alt nivell de processament.**
  Aquesta característica ens permetrà mantenir un rendiment òptim en el sistema encara que el nivell de transaccions i log sigui extrem.

- **Dissenyat per oferir un nivell de durabilitat de 99,999999999% dels objectes en diverses zones de disponibilitat**.
  aquesta característica ens garanteix que els objectes no es perdran, és a dir ens proporciona seguretat davant perdudes.

- **Dissenyat per oferir una disponibilitat de 99,99% durant un període d'un any.**
  Aquesta característica ens indica que podrem disposar dels nostres objectes en qualsevol moment.

- **Admet SSL per a dades en trànsit i xifrat de dades en repòs.**
  Aquesta característica garanteix la seguretat de les dades en trànsit i un cop allotjats.

- **Administració del cicle de vida de S3 per a la migració automàtica d'objectes a altres classes d'emmagatzematge S3.**
  Aquesta característica garanteix la migració a futurs serveis S3, és molt

important en escenaris en els quals no es té el control total de la versió del programari que s'està utilitzant en aquest cas S3.

A més, el seu preu d'emmagatzematge és bastant baix 0,023 USD per GB per al primer 50 TB/mes a la regió de la UE (Irlanda).

4.3 Anàlisis de dades.

En l'ecosistema AWS hi ha diverses solucions com: Athena, EMR, ElasticSearch Service, Data Pipeline, AWS Glue, MSK, Kinesis, QuickSight.

Breu descripció:

- Athena: Servei de consulta interactiva que facilita l'anàlisi de dades a Amazon S3 amb l'estàndard SQL.
- EMR: Marc Hadoop administrat que permet processar enormes volums de dades en instàncies Amazon EC2.
- Data Pipeline: Processament de dades i transferència a intervals definits entre diferents serveis d'emmagatzematge i informàtica de AWS.
- AWS Glue: Servei d'extracció, processament i càrrega (ETL).
- MSK: Managed Streaming for Kafka.
- QuickSight: Servei BI amb pagament per sessió que inclou panells, informes i anàlisi de dades integrat.

Tenint en compte que el nombre d'usuaris que consultaran els informes generats és baix i que és urgent la implementació del sistema BI, **optarem pel servei QuickSight** ja que s'integra fàcilment amb la font de dades S3 i el seu cost serà baix ja que es paga per sessió.

Característiques principals QuickSight

- **Pagament per sessió.**

  Aquesta característica ens permet estalviar en el nostre escenari ja que com s'ha comentat, seran pocs els usuaris que consultin els informes i amb tota segurament només el realitzen un parell de vegades al dia. A més cal tenir en compte que no s'hauran de realitzar desemborsaments inicials ni existiran compromisos anuals ni càrrecs pels usuaris actius, d'aquesta manera encara que el nombre d'usuaris s'incrementi o hi hagi certa rotació en aquests usuaris a l'empresa, el preu no es veurà incrementat. Només s'incrementarà pel nombre de consultes.

- **Arquitectura sense servidor.**

  Aquesta característica ens permet oblidar-nos de necessitat de planificar la capacitat del sistema, l'administració, manteniment o Backups.

- **Anàlisi de dades d'autoservei integrat**.

  Aquesta característica ens permet en cas que algun client ho sol·liciti, integrar elements visuals de QuickSight (Informes) en el seu sistema o en el seu accés a l'extranet del sistema des API pròpia.

5. Implementació

5.1 Captura dades de mostra.

El nostre sistema ha implementat un sistema de logs que està generant arxius de dos tipus:

1. Arxius que contenen informació de consulta de disponibilitat de serveis i el format del nom és: box_ddmmyyyyyhhmm. csv
   La informació es desa en format CSV i està separada per comes ",".
   Cada fila conté la informació següent:
   - Data i hora.
   - Mode execució.
   - Sistema.
   - Nom de la transacció.
   - Resultat de la transacció.
   - Temps d'execució.
   - Missatge de sol·licitud de transacció.
2. Els arxius que contenen informació de transaccions de bloqueig i tancament de reserva i el format de nom és: boxbookings_ddmmyyyyhhmm.csv
   La informació es desa en format CSV i està separada per comes ",".
   Cada fila conté la informació següent:
   - Data i hora.
   - Mode execució.
   - Sistema.
   - Nombre transacció.
   - Resultat de la transacció.
   - Id de la sesió.
   - Temps d'execució.
   - Missatge de sol·licitud de transacció.
   - Missatge de resposta de transacció.

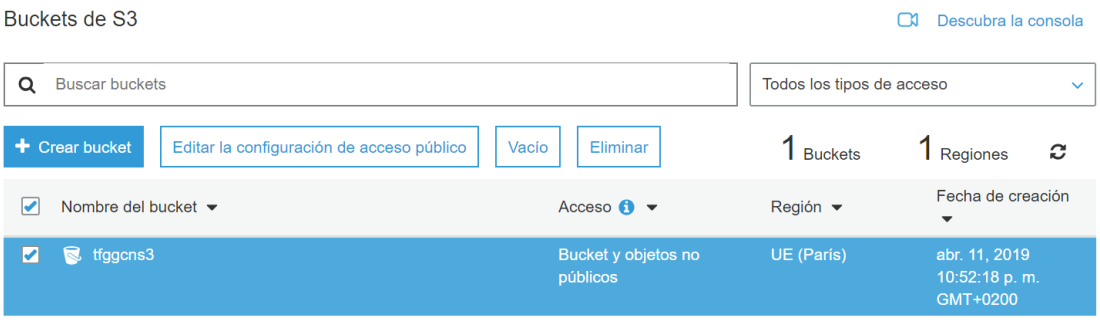
Hem obtingut una mostra tipus d'una setmana de duració de peticions de disponibilitat i tancaments de reserva rebudes a través del nostre sistema xml.

Aquesta mostra s'ha pujat directament a Amazon S3 mitjançant el client proporcionat per Amazon per .NET Amazon.S3.AmazonS3Client

5.2 Processament de dades i guardat.

Les dades estan guardats a Amazon (S3) en un bucket amb nom tfggcns3:

Actualment no han patit cap acció de processat ja que queda pendent de definir el tipus de processament segons els informes que es defineixin properament.



Il·lustració 3. Bucket TFG

5.3 Definició els diversos informes que conformaran l'anàlisi de dades.

El sistema hauria d'oferir informes que responguin a les següents preguntes:

- A quina hora es realitzen més peticions de disponibilitat de serveis de vacances?
- Que zona geogràfica és la més demandada?
- Quina zona geogràfica és la menys demandada?
- Quin tipus de distribució és la més demandada (nombre d'habitacions)?
- ¿Quin ha estat el temps mitjà de resposta segons l'hora del dia?
- ¿Quin ha estat el ràtio de conversió entre disponibilitat i bloqueig de reserva?
- ¿Quin ha estat el ràtio de conversió entre bloqueig de reserva i confirmació de reserva?
- ¿Quin ha estat el ràtio d'error en disponibilitat, bloqueig i tancament?

Les preguntes anteriors haurien de poder respondre a nivell general i també a nivell de client específic, per tenir informació comercial sobre quins clients potenciar.
Per tant els informes seran:

- Informe de número de consultes per hora.
- Informe de número de consultes per zona y dia.
- Informe de demanda per tipus de distribució (nombre d'habitacions)
- Informe de temps de resposta de consulta per hora.
- Informe ràtio conversió disponibilitat / bloqueig de reserva per dia.
- Informe ràtio conversió disponibilitat / confirmació de reserva per dia.
- Informe ràtio d'error en disponibilitat per dia.
- Informe ràtio d'error en fase de bloqueig.
- Informe ràtio d'error en fase de confirmació de reserva.

5.4 Implementar lectura de dades.

En el nostre sistema com es va comentar anteriorment les dades s'han emmagatzemat en el Servei d'emmagatzematge d'objectes S3 i està previst generar informes a través QuickSight.

Per poder generar els informes prèviament cal llegir les dades per al seu posterior processat, QuickSight utilitza un motor de càlcul en memòria anomenat "SPICE" per carregar les dades en "SPICE" des S3 és necessari crear un fitxer de manifest JSON a on es pot especificar els arxius o carpetes a carregar des de l'origen de dades en aquest cas S3.

El fitxer manifest creat per al nostre propòsit és:

```
{
        "fileLocations": [
                {
                        "URIPrefixes": [
                                "https://s3.eu-west-3.amazonaws.com/tfggcns3/"
                        ]
                }
        ],
        "globalUploadSettings": {
                "delimiter": ","
        }
}
```

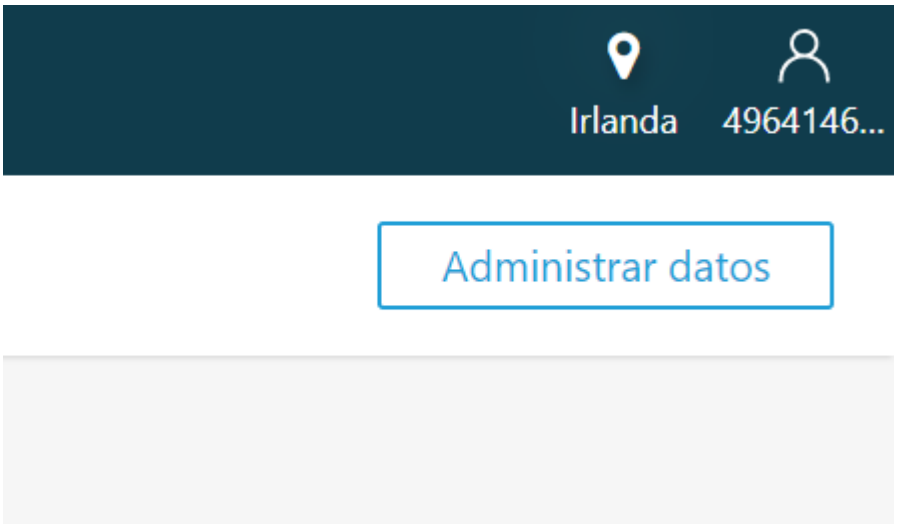
Explicació de les diferents seccions de l'arxiu manifest:

- fileLocations: especifica el bucket i la carpeta on s'allotgen les dades a importar.
- globalUpLoadSettings: especifica paràmetres per a la interpretació de les dades, com a tipus d'arxiu, delimitadors de dades, etc ... En el nostre cas s'especifica que el delimitador de camps és ",".

Un cop s'ha creat el procés de càrrega seria el següent:

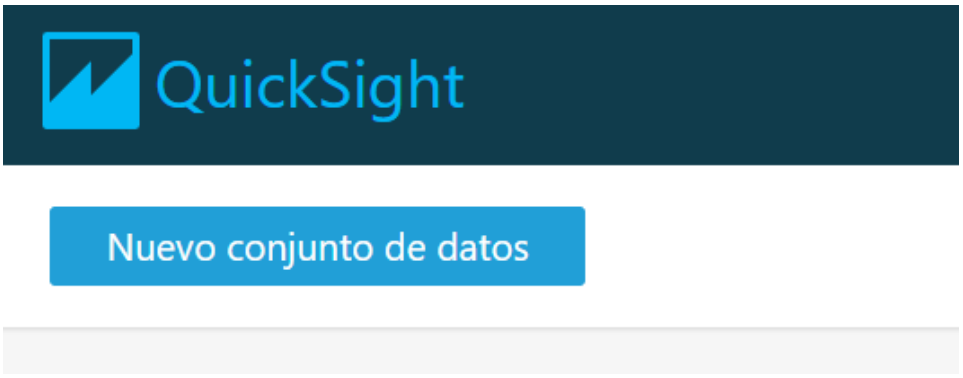
1. Accedir a: https://eu-west-1.quicksight.aws.amazon.com/sn/start

2. Seleccionar “administrar dades”:



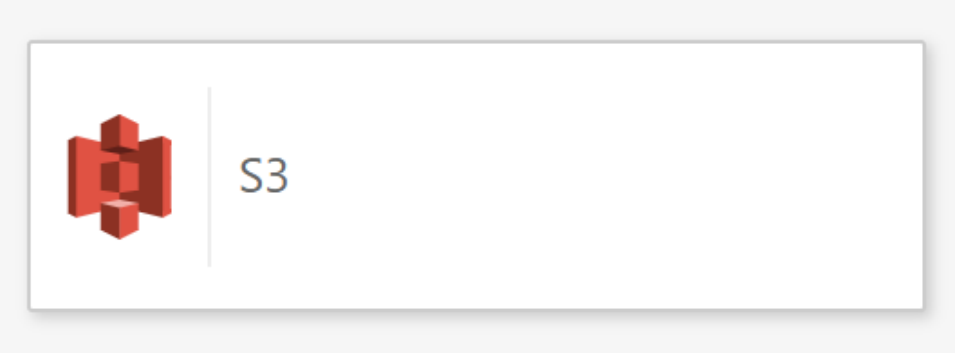
Il·lustració 4. Detall administrar dades

3. Seleccionar “Nou conjunt de dades”:



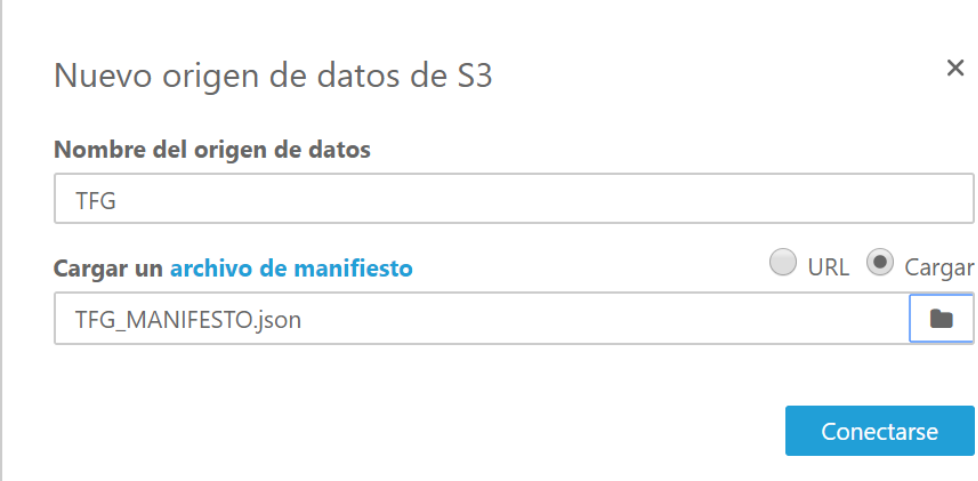
Il·lustració 5. Detall nou conjunt de dades

4. Seleccionar l'origen de dades, en el nostre cas “S3”:



Il·lustració 6. Detall origen S3

5. Donar nom a l'origen de dades i seleccionar l'arxiu manifest des del sistema d'arxius, també es podria carregar el fitxer manifest des d'una URL el que possibilita l'actualització dinàmica d'aquest manifest, aquesta opció és útil si l'origen de les dades pot ser variable
:



Il·lustració 7. Detall seleccionar arxiu manifest

Un cop realitzada la seqüència d'accions anterior, les dades quedaran guardats en "SPICE" per al seu posterior processament.

5.5 Implementar processament de dades per a l'obtenció dels informes.

Un cop carregats les dades en "SPICE" procedim a definir els càlculs necessaris a realitzar amb aquestes dades en relació amb els informes que es volen Obtenir.

Per motius de rendiment la informació guardada en els arxius csv que componen l'origen de dades no ha sofert canvis, per exemple, un dels camps guardats més importants en aquest arxiu csv és el camp que conté el missatge de transacció que s'envia al proveidor de Serveis. Aquest missatge conté informació rellevant del tipus d'ocupació, nombre de passatgers, data, etc ..., aquesta informació no està separada en diferents camps la qual cosa facilitaria bastant la interpretació i la creació d'informes.

Cal realitzar el Treball de processament de la informació per poder separar informació rellevant en camps "calculats" per tal que el sistema pugui realitzar informes de forma senzilla.

Nous camps calculats:

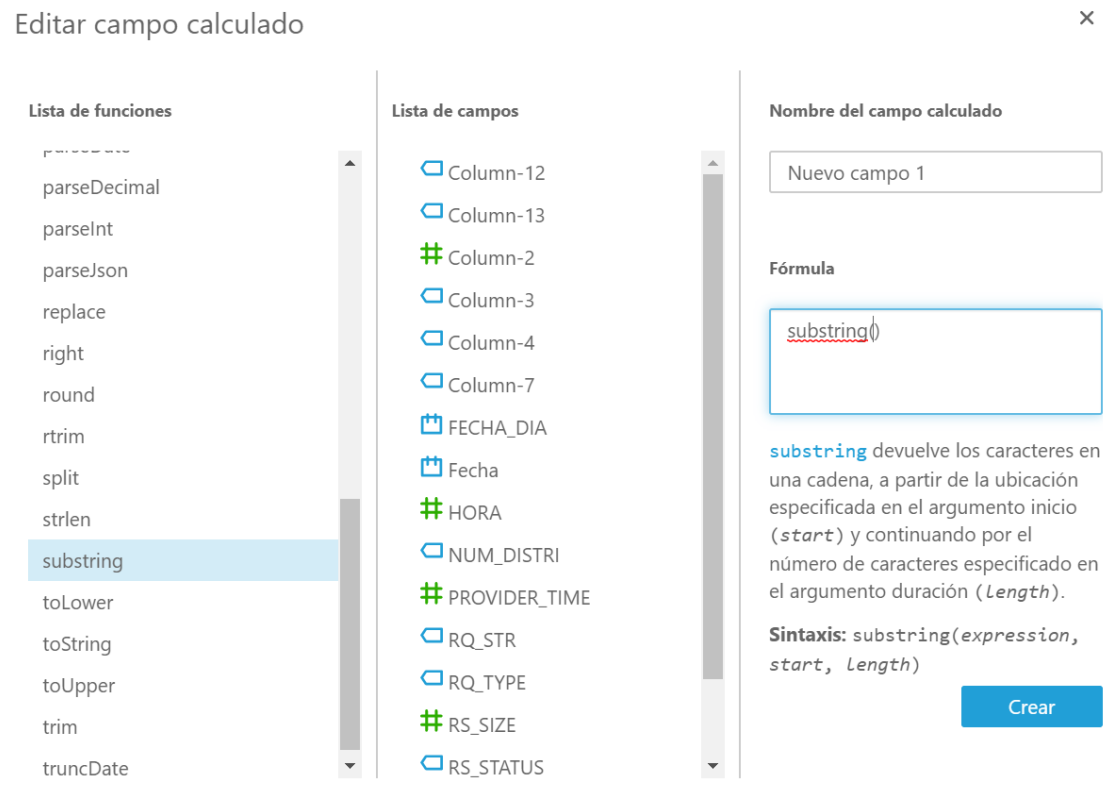
- NUM_DISTRI
  Conté informació del nombre d'habitacions sol·licitada en la consulta de disponibilitat.
  El format tipus de petició de disponibilitat de l'API del nostre sistema és:

```
<?xml version="1.0" encoding="UTF-8"?>
<DisponibilidadHotelPeticion>
    <ideses>30A68DBC435343A4B4929881D8BD63E8</ideses>
    <codtou>EPL</codtou>
    <fecini>07/04/2019</fecini>
```

```
    <fecfin>13/07/2019</fecfin>
    <flgdec>D</flgdec>
    <codser>705</codser>
    <distri id = "1">
        <numuni>1</numuni>
        <numadl>2</numadl>
        <numnin>2</numnin>
        <numbeb>0</numbeb>
        <edanin>7</edanin>
        <edanin>4</edanin>
    </distri>
</DisponibilidadHotelPeticion>
```

El tag distri indica la "distribució" sol·licitada i la propietat "id" indica el nombre de distribució (habitació) sol·licitada, per tant el camp NUM_DISTRI està definit amb una recerca del tag distri a través de la funció "substring "disponible, per això n'hi ha prou amb utilitzar l'opció" Afegir camp calculat ", seleccionar la funció a implementar,

els camps involucrats i escriure la fórmula:



Il·lustració 8. Detall modifica camp calculat

En el cas de NUM_DISTRI la fórmula és:

```
ifelse(substring(RQ_STR,11 + locate(RQ_STR,'distri id="5"'),1) =
'5','5',ifelse(substring(RQ_STR,11 + locate(RQ_STR,'distri id="4"'),1) =
'4','4',ifelse(substring(RQ_STR,11 + locate(RQ_STR,'distri id="3"'),1) =
'3','3',ifelse(substring(RQ_STR,11 + locate(RQ_STR,'distri id="2"'),1) =
'2','2',ifelse(substring(RQ_STR,11 + locate(RQ_STR,'distri id="1"'),1) = '1','1','0')))))
```

D'aquesta manera tenim un camp calculat que ens informa exactament el nombre de distribucions sol·licitades:

Il·lustració 9. Detall camp calculat distribucions sol·licitades

- FECHA_DIA

  Per a cada registre que compon els arxius de l'origen de dades, el sistema guarda la data i l'hora amb una precisió de milisegons, aquesta informació és massa precisa per confeccionar informes per dies de forma senzilla, per la qual cosa es fa necessari crear un nou camp calculat en què només s'inclou el dia. Per a això, seguint el procediment descrit anteriorment s'ha definit la següent formula:

  ```
  formatDate(Fecha,'dd MM yyyy')
  ```

  D'aquesta manera tenim un camp calculat que ens informa només de la data i no de l'hora:

Il·lustració 10. Detall camp calculat data

- HORA

  Atès que necessitem crear un informe que representi les peticions agrupades per hores, serà necessari tenir un camp calculat que inclogui l'hora. Per a això, s'ha definit la següent formula:

  ```
  extract('HH',Fecha)
  ```

  Aquesta formula extreu l'hora del camp Data de manera que tenim un camp calculat que ens informa exclusivament de l'hora:

Il·lustració 11. Detall camp hora

- ZONA

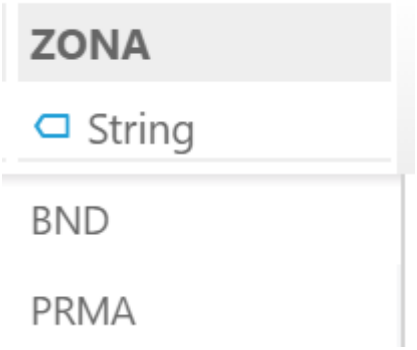
  Les consultes de disponibilitat es poden realitzar directament contra el codi d'un hotel o contra un codi de zona, en el segon cas es retorna una llista d'hotels. Per efectuar una consulta de disponibilitat per zona s'informa el tag <codzge> en comptes del tag <codser>:

```
<?xml version="1.0" encoding="UTF-8"?>
<DisponibilidadHotelPeticion>
    <ideses>30A68DBC435343A4B4929881D8BD63E8</ideses>
    <codtou>EPL</codtou>
    <fecini>07/04/2019</fecini>
    <fecfin>13/07/2019</fecfin>
    <flgdec>D</flgdec>
    <codzge>MAD</codzge>
    <distri id = "1">
        <numuni>1</numuni>
        <numadl>2</numadl>
        <numnin>2</numnin>
        <numbeb>0</numbeb>
        <edanin>7</edanin>
        <edanin>4</edanin>
    </distri>
</DisponibilidadHotelPeticion>
```

La petició anterior realitza una petició de disponibilitat per a la zona de Madrid, per tant per crear el camp calculat "ZONA" realitzarem una recerca a través de la funció substring de forma anàloga a com vam realitzar per calcular el camp NUM_DISTRI:

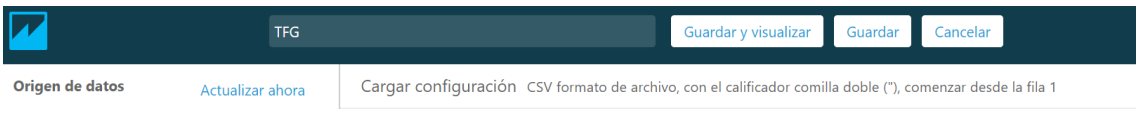
```
ifelse(RQ_TYPE = 'DisponibilidadHotelPeticion',replace(substring(RQ_STR,locate(RQ_STR,'codzge')
+ 7,4),'<',''),'no')
```

D'aquesta manera tenim un camp calculat que ens indica la zona:



Il·lustració 12. Detall camp calculat zona

Un cop tenim tots els camps necessaris guardem les dades a través de l'opció "Guardar i visualitzar":



Il·lustració 13. Detall opció "guardar i visualitzar"

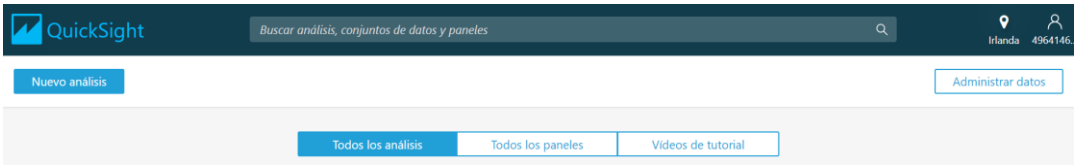
D'aquesta manera es crea un nou conjunt de dades anomenat "TFG".

5.6 Creació i refinament informes.

A partir d'un conjunt de dades prèviament creat, podem realitzar fàcilment una anàlisi amb els següents passos:

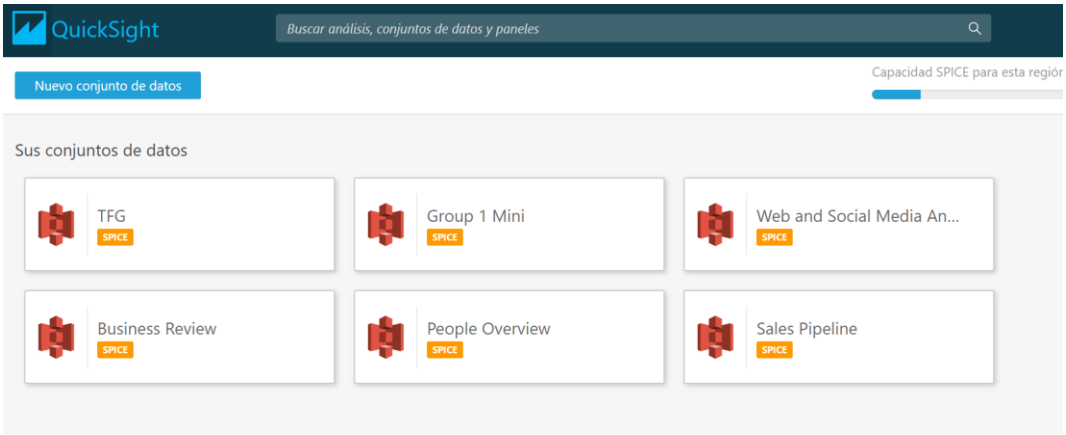
1. Nou anàlisi:

   Des de la pàgina principal de QuickSight tenim l'opció "Nou anàlisi":



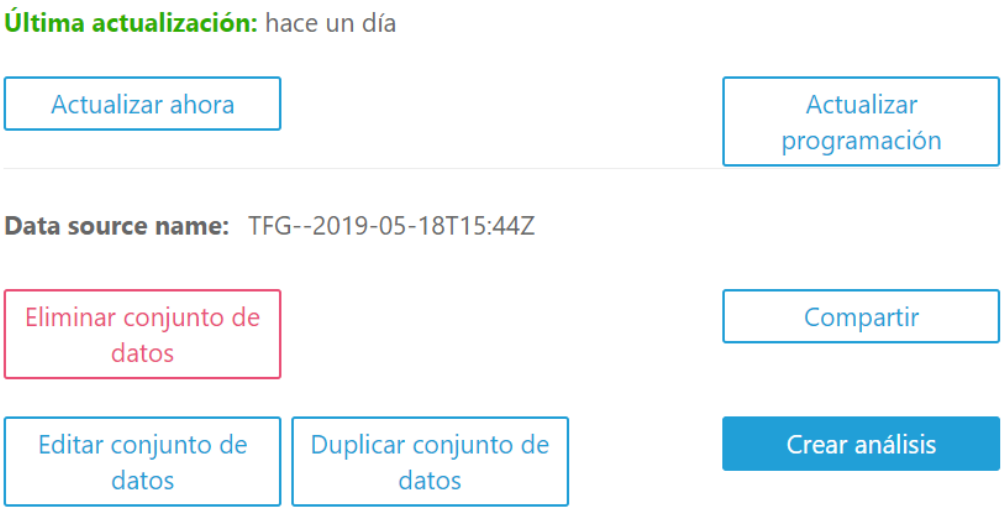
Il·lustració 14. Detall opció "Nou anàlisi"

2. El sistema ens mostrarà una sèrie de conjunt de dades disponibles per a realitzar l'anàlisi, en el nostre casos vam seleccionar TFG:



Il·lustració 15. Detall conjunt de dades disponibles

3. tot seguit “Crear anàlisis”:



Il·lustració 16. Detall "crear anàlisi"

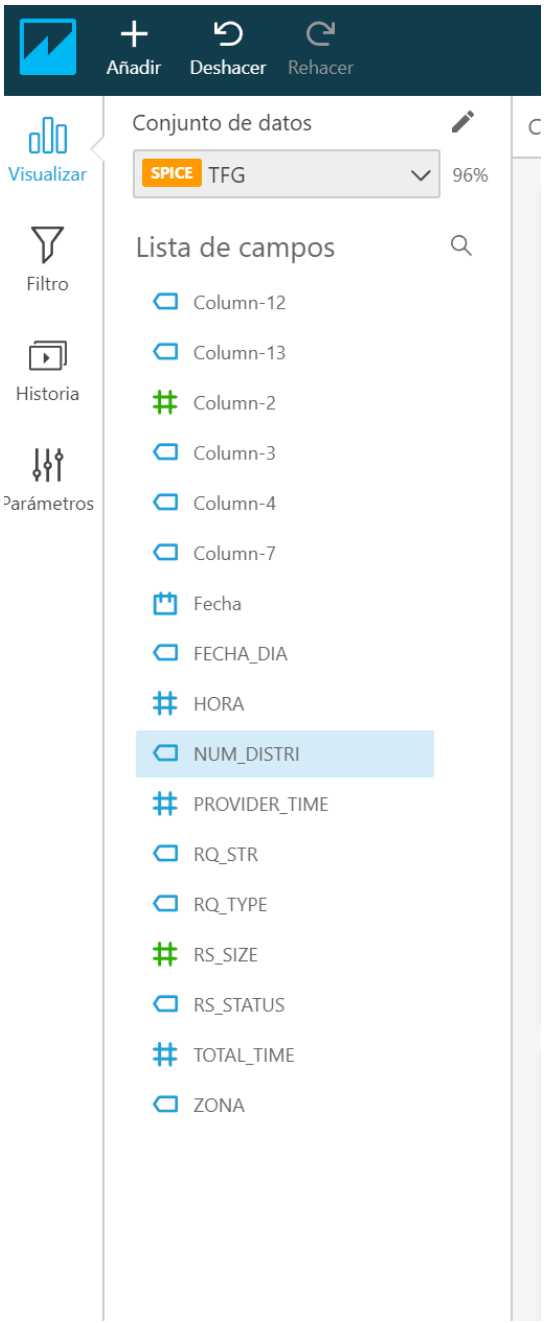
4. El sistema crearà una nova anàlisi.

5. Afegir un element visual:

   Per poder afegir nous elements visuals com un gràfic de barres, gràfic circular, gràfic de rectangles, etc ... haurem utilitzar l'opció "afegir".

   Posteriorment podrem seleccionar els camps que conformaran les dades a

mostrar en els elements visuals que anem creant, així com els filtres a aplicar.

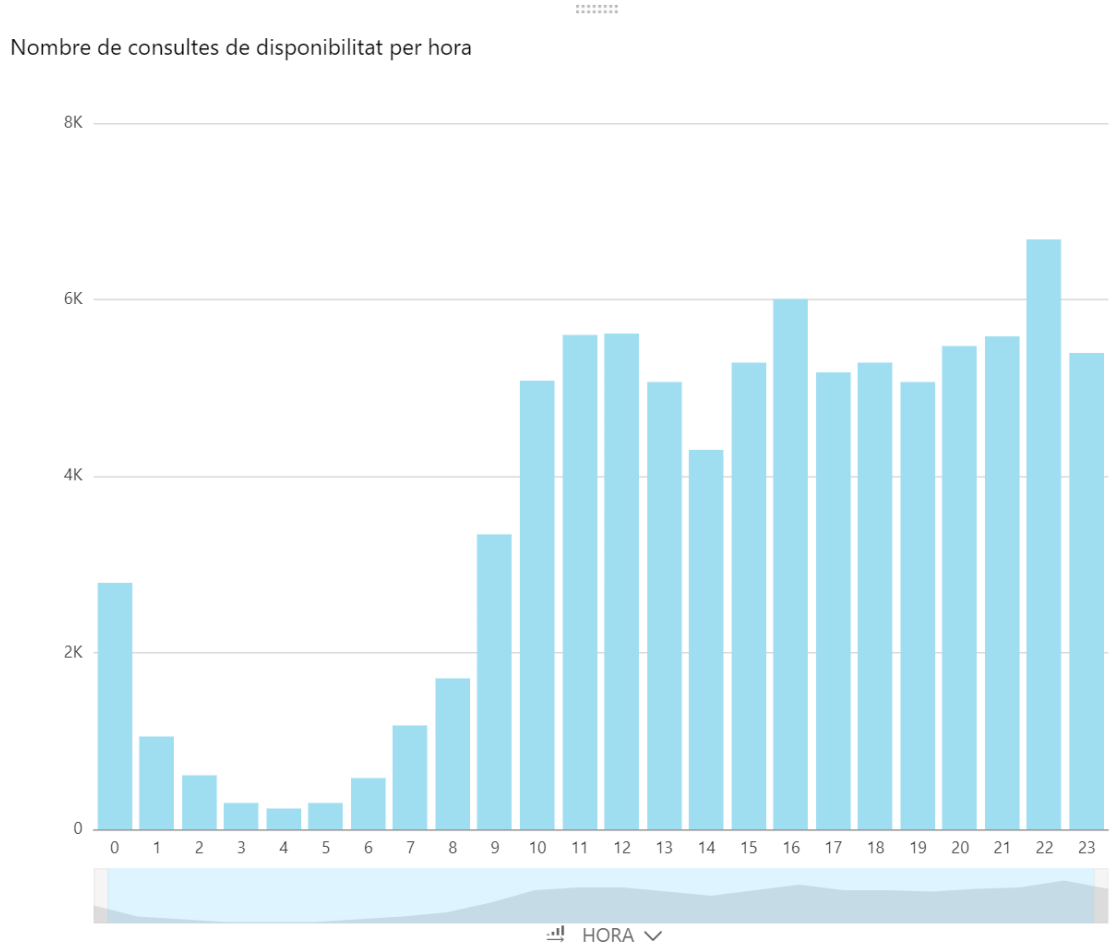


Il·lustració 17. Detall afegir element visual

5.6.1 Informes creats segons la pregunta a respondre

a. **A quina hora es realitzen més peticions de disponibilitat de serveis de vacances?**

S'ha creat l'element "***Nombre de consultes de disponibilitat per hora***"



Il·lustració 18. Informe nombre de consultes de disponibilitat per hora

En el qual es pot comprovar que l'hora amb més consultes per dia és entre les 22:00 i les 23:00.

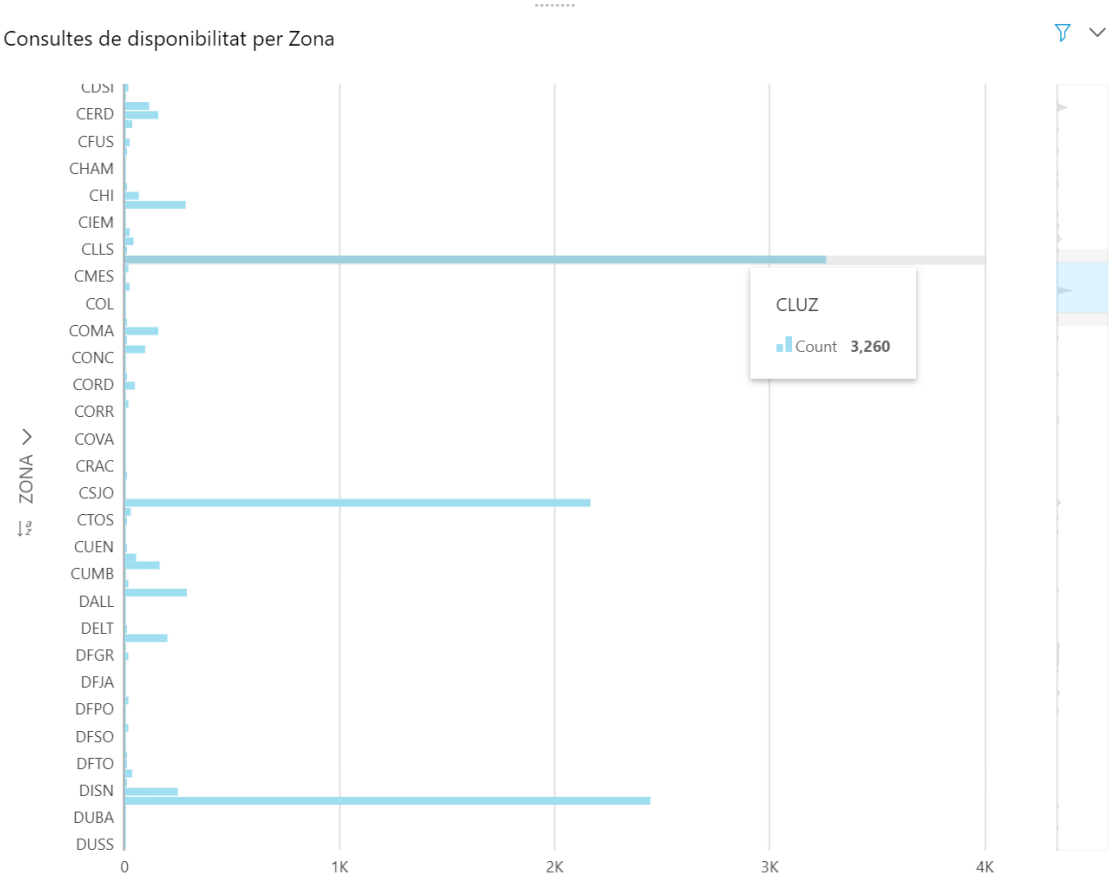
Aquesta informació serà molt útil per saber a quines hores hem de ser capaços

de absorbir major càrrega de treball assignant més recursos o dimensionant de forma dinàmica incrementant la capacitat just a aquelles hores.

D'altra banda també serà molt útil per determinar la mida de l'equip humà que ha de supervisar l'activitat del sistema, per exemple a les 4:00 del matí segons la gràfica no és pràcticament necessari que ningú revisi l'activitat de transaccions de disponibilitat perquè dit nombre de transaccions és pràcticament residual. D'altra banda a les 22:00 el sistema es troba en màxima càrrega de treball pel que serà necessari disposar un equip humà de supervisió per comprovar que s'està retornant producte adequat i / o atractiu per a la venda.

b. Quina zona geogràfica és la més demandada?

S'ha creat l'element "***Consultes de disponibilitat per Zona***":



Il·lustració 19. Informe consultes de disponibilitat per zona

En el qual es pot comprovar que la zona que més peticions rep és "CLUZ" (Costa de la llum). Així mateix amb aquesta gràfica es pot comprovar que hi ha moltíssimes zones que només han rebut una petició en tota una setmana, per tant per la part baixa del rànquing de peticions / zona hi ha empat entre moltíssimes zones diferents.

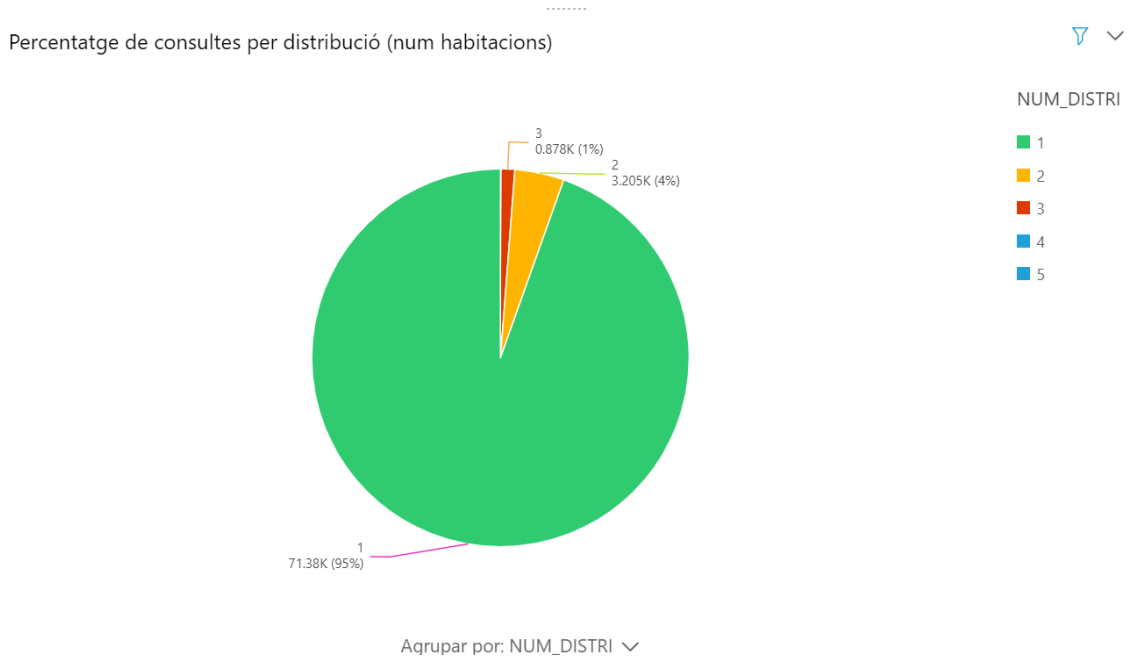
Aquesta informació és molt útil per a determinar per dos motius:

En primer lloc perquè ens dona informació de les preferències del mercat, sembla ser que hi ha cert interès per part de clients a visitar la costa de la llum ja que és la zona més demandada, per tant si aquestes consultes realment s'estan convertint en venda , podríem intentar rendibilitzar-incrementant el preu (llei de l'oferta i la demanda).

D'altra banda podríem potenciar la venda d'altres zones aplicant polítiques de millora dels productes oferts, potser els hotels oferts no són atractius per la seva ubicació o característiques pel podrien ser eliminats del sistema i substituïts per altres o bé es podria aplicar una política de baixada de preus per reactivar la venda d'aquesta zona.

c. Quin tipus de distribució és la més demandada?

S'ha creat l'element "***Percentatge de consultes per distribució (nombre d' habitacions)***":



Il·lustració 20. Informe percentatge de consultes per distribució

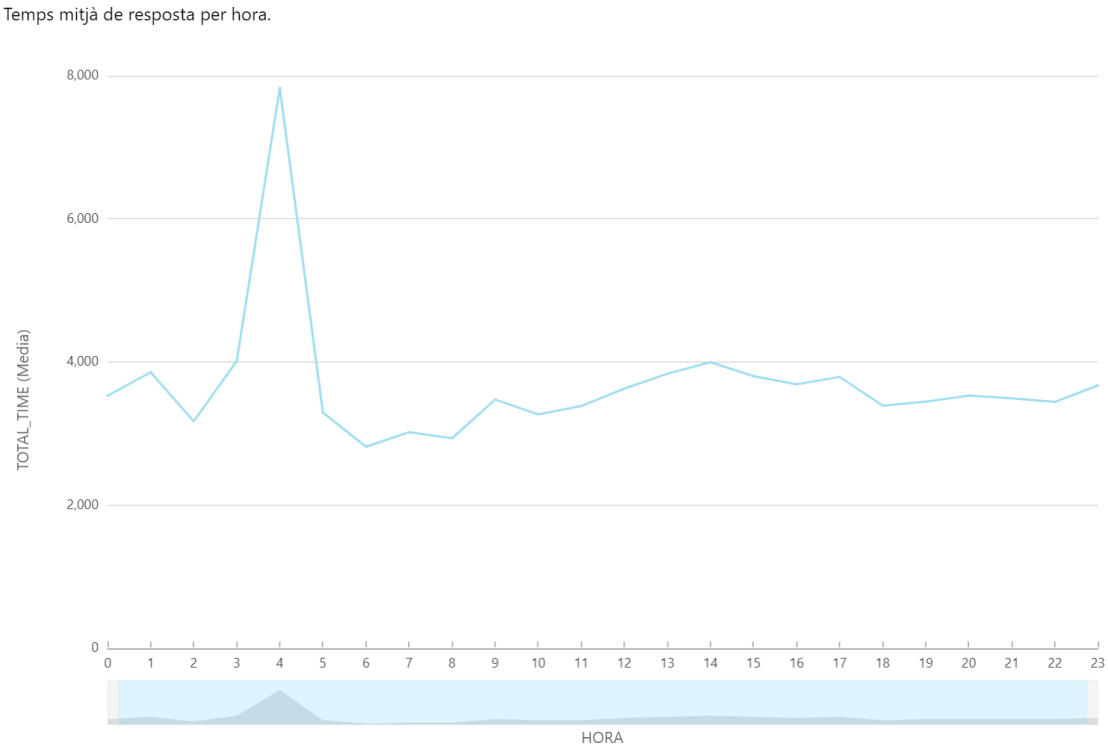
En el qual es pot comprovar que la majoria de les consultes són per a una sola habitació.

Aquesta informació és molt valuosa ja que la pot utilitzar el departament de producte de l'empresa per posar més afany a crear producte amb aquest tipus de distribucions i descartar sobre carregar el sistema amb productes que incloguin distribucions que amb prou feines se sol·liciten. També és un indicador de prioritat, és a dir si es detecta algun problema en el sistema relacionat amb distribucions d'una habitació, sabem que està afectant més del 95% del producte posat a la venda, de manera que tindrà màxima prioritat davant algun problema que afecti distribucions de cinc habitacions

que no es demanen.

d. Quin ha estat el temps mitjà de resposta segons l'hora del dia?

S'ha creat l'element "***Temps mitjà de resposta per hora***":



Il·lustració 21. Informe temps mitjà per hora

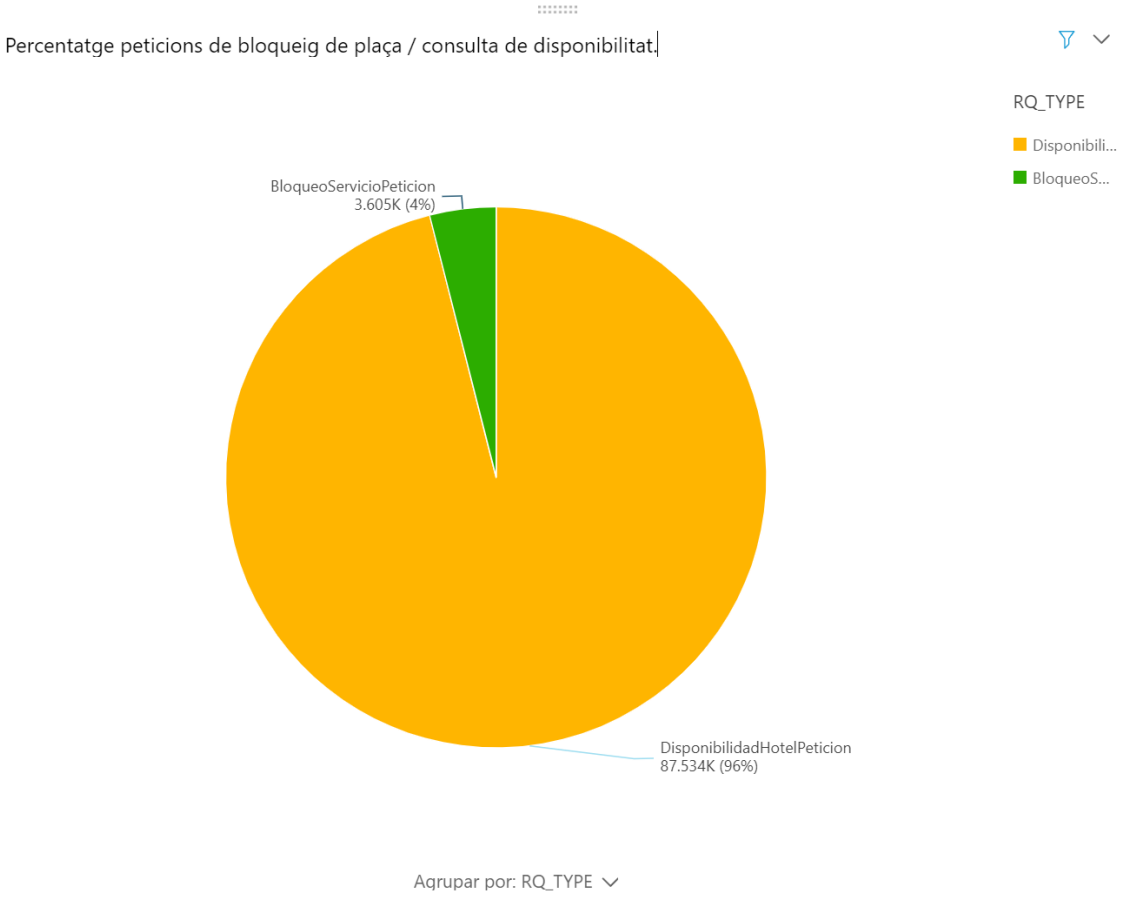
En el qual es pot comprovar que el sistema es manté més o menys estable pel que fa als temps de respostes 3s-4s, menys a les 4:00 hora en què els sistemes es reinicien.

Aquesta és informació és crucial per al departament d'IT de l'empresa ja que és un indicador de la "salut" del sistema. Un sistema d'aquestes característiques és molt sensible davant de qualsevol problema ja que es reflecteix ràpidament en els temps de respostes. Qualsevol canvi aplicat en el programari que no hagi passat les pertinents proves de rendiment podrien fer que el sistema s'alenteixi amb el consegüent increment en el temps de resposta.

Un temps de resposta elevat (> 5s) en aquest tipus de negoci suposa que els clients "desconnecten" la integració amb el teu sistema per no veure penalitzats per aquests temps elevats. La desconnexió implica que deixen d'oferir els teus productes per la qual cosa deixes de vendre automàticament. És una gràfica que s'haurà de revisar diàriament per detectar problemes en el nostre sistema, a més en alguns casos el sistema pot fer trucades a proveïdors tercers per oferir més producte per tant també ens veurem afectats si un proveïdor té problemes, nosaltres hauríem desconnectar a proveïdors de serveis que triguin a contestar a les peticions que el sistema li llança ja que afectarà el nostre temps de resposta.

e. **Quin ha estat el ràtio de conversió entre disponibilitat i bloqueig de reserva?**

S'ha creat l'element "***Percentatge peticions de bloqueig de plaça / consulta de disponibilitat***":



Il·lustració 22. Informe ràtio bloqueig de plaça / consultes de disponibilitat

En el qual es pot comprovar que la ràtio és aproximadament d'1 bloqueig per 24 consultes de disponibilitat.

Aquesta informació és molt valuosa per al departament de producte ja que és un indicador de l'interès real de formalitzar una reserva. Cal tenir en compte que el procés de confirmació d'una reserva sempre consta de tres passos:

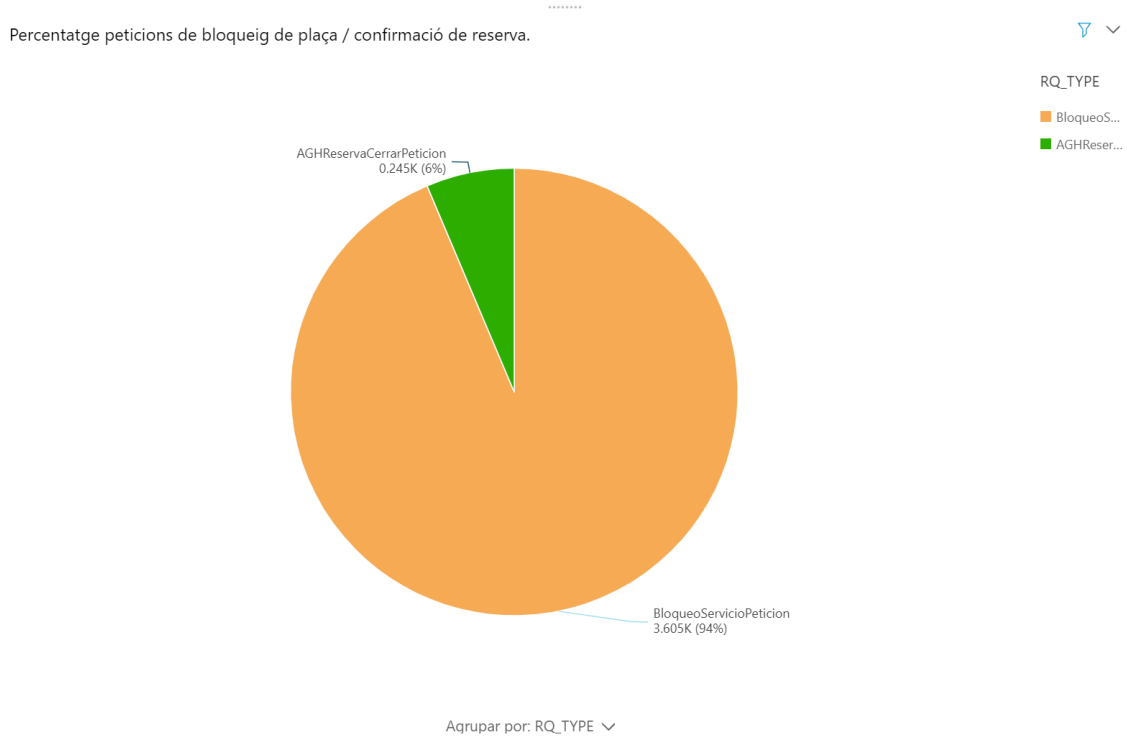
- disponibilitat
- bloqueig
- confirmació

Si es produeixen molt pocs bloquejos sobre moltes consultes de disponibilitat això vol dir que o bé ens estan escanejant preus o bé el client està aplicant una política de preus que finalment fa que no sigui atractiu el producte, en tots dos casos serà potestat del departament de producte posar-se en contacte amb el client per exposar la situació i tractar de corregir-la ja que totes les transaccions tenen un cost i una ràtio baix de transaccions de bloquejos fa inviable la sostenibilitat dels costos de les consultes ja que resulten improductives.

En aquest cas una ràtio de 1/24 és una ràtio excel·lent.

f. Quin ha estat el ràtio de conversió entre bloqueig de reserva i confirmació de reserva?

S'ha creat l'element "***Percentatge peticions de bloqueig de plaça / confirmació de reserva***":



Il·lustració 23. Informe ràtio bloqueig de plaça / confirmació de reserva

En el qual es pot comprovar que la ràtio és aproximadament d'1 confirmació de reserva per cada 14,7 peticions de bloqueig per servei.

Aquesta informació és molt important per al departament de producte ja que és l'indicador de negoci real ja que informa de la ràtio de conversió entre confirmació i bloqueig. Una ràtio de conversió molt baix denotaria que hi ha problemes de quota o que les condicions finals de la reserva no són atractives per als clients.

Els problemes de quota succeeixen quan ja queden poques places d'un determinat allotjament, cada bloqueig reserva una plaça de quota però hauria de garantir la confirmació.

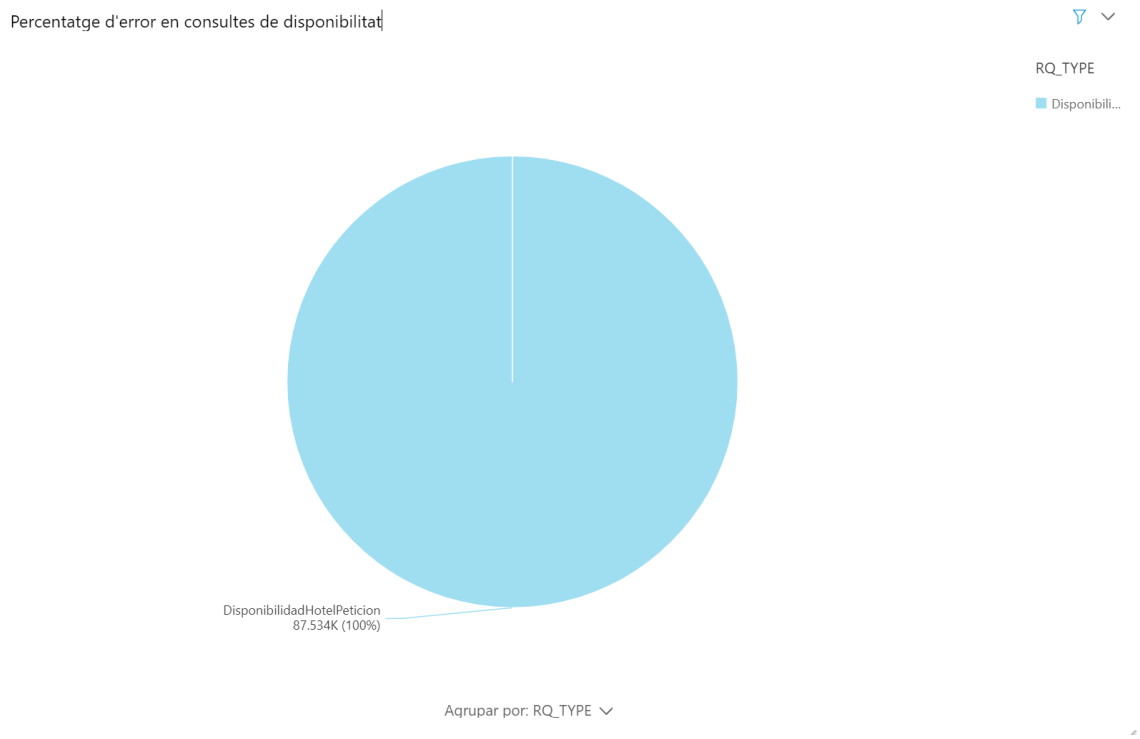
Les condicions finals de reserva són condicions que generalment només es poden consultar en fase de bloqueig, com ara despeses de cancel·lació, es pot donar el cas que les condicions de cancel·lació de la reserva no siguin atractius i per aquest motiu no s'acabi confirmant aquesta reserva.

Amb aquesta informació el departament de producte podrà realitzar les accions necessàries per reactivar la venda.

En aquest cas la ràtio aproximadament 1/14 és excel·lent segons els estàndards del mercat.

g. Quin ha estat el ràtio d'error en fase de disponibilitat?

S'ha crear l'element "***Percentatge d'error en consultes de disponibilitat***":



Il·lustració 24. Informe percentatge d'error en consultes de disponibilitat

En el qual es pot comprovar que el percentatge d'error en fase de disponibilitat és extremadament baix, menys de l'1% ja que les peticions de disponibilitat que

han acabat amb èxit són del 100%.

Aquesta informació és molt important per al departament d'IT.

El motiu és que els temps de resposta poden ser bons, i les ràtios de conversió també, però podrien passar desapercebuts errors en la resposta de disponibilitat.

Cal tenir en compte que per norma general un error de disponibilitat es respon abans que un procés de disponibilitat normal, encara que aquest fet depèn en gran mesura d'on es produeixi l'error, en la majoria dels casos, el sistema no haurà de realitzar certs treballs de traducció o càlcul de preus perquè el programa ha acabat de forma sobtada. Això provoca que es registri per a aquesta transacció un temps de resposta curt però que en realitat està relacionat amb un error i no amb una transacció correcta. A més aquests errors poden distorsionar les ràtios de conversió de bloqueig i confirmació.

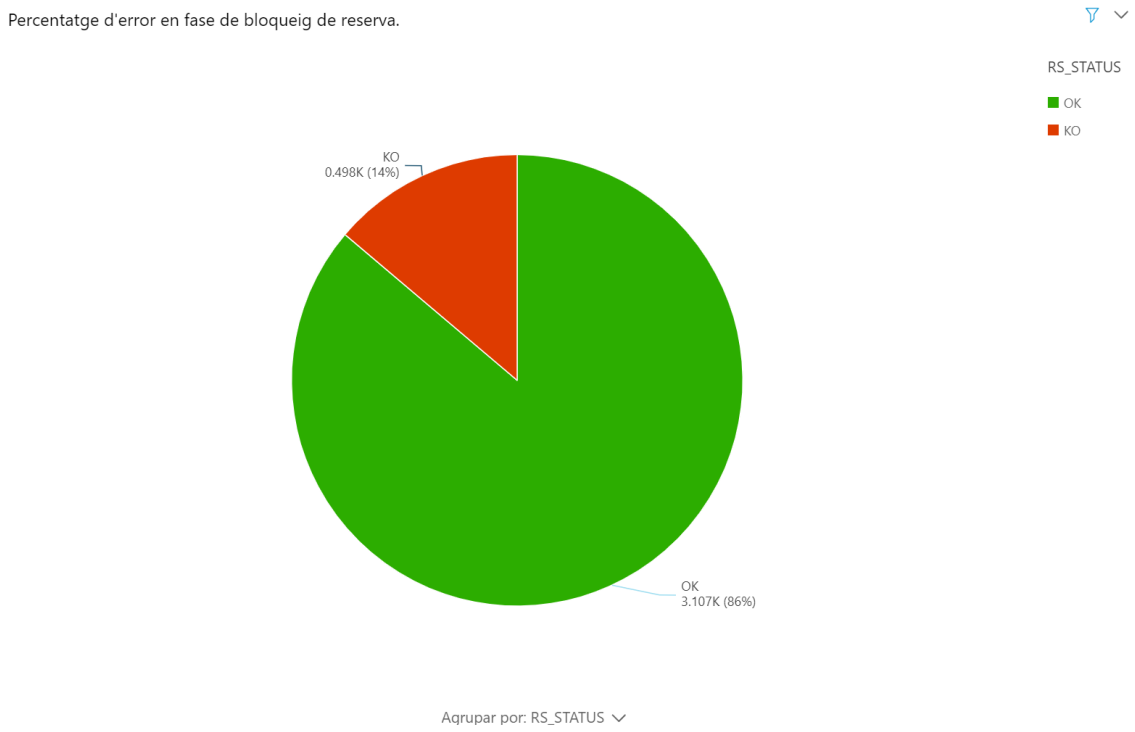
D'altra banda també pot provocar la desconnexió per part dels clients ja que aquests errors poden provocar problemes en els seus sistemes.

Per aquest motiu el percentatge d'error en fase de disponibilitat deu ser el mínim possible

En aquest cas registrar prop de 100% lliure d'error és una dada excel·lent.

h. Quin ha estat el ràtio d'error en fase de bloqueig de reserva?

S'ha creat l'element "***Percentatge d'error en fase de bloqueig de reserva***":



Il·lustració 25. Informe percentatge d'error en fase de bloqueig de reserva

En el qual es pot comprovar que la ràtio és aproximadament d'1 error cada 6,2 peticions de bloqueig (14% d'error).

Aquesta informació serveix en major mesura per controlar els proveïdors d'estades que el sistema consulta a través dels connectors disponibles a aquest efecte.

Els errors de bloqueig per producte propi es donen amb poca freqüència i són fàcilment detectables pel departament de producte a través dels informes específics, també són fàcilment esmenables ja que el departament té el control total sobre el producte. No obstant això, errors en fase de bloqueig contra proveïdors són més complicats de gestionar ja que el departament de producte no té control sobre aquest producte extern i la seva única acció disponible és desconnectar aquest proveïdor per no oferir el seu producte i no generar errors.

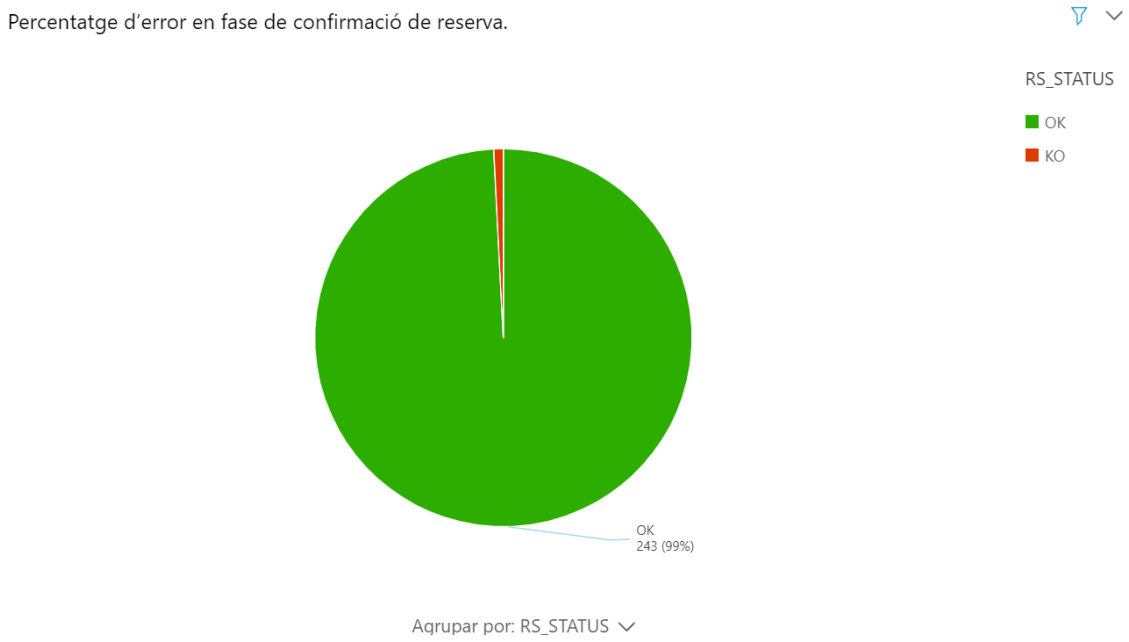
Si es detecten percentatges d'error elevats cal posar-se en contacte amb el proveïdor i exposar aquesta circumstància perquè el solucionin al més aviat possible ja que d'una altra manera serà necessària la desconnexió per evitar que al seu torn ens desconnectin els clients.

També pot servir al departament d'IT per monitoritzar la repercussió que pot estar tenir alguna modificació o circumstància excepcional en el sistema.

En aquest cas el percentatge d'error és certament elevat i s'hauran de prendre les mesures necessàries per corregir-ho.

i. **Quin ha estat el ràtio d'error en fase de confirmació de reserva?**

   S'ha creat l'element "**Percentatge d'error en fase de confirmació de reserva**":



Il·lustració 26. Informe percentatge d'error en fase de confirmació de reserva.

En el qual es pot comprovar que la ràtio és aproximadament d'1 error cada 121,5 peticions de confirmació de reserva (1% d'error).

De forma anàloga a com passa en fase de bloqueig, els errors en la confirmació de reserva es donen amb poca freqüència, molt menys encara que a la fase de bloqueig.

Això és degut a que la informació que maneja el sistema en fase de tancament és menys complexa que en la fase de bloqueig, en fase de bloqueig ja es manegen totes les característiques del producte: descripció completa, polítiques de preus, polítiques de cancel·lació, extres , etc ..

A més el sistema és molt més permissiu quant timeouts per evitar problemes per sobrecàrrega, de fet és fàcil trobar configuracions de timeout de 120 segons per tal d'evitar en mesura del possible aquest tipus d'error.

L'error més vigilat per les empreses de touroperació són els errors de confirmació ja que és una venda que ja no era potencial, és una venda que ha acceptat el client, però el sistema ha rebutjat per problemes tècnics.

Per tant com en el cas dels bloquejos s'haurà de determinar si és producte propi o de tercers i actuar en conseqüència.

Els departaments d'IT solen tenir alarmes que avisen directament als responsables de producció d'aquest tipus d'eventualitats per intentar recuperar la reserva i confirmar-la per tots els mitjans.

En aquest cas un percentatge d'1% és assumible però haurà de ser revisat ja que en aquest tipus de transacció es pretén estar el més a prop possible del 0%

6. Problemes detectats:

- **Per trobar l'eina**: Amazon proporciona un ventall molt gran d'eines en el seu ecosistema big data, en alguns casos aquestes eines es solapen en quant a funcionalitat sobretot per la seva capacitat de posar en producció noves eines constantment. Això propicia certa dificultat a l'hora de triar quina és l'eina adequada ja que en alguns casos la informació aportada no està degudament cohesionada com ara la no inclusió d'algunes eines en esquema d'arquitectura de casos d'ús d'exemple.
- **Per identificar el client**: Amb la informació que guarda actualment el sistema, no és possible identificar el client (agència de viatges) que ha realitzat la compra, això és degut a que la informació que identifica el client que està realitzant la compra es troba en el missatge que obre sessió , del tipus:

```
<SesionAbrirPeticion>
        <codsys>AGH</codsys>
        <idtusu>9028794xx</idtusu>
        <pasusu>876700xx</pasusu>
        <codage>876700xx</codage>
        <tipcre>X</tipcre>
        <chkpol>S</chkpol>
</SesionAbrirPeticion>
```

On el tag <idtusu> és l'identificador de client, aquesta transacció no es està guardant en els logs per problemes de costos i rendiment.

Es proposa realitzar canvi en la implementació del sistema agregador que tracta les peticions per inclogui el <idtusu> en el registre de logs per tal de poder identificar de manera inequívoca l'origen de cada transacció sense haver d'anar a buscar per l'identificador de sessió <idses> que s'informa en tot el flux de reserva. Cal destacar que el sistema si és capaç de determinar a quin client correspon cada reserva a través del dashboard de control de reserves perquè guarda aquesta informació en la base de dades de reserves.

7. Conclusions

Comencem l'apartat amb les conclusions que podem obtenir a partir de les anàlisis obtinguts a manera de resum final.

- La màxima càrrega de treball es rep a les 22:00 i la mínima a les 04:00.
- La zona més demandada és "Costa de la llum".
- La distribució més demandada és d'una habitació.
- El temps de resposta de les transaccions de consulta de disponibilitat és molt estable 3s-4s.
- La ràtio de bloqueig és excel·lent, 1 bloqueig per cada 24 consultes de disponibilitat
- La ràtio de confirmació és excel·lent, 1 tancament per cada 14 bloquejos.
- El percentatge d'error en les peticions de disponibilitat és excel·lent, per sota de l'1%
- El percentatge d'error en les peticions de bloqueig indica que hi ha un problema en el sistema o proveïdors tercers 14% d'error.
- El percentatge d'error en les peticions de tancament és assumible però cal revisar 1%.

Segons aquestes dades la conclusió que podem treure és que el sistema respon en temps òptims, hi ha un producte atractiu, però en fase de bloqueig i tancament es produeixen més errors dels desitjables per la qual cosa s'haurien de revisar, a més de potenciar noves zones de disponibilitat ja que hi ha un notable desequilibri entre aquestes.

Pel que fa al desenvolupament del projecte, com lliçó principal d'aquest treball, m'agradaria emfatitzar la increïblement ràpid i senzill que pot ser implementar un sistema BI complet utilitzant eines cloud, el desenvolupament d'aquest TFG s'ha realitzat en base a tecnologies AWS però les tecnologies aportades per altres empreses com pot ser Microsoft (Azure) haurien de tenir un procés anàleg de posada en marxa.

Una altra lliçó a destacar és el fet que sense dades no hi ha anàlisi pel que és molt important registrar totes les dades que genera el sistema, serà necessari realitzar modificacions en certs components perquè registrin més i millor estructurats en mesura del possible aquestes dades .

També s'ha après a que en circumstàncies normals no serà possible implementar un sistema BI sense certa coordinació amb el departament IT de l'empresa ja que és molt probable que s'hagin d'abordar canvis en els sistemes en relació a la informació guardada o en relació a ajustaments relacionats amb l'arquitectura del sistema.

D'altra banda, els objectius del treball s'han aconseguit només a nivell general, és a dir, sense poder determinar a quin client pertany cada reserva o nombre de consultes de disponibilitat realitzada això és degut en gran part al fet que com es comenta en el paràgraf anterior, és molt difícil implementar un sistema BI sense que des del departament d'IT hagin de realitzar certs ajustos en el sistema.

En concret en aquest cas cal incloure més informació amb l'objectiu de poder identificar als clients que realitzen les peticions.

Aquesta circumstància confirma que els objectius no eren realistes amb l'especificació de l'API actual, en la fase d'anàlisi d'aquest API, s'hauria d'haver detectat que aquest objectiu no era assequible sense fer canvis de programació.

Cal tenir en compte per a futurs projectes, que en aquestes circumstàncies cal comptar amb recursos de l'àrea de desenvolupament del departament d'IT per tal de poder abordar modificacions en el sistema.

Pel que fa a la planificació del treball, van quedar definides les tasques de processament i definició d'informes en PACS separades (PAC2 i PAC3) això ha suposat un problema ja que el processat de les dades va lligat al tipus d'informe perquè el que en principi aquestes tasques haurien de realitzar en una mateixa PAC. Com a resultat es va produir una desviació en la tasca de processament de dades ja que es va realitzar a la PAC3 alhora de la definició dels informes. Aquesta tasca de processament es va realitzar en paral·lel a "Implementar lectura de dades per a l'Obtenció dels informes".

D'altra banda la metodologia de treball definida en l'apartat ha resultat eficaç i no ha estat necessari cap canvi al respecte, si bé és cert que durant la captura de dades es va produir un fet fora de control i és vaig haver de fer un viatge precisament de treball que m'impedia obtenir aquesta informació el que va impedir tenir una mostra més gran de dades.

Com a línies de treball a futur sería interessant aplicar tècniques de machine learning per tal que el sistema sigui capaç d'optimar el producte de forma autònoma en conjunt amb tècniques de predicció, d'aquesta manera es podrien configurar productes atractius de forma dinàmica en dies en concret o fins i tot a certes hores .

De la mateixa manera també es podrien aplicar aquestes tècniques per prevenir càrregues de treball i dimensionar de forma dinàmica la capacitat del sistema aprovisionant més instàncies EC2 o incrementant les instàncies serverless dels serveis que ho requereixin just abans que es produeixin puntes de treball amb el consegüent estalvi en costos.

8. Glossari

- **API:** Interfície de programació d'aplicacions, és un conjunt de funcions, subrutines, procediments o mètodes que ofereix certa biblioteca per ser utilitzat per un altre programari com una capa d'abstracció.
- **AWS:** Amazon Web Services, és una col·lecció de serveis Cloud oferta per la companyia Amazon.
- **Azure:** Servei en el núvol oferta com a servei i allotjat en els Data Centers de Microsoft.
- **BI:** Bussines intelligence es refereix a la utilització de les dades que genera una empresa per a la presa de decisions. Per a això es defineixen una sèrie d'estratègies i s'utilitzen una sèrie d'eines per tal de análizar les dades existents.
- **Bucket:** En AWS contenidor de dades a manera d'objectes.
- **Cloud:** En termes informàtics ens referim a un paradigma que permet oferir serveis de computació a través d'una xarxa, que normalment és Internet.
- **CSV:** Fa referència a un tipus de fitxer que conté valors separats per coma.
- **EC2:** Amazon Elastic Compute Cloud es un servei web que proporciona capacitat informàtica en el núvol per allotjar màquines virtuals**.**
- **ETL:** Extreure, transformar i carregar. Fa referència a un procés que mou dades des de múltiples fonts i els processa per carregar-los en una altra base de dades per a ser analitzats per les organitzacions.

- **Hosting:** És un servei mitjançant el qual es pot publicar una pàgina web o un altre tipus de serveis perquè puguin ser accedits per usuaris des d'internet.

- **HPC:** Computació d'alt de rendiment, fa referència a l'estratègia d'agregació de potència de càlcul aplicats a gestió, ciència i enginyeria.

- **Java:** Llenguatge de programació.

- **JSON:** acrònim de JavaScript Object Notation, és un format d'arxiu de text per a l'intercanvi de dades.

- **Machine Learning:** És un mètode d'anàlisi de dades que automatitza la construcció de models analítics. Es basa en la creença que els sistemes poden aprendre de les dades identificant patrons i prenent decisions amb mínima intervenció.

- **MS Project:** Software de gestió de projectes comercialitzat per Microsoft.

- **.NET:** Entorn gestionat d'execució d'aplicacions, llenguatges de programació i compiladors creat per Microsoft que permet el desenvolupament de tot tio de programes o serveis Windows així com aplicacions per a dispositius mòbils.

- **Latència:** És el temps que transcorre entre que es llança una petició i es rep el primer bit de la resposta.

- **Operador turístic:** Empresa majorista de turisme que ven viatges organitzats.

- **Serverless:** És un model de computació en el núvol en què el proveïdor de serveis cloud executa el servidor i administra de forma dinàmica els seus recursos.

- **SLA:** Acord de nivell de servei en els quals es defineixen els nivells de servei als quals es compromet el proveïdor de serveis amb el client.

- **SQL:** Structured Query Language, llenguatge dissenyat per administrar i recuperar informació de sistemes de gestió de bases de dades relacionals.
- **SPICE:** En AWS sistema d'emmagatzematge de dades en memòria d'alt rendiment.
- **Stream:** Canal a través del qual flueixen les dades.
- **Timeout:** El temps màxim permès per a que un procés acabi de forma normal.
- **Webservices:** Tecnologia que utilitza un conjunt de protocols i estàndards que serveixen per intercanviar dades entre aplicacions.

9. Bibliografia

Llibres:

Josep Curto Díaz, *Introducció al Bussiness Intelligence*, Editorial UOC, Barcelona, 2010.

Material UOC:

- Jordi Conesa Caralt, José Luis Gómez García, Josep Curto. *Fundamentos y usos del big data* Editorial Oberta UOC Publishing SL , 2016
- Jordi Conesa Caralt, José Ramón Rodríguez, Josep Curto. *Fundamentos de intel·ligència de negocio*, Editorial Oberta UOC Publishing SL , 2016

Referències web:

- https://docs.aws.amazon.com/quicksight/latest/user/amazon-quicksight-user.pdf (Junio 2019)
- https://aws.amazon.com/es/s3/storage-classes/?nc=sn&loc=3 (Junio 2019)

- http://lema.rae.es/dpd/srv/search?id=N9YfVc98xD6F5Wqog2 (Junio 2019)
- https://es.wikipedia.org/wiki/Servicio_web (Junio 2019)
- https://es.workmeter.com/blog/bid/177356/qu-es-el-business-intelligence (Junio 2019)
- https://es.wikipedia.org/wiki/Computaci%C3%B3n_de_alto_rendimiento (Junio 2019)
- https://es.wikipedia.org/wiki/Extract,_transform_and_load (Junio 2019)
- https://es.wikipedia.org/wiki/Latencia (Junio 2019)
- https://sistemas.com/hosting.php (Junio 2019)
- https://es.wikipedia.org/wiki/Acuerdo_de_nivel_de_servicio (Junio 2019)
- https://aws.amazon.com/es/ec2/ (Junio 2019)
- https://www.campusmvp.es/recursos/post/que-es-la-plataforma-net-y-cuales-son-sus-principales-partes.aspx (Junio 2019)
- https://es.wikipedia.org/wiki/SQL (Junio 2019)
- https://es.wikipedia.org/wiki/Interfaz_de_programaci%C3%B3n_de_aplicaciones (Junio 2019)
- https://docs.aws.amazon.com/es_es/AmazonS3/latest/gsg/AmazonS3Basics.html (Junio 2019)