ESERCITAZIONE

12 Agosto 2022

Davide Caruso

Consegna prevista per le 18:00, una volta terminata la prova bisognerà inviarlaa soluzioni.esercizi2022@gmail.com

La prova va svolta necessariamente in autonomia. Buon lavoro!

Teoria

- **Quali sono le principali architetture di un Data Warehouse? Descrivile nel dettaglio.**
  L'architettura di un data warehouse è determinata dalle esigenze specifiche dell'organizzazione. Di seguito sono indicate alcune delle architetture comuni:
  - *Semplice. Tutti i data warehouse condividono una struttura di base dove metadati, dati di riepilogo e dati non elaborati sono archiviati nel repository centrale del warehouse. Il repository è alimentato da fonti di dati da un lato ed è accessibile da parte degli utenti finali ai fini di analisi, reporting e data mining dall'altro.*
  - *Semplice con un'area di gestione temporanea. I dati operativi devono essere puliti ed elaborati prima di essere messi nel warehouse. Sebbene tale operazione possa essere eseguita a livello di programmazione, molti data warehouse aggiungono un'area di gestione temporanea*

per i dati prima che vengano introdotti nel warehouse, per semplificare la preparazione dei dati.

- *Hub e spoke. L'aggiunta di data mart tra il repository centrale e gli utenti finali consente a un'organizzazione di personalizzare il proprio data warehouse in modo da poter gestire diverse linee di business. Quando i dati sono pronti per l'uso, vengono spostati nel data mart appropriato.*
- *Sandbox. Le sandbox sono aree private, protette e sicure che consentono alle aziende di esplorare in modo rapido e informale nuovi set di dati o metodi di analisi dei dati senza dover rispettare la compliance con le regole formali e il protocollo del data warehouse.*

- **Cosa sono i Data-Mart? Qual è la loro importanza?**

Un Data Mart è un database strutturato in base ad un unico argomento, una singola unità aziendale, ha relativamente poche sorgenti (rispetto all' intero datawarehouse di cui fa parte),al livello di integrazione dati le informazioni contenute riguardano una singola area di interesse, il grado di intervallo temporale è molto dettagliato (minuti, settimane o massimo mesi)

Organizzare un grande data werehouse in vari data mart offre vari vantaggi tra cui:

accesso efficiente dai vari set di dati,semplificazione nella fase di sviluppo e miglioramento delle prestazioni (potendo distribuire il carico di lavoro di elaborazione su piu architetture hardware separate).

- **Quali due teorie danno le due definizioni dei Data Warehouse viste a lezione? Descrivi entrambe le teorie e sottolineane le differenze.**

Esistono due differenti tipoligie di Progettazione e sviluppo di un Data warehouse:

Il modello Inmon e il modello Kimball

Il modello Inmon:

ha un approccio Top-Down, i dati riguardano una larga area di Business, per l'implementazione è necessario un Team di

Specialisti,la fase di avvio richiede molto tempo ed ha un elevato costo che si riduce nelle fasi successive, il mantenimento è semplice,l'implementazione ha un grado di complessità elevato)
Il modello Kimball:
Ha un approccio Buttom-Up,i dati riguardano una singola area di Business, è sufficiente un generico team di sviluppo per l'implementazione,la fase iniziale può essere relativamente veloce, il costo dello sviluppo è costante e basso anche in fase di avvio, il mantenimento è difficile, complessivamente richiede un minore tempo di sviluppo).

- **Descrivi la gerarchia parent-child nei DWH.**

All'interno delle tabelle delle dimenzioni di un datawerehouse multidimensionale è possibile implementare delle gerarchie Padre-Figlio che consentono in fase di visualizzazione della matrice di dati multidimenzionale di variare la granularità dei dati (esempio nella dimenzione Date si può raggruppare la data-tempo in giorno gg/mm/aaaa oppure, in anno/mese, oopure in anno).

- **Quali sono i vantaggi dell'uso di un DWH?**

I data warehouse danno la possibilità di contenere grandi volumi di dati strutturati, che sono utilizzati in fase di analisi sia storiche sia previsionali, per prendere decisioni di Business (es. migliorare la propria strategia di marketing, incrementare il fatturato minimizzando gli investimenti, ecc.).

- **In che modo i DWH, i database e i data lake funzionano insieme?**

I data lake sono spesso utilizzati per archiviare i dati grezzi in fase di importazione sul Cloud, prima di essere poi gestiti in un processo di ETL che li elabora e memorizza su un datawerehouse strutturato.

- **Quali sono le differenze tra DWH e Data Lake? Descrivile nel dettaglio.**

I Data Lake: hanno una struttura dati grezza, la finalità dei dati è

incerta,sono utilizzati dai tecnici informatici, offrono elevate performance in fase di accesso e modifica.
I data WareHouse: hanno una struttura dati elaborata, tutti i dati contenuti sono utili ed utilizzabili,vengono utilizzati dai professionisti di Business aziendale,hanno un grado di complessita e costi elevato in fase di creazione o di modifiche strutturali.

- **Come funziona lo storage dei dati in un DWH?**

I dati grezzi entrano con un processo di caricamento sul data lake. Successivamente vengono puliti da eventuali informazioni di dati sensibli, o non legali o non sicuri attraverso la fase Tokenized Data Raw, poi vengono raffinati tramine delle validazioni di integrita referenziale e memorizzati in modo strutturato. In fine nella fase di Master Data vengono ulteriormente filtrati e verificati in base a controlli di complessita piu elevata o manuale, per poi in fine essere resi disponibili ai Data Scientists.

- **Descrivi nel dettaglio il processo di ETL.**

ETL (estrazione, trasformazione e caricamento) è una pipeline di dati usata per raccogliere dati da diverse origini. Trasforma quindi i dati in base alle regole business e li carica in un archivio dati di destinazione. Le operazioni di trasformazione nella pipeline ETL vengono eseguite in un motore software specializzato e spesso comportano l'uso di tabelle di staging per archiviare temporaneamente i dati che vengono trasformati e infine caricati nella rispettiva destinazione.

- **Cosa sono le Late Arriving Dimension? Come vengono gestite?**

Nel processo ETL di Data Warehouse, la dimensione in arrivo in ritardo rappresenta il rilevamento dell'arrivo di record in una tabella dei fatti e la dimensione corrispondente non esiste. Di conseguenza, una dimensione che arriva in ritardo ha una chiave naturale che esiste nei nuovi dati di fatto, che non esiste ancora nella dimensione.

Ci sono tre differenti soluzioni a questo problema:

Never Process Fact: I dati che non verificano l'integrità referenziale vengono scartati.

Park and Retry: I dati vengono accantonati per un successivo tentativo di reinserimento (in attesa di un aggiornamento delle dimensioni)

Inferred Flag (inseriti con particolari valodi di default o flag, del tipo N/A,ecc)

- **Come funziona e a cosa serve il Master Data Management?**

Il Master Data Management è un insieme di strumenti e i processi utilizzati da un'organizzazione per stabilire un'unica fonte di verità per tutti i suoi dati critici. Attraverso la gestione dei master data, un'organizzazione può diffondere dati Master coerenti e accurati nell'intera azienda.

- **Descrivi cos'è la Data Quality.**

la data quality è un processo che ha come scopo quello di ottenere un patrinomio dati con standard elevati di qualità attraverso le varie fasi di profilatura e pulizia dei dati e l'utilizzo di regole di qualità ben definite.

- **Cos'è la piattaforma AWS? Descrivila in generale.**

Amazon Web Services (AWS) è una delle piattaforma cloud più complete e utilizzate del mondo, offre più di 200 servizi completi da data center a livello globale. I clienti variano da piccole start-up a grandi aziende, incluse le agenzie governative leader di settore, in quanto gli consente di diminuire i costi, diventare più agili e innovarsi.

- **Descrivi il modello Serverless Computing.**

Il serverless computing è un modello di sviluppo cloud native che consente agli sviluppatori di creare ed eseguire applicazioni senza gestire i server.

- **Cos'è Amazon S3? Descrivine le caratteristiche e la sua architettura.**

Amazon Simple Storage Service (Amazon S3) è un servizio di

archiviazione di oggetti che offre scalabilità, disponibilità dei dati, sicurezza e prestazioni all'avanguardia nel settore. I clienti di tutte le entità e settori possono archiviare e proteggere qualsiasi quantità di dati per qualsiasi caso d'uso, come data lake, applicazioni native per il cloud e app mobili.

I dati sono archiviati come oggetti all'interno di risorse chiamate "bucket"; un singolo oggetto può avere dimensioni di massimo 5 terabyte. Le caratteristiche di S3 includono funzionalità per aggiungere tag di metadati agli oggetti, spostare e archiviare i dati tra le classi di archiviazione S3, configurare e rinforzare i controlli di accesso ai dati, proteggere i dati da utenti non autorizzati, eseguire analisi dei Big Data, monitorare i dati a livello di oggetto e di bucket e visualizzare l'utilizzo dell'archiviazione e le tendenze delle attività all'interno dell'organizzazione.

- **Come vengono gestiti gli accessi in Amazon S3?**

Per impostazione predefinita, tutti i bucket e gli oggetti Amazon S3 sono privati. Solo il proprietario della risorsa che è l'account AWS che ha creato il bucket può accedere a quel bucket. Il proprietario della risorsa può, tuttavia, scegliere di concedere autorizzazioni di accesso ad altre risorse e utenti. Un modo per farlo è scrivere una politica di accesso.

- **Cos'è un Bucket? Indica i principali passi da portare a termine per crearne uno.**

Un Amazon S3 bucket è una risorsa di storage su cloud pubblico disponibile in Simple Storage Service (S3) di Amazon Web Services (AWS). I bucket Amazon S3, che sono simili alle cartelle di file, archiviano oggetti, che consistono in dati e relativi metadati descrittivi. I passi da seguire per la creazione sono:

Accedere a AWS Management Console, cliccare su Create bucket e seguire la procedura guidata: Impostando un nome univoco, scegliendo la Regione in cui allocare fisicamente i dati tra quelle disponibili ed impostando le proprietà di controllo degli accessi ACL.

- **Cosa sono gli Oggetti in Amazon S3? Come vengono identificati?**

Amazon S3 è un negozio di oggetti che utilizza valori-chiave univoci per archiviare tutti gli oggetti che desideri. Archivi questi oggetti in uno o più bucket e ogni oggetto può avere una dimensione massima di 5 TB. Un oggetto è costituito da quanto segue: Il nome assegnato a un oggetto. Utilizzare la chiave dell'oggetto per recuperare l'oggetto.

- **A cosa serve Amazon Redshift? Descrivine l'utilizzo.**

Amazon Redshift è un sistema di gestione di database relazionali (RDBMS, Relational Database Management System) ottimizzato per analisi e creazione di report ad alte prestazioni di set di dati di dimensioni molto grandi.
Amazon Redshift si integra con vari strumenti di caricamento dei dati ed ETL (estrazione, trasformazione e caricamento) e strumenti di reportistica, data mining e analisi di business intelligence (BI).

- **Descrivi cos'è un Cluster, da quali elementi è composto e come questi funzionano.**

Un Amazon Redshift - Cluster è un un insieme strutturato di risorse di calcolo, denominate nodi.
Un cluster contiene uno o più database. I dati utente vengono archiviati nei nodi di calcolo. Il client SQL comunica con il nodo principale, che a sua volta coordina l'esecuzione di query con i nodi di calcolo.

- **A cosa servono i parametri in Amazon Redshift?**

In Amazon Redshift, associ un gruppo di parametri a ogni cluster che crei. Un gruppo di parametri è un gruppo di parametri che si applicano a tutti i database creati nel cluster. Questi parametri configurano le impostazioni del database come il timeout della query e lo stile della data.

- **Quali sono le operazioni più comuni che possono essere effettuate sui Cluster?**

Le operazioni piu comini sono:
Resize operation (Elastic resize o Classic resize) (Ad esempio in caso di notevole incremeno del volume di dati), Renaming clusters, Shutting down and deleting clusters (cioè rinomina, spegnimento dello storage e cancellazione).

- **Cos'è AWS Glue? Come funziona? Quali sono le tre componenti di AWS Glue? Descrivine l'utilizzo e l'architettura.**

AWS Glue è un servizio di integrazione dati serverless che semplifica l'individuazione, la preparazione e la combinazione dei dati per l'analisi, il machine learning e lo sviluppo di applicazioni.
Le sue componenti sono:

1. *L'**AWS Glue Data Catalog** è un archivio di metadati tecnici permanenti nel cloud AWS.*
2. ***AWS Glue crawlers and classifiers.** AWS Glue consente inoltre di configurare crawler in grado di scansionare i dati in tutti i tipi di repository, classificarli, estrarne informazioni sullo schema e archiviare automaticamente i metadati nel Catalogo dati di AWS Glue. Il Catalogo dati di AWS Glue può quindi essere utilizzato per guidare le operazioni ETL.*
3. ***AWS Glue ETL operations e jobs system**. Utilizzando i metadati nel Catalogo dati, AWS Glue può generare automaticamente script Scala o PySpark (l'API Python per Apache Spark) con estensioni AWS Glue che puoi utilizzare e modificare per eseguire varie operazioni ETL.Il sistema AWS Glue Jobs fornisce un'infrastruttura gestita per orchestrare il flusso di lavoro ETL.*

- **Quando si utilizza AWS Glue?**

I casi piu comuni di utilizzo sono:

1. *Per eseguire Query su un Data Lake Amazon S3*
2. *Analizzare i dati di registro nel data warehouse*
3. *Visualizzazione in modo unificato i tuoi dati aziendali storicizzati su più data store*

4. *creare Pipeline ETL basate sugli eventi*

- **Cos'è un AWS Glue Data Catalog? Qual è la sua importanza?**

AWS Glue Data Catalog è un metastore compatibile con Apache Hive utilizzato da AWS Glue come repository uniforme di metadati provenienti da sistemi diversi. Oltre a essere un catalogo di dati, AWS Glue Data Catalog offre anche funzionalità di audit e governance dei dati.

- **Cosa sono e come funzionano i Crawler?**

Un Crawler è un Programma che si connette a un datastore ed attraverso un elenco di classificatori ordinato per priorità, determina lo schema dei dati e crea tabelle di metadati nel AWS Glue Data Catalog.

- **Cos'è un Classificatore? Quali tipi di Classificatori conosci? Cosa c'entra con i Crawler?**

Un classificatore è un oggetto software che ha lo scopo di far apprendere al motore di Aws glue quale schema associare ai dati.
I classificatori predefiniti per i tipi di file più comuni CSV, JSON, XML, AVRO e molti altri. Esistono anche
classificatori per i più comuni sistemi di gestione di
database relazionali utilizzando una connessione
JDBC. E' possibile creare un classificatore personalizzato
utilizzando un pattern grok o specificando un tag di riga
in un documento XML.
Il Crawler utilizza un elenco di classificatori per il suo funzionamento.

- **Descrivi il concetto di Schema Similarity.**

Durante la prima esecuzione del crawler, questo legge i primi 1.000 registri o il primo megabyte di ciascun file per dedurre lo schema. La quantità di dati letti dipende dal formato del file e dalla disponibilità di un registro valido. Il crawler confronta gli schemi dedotti da tutte le sottocartelle e i file, quindi valuta se creare o meno una o più tabelle. Quando un crawler crea una tabella,

considera i seguenti fattori:

Compatibilità dei dati *per verificare se i dati sono dello stesso formato, tipo di compressione e percorso di inclusione*

Somiglianza dello schema *per verificare la somiglianza degli schemi in termini di soglia di partizione e numero di schemi diversi.*

- **Spiega l'utilità di un Crawler nel partizionamento di una tabella.**

Per sua natura un Crawler è creato al fine da massimizzare le prestazioni all'interno dei Buckets AWS S3, bilanciando i dati all'interno delle partizioni in modo equo per tutta la gerarchia.

- **Cosa sono e come funzionano i Job Bookmark?**

I Job Bookmark sono dei dati di log persistenti che riguardano i Job ETL eseguiti e contengono le informazioni di corretta esecuzione o i vari errori verificati in fase di esecuzione.

- **Elenca e descrivi due tipi di Trasformazioni Built-In di AWS Glue.**

ApplyMapping

Mappa colonne e tipi di dati di origine da un DynamicFrame a colonne e tipi di dati di destinazione in un DynamicFrame restituito. Specificando l'argomento di mappatura, che è un elenco di tuple che contengono la colonna di origine, il tipo di origine, la colonna di destinazione e il tipo di destinazione.

DropFields

Rimuove un campo da un DynamicFrame. L'output DynamicFrame contiene meno campi dell'input. Specificando quali campi rimuovere e utilizzando i vari paths argument. I vari paths argument puntano ciascuno a un campo nella struttura ad albero dello schema usando la notazione con punti. Ad esempio, per rimuovere il campo B, che è figlio del campo A nell'albero, digitare A.B per il path.

- **Cos'è un Trigger? Quali tipi di Trigger conosci? Descrivine il funzionamento.**

AWS Glue Trigger è una risorsa per Glue di Amazon Web Service, che è possibile utilizzare per avviare, manualmente o automaticamente, uno o più crawler o Jobs ETL, anche a cascata.

- **A cosa servono i notebook Apache Zeppelin?**

***Data Ingestion** (Importazione di dati in un unico punto)*
***Data Discovery** (Ricercare e scoprire informazioni)*
***Data Analytics** (analizzare e confrontare i dati)*
***Data Visualization & Collaboration** (visualizzarle e condividere i dati tra i collaboratori)*

- **Cos'è l'Endpoint di Sviluppo?**

Un endpoint di sviluppo è un ambiente viene utilizzato per sviluppare e testare gli script AWS Glue. AWS Glue consente di per creare, modificare ed eliminare gli endpoint di sviluppo.

- **Cos'è Amazon Athena? Qual è la sua utilità?**

Amazon Athena è un servizio che consente agli analisti di dati di eseguire query interattive nel servizio di storage cloud basato sul Web, Amazon Simple Storage Service (S3). Athena viene utilizzato con set di dati su larga scala. Athena è un servizio senza server. Atena è facile da usare: Basta indicare al servizio i dati in salvati Amazon S3, definire lo schema e iniziare a eseguire query utilizzando lo standard SQL.

- **Cos'è SerDe?**

SerDe è un framework per serializzare e deserializzare le strutture di dati Rust in modo efficiente e generico.
Supporta vari formati JSON, Postcard,Avro, ecc.

- **Come funziona Amazon Redshift Spectrum? Quali sono le differenze con Amazon Athena? Descrivi secondo quali criteri sceglieresti l'uno piuttosto che l'altro.**

Amazon Redshift Spectrum è una funzionalità del

servizio di data warehousing Redshift di Amazon Web Services che consente a un analista di dati di condurre analisi rapide e complesse sugli oggetti archiviati nel cloud AWS.

Le differenze tra Amazon Redshift Spectrum e Amazon Athena sono:

1. *Redshift Spectrum viene eseguito in tandem con Amazon Redshift, mentre Athena è un motore di query autonomo per eseguire query sui dati archiviati in Amazon S3*
2. *Con Redshift Spectrum si ha il controllo sul provisioning delle risorse, mentre nel caso di Athena, AWS alloca le risorse automaticamente*
3. *Le prestazioni di Redshift Spectrum dipendono dalle risorse del cluster Redshift e dall'ottimizzazione dello storage S3, mentre le prestazioni di Athena dipendono solo dall'ottimizzazione S3*
4. *Redshift Spectrum può essere più coerente in termini di prestazioni mentre le query in Athena possono essere lente durante le ore di punta poiché viene eseguito su risorse in pool*
5. *Redshift Spectrum è più adatto per eseguire query grandi e complesse, mentre Athena è più adatto per semplificare query interattive*
6. *Redshift Spectrum richiede la gestione dei cluster, mentre Athena consente un'architettura veramente serverless*

- **Descrivi l'architettura di Spectrum.**

Redshift Spectrum ha un architettura distribuita e scalabile in modo trasparente all' utente utilizzatore. Spectrum Suddivide una query utente in sottoinsiemi filtrati che vengono eseguiti contemporaneamente. Tali richieste sono distribuite su migliaia di nodi gestiti da AWS per mantenere la velocità delle query e prestazioni coerenti.

- **Cos'è e come funziona Amazon QuickSight? Quali sono gli step principali per l'utilizzo di QuickSight?**

Amazon QuickSight è un servizio di business intelligence basato sull'apprendimento automatico creato per il cloud sotto l'ombrello di Amazon Web Services. Consente alle aziende di prendere decisioni

più intelligenti basate sui fatti (... e non sulle opinioni). Altri tools di BI competitors sono: Tableau e Microsoft BI.
SPICE (Super-fast, Parallel, In-memory Calculation Engine) è il robusto motore in-memory utilizzato da QuickSight. È progettato per eseguire rapidamente calcoli avanzati e fornire dati.
Nell'edizione Enterprise, i dati archiviati in SPICE vengono crittografati a riposo.
Gli step di utilizzo sono: selezione dell'origine dei dai, definizione dei vari dataset e relative relazioni,creazione dei report di visualizzazione dati,pubblicazione dei risultati e condivisione.

- ***Cosa s'intende per motore SPICE? Descrivilo in dettaglio.***

SPICE engine è progettato, specificamente, per una visualizzazione rapida e ad hoc dei dati. SPICE archivia i dati in un sistema progettato per l'alta disponibilità, dove vengono salvati fino a si decide di eliminarli. È possibile migliorare le prestazioni dei set di dati del database importando i dati in SPICE invece di utilizzare una query diretta al database. Ha i seguenti vantaggi:

1. E' Altamente compatibile con diverse fonti di dati
2. PErmette l'accesso ai dashboard su qualsiasi browser Web o dispositivo iOS
3. Visualizzazioni di immediata comprensione e utilizzo.
4. Possibilità di utilizzare creazioni guidate code-less
5. Opzioni di pagamento convenienti
6. Scalabilità molto performante.