

GCP Notes



Google Cloud Platform

hBcunet

- To quickly redact sensitive information can be used **DATA LOSS PREVENTION**
- Quando il requisito è di ridurre i job Hadoop con il minino management il consiglio è di utilizzare **Dataproc** con un connettore GCS in caso di dati persistenti.
- Incrementando la size di una tabella **BigTable** aumentano i costi.
- Usare uno schema con row-key che contenga spesso riduce le performance
- Sequential IDs danneggia le performance
- Per avere buone performance da BigTable è necessario avere uno schema che distribuisce reads & writes su tutte le tabelle
- Il miglior modo per minimizzare gli stessi dati tra BQ e sop Hadoop & Spark è sincronizzare in GCS
- JobUser permette di eseguire e cancellare i propri job in un progetto
- User permette di eseguire query tra i vari progetti sotto un'organizzazione
- Owner permette anche di cancellare job di altri
- Viewer non permette di eseguire job
- Pub/Sub non mantiene lo storico temporale, quindi è necessario associare un timestamp nel publisher e quindi in seguito con **Dataflow**
- Per pipeline streaming è necessario usare Dataflow considerando di arricchimento e trasformazione
- Quando la velocità di rete non è abbastanza per trasferire le mole dei dati si deve usare **Google TRANSFER APPLIANCE**
- Per allargare una rete neutrale si devono utilizzare **bases & weights**
- Compute Engine is no Ops free. Se non c'è bisogno di processing o storage si usa **App Engine**
- Per un job su testo (NL) usare **CLOUD NATURAL LANGUAGE API**
 " " " " " " " " " " **CLOUD VISION API**
- Per organizzare dipendenze fra dati, creare workflow o servizi si usa **Cloud Composer**
- Per dare accesso solo a determinati dati bisogna avere **Authorized view** ed assegnare user role a tutti di progetto.
- Per ricevere audio si usa **Cloud Speech-to-Text API**. Si ricevono file audio < 2 min, assicurarsi di rimediare
- Bigquery ha supporto a Linear Regression, Binary Logistic e Multiclass Logistic.
- BigQuery è l'ideale per analisi, supporta OLAP
- Dataproc è l'ideale per **Semi-structured** e **NoSQL**. Ideale per cataloghi disponibili.
- Dataproc permette di effettuare operazioni e trasformazioni sui dati tramite **HQL**, senza conoscere linguaggi.
- Cloud Storage non fornisce alcun servizio, è solo uno storage per qualsiasi tipo di dato.
- BigQuery fornisce 2 metriche per gli slot: **Allocated** e **Available**

- Per mantenere sincronizzati due cloud storage si deve usare **gattile sync** (**Storage Transfer Service**)
- Per installare dipendenze al un job con solo permessi interni bisogna caricare le dipendenze su GCS e installarle da lì.
- **Transfer Appliance** è file per singoli trasferimenti, **Storage Transfer** può essere utilizzato per quelli periodici.
- **Catchline** è per dati per cui provvediamo un accesso comune.
- In caso di task critici con **DataProc** è consigliato usare cluster in High Availability, Standard Attenuanti. } **DATAPROC**
 - Preemptible nodes vengono usati solo se realmente necessari.
- Per questione di costi è sempre preferibile usare bucket con region.
- Il migliore modo per operazioni semplici è preparare i dati e usare **Data Prep** ed essere le **Transformazioni** consigliate per colonne.
- Data Proc è file anche per individuare anomalie nei dati.
- **Bigtable**, **Fall and Recovery** tables sono ideali per minimizzare i costi di query per timeseries su valori individuali.
- **Pub/Sub** è il sostituto nativo di Kafka. È scalabile e globale, una sola istanza può essere usata per più regioni.
- **Rehydrator** può essere usato per decrittare i dati.
- **StackDriver** è un tool di monitoring, logging, alerting e debugging.
- **Datastore** è scalabile, robusto e transazionale.
- Per spostare dati da Cloud SQL a BigQuery bisogna passare da GCS
- Partitioning o Clustering possono essere usate per migliorare le performance Che l'hanno fatto???
- Per accedere al DataProc all'esterno si può usare un NAT gateway
- Kubernetes non è l'ideale per ML!
- Per sostituire le Analytics a BigQuery bisogna usare BigQuery Transfer Service
- Online predictions vengono memorati nella retezione, non sono consigliate per predizioni complete.
- Per ricevere notifiche da StackDriver bisogna usare StackDriver API, creare un sink per esportare in Pub/Sub e riceverne
- Su Pub/Sub si generano duplicati quando viene l'Ack
- La notificazione per salvare gli indici in database è immediata, separata.
- **Dedicated Interconnect** è stile per dati sopra i 20GBPS o SLA PARTNER per infrazioni.
- **Read Replicas** permette di ridurre il carico di lavoro sul DB principale.
- **Dialogflow** permette di disegnare interface mobile, web applicativa, bot ecc...

- Linear regression può essere usata per predire.
- Reinforcement learning può essere usato per integrare modelli basati sul reward
- ML: Classification → Prevedere 1 tra N classi **S**
 Regression → Prevedere valori numerici **S**
 Clustering → Riconoscere **VNS**
 Association → Inferire associazioni fra pattern **VNS**
 Structured Output → Creare output complessi (image recognition, etc ecc)
 Ranking → Classificare, identificare una posizione in una scala
- **SUPERVISED:** CLASSIFICATION, REGRESSION **UNSUPERVISED:** Clustering, Association, Dimensionality reduction
- Per muovere da **ADS** a **A2 Manager** a **BQ** bisogna usare **BigQuery Transfer**
- Cloud Scheduler permette di schedolare ogni job.
- BigQuery non supporta: cambiamenti, cambi tipi, tenere register, cancellare una colonna
- In BigTable si usano tabella **full e narrow** quando - una riga per evento rende più facile le query (dimensioni) - più righe per eventi supererebbero la size.
- Le **external tables** permettono di creare tabella da esterni a BQ (es. CSV)
- BigTable garantisce dt carico di riduttive e Golde Datastore, BQ no
- Se vogliono **limiteare** le quote in base a diverse condizioni si setta le **custom quotas**.
- Per aumentare la capacità **TP** di determinare **more users**.
- Quando si migra da Redshift a **Cloud Spanner** è opportuno usare **views** come **placeholder** per le query.
- Se importando un CSV in BQ i file sono distanti e non rispettano le colonne il file non è **valid**
- Per ^{informazione} creare test su BT non usare max 200mb, esegui un **pre-test**, almeno 10 minuti.
- Se le performance di BT sono basse le righe possono avere **size large** o le istanze non hanno abbastanza dati.
- Se bisogna migliorare i tempi positive bisogna **incrementare la regularization** e **diminuire le feature parameters**
- Dropout regularization va usata quando sono numerose selezioni random da un livello di una rete neurale
- Precision è la formula per verificare quanto è accaduto un modello se la maggior parte dell'output è positivo.
- Gradient descent viene usato per trovare il minimo in una funzione
- Recall è la formula per verificare quanto è accaduto un modello se la maggior parte dell'output è negativo
- Per **hyperparameter tuning** si usano **number of nodes** e **number of hidden layers**
- Per eseguire modelli già pronti al deploy si usa **hostile machine LEARNING ENGINE**

- Google ML Dev Learning permette di creare un pre-configure per ML applications
 - Google Container Registry è un servizio che codice images
 - Google Kubernetes Engine è un servizio per deploy e scale di Docker Containers.
- Per Tensorflow è consigliato usare API
- E' possibile usare environment esterne nelle API.
 - Google supporta ricreazione snapshot direttamente.
- Per backup di database si usa gcloud con cron
- Per vedere i dati sulle dashboard usare Stackdriver.
- Per ridurre dimensione delle complesse sono Neurons e Hidden layer
 - I tipi di storage di BigTable non può essere cambiato
- Si può accedere a TAPI da DATABASE Thrift Socks
- In BigQuery si page storage, query e streaming
 - Cloud Spanner è per quando ci è Real Write in più region, l'alternativa in sola lettura è READ REPLICAS
 - MultiRegional e ColdLine permettono di creare multi-region georedundant
 - Wide model è usato per memorizzare mentre Deep model per generalizzare. Utili per long-scale classification
 - Cloud Video Intelligence può essere usato per analizzare i video
 - DFL sono limitate dalla quote. Il numero di uscite, sezioni e nesse è 200/giorni per tabella
 - Preemptible workers non possono memorizzare dati, un cluster non può avere solo preemptible.
- Customtier permette di specificare specifiche per il cluster:
- TrainingInput, masterType → tipo del master
 - " " " . workersCount → nome dei worker
 - parameterServerCount → nome dei parameter server
- Machine Learning Engine supporta Tensorflow!
- In Tensorflow se c'è un solo valore di una colonna categorial si usa categorical_column_with_vocabulary_list, categorical_with_hash_bucket & dimensioni. Anche
 - Tutti Cloud Machine Learning possono essere runnati localmente
 - A sparse vector contiene ONLY one second entries only 0 and 1.
 - Big Query può essere usato come sink batch e streaming
 - Crossed feature e Bucketization sono feature engineering

- In Dataflow Trigger determina quando specifiche chiavi e windows vengono scritte.
- Dataflow supporta i seguenti 4 tipi di trigger: onTime, onElement count e combinazione.
- Quando si crea un dataset Dataflow si devono specificare: progetto, regione, nome e zona.
- Per BigTable si possono dare i primi di visione di una sola tabella dentro all'intero Editor Table
- Il miglior modo per evitare lotlosting in BigTable è FIELD Partitioning
- Quando si esegue schema BigTable è raccomandato riduzione ~~dimensione~~ ^{remove fields} delle righe.
- Hbase shell è una cli-tools per eseguire task amministrativi su BigTable.
- Regioni per cui le query non vengono dividite in BQ sono 50 di Wildcards e 60 di funzione CURRENT (es. CURRENT_DATE())
- Per aggregare eventi distanti SLIDING-WINDOWS
- Quando la matrice di confusione dei falsi positivi si può: ordinare in "entroso" e vedere se solo ~~l'ultimo condizionante~~ ^{remove fields}
- To connect to service in DataProc you can use SSH 4 times.
- COLDLINE → al massimo un'ora / NEARLINE → poche volte all'anno / 1 di mese
- Quando in AdMob abbiamo i overflow bisogna prendere le immagini da più angoli.
- Quando Scanner ha performance buone cause possono essere: UUID come chiavi, generare cache
- È possibile impostare CloudSQL a Blazquez
- BigTable può passare da Development a Production.
- In BQ la cache non funziona se si usano wildcards o se in uno job due usano le tabelle
- Storage Transfer permette di trasferire dati ONLINE in GCS, applicare OFFLINE
- Per determinare quando aumentare la size in BigTable bisogna monitorare la latenza delle operazioni di write (write latency)
- monitorare l'utilizzo storage ed aumentare quando l'utilizzo supera il 70%
- Dataflow non richiede alcuna conoscenza di linguaggi SQL.
- gestisce tutte uguali una singola (directory,bucket) di una destinazione.
- Dataflow può essere funzionato come Dataflow Template
- In PUBSUB se un messaggio supera i 10ms non può essere pubblicato.
- Diabolflow è uso per analisi vocali per informazioni umane
- Denormalization aiuta a velocizzare le query. UPDATE no buono in SQL
di può applicare comando 2 tabella ad attributo REPLICATED.

- Con multi-region si ottiene una alta disponibilità.
- Con un sharding bisogna migliorare il modello. Con partitioning problem bit min.
- In Dataflow:
 - Local Execution file per reading a fast
 - Usare flatten per avere un'unica operazione di write
 - Se ci sono molti file dataset si usa CoGroupByKey
- Per evitare hotspot in Spanner si dovrebbe usare una chiave random.
- Usare Spanner quando bisogna salvare per molti dati, classificare attivati.
- In BigQuery si può esportare in CSV, JSON, AVRO (non senza header)
- Se bisogna dare accesso solo a N colonne usare authorized views
- Non si può usare la web interface - file > 10mb, più file per volta, ~~500~~ file.
- Per traffico web fare download in browser si può usare SSW con ^{socket} protocol.