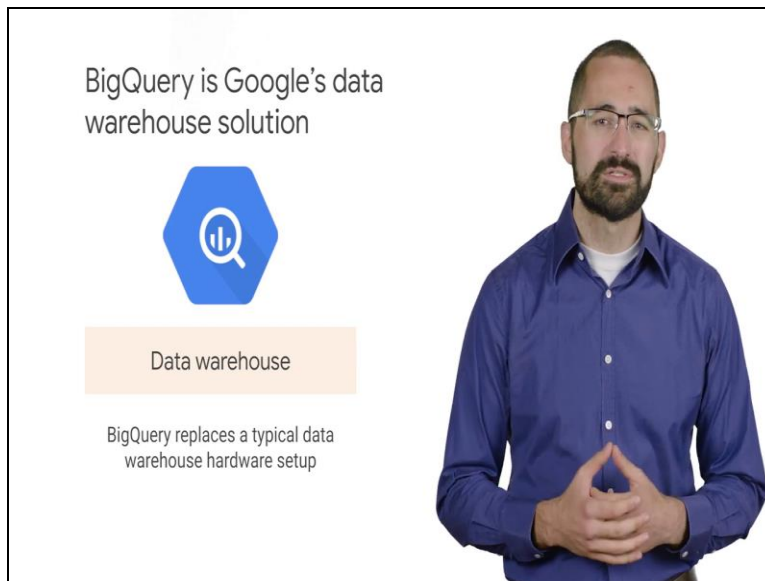


Google Cloud Computing Foundation Course
Evan Jones
Technical Curriculum Developer
Google Cloud

Lecture-73
BigQuery Googles Enterprise Data Warehouse

In this last topic you learned about bigquery. Bigqueries are fully managed petabyte scale low-cost analytics data warehouse. Bigquery serverless there is no infrastructure to manage and you do not need a database administrator. Bigquery is a powerful big data analytics platform used by all types of organizations from startups to Fortune 500 companies. A short animated video follows that introduces bigquery and how it helps to handle the complexity of today's data.

(Refer Slide Time: 00:31)



The bigquery service replaces the typical hardware setup for a traditional data warehouse that is it serves as a collective home for all your analytical data inside of your organization.

(Refer Slide Time: 00:43)

BigQuery is Google's data warehouse solution



Data mart


BigQuery organizes data tables into units called datasets



Data sets are collections of tables, views and now even machine learning models that can be divided along business lines or a given analytical domain. Each data set is tied to a GCP project.


(Refer Slide Time: 00:56)

BigQuery is Google's data warehouse solution



Data lake

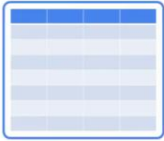
BigQuery defines schemas and issues queries directly on external data sources



A data Lake might contain files and Google Cloud storage or Google Drive or transactional data and cloud BigTable, Bigquery can define a schema and issue queries directly against these external data sources called federated queries.


(Refer Slide Time: 01:12)

BigQuery is Google's data warehouse solution



Tables and views


Function the same way as in a traditional data warehouse



Database tables and views function the same way in bigquery as they do in a traditional data warehouse allowing bigquery to support queries that are written in a standard sequel dialect that's called ANSI 2011 compliance.


(Refer Slide Time: 01:26)

BigQuery is Google's data warehouse solution



Grants

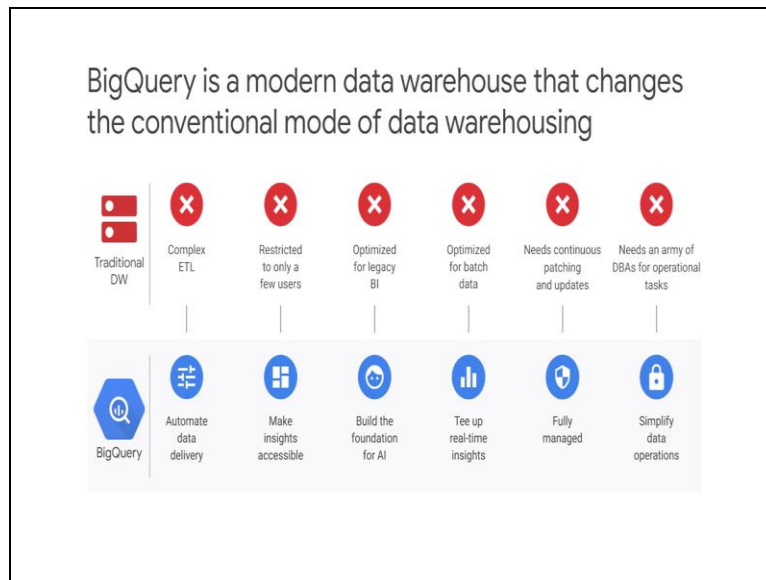
Cloud IAM grants permission to perform specific actions



Cloud Identity and Access Management is used to grant permission to perform specific actions inside a big query. This replaces; the sequel grant and revoke statements that you might have seen before to manage access permissions and traditional sequel databases. Traditional data warehouses are hard to manage and operate. They were designed for a batch paradigm and data analytics for operational reporting needs. The data in the data warehouse was meant to only be used by a few management folks for just reporting purposes.

Bigquery by contrast is a modern data warehouse that changes the conventional mode of data warehousing. Let us look at some of those key comparisons between the traditional data warehouse and what you get with bigquery.

(Refer Slide Time: 02:12)



Bigquery provides mechanism for the automated data transfer and powers applications that your team's already know and use so that everyone has access to data insights. You can create read only shared data sources that both internal and external users can query and then make those query results accessible to anyone through user friendly tools such as Google sheets, looker, tableau, click or Google Data studio.

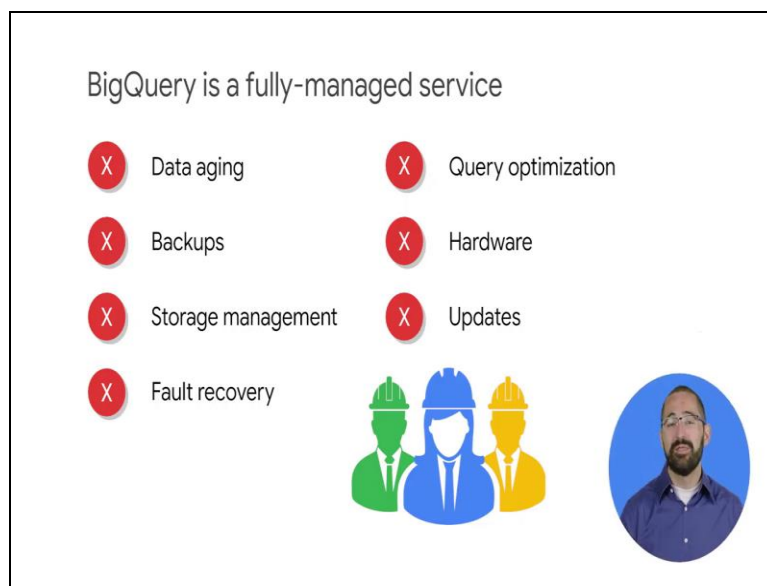
Bigquery lays the foundation for AI it is possible to train tensorflow and Google Cloud machine learning models directly with datasets stored in bigquery. And bigquery ML can be used to build and train machine learning models with using just sequel it is my favorite feature. Another extended capability is bigquery GIS which allows organizations to analyze Geographic data in bigquery essential to many critical business decisions that revolve around location data.

Bigquery also allows organizations to analyze business events in real time by automatically ingesting data and making it immediately available to query inside of their data warehouse. This is supported by the ability of bigquery to ingest up to 100,000 rows of data per second as in this

recording and for petabytes of data to be queried at lightning-fast speeds. Due to our fully managed serverless infrastructure and globally available Network bigquery eliminates the work associated with provisioning and maintaining a traditional data warehousing infrastructure.

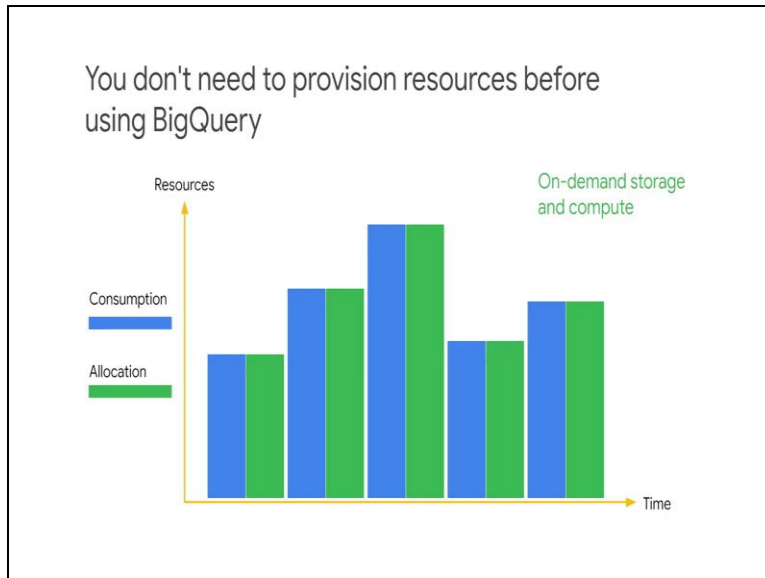
Bigquery also simplifies data operations through the use of Identity and Access Management or IAM to control users access to your resources. By creating roles and groups and assigning permissions for running those bigquery jobs and queries in a project and also providing automatic data backup and replications.

(Refer Slide Time: 04:09)



Bigquery is a fully managed service which means that the bigquery engineering team here at Google takes care of all the updates and the maintenance upgrades should not require downtime or hinder a system performance. This frees up real people hours for not having to worry about these common maintenance tasks.

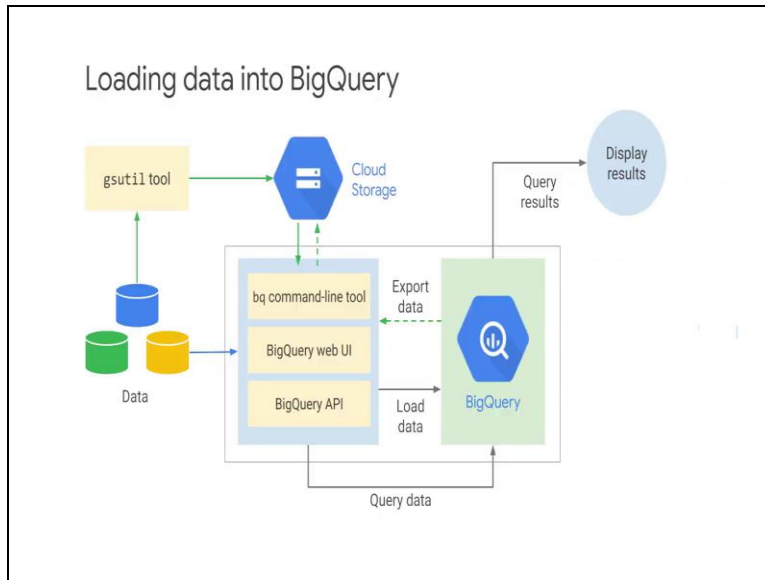
(Refer Slide Time: 04:26)



Users do not need a provision resource before using bigquery unlike many RDBMS systems bigquery allocates storage and query resources dynamically based on your usage patterns. Storage resources are allocated as users consume them and then D allocated as they remove data or drop those tables. Query resources are allocated according to the query type and the complexity of that sequel. Each query uses a number of what are called slots their units of computation that comprise a certain amount of CPU and RAM.

Users do not have to make a minimum usage commitment to you as Qigquery the service allocates and charges her resources based on their actual usage. By default all bigquery users have access to 2,000 slots for query operations they can also reserve a number of fixed slots for their project if you want.

(Refer Slide Time: 05:21)



Well there are situations where you can query data without loading it for example when using a public or shared dataset, stackdriver log files are those external data sources. For all other situations you must first load your data into bigquery before we can run your queries. In most cases you load data into bigquery native storage and if you want to get data back out a big query you can then export the data.

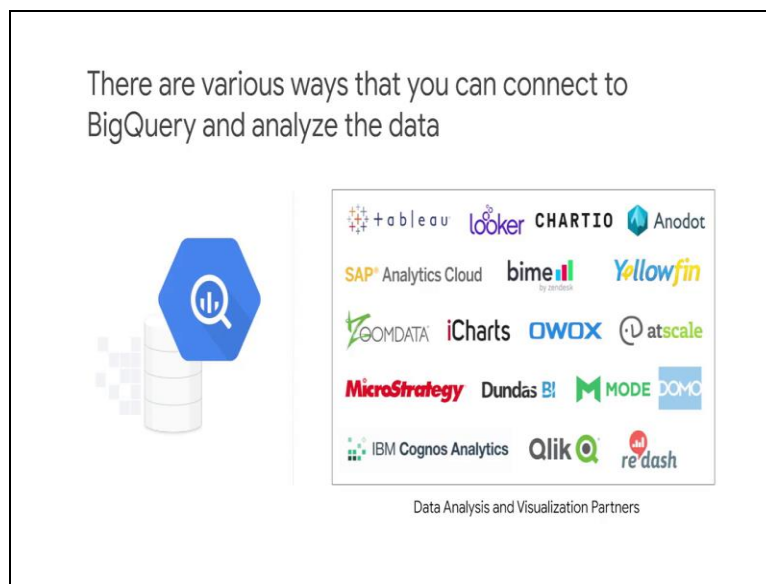
The gsutil tool is a Python application that lets you access cloud storage from the command line you can use gsutil to do a wide range of bucket and object management tasks including uploading downloading and then deleting those objects. The officially supported installation update method for gsutil is to do so as part of the Google Cloud SDK. The bigquery command line tool is another Python based command line tool and it is also installed through the SDK.

The BQ command line tool serves many functions within bigquery but for loading it is good for large data files scheduled uploads creating those tables defining schemas and loading data with one single command. You can use the Bigquery web interface in the GCP console as a visual way to complete various tasks including low exporting data as well as running your queries. The bigquery API allows a wide range of services such as cloud dataflow and cloud dataproc like we talked about earlier to load or extract data to and from bigquery.

The bigquery data transfer service for cloud storage allows you to schedule recurring data loads from cloud storage to bigquery. It also automates data movement from a range of software-as-a-service applications to bigquery on a scheduled and managed basis. The bigquery data transfer service is accessible through the GCP console the bigquery web UI the BQ command-line tool or the bigquery data transfer services API.

Another alternative to loading data is to stream the data one record at a time. Streaming is typically used when you need the data to be immediately available such as a fraud detection system or a monitoring system. While low jobs are free in bigquery there is a charge for streaming data therefore it is important to use streaming in situations where the benefits outweigh the costs.

(Refer Slide Time: 07:52)



To take full advantage of bigquery as an analytical engine you should store your data inside a big queries native storage. However your specific use case might benefit from analyzing external sources either by themselves or join together within bigquery storage. Google Data studio as well as many partner tools that are already integrated with bigquery can be used to draw analytics from bigquery and build sophisticated interactive data visualizations and dashboards for your teams.