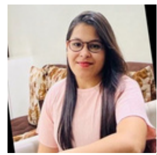
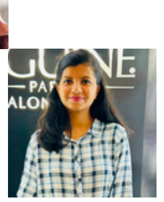
Google Cloud Platform

BIG QUERY INTERVIEW QUESTIONS AND ANSWERS

POOJA JAIN



DIKSHA CHOURASIYA

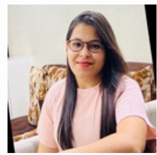


BigQuery

EASY

- What is BigQuery and What is the purpose of using BigQuery ?
- What is the Architecture of BigQuery?
- What are the different types of file formats supported in BigQuery?
- What is Materialized views in BigQuery?
- What are the Optimation techniques in BigQuery?
- What are the different ways of loading data in BigQuery Table?
- What is the difference between Row level and columnar base Datawarehouse?
- What is the Difference between Redshift and BigQuery?
- What is the Difference between Hive and BigQuery?
- What are the advantages and limitations of using BigQuery?

POOJA JAIN



DIKSHA CHOURASIYA

BigQuery

WHAT IS BIGQUERY AND WHAT IS THE PURPOSE OF USING BIGQUERY ?

Bigquery is one of the cloud based datawarehouse solution.It is serverless,fully managed, scalable,cost effective datawarehouse.

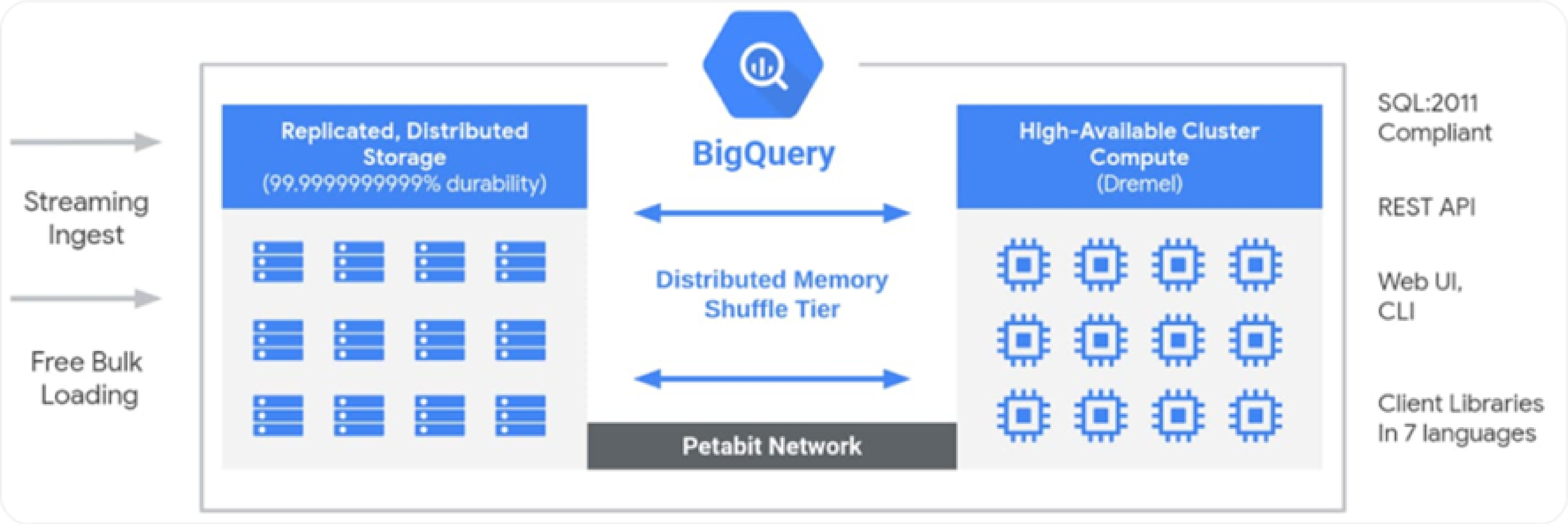
- Bigquery supports both Batch and streaming data Ingestion
- Bigquery provides on-demand pricing model i.e You only have to pay for the usage of Bigquery (pay as you go)
- It caches the same queries so that you do not have to pay for the same query again (one thing to be keep in mind that the query should be exactly the same then only you can leverage the feature of caching)
- It supports easy transfer of data from different sources like S3 Amazon bucket,Teradata etc.
- We can provide access for read/write jobs in Bigquery using IAM service.
- Bigquery automatically backup the data and keeps 7 days of history data.
- We can also manage Monitoring,Logging, Alerting of Bigquery dataset using Cloud Audit service of GCP.
- We can also collect logs of other various GCP services to Bigquery and can be used for analysis.
- We can also leverage fedrated queries in Bigquery (Federated queries are those in which we can process the data without actually storing them in Bigquery)
- Bigquery supports AI and ML services to be integrated along.

BigQuery

WHAT IS THE ARCHITECTURE OF BIGQUERY?

BigQuery is GCP’s serverless, highly scalable, and cost effective cloud data warehouse. It allows for super-fast queries at petabyte scale using the processing power of Google’s infrastructure.

1. Colossus
2. Dremel
3. Jupiter Network
4. Borg


Streaming Ingest
Free Bulk Loading
Replicated, Distributed Storage (99.9999999999% durability)
BigQuery
Distributed Memory Shuffle Tier
Petabit Network
High-Available Cluster Compute (Dremel)
SQL:2011 Compliant
REST API
Web UI, CLI
Client Libraries In 7 languages

BigQuery

1. **Colossus**

Storage is Colossus, Google’s global storage system.

- Colossus also handles replication, recovery (when disks crash) and distributed management (so there is no single point of failure).
- Colossus allows BigQuery users to scale to dozens of petabytes of data stored seamlessly

2. **Dremel**

Compute is Dremel, a large multi-tenant cluster that executes SQL queries.

- Dremel dynamically apportions slots to queries on an as-needed basis
- Maintaining fairness for concurrent queries from multiple users. A single user can get thousands of slots to run their queries

3. **Jupiter Network.**

Compute and storage talk to each other through the petabit Jupiter network.

- In between storage and compute is ‘shuffle’, which takes advantage of Google’s Jupiter network to move data extremely rapidly from one place to another.

4. **Borg**

BigQuery is orchestrated via Borg, Google’s precursor to Kubernetes.

- The mixers and slots are all run by Borg, which allocates hardware resources.

BigQuery

WHAT ARE DIFFERENT TYPES OF FILE FORMAT SUPPORTED IN BIGQUERY?

The records can be in Avro, CSV, JSON, ORC, or Parquet format.

WHAT ARE THE DIFFERENT WAYS OF LOADING DATA IN BIGQUERY TABLE?

1. Batch load a set of data records.

   a.Load Job

   b.SQL

   c.Bigquery Data Transfer Service

2. Stream individual records or batches of records. Options for streaming in BigQuery include the following:

   a. pub-sub

   b. dataflow

   c. Storage write API

.

BigQuery

WHAT IS MATERIALIZED VIEW?

"Materialized view is a snapshot of the data from a query result stored on the disk"

- They store the result physically on disk and thus occupy the space as it stores the data.
- Materialized views are pre-computed and does not refer to the base table every time we query the table.
- The data stored in the Materialized view can be manually updated or can be updated with an AUTO refresh mechanism.
- Materialized views inherits the expiration time from its base table.
- Materialized views are comparatively faster as it consists pre-computed data (There is no processing on the main base table)
- Materialized view are mostly used for the Optimization purpose.
- Materialized view should be present in the same dataset not like Normal view in different dataset.
- Materialized view cannot be used with JOINS.
- Materialized view can be PARTITIONED and CLUSTERED if the base table possess the same feature of Partitioning/Clustering

WHAT ARE THE OPTIMATION TECHNIQUES IN BIGQUERY?

In Bigquery you can optimise your data in two ways:

1. Partitioning
2. Clustering

Both Partioning and Clustering scans the only data we need to save time and money

1. Partitioning

- Partioning basically breaks down the big table into small sets of Partition
- Queries will run much faster and cheaper
- Bigquery will run subset of the data (only specified column/range's data will be scanned instead of whole scanning of the table)

There are 3 Ways to create Partition in BigQuery

- Partition creates at the time of Ingesting data
- Partition creates on the basis of DATE
- Partition creates on the basis of INTEGER

BigQuery

WHAT ARE THE OPTIMATION TECHNIQUES IN BIGQUERY?

1.Partition creates at the time of Ingesting data.
In this the Partition will be created at Run time.Ingestthe data at the time of arrival of the data

2.Partition creates on the basis of DATE, Specified on the basis of DAY

3.Partition creates on the basis of INTEGER
We have to explicitly give the start and end and interval of the integer value

Points to Remember

1.If a Null value is Present inside the data then NULL Partition will be created.

2.If unmatched value is PReset or some out of range value is present inside the data then _Unmatched Partition will be created

WHAT ARE THE OPTIMATION TECHNIQUES IN BIGQUERY?

Clustering

- Clustering is a efficient way to optimize the data
- Clustering is a technique in which similar data puts together to make the query more effficient
- Clustering can be done on more than one column
- Clustering should be done on the high cardinality column
- (High cardinality means more number of distinct values whereas low cardinality means many repeated values)
- Cluster contains similar kind of data

BigQuery

WHAT ARE DIFFERENT WAYS OF LOADING DATA IN BIGQUERY TABLE?

Following are the ways to load the data in BigQuery.

- Batch load
- Stream Load
- Loading through Existing generated Data.

Batch Load
load the source data into a BigQuery table in a single batch operation

Options for batch loading in BigQuery include the following:

1. Load jobs
2. SQL
3. BigQuery Data Transfer Service
4. BigQuery Storage Write API

BigQuery

WHAT ARE DIFFERENT WAYS OF LOADING DATA IN BIGQUERY TABLE?

Load Jobs :
Load data from Cloud Storage or from a local file by creating a load job. The records can be in Avro, CSV, JSON, ORC, or Parquet format.

SQL :
The SQL statement loads data from one or more files into a new or existing table. You can use the LOAD DATA statement to load Avro, CSV, JSON, ORC, or Parquet files.

BigQuery Data Transfer Service
To automate loading data from Google Software as a Service (SaaS) apps or from third-party applications and services.

BigQuery Storage Write API
The Storage Write API lets you batch-process an arbitrarily large number of records and commit them in a single atomic operation.

WHAT ARE DIFFERENT WAYS OF LOADING DATA IN BIGQUERY TABLE?

Stream Load:
With streaming, you continually send smaller batches of data in real time

1.Pub-Sub
2.Dataflow
3.Datastream

Genrated Data:
You can use SQL to generate data and store the results in BigQuery. Options for generating data include:

- Use data manipulation language (DML) statements to perform bulk inserts into an existing table or store query results in a new table.
- Use a CREATE TABLE ... AS statement to create a new table from a query result.
- Run a query and save the results to a table. You can append the results to an existing table or write to a new table

WHAT IS THE DIFFERENCE BETWEEN ROW LEVEL AND COLUMNAR BASE DATAWAREHOUSE?

A data store is a place for storing collections of data, such as a database, a file system, or a directory. In a Database system,
they can be stored in two ways. These are as follows:

1. Row-Oriented Store
2. Column-Oriented Store

The basic difference between them is that a row-oriented database stores the data in a table row by rows, whereas a column-oriented database stores the data tables by columns.

Examples of Row Oriented Store : MySQL,PostGresSQL

Examples of Column Oriented Store : Apache Cassandra, Microsoft Azure Cosmos DB

BigQuery

WHAT IS THE DIFFERENCE BETWEEN REDSHIFT AND BIGQUERY?

S.No	**Redshift**	**BigQuery**
1	Redshift is Amazon based Datawarehouse	BigQuery is a Google based Datawarehouse
2	RedShift's pricing model is extremely simple	BigQuery's pricing is much more complicated
3	RedShift costs $306 per TB per month for storage AND unlimited processing on that storage.	BigQuery costs $20 per TB per month for the storage line and $5 per TB processed on that storage line.
4	RedShift, on the other hand, is limited by the node you're running. But, that's not the only factor that goes into query performance.	BigQuery simply abstracts prices based on how much data you process, you're not locked into a specific resource when you run a query.

BigQuery

WHAT IS THE DIFFERENCE BETWEEN REDSHIFT AND BIGQUERY?

5	RedShift is great at handling everyday business processes. This means spinning a node during work hours for BI tools and interfaces. It's less expensive, has plenty of power to handle semi-complex schemas, and it's easy-to-use	BigQuery is great at handling niche business workloads that query big chunks in a small timeframe and for data scientists and ML/data mining.
6	RedShift supports standard SQL data types	BigQuery works with some standard SQL data types and a small range of sub-standard SQL. One of the biggest benefits of BigQuery is that it treats nested data classes as first-class citizens due to its Dremel capabilities
7	Encryption of data must be enabled	Encrypts data by default
8	AWS data loss prevention (DLP) service, Macie, does not support Redshift	Google Cloud DLP service supports BigQuery

WHAT ARE THE ADVANTAGES AND LIMITATIONS OF USING BIGQUERY?

Advantages

- Fully managed platform that offers high availability and geo-redundancy without requiring downtime for upgrades.
- Low storage costs combined with industry-leading performance for very large data sets
- BigQuery Omni allows you to query data from Azure, AWS, and Google Cloud Platform.
- It excels at evaluating massive data volumes and employs artificial intelligence to optimize storage automatically.

BigQuery

WHAT ARE THE ADVANTAGES AND LIMITATIONS OF USING BIGQUERY?

Limitations

- Queries that haven’t been adjusted for speed or that return a lot of redundant data can soon become expensive.
- Flat tables work best, which can make managing an enterprise data model more challenging.
- In comparison to other platforms, tooling support outside of the GCP ecosystem is frequently weak.

THANK YOU

POOJA JAIN

DIKSHA CHOURASIYA

BigQuery