**1.What is OLTP?**

OnLine Transaction Processing.

RDBMS (Relational Database Management Systems).

Contain homogeneous data, usually normalized till 3NF.

Databases fill up quickly, data needs to be sent somewhere else (Data Warehouses)

They contain current, transactional data.

Examples: Oracle, MySQL, Postgres, DB2, SQL Server, IBM Informix"

**2.What is OLAP?**

OnLine Analytical Processing

A form of decision support statement (DSS).

Generates a pre-prepared report based on data from DWH

Report can be written summary, or visuals can be sent like charts, dashboard, etc.

Examples: Sap Business Objects, IBM Cognos, MicroStrategy, QlikView, Tableau

**3.What is a Data Warehouse?**

Stores historical data from OLTP databases periodically

Stores for Reports, Queries, Analysis, Business decisions, Analytics

Data in DWH is stored within a slowly changing dimension (SCD) [ETL should cleanse and summarize data before coming in here).

Examples: Teradata (NCR), Exadata (Oracle), Vertica, Netezza, Greenplum

**4.What is the standard structure of tables in a Data Warehouse?**

Star Schema. Should have a center Fact Table which contains Key Performance Indicators (KPI) and foreign keys to Dimension tables (specific summarized data holders).

This helps reporting tools to perform reports in a fast manner without complex queries.

**5.What is a KPI (Key Performance Indicator)?**

It's a calculated value which can quickly answer business questions. An example would be: Total Sales Last Year. It is used to avoid querying the Dimension Tables when it is not needed.

**A Data Warehouse should follow which properties?**

SINTA

Subject-oriented

 Data is based around main objective of company using data

Integrated

 Data is integrated, summarized, compact

 Non-volatile

 Data will persist and won’t change. Write once, read many times.

 Time-variant

Data should have a time stamp to keep track of period of data retrieval. Data like ""Age"" will become stale.

Archive

Data is archived to remember location of data in DWH and high volumes.

**6.What is ETL?**

Extract-Transform-Load where data from the OLTP databases are extracted, transformed into either a summary of the data or the summary and an inclusion of more detailed information regarding the data, then loaded into the DWH.

Examples: Informatica, DataStage, SSIS, Abnitio, Oracle Data Integrator (ODI)

**7.What is a Staging Database?**

Before entering DWH, data is quality checked in the “Staging Area”, otherwise known as Operational Data Store (ODS)

Quality checking also known as “Data Cleansing”

Includes removing trailing spaces, checking proper value of “gender” (M or F, nothing else), filtering out unnecessary data (“Mr.” from “Mr. Ram”)

**8.What is a Data Mart?**

Data Warehouse cannot with hold the entirity of historical data. The standard maximum should be 20 years.

After 20 years, data can be split into smaller databases called DataMarts, which hold 5 years ~ 10 years data.

These periodically receive select portions of the DWH that reflect business data.

**9.What is Big Data?**

Big Data is a terminology used to describe a problem. Data has some specific properties defined by Gartner:

 Volume: How much data are companies currently holding? A lot. And where do we store it?

 Velocity: How fast can we process so much growing data?

 Variety: Data comes in different kinds, they can go from flat files to relational tables that actually have a schema.

**10.What kind of data do we have available?**

Structured Data

 Defined Format (like RDBMS)

Semi-structured Data

 XML, Email, Excel Spreadsheet

 Has an apparent pattern, enabling analysis

Quasi-structured Data

 Erratic format that can be formatted with tools

 Clickstream Data

Unstructured Data

 No inherent structure

 Text Documents, PDFs, image, videos

**11. Define  BigQuery?**

BigQuery is a big data analytics web service that runs in the cloud and is designed to process very massive read-only data collections. It is a fully managed, serverless data warehouse that enables petabyte-scale data processing. BigQuery's serverless architecture allows you to perform SQL queries to resolve your business's most pressing issues. Using BigQuery's distributed analytical engine, you may query terabytes in seconds and petabytes in minutes.

**12.Explain the architecture of BigQuery.**

BigQuery architecture consists of the majority of 4 parts. They are

Dremel- It makes creating execution trees from SQL queries much easier.

Colossus- It enables columnar storage and comes equipped with a compression mechanism, both of which are beneficial for data storage.

Jupiter- It is helpful because it improves the CPUs and storage connection.

Borg- It contributes to the regulation of error tolerance for the processing power of Dremel jobs.

**13.What are the benefits of BigQuery?**

BigQuery benefits include:

 BigQuery's Storage API makes it possible to read Spark and Beam workloads, which is a big assist for integration.

BigQuery reduces the need to rewrite the code by supporting the standard SQL Dialect.

Data can be replicated using BigQuery, and a seven-day history of changes can be kept to aid restoration and comparison.

**14.How data can be loaded into BigQuery?**

BigQuery Data Transfer Service is the tool that should be utilized for the most successful loading of data into BigQuery. we will be able to swiftly and efficiently import data into BigQuery from various sources, including other services offered by the Google Cloud Platform.BigQuery supports multiple input formats when receiving data.

BigQuery's web-based user interface is another option for transferring data files. In addition to importing data from a local file or a Google Cloud Storage bucket, the BigQuery command-line tool can do the same for a Google Cloud Datastore bucket. BigQuery's application programming interface (API) then lets you import records from numerous sources.

**15.What is BigQuery Storage?**

Data can be represented in BigQuery Storage using rows, columns, and tables, and the columnar Storage format, which is optimized for analytical queries, can be used to store the data. BigQuery Storage also assists with storing the data. It supplies comprehensive assistance for database transaction semantics (ACID). It is possible to replicate it across many sites to provide high availability.

**16. What is Sharding?**

Sharding is the process of breaking data into smaller pieces so that it can be handled and managed more quickly and easily. When working with BigQuery, sharding, which is the process of dividing the data across multiple processors, can be used to speed things up overall.

**17.How to convert a stringified array to an array?**

We can convert a stringified array to an array from a BigQuery Table:

COMMAND:

#standardSQL

WITH k AS (

  SELECT 1 AS id, '["a", "b", "c"]' AS x UNION ALL

  SELECT 2, '["x", "y"]'

)

SELECT

  id,

  ARRAY(SELECT \* FROM UNNEST(SPLIT(SUBSTR(x, 2 , LENGTH(x) - 2)))) AS x

FROM k

**18.How can I determine the BigQuery storage size for a single table?**

We can determine the BigQuery storage size for a single table:

COMMAND:

select

  sum(size\_bytes)/pow(10,9) as size

from

  <your\_dataset>.\_\_TABLES\_\_

where

  table\_id = '<your\_table>'

**19. Write down a command to fetch each user between two dates?**

We may fetch each user between two dates from a BigQuery Table:

COMMAND:

SELECT

  timestamp\_trunc(timestamp, DAY) as Day,

  user\_id,

  count(1) as Number

FROM `table`

WHERE timestamp >= '2023-06-28 00:00:00 UTC'

AND timestamp <= '2023-06-27 23:59:59 UTC'

GROUP BY 1, 2

ORDER BY Day

**20.Why we have to use Google Cloud Storage as a secondary storage layer when loading data into BigQuery?**

Google Cloud Storage is utilized as an intermediary storage layer to import data into BigQuery because of the reasonable pricing of the cloud data storage that it provides. we can significantly reduce the high expenses associated with cloud storage if you use Google Cloud Storage rather than one of the many other cloud storage providers.

**21.How to access BigQuery once it is configured?**

Once BigQuery has been configured, it can be accessed in several ways.

-Most users access BigQuery via the Google Cloud Console, a web-based administration and data analysis interface.

-We can use the BigQuery command-line tool, which lets you communicate with BigQuery via the command line and issue queries.

- BigQuery may be integrated with various third-party tools that offer additional features and capabilities.

**22.What are views in BigQuery?**

We can use the command line interface (CLI), the BigQuery online UI, or the API to accomplish this. Before developing a view, we will need first to make a dataset and can generate a view after that dataset has been created.

**23. Differentiate BigQuery and SQL?**

BigQuery is a cloud-based architecture that offers remarkable performance due to its ability to auto-scale up and down depending on the amount of data load and rapidly perform data analysis.

SQL Server employs a client-server architecture and, unless the user scales it manually, maintains a constant level of performance throughout the system.

**24.Explain the working of BigQuery columnar database?**

A database that organizes its data into columns instead of rows is known as a columnar database. BigQuery is a columnar database that contains data in columns rather than rows, like traditional relational databases do and because of this, it is an excellent choice for storing massive volumes of data and conducting queries on that data.

**25.What are the features of  BigQuery?**

The BigQuery features  includes:

BigQuery Omni: It's a multi-cloud analytics solution that's fully managed, so we can use it to perform analyses across AWS and Azure, for example.

BigQuery ML: It makes it possible for all data analysts to construct and operationalize machine learning models on structured or semi-structured data at planet size directly inside BigQuery, using simple SQL and doing so in a fraction of the time previously required.

BigQuery GIS: It is utilized to connect the serverless architecture of BigQuery with the native support for geospatial analysis, which enables you to enhance your analytics processes with location intelligence.

BigQuery BI Engine: It is used in the analyzing service created in BigQuery, enabling users to interactively examine huge and complicated datasets with a query response time of less than one second and high concurrency.

**26.How do you fix the most common SQL errors in BigQuery?**

We have to check that the Query follows the proper syntax using the Query Validator. If you try to run a query that already has errors, the attempt will fail, and the error will be logged in the Job details. The query validator will show a tick in the green box whenever there are no problems with the Query. Click the Run button to execute the Query and see the results after the checkmark appears in the green box.

**27. Explain what makes legacy SQL different from standard SQL?**

Using standard SQL to query data in BigQuery is the most up-to-date and recommended way. The SQL:2011 standard, on which it is based, provides numerous enhancements over older versions of the language. Performance enhancements, more assistance for SQL standard features, and enhanced compatibility with other SQL-based systems are only some of how this has been enhanced.

Legacy SQL is a way of querying data in BigQuery that predates the SQL:2003 standard. While traditional SQL is still supported for compatibility reasons, it is strongly recommended that you use more modern forms of the language whenever possible.

**28.What kinds of reports can be generated using BigQuery data?**

We can generate the following reports using BigQuery:

Inventory reports

Marketing reports

Sales reports

Product reports

Customer reports

Financial reports

**29.What are BigQuery's Window Functions?**

Functions that aggregate to a single value for a set of rows are called Window functions.They're helpful for computing values over a set of rows and returning just that result.

It has three different kinds of functions, such as:

Navigation function: The navigation function returns a value based on a specific location.

Numbering function: Numbering functions are used to give each row a number that is based on where it is in the window.

Analytic function: Analytic functions are used to do the math on a set of values.

**30.How can I get into the BigQuery Cloud Data Warehouse?**

The following are methods for connecting to the BigQuery Cloud Data Warehouse:

ODBC Drivers

Web User Interface

JDBC Drivers

Python Libraries

BQ Command-line Client

**31.What is a Big Query time decorator?**

Time decorators in Big Query enables access to historical data. For instance, if you accidentally deleted a table an hour ago, you may still retrieve the data from the table using a time decorator.

**32.How much information can BigQuery deal with?**

BigQuery was explicitly developed to manage massive data collections. It can store up to 10 petabytes of data and can analyze up to 100 terabytes of data every single day.

**33.How to permit sharing data and queries with the public in BigQuery?**

We can publicize your searches and data on Google BigQuery. We can implement this by launching a project, discussing it with specific people, or making it available to the general public.

**34.Is BigQuery PaaS or SAAS?**

BigQuery is a fully managed serverless data warehouse that supports scalable analysis of data sets up to several petabytes in size. A cloud computing environment that supports ANSI SQL queries.

**35.Is BigQuery an OLAP or OLTP?**

BigQuery is a solution for OLAP, which stands for online analytical processing.

BigQuery is best suited for large workloads, such as regular OLAP reporting and archiving activities. It is because query latency is significant in BigQuery. BigQuery's architecture discourages OLTP-style queries.

**36.Is BigQuery an ETL tool?**

BigQuery is not a dedicated ETL tool. While BigQuery is a powerful tool for querying and analyzing large datasets, it is not a traditional ETL (Extract, Transform, Load) tool, but can be used in conjunction with ETL tools as part of a data pipeline.

**37.Why is BigQuery faster than SQL?**

The query engine can process petabyte-scale data using standard SQL queries in seconds, and terabyte-scale data takes only minutes. BigQuery delivers tremendous efficiency with no requirement for index creation, infrastructure maintenance, or rebuilding. Because of its speed and scalability, BigQuery is well-suited for processing massive datasets.

**38.Why does BigQuery use SQL?**

BigQuery supports the Google Standard SQL dialect. If you are unfamiliar with BigQuery, you should start with Google Standard SQL because it has the most comprehensive features. For instance, features such as DDL and DML statements are only supported when using Google Standard SQL.

**39.Why BigQuery needs a schema?**

When creating an empty table or loading data into an existing one, BigQuery allows you to choose the table's schema. On the other hand, you can utilize schema auto-detection to find out what file formats can be used to store your data.

**40.What type of storage is BigQuery?**

BigQuery's data storage is a fully managed service. No need to set aside storage space or reserve a certain amount of storage capacity. When you upload data to BigQuery, the service immediately begins allocating storage space. Only the space you actually use will be charged to you.

**41.BigQuery structured or unstructured data?Discuss?**

BigQuery is intended for typical SQL queries on structured and semi-structured data. It is extremely cost-effective and highly optimized for query performance. BigQuery is a fully managed cloud service; therefore, there is no operational overhead.

**42.How many slots does BigQuery have?**

It Depends on the size and complexity of each query, BigQuery determines how many slots are needed. At each level of the query, separate units of work are carried out by BigQuery slots. For a certain step of a query, BigQuery can ask for an unlimited number of slots.

For example, If BigQuery finds that the best parallelization factor for a stage is 10, it asks 10 slots to process that stage.

**43. Your data team is building a new real-time data warehouse for a client. The client wants to use Google Big Query for performing streaming inserts. You get a unique ID and an event timestamp whenever data gets inserted in the row but it is not guaranteed that data will only be sent in once. Which clauses and functions you will use to write a query which ensures that duplicates are not included while interactively querying data?**

To ensure that duplicates are not included, use the ROW\_NUMBER window function with PARTITION BY based on unique ID WHERE row equals to1.

**44.An analytics company handles data processing for different clients. Clients use their own suite of analytics tools. Some clients have allowed direct query access via Google Big Query. You want to ensure that clients cannot see each other’s data. What steps can you perform inside Big Query to ensure the data security of clients?**

To ensure that clients could not see each other’s data, the following steps could be taken:

a. For each client, load data into a different dataset.

b. Restrict a client’s dataset such that only approved users can access their dataset

c. For further security, use the relevant identity and access management (IAM) roles for each client’s users.

**45.A client provides your company with a daily dump of data that flows into Google Cloud Storage as CSV files. How would you build a pipeline that will analyze the data stored in Google Cloud Storage in the Google Big Query when the data may contain rows which are formatted incorrectly or corrupted?**

To build a pipeline for the above scenario follow the below steps:

a. Import the data from Google Cloud Storage to the Big Query by running Google Cloud Dataflow.

b. Push the corrupted rows to another dead-letter table for analysis.

**46.You work as an analyst in an e-commerce company. You use Google Big Query to correlate the customer data with the average prices of the 40 most common products sold, including laptops, mobile phones, television, etc. After every 25 minutes, the average prices of these goods are updated. What steps you should follow to ensure that this average price data stays up to date so that you can easily combine it with other data in Big Query as cheaply as possible?**

Follow the below steps to ensure that this average price data stays up to date so that you can easily combine it with other data in Big Query as cheaply as possible:

a. Create a regional Google Cloud Storage Bucket to store and update the average price data

b. Then, use the Cloud Storage Bucket as a federated data source in Big Query.