====================**Azure Data Factory Questions**========================

1.What are major components of Azure Data Factory.

- **Integration Runtimes**: compute infrastructure to execute activities and performing data movements from different networks.
- **Linked Services:** Connection Information to any data source.
- **Data Sets:** Preparation of data to move and transform and load
- **Pipelines:** Designing of ETL process by using different activates and defining execution flow.
- **Activities:** To perform a specific operation like Copy data activity, delete activity, etc
- **Triggers:** Schedule of pipelines on specific time automatically, we need triggers.

2.What are different kind of Integration Runtimes available.

- **Self hosted**: it is the compute infrastructure that Azure Data Factory uses to provide data-integration capabilities across different network environments.
- **Auto Resolve or Azure Integration Services**: It is an offering by Microsoft Cloud for performing mission-critical integrations
- **SSIS Integration Runtime**: It is switch to the Manage tab and then switch to the Integration runtimes tab on the Connections pane to view existing integration runtimes in your data factory.

3.What is diff between Linked Service and Dataset.
A. Linked services are using to store connection information
Datasets are preparing of data using Linked services
4.What is Integration Runtime used to connect on premises Data sources.
A. Self hosted

5.What is diff between Self hosted and Azure Integration runtime
A. Self hosted Integration Runtime is for On premises or Private network data sources
Azure integration runtime is for Cloud based services

6.What is diff between Auto resolve and Azure Integration runtime.
A. Auto Resolve is Default Integration Runtime is used to connect cloud
Azure Integration runtime is Customized runtime is used to connect cloud

7.What are Activities used in u r project
A. Get Metadata--To get all list of filename in storage by specifying Childitems property
Copy Data activity---Copying of data from any source to destination
Lookup Activity--Retrieving of data from data sources supported by data factory
and passing that output as input to other activities
Ex: Retrieving watermark value while implementing incremental Loading
Stored Procedure--Executing stored procedure
Ex: We are executing stored procedure, to update watermark value in control table
For Each Loop: Looping and executing several activities
Ex: Copying of multiple files using single copy data activity
If Condition :based on condition true or false activities will be executed
Ex: If file exist run copy activity else send email notification to place file using web activity
Web Activity: We are using to send email notification by using Logic apps
Data Flows: we are using to apply transformations

8.What are Transformations used in Data Flows.

A. Derived Column--Deriving new columns by writing expressions

Lookup---Comparing two inputs finding matching non matching based on common column(left outer join)

Alter row---categorizing the rows to perform upsert operations, which row needs to insert and update and delete

Join---Joining of two inputs like inner, left, fullouter, customer

Conditional Split---base don condition data will be spitted

Ex: Male and female two outputs

Filter: filtering of data by putting filter condition

Assertions--Performing some data validations by specifying data rules

Windows--to find duplicate or [performing rank, rownumber, running total

Aggregate: performing aggregate and group by

Union: appending of data into one output

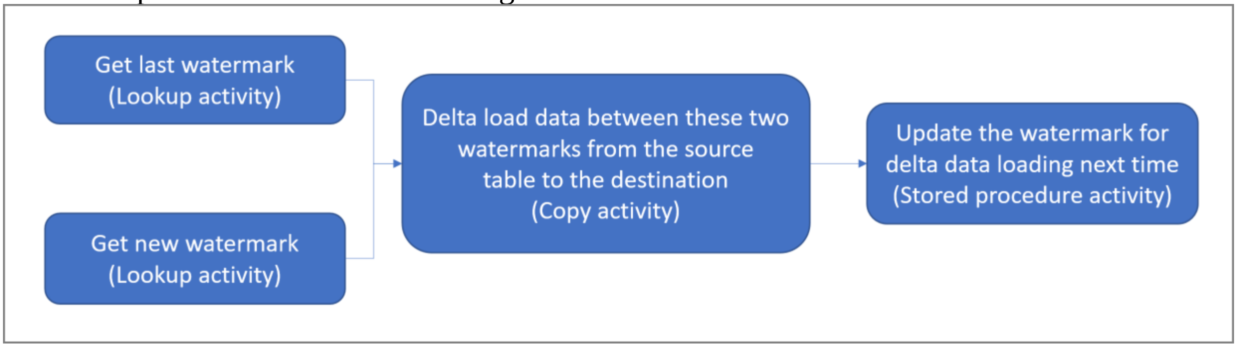
9.What is diff between Lookup and Stored Procedures activity.

A. **Lookup activity**: It can retrieve a dataset from any of the data sources supported by data factory and Synapse pipelines. You can use it to dynamically determine which objects to operate on in a subsequent activity, instead of hard coding the object name. Some object examples are files and table

Stored Procedure: You use data transformation activities in a Data Factory or Synapse pipeline to transform and process raw data into predictions and insights. The Stored Procedure Activity is one of the transformation activities that pipelines support.

10.How to implement Incremental Loading.

A.


Get last watermark
(Lookup activity)
Get new watermark
(Lookup activity)
Delta load data between these two watermarks from the source table to the destination
(Copy activity)
Update the watermark for delta data loading next time
(Stored procedure activity)

- **Select the watermark column:** Select one column in the source data store, which can be used to slice the new or updated records for every run. Normally, the data in this selected column (for example, last_modify_time or ID) keeps increasing when rows are created or updated. The maximum value in this column is used as a watermark.
- **Prepare a data store to store the watermark value:** In this tutorial, you store the watermark value in a SQL database.
- **Create a pipeline with the following workflow**:
- The pipeline in this solution has the following activities:
    - Create two Lookup activities. Use the first Lookup activity to retrieve the last watermark value. Use the second Lookup activity to retrieve the new watermark value. These watermark values are passed to the Copy activity.
    - Create a Copy activity that copies rows from the source data store with the value of the watermark column greater than the old watermark value and less than the new watermark value. Then, it copies the delta data from the source data store to Blob storage as a new file.

- Create a StoredProcedure activity that updates the watermark value for the pipeline that runs next time.

11.How to implement Slowly changing type1
12.How to load multiple files into single table
13.How to load multiple files into different tables
14.How to implement security mechanism
Azure Key vault
15.How to send email notifications
Web activity by using Logic apps
Alerts on Data factory level
Questions 11, 12, 13, 14, 15 Go to Azure Data Factory Syntax

16.What are different kind triggers scheduling

- **Schedule Trigger:** As per schedule trigger recursively executing pipelines.
- **Tumbling Window Trigger:** It is executing in periodic time interval from specified start time, historical manner, etc
- **Event-based Trigger:** It will executing pipelines based on event specified like file uploaded or deleted blob

17.How to deploy pipelines

- Create a pipeline and select the ASP.NET Core template.
- Save the pipeline and queue a build to see it in action.
- Create a release pipeline and select the Azure App Service Deployment template for your stage.
- Link the build pipeline as an artifact for this release pipeline.

18. what is difference between data lake and data warehouse
A. Data lakes and data warehouses are both widely used for storing big data, but they are not interchangeable terms. A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose.

19. What is diff between SSIS and Data Factory
A.**SSIS**:

- It is one of components in MSBI
- It is ETL Tool available in On premises
- This is not auto scale up and down compute based on workload
- Very less number of cloud integrations
- SSIS contains more number of transformations

Data Factory :

- It is cloud based ETL Tool
- It is having auto compute scale up and down
- It has more cloud integrations
- It has all options in one place, monitoring, scheduling
- comparing to ssis data factory has less number of transformations

\

20. What is the difference between Data factory v1 and Data Factory v2

A. **Data Factory V1:**

- GIT HUB Integration not available
- So many Activities and transformations are missing
- Triggers are not available for scheduling

Data factory V2:

- connection information
- dataset
- Actions or activates
- Compute Infrastructure
- workflow
- scheduling

21. What is The difference between Parameters and variables

A. **Parameter :** Passing value to Pipeline from outside at the time of running pipeline

Variable: It is Storing or assigning values internally for re using of that values in Pipeline

Set Variable---Assigning values to variable of data type as a string or Boolean

Append Variable---assigning values to variable of data type as a array list(more then one value)

22. What is a Data Flows

A. Mapping data flows are **visually designed data transformations** in Azure Data Factory. Data flows allow data engineers to develop data transformation logic without writing code. The resulting data flows are executed as activities within Azure Data Factory pipelines that use scaled-out Apache Spark clusters.

23.What is the difference between Azure Data Factory and Azure Data Bricks

A. ADF is primarily used for Data Integration services to perform ETL processes and orchestrate data movements at scale. In contrast, Data bricks provides a collaborative platform for Data Engineers and Data Scientists to perform ETL as well as build Machine Learning models under a single platform.

24.What are the Transformations used in Data Flows

A. **DERIVED COLUMN**: Deriving New Columns or adding new columns in Dataflow by writing of Expressions

Ex: Firstname,LastName column,we can derive new column Full Name by Concat both firstname,lastname

JOIN: If you want to perform all kind of join Inner join, Left outer Join, Right Outer Join..customer join in between two tables

Ex:-Customer, Customer Address join both based on Location ID

SELECT: Limiting Number of columns in Dataflow

Ex: 10 columns from source, But want to load only 5 columns, Then use select transformation Limit only required columns

FILTER: Applying filter on data coming from source

Ex: Load only city= hyderabad customers into target

CONDITIONAL SPLIT: I have one input customer data Male and female data and transgender data

I want to load above customer data into 3 tables

Male customer
Female Customer
Trans gender

UNION: Appending data from multiple inputs into one output

ex: Emphyd
EmpBang

combine all data from both data sources

WINDOWS: If you want to apply Row_Number, Rank , Dense_Rank functionalities use this windows

Ex: Finding duplicates using rownumber partition and order by and filtering using Filter Transformation

AGGREGATE: Performing Aggregate operations includes Min , Max, Sum, Avg......

Ex: Each gender wise yearly income, avg income...........

SORT: Just sorting order of data in acesending or descending based on column

Ex :Sort data based on productid

ASSERTION: Applying data validation rules on incoming data to identify unique, duplicates, condition met or not

Ex: Phone number contains 10 digits or not

LOOKUP: comparing two inputs based on common column and finding matched and non matched records(Left Outer Join)

Ex: Comparing customer source data
Destination Dimcustomer based on customerid

ALTER ROW: categorizing records to perform upsert operations it includes insert, update, delete

BRANCHING: It will take one input and giving multiple outputs of data

customer data source table contains 10 rows, i want to load same data into multiple tables

customer---dimcustomer
customer---dimcustomerbackup
customer---dimcustomercopy

EXIST: Comparing two source inputs and verifying a record from first source input to second source input a custid in customer table exist or not exist in other table dimcustomer

25. What is the difference between Mapping Dataflows and Warngling Datflows

A. Mapping Data Flows: Use Mapping Data Flows to visually *transform* data without having to write any code. You can focus on the transformations and logic, while Azure Data Factory does the heavy lifting behind the scenes. It translates your transformations and logic to code that runs on scaled-out Azure Databricks clusters for maximum performance.

A Mapping Data Flow can look something like this:
The focus in this interface is on the *flow*. You can quickly identify your sources, transformations, branches, joins, and sinks. To see the actual data, you need to enable Data Flow Debug and preview the data per transformation.

Wrangling Data Flows: Use Wrangling Data Flows to visually *explore* and *prepare* datasets using the Power Query Online mashup editor. You can focus on the modeling and logic, while Azure Data Factory does the heavy lifting behind the scenes. It translates the underlying M code to code that runs on a managed Spark environment for maximum performance.

A Wrangling Data Flow can look something like this:
The focus in this interface is on the *data*. You can quickly see what the final dataset will look like. To see the actual sources, transformations, and joins, you need to go through the list of Applied Steps.

Comparing Mapping and Wrangling Data Flows
When comparing Mapping and Wrangling Data Flows, we see that there is some overlap, but also some key differences.

Data Sources and Sinks: Mapping and Wrangling Data Flows currently both support the same sources and sinks. These include Azure Blob Storage, Azure Data Lake Storage Gen1 and Gen2, Azure SQL Database, and Azure SQL Data Warehouse.

Transformations: You can do many of the same transformations in Mapping and Wrangling Data Flows. For example filtering rows, adding and renaming columns, merging / joining datasets, grouping, and sorting.

However, if you are loading data into a database, Mapping Data Flows can also handle inserts, updates, deletes, and upserts. These row operations are managed behind the scenes, so all you have to do is enable the features and define the rules. There are also pre-defined templates available with common ETL patterns like SCD1 and SCD2.

Schema Drift: You can create Mapping Data Flows to handle schema drift if your source changes frequently. For example, if columns are added or removed, the destination can be automatically updated to include or exclude those columns. In Wrangling Data Flows, you need to make these changes manually.

File and Table Handling : Mapping Data Flows has built-in support for file handling, such as moving files after they have been read. You can also choose to recreate sink tables during execution. This means that you won’t have to manually create and execute T-SQL scripts before loading data.

================= **Azure Data Factory Syntax**==========================

UPLOAD A MULTIFILES:

- create a getmetadata activites
- one dataset in getmetadata
  FOREACH ACTIVITY:
- **@activity('Get Metadata1').output.childitems**-- in settings
- copyactivity create a new dataset and oj dataset create a new parameter and create a new name
- and in sink add a value **@item().name**

MULTIPLES FILE TO MULTIPLES TABLES:

- create one getmetadata activity
- connect to ForEach activity
- in Foreach take one Copy data activity
- create new parameters both source and sink
- source value is **item().name**
- sink value is **@concat(replace(item().name,'.txt',''))**
- because already table is exists in sink side

MULTIPLETABLES TO MULTIPLE FILES:

- create one lookup activity
- and write a query:

```
SELECT  TABLE_SCHEMA As MySchema,
 TABLE_NAME As MyTable
 FROM INFORMATION_SCHEMA.TABLES
 WHERE TABLE_TYPE = 'BASE TABLE'
```

- connect to FOREACH Activity
- create new two parameters in source one is schema and another is table
- values will **@item().MySchema,@item().Mytable**
- create new paremeter in sink
- sink value is **@concat(item().mytable)**

KEY VAULTS:

- we can keep the secrete of keys, certificates, serve name etc for that we are using a key vault.
- Server=tcp:eclasess.database.windows.net,1433;Initial Catalog=key vaults ;Persist Security Info=False; User ID=eclasessgrp;Password=Eclasess@123;MultipleActiveResultSets=False;Encrypt=True;TrustServerCertificate=False;Connection Timeout=30; --This is Available in Azure sql database in connection string
- after that a open a Key vaults copy the key in secrete and open the access polices add the Adf and save it.
- create a new linked services for key vault and attach the linked service the key vaults to azure sql database

INCREMENTAL LOADING:

First Create a One Azure Sql Data Base and create the tables are Tables, Watermarktable, Stored procedure

```
create table watermarktable
(

TableName varchar(255),
WatermarkValue datetime,
);
insert into watermarktable values('data_source_table','1990-09-05 08:06:00.000')

create procedure updatewatermarkvalue
as
  begin
  declare @LastModifytime datetime
  select @LastModifytime=max(LastModifytime) from [dbo].[data_source_table]
  update watermarktable set WatermarkValue=@LastModifytime where
 TableName='data_source_table'
  end
```

- Create one LookUp Activity
- And write a Query is:
  select WatermarkValue from watermarktable where TableName='data_source_table'
- Connect to Copydata Activity in that add a Dynamic Contect is
  select * from data_source_table where LastModifytime > '@{activity('LookupOldWaterMarkActivity').output.firstRow.WatermarkValue}'
- Connect to Stored Procedure in that dataset automatically will show a Stored Procedure file.

SEND THE MAIL:

- Create a New Getmetedata Activity in that Add a Exists.
- Connect to If Condition Activity in that there is True and False
- In If condition activity give Expressions a Query is
  @bool(activity(Getmetadata).output.exists)
- In true side keep one web activity connect to copy data activity
- In false side keep only web activity
- By using a Web activity we should have a Service is Logic apps and create a New
- While Creating a Logic app we want change a Plan Type Standard to Consumption because we will get a All services on it.
- After creating click WHEN A HTTP REQUEST IS RECEIVED and add the syntax

```
{
    "properties": {
        "DataFactoryName": {
            "type": "string"
        },
        "EmailTo": {
            "type": "string"
        },
        "ErrorMessage": {
            "type": "string"
        },
        "PipelineName": {
            "type": "string"
        },
        "Subject": {
            "type": "string"
        }
    },
    "type": "object"
}
```

- After click on the New Step in that Open the Gmail, in that Choose the any option(send gmail, reply gmail, etc)
- First give a From Mail-ID and after give To Mail-id, in that Give Subject, Body what your like it.
- In That Body add a (Error message, PipelineName, DataFactory, etc)
- After a click on Save, While clicking on save button there will create a URL and copy that
- Now open the web activity in Copy the URL
- And give the body for both the true and false side web activity is:

```
{"DataFactoryName":"yugandharemail"
,"PipelineName":"pipeline1"
,"Subject":"pipe line runned sucessfully"
,"ErrorMessage":"copy data acticivity sucessfully"
,"EmailTo":"tatanagasivaram@gmail.com"}
}
```

- Open the Copy data activity give source and sink and debug
- After completion of Debug check the mail.

=================**DATAFLOWS in ADF**=====================

Joins:

- create a one dataflow
- create two data flows one is emp and another is dept
- add a joins in joins give the details
- add a sink in sink side setting a filename option is output to single and optimize i single partition

UNION:

- create a one dataflow
- create two dataflows one is emp and another is dept
- add a union
- add a sink in sink side setting a filename option is output to single and optimize i single partition

REMOVE DUPLICATE:

- create a one dataflow
- create One databases one is Emp from source
- Add a Windows in that give Over (it is like partition) give a One column name , sort(it is like Grouping), Windows Column (create a New column) as Rank and write a expression is **rownumber()**
- After add a filter (it will filter the rows) and give the expression is Rank==1
- add a sink in sink side setting a filename option is output to single and optimize i single partition.

SLOWLY CHANGING DEMENSION TYPE-1(SCD 1):

It will show update of the existing file but not Historical values.

- create a two database one is source and other is target
- in source create table with tablename and insert values
- in target create demtable
- after go to the data factory create a data flows
- in that add a two sources attach a Lookup in give lookup conditions same
- after attach a Alter row and give a Alter row condition is update or delete or upsert
- after attach a sink and go to settings select one update method and give one column in List of columns and do the mapping manually.

SLOWLY CHANGING DEMENSION TYPE-2(SCD 2) :

It will show update and Historical values of the existing file. In this Historical values it will show a isactive is ‘**0’** and updated values isactive is **‘1’.**

- create a two database one is source and other is target
- in source create table with table name and insert values
- in target create demtable
- after go to the data factory create a data flows
- in that add a one source add a source table after attach a derived column and add a column is isactive and exp is 1
- after attach a select changes a column name (it is understanding a different from both tables) after attach a sink1
- add a another source add a target table
- add a new branch from source attach a Lookup activity after attach a filter(filtering the common row)
- after attach select , remove the source column and attach a derived column upgrade a isactive & exp is 0
- after attach alter row and give a Alter row condition is update
- after attach a sink2 and go to settings select one update method and give one column in List of columns and do the mapping manually
- after go to settings of data flows
- and keep 1 in sink2 and 2 in sink 1.

=============**AZURE DEVOPS GIT CONFIGURATION**===============

- create a one azure devops account one and add one new project
- go to REPOS in this there is a files and create branch on it
- after that open a data factory in that data factory create a new azure git configuration
- in that branch create a new pipeline and create another branch add a any dataset or activity
- after go to azure deveops create a new pull request because it is a requesting a main person to check my work
- add details like title, etc and give APPROVE AND COMPLETE in this complete disable the option is delete the task because after completion the branch will not delete

Now **CI/CD METHODLOGY**

- in here we can do two types one is a automatically and manually
- in automatic we are using a
- pipeline in this there is a release pipeline
- in this release pipeline there is a AIRCRAFT AND STAGE
- in Aircraft keep details like project name, branch name etc
- in Stage keep a Empty job because there is no ARM templates
- after stage go to task add a agent is a ARM Template Deployment and add details like display name, resources group ,azure manage,
- attach the templates and templates parameters.

If automatically not execute do **MANUALLY** go to data factory

- open the manage in there is a option is ARM Templates
- in that there is a two options one is Export and import
- First Export the file and after the import the file in the keep a ARM Template of the Azure Develops Git.

=================**AZURE DATA BRICKS QUESTIONS**===============

1.What is Spark

A. Spark it is Inmemory parallel space and it will use ram inmemory for read & write operations

2.What is Azure Data bricks

A. Azure Data bricks is a data analytics platform optimized for the Microsoft Azure cloud services platform. Azure Data bricks offers three environments for developing data intensive applications: Data bricks SQL, Data bricks Data Science & Engineering, and Data bricks Machine Learning.

3.What are components of Azure data bricks

- User --A unique individual who has access to the system.
- Group --- A collection of users
- Access control list (ACL)
- Notebook
- Dashboard
- Library
- Repo
- Experiment

4.What is diff between Map reduce and Spark

S.No.	**Map Reduce**	**Spark**
1.	It is a framework that is open-source which is used for writing data into the Hadoop Distributed File System.	It is an open-source framework used for faster data processing.
2.	It is having a very slow speed as compared to Apache Spark.	It is much faster than Map Reduce.
3.	It is unable to handle real-time processing.	It can deal with real-time processing.
4.	It is difficult to program as you required code for every process.	It is easy to program.
5.	It supports more security projects	Its security is not as good as Map Reduce and continuously working on its security issues.
6.	For performing the task, It is unable to cache in memory.	It can cache the memory data for processing its task.
7.	Its scalability is good as you can add up to n different nodes	It is having low scalability as compared to Map Reduce
8.	It actually needs other queries to perform the task.	It has Spark SQL as its very own query language.

5.What is Azure Data Bricks Runtime Integration

A. Data bricks Runtime ML is a variant of Data bricks Runtime that adds multiple popular machine learning libraries, including Tensor Flow, Keras, PyTorch, and XGBoost. Photon. Photon is the Azure Data bricks native vectorized query engine that runs SQL workloads faster and reduces your total cost per workload.

6.What is the architecture of spark

- Driver Node
- Cluster Manager
- Worker Node: In This, there is a Executor and Task

7.What are different kind of clusters in Azure Data bricks

- **Single Node:** Doesn't have any worker nodes and recommended for the single user cluster the small data volume.
- **Standard:** Recommend for a Single user cluster and can run SQL, Python, R, SCALA workload .
- **High Concurrency:** Optimized to run concurrency SQL, Python, R workloads rund and scala is not supported.

8.what are different languages supported in Data bricks

- Python or Pyspark
- Scala
- Sql
- R-Language

9.What is DAG

A. A directed acyclic graph (DAG) is a conceptual representation of a series of activities. The order of the activities is depicted by a graph, which is visually presented as a set of circles, each one representing an activity, some of which are connected by lines, which represent the flow from one activity to another.

10.What is Languages using in your project

A. Pyspark and Scala

11.What is RDD, Data frame, Data Set

- **RDD**: A Resilient Distributed Dataset (RDD), the basic abstraction in Spark. Represents an immutable, partitioned collection of elements that can be operated on in parallel.
- **Data Frame**: A Data Frame is equivalent to a relational table in Spark SQL, and can be created using various functions
- **Dataset:** A Dataset is a distributed collection of data. Dataset is a new interface added in Spark 1.6 that provides the benefits of RDDs.

12.How to convert RDD into Data Frame

A. Convert Spark RDD to Data Frame and Dataset as these provide more advantages over RDD. For instance, Data Frame is a distributed collection of data organized into named columns similar to Database tables and provides optimization and performance improvement.

Syntax:

```
val dfFromRDD1 = rdd.toDF()
dfFromRDD1.printSchema()
```

13.What is Transformation and Action

- **Transformation:** Transformation refers to the operation applied on a RDD to create new RDD. Filter, groupBy and map are the examples of transformations.
- **Actions:** Actions refer to an operation which also applies on RDD, that instructs Spark to perform computation and send the result back to driver.

14.How to create Data Frame

A. # Create DataFrame.
```
df = pd.DataFrame(data)
# Print the output.
print(df)
```

15.How to read Data from csv file to create data frame

A. usercsvdf=spark.read.csv("/FileStore/tables/userdata1.csv",header=True)

16.How to write Data into storage from Data Frame

A. csvdf=userdf.write.csv("/FileStore/tables/userdata1.csv",header=True)

17.What are different kind of joins available in Pyspark

- INNER JOIN.
- CROSS JOIN.
- LEFT OUTER JOIN.
- RIGHT OUTER JOIN.
- FULL OUTER JOIN.
- LEFT SEMI JOIN.
- LEFT ANTI JOIN.

18.How to remove duplicate values from Dataframe

A. dropdf = df.dropDuplicates(["department","salary"])
```
  print("Distinct count of department & salary :"+ str(dropdf.count()))
```

19.How to add new column into Dataframe

A. df. withcolunm("salary",df.salary+500).show()

20.How to write sql or any other langue in python Notebook

A. %Scala & % R

21.How to write sql commands on Dataframe

A. %sql
```
select * from user
```

22.How to select only few columns from dataframe

A. In PySpark we can select columns **using the select() function**. The select() function allows us to select single or multiple columns in different formats.

Syntax:
```
df.select("firstname","lastname").show()
df.select(df.firstname,df.lastname).show()
df.select(df["firstname"],df["lastname"]).show()
#By using col() function
from pyspark.sql.functions import col
df.select(col("firstname"),col("lastname")).show()
#Select columns by regular expression
df.select(df.colRegex("`^.*name*`")).show()
```

23.How to drop columns from dataframe

A. "" import col is required """
```
df.drop(col("firstname")) \
  .printSchema()
```

24.what is diff between repartition and coalesce

A. The repartition() the number of partitions can be increased/decreased, but with coalesce() the number of partitions can only be decreased and The repartition algorithm does a full shuffle of the data and creates equal sized partitions of data. coalesce combines existing partitions to avoid a full shuffle.

25.How to create mounting point to Datalake gen2

Step 1: Create a container in Azure Data Lake Gen2 Storage. Here, creating a container named blob-container. ...

Step 2: Get ADLS Gen2 Access Key. ...

Step 3: Create Secret for Access Key in Azure Key Vault. ...

Step 4: Create Mount in Azure Databricks. ...

Step 5: List Created Mount Point.

26.what is caching and persistence

- A. Caching or persistence are optimization techniques for (iterative and interactive) Spark computations. They help saving interim partial results so they can be reused in subsequent stages. These interim results as RDDs are thus kept in memory (default) or more solid storage like disk and/or replicated.
- These functions can be used to adjust the storage level of a RDD. When freeing up memory, Spark will use the storage level identifier to decide which partitions should be kept. The parameter less variants persist() and cache() are just abbreviations.

27.What is delta Lake, How to create delta lake tables

A. Delta lake is an open-source data format that provides ACID transactions, data reliability, query performance, data caching and indexing, and many other benefits. Delta lake can be thought of as an extension of existing data lakes and can be configured per the data requirements.

28.what are broad cast variables and broadcast joins

- Broadcast variables allow the programmer to keep a read-only variable cached on each machine rather than shipping a copy of it with tasks.
- Broadcast join is an important part of Spark SQL's execution engine. When used, it performs a join on two relations by first broadcasting the smaller one to all Spark executors, then evaluating the join criteria with each executor's partitions of the other relation.

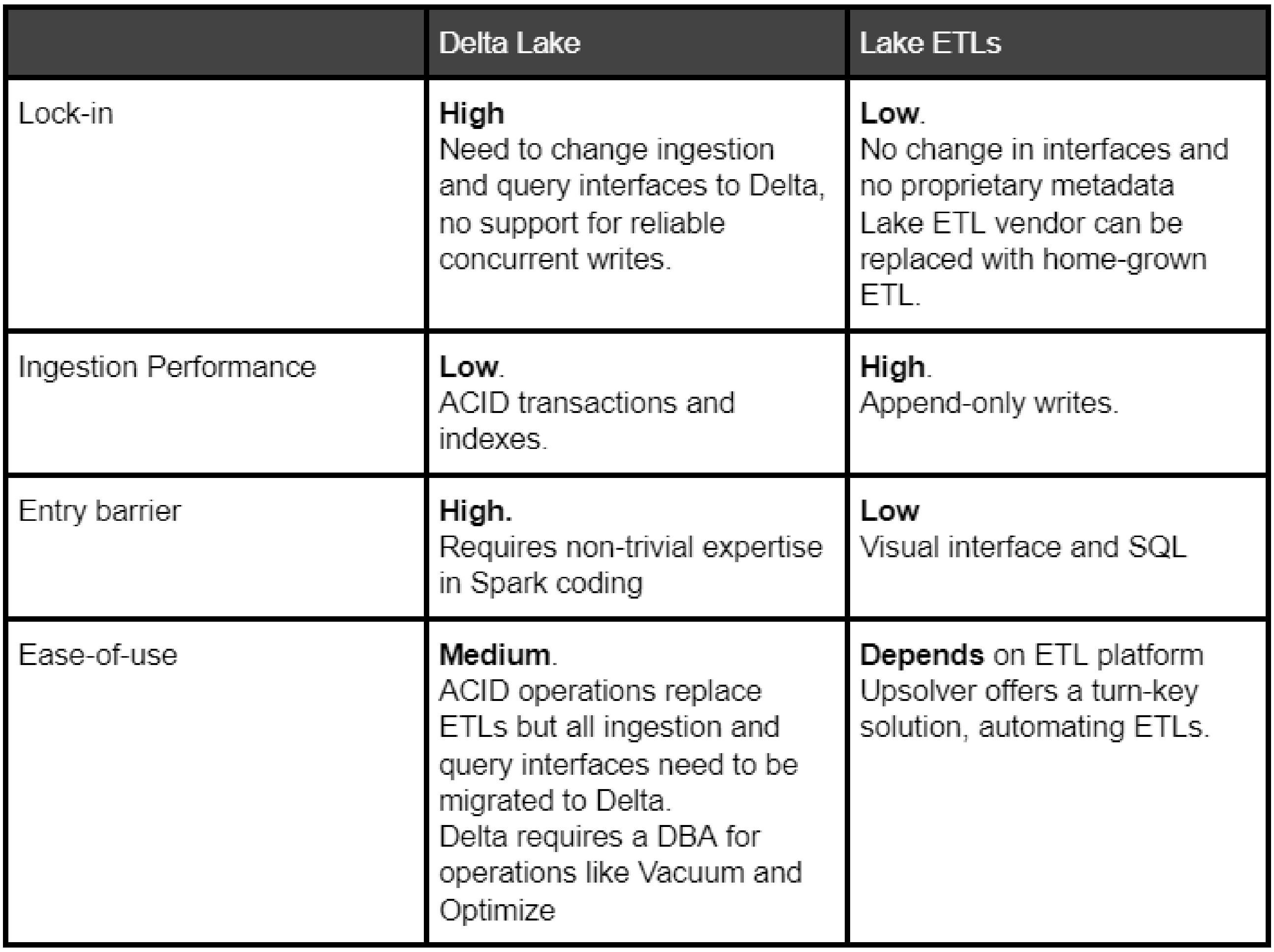
29.How to call one notebook in another notebook

A. The %run command allows you to include another notebook within a notebook. You can use %run to modularize your code, for example by putting supporting functions in a separate notebook. You can also use it to concatenate notebooks that implement the steps in an analysis.

30.How to pass parameters inside Notebook

A. If you are running a notebook from another notebook, then use dbutils. notebook. run(path = " ", args={}, timeout='120'), you can pass variables in args = {}. And you will use dbutils.

31. What is Difference Between Data Lake and Delta lake



	Delta Lake	Lake ETLs
Lock-in	**High** Need to change ingestion and query interfaces to Delta, no support for reliable concurrent writes.	**Low**. No change in interfaces and no proprietary metadata Lake ETL vendor can be replaced with home-grown ETL.
Ingestion Performance	**Low**. ACID transactions and indexes.	**High**. Append-only writes.
Entry barrier	**High.** Requires non-trivial expertise in Spark coding	**Low** Visual interface and SQL
Ease-of-use	**Medium**. ACID operations replace ETLs but all ingestion and query interfaces need to be migrated to Delta. Delta requires a DBA for operations like Vacuum and Optimize	**Depends** on ETL platform Upsolver offers a turn-key solution, automating ETLs.

=============AZURE DATA BRICKS TABLES AND SYNTAX===================

https://sparkbyexamples.com/pyspark/pyspark-withcolumn/ --- This link is reference of any commands

CREATE A DATAFRAME BY USING A SPARK

It means that Output will be visual in Table

Syntax : DF=spark.createDataFrame(data=data, schema=colunm)

Joins:

There is a Different joins like Inner, Left Join(left outer, left semi), Right Outer Join, Semi

Table Syntax:

```
emp = [(1,"Smith",-1,"2018","10","M",3000), \
   (2,"Rose",1,"2010","20","M",4000), \
   (3,"Williams",1,"2010","10","M",1000), \
   (4,"Jones",2,"2005","10","F",2000), \
   (5,"Brown",2,"2010","40","",-1), \
     (6,"Brown",2,"2010","50","",-1) \
  ]
empColumns = ["emp_id","name","superior_emp_id","year_joined", \
       "emp_dept_id","gender","salary"]
```

empDF = spark.createDataFrame(data=emp, schema = empColumns)

empDF.show() --- Creating a Data Frame

```
dept = [("Finance",10), \
   ("Marketing",20), \
   ("Sales",30), \
   ("IT",40) \
  ]
deptColumns = ["dept_name","dept_id"]
```

deptDF = spark.createDataFrame(data=dept, schema = deptColumns)

deptDF.show() ---- Creating a Data Frame and show

Syntax:

empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"Inner").show() --- it is a Inner join

empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"leftsemi").show() -- it is a Left join

<u>SELFJOIN</u>

from pyspark.sql.functions import col --- (If col is not define on that we define this is statement)

```
   empDF.alias("emp1").join(empDF.alias("emp2"), \
                col("emp1.superior_emp_id") == col("emp2.emp_id"),"inner") \
   .select (("emp1.emp_id"), ("emp1.name"), \
                  col("emp2.emp_id").alias("superid"), \
          col("emp2.name").alias("supername")).show()
```

<u>WTIH COLUMN</u>

It menas to define a New colunm and add a any expression like add, multiple etc\

Table:

```
data = [('James',",'Smith','1991-04-01','M',3000),
 ('Michael','Rose',",'2000-05-19','M',4000),
 ('Robert',",'Williams','1978-09-05','M',4000),
 ('Maria','Anne','Jones','1967-12-01','F',4000),
 ('Jen','Mary','Brown','1980-02-17','F',-1)
]
columns = ["firstname","middlename","lastname","dob","gender","salary"]
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()
df = spark.createDataFrame(data=data, schema = columns)
df.show()
```

Syntax : **df. withcolunm("salary",df.salary+500).show()**

<u>HOW TO LOAD THE IN DATA BRICKS</u>

Go to Data and click on create table and Upload the data

Syntax: **loaddf=spark.read.parquet("/FileStore/tables/userdata1.parquet")**

<u>DROP THE COLUNM</u>

Syntax: **userdf=loaddf.drop("ip_address,cc")**

CONVERT THE PARQUET TO CSV FILE

Syntax: csvdf=userdf.write.csv("/FileStore/tables/userdata1.csv",header=True) -- it will convert and store a New file

usercsvdf=spark.read.csv("/FileStore/tables/userdata1.csv",header=True) --- it will read the data and remember always give new dataframe

usercsvdf.createOrReplaceTempView("user") --- it will convert the filename into Table name

%sql --- it is important because there is a default is python, so by using a SQl Command to use (if R language %R)

select * from user

MOUNT FROM DATA LAKE GEN2 TO DATABRICKS

- create a data lake gen2 and upload a parquet file
- create a Azure Active Dictory and create a New App Registration and also create a new cerfitications and secret
- Remember three things to copy id's is cilent.id(Application id), client.secret(value of cerfitications and secret) and client.endpoint(tenant id)
- Give a API permission and add the permission is Azure blob storage, Azure data lake
- Once go back to data lake gen2 and give Access Control(IAM) and
- add Role assignment in that Give role is Storage blob data reader(for to read the data in databaricks) ,
- add memebers give the App Name (when you create a Azure Activity)
- Create New Notebook and remember select the language is Scala for understanding purpose
- syntax:
- This is all id of datalake gen2 and azure active dictory for to load the data

val configs = Map(

"fs.azure.account.auth.type" -> "OAuth",

"fs.azure.account.oauth.provider.type" -> "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",

"fs.azure.account.oauth2.client.id" -> "b5e7b223-6ccb-4aab-86b2-fd9773f11fe1", --(Application id)

"fs.azure.account.oauth2.client.secret" -> "_KI8Q~FStvRvFTPguBS~wwAVthPsxQ4zy4sxndnX", ---(value of certificate and secret)

"fs.azure.account.oauth2.client.endpoint" -> "https://login.microsoftonline.com/dee06880-2584-4c56-bf98-b4023454c861/oauth2/token"); --- (tenant id)

dbutils.fs.mount(------------This is For Mount Purpose

source = "abfss://employee@storimgdatalakr.dfs.core.windows.net/",

mountPoint = "/mnt/storingthedata",

extraConfigs = configs)

%python

df=spark.read.parquet("/mnt/storingthedata") ---it will read the data and remember always give new dataframe

%python

display(df) -- it will the display

dbutils.fs.unmount("/mnt/eclasessbatch112") ----- it will unmount the data

Delta Lake

Delta lake it is like a Data Lake but in this Any changes in existing file it will store the data .

- schema enforcements
- Time travel /versions
- update
- delete
- merge

create a Dataframe

Syntax:
df=spark.createDataFrame([(100,'krish',1000),(200,'mani',2000),(300,'siva',3000)],["eno","ename","sal"])

df.show()

df.write.format('parquet').save("/Filestore/tables/storing/newfile") --- coverting the parquet file

create another dataframe

syntax:
df1=spark.createDataFrame([(100,'krish',1000,'guntur'),(200,'mani',2000,'vijaya'),(300,'siva',3000,'ongole')],["eno","ename","sal","place"]).show()

df1.write.mode('overwrite').format('parquet').save("/Filestore/tables/storing/newfile") -- it is to overwrite the data

---But it will not happend in Delta----------

df.write.format('delta').save("/Filestore/tables/storingdelta/filedelta") --- converting the parquet file

create another dataframe

syntax:
df1=spark.createDataFrame([(100,'krish',1000,'guntur'),(200,'mani',2000,'vijaya'),(300,'siva',3000,'ongole')],["eno","ename","sal","place"])

df1.show()

df1.write.mode('overwrite').format('delta').save("/Filestore/tables/storingdelta/filedelta") – **(it will get error so change the syntax)** df1.write.mode("append").format('delta').option("mergeschema",True).save("/Filestore/tables/storingdelta/filedelta") --- it will the update the table

spark.read.format('delta').load("/Filestore/tables/storingdelta/filedelta",timestampAsof="2022-06-21 00:00:00.0").show() -- it will add the Time stamp

spark.read.format('delta').load("/Filestore/tables/storingdelta/filedelta",versionAsof=1).show() –

if version is 0, it will execute the first table(df)

if version is 1, it will execute the updated table

----In sql-------

from delta.tables import DeltaTable   --   (it will convert the file to Table)

spark.sql("CREATE TABLE dbtablenew USING DELTA LOCATION '/Filestore/tables/storingdelta/filedelta'")   -- (it will create a new table)

%sql

select * from dbtablenew

%sql   -- it will update and insert when mergering the files

MERGE INTO dbtablenew

USING dbtable

ON dbtablenew.eno = dbtable.eno

WHEN MATCHED THEN

  UPDATE SET dbtablenew.data = dbtable.data

WHEN NOT MATCHED

  THEN INSERT (eno, ename, sal) VALUES (eno, ename, sal)

=========**DATA WAREHOUSE CONCEPTS**============

- Data Ware house it is used for the ETL because fro analysis purpose.
- We want to design a OLTP system , we should a have modeling is E-R Modeling (Entity Relational Modeling) and have a Primary key and Foreign Key.
- Entity relational is nothing but a tables like customer tables, location tables, category, sub category and making a relational between two tables.
- It is an real time object and this tables we can call as Master tables.
- There is a Transaction table(sales tables, etc) in that we can see the information but we want use a joins.
- OLTP concept it is for any update or insert should be a normalized because there is tables in a Small and performance is less.
- In that OLTP want get a information is difficult because there is a different tables like customer table and Details customer table .
- We want to implement the business intelligence, so we have to apply to data because it is not
  a format data.

Master Tables: Product ID and Customer ID are Primary Keys

Product ID	**Product name**
1	Honda Bike
2	Tvs Bike

Customer ID	**Customer Name**
1	Rama
2	Siva

Transaction Table:

Product ID	**Customer ID**	**Quantity**	**Amount**
1	1	6	123455
2	2	5	7654567

- By using a ETL to loading the data OLTP system to another system, so we called a Data Ware Housing or OLAP.
- By Designing a OLAP we have modeling is Dimensional and tables we can called as Dimensional Tables(Dimproduct, dimcustomer, etc) in that we have a Bussiness key.
- It will give the information from table will get a combination of Dimcustomer and Dim detalis customer.
- In this dimensional it will check the is existing or not, if not existing it will update or insert.
- In this Dimensional table will have a Surrogate key, while inserting the data automatically it will give a Unique number and easily we can identify the data.
- There is Fact Table it is a connection of all dimensions tables.
- We can merge and also we can add a Values like Quantity, amount etc.

Dimensional Tables: Product ID and Customer ID are Business Keys and Pkey and Ckey are called a Surrogate key.

PKey	Product ID	Product name
1	1	Honda Bike
2	2	Tvs Bike

CKey	Customer ID	Customer Name
1	1	Rama
2	2	Siva

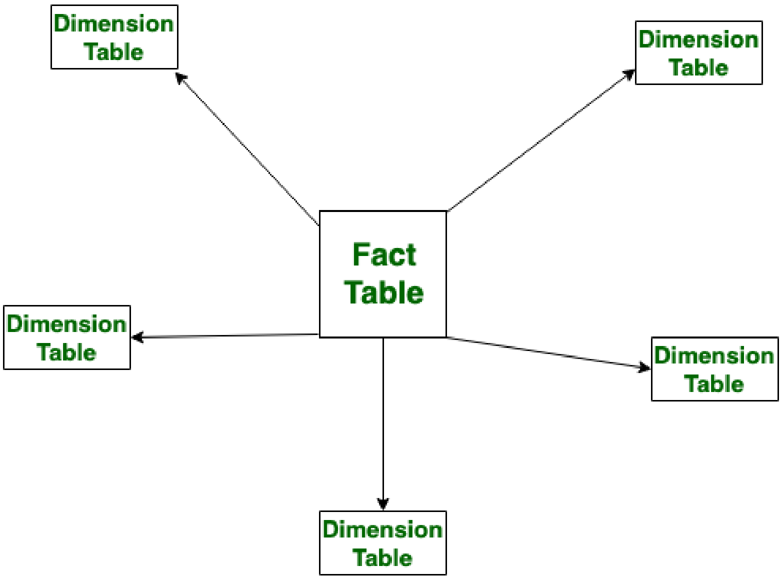
Fact Table:

Pkey	Ckey	Quantity	Amount
1	1	6	123455
2	2	5	7654567

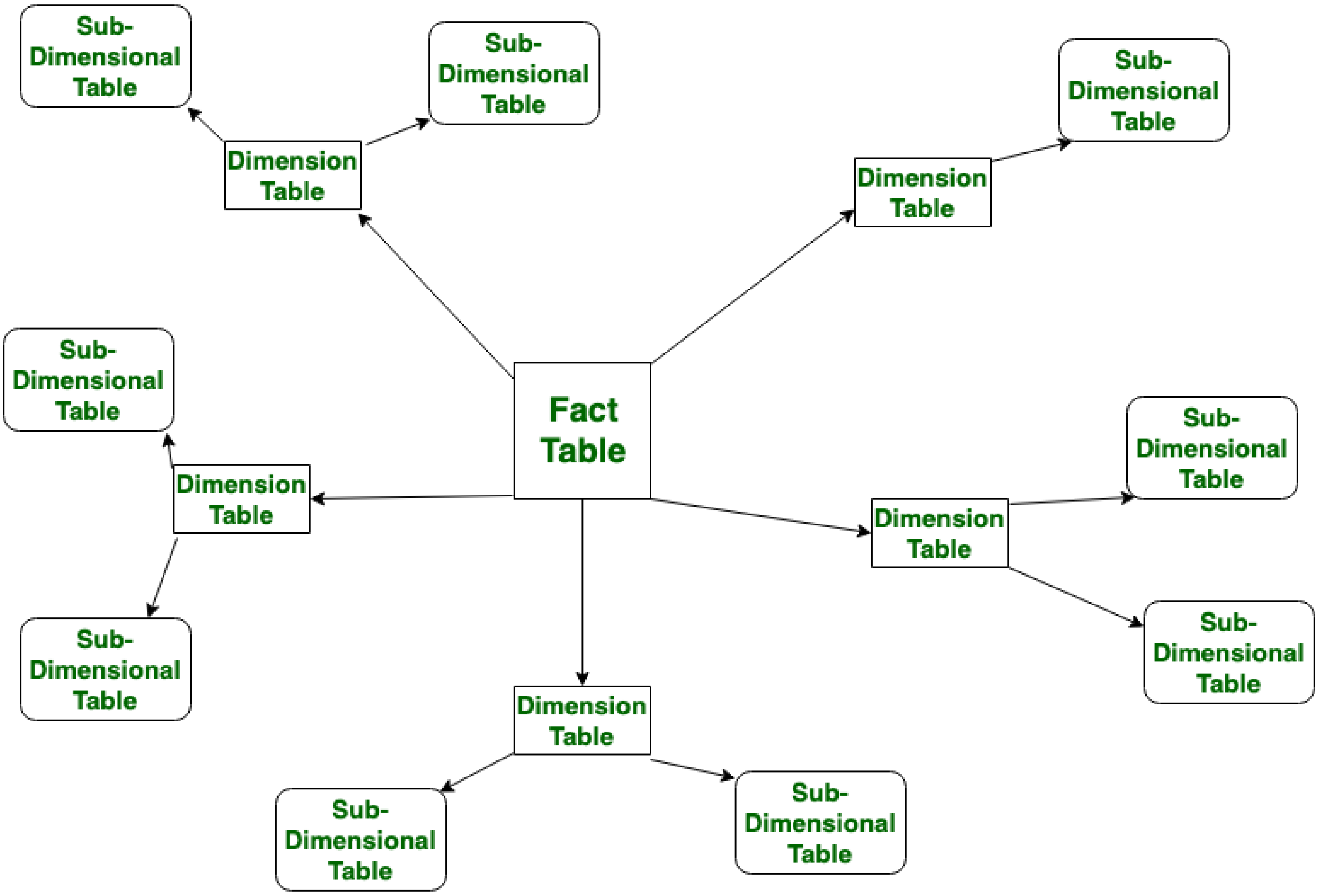
- In this Dimensional there is two schema
  1. Star Schema
  2. Snow Flake Schema

S.NO	Star Schema	Snowflake Schema
1.	In star schema, The fact tables and the dimension tables are contained.	While in snowflake schema, The fact tables, dimension tables as well as sub dimension tables are contained.
2.	Star schema is a top-down model.	While it is a bottom-up model.
3.	Star schema uses more space.	While it uses less space.
4.	It takes less time for the execution of queries.	While it takes more time than star schema for the execution of queries.
5.	In star schema, Normalization is not used.	While in this, Both normalization and denormalization are used.
6.	It's design is very simple.	While it's design is complex.
7.	The query complexity of star schema is low.	While the query complexity of snowflake schema is higher than star schema.
8.	It's understanding is very simple.	While it's understanding is difficult.
9.	It has less number of foreign keys.	While it has more number of foreign keys.
10.	It has high data redundancy.	While it has low data redundancy.

Star Schema Diagram:



Snow Flake Diagram:



==================**Azure Questions**================================

1. What is Cloud or Cloud Computing?

Cloud computing is renting & Accessing resources through internet browser, like storage space and CPU, Software's & hard wares You only pay running on another company's computers. You only payfor you use.

Or

Accessing services remotely provided by third party vendors through web browsers from datacentres.

Examples of cloud providers Microsoft (Azure), Amazon (AWS), Google (GCP).

2.What are the different kind of deployment modes available?

1. Pubic
2. Private
3. Hybrid

3.What are the different category of services available?

1. IAAS---Infra structure as a service { CPU, RAM, Hard disk, Memory….}
2. PAAS--- Platform as a service {Operating system, SQL server….}
3. SAAS---Software as a service{office 365…….}

4.What is storage account, different kind of storage accounts?

Azure storage account is cloud based storage solution to store any kind of data like structured, Unstructured & semi structured

1. Containers or blob storage
2. Tables
3. Queues or messages
4. File share

5.What is blob storage or containers what are the different types?

Blob stands for Binary large objective file It is basically general purpose storage for unstructureddata that include pictures, videos, Music files, documents, raw data, and log data…..

1. **Block blob**: You can use block blobs for documents, image files, and video file storage.
2. **Append blob**: Append blob is similar to block blobs, but more often used append operationslike logging
3. **Page blob**: Page blobs are used for objects meant for frequent read, write operations.therefore used in Azure VMs to store OS & data disks

6.What are different Access tiers available in storage Accounts?

1. **Hot tier**: It is suitable for frequently accessing data, Cost also high & Performance also more.
2. **Cool tier**: It is suitable for in frequently accessing data. Cost also less & Performance alsoless.
3. **Archive tier**: It is suitable Po backup data.

7.What are the different kind of security of Blob or storage account?

1. Connection strings
2. Access keys
3. SAS (Share Access Signature

8.What are the advantage of SAS (Share Access Signature)?

1. Providing access to only particular resources
2. Providing specific role
3. Providing only particular IP system

9.How to change Access Tiers from Hot to cool Tier if files are older than 30 Days?
By using life cycle management under storage account we can add rule

10.What is RBAC?
Role back access control
Click on Access control (IAM) then click on add role assignment

11.What is data lake and what is difference between data lake and blob storage?
Azure Data Lake is big data storage system used to store high volume, velocity, varities of data.It will store metadata hierarchal format.It is compatibility with Hadoop HDFS Frame work.

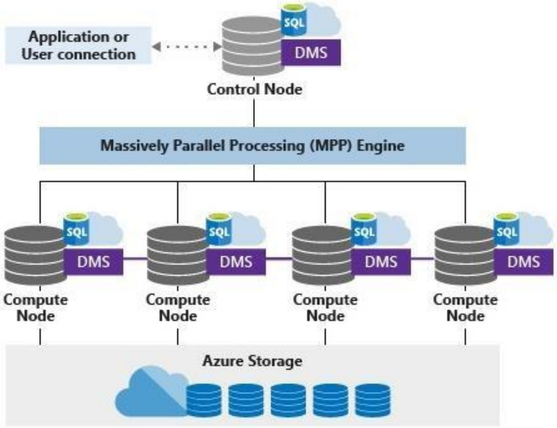
Azure Blob Storage	Azure Data Lake Storage
General purpose Storage	Optimized storage for Big data analytics purpose
Object based model with container	Hierarchical name space model
Not Compatible with Hadoop	compatible with Hadoop
Less Cost of storage	Cost is expensive
High cost for analytics	Low cost for analytics

12.What is Elastic pool, what are different compute Tiers?
Group of azure SQL data bases to share resources among all

1. DTU(Data Transaction Unit)
2. Vcore processors

13.What is Architecture of Synapse Data ware house? And what are different kind of Distributions?

Application connect issue T-SQL commands acontrol node.
The control node hosts the MPP Engine, which optimise queries for parallel processing and thenpasses operations to compute node to do their work in parallel.
The compute nodes store all users data in Azurestorage and run the parallel Queries



Control Node: It will act lick human brain it will take request from application and passing to MPP**MPP (Massive Parallel Processing)**: Splitting of Query into multiple smaller queries and execute queries parallel.

Compute node: It is place where each query will execute. 0-60 compute nodes

Distribution Method: Based on Distribution method created tables, data will be distributed acrossnodes

1. Round Robin: It will dis tribute data randomly row by row on each node. Less volume of datatables , dimension tables(customer, product)
   Disadvantage is data need to sorted at the time of combining from all nodes
2. Hash: It will distributed based on column similar values like same kind of product ID willshare one node. High volume data tables , Fact tables(Sales table)
   Advantage is no need of sorting while combining of data from nodes
3. Replication: Entire data will be distributed each node. Very small tables likes lookup tables.

14.What is Difference between Azure SQL and Synapse DWH ?

Azure SQL database	Azure Synapse DW
Insert, update, delete options mostly	Selecting data from reporting
OLTP	OLAP
Dynamic masking possible	Dynamic masking not possible
Polybase will not support to blob storage or Datalake	Polybase will support to blob storage or datalake
4TB	1PB
30,000 concurrent connections	1024
concurrent quires 6400	32
not support 3 part naming	not support 3 part naming
Not support MPP	Support MPP

15.What is difference between server less pool and dedicated SQL pool?

Dedicated pool	Server less pool
Dedicated pool needs of providing of 'DWU' at the time of creating	Server less pool no need to provide, it is auto scale up and down
Dedicated pool used storage of synapse SQL	Server less pool used to store data lake
Appling charges hourly	Server less pool charges only for executing queries

16.what is polybase what are steps needs to follow to implement polybase and advantage of polybase?

Querying the data from external data sources like blob, Data Lake into SQL synapse Data warehouseby creating external data source and external tables is called Poly base.

1. Create master key
2. Create database scoped credentials
3. Create external data source
4. Create external file format
5. Create external tables

Advantages:

- Azure blob storage is convenient place to store data for use by Azure service. Polybase it easy toAccess the data by using T-SQL.
- Polybase makes it easy to query the data by using T-SQL. Query data stored in Azure Blob storage.

1.One Database to Another Database

```
create master key encryption by password='Eclasess@456'  --Creating a New password

create database scoped credential ploybase  --Creating a New Database
with
identity='eclasessgrp',     ---This is the Username and password from Source side
secret='Eclasess@123'

create external data source ploy  -----Creating a New external Datasource
with
(
type=rdbms,
location='eclasess1.database.windows.net',   -----This is Server of the Source side
database_name='source1',
credential=ploybase)

create external table empext      -----Creating a new External tables
(
id int,
person varchar(100),
salary float)
with(
data_source=ploy,
schema_name='dbo',
object_name='emp'      ------The table from the source side
)
select * from empext
```

Ploybase from Blob storage to SQL Database

```
create master key encryption by password='eclasess@123' --- Creating a New Password

create database scoped credential ploybase   ---Creating a New Database
with
identity='avinash',     -----You can give a any name
secret='XabXvCkBrzcio24wEMxhyHkj1ZX7pK19SfUe3/Ux1xj65tr8LKA+HFSCOurjaDgT0EE1mP
KcyYlM+AStQwpk+Q=='     --------it is of secrete  of  blob storage available in the ACCESS KEY

create external data source ploy_source ---creating a external data source
with
(
type=hadoop,
location='wasbs://eclasess456@eclasess123.blob.core.windows.net',    -----(in this eclasess456 is a
Container name and eclasess123 is a Blob storage name)
credential=ploybase
)
```

```
create external file format file1  -- creating a external format
with(
format_type=delimitedtext,
format_options(field_terminator=',',first_row=2))

create external table product1
(
eno int,
enmae varchar(10),
epro varchar(10))
with(
data_source=ploy_source,
file_format=file1,
location='/product/')
select * from product1
```

17.What is OLTP?

A.OLTP or Online Transaction Processing is a type of data processing that consists of executing a number of transactions occurring concurrently—online banking, shopping, order entry, or sending text messages

18.What is OLAP?

A.Online analytical processing (OLAP) is a technology that organizes large business databases and supports complex analysis. It can be used to perform complex analytical queries without negatively affecting transactional systems.

19.What is DW?

A.Data ware house is nothing but collect the data, store the data, analysis the data and consume the data.

A data warehouse is specially designed for data analytics, which involves reading large amounts of data to understand relationships and trends across the data. A database is used to capture and store data, such as recording details of a transaction.

20.What is data mart?

A data mart is a simple form of data warehouse focused on a single subject or line of business.

21.What is ER modelling?

A.An entity–relationship model describes interrelated things of interest in a specific domain of knowledge. A basic ER model is composed of entity types and specifies relationships that can exist between entities.

22.What is master and transaction tables?

A.Transaction table: This tables gets impacted with insert or update delete we perform business transactions

Master table: This will be read and used during transactions, but this table will not get impacted for inset and update or delete.

A transaction is an activity performed by entities (master tables) within the system. These activities are captured in transaction tables and usually, these transaction entries have foreign keys to master records. Transaction tables are designed to store events in the system.

23.What is primary and foreign key?

A.A primary key is used to assure the value in the particular column is unique. The foreign key providesthe link between the two tables.

24.What is dimension modelling?

A.Dimensional Data Modelling is one of the data modelling techniques used in data warehouse design.To improve the data retrieval.
Dimensional modeling involves the use of fact and dimension tables to maintain a record ofhistorical data in data warehouses.

25.What is star schema and snowflake schema?

A.In star schema, The fact tables and the dimension tables are contained. In this schema fewer foreign-key join is used. This schema forms a star with fact table and dimension tables.
In snowflake schema, The fact tables, dimension tables as well as sub dimension tables are contained. This schema forms a snowflake with fact tables, dimension tables as well as sub-dimension tables.

26.What is surrogate key?

A.Surrogate key also called a synthetic primary key, is generated when a new record is inserted into atable automatically by a database that can be declared as the primary key of that table.

27.What are different kind of dimensions?

A.Fact means measures and dimension means description related to facts.

1. **Slowly changing dimension**: Changing over a period of time.is nothing but addressThese are three types
   Type1: need to override address,
   Type2: store all address
   Type3: need to store only current and previous address only or current previous and beforeone
2. **Conformed dimension**: A dimension table which can be used multiple fact tables is nothingbut conformed dimension.
3. **Degenerate dimension**: Fact table contain dimension table information that is called asdegenerate dimension table.
4. **Junk Dimension**: If a model has to many small dimension, we can take all this model in onedimension table is nothing but junk dimensions
   Ex: ‘True or False’, ‘Yes or No’, ‘Male and Female’, ’Single or married’
5. **Role-Plying dimension**: Fact table is more than two times relationship with dimension tableis nothing but Role-Play dimension
6. **Static dimension**: Something which never change like gender is nothing but static dimension

28.What is different kind of facts?

1. **Additive facts**: All data in the fact can be aggregate, we can apply all aggregate functions ondata example sales.
2. **Semi additive facts**: we cannot apply all aggregate function but you will able to apply someaggregate functions example current account balance.
3. **Non additive facts**: we cannot apply any aggregate function like percentage tables

29.What are the different kind of fact tables in DWH?

1. **Fact less fact table**: It does not contain any measure in it contain only foreign fromdimension table
2. **Snapshot fact table**: Store the data in relation to date or time
3. **Centipede fact table**:
4. **Conformed fact table**:

30.What is difference between ADF V1 and V2?

1. V1 doesn't contain azure devops git repository integration.
2. V1 doesn't contain many of activites.
3. V1 doesn't contains CICD methodology of deployment

=====================SQL QUESTIONS===================

1.What are dml and ddl comands?

- DML COMMANDS: INSERT,UPDATE,DELETE
- DDL COMMANDS: CREATE,DROP,ALTER,TRUNCATE

2.What is the difference between delete and truncate?

- **DELETE**: it is used to delete records from a database table. There is a possibility to rollbackthe data.
- **TRUNCATE**: it is used to remove records from a database table. Once truncate the data. itwill delete permanently.

3.What is the difference between where and having clause?

- **WHERE**: Where clause is used to filter records. where clause cannot be used in aggregatefunction. where clause used to filter the records from table or used while joining more than one table only those records will be extracted who are satisfying the specified condition in WHERE clause. It can be used with select, update, delete statement.
- **HAVING**: Having clause is used to filter the records from the groups based on the givencondition in the having clause those groups who will satisfy the given condition will appear in the final result having clause can only be used with select statement.

4.What are distinct and top key word in SQL?

- **Distinct statement**: It is used to return only distinct or different values inside a table, a columncontains many duplicate values and sometime you only want to list the different values
- **Top keyword**: Advertisements. The SQL TOP clause is used to fetch a TOP N number or X percentrecords from a table.
  Note − All the databases do not support the TOP clause.
  For example MySQL supports the LIMIT clause to fetch limited number of records while Oracle usesthe ROWNUM command to fetch a limited number of records.

5.What is sql query order of writing

Six Operations to order of Writing: SELECT, FROM, WHERE, GROUP BY, HAVING, and ORDER BY.

By using examples, we will explain the execution order of the six most common operations or piecesin an SQL query.

6.What is order of execution of SQL query?

Six Operations to Order: to order of executing:
FROM,WHERE,GROUPBY,HAVING,SELECT,ORDERBY By using examples, we will explain the execution order of the six most common operations or pieces in an SQL query. Because the database executes query components in a specific order,it's helpful for the developer to know this order

7.What are aggregate functions in SQL?

COUNT Function: COUNT function is used to Count the number of rows in a database table. ...
SUM Function: Sum function is used to calculate the sum of all selected columns. ...
AVG Function: The AVG function is used to calculate the average value of the numeric type. ...

8.What is difference between isnull, nullif, coalesce function?

ISNULL(): You can only provide one alternate value but with **COALESCE** you can provide more thanone

e.g. if col1 IS NULL then take value from column2, if that is NULL then take the default value**NULLIF()**: Function returns NULL if two expressions are equal, otherwise it returns the first expression.

9.What are constrains available in SQL?

SQL Server contains the following 6 types of constraints:

1. Not Null Constraint
2. Check Constraint
3. Default Constraint
4. Unique Constraint
5. Primary Constraint
6. Foreign Constraint

10.What is difference between the primary key and unique key?

A. Both Primary key and Unique Key are used to uniquely define of a row in a table.

Primary Key creates a clustered index of the column whereas a Unique creates an unclusteredindex of the column .

Primary Key doesn't allow NULL value, however a Unique Key does allow one NULL value

11.What is view and materilized view?

A. A view uses a query to pull data from the underlying tables. A materialized view is a table on diskthat contains the result set of a query.

Materialized views are primarily used to increase application performance when it isn't feasible ordesirable to use a standard view with indexes applied to it.

12.What are difference between table variable and temp tables, tables and cte?

A. This biggest difference is that a CTE can only be used in the current query scope whereas a temporary table or table variable can exist for the entire duration of the session allowing you to perform many different DML operations against them.

13. what is the difference between storaged procedure and and functions?

A: The function must return a value but in Stored Procedure it is optional. Even a procedure canreturn zero or n values.

Functions can have only input parameters for it whereas Procedures can have input or outputparameters.

Functions can be called from Procedure whereas Procedures cannot be called from a Function

14.What are different kind of index available and difference between those?

A. Expression-based indexes efficiently evaluate queries with the indexed Expression.

- ➢ Unique and Non-Unique indexes
- ➢ Clustered and Non- Clustered indexes
- ➢ Partitioned and Non Partitioned indexes
- ➢ Expression-Based indexes

Difference between those:

- In addition to enforcing the uniqueness of data values,
- A unique index can also be used to improve data retrieval performance during query processing.Non-unique indexes are not used to enforce constraints on the tables with which they are associated.
- In Non-Clustered index leaf nodes are not the actual data itself rather they only contains included columns
- In Clustered index leaf nodes are actual data itself
- A non partitioned index is a single index object that refers to all rows in a partitioned table. Non partitioned indexes are always created as independent index objects in a single table space,even if the table data partitions span multiple table spaces.
- When you create an index for a partitioned table, the index is a partitioned index by default unless you create one of the following types of indexes:
- A unique index where the index key does not include all of the table-partitioning columns A spatial index

15.What are different kind of join available and syntax difference between eachone?

A SQL Join statement is used to combine data or rows from two or more tables based on a common field between them.

Different types of Joins are

INNER JOIN: The INNER JOIN keyword selects all rows from both the tables as long as the condition satisfies.This keyword will create the result-set by combining all rows from both the tables where the condition satisfies

i.e value of the common field will be same.

syn: SELECT table1.column1,table1.column2,table2.column1,....
```
FROM table1
INNER JOIN
table2
ON table1.matching_column = table2.matching_column
```

LEFT JOIN: This join returns all the rows of the table on the left side of the join and matching rows for the table on the right side of join.

The rows for which there is no matching row on right side, the result-set will contain null.LEFT JOIN is also known as LEFT OUTER JOIN

Syn: SELECT table1.column1, table1.column2, table2.column1,....
```
FROM table1
LEFT JOIN
table2
ON table1.matching_column = table2.matching_column;
```

RIGHT JOIN: RIGHT JOIN is similar to LEFT JOIN. This join returns all the rows of the table on the right side of the join and matching rows for the table on the left side of join.

The rows for which there is no matching row on left side, the result-set will contain null.

RIGHT JOIN is also known as RIGHT OUTER JOIN

syn:
```
SELECT table1.column1,table1.column2,table2.column1,....
       FROM table1
       RIGHT JOIN
       table2
       ON table1.matching_column = table2.matching_column
```

FULL JOIN: FULL JOIN creates the result-set by combining result of both LEFT JOIN and RIGHT JOIN.The result-set will contain all the rows from both the tables. The rows for which there is no matching, the result-set will contain NULL values

syn:
```
SELECT table1.column1,table1.column2,table2.column1,....
      FROM table1
      FULL JOIN
      table2
      ON table1.matching_column = table2.matching_column
```

====================**SQL Commands**==============================

1. How to find Duplicate records in Employee Table

---Solution1

```
SELECT empno,ename,sal, count(*)FROM
employeeGROUP BY empno,ename,sal having
count(*)>1
```

---Solution2

```
select *, row_number() over (partition by empno order by empno) as rownumber from employee
```

2. How to remove duplicate Records in Employee Table using CTE

```
with ABC as
(
  SELECT*, row_number() over (PARTITION BY empno ORDER BY empno) as
  RowNumberFROM Employee
  )
 DELETE FROM ABC WHERE RowNumber>1
```

3. How to Find 2nd highest Salary employees in emp Table using CTE

---Solution1

```
WITH RESULT
AS (

   SELECT SAL,
        DENSE_RANK() OVER (ORDER BY SAL DESC)
   AS TOPRANKFROM EMPLOYEE
)
SELECT TOP 2 SAL FROM
RESULTWHERE TOPRANK
= 2
```

---solution2

```
WITH RESULT
AS (

   SELECT *,
        DENSE_RANK() OVER (ORDER BY SAL desc)
    AS TOPRANKFROM EMPLOYEE
)
SELECT ename,sal FROM
RESULTWHERE TOPRANK
= 2
```

4. How to find 2nd Highest salary Employees in emp Table with out using rank andcte use subquery

---Solution1

```
SELECT ename,sal FROM employee e1
WHERE 1=(SELECT COUNT (distinct sal) FROM employee e2 WHERE e2.sal > e1.sal)
```

---Solution2

```
SELECT TOP 1
salFROM (

SELECT DISTINCT TOP 2 SAL
FROM Employee
ORDER BY SAL
DESC
)  result
ORDER BY
SAL
```

---Solution3

```
SELECT MAX(Sal) From Employee WHERE Sal < ( SELECT Max(Sal) FROM Employee)
```

---Solution4 find two top salaries

```
select distinct top 2 sal from EMPLOYEE order by Sal desc
```

5. How to Find 2nd highest Salary employees by each department using CTE

```
WITH RESULT
AS(
    SELECT *,
        DENSE_RANK() OVER (partition by deptno ORDER BY SAL desc)
    AS TOPRANKFROM EMPLOYEE
)
SELECT * FROM
RESULTWHERE
TOPRANK = 2
```

6.Write query to display running Total or cumulative sum in employee

---Solution1 -- department wise cumulative sum

```
 select empno,ename,job,deptno,sal,sum(sal) over(partition by deptno order by deptno,ename)
ashighsal
 from employee
```

---Solution2---total employees wise cumulative sum

```
 select empno,ename,job,deptno,sal,sum(sal) over (order by deptno,ename) as
```

highsal from employee

7. Find Employees from employees tables, doen't belongs to any department in dept table

---Solutio1
select * from employee e right outer join dept d on e.deptno=d.deptno where e.deptno is null

---Solution2
select e.ename,dname,loc from employee e right join dept d on e.deptno=d.deptno where e.deptno is null

---Solution3
select e.deptno,ename,dname,loc from employee e left join dept d on e.deptno=d.deptno where d.deptno is null

8. Tell me how many records will come for
<u>INNER JOIN</u>
---solution1 select * from dbo.table1 as a inner join table2 b on a.id=b.id

---solution2 SELECT A.id FROM table1 A INNER JOIN Table2 B ON A.id=B.id
<u>LEFT JOIN</u>
SELECT A.id FROM table1 A left join Table2 B ON A.id=B.id

<u>RIGHT JOIN</u>

SELECT A.id FROM table1 A right join Table2 B ON A.id=B.id

<u>FULL JOIN</u>
SELECT * FROM table1 A full join Table2 B ON A.id=B.id

<u>CROSS JOIN</u>
select * from table1, table2

9. Write query to make combination of cricket match between below countries
---Solution1
select a.team+' Vs '+b.team from abc a,abc
b where a.team<b.team order by a.team
---Solution2
SELECT * FROM abc A JOIN abc B on A.team<>B.team

10. Write query to find all employee having sal greater then the average salary of each dept
select e.ename,e.sal,e.deptno,s.salavg from employee e,
(select deptno,avg(sal) salavg from employee group by
deptno)s where e.deptno=s.deptno and e.sal>s.salavg

11. Write query to find employee and manager name in below table
---Solution1
select e.eno,e.ename,m.ename managername from employee_test e,employee_test

```
m where e.mgrid=m.eno
```

---Solution2

```
select e.ename,m.ename as manager
from employee_test e
left join employee_test m
on e.mgrid=m.eno
```

https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver16&tabs=ssms ----Adventure scripts link.

12. Connecting the multiples files.

======Adventure scripts ===============

```
select * from dimproductsubcategory
select * from dimproductcategory
select * from FactInternetSales
select * from dimproduct
select * from DimProductSubcategory
```

---1.Display each category wise totalsalesamount

```
select EnglishProductCategoryName, sum(salesamount) as totalsalesamount
from DimProductCategory as c join DimProductSubcategory as s
on c.ProductCategoryKey=s.ProductCategoryKey
join DimProduct as p on p.ProductSubcategoryKey=s.ProductSubcategoryKey
join FactInternetSales as f on f.ProductKey=p.ProductKey
group by c.EnglishProductCategoryName
```

------2.Display 2nd highest sales product

```
with cte as
(
select EnglishProductCategoryName, sum(salesamount) as totalsalesamount, DENSE_RANK() over
(order by sum(salesamount) desc) as rank
from DimProductCategory as c join DimProductSubcategory as s
on c.ProductCategoryKey=s.ProductCategoryKey
join DimProduct as p on p.ProductSubcategoryKey=s.ProductSubcategoryKey
join FactInternetSales as f on f.ProductKey=p.ProductKey
group by c.EnglishProductCategoryName )
select * from cte where rank =2
```

----3.display product names dont have any sales

```
select EnglishProductCategoryName, sum(salesamount) as totalsalesamount
from DimProductCategory as c join DimProductSubcategory as s
on c.ProductCategoryKey=s.ProductCategoryKey
join DimProduct as p on p.ProductSubcategoryKey=s.ProductSubcategoryKey
left join FactInternetSales as f on f.ProductKey=p.ProductKey
group by c.EnglishProductCategoryName
having sum(salesamount) is null
```

------4.display list of products belongs to categoryname is "BIkes"

```
select p.EnglishProductName, c.EnglishProductCategoryName
from DimProductCategory as c join DimProductSubcategory as s
on c.ProductCategoryKey=s.ProductCategoryKey
join DimProduct as p on p.ProductSubcategoryKey=s.ProductSubcategoryKey
```

```
where EnglishProductCategoryName='Bikes'
```

To Know common column of the two tables

```
select name
 from sys.columns
 where object_id=object_id('dbo.tablename1')
 intersect
 select name
 from sys.columns
 where object_id = object_id('dbo.tablename2')
```

44a4bxedxPwGQtSnEzGzdejnCdtrDJqkJ6uqWwGbXmdx5zrjGdHivmr9pyd7x8rVQMgumwGPCfuwseieCg8tuN5jNNeMJPU – This is the link of Emp and Dept scripts.

13. SELECT VIEW:

```
select * from sys.databases --display all database
select * from sys.tables --display all tables
select * from cons.sys.tables --display only given database name
select * from sys.views -- display all the views
select * from sys.columns --display all the coloumsin all database
select * from sys.columns where name like '%emp%' --display only given table name
```

14.STORED PROCEDURE: A group of statements that to be existing.

-- INNER PARAMETERS

```
create procedure int_par1(@ename varchar(100),@var int)-- create a proce name
as
begin
select empno,ename from emp where @ename=ename And @var=empno--on what table values you want
end
exec int_par1 'smith',7369
```

OUTPUT PARAMETER

```
create procedure out_Par6(@ename varchar(100),@var int output)-- give name output
as
begin
select  @var=count(Empno) from emp where @ename=job
end
declare @totalcount2 varchar(100)-- give the specfic name of the output
execute out_par6  'clerk', @totalcount2 out
print @totalcount2
```

RETURN PARAMETERS

```
create procedure out_Par7
as
begin
 return (select  count(Empno) from emp)
end
declare @totalcount3 int
execute  @totalcount3 = out_Par7
print @totalcount3
```

```
 select * from emp
 create procedure sum2(@var1 int,@var2 int,@var3 int)
 as
 begin
 declare @sum int
 set @sum=(@var1+@var2+@var3)
 select @sum as colunm
 end
 exec sum2 1000,2000,3000
```

15.SET OPREATORS: Set operators are used to combine the result of 2 or more tables as a single set of values.

```
create table emp_hyd(
empid int,
personname varchar(100),
salary float
)
insert into emp_hyd
values(100,'name1',1000),(102,'name2',1000),(103,'name3',3000),(104,'name4',4000)
select * from emp_hyd
create table emp_pune(
empid int,
personname varchar(100),
salary float
)
insert into emp_pune
values(100,'name1',1000),(107,'name7',7000),(108,'name8',8000),(109,'name9',9000)
select * from emp_pune
--operators
--union: it will return the values but duplicate will not execute
select * from emp_hyd
union
select * from emp_pune
-- union all: it will return values and it will execute the duplicates also
select * from emp_hyd
union all
select * from emp_pune
-- intersect: it will execute only the common value
select * from emp_hyd
intersect
select * from emp_pune
-- EXPECT :it will return values from left hand table which are not right hand table
select * from emp_hyd
except
select * from emp_pune
select * from emp_pune
except
select * from emp_hyd
```

16. TRIGGERS:

```
--trigger; it is automatically perform
--dml tigger:in response to upadte, insert, delete
select * from res
create trigger res_trig on res
for
insert
as
print 'hi'
insert into res values('ai05',54,'ytre')
--two magic
create trigger res_trig2 on res
for
insert,update,delete
as
begin
select * from inserted-- it will consider both update,insert
select * from deleted-- it will consider delete
end
insert into res values ('aio6',45,'poiu')
insert into res values ('aio8',45,'poiu')
delete  from res where aircrafe_code='aio6'
create table res1(
aircrafe_code varchar(100),
seat_no int,
clasess varchar(100))

-- if seat>25 fro other new table
alter trigger res_trig2 on res
for
insert, update,delete
as
begin
declare @seat_no int
select @seat_no=seat_no from inserted
if @seat_no>25
begin
insert into res1 select * from inserted
print 'it is suceed'
end
else
print ' it is not'
end
insert into res values('aio10',21,'poiuy')
insert into res values('aio10',29,'poiuy')
select * from res1
--delete triggers
create trigger trig_del1 on res1
instead of delete
```

```
as
print 'it have been deleted'
set nocount on
delete from res1
--ddl trigger-- in respones yo create,alter,drop
--database scoped ddl trigger
create trigger table_trig--it is a creating a database
on database
for
create_table,alter_table,drop_procedure
as
print 'table is created'
create table t1 (
col1 int,
col2 int)

create trigger table_drop
on database
for
drop_table,alter_table
as
print 'table is deleted'
rollback
drop table t1
--server scope ddl trigger -- it is a servers
create trigger server_trig on all server
for
create_database,drop_database
as
print 'server has been created'
create database data_base
drop trigger res_trig
drop trigger table_trig on database
drop trigger data_base on all server
```

17. SUBQUERY: A sub query is a form of a SQL Statement that appears inside another sql statement. It is also treate as nested query. The statement containing a Subquery is called Parent statements . The parent statement uses the rows returned by the subquery.

```
create table sales_orders(
order_no varchar(10),
client_no varchar (10))
insert into sales_orders values('poo1','coo1'),
('poo2','coo2'),('poo3','coo1'),('poo4','coo3'),('poo5','coo3'),
('poo6','coo3'),('poo7','coo4')
create table cilents(
client_no varchar(20),
person_name varchar(10),
quality numeric (10))
insert into cilents values('coo1','name1',1),('coo2','name2',1),
('coo3','name3',1),('coo4','name4',1),
('coo5','name1',1),('coo6','name6',1)
select * from sales_orders
select * from cilents
select * from cilents where person_name='name3'
-- it is form of a sql statement appears inside the another  sql statement
select * from sales_orders
where client_no=(select client_no from cilents where person_name='name3')
select * from sales_orders
where client_no not in (select client_no from cilents where person_name='name3')
```

18.GROUPING AND HAVING :

Group By: It is used to group rows based on distinct values that exist for specified columns ie it creates a data set containing several set of records grouped together based on a condition.

Having : It is always used in conjuction with group by, It is imposes a condition on the group by clause which further filters the group created by the group by clause.

```
--groups by
create table sales_pro(
product_name varchar(100),
product_num varchar(10),
quality_num int,
quality_dis int
)
insert into sales_pro values('pd1','pooo1',10,10),
('pd1','pooo1',9,9),('pd1','pooo1',8,8),
('pd2','pooo2',7,7),
('pd2','pooo2',8,8),
('pd3','pooo3',10,10),
('pd4','pooo4',5,5),
('pd4','pooo4',6,6),
('pd5','pooo5',9,9),
('pd5','pooo5',8,8),
('pd1','pooo1',10,10)
select *from sales_pro
select product_num,sum(quality_num) as "total quality"
from sales_pro group by product_num -- group by it add rows having a same values in a colunm
-- having condition
select product_num,sum(quality_num) as "total quality"
from sales_pro group by product_num
having product_num in('pooo1','pooo5')-- this HAVING gives the filiters
```

19.FUNCTIONS:

```
--functions: it is a return values
--scaler function:it is a single value
create function fun_sum(@var1 int,@var2 int,@var3 int)
returns int
as
begin
declare @sum int
set @sum=@var1+@var2+@var3
```

```
return @sum
end
select dbo. fun_sum (100,200,300) as total
alter function nrt_fun(@customerid varchar(100))
returns varchar(100)
as
begin
declare @nrt_fun1 varchar(100)
select @nrt_fun1=country from Customers where customerid=@customerid
return @nrt_fun1
end
select dbo.nrt_fun('usa')
--table function retuen values
select * from res
create function tab_fun(@seat_no int)
returns table
as
return (select * from res where @seat_no=seat_no)
select * from tab_fun(87)
```

20.System Functions:

```
--SYSTEM FUNCTIONS
SELECT * FROM Products
SELECT SUM(UnitPrice) AS TotalPrice FROM Products
SELECT COUNT(UnitPrice) AS Total FROM Products
SELECT MAX(UnitPrice) AS Maximum FROM Products
SELECT MIN(UnitPrice) AS Minimum FROM Products
SELECT AVG(UnitPrice) AS Average FROM Products

-- STRING FUNCTIONS
SELECT * FROM Customers
SELECT UPPER(ContactName) as ContactName FROM Customers--uppercase letter
SELECT LOWER(ContactName) as ContactName FROM Customers--lower case letter
SELECT LEFT(ContactName,5) as ContactName FROM Customers-- it will give first five letter
SELECT RIGHT(ContactName,4) as ContactName FROM Customers--it will give last four letter
SELECT SUBSTRING('WELCOME',4,4)-- it will start from 4 letter
SELECT REPLACE('Ahmedabad','Ahmed','Allah') as Result--  it will replace the letters
SELECT LEN('ELEPHANT') AS Result-- it will count the letters
SELECT LTRIM('    WELCOME')--it will erase the space from the left
SELECT RTRIM('WELCOME      ')--it will erase the space from the right
SELECT TRIM('     WELCOME      ')--it will erase the space
SELECT REPLICATE('INDIA ',5)-- it will repate the words
SELECT REVERSE('WELCOME')--it will reverse the words

-- NUMERIC FUNCTIONS
SELECT ABS(-20)-- it will convert the -ve to +ve
SELECT ABS(20)
SELECT SQRT(25)-- square root of the value
SELECT FLOOR(10.8756)-- it will round figure in a lower value
SELECT CEILING(10.8756)--it will round figure in a upper value
SELECT ROUND(10.8756,2)--it will round the value given operation
-- DATE FUNCTIONS
SELECT GETDATE()
SELECT YEAR(GETDATE())
SELECT MONTH(GETDATE())
SELECT DAY(GETDATE())
SELECT DATENAME(MONTH,GETDATE())
SELECT DATENAME(DW,GETDATE())--date name
SELECT DATEPART(HOUR,GETDATE())
SELECT DATEPART(MINUTE,GETDATE())
SELECT DATEPART(SECOND,GETDATE())
SELECT DATEPART(DAY,GETDATE())
SELECT DATEPART(MONTH,GETDATE())
SELECT DATEPART(YEAR,GETDATE())
SELECT DATEADD(DAY,1,GETDATE())
SELECT DATEADD(MONTH,2,GETDATE())
SELECT DATEADD(YEAR,5,GETDATE())
SELECT DATEDIFF(DAY,GETDATE(),'2021-12-24')
SELECT DATEDIFF(DAY,'2021-12-24',GETDATE())
SELECT GETUTCDATE()
```

21.TRANSACTIONS: A SET OF T-SQL STATEMENTS TO BE EXECUTED AS A SINGLE UNIT

```
  create table emp_tran(
empid int,
person_name varchar(100),
salary int)
select * from emp_tran
--explicit transaction: it will commit and rollback
begin tran-- only in this explicit
insert into emp_tran  values (100,'name1',1000),(200,'name2',2000),(100,'name1',1000),
(100,'name1',1000),(100,'name1',1000)
declare @row_count int
select @row_count=count(*) from emp_tran
if @row_count <=10
commit tran
else
rollback tran
select count(*) from  emp_tran
truncate table emp_tran
--implict tran: it is like explicit
set  implicit_transactions on--: imp command
insert into emp_tran  values (100,'name1',1000),(200,'name2',2000),(100,'name1',1000),
(100,'name1',1000),(100,'name1',1000)
declare @row_count int
select @row_count=count(*) from emp_tran
if @row_count <=10
commit tran
else
rollback tran
```

```
select count(*) from  emp_tran
truncate table emp_tran
--autocomtt tran & open transbegin tran
insert into emp_tran  values (100,'name1',1000),(200,'name2',2000),(100,'name1',1000),
(100,'name1',1000),(100,'name1',1000)
--isolation level
--isolation level read and uncommitted
set  transaction ISOLATION level READ UNCOMMITTED
insert into emp_tran  values (100,'name1',1000),(200,'name2',2000),(100,'name1',1000),
(100,'name1',1000),(100,'name1',1000)

set  transaction ISOLATION level READ COMMITTED
insert into emp_tran  values (100,'name1',1000),(200,'name2',2000),(100,'name1',1000),
(100,'name1',1000),(100,'name1',1000)
set  transaction ISOLATION level SERIALIZABLE
insert into emp_tran  values (100,'name1',1000),(200,'name2',2000),(100,'name1',1000),
(100,'name1',1000),(100,'name1',1000)

--snapshot
use master
alter database [proce]
set allow_snapshot_isolation on

use [proce]
set transaction isolation level snapshot

begin tran
insert into emp_tran  values (100,'name1',1000),(200,'name2',2000),(100,'name1',1000),
(100,'name1',1000),(100,'name1',1000)
select * from emp_tran
```

22.QUEREY OPEATORS

```
-- TO HELP T-QUERIES PERTAINING TO SEARCHING AND SORTING ACTIVITIES,
          -- WE GENERALLY CREATE INDEXES

CREATE TABLE EMP_DETAILS
(
EMP_ID INT,
EMP_NAME VARCHAR(20),
EMP_SAL INT
)
INSERT INTO EMP_DETAILS VALUES(1001,'SUNNY',1000)
INSERT INTO EMP_DETAILS VALUES(1002,'PRIYANKA',1200)
INSERT INTO EMP_DETAILS VALUES(1003,'RAVI',1060)
INSERT INTO EMP_DETAILS VALUES(1004,'JEFF',2000)
INSERT INTO EMP_DETAILS VALUES(1005,'JONATHAN',1220)
INSERT INTO EMP_DETAILS VALUES(1006,'PETER',1044)
INSERT INTO EMP_DETAILS VALUES(1007,'AMITABH',1456)
INSERT INTO EMP_DETAILS VALUES(1008,'RAJENDRA',2100)
INSERT INTO EMP_DETAILS VALUES(1009,'ANUSHKA',1220)
INSERT INTO EMP_DETAILS VALUES(2001,'KAREENA',1880)

SELECT * FROM EMP_DETAILS
```

Index: It is used by queries to find data form table quickly. Indexs are created on tables & view , is very similar to an index that we find in a book.

```
create index emp_details_index
on emp_details(emp_name)
SELECT EMP_NAME FROM EMP_DETAILS--index scan

create nonclustered index emp_detalis_index_v2
on emp_details(emp_name)
include(EMP_ID,EMP_SAL)
select * from EMP_DETAILS where EMP_NAME='sunny'--index seek
select * from EMP_DETAILS where EMP_NAME='xyz'--index seek
```

23.NULL AND COALESCE:

```
--isnull : it will take a only first if both colunms is null is shown a null.
create table table_1 ( col1 varchar(10), col2 varchar(10))
insert into table_1 values('siva','krish'),(null,'ram'),('radhe',null),(null,null)
select* from table_1
select isnull(col1,col2) as name from table_1

--nullif: in this both parameters are same it will give null and it is not same returns name only
create table table_3 ( col1 varchar(10), col2 varchar(10))
insert into table_3 values('avi','avi'),('siva','krish'),(null,'ram'),('radhe',null),(null,null)
select* from table_3
select nullif(col1,col2) as name from table_3

--coalesce: it is a like a null but in this we can use number of
create table table_2 ( col1 varchar(10), col2 varchar(10),col3 varchar(10))
insert into table_2 values('hari',null,'varam'),('siva','krish','prasad'),(null,'ram','krishna'),
(null,'radhe',null),(null,null,'aray'),(null,null,null)
select * from table_2
select Coalesce(col1,col2,col3) as name from table_2
use proce
select * from res
```

24.MERGERS: It is a Comparing the table and dimension table if dimension table is not existiong the vale it will update , insert delete.

```
create table customer (
cno int,
cname varchar(10),
price int)
insert into customer values (1,'abs',200),(2,'tre',400),(3,'cbe',600),(4,'oiu',500)
select * from customer
create table dimcustomer (
cno int,
cname varchar(10),
price int)
insert into dimcustomer values (1,'abs',200),(2,'tre',400),(3,'cbe',600)
update customer set price=800 where price=400
delete customer where price=900

merge dimcustomer as target
using customer as source
on source.cno=target.cno
when not matched by target then --- it will be insert
insert(cno,cname,price)
values(source.cno,source.cname,source.cno)  ii ti will be update
when matched and target.cno=2  then update set
target.cname=source.cname,
target.price=source.price
when not matched by source then ----  it will delete .
delete;
```

25. CURSOR: Cursor is a memory location for storing database tables.

- It is a temporary work area allotted to the client at server when a select statements is Executed.
- A cursor contains information on a select statement and rows of data accessed by it.
- The temporary work area is used to store the data retrived from the data base and manipulate the data.

Types of Cursors:

1.
2. **Implicit Cursors:** These cursor created by SQL server by default when select statement is executed.
3. **Explicit Cursors:**
    - When a user create a memory location to store the tables.
    - These cursor will access the records in the table row by row Or one by one.

Step 1.declare the cursor

```
DECLARE @ACODE VARCHAR(10)
DECLARE @SEATS INT
DECLARE @CCODE VARCHAR(10)
DECLARE CRS_RET_VARIABLE CURSOR--creating a cursor
FOR
SELECT * FROM RES
```

Step 2.open the cursor

```
OPEN CRS_RET_VARIABLE
```

Step 3 fetching data the cursor

```
FETCH NEXT FROM CRS_RET_VARIABLE INTO @ACODE,@SEATS,@CCODE
WHILE @@FETCH_STATUS=0--it is holding the data
        BEGIN
                PRINT 'AIRCRAFT_CODE :' + @ACODE + ',seat_no:' + CAST(@SEATS AS
VARCHAR(10)) +',CLAsess:' + @CCODE
                FETCH NEXT FROM CRS_RET_VARIABLE INTO
@ACODE,@SEATS,@CCODE
        END
```

(**@@ Fetch_status:** It is a environmental variable use to check whether cursor variable is holding the record or not. If record is there then the is value is'0')

Step 4 closing the cursor

```
CLOSE CRS_RET_VARIABLE
```

Step 5 deallocate the cursor

```
DEALLOCATE CRS_RET_VARIABLE
```

```
--count a number of row in each table
DECLARE @TAB_NAME VARCHAR(MAX)
DECLARE CURSOR1 CURSOR
SCROLL
FOR
SELECT name FROM sys.tables

OPEN CURSOR1

FETCH FIRST FROM CURSOR1 INTO @TAB_NAME
WHILE @@FETCH_STATUS=0
        BEGIN
                DECLARE @QUERY NVARCHAR(MAX)
                SET @QUERY=N'SELECT +''' + @TAB_NAME +''' AS
TABLE_NAME,COUNT(*) AS ROW_COUNT FROM ' +          @TAB_NAME
                EXEC sp_executesql @QUERY
                FETCH next FROM CURSOR1 INTO @TAB_NAME
        END
CLOSE CURSOR1

DEALLOCATE CURSOR1
```

=======**PROJECT ARTICHTERURE**====================

Copying the data from On premises to Synapses by using a Azure Data Factory.

- Taking the data from On premises files like Oracle db or Some CSV file from FTP site.
- Load the data in to Data Lake gen 2 by using a Azure Data Factory with Pipelines and different activities like Copydata activity , get metadata activity, lookup activities for ETL process.
- After loading a data use a ploybase in Synapses Data Ware House creating a external tables for implementing slowly changing dimensions .
- Loading the data into Dimension and Fact tables which are available in Synapses.
- After that will connect to the Powerbi for designing Visualizations and Dashboards.

Without using a Ploybase how we can copy the data.

- In source side there is a Dataset is Delimiter in that we want to change a three options.
- Firstly to change Option is **Row delimiter** keep a **Line feed(\n).**
- And also there is two options is **Escape character and Quota character** is should be same. So click edit and write a double quotation symbol is **“** both the options.

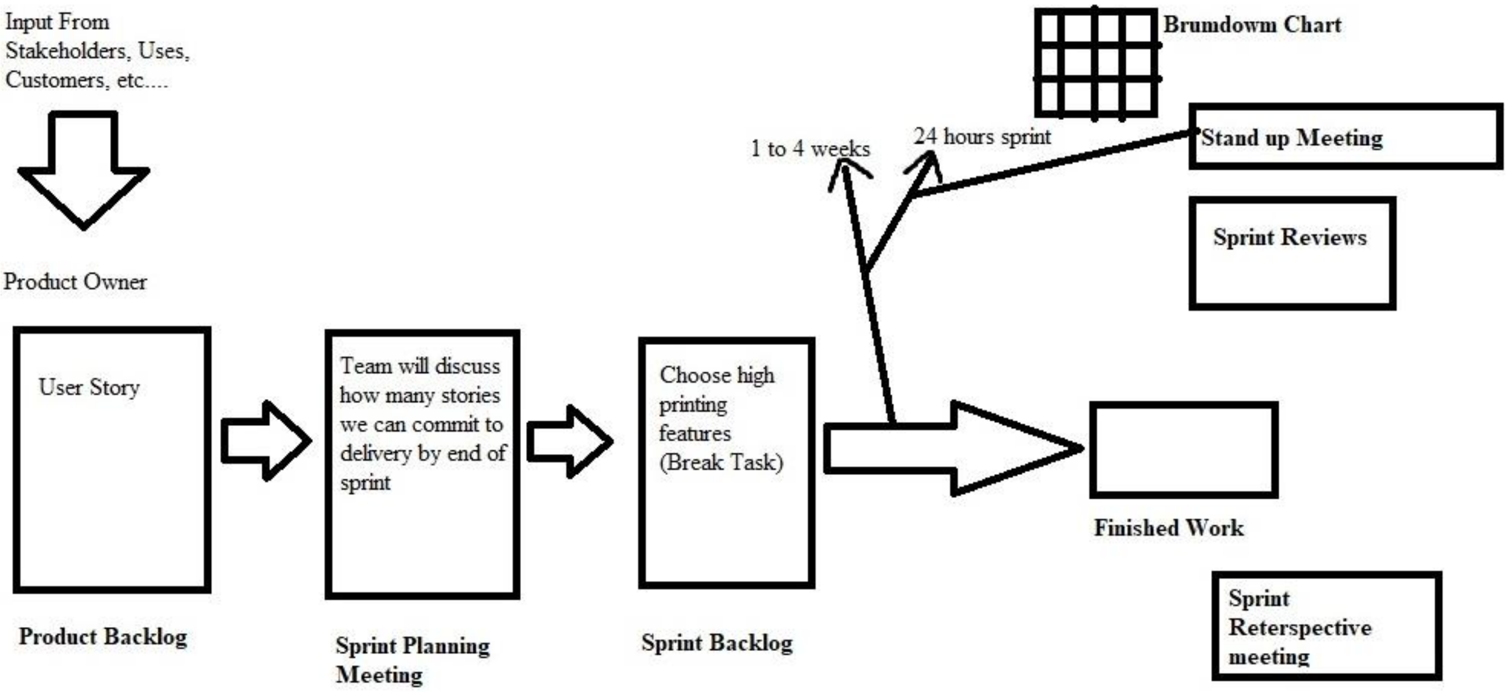
How to run the Data Bricks Notebooks in Azure Data Factory

- In Pipeline there is a Notebook and add a Azure Databricks create a New .
- Give the details and remember if already created the cluster click the option is Existing interactive cluster.
- There is a option a Access token. So to give a Token go to workspace click on workspace name and clock on User settings.
- Generate New token give a New Comment Name and Lifetime(days) according to the client.
- And there create a new token copy that and paste into a Access Token.
- Go to settings and browse the file.

================AGILE METHODLOGY======================

With Simple Sentence of explanation of Agile and scrum work process:

- Agile it is a Methodology, It's implement the project and processing be a iterative work(repeating the work) and incremental
- Previously there will use a Water Fall but there is a some draw backs like Taking the long time for delivery the project and we can't add any update feature after few days of completion project.
- Mainly in this agile is we can delivery the project in a Specific features within short time period what the customer requirements firstly.
- The main Principles are Customer satisfaction (customer can't wait long time) and Collaboration(designing person, developer, product owner, scrum master).
- **Product Backlog**, there is a product owner he will collect the all the information like requirements and features of the project from the End User.
- **Sprint Planning Meeting** in this scrum master, production, developer, testing are all involved and in this there discuss of the work in a period of time.
- Scrum Master will check the principles of Agile end of the project.
- **Sprint Backlog,** here do the work what discussed in Sprint Planning
- Scrum Master will discuss every day with the team 10 to 15 mins of standup meeting for progress of the work and he will prepare the Brumdown chart(means the graph) progress of the work.
- **Sprint Review** , in this the product owner and Scrum master review the Project of the Customer requirements is completed or not.
- **Sprint Retrespective Meeting,** in this all members involved and discuss of progess of the work and also any changes of the work for the next sprint.
- And I will explain with example suppose what discuss in Sprint completed 90% remaining 10% it will add next sprint but in Water Fall can't do like this.


Input From Stakeholders, Uses, Customers, etc....
Product Owner
User Story
Product Backlog
Team will discuss how many stories we can commit to delivery by end of sprint
Sprint Planning Meeting
Choose high printing features (Break Task)
Sprint Backlog
1 to 4 weeks
24 hours sprint
Brumdowm Chart
Stand up Meeting
Sprint Reviews
Finished Work
Sprint Reterspective meeting