

-----Interview Preparation sessions-----

1.Tell Me about your self,exp,skillset,education

Experience

Skill

Hi,My Name is Rajesh

I from bangalore

I have completed my bachler of engineering from JNTU University

I Have around 4 Years of work experience in IT Industery

I have been working for TCS from last 4 years.

I have very good experience on SQL Server and Azure data Engineering

As part of this azure engineering i have worked on different services like

Blob

datalake,

Azure Syanpse,

Data Factory,

Data Bricks

and also i have very good experience on writing of stored procedures and functions and sql queries.

We are following Agile methodolgy

for every two weekes we are conducting one sprint

for deployment of project we are following Azure devops CICD mechanisam.

2.Ratings on Skill

SQL Server---8

Data lake--7

Synapse --used dedicated sql pool---7

Data Factory---8

Data Bricks--Pyspark---7

2.can you please tell me about your project.

Rajesh---CapGemini--HP Client

Bhrmareddy--xyz(consultency)--Capgemini--Cococola

Project Name--client Name

BI Analytics---Flipkart

website--customer--products

Business of Client

What is Domain of client

Sales

Insurance

Healthcare

Banking

HR

Finance

Find out some dimension tables

Fact tables Names in Project

Business Domain---Sonatrach--Oil and Gas Business sales

Data from source

Transformations

Loading Data Target

I am working for sonatrach(abc business), The business of customer is oil and gas.

our data source system is on premises Oracle Database.

first we are copying data from on premises oracle to Data Lake Gen2 in formate of CSV files

by following incremental Loading.

To implement incremental loading we are mainting watermark table to filter data while extracting based on lastruntime of pipeline. we have used lookup activity,copy data activity,stored procedure activity

Once data is available in staging are of data lake gen2

we are applying transfromations and implementing slowly changing dimension type 1 mechanisam

and finally sinking data into target synspase dedicated sql pool of data ware house

by using Mapping Data Flows.

once data is available in target, power bi team desigining reports and dashboards.

1.What is methodology used to develop project

Agile Methdology

Duration Sprint: 2 Weeks

each task ---calling as a User story

2.What are tools used to track project.

Jira or Azure Devops

3.What is backlog

Task which ever not completed in previous sprint then moved to next sprint.

4.User story can be didvided into multiple story points

Needs to give estimation hours

5. Agile meetings

Sprint Planning

Sprint Review

Sprint retrospective

Sprint daily Standup

-----SQL Server-----

SQL Server 2019

SQL Server

SQL Server Managmenet Studio

1. What are system Database

Master

Model

MSDB

TempDB

2. What are DML And DDL Commands

DML---Insert ,delete,update,Select

DDL---create ,Alter,Drop,Truncate

3. What are diff between Delete and Truncate and drop?

delete from employee where empid=200

Truncate table employee

drop table employee

delete command can be used to delete a specific record from table by using where condition

Truncate cant be used to delete a specific record from table by using where cluse

Data deleted by using delete command can be rollbacked

But Truncate command cant be rollbacked

Delete is DML Command

Truncate is DDL command

idntity property will not persist to starting value when deleteting record from table.

Identity property will be persisting to starting value when truncating data from table.

drop command completely removing of
objects(Table,database,storedprocedure,function)

create table employeee(eno int identity(1,1),ename varchar(100),sal int)

eno,ename,sal

1,ram,1000

insert into employee values(1,'Ram',1000)

```
select * from employeee
```

```
delete from employeee where eno=2
```

2.what is diff between Union all and union?

```
select eno,ename,sal from employee  
union all  
select eno,ename,sal from employeee
```

```
select eno,ename,sal from employee  
union  
select eno,ename,sal from employeee
```

Union all simply combining the rows coming from tables.

Union is simply removing duplicates and combining rows coming from tables.

comparing to union, union all is more faster.

3.what are different constraints available in sql server?

Not Null

Check constraint

Default

Primary key

Foreign key

unique

4.What is diff between Primary key and foreign Key and unique key.

Primary key will not allow any null values and any duplicate values.

Foreign key allowing duplicates and refereing the primary key column of other table

unique and primary key both are same but unique will allow only one single null value

Emp

eno(Primary key),ename,dno(foreign Key)--refering dno column from dept

100,ram,10

200,shyaam,20

300,vihaan,10

Dept

dno(Primary key),dname

10,sql

20,powerbi

5.what is order of sql server query to write and to execute.

select eid,sum(sal)

from

employee

where eid=100

group by

eid

having sum(sal)>100

order by eid desc

From

Where

Group by

having

order by

from

where

group by

having

select

order by

Questions:

Table employee: eno, ename, sal, gender

i want drop column Gender

I want add new column employee table

Joining Date:

Create table with identity column

Employee(en0 int, ename varchar(100), sal int, location varchar(100))

add identity property on eno column

add default constraint on 'hyderabad'

add check constraint if sal >500

Two Tables

DROP TABLE EMP

DROP TABLE DEPT

CREATE TABLE EMP

(
EMPNO NUMERIC(4) NOT NULL,
ENAME VARCHAR(10),
JOB VARCHAR(9),
MGR NUMERIC(4),
HIREDATE DATETIME,
SAL NUMERIC(7, 2),
COMM NUMERIC(7, 2),
DEPTNO NUMERIC(2)
)

INSERT INTO EMP VALUES

(7369, 'SMITH', 'CLERK', 7902, '17-DEC-1980', 800, NULL, 20)

INSERT INTO EMP VALUES

(7499, 'ALLEN', 'SALESMAN', 7698, '20-FEB-1981', 1600, 300, 30)

INSERT INTO EMP VALUES

(7521, 'WARD', 'SALESMAN', 7698, '22-FEB-1981', 1250, 500, 30)

INSERT INTO EMP VALUES

```
(7566, 'JONES', 'MANAGER', 7839, '2-APR-1981', 2975, NULL, 20)
INSERT INTO EMP VALUES
(7654, 'MARTIN', 'SALESMAN', 7698, '28-SEP-1981', 1250, 1400, 30)
INSERT INTO EMP VALUES
(7698, 'BLAKE', 'MANAGER', 7839, '1-MAY-1981', 2850, NULL, 30)
INSERT INTO EMP VALUES
(7782, 'CLARK', 'MANAGER', 7839, '9-JUN-1981', 2450, NULL, 10)
INSERT INTO EMP VALUES
(7788, 'SCOTT', 'ANALYST', 7566, '09-DEC-1982', 3000, NULL, 20)
INSERT INTO EMP VALUES
(7839, 'KING', 'PRESIDENT', NULL, '17-NOV-1981', 5000, NULL, 10)
INSERT INTO EMP VALUES
(7844, 'TURNER', 'SALESMAN', 7698, '8-SEP-1981', 1500, 0, 30)
INSERT INTO EMP VALUES
(7876, 'ADAMS', 'CLERK', 7788, '12-JAN-1983', 1100, NULL, 20)
INSERT INTO EMP VALUES
(7900, 'JAMES', 'CLERK', 7698, '3-DEC-1981', 950, NULL, 30)
INSERT INTO EMP VALUES
(7902, 'FORD', 'ANALYST', 7566, '3-DEC-1981', 3000, NULL, 20)
INSERT INTO EMP VALUES
(7934, 'MILLER', 'CLERK', 7782, '23-JAN-1982', 1300, NULL, 10)
```

```
CREATE TABLE DEPT
(DEPTNO NUMERIC(2),
DNAME VARCHAR(14),
LOC VARCHAR(13) )
```

```
INSERT INTO DEPT VALUES (10, 'ACCOUNTING', 'NEW YORK')
INSERT INTO DEPT VALUES (20, 'RESEARCH', 'DALLAS')
INSERT INTO DEPT VALUES (30, 'SALES', 'CHICAGO')
INSERT INTO DEPT VALUES (40, 'OPERATIONS', 'BOSTON')
```

All Kind of joins write sql queries

Inner join

Left outer join

Right outer join

Full Outer join

Cross join

Self join--employee

eno,ename,managename,sal

6.What are diff kind of jons available in SQL Server.

Inner join

Left Outer join

Right Outer join

Full Outer join

Cross Join

7.What is diff between Inner and Left Join.

Employee

eno,ename,dno

1,ram,10

2,shyaam,20

3,vihaan,10

4,Rajesh,40

Dept

Dno,Dname

10,SQL

20,Azure

Inner join

ename,dname

ram,sql

shyaam,azure

vihaan,sql

Left join

ename,dname

ram,sql

shyaam,azure

vihaan,sql

Rajesh,null

Select ename,dname from employee as e inner join Dept as d on e.dno=d.dno

Select ename,dname from employee as e left join Dept as d on e.dno=d.dno

Select ename,dname from employee as e right join Dept as d on e.dno=d.dno

Inner join will give you only matching records from both tables

Left Join will give you all records from left side table,if any record matching null value returned.

Q)Please find no of records returned when i am join below tables on id

Inner ,left,right,full,cross

Table1

id

1

1

1

2

Null

Table2

id

1

1

2

2

null

null

Right Join

id,id

1,1

1,1

1,1

1,1

1,1

1,1

2,2

2,2

null,null

null,null

Left join

id,id

1,1

1,1

1,1

1,1

1,1

1,1

2,2

2,2

Null,Null

Inner join

id,id

1,1

1,1

1,1

1,1

1,1

1,1

2,2

2,2

Table1-id

1

1

3

null

Table-2-id

1

2

null

null

Table1

id

1

1

1

1

Null

5

table2

id

1

1

6

null

null

Table1

id

1

1

2

null

table2

id

1

1

null

null

-----SQL Quireis-----

Create Database

Create table employee(eno,ename,sal ,joiningdate,dno)

Insert few records

when creating table add identity column and primary key on eno column

and sal add check constraint sal>1000

Task2:

Create department

dno,dname

perform all kind of joins

```
CREATE TABLE [dbo].[DEPT]
```

```
(DEPTNO INT PRIMARY KEY ,  
DNAME VARCHAR(20),  
LOC VARCHAR(20) );  
GO
```

```
INSERT INTO [dbo].[DEPT] VALUES (10, 'ACCOUNTING', 'NEW YORK');  
INSERT INTO [dbo].[DEPT] VALUES (20, 'RESEARCH', 'DALLAS');  
INSERT INTO [dbo].[DEPT] VALUES (30, 'SALES', 'CHICAGO');  
INSERT INTO [dbo].[DEPT] VALUES (40, 'OPERATIONS', 'BOSTON');
```

```
CREATE TABLE [dbo].[EMP]  
(EMPNO INT PRIMARY KEY,  
ENAME VARCHAR(20),  
JOB VARCHAR(20),  
MGR INT,  
HIREDATE DATE,  
SAL MONEY,  
COMM MONEY,  
DEPTNO INT FOREIGN KEY REFERENCES DEPT(DEPTNO) );  
GO
```

```
INSERT INTO [dbo].[EMP] VALUES  
(7369, 'SMITH', 'CLERK', 7902, '17-DEC-1980', 800, NULL, 20);  
INSERT INTO [dbo].[EMP] VALUES  
(7499, 'ALLEN', 'SALESMAN', 7698, '20-FEB-1981', 1600, 300, 30);  
INSERT INTO [dbo].[EMP] VALUES  
(7521, 'WARD', 'SALESMAN', 7698, '22-FEB-1981', 1250, 500, 30);  
INSERT INTO [dbo].[EMP] VALUES  
(7566, 'JONES', 'MANAGER', 7839, '2-APR-1981', 2975, NULL, 20);  
INSERT INTO [dbo].[EMP] VALUES  
(7654, 'MARTIN', 'SALESMAN', 7698, '28-SEP-1981', 1250, 1400, 30);
```

```

INSERT INTO [dbo].[EMP] VALUES
(7698, 'BLAKE', 'MANAGER', 7839, '1-MAY-1981', 2850, NULL, 30);
INSERT INTO [dbo].[EMP] VALUES
(7782, 'CLARK', 'MANAGER', 7839, '9-JUN-1981', 2450, NULL, 10);
INSERT INTO [dbo].[EMP] VALUES
(7788, 'SCOTT', 'ANALYST', 7566, '09-DEC-1982', 3000, NULL, 20);
INSERT INTO [dbo].[EMP] VALUES
(7839, 'KING', 'PRESIDENT', NULL, '17-NOV-1981', 5000, NULL, 10);
INSERT INTO [dbo].[EMP] VALUES
(7844, 'TURNER', 'SALESMAN', 7698, '8-SEP-1981', 1500, 0, 30);
INSERT INTO [dbo].[EMP] VALUES
(7876, 'ADAMS', 'CLERK', 7788, '12-JAN-1983', 1100, NULL, 20);
INSERT INTO [dbo].[EMP] VALUES
(7900, 'JAMES', 'CLERK', 7698, '3-DEC-1981', 950, NULL, 30);
INSERT INTO [dbo].[EMP] VALUES
(7902, 'FORD', 'ANALYST', 7566, '3-DEC-1981', 3000, NULL, 20);
INSERT INTO [dbo].[EMP] VALUES
(7934, 'MILLER', 'CLERK', 7782, '23-JAN-1982', 1300, NULL, 10);

```

-----common Quries-----

Display all employee sal more then 3000

Display all employee sal in between 2000 and 3500

Display what are different kind of job types in emp table

Dispaly all emp belongs to department name 'sales'

Display top 1 Salary employeename

Display employee name who is 'president' of this company

Display Employees who are commison is null or not eligible.

-----Duplicate Queries-----

Find Duplicate employees in employee table if empid is repeating more than one time

(How to find duplicate values in Table with out using CTE)

Delete duplicate values in Employee table by using CTE(Row_Number)

-----Highest salary Queries-----

Find Top 1 Salary employee in Employee table using TOP Keyword

Find 2nd highest salary employee in Employee table using dense_rank and CTE

Find 2nd highest salary employee in employee table using Correlated subquery.

How to find duplicate values and delete in sql server employee tables

There are several ways to find and delete duplicate values in a SQL Server table, but one common method is to use a self-join and a subquery. Here is an example of how this could be done for an "employee" table:

Step 1: Finding duplicate values:

sql

```
SELECT
    first_name, last_name, COUNT(*) as count
FROM
    employee
GROUP BY
    first_name, last_name
HAVING
    COUNT(*) > 1
```

This query groups the rows in the "employee" table by the "first_name" and "last_name" columns, and then uses the COUNT() function to count the number of occurrences of each group. The HAVING clause is used to filter out groups that have a count greater than 1, thus identifying duplicate values.

Step 2: Deleting duplicate values:

sql

```
WITH CTE AS (  
    SELECT  
        ROW_NUMBER() OVER (PARTITION BY first_name, last_name ORDER BY id)  
    as row_num,  
        id  
    FROM  
        employee  
)  
DELETE  
FROM  
    CTE  
WHERE  
    row_num > 1
```

What are Temporary tables

Local

Global

What is diff between cte and temporary tables and table variables

what is diff between stored procedure and function

what is index and what are different kind of index available

-----Quries-----

1.Create two temporary tables synatx

emplocal

empglobal

eno ,ename,sal

employee ,dept

2.Create CTe to find 2nd highest salary by each department wise

3.create table variable

name: emptable

eno,ename,sal

4.i have emp

needs to create new table name:empnew(eno,ename,sal,job)

how to create new table using existing table

5.emp(columns)

empold(columns)

insert data from emp into empold

What is diff between row_number,Rank,Dense_Rank

eno,ename,sal,rownumber,rank,dense_rank

1,ram,1000,1,1,1

2,shyaam,800,2,2,2

3,vihaan,800,3,2,2

4,'rajesh',700,4,4,3

rownumber will give u continuous values

rank will maintain gaps when u have same value

dense_rank will not maintain any of gaps, even u have same values.

Temp tables are using Temp Database

Temp tables are using u r harddisk memory

Temp Tables are eligible to create index

Temp tables are used to store high volume of data bcz optimization available on this tables

Table variables are not storing any of database

Table variables are using in memory means RAM in computer

Table variables are not eligible to create index

Table variables are used to store small volume of data because no optimization on this tables

CTE--Common Table Expression

cte is used to save query result and needs to use that result immediately

it is also using in memory

Recursiveness feature..cte result can be used within same cte

what are diff between stored procedure and function.

Stored procedure contains set of sql statements precompiled and saved inside database.

sp is not mandatory to return data, it may or may not return data.

function should return data.

sp allowing a function to execute inside

function will not allow to execute a procedure inside.

error handling try and catch block is possible in sp

but function will not support

inside sp u can use temp tables

function will not support

sp result will not join with any table

function result can be perform inner left kind of joins with other tables.

-----self join-----

find out employee name and manager name of each employee using self join in emp table.

```
select m.ENAME as employee_name, e.ENAME from emp as e inner join emp as m
on e.EMPNO=m.MGR
```

what is diff between subquery and correlated subquery

Non correlated query is one direction inside query will execute first

then based on result outside query will be executed.

```
select * from emp where deptno in (select DEPTNO from dept where  
dname='sales')
```

Correlated sub query:

correlated subquery is two directional query

first outer query will execute

based on result inner query will execute for each row

based on condition outer query will return result

correlated query is performance is not good comparing to non correlated

Finding nth highest salary in employee table

```
select * from emp as e where
```

```
1=(select count(distinct sal) from emp as e1 where e1.sal>=e.sal)
```

what is index, diff between clustered and non clustered index

index is for improving query search performance by creating pointers on table.

Two kind of index

clustered:

Only one clustered index on table on single column.

It will sort data in table when clustered index created.

clustered index contains actual data in leaf nodes.

Non clustered

as many you can create on table

it will not sort any data

in leaf nodes it doesn't contain actual data, it contains key lookup value to identify actual data.

How to improve long running sql queries or stored procedures in SQL Server.

we have to look on execution plan.

by finding cost operation, where it is taking too much time

and identifying kind of join and kind of scan happening

whether table scan, index scan based on that needs to convert index scan to index seek operation

by creating some non clustered indexes.

index seek will reduce number of page search operations.

then we are following best practices while writing sql queries like
using temp tables, table variables cte based on volume of data
making joins properly.

what is view.

views are contains some select statements returning result from base tables

views doesn't store any data.

Views are calling as imaginary tables also

view can be used to update or delete or insert when writing simple select statements inside.

it will effect base table

create view empdepinfo

as

```
select ename,dname from emp as e inner join dept as d on e.dno=d.dno
where dname='sales'
```

what is materilized view

when creating index on view

it will occupy some memory and data.

-----how to fill missing dates in below table

```
--date,salesamount
```

```
-----
```

```
--2023-01-01,100
```

```
--2023-01-03,300
```

```
--2023-01-07,700
```

-----how to generate all dates between two dates

```
declare @startdate date='2023-01-01'
```

```
declare @enddate date='2023-01-07'
```

```
;with cte as
```

```
(
```

```
select @startdate as startdate
```

```
union all
```

```
select dateadd(dd,1,startdate) from cte where startdate<@enddate
```

```
) select c.startdate,isnull(s.salesamount,0) from cte as c left join sales as  
c.date=s.startdate
```

```
customer,mobilenumber,landnumber,officenumber
```

```
ram,988653534,21662232,null
```

```
shyaam,null,2326532,73643764
```

```
vihaan,null,null,353586585
```

-----what is diff between isnull and coalesce

```
select customer,coalesce(mobilenumber,landnumber,officenumber) as  
contact from table
```

```
select empno,ename,isnull(comm,0) as comm from emp
```

-----Data Lake-----

what is diff between Data Lake gen1 and Data Lake Gen2

what is diff between blob storage and Data Lake Gen2

what is Hierarchical Namespace

Azure Data Engineering

What is cloud computing

accessing services remotely through internet by using web browser from datacenter

what are different cloud services categorized

IAAS--infrastructure as a service

PAAS--platform as a service

SAAS--software as a service

what are different deployment modes of cloud computing

public

private

hybrid

What is Storage account

this is used to store any kind of data.

structured---csv table format

semi structured---json,xml

unstructured---video ,audio,images

what are different storage modes in storage account

container or blob

Table

file share

Queue

what are diff blob types

block

page

append

what are diff access tiers

hot--

cool

archive

What are diff security mechanisms

Access Keys

Connections Strings

Share access signatures

RBAC---Role based access control

Blob Storage

Data Lake

Azure SQL

Synapse Dedicated SQL Pool

Data Factory

Data Bricks

-----Synapse Datawarehouse-----

-----Synapse Dedicated SQL pool-----

Synapse Analytics

Serverless SQL Pool

Pipelines

Notebooks

Sparkpools

-----DWH Concepts-----

what is datawarehouse

what is data mart

what is dimension table

what is fact table

what is are diff kind of dimensions

roleplaying,conformed,junk,degerated,slowly changing diemnsion

what are diff fcats

additive

semi additive

non additive

what is star schema

what is snowflake schema

what is diff between star and snowflake

what is factless fcat table

what is surrogate key

what is oltp and olap

-----Synapse Dedicated SQL Pool-----

What is Archture of Synapse Dedicated SQL Pool

what are diff distributions available

what is default distribution method

when can we use round robin ,hash,replicate

what is polybase.

what is diff between azure sql vs synapse dedicated sql pool

what is diff between serverless pool vs dedicated sql pool

-----Practicals-----

create synapse dedicated sql pool

create external table to read data from blob storage contains employee file using polybase

what is CTAS and create table and load from external table using CTAS also assign roundrobin dist

on target table.

-----OLTP

Capturing of day to day live transactions data from application into database

ER--Entity Relational Modeling

Master tables--customer,customer_address,product

Transactions--sales

primary key and foreign key relation

customer

cid,firstname,lastname.....

100,ramboyapati

transatuionsales

cid,pid,qty,amount

100,6,7,1000

ETL

-----DWH

Dimensional Modeling

Dimesnions--dimcustomer,dimproduct

Fact--fcatsales

ckey,cid,cfullname,loaction,flag,startdate,enaddate

1,100,ram boayapti,bangalore,old,2023-01-01,2023-02-22

2,100,ram boyapati,hyderabad,current,2023-02-22,Null

Factsales

pkey,ckey,amount,qty

2,1,1000,20

5,2,2000,30

2 lakh

6 lakh

1 lakh

5 lakh

4 lakh

1

10

2

13

20

33

What is Spark

What is archeture of spark

what is diff between mapreduce and spark

what are diff kind of cluster types in databricks

what are diff cluster modes

what are List,Tupple,Set,Dictnories

what is diff between RDD,Data Frame,Dataset

What is DAG

what is Lazy Evalution

What are Traansformations

What are actions

What are Narrow and wide Transformations

MAP

Shuffle

Reduce

Narrow---Filter,Map

narrow transformations are operations that do not require data shuffling or data movement across partitions

Wide---Groupby ,Join

wide transformations require data shuffling or data movement across partitions, and require data to be collected and redistributed across multiple partitions

What are Coalesce and Repartition

What are list of transformations used in Pyspark

withcolumn

Filter

firstname,lastname

Ram,Boyapati

Ram-Boyapati

```
from pyspark.sql import *
```

```
df2=df.withColumn("FullName",concat(df.firstname,lit('-'),df.lastname))  
      .withcolumn("currenttime",Current_Timestamp())
```

filter

select

Display(Df)

df.show()

df.printschema()

Distinct

sum()

Groupby()

JOINS

emp			Dept	
eno	ename	dno	dno	dname
1	Ram	10	10	MSBI
2	Shyaam	20		

Left ANTI---It will returns all records from left side which ever non matching with right side

Left Anti	eno	ename	dno
	2	Shyaam	20

Left Semi---It will returns all matching records from left side and only columns from left side

Left SEMI

eno	ename	dno
1	Ram	10

How to write SQL commands in NoteBook Default language as Pyspark

%SQL

What are diff kind of joins available in Pyspark

inner

left

right

full

left anti

left semi

```
dfemp.join(dfdept,dfemp.dno==dfdept.dno,"inner").select
(dfdept.dno,dfemp.eno,dfemp.ename).show()
```

What is Broadcast Join and Broadcast variable

```
dfemp.join(Broadcast(dfdept),dfemp.dno==dfdept.dno,"inner").select  
(dfdept.dno,dfemp.eno,dfemp.ename).show()
```

Replicating small dataset over all cluster nodes and saved into memory.
without data shuffling join operation between datasets will be performed

How to Create Mount to Data Lake

first we are creating App in Active directory

app contains secretkey,tenantid,clientid

then creating mount point by using dbutil.fs.mount(containername and storage
account name)

How to create Connection to Azure SQL database

How to write sql on Data Frame

empdf

Create or replace tempview("emp")

%sql

select gender,sum(sal) from emp

group by gender

How to pass parameters to notebook

dbutils.widget.

How to Call notebook from another notebook

what is caching and persistence

caching and persistence both are used to save data in memory

for more faster access

in addition to caching persistence will support to specify external storage to cache data

what is delta Lake

What is Parquet

more faster to read and access

compatible to HDFS Big Data

columnar format, compressing data

for write operations avro file format more suitable

How to remove duplicates write code in Pyspark

Find 2nd highest salary employee write pyspark code

How to convert RDD into Dataframe

```
dataframe=rdd.toDF()
```

What is diff Pyspark Dataframe and Pandas Dataframe

pandas df is on single machine

pyspark df is running on multiple distributed clusters

How to find History of Delta Lake Tables

Find Current version of Delta Lake Table

How to read multiple files in Pyspark into DataFrame

How to Read and write data in Pyspark

-----what are major components of power bi-----

Power BI Desktop

Power BI Online

Power BI report server

Power Mobile APP

-----what are diff connectivity modes-----

Import

Direct query Mode

Live Connection

-----what are datasources used in u r project-----

Azure synapse dedicated sql pool---Import

Azure analysis service tabular--Live

-----what are diff kind of filters available in power bi----

ALL Page level---Report level

Single page level

Visual level

Slicer

Drill Through

-----what are visuals used in u r project-----

column chart

bar chart

table

matrix

pie chart

slicer

line chart

-----where did u publish reports-----

published into power bi online services

www.app.powerbi.com

workspace---publish

-----what is report and dashboard

-----what is power bi gateway

-----what is row level security

-----diff kind of licence types in power bi-----

Azure data engineering with SQL Real Time Preparation

SQL Server---30 %

3 questions on Theory

5 Questions Queries

1.A) What are diff system databases in SQL Server

Master

Model

MSDB

TempDB

1.what are DML and DDL Commands

in sql server we have dml and ddl commands

DML--data manipulation commands

insert,update,delete,select

DDL--Data definition commands

Create ,Alter,Drop,Truncate

DCL--Data Control Language

Grant,Revoke,Rename

2.what is diff between Delete and Truncate and drop

-----Delete

Delete is a DML Command

Performance is not good

while deleting log files will be created

u can delete based on condition

delete from employee where eno=100

identity property will not persist to starting value

-----truncate

Truncate is DDL Command

Performance is good

It will not create any log Data

you cant write any conditional delete

truncate table employee

identity property value persist to starting value

3.what is diff between where and Having

Where condition filtering rows on a table columns

select * from emp where dno=10

Having clause is used to filter rows after aggregation

```
SELECT StudentName, CourseName
FROM Students GROUP BY StudentName, CourseName
HAVING COUNT(*) > 10;
```

4. what is diff union all and union

Union and union all both are used to combine rows from different tables into one table

union all combines data even rows are repeating

union combines data and removing any repetition of rows

union all is more performance

union is slow in performance bcz it needs to remove duplicate

table1

eno,ename,sal,dname

table2

eno,ename,sal

```
select eno,ename,sal,dname from emp1
union all
select eno,ename,sal," as dname from emp2
```

5. what are diff kind of constraints

not null

default

check

primary key

unique key

foreign key

6.what is diff between Primary Key and Unique Key

only one primary key can be created

No duplicates and No null values in primary key column

Many unique keys can be created

No duplicates But only single null value

7.what are diff kind of Joins available in sql server

if u want to join two tables there should be one common column in both tables

INNER JOIN--ONLY MATCHING RECORDS FROM BOTH TABLES

LEFT OUTER JOIN---ALL RECORDS FROM LEFT SIDE TABLE ,ONLY MATCHING VALUES FROM RIGHT SIDE TABLE

RIGHT OUTER JOIN--ALL RECORDS FROM RIGHT SIDE TABLE, ONLY MATCHING FROM LEFT SIDE TABLE

IF ANY RECORDS NOT MATCHING CORRESPONDING VALUE DISPLAY AS NULL.

FULL OUTER JOIN---ALL RECORDS FROM BOTH SIDES,MATCHING AND NON MATCHING

IF NON MATCHING NULL VALUES WILL DISPLAY

CROSS JOIN---CARTESIAN PRODUCT OF TWO TABLES MEANS COMBINATION OF TWO TABLES

8.what are temporary tables

TEMPORARY TABLES ARE USED TO STORE CALCULATED DATA TEMPORARLY IN TEMP DATABASE

LOCAL--#

CREATE TABLE #EMP(ENO INT,ENAME VARCHAR(100),SAL INT)

WITH IN SESSION ONLY THAT TABLE AVAILABLE,ONCE SESSION CLOSE TABLE WILL DROP

GLOBAL--##

CREATE TABLE ##EMP(ENO INT,ENAME VARCHAR(100),SAL INT)

WITH IN CONNECTION ONCE CONNECTION CLOSE TABLE WILL BE DROPPED

9.what is view and what is materilized view

VIEW IS IMAGINARY TABLE ,IT DOESNT CONATIN ANY DATA.IT WILL JUST STORE QUERY

IF I AM CREATING INDEX ON VIEW ,IT IS CALLED MATERIALIZED VIEW

CREATE VIEW EMPDEPT

AS

WITH ENCRYPTION

SELECT ENAME,SAL,DNAME FROM EMP AS E INNER JOIN DEPT AS D ON E.DEPTNO=D.DEPTNO WHERE DNO=10

SELECT * FROM EMPDEPT

10. what is Cte and table variable

TABLE VARIABLE ALSO TEMPORARLY STORING OF DATA,EXISTINCE OF TABLE VARIABLE IS WITH IN THAT
QUERY EXECUTION

```
DECLARE @Employees TABLE
(
    EmployeeID INT,
    EmployeeName VARCHAR(50)
)
```

11.what are diff between cte ,temp tables ,TABLE VARIABLES.

CTE(Common Table expression)---immediatly needs to use result

Table variables---existency of table variables are with in that scope of query execution

Temp tables---are storing of data in temp db

local ---existency of local is with in session

Global---with in connection

Recursiveness Feture---result of cte can be used immediatly with in same cte

when having less data around 10000 rows use Table Variables

when having high volume data use temp tables

Temp tables can use to create index

Temp tables can use perfomence improvement statstics

Date Table

Datekey,month,year.....date

```
DECLARE @StartDate DATE = '1999-01-01';
```

```
DECLARE @EndDate DATE = '2023-01-10';
```

```
-- Generate dates using a CTE and insert them into the temporary table
```

```
WITH DateRange AS (
```

```
    SELECT @StartDate AS CurrentDate
```

```
    UNION ALL
```

```
    SELECT DATEADD(DAY, 1, CurrentDate)
```

```
    FROM DateRange
```

```
    WHERE DATEADD(DAY, 1, CurrentDate) <= @EndDate
```

```
)
```

```
SELECT CurrentDate, year(CurrentDate) as Yearname, Month(CurrentDate) as  
Month, Datename(mm, CurrentDate) as MonthName
```

```
FROM DateRange
```

```
OPTION (MAXRECURSION 0);
```

12. what is diff between stored procedure and Functions

Stored procedure may or may not return data.

Function will always return data.

Stored procedure can have error handling Try and catch blocks

Function will not have error handling Try and catch block

stored procedure can not be call and execute inside function

function can be call and execute inside stored procedure

Function can join with table

stored procedure cannot join with tables

13.what are diff kind of index types available

Index can be created on table to query data more faster

clustered ---only one clustered index can be created

 eno --Data will be sorted on index

 Btree structer---leaf nodes it contains actual data.

Non clustered---as many non clusetered index can be created on single table

 leaf nodes it doesnt contain actual data, it contains key values

Row index

columnor storage index

14.How to improve SQL Server query performance

first we are looking in query execution plan, inside there is a cost operator
will provide information where query is taking more time.

we can look on index missing, if any indexex are missing we can create
we have to look on scan type

Table scan---if no index

Index scan---when table contains clustered index,

Index seek---when table contains clustered and non clustered then index seek will happen.

Index seek operations can reduce no of page search operations.

also we need to follow some best practises while writing sql queries

making proper joins

Use appropriate indexes: Analyze your query execution plans and identify which columns are frequently used for filtering, joining, or sorting. Create indexes on those columns to speed up data retrieval.

Use explicit JOIN syntax: Instead of using implicit joins (comma-separated tables in the FROM clause), use explicit JOIN syntax (INNER JOIN, LEFT JOIN, etc.). It makes your queries easier to read and maintain.

Keep column names in place of star in select statement

Avoid using scalar functions in WHERE clauses: Scalar functions can hinder performance as they need to be executed row by row. Consider rewriting queries to eliminate or reduce the usage of scalar functions.

Minimize network round trips: Combine multiple queries into a single query when possible. Reduce the number of calls between the application and the database server.

Use appropriate data types: Choose the correct data types for your columns. Avoid using larger data types than necessary, as it increases storage and memory requirements.

Update statistics: Keep your statistics up to date to help the query optimizer generate efficient execution plans. Use the "UPDATE STATISTICS" command to update statistics on tables and indexes.

15.what are windows functions and what diff between Rank() and dense_Rank()

RowNumber

Rank

Dense_Rank

cid,sales,rank,denserank

1,1000,4,3

2,3000,2,2

3,3000,2,2

4,4000,1,1

16.what is order of sql query and order of execution

select---5

from---1

where---2

group by---3

having---4

order by---6

17.what is diff between isnull,nullif,coalsce

select *,isnull(commisison,10) from employee where commisiion is null

isnull()---finding null values in column and replacing with alternative value

Nullif()---comparing two values if both same will return null else first value will return

coalsce()---first non null value among multiple columns

18.what is subquery ,diff correlated and non correlated sub query

Non coreelated sub query will execuste first then outer query will execute

This is

```
select * from employee where dno in (select dno from dept where
dname='MSBI')
```

correlated sub query:

two directional query first outer query will execute and then inner query will execute based on outer result

again outer will execute and give u final

Quries:

1.find duplicated employees in table

eid,ename,sal

1,ram,1000

2,shyaam,2000

1,kiran,2000

```
select eid,count(*) as dup from employee
```

```
group by eid
```

```
having count(*)>1
```

2.How to find 2nd highest salary employee with out using CTE

-----using Top

```
SELECT TOP 1 Salary
FROM (
    SELECT DISTINCT TOP 2 Salary
    FROM Employee
    ORDER BY Salary DESC
) AS SalaryTable
ORDER BY Salary ASC
```

-----correlated sub query

```
SELECT DISTINCT Salary
FROM Employee e1
WHERE (N-1) = (
    SELECT COUNT(DISTINCT Salary)
    FROM Employee e2
    WHERE e2.Salary > e1.Salary
)
```

3.How to find 2nd highest salary employee with by using CTE

```
WITH SalaryCTE AS (
    SELECT Salary, DENSE_RANK() OVER (ORDER BY Salary DESC) AS Rank
    FROM Employee
)
SELECT Salary
FROM SalaryCTE
```

WHERE Rank = 2;

4.Find 2nd highest salary employees in table by each department wise

WITH SalaryCTE AS (

 SELECT Salary, DENSE_RANK() OVER (partition by deptname ORDER BY Salary
DESC) AS Rank

 FROM Employee

)

SELECT Salary

FROM SalaryCTE

WHERE Rank = 2;

5.Remove duplicate employees in table using CTe

WITH DuplicateCTE AS (

 SELECT *, ROW_NUMBER() OVER (PARTITION BY EmployeeID ORDER BY
EmployeeID) AS RowNum

 FROM Employee

)

DELETE

FROM DuplicateCTE

WHERE RowNum > 1;

6.Convert Upivoted Data into Pivot formate

City,Year,sales

hyd,2022,1000

hyd,2023,1000

Bang,2022,3000

Bang,2023,2000

expected Result

City,2022,2023

Hyd,1000,1000

Bang,3000,2000

```
SELECT *  
FROM (  
    SELECT City, [Year], Sales  
    FROM YourTable  
) AS SourceTable  
PIVOT (  
    SUM(Sales)  
    FOR [Year] IN ([2022], [2023])  
) AS PivotTable;
```

7.empid,ename,gender

1,Ram,Male

2,radha,Female

update above table gender column Male into female vice versa

```
UPDATE YourTableName  
SET gender =  
CASE  
    WHEN gender = 'Male' THEN 'Female'  
    WHEN gender = 'Female' THEN 'Male'  
    ELSE gender  
END;
```

8.find employee in each department who ever having there salaries more then average salary of department

```
SELECT d.department_name, e.employee_name, e.salary
FROM employees e
INNER JOIN departments d ON e.department_id = d.department_id
WHERE e.salary > (
    SELECT AVG(salary)
    FROM employees
    WHERE department_id = e.department_id
)
ORDER BY d.department_name, e.employee_name;
```

9.I have two Tables contains

Table1

ID

1

2

1

1

null

null

Table2

ID

1

1

Null

2

Inner Join---

Left Join---

Right Join

Full Join

Cross Join

id,id

Table1

COUNTRY

Ind

SL

Pak

Result:

Ind-SL

Ind-Pak

SL-Pak

Please write a query to get the output result

SELECT t1.Country + '-' + t2.Country AS Result

FROM YourTable t1

JOIN YourTable t2 ON t1.Country < t2.Country

-----Miscellaneous Questions-----

- 1.can i insert data into View
- 2.Can i call stored procedure in Function
- 3.Can i restore data once i truncated
- 4.Can i insert values into identity column
- 5.can i delete row contains foreign key reference
- 6.Can i create more than one clustered index on single table
- 7.Can i create more than one primary key on single table

-----Today Expectation-----

- 1.Install SQL Server and SQL management studio
- 2.create data base
- 3.create two tables
employee table

eno,ename,sal,jobtype,dno

Department table

dno,dname

create

alter

insert

update

delete

truncate

drop

Perform all kind of Joins in SQL

employee ,deptment

-----syllabus of SQL Server-----

DML and DDL---create,delete,update,select,alter.....

all keyword---top,distinct,between....union ,union all

all cluses---where,having,group by,order by

all aggregate functions---sum,max,min,avg

all string functions---

left,right,substring,charindex,nullif,isnull,coalesce,concat,ltrim,rtrim,reverse

all date and time functions---

year,month,day,datepart,datediff,dateadd

all constraints---check,not null,default,primary key,foreign key,unique key

subqueries---

views

cte

table variables

temp tables---Local and global

stored procedures

functions

window functions

index---clustered,non clustered

Azure Data Engineer--70%

-----Interview Preparation sessions-----

1.Tell Me about your self,exp,skillset,education

Experience

Skill

Hi,My Name is Rajesh

I from bangalore

I have completed my bachelor of engineering from JNTU University

I Have around 4 Years of work experience in IT Industry

I have been working for TCS from last 4 years.

I have very good experience on SQL Server and Azure data Engineering

As part of this azure engineering i have worked on different services like

Blob

datalake,

Azure Synapse,

Data Factory,

Data Bricks

and also i have very good experience on writing of stored procedures and functions and sql queries.

We are following Agile methodology

for every two weeks we are conducting one sprint

for deployment of project we are following Azure devops CI/CD mechanism.

Create Storage Account BLOB storage

Upload few files

Create Azure SQL database

Create two Tables

and insert data

and connect to azure sql from local sql management studio

create data lake gen2 create 3 lawyers

create synapse dedicated sql pool

create polybase external table

-----Data Factory,Data Bricks-----

2.Ratings on Skill

SQL Server---8

Data lake--7

Synapse --used dedicated sql pool---7

Data Factory---8

Data Bricks--Pyspark---7

2.can you please tell me about your project.

Rajesh---CapGemini--HP Client

Bhrmareddy--xyz(consultancy)--Capgemini--Cococola

Project Name--client Name

BI Analytics---Flipkart

website--customer--products

Business of Client

What is Domain of client

Sales

Insurance

Healthcare

Banking

HR

Finance

Find out some dimension tables

Fact tables Names in Project

Business Domain---Sonatrach--Oil and Gas Business sales

Data from source

Transformations

Loading Data Target

I am working for sonatrach(abc business), The business of customer is oil and gas.

our data source system is on premises Oracle Database.

first we are copying data from on premises oracle to Data Lake Gen2 in format of CSV files

by following incremental Loading.

To implement incremental loading we are maintaining watermark table to filter data while extracting based on last runtime of pipeline. we have used lookup activity, copy data activity, stored procedure activity

Once data is available in staging area of data lake gen2 we are applying transformations and implementing slowly changing dimension type 1 mechanism and finally sinking data into target synapse dedicated sql pool of data warehouse by using Mapping Data Flows.

once data is available in target, power bi team designing reports and dashboards.

1. What is methodology used to develop project

Agile Methodology

Duration Sprint: 2 Weeks

each task --- calling as a User story

2. What are tools used to track project.

Jira or Azure DevOps

3. What is backlog

Task which ever not completed in previous sprint then moved to next sprint.

4. User story can be divided into multiple story points

Needs to give estimation hours

5. Agile meetings

Sprint Planning

Sprint Review

Sprint retrospective

Sprint daily Standup

-----Azure Data Engineering-----

What is cloud computing

what are diff cloud vendors

what are diff cloud deployments

public

private

Hybrid

what are diff categories of cloud services available

IAAS--infra structer as a services

PAAS--platforma as a services

SAAS--software as services

Storage Account

what is storage account

text, csv, excel, json, xml, audio, video, image

Different kind of storage modes

Containers or blobs

Table

File share

quese

what are different blob types.

Block blob

Page Blob

Append Blob

what are diff access tiers

Hot---frequently

Cool--un frequently

Archive---

what is soft delete---7 resotre

how to delete file if created or modified date gretaer then specific number of days

how to move file hot tier to cool or archive

if created or modified date gretaer then specific number of days

Life cycle Managemnt---Rules

what are diff security mechanisam

Access Keys

Connection Strings

Shared Access Siganture

RBAC---Role based access controle

what is advantages of SAS(Shared access signature)

what is lease acquire

file blob---

Data Lake Gen1---Not Available

Data Lake Gen2

What is Data Lake

Data Lake is Big data storage and processing,

It has compatability with HDFS(Hadoop Distributed File System)

What is Data Lake gen used in u r project

what is diff between Gen1 and Gen2

independent service

doesn't have RBAC

Redudency

What is diff between Blob storage and Data Lake Gen2

storage account+ enable Hierachacal Namespace

blob storage used to store backup

flatter formate

doesnot have hdfs compatability

Big data storage

Hierarchical format
HDFS compatibility
File level access control

In your project in Data Lake gen2 How many Layers created

Raw
Refined
Archive

brown
silver
gold

-----Azure SQL DB-----

Azure SQL database as a service

What is elastic pool

sharing of resources across multiple database grouped under one pool

how can you provide computation
DTU---data Transaction units
VCore---

Azure SQL managed instance
Azure SQL in virtual Machine

-----dwh concepts-----

what is OLTP vs OLAP

what is DWH

What is Datamart

what is Star schema and snowflake schema

what is dimension table

what is fact table

what are diff kind of dimensions

role playing

conformed

junk

degenerated dimension

what are diff kind of facts

what are diff kind of measures

what is surrogate key

what is fact less fact table

<https://youtu.be/FpYqPC4F5ww>

-----Azure Synapse Dedicated SQL Pool-----

what is diff between dtu vs DWU

DTU--Data Transaction Units---Azure SQL DB

DWU--Data Ware Housing---Synapse dedicated SQL Pool

what is diff between Azure sql DB vs Synapse dedicated sql db

	SQL Database	SQL Data Warehouse
Scalability	Scale Up (Vertical)	Scale Out (Horizontal)
Units of Performance	Database Transaction Units (DTU)	Data Warehouse Units (DWU) and Compute DWU (cDWU)

Max Size on disk)	4TB	Up to 1 Petabyte with compression (240TB
Designed for	OLTP	OLAP
Architecture	SMP MPP	
Concurrent Sessions	30000	1024
Concurrent Queries	6400	32
PolyBase Support	NO	YES

what is archeture of synapse dedicated sql pool

Controle Node

Compute Nodes

MPP engine(Massive Parlell Processing Engine)

DMS(Data Movement Service)

Distributions

Round Robin---less data dimension tables---customer,product..

Hash---fact big volume --sales

Replication---lookup tbales--citycode,cityname

what is diff between serverless and dedicated sql pool

Architecture:

Serverless SQL pool:no need to provision of resources at the time of creation.

Dedicated SQL pool: need to specify no of DWU at the time creation

Scaling:

Serverless SQL pool: It automatically scales up and down

Dedicated SQL pool: It requires manual provisioning of compute resources

Cost:

Serverless SQL pool: You are billed based on the amount of data processed.

Dedicated SQL pool: You are billed based on the provisioned compute resources

Querying and Workload Type:

Serverless SQL pool: It supports querying structured and unstructured data using T-SQL.

Dedicated SQL pool: It provides a full-featured SQL engine and supports complex analytical queries.

Concurrency:

Serverless SQL pool: It supports multiple concurrent queries but has some limitations on concurrency

Dedicated SQL pool: It provides dedicated compute resources, allowing higher concurrency and better performance for concurrent queries.

Data Storage:

Serverless SQL pool: Data Lake Storage Gen2.

Dedicated SQL pool: It stores data in dedicated storage, in columnar format

what are different kinds of distribution methods available

round robin---dimension

hash---high volume table like fact tables

replicate---very very small lookup tables

what is max compute nodes---60 compute nodes

what is CTAS

CREATE TABLE new_table

```
[WITH (distribution = round_robin, ...)]
```

```
AS
```

```
SELECT *
```

```
FROM source_table
```

what is polybase , steps to implement

create master key encryption by password='Appuro@100'

```
CREATE DATABASE SCOPED CREDENTIAL BlobStorageCredential
```

```
WITH
```

```
    IDENTITY = 'ram',
```

```
    SECRET =
```

```
'+GKcvqd8BuwNdr3AVcB/KgBIXe34Mhwxn2Xt2ZU61woa9DLAziCuB0566oTYydO  
HCy9FQvYvJzVP+AStNc95vw==';
```

```
CREATE EXTERNAL DATA SOURCE BlobStorageDataSource
```

```
WITH (
```

```
    TYPE = HADOOP,
```

```
    LOCATION = 'wasbs://customer@eclasesscustomerinfo.blob.core.windows.net',
```

```
    CREDENTIAL = BlobStorageCredential
```

```
);
```

```
CREATE EXTERNAL FILE FORMAT TextFileFormat
```

```
WITH (
```

```
    FORMAT_TYPE = DELIMITEDTEXT,
```

```
    FORMAT_OPTIONS(
```

```
        FIELD_TERMINATOR = ',',
```

```
        STRING_DELIMITER = '"',
```

```
        FIRST_ROW = 2
```

```
)  
);
```

```
CREATE EXTERNAL TABLE extcustomer  
(  
    cno INT,  
    Cname VARCHAR(100),  
    sal int,  
    city varchar(100)  
)  
WITH (  
    LOCATION = '/cust',  
    DATA_SOURCE = BlobStorageDataSource,  
    FILE_FORMAT = TextFileFormat  
)  
);
```

```
select * from extcustomer
```

-----Data Factory-----

- 1.What is ETL
- 2.What are diff between SSIS vs Data Factory
- 3.what are diff versions of Data Factory V1 and V2
- 4.What are Major Components of Azure Data Factory.
 - 1.Integration Runtimes
 - 2.Linked Services

3.Datasets

4.Activities

5.Pipelines

6.Triggers

5.What are different types of Integration Runtimes available.

Providing compute infrastructure to execute pipeline activities and connecting to different networks using linked services.

1.Self Hosted--Connecting to on premises or private Networks or Virtual Machine

2.Auto Resolved or Azure IR--Cloud Based,Auto Resolve is Default one,Azure IR is created developer You can Customize

3.SSIS IR--SSIS Project in On premises, If you want to shift into cloud and scheduling inside Data Factory

6.What is Linked services.

Linked services are storing of connection information and establishing connection to datasources with help of IR

7.what is Datasets

Datasets are preparation of data with help of Linked services based on activity.

8.What are activities used in Data Factory in Your Project.

Copy Data Activity--Simply Copying Data from source to Destination

Get Metadata Activity--To Get metadata information like filename in folder,file exist or not,size of file

Stored Procedure Activity--To execute Any sql commands in Stored Procedure,You can call Procedure.

Web Activity--To send email notifications with help of logic apps

For Each Loop Activity--to loop any activities multiple times based on input.

Lookup Activity--to extract data from any datasource and return that output as input to other activity

If Condition Activity--based on condition expression true or false

Set Variable--to assign any value to variable

Switch Activity--multiple conditional based execution of activities

Data Flows--to perform Extraction and Transformation Loading data from source to destination

Wait---to stop execution of pipeline some time.

9. Did you use Mapping Data Flows, What are transformations?

Source--connecting to any kind of data sources supported by data factory

1. Derived Column--To derive new columns based on expressions

2. Conditional Split--Loading data by splitting based on condition

3. Join--Performing joins based on common column like inner, left, right, full, cross

4. Lookup--comparing two datasources based on common column find matched and unmatched records

5. Select--filter of columns, restricting of incoming columns

6. Filter--filtering of Data or rows

7. Sort--sorting data based on column in asc and desc order

8. Union--combining or appending of two datasources

9. Window---to perform row number, rank, dense rank expressions

10. Aggregate---to group data and performing aggregations like sum, min, max.....

11. Alter Row---identifying which records need to insert and need to update..performing upsert operations

13. New Branch--incoming data can be duplicated into multiple destinations

14. Surrogate Key---to generate continuous key values

Sink---Loading data into destination supported by data factory

10. Can we connect Data flows into on premises

No, bcz self hosted IR will support so needs to use copy data activity then copy data and then load into

blob or data lake then use data flow to perform ETL

11. inside for each loop activity one more for each nothing but nested for each.

you cant use nested for each loop

so alternative solution is call pipeline using execute pipeline activity and use another for each

12. can we use nested if condition in same pipeline

You cant

so alternative solution use execute pipeline activity inside another pipeline

13. Scenarios:

1. How to copy data from on premises sql server to cloud blob storage

2. How to copy blob to blob

3. How to copy multiple files into azure sql single table

4. How to implement incremental Loading

5. How to copy data when file exist else send email notification

7. How to send email notifications in Data Factory

8. How to implement SCD1 using mapping Data Flows

11. what is your Data Factory pipeline flow

we are taking Data from source on premises SQL Server

Copying into Staging area of blob or Data lake like csv files

Then by using mapping data flows we are performing SCd1 mechanism like insert and update on target azure sql db

Then power bi team designing of reports and dashboards

12.How to schedule Pipelines

by using triggers

1.scheduling

2.event based trigger

3.tumbling window--historical manner to execute pipeline

source sql database--sales--last 1 year--month by month sales

Jan-01-2023--31-Jan-2023

Feb-01-2023--28-Feb-2023

13.deployment of pipeline

we are using azure git repo and deploying pipelines using cicd release pipelines