****

**Munavath Venu**

**Azure Data Engineering Interview Questions & Answers**

* **SQL SERVER**
* **Azure Data Factory**
* **Azure Synapse Analytical Services**
* **Azure Synapse Data Warehouse**
* **Azure Data Bricks**
* **Azure CICD/Agile Methodology**
* **High level Interview Questions**

**Contact:+91 9951103370**

***Special Thanks by Ram Boyapati***

***Dr.Ram Bathula***

**Azure data Engineering with SQL Real Time Preparation**

--------------------------------------------------------------------------------------

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

Master.

Model

MSDB

TempDB

.........................................................................................

**what are the recovery model. in sql**

**1,Simple Recovery Model**

**2,Full Recovery Model**

**3,Bulk-Logged Recovery Model**

**1.what are DML. and DDL. and DCL. Commands**

in sql server we have dml and ddl commands

DML--data manipulation language

insert,update,delete,select

DDL--Data definition language

Create ,Alter,Drop,Truncate

DCL:data control language command

grant,rename,revoke

----------------------------------------------------------------------------------------------

**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=10)

Identity property will not persist to the starting value when we delete records from the table.

It rollback the data

**Truncate:**

Truncate is DDL Command and it will delete only data inside the table, not the table structure columns .

Performance is good

It will not create any log Data

truncate table employee

By using truncate it will delete all the records at a time.

identity property value persist to starting value when we delete record from table(duplicate)

it does not rollback the data

**Drop:** it will remove all object like (data base ,tables,and sp and views)

--------------------------------------------------------------------------------------------------------------

3.**what is diff between where Clause. and Having. Clause**

Where Clause: condition filtering rows on a table columns

**select \* from emp where dno=10**

**Having:**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:** union all combines data even doesn't removing duplicate

union all is more performance

**Union:** union combines data and removing duplicate

union is slow in performance because it needs to remove duplicate

table1

------------------------------------------------------------------------------------------

eno,ename,sal,dname

table2

--------------------------------------------------------------------------

eno,ename,sal

**QUERY:**

select eno,ename,sal,dname from emp1

union all or Union

select eno,ename,sal,'' as dname from emp2

=======================================================================

**5.what are diff kind of constraint. IN SQL**

**Constraints:** it will apply some conditions based on columns.

not null

default

check.

primary key

unique key

foreign key

-------------------------------------------------------------------------------------------------

**6.what is diff between Primary Key. and Unique Key. and foreign key**

**Primary key.**

only one primary key can be created ON COLUMN of the table

No duplicates and No null values in primary key column

**Unique key.:**

Many unique keys can be created

No duplicates But only single null value

**foreign key.:**it allow duplicate value and it referring the primary key column of other table

\*it combine two tables of data

----------------------------------------------------------------------------------------------------

**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 OF TWO TABLES MEANS COMBINATION OF TWO TABLES

select \* from table1,table2

**self\_join**-- the table join itself is known as self join

select a.ename as employee,b.ename as manager from emp as a inner join emp as b on a.MGR=b.EMPNO

-----------------------------------------------------------------------------------------------------------------------

**8.what are temporary. tables**

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

2 types

**1, LOCAL.--#**

It indicates hash(#) symbol

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

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

-------------------------------------------------------------------------------------------------------------------------------

**2, GLOBAL.--##**

It indicates double hash(##) symbol

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

insert into ##EMP values (1,’VENU’,300)

WITH IN CONNECTION ONCE CONNECTION CLOSE TABLE WILL BE DROPPED

– **-insert into #localemp**

**select \* from ##globalemp**

**—------------------------------------------------------------------------------------**

**we have oldtable than how can create newtable**

**select \* into newtable from oldtable**

----------------------------------------------------------------------------------------

**9.what is view. and what is materialized view**

VIEW IS IMAGINARY TABLE ,IT DOESN'T STORE ANY DATA.IT WILL JUST STORE QUERY VIEWS CAN BE USED TO UPDATE, DELETE OR INSERT WHEN WRITING SIMPLE SELECT STATEMENT INSIDE THE VIEW, IT WILL AFFECT BASE TABLE.

create view priyaview

as

select \* from emp

select \* from priyaview

**MATERIALIZED VIEW:**

WHEN WE CREATING INDEX ON VIEW ,IT WILL OCCUPY SOME MEMORY AND DATA IS KNOWN MATERIALIZED VIEW.

A materialized view is physically stored in a database.This stored data is refreshed to keep it up to date with the source data.

-----------------------------------------------------------------------------------------------------------

**11.what are diff between temp table ,TABLE VARIABLE.**

**Table variable.--**-Temp variable does not store any kind of data in database

It has in memory RAM memory

It is not eligible to create indexes

It store small value of data because optimization is not available in table

**Temp table.**---are storing of data in temp db

local ---existence of local is with in session

Global---with in connection

**Temp tables:** it stores any kind of data in temp database

It uses harddisk memory

It is eligible to create indexes

It stores huge amount of data because optimization is available in table

**10. what is Cte. and table variable**

**CTE.** : common table expression

it contains in memory

it have recursive feature (the result cte used in same cte)

It will save the query result,when you want it will immediately return the value.

Date Table

--------------------

Datekey,month,year.........date

DECLARE @StartDate DATE = '2003-01-01';

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

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

WITH cte AS (

SELECT @StartDate AS CurrentDate

UNION ALL

SELECT DATEADD(DAY, 1, CurrentDate)

FROM cte

WHERE CurrentDate <= @EndDate

)

SELECT CurrentDate,year(CurrentDate) as Yearname,Month(CurrentDate) as Month,Datename(mm,CurrentDate) as MonthName,DATENAME(DW,CurrentDate) AS DAYS

FROM cte

OPTION (MAXRECURSION 0);

**12.what is diff between stored. procedure. and Function.**

**Stored procedure:** it may or may not return data.

Stored procedure can have error handling Try and catch blocks

stored procedure can be call to function and execute inside function

stored procedure cannot join with tables

Stored procedure allows input and output parameter

Inside the Stored procedure we are using temp tables

**Function.** : Function will always return data.

Function will not have error handling Try and catch block

function can not be call and execute inside stored procedure

Function can join with table

Function have only input parameter

Inside the function we do not use temp tables

------------------------------------------------------------------------------------------

**13.what are diff kind of index. and types available**

**Index:** it improves query performance by creating pointers and tables

Index is used by queries to find data from tables

indexes are created on tables and views

Index can be created on table to query data more faster

2 types of indexes

Clustered. indexes

Non cluster Indexes

**Clustered indexes** —

only one clustered index can be created on single table

Data will be sorted on cluster index

It is a B Tree structure---leaf nodes it contains actual data.

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

leaf nodes it does not contain actual data, it contains key values

Row index

columnar storage index

------------------------------------------------------------------------------------------------------------

**14.How to improve SQL Server query performance or optimization in sql or tuning in sql**

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 indexes 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 Stored Procedures:**

Stored procedures can be precompiled and stored in the database, reducing overhead and improving performance for frequently executed queries.

**Normalize Database Design:**

* Normalize your database schema to reduce redundancy and improve data integrity. However, be cautious not to over-normalize, as it can lead to more complex queries.

**Partitioning:**

* If dealing with large tables, consider partitioning them based on specific criteria. This can improve query performance by allowing the database to scan only relevant partitions.

--------------------------------------------------------------------------------------------------------------------

**15.what are windows functions and what diff between Rank() and dense\_Rank(),rownumber.**

**RowNumber**:it assign unique rank without any duplicate

**Rank.:** it assign duplicate rank and it will maintain gaps when same record have in table

**Dense\_Rank :**it assign duplicate rank and it will does not maintain gaps when same record have in table

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

-----writing. order---

select---1

from---2

where---3

group by---4

having---5

order by---6

-------------------**execution.**--------------

from---1

where---2

group by---3

having---4

select--5

order by---6

------------------------------------------------------------------------------------------

**17.what is diff between isnull.,nullif.,coalesce.**

select \*,isnull(commission,10) from employee where commission 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

**coalesce()**---coalesce function is used to return the first non-null expression among its arguments.It takes multiple parameters and returns the value of the first non-null expression in the list

or

first non null value among multiple columns

select customer,coalesce(mobile number,land number,office number) as contact from table

-------------------------------------------------------------------------------------------------------------

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

Query within another query is known as subquery

**NON QUERY DATA:**

IT is one directional query,Non correlated sub query will execute first inner query then outer query will execute

Its performance is better than correlated query.

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

the performance correlated query not good compared to non\_correlated query

**select \* from emp as e where 1=(select count(distinct sal) from emp as z where e.sal<=z.sal)**

**What is a cursor.**

It is a simple temporary sql memory area

**We have two types**

1.Implicit Cursor

2.Explicit cursor

**What is Normalization.**

It is used to reduce the redundancy and dependency of the data

**check if the mail address is valid or not ?**

select \* from emp where email like '%@%.%'

**What is lead. and lag. in the sql server?**

**LEAD()**: This function allows you to access the value of a specified column from the next row within the current result set

**LAG():** This function allows you to access the value of a specified column from the previous row within the current result set

—-------------------------------important query—-----------------------------

select \* from sys.databases ------display all databases

select \* from sys.tables-----display all tables

select \* from sanju.sys.tables----displays only selected database tables

select \* from sys.views------display all views

select \* from sys.columns----displays all columns

select \* from sanju.sys.columns----displays all columns in selected database

select \* from INFORMATION\_SCHEMA.tables---it will give information about selected database tables

select \* from INFORMATION\_SCHEMA.columns--it will give information about selected database columns

-----------------return all columns names from table--------------------

SELECT COLUMN\_NAME

FROM INFORMATION\_SCHEMA.COLUMNS

WHERE TABLE\_NAME = 'DimCustomer'

--------------identify column name—--------------------

SELECT a.TABLE\_SCHEMA,b.TABLE\_NAME,COLUMN\_NAME,DATA\_TYPE

FROM INFORMATION\_SCHEMA.TABLES as a inner join INFORMATION\_SCHEMA.COLUMNS as b

on a.TABLE\_NAME=b.TABLE\_NAME and a.TABLE\_SCHEMA=b.TABLE\_SCHEMA

WHERE COLUMN\_NAME LIKE '%firstname%' and TABLE\_TYPE = 'BASE TABLE'

**Queries practice :**

-------------------------------**how to remove duplicate.**-------------------------

CREATE TABLE Duplicates (

ID INT IDENTITY(1,1),

Value VARCHAR(50)

);

INSERT INTO Duplicates (Value) VALUES ('Apple');

INSERT INTO Duplicates (Value) VALUES ('Banana');

INSERT INTO Duplicates (Value) VALUES ('Cherry');

INSERT INTO Duplicates (Value) VALUES ('Apple');

INSERT INTO Duplicates (Value) VALUES ('Grape');

INSERT INTO Duplicates (Value) VALUES ('Orange');

INSERT INTO Duplicates (Value) VALUES ('Apple');

INSERT INTO Duplicates (Value) VALUES ('Strawberry');

INSERT INTO Duplicates (Value) VALUES ('Grape');

INSERT INTO Duplicates (Value) VALUES ('Apple');

select \* from Duplicates

--------------- Query to retrieve duplicate. values

**select distinct value from Duplicates**

or

**select value,count(\*) as bb from Duplicates group by (value) having count(\*)>1**

or

**with empcte**

**as**

**(**

**select \*,ROW\_NUMBER() over(partition by value order by value ) as re from Duplicates**

**)**

**delete from empcte where re>1**

=======================================================================

**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;

**6.Convert Unpivoted Data into Pivot. format**

create table emp\_po(eno int identity(1,1),City varchar(90),year varchar(90),sales money)

insert into emp\_po values('hyd',2022,1000),

('hyd',2023,1000),

('Bang',2022,3000),

('Bang',2023,2000)

select \* from emp\_po

expected Result

---------------

City,2022,2023

Hyd,1000,1000

Bang,3000,2000

SELECT \*

FROM (

SELECT City, [Year], Sales

FROM YourTable

) AS vv

PIVOT (

SUM(Sales)

FOR [Year] IN ([2022], [2023])

) AS rr;

—----------------------------- **in spark--**------------------------------------------------------------------------------

Input:df assume

----------------------------

| Product | Amount | Country |

|---------|--------|---------|

| Banana | 1000 | USA |

| Carrots | 1500 | USA |

| Banana | 400 | China |

Output:

------------------------

| product | China | USA |

|---------|-------|------|

| Banana | 400 | 1000 |

| Carrots | null | 1500 |

----------------------------------------------------------------------------------

**query in pyspark:**

new\_df=df.groupby("Product").pivot(" Country").SUM("Amount").SHOW()

Display(new\_df)

======================================================================

7.create table emp\_gender(empid int,ename varchar(90),gender varchar(90)

insert into emp\_gender values(1,'Ram','Male'),

(2,'radha','Female')

select \* from emp\_gender

**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;

----------------------------------------------------------------------------------------

**We have 2 column here , generate debit when it debited,generate credit when it credited**

-- Create the transactions table

CREATE TABLE transactions (

ename VARCHAR(255),

amount INT

);

-- Insert sample data into the table

INSERT INTO transactions (ename, amount)

VALUES

('Alice', 100),

('Bob', -50),

('Charlie', 200),

('David', -75),

('Eva', 150);

-- Generate debit and credit columns based on conditions

SELECT ename,

amount,

CASE WHEN amount > 0 THEN amount ELSE 0 END AS debit,

CASE WHEN amount < 0 THEN -amount ELSE 0 END AS credit

FROM transactions;

+---------+--------+-------+-------+

**output:**

| ename | amount | debit | credit|

+---------+--------+-------+-------+

| Alice | 100 | 100 | 0 |

| Bob | -50 | 0 | 50 |

| Charlie | 200 | 200 | 0 |

| David | -75 | 0 | 75 |

| Eva | 150 | 150 | 0 |

+---------+--------+-------+-------+

—------------------------------------------------------------------------------------------------------------------------------ Create a new table called 'transactions'

CREATE TABLE transactions (

id INT,

amount INT,

transaction\_type VARCHAR(10)

);

-- Insert 8 rows of data into the 'transactions' table

INSERT INTO transactions (id, amount)

VALUES

(1, 100),

(2, -50),

(3, 200),

(4, -75),

(5, 0),

(6, 150),

(7, -30),

(8, 0);

-- Update the 'transaction\_type' column using a CASE statement

UPDATE transactions

SET transaction\_type = CASE

WHEN amount > 0 THEN 'Credit'

WHEN amount < 0 THEN 'Debit'

ELSE 'No Transaction'

END;

SELECT \* FROM transactions;

**8.find employee in each department who ever having there salaries**

**more then average salary of department**

select e.\* from emp as e where e.sal<(select avg(sal) from emp where deptno=e.deptno)

select e.\* from emp as e where e.sal>(select avg(sal) from emp where deptno=e.deptno)

—-----------------------------------------------------------------------------------------------------------------

**cricket. match schedule:**

**create table matchs(country varchar(90))**

**insert into matchs values('ind'),('sl'),('eng'),('nz'),('aus'),('pak'),('nepal')**

**select \* from matchs**

**select a.country+ ' vs ' +b.country as shcdule from matchs as a join matchs as b on a.country>b.country**

---------------Miscellaneous. Questions---------------------------

1.can i insert data into View::yes

2.Can i call stored procedure in Function::no

3.Can i restore data once i truncated::no

4.Can i insert values into identity column::yes

5.can i delete row contains foreign key reference::yes

6.Can i create more than one clustered index on single table::no

7.Can i create more then one primary key on single table::no

8.can we write CTE in VIEW:YES

9.can we write VIEW in CTE:NO

10.can we call foreach loop in another for each:No

11. can we call nested loop inADF:NO

---------------------------------

select \* from emp

select \* from dept

**---------------------------------------cumulative. query------------------------**

select deptno,ename,sal,sum(sal) over(order by sal ) as cumm from emp

select deptno,ename,sal,sum(sal) over( partition by deptno order by sal ) as cumm from emp

**------------------------- to get above and below sal particular. employee----------------------------**

select \* from emp where sal <(select sal from emp where empno=7698) order by sal

select \* from emp where sal >(select sal from emp where empno=7698) order by sal

**---------------------------------how to get max salary from table-------------------------**

select \* from emp where sal=(select max(sal) from emp)

select max(sal)as maxsal from emp

select top 1 \* from(select top 1 \* from emp order by sal desc)a order by sal asc

with priyacte

as

(

select \*,ROW\_NUMBER() over(order by sal desc) as rr from emp

)

select \* from priyacte where rr=1

select\* from(select \*,ROW\_NUMBER() over(order by sal desc) as rm from emp)z where rm=1

select \* from emp as e where 1=(select count(distinct sal) from emp as z where e.sal<=z.sal)

**---------------------------how to get 2nd highest. sal---------------------------------------------------**

select max(sal)as secondhigh from emp where sal<(select max(sal) from emp)

select top 1 \* from(select top 2 \* from emp order by sal desc)a order by sal asc

with priyacte

as

(

select \*,dense\_rank() over(order by sal desc) as rr from emp

)

select \* from priyacte where rr=2

with priyacte

as

(

select \*,dense\_rank() over(partition by dept order by sal desc) as rr from emp

)

select \* from priyacte where rr=2

select\* from(select \*,dense\_rank() over(order by sal desc) as rm from emp)z where rm=2

select \* from emp as e where 2=(select count(distinct sal) from emp as z where e.sal<=z.sal)

**each dept.**

with cte

as

(

select \*,dense rank() over(partition by dept order by sal desc)as eachdepthighsal from emp as e inner join dept as d on e.id=d.id

)

select \* from cte where eachdepthighsal=2

**--------------------in table the employee whose sal greater or less of avg**

select avg(sal) from emp

select \* from emp where sal <(select avg(sal) from emp)

select \* from emp where sal >(select avg(sal) from emp)

**------------------------avg of each. department less or greater each employee----------------------**

select deptno, avg(sal) as xyz from emp group by DEPTNO

select e.\* from emp as e where e.sal<(select avg(sal) from emp where deptno=e.deptno)

select e.\* from emp as e where e.sal>(select avg(sal) from emp where deptno=e.deptno)

========================================================================

-----------------------practice-sql-----------------------------------------------------------------

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')

-----------------------------------------------------------------------------

select \* from emp

select \* from dept

select \* from emp where sal <2450

select \* from emp where sal >2450

select \* from emp where sal between 1250 and 3000

select \* from emp where job='salesman'

select \* from emp where job='salesman' or job='manager'

select \* from emp where comm is null

select \* from emp where comm is not null

—-------------------------------------------------------------------------------------

-- Create the dept (department) table

CREATE TABLE dep (

dept\_id INT ,

dept\_name VARCHAR(50)

);

-- Insert values into the dept table

INSERT INTO dep (dept\_id, dept\_name)

VALUES

(1, 'HR'),

(2, 'Finance'),

(3, 'Engineering'),

(4, 'Marketing'),

(5, 'Sales'),

(6,'azure');

select \* from dep

-- Create the emp (employee) table

CREATE TABLE em (

emp\_id INT ,

emp\_name VARCHAR(50) NOT NULL,

emp\_salary DECIMAL(10, 2),

dept\_id INT,

hire\_date DATE

);

-- Insert values into the emp table

INSERT INTO em (emp\_id, emp\_name, emp\_salary, dept\_id, hire\_date)

VALUES

(101, 'John Smith', 60000.00, 3, '2020-01-15'),

(102, 'Alice Johnson', 55000.00, 2, '2019-05-20'),

(103, 'Michael Brown', 70000.00, 9, '2021-02-10'),

(104, 'Emma Davis', 62000.00, 1, '2018-11-30'),

(105, 'David Wilson', 58000.00, 3, '2022-03-05'),

(106, 'Olivia Lee', 62000.00, 7, '2019-08-12'),

(107, 'James Taylor', 54000.00, 2, '2020-07-25'),

(108, 'Sophia White', 63000.00, 5, '2017-04-18'),

(109, 'Daniel Miller', 71000.00, 1, '2021-09-03'),

(110, 'Ava Martinez', 59000.00, 5, '2019-06-28'),

(111, 'Liam Jones', 62000.00, 3, '2022-01-08'),

(112, 'Mia Harris', 56000.00, 7, '2018-12-14'),

(113, 'Noah Clark', 69000.00, 5, '2020-04-22'),

(114, 'Isabella Turner', 61000.00, 2, '2021-11-17'),

(115, 'Ethan Moore', 55000.00, 1, '2019-10-01'),

(116, 'Charlotte Garcia', 64000.00, 3, '2022-05-09'),

(117, 'Liam Jackson', 60000.00, 2, '2020-08-14'),

(118, 'Ava Wilson', 67000.00, 5, '2018-07-03'),

(119, 'William Anderson', 59000.00, 1, '2021-03-12'),

(120, 'Harper Martinez', 71000.00, 5, '2019-09-19');

select \* from em

—-------------------------------------------------------------------------------

select \* from dept as a inner join emp as b on a.DEPTNO=b.DEPTNO

select \* from dept as a left join emp as b on a.DEPTNO=b.DEPTNO

select \* from dept as a right join emp as b on a.DEPTNO=b.DEPTNO

select \* from dept as a full join emp as b on a.DEPTNO=b.DEPTNO

select \* from venu,emp

--------------------------------------------------------

aggregation

---------------------------------------------------

max()

min()

count()

avg()

sum()

-------------------------------------------------------------------------------------------

identity(1,1)

create table venue(vno int identity(1,1),vname varchar(90))

select \* from venue

insert into venue values('anu'),

('siva'),

('irfan')

delete from venue where vno=3

insert into venue values('priya')

truncate table venue

insert into venue values('manu'),

('begum'),

('syed')

---------------------------------------------------------------------------------------------------------

union and union all

select \* from venue

create table manu(vno int,vname varchar(90))

insert into manu values(4,'venu'),(5,'manu')

select \* from manu

select \* from venue

---------------------------------------------------------

SELECT \* FROM manu

UNION ALL

SELECT \* FROM venue;

-----------------------------------------------

SELECT \* FROM manu

UNION

SELECT \* FROM venue;

---------------------------------------------------------------------------

**----------------------------union and union all --------------------------------**

create table venu(eno int,ename varchar(90),ecity varchar(90))

insert into venu values (1,'venky','krnl'),(2,'mono','hyb'),(3,'sai','vzg')

select \* from venu

create table manu(eno int,ename varchar(90),ecity varchar(90))

insert into manu values (1,'venky','krnl'),(2,'teja','hyb'),(3,'geetha','vzg')

select \* from manu

select \* from venu

union

select \* from manu

select \* from venu

union all

select \* from manu

—-------------------------------------------------------------------------------------------------

select \* from emp

select \* from dept

select \* from emp as a right join dept as d on a.DEPTNO=d.DEPTNO where a.DEPTNO is null

result\_df = emp\_df.alias("a").join(dept\_df.alias("d"), emp\_df.DEPTNO == dept\_df.DEPTNO, "right") \ .where(emp\_df.DEPTNO.isNull())

—-------------------------------------------------------------------------------

select \* from emp as a left join dept as d on a.DEPTNO=d.DEPTNO where d.DEPTNO is null

=-----------------------------------------------------------------------------------------------

select \* from emp

SELECT \* FROM DEPT

-----**-----------SUB QUERY TO GET SPECIFIC COLUMNS-----------------**---------------------

select \* from (select job,mgr,sal,deptno from emp)E

-**------TO GET PARTICULAR COLUMN IN BOTH TABLE**

SELECT A.DEPTNO,A.JOB,B.DNAME FROM EMP AS A INNER JOIN DEPT AS B ON A.DEPTNO=B.DEPTNO

**WE DO NOT PURCHASE ANY ITEMS IN DEMART-------**

select \* from emp as a right join dept as d on a.DEPTNO=d.DEPTNO where a.DEPTNO is null

or in pyspark

result\_df = dept.alias("d")\ .join(emp.alias("a"), ["DEPTNO"], "right")\ .where(emp.DEPTNO.isNull())

select \* from dept where DEPTNO not in(select DEPTNO from emp)

**EMPLOYEE DO NOT PURCHASE ANY ITEM IN MARKET------------------------**

select \* from emp as a left join dept as d on a.DEPTNO=d.DEPTNO where d.DEPTNO is null

select \* from EMP where DEPTNO not in(select DEPTNO from DEPT)

**-------------------------------AGGREGATION------------------------**

select \* from emp

SELECT \* FROM DEPT

SELECT MAX(SAL) AS MAXSALINTABLE FROM EMP

SELECT MIN(SAL) AS MINSALINTABLE FROM EMP

SELECT AVG(SAL) AS MINSALINTABLE FROM EMP

SELECT COUNT(EMPNO) AS MINSALINTABLE FROM EMP

SELECT SUM(SAL) AS MINSALINTABLE FROM EMP

—---------------------------------

from pyspark.sql import SparkSession

from pyspark.sql.functions import max, min, avg, count, sum

# Create a Spark session

spark = SparkSession.builder.appName("AggregateExample").getOrCreate()

# Sample data for demonstration

emp\_data = [(1, "John", 1000), (2, "Jane", 1200), (3, "Alice", 1500), (4, "Bob", 1300)]

# Create a DataFrame

emp\_columns = ["EMPNO", "ENAME", "SAL"]

emp\_df = spark.createDataFrame(emp\_data, emp\_columns)

# Calculate the required aggregate values

max\_sal = emp\_df.select(max("SAL").alias("MAXSALINTABLE")).collect()[0]

min\_sal = emp\_df.select(min("SAL").alias("MINSALINTABLE")).collect()[0]

avg\_sal = emp\_df.select(avg("SAL").alias("AVGSALINTABLE")).collect()[0]

count\_empno = emp\_df.select(count("EMPNO").alias("COUNTINTABLE")).collect()[0]

sum\_sal = emp\_df.select(sum("SAL").alias("SUMSALINTABLE")).collect()[0]

-------------------------------------------------

SELECT DEPTNO, MAX(SAL) AS MAXSALINTABLE FROM EMP GROUP BY DEPTNO

Or in pyspark

max\_sal\_per\_dept = emp\_df.groupBy("DEPTNO").agg(max("SAL").alias("MAXSALINTABLE"))

—-

SELECT DEPTNO, COUNT(\*) AS MAXSALINTABLE FROM EMP GROUP BY DEPTNO

SELECT A.DEPTNO,COUNT(\*) AS TCOUNT FROM EMP AS A INNER JOIN DEPT AS D ON A.DEPTNO=D.DEPTNO GROUP BY A.DEPTNO

SELECT \* FROM EMP AS A INNER JOIN DEPT AS D ON A.DEPTNO=D.DEPTNO WHERE A.JOB='CLERK'

SELECT \* FROM EMP AS A INNER JOIN DEPT AS D ON A.DEPTNO=D.DEPTNO

SELECT DEPTNO, MAX(SAL) AS MAXSALINTABLE

FROM EMP

WHERE SAL > 2900

GROUP BY DEPTNO;

**---TO PRACTICE IN PLACE OF COUNT-------------WE PLACE MAX,AVG,MIN,SUM------**

SELECT DEPTNO, COUNT(\*) AS MAXSALINTABLE FROM EMP GROUP BY DEPTNO

SELECT A.DEPTNO,COUNT(\*) AS TCOUNT FROM EMP AS A INNER JOIN DEPT AS D ON A.DEPTNO=D.DEPTNO GROUP BY A.DEPTNO

or in pyspark

result\_df = EMP.alias("A")\ .join(DEPT.alias("D"), col("A.DEPTNO") == col("D.DEPTNO"))\ .groupBy("A.DEPTNO")\ .agg({"\*": "count"})\ .withColumnRenamed("count(1)", "TCOUNT")

**------------AT A TIME WE WRITE MAX,MIN,SUM------------------------------------------**

SELECT MAX(SAL) AS MAXSAL,MIN(SAL) AS MINSAL,SUM(SAL) AS TOTALSAL FROM EMP

---------------------------------

**--------------------how to know to date--------------------------**

select getdate() as aa

select getutcdate() as aa

select year(getdate()) as aa

select month(getdate()) as aa

select day(getdate()) as aa

select datename(mm,getdate()) as aa

select datename(dw,getdate()) as aa

select year(getdate()) as aa,month(getdate()) as aa,day(getdate()) as aa, datename(mm,getdate()) as aa,datename(dw,getdate()) as aa

--------------------------date add--------------------------------------------------------

select DATEADD(dd,1,getdate()) as xx

select DATEADD(mm,2,getdate()) as xx

select DATEADD(yy,2,getdate()) as xx

select DATEADD(mm,-2,getdate()) as xx

------------**datedifferent**-------------------------------------------------------

select DATEDIFF(yy,'1947-08-15',getdate()) as ff

select DATEDIFF(mm,'1947-08-15',getdate()) as ff

select DATEDIFF(dd,'1947-08-15',getdate()) as ff

**--------------------------dob and doj sample data------------------------------------**

-- Create a table to store employee data

CREATE TABLE Employee\_date (

EmployeeID INT IDENTITY(1,1) PRIMARY KEY,

EmployeeName NVARCHAR(100),

DOB DATE,

DOJ DATE

);

-- Insert up to 10 records

INSERT INTO Employee\_date (EmployeeName, DOB, DOJ)

VALUES ('John Doe', '1990-05-15', '2010-08-20'),

('Jane Smith', '1985-02-28', '2009-07-10'),

('Mike Johnson', '1992-11-10', '2015-04-05'),

('Sarah Davis', '1988-09-03', '2012-11-30'),

('David Wilson', '1995-06-20', '2018-02-15'),

('Emily Thompson', '1991-03-12', '2013-09-25'),

('Michael Brown', '1987-12-07', '2011-05-01'),

('Jessica Lee', '1994-08-18', '2017-01-10'),

('Andrew Miller', '1989-04-25', '2014-06-05'),

('Olivia Clark', '1993-01-08', '2016-11-20');

select \* from Employee\_date

**---------------------------how to get experience. of employee------------------------------------**

select \*,DATEDIFF(yy,doj,getdate()) as uu from Employee\_date

select \*,DATEDIFF(yy,doj,getdate()) as uu from Employee\_date where DATEDIFF(yy,doj,getdate())>10

select EmployeeName,DATEDIFF(yy,doj,getdate()) as uu from Employee\_date where DATEDIFF(yy,doj,getdate())>10

**---------------------how to get age of employee---------------------------**

select employeename,DATEDIFF(yy,dob,getdate()) as vv from Employee\_date

select employeename,DATEDIFF(yy,dob,getdate()) as vv from Employee\_date where DATEDIFF(yy,dob,getdate())<30

**---------------------------how to get exper and age ----------------------**

select \*,DATEDIFF(yy,dob,getdate()) as vv,DATEDIFF(yy,doj,getdate()) as exvv from Employee\_date

select \*,DATEDIFF(yy,dob,getdate()) as vv,DATEDIFF(yy,doj,getdate()) as exvv,getdate() as todate from Employee\_date

select EmployeeName,employeeid,DATEDIFF(yy,dob,getdate()) as age,DATEDIFF(yy,doj,getdate()) as expe,getdate() as todate from Employee\_date

**----------------------------------cross join--------------------**

create table cross1(id varchar(90))

insert into cross1 values(1),(2),(null),(3)

select \* from cross1

create table cross2(id varchar(90))

insert into cross2 values(1),(2),(null),(null)

select \* from cross2

select \* from cross1,cross2

SELECT \* FROM CROSS1 AS Z INNER JOIN CROSS2 AS Y ON Z.ID=Y.ID

SELECT \* FROM CROSS1 AS Z LEFT JOIN CROSS2 AS Y ON Z.ID=Y.ID

SELECT \* FROM CROSS1 AS Z RIGHT JOIN CROSS2 AS Y ON Z.ID=Y.ID

SELECT \* FROM CROSS1 AS Z FULL JOIN CROSS2 AS Y ON Z.ID=Y.ID

table\_A TABLE\_B

1 1

2 NULL

1 3

NULL

-------------------------------------------------------------------------

INNER\_JOIN

1,1

1,1

-------------------------------------------------

LEFT\_JOIN

1,1

2,NULL

1,1

NULL,NULL

-------------------------------------------------------------------

RIGHT\_JOIN

1,1

1,1

NULL,NULL

NULL,3

------------------------------------------------------------------

FULL\_JOIN

FULL=(RIGHT+LEFT)-INNER

6

---------------------------------------------------------------------------------------

TABLE\_A TABLE\_B

1 1

2 2

NULL NULL

3 NULL

--INNER\_JOIN

1,1

2,2

----LEFT\_JOIN

1,1

2,2

NULL,NULL

3,NULL

---RIGHT\_JOIN

1,1

2,2

NULL,NULL

NULL,NULL

---FULL\_JOIN

FULL=6

---------------------------**self**.------------------------------------

select \* from em

select a.ename as employee,b.ename as manager from emp as a inner join emp as b on a.MGR=b.EMPNO

**---------------------------temporary tables---------------------------**

------------------local

create table #localemp(eno int,ename varchar(90))

select \* from #localemp

----------------global

create table ##globalemp(eno int,ename varchar(90))

insert into ##globalemp values (1,'hima'),(2,'anu')

select \* from ##globalemp

**------------------------ how to move data global to local------------------------**

insert into #localemp

select \* from ##globalemp

**----------------------------------view-------------------------**

create view priyaview

as

select \* from emp

select \* from priyaview

**----------------------------cte------------------**

with priyacte

as

(select \* from emp)

select \* from priyacte

**—---------cte.---------------**

DECLARE @StartDate DATE = '2003-01-01';

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

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

WITH cte AS (

SELECT @StartDate AS CurrentDate

UNION ALL

SELECT DATEADD(DAY, 1, CurrentDate)

FROM cte

WHERE CurrentDate <= @EndDate

)

SELECT CurrentDate,year(CurrentDate) as Yearname,Month(CurrentDate) as Month,Datename(mm,CurrentDate) as MonthName,DATENAME(DW,CurrentDate) AS DAYS

FROM cte

OPTION (MAXRECURSION 0);

------------------------------------------------------------------------

===================================================================================

Create Table Sale(

Dates Date,

Sale INT

)

Insert into Sale values('2022-01-01',500),('2022-01-02',600),('2022-01-05',700),('2022-01-09',500),('2022-01-13',1300)

Select \* from Sale

DECLARE @StartDate DATE = '2022-01-01';

DECLARE @EndDate DATE = '2022-01-13';

WITH CTE AS (

SELECT @StartDate AS StartKey

UNION ALL

SELECT DATEADD(DAY, 1, StartKey)

FROM CTE

WHERE StartKey < @EndDate

)

SELECT C.StartKey, ISNULL(S.Sale, 0) FROM CTE AS CLEFT JOIN Sale AS S ON

C.StartKey = S.Dates;

**--------------------------------------------window function ,row\_number,rank,dense\_rank------**

CREATE TABLE Marks (

StudentName VARCHAR(50),

Marks INT

);

INSERT INTO Marks (StudentName, Marks) VALUES ('John', 80);

INSERT INTO Marks (StudentName, Marks) VALUES ('Alice', 75);

INSERT INTO Marks (StudentName, Marks) VALUES ('Bob', 90);

INSERT INTO Marks (StudentName, Marks) VALUES ('Sarah', 95);

INSERT INTO Marks (StudentName, Marks) VALUES ('Michael', 95);

INSERT INTO Marks (StudentName, Marks) VALUES ('Emma', 88);

INSERT INTO Marks (StudentName, Marks) VALUES ('David', 93);

INSERT INTO Marks (StudentName, Marks) VALUES ('Olivia', 80);

INSERT INTO Marks (StudentName, Marks) VALUES ('James', 95);

INSERT INTO Marks (StudentName, Marks) VALUES ('Sophia', 93);

INSERT INTO Marks (StudentName, Marks) VALUES ('William', 84);

select \* from marks

select \*,ROW\_NUMBER() over(order by marks desc) as rr from marks

select \*,rank() over(order by marks desc) as rr from marks

select \*,dense\_rank() over(order by marks desc) as rr from marks

select \*,ROW\_NUMBER() over(order by marks desc) as rr,rank() over(order by marks desc) as rm

,dense\_rank() over(order by marks desc) as rl from marks

**-----------sub\_query-------------------------------**

select \* from emp

select ename,job from emp

select \* from(select \* from emp) as dd

select ename from(select \* from emp) as dd

-------------------------------------------------------------------------------

select \* from emp

select \* from dept

**------------------------------- to get above and below sal a particularly employee----------**

select \* from emp where sal <(select sal from emp where empno=7698) order by sal

select \* from emp where sal >(select sal from emp where empno=7698) order by sal

**---------------------------------how to get max salary from table-------------------------**

select \* from emp where sal=(select max(sal) from emp)

select max(sal)as maxsal from emp

select top 1 \* from(select top 1 \* from emp order by sal desc)a order by sal asc

with priyacte

as

(

select \*,ROW\_NUMBER() over(order by sal desc) as rr from emp

)

select \* from priyacte where rr=1

select\* from(select \*,ROW\_NUMBER() over(order by sal desc) as rm from emp)z where rm=1

select \* from emp as e where 1=(select count(distinct sal) from emp as z where e.sal<=z.sal)

**---------------------------how to get 2nd highest. sal----------**

select max(sal)as secondhigh from emp where sal<(select max(sal) from emp)

select top 1 \* from(select top 2 \* from emp order by sal desc)a order by sal asc

with priyacte

as

(

select \*,dense\_rank() over(order by sal desc) as rr from emp

)

select \* from priyacte where rr=2

select\* from(select \*,dense\_rank() over(order by sal desc) as rm from emp)z where rm=2

select \* from emp as e where 2=(select count(distinct sal) from emp as z where e.sal<=z.sal)

**----------------------------WHERE AND GROUP BY AND HAVING -------------------------**

CREATE TABLE MONTH\_DATE(MONTH VARCHAR(50),AMPUNT INT)

INSERT INTO MONTH\_DATE VALUES('JAN 10',100),('FEB 10',200),('MARCH 10',300),('JAN 10',300),('FEB 10',200),('MARCH 10',100)

SELECT \* FROM MONTH\_DATE

---QUERY

SELECT MONTH,SUM(AMPUNT) AS TOTAL\_AMOUNT FROM MONTH\_DATE GROUP BY MONTH

SELECT MONTH,SUM(AMPUNT) AS MONTHS\_WISE FROM MONTH\_DATE GROUP BY MONTH HAVING MONTH='JAN 10' OR MONTH='MARCH 10'

or in pyspark

result\_df = month\_date\_with\_month\_col\ .groupBy("MONTH")\ .agg({"AMPUNT": "sum"})\ .withColumnRenamed("sum(AMPUNT)", "MONTHS\_WISE")\ .filter((col("MONTH") == "JAN 10") | (col("MONTH") == "MARCH 10"))

SELECT MONTH,SUM(AMPUNT) AS MOHLY\_TOTAL FROM MONTH\_DATE WHERE MONTH BETWEEN 'JAN 10'

AND 'MARCH 10' GROUP BY MONTH

========================================================================

**how to find %sale. compare this year to previous year**

Create Table City(

Cno INT,

Cname Varchar(50),

Pop INT,

Years INT

)

Insert Into City Values(1,'Hyd',1000,2023),(2,'Warangal',500,2023),(3,'Vizag',600,2023),(4,'Hyd',2000,2024),

(5,'Warangal',2000,2024),(6,'Vizag',2400,2024)

Select \* from City

WITH CTE AS (

SELECT \*, LEAD(Pop, 1, 0) OVER (PARTITION BY Cname ORDER BY Pop) AS P

FROM City

)

SELECT \*, ((P - Pop) \* 100) / Pop AS D

FROM CTE

WHERE Years = 2023;

========================================================================

**stored. procedure:**

-

create database eclasss

use eclasss

--creation of store procedure

--mean it:Stored Procedure:it is a precompiled statements with structured query which is called as store procedure.

create Table emp(

eno INT,

Enmae Varchar(10),

Esal Money

);

Insert into emp Values (1,'Venu',100),

(2,'Venky',200)

Select \* from emp

Create Proc Proc\_emp

As

Begin

Select \* from emp

end

exec Proc\_emp

---------------------------------------

Create Proc View\_Proc(@eno INT)

As

Begin

Select \* from emp where @eno=eno

end

View\_Proc 1

**-------------------------------INSERT—-------------------------------------------------**

Create Proc Insert\_Proc(@eno INT,

@Enmae Varchar(10),

@Esal Money)

As

Begin

Insert into emp values(@eno,@Enmae,@Esal)

end

Insert\_Proc 3,'mahesh',500

Select \* From emp

-------------------------------**UPDATE**------------------------------------

Create Proc Update\_Proc(@eno INT,@sal Money)

As

Begin

Update emp set Esal=@sal where eno=@eno

end

Update\_Proc 3,555

Drop Proc Update\_Proc

select \* from emp

-------------------------------------**DELETE**----------------------------

Create Proc Delete\_proc(@eno INT)

As

Begin

Delete from emp where eno=@eno

end

Delete\_proc 3

select \* from emp

-------------------------------**DATA FACTORY SYSTEM VARIABLE**-------------------------

create table pipeline\_log\_detailes

(datafactoryname varchar(500),

pipelinename varchar(500),

pipelinegroupid varchar(500),

pipelinerunid varchar(500),

pipelinetriggername varchar(500),

pipelinetriggerid varchar(500),

pipelinetriggertime varchar(500),

pipelinetriggertype varchar(500))

select \* from pipeline\_log\_detailes

create proc pipeline\_log(

@datafactoryname varchar(500),

@pipelinename varchar(500),

@pipelinegroupid varchar(500),

@pipelinerunid varchar(500),

@pipelinetriggername varchar(500),

@pipelinetriggerid varchar(500),

@pipelinetriggertime varchar(500),

@pipelinetriggertype varchar(500))

as begin

insert into pipeline\_log\_detailes values(@datafactoryname,

@pipelinename,

@pipelinegroupid,

@pipelinerunid,

@pipelinetriggername,

@pipelinetriggerid,

@pipelinetriggertime,

@pipelinetriggertype

)

end

================= **how to split the fullname.** ==========================

Declare @FullName Varchar(50)='Sikilambatla Pranay Kumar';

Declare @Firstname Varchar(50);

Declare @MiddleName Varchar(50);

Declare @LastName Varchar(50);

Set @Firstname=PARSENAME(Replace(@FullName,' ','.'),3)

Set @MiddleName=PARSENAME(Replace(@FullName,' ','.'),2)

Set @LastName=PARSENAME(Replace(@FullName,' ','.'),1)

Select @Firstname as F,@MiddleName as M,@LastName as L

========**ROLLING. AVG**==================================

CREATE TABLE posts (

user\_id INTEGER,

date DATE,

post\_count INTEGER

);

INSERT INTO posts (user\_id, date, post\_count)

VALUES

(1, '2023-07-01', 5),

(1, '2023-07-02', 7),

(1, '2023-07-03', 3),

(1, '2023-07-04', 10),

(1, '2023-07-05', 6),

(2, '2023-07-01', 8),

(2, '2023-07-02', 4),

(2, '2023-07-03', 2),

(2, '2023-07-04', 9),

(2, '2023-07-05', 5);

SELECT

user\_id,

date,

post\_count,

ROUND(AVG(post\_count) OVER (PARTITION BY user\_id ORDER BY date ), 2) AS rolling\_average

FROM

posts;

OUTPUT:

user\_id | date | post\_count | rolling\_average

--------+------------+------------+-----------------

1 | 2023-07-01 | 5 | 5.00

1 | 2023-07-02 | 7 | 6.00

1 | 2023-07-03 | 3 | 5.00

1 | 2023-07-04 | 10 | 6.25

1 | 2023-07-05 | 6 | 6.20

2 | 2023-07-01 | 8 | 8.00

2 | 2023-07-02 | 4 | 6.00

2 | 2023-07-03 | 2 | 4.67

2 | 2023-07-04 | 9 | 6.00

2 | 2023-07-05 | 5 | 5.60

**===========MEAN. ,MODE. MEDIAN. QUERY**===================

-- Step 1: Create the table

CREATE TABLE employee (

Emp\_id INT PRIMARY KEY,

salary DECIMAL(10, 2)

);

-- Step 2: Insert up to 7 records

INSERT INTO employee (Emp\_id, salary) VALUES

(1, 45000),

(2, 60000),

(3, 55000),

(4, 70000),

(5, 50000),

(6, 60000),

(7, 55000);

QUERY:

SELECT

AVG(salary) AS mean\_earning,

PERCENTILE\_CONT(0.5) WITHIN GROUP (ORDER BY salary) AS median\_earning,

MODE() WITHIN GROUP (ORDER BY salary) AS mode\_earning

FROM

employee;

**ADF INTERVIEW QUESTIONS**

**What are your role. and responsibilities in your project?**

We are taking user stories and performing analysis,Designing and implementing data integration pipelines using Azure Data Factory to extract, transform, and load data from various sources into the analytics platform. This involves ensuring data quality, data consistency, and efficient data movement. development after completing development we perform unit testing. We are Attending every Day status or scrum calls to discuss progress of work or any blockers to task.

**if I get any chance I will become a functional Architect:**

So I worked as a Date Engineer in coginzent AS a team of 8 members, we used to work as a team whenever we got any errors we used sort it as a team.Mainly I used to take user stories and performing analysis, development after completing development we perform unit testing we are Attending every Day status or scrum calls to discuss progress of work or any blockers to task.

if I get any chance I will become a functional Architect

**What is the methodology used to develop the project?**

Agile methodology

**What are tools used to track the project?**

Jira tool or Azure devops

**What is the use of Azure Devops**.

for code commit and deployment

**What is your team. size**

**8 members**

**What is the backlog.**

Task which is not completed in previous sprint, then we have to move to next sprint

The sprint duration is 2 weeks.

**What are meetings conducted in agile methodology.?**

Sprint planning

Sprint review

Sprint Retrospect

Daily standup call or Scrum call

**How many Activities can be used in a single pipeline?**

40 maximum

**How many pipelines are created in your project?**

maximum 35

**What are Dimension table. and Fact table. created in in your project**

**Dimension Table:**Dim customer , Dim product, Dim time, Dim location ,Dim date, Dim promotion

**Fact Table:**Sales fact table, Marketing Fact,Production Fact ,Inventory Fact

**What is the KPI. and insight in your project?**

**Total sales:**The KPI shows the total sales for the selected period

**Sales Region:**  The KPI shows the total sales breakdown by region sales india,america

**Growth:**This KPI shows the sales growth, compared to the previous period.

Sales breakdown by product category such as beverages,cool drinks,kitkat,maggie etc..

**What Is a cloud.**

cloud is a rental based system

accessing services remotely through the internet by using a web browser from the data center.

**What Is a Resource group.**

Whatever Resource Is created inside Azure needs to go to groups as a single entity that is called resource group manager.

advantages:

->there are multiple resources based on requests we can create.

->we can delete the resource group and all the resources will be deleted.

->when we want to deploy a resource like an ADFpipeline from one environment to another environment it will move easily.

**Subscription**:?

Any services inside azure portal is known as subscription

**What are the different deployment mode. of the project?**

->we have 3 types of deployment modes,

1)Public:Azure,AWS,GCP

2)private:TCS

3)Hybrid:Public +Private

**What are different categories of service. available in the cloud?**

->3 types of services available in cloud they are,

1) IAAS. : Infrastructure as a services(eg: CPU,RAM,MEMORY)

2) PAAS. : Platform as a services(eg:SQL SERVER,Operating system,VM(virtual machines))

3) SAAS. : Software as a services(eg: office 365)

**What are redundancy. available in azure or ADF**

LRS.

GRS

ZRS

GZRS

**What Is a storage account or BLOB.**

Storage account is one of the service available in azure to store any kind of data like backup

storage ,it will store data in flatten format and structured, unstructured, semi structured

eg:structured->csv,columns and rows

unstructured->audio,video,image

Semi structured ->xml,json

**What are storage mode. in a storage account?**

there are 4 types of modes in storage account they are,

1)container. or BLOB(Binary Large Object)

2)Table

3)File share

4)Queue

**1)Container or BLOB(Binary Large Object):**whatever the file inside the memory is known as container or blob

**What blob type. do we have ?**

3 types of blob storage:

1) Block. Blob.->it is default storage, it stores data like audio,video,image

2) Page Blob->VM(virtual machines) related will be store

3)Append Blob->storing of log data , it will do append operations

**What are different kinds of access. tires?**

3 types of access tires,

1)Hot tire.->Regularly or it accessing data frequently,the cost of hot tier is expensive,high performance

2)cool->Rarely or it accessing data unfrequently,the cost of cool tier is not expensive,low performance

3)Archive->very rarely

**What are security. mechanisms?**

3 types of security mechanism

1)Access key.

2)Connection string.

3)shared Access Signatures(sas.)

**Access keys:**

**Disadvantages**:Access keys and Connection strings

1)By default we will get access to all storage mode in account

2)set of expiration and activation not possible

3)cannot restrict permission like read, write,, view, delete

4) cannot restrict IP address

**Shared access signature**

**Advantages**:

1)it will get restrict the access to all storage mode in account

2)set of expiration and activation possible

3)it restrict permission like read, write,, view, delete

4)it restrict the IP address

**how to delete files automatically in 30 days?**

-> By using Life cycle management system

**how to give control access to another person?**

->By using RBAC.(Role Based Access Control) which access like to Reader,Contributor,and   
 Admin.

->IAM-Access Control

**what is the degree. of parallelism in copy data activity**

“Degree of Parallelism” in the copy data activity refers to the number of parallel connections used to copy data from a source to a destination.

maximum number of parallelism is 32

we can in between 10 to 20

-=-====================================================================== **ADLS2.**

**What is data lake GEN2. or ADSL2.**

It is a big data storage system and to store any kind of data like structured, semistructured, unstructured format

-> it is designed top of the hadoop framework

->data lake has more compatibility with HDFS

->data will be store in data lake in hierarchical name space formate

**What is hierarchical. namespace use?**

hierarchical namespace refers to a feature that provides a more efficient and structured way to organize and access data within the storage account.

**What is big data.?**

it deals with a dataset that are too large data or complex data

volume->large amount of data (59zb now)(175zb predict in 2025)

variety->structured, semistructured, unstructured

velocity->performance

**What is Hadoop.?**

Hadoop is data frame it stores the data and it process the data, for storing purpose we are using HDFS(Haddop file system) for processing map reduce

**What is mapreduce.?**

it will do parallel processing

it split the task into multiple tasks

it contains hard disk memory

the read and write operation compare to spark is slow

**What is spark.?**

it will do parallel processing

it split the task into multiple tasks

it contains in memory RAM space

the read and write operation compare to Map reduce is more

**Difference between GEN1. and GEN2?**

**GEN1:**

->Inside GEN1 does not have RBAC and IAM

->Inside GEN1 does not have Replication

**GEN2:**

->Inside GEN2 does have RBAC and IAM

->Inside GEN2 we have Replication

**GEN2=GEN1+BLOB**

**Difference between blob and Data Lake GEN2.?**

**Blob**:Storage account is one of the service available in azure to store any kind of data like backup storage ,it will store data in flatten format and structured, unstructured, semi structured

eg:structured->csv,columns and rows

unstructured->audio,video,image

Semi structured ->xml,json

->Blob is object based model with container

->Not compatible with Hadoop

->the data access can be done using access keys and shared access signature

->less cost storage

**GEN2**:It is a big data storage system and to store any kind of data like structured, semistructured, unstructured formate

-> it is designed top of the hadoop frame work

->data lake has more compatibility with HDFS

->data will be store in data lake in hierarchical name space formate

->data access can done using active directory authentication

->cost is expensive

========================================================================

**what is the architecture of a synapse data warehouse?**

**What are the terminologies in synapse DWH. architecture?**

**->1)control node**

**2)MPP. Engine(massive parallel processing)**

**3)Compute Nodes**

**4)DMS.(Data Movement Service)**

**5)Distribution Methods**

Dedicated SQL Pool or(SQL synapse dw) is uses a " Node base architecture"

**Control Node: it** looks like a human brain **.** It is taking request from application and host MPP engine

to distribute and optimize queries for parallel processing

**MPP Engine:** Massive parallel processing of SQL queries by splitting into multiple

smaller tasks

**Compute Nodes:** Providing infrastructure to execute each task on each compute node(0-60)

**DMS.:**Data Movement Service to move the data across nodes

**DISTRIBUTION. METHODS:**

Based on distribution method created on tables, data will be distributed across nodes

1)Round Robin.

2)Hash.

3)Replication

**Round Robin.:**Data will be distributed equally (or) randomly across nodes recommended for less volume of data tables like "dimension tables"

Data Distribution below 2 gb

**Disadvantages:** data need to be sorted at time of combining data from all nodes

**Hash.:**

->Data will be distributed based on columns specified in hash

->some kind of product ids will be shared to one compute node

Data Distribution above 2 gb

Advantages:

-> No need sorting while combining data from all node

->Mostly recommended for large tables like fact tables

**Replication:**

->All the data will be shared among all nodes

->Recommended for small lookup tables

Data Distribution method less than 500 rows or 10 mb

**difference between Azure sql database and Azure synapse dwh.?**

**Azure sql db.:**

->insert, update, delete, mostly used this option

->it is a OLTP(Online Transaction Processing) database

->Dynamic mask possible

->Polybase will not support to BLOB or Data Lake Gen2

->its contain 4TB data

->its have 30000 concurrent connection

->its have 6400 concurrent queries

->Not supporting MPP

**Synapse. DWH:**

>it will stored structure data

->Mostly select command is used dml

-> it is OLAP(Online Analytical Processing)

->Dynamic mask not possible

->Polybase will support to BLOB or Data Lake Gen2

->its contain 1PB data

->its have 1024 concurrent connection

->its have 32 concurrent queries

->supporting MPP

**POLYBASE.:**

querying the data from external data source like BLOB or Data lake GEN2 into sql synapse DWH by creating external data source and external file format and external tables is called polybase

There is a 5 steps to create polybase

1) create master key

2)create database scoped credential

3)create external data source

4)create external file format

5)create external tables

**Advantage:**

->Azure BLOB storage is convenient place to store the data by using azure services and polybase and it is easy to access data from data lake GEN2

**What is an elastic. pool? What are different compute tires?**

->Group of azure Sql databases to shares resources among to all this is known as elastic pool

2 types of elastic pool:

1)DTU(Data Transaction Unit)

2)vcore processors

2 types of vcore processors

->provisioned

->serverless

**What is data mask.?**

data masking in securing the sensitive data or hiding the sensitive data

**difference between Synapse dedicated sql pool formerly known as DWH and Serverless pool?**

**Synapse dedicated sql pool formerly known as DWH.:**

->Dedicated pool of providing of 'DWH' at the time of creating

->it is used storage of synapse sql pool

->Application charges hourly/monthly

**Serverless. pool**

->serverless pool no need to provide DWH

->it is auto scale up and down

->it is used to store data lake

->charges only for execution queries

-**>how many compute nodes can be created------>0-60 max**

**->several distributions available?**

1)Round Robin

2)hash

3)Replication

->Default distribution available->Round Robin

========================================================================

**Synapse analytical. Service**

->It is one type of storage

->it have more analytical power

->it used in memory RAM space formate

-> it compress the data

—---------------------------------------------------------------------------------------------------

**Authentication. type in adf at linked services level ?**

**\* Account key**

**\*sas url**

**\*service principle**

**\*system assigned managed identity**

**\*user managed identity**

------------------------------------------------------------------------------------------------------

**How to change the data types in SQL,ADF,data flows and Data Bricks ?**

SQL:we are using ‘CAST.’ function

ADF:at pipeline level we have import schema option or right side top we have json{ } file we can modify the data types in pipeline level

Data flows: we have project tab

Data Bricks: with the help of inferschema=’True’ and struct type and struct field

**Azure Data Factory**

**How much GB. of data had Handled in your project and daily incremental of data**

60 GB of data and we used to get 3k of tables daily where we used to maintain two tables for incremental data.and we generated up to 160 reports.

**What is ADF.?**

->it is cloud based ETL tool and zero coding inside the cloud ETL

->it have some predefined activities

->ETL(Extraction Transformation loading of data)

**How many version types in ADF?**

two types of versions they are,

Azure data factory v1

Azure data factory v2.

**Azure data factory v1:** Github integration not available

trigger is not available for scheduling

some activities and transformation are missing

**Azure data factory v2:** Github integration available

trigger are available for scheduling

more activities and transformation

**What are the major component. used in adf?**

1)integration runtime

2)linked service.

3)Data set.

4)activities

5)pipeline.

6)Trigger

**What is integration. runtime and its types?**

this providing compute infrastructure to run pipeline and some activities

**3 types :**

1)Auto resolve IR or Azure IR

2)Self Hosted IR

3)SSIS. IR

**1)Auto. resolve IR or Azure IR**:to connect cloud system to cloud

**2)Self Hosted IR:to** connect on-premises network or (private network) to cloud network

3)**SSIS IR**:shifting lifting of SSIS packages to on-premises to cloud

**What are linked service. ?**

it is a connection string between any data sources like source and destination

**What Is the data set.?**

preparation of data to move and transform and load the data

**What Is a pipeline?**

it is ETL process

**What are the optimization in ADF or challenging in adf ?**

complex data transformations

Data quality

Performance Optimization

Error handling and Loggings

Security and complianc**e**

check appropriate data types

check partition

increasing cluster nodes

If our pipeline if failed due to network issue after executing some activities in pipeline,we can rerun the pipeline from the failed activities.

**What Is the trigger.?**

Schedule of pipeline on a specific time automatically we need trigger

3 types of trigger

1)Normal schedule trigger

2)Tumbling trigger

3)Event based trigger

**1)Normal schedule trigger:**As per schedule it uses daily bases,present or future.

->it will execute pipeline recursively

**2)Tumbling trigger:** it used in historical manner

->it will execute a pipeline periodic time interval from specific time

**3)Event based trigger**:when the event occurred the blob create or delete

->it will execute pipeline

**Which triggers are used to run the pipeline on every 2nd and 4th saturday?**

Scheduling and tumbling window triggers.

**How to send EMail. Notification**

* we are using a web activity and logic app, there is a http request using get and post method.

**What is dependent on adf activity?**

->success

->failure

->skip

->completion

**What are the activity. used in a project?**

->Getmeta data

->Lookup.

->Stored procedure

->for each

->if condition

->web activity

->Dataflows

->execute pipeline

->Delete activity

**What are the transformation. used in data flow.?**

->look up

->derived column

->join

->aggregate

->windows

->alter row

->sort

->unions

->select

->condition split

->new branch

**difference between Getmeta. data activity and lookup. activity?**

Getmeta data activity:it is used to getting the information about the files

lookup activity:it is used to getting the data in table from a database

**difference between lookup activity and stored procedure activity?**

**lookup. activity**:it will give some output

->somewhere we have data in the database table that data extracted by lookup activity that gives output for other activity

**stored procedure activity:it** will not written any output

it will execute sql statement

**What is dataflow.?**

it is visual designed data transformation in data factory

to develop the transformation without coding

->look up

->derived column

->join

->aggregate

->windows

->alter row

->sort

->unions

->select

->condition split

->new branch

->sink

->pivot

**Can we connect dataflow to on premises?**

**Ans:**No,Because data flow will not support the self hosted Integration runtime

**What is the parameter.?**

at running of the pipeline we pass a value from outside is known as parameter

**What is variable.?**

at running of the pipeline we pass a value from inside is known as variable

**What is the set variable.?**

assigning the value for variable of data type as string or boolean

**What is the append variable.?**

assigning the value for variable of data type as array

**What is the global variable.?**

it is predefined parameter that can be used across the multiple components activity with in adf pipeline

**What is the drift schema.?**

our source often change the meta data.fields, columns,data types can be added or removed without handling drift schema our data flow becomes vulnerable

**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

**.What are Activity. 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

**What are Transformation. 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---based on condition data will be splitted

Ex: Male and female two outputs

Filter: filtering of data by putting filter condition

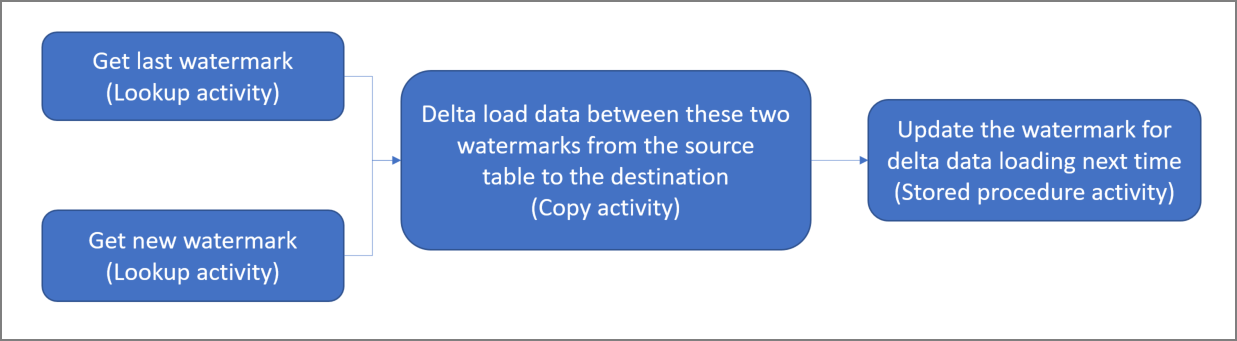
Assertions--Performing some data validations by specifying data rules

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

Aggregate: performing aggregate and group by

Union: appending of data into one output

**10.How to implement Incremental. Loading.**

**A.** 

* **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.

**What is incremental. loading?**

When new and modified records and updated records coming into our database then we create the watermark value and control table and we filter the record based on the last modified datetime by using some activities overthere like lookup,copy data activity and stored procedure.

**select \* from data\_source\_table where LastModifytime > '@{activity('LookupOldWaterMarkActivity').output.firstRow.WatermarkValue}'**

**—------------------------------------------------------------------------------------------------------**

**What is CDC in Incremental loading?**

CDC stands for change data capture in incremental loading.

Change Data Capture is a technique used to identify and capture changes made to the data in the data warehousing and database.

finally changed data will be transferred instead of transferring entire data to the sink/destination.

**INCREMENTAL LOADING:**

First Create a One Azure Sql Database and create the tables are Tables, Watermark Table, 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.
* **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.

**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.:**

* GITHUB 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

**22. What is a Data Flow.**

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 row\_number 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 ascending 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---dimcustomer backup

customer---dimcustomer copy

**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 Data Flows and Wrangling Data Flows**

## 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.

**Wrangling. dataflow:**

Data wrangling in the context of Azure Data Factory (ADF) refers to the process of transforming and cleaning raw data using visual design tools within ADF's Data Flow feature. ADF Data Flows provide a graphical interface for building data transformation logic without the need for extensive coding.

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

**UPLOAD A MULTIFILES:**

* create a getmetadata activities
* one dataset in getmetadata

FOREACH ACTIVITY:

* **@activity('Get Metadata1').output.child items--** 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

**MULTIPLE TABLES 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 parameter in sink
* sink value is **@concat(item().mytable)**

**KEYVAULT.:**

* We use key vault to keep the secrets of keys, certificates, server name etc
* 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

**SEND THE MAIL:**

* Create a New Getmetadata Activity that Add 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 to 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 you 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":"pipeline 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 **row\_number()**
* 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 updates of the existing file but not Historical values.

* create a two database one is source and other is target
* in source create table with table name and insert values
* in target create dim table
* 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 METHODOLOGY**

* 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 AIRTICRAFT 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.

DATA WAREHOUSING INTERVIEW QUESTIONS

- **what is DWH. ?**

Data warehouse it is used for ETL because of analytics purpose

DW we have a lot of historical information

combine all data marts is called data warehousing

**- What is data mart. ?**

Data Mart: A subset of a data warehouse that is focused on a specific business function or

department,making it easier to manage and analyze data for that area.

(D.M is the one Business area or one of work in the entire area or Organization.)

...................................................................................................................

**- What is a dimensional. table. ?**

D.T having descriptive textual information is called D.T.

**- What is a fact. table. ?**

F.T having Key values and Measure Values is called F.T

...............................................................................................................

**- What are the diff kinds of dimension. ?**

role playing. :a Fact table having more than one time relationship with single dimension table is called roleplaying

conformed :one dimension table having relationship with different fact table is known as confirmed

junk. :Repeating unnecessary information in table is called junk dimension

degenerated:it doesn't have any source data.fact table contain dimension table information is known as degenerate

slowly changing dimension:

.................................................................................................................

- **what are diff fact Measure.**

additive (use quantity values with any kind of dimension this kind of measure is called Add.M)

semi additive (Sum of the measure values are calculating with corresponding with the sum

of the dimension table)

non additive (Sum of the measure values are NOT calculating with corresponding with the

sum of the dimension table)

**- what is star. schema.**

All dimensional tables should have a direct relationship with the fact table with some kind of surrogate key and Foreign Key.

Star Schema top to down model

Inside the star schema normalization not used

This design is very simple and it have high data redundancy  
It will have less number of foreign keys

**- what is a snowflake. schema.**

sum of the dimensional tables should not have direct relationship with the fact table.

-Its is down to top model

-Inside the Snowflake schema normalization will used

-This design is very complex

-Low data redundancies

-High number of foreign Keys

**- what is surrogate. key.**

It generates the sequence number when ever you insert new row it will give new number,

This will identify each row uniquely.

**- what is oltp. and olap.**

**OLTP:** (Online transaction processing)

in OLTP we have a master table and transaction tables.

oltp : capturing of the day to day live transactions data from application into database.

(Insert , update , delete)

eg:online banking,shopping

OLAP: (Online Analytical Processing)

in OLAP we have a Dimensional table and fact tables.

olap:online analytical processing,that organizes large business database and supports complex analysis

..............................................................................................................

**What is the slowly changing machsim ?**

SCD0- inserting the new record which ever not available in the target

SCD1- inserting the new record, it updates the records(particular value in record) (overwrite)

SCD2- inserting the new record, maintaining the historical information by keeping a fLAG column

or start time or end time

......................................................................................................................................

**What is ER. modeling. ?**

ER-modelling:- Entity Relational modeling

Entity Relational modeling PK and FK relationship

To describe the relationship between different entity in system

....................................................................................................................

**dimension. -modelling**:it is a one of data modelling techniques used in data warehouse design .to improve the data retrieval

To maintain historical data in data warehouses

—----------------------------------------------------------------------------------------------------------------------------**-What is Medalioin. Architecture ?**

A medallion architecture is a data design pattern used to logically organize data in a [lakehouse](https://www.databricks.com/glossary/data-lakehouse), with the goal of incrementally and progressively improving the structure and quality of data as it flows through each layer of the architecture (from Bronze ⇒ Silver ⇒ Gold layer tables). Medallion architectures are sometimes also referred to as "multi-hop" architectures.

**Bronze layer (raw data):** The Bronze layer is where we land all the data from external source systems.The focus in this layer is quick Change Data Capture and raw data load in in this layer

Silver layer (cleansed and conformed data):In the Silver layer of the lakehouse, the data from the Bronze layer is matched, merged, conformed and cleansed.and we are doing some transformation

**Gold layer (curated business-level tables**):Data in the Gold layer of the lakehouse is typically organized in consumption-ready "project-specific" databases. The Gold layer is for reporting and uses more de-normalized and read-optimized data models with fewer joins. The final layer of data transformations and data quality rules are applied here we are doing aggregation like

max,min,avg,sum

-------------------------------------------------------------------------------------------

Master tables----customer,product

transaction tables----sales

ETL

----------------------dwh

Dimensional Modeling

-----

Dimension table-- dimcust,dimprodt

Fact --- fact table

---------------------------------------------

-------------------------------------------------

--------------------------------------------------------------

oltp - capturing of the day to day live transactions data from application into database.

(Insert , update , delete)

will do ETL to dwh

ER- Entity Relational modeling

Entity Relational modeling PK and FK relationship

------------------------------------------------

Master tables----customer,product

transaction tables----sales

ETL

----------------------dwh

Dimesional Modeling

-----

Dimesion table-- dimcust,dimprodt

Fact --- fact table

--------------------------------------------

-------------------------------------------------------------

-country

cid cname

1 india

2 pakistan

-state

sid sname cid

1 telangana 1

2 karachi 2

-city

ctid cname sid

1 Hyderabad 1

2 Sindh province 2

-customer

custid custname

-sales

custid amount quanty

-sales

custid cid sid ctid amount quanty

**Scenario based question**

**1.How to load the data(file) into on-premises file system into Data Lake gen2**

- Create a self-hosted integration runtime on a machine in your on-premises environment.

- Configure the linked service for the on-premises file system and Azure Blob Storage in Azure Data Factory and then create dataset for source and sink

- Create a pipeline with copy data activity that uses the on-premises file system as the source and Azure Blob Storage as the destination.

-Choose getmeta activity,inside the getmeta data activity we have to choose the field list as child items and take foreach loop activity,inside foreach activity take the copy data activity

---------------------------------------------------------------------------------------------------------

**2.How to load the data(table) into on-premises sql server into Data Lake gen2**

- Create a self-hosted integration runtime on a machine in our on-premises environment.

- Configure the linked service for the on-premises SQL Server database and Azure Blob Storage in Azure Data Factory.

- Create a pipeline with a copy activity that uses the on-premises SQL Server database as the source and Azure Blob Storage as the destination.

-Choose lookup activity and take foreach loop activity and pass the information of lookup activity to foreach activity inside foreach activity there is copy data activity

—-------------------------------------------------------------------------------------------------------------------------

**How to create/setup a pipeline.?**

create datafactory

inside datafactory we have the four major options like author,manage,monitor and home.

in the manage create integration runtime like auto-resolve(or)self hosted and create linked service for source and destination.

in the author option create the dataset for the data which you want to extract, finally setup the pipeline by using linked services,datasets.

—-----------------------------------------------------------------------------------------------------------------

**how to load multiple excel sheets?**

create azure datafactory ETL pipeline---create linked services,datasets for the pipeline

source is blob/data lake gen2 and sink is azure sql.create a table in azure sql, insert the excel sheet names as values in the table.

Then take the lookup activity and give the created table to lookup,then take the foreach activity,for foreach activity the information will get from the lookup activity.

Inside foreach take the copy data activity add the source and sink datasets to the copy activity make sure of enabling the sheet index option and use the dataset parameter for sheet index in copy data source dataset.

**how to separate the different files like csv,txt and load them into different sinks separately?**

create azure datafactory ETL pipeline---create linked services,datasets for the pipeline

source is blob/data lake gen2 and sink is different azure sql databases(one database is for txt and another database is for csv files).

take the getmeta activity to read all the file information in blob/datalakegen2,take two filter activities for csv and txt files, give connection with getmeta activity to both filter activities.

for separating files write the condition in both filter activities "@endswith(item().name,'.csv','')".

take two foreach activities and connect with filter both activities,inside the foreach activities take the copy data activities and give the source and sink datasets for both copy data activities.

**3.How do you connect two Data Factories?**

**-** By using web activity

**4.How to connect any server to data lake gen2?**

* we are using a web activity and logic app, there is a http request using get and post method.

**Project workflow environment**

we have 2 types of methodology

1. waterfall model

2.Agile methodology

Waterfall Model-

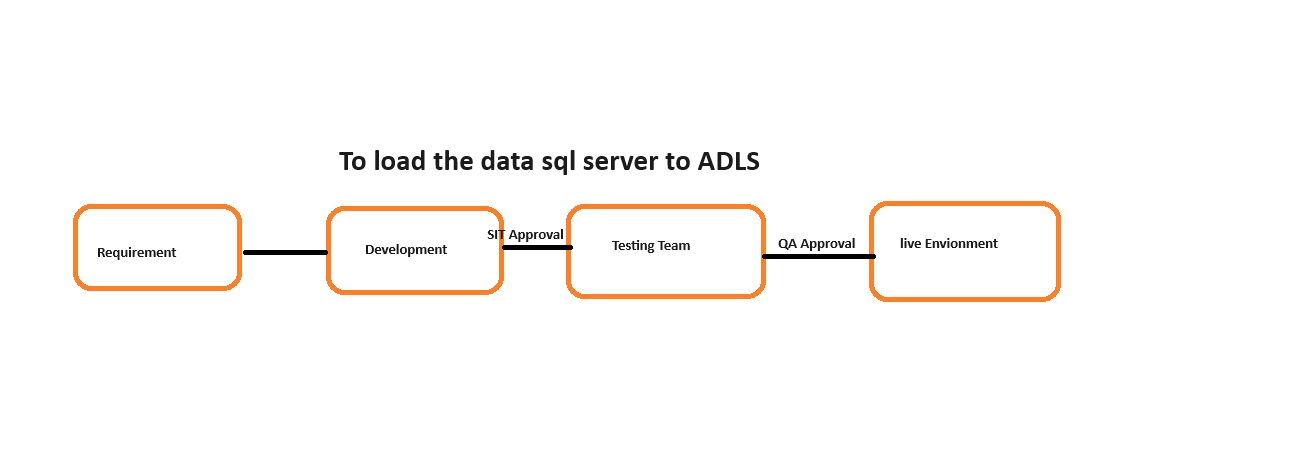
1.Waste of time

2.cost also high

Agile Methodology:

1.it handle small tasks

2.low cost



**Dev Environment:**

it will handled by developers

they can change as per business requirement

**SIT Environment:**

it will handle by client

if any changes occur , he can make changes

in this environment developer can only view the data

**QA or UAT OR pre prod:**

QA(Quality assurance)

UAT(User Acceptance testing)

First we are doing unit testing, then find any blockers or bottlenecks or impediments there, if we get any blockers again we will send them to developer.

AZURE DATA BRICKS INTERVIEW QUESTIONS

=====================================ADB================================

**Difference between list. ,tuple.,set,dictionary?**

**LIST:::**it store element in single row and multiple column

It represented like– [ ] or square bracket

It allow duplicate element

It is a mutable—( changeable and updating)

It is a order

**TUPLE.:**:it store element in single row and multiple column

It represented like— ( ) or parenthesis bracket

It does allow duplicate element

It is a immutable—(not changeable and we can’t update)

It is a order

**SET::**it store element in single row

It represented like– { } or curly bracket

It does n’t allow duplicate element

It is a mutable

It is a unorder

**DICTIONARY::**it store key values pairs

It represented like– { } or curly bracket

It does n’t allow duplicate element

It is a mutable

It is a order

**1. What is Spark and MapReduc.?**

**Spark.** :it is doing parallel processing and splitting the task into multiple tasks. and it will use ram in memory for read & write operations

**Map reducing:**it is doing parallel processing and splitting the task into multiple tasks. and it will use harddisk in memory for read & write operations,comparatively spark its low performance

**2. What is Azure Data Bricks adb.?**

Azure Data bricks is a data analytics platform optimized for the Microsoft Azure cloud services platform.and it is a auto scaling up and down

it is integrated with other azure services like ADF, power BI ,ADLs gen2 in an interactive workspace and reduce the cost and complexity with a managed platform

The data bricks has multiple language like

python

scala.

sql

R.

**3. What are components of Azure data bricks or What is the control panel. of 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

**What is the Default number of Partition. in Spark**

Partition is nothing but can be performed on Large Data sets which can be break down into small chunks to Process the data faster.

**Default number of Partition:**

Max No of Partitions = 200.

Min No of Partitions = 2

**How do we connect Pyspark with Databricks**

Internally Pyspark will connect with Scala by using PY4J Connector

**What is PY4J. Error**

It is a connection Error.

**What is RDD., Data frame, Data set.**

**RDD:** A Resilient Distributed Dataset (RDD), it is a distributed collection of data across the multiple cluster node

RDD is a immutable and it is fault tolerant i.e it means they can recover from node failure without any data loss

do not have fixed schema and does not have optimization technique

performance is slower than dataframe

**DataFrame.:** A Data Frame is a table structure and it contains named columns

It contains inbuilt optimizing techniques.

it is a distributed collection of data across the multiple cluster node

it have fixed schema

It is similar to RDBMs

**Dataset.:** A Dataset it is a distributed collection of data across the multiple cluster node

dataset have well defined schema

datatype comparing to dataframes

**Resilient Distributed Dataset (RDD)**

Spark 1.0, 2011

No Schema

Java, Scala, Python, R

Perform slow in aggregation

Low level API

**Data Frame**

Spark 1.3, 2013

Schema Structured

Perform fast in aggregation

Java, Scala, Python, R

High level API

**Data Set**

Spark 1.6, 2015

Schema Structured

Java and Scala

Perform fast in aggregation

High level API

**What is Transformation and Action or what are operations performing on rdd and data frame and data set?**

**Transformation.:** Transformation refers to the operation applied on a RDD to create a new RDD Filter, group By and map are the examples of transformations. It is producing new data frame

**Action.:** It is not producing data frame, those returning result to driver node or writing data into external storage

**What is diff between Map reduce and Spark**

**1.MapReduce.**

it is doing parallel processing and splitting the task into multiple tasks

It is having a very slow speed as compared to Apache Spark.

It is unable to handle real-time processing.

it contain harddisk memory

**2.Spark.**

it is doing parallel processing and splitting the task into multiple tasks

It is much faster than Map Reduce.

It can deal with real-time processing.

It is difficult to program as you require code for every process.

It is easy to program.

it contain in memory RAM space

**What are different kinds of cluster node. in Azure Data Bricks?**

**Single Node:** Doesn’t have any worker nodes and recommended for the single server

**Standard.:** Recommended for a Single user cluster and can run SQL, Python, R, SCALA workload. Execute JOB in sequential

**High Concurrency:** Optimized to run concurrency SQL, Python, R workloads runs and scala is not supported. Execute JOB in parallel

**What is DAG?**

A. A directed acyclic graph (DAG) it will save all the information of parent and child RDD inheritance

When we call an Action on Spark RDD at a high level, Spark submits the operator graph to the DAG Scheduler.

Dag Divide the operators into stages of the task in the DAG Scheduler. A stage contains task based on the partition of the input data. The DAG scheduler pipelines operators together.

The stages pass on to the Task Scheduler. It launches task through [cluster manager](https://data-flair.training/blogs/apache-spark-cluster-managers-tutorial/). The dependencies of stages are unknown to the task scheduler.

**1..Difference between instance. and cluster**

An instance, also known as a "worker node," represents a single virtual machine in the Data bricks cluster. Instances are the individual compute nodes that perform the actual data processing and computation.

**cluster:**

A cluster in Azure Data bricks is a collection of instances (driver and worker nodes) that are used to process and analyze data. When you start a Data bricks cluster, you define the number and type of instances to use, the cluster configuration, and other settings.

**Different Type of Clusters**

**All-Purpose interactive Clusters:**

All purpose interactive clusters can be restart able

After job is executed then back end cluster running so cost issue will occurs

It is using in development

**Job. Clusters:**

Run job cluster your job is executed it is terminated automatically after runs

You can’t restart this cluster

It is using in production

**What is the spark session and spark context?**

**SPARK SESSION.**:Spark Session is the entry point to the spark application and provides a way to interact with spark clusters. It is responsible for creating and managing the spark context.

**Spark context.:it** is a low level interface that represents the connection to a spark cluster and can be used to create RDD and perform actions and transforms.

**What is Lazy Evolution?**

the transformation are not evaluating action is called Lazy point,

Advantages: Reduce the amount of memory usage

**What is auto scaling.?**

Auto Scaling is Nothing But Based on The workLoad it will increase and decrease the cluster nodes whichever direction you need.

**What is dbfs.**

It is data bricks file system

It have high performance tiers

Replication available

Massive data we storage

—----------------------------------------------------------------------------------------------------------------------------

**What is the difference Between Caching and Persistence or Optimization in adb ?**

**Broadcast Join**

**Broadcast Variable**

**Cache & Persist**

**Coalsec & Repartition**

1)caching and persistence both are used to save the data in memory for faster access

2) in addition to persistence will support the specify external storage to cache the data

Caching has default memory persistence is customized memory.

—------------------------------------------------------------------------------------------------------------------------

**How can I handle bad data. in adb?**

"permissive,"

"dropMalformed,"

"failFast"

**What is Delta Lake.?**

1)Delta lake is an open-source data format that provides ACID transactions, data reliability,

2) It is another storage level top of the data lake only

Advantages:

its treat as a OLTP database

time traversing

version controlling

schema and enforcement

Vacuum: It will erase the data with in 14 days

**We create new dataframe df**

**eno ename esal**

**1 venu 100**

**2 anu 200**

**df.write.format(‘parquet’).save(“filestore/tables/geetha/vsit31”)**

then we will create the new dataframe df1

**eno ename esal cont no**

**1 venu 100 23454**

**2 anu 200 34562**

**df1.write.mode(‘overwrite’).format(‘parquet’).save(“filestore/tables/geetha/vsit31”)**

**delta mode:**

**df.write.format(‘delta’).save(“filestore/tables/geethadelta/vsit31”)**

**again we create the dataframe df1**

**eno ename esal cont no**

**1 venu 100 23454**

**2 anu 200 34562**

**df1.write.mode(‘append’).format(‘delta’).option(“mergeschema”,True).save(“filestore/tables/geethadelta/vsit31”)**

**spark.read.format(‘delta’).load(“filestore/tables/geethadelta/vsit31”,TimestampAsof=”2022-12-24 00:00:00”).show()**

**spark.read.format(‘delta’).load(“filestore/tables/geethadelta/vsit31”,versionAsof1=1).show()**

—-----------------------------------------------------------------------------------

**Difference between Internal Delta Table. and External Delta Table. ?**

* **Internal Delta Table:** Delta Lake manages both the data and metadata within its own transaction log and directories.
* CREATE TABLE internal\_delta\_table
* USING delta
* AS
* SELECT \* FROM source\_data;

s

* **External Delta Table:** Delta Lake manages the metadata (schema, table structure, transaction log) while the data resides in an external storage system, allowing you to leverage Delta Lake's features without moving the data.
* CREATE TABLE external\_delta\_table
* USING delta
* LOCATION '/path/to/external/data';

**What is data skew.?**

Data skew is a condition in which a table's data is unevenly distributed among partitions in the cluster. Data skew can severely downgrade performance of queries, especially those with joins. Joins between big tables require shuffling data

**What are Different of transformations**

**Narrow Transformations:**

These types of transformations convert each input partition to only one output partition. When each partition at the parent RDD is used by at most one partition of the child RDD or when each partition from child produced or dependent on single parent RDD

that do not require data shuffling

Example: Filter,Union , MAP,distinct,sortBy,pivot,select

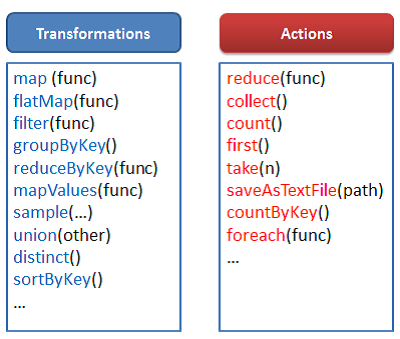
**Wide Transformations:**

This type of transformation will have input partitions contributing to many output partitions. When each partition at the parent RDD is used by multiple partitions of the child RDD or when each partition from child produced or dependent on multiple parent RDD

that do require data shuffling

Example: Join, group by, aggregate,order by

**transformation and action.**

****

**How to convert RDD into Dataframe**

We create a new rdd(i.e venurdd)

new\_df=venurdd.todf()

**What is Broadcast join. and variable?**

Broadcast join is replicating small data set overall cluster nodes and saved into memory without data shuffling join operation between datasets will be performed.

Syntax:

**BROADCAST VARIABLE**

Broadcast variable is replicating small data set overall cluster nodes and saved into memory without data shuffling join operation between datasets will be performed.

Syntax:

**df2=df.join.(Broadcast(df1),df.id==df1.id,”inner”).show()**

**What is the threshold limit of broadcast join?**

It refers to the large size of the data on a table in a broadcast join known as threshold limit.

### Accumulator.:

Accumulators are variables .and it is immutable.They are used to accumulate results from worker nodes back to the driver program in a parallel operation. Accumulators do not provide a way to read their values directly;only for write

**Difference between Coalesce. and repartition.**

**Repartition:** The repartition operation is used to increase or decrease the number of partitions in a DataFrame.it will go for shuffling it will copy the data between nodes, it will give 30 partition equal size

Syntax:

**df\_repartitioned = df.repartition(numPartitions)**

**new\_def=df.write.Repartition.mode(“overwrite”).option(inferschema,’TRUE’).save(“Path”)**

**Coalesce:** The coalesce operation is used to reduce the number of partitions in a DataFrame.

coalesce do not go for data shuffling and it will do local merging

there is no guarantee of equal size partition.

coalesce is faster than repartition.

Syntax:**df\_coalesced = df.coalesce(numPartitions)**

**new\_def=df.write.coalese.mode(“overwrite”).option(inferschema,’TRUE’).save(“Path”)**

**HOW TO CREATE A NEW SPARK SESSION?**

from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("RemoveDuplicatesExample").getOrCreate()

**–>CREATE A DATAFRAME BY USING A SPARK**

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)

**How to change Schema ?**

from pyspark.sql.types import \*

schema1=StructType([

StructField("EmployeeID",IntegerType(),True), StructField("FirstName",StringType(),True),

StructField("LastName",StringType(),True),

StructField("DepartmentID",IntegerType(),True),

StructField("JobTitle",StringType(),True),

StructField("Salary",IntegerType(),True)

])

Schema2=StructType([

StructField("DepartmentID",IntegerType(),True),

StructField("DepartmentName",StringType(),True),

StructField("ManagerID",IntegerType(),True)

])

**Empdf=spark.read.csv(path='dbfs:/FileStore/tables/Employee-2.csv', schema=schema1, header=True)**

**Deptdf=spark.read.csv(path='dbfs:/FileStore/tables/Department.csv', schema=Schema2, header=True)**

display(Empdf)

display(Deptdf)

**Filter transformation in Dataframe**

Df\_101 = Empdf.filter(Empdf["DepartmentID"] == 101)

display(Df\_101)

**How to add new column into Dataframe**

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

Or

**How to add new column into Dataframe**

from pyspark.sql.functions import col

updated\_df = Empdf.withColumn("Salary", when(col("EmployeeID") == 3, 60000).otherwise(col("Salary")))

**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.

**df2=df.join.(df1,df.id==df1.id,”inner”).show()**

**df3=df.join.(df1,df.id==df1.id,”left”).show()**

**df3=df.join.(df1,df.id==df1.id,”right”).show()**

**How to read Data from csv file to create data frame**

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

**How to write Data into storage from Data Frame**

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

**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.

**display(Empdf.join(Deptdf, on="DepartmentID", how="inner"))**

**Empdf.join(Deptdf, Empdf.DepartmentID==Deptdf.DepartmentID, 'Inner').show()**

**How to remove duplicate values from Dataframe**

**dropdf = df.dropDuplicates(["department"])**

**print("Distinct count of department :"+ str(dropdf.count()))**

**Remove duplicate. in pyspark?**

from pyspark.sql.window import Window

from pyspark.sql.functions import row\_number

Df\_duplicate=Empdf.withColumn("rn", row\_number().over(Window.partitionBy("DepartmentID").orderBy(desc("Salary"))))

Df\_duplicate1=Df\_duplicate.filter(Df\_duplicate.rn == 2)

—---------------------------------------------------------------------------------------------------------------------------

**To know the high salary. in pyspark?**

from pyspark.sql.window import Window

from pyspark.sql.functions import row\_number

Df\_highsal=Empdf.withColumn("rn", row\_number().over(Window.partitionBy("DepartmentID").orderBy(desc("Salary"))))

Df\_secondhighsal=Df\_highsal.where(Df\_highsal.rn == 2)

—----------------------------------------------------------------------------------------------------------------------------

**How to write sql commands on Dataframe?**

%sql

Select \* from user

**How to select only a few columns from a data frame?**

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()

Difference between Select & Select expression?

**Select**:

is a method that is used to select one or more columns from a DataFrame.

You can use it to create a new DataFrame with a subset of the existing columns, optionally renaming them.

**selectExpr:**

provides more flexibility for performing complex operations on columns using SQL-like expressions and functions.

**How to drop columns from a dataframe?**

from pyspark.sql.functions import col

df.drop(col("firstname")) \

.printSchema()

Or

d**f1=df.na.drop(subset[“any column”]).show()**

**What is azure DBFS.?**

It is big data massive we can store, Replication available in DBFS and performance type available in th DBMS

**What Is the interface.?**

It is supported for accessing your assets.

**What is a notebook.?**

It is a web based interface to documents containing a series of runnable cells that operate on files and tables.

**What are libraries.?**

It makes third parties and it has locally built code available to notebook at library level where we import excel files.

**What are the types of jobs?**

Two types

1.Manual

2.Schedule

**What is a databricks runtime.?**

It including apache sparks

1.databricks for machine learning

2.databricks runtime for genomics

12ML scala 2.1.2 and spark 3.0

**What is the architecture of spark?**

1.driver node.

2.worker node.

3.cluster manager

**What is a driver node.?**

It is a main key to start our spark application

the entire task divided by driver node and split into sub tasks and sent to worker nodes.

**What Is a worker node?**

It contains more than one worker node

Each worker node execute one cluster

Worker node contains some multiple executors and each executor contains some tasks and each task contains multiple partitions to save the data and process the data we need RDD and dataframe.

**What is the storage capacity of a worker node?**

the storage capacity of worker nodes 64GB and 1 partition is 128MB

2GB of data is 20 partition

**What is a cluster manager.?**

It is a mediator in between driver node and worker node

**Difference between ADF and ADB?**

ADF is a zero coding ETL inside the cloud

It contains some predefined activities

ADB is fully coding running on spark cluster

It also used to do ETL streaming analytics and machine learning kind of activities.

**What are the transformations used in pyspark?**

filter,select,union,map,flatmap,distinct,group by,join.

**What are the operations performing on RDD,dataframe and dataset?**

1.transformation

2.action

**What is left anti join and left semi join?**

**left anti join:** it will return all the records from the left side table which are not matching from the right side.

syntax:**df3=df.join.(df1,df.id==df1.id,”anti”).show()**

**left semi join:**It will return all matching records from the left side and it will give only columns from the left side.

syntax:**df3=df.join.(df1,df.id==df1.id,”semi”).show()**

**What is mount. point?**

**Mount point** which is used to give the connection between ADLS and Databricks Notebook Connection.

We create app in azure active directory the app contain tenant id,client id and secret value and in side notebook we give security informations

**dbutils.fs.mount(container name and storage name)**

.

**How to write a sql query on dataframe?**

df.CreateOrReplaceTempView(“newdf”)

**What are the ’dbutils’ in ADB?**

It is a utility provided by Databricks that allows you to **interact with the Databricks File System** (DBFS) and perform various file and data manipulation tasks within Databricks notebooks.

**What is the difference between a csv. file parquet. file?**

csv file:It is a text file and it is separated by delimiter

parquet file:more compatible with big data hadoop it's going to maintain the data in column storage.

It compresses the data and is faster to read.

**What is avro. file?**

It is suitable for writing operations.

**What is the difference between dataframe. and pandas. dataframe?**

dataframe:It running on multiple distributed cluster

pandas dataframe:it have single machine

**What is the inferschema.?**

Infer: it means to conclude information based on available data

**What are pools?**

It is a group of virtual machine that are used to execute or run job and notebooks.we have two types of pools

1.standard pools

2.high concurrency pools

**What is map. and flatmap.?**

**Map**:It is also a transformation that takes elements in RDD and the result of the function is a new value of each element in the RDD

**flatmap:It** is also one transformation used to produce multiple output elements for each input element.

**How to add a column to generate a unique key?**

* **by using monotonically.\_increasing\_id. it will generate unique number**

from pyspark.sql.functions import monotonically\_increasing\_id

**df1=df.withcolumn(“unique key”,monotonically\_increasing\_id()).show**

**How to connect sql to notebook?**

We are using a jdbc connection to connect Azure SQL DB or from On premises SQL to Azure DataBricks Notebook.

**READ:df=spark.read.jdbc(url,”path”)**

**WRITE:df.write.mode(“overwrite”).jdbc(url,”path”)**

---------------------------------status change y to active=====================

Input:

df\_emp

EmpName DateOfJoining Status

Deepak 05-01-2021 Y

Dinesh 06-01-2021 N

Ravi 07-01-2021 Y

output:

df\_emp

EmpName DateOfJoining Status

Deepak 05-01-2021 Active

Dinesh 06-01-2021 InActive

Ravi 07-01-2021 Active

result\_df = df\_emp.withColumn("Status", when(col("Status") == "Y","Active").otherwise("InActive"))

------------------------------- or------------------------------------------

result\_df = df\_emp.withColumn("Status",

when(col("Status") == "Y", "Active")

.when(col("Status") == "N", "InActive")

.when(isNull(col("Status")), "Unknown")

.otherwise(col("Status"))

==============================remove number from word=====================

**from pyspark.sql import SparkSession**

**from pyspark.sql.functions import regexp\_replace**

# Create a Spark session

**spark = SparkSession.builder.appName("RemoveNumbers").getOrCreate()**

-----------------# Sample DataFrame with a "names" column

data = [("Yughandar123",),

("John456",),

("Alice",),

("Bob789",)]

columns = ["names"]

df = spark.createDataFrame(data, columns)

--------------# Remove numbers from the "names" column using regexp\_replace

df\_without\_numbers = df.withColumn("names\_without\_numbers", regexp\_replace("names", r'\d+', ''))

df\_without\_numbers.show()

#---------------- Stop the Spark session

spark.stop()

—--------------------------------------------------------------------------------------------------------------------

**How to Create the new folder or directory**

**dbutils.widgets.mkdirs(‘dbfs:venucontainer/anu/inputfolder’)**

**dbutils.widgets.mkdirs(‘dbfs:manucontainer/sri/outputfolder’)**

**dbutils.widgets.put (‘dbfs:venucontainer/anu/inputfolder’,’Hello Venu’)**

**dbutils.widgets.Head (‘dbfs:venucontainer/anu/inputfolder’)**

**Ans:Hello Venu**

**CP:Copy the data one point to the another point  
Move:It Moves Total Data**

**dbutils.widgets.cp (‘dbfs:venucontainer/anu/inputfolder’,(‘dbfs:manucontainer/sri/outputfolder’)**

**dbutils.widgets.mv (‘dbfs:venucontainer/anu/inputfolder’,(‘dbfs:manucontainer/sri/outputfolder’)**

**How to pass the parameter?**

**day=dbutils.widgets.text(‘current date’,’4/10/2023’)**

**dbutils.widgets.get(‘current day’)**

**print(date)**

**or**

**How to Know Today's Date by Using Parameters?**

**Ans:** Today date is 4-10-2023

dbutils.widgets.dropdown(‘days,’1’ [str(X) in range (1-32)])

today=dbutils.widgets.get(‘days’)

print(today**)**

**How to call the notebook?**

**dbutils.notebook.run(‘Name of the Notebook’)**

================**Aggregation using sql**=====================================

select \* from emp

SELECT \* FROM DEPT

SELECT MAX(SAL) AS MAXSALINTABLE FROM EMP

SELECT MIN(SAL) AS MINSALINTABLE FROM EMP

SELECT AVG(SAL) AS MINSALINTABLE FROM EMP

SELECT COUNT(EMPNO) AS MINSALINTABLE FROM EMP

SELECT SUM(SAL) AS MINSALINTABLE FROM EMP

—--------------------**Aggregation using pyspark—---------------------------**

from pyspark.sql import SparkSession

from pyspark.sql.functions import max, min, avg, count, sum

# Create a Spark session

spark = SparkSession.builder.appName("AggregateExample").getOrCreate()

# Sample data for demonstration

emp\_data = [(1, "John", 1000), (2, "Jane", 1200), (3, "Alice", 1500), (4, "Bob", 1300)]

# Create a DataFrame

emp\_columns = ["EMPNO", "ENAME", "SAL"]

emp\_df = spark.createDataFrame(emp\_data, emp\_columns)

**# Calculate the required aggregate values**

**max\_sal = emp\_df.select(max("SAL").alias("MAXSALINTABLE")).collect()[0]**

**min\_sal = emp\_df.select(min("SAL").alias("MINSALINTABLE")).collect()[0]**

**avg\_sal = emp\_df.select(avg("SAL").alias("AVGSALINTABLE")).collect()[0]**

**count\_empno = emp\_df.select(count("EMPNO").alias("COUNTINTABLE")).collect()[0]**

**sum\_sal = emp\_df.select(sum("SAL").alias("SUMSALINTABLE")).collect()[0]**

GIT REPOSITORY AND CICD MECHANISM

CICD Methodology:

Here we can do two types of deployments

1.Automatic

2.Manual

Manual:Go to Datafactory and inside Manage choose ARM Templates in that two options Export and import

fist Export the File into our local PC and after that import the file into ARM Template By using Azure Devops GIT

Automatic:

Azure Devops GIT

1.Click the Pipeline there is a release pipeline and there is a articraft and stage in articraft give the details like project name,branch name,main branch,feature branch,publisher branch and go to stage level there is a option Job and click the empty jobs there is a ARM Templates after the stage it will ask what is your subscription name what is your development name and what is the region and click the ARM templates Deployment

attache the templates to the template parameters

===========================================================================================================

Git Repository:

Create a Azure Devops Account

Create a Organization name inside the organization we add new project the project should be in public level

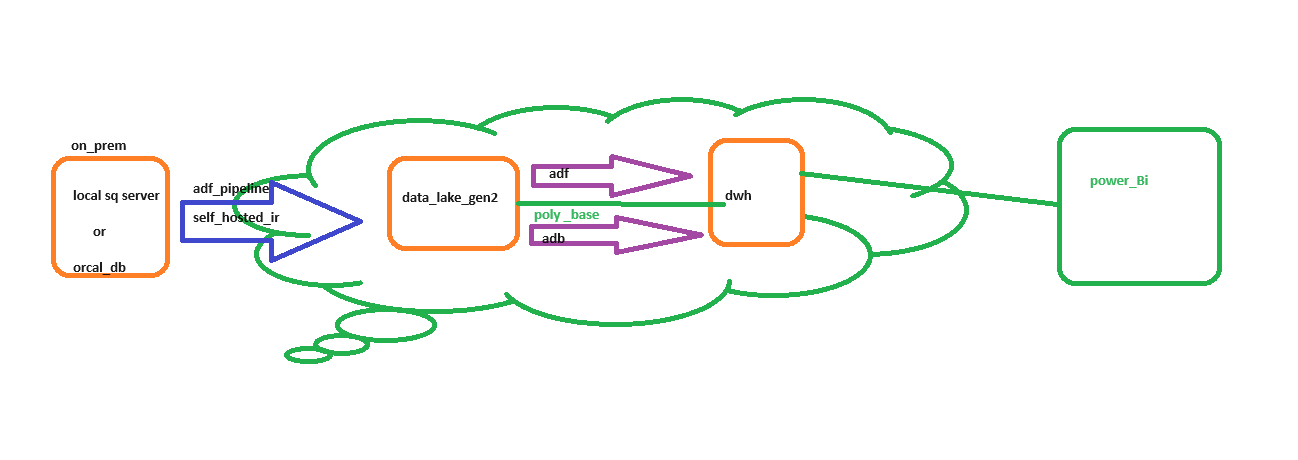
go for boards options and check the any tasks are assigned for us by scrum master or project manager and go to Repos and there is a file and create a branch branch name is 123 on it after the creating branch open the azure data factory

in that create a new azure git configuration

Inside give the repos name, branch name, project name and create a new Pipeline and create another branch

branch name is 124 and create a dataset or any activity after going to the azure devops create a new pull request because it is requesting the main person to check my work, add details like title etc.... then after Approve and complete,after completing the pull request our feature branch will be deleted

PROJECT FLOW

****

**=======PROJECT ARCHITECTURE IN EASY WAY**====================

Project name:xyz

Example:Client Name:KIA

Domain:Manufacturing and sales

Myself Venu , I have around 4+ years of experience in IT and during the period I have worked on SQL Server and Azure data engineering. As part of Azure data engineering we are using different activities inside the cloud environment like Blob,ADLS Gen 2, Azure Sql, Synapse dedicated SQL pool formally known as data warehouse, Azure data factory.

And I have worked on Agile methodology and every two weeks we are conducting 1 sprint and we are using JIRA tool for assigning user stories and also we are using azureDEVops Git repository for the deployment of our pipeline into different environment like Development to production by using CI/CD (continuous integration and continuous Deployment)

**project flow:**

Copying the data from On premises to gen2 by using an Azure Data Factory.

Taking the data from On premises files like Oracle db or Some CSV file from file system

Load the data into Data Lake gen 2 by using a Azure Data Factory with help of self hosted \_ir

and using Pipelines and different activities like Copy Data activity , get metadata activity, lookup activities for ETL process.

we are implementing incremental loading creating control table and water mark value when the new and modified records are coming into database we filter the records based on the last runtime of the pipeline and we extract the data by using lookup activity and copy data and stored procedure,once the data is available in data lake gen2 we are using staging area

After loading data use a polybase in Synapses Data WareHouse creating external tables for implementing slowly changing dimensions .

Loading the data into Dimension and Fact tables which are available in Synapses. finally we load the data into gen2 to synapse dedicated sql pool by using mappings and dataflows

After that will connect to the Powerbi for designing Visualizations and Dashboards.

**ROLES AND RESPONSIBILITY:**

Taking user stories and performing analysis, development and code reviews and unit testing

Attending every Day status calls to discuss progress of work or any blockers to task.

and I am completely involved in this project with 70% of effort on Azure Data Factory Pipelines development.

Rest involved on sql related and power bi reports designing.

============**Modern way of project flow**=================================

I am delighted to introduce myself as an experienced Azure Data Engineer with over 4+ years of overall experience.

Within this timeframe, I have dedicated 2.4 years to working extensively in the field of Azure data engineering,

honing my skills and expertise in this domain.

My proficiency in SQL, Azure Data Factory (ADF),Azure Data Bricks,Blob ,Data Lake Gen2, and Azure Synapse Analytics Dedicated SQL Pool

**Project Flow:**

==================================================================================================

Project Name is KIA This is a South Korean multinational automotive manufacturer.Its the second largest car

manufacturer in south korea.

In the Manufacturing and sales domain, our customer's source data system consists of an on-premises SQL Database and CSV files obtained from a file system

To migrate this data to the cloud environment, we utilize Azure Data Factory with the assistance of a Self-Hosted Integration Runtime.

This allows us to efficiently transfer data in the form of CSV files to Azure, where we stage it in Data Lake Gen2

To achieve incremental loading, we implement a filtering process based on the last runtime of the pipeline,

ensuring only new and modified records are extracted from the database. This is facilitated through activities such as LookUp,

Copy Data, and Stored Procedures.

After staging the data, we apply various transformation techniques, including implementing Slowly Changing Dimension (SCD1) mechanisms,

to refine and prepare the data. Once the data is ready, we load it into the Synapse Dedicated SQL Pool Data Warehouse using mapping and data flows, making it readily available in the target or destination.

Finally, our Power BI team designs insightful reports and interactive dashboards to provide valuable business insights and analytics based

on the data available in the Synapse Dedicated SQL Pool Data Warehouse. This enables our customer to make informed decisions and optimize their retail operations effectively.

**HIGH LEVEL INTERVIEW QUESTIONS**

What is the advantage of databricks over data flow?

What kind of activities do you use in adf?

What is purpose of adf?

How job parallelized across nodes in Spark?

How you can pass the value to notebook activity?

How can you return output from ADB to ADF?

What is the advantage of parquet over other formats?

What is lazy loading?

What is data shuffling? How data shuffles across nodes ? Is it good or bad?

What is hierarchical name space? Explain?

Difference between data transformation and action in Spark?

How to know datafactory pipeline failed without using azure monitor?

How databricks notebook return the value to azure data factory?

How stored procedure works?

What is the difference between delta and parquet?

How you shift one language to another language?

Why you choose spark engine over others?

What delta log stores?

If partition is making smaller chunks of the data what is the advantage to do this?

Did you work with azure synapse?

Difference between parquet and delta? Advantages / disadvantages?

What is incremental load in delta? What statement do you use?

How incremental load works?

What database you worked with ? did you work with azure sql?

What sql statement do you use?

How you make a class?

How to define a constructor?

How to do inheritance?

Azure Databricks Which is the main architecture behind data bricks?

How the job is parallelized across nodes?

What is a data bricks cluster?

What do data bricks notebooks serve?

Which possible programming languages you can use in a notebook? Which languages have you used in your projects?

How you can shift from one language to another within a notebook?

How do you go about parameterizing the data bricks notebook?

What is the delta table used for?

Use spark cluster inside the notebook: how is it used?

How to implement CI/CD and versioning on a data bricks notebook? Have you used Devops with Databricks?

Azure SQL Would it be possible to secure SQL Table at the row level?

Are you familiar with SQL queries, SQL store procedures, and T-SQL language?

What is a lock over a SQL Table?

Are you familiar with SQL tuning and optimizations?

Differences between cluster index and non cluster index

Differences between a table and a view Have you ever used Azure SQL Synapse?

ADF What kind of ADF activities have you used in your projects?

Which kind of approach do you follow when building an orchestrator pipeline in ADF?

How do you retrieve info or configuration data in a SQL table using ADF? With Lookup activity?

Do you have experience using Azure DevOps for deployment? How have you used it?

How did you perform deployment from an environment to another during your projects?

Azure devops How did you manage version controls in your past project experiences?

How is the job parallelized across nodes?

What do data bricks notebooks serve?

Which possible programming languages you can use in a notebook? Which languages have you used in your projects?

How you can shift from one language to another within a notebook? Is it possible in 1 notebook? magic command.

How do you go about parameterizing the data bricks notebook?

What is the delta table used for? Delta table vs parquet table.

Which framework you used in the notebook?

Are you familiar with SQL queries DDL,DCL,DQL etc.

SQL store procedures(query of it), how it works in adf?

T-SQL language? What kind of transaction you have done?

What is a lock over a SQL Table?

ADF What kind of ADF activities have you used in your projects?

Look up compare copy activity use case?

Use case of for each, notebook activity?

How do you retrieve info or configuration data in a SQL table using ADF? With Lookup activity?

How to pass in adf parameter?

Which database you have used?

What is broadcasting?

What is the data transformation in databricks vs adf?

What is purpose of adf?

How you can pass the parameter in adf?

How can you return output from ADB to ADF?

What is the advantage of parquet over other formats?

How to do read write operation in parquet?

Parquet vs csv?

What is data shuffling? How data shuffles across nodes ? Is it good or bad? When will you do it?

If we have blob storage then why we need azure data lake gen 1 or gen2?

Difference between data transformation and action in Spark?

Does parquet format not follow acid property?

If partition is making smaller chunks of the data what is the advantage to do this? When will you do partitioning?

Partitioning is done based on which technique?

Did you used blob or data lake?

What database you worked with ? did you work with azure sql?

What sql query do you use?

Composite model in powerbi?

Dataset used in powerbi?

What kind of data model , report you have done in powerbi?

How you have taken data from sources like direct access or importing?

What operations you have done in powerbi?

Size of data you have worked in?

Max Row count of data you have worked in?

Give some sample of your transformation of data?

Are you comfortable with javascript?

Python vs pyspark?

Magic command , widgets command- how to pass data, what kind of data you can pass.

What is the archicture of your project ?

what kind of CICD tools you have used?

what is tool used for users to report the data?

how we can connect ADB over machine learning tools ?

what is infrasturce u have to build the project ?

what is the deployment tool used in u r project ?

how can you identify the pipeline failure ?

what are the rules u follow while building the pipeline ?

if we give blank data how can you migrate that data?

what type of data u r worked with ?

***Special Thanks by Bhukya Mahender***

***Anita karuturi***

***S. Pranay Kumar***

***Boddu Yugandhar***

***syed irfan***

***munavath venkanna***