1) (C) modify the settings of the Research division workspace to use an Azure Repos repository

2) Domain → The Fabric Admin portal

3) (D) an Azure Data factory Pipeline That executes dataflow to retreive The maximum value of orderId column in destination

dataflow: Because procedure doesnot return value

maximum value: to identify The latest record already loaded in The Table

4) (B) Spark.sql("Select * From Lakehouse1.ResearchProduct")

Spark.read.format("delta").load("Tables/ResearchProduct")
extrenal_table(ResearchProduct)
| take 100

---

(64) (D) F All fabric Capacities start with f

(65) delta → tables / Producttine1

- CSV is not compatible with managed tables (not supported) which is crucial in maintaing metadata & data life cycle in lakehouse
- Parquet: only delta is allailable for queriyng via sql endpoint
- If you use saveAsTable() you don't need to specify The path "table", if you use save() you specify The full Path

(66) (A), (B) and (D) are correct: microsoft Power Bi Desketop
Power Bi Service
tabular Editor

(67) Viewer access → Read lakehouse & Data warehouse and shortcats with T-SQL Through TDS endpont

Contributer → Read lakehouse & Through lakehouse explorer
data

(5) Read All Apache Spark
Build Reports on default dataset
Read All Apache Spark

1) ReadAll Permission (Read all data using Apache Spark):

- The shared reciepient has read access to the underlying Parquet files in onelake, which can be consumed using Spark
- should be provided only if the shared recipient wants complete access to your warehouse's files using a spark engine

2) Build permission (Build reports on default dataset):
The shared recipient can build reports on top of default semantic model that is connected to your warehouse.

(6) AVerage('survey'[ResponseValue]), Period → The Period to aggregate on.
↳ only logical numeric value to aggregate

(7) View → As it shows data as soon as the data is updated (imediattely) in datastore

Case
When (Listprice > 50 And Listprice ≤ 1000 Then 'Medium')

note within the warehouse, the user can add warehouse objects:
tables or views to the default Power BI semantic model

(8) dataflow (not optimal as we can use the copy data tool to do the ingestion but it's not mentioned).

stored procedure → not making sense
xql → used for real time data
Spark notebook → mentioned use low-code as possible

(59) (C) Lakehouse - Suitable for semi-structured & unstructured
- mentioned that Files loaded by the data engineers to one lake will be stored in the Parquet format and will meet Delta Lake speci.

Hive metastore: centralized repository which store metadata for tables, Partitions, columns, schemmas and other data structure used by hive & spark

(60) (B) and (E) focousing on security & Efficiency

(61) SSO, Direct query.

Direct query Provide fresh data as it query directly from source

Direct Lake Provide almost fresh data, slightly lagging as it depends on frequency of updates to the Lakehouse files.

Single Sign on (SSO): Prefered Authintication method, it allow users to authinticate using organization credentials.
- Simplify access accross multiple systems
- " security by reducing credentials need to manage

BASIC : less roboust Authentication using username Password

Service Principal : used non-interactive with Applications and automated tools

(62) (A) & (D) Dataflows are ideal for creating date dimensions from scratch, They are suitable for calculations and transformations before storage

(63) Create a pipeline That has dependencies between activities and schedule The Pipeline

This solution is ideal as it ensure all data load activities are executed one after another in sequence

note (dataflow lacks The comprehensive orchestration capabilites required That handle dependencies for multiple workflows or different operations)

(10) (A) * where can come with group by
* Having Clause can not be used by non-aggregate coloums

(11) Row_Number() , where x=1

(12) display(df, summary= true) → Check The statistics of a given Apache Spark dataframe ✓ summary

(13) - Sql analytics endPoint operates in read-only mode over lakehouse Delta tables. You can read only data from Delta tables using The sql analytics endPoint, They can save functions, views and set sql object-level security

- External tables created by Spark can not be access by Sql EndPoint

(13) Managed tables ~~Created by~~ are fully-controlled by Lakehouse including storage, metadata management

meaning This tables are fully supported For both read & write operations Via sql Endpoints

(14) table.Profile. Provide a statistical summary of table's Coloumn (min, max, Avg, std, count)

table.Range Extracts range of rows from a table starting from specified index

table.Max Returns The maximum value in a Coloumn based on specified comparison criteria

table.MaxN Return The Top N rows with The maximum values in a Coloumn, based on The specified Comparison criteria

(15) Spark sql
Pyspark You can invoke The predict function using one of The following 3 ways:

1) Transformer API from SynapseML
2) Spark sql API
3) Pyspark user-defined function (UDF)

(16) display by interactive visualization in notebook, displa
shows data in barchart (and other various
rich format data
visualizati

(2) displayHTML → show HTML content
(3) show → show/Print df in The Console

(17) to make duplicate rows displayed in Power BI report, we can add unique field to each row

(18) Define - introduce one or more entity definations that can be applied to one or more Evaluate statement

Evaluate - Introduce a statement containing a table expression required to execute a DAX query

Summarize - Returns a summary table for the requested totals over set of groups

measure - to define a measure

Var - Store the result of an expression in a variable for reuse in calculations.

(19) (A) Distinct doesnot mean it occour only once

(20) Capacity Settings. to enable the XMLA read-write, you need to modify the capacity settings as XMLA must be enabled at the capacity level
↳ essential for creating & publishing semantic models (new way to author, collaborate and save projects)

(21) PBIP → - Store semantic model using TMDL Format in visual studio code (Microsoft) for bulk changes

PBIX → - essential for regular report development as it contains whole model & report file on Desktop

PBIT → - format used for Power Bi templates, it include the report layout & inqueries. not include actual data

~~PBIDS~~ → - useful for creating reusable templates

(22) PBIDS → File format used for Power BI data source files. it define the data source & their conn. details, but does not contain data model or report

(23) Grant Alter: give the permissions to perform DDL operations including creating, Truncating, Altering

Execute Permission: Grant executing Functions & stored procedures

(25) Automatic aggregations
use state of the art machine learning to continously optimize Direct query semantic models for maximum report query performance (self training & optmizing) Created by Power BI

user-defined- aggregations.

Aggregations explicity created & managed by model designer to optmize query performance over large Direct query semantic models.

By using these aggregations, you cache data at the aggregated level in-memory

Query Caching : is not applicable for Direct query or LiveConnect Semmantic models

(26) - Query folding : ~~Proceec~~

Process of PowerQuery query is translated into native query That can be executed by the underlying data source, such as relational database; allowing most of data transforma logic to be pushed back to the data source, leveraging its Computational power

- Note - Queries That rely solely on unstructured data Sources or That don't have a Compute engine such as CSV or Excel files, don't have query folding

- only refresh Complete days Setting :

ensure all rows of entire day are included in The refresh operation

- loading data takes an excessive amount of time and resources

Cause : Data type mismatch

The data type & format must be The same. If not : query folding cannot occu

(27) to scale out semantic model: set large dataset storage format to on

(28). Row-level security only applies to queries on a warehouse or sql analytics endpoint in fabric.

- Power Bi queries on a warehouse in Direct Lake mode will fall back to Direct query mode to abide by row-level security

(30) Vertipaq Analyzer

is a tool specifically design to analyze the memory usage of tables and columns in Power BI Models

$ system.Discover_storage_table_column_segments
dynamic management view (DMV)

- This DMV provides detailed information about the storage and memory usage of table columns within a Power BI Models

- Return information about the column segments used for storing data for in-memory tables

(32) - spliting Date/time Coloumn into separate date & time Coloumns is benefical in compression

- Each component date & time has fewer unique Values Than a Combined date/time Coloumn leading to reduce memory usage and potentially fast query Performance; If you frequently query by only one component

- Remember Calculated Coloumns are Computed & stored in The model leading to additional memory consumption

unlike measure That is Computed in-memory (calculated at query time)

(33) in Power BI Desktop and tabular Editor, From Settings, we Can UnEnable Implicit measures

(36) - to manage and utilize sql endpoint. it's essential to make shortcut in table's section to work with structured data

only tables in delta format are available in SQL analytics endpoint. Parquet, CSV and other formats can not be queried using sql analytic endpoint

(37) load to tables option capability:

tables are always loaded using the delta lake table format with V-order optimization enabled

optimize feature

used to improve the performance of existing delta table by compacting files and optimizing their layout

(38) to ~~enable~~ specify the partition column in the destination settings:

1) from destination tab: set Mode to overwrite
2) from destination tab: Enable partition
3) " " : specify the partition column

(39) Create Table AS Clone of :

only the metadata of the table is copied. The underlying data of the table, stored as parquet files, is not copied

(Zero-Copy Clone)

(40) data stores that support delta-table and V-order optimized lakehouse & warehouse

Azure db ⊗ , kql db ⊗

(41) Spark employs compression by default and its partition can be read in parallel across multiple nodes

(45) in Spark, on the existing of read & select, the spark read all the columns first, then select the specified columns, but on existing option of infer Schemma, it will read only the selecting columns and infer datatype of them.

(47) The function Value.NativeQuery in Power Query is used to execute a native sql query against a database

Enablefolding a feature in powerquery ensure query folding

a Process where Power query Pushes transformations back to The source system so, They are executed in The database or data source rather than locally.

Improve Performance by offloading heavy computations to data source

(48) In Delta tables:

1) Bin-Compaction is achieved by The optmize Command , It merges all Changes into bigger, consolidating Parquet files

2) Dereferenced Storage Clean-up is achieved by The Vacuum Command.

(49) Type 1 SCD:

always reflect The latest Values, and when Changes in Source data are detected, The dimension table data is overwritten

note managing effective end date of rows is Characterstic of Type 2 SCD

For Azure Data Lake Storage (ADLS) Gen2 account, if we need to make a shortcut to it:

Connection: https

Endpoint: dfs

(51) In Broadcast join: It minimize data schuffling

- is an optmization technique used in Spark sql engine.
- it utilized when one of dataframes is small enough to be stored in the memory of all executor nodes
- performance impoving by of join operation by minimizing data schuffling across the network
- It's advantageous when working with large dataset

(52) Automatic: allow Direct Lake with fall back to Direct Query (default behavior)

(54) df.explain() - Print (logical & physical) Plans to the Console for debugging Purposes.

df.show() - Prints the first n rows of df to Console

df.summary() - Computes specified statistics for numeric and string Coloumns (mean - std - min - max - ..

(57) Retrieve Delta table History: → Help in maintaince and reviewing

You can retrieve information including The operations, user, timestamp for each write to a Delta table by running The history Command for (30 days)

Describe History table_name -- get The full history of The table

Describe History table_name limit 1 -- get The last operation only

Describe Detail -- get information about schema, Partitioning, table size and so on.

Refreash Table table_name

Command to refresh The metadata and date of a specific table. It ensures That The table's metadata and data in Spark's catalog are synchronized with The underlying data source, especially after external changes

Explain table table_name

used to provide logical/Physical plans for an input statem

(73) Dataflow support Power Query M formula language expressions That has transformations capabilities during ingestion (copy data activity donot)

Error if Exists: on saving dataframe, if data already exists: an exception to be thrown

Ignore: " " "

" : The save operation is excpected to not to save the contents of the df and not to change existing data

Append

Overwrite

(75) In sql: Function LEAST (expr [, ... exprN])

Return the minimum value from a list of one or more expressions

(79) Fuzzy merge: Type of data join that allows for inexact matching between columns in 2 dataset

useful when keys don't match because formatting or typos

Fuzzy Append — Similar to merge, but focuses on appending new rows to a dataset based on inexact matching

(79) in sql analytic Endpoint:

You can modify the default behaviour to not automaticly add tables and Views To the default Power BI Semantic model

(83) lookup activity in data factory and Synapse analytics:

- read and returns the contents of a configuration file or table (data from a source)
- also returns the result of executing a query or stored Procedure

It can be consumed in a subsequent copy, transformation, or Control flow activities (ForEach activity)

Switch activity:

evaluates a set of activites Corresponding to a case That matches the Condition evaluation

Append Variable activity:

to Add a Value to an existing array variable defined in a data Factory or synapse Analytic pipeline

(90)

Calculate

Function eValuates an expression in a modified Filter context

Calculate(sum(sales[Amount]), Sales[Region] = "North")

expression ↙ Filter 1 ↙

Generate

used to create a Cartesian Product between 2 tables

Generate (Products, Customers)

CombineValues

Concatenate Values from multiple Coloumns or expressions into a single text string, separated by a specified delimeter

CombineValues ("", Employees[FirstName], Employees[lastName])

delimeter ↙ expression 1 ↙ expression 2 ↙

SelectedMeasure

Function returns the measure currently being evaluated in a calculation group. Used within calculation items in a calculation group SelectedMeasure()

SelectedValue

returns the value when there is a single value in the context; otherwise, it return an alternative result (Blank)

SelectedValue(Products[ProductName], "No Products selected")

alternative Result

(94)

	Boolean Expression	table Expression
output	- True or false value for each row	- A table with rows & columns
usage scope	- operates on one/more columns	- operates on entire table or subset of it
usage	- Direct filtering in calculate, keepfilters	- functions requiring a table Filter, sumx
Example	Sales[Region] = "East"	Filter(Sales, Sales[Amount] >100)

The Calculate Function accepts table expression by Filter Dax / Which eValuates its filter expression for each row of The table

However, it Could be achieved much more efficently by using Boolean expression

Here is Code That use Boolean expression using KeepFilters Dax function which ensures any existing filters applied and preserved, not overwritten

```
RedSales =
Calculate(
[Sales],
KeepFilters('Product'[Color] = "Red")
)
```

It's recommended to Pass Filter arguments as Boolean expression

It's Because Import model tables are in-memory Coloumn stores. They are explicity optmized to efficiently Filter Coloumns in this way

Can not refrence Coloumns from multiple tables

(99) DateTrunc

used to truncate a date or datetime value to a specified Precision, such as year, quarter, month, ...etc (remove smaller components of date)

select Datetrunc (month, orderDate);

↙ to first of that month in that orderDate

DateName

returns a character string representing the specified datepart of specified date

DateName (weekday, orderdate);

↳ name of weekday (sunday) in that orderdate

day → number of day (not name)

- Dataflow and Piplelines Can not be Shared at item level

1- Read all data using sql (Readdata)
⇒ Read all objects within warehouse using T-sql

2- Read all oneLake data (ReadAll)
⇒ Read all the warehouse's underlying oneLake files using Apache Spark, Pipelines, or other apps that access the oneLake data directly

3- Build reports on the default semantic model Connected to warehouse (Build)

- to implement Row-level security, user should have access either → Readdata
→ object_level security

* to Create a RLS:

1) create schema
2) Create a function That Filter a table based on logged-user
3) Create a security policy to apply The filtering function

* to implement OLS & CLS

⇒ The user should have access Read permission only to warehouse

⇒ Can be done at warehouse or lakehouse t-sql endpoint

* using Manage onelake Data Access (PREVIEW)

you can: 1) Define roles
2) For each role give ↓ access to specific tables / Files

* to Configure access, you have to be: Admin
Member
Contributer
of workspace

Data Governance

* Sensivity Labels :

Labels needs to be created and managed in microsoft Purview (before They become accessible in Fabric)

in some industries, labelling data and information with a sensitivity label is necessary for compliance with information protection regulations

* Endorsement :

Endores quality content to help Colleagues discover it

Promoted ⇒ recommended for others to use it

Certified ⇒ certified by your org as trusted source

Master Data ⇒ core source (single truth) of organizational data records

Reusable Power BI assets

1) Power BI template can act as a starting point for a new report's layout, data model, and queries

- It's a reusable asset which can improve efficiency & consistency

- Export an existing report (.Pbix file) as a Power BI Template
- If there are any Parameters in the template, you are asked to input these at the time of creating a new report

2) The (.Pbids file) is a power BI data source file, which include connection details to a data source

- Can help to quickly transfer all the data connections from one Project to another

- Pbids file can be exported from an existing report file

- file can be edited in a text editor accordingly

* you can share semantic model from either:
  1) semantic model's options menu on OneLake data hub (catalog)
  2) data details page

options ... modify ... share ... manage ...

- we make version control at workspace level (settings)
  - Git in Azure RePos → Same tenant as Fabric tenant
  - GitHub
  - GitHub Enterprise

- The semantic model can be represented

1) TMSL — model.bim file
2) TMDL — defination folder

Semantic Model deployment (XMLA endpoint)

XMLA Endpoint: Support open-platform connectivity from microsoft and third-party client applications and tools to the semantic models workload

By default, read only connectivity using the endpoint is enabled

Read-write Provides more semantic model management, governance, advanced semantic modeling, debugging and monitoring

They are available to Premium or Embedded only

* Ingest batch with Dataflows Gen 2, Pipelines or notebooks

* Can ingest data for time series analysis in GUI-based way with eventstreams

*

lakehouse → storing raw files and transforming to delta tables, That can be manipulated by spark (pyspark - sparksql)

warehouse → use T-sql to work with large Volumes of data

EventHouse → storing timeseries data That can be queried by KQL

* You can convert coloum data types:

1) Dataflow Gen2
2) Pyspark in notebooks

Direct Lake for Power Bi :

Prequisite: Fabric Capacity / Power Bi premium
lakehouse + Sql Endpoint / warehouse
Delta tables

limitations

- Quering one single lakehouse or warehouse
- T-Sql View are not supported (will Fall back to DQ)
- Dax querics exceeding limits or using unsupported Features Fall back to DQ mode

- No Dax Calculated cols
- No Composite model
- No Date time relationships

* Relation_ship Filter direction
Single Both

Dax Iterator Function

- Iterate over a table, evaluate the expression for each row, and then aggregate result
- Ends with X
- must whenever morethan 1 Coloumn is used in the expression

Aggregation

When you create Aggregation in composite model

original table → must be direct query

with Dim tables → you can set Dual storage mode to overcome limited relationships

Aggregate tables → in import mode gives the maximum performance

* Unlike CSV files, Parquet files have the schema and data types embedded within them

Parquet files are the format supported by deltaLake which allows for a transactional/data warehousing style capability to be used alongside the files

Parquet are better for compression