4. Python (Added Advantage)

1. Why is Python popular in data engineering and analytics?
2. Difference between **list, tuple, and dictionary**.
3. What are Python libraries commonly used for data processing?
4. Explain basic data cleaning steps in Python.
5. What is Pandas DataFrame?
6. Write a simple Python program to read a CSV file and display first 5 rows.

5. Data Engineering & Analytics (Key Focus Area)

Conceptual Questions

1. What is **Data Engineering**?
2. Difference between **OLTP** and **OLAP** systems.
3. What is **ETL**? Explain with an example.
4. What is a **data pipeline**?
5. Difference between **structured, semi-structured, and unstructured data**.

Azure Data Factory (ADF)

1. What is Azure Data Factory?
2. What are **pipelines, datasets, and linked services** in ADF?
3. What is a **trigger** in ADF?
4. Difference between **Copy Activity** and **Data Flow**.
5. How do you handle failures in ADF pipelines?

Azure Data Lake

1. What is **Azure Data Lake**?
2. Difference between **Data Lake** and **Data Warehouse**.
3. What are the benefits of storing data in a Data Lake?
4. What file formats are commonly used in Data Lakes?

Microsoft Fabric (Basic Awareness)

1. What is **Microsoft Fabric**?
2. How is Fabric related to Power BI and Azure?
3. What is **OneLake**?
4. Why is Fabric important for modern analytics?

6. Power BI (Very Important)

1. What is Power BI?

2. Difference between **Power BI Desktop** and **Power BI Service**.
3. What are **facts and dimensions**?
4. What is **DAX**?
5. Explain **measures vs calculated columns**.
6. How do you optimize a Power BI report?
7. What is **data refresh** in Power BI?
8. Have you created any dashboards? Explain one use case.

7. MS Excel (Mandatory)

1. Difference between **VLOOKUP, HLOOKUP, and XLOOKUP**.
2. What are **Pivot Tables** and where are they used?
3. Explain **IF, COUNTIF, SUMIF** functions.
4. How do you remove duplicates in Excel?
5. What is conditional formatting?
6. How is Excel used in data analysis?

8. SQL & Data Basics (Expected Even if Not Explicit)

1. What is a database?
2. Difference between **SQL and NoSQL**.
3. What are **primary key and foreign key**?
4. Write a query to fetch top 5 records from a table.
5. Difference between **WHERE** and **HAVING**.
6. What is normalization?

🔹 **Power BI – Core Concepts (With Answers)**

Q1. What is Power BI and why is it used?

Answer:
Power BI is a business analytics tool by Microsoft used to **collect, transform, model, and visualize data** to support data-driven decision-making.

Q2. What are the main components of Power BI?

Answer:

1. Power BI Desktop – Report creation
2. Power BI Service – Cloud publishing & sharing
3. Power BI Mobile – Mobile access
4. Power BI Gateway – On-premises data connectivity

Q3. What is the Power BI workflow?

Answer:
Data Source → Power Query → Data Model → DAX → Reports → Dashboard → Share

Q4. What is Power Query?

Answer:
Power Query is used for **data extraction, cleaning, and transformation** before loading data into Power BI.

Q5. What is DAX?

Answer:
DAX (Data Analysis Expressions) is used to create **measures, calculated columns, and tables** for data analysis.

🔹 **Data Modeling & Relationships**

Q6. What is data modeling in Power BI?

Answer:
It is the process of defining **relationships** between tables to enable accurate reporting and analysis.

Q7. Difference between Star Schema and Snowflake Schema?

Answer:

- **Star Schema**: One fact table connected to dimension tables (preferred in Power BI)
- **Snowflake Schema**: Dimension tables further normalized

Q8. What is a fact table?

Answer:
A fact table stores **measurable data** like sales amount, quantity, revenue.

Q9. What is a dimension table?

Answer:
Dimension tables store **descriptive attributes** like date, product, customer, location.

Q10. What are relationships in Power BI?

Answer:
Relationships define how tables are connected (one-to-many, many-to-many).

🔹 **DAX – Important Interview Questions**

Q11. Difference between Calculated Column and Measure?

Answer:

Calculated Column	Measure
Stored in table	Calculated at runtime
Uses row context	Uses filter context
Increases model size	Efficient

Q12. What is row context?

Answer:
Row context means DAX calculation applies **row-by-row**.

Q13. What is filter context?

Answer:
Filter context is applied by **filters, slicers, and visuals**.

Q14. Explain CALCULATE() function.

Answer:
CALCULATE() modifies filter context to perform dynamic calculations.

Example:

Total Sales = CALCULATE(SUM(Sales[Amount]), Sales[Region] = "East")

Q15. What is SUMX()?

Answer:
An iterator function that performs calculations **row-by-row** and then aggregates.

Q16. What is TIME INTELLIGENCE in Power BI?

Answer:
Functions that analyze data over time like YTD, MTD, SAMEPERIODLASTYEAR.

Q17. What is ALL() in DAX?

Answer:
Removes filters from a table or column.

Q18. What is RELATED()?

Answer:
Fetches related values from another table using relationships.

🔹 **Power BI Visuals & Reports**

Q19. Difference between Report and Dashboard?

Answer:

- Report: Multi-page, interactive (Power BI Desktop)
- Dashboard: Single page, summary view (Power BI Service)

Q20. What is a slicer?

Answer:
A visual filter that allows users to dynamically filter data.

Q21. What are custom visuals?

Answer:
Third-party or custom-built visuals from AppSource.

Q22. What is drill-down and drill-through?

Answer:

- Drill-down: Navigate within same visual
- Drill-through: Navigate to another report page

Q23. What is tooltip?

Answer:
Displays additional information when hovering over visuals.

🔹 **Performance Optimization**

Q24. How do you optimize Power BI performance?

Answer:

- Use star schema
- Reduce columns
- Prefer measures over calculated columns
- Avoid bi-directional relationships
- Use import mode when possible

Q25. What is cardinality?

Answer:
Number of unique values in a column. High cardinality impacts performance.

🔹 **Security & Sharing**

Q26. What is Row-Level Security (RLS)?

Answer:
Restricts data access at row level for different users.

Q27. What is Power BI Gateway?

Answer:
Connects on-premises data sources to Power BI Service.

Q28. What is workspace?

Answer:
A collaboration area to manage datasets, reports, and dashboards.

🔹 **Real-World / Scenario Questions (Very Important)**

Q29. How would you build a sales dashboard?

Answer:

1. Connect data
2. Clean in Power Query

3. Create relationships
4. Write DAX measures
5. Build visuals
6. Publish & share

Q30. How does Power BI fit into Azure Data Factory?

Answer:
ADF handles ETL, Power BI handles visualization and reporting.

🔹 **Power BI Interview One-Liners (HR Loves These)**

- "Power BI turns raw data into actionable insights."
- "Measures are preferred over calculated columns for performance."
- "Star schema improves query efficiency."
- "DAX works on row and filter context."

🔹 **Common Fresher Mistakes (Avoid Saying)**

❌ "Power BI is only for charts"
❌ "Calculated columns and measures are same"
❌ "RLS is optional in enterprise projects"

🔹 **Excel – Core Fundamentals (With Answers)**

Q1. What is Microsoft Excel used for?

Answer:
Excel is used for **data storage, analysis, calculation, reporting, and visualization** using rows, columns, formulas, and charts.

Q2. What is a workbook and a worksheet?

Answer:

- **Workbook**: Excel file (.xlsx)
- **Worksheet**: Individual sheets inside a workbook

Q3. What is a cell address?

Answer:
Combination of column letter and row number (example: A1).

🔹 **Formulas & Functions (Very Important)**

Q4. Difference between formula and function?

Answer:

- **Formula**: User-written calculation (=A1+B1)
- **Function**: Built-in formula (=SUM(A1:A10))

Q5. What are relative, absolute, and mixed references?

Answer:

- Relative: A1
- Absolute: A1
- Mixed: $A1 or A$1

Q6. What is IF() function?

Answer:
Used for conditional logic.

=IF(A1>=70,"Pass","Fail")

Q7. Difference between COUNT, COUNTA, COUNTBLANK?

Answer:

- COUNT: Numbers only
- COUNTA: Non-empty cells
- COUNTBLANK: Empty cells

Q8. Explain SUMIF and COUNTIF.

Answer:
Used for conditional aggregation.

=SUMIF(A:A,"East",B:B)

🔹 **Lookup & Reference Functions (Highly Asked)**

Q9. What is VLOOKUP?

Answer:
Searches a value vertically and returns matching data.

=VLOOKUP(A2,Table,2,FALSE)

Q10. Limitations of VLOOKUP?

Answer:

- Works only left to right
- Breaks if columns change
- Slower on large datasets

Q11. What is XLOOKUP?

Answer:
Modern replacement for VLOOKUP with no direction limitation.

=XLOOKUP(A2,A:A,B:B)

Q12. Difference between INDEX + MATCH and VLOOKUP?

Answer:
INDEX+MATCH is more flexible and efficient.

🔹 **Data Cleaning & Preparation**

Q13. How do you remove duplicates?

Answer:
Data → Remove Duplicates

Q14. What is Text to Columns?

Answer:
Splits data using delimiter (comma, space, etc.).

Q15. What are TRIM, CLEAN, PROPER functions?

Answer:
Used to clean text data.

Q16. How do you handle missing values in Excel?

Answer:

- Filter blanks
- Use IFERROR
- Replace with mean/median

🔹 **Sorting, Filtering & Validation**

Q17. What is data validation?

Answer:
Restricts input to avoid incorrect data entry.

Q18. Difference between filter and sort?

Answer:

- Filter: Show specific data
- Sort: Rearrange data order

Q19. What is conditional formatting?

Answer:
Formats cells based on rules (color scale, icons).

🔹 **Pivot Tables (Extremely Important)**

Q20. What is a Pivot Table?

Answer:
Summarizes large datasets quickly.

Q21. What are rows, columns, values, filters in Pivot?

Answer:
They define how data is grouped and summarized.

Q22. Difference between Pivot Table and Normal Table?

Answer:
Pivot table provides dynamic summarization.

Q23. What is a Pivot Chart?

Answer:
Graphical representation of Pivot Table data.

🔹 **Excel Charts & Visualization**

Q24. Types of charts in Excel?

Answer:
Bar, Line, Pie, Column, Scatter, Area

Q25. When do you use a line chart?

Answer:
To show trends over time.

- **Advanced Excel (Interview Gold)**

Q26. What is Power Query in Excel?

Answer:
Tool for data extraction, transformation, and loading (ETL).

Q27. What is Power Pivot?

Answer:
Used for data modeling and large datasets.

Q28. What are macros?

Answer:
Automated tasks using VBA.

Q29. What is IFERROR()?

Answer:
Handles formula errors.

=IFERROR(A1/B1,0)

Q30. What is CSV file?

Answer:
Comma-Separated Values, plain text format.

- **Real-World Scenario Questions**

Q31. How do you prepare Excel data for Power BI?

Answer:
Clean headers, remove blanks, use proper data types, normalize tables.

Q32. How do you analyze sales data in Excel?

Answer:
Pivot tables + charts + filters.

Q33. How do you find top 5 values?

Answer:
Using sort, filters, or LARGE() function.

🔹 **Excel Interview One-Liners**

- "Pivot tables summarize large datasets quickly."
- "XLOOKUP is more flexible than VLOOKUP."
- "Power Query enables ETL inside Excel."
- "Excel is often the first step in data analysis."

🔹 **Common Fresher Mistakes**

❌ Hard-coding values
❌ Not locking cell references
❌ Using merged cells in data tables

Below are **more domain-specific interview questions with clear, fresher-friendly answers**, focused on **Sasken's business areas** and **Data / Software roles**. These are ideal for **offline technical interviews** where understanding + application is tested.

🔹 **1. Automotive Domain (Connected, Intelligent & Autonomous Vehicles)**

Q1. What is a connected vehicle?

Answer:
A connected vehicle uses internet and wireless communication to exchange data with other vehicles, infrastructure, cloud, and devices. It enables features like GPS navigation, remote diagnostics, OTA updates, and emergency assistance.

Q2. What role does data analytics play in automotive systems?

Answer:
Data analytics helps in:

- Predictive maintenance (detect faults early)
- Driver behavior analysis
- Traffic pattern analysis
- Fuel efficiency optimization
- Autonomous driving decisions

Q3. What is ADAS?

Answer:
ADAS (Advanced Driver Assistance Systems) are safety features like lane assist, adaptive cruise control, collision avoidance, and parking assist that help reduce human error.

Q4. What is an OTA (Over-The-Air) update?

Answer:
OTA updates allow vehicle software to be updated remotely without visiting a service center, improving features and fixing bugs.

Q5. What is V2X communication?

Answer:
V2X (Vehicle-to-Everything) enables communication between vehicles and infrastructure (V2V, V2I, V2P), improving safety and traffic efficiency.

🔹 **2. Telecom Domain (5G, Networks & Infrastructure)**

Q6. How is 5G different from 4G?

Answer:
5G offers:

- Higher speed
- Lower latency
- Higher device connectivity
- Better support for IoT, AR/VR, and autonomous vehicles

Q7. What is latency and why is it important?

Answer:
Latency is the time taken for data to travel from source to destination. Low latency is crucial for real-time applications like autonomous driving and remote surgery.

Q8. What is a core network?

Answer:
The core network manages authentication, routing, data sessions, and billing in telecom systems.

Q9. What is network analytics?

Answer:
Network analytics uses data to monitor performance, detect faults, optimize bandwidth, and improve customer experience.

Q10. How does data engineering help telecom companies?

Answer:
It helps process massive call, usage, and network data for:

- Fraud detection
- Network optimization
- Customer churn analysis
- Capacity planning

🔹 3. Semiconductor Domain (Chip Design & Embedded Systems)

Q11. What is a semiconductor?

Answer:
A semiconductor is a material (like silicon) whose conductivity lies between conductor and insulator and is used in chips and electronic devices.

Q12. What is chip verification?

Answer:
Chip verification ensures that a chip design behaves correctly before manufacturing, reducing costly errors.

Q13. What is an embedded system?

Answer:
An embedded system is a dedicated computer system designed for a specific function, such as a car ECU or washing machine controller.

Q14. Why is software important in semiconductor design?

Answer:
Software is used for simulation, testing, firmware development, and controlling hardware behavior.

Q15. What is firmware?

Answer:
Firmware is low-level software that directly interacts with hardware, often stored in ROM or flash memory.

🔹 4. Industrial & IoT Domain (Manufacturing, Energy)

Q16. What is Industrial IoT (IIoT)?

Answer:
IIoT connects machines and sensors in industries to monitor performance, automate processes, and predict failures.

Q17. What kind of data is collected in IIoT?

Answer:

- Temperature
- Pressure
- Vibration
- Energy consumption
- Machine runtime

Q18. What is predictive maintenance?

Answer:
Predictive maintenance uses data and analytics to predict equipment failure before it happens, reducing downtime.

Q19. How does cloud computing help IIoT?

Answer:
Cloud enables scalable storage, real-time analytics, and remote monitoring of industrial data.

Q20. What is SCADA?

Answer:
SCADA (Supervisory Control and Data Acquisition) systems monitor and control industrial processes.

🔹 5. Consumer Electronics Domain

Q21. What is meant by user experience (UX)?

Answer:
UX refers to how easy and intuitive a device is for users, including performance, interface, and responsiveness.

Q22. How is data used in consumer electronics?

Answer:
Data is used to:

- Analyze user behavior
- Improve product features
- Detect usage patterns
- Enhance personalization

Q23. What is edge computing?

Answer:
Edge computing processes data closer to the device instead of the cloud, reducing latency and bandwidth usage.

🔹 **6. Satellite Communication (Satcom)**

Q24. What is satellite communication?

Answer:
Satcom uses satellites to transmit data, voice, and video over long distances, especially where terrestrial networks are unavailable.

Q25. What are ground stations?

Answer:
Ground stations communicate with satellites for data transmission, control, and monitoring.

Q26. Why is low latency important in Satcom?

Answer:
Low latency ensures real-time communication for navigation, defense, and emergency services.

🔹 **7. Transportation & Smart Mobility**

Q27. What is smart transportation?

Answer:
Smart transportation uses data, IoT, and analytics to improve traffic management, public transport, and logistics.

Q28. How does analytics help fleet management?

Answer:
It helps in:

- Route optimization
- Fuel efficiency
- Vehicle health monitoring
- Driver performance analysis

Q29. What is real-time data processing?

Answer:
Processing data immediately as it arrives, essential for live tracking and alerts.

Q30. Why is data security important in transportation systems?

Answer:
Because sensitive data like location, vehicle control, and user information must be protected from cyber threats.

🔹 **Interview Tip for Freshers (Very Important)**

👉 **Link domain + data/software** in every answer.
Example:

> "In automotive systems, data engineering helps process sensor data using cloud pipelines for real-time decision making."

🔹 **Azure Data Factory – Core Concepts**

Q1. What is Azure Data Factory (ADF)?

Answer:
Azure Data Factory is a **cloud-based ETL/ELT service** used to **extract, transform, and load data** from multiple sources into data stores like Azure Data Lake, SQL, or Synapse.

Q2. What problems does ADF solve?

Answer:

- Automates data movement
- Integrates data from multiple sources
- Schedules workflows
- Enables scalable data pipelines

Q3. Difference between ETL and ELT?

Answer:

- **ETL**: Transform data before loading
- **ELT**: Load raw data first, then transform
  ADF supports **both**.

🔹 **ADF Architecture & Components (Very Important)**

Q4. What are the main components of ADF?

Answer:

1. Pipelines
2. Activities
3. Datasets
4. Linked Services
5. Integration Runtime (IR)

Q5. What is a Pipeline?

Answer:
A pipeline is a **logical group of activities** that perform a complete ETL workflow.

Q6. What is an Activity?

Answer:
An activity defines **what action to perform** (copy, transform, execute).

Q7. What is a Dataset?

Answer:
A dataset represents **data structure** inside a data store (table, file, folder).

Q8. What is a Linked Service?

Answer:
Linked service defines **connection information** to data sources.

Q9. What is Integration Runtime (IR)?

Answer:
IR provides the **compute environment** to move and transform data.

Types:

- Azure IR
- Self-Hosted IR
- Azure-SSIS IR

🔹 **Data Movement & Transformation**

Q10. What is Copy Activity?

Answer:
Used to **copy data** from source to destination.

Q11. What is Mapping Data Flow?

Answer:
A visually designed data transformation service inside ADF.

Q12. Difference between Data Flow and Power Query?

Answer:

- Data Flow: Scalable, Spark-based, production ETL
- Power Query: Interactive, smaller transformations

Q13. What transformations are available in Data Flow?

Answer:
Filter, Derived Column, Join, Aggregate, Lookup, Sort.

🔹 **Control Flow & Orchestration**

Q14. What is Control Flow in ADF?

Answer:
Manages **execution logic** like conditions, loops, dependencies.

Q15. What is If Condition activity?

Answer:
Executes activities based on true/false conditions.

Q16. What is ForEach activity?

Answer:
Iterates over a collection and executes activities repeatedly.

Q17. What is Until activity?

Answer:
Repeats execution until a condition is met.

🔹 **Triggers & Scheduling**

Q18. What are triggers in ADF?

Answer:
Triggers define **when pipelines run**.

Q19. Types of triggers?

Answer:

- Schedule trigger
- Tumbling window trigger
- Event-based trigger

Q20. What is Tumbling Window trigger?

Answer:
Executes pipelines in **fixed, non-overlapping time intervals**.

🔹 **Parameters, Variables & Expressions**

Q21. What are parameters?

Answer:
Used to pass values dynamically to pipelines, datasets, activities.

Q22. What are variables?

Answer:
Used to store temporary values during pipeline execution.

Q23. What is an expression in ADF?

Answer:
Dynamic content using functions like @concat(), @utcnow().

🔹 **Error Handling & Monitoring**

Q24. How do you handle errors in ADF?

Answer:

- Use failure paths
- Set retry policies
- Log errors to storage

Q25. What is monitoring in ADF?

Answer:
Tracking pipeline runs, activity status, execution time, failures.

Q26. What is retry policy?

Answer:
Defines number of retries and interval if an activity fails.

🔹 **Security & Best Practices**

Q27. How do you secure credentials in ADF?

Answer:
Using **Azure Key Vault**.

Q28. What is Managed Identity?

Answer:
Provides secure access to resources without storing credentials.

Q29. Best practices in ADF?

Answer:

- Parameterize pipelines
- Use Key Vault
- Separate DEV/TEST/PROD
- Monitor failures

🔹 **Performance & Optimization**

Q30. How do you improve ADF performance?

Answer:

- Parallel copy
- Partition data
- Optimize file sizes
- Use appropriate IR

Q31. What is data partitioning?

Answer:
Splitting data to process it in parallel.

🔹 Real-World Scenario Questions (High Value)

Q32. Design a simple ADF pipeline.

Answer:
Source → Copy Activity → Azure Data Lake → Power BI

Q33. How does ADF integrate with Power BI?

Answer:
ADF prepares data; Power BI consumes it for reporting.

Q34. How does ADF fit in Azure Data Lake architecture?

Answer:
ADF ingests raw data into Data Lake zones (raw, curated).

Q35. Difference between ADF and SSIS?

Answer:
ADF is cloud-native; SSIS is traditional ETL.

🔹 Interview One-Liners (Memorize)

- “ADF is an orchestration service, not a data store.”
- “Copy activity moves data; data flows transform data.”
- “Integration Runtime defines compute.”
- “ADF supports both ETL and ELT.”

🔹 Common Fresher Mistakes

❌ Thinking ADF stores data
❌ Hard-coding paths
❌ Ignoring monitoring
❌ Not using parameters

Azure Data Lake – Core Fundamentals

Q1. What is Azure Data Lake?

Answer:
Azure Data Lake is a **scalable cloud storage service** optimized for **big data analytics**, capable of storing structured, semi-structured, and unstructured data.

Q2. Difference between Azure Data Lake and Azure Blob Storage?

Answer:
Azure Data Lake Gen2 is built on Blob Storage but adds:

- Hierarchical namespace
- Better performance for analytics
- Fine-grained security (ACLs)

Q3. What is ADLS Gen2?

Answer:
ADLS Gen2 combines **Blob Storage + Data Lake capabilities**, making it ideal for analytics workloads.

🔹 **Data Types & Storage**

Q4. What types of data can be stored in Data Lake?

Answer:

- Structured (tables, CSV)
- Semi-structured (JSON, XML)
- Unstructured (images, logs, videos)

Q5. Why is Data Lake preferred over Data Warehouse for raw data?

Answer:
Because Data Lake:

- Stores raw data
- Is schema-on-read
- Cheaper storage
- Supports large-scale analytics

Q6. What is schema-on-read?

Answer:
Schema is applied **when data is read**, not when it is stored.

🔹 **Folder Structure & Zones (Very Important)**

Q7. What are Data Lake zones?

Answer:
Logical separation of data for better governance:

- **Raw (Bronze)** – Unprocessed data
- **Curated (Silver)** – Cleaned data
- **Analytics (Gold)** – Aggregated data

Q8. Why do we use zones?

Answer:
For:

- Data quality control
- Easier debugging
- Better security
- Clear data lifecycle

Q9. Example of Data Lake folder structure?

Answer:

/raw/sales/2025/01/

/curated/sales/

/analytics/sales/

🔹 **Security & Access Control**

Q10. How is security handled in Azure Data Lake?

Answer:

- Azure RBAC
- Access Control Lists (ACLs)
- Managed Identity
- Azure Key Vault

Q11. Difference between RBAC and ACL?

Answer:

- **RBAC**: Account-level access
- **ACL**: File/folder-level access

Q12. What is Managed Identity?

Answer:
Allows Azure services like ADF to access Data Lake **without storing credentials**.

🔹 **Integration with Azure Services**

Q13. How does ADF use Data Lake?

Answer:
ADF ingests and transforms data into Data Lake zones.

Q14. How does Power BI connect to Data Lake?

Answer:
Using:

- DirectQuery
- Import mode
- Via Synapse or SQL endpoints

Q15. How does Data Lake integrate with Azure Synapse?

Answer:
Synapse reads data directly from Data Lake for analytics.

🔹 **File Formats & Performance**

Q16. Common file formats used in Data Lake?

Answer:
CSV, JSON, Parquet, Avro, ORC.

Q17. Why is Parquet preferred?

Answer:

- Columnar format
- Faster queries

- Smaller file size

Q18. What is partitioning in Data Lake?

Answer:
Organizing data by folders (date, region) to improve performance.

Q19. Difference between small files and large files issue?

Answer:
Too many small files degrade performance; recommended size is 100MB–1GB.

🔹 **Data Governance & Management**

Q20. What is data governance?

Answer:
Ensuring data quality, security, and compliance.

Q21. What is data lineage?

Answer:
Tracking data movement from source to destination.

Q22. How do you handle data retention?

Answer:
Using lifecycle management policies.

🔹 **Real-World Scenarios (Highly Asked)**

Q23. Design a simple Data Lake architecture.

Answer:
Source → ADF → Data Lake (Raw → Curated) → Power BI

Q24. How do you store IoT data in Data Lake?

Answer:
In partitioned folders by device and date.

Q25. How do you handle incremental loads?

Answer:
Using watermark columns (date/time).

🔹 **Data Lake vs Other Azure Services**

Q26. Data Lake vs SQL Database?

Answer:
Data Lake stores files; SQL DB stores structured tables.

Q27. Data Lake vs Data Warehouse?

Answer:
Data Lake = raw, flexible
Warehouse = structured, analytics-ready

🔹 **Interview One-Liners (Memorize)**

- "Data Lake stores raw data at scale."
- "Schema-on-read provides flexibility."
- "Parquet improves performance."
- "ADF orchestrates, Data Lake stores."

🔹 **Common Fresher Mistakes**

❌ Treating Data Lake as a database
❌ No folder structure
❌ Not securing raw data
❌ Too many small files

Below is a **deep, interview-oriented guide on Microsoft Fabric** with **clear, fresher-friendly questions and answers**, aligned to **Data Engineering, Analytics, Power BI, ADF, and Azure Data Lake**—exactly the stack Sasken looks for.

🔹 **Microsoft Fabric – Core Overview**

Q1. What is Microsoft Fabric?

Answer:
Microsoft Fabric is an **end-to-end SaaS analytics platform** that unifies **data engineering, data integration, data warehousing, real-time analytics, data science, and Power BI** in one environment.

Q2. Why was Microsoft Fabric introduced?

Answer:
To eliminate data silos by providing:

- One unified platform
- One storage layer (OneLake)
- One security model
- One experience for analytics

Q3. How is Fabric different from traditional Azure services?

Answer:
Traditional Azure uses **multiple services** (ADF, ADLS, Synapse, Power BI separately).
Fabric **combines them into a single platform**.

🔹 **OneLake – The Heart of Fabric (Very Important)**

Q4. What is OneLake?

Answer:
OneLake is the **single, unified data lake** for all Fabric workloads—similar to "OneDrive for data".

Q5. How is OneLake different from ADLS Gen2?

Answer:

ADLS Gen2	OneLake
Standalone storage	Built-in Fabric storage
Manual integration	Automatic integration
Separate security	Unified security

Q6. What are Shortcuts in OneLake?

Answer:
Shortcuts allow Fabric to **access external data (ADLS, S3)** without copying it.

🔹 **Fabric Workloads (Highly Asked)**

Q7. What are the main workloads in Fabric?

Answer:

1. Data Engineering
2. Data Factory
3. Data Warehouse
4. Data Science
5. Real-Time Analytics
6. Power BI

🔹 **Data Engineering in Fabric**

Q8. What is Data Engineering in Fabric?

Answer:
It enables building **Spark-based data pipelines** using notebooks and lakehouses.

Q9. What is a Lakehouse?

Answer:
A lakehouse combines:

- Data Lake storage
- Data Warehouse analytics

Supports both files and tables.

Q10. What languages are used in Fabric notebooks?

Answer:

- PySpark
- SQL
- Scala

🔹 **Data Factory in Fabric**

Q11. What is Data Factory in Fabric?

Answer:
It is the **modern version of ADF**, used for **data ingestion and orchestration** inside Fabric.

Q12. Difference between ADF and Fabric Data Factory?

Answer:
Fabric Data Factory is **fully integrated** with OneLake and Power BI.

🔹 **Data Warehouse in Fabric**

Q13. What is Fabric Data Warehouse?

Answer:
A cloud-native SQL-based analytics engine optimized for reporting and BI.

Q14. Difference between Warehouse and Lakehouse?

Answer:

Warehouse	Lakehouse
Structured data	Structured + raw
SQL only	SQL + Spark
BI focused	Engineering + BI

🔹 **Real-Time Analytics**

Q15. What is Real-Time Analytics in Fabric?

Answer:
It processes **streaming data** like IoT, logs, and events in near real time.

Q16. Use cases of Real-Time Analytics?

Answer:

- IoT monitoring
- Telemetry data
- Fraud detection
- Network monitoring

🔹 **Power BI in Fabric**

Q17. How is Power BI integrated with Fabric?

Answer:
Power BI is **natively embedded**—no separate publishing or refresh required.

Q18. What is Direct Lake mode?

Answer:
Direct Lake allows Power BI to **query OneLake data directly** without import or DirectQuery.

Q19. Benefits of Direct Lake?

Answer:

- Faster performance
- No data duplication
- Near real-time analytics

🔹 **Security & Governance**

Q20. How is security managed in Fabric?

Answer:

- Entra ID (Azure AD)
- Workspace-level access
- Unified permissions
- Row-Level Security (RLS)

Q21. What is data governance in Fabric?

Answer:
Ensures data quality, lineage, compliance, and access control.

🔹 **Performance & Optimization**

Q22. How do you optimize Fabric performance?

Answer:

- Use Parquet/Delta formats
- Partition data
- Use Direct Lake
- Avoid data duplication

🔹 **Fabric vs Azure Stack (Interview Favorite)**

Q23. Fabric vs ADF + ADLS + Synapse + Power BI?

Answer:
Fabric replaces the **entire analytics stack** with a single SaaS platform.

🔹 **Real-World Scenario (Very Important)**

Q24. Design an end-to-end Fabric solution.

Answer:
Source → Fabric Data Factory → OneLake → Lakehouse → Power BI (Direct Lake)

Q25. How does Fabric help Automotive / Telecom / IoT?

Answer:
It enables real-time analytics, scalable data pipelines, and fast BI insights.

🔹 **Interview One-Liners (Memorize)**

- “Fabric is an all-in-one analytics platform.”
- “OneLake is the single source of truth.”
- “Direct Lake eliminates data duplication.”
- “Fabric simplifies the modern data stack.”

🔹 **Common Fresher Mistakes**

❌ Thinking Fabric is only Power BI
❌ Confusing OneLake with ADLS
❌ Ignoring Lakehouse concept
❌ Not understanding Direct Lake

Below is a **clear, interview-ready deep dive on Data Agents**, explained in a **fresher-friendly + enterprise-oriented way**, aligned with **Microsoft Fabric, Power BI, Azure ecosystem**, and **modern data platforms**.

> ⚠️ Note: "Data Agents" is often used as a **conceptual role/component**, not a single Azure service. Interviewers test **understanding**, not tool names.

🔹 What Are Data Agents? (Core Understanding)

Q1. What is a Data Agent?

Answer:
A Data Agent is a **software component or service** that **collects, monitors, moves, transforms, or manages data** between systems automatically.

> Simple definition:
> **"A data agent acts as a bridge between data sources and data platforms."**

Q2. Why are Data Agents needed?

Answer:
Because modern systems generate:

- Huge volumes of data
- Real-time data
- Distributed data

Manual handling is impossible, so **agents automate data operations**.

🔹 Types of Data Agents (Very Important)

Q3. What are the types of Data Agents?

Answer:

1. **Ingestion Agents**
2. **Monitoring Agents**
3. **Transformation Agents**
4. **Security & Governance Agents**
5. **AI / Intelligent Data Agents**

🔹 1️⃣ Data Ingestion Agents

Q4. What is a Data Ingestion Agent?

Answer:
An ingestion agent **collects data from sources** and sends it to a data platform like Data Lake or OneLake.

Q5. Examples of ingestion agents?

Answer:

- ADF Copy Activity
- Fabric Data Factory pipelines
- IoT Edge agents
- Log collectors

Q6. Real-world example?

Answer:
Vehicle sensors send data → Agent collects → Stores in Data Lake.

🔹 2️⃣ **Data Monitoring Agents**

Q7. What is a Monitoring Agent?

Answer:
It tracks:

- Data freshness
- Pipeline success/failure
- Performance issues

Q8. Why monitoring agents are critical?

Answer:
Because **data failures directly impact business decisions**.

Q9. Example?

Answer:
An agent alerts if yesterday's sales data did not load.

🔹 3️⃣ **Data Transformation Agents**

Q10. What is a Transformation Agent?

Answer:
Transforms raw data into **clean, structured, analytics-ready data**.

Q11. Where do we see this in Azure/Fabric?

Answer:

- Mapping Data Flows (ADF)
- Spark jobs in Fabric
- Lakehouse transformations

Q12. Example?

Answer:
Raw CSV → Cleaned Parquet → Aggregated sales table.

🔹 4️⃣ **Security & Governance Data Agents**

Q13. What are Security Data Agents?

Answer:
They enforce:

- Access control
- Masking
- Compliance rules

Q14. Why are they important?

Answer:
To protect:

- Customer data
- Location data
- Financial data

Q15. Example?

Answer:
Agent restricts HR salary data access based on user role.

🔹 5️⃣ Intelligent / AI Data Agents (Trending Topic)

Q16. What are AI Data Agents?

Answer:
Agents that:

- Understand data context
- Respond to queries
- Automate decisions

Often powered by **AI / Copilot**.

Q17. Example in Microsoft ecosystem?

Answer:

- Fabric Copilot
- Power BI Copilot
- Semantic model agents

Q18. Use case?

Answer:
User asks:

> "Show last quarter revenue by region"
> Agent fetches + analyzes + visualizes data.

🔹 Data Agents in Microsoft Fabric (Interview Favorite)

Q19. How do Data Agents fit into Microsoft Fabric?

Answer:
Fabric uses agents to:

- Ingest data into OneLake
- Monitor pipelines
- Optimize queries

- Enable AI-driven insights

Q20. Data Agent vs Pipeline?

Answer:

Pipeline	Data Agent
Executes workflow	Acts continuously
Scheduled	Event-driven
Static	Intelligent

🔹 **Data Agents vs ETL Tools**

Q21. Data Agent vs ADF?

Answer:
ADF is a **tool**; Data Agent is a **behavior or role**.

Q22. Can ADF act as a Data Agent?

Answer:
✅ Yes.
ADF pipelines **behave like ingestion agents**.

🔹 **Real-World Interview Scenarios**

Q23. How do Data Agents help Automotive / IoT?

Answer:
They collect sensor data, validate it, and route it to analytics platforms.

Q24. How do Data Agents help Telecom?

Answer:
They process call records, monitor network health, and detect anomalies.

Q25. How do Data Agents help Enterprise Analytics?

Answer:
They ensure:

- Reliable data ingestion
- Secure access
- Timely reporting

🔹 **Interview One-Liners (Memorize These)**

- "A Data Agent automates data movement and management."
- "ADF pipelines behave like ingestion agents."
- "Fabric uses intelligent agents for analytics automation."
- "Agents reduce manual intervention and errors."

🔹 **Common Fresher Mistakes**

❌ Treating Data Agent as a single Azure service
❌ Ignoring monitoring role
❌ Confusing agent with storage
❌ Missing AI angle

Python – Core Fundamentals

Q1. What is Python and why is it popular?

Answer:
Python is a **high-level, interpreted, object-oriented language** known for its **simple syntax**, large libraries, and strong support for **data engineering, analytics, and automation**.

Q2. Python is interpreted – what does it mean?

Answer:
Code is executed line by line, which makes debugging easier but slightly slower than compiled languages.

Q3. What are Python use cases in data roles?

Answer:

- Data cleaning & transformation

- Automation
- ETL pipelines
- Analytics & visualization
- Machine learning

🔹 **Data Types & Variables (Very Important)**

Q4. Built-in data types in Python?

Answer:
int, float, str, bool, list, tuple, set, dict

Q5. Difference between list, tuple, set?

List	Tuple	Set
Mutable	Immutable	Unordered
Allows duplicates	Allows duplicates	No duplicates

Q6. What is dynamic typing?

Answer:
Python does not require declaring variable types explicitly.

🔹 **Control Flow**

Q7. What are loops in Python?

Answer:
Used to repeat execution: for, while.

Q8. Difference between break, continue, pass?

Answer:

- break → exits loop
- continue → skips iteration

- pass → placeholder

🔹 **Functions & Modules**

Q9. What is a function?

Answer:
A reusable block of code that performs a task.

Q10. What is lambda function?

Answer:
Anonymous one-line function.

lambda x: x*x

Q11. What is a module?

Answer:
A file containing Python code that can be imported.

🔹 **Object-Oriented Programming (OOP)**

Q12. OOP concepts in Python?

Answer:
Encapsulation, Inheritance, Polymorphism, Abstraction

Q13. What is __init__?

Answer:
Constructor method executed when an object is created.

Q14. What is inheritance?

Answer:
Child class acquires properties of parent class.

🔹 **Exception Handling**

Q15. Why is exception handling needed?

Answer:
To handle runtime errors without crashing the program.

Q16. Example?

```
try:
    print(10/0)
except ZeroDivisionError:
    print("Cannot divide by zero")
```

🔹 **File Handling (Very Important for ETL)**

Q17. How do you read/write files?

```
with open("data.txt","r") as f:
    data = f.read()
```

Q18. Difference between read(), readline(), readlines()?

Answer:

- read() → full file
- readline() → one line
- readlines() → list of lines

🔹 **Python for Data Engineering**

Q19. What is NumPy?

Answer:
Library for numerical computing.

Q20. What is pandas?

Answer:
Library for data manipulation and analysis using DataFrames.

Q21. Common pandas operations?

Answer:

- Read CSV
- Filter rows

- Handle missing values
- GroupBy operations

Q22. Example: read CSV & find mean?

```
import pandas as pd

df = pd.read_csv("sales.csv")

df["amount"].mean()
```

🔹 **Python + ETL Scenario Questions**

Q23. How does Python fit into ADF / Fabric?

Answer:
Python scripts are used in:

- Spark notebooks
- Data transformation
- Automation tasks

Q24. Python vs SQL?

Answer:

- SQL → querying structured data
- Python → processing, automation, logic

🔹 **Advanced Python (Interview Bonus)**

Q25. What is a generator?

Answer:
Generates values one at a time using yield.

Q26. What is decorator?

Answer:
Modifies function behavior without changing code.

Q27. What is virtual environment?

Answer:
Isolated Python environment for dependency management.

🔹 **Real-World Scenarios (High Value)**

Q28. How do you clean missing data in Python?

Answer:
Using pandas: dropna(), fillna().

Q29. How do you process large datasets?

Answer:

- Chunking
- Vectorized operations
- Spark

Q30. Python in Automotive / IoT?

Answer:
Processes sensor data, analyzes patterns, triggers alerts.

🔹 **Common Fresher Mistakes**

❌ Not using virtual environments
❌ Writing inefficient loops
❌ Ignoring exception handling
❌ Hard-coding paths