Palantir Foundry Data Engineering Certification - 50 Authentic Practice Questions with Detailed Answers

Section 1: Foundational Concepts & Architecture (10 Questions)

Question 1
What is the PRIMARY purpose of a Resource Identifier (RID) in Foundry?

A) To provide a human-readable name for resources
B) To enable fast searching of datasets
C) To uniquely identify a specific version of a resource
D) To control access permissions

Answer: C
Explanation: RIDs are immutable identifiers that point to exact versions of resources. This ensures reproducibility - a pipeline using a specific RID will always get the same data version, even if the resource is updated later.

Question 2
Which statement BEST describes the relationship between branches and cuts in Foundry?

A) Branches are for development; cuts promote changes between branches
B) Branches are for backup; cuts are for deletion
C) Branches are for testing; cuts are for scheduling
D) Branches are for security; cuts are for versioning

Answer: A
Explanation: Developers work in feature branches. When ready, they cut (promote) changes to staging or main branches. This workflow enables parallel development with controlled releases.

Question 3
You need to store sensitive customer data that should only be accessible to the data engineering team. Where should you place this dataset?

A) `/Global/Customers/`
B) `/Team/Data-Engineering/Customers/`
C) `/Private/Customers/`
D) `/Shared/Customers/`

Answer: B

Explanation: `/Team/Data-Engineering/` restricts access to members of that team. `/Global/` is organization-wide, `/Private/` is individual-only, and `/Shared/` isn't a standard Foundry folder.

Question 4
What happens when you delete a dataset in Foundry?

A) It's permanently erased from storage immediately
B) It's moved to a recycle bin for 30 days
C) It's marked as deleted but all versions remain accessible via RIDs
D) Only the metadata is deleted; data files remain in storage

Answer: C
Explanation: Foundry never truly deletes data. The dataset is marked as deleted in the UI, but all historical versions remain accessible via their RIDs for compliance and reproducibility.

Question 5
Which component is NOT part of Foundry's core architecture?

A) Ontology Service
B) Transform Service
C) Workflow Service
D) Blockchain Service

Answer: D
Explanation: Foundry uses Ontology, Transform, and Workflow services, but does not have a Blockchain service as part of its standard architecture.

Question 6
What is the PRIMARY advantage of using Delta Lake format in Foundry datasets?

A) It enables real-time streaming
B) It provides ACID transactions and schema enforcement
C) It compresses data better than Parquet
D) It allows direct querying from external tools

Answer: B
Explanation: Delta Lake provides ACID transactions, schema enforcement, and time travel capabilities, which are valuable for data engineering workflows.

Question 7
When should you use a Spark dataset versus a SQL dataset?

A) Spark datasets for all use cases
B) SQL datasets for large-scale processing, Spark for small datasets
C) Spark datasets for large-scale distributed processing, SQL for smaller transforms
D) SQL datasets for real-time streaming

Answer: C
Explanation: Spark datasets handle terabyte-scale distributed processing. SQL datasets are better for smaller, SQL-based transformations that don't require distributed computing.

Question 8
What does "schema-on-read" mean in Foundry?

A) Schema must be defined before writing data
B) Schema is inferred when data is read
C) Schema is automatically optimized for reading speed
D) Schema cannot be changed after data is written

Answer: B
Explanation: Schema-on-read allows flexible ingestion where Foundry infers the schema when data is read, useful for exploratory work. Schema-on-write requires upfront definition.

Question 9
Which statement about Foundry's versioning is CORRECT?

A) Only the latest version of a dataset is stored
B) Every change creates a new version with a new RID
C) Versioning is optional and must be enabled
D) Versions are deleted after 90 days

Answer: B
Explanation: Every change (write, update, delete) creates a new immutable version with a unique RID. All versions are preserved indefinitely.

Question 10
What is the purpose of the "Lineage" feature in Foundry?

A) To track data movement between systems
B) To visualize data flow and dependencies between resources
C) To monitor real-time data streaming
D) To optimize query performance

Answer: B
Explanation: Lineage shows how data flows through transforms, workflows, and datasets, helping understand dependencies and impact analysis.

Section 2: Data Ingestion & Integration (10 Questions)

Question 11
You need to ingest CSV files from an S3 bucket daily at 2 AM UTC. The files have inconsistent headers. Which approach is BEST?

A) Use a Loader with schema inference enabled
B) Use Contour to manually upload each file
C) Use a Foundry Function triggered by S3 events
D) Write a custom Python script with hardcoded schema

Answer: A
Explanation: Loaders support scheduled ingestion and can handle schema inference. For inconsistent headers, schema inference is better than hardcoded schemas.

Question 12
What is the MAIN difference between Connectors and Loaders?

A) Connectors move data; Loaders transform data
B) Connectors define connections; Loaders define ingestion jobs
C) Connectors are for batch; Loaders are for streaming
D) Connectors are free; Loaders require licensing

Answer: B
Explanation: Connectors configure connections to source systems (credentials, endpoints). Loaders define what to ingest, when, and where.

Question 13
When is Contour the MOST appropriate ingestion tool?

A) For scheduled nightly imports from production databases
B) For one-time upload of historical data for analysis
C) For real-time streaming from IoT devices
D) For automated ingestion from external APIs

Answer: B
Explanation: Contour is designed for ad-hoc, manual uploads like historical data. Scheduled, automated ingestion should use Loaders.

Question 14
You're setting up a Loader for Salesforce data. The source has 1 million records updated daily. How should you configure it?

A) Full refresh with overwrite mode daily
B) Incremental load based on LastModifiedDate
C) Load only new records manually each day
D) Schedule hourly loads to reduce batch size

Answer: B
Explanation: For large datasets with daily updates, incremental loading based on modification timestamp is most efficient.

Question 15

Which file format is MOST efficient for analytical queries in Foundry?

A) CSV
B) JSON
C) Parquet
D) XML

Answer: C
Explanation: Parquet is columnar, compressed, and optimized for analytical queries in systems like Spark.

Question 16
You need to ingest real-time data from a Kafka topic. Which approach is BEST?

A) Use a Loader with 1-minute schedule
B) Use the Kafka connector with streaming enabled
C) Use Contour with auto-refresh
D) Write a custom Spark Streaming job

Answer: B
Explanation: Foundry's Kafka connector supports streaming ingestion directly into datasets.

Question 17
What is the purpose of the "Bronze-Silver-Gold" pattern?

A) To categorize data by importance
B) To implement a data quality framework
C) To create a data lake architecture with raw, cleaned, and business-ready layers
D) To classify data by retention period

Answer: C
Explanation: This pattern creates layers: Bronze (raw), Silver (cleaned), Gold (business-ready aggregated data).

Question 18
You're ingesting sensitive data. Which feature should you enable?

A) Automatic partitioning
B) Schema inference
C) Data encryption
D) Compression

Answer: C
Explanation: Foundry provides encryption at rest and in transit, which should be enabled for sensitive data.

Question 19

A Loader job fails due to network issues. What happens by default?

A) The job stops and requires manual restart
B) The job retries 3 times before failing
C) The job continues with available data
D) All dependent workflows are cancelled

Answer: B
Explanation: Loaders have configurable retry logic (default is usually 3 retries) for transient failures.

Question 20
Which statement about incremental ingestion is TRUE?

A) It requires a watermark column in the source
B) It always performs better than full refresh
C) It can miss updates if not configured properly
D) It's not supported in Foundry

Answer: C
Explanation: Incremental ingestion requires careful configuration of change detection columns; improper setup can miss updates or cause duplicates.

Section 3: Transformation & Processing (10 Questions)

Question 21
Your PySpark job is failing with "OutOfMemoryError." What should you check FIRST?

A) Increase spark.driver.memory in configuration
B) Check for data skew in Spark UI
C) Reduce the number of output partitions
D) Switch to SQL transforms

Answer: B
Explanation: Data skew (uneven data distribution) is the most common cause of OOM errors. Always diagnose in Spark UI before applying fixes.

Question 22
When should you use broadcast join in PySpark?

A) When both tables are larger than 1 GB
B) When joining a large table with a small table (< 1 GB)
C) For all join operations
D) Only for streaming joins

Answer: B

Explanation: Broadcast join is efficient when one table is small enough to be sent to all executors.

Question 23
What is the PRIMARY benefit of using SQL transforms over PySpark transforms?

A) Better performance for all operations
B) Easier for business users to understand and maintain
C) More advanced machine learning capabilities
D) Real-time processing support

Answer: B
Explanation: SQL transforms are more accessible to users familiar with SQL and often sufficient for business logic transformations.

Question 24
You need to process 10 TB of data daily. Which compute option is MOST appropriate?

A) Foundry Functions
B) SQL transforms
C) PySpark transforms with auto-scaling
D) Prepare transforms

Answer: C
Explanation: PySpark with auto-scaling handles large-scale distributed processing efficiently.

Question 25
What does the `@transform` decorator do in Foundry PySpark code?

A) Marks a function as a data transformation
B) Optimizes Spark execution plan
C) Enables real-time processing
D) Creates a workflow automatically

Answer: A
Explanation: The `@transform` decorator identifies functions that should be executed as Foundry transforms with input/output tracking.

Question 26
You notice a job is creating thousands of small output files. How should you fix this?

A) Increase the number of partitions
B) Use `coalesce()` or `repartition()` before writing
C) Switch to CSV format
D) Enable compression

Answer: B

Explanation: `coalesce()` or `repartition()` controls the number of output files, preventing the "small files problem."

Question 27
When should you use the "Prepare" tool instead of code transforms?

A) For complex machine learning pipelines
B) For simple data cleaning and exploration
C) For real-time stream processing
D) For production ETL with strict SLAs

Answer: B
Explanation: Prepare is a no-code UI tool ideal for simple transformations, data profiling, and exploratory work.

Question 28
What is the purpose of partitioning in Foundry datasets?

A) To improve query performance through data pruning
B) To encrypt sensitive data
C) To enable real-time updates
D) To reduce storage costs

Answer: A
Explanation: Partitioning organizes data so queries can skip irrelevant partitions, improving performance.

Question 29
Which data quality check is MOST important for a customer ID column?

A) Value range validation
B) Uniqueness constraint
C) Pattern matching
D) Null check

Answer: B
Explanation: Customer IDs should typically be unique; duplicates could cause serious data integrity issues.

Question 30
You need to join two datasets where 80% of records match on a common key. Which join type should you use?

A) Inner join
B) Left outer join
C) Full outer join
D) Cross join

Answer: A
Explanation: Inner join returns only matching records, appropriate when you want only complete matches.

Section 4: Orchestration & Workflows (10 Questions)

Question 31
Job B depends on Job A. Job A fails. What happens to Job B?

A) Job B runs with whatever data is available
B) Job B is automatically skipped
C) Job B fails immediately
D) The entire workflow stops

Answer: B
Explanation: In Foundry workflows, dependent jobs are skipped (not failed) when their dependencies fail.

Question 32
You need to run a pipeline at 2 AM daily, but only on weekdays. Which cron expression is correct?

A) `0 2 * * 1-5`
B) `0 2 * * 0-4`
C) `0 2 * * 6,7`
D) `0 2 * * *`

Answer: A
Explanation: Cron format: minute hour day month day-of-week. `1-5` = Monday-Friday (1=Monday, 7=Sunday).

Question 33
What is the MAIN advantage of using Workflows over manually running transforms?

A) Better performance
B) Automated scheduling and dependency management
C) Lower cost
D) Real-time processing

Answer: B
Explanation: Workflows automate execution, handle dependencies, retries, and scheduling.

Question 34
A workflow job times out after 1 hour. How should you address this?

A) Increase the timeout setting
B) Reduce the data volume
C) Optimize the transform code first, then adjust timeout if needed
D) Split the job into smaller jobs

Answer: C
Explanation: Always optimize code first. Increasing timeout without optimization just delays the problem.

Question 35
You need to process data for each region in parallel. Which workflow pattern should you use?

A) Sequential execution
B) Fan-out, fan-in
C) Conditional branching
D) Looping

Answer: B
Explanation: Fan-out processes regions in parallel; fan-in combines results.

Question 36
When should you use "Conditional Execution" in a workflow?

A) To run jobs based on data quality results
B) To improve job performance
C) To reduce costs
D) To enable real-time processing

Answer: A
Explanation: Conditional execution runs jobs based on conditions like data quality checks passing.

Question 37
What is the purpose of workflow notifications?

A) To optimize job performance
B) To alert on job success/failure
C) To track data lineage
D) To manage compute resources

Answer: B
Explanation: Notifications alert teams about workflow status via email, Slack, etc.

Question 38
A workflow has 5 jobs running sequentially. Job 3 fails. What happens to jobs 4 and 5?

A) They run normally

B) They are skipped
C) They fail immediately
D) They wait for manual intervention

Answer: B
Explanation: In a sequential workflow, when one job fails, subsequent dependent jobs are skipped.

Question 39
You need to trigger a workflow when a dataset is updated. Which trigger type should you use?

A) Cron schedule
B) Dataset update trigger
C) Manual trigger only
D) API trigger

Answer: B
Explanation: Foundry supports triggering workflows when datasets are updated.

Question 40
What is the MAXIMUM number of retries recommended for a failing job?

A) 1
B) 3
C) 10
D) Unlimited

Answer: B
Explanation: 3 retries is generally sufficient. More retries may indicate underlying issues needing investigation.

Section 5: Ontology & Data Modeling (10 Questions)

Question 41
What is the PRIMARY purpose of the Ontology in Foundry?

A) To store raw data
B) To define business concepts and relationships
C) To schedule data pipelines
D) To monitor system performance

Answer: B
Explanation: The Ontology creates a unified model of business concepts (Customers, Products, etc.) and their relationships.

Question 42
You have a dataset with customer data. What must you do to query it as `Customer` objects?

A) Create a SQL view
B) Apply links to the Customer Object Type
C) Convert it to Parquet format
D) Move it to the Global folder

Answer: B
Explanation: Applying links connects dataset columns to Ontology Object Type properties, enabling semantic queries.

Question 43
What does "cardinality" specify in an Ontology relationship?

A) Data type of the property
B) Whether the property is required
C) Number of related objects (one-to-one, one-to-many)
D) Encryption method for the property

Answer: C
Explanation: Cardinality defines relationship type: one-to-one, one-to-many, or many-to-many.

Question 44
How do you query all Orders for a specific Customer using Ontology syntax?

A) `SELECT * FROM Order WHERE customer_id = '123'`
B) `SELECT customer.orders FROM Customer WHERE customer.id = '123'`
C) `SELECT * FROM Customer JOIN Order ON Customer.id = Order.customer_id`
D) `SELECT Order FROM Customer WHERE id = '123'`

Answer: B
Explanation: Ontology queries use navigation syntax: `customer.orders` traverses the relationship.

Question 45
What happens during Ontology Sync?

A) Data is cleaned and standardized
B) The ontology graph is materialized into queryable datasets
C) Object Types are backed up
D) Links are validated for consistency

Answer: B
Explanation: Sync materializes the ontology graph into SQL-queryable datasets.

Question 46
Which property type should you use for a customer's email address?

A) Integer
B) String with email format validation
C) Boolean
D) Link

Answer: B
Explanation: Email is a string with specific format; Foundry supports format validation.

Question 47
You need to model "Employee works in Department." How should you represent this?

A) Add department_id string to Employee
B) Add works_in link from Employee to Department
C) Add employees link from Department to Employee
D) Both B and C are correct

Answer: D
Explanation: You can model from either direction. Typically, Employee has `works_in` link to Department (cardinality: one), and Department has `employees` link to Employee (cardinality: many).

Question 48
What is the difference between an Object Type and an Entity?

A) Object Type is the blueprint; Entity is an instance
B) Object Type is for storage; Entity is for processing
C) Object Type is immutable; Entity can change
D) There is no difference

Answer: A
Explanation: Object Type defines structure (like a class); Entity is an instance (like an object).

Question 49
Why would you create a Property as a "link" type?

A) To store URL addresses
B) To connect to external websites
C) To create relationships between Object Types
D) To enable data encryption

Answer: C
Explanation: Link properties create relationships between Object Types in the ontology graph.

Question 50
What is required for an Object Type to be queryable?

A) It must have at least one property
B) It must have a primary key defined
C) It must be linked to at least one dataset
D) It must be in the Global ontology

Answer: C
Explanation: Object Types need linked datasets with data to be queryable. Empty Object Types return no results.

Answer Key Summary:

1. C 2. A 3. B 4. C 5. D 6. B 7. C 8. B 9. B 10. B
11. A 12. B 13. B 14. B 15. C 16. B 17. C 18. C 19. B 20. C
21. B 22. B 23. B 24. C 25. A 26. B 27. B 28. A 29. B 30. A
31. B 32. A 33. B 34. C 35. B 36. A 37. B 38. B 39. B 40. B
41. B 42. B 43. C 44. B 45. B 46. B 47. D 48. A 49. C 50. C