



Data Engineering Fundamentals Testing

Which tool uses dataframe abstraction?

PySpark ☒

Airflow

Hive

Hadoop

Why is distributed computing not useful for small tasks?

Communications limit the execution speed. ☒

Storage requirements become too large.

Small tasks are difficult to partition.

- This is because the overhead of managing communication and coordination between nodes in a distributed system can outweigh the benefits of parallelization for small tasks.
 - The time taken to distribute the task, process it, and then collect the results can be longer than simply processing the task on a single machine.
-

Which family of tools addresses the storage problem?

schedulers

processing frameworks

databases ☒

How do flat or plain files often accommodate structures such as arrays?

by using OLAP


by using an API

by using OLTP

by using JSON ☒

- Flat or plain files accommodate structures such as arrays by using JSON.
 - JSON, which stands for JavaScript Object Notation, is a lightweight data-interchange format that is easy for humans to read and write and easy for machines to parse and generate.
 - It's commonly used for representing structured data in plain text format,
 - Making it an excellent choice for flat or plain files that need to store complex data structures like arrays, objects, and nested information.
 - JSON's format is both flexible and hierarchical, allowing for the effective representation of data structures within a flat file context without the need for a database or specialized data management system.
-


What is the purpose of Dask?

to eliminate redundant nodes
to simulate multiprocessor systems
to test each partition separately
to simplify coding for distributed processing 


- Dask is an open-source library for parallel computing in Python.
 - It enables efficient parallel computations on single computers or across clusters of computers by breaking down complex tasks into smaller, manageable tasks that can be executed in parallel.
 - This simplification helps in handling large datasets that don't fit into memory, by using efficient scheduling and by leveraging multiple processors or distributed systems.
 - Unlike tools that are primarily focused on streamlining specific aspects of distributed systems, such as eliminating redundant nodes or simulating multiprocessor systems,
 - Dask provides a general framework that extends Python's concurrent programming tools and libraries like NumPy, pandas, and scikit-learn for scalable, distributed computing.
-

Which skills are most helpful for becoming a data engineer?

mathematics and statistics
hardware connections and networking

database management and acceleration
coding and formatting 

If Chloe wants to migrate to data engineering from data science, what should she study?

storage systems 

machine learning

business metrics

pattern recognition


- Data engineering primarily focuses on the architecture, design, and management of the data infrastructure that allows for efficient data collection, storage, and retrieval.
 - Understanding storage systems—including databases (both SQL and NoSQL), data lakes, and data warehouses—is crucial for designing scalable and reliable data pipelines.
 - This knowledge enables data engineers to ensure that data is accessible, secure, and organized in a way that supports efficient analysis and processing.
 - While knowledge in areas like machine learning, business metrics, and pattern recognition is valuable for data scientists, a deep understanding of storage systems is fundamental for transitioning into a data engineering role.
-

Which choice is an RDD transformation function?

collect()

count()

first()

filter() 

- An RDD (Resilient Distributed Dataset) transformation function in the context of Apache Spark is "filter()".
- Transformations are operations on RDDs that return a new RDD, such as mapping data or filtering it based on a condition, but they do not return a single value or result immediately.

- Instead, transformations are lazy and only compute their results when an action (like `collect()`, `count()`, or `first()`) is called. The `"filter()"` function is used to create a new RDD by selecting elements from the current one that meet a specified condition.

Manish is starting a new database for a small company. Why should he invest effort in the schema design?

to allow multiple users


to have efficiency of storage and operation 

to implement better encryption and security

to permit machine-learning applications

- Proper schema design is crucial for databases because it ensures that data is stored in a structured and efficient manner, which can significantly impact the performance of database operations, including data retrieval, insertion, updating, and deletion.
- A well-designed schema helps in optimizing storage space, improving query performance by reducing the amount of data scanned, and ensuring data integrity and consistency.
- While allowing multiple users, implementing better encryption and security, and permitting machine-learning applications are important considerations for a database, the primary reason for investing in schema design is to ensure efficient storage and operation.

Why is organization helpful in a database?

for speed of operation 

for reduced storage space

for simplicity of construction

- A well-organized database can dramatically improve the speed at which data can be accessed, queried, and analyzed.
- Proper organization involves structuring data in a way that aligns with how it will be accessed and used, which can include indexing, partitioning, and designing efficient schema.
- These strategies help in reducing the time it takes to perform operations on the data, such as searches, updates, and retrieval, by minimizing the

amount of data that needs to be scanned or processed for each query.

- While reduced storage space and simplicity of construction can be benefits of good database design, the primary advantage of organization is to enhance the speed and efficiency of database operations.