Walmart Data Engineer Interview Guide – Experienced

Round 1: Preliminary Round (Screening Round) – Telephonic Interview (45 minutes)

In this initial phase, the focus was on discussing my prior experiences, particularly related to data engineering tools and platforms I've worked with. I was also asked to elaborate on some of the core data concepts and my work with them.

Key Discussion Points:

- **Overview of Previous Projects**: I discussed my involvement with tools like Mixpanel, Kafka, ETL processes, Datahub, Spark, and Presto architecture.
- **Data Modeling**: Detailed insights on how I created a data model during experimentation and A/B testing.
- **Why Walmart?** I explained my motivation for applying to Walmart, citing their global presence, innovative data practices, and impactful role in the retail industry.

Round 2: Technical Interview 1 (Coding & DSA Round)

Overview:

The second round of the interview focused primarily on assessing core technical skills relevant to data engineering, particularly in coding, data structures, algorithms, and large-scale data processing frameworks. This round lasted for about 1 hour and 30 minutes and was conducted by a senior data engineer. The discussion covered various domains, including coding proficiency, SQL expertise, big data technologies, cloud platforms, and key engineering practices.

Topics Covered:

1. **Data Structures and Algorithms (DSA)**:
   - Medium-level data structures, including arrays, stacks, linked lists, and trees.
   - Problem-solving related to fundamental algorithms.
2. **SQL**:
   - Complex SQL queries, particularly using advanced features such as window functions.
   - Writing efficient queries for scenarios involving large datasets.
3. **Big Data Concepts**:
   - Understanding of distributed processing systems, particularly related to Apache Spark and Hadoop.
   - Architectural and operational aspects of these big data tools.

4. **Cloud Computing**:
   - Scenarios related to cloud platforms such as AWS.
   - Cloud service management and optimization techniques.
5. **Python Coding**:
   - Writing Python code for automation and data engineering tasks.
   - Coding challenges with a focus on algorithmic problem-solving.
6. **DevOps & SDLC**:
   - Basic knowledge of DevOps practices, continuous integration/deployment (CI/CD) pipelines.
   - Understanding of the software development lifecycle (SDLC) and agile methodology.

Detailed Breakdown of Interview Questions:

1. **Data Structures and Algorithm Questions:**
   - **Coin Change Problem**: A typical dynamic programming question asking for the minimum number of coins required to make a change for a given amount. This problem tests the understanding of optimal substructure and overlapping subproblems, which are central to dynamic programming.
   - **Partitioning a Linked List**: This problem requires partitioning a linked list based on a value, ensuring that all nodes with values less than a given value x appear before those with values greater than or equal to x. It evaluates knowledge of linked list manipulation and partitioning techniques.
2. **SQL Questions**:
   - **Finding nth Highest Salary**: Given a table of employees and departments, the task is to find the nth highest salary within each department. This is typically solved using SQL window functions, such as DENSE_RANK(), which assigns a rank to each row within a partition of the dataset. The candidate is also expected to explain why they chose DENSE_RANK() over RANK()—with the former allowing ties to be handled without skipping ranks.
   - **SQL Query Design**: The task involved writing SQL queries to identify employees with the highest salaries within each department. This challenges both logical thinking and SQL proficiency, as the solution requires correctly implementing window functions or alternative approaches.
3. **Big Data and Spark Optimization**:
   - **Airflow and Kubernetes**: Questions related to how Airflow operates in a Kubernetes environment, including the use of Pods and the interaction between the Airflow scheduler, web server, and worker machines. Understanding containerization and Kubernetes' role in managing distributed workloads is essential for big data engineers working in the cloud.
   - **Optimizing Spark Jobs**: When Spark jobs take longer than expected, performance bottlenecks are common. The candidate is expected to suggest methods for identifying these issues, such as analyzing Spark UI logs, checking for skewed data, and tuning configurations for resource allocation.

    - **Cluster Resource Allocation**: Given limited resources in a Spark cluster, the interview explores how to optimize resource distribution. This could include strategies for configuring executors, adjusting memory allocations, and managing job priorities
    - **Spark Job Code**: A coding task was presented, where the candidate was asked to write a code snippet to upload Parquet files to an S3 bucket using the boto3 Python library. This test evaluates knowledge of both Python and cloud-specific APIs for interacting with data storage.

4. **Airflow and Logging**:
    - **Airflow Log Storage**: The interviewer asked about how Airflow stores logs and the role of its backend database. This tests an understanding of Airflow's architecture and how it manages operational logs in a distributed environment.

5. **Cloud Computing and AWS**:
    - **AWS-Based Scenarios**: Scenarios revolving around AWS services were discussed, focusing on real-world applications and how cloud platforms can be leveraged to solve big data problems efficiently. This includes understanding various AWS services like S3, EC2, Lambda, and their interaction with big data tools like Spark.

Key Insights for Candidates:

- **Coding and Algorithms**: Proficiency in algorithms and data structures is essential for problem-solving in coding interviews. Candidates should practice common problems related to dynamic programming, linked lists, trees, and sorting algorithms.

- **SQL and Database Optimization**: SQL queries should not only be accurate but also optimized for performance, especially when working with large datasets. Candidates should be familiar with advanced SQL features like window functions and be prepared to explain their choice of techniques.

- **Big Data Tools and Cloud Platforms**: Knowledge of tools like Spark, Hadoop, Kubernetes, and Airflow is crucial. Understanding how these tools interact in distributed systems and cloud environments is key to succeeding in interviews for data engineering roles.

- **Practical Application**: It's important to be able to write code under pressure, especially for cloud and big data tasks like working with AWS and performing operations in Spark. Prepare for hands-on coding exercises that test both theoretical understanding and practical implementation.

Round 3: Technical Interview 2 (Data Modeling/System Design with Big Data Concepts)

Overview:

The third technical round, which lasted approximately 1 hour and 45 minutes, was focused on data modeling, system design, and big data concepts. The interview was conducted by a Staff Data Engineer from Walmart. This round required the candidate to demonstrate in-depth knowledge in system architecture, big data tools, Java, and advanced data engineering concepts, with a focus on both theoretical understanding and practical coding abilities.

Topics Covered:

1. **System Design**:
   - Event-driven architectures and large-scale system design.
   - Specific focus on designing systems like Mixpanel.
   - Detailed exploration of load balancing, request handling, and system components.
2. **Big Data & Spark**:
   - Coding tasks with Spark, focusing on data ingestion and transformation using Delta Lake.
   - Optimizations for Spark jobs, including skewed joins, broadcast joins, and Spark's Catalyst Optimizer.
3. **Java & Advanced Java Concepts**:
   - Questions on multithreading, synchronization, garbage collection, and serialization.
   - Deep dive into Java collections, including interfaces, maps, and linked lists.
4. **ETL and Data Warehousing**:
   - Understanding of data warehouse concepts, schema design, and ETL best practices.
   - Key topics included Snowflake vs. Star schema, normalization, and Slowly Changing Dimensions (SCD).

Detailed Breakdown of Interview Questions:

1. System Design and Event-Driven Architecture:

- **Designing Mixpanel**: The candidate was tasked with designing the Mixpanel system, an event-driven analytics platform. The candidate leveraged tools like draw.io to illustrate how events are captured from different sources (Android, Web, and iOS apps). This exercise assessed the candidate's understanding of event-driven architectures and the flow of data between various components in a distributed system.
- **Request Handling in Distributed Systems**: The candidate was asked to explain how a request would travel from a client (like opening a Presto URL in a browser) through various system layers, including DNS resolution, load balancing, and routing through the Presto coordinator. This question tested the candidate's understanding of networking and how requests are handled in complex, distributed systems.
- **Custom API with Spring Boot**: A hands-on coding exercise where the candidate was asked to write a simple service and controller class in Spring Boot, simulating the development of a REST API. This tested the candidate's knowledge of Java, Spring Boot, and their ability to implement API logic effectively.

2. Spark Optimization and Big Data Concepts:

- **Upsert in Delta Lake**: The candidate was asked to write code to read data from a Delta Lake stored in an S3 bucket and perform an upsert (update existing records and insert new ones) based on a primary key. This task involved using Spark DataFrames, emphasizing knowledge of data ingestion, transformation, and Delta Lake's capabilities.
- **Spark Optimizations**: The interview covered various Spark optimization techniques, such as handling skewed joins, using broadcast joins for small tables, leveraging the Catalyst Optimizer for query optimization, and understanding the differences between repartition() and coalesce() for managing data partitions in Spark.
- **Spark Tungsten & Catalyst Optimizer**: The candidate was asked about Spark's Tungsten and Catalyst components, which are critical for optimizing query execution. Tungsten manages memory and execution for Spark, while Catalyst performs query optimization. The candidate needed to explain how these technologies improve performance in distributed big data processing.

3. Java & Advanced Java Concepts:

- **Garbage Collection in Java**: The candidate was asked to write code to manually invoke the garbage collection process using the Java GC thread. This tested knowledge of Java memory management and understanding of how garbage collection is handled in the JVM.

- **Multithreading and Synchronization**: The interviewer probed deeper into Java multithreading concepts, asking the candidate to explain synchronization and write code to manage synchronized threads. This task required understanding thread safety and how to use synchronization mechanisms (like synchronized keyword) to avoid concurrency issues such as race conditions.
- **Serialization vs Deserialization**: The candidate was questioned on the differences between serialization (converting objects to a byte stream) and deserialization (converting byte streams back to objects), and how they are used in distributed systems for transmitting objects over networks.
- **The transient Keyword in Java**: The candidate was asked to explain the use of the transient keyword, which marks fields to be excluded from serialization. This concept is essential for optimizing data transmission and storage in distributed systems.

4. System Design and Synchronization:

- **Semaphore in Java**: The candidate was asked to explain and implement the concept of a semaphore in Java, which is used for controlling access to shared resources in concurrent programming. They were tasked with completing code for a semaphore, managing processes and ensuring synchronization to avoid deadlocks.
- **Deadlock Prevention**: The interviewer tested the candidate's understanding of deadlock prevention techniques, asking how deadlocks occur in multithreaded systems and how to prevent them, specifically using semaphores and other synchronization mechanisms.

5. ETL & Data Warehousing:

- **Snowflake vs. Star Schema**: The candidate was asked to explain the difference between the Snowflake and Star schema designs in data warehousing. The Star schema involves a central fact table connected to dimension tables, while the Snowflake schema normalizes these dimension tables into multiple related tables.
- **Data Warehouse Design**: The interview shifted towards designing a data warehouse from scratch, with the candidate expected to consider new requirements and structure a solution that fits the needs of a modern data architecture.
- **Normalization & Slowly Changing Dimensions (SCD)**: The candidate was asked to explain normalization and how to manage historical data in data warehouses using Slowly Changing Dimensions (SCD) Type 2. SCD Type 2 tracks changes over time by creating multiple records for a dimension, maintaining historical accuracy.
- **Onboarding Delta Lake Catalog to Presto**: The candidate was asked how to onboard a Delta Lake catalog to Presto, highlighting the need for integration between big data tools for efficient querying and analytics.

- **Agile vs. Waterfall**: The candidate was asked to explain why Agile is preferred over the traditional Waterfall model. This question focused on the iterative, flexible nature of Agile, which is well-suited for data engineering projects where requirements can evolve over time. The candidate detailed the Scrum framework, sprints, Jira boards, and the benefits of incremental progress.

Key Insights for Candidates:

- **System Design & Architecture**: A solid understanding of system design principles, particularly in the context of event-driven architectures and distributed systems, is essential. Familiarity with tools like Spring Boot and techniques for handling large-scale data processing in Spark is critical.
- **Big Data & Spark Optimizations**: Proficiency in Spark, including its optimization techniques (e.g., skewed joins, broadcast joins, and the Catalyst optimizer), is crucial for tackling performance issues in big data workflows.
- **Java & Concurrency**: A deep understanding of Java, especially regarding multithreading, synchronization, garbage collection, and serialization, is essential for solving concurrency-related problems and optimizing memory management.
- **ETL & Data Warehousing**: Knowledge of data modeling concepts like Snowflake and Star schemas, normalization, and Slowly Changing Dimensions (SCD) is key for building scalable and efficient data warehouses.
- **Agile Methodology**: Understanding of Agile principles, particularly Scrum, is necessary for managing projects in a fast-paced, iterative environment. Familiarity with tools like Jira and understanding Agile's flexibility in adapting to changes is critical for success in modern engineering teams.

Round 4: Techno-Managerial Interview (Managerial Round): 1 hour 10 minutes

1. Introduction & Skillset Overview:

- Introduction of expertise and technical skillset.

2. Data Modeling & ETL Design:

- Questions on Data Modeling, Databricks, Datahub, PySpark, and architecture design (ETL Design).
- Explanation of the Mixpanel project and data model creation using delta tables.
- Detailed explanation of the data pipeline on Databricks for creating aggregated tables based on business requirements.

3. Contributions to Open-Source Projects:

- Discussion on contributions to open-source projects, including Datahub and Spark Lineage.
- Explanation of Spark jar creation with Spark listeners and the Spline package.

4. Cost Optimization:

- Questions on cost optimization in cloud technologies:
    - Can you share an example of a project you worked on that had a significant impact on your organization?
    - How did you contribute to cost optimization initiatives while working with cloud technologies?
    - Could you describe a specific cost optimization strategy you implemented in the cloud and its results?

5. Databricks & Spark Monitoring:

- Questions on capturing event logs and user activities on Databricks, including cluster creation and job execution.
- Questions on Spark monitoring and performance management.

6. Agile Methodology (JIRA & Scrum):

- Questions on managing multiple tasks using Agile methodology.

Round 5: Director Round (Behavioral & Technical Round): 45 minutes

1. Introduction & Experience Discussion:

- **Introduction**: Introduction of self and brief overview of professional background.
- **Project Experience**:
  - Discussion on the Datahub Spark Lineage Project and role as Data Engineer at Meesho.

2. Core Principles & Values:

- Questions on core principles or core values of Walmart and personal inspirations.

3. Team Management & Leadership:

- Situation-based questions like:
  - Tell me about a time when you faced a challenging situation at work and how you handled it.
  - Questions on team management and leadership qualities.

4. Technical Expertise:

- Discussion based on Presto vs. Spark distributed architecture, Databricks, AWS, Delta Lakes, and Data Governance.
- Specific questions included:
  - What is Avro file format & what is its significance in delta tables?
  - Difference between Presto vs. Spark underlying architecture.
  - Can Presto work with Near Real-Time Data (Streaming Data Source)?
  - How did you develop the Datahub using Open Source Projects such as Spline & Datahub?
  - What do you think about Data uncertainty?

Round 6: HR Round (General Discussion & Salary Discussion): 30 minutes

1. General Discussion:

- Questions about experience with Big Data projects, hobbies, and strengths & weaknesses.
- Inquiry about family background, previous interview experiences, and life goals.

2. Final Questions:

- "Why should we hire you?"
- "What inspires you to join Walmart?"

3. Salary Discussion:

- Discussion around salary and benefits.

4. Outcome:

- Positive feedback from HR, resulting in selection for the position of Senior Data Engineer (Data Engineer-3) at Walmart.

Glassdoor Walmart Review –

https://www.glassdoor.co.in/Reviews/Walmart-Reviews-E715.htm

Walmart Careers –

https://careers.walmart.com/

Subscribe to my YouTube Channel for Free Data Engineering Content –

https://www.youtube.com/@shubhamwadekar27

Connect with me here –

https://bento.me/shubhamwadekar

Checkout more Interview Preparation Material on –

https://topmate.io/shubham_wadekar