Welcome to Your Data Engineering Adventure !

You're about to step into the role of a data engineer at the heart of our analytics ecosystem. Each day, you'll transform raw, real-world data into actionable insights that empower BI teams, Data Scientists, Monitoring groups, and key stakeholders to make smarter decisions.

Ready to dive in? Here's your mission:

Task 1: Repository Setup

1. Set up your Git repository: This is your project's home base. organize your code, track your progress, and collaborate.
2. Create a README.md: Imagine you're onboarding a teammate. Write clear instructions on how to run your pipeline, describe the project structure, and share any assumptions you've made. Make it welcoming and easy to follow!

Task 2: Become a Data Detective

Explore the provided CSV files: Use your SQL skills to uncover key metrics:

1. What's the total maintenance cost?
2. How many minutes of downtime have there been?
3. How many maintenance events occurred?
4. How many breakdowns (unplanned) happened?
5. What's the average downtime per event?

Challenge: Can you write queries that are both efficient and easy to understand? Share your thought process!

Task 3: Build & Transform with PySpark

1. Design a robust pipeline: Read tables or CSV files (think warehouse-scale data).
2. Clean and transform: Implement at least two data cleaning or transformation functions. Get creative—how will you ensure data quality?
3. Produce a fact_table: This is your analytics-ready table, foundation for downstream users.
4. Export your results: Save your fact_table in both Parquet and CSV formats.
5. Bonus: Document your pipeline steps—what choices did you make and why?

Task 4: Test Your Ingenuity

1. Write at least two unit tests: Validate your transformation logic. How do you ensure your pipeline is reliable and robust?
2. Reflect: What edge cases did you consider? How would you extend your tests for production?

How to Submit

- Please write up your solution in your GitHub repository, including all code, documentation, and tests.
- Alternatively, you may share your work as a PDF document.
- You are welcome to submit both formats if convenient.

Your Impact: This isn’t just a technical exercise—it’s a chance to showcase your creativity, problem-solving, and communication skills. We’re excited to see how you approach real-world data challenges and deliver solutions that drive business value.

Are you ready to make an impact? Let’s get started!