Data Engineer Interview Questions

These are the minimum questions to ask a data engineering professional. I do not provide “correct” answers because correctness is not important. Anybody worth a darn will be able to answer these questions to some extent. The key is to understand to what depth they understand the concepts. Length and substance of answers are more important than if the answers are technically correct.

An experienced professional might be slightly insulted being asked these questions, some even exhibiting impatience or boredom at having to answer them. That’s a good sign.

1. Explain the difference between Kimball and Inmon methodologies. (You do not really care about the answer. You just want to see if they are familiar with the concepts.)

2. Discuss the different types of slowly changing dimensions. (Being able to explain I, II, and III is the minimal acceptable answer; bonus if they know that there are more than three.)

3. When you are pulling large amounts of data from a transactional source system, how do you keep from locking users out? (Amateur answer: "With no lock." Professional answer: "Set transaction isolation level read committed." Bonus if they can provide more color around these answers.)

4. There is a number we look at that is the result of a calculation. Where in the ETL process does the math for that calculation belong? (Answer: It does not belong in the ETL at all. It goes in the report. Bonus points if they can explain why. Extra, extra bonus points if they answer, "Somewhere in the ETL," and can specify where, why, and justify their answer.)

5. We are getting started and this is what we are planning. [List off the things you are planning.] What else do you recommend we do? (Use this answer to calibrate experience, expectations, and separate candidates.)

6. What are the challenges surrounding managing the history of a Type II SCD? (There is no correct answer here. Just sit back and enjoy the show.)

7. We have customers stored in three different systems. Because they are entered three different ways, we often cannot tell when a customer is entered twice. We don’t want duplicates in the warehouse. How would you solve this problem? (Again, sit back and enjoy the show. Bonus points if the candidate asks clarifying questions. Be prepared to flesh out the scenario for your specific organization.)

8. How do you decide where to put indexes on tables? (Indexes speed things up, yet they also take up space and slow things down. You don’t want to hire someone that is index happy and doesn’t have a methodical way to decide where they go.)

9. Describe how you go about transforming a transactional system into a dimensional model. (They may not have an answer for this. If they don’t, that is a sign of inexperience. It will be up to you if that is acceptable or not.)

10. We get bank reconciliation data in a feed. The data comes in at the transaction level on a daily basis but it also gives us a month-to-date rollup. How would you go about storing this? (There is more than one way to skin this cat. Make them describe in detail how they would handle this.)