Mostafa Mohamed ismail ahmed

Data Quality Tests

These tests assess the fundamental quality of the data, paying special attention to accuracy, completeness, and consistency.

1. **Data completeness testing:**

   This ensures that all necessary information is present. If a field is marked as required, this test verifies that it isn't left blank. It's especially important when moving data between systems or bringing in new data sources, since missing pieces can lead to flawed insights.

2. **Data consistency testing:**

   This verifies that the data across various systems or databases is aligned and adheres to the same standards. When dealing with multiple data sources, this test checks that things like data formats and naming conventions match up, helping to avoid errors in reports and analysis.

3. **Data accuracy testing:**

   This confirms that the data accurately represents the real world it's intended to depict. It's critical for organizations that rely on data for important decisions—think financial or healthcare sectors. You can verify accuracy by comparing your data to a trusted source.

4. **Data Quality & Integrity testing:**

   This ensures that the data remains unchanged and keeps its consistency and accuracy throughout its lifecycle. It's all about safeguarding the data from corruption or unauthorized alterations, which is crucial when introducing new systems or databases where data is being relocated or transformed.

   i. **Uniqueness Tests**:These tests confirm that a column contains only unique entries, without duplicates. This is especially important for fields used as identifiers—like customer IDs, social security numbers, or product codes—where each entry needs to be distinct.
   ii. **Null Values Test**:This test looks for any missing or empty values in your dataset. It's a basic check for data completeness; a null value might suggest an error in data entry, a gap in the data pipeline, or simply an unknown value.
   iii. **Freshness Checks**: Freshness checks gauge how current your data is. This is crucial for time-sensitive information, like stock prices or sensor readings, ensuring that your data reflects the present situation and isn't outdated, which could lead to poor decision-making.
   iv. **String Patterns**: This validation test checks if text data follows a specified format or pattern. For instance, a string pattern test can confirm that all email addresses are structured correctly, like "name@domain.com," or that phone numbers adhere to a particular format.

5. **Data validation testing:**

   This ensures that the data entered into a system meets set rules. For example, it will check that a date is entered in the correct format or that a number falls within an acceptable range. This is key when user input is involved, as it helps to block bad data from entering the system.

6. **Data regression testing:**

   This involves re-evaluating data-related components after any changes have been made to the system, like a software update or bug fix. Its aim is to ensure that these changes don't introduce new issues or bring back old ones. It's an essential practice for maintaining system stability following any adjustments.

7. **Data performance testing:**

   This evaluates how well a system manages a large volume of data. It looks at response times and resource usage to ensure the system can handle the anticipated load. This step is crucial when building systems designed to process significant amounts of data or that have strict performance expectations.

8. **Data Volume & Distribution Tests:**

   These tests assess the characteristics of the data and its behavior within a system.

   i. **Volume Tests**: A volume test checks a system's capacity to handle large datasets, seeing how it performs when processing, storing, and retrieving significant amounts of data to ensure it doesn't lag or crash under heavy loads.
   ii. **Numeric Distribution Tests**: This test examines the statistical distribution of numeric data, helping to identify if the data is normally distributed or skewed. This step is vital before conducting statistical analyses like t-tests or ANOVAs, which often require specific data distributions.