NATIONAL INSIDER THREAT TASK FORCE

NITTF Tech Bulletin TB20180105: Data Quality for Insider Threat Programs

Abstract:

Executive branch departments and agencies should not overlook the importance of data quality to their insider threat programs. Inaccurate or 'poor-quality' data can hinder a program's ability to identify theat behaviors and conduct an effective inquiry.

BACKGROUND and GUIDANCE:

Insider threat programs should not overlook the importance of their organization's data quality. Data quality refers to the accuracy, usefulness, and availability of an organization's data records or data resources. Numerous studies have shown that data quality has a significant impact on an organization's ability to understand its business, to conduct business, and sometimes even to remain in business.

Data-quality experts often describe data quality in terms of "good-quality" data and "poor-quality" data. The term "good-quality" data refers to data records that are accurate (there are few, if any, errors in the data records), available (because of the accuracy, the data records are easily searched), and suitable (the data records contain the information you need to complete the mission). The most effective insider threat programs – the programs that can identify and resolve threats, protect employees and facilities, and conduct insightful analysis and inquiries – are those that base their efforts on good-quality data.

The term "poor-quality" data refers to data records that have data errors in them. These data errors can be anything from misspellings to name variations (e.g., Smith, Smiht, Smit, Smyth, Smythe, B. Smith, Smith B., etc.), wrong or incomplete addresses, inconsistent data formats (e.g., telephones formatted as 9999999999 or 999-9999 instead of 1-999-999-9999), missing or duplicated information, and even duplicated data records. There are numerous reasons for poor-quality data, but the most important thing to bear in mind is that basing decisions on poor-quality data can have a negative effect on insider threat programs. One common data-quality problem is the misspelling or variation of a subject's name. A simple misspelling will often generate a new database record (called a duplicate record), but the information associated with this misspelling will not be available to any other data record of this subject. If multiple or duplicate records exist when there should only be one record, then the insider threat program may have difficulty locating all the relevant information on a subject and conducting an effective inquiry.

As a best practice, the NITTF recommends that executive branch departments and agencies (D/As) periodically review the quality of their data. A review typically entails taking a small cross-section of the D/A's data and checking the data records for accuracy, completeness, and duplication. While correcting all data records is a more significant undertaking, the review will inform the D/A of some of the pitfalls of its insider threat program. D/As can also have their data problems corrected (or "cleansed") using a commercial data-quality tool.

For more information about data quality, contact the NITTF Tech Team at nittftechnical@dni.ic.gov.

Common Data Quality Problems:

- Spelling and punctuation errors
- Incorrect data (e.g., incorrect birth dates, social security numbers, addresses, etc.)
- Missing data (blank spaces in data records)
- Inconsistent data formats (e.g., zip codes formatted as 99999, 99999-9999, 99999999, or names variously entered as Smith, Robert; Robert Smith; R Smith; Bob Smith, etc.)
- Data in the wrong record fields (e.g., address data in name fields)
- Duplicate data (multiple data records on the same subject due to inconsistent formats, incorrect information, etc.)
- Incomplete data records (records in which all attributes of an entity are not available)