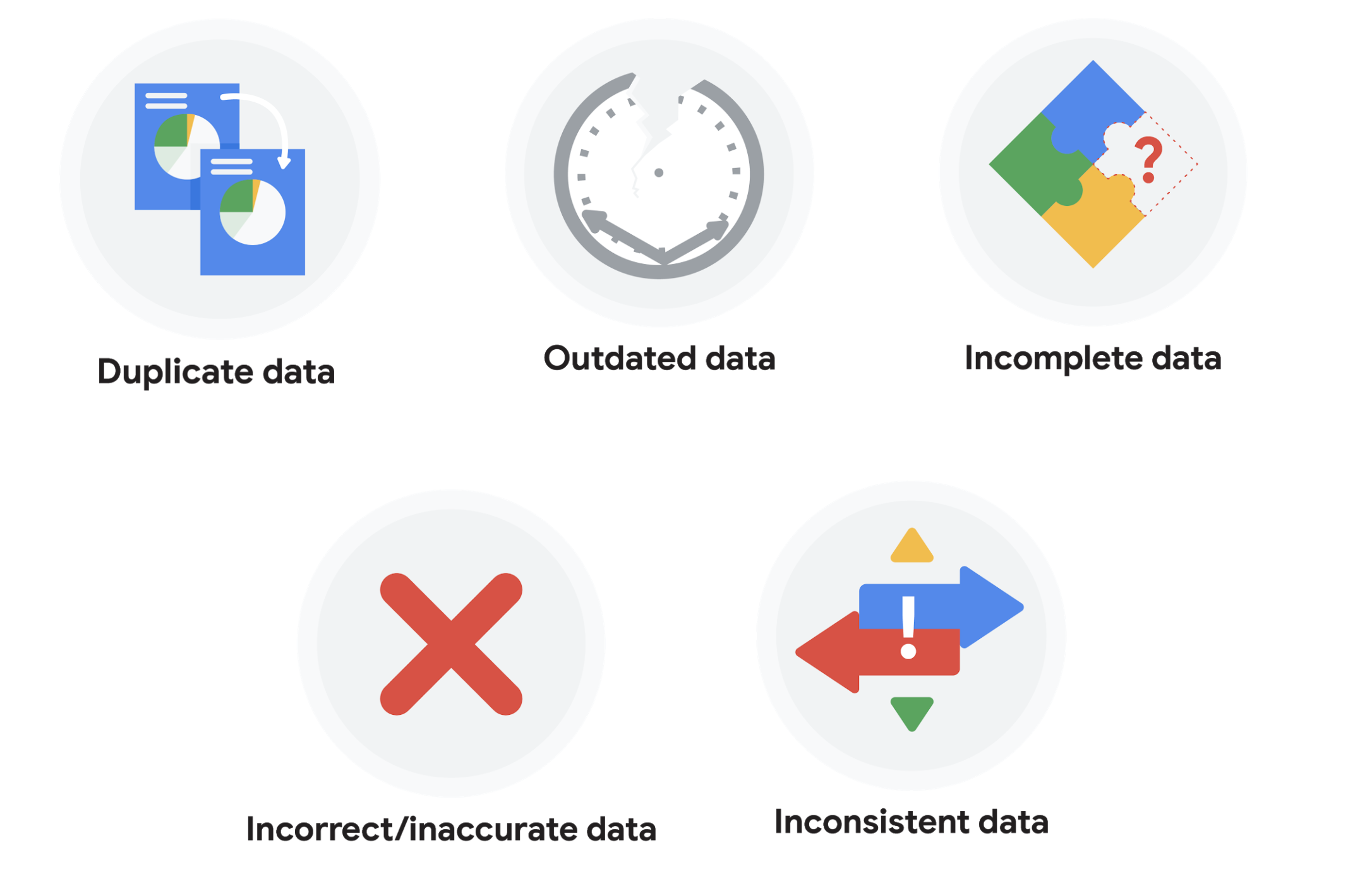
## **Summary:**

# What is dirty data?

Earlier, we discussed that **dirty data** is data that is incomplete, incorrect, or irrelevant to the problem you are trying to solve. This reading summarizes:

* Types of dirty data you may encounter
* What may have caused the data to become dirty
* How dirty data is harmful to businesses

## **Types of dirty data**

****

### **Duplicate data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data record that shows up more than once | Manual data entry, batch data imports, or data migration | Skewed metrics or analyses, inflated or inaccurate counts or predictions, or confusion during data retrieval |

### **Outdated data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that is old which should be replaced with newer and more accurate information | People changing roles or companies, or software and systems becoming obsolete | Inaccurate insights, decision-making, and analytics |

### **Incomplete data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that is missing important fields | Improper data collection or incorrect data entry | Decreased productivity, inaccurate insights, or inability to complete essential services |

### **Incorrect/inaccurate data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that is complete but inaccurate | Human error inserted during data input, fake information, or mock data | Inaccurate insights or decision-making based on bad information resulting in revenue loss |

### **Inconsistent data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that uses different formats to represent the same thing | Data stored incorrectly or errors inserted during data transfer | Contradictory data points leading to confusion or inability to classify or segment customers |

### **Business impact of dirty data**

For further reading on the business impact of dirty data, enter the term “dirty data” into your preferred browser’s search bar to bring up numerous articles on the topic. Here are a few impacts cited for certain industries from a previous search:

* **Banking**: Inaccuracies cost companies between 15% and 25% of revenue ([source](https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/)).
* **Digital commerce:** Up to 25% of B2B database contacts contain inaccuracies ([source](https://www.demandgen.com/dirty-data-what-is-it-costing-you/)).
* **Marketing and sales**: 8 out of 10 companies have said that dirty data hinders sales campaigns ([source](https://www.dqglobal.com/2011/05/04/obsolete-or-dirty-data/)).
* **Healthcare**: Duplicate records can be 10% and even up to 20% of a hospital’s electronic health records ([source](https://searchhealthit.techtarget.com/feature/Hospitals-battle-duplicate-medical-records-with-technology)).

**Data Engineers:->**

1. Data engineers are responsible for designing, building, and maintaining the infrastructure and systems that store, process, and analyze large volumes of data.
2. They work closely with data scientists and analysts to ensure that data is properly collected, cleaned, and structured for analysis and reporting.
3. Some key responsibilities of a data engineer include designing and implementing data pipelines, managing and optimizing data storage systems, and ensuring data security and privacy.

**DataWarehouse specialists:->**

1. Data warehouse specialists are responsible for designing and maintaining the systems that store and organize large amounts of data for reporting and analysis.
2. They work closely with business stakeholders and data analysts to understand the organization's data needs and design the data warehouse accordingly.
3. Some key responsibilities of a data warehouse specialist include designing and implementing data models, ETL processes, and data governance policies, as well as maintaining and optimizing the performance of the data warehouse.

**Field and Field Length:->**

In a database, a field refers to a specific piece of information or data that is stored in a specific column of a table. A field can contain a wide range of data types, such as text, numbers, dates, and Boolean values.

Field length refers to the maximum number of characters or bytes that a field can contain. For example, a field that is defined as a text field with a length of 100 can store up to 100 characters of text. Field length is an important consideration when designing a database, as it can affect the performance and storage requirements of the database.

Field length can also be used to validate user inputs, if the input is longer than the field length it can be truncated or rejected.

For example, a VARCHAR field in a SQL database typically has a specified maximum length, such as VARCHAR(50) which means the field can hold up to 50 characters

.