

データ品質

- データ品質とは
- データ品質の評価指標
- データ品質リスク
- データ品質対策基準

データ品質

- データ品質とは
- データ品質の評価指標
- データ品質リスク
- データ品質対策基準

データ品質とは

データが価値を持つ前提条件は、データそのものが確かで信頼できるかということである。

データ品質が低いと....

- 誤請求
- 顧客サービスコールの増加とそれを解決する能力の低下
- 事業機会の逸失による収益損失
- 合併・買収の間に発生する業務統合の遅延
- 不正なデータに起因する業務上の意思決定不備がもたらす損失
- 良好な信用力の欠如による事業の損失

データマネジメント知識体系 第二版 (DMBOK2)

データ品質

- データ品質とは
- データ品質の評価指標
- データ品質リスク
- データ品質対策基準

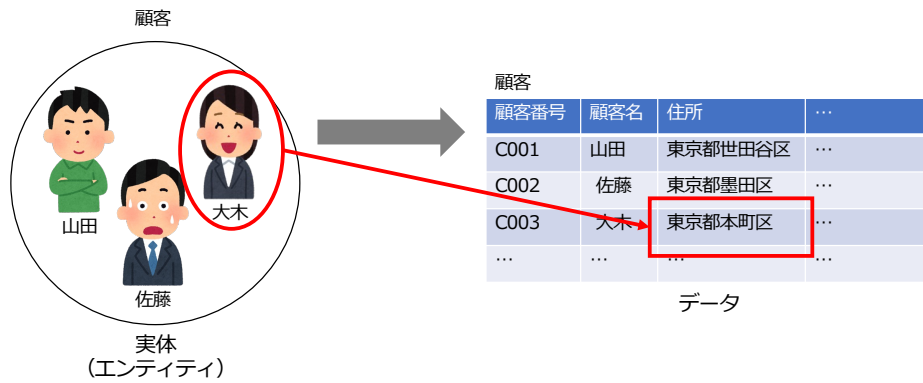
データ品質の評価指標

- 正確性
- 完全性
- 一意性
- 一貫性
- 適時性
- 有効性

データマネジメント知識体系 第二版 (DMBOK2)

正確性

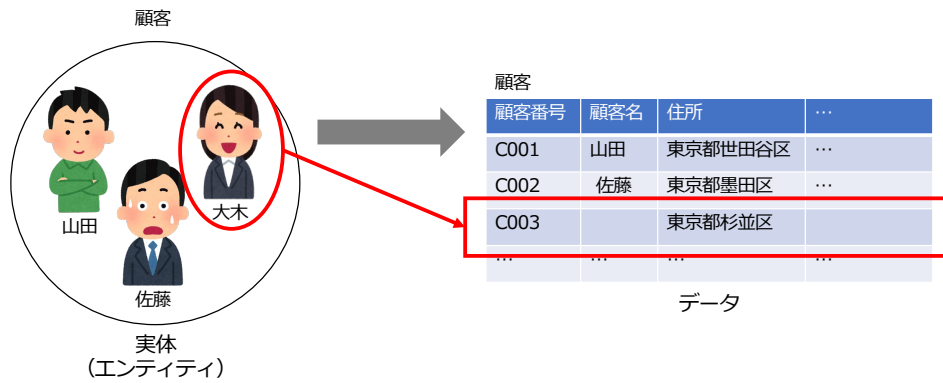
データが現実の実体を正しく表している程度を表す。



データマネジメント知識体系 第二版 (DMBOK2)

完全性

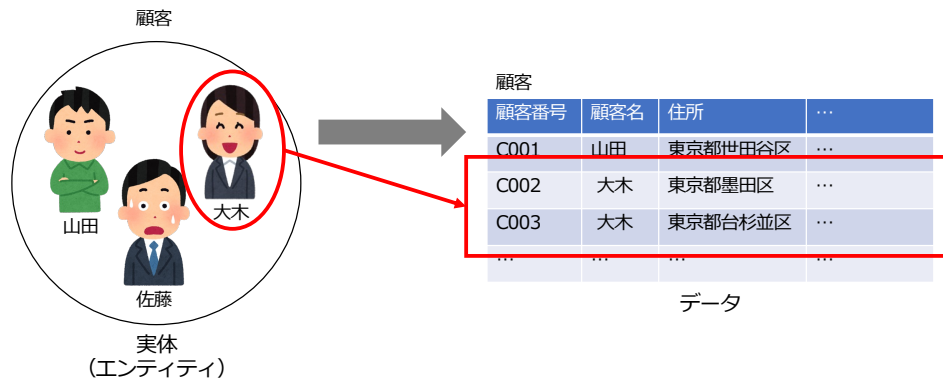
必要なデータが全て存在しているかどうか、その程度を表す。



データマネジメント知識体系 第二版 (DMBOK2)

一意性

同じ実体を表すデータが同じデータセット内に複数存在していないか、その程度を表す。



データマネジメント知識体系 第二版 (DMBOK2)

一貫性

データが、特定のデータベースなどデータセット内で一貫して（正しいルールに貫かれて）表現されているか、あるいは、データセット間で一貫して関連付けられ、一貫して表現されているか、その程度を表す。

外部キーを介したデータセット間の一貫性（例えば参照先のデータだけがないという欠落がない）を参照整合性という。

- データ項目の一貫性
- データ項目間の一貫
- エンティティ間の一貫

データマネジメント知識体系 第二版 (DMBOK2)

データ項目の一貫性

同じデータ項目のデータの値に一貫性があるかどうか、その程度を表します。
あるレコードのデータ項目の値と、異なる時点における同じレコードのデータ項目の値との間の一貫性も含まれます。（経時的一貫性）

注文番号	注文日付	注文金額	...
S001	2021-1-3	1000円	...
S002	2021/3/5	2000円	...
S003	10/3/2021	5000円	...
...	...	...	...

データマネジメント知識体系 第二版（DMBOK2）

データ項目間の一貫性

1レコード内にあるデータ項目の値と別のデータ項目の値との間に一貫性はあるか、その程度を表します。

出荷番号	注文日付	出荷予定日	...
P001	2021/1/3	2021/3/4	...
P002	2021/3/5	2021/5/6	...
P003	2021/10/3	2021/7/8	...
...	...	...	...

データマネジメント知識体系 第二版 (DMBOK2)

エンティティ間の一貫性

あるレコードのデータ項目の値と別のレコードのデータ項目の値の間に一貫性はあるか、その程度を表します。

外部キーを介したデータセット間の一貫性（例えば参照先のデータだけがないという欠落がない）を参照整合性といいます。

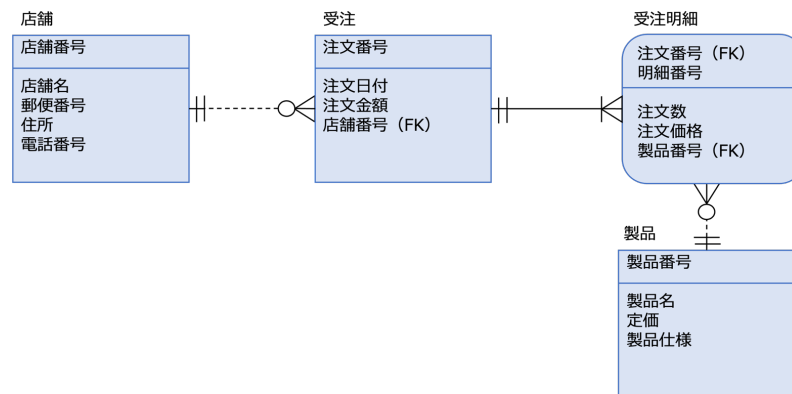
エンティティ間の一貫性を担保するための制約

- 存在制約
あるインスタンスが存在するには、他のインスタンスの存在が必要であるという制約。
- 参照整合性制約
外部キーの値が主キーの値として存在することを保証する制約。

存在制約

論理データモデルの例

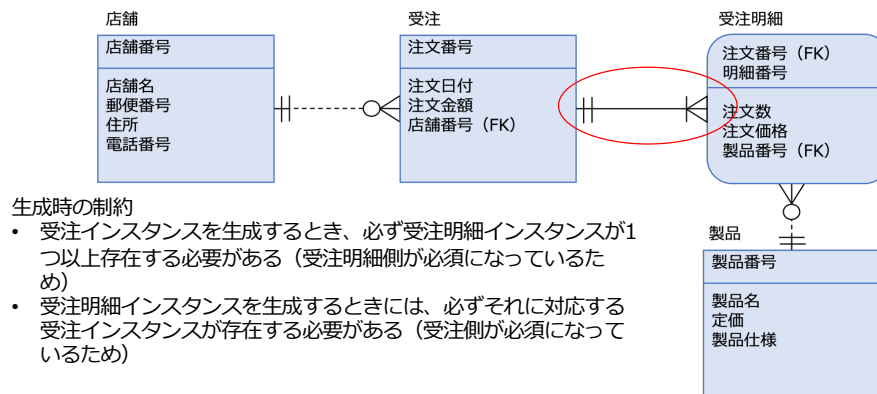
多重度のオプションナリティやリレーションシップで存在制約を表現することができる。



存在制約

論理データモデルの例

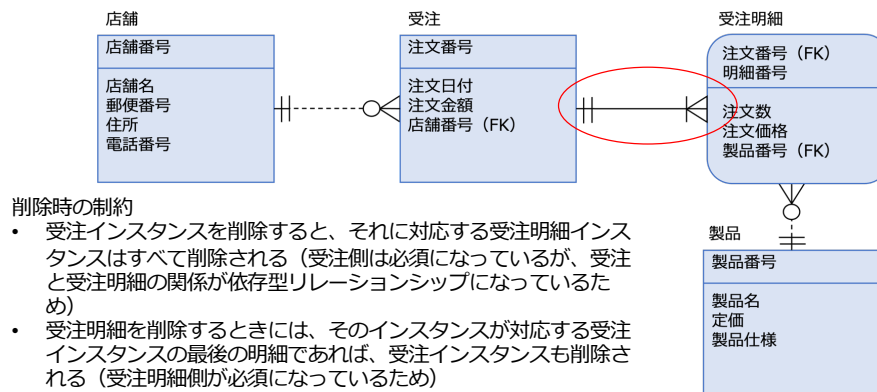
多重度のオプションナリティやリレーションシップで存在制約を表現することができる。



存在制約

論理データモデルの例

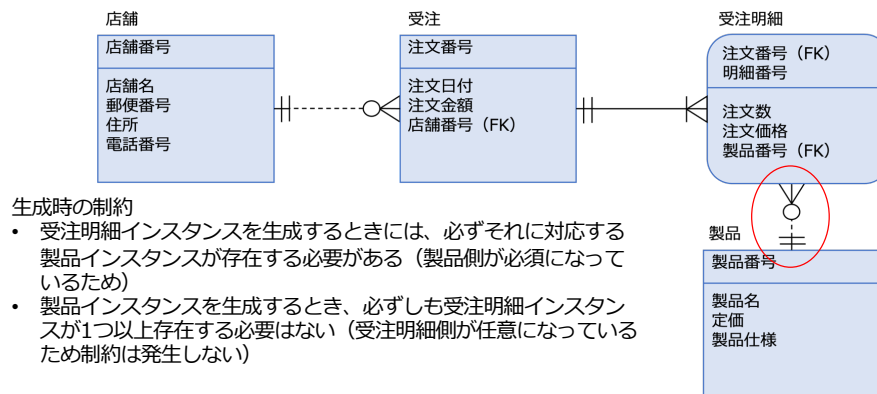
多重度のオプションナリティやリレーションシップで存在制約を表現することができる。



存在制約

論理データモデルの例

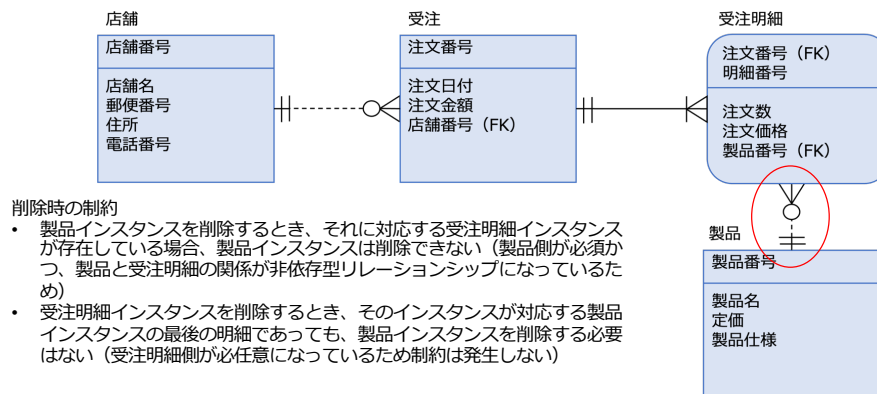
多重度のオプションナリティやリレーションシップで存在制約を表現することができる。



存在制約

論理データモデルの例

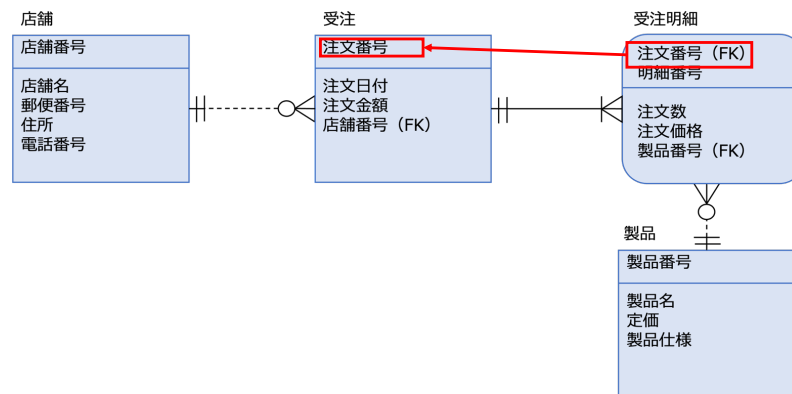
多重度のオプションリティやリレーションシップで存在制約を表現することができる。



参照整合性制約

論理データモデルの例

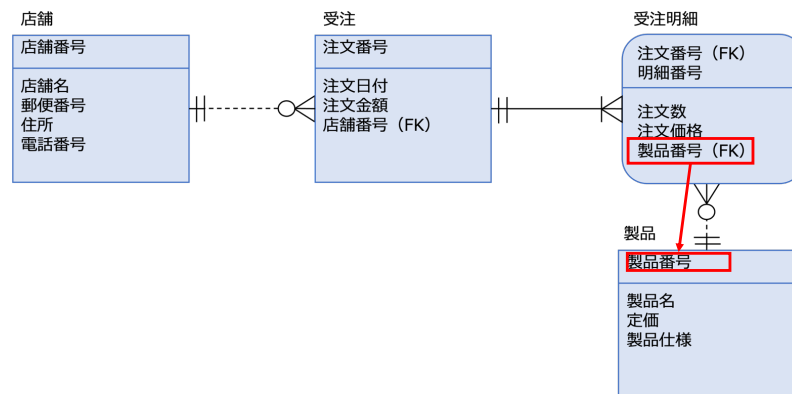
外部キーである注文番号の値が主キーの値として存在しなければならない。



参照整合性制約

論理データモデルの例

外部キーである製品番号の値が主キーの値として存在しなければならない。



適時性

データが適切な時点のものであるか、その程度を表す。

多くのデータの適時性は、データが最新かどうかによって測られますが、必ずしもそうではありません。

例えば、過去のある時点の状態を知りたい場合、過去の時点のデータが適時性の高いデータになります。

適時性の評価基準

データが発生して利用可能になるまで遅延する場合、遅延の程度によってデータの適時性を定義することになります。

データ利用者がデータのレイテンシをどの程度許容できるのかデータニーズとして把握し、それを品質評価基準（許容値）として設定する。

※レイテンシ

ソースシステムでデータが生成されてからターゲットシステムでデータが利用可能になるまでの時間差。

例

株式市場のデータが取引されてから 5 分以内に届くこと。

適時性の評価基準

エンティティ番号	データ品質要件				
	データ品質評価指標	ビジネスルール	品質測定方法	品質評価基準	品質評価結果
E001	適時性	株式市場のデータが取引されてから5分以内に届くこと。	不完全なトランザクションの数を、試行したトランザクションの総数で除算して100を掛けてパーセントで表す。	許容できない結果 1%以上が5分以内に届かない。	—

有効性

データが定義域（データドメイン）に準拠しているか、その程度を表す。

定義方法	内容
データタイプ型	属性に当てはまる標準データタイプを指定するドメイン。例えば、整数型、文字型、日付型はすべてデータタイプ型ドメイン。
データ形式型	郵便番号や電話番号にあるようなテンプレート、有効な値を定義するための文字制限（英数字のみ、英数字に加えて特殊文字を使用できるなど）のパターンを使用したドメイン。
リスト型	有限個の値を持つドメイン。ドロップダウンリスト機能としてよく知られている。例えば、受注ステータスコードのリスト型ドメインは受注済、出荷済、処理済、返品済といった値のみに制限できる。列挙型とも呼ばれる。
範囲型	同じデータタイプで、最小値と最大値の組合せを一つか複数持ち、その間にあるすべての値を許すドメイン。例えば、発送日は注文日から3ヶ月以内など。範囲に明示的な終了日を設定しなくてもよい。
ルールベース型	ルールにより定義されたドメインで、値が有効であるためにはルールに従わなければならない。ルールには、入力値を計算された値と比べたり、リレーションシップを持つ他エンティティの属性値と比べたりするものもある。例えば、売値は仕入値より高くなければならないなど。

データマネジメント知識体系 第二版（DMBOK2）

データ品質・演習①

次の例のように、注文日付の書式がバラバラになっています。
これは、データ品質上、何が問題でしょうか。

注文番号	注文日付	注文金額	...
S001	2021-1-3	1000円	...
S002	2021/3/5	2000円	...
S003	10/3/2021	5000円	...
...	...	...	...

データ品質・演習①・解答例

次の例のように、注文日付の書式がバラバラになっています。
これは、データ品質上、何が問題でしょうか。

注文番号	注文日付	注文金額	...
S001	2021-1-3	1000円	...
S002	2021/3/5	2000円	...
S003	10/3/2021	5000円	...
...	...	...	...

データ項目の一貫性に問題があります。
なので、例えば、ある特定日以降の注文を検索する場合など、データの処理がうまくできない可能性があります。

データ品質・演習②

次の例のように、注文日付の書式がバラバラになっています。
データ項目の一貫性を確保するための方法を考えてください。

注文番号	注文日付	注文金額	...
S001	2021-1-3	1000円	...
S002	2021/3/5	2000円	...
S003	10/3/2021	5000円	...
...	...	...	...

データ品質・演習②・解答例

次の例のように、注文日付の書式がバラバラになっています。
データ項目の一貫性を確保するための方法を考えてください。

注文番号	注文日付	注文金額	...
S001	2021-1-3	1000円	...
S002	2021/3/5	2000円	...
S003	10/3/2021	5000円	...
...	...	...	...

日付の書式（例えば、yyyy/mm/ddなど）を定義したデータ形式型ドメインを設け、注文日付を、それに準拠するようにする。

データ品質・演習③

日付の書式を定義したデータドメインを設け、注文日付を、それに準拠するようにしましたが、まだ、注文日付の書式がバラバラになっています。

これは、データ品質上、何が問題でしょうか。

注文番号	注文日付	注文金額	...
S001	2021-1-3	1000円	...
S002	2021/3/5	2000円	...
S003	10/3/2021	5000円	...
...	...	...	...

データ品質・演習③・解答例

日付の書式を定義したデータドメインを設け、注文日付を、それに準拠するようにしましたが、まだ、注文日付の書式がバラバラになっています。

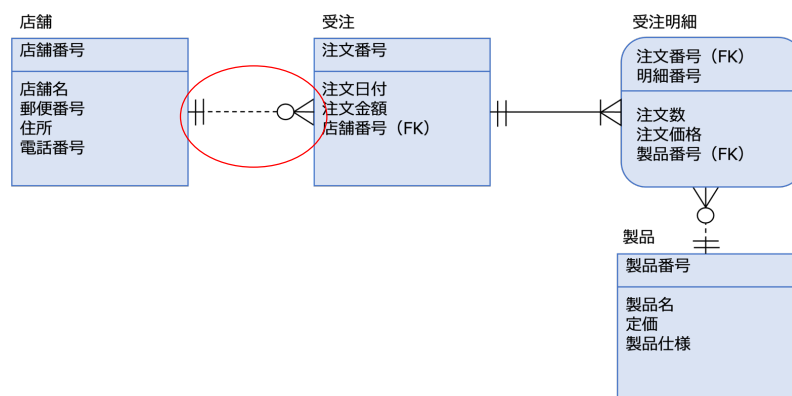
これは、データ品質上、何が問題でしょうか。

注文番号	注文日付	注文金額	...
S001	2021-1-3	1000円	...
S002	2021/3/5	2000円	...
S003	10/3/2021	5000円	...
...	...	...	...

データの有効性に問題があります。
なので、例えば、アプリケーションやデータベースでデータドメインが有効になるような処理を実装するなどの対策を講じる必要があります。

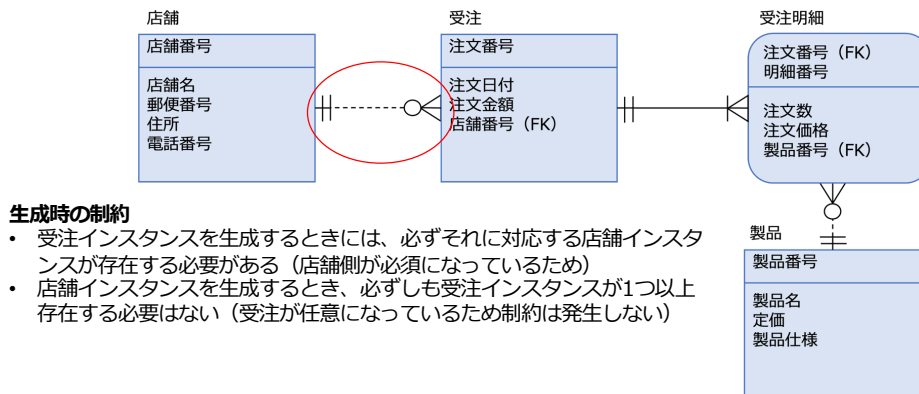
データ品質・演習④

次の論理データモデルの受注エンティティと店舗エンティティのリレーションシップからどのような存在制約が考えられますか。



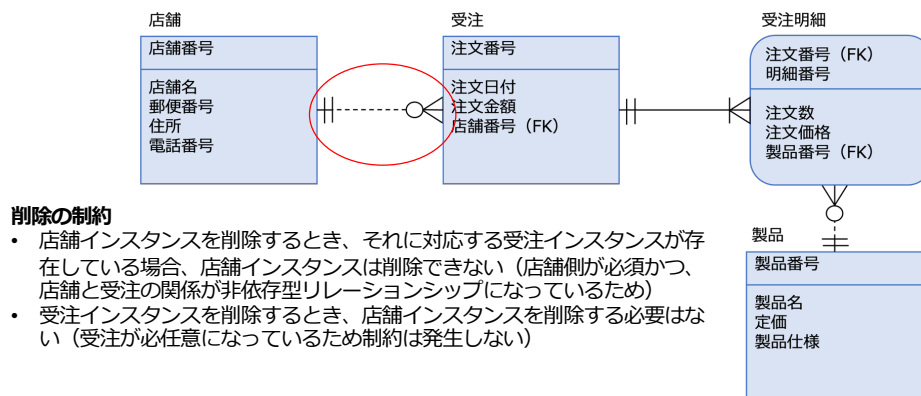
データ品質・演習④・解答例

次の論理データモデルの受注エンティティと店舗エンティティのリレーションシップからどのような存在制約が考えられますか。



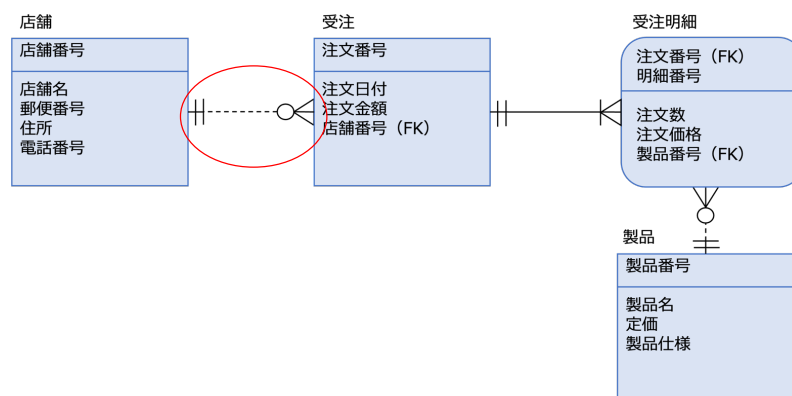
データ品質・演習④・解答例

次の論理データモデルの受注エンティティと店舗エンティティのリレーションシップからどのような存在制約が考えられますか。



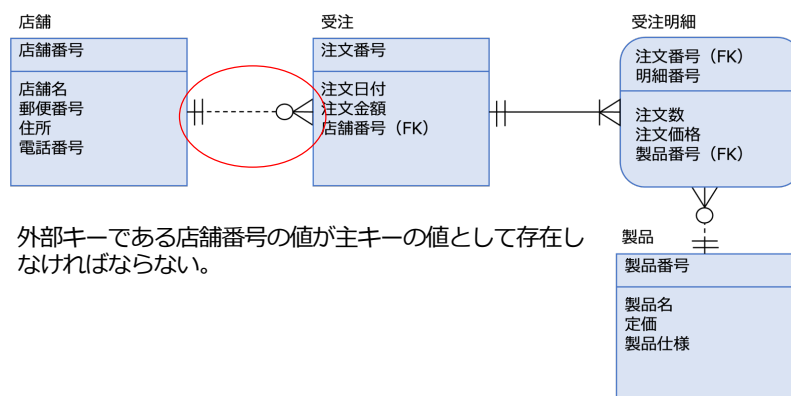
データ品質・演習⑤

次の論理データモデルの受注エンティティと店舗エンティティのリレーションシップからどのような参照整合性制約が考えられますか。



データ品質・演習⑤・解答例

次の論理データモデルの受注エンティティと店舗エンティティのリレーションシップからどのような参照整合性制約が考えられますか。



データ品質

- データ品質とは
- データ品質の評価指標
- データ品質リスク
- データ品質対策基準

。

データ品質リスクの例

- **参照整合性制約の不備**

参照整合性制約に不備がある場合、存在しないデータの参照など、データ品質上の問題が発生する可能性があります。

- **存在制約の不備**

存在制約に不備がある場合、必要なデータがないのにデータが生成されたり、一緒に削除すべきデータが削除されなかったりする可能性があります。

- **一意性制約の不備**

一意性制約に不備がある場合、データが一意であることを想定しているテーブルに同じデータが複数存在するなど、データ品質上の問題が発生する可能性があります。

- **時系列データの不一致**

複数のシステムで異なる日付形式や時間形式が実装されると、システム間でデータの同期が行われると、データの不一致やデータの損失が発生するなど、データ品質上の問題が発生する可能性があります。

- **プログラムの欠陥**

不正確なプログラミングがあると、計算方法が間違っていたり、データが不適切なフィールド、キーに割り当てられたり、間違った参照先にリンクされたりし、データ品質上の問題が発生する可能性があります。

- **入カインターフェースの問題**

データ入力インターフェースに不正なデータがシステムに書き込まれることを防ぐための編集機能や制御が有効でない場合、データ品質上の問題が発生する可能性があります。

データマネジメント知識体系 第二版（DMBOK2）

データ品質リスクの例

- **モデルと実装の乖離**

データモデル内に設けられた前提と、実際のデータが異なる場合、実際のデータがフィールド長を超えたため一部失われたり、データが不適切なIDやキーに割り当てられたりするなど、データ品質上の問題が発生する可能性があります。

- **フィールドの多目的利用**

様々な目的でフィールドを利用すると、紛らわしい値、不明瞭な意味、誤ってキーを割り当てるなど、データ品質上の問題が発生する可能性があります。

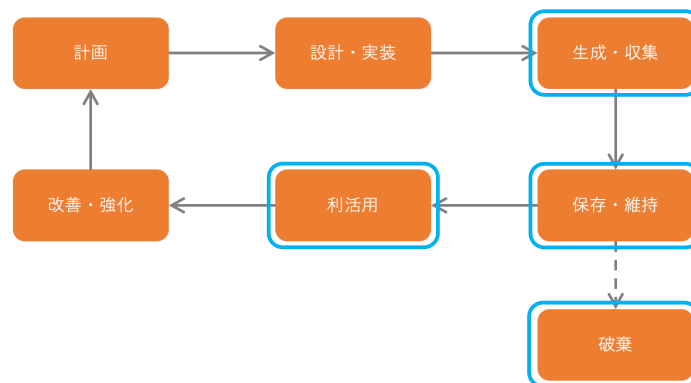
- **データ統合の不備**

データの統合に不備があると次のような問題が発生します。
例えば、同じデータベース内の複数テーブルに同じ顧客を表すデータがある場合、どちらのデータが正確であるか、知るのが困難になります。
また、複数のシステムから連携される単一の顧客データを処理する場合、重複が発生する可能性があります。

- **データ操作の教育不足**

入力プロセスの知識が不足していると、プログラムで入力の制御や編集が行われていても、誤ったデータが入力されてしまう可能性があります。

データ品質リスクとデータライフサイクル



データマネジメント知識体系 第二版 (DMBOK2)

データ品質

- データ品質とは
- データ品質の評価指標
- データ品質リスク
- データ品質対策基準

データ品質対策基準の例

- 参照整合性制約の定義
- 存在制約の定義
- 一意性制約の定義
- データドメインの適用
- データの正確な意味定義
- データ入力の自動化
- データ処理の検証
- データ操作の教育
- データのプロファイリング
- データのクレンジング
- データの強化

データマネジメント知識体系 第二版 (DMBOK2)

データのプロファイリング

- データプロファイリングは、データを検査し、品質を評価するために行われるデータ分析の一形式です。
- データプロファイリングは、統計的手法を活用して、収集したデータの真の構造、コンテンツ、品質を洗い出します。

データマネジメント知識体系 第二版 (DMBOK2)

プロファイリングの例

- NULL数
NULLが存在することを検出し、許容可能かどうかを検査できるようにします。
- 最大値・最小値
マイナス値などの異常値を検出します。
- 最大長・最小長
特定の桁数要件を持つフィールドの外れ値や異常値を検出します。
- 個々のカラムに存在する値の度数分布
トランザクション内の国コード分布、値の発生頻度検査、初期値が設定されているレコードの割合など値の妥当性を評価します。
- データタイプとフォーマット
小数点以下の桁数、スペースの混入などフォーマット要件への不適合レベルを検出します。

データマネジメント知識体系 第二版 (DMBOK2)

データのクレンジング

データの洗浄。

破損または不正確なデータを検出して修正するプロセス。

データの名寄（データベースに存在するデータを名前、住所、電話番号などの情報から読み取り、同一人物や同一企業等の重複しているデータをひとつにまとめる作業）もクレンジングの一種と考えます。

データのスクラビング（バックグラウンドタスクを使用してメインメモリまたはストレージのエラーを定期的に検査し修正する手法）もクレンジングの一種と考えます。

データマネジメント知識体系 第二版 (DMBOK2)

データの強化

次の例のようにデータを強化することで品質を上げることができます。
データの強化はビジネスメタデータとして設定することができます。

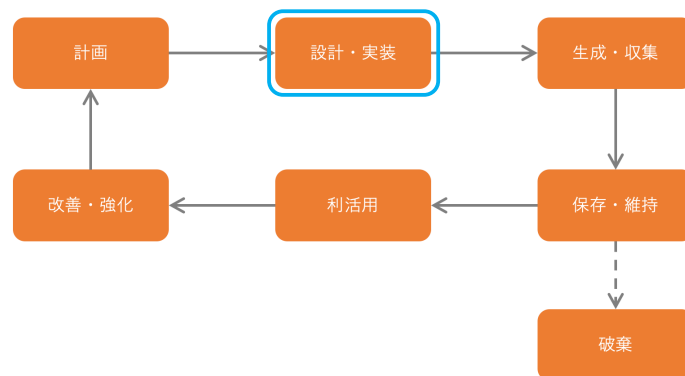
- **時間・日付スタンプ**
データ品質に問題が発生した場合、問題が発生した時間帯を分離できるため、データの生成、更新、破棄された日時を履歴として記録するようにします。
- **データリネージュ**
データの発生から現在に至るまでの一連の情報、データの来歴。
データリネージュを記録することで、データのライフサイクルが可視化され、どこで問題が生じたかなど追跡できるようになります。
- **参照用語**
業界固有の用語も含めた用語集を整備することにより、データに対する理解と統制が強化されます。
- **コンテキスト情報**
データに対するレビューと分析をするために、場所、環境、アクセス方法、データのタグなどコンテキスト情報をデータに付加します。
- **地理情報**
住所の標準化と位置座標化によって地理情報を強化することができます。
- **人口統計情報**
顧客データは、年齢など人口統計情報によって強化することができます。
- **心理学的情報**
商品やブランドの嗜好など心理学的情報で顧客データを強化することで市場をセグメント化することができます。
- **評価情報**
資産評価、在庫管理、販売促進には、この種の属性を拡張することで強化することができます。

データマネジメント知識体系 第二版 (DMBOK2)

データ品質対策基準とデータライフサイクル

- 参照整合性制約の定義
- 存在制約の定義
- 一意性制約の定義
- データドメインの適用
- データの正確な意味定義
- データ入力の自動化
- データ処理の検証
- データ操作の教育
- データのプロファイリング
- データのクレンジング
- データの強化

データ対策基準とデータライフサイクル

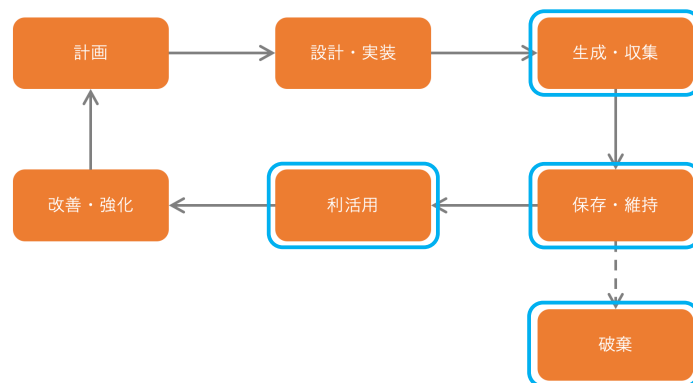


データマネジメント知識体系 第二版 (DMBOK2)

データ品質対策基準とデータライフサイクル

- 参照整合性制約の定義
- 存在制約の定義
- 一意性制約の定義
- データドメインの適用
- データの正確な意味定義
- データ入力の自動化
- データ処理の検証
- データ操作の教育
- データのプロファイリング
- データのクレンジング
- データの強化

データ品質対策基準とデータライフサイクル

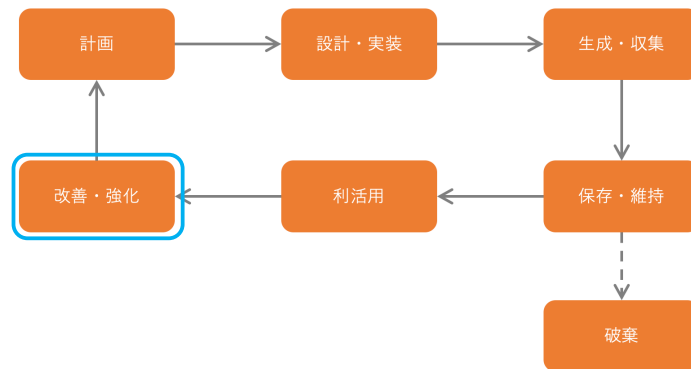


データマネジメント知識体系 第二版 (DMBOK2)

データ品質対策基準とデータライフサイクル

- 参照整合性制約の定義
- 存在制約の定義
- 一意性制約の定義
- データドメインの適用
- データの正確な意味定義
- データ入力の自動化
- データ処理の検証
- データ操作の教育
- データのプロファイリング
- データのクレンジング
- データの強化

データ品質対策基準とデータライフサイクル



データマネジメント知識体系 第二版 (DMBOK2)