

Professional Data Engineer

認定試験ガイド

Professional Data Engineer は、さまざまなアプリケーションのデータを収集、変換、保存、配信することで、データドリブンな意思決定を可能にします。Professional Data Engineer は、パフォーマンスとセキュリティを最適化しながら、堅牢なデータ インフラストラクチャを設計、構築します。ビジネスニーズと規制ニーズを満たすソリューションを評価して選択し、データ プラットフォームを効果的に管理します。Professional Data Engineer は、データ処理、クリーニング、拡充、クエリの生成と変換に最新のテクノロジーを活用します。Professional Data Engineer は、データ ストレージとデータ処理の複雑さを理解し、複雑なワークロードの設計、構築、デプロイ、モニタリング、メンテナンス、最適化、保護に長けています。

セクション 1: データ処理システムの設計 (試験内容の 22% 以下)

セキュリティとコンプライアンスを考慮した設計考慮事項:

- Identity and Access Management (Cloud IAM と組織のポリシーなど)
- データ セキュリティ (暗号化と鍵管理)
- プライバシー (個人を特定できる情報を処理する戦略など)
- データアクセスと保存に関する地域的な考慮事項 (データ主権)
- 法令遵守、規制遵守
- 適切なデータ ガバナンスを確実化するためのプロジェクト、データセット、テーブル アーキテクチャの設計
- マルチ環境のユースケース (開発環境と本番環境)

1.2 信頼性と確実性を考慮した設計。以下のような点を考察します。

- データの準備とクリーニング (Dataform、Dataflow、Cloud Data Fusion、クエリ生成のための LLM のプロンプト)
- データ パイプラインのモニタリングとオーケストレーション
- 障害復旧とフォールトトレランス
- Atomicity (原子性)、Consistency (一貫性)、Isolation (独立性)、Durability (永続性) (ACID) に対するコンプライアンスと可用性に関連する意思決定
- データの検証

1.3 柔軟性とポータビリティを考慮した設計。以下のような点を考察します。

- アーキテクチャへの現在と将来のビジネス要件のマッピング
- データとアプリケーションのポータビリティを考慮した設計（例: マルチクラウド、データ所在地の要件）
- データのステージング、カタログ化、検出（データガバナンス）

1.4 データ移行の設計。以下のような点を考察します。

- 現在の関係者のニーズ、ユーザー、プロセス、技術の分析と望ましい状態を実現するための計画の策定
- Google Cloud への移行と検証の計画（BigQuery Data Transfer Service、Database Migration Service、Transfer Appliance、Google Cloud ネットワーキング、Datastream など）

セクション 2: データの取り込みと処理（試験内容の 25% 以下）

2.1 データパイプラインの計画。以下のような点を考察します。

- データソースとシンクの定義
- データ変換とオーケストレーションのロジックの定義
- ネットワーキングの基礎
- データ暗号化

2.2 パイプラインの構築。以下のような点を考察します。

- データ クレンジング
- サービスの特定（例: Dataflow、Apache Beam、Dataproc、Cloud Data Fusion、BigQuery、Pub/Sub、Apache Spark、Hadoop エコシステム、Apache Kafka など）
- 変換
 - バッチ
 - ストリーミング（例: ウィンドウ処理、受信遅延データなど）
 - 処理ロジック
 - AI によるデータ拡充
- データの取得とインポート
- 新しいデータソースとの統合

2.3 パイプラインのデプロイと運用化。以下のような点を考察します。

- ジョブの自動化とオーケストレーション (例: Cloud Composer と Workflows など)
- CI / CD (継続的インテグレーションおよび継続的デプロイ)

セクション 3: データの保存 (試験内容の 20% 以下)

3.1 ストレージ システムの選択。以下のような点を考察します。

- データアクセス パターンの分析
- マネージド サービスの選択 (例: BigQuery、BigLake、AlloyDB、Bigtable、Spanner、Cloud SQL、Cloud Storage、Firestore、Memorystore)
- ストレージの費用とパフォーマンスの計画
- データのライフサイクル管理

3.2 データ ウェアハウスを使用するための計画。以下のような点を考察します。

- データモデルの設計
- データ正規化の度合いの決定
- ビジネス要件のマッピング
- データアクセス パターンをサポートするアーキテクチャの定義

3.3 データレイクの使用。以下のような点を考察します。

- レイクの管理 (データの検出、アクセス、費用管理の構成)
- データの処理
- データレイクのモニタリング

3.4 データプラットフォームを考慮した設計。以下のような点を考察します。

- 要件に基づくデータ プラットフォームを Google Cloud のツール (例: Dataplex、Dataplex Catalog、BigQuery、Cloud Storage) で構築する
- 分散データシステム用の連携ガバナンス モデルを構築する

セクション 4: 分析用データの準備と使用(試験内容の 15% 以下)

4.1 可視化用データの準備。以下のような点を考察します。

- ツールへの接続
- フィールドの事前計算
- ビジネス インテリジェンス向けの BigQuery の機能 (例: BI Engine、マテリアライズド ビュー)
- パフォーマンスの悪いクエリのトラブルシューティング
- セキュリティ、データマスキング、Identity and Access Management (IAM)、Cloud Data Loss Prevention (Cloud DLP)

4.2 AI と ML のためのデータの準備。以下のような点を考察します。

- 特徴量エンジニアリング、ML モデルのトレーニングと提供のためのデータ準備 (例: BigQuery ML)
- エンベディングと検索拡張生成 (RAG) のための非構造化データの準備

4.2 データの共有。以下のような点を考察します。

- データ共有のルール定義
- データセットの公開
- レポートと視覚化の公開
- BigQuery Sharing (Analytics Hub)

セクション 5: データ ワークロードの管理と自動化(試験内容の 18% 以下)

5.1 リソースの最適化。以下のような点を考察します。

- データに関連するビジネスニーズに従って費用を最小限に抑える
- ビジネス クリティカルなデータプロセスにとって十分なリソースを使用できるようにする
- 永続的なデータクラスタとジョブベースのデータクラスタ (例: Dataproc) のどちらを使用するかを決定する

5.2 自動化と反復性の設計。以下のような点を考察します。

- Cloud Composer の有向非巡回グラフ (DAG) の作成
- ジョブを反復可能な方法でスケジューリングおよびオーケストレーションする

5.3 ビジネス要件に基づくワークロードの最適化。以下のような点を考察します。

- 容量管理 (BigQuery エディションと予約)
- インタラクティブ方式またはバッチ方式のクエリジョブ

5.4 プロセスのモニタリングとトラブルシューティング。以下のような点を考察します。

- データプロセスのオブザーバビリティ (例: Cloud Monitoring、Cloud Logging、BigQuery 管理パネル)
- 計画された使用量のモニタリング
- エラー メッセージ、請求に関する問題、割り当てのトラブルシューティング
- ジョブ、クエリ、コンピューティング容量 (予約) などのワークロードの管理

5.5 障害への意識の持続と影響の軽減。以下のような点を考察します。

- フォールトトレランスを念頭に置いてシステムを設計し、再開を管理する
- 複数のリージョンまたはゾーンでジョブを実行する
- データの破損や欠落に備える
- データのレプリケーションとフェイルオーバー (例: Cloud SQL、Redis クラスター)