【データ分析基盤】

データ分析を行うために、継続的にデータを収集して蓄積する基盤が必要となります。そのため、下図のとおり、生成、収集、蓄積、活用を行います。下図に出てくる用語は、データベーススペシャリスト試験の必須用語です。

生成段階：データ分析のためにデータを生成する段階である。つまり、既存データを活用して分析するのではなく、分析を前提としてデータを生成している。生成されたデータは、**データソース**と呼ばれ、基幹システムのDBのデータ、アクセスログ、センサーデータなどがある。

↓

収集段階：生成されたデータを収集する段階であり、**データレイク**に保管しておく。
※データレイク：必要な時に加工するために、収集した多種多様なデータを発生したままの形で保管する場所

収集段階から蓄積段階への過程
ETL（Extract Transform Load：抽出、変換、書き出し）：以下の一連の処理を行うツール
①**データレイク**に格納されたデータを抽出
②分析に適するように変換（**データクレンジング**）
③**データウェアハウス**への書き出し
※データクレンジング：データの重複、誤記及び表記の揺れなどに対して、削除、修正及び正規化などを行うことによってデータの品質を高めること。つまり、データ属性やコード体系を統一する処理
※データウェアハウス（分析用のデータベース）：分析に適するように構造化（整形、加工など）されたDB

蓄積段階：**データウェアハウス（分析用のデータベース）**に蓄積する段階である。本DBは、複数の属性項目を軸（次元）にして種々の分析を行うことができる多次元DBである。本DBは、蓄積専用（追加はあるが更新及び削除はない）なので、更新時異常は発生しない。よって、本DBは正規化不要の**ファクトテーブル（FT）**に格納される。その**FT**が参照するのが**ディメンション（次元）テーブル（DT）**であり、そのE-R図は**FT**を中心に**DT**がスター型に配置され、**スタースキーマ**と呼ばれる。

↓

活用段階：データを分析して活用する段階である。**データウェアハウス**や**データマート**からデータを取り出し、**OLAP**を行う。
※データマート：データウェアハウスに格納されたデータから、特定の用途や部門用に切り出したDB
※OLAP(OnLine Analytical Processing：オンライン分析処理)：多次元DBを用いた分析技術であり、例えば、3つの属性で分析する場合には、ルービックキューブの様な立方体構造になる。そして、分析軸を切り替えるために回転させる**ダイズ(ダイジング)**、分析軸の分析単位を小さくする**ドリルダウン(ロールダウン)**、その逆の**ドリルアップ(ロールアップ)**がある。
なお、活用段階として、データマイニングもある。データマイニングとは、データレイクなどにある大量の整理されていない生データを分析し、単なる検索だけではわからない隠れた規則や相関関係を見つけ出す（マイニング）こと。つまり、OLAPとは別系統の活用方法である。また、データディクショナリはDBを管理するための様々な情報のことなのでここでは無関係。