【ビックデータ】

ビックデータの定義

巨大な複雑なデータの集合であり、**Variety（データの種類が多様）**、**Volume（データの量が膨大）**、**Velocity（データの発生速度又は発生頻度が大きい）**という３つのVの特徴を持ちます。

NoSQL

ビックデータの中心的な技術基盤である**NoSQL（Not Only SQL）**は、RDBMS以外のDBMSであり様々な種類がありますが、共通して言えることは、**BASE特性**を持っているということです。

BASE特性

RDBMSにおけるACID特性の１つである一貫性は、**直列化可能性**（トランザクションを複数並行で実行しても、完全独立して順番に実行しても、結果は同じこと）を確保することですが、ロックによる排他制御（同時実行制御）を行うため、スループットは格段に悪くなります。

よって、ビックデータを高速に処理するために、NoSQLは、リアルタイムでの一貫性をあきらめ、一定の時間が経過すれば最終的に一貫性が確保できる結果整合性を採用しています。この結果整合性を保証するものが、RDBMSのACID特性に対応するものである下表の**BASE特性**です。

BASE特性	左記の意味
BA(Basically Available)	可用性が高く、基本的に利用可能
S(Soft-State)	厳密な状態を要求しない。厳密ではない状態遷移。
E(Eventually Consistent)	最終的には一貫性が保たれる。結果整合性

CEP（Complex Event Processing：複合イベント処理）

ビックデータの処理において、連続して発生する大量のデータを瞬時に処理するために、大量のデータをメモリ上に展開し、予め設定した条件を満たす場合だけ、処理を実行する方法です。

ビックデータのためのミドルウェア

様々な分散処理を組み合わせるためのミドルウェアとして、以下のものがあります。

①**Hadoop：**大規模なデータを分散処理するためのソフトウエアライブラリ
②**Apache Spark：**RDD（分散共有メモリの仕組み）を使用したフレームワーク（ベースプログラム）

その他

JupyterLab：Webブラウザ上で、Python等でデータ解析や機械学習を行う際に便利な対話式の統合開発環境であり、実行結果を見ながら、データ処理ができるのでプログラムのテスト版を作成することに適しています。JupyterNotebookの後継です。