**Définition Big Data :**

L’expression « Big Data désigne la masse hétérogène des données numériques produites par les entreprises et les particuliers dont les caractéristiques (très grand volume, diversité de forme, vitesse de traitement) requièrent des outils d’analyse informatiques spécifiques.

**Avant le Big Data « Calcul répartie et HPC » :**

Effort concentré sur les problèmes « calcul intensive »

Le réseau n’étant pas si performant, on évitait de faire des transferts de données

* **Bases de données :**

Le modèle relationnel s’est imposé dans les années 80 u Données structurées (tableaux), formes normales

Très forte optimisation des SGBD u Bases de données fédérées : trop difficile

* **Data mining**

Méthodes statistiques pour l’extraction des connaissances

D’abord un modèle, qui sera ensuite validé

Echantillonnage des données pour tenir dans la mémoire

**Les 3V :**

* **Volume (Volume) :**

Volumes de données croissants de tous types, qui se comptent en téraoctets ou même en pétaoctets

* **Variété (Variety) :**

Gérer la complexité de plusieurs types de données et de schémas structurés ou non structurés u texte, données de capteurs, son, vidéo, logs, ...

* **Vitesse (Velocity) :**

Parfois, les données doivent être saisies et traitées au fur et à mesure de leur collection

**D’autre exemple de V :**

* **Véracité (Veracity) :**

La qualité et la précision des données sont aussi importantes

* **Valeur (Value) :**

La valeur ajoutée des données ou des informations extraites u Sans une réelle valeur, ce n’est qu’un gaspillage de ressources

**Les outils big data :**

* **Hadoop** : Framework java libre permettant de crées des applications distribuées et scalables. Basé sur un système de fichiers distribués (HDFS). Conçu pour stocker de très grand volume de données sur un grand nombre de machines. intègre le patron d’architecture MapReduce dans lesquels sont effectués les calculs parallèles

Distributions : Cloudera, haltonwork, mapR technologies

* **HDFS :** est un système de fichiers distribué. C’est à dire :

Les fichiers et dossiers sont organisés en arbre (comme Unix)

Ces fichiers sont stockés sur un grand nombre de machines de manière à rendre invisible la position exacte d’un fichier. L’accès est transparent, quelle que soient les machines qui contiennent les fichiers.

Les fichiers sont copiés en plusieurs exemplaires pour la fiabilité et permettre des accès simultanés multiples HDFS permet de voir tous les dossiers et fichiers de ces milliers de machine

* **NoSQL :**(Not only SQL) système de gestion de bases de données à faible structuration relationnelle basé sur un modèle clé valeur, facilement scalable en multipliant les serveurs

Produits : Cassandra(Twitter), BigTable (Google), Hbase(Facebook), MongoDB, Neo4