# **Portail de rapport Big Data**

# Introduction

Chaque jour, 2,5 de téraoctets de données sont générées dans le monde. D’ici 2020, il est prévu que la taille des données se multiplie de 50 fois. Google reçoit 40 000 requêtes toutes les secondes, 72 vidéos sont mises en lignes chaque minute sur YouTube et 217 nouveaux utilisateurs de Smartphone sont comptés toutes les minutes (Miranda, 2013). Aujourd’hui l’information nous provient de toute part : des capteurs de géolocalisation, des données émanant des smartphones (logs de connexion, appels…), des données postées sur les réseaux sociaux, des vidéos et des images digitales, des transactions des clients, des capteurs de formes ou de mouvement des objets connectés… Le développement et l’accès à ces données représentent ce qui est appelé le phénomène « Big Data ». Ce phénomène impacte en particulier les entreprises qui sont amenées à manipuler des téraoctets voire des penta octets de données nécessitant une infrastructure spécifique pour leur création, leur stockage, leur traitement, leur analyse et leur récupération. En d’autres termes, il s’agit du développement en temps réel d’une masse de données volumineuse qui dépasse la capacité des outils de traitement et d’analyse traditionnels (bases de données relationnelles, requêtes SQL …). Le Big Data est une évolution de la Business Intelligence (BI). Alors que la BI traditionnelle repose sur des entrepôts de données (datawarehouses) limités en taille (quelques téraoctets), gérant difficilement des données non structurées et les analyses en temps réel, l’ère du Big Data ouvre une nouvelle période technologique offrant des architectures et infrastructures évoluées permettant des analyses sophistiquées en prenant en compte ces nouvelles données intégrées à l’écosystème de l’entreprise (Chen et al., 2012).

En effet, le « phénomène Big Data » pour les entreprises recouvre ainsi deux réalités : d’une part cette explosion des données de façon continue, d’autre part la capacité technologique de traiter et d’analyser cette grande masse de données pour en tirer un profit. Grâce au Big data, les entreprises peuvent désormais gérer et traiter des données massives pour en extraire de la valeur, décider et agir en temps réel. L’enjeu central pour elle est bien « d’améliorer l’efficacité des prises de décision par l’exploitation d’informations protéiformes » (Cigref, 2013). Elles deviennent ainsi plus réactives et plus compétitives. De fait la capacité à traiter de grande masse de données, à tisser des liens et corrélations entre des informations hétérogènes et à faire communiquer entre elles des données structurées ou non, ouvre la voie à des traitements probabilistes qui permettent l’amélioration des opérations managériales et la conception de produits et services innovants pour les clients.

Ainsi, ce présent mémoire qui est l’illustration du travail effectué dans ce stage, relate les différentes phases ainsi que les moyens auxquels nous avons eu recours pour atteindre les objectifs fixés. Ce rapport est articulé autour de cinq chapitres

* Un premier chapitre « PRESENTATION GENERALE » dans lequel nous présentons la structure d’accueil, le contexte, la problématique ainsi que les objectifs du sujet.
* Un deuxième chapitre « GENERALITES » définissant les différents concepts généraux apparaissant dans notre architecture ainsi que les outils utilisés dans la réalisation du sujet.
* Un troisième chapitre « SPECIFICATION FONCTIONNELLE » dans lequel nous présentons les fonctionnalités du système, leur cas d’utilisation ainsi que les règles métiers qui sont associés. Ce chapitre constitue la colonne vertébrale de notre sujet.
* Un quatrième chapitre « CONCEPTION DU SYSTEME » qui présente notre travail d’analyse et de conception à travers des spécifications des modèles nécessaires à chaque étape. Il présente également les différentes architectures du système.
* Un cinquième chapitre « REALISATION DE LA SOLUTION » qui présente les outils que nous avons utilisés pour la réalisation de notre **portail de rapport big data.**

# Présentation du sujet

Dans cette section, nous présentons explicitement le sujet de notre mémoire s’intitulant « **Portail de rapport Big Data** ». Il sera question de présenter d’abord le contexte du projet ensuite énumérer la problématique et enfin exposer les différents objectifs fixés.

## Contexte

Chaque jour, nous générons 2,5 trillions d’octets de données. A tel point que 90% des données dans le monde ont été créées au cours des deux dernières années seulement. Ces données proviennent de partout : de capteurs utilisés pour collecter les informations climatiques, de messages sur les sites de médias sociaux, d'images numériques et de vidéos publiées en ligne, d'enregistrements transactionnels d'achats en ligne et de signaux GPS de téléphones mobiles, pour ne citer que quelques sources. Ces données sont appelées **Big Data ou volumes massifs de données**. Le Big Data contraint à de nouvelles manières de voir et analyser le monde. De nouveaux ordres de grandeur concernent la capture, le stockage, la recherche, le partage, l'analyse et la visualisation des données.

Pour caractériser le Big Data, on parle de la règle des 3V devenue par extension 5V qui fait référence à **Volume**, **Variété**, **Vitesse**, **Véracité** et **Valeur**. S’agissant de la notion de **volume**, on parle de Big Data au-delà de 100 To (téraoctets) de données structurées, semi-structurées ou non structurées. La **variété** nous rappelle que la diversité des données est fondamentale pour le Big Data. La **vitesse** de traitement est enfin cruciale et propre à la notion de Big Data. La véracité qui suppose à vérifier systématiquement la validité des données. Ces dimensions supposent que l’on soit capable de stocker et d’exploiter des volumes énormes de données, d’accéder à de multiples sources par un réseau performant, de traiter dans des temps toujours plus courts des volumes considérables de données.

Face à cette situation, le groupe Atos leader international dans le secteur informatique, veut mettre en place un **portail de rapport Big Data** au niveau de ses puissants serveurs afin d’améliorer de façon significative le processus de prise de décision et de bien exploiter ses données qui représentent une source précieuse d’informations. Par ailleurs avec toute cette masse de données on évoque de nouvelles possibilités d’exploration pour pouvoir en apprendre un peu plus sur ces données. C’est dans ce cadre que le groupe Atos a proposé un projet de mettre en place un **portail de rapport Big Data**.

## Problématique

Aujourd’hui, ce sont plus de 2,5 quintillions octets, soit 1030 données qui sont créés par jour. L’hyper-connectivité des individus, la multiplication des terminaux (ordinateurs, téléphones, tablettes), et des objets connectés, ont fait se multiplier les données transitant sur le réseau. Ainsi, on estime que le nombre de données produites par les internautes doublerait tous les 18 à 24 mois. Et la tendance ne risque pas de s’atténuer car avec les objets connectés, c’est encore plus de données qui vont être disponibles. Demain, nous posséderons en moyenne 8 objets connectés à titre personnel.

Face à cette croissance exponentielle du volume des données et la diversité de leurs formats, les organisations sont confrontées à des problèmes tels que la collecte, le stockage, l’analyse et l’exploitation de grands volumes de données pour créer de la valeur ajoutée. D’ailleurs, les bases de données traditionnelles ne suffisent plus pour gérer les données. Il faut imaginer de nouvelles architectures, de nouvelles technologies et des nouveaux algorithmes pour stocker, traiter et transporter ces masses considérables de données.

Mais comment gérer des milliers ou des millions de données ? D'abord il faut centraliser toutes ces données dans un ou plusieurs serveurs.

Donc, tout l’enjeu pour les entreprises et les administrations du XXIème siècle consiste à stocker à long terme ces volumes de données ainsi de les traiter et de les analyser pour ne pas passer à côté d’informations précieuses cachées dans la masse.

C’est dans ce cadre que le Groupe AtoS leader international dans le secteur informatique, veut étendre son expertise en service Big Data, en fournissant les bénifices d’un portail de rapport Big Data. Les besoins exprimés et les contraintes à respecter en vue de mettre en place un tel système conduit à répondre à la problématique suivante :

Comment centraliser et fournir un accès facile à ces données pour en tirer profit ?

## Objectifs

### Objectif général

L'objectif principal de cet étude est de mettre en place un outil permettant de pouvoir collecter ces données, les traitées et les observées via une plateforme web de visualisation. Cet objectif permettra la simplification d’étude et aidera à prendre des décisions.

### Objectifs spécifiques

Pour atteindre cet objectif général décrit plus haut, nous allons de façon plus concrète en définir de plus spécifiques qui nous permettront d'atteindre le premier nommé.

IL s’agit de:

* Centraliser la collecte de toutes les données dans un Data Lake;
* Fournir un accès facile aux données en utilisant le graphique
* Exécuter des recherches rapides et avancées
* Visualiser les données d’un data Lake (Lac de données) en utilisant des modules de visualisation
* Définir une utilisation interactive pour la manipulation des données et la visualisation des données
* Créer des Tableaux de bords.