**Définir le BigData**

==> Pourquoi le BigDate ... Définir les 5V du Big Data ...

<https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know/>

<https://www.definitions-marketing.com/definition/5v-du-big-data/>

==> Définir le NoSQL :

- données structurées / semi / non

- Clé/Valeur, colonnes, documents, graphes

- théorème de CAP de Brewer

NoSQL/  
  
Stockage en JSON? Source Amazon (<https://aws.amazon.com/fr/nosql/>) → Orienter document si on regarde le second lien.

Plusieurs type de NOSQL?? (<https://www.grafikart.fr/blog/sql-nosql>) ah bah cétais dans le hat donc oui plusieurs type qui sont décris là.

On a plusieurs type de base de données NOSQl (orienté ligne, orienter document, etc)  
  
Le NoSQL est moins “fiable” pour plus de rapidité, certains points des BD relationnels ont été tronqués.

Théorème de CAP de Brewer

<http://davidmasclet.gisgraphy.com/post/2010/06/09/10-minutes-pour-comprendre...NoSQL>

==> Définir le DataLake

C’est globalement un dataWarehouse mais sans les contraintes de formes. On peut y mettre un peu tout le bordel qu’on veut dedans. Mais ça pose des problèmes de fiabilité. Pas de soucis de type de donnée.

**Présentation des outils**

==> UN EXEMPLE d'architecture BigData à étudier :

- Hadoop HDFS : intégration et stockage des fichiers et données brutes (non transformées) dans HDFS

- Yarn : l’exécution simultanée de multiples applications + mixer des workloads en batch, interactif et en temps réel

- MapReduce / **Spark** : pour réaliser les traitements d’intégration et de calculs (entre HBase et Hive)

- HBase (BD NoSQL colonnes) / Phoenix (permet de faire du OLTP/ACID) : stockage data

- Hive : reporting / analyse de données / SQL (permet de faire du SQL sur HBase)

Pour compléter, voici une vidéo que j'avais trouvé très instructive quand j'avais étudié le sujet :

<https://www.youtube.com/watch?v=CNUlNk0KtKs>

==> Présenter les distributeurs Hadoop sur la place (Cloudera, MAPR, Hortonworks,...)

<https://www.lebigdata.fr/hadoop-top-des-distributeurs>

==> Présenter le mode de fonctionnement de Hadoop HDFS (NameNode/Job Tracker, DataNode/Task Tracker, ...)

**Démonstration**

C'est là le gros défis de ce projet.

En fonction de l'architecture technique que vous allez présenter, on vous invite à nous présenter concrètement :

-> comment s'installe l'outil (DFS, DRILL, ...)

-> à quoi ressemble l'interface de développement

-> à quoi ressemble le développement

<https://www.it-connect.fr/cest-quoi-le-dfs-windows-server/>

<https://drill.apache.org/docs/why-drill/>

DRIll : moteur SQL facile à installer avec du "real SQL" s'exécutant sur toutes les plateformes

Pour commencer, je vous invite tous à passer du temps sur internet à lire de la documentation sur le BigData et les architectures techniques possibles.

L'idée est de vous familiariser avec le sujet.

L'échéance étant à fin juin, vous devez au plus vite avoir une vue d'ensemble des outils disponibles sur le marché et en choisir un pour votre présentation.