Big Data – Thierry Boulanger – Capgemini 2023

Big Data

## Introduction

Une définition :

« On parle depuis quelques années du phénomène de big data , que l'on traduit souvent par « données massives ». Avec le développement des nouvelles technologies, d'internet et des réseaux sociaux ces vingt dernières années, la production de données numériques a été de plus en plus nombreuse : textes, photos, vidéos, etc. »

Issu du site de CNIL : https://www.cnil.fr/fr/definition/big-data#:~:text=On%20parle%20depuis%20quelques%20ann%C3%A9es,%2C%20photos%2C%20vid%C3%A9os%2C%20etc.

## La frontière entre Data et BigData

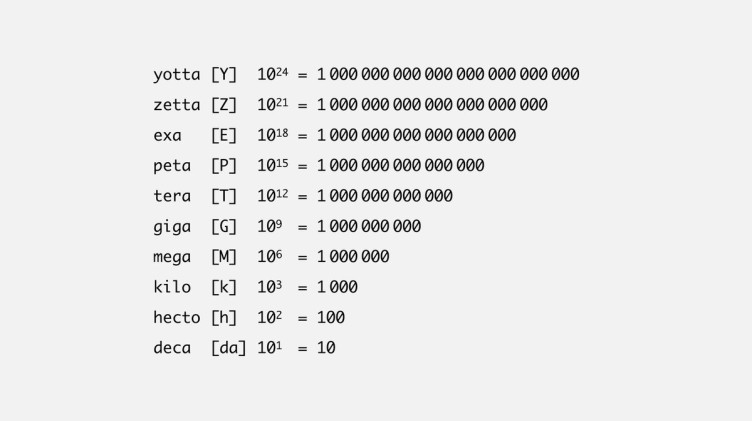
Ce qui n’est pas big data

* Ce qu l’on peut ouvrir avec Excel
* Les capteurs utilisé dans l’IOT
* Les données publiques disponible en téléchargement

Ce qui est Big Data

* Les Logs des grandes enseigne de la distribution
* Le traffic des gros sites web
* Les données de localisation GSM s’un opérateur de téléphone sur une journée

## Les chiffres du Big Data



1 mégaoctet (Mo) = 1 million d'octets

2 Mo = une photo numérique en haute résolution

10 Mo = 1 minute de son haute fidélité

100 Mo = le contenu d'une pile de livres de 1 mètre de haut

1 gigaoctet (Go) = 1 milliard d'octets

20 Go = un enregistrement de l'oeuvre complète de Beethoven

500 Go = le plus gros site FTP

1 téraoctet (To) = 1000 milliards d'octets

2 To = tous les ouvrages d'une bibliothèque universitaire

10 To = tous les imprimés de la bibliothèque du Congrès américain

400 To = la base de données du National Climatic Data Center

1 petaoctet (Po) = 1 million de milliards d'octets

2 Po = les fonds de toutes les bibliothèques universitaires des Etats-Unis

8 Po = les données disponibles sur le Web

20 Po = l'ensemble des disques durs produits en 1995

200 Po = l'ensemble des bandes magnétiques produites en 1995

1 exaoctet (Eo) = 1 milliard de milliards d'octets

2 Eo = le volume annuel des informations générées dans le monde

5 Eo = tous les mots prononcés depuis le début de l'humanité

## Historique du Big Data

Avant 2000, on ne parle pas encore de big data. En ce qui concerne les data on cherche l’ETL => Extract Transform Load

Stockage ⇒ Analyse ⇒ Restitution

Un engouement pour le e-commerce donne naissance au WebAnalytics

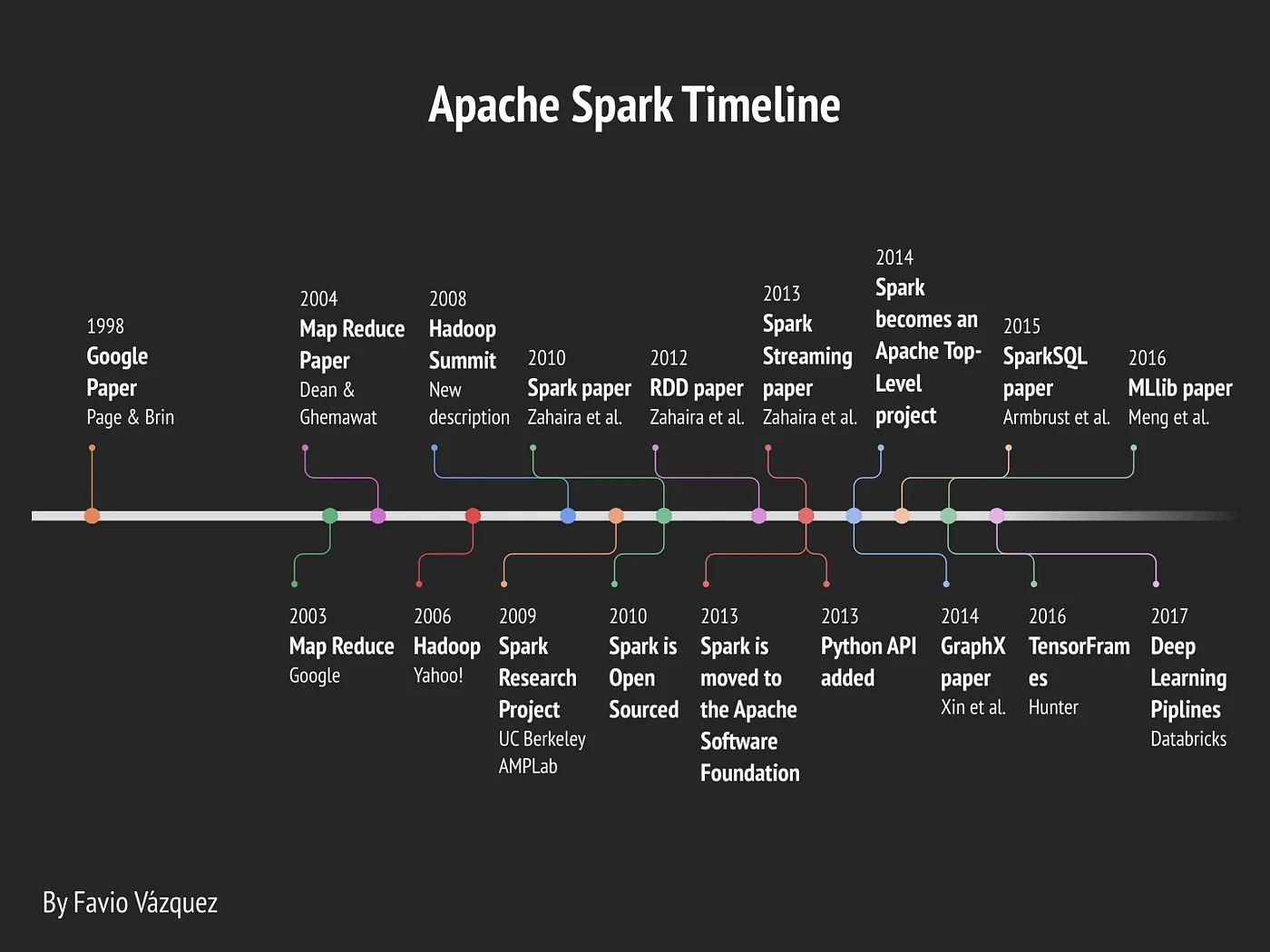
Milieu des années 2000

BI 1.0

BI 2.0

Apparition de la Business Intelligence

2004 : Google dégaine Map Reduce

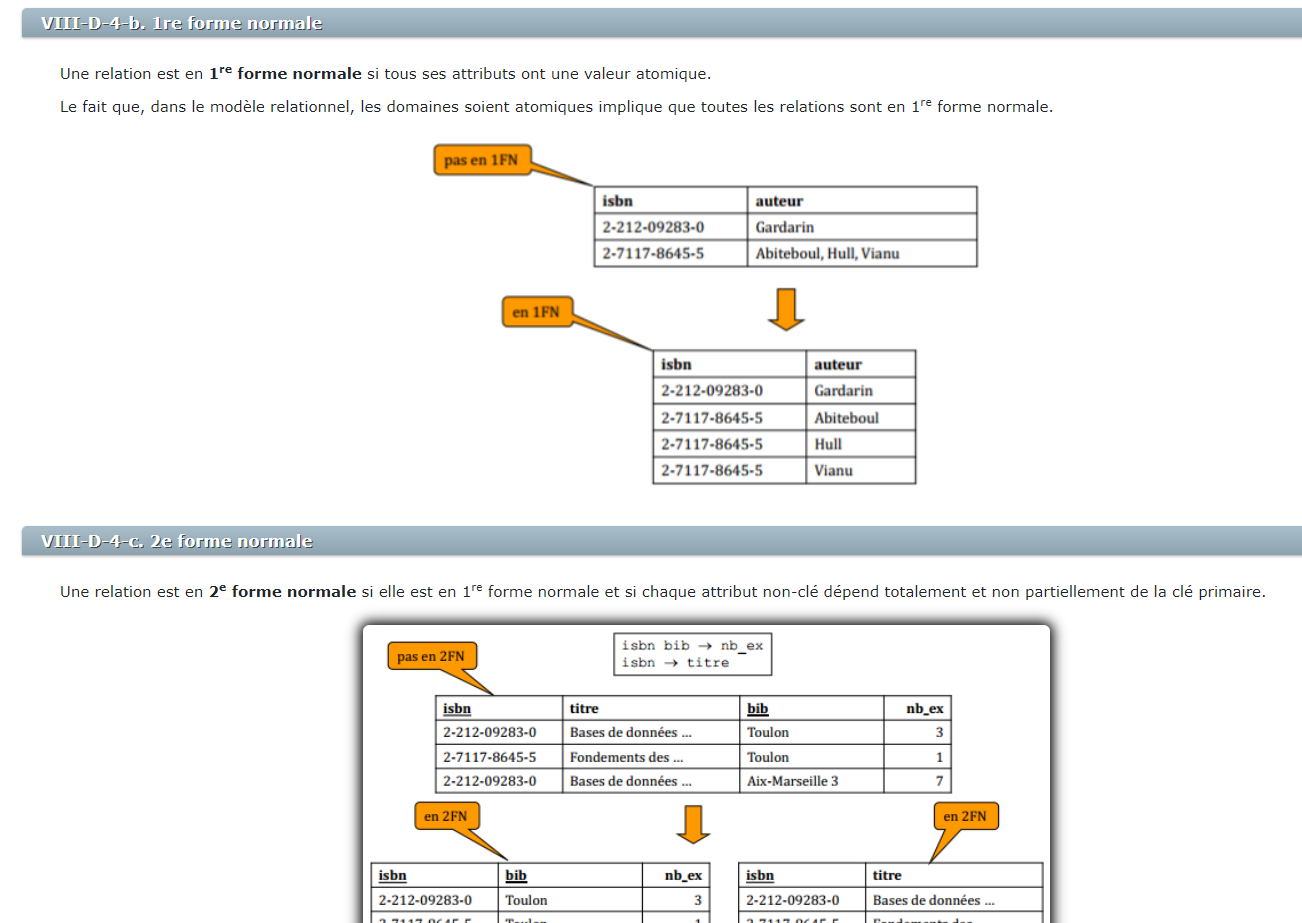


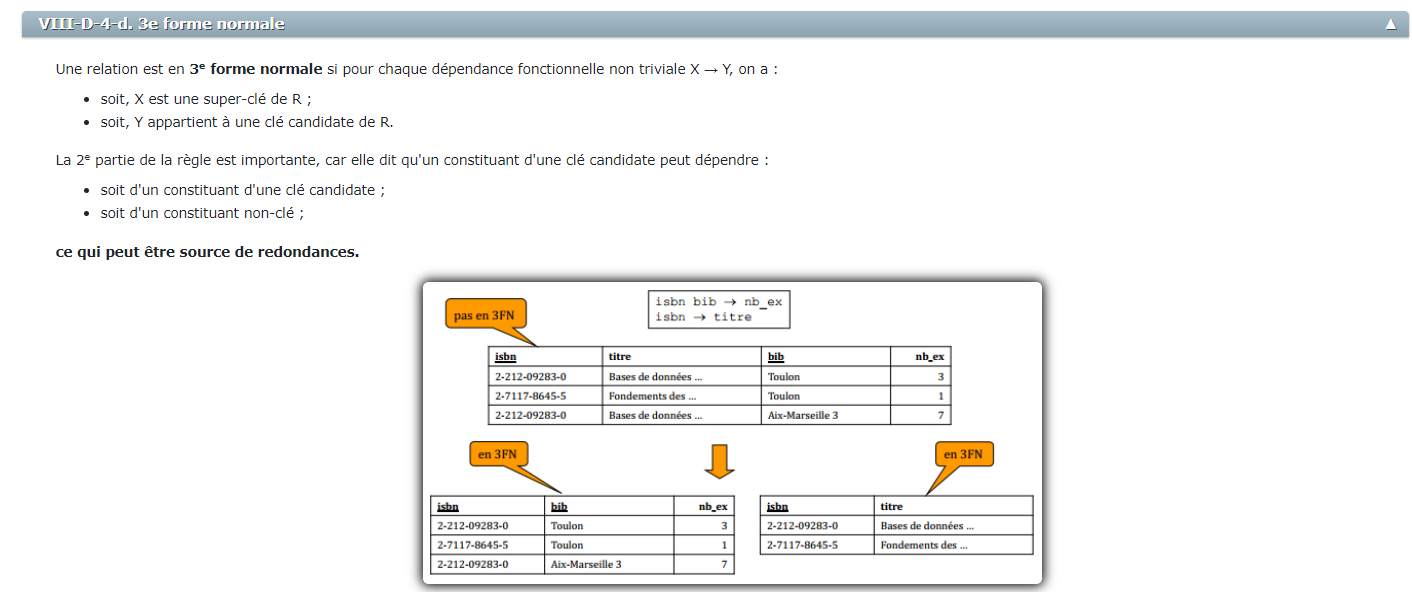
## 

## La data et ses formes

### Toujours optimiser

### Rappel sur les formes normales





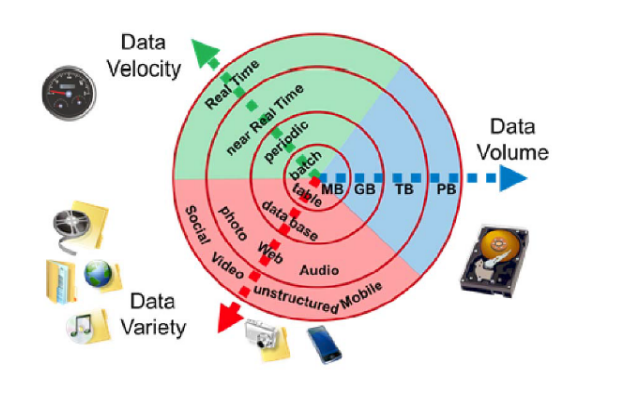
## 

Calcul avec index

<https://www.sqlfacile.com/apprendre_bases_de_donnees/bench_mysql>

### Lorsque que cela n’est plus possible => Penser les data autrement

## Les 3 V



### Volume

Fait référence à l'ampleur des données

### Vélocité

Il s'agit de la vitesse à laquelle les données sont générées, ce qui peut être fait en

les manières suivantes :

• Le temps réel correspond au moment où les données sont générées immédiatement, telles que le streaming, les systèmes radar, les systèmes de service client et les guichets automatiques bancaires.

• Le quasi-temps réel, c'est lorsque la vitesse est importante mais que les données ne sont pas générées immédiatement, principalement pour la production d'intelligence opérationnelle, qui est une combinaison de traitement de données et de CEP (Complete Event Processing) combinant les données de

plusieurs sources afin de détecter des modèles tels que : le traitement des données des capteurs, la surveillance des systèmes informatiques, le traitement des transactions financières.

• Le batch c’est lorsque les données sont générées dans des délais tels que la paie, la facturation, l'analyse des données à partir des données opérationnelles, des données historiques et archivées, des données des médias sociaux, des services

données, etc…

### Variété

Fait référence aux différentes formes de données de plus en plus nombreuses telles que le texte, les images, la voix,

et géospatiale.

et des données semi-structurées.

• Les données structurées font référence aux données stockées dans des bases de données de manière ordonnée.

données dans les catalogues de bibliothèques (telles que : date, auteur, lieu, sujet, etc.) et économiques

les données (telles que : PIB, PPI, ASX) sont considérées comme structurées.

• Les données non structurées font référence à toutes les données dont la forme/structure est inconnue.

forme de médias (mp3, photos numériques, audio, vidéo), texte (traitement de texte, tableurs, présentations) et médias sociaux (données de Facebook, twitter, LinkedIn) sont

considéré comme non structuré.

• Les données semi-structurées font référence à une forme de données structurées qui n'est pas conforme

la structure formelle des modèles de données comme dans le SGBDR ou d'autres tables de données, mais contient

balises ou autres marqueurs qui séparent les éléments sémantiques et appliquent des hiérarchies de

enregistrements et champs dans les données.

#### Retour d’expérience

**Chez L'oréal**

Sur une des applications la solution technique a été de créer autant de répertoires dans le système de fichiers que d’utilisateur. En plus de cela une base de données stocke également des données. Dans certaines tables, des données sont stockées sous la forme de chaines de caractères où l’on retrouve des séparateurs.

=> des données structurées

=> des données non structurées (fichier textes, résultat de machine learning en R, fichiers xlsx, xslxm…)

=> de données semi structurées

**Chez Mentor**

On fait du crédit et un crédit ça veut dire beaucoup de documents. Fiche de paie; relevé bancaire…

Une demande de crédit : 20 pdf => 20 miniatures X 2 (small et big) => 40 png  
  
200 demandes par jours => 200X40 fichiers png = 8000

Sur 30 jours => 24000

Sur 6 mois => 144 000

Sur 1 an => 288 000

Sur 5 ans => 1 440 000 fichiers

sur 5 ans en mo = 7200000 mo => 72000 go

5 marques blanches = > 72000\*5

Pour éviter la fraude des algorithmes sont utilisés pour repérer les incohérences dans les fiches de paie et les relevés de comptes.

Algorithme de Cluster comme k-means

## 

## Hadoop

selon Wikipedia

“Hadoop est un framework libre et open source écrit en Java destiné à faciliter la création d'applications distribuées (au niveau du stockage des données et de leur traitement) et échelonnables (scalables) permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données. Ainsi chaque nœud est constitué de machines standard regroupées en grappe. Tous les modules de Hadoop sont conçus selon l'idée que les pannes matérielles sont fréquentes et qu'en conséquence elles doivent être gérées automatiquement par le framework.”

https://fr.wikipedia.org/wiki/Hadoop

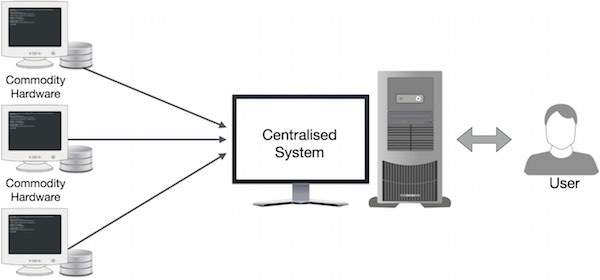
Hadoop a été inspiré par la publication de MapReduce, GoogleFS et BigTable de Google. Hadoop a été créé par Doug Cutting et fait partie des projets de la fondation logicielle Apache depuis 2009.

### MapReduce

Pourquoi MapReduce ?

Les systèmes d'entreprise traditionnels ont normalement un serveur centralisé pour stocker et traiter les données. , le système centralisé crée trop de goulot d'étranglement lors du traitement simultané de plusieurs fichiers.

Google a résolu ce problème de goulot d'étranglement à l'aide d'un algorithme appelé MapReduce. MapReduce divise une tâche en petites parties et les affecte à plusieurs ordinateurs. Plus tard, les résultats sont collectés en un seul endroit et intégrés pour former l'ensemble de données de résultats.



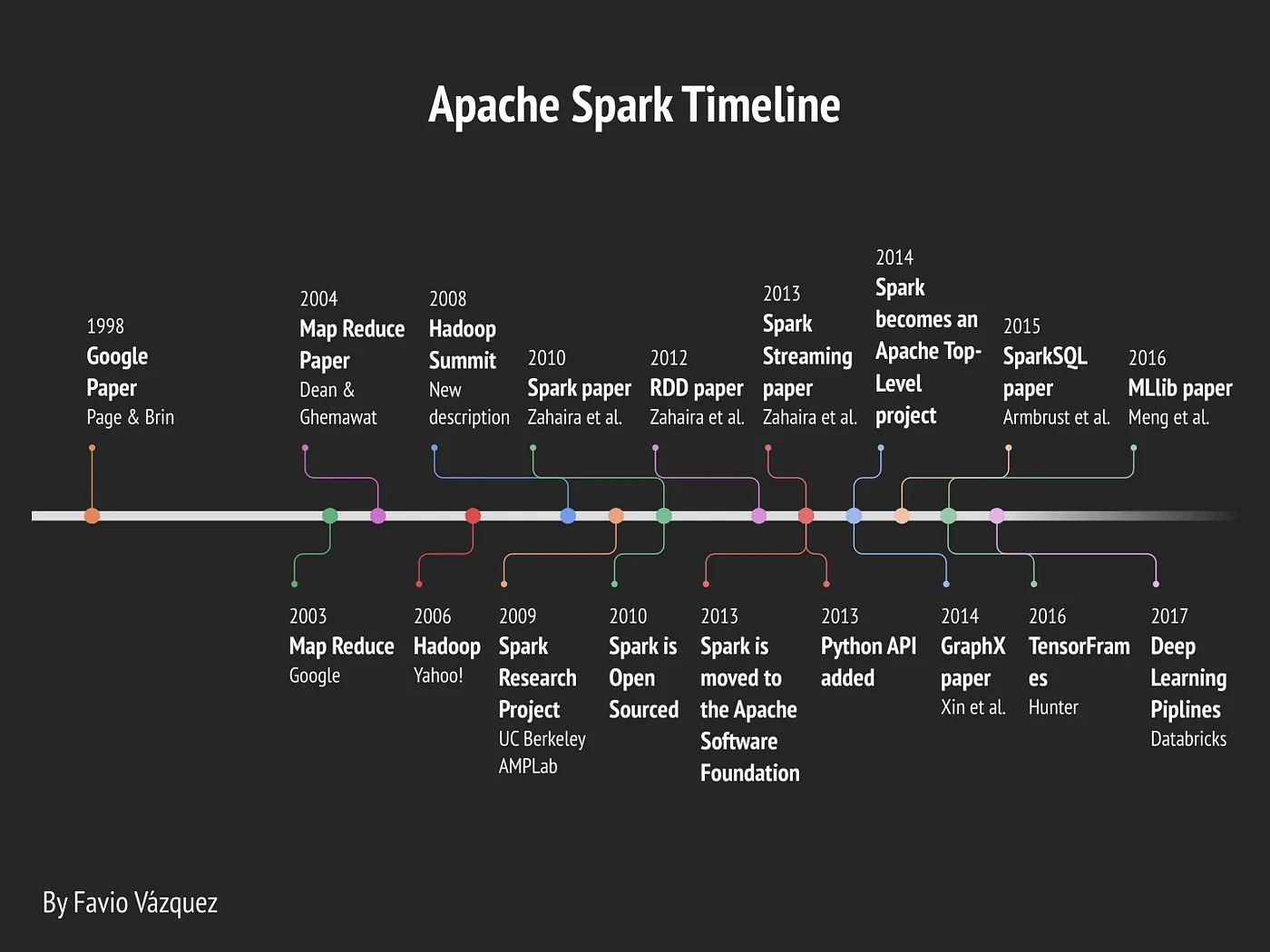
## Spark

<https://towardsdatascience.com/deep-learning-with-apache-spark-part-2-2a2938a36d35>

<https://towardsdatascience.com/beginners-guide-to-pyspark-bbe3b553b79f>

Selon Wikipedia

C’est un cadre applicatif de traitements des [mégadonnées](https://fr.wikipedia.org/wiki/Mégadonnées) (*big data*) pour effectuer des analyses complexes à grande échelle.



## Sources

<https://itandsi.files.wordpress.com/2015/04/les-bases-du-big-data.pdf>

<http://www.metz.supelec.fr/metz/personnel/vialle/course/BigData-2A-CS/poly-pdf/Poly-chap1.pdf>

<https://cours.etsmtl.ca/log660/public_docs/acetates/BigData_Technologies_PL.pdf>

<https://blog.developpez.com/sqlpro/p10070/langage-sql-norme/base_de_donnees_et_performances_petites>

<https://www.techtarget.com/searchdatamanagement/feature/15-big-data-tools-and-technologies-to-know-about>

<https://sgbd.developpez.com/tutoriels/cours-complet-bdd-sql/?page=normalisation>