WENVISION

# TechWaves Data :

## Les bons choix de technologies pour valoriser votre patrimoine de données

# Sommaire



Les TechWaves Data
TechWaves

WENVISION
TechWaves Data

# 1 Qu’est-ce qu’une TechWave ?

WENVISION
TechWaves Data

# Les TechWaves
## pour surfer sur les bonnes vagues technologiques


Une vague technologique chasse l'autre. Certaines se contentent de produire un peu d'écume, tandis que d'autres sont des lames de fond. Il faut savoir laquelle surfer, et pour quelles raisons : veut-on prendre chaque vague, pour le plaisir de se lancer dans un nouveau langage, une nouvelle architecture ? souhaite-t-on au contraire attendre la bonne vague, celle qui nous permettra d'innover tout en nous portant longtemps ?

Chacun aura sa propre conception de "la bonne vague", en fonction de son contexte, de ses compétences, de ses perspectives. Le plus important étant de pouvoir identifier et classer les technologies émergentes, de façon à se positionner correctement.

Chez SFEIR, c'est un exercice que nous pratiquons depuis plusieurs années, sous le nom de "TechWaves". Cela nous a permis de nous positionner sur des technologies innovantes à fort potentiel et a contribué au succès que nous connaissons. Aujourd'hui, notre méthodologie est bien rodée et nous considérons que la partager avec vous s'inscrit pleinement dans la vocation d'Envision by SFEIR, l'entité conseil du groupe SFEIR.



## Identifier le potentiel des technologies sur le marché[1]

(1) Éviter de créer de la dette technologique



→ Les cabinets d'analystes nous ont habitués à classer le potentiel des grandes tendances technologiques ainsi que les entreprises en fonction de leur rapidité d'adoption de ces tendances. Toutefois, il s'agit d'une approche macroscopique, qui ne se préoccupe pas des choix technologiques au sein d'une tendance.

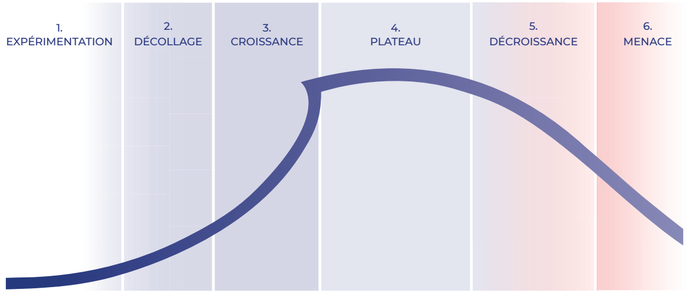
Quand on parle de diffusion d'information au fil de l'eau, vaut-il mieux miser sur Kafka, sur Confluent ou sur les services managés des fournisseurs de Cloud ? Est-ce que le projet Apache Pulsar va changer la donne ? De même, pour des applications transactionnelles classiques, est-il prudent de rester sur des offres traditionnelles d'Oracle, Microsoft ou IBM ?

Est-il pertinent, pour un datawarehouse, d'aller vers une offre comme Panoply ? Existe-t-il des compétences sur le marché ? Y en aura-t-il dans quelques années pour assurer l'évolution et la pérennité de l'application ? Ce sont des questions extrêmement importantes pour ne pas se retrouver handicapé par un choix malheureux.

→ Nos TechWaves s'attaquent à ces questions en positionnant les technologies elles-mêmes (par exemple MongoDB), plutôt que les grandes tendances (par exemple, les bases NoSQL), sur une courbe d'adoption, depuis leur émergence jusqu'à leur décroissance - c'est-à-dire le moment où continuer de les utiliser crée de la dette technique.

→ Nous publierons des TechWaves consacrées aux principaux domaines technologiques des SI modernes. Voici nos TechWaves sur la Data.

→ La courbe de diffusion des technologies que nous utilisons pour nos TechWaves s'appuie sur le modèle classique d'adoption des technologies[1] à ceci près que, dans la mesure où nous considérons l'usage de la technologie sur la durée et pas uniquement son adoption, nous avons introduit un long plateau avant la phase de décroissance.




Notre courbe distingue donc 6 phases :

## 1.
### L'EXPÉRIMENTATION

Cela concerne les développeurs et entreprises innovantes, start-up, qui cherchent des moyens d'obtenir un avantage décisif grâce à la technologie. Ils testent beaucoup, réalisent des "proofs of concept".

## 2.
### LE DÉCOLLAGE

La technologie est adoptée par les start-up et DSI qui sont prêts à miser beaucoup dessus afin d'obtenir un avantage différenciant. Pour eux, il est vital que le projet aboutisse.

## 3.
### LA CROISSANCE

Les entreprises un peu plus prudentes, mais qui souhaitent tout de même pouvoir s'appuyer rapidement sur des technologies innovantes, adoptent à leur tour la technologie.

## 4.
### LE PLATEAU

La technologie a trouvé son marché. Les plus conservateurs, qui souhaitent partir sur des technologies éprouvées, l'ont adoptée. Le marché est capable de répondre à la demande.

## 5.
### LA DÉCROISSANCE

De moins en moins de nouveaux projets adoptent cette technologie. L'usage est en décroissance. L'utilisation de cette technologie peut être assimilée à de la dette.

## 6.
### LA MENACE

L'utilisation de cette technologie sur un nouveau projet est à proscrire. Peu d'acteurs sont encore impliqués dessus.

(1) "technology adoption lifecycle"
https://en.wikipedia.org/wiki/Technology_adoption_life_cycle







# 2 Panorama des TechWaves Data

WENVISION
TechWaves Data

## Solutions orientées “engagement client”

### Bases relationnelles & transactionnelles

Amazon Aurora
Cloud SQL
Cloud Spanner
MariaDB
Microsoft Azure SQL Server Database
PostgreSQL
SAP SYBASE
MySQL
IBM Informix software
ORACLE
SQL Server
IBM DB2

### Bases clés-valeurs

AEROSPIKE
Memcached
Amazon DynamoDB
Redis

### Bases orientées documents

Amazon DynamoDB
Couchbase
CouchDB
Azure Cosmos DB
Firestore
mongoDB.
mongoDB. Atlas
Firebase Realtime Database

### Bases de recherche

Search Service
algolia
APACHE LUCENE
splunk>
elasticsearch
Amazon Opensearch
Solr

## Solutions pour l’analyse de données et l’algorithmie

### Frameworks de machine learning

H2O.ai
TensorFlow
APACHE Spark
mxnet
PyTorch
dmlc XGBoost
learn

### Plateformes MLOps

DataRobot
Kubeflow
Azure Machine Learning
mlflow
Sagemaker

### Bases orientées graphe

neo4j
OrientDB
Dgraph
Azure Cosmos DB
JanusGraph
Amazon Neptune

### Entrepôts de données (datawarehouses)

Google BigQuery
amazon REDSHIFT
snowflake
Azure synapse Analytics
ClickHouse
ORACLE AUTONOMOUS DATA WAREHOUSE CLOUD
Oracle access manager
VERTICA
teradata.
HIVE
panoply
NETEZZA an IBM Company
CLOUDERA

## Solutions Administrateurs

### Bases de stockage objet

Cloud Storage
Amazon S3
Azure blob Storage
OVHcloud
hadoop
IBM Cloud Object Storage

### Catalogues de données

Data Catalog
Aws Glue
Informatica
Collibra
DataGalaxy
Alation
zeenea

### Diffusion d’événements

Cloud Pub/Sub
Rabbit MD
Red Hat
PULSAR
CONFLUENT
Solace PubSub+
kafka
IBM
Amazon Kafka
Azure Event Hub
aiven
TIBCO
Red Panda
WEBMETHODS.io Integration

## Solutions d’enregistrement de flux de données

### Bases “wide column”

cassandra
APACHE HBASE
DATASTAX
Cloud BigTable
SCYLLA
Amazon Keyspaces
Azure Cosmos DB

### Bases pour séries temporelles

Amazon Timestream
influxdb
kx
Prometheus Software



WENVISION
TechWaves Data

3

# Les fiches TechWaves Data

WENVISION
TechWaves Data

# Les bases relationnelles & transactionnelles

## Définition

Dans les bases de données relationnelles et transactionnelles (SGBDR), les données sont stockées en tant qu'entités relationnelles dans une forme normalisée. Les interactions sont réalisées grâce à un langage spécifique permettant d'effectuer des opérations d'algèbre relationnelle (unions, jointures, projections, etc.). Ces bases doivent respecter les principes ACID : atomicité, cohérence, isolation et durabilité, des propriétés qui garantissent qu'une transaction informatique est exécutée de façon fiable.

## Cas d'usage

Ces bases sont optimisées pour des transactions courtes contenant un petit nombre d'opérations (insert, update, delete) et des requêtes relativement simples. Elles sont utilisées dans des domaines où le modèle transactionnel est critique, par exemple pour des transactions commerciales, où le modèle ACID garantit que l'intégralité des étapes aboutissant au paiement a été prise en compte.

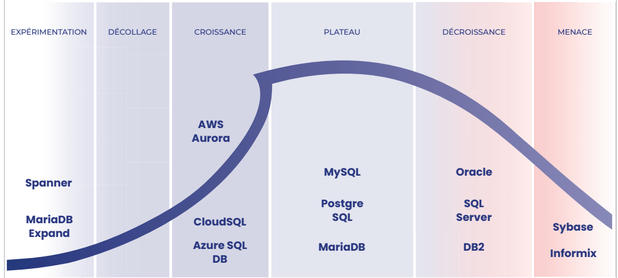
## Notre œil d'expert

Cette TechWave retrace en quelque sorte, de la droite vers la gauche, l'histoire des bases de données, depuis les systèmes traditionnels jusqu'aux bases SQL distribuées dans le Cloud, en passant par l'open source.

En effet, ce domaine des SGBDR transactionnels a longtemps a été l'apanage de bases qu'on retrouve désormais dans la catégorie des produits en ***menace***, abandonnés par leurs éditeurs, mais aussi en ***décroissance***, où l'on trouve des produits dont les éditeurs proposent désormais des versions managées dans le Cloud.

L'open source a pris le relais, et nous retrouvons ces solutions, qui ont désormais une base installée importante, dans la partie ***plateau*** de notre TechWave.

Dans la catégorie ***croissance***, nous retrouvons en général des solutions soit de type éditeur (sous licence) soit de type open source, dont l'évolution est assurée par hyperscaler (scalabilité verticale). Nous n'identifions pas véritablement de produit en mode de ***décollage*** aujourd'hui, mais nous regardons avec intérêt les produits à scalabilité horizontale qui se profilent dans la partie ***expérimentation***.




## Focus sur...

### Amazon Aurora

AWS Aurora est la version managée d'Amazon RDS. Compatible avec les moteurs PostgreSQL et MySQL, elle affiche des gains de performance x3 et x5 respectivement par rapport à ces bases de données. Sa fonctionnalité vedette est le "Parallel Query", qui permet d'accélérer les requêtes plus compliquées ou utilisées dans des pipelines sensibles au temps comme la détection des fraudes.

### Microsoft Azure SQL Server Database

Base de données relationnelle dans le Cloud Azure compatible avec Microsoft SQL Server, Azure SQL DB est une base de données intelligente : elle utilise de l'IA pour optimiser ou détecter en amont les dégradations des performances, pour l'analyse des causes ou pour la détection des menaces. C'est le choix logique pour les entreprises déjà engagées dans l'écosystème Microsoft.

### Google Cloud Spanner

Spanner est une base de données managée et globalement distribuée. Grâce à la technologie propriétaire TrueTime, Google arrive à fournir une solution de stockage des données à cohérence forte et couvrant des vastes zones géographiques. Le choix est parfait pour les entreprises ayant des besoins transactionnels transnationaux - d'autant que c'est quasiment impossible à réaliser en interne.



WENVISION
TechWaves Data

## Les bases orientées documents

### Définition

Les bases de données orientées documents remplacent les notions de tables et relations par des documents, autrement dit des objets dans un format de données textuelles JSON ou XML. Là où les bases à vocation transactionnelle adhèrent aux propriétés ACID (cf. notre TechWave sur les SGBDR), les bases orientées documents suivent les principes BASE : Basic Availability (disponibilité), Soft state (pas d'obligation de cohérence à tout moment entre deux instances) et Eventual consistency (la cohérence sera assurée plus tard, au moment de la lecture ou de la sauvegarde).

### Cas d'usage

Ces bases servent pour des applications nécessitant la distribution de documents sans avoir à réaliser de coûteuses opérations de jointure pour rassembler des informations contenues dans plusieurs tables. Elles sont typiquement privilégiées pour des plateformes de contenu ainsi que pour les catalogues de sites e-commerce. Tous les attributs d'un produit étant rassemblés dans un même document, les pages peuvent se charger très rapidement.

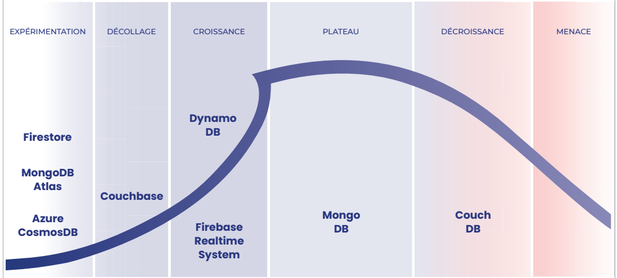
### Notre œil d'expert

Les bases de données orientées documents font partie des premières bases de la mouvance NoSQL. Les premières solutions étaient "on-premises", ce sont celles qu'on retrouve logiquement dans les catégories ***plateau*** et ***décroissance*** de notre TechWave.

La montée en puissance du Cloud a donné naissance à des bases "Cloud natives", ou "fully managed", dont DynamoDB, en pleine ***croissance***. Ces solutions managées par les éditeurs ou les fournisseurs de Cloud sont celles qui sont de plus en plus choisies par l'industrie.

Nos clients plébiscitent la simplicité d'administration de ces bases, que cela concerne l'installation, la montée en charge ou encore la maintenance. Les bases 'on-premises' ne vont pas nécessairement disparaître, mais leur usage pourrait s'avérer coûteux sur le long terme : il est probable que l'offre de service associée soit de plus en plus chère à mesure que l'industrie se tourne vers les solutions managées.

À noter également la ***croissance*** de Firebase, qui fournit des outils d'intégration particulièrement riches et pertinents pour les environnements mobiles.






mongoDB. Atlas

Version "managée", multi-Cloud et géographiquement distribuée de MongoDB, Atlas propose auto-scaling et réplication intelligente pour optimiser l'accès aux données. Atlas est capable d'analyser les requêtes et proposer de nouveaux index ou schémas. Sa capacité "workload isolation" permet d'exécuter des requêtes analytiques lourdes sur des nœuds à part, pour ne pas impacter l'opérationnel.

Azure
Cosmos DB

Base polyvalente (orientée documents, clés-valeurs, colonnes larges, graphe et spatiale), Cosmos DB offre 5 niveaux de cohérence et une indexation automatique sur tous les champs, sans aucune contrainte au niveau du schéma. Elle peut servir dans les projets temps réel (IoT) ou pour des apps distribuées critiques à fort besoin de sécurité ou de continuité d'activité, sans aucun impact sur les performances.

Couchbase

Proposée as-a-service sur AWS et Azure, Couchbase est une base "en mémoire" idéale pour les applications critiques et facile à adopter grâce à N1QL, un dialecte ANSI SQL adapté aux documents.
La réplication des données est possible inter et intra datacenters pour une vaste présence géographique mais également pour du "peer-to-peer" : une synchronisation avec les micro-clients "on the edge", parfaite pour mobiles et IoT.



## Les bases "wide column"

### Définition

Le magasin de données à colonnes larges (ou super colonnes) peut être vu comme une implémentation à part d'un système de stockage clé-valeur dont la clé représente une concaténation des valeurs fréquemment recherchées ou juste un horodatage. Les colonnes ne sont pas définies par un schéma, les enregistrements peuvent avoir différentes colonnes (colonnes éparses) ou la même colonne mais un type différent. Le nombre de colonnes dans une table peut franchir les centaines (dont le nom). Les colonnes peuvent être regroupées logiquement dans des familles de colonnes pour un accès plus rapide.

### Cas d'usage

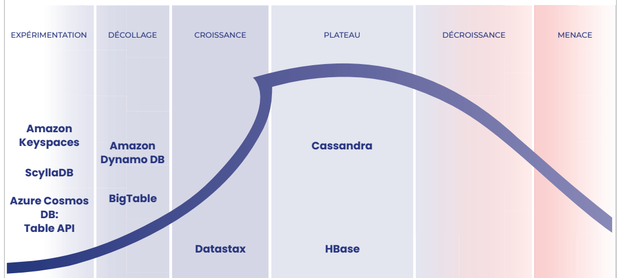
Ce sont des bases de données de niche dont la simplicité garantit une performance très élevée. Les magasins de données à colonnes larges ont la capacité d'ingérer vite des grosses volumétries de données multi-structurées avec une latence basse. C'est une solution utilisée dans le cas des applications en temps réel ou de l'IoT. Elles gèrent aussi très bien le "one-to-many" : par exemple, un utilisateur avec une liste d'actions, ou un panier avec une liste d'items.

### Notre œil d'expert

Ce domaine voit s'affronter deux philosophies. En premier lieu, les bases open source, largement diffusées et utilisées lors de l'avènement de la stack Hadoop, pendant le décollage du phénomène Big Data. Ce sont celles que nous avons positionnées au niveau du ***plateau*** - plus Datastax, l'éditeur de la version commerciale de Cassandra, en phase de ***croissance***. Le besoin qui a alimenté la croissance de ces bases est aujourd'hui couvert par de nombreuses bases de données analytiques, désormais capables de gérer de grands volumes de données.

En deuxième ligne, en phase d'***expérimentation*** et de ***décollage***, nous trouvons plutôt des technologies propriétaires, ré-interprétations par les grands acteurs du Cloud des bases open source du domaine, voire une version managée de Cassandra par Amazon : Keyspaces.

En tant que solution open source émergente, ScyllaDB fait figure d'exception ; il est difficile à ce stade de prévoir son impact.






Google
Cloud Bigtable

Base "managée" proposée par Google, utilisée en interne pour indexer le Web pendant sa croissance exponentielle, au début des années 2000, BigTable est positionnée pour des besoins de faible latence (quelques millisecondes) et de volumétrie dépassant les centaines de gigaoctets par nœud, sans limites supérieures. Les performances augmentent linéairement avec le nombre de nœuds dans le cluster.

Azure
Cosmos DB Table API

Il est la solution Cloud offerte parMicrosoft. Azure Table Storage est capable de manipuler des pétaoctets de données semi-structurées d'une manière sécurisée et flexible. Maintenant que Table Storage est intégrée à Cosmos DB, elle offre la possibilité d'avoir une base répliquée avec une disponibilité en lecture et en écriture de 99,999 % dans le monde entier.

DATASTAX

Datastax est une version as a service et Cloud de Cassandra. Apache Cassandra est un système de base de données distribuée initialement créé par Facebook, désormais open source. DataStax sait gérer virtuellement autant de données que l'on veut du moment qu'on a assez de serveurs. L'autre avantage est la possibilité de traiter facilement de la donnée structurée et semi-structurée.

# Les bases orientées graphes

## Définition

Les bases des données orientées graphes sont des plateformes NoSQL dont les relations entre les données sont explicites et traitées comme des enregistrements. Ces relations sont labellisables et quantifiables, et ce sont elles qui font l'objet des requêtes. Alors qu'il faut plusieurs jointures dans une base relationnelle (où les relations sont implicites) pour obtenir l'ensemble des informations relatives à un item ou un individu, un graphe permet d'extraire directement toutes les données.

## Cas d'usage

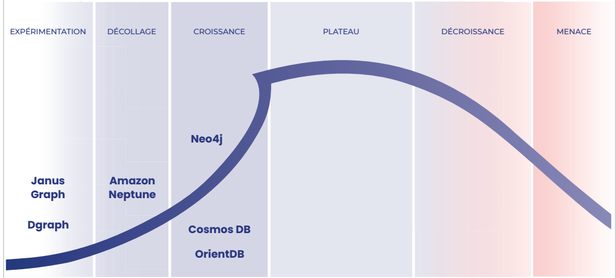
Ces bases représentent le meilleur choix quand la logique métier peut être modélisée comme un graphe et les relations entre les données ne sont pas monotones. Par exemple, pour un système contre le blanchiment d'argent où les nœuds sont les comptes et les sommets du graphe représentent les montants d'argent transférés, de façon à visualiser clairement les interrelations.
Ou bien pour les réseaux sociaux ou applications de rencontre qui cherchent à mettre en évidence et monétiser les relations actuelles ou possibles entre individus.

## Notre œil d'expert

Les systèmes de graphes sont des bases très spécifiques qui ne sont nécessaires que dans certains cas. Historiquement, Neo4j est l'une des bases de graphes les plus anciennes et continue sa phase de ***croissance*** : elle est bien connue et adoptée dans de nombreux projets en production. Une autre base open source, OrientDB, dont la version enterprise a été rachetée par SAP en 2018, voit sa croissance portée par l'utilisation de SAP dans beaucoup de secteurs.

Deux acteurs Cloud proposent une base de graphes : Microsoft avec Cosmos DB, en croissance, et Amazon avec Neptune, en phase de ***décollage***.

Le choix entre ses solutions sera conditionné à l'environnement sur lequel on évolue. CosmosDB a l'avantage d'être compatible nativement avec plus d'API que Neptune. De son côté, Google Cloud a noué un partenariat avec Neo4J, ou propose d'exécuter la base open source JanusGraph avec sa solution BigTable. Également en expérimentation, DGraph est une autre solution open source mettant en avant sa simplicité d'utilisation et ses performances ; à surveiller avec intérêt car elle semble réellement prometteuse.






Amazon Neptune

Base "managée" en totalité par AWS, Neptune arrive à parcourir de gros graphes efficacement au travers des API connues comme Gremlin ou SPARQL pour RDF. Les ressources de calcul et de stockage sont élastiques et peuvent facilement être augmentées ou diminuées, selon les besoins. Avec jusqu'à 15 réplicas en lecture, les données peuvent être accédées rapidement.

Dgraph

Base GraphQL native, distribuée, open source mais aussi dans le Cloud, Dgraph peut traiter 150 000 requêtes complexes par seconde et offre un support pour des transactions ACID. La base propose également une fonctionnalité recherche plein texte, avec filtrage de mots vides ("stop words" en NLP) et extraction des radicaux pour plusieurs langues. Le changement de schéma peut être effectué à la volée sans temps d'arrêt.

neo4j

Neo4j est une base de données développée dès le départ comme une base orientée graphes, capable de gérer plusieurs milliards d'éléments et de relations. Elle supporte les transactions ACID. Certaines applications revendiquent 100 000 écritures transactionnelles par seconde. Depuis 2019, elle est disponible au sein de Google Cloud en tant que service managé. Sa version Aura Enterprise est depuis peu sur AWS.





# Les bases clés-valeurs

## Définition

La base de données clés-valeurs est un système simple de stockage des données qui se base sur une table de hachage ou un dictionnaire pour facilement stocker et récupérer les données. Comme indiqué par le nom, ces bases utilisent des clés uniques pour identifier des objets. Le type d'un objet peut varier, à partir d'un simple "string" jusqu'à des types plus complexes comme les listes ou les "maps". Elles sont ainsi beaucoup plus flexibles que les bases relationnelles, mais aussi moins gourmandes en mémoire.

## Cas d'usage

Les magasins clé-valeur sont utilisés pour stocker des données qui doivent être facilement et rapidement accédées. Cette technologie NoSQL peut être utilisée, par exemple, pour sauvegarder le panier d'un client sur une boutique online, les configurations des différents services éphémères ou encore comme un cache pour des objets.

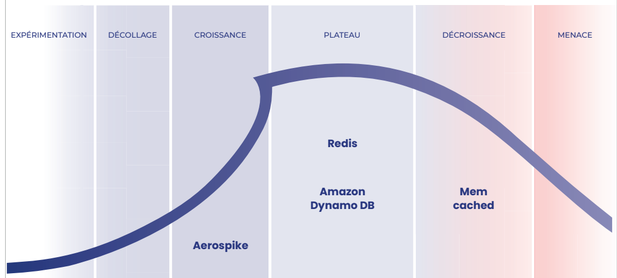
## Notre œil d'expert

Le marché des bases clés-valeurs est très compétitif et beaucoup orienté vers l'utilisation de ces outils pour le cache de sites Web ou d'applications.

Memcached et Redis sont toutes deux extrêmement utilisées pour accélérer l'accès à la donner dans les applications Web. Bien qu'elles offrent des performances brutes similaires, exceptionnelles dans beaucoup de cas, nous considérons Memcached en ***décroissance*** et Redis encore en ***plateau***, cette dernière offrant des fonctionnalités spécifiques recherchées.

Dans la même catégorie, DynamoDI , base de données créée et managée par Amazon, offre une alternative intéressante pour les entreprises recourant à ce fournisseur de Cloud. À noter que DynamoDB est basée sur des disques SSD, à la différence de Redis et Memcached qui sont stockées en RAM. Mais avec DAX (Amazon DynamoDB Accelerator), Amazon dit atteindre des performances en lecture/écriture équivalentes.

Enfin, une solution en ***croissance***, open source et flexible, concerne tout particulièrement les projets manipulant énormément de données : Aerospike.






Amazon DynamoDB

L'un des services les plus populaires offerts par Amazon, cette base "managée" arrive à traiter jusqu'à 10 trillions de requêtes par jour avec une charge maximale de 20 millions de requêtes par seconde et des latences de l'ordre des millisecondes. Dynamo supporte plusieurs types des données, dont le type "document", mais il n'y a aucune contrainte au niveau du schéma. Les différentes valeurs peuvent avoir des types différents.

Redis

Redis est un système de base de données structurées stockées en mémoire open source (licence BSD). Cet outil peut être utilisé pour différents cas d'usages : base de données, système de cache ou encore "message broker". Redis peut être distribué sur plusieurs serveurs. Coté performances, StackOverflow, qui utilise Redis comme cache, était capable d'absorber plus de 1500 millions de commandes par jour en 2019.

Aerospike

Aerospike est une base de données open source distribuée NoSQL de type clé-valeur. Elle prend en charge des schémas de données flexibles et des transactions ACID. Aerospike permet de stocker des milliards d'éléments et peut effectuer des opérations de lecture/écriture en moins de 1 milliseconde. Un des avantages de Aerospike et d'être autant compatible avec un stockage en RAM que sur disques SSD.

# Les entrepôts de données (datawarehouses)

→ **Définition**

L'entrepôt de données est utilisé pour centraliser et conserver en l'état les données de l'entreprise pour avoir une vision globale sur le métier, à des fins analytiques. Les données proviennent de différents systèmes opérationnels et sont conservées sur une longue durée. Le datawarehouse est optimisé pour la lecture et l'analyse de grosses quantités des données au travers de requêtes complexes. Pour avoir de meilleures performances, les données peuvent être agrégées par domaine au sein de datamarts, avec un modèle en étoile préparé pour l'analyse (les cubes OLAP).

→ **Cas d'usage**

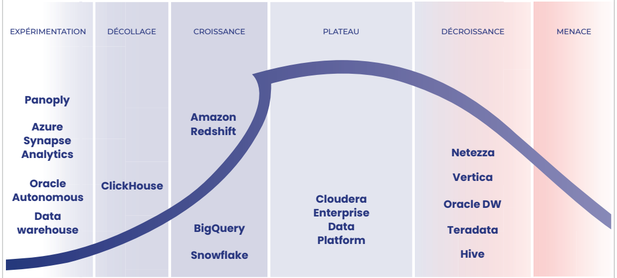
L'entrepôt de données est le socle de l'infrastructure décisionnelle. C'est sur lui que s'appuient les outils d'analyse et de reporting. L'analyse descriptive est la fonction la plus fréquente : il s'agit d'analyser le passé au travers d'agrégations et statistiques simples (chiffre d'affaires, unités vendues, etc.). Les datawarehouses modernes permettent aussi des analyses de type diagnostic (expliquer le pourquoi), prédictif (estimer des valeurs futures) et prescriptif (recommandations basées sur du machine learning).

→ **Notre œil d'expert**

Le monde du datawarehouse est naturellement très peuplé : voilà des décennies qu'il constitue le socle de tout SI décisionnel. Les solutions des acteurs historiques tels qu'Oracle et Teradata, ou encore Netezza, précurseur des "appliances" racheté par IBM, sont en forte ***décroissance***.

Vu le poids des entrepôts de données, de nombreux acteurs plus agressifs sont arrivés, avec des solutions plus généralistes pour le Big Data, comme Hive et Cloudera, aujourd'hui en ***plateau***. Surtout, les acteurs du Cloud proposent désormais des solutions natives, ce sont les offres désormais en ***croissance*** : on citera Amazon Redshift, Google BigQuery ou Snowflake.

La puissance du Cloud leur confère une grande rapidité d'exécution sans les problématiques de maintenance associées. Certaines offres (Snowflake et depuis peu BigQuery) apportent même la possibilité de monnayer ses données sur une marketplace publique afin d'offrir de nouveaux axes de valorisation de la donnée. Dans les solutions en ***expérimentation***, on observe toujours ce virage vers le Cloud Native, que ce soit sur des nouveaux acteurs tel que Panoply ou avec des acteurs historiques comme Oracle. Ces solutions sont à surveiller.






Google BigQuery

BigQuery est le service au centre de l'écosystème "data" de Google, acceptant de gros volumes de données (batch ou stream) depuis les autres services GCP ou plus de 155 sources. Serverless, BigQuery peut analyser des téraoctets en quelques dizaines de secondes sans rien devoir provisionner. Les requêtes fédérées permettent d'analyser des données stockées hors de l'entrepôt. Il est possible d'y créer des modèles ML en SQL.

snowflake®

Snowflake est un entrepôt Cloud managé pour AWS, Azure et GCP, idéal pour le multi-Cloud. Il supporte les données structurées et semi-structurées et permet les requêtes sur les données brutes. Proche du zéro-administration, il peut ajuster son dimensionnement pour optimiser les performances (durant les pics de charge) ou les coûts (durant les heures creuses).

Azure Synapse Analytics

Azure Synapse Analytics est la dernière évolution du datawarehouse Cloud développé par Microsoft, avec l'ambition de réconcilier tous les usages de la data. Elle permet de faire l'analyse de grands volumes de données en T-SQL et même d'exécuter directement du code Spark. Elle regroupe beaucoup d'outils disponibles dans la plateforme Azure et assure une compatibilité avec des services tels que Cosmos DB ou AzureML.



# Les bases pour séries temporelles

→ **Définition**

Les bases pour séries temporelles sont optimisées pour le stockage et les requêtes de données horodatées. Le tampon (date, heure) sert d'index, mais peut être complété par d'autres éléments pour optimiser les performances. Ce type de données étant relativement homogène, il s'agit de bases dont les taux de compression sont particulièrement élevés.

→ **Cas d'usage**

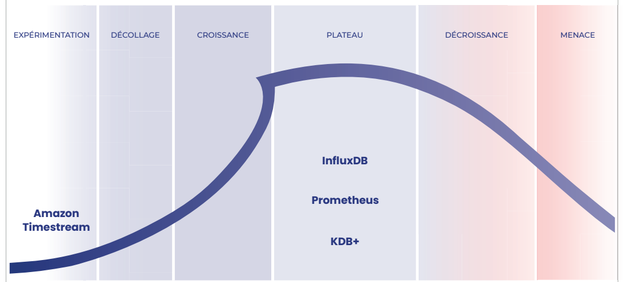
Les bases pour séries temporelles sont recommandées pour les relevés fréquents de données, leur mode d'organisation s'avérant bien plus efficace qu'un base relationnelle ou clés-valeurs. On les trouvait à l'origine dans le domaine financier, pour le trading, mais elles sont largement utilisées aujourd'hui pour tous types de relevés : dans l'IoT, les opérations IT (pour les logs applicatifs, réseaux, etc.) et plus largement partout où des capteurs font des relevés.

→ **Notre œil d'expert**

Le domaine des bases temporelles est dominé par 3 acteurs sur notre ***plateau***. InfluxDB est une base de données très utilisée pour des use cases de monitoring, IoT... Contrairement à Prometheus qui dispose d'outils natifs pour faire du monitoring/alerting, InfluxDB nécessitera l'ajout de programmes tiers si on veut l'utiliser à cet usage. À noter que Prometheus et InfluxDB sont tous deux open source. KDB+ est quant à elle une base sous licence propriétaire, elle offre plusieurs API nativement pour permettre de s'interfacer avec d'autres outils, notamment des messageries interapplicatives.

À l'opposé, la position open source de Prometheus et InfluxDB permet à d'autres acteurs de proposer des plug-in pour leurs outils (PubSub, Kinesis...).

Classé en ***expérimentation***, Amazon Timestream est la solution manageé d'Amazon pour son infrastructure AWS. Amazon Timestream est relativement jeune par rapport aux autres solutions, sa disponibilité générale ne date que de 2020, mais elle s'avère attrayante pour tous les utilisateurs AWS. Côté Google Cloud, il faut concevoir un schéma de données spécifique au sein de BigTable.






influxdb

Base open source conçue pour les séries temporelles, InfluxDB est aussi disponible sur AWS, GCP et Azure. Les requêtes s'écrivent en InfluxQL, une variante de SQL qui permet de réaliser des agrégations. Les requêtes fréquemment utilisées peuvent être précalculées, pour réduire les temps de latence. L'espace de stockage peut être optimisé en supprimant automatiquement les données périmées.

kx

Base en colonnes pourvue d'un moteur "in-memory", kdb+ revendique la capacité d'ingérer 30 millions de lectures de capteurs par seconde, récupérer plus de 100 millions de données par seconde avec des latences inférieures à la milliseconde. L'interfaçage se fait via un langage de requête et de programmation maison, «q». KDB+ est disponible chez les principaux fournisseurs de Cloud et installable sur VM via les places de marché.

Amazon Timestream

La base d'AWS propose un stockage en mémoire pour les requêtes à faible latence et un stockage sur disque pour optimiser le coût. Ainsi qu'un système de gestion du cycle de vie pour déplacer les données de l'un vers l'autre. Les données peuvent être analysées de manière transparente, sans avoir à spécifier le type de stockage. Les outils d'analyse intégrés proposent des agrégats avancés et des types de données complexes.

# Les bases de stockage objet

## Définition

Les entrepôts d'objets sont utilisés pour stocker des données non structurées (images, vidéos, fichiers de forme libre). Les données sont stockées sous la forme d'un objet binaire de grande taille (blob) accompagné de certaines métadonnées (taille, date de création, version de l'objet, etc.) et peuvent être retrouvées grâce à un identifiant unique. Bien qu'une base objets ressemble davantage à une base clés-valeurs avec une hiérarchie plate, l'utilisation d'une convention de nommage peut simuler une hiérarchie imbriquée de dossiers et de fichiers. La nature immuable des objets signifie que certaines opérations, comme un renommage, sont plus coûteuses (un renommage implique le listage, la copie des fichiers sous une clé différente et la suppression des antécédents).

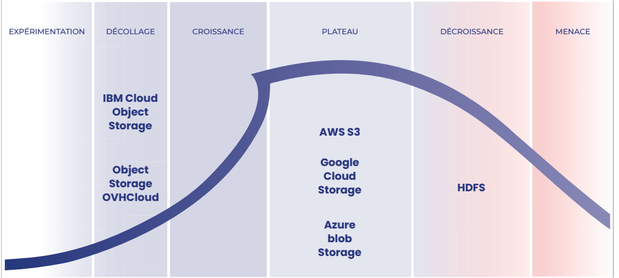
## Cas d'usage

Les bases d'objets sont un candidat idéal pour un datalake d'entreprise. Elles peuvent également servir de zones d'archivage ou de transit pour les pipelines ETL. Une base objet dans le Cloud peut s'utiliser comme un CDN, pour la distribution de contenu. L'architecture peut s'étendre sur plusieurs nœuds, en faisant abstraction des couches de système d'exploitation et de matériel, et en répliquant les données sur ces nœuds pour obtenir redondance et disponibilité maximale pour des configurations multirégionales.

## Notre œil d'expert

Le marché de la base objets est majoritairement détenu par les trois grands acteurs du Cloud, dans notre ***plateau***. Les technologies sont relativement similaires, avec quelques spécificités pour chaque outil. Le choix sera donc surtout déterminé par le fait d'être déjà client d'une de ces plateformes. De même, des acteurs comme IBM ou OVH proposent leur propre solution, que nous positionnons en ***décollage***.

Il existe bien une solution "on-premises", HDFS, mais la tendance pour cette technologie est plutôt à la ***décroissance*** au vu du mouvement de migration vers le Cloud et l'utilisation des services managés.






Amazon S3

Le Simple Storage Service d'Amazon est le premier et l'un des services les plus connus du Cloud. Il propose plusieurs classes de stockage afin d'optimiser le coût des données moins fréquemment consultées. En revanche, pour la classe d'archivage à long terme, la récupération des données peut prendre plusieurs heures. S3 dispose de mises à jour atomiques des objets et de garanties de lecture après écriture.

Google Cloud Storage

GCS est un service clé dans l'architecture serverless GCP. Il dispose de plusieurs classes de stockage, mais il est possible de récupérer immédiatement les données, même en classe archives. Le cycle de vie des objets peut être automatisé à l'aide de règles. Il est aussi possible de créer des politiques de rétention au niveau du bucket ou de l'objet, garantissant que les données ne seront pas altérées pendant une période donnée.

Microsoft Azure blob

Le service de Microsoft Azure blob Storage permet aux utilisateurs de stocker de grandes quantités de données non structurées. Les blob sont des Binary Large Objects, autrement dit des objets tels que des images ou des fichiers multimédias. Intégré à l'environnement Microsoft, notamment à Active Directory, il permet par exemple de partager des fichiers en interne de façon sécurisée.



# Les frameworks de machine learning

## Définition

Les frameworks de machine learning sont aujourd'hui au centre des problématiques liées à la data : ces outils ont pour but d'accélérer les initiatives data science en proposant une boîte à outils pré-codée et optimisée pour s'exécuter sur différents supports. Tout projet de data science repose sur ces outils, dont il existe deux grandes catégories : ceux basés sur des algorithmes "classiques" comme Sklearn ou XGBoost, et ceux fait pour développer des modèles de type réseau de neurones (TensorFlow, PyTorch).

## Cas d'usage

Ces frameworks sont la base de tout projet de data science. Il permettent de développer rapidement des modèles optimisés et aident à l'industrialisation des projets de machine learning. Ils se retrouvent ainsi dans les projets d'analyse d'images, de reconnaissance d'objets, de traitement du langage naturel ou encore l'analyse de sentiment d'un commentaire. On retrouve aussi ces technologies dans les moteurs de recommandation, en particulier dans le domaine marketing. Enfin, les moteurs prédictifs sont généralement développés avec ces technologies, pour la prévision de prix ou de stock.

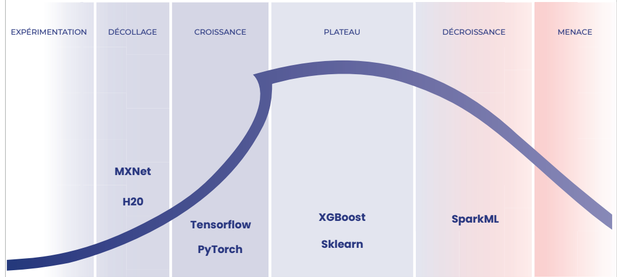
## Notre œil d'expert

Historiquement, les frameworks "classiques" de machine learning sont encore beaucoup utilisés (SKlearn et XGBoost) et restent en ***plateau***. On observe que pour la plupart des problèmes touchant à de la donnée structurée, ces frameworks fonctionnent très bien et sont souvent plus efficaces et moins compliqués à mettre en œuvre que les frameworks de deep learning. Ces outils sont d'autant plus valables sur des volumétries peu importantes (<100Go).

Sur un environnement Spark il est possible d'utiliser SparkML permettant de distribuer l'entraînement des modèles sur un cluster.

Cette solution, bien qu'encore valable, devient obsolète et en ***décroissance*** face à Tensorflow ou PyTorch, en ***croissance***, qui permettent nativement un entraînement distribué et peuvent s'utiliser sur du matériel spécifique pour l'entraînement (TPU), ce qui accroît encore cet avantage. Ils sont particulièrement recommandés pour les données de type photos, vidéos ou encore textes libres.

Les outils en ***décollage*** tels que H20 ou MxNet sont aujourd'hui peu utilisés par l'industrie mais s'avèrent intéressants pour certaines fonctionnalités comme l'AutoML de H20.ai.






Tensorflow

Initialement développé par Google et rendu open source en 2015, Tensorflow est aujourd'hui le framework de ML le plus utilisé. Il offre beaucoup d'avantages et est en constante évolution. La bibliothèque est capable de fonctionner sur plusieurs architectures CPU ou GPU et est disponible sur plusieurs plateformes, y compris sur mobile. En 2016, Google a même créé des processeurs spécifiques appelés TPU.

PyTorch

PyTorch est une bibliothèque d'apprentissage automatique pour Python utilisée principalement pour le traitement du langage naturel. Ce framework open source a été développé par les équipes d'intelligence artificielle de Facebook en 2016. PyTorch est souvent utilisé pour développer des réseaux de neurones profonds.

H2O.ai

H2O est une plateforme d'apprentissage automatique distribuée, open source, qui fonctionne sur différent langages, R ou Python principalement. H2O prend en charge les algorithmes statistiques et d'apprentissage automatique les plus utilisés. H2O dispose également d'une fonctionnalité AutoML pour automatiser l'entraînement et le réglage des algorithmes.



WENVISION
TechWaves Data

# Plateformes MLOps

## Définition

Le DevOps, contraction des mots développement (Dev) et opérations (Ops), vise à fournir un cadre avec différentes étapes clés lors du développement, du déploiement et du suivi d'une application. Cependant, l'approche DevOps dans le domaine du machine learning n'est pas si simple. C'est pourquoi le terme MLOps, cette fois une contraction de M-machine learning et Operations, est récemment apparu. Le MLOps se veut une adaptation du DevOps aux problématiques spécifiques du machine learning. Le développement de ces méthodes MLOps répond aux besoins croissants des entreprises pour mener des projets de données et industrialiser les pipelines ML, de l'idée au déploiement en conditions réelles, en adoptant des méthodes efficaces pour le développement, le déploiement et le contrôle d'un système de machine learning.

## Cas d'usage

Aujourd'hui toute initiative data science doit être accompagnée d'une réflexion autour du MLOps. Ce prérequis va permettre de limiter l'effet "POC" dont ont souffert trop de projets de machine learning, à savoir d'avoir de très beaux modèles d'inférence mais complètement inutilisables car décorrélés de l'IT et des données de production.

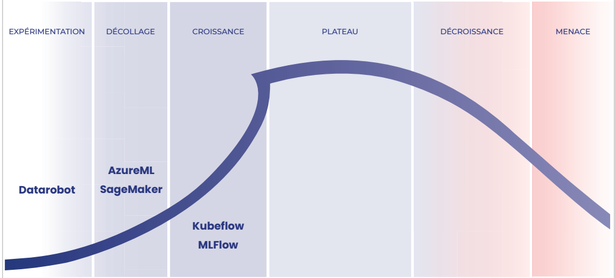
## Notre œil d'expert

Le domaine du MLOps est encore jeune, mais s'avère essentiel pour la réussite des projets autour du machine learning. Dans ce domaine, les solutions les plus utilisées aujourd'hui de notre point de vue, KubeFlow et MLFlow, affichent une belle ***croissance***. Elles offrent la meilleure couverture concernant les fonctionnalités et la portabilité.

Du côté des grands acteurs du Cloud, nous avons des initiatives chez Amazon (SageMaker) et Microsoft (AzureML), en ***décollage***.

Ces solutions restent moins utilisées par l'industrie car elles sont propriétaires et moins portables. Google, de son côté, propose une architecture de référence MLOps pour industrialiser la démarche d'utilisation de ses différents produits consacrés au machine learning.

À surveiller en ***expérimentation***, Datarobot, entreprise créée en 2012, qui offre beaucoup de solutions ML, du MLOps à l'AutoML.






Kubeflow

Le projet Kubeflow vise à rendre les déploiements de projets ML sur Kubernetes simples, portables et évolutifs. Kubeflow s'appuie sur l'infrastructure K8s pour proposer un framework permettant de déployer, monitorer les meilleurs systèmes d'apprentissage. Partout où vous utilisez Kubernetes, vous devriez être en mesure d'utiliser Kubeflow.

mlflow

MLflow est une plateforme permettant de rationaliser le développement de l'apprentissage automatique, y compris le suivi des expériences, le conditionnement du code en exécutions reproductibles, ainsi que le partage et le déploiement des modèles. MLflow peut être utilisée avec n'importe quelle bibliothèque d'apprentissage automatique existante quel que soit l'endroit où vous l'exécutez.

AzureML

AzureML est un service Cloud de Microsoft. Il met à disposition des outils pour aider à développer et déployer des solutions de machine learning. Il est conçu pour aider les data scientists et les ingénieurs ML à industrialiser les approches autours des initiatives ML. L'avantage d'AzureML est d'utiliser nativement les outils Cloud mis à disposition sur la plateforme Azure.



# Catalogues de données

## Définition

Les solutions de catalogues de données ont pour objectif de répertorier tous les actifs de données d'une organisation ainsi que les métadonnées associées. Ces outils sont conçus pour aider les utilisateurs, aussi bien les métiers que les informaticiens ou les spécialistes de la donnée, à trouver rapidement les données et les informations autour de celle-ci : sémantique, lignage, valeur, propriétaire, etc. À la différence de listes maintenues (difficilement) par des individus, leur suivi de la donnée dans le temps est automatisé.

## Cas d'usage

Pour aider les entreprises dans leurs processus d'innovation et de valorisation de la donnée, le catalogue de données est un élément essentiel. Il permet de répandre la connaissance à travers l'entreprise et de ne plus avoir une chasse gardée de la donnée aux seules personnes techniques. Le but est de supprimer les silos et d'apporter une connaissance claire de l'écosystème data d'une entreprise. Des fonctionnalités spécifiques sont possibles comme le suivi de la donnée, la recherche ou encore les registres de traitements.

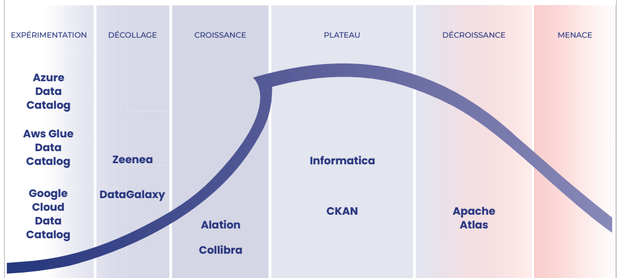
## Notre œil d'expert

La catégorie expérimentation voit apparaître les outils des fournisseurs de Cloud tels que Google Cloud Data Catalog ou AWS Glue Data Catalogue. Ces produits sont aujourd'hui particulièrement utiles pour les utilisateurs de ces fournisseurs. Cependant, ils restent fermés et s'accordent moins avec une approche multi-Cloud ou avec des sources "on-premises".

Dans ces cas de figure, ou si les besoins sont complexes, il faudra plutôt privilégier des outils spécifiques de catalogues de données tels que Zeena, une offre française en ***décollage***, ou Collibra, en ***croissance***.

Ces produits offrent aujourd'hui le panel le plus complet sur ce sujet et utilisent des algorithmes de machine learning pour aider un maximum les processus de recherche et annotation de la donnée

Il existe également des solutions open source comme CKAN, sur notre plateau, ou Apache Atlas, en ***décroissance***. Ces produits offrent une possibilité de limiter les coûts logiciels mais avec une approche très manuelle du catalogue de données.






Collibra

Fondée en 2008, Collibra se concentre sur la gouvernance des données. Collibra Data Governance Center offre un référentiel interactif pour gérer les actifs de données d'une organisation et la terminologie associée, ainsi que les politiques et règles de gouvernance des données (accès, confidentialité...). Le lignage est automatisé et du machine learning est embarqué pour la classification des données.

Alation

Précurseur du domaine, Alation s'appuie sur de solides capacités de découverte, d'accès et de catalogage de données. La vision de l'entreprise californienne fondée en 2012 est de favoriser une forte culture de la donnée chez ses clients en facilitant l'accès à la donnée. Alors qu'Alation Data Governance a traditionnellement été déployée "on-premises", l'offre Cloud plus récente de la firme gagne de plus en plus de terrain.

DataGalaxy

Fondée en 2015 à Lyon, DataGalaxy est une start-up spécialisée dans la gouvernance collaborative des données, partenaire notamment d'Azure. DataGalaxy propose une plateforme orientée sur les principes agiles et la collaboration entre les équipes. Ce catalogue promet des fonctionnalités qui permettent de réduire le temps passé à collecter et tagger les données grâce à des algorithmes de machine learning.

# Les bases de recherche

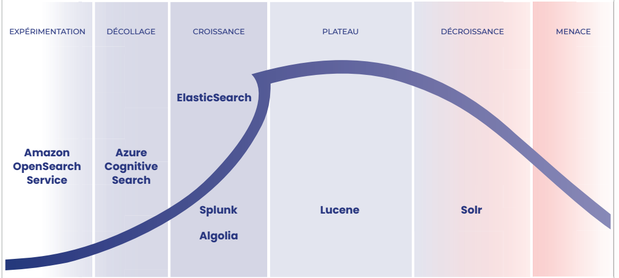
**Définition**
Les moteurs de recherche sont des systèmes de gestion de bases de données NoSQL dédiés à la recherche d'informations. Les bases de données de moteurs de recherche sont optimisées pour traiter des données qui peuvent être longues, semi-structurées ou non structurées, et elles offrent généralement des méthodes spécialisées telles que la recherche plein texte, les expressions de recherches complexes et le classement des résultats de recherche.

**Cas d'usage**
Le cas d'usage le plus fréquent est la recherche plein texte. Grâce à leur architecture particulière, les bases de type "search" sont capables d'exceller dans les recherches approximatives, avec de l'auto complétion ; idéal pour un portail où l'utilisateur ne connaît pas forcément la terminologie. Ces fonctionnalités peuvent aussi être exploitées pour de l'analyse de logs. La structure de ces bases leur permet d'ingérer énormément de logs venant d'applications différentes et d'offrir des options de recherche avancées.

**Notre œil d'expert**
Le moteur de recherche le plus ancien, Lucene, constitue le cœur de produits comme Solr et ElasticSearch. Il est aujourd'hui dans une phase de stabilisation, en ***plateau***, tandis que Solr entame sa ***décroissance***.

En ***croissance***, nous positionnons Elasticsearch, énormément utilisé par l'industrie et qui est aujourd'hui la base de recherche la plus populaire, Algolia, une licorne française, ainsi qu'un outil ancien mais qui continue d'innover et d'être adopté par de nombreuses entreprises pour ses capacités spécifiques en matière d'analyse des logs, Splunk.

En ***décollage***, Azure Cognitive Search, développée par Microsoft, offre de nombreuses fonctionnalités très intéressantes, notamment l'utilisation du machine learning pour l'analyse de document. Cet avantage par rapport à ses concurrents en fait pour nous une solution à observer avec attention. La solution d'AWS, Amazon OpenSearch, basée sur ElasticSearch et compatible avec les anciennes versions de celle-ci, reste en ***expérimentation*** fait du différend entre Amazon et Elastic, l'éditeur d'ElasticSearch. Google, de son côté, n'offre pas de solution comparable, puisqu'on peut intégrer directement son moteur de recherche (pour les sites externes) ou Cloud Search, au sein du domaine d'une entreprise.






## elasticsearch

ElasticSearch est un moteur de recherche et d'analyse RESTful distribué. Au cœur de la stack Elastic, il stocke les données de manière centralisée pour des recherches rapides. Souvent complétée avec d'autres outils tel que Kibana ou Logstash, cette base de recherche est l'une des plus utilisées dans l'industrie. Scalable horizontalement, elle permet de stocker et de requêter des pétaoctets de données.

## Azure Search

Azure Search est la solution de search-as-a-service développée par Microsoft. L'une des puissantes fonctionnalités d'Azure Search consiste à utiliser les capacités de l'intelligence artificielle pour extraire les informations des fichiers, que ce soit le texte, les caractéristiques d'image, ou les entités et les phrases clés du texte brut. Cette fonctionnalité permet de regrouper dans un seul produit différentes typologies de fichiers.

## Amazon OpenSearch Service

Amazon OpenSearch Service est le successeur du service Amazon ElasticSearch Service, basé sur la technologie OpenSearch. OpenSearch est une suite de recherche et d'analyse distribuée, open source, dérivée d'ElasticSearch. Ce service permet d'avoir en quelques clics un moteur de base de recherche permettant de scanner et de requêter des pétaoctets de données textuelles non structurées.



# Les services de diffusion d'événements au fil de l'eau

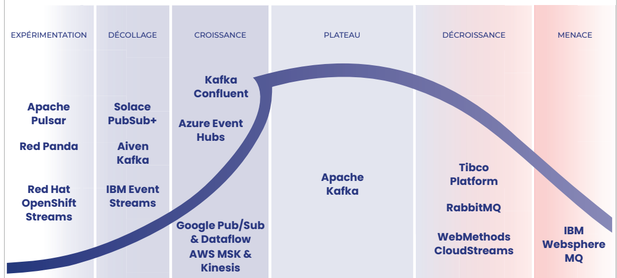
**Définition**
Les services de diffusion d'événements ("event streaming") servent à récupérer les informations de divers systèmes (les producteurs) et à les redistribuer au fil de l'eau à d'autres systèmes, abonnés à ces flux d'informations (les consommateurs). Ainsi, l'information circule, sans nécessiter de couplage fort entre un système et un autre.

**Cas d'usage**
Plusieurs outils de type pub/sub (publish & subscribe) proposent un mécanisme simple, assurant la distribution de la donnée, tandis que des plateformes spécifiques proposent un environnement complet, incluant la possibilité de conserver les données transmises, voire les traiter ou en assurer la gouvernance. De ce fait, les outils et plateformes de diffusion d'événements au fil de l'eau ont une large palette d'utilisation, de la simple ingestion de données d'un outil transactionnel ou d'un capteur IoT vers une base ou une autre application (pour déclencher un processus, éditer une facture, etc.), par exemple, à du traitement complexe en temps réel, avec persistance des données à des fins d'audit ou de constitution d'un modèle de machine learning.

**Notre œil d'expert**
Le domaine de la diffusion d'événements est longtemps resté l'apanage des spécialistes du middleware, comme IBM, dont l'offre historique se situe au bout de notre vague, en ***menace***, ou bien SoftwareAG et son offre autour de WebMethods, ou Tibco, référence du domaine, plutôt en ***décroissance***.

Ces grands éditeurs avaient adapté leurs intergiciels pour échanger des messages, mais se sont fait rattraper par un projet open source spécifiquement conçu pour ce besoin : Kafka, sur notre ***plateau***. Sa mise en œuvre, par de très grandes entreprises, a mis en exergue sa complexité.

D'où l'émergence d'offres managées, telles que celles de Confluent, en ***croissance***, d'Aiven, en ***décollage***, ou Red Hat, en ***expérimentation***. De même, les fournisseurs de Cloud ont packagé leurs propres offres. Chez Microsoft, Event Hubs sert à l'ingestion de données, le traitement étant assuré par Stream Analytics. Même schéma chez Google, avec Pub/Sub et Dataflow. Dans les deux cas, une connexion à Kafka est possible. AWS propose un déploiement automatisé de Kafka avec MSK et un service de traitement temps réel avec Kinesis.
À noter, le projet Pulsar de la Fondation Apache fait évoluer l'approche initiée par Kafka et oriente l'évolution du marché.






## kafka

Initialement développé au sein de LinkedIn, Kafka est désormais un projet open source géré par la Fondation Apache. Il est automatiquement configuré en cluster, avec réplication et tampon. C'est le système de référence aujourd'hui, utilisé par Salesforce, Uber, Netflix, Spotify, Leboncoin... ainsi que des organismes financiers, car son mécanisme de persistance rend les flux totalement auditables.

## CONFLUENT

Fondée par les concepteurs de Kafka, Confluent vise à en éliminer les deux défauts principaux : le temps de mise en place et la gestion au quotidien. L'offre de services managés peut être déployée sur le site du client, dans son Cloud préféré ou en mode hybride. Elle propose un accès SQL pour simplifier l'usage par les data engineers et ajoute de la propriété intellectuelle, comme un jeu de 150 connecteurs ou un cockpit de pilotage des flux.

## Cloud Pub/Sub

Google propose Pub/Sub, une offre serverless extrêmement simple pour diffuser des événements cross-zones et cross-régions, répondant à des cas d'usage limités. Elle peut être complétée de Dataflow, autre service serverless, pour initier des traitements rapides sur de la donnée en mouvement. L'offre est intégrée aux services maison (Cloud Storage, Gmail...) et à des offres de tiers : Splunk, Datadog, Confluent...

# Remerciements

Ce document est la somme du travail réalisé par l'ensemble des personnes travaillant sur le sujet Data au sein du groupe SFEIR et WENVISION, nous tenons à les remercier chaleureusement pour leurs nombreuses contributions, que ce soit en amont pour l'identification des solutions, au cours des recherches que nous avons menées ou bien encore en aval pour leur relecture et leurs commentaires.

Il serait trop long de citer tout le monde, mais nous pouvons remercier en particulier Alaeddine Bouattour, Amel Hafsa, Aurélien Allienne, Christian Kravanja, David Duarte, Florent Legras, Ionel Munteanu, Oussama Mahjoub, Sébastien Coulle, Simon La Personne et Yulianna Khorolich.

Toutes les erreurs ou omissions sont de notre fait ; n'hésitez pas à nous faire part de vos retours, nous en tiendrons compte pour la mise à jour !

**Olivier Rafal**
Consulting Director - Strategy
WENVISION

**Didier Girard**
Co-CEO de SFEIR
Founder WENVISION

WENVISION
TechWaves Data

WENVISION
TechWaves Data

WENVISION

www.wenvision.com

## Paris

48, rue Jacques Dulud
92200 Neuilly-sur-Seine
+33 1 41 38 52 00

## Nantes

Halle 6 Est
40 rue de la Tour
d'Auvergne
44200 Nantes
+33 2 55 59 07 00

## Belgique

Avenue des Arts 6
1210 Bruxelles
+32(0)2 899 83 70

## Strasbourg

Crystal Park
1, avenue de l'Europe
67300 Schiltigheim
+33 3 88 47 04 38

## Lille

74, rue des Arts
59800 Lille
+33 3 66 72 61 32

## Bordeaux

Centre les Grands
Hommes
Place des Grands
Hommes
33300 Bordeaux

## Luxembourg

5 Place de la Gare
1616 Luxembourg
+352 26 54 471

## Nous contacter :

**WENVISION**
contact@wenvision.com