Le document présent résume les technologies utilisées aujourd'hui dans le monde du Big Data, il permet de se familiariser avec l'écosystème Hadoop, de construire une idée globale afin de pouvoir ensuite aller en profondeur et mettre ces notions en pratique.

Afin de bien comprendre l'utilité de chaque technologie dans l'écosysteme Hadoop, j'ai jugé utile de colorer chaque outil et de le réfrérencer à une famille.
Ex : ElasticSearch est en orange car il appartient à la famille "Base de données / Stockage / Indexation"

Orange : Base de données / Stockage / Indexation
Bleu : Traitement - Analyse de données - Processing
Vert : Accès - Récupération, parsing, agrégation de données
Rouge : Sécurité & Monitoring
Mauve : Data Management & Coordination – Sérialisation
Gris : DataViz
Noir : Les distributions

Acunu (Apple) : est une plate-forme analytique à faible latence pour la surveillance et le contrôle des données à grande vitesse utilisées dans les environnements de production.

Ambari (Apache) : Il est utilisé pour le management, réapprovisionnement et le monotoring des clusters Hadoop.

Avro (Apache) : Un compétiteur de **Thrift**, Avro est un système de sérialisation de données et un RPC en même temps, il utilise JSON pour définir les types de données et les protocoles.

Azkaban (Apache): Est un système de contrôle de Job dans Hadoop (comme Oozie).

BackUp : Réplique de données, sauvegarde de données.

BigSheets (IBM) : Est une application Web qui permet aux utilisateurs non techniques de rassembler des données non structurées à partir de sources internes et en ligne et de les analyser pour créer des rapports et des visualisations.

BigTable (Google) : Est une base de données NoSQL orientée colonne, connu pour sa haute performance.

Caffeine (Google) : Considéré comme un remplaçant de MapReduce, Caffeine Project est système d'indexation.

Cassandra (Apache) : Est une base de données NoSQL orientée colonne avec une réplication asynchrone sans master.

Cloudera : Une star-up qui se consacre au développement de solutions de type Big Data basées sur le framework Hadoop.

Cloudera Manager : Un outil pour faire le Management des cluster Hadoop.

Cluster HDFS : Une architecture de machine HDFS

- NameNode : Gère l'espace de noms, arborescence de système de fichiers et les méta-données des fichiers et des répertoires .
- DataNode : Stock et restitue les blocks de données, il interroge NameNode pour connaitre quel DataNode est disponible (celui avec la plus grande bande passante).

CouchDB (Apache) : Est une BDD NoSQL orientée document, conçu pour pouvoir être réparti sur de multiples serveurs.

D3.js : Est une bibliothèque graphique en JavaScript qui permet l'affichage des données numériques sous une forme graphique et dynamique.

DataMeer : Société qui apporte des outils de gouvernance pour l'analyse Hadoop.

Data Center : Est un site physique sur lequel se trouve regroupés des équipements constituant les SI des entreprises.

Drill (Apache) : Est un framework logiciel open-source qui supporte les applications temps réel distribuées pour l'analyse interactive des jeux de données à grande échelle.

ElasticSearch : Est un moteur de recherche et une BDD orientée document.

Flume (Apache) : Fait la collecte et l'agrégation des gros volumes de données (ex : les logs).

Hbase (Apache) : Est une BDD NoSQL orientée colonne, elle dispose d'un stockage pour les grandes tables.

HDFS : Système de fichier distribué, conçu pour stocker de très gros volumes de données.

Hive (Facebook) : Logiciel d'analyse de données permettant d'utiliser Hadoop avec une syntaxe proche du SQL.

HortonWorks : Une société qui se concentre sur le développement et le soutien Hadoop.

Hue (Apache) : Une interface Web qui permet de faire de l'analyse des données.

HyperTable : Est une base de données NoSQL proche de **BigTable,** elle est orientée colonne.

Ganglia : **E**st un outil scalable et distribué de Monotoring pour les systèmes informatiques, les clusters et les réseaux à hautes performances.

Gephi : Gephi est un logiciel libre d'analyse et de visualisation de réseaux.

Giraph (Apache) : Est un projet Apache destiné à réaliser du traitement de graphes sur des volumes importants de données. Giraph utilise l'implémentation de MapReduce réalisée par Apache Hadoop afin de traiter les graphes.

Grafana : Est une solution libre permettant de réaliser des dashboards depuis des métriques Graphite, InfluxDB et OpenTSDB.

GraphViz : Est un outil de visualisation de graphique en ligne de commande. Il est surtout utilisé pour le diagramme de flux et d'arbre à usage général plutôt que les graphes moins structurés que Gephi.

GreenPlum : Est un outil qui permet de faire du Data Analytics, Greenplum offre un moyen intéressant de combiner un langage de requête flexible avec des performances distribuées.

Impala (Cloudera) : Est moteur de requête SQL open source de Cloudera, il est utilisé pour les données stockées dans des cluster d'ordinateurs exécutant Apache Hadoop.

Index : Correspond au nom de la BDD en relationnel.

InfluxDB : Est une base de données NoSQL orientée Time Series.

Kafka (Apache) : Est un projet qui vise à fournir un système unifié, temps réel à latence faible pour la manipulation de flux de données en temps réel. La conception est fortement influencée par les transactions logs.

Kibana : Est un plugin de visualisation de données open source pour Elasticsearch. Il fournit des fonctionnalités de visualisation en plus du contenu indexé sur un cluster Elasticsearch.

Kerberos : Un protocole d'authentification réseau qui repose sur un mécanisme : clé secrète (chiffrement symétrique).

LDAP : Un protocole permettant de l'interrogation et la modification des services annuaires.

MapR : Une distribution en Big Data, Outil de traitement de données qui fait partie de l'écosystème Hadoop

Mahout (Apache) : Est un Framework open source qui permet d'exécuter des algorithmes d'apprentissage machine communs sur des ensembles de données massives.

MapReduce : Est un patron d'architecture, dans lequel sont effectués des "calculs parallèles" sur des gros volumes de données.

- Map : Pour le mapping de la donnée
- Reduce : Filtre et agrégation du résultat.

Marvel : Est un outil qui fait du Monotoring et le Managment dans ElasticSearch, Il donne une visibilité complète sur votre déploiement d'Elasticsearch. Anticipez les problèmes, accédez plus rapidement et optimisez la performance de cluster Elasticsearch.

Mllib (Spark) : Une bibliothèque Machine Learning de Spark. Pour les calculs parallélisés, Tous les algorithmes de cette librairie sont conçus de manière à être optimisés pour le calcul en parallèle sur un cluster.

MongoDB : Une base de données NoSQL orientée document.

MrJob : Est un framework qui permet d'écrire le code pour votre traitement des données, puis de l'exécuter de manière transparente soit localement, sur Elastic MapReduce, soit sur votre propre cluster Hadoop.

Nagios : Est une application permettant la surveillance système et réseau. Elle surveille les hôtes et services spécifiés, alertant lorsque les systèmes ont des dysfonctionnements et quand ils repassent en fonctionnement normal.

Neo4J : Une base de données NoSQL orientée Graphe.

Oozie (Apache) : Est un système de contrôle de Job similaire à Azkaban, mais exclusivement axé sur Hadoop.

Pig (Apache) : Est une plateforme de haut niveau pour la création de programme MapReduce utilisée avec Hadoop. Le langage de cette plateforme est appelée le Pig Latin.

Ranger (Apache) : Offre un cadre de sécurité centralisée pour gérer le contrôle d'accès à grande échelle à travers.

Redis : Une base de données NoSQL orientée Clé-valeur.

R Connectors (Oracle) : Fournit un accès à un cluster Hadoop à partir de R, permettant la manipulation des données résident HDFS.

Riak : Une base de données NoSQL orientée Clé-valeur.

S4 (Yahoo !) : Yahoo! a initialement créé le S4 pour prendre des décisions sur le choix et le positionnement des annonces, après la société à décider de l'utiliser pour traiter des flux d'événements arbitraires.

SearchNode : Alège le Master, joue le rôle d'un proxy.

Sentry : Est un système permettant d'appliquer une autorisation basée sur les rôles aux données-méta-données stocké sur les clusters Hadoop.

Shard : Une partition d'un index (400 shards/datanode).

Solr/Lucene (Apache) : Lucene est une bibliothèque Java qui gère l'indexation et la recherche de grandes collections de documents.

Spark (Apache) : Est un FrameWork de calcul distribué qui effectue les opérations sur la mémoire vive (rapidité).

Sqoop (Apache) : Est une application d'interface de ligne de commande pour le transfert de données entre les bases de données relationnelles et Hadoop.

SSL : Un protocoles de sécurisation des échanges sur Internet (mode de chiffrement).

Storm (Apache) : Est un système de calculs temps réel distribué en temps réel "tolérant aux pannes".

Type d'Index : Correspond à la table dans le monde relationnel (BDD).

Tableau : Est une société de logiciel américaine dont le siège se trouve à Seattle. Elle conçoit une famille de produits orientées visualisation de données.

TinkerPop (Apache) : Est un framework de calcul graphique pour les bases de données graphiques (OLTP) et les systèmes analytiques de graphes (OLAP).

Thrift (Apache) : Avec Thrift, on pourra prédéfinir à la fois la structure des objets de données et les interfaces qu'on utilise pour interagir avec eux.

Voldemort (LinkedIn) : Est une BDD NoSQL, elle est orientée Clé-Valeur utilisée par LinkedIn pour un stockage de haute scalabilité.

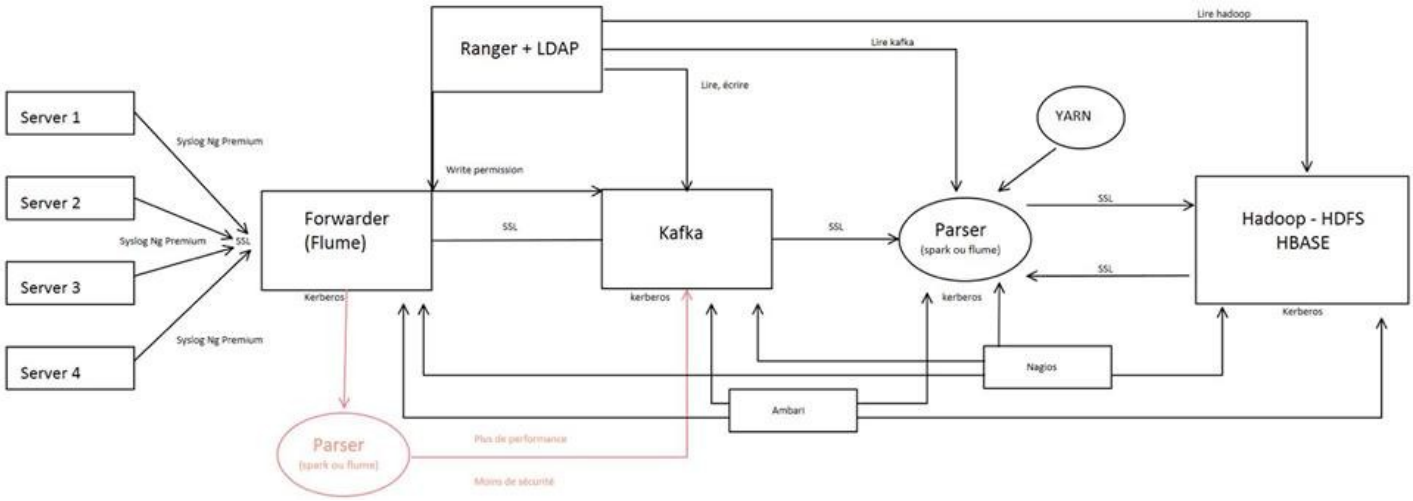
Watcher : Plugin d'alerting et de notification pour Elasticsearch qui permet de détecter les modifications et anomalies dans nos données pour des applications telles que la journalisation, la sécurité.

Weka : Est une suite de logiciels d'apprentissage automatique.

Yarn : (Yet Another Resource Negotiator) est une technologie de gestion de clusters. Elle rend l'environnement Hadoop mieux adapté aux applications opérationnelles qui ne peuvent pas attendre la fin des traitements par lots.

Zookeeper (Apache) : Il s'agit d'un logiciel de gestion de configuration pour systèmes distribués. Il fait la gestion de la haute disponibilité de noeuds.

Un exemple d'une solution Big Data :



Ecosystème Hadoop :