**ROADMAP du projet**

Réunion de lancement du projet, clarification des objectifs et des attentes du client.

**Phase 1 : Collecte et nettoyage des données**

* Extraction initiale des données à partir des fichiers plats CSV :
  + Un fichier CSV contenant l'historique des commandes des clients
  + Un fichier CSV contenant des informations sur les différents modèles de vélos
  + Un fichier CSV contenant des informations sur les distributeurs
* Les données seront nettoyées pour garantir leur qualité et leur cohérence. Cela comprend la suppression des doublons, la correction des valeurs aberrantes et le traitement des valeurs manquantes.

**Phase 2 : Intégration des données et modélisation de l'Architecture et Alimentation Batch**

* Les données seront intégrées dans un data lakehouse avec Spark et Delta Lake.
* Mise en place de l'alimentation en mode batch à partir des fichiers CSV dans le Data Lake House et enrichissement des données avec des informations sur le produit.
* Un schéma en étoile sera utilisé pour assurer des requêtes rapides et une structure simplifiée.
* Les données seront enrichies en les croisant avec des informations sur le produit, telles que le nom du produit, le type du produit, les composantes internes du produit, etc.

**Phase 3 : Mise en place du POC**

Un POC (Proof of Concept) sera mis en place pour valider la faisabilité de la solution.

Le POC sera basé sur une architecture Docker pour faciliter le déploiement et la gestion de la solution

* Choix des Images Docker
* Configuration des conteneurs
* Création de scripts Docker-Compose

**Phase 4 : Développement des tableaux de bord**

* Des tableaux de bord seront développés pour visualiser les données et permettre aux utilisateurs de prendre des décisions éclairées.
* Les tableaux de bord comprendront des indicateurs clés tels que les nouveaux clients, la valeur vie client, le taux de rétention, le taux de désabonnement, etc.

**Phase 5 : Campagne de tests**

* Des scénarios de test seront définis pour valider la performance et la fiabilité de la solution.
* Les scénarios de test couvriront les aspects suivants :
  + Déploiement et configuration
  + Intégration des données
  + Scalabilité
  + Recovery
* Une campagne de tests sera menée pour valider la performance et la fiabilité de la solution.

**Phase 6 : Gestion de projet**

* Des outils de gestion de projet seront utilisés pour suivre l'avancement du projet et garantir la conformité au plan.
* Les outils de gestion de projet utilisés est Trello.

**Phase 4 : Mise en production**

* La solution sera mise en production et les utilisateurs seront formés à son utilisation.
* La solution sera surveillée et maintenue pour garantir son bon fonctionnement.

**Calendrier**

* 4 semaines du 13/09/2023 au 12/10/2023

**Budget**

* Étant un projet fil rouge, il est difficile d’estimer un budget.

**Ressources**

* L'équipe du projet sera composée d’un ingénieur en données, d’un analyste de données et d’un business intelligence.
* L'équipe utilisera des outils open source tels que Hadoop, Spark, Delta Lake, MLlib, Prometheus, Grafana…

**Suivi**

* Le projet sera suivi par un comité de pilotage composé de l'équipe de projet.
* Le comité de pilotage se réunira toutes plusieurs fois par semaine pour examiner l'avancement du projet et prendre des décisions.