Recap

Partie 1 : D’où vient la data ?

I. La base de données de production

Quasi toutes les applications (web ou mobile) s'appuient sur une ou plusieurs bases de données. Par exemple, pour une application à laquelle un client doit se connecter, il y aura obligatoirement une base de données pour stocker ses informations. La même chose si vous vendez des produits, vous devez stocker les quantités qu’il vous reste en stock. Et ainsi de suite pour chaque donnée que vous devez enregistrer.

Dans de nombreux cas, ces données sont très utiles pour l'analyse. Elles répondent par exemple à la question "combien d'utilisateurs actifs avons-nous ce mois-ci ?".

II. Les évènements

Des événements sont crées à chaque fois qu'un utilisateur réalise une action dans le produit. Ces événements sont donc enregistrés dans une base de données et peuvent être utilisés par la suite.

Par exemple si un utilisateur clique sur l'onglet "Paramètres" de votre application, vous pourriez déclencher automatiquement un évènement.

Maintenant, si je suis curieux de savoir combien de fois les utilisateurs cliquent sur l'onglet "paramètres", je peux simplement agréger ces événements avec une requête SQL. L'obtention d'informations à partir d'événements plutôt qu'à partir d'une base de données de production est appelée Event Driven Analytics.

III. Les outils SaaS

Les entreprises s'aident fréquemment d'outils tels que Salesforce, Hubspot et Stripe, qui produisent des données précieuses pour répondre aux questions de l'entreprise.

Par exemple, si vous gérez votre facturation via Stripe, c'est de là que proviendront vos données de revenus.

Salesforce lui dispose d'informations intéressantes sur la manière dont vos prospects passent dans le processus de vente, de données démographiques sur vos prospects, etc.

Le plus difficile est de faire sortir les données de ces outils Saas et de les placer dans un endroit où vous pouvez les analyser avec toutes vos autres données.

IV. La donnée publique

C'est moins courant, mais parfois les entreprises vont aller chercher des données auprès de sources publiques ou utiliser des fournisseurs d'enrichissement comme Clearbit pour obtenir des informations sur leurs clients ou les visiteurs de leur site.

Partie 2 : Où va la data ?

I. Pour la data analyse, dans un data warehouse

Il est aujourd'hui communément admis que pour la data analyse, il faut un Data Warehouse contenant toutes les données dont on a besoin, centralisées en un seul endroit et optimisées pour que les requêtes SQL s'exécutent rapidement.

Les trois grands acteurs du secteur sont Snowflake, BigQuery (de Google), Redshift (d'Amazon), et un nouveau venu - Firebolt. Ces solutions sont coûteuses, mais elles simplifient le tout considérablement en évitant de devoir gérer des serveurs et une infrastructure.

Partie 3 : Comment bouger la data ?

Une fois qu'on a une idée d'où viennent les données et vers où elles vont, on doit réfléchir à la manière dont on les déplace. L'objectif ici est d'extraire les données des systèmes cloisonnés et de les transférer dans notre Data Warehouse, où on va pouvoir les utiliser à notre guise.

I. Les orchestrateurs

Transférer des données vers un Data Warehouse n'est généralement pas une opération simple. Pour se faire, il faut construire des "pipelines de données" qui s'appuient sur plusieurs tables différentes.

Les orchestrateurs aident les équipes à construire des pipelines de données complexes, à s'assurer qu'ils s'exécutent dans les temps et à les alerter lorsque les choses ne vont pas bien. L'outil open source le plus populaire dans ce domaine est Airflow.

II. Les logiciels SaaS

Plus facile, qu'Airflow, il existe aussi des outils SaaS que vous pouvez utiliser pour déplacer des données sans avoir à écrire une tonne de code.

Ces outils permettent principalement d'extraire des données d'outils dits "classiques", qui représentent une grosse partie du marché (Salesforce, Stripe, Google Analytics, Shopify, ...).

Les données qui se trouvent dans ces outils ont l'avantage d'un schéma fixe - la

façon dont les données sont stockées est identique pour tout le monde. Toutes les entreprises qui utilisent Stripe ont (à quelques détails près) des noms de table, des noms de colonne et une structure de base de données identiques, car Stripe a tout mis en place. Il en va de même pour Salesforce et Hubspot. Cette cohérence simplifie considérablement le déplacement des données sans code personnalisé.

Les outils les plus populaires pour ce type de déplacement sont Stitch, Fivetran et Airbyte (qui lui est open source).

Partie 4 : Comment préparer la data ?

I. La modélisation

Chaque équipe d'analyst a ce qu'on appelle un modèle de données - c'est la façon dont elle fait correspondre le monde de l'entreprise (utilisateurs, clients, fidélité, churn, ...) aux données récoltées.

La modélisation couvre quelques éléments importants :

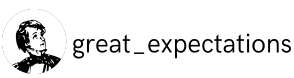
Les définitions : que signifie un utilisateur ? Que signifie "actif" ? Que signifie récurrent ?

Les tables intermédiaires : prendre des requêtes SQL exécutées fréquemment, les programmer et les matérialiser sous forme de nouvelle table (par exemple, daily_active_users)

Les performances et la structure : comment organiser des grandes tables dans un schéma qui optimise la vitesse et le coût des requêtes.

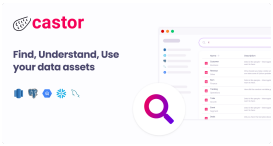
La modélisation est une discipline vieille comme le monde et, aujourd'hui encore, les équipes paient des consultants pour qu'ils examinent le business et qu'ils définissent intelligemment le modèle de données dans lequel les données de l'entreprise vont s'inscrire.

II. Les tests



Les pipelines et les tables de données doivent être testés régulièrement pour éviter les erreurs et les inexactitudes dans les données, et contrôlés en permanence pour maintenir le temps de fonctionnement aussi proche de 100 % que possible.

III. La documentation



Collibra

La documentation des données permet à l'entreprise de :

- comprendre d'où vient la donnée
- ce qu'elle représente (sa définition et son périmètre)
- comment elle peut être utilisées (via des script SQL ou des dashboards)

Partie 5 : Comment est utilisée la data par les data analyst ?

I. Des requêtes SQL ad'hoc

DBeaver

POPSQL

Les data analysts, pour récupérer la donnée dont ils ont besoin, utilisent des logiciels pour écrire du SQL, comme PopSQL ou DBeaver. Via leurs requêtes ils vont pouvoir extraire la donnée nécessaire aux réponses qui leur sont posées.

II. De la visualisation

tableau Looker Google Data Studio Power BI

Les logiciels de dashboarding modernes vont permettre de se connecter à notre Data Warehouse pour ainsi automatiser tout le reporting. Les graphes réalisés vont en effet se mettre à jour avec la nouvelle data et ainsi permettre chaque semaine

d’avoir automatiquement son dashboard actualisé avec de la donnée “fraiche”. En plus d’automatiser le reporting, ils permettent de réaliser des analyses plus visuelles et plus facilement transmissibles que des requêtes SQL.