Data Engineer

Definition of the profession

Définition DataEngineer

L'ingénieur de données, ou Data Engineer, est un expert qui joue un rôle crucial dans la gestion et l'exploitation des données au sein d'une entreprise. Il est responsable de concevoir, construire et maintenir l'infrastructure nécessaire pour collecter, stocker, traiter et analyser les données. Son travail permet aux data scientists et aux autres utilisateurs de données d'accéder aux informations dont ils ont besoin pour prendre des décisions éclairées.

Sources:

Oracle.com - Définition Data Engineer Apec.fr - Tous nos métier, Data Engineer Mycommunit.io - Fiche métier, Data Engineer

What qualities do you need to have for this job?

Soft Skills:

1. **Capacité à résoudre des problèmes de manière créative** et à prendre des décisions techniques.
2. Bonnes compétences en **communication** pour collaborer avec d'autres membres de l'équipe et expliquer des concepts techniques à des non-techniciens.
3. Capacité à **rester à jour sur les nouvelles technologies** et les tendances en matière de données.

4. **Attention aux détails** et capacité à assurer la qualité et la fiabilité des données traitées.

Hard Skills:

1. **Compétences techniques solides en programmation** (comme Python, Java, Scala), en manipulation de données et en développement de bases de données. `print("HelloWorld")`
2. Une bonne **compréhension des concepts** liés aux **bases de données** et aux **technologies de stockage de données**.
3. **Capacit à travailler** avec de **grands ensembles de données** et à les analyser en utilisant des outils et des technologies appropriés.
4. **Compétences** en ingénierie logicielle pour concevoir et **construire des pipelines** de données fiables et évolutifs.

Ces qualités sont essentielles pour réussir en tant qu'ingénieur de données et pour contribuer de manière significative à la gestion et à l'analyse des données dans une organisation.

Sources:

Datacamp.com - Become DataEngineer Clementine.jobs - Metiers Big Data

Mais qu'est qu'un Pipeline

Un pipeline de données est une série d'étapes de traitement visant à préparer les données de l'entreprise pour l'analyse. Les organisations disposent d'un grand volume de données provenant de diverses sources telles que les applications, les appareils de l'Internet des objets (IoT), et d'autres canaux numériques. Cependant, les données brutes sont inutiles ; elles doivent être déplacées, triées, filtrées, reformatées et analysées pour l'informatique décisionnelle. Un pipeline de données comprend diverses technologies permettant de vérifier, de résumer et de trouver des modèles dans les données afin d'éclairer les décisions métier. Des pipelines de données bien organisés prennent en charge divers projets big data, tels que les visualisations de données, les analyses de données exploratoires et les tâches de machine learning.

Sources:

aws.amazon.com - DataPipeline

What languages/tools do I need to master?

- Les langages de programmation comme Python, Java et Scala ainsi que l’environnement UNIX. (Python et Java reste donc la priorité ! Mais Scala est un bon bonus.)
  - **Python**: C'est le langage le plus populaire pour l'ingénierie de données. Il est polyvalent, facile à apprendre et dispose d'une large communauté de développeurs et d'utilisateurs.
  - **Java**: Ce langage est robuste et fiable, et il est souvent utilisé dans les grandes entreprises.
  - **Scala**: Ce langage est particulièrement bien adapté pour le traitement distribué de données volumineuses. Il est utilisé avec des frameworks comme Apache Spark.
  - **Bash**: Ce langage est utilisé pour écrire des scripts pour automatiser des tâches.

Data Lakes et les Data Warehouses

Le stockage de données sous différentes formes et à plusieurs volumétries, dans les Data Lakes et les Data Warehouses. Le stockage de données c’est LA clef de voûte du métier de Data Engineer donc connaître les bases SQL et NoSQL reste très important.

Pourquoi apprendre les Data Lakes et les Data Warehouses ?

- Les Data Lakes et les Data Warehouses sont des technologies essentielles pour l'analyse de données.
- La demande en professionnels capables de gérer et d'analyser ces données est forte.
- Apprendre ces technologies vous permettra d'accéder à des postes plus rémunérateurs et plus intéressants.

Le traitement de données Big Data

- En particulier avec les frameworks comme Hadoop et Spark, pour manipuler des données de plusieurs dizaines/centaines de Go.
    - Apache Hadoop: Un framework open source pour le stockage et le traitement distribué de données volumineuses.
    - Apache Spark: Un framework open source pour le traitement distribué de données volumineuses. Les flux de données en temps réel, où la nécessité de diffuser et analyser des données en temps réel devient de plus en plus présente
        - Apache Hive: Un entrepôt de données SQL construit sur Hadoop.

Mais qu’est-ce qu’un frameworks ?

Un framework en Data Engineering est un ensemble d'outils qui facilite la création et la gestion de pipelines de données. Il permet aux Data Engineers de se concentrer sur la logique métier de leurs projets sans se soucier des détails techniques.

Les Environnements Cloud

- (Azure, GCP, AWS) et les services associés. Un environnement Cloud, est un système informatique décentralisé qui utilise internet pour fournir des services informatiques à la demande.
- *Les environnements Cloud sont devenus essentiels pour les Data Engineers car ils offrent:**
  - **Flexibilité et évolutivité** pour répondre aux besoins fluctuants en matière de traitement de données.
  - **Coûts réduits** en ne payant que les ressources utilisées.
  - **Fiabilité** et sécurité grâce aux infrastructures des fournisseurs Cloud.
  - **Innovation** avec de nouveaux services et fonctionnalités.
  - **Collaboration** facilitée entre les membres de l'équipe. Exemple
  - **Microsoft Azure**: Un fournisseur de services de cloud computing.
  - **Google Cloud Platform (GCP)**: Un fournisseur de services de cloud computing.
  - **Amazon Web Services (AWS)**: Un fournisseur de services de cloud computing.
    - **Apache Kafka**: Une plateforme de streaming de données.

Les flux de données en temps réel, où la nécessité de diffuser et analyser des données en temps réel devient de plus en plus présente

- *Compétence plus inattendue mais ô combien nécessaire pour les applications (ex: Uber), les services financiers et même des géants comme Carrefour ! Donc c'est flux de données traitent les données au moment où elles sont générées, sans les stocker au préalable. Cela permet d'obtenir des insights plus rapides et de réagir plus rapidement aux événements, ce qui es très recherchés par les entreprises.*

Les pipelines ETL (Apache Airflow, Luigi, Kafka Connect)

L'automatisation, qui s'applique aussi bien sur les pipelines de données que sur les cycles de vie en Machine Learning.

- **Airflow :** est un framework open-source et flexible destiné à la gestion des workflows de données. Il utilise une interface web intuitive pour visualiser et

gérer les pipelines, et permet de définir des dépendances entre les tâches.

- **Luigi** : est un framework Python open-source plus léger et conçu pour les workflows basés sur des tâches. Il se concentre sur la construction de pipelines robustes et reproductibles.
- **Kafka Connect** : est un framework open-source permettant de connecter facilement des sources de données et des destinations à Apache Kafka, une plateforme de streaming distribuée. Il agit comme un adaptateur flexible pour connecter diverses sources et formats de données à Kafka.

Les trois étapes principales d'un pipeline ETL (Extraction/Transformation/Chargement) :

1. **Extraction**: Récupérer les données de sources variées, comme des bases de données, des fichiers CSV, des API, etc. Gérer les formats de données différents et les structures non standardisées. Assurer la sécurité et la confidentialité des données pendant l'extraction.
2. **Transformation**: Nettoyer les données en corrigeant les erreurs et en supprimant les doublons. Transformer les données pour les rendre conformes au format cible. Enrichir les données en ajoutant des informations supplémentaires provenant d'autres sources. Appliquer des règles métier et des transformations complexes.
3. **Chargement**: Charger les données transformées dans la destination finale. Optimiser le processus de chargement pour garantir la performance et l'intégrité des données. Gérer les conflits et les erreurs de chargement.

L’intelligence artificielle :

 DeepLearning vs MachineLearning

Comme le Machine Learning et le Deep Learning. Il n’est pas nécessaire de disposer de connaissances avancées. Mais son travail étant de faciliter celui des data scientists, il doit comprendre les concepts clés de la science des données.

Sources:

Malt.fr - Compétence demandée Datascientest.com - Tout savoir

How long does training take before being ready to enter the job market?

La durée de la formation nécessaire pour être prêt à entrer sur le marché du travail en tant qu'ingénieur de données peut varier en fonction de votre parcours, de votre expérience antérieure et des compétences et des technologies spécifiques que vous devez maîtriser. En général, cela peut prendre de plusieurs mois à quelques années pour développer les compétences et les connaissances nécessaires pour entrer sur le marché du travail en tant qu'ingénieur de données.

Voici quelques étapes typiques de la formation pour devenir ingénieur de données :

1. **Formation académique** : De nombreux ingénieurs de données ont un diplôme en informatique, en science des données ou dans un domaine connexe (Master). Si vous n'avez pas de diplôme pertinent, vous devrez peut-être suivre des cours supplémentaires ou obtenir des certifications pour acquérir les connaissances nécessaires.
2. **Développement des compétences techniques** : Vous devrez développer de solides compétences en langages de programmation tels que Python, SQL, et des outils comme Apache Hadoop, Apache Spark et les bases de données relationnelles. Cela peut se faire par l'auto-apprentissage, des bootcamps, des cours en ligne ou des programmes d'enseignement formels.
3. **Expérience pratique** : Réaliser des projets concrets et acquérir de l'expérience pratique dans des tâches d'ingénierie de données est crucial. Vous pouvez utiliser des jeux de données disponibles en ligne ou travailler sur des projets personnels pour perfectionner vos compétences.
4. **Stages ou postes débutants** : Faire des stages ou occuper des postes débutants dans des rôles liés aux données peut fournir une expérience précieuse et vous aider à passer à un rôle d'ingénieur de données.

5. **Apprentissage continu** : Le domaine de l'ingénierie de données évolue constamment, il est donc important de rester à jour sur les nouvelles technologies et les tendances en assistant à des ateliers, des conférences et des cours en ligne.

En **fin de compte**, le temps nécessaire pour être prêt pour un emploi en ingénierie de données dépendra de votre engagement en matière d'apprentissage, des ressources à votre disposition et du niveau de compétence que vous visez. Certaines personnes peuvent être prêtes à entrer sur le marché du travail après environ un an de formation ciblée, tandis que d'autres peuvent mettre plus de temps. L'essentiel est de construire une base solide de compétences et d'expérience pour réussir dans ce domaine.

Le temps "classique" est de 4 à 5 ans avant d'avoir son premier job comme Data Engineer.

Source:

Datacamp.com - Become data enginneer

Job in short supply? Salary for a beginner/junior?

 Définition DataEngineer

En ce qui concerne le salaire pour les débutants ou les juniors en tant qu'ingénieur de données, cela peut **varier en fonction de facteurs** tels que l'emplacement, le niveau d'expérience et l'ensemble de compétences spécifique.

Cependant, voici quelques chiffres généraux :

- **Aux États-Unis**, le salaire moyen pour les ingénieurs de données débutants/juniors peut varier de 60 000 à 90 000 dollars par an.
- **En Europe**, les salaires pour les ingénieurs de données débutants peuvent varier de 30 000 à 50 000 euros par an, en fonction du pays et de l'entreprise (entre 39 000 et 42 000 en moyenne en Belgique).
- **En Asie**, les salaires pour les ingénieurs de données juniors peuvent varier de 20 000 à 40 000 dollars par an, en fonction de l'emplacement et de l'organisation.

Pays/Salaire	États-Unis	Europe	Asie	Belgique
Min/an	60k $	30k $	20k $	39k €
Max/an	90K $	50k $	40k $	42k €

Ces chiffres sont approximatifs et peuvent fluctuer en fonction des conditions du marché et des circonstances individuelles. À mesure que vous acquérez plus d'expérience et d'expertise dans le domaine de l'ingénierie de données, votre salaire est susceptible d'augmenter en conséquence.

En Belgique, en tant que Junior Data Engineer, vous pouvez vous attendre à un salaire moyen de 33.100€. Il y a pas mal d'opportunités à Bruxelles, Malines, Zaventem. Sur Stepstone, il y a actuellement 330 offres d'emploi disponibles.

Les certifications spécialisées.

Telles que le **certificat Cloudera Certified Data Engineer**, démontrent des compétences spécifiques recherchées par les employeurs. Les certifications d'IBM, Amazon, Google, Oracle et Microsoft peuvent également vous aider à gagner davantage et à obtenir des emplois dans des rôles spécialisés. Voici quelques certifications d'ingénieur de données à considérer.

- Google Certified Data Engineer Certification
- Data Engineer Certifications by Microsoft
- SAS Certified Big Data Professional
- Google Professional Data Engineer
- Data Science Council of America (DASCA) Associate Big Data Engineer
- Data Science Council of America (DASCA) Senior Big Data Engineer
- Amazon Web Services (AWS) Certified Data Analytics – Specialty
- Cloudera Data Platform Generalist Certification

Sources:

Stepstone.com - Junior Data Engineer Coursera.org - Data Engineer Salary