Кто такой Data Engineer?

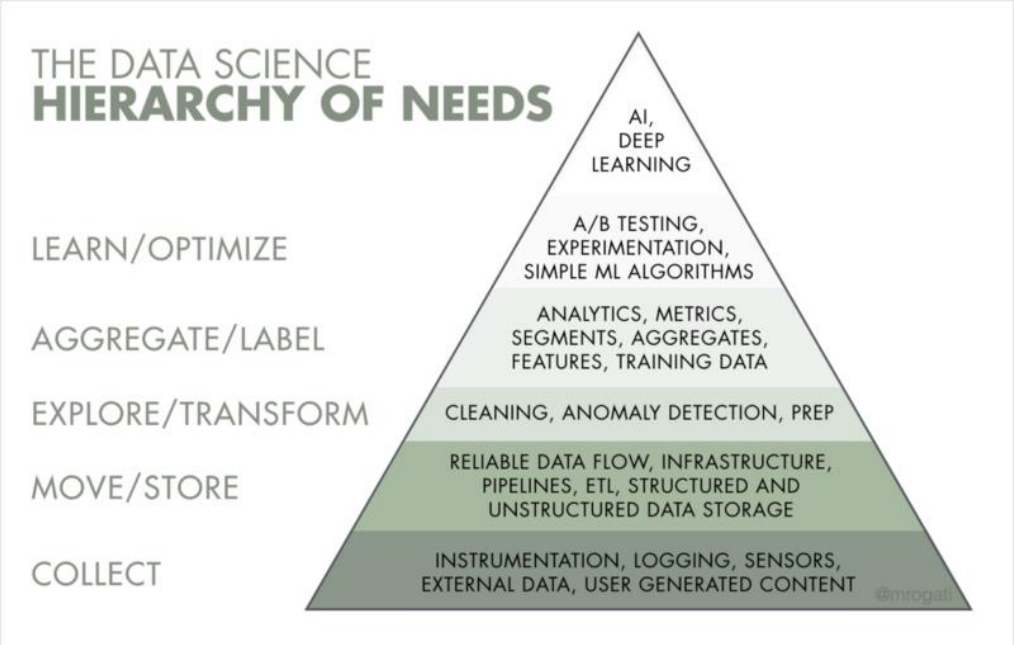
Инженер данных — IT-специалист, который отвечает за извлечение, преобразование, загрузку данных и их обработку.

Его главная задача – сделать процесс анализа данных в компании максимально удобным для аналитиков, обеспечить их очищенными данными в должном количестве и в должный срок, а затем вывести модель в продукт.

В чём отличие Data Engineer от Data Scientist?

Профессии Data Scientist и Data Engineer часто путают. Их главное различие в том, что обычно у них разные цели.
Data Engineer разрабатывает, тестирует и поддерживает инфраструктуру работы с данными: базы данных, хранилища и системы массовой обработки, а также очищает данные для использования аналитиками и дата-сайентистами, то есть создаёт конвейеры обработки данных.

Data Scientist создает и обучает предиктивные модели с помощью алгоритмов машинного обучения и нейросетей, помогая бизнесу находить скрытые закономерности, прогнозировать развитие событий и оптимизировать ключевые бизнес-процессы.



https://netology.ru/blog/08-2019-kto-takoy-data-engineer

В иерархии Data Science потребностей инженерия данных занимает первые ступени(сбор, перемещение и хранение, подготовка данных), что говорит о том, что любая работа по Data Science начинается с помощью инженера данных.

Обязанности Data Engineer:

Понимание сути и сбор данных	Построение архитектуры пайплайна обработки данных	Вывод моделей в готовый продукт (главная функция)
вникнуть в суть данных, понять, что нужно аналитику для построения модели, определить основные источники данных, получить доступ к ним(понять, как они хранятся и в каком виде можно извлечь)	владеть всем необходимым инструментарием хранения и обработки больших массивов данных	безболезненно интегрировать модель в уже устоявшиеся внутренние процессы

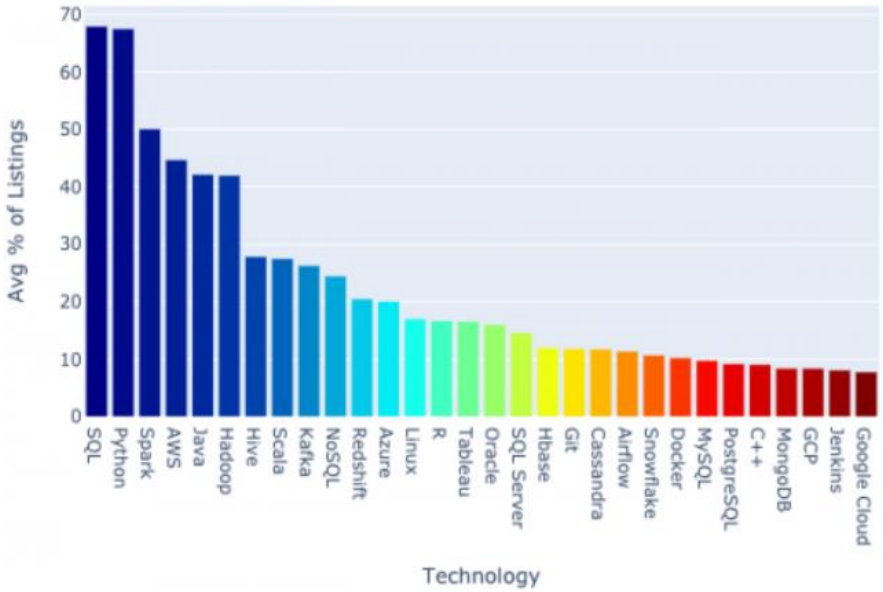
Необходимые профессиональные навыки и знания Data Engineer:

- алгоритмы и структуры данных;
- принципы хранения информации в SQL и NoSQL базах данных(MySQL, MSSQL, PostgreSQL, MongoDB, SQL Server, Oracle, HP Vertica, Amazon Redshift)
- ETL-системы (Informatica ETL, Pentaho ETL, Talend);
- облачные платформы для Big Data решений (Amazon Web Services, Google Cloud Platform, Microsoft Azure);
- стек Apache Hadoop (HDFS, HBase, Cassandra);
- SQL-движки для анализа данных, хранящихся в распределенных файловых системах типа HDFS (Apache Hive, Impala);
- кластеры Big Data на базе Apache (Hadoop, Kafka, Spark);
- языки программирования (Python, Java, Scala) для работы с Big Data системами;
- большим плюсом будет опыт работы с инструментами визуализации данных, такими как Tableau или ElasticSearch.

Личные качества Data Engineer:

- умение работать в условиях многозадачности;
- готовность разбираться в новых технологиях;
- стрессоустойчивость;
- желание работать в команде.

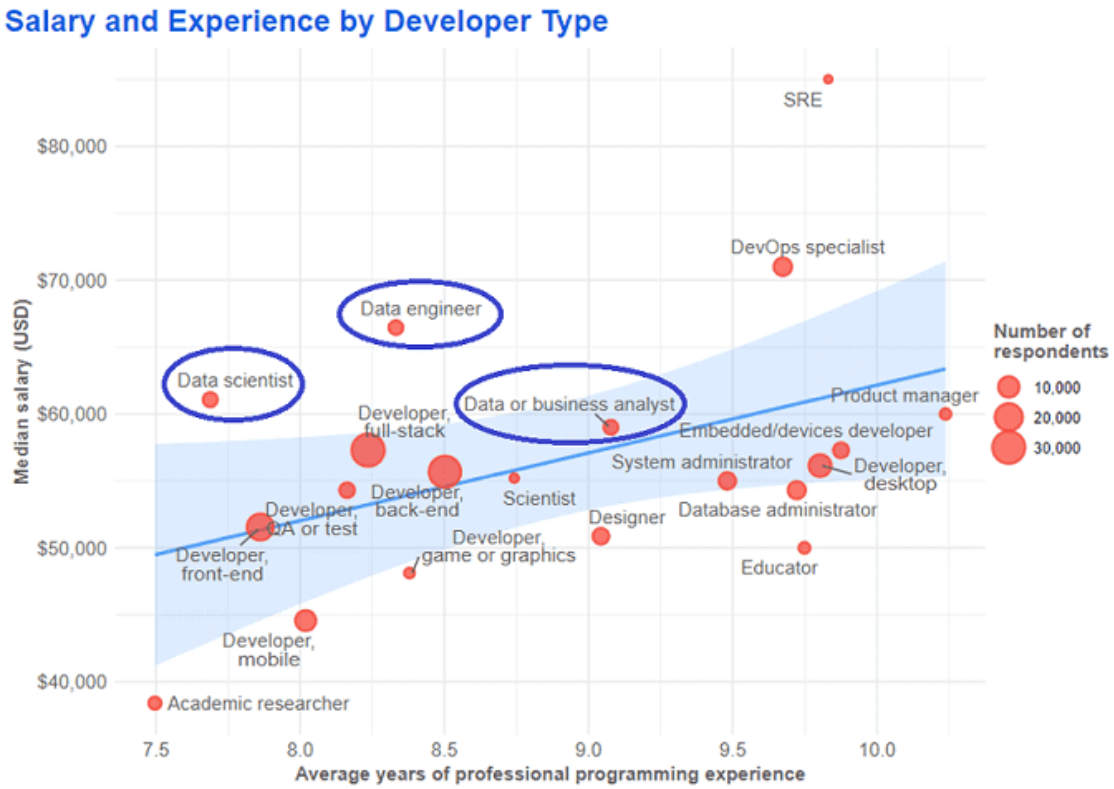
Посмотрим на распределение технических терминов из сферы Data Engineering с самыми высокими показателями по сайтам вакансий 2020 года:



https://prohoster.info/blog/administrirovanie/samye-vostrebovannye-navyki-v-professii-data-engineer

Сколько платят Data Engineer?

Для начала посмотрим на распределение зарплат в зависимости от направления Data Science, согласно ежегодному исследованию Stack OverFlow:

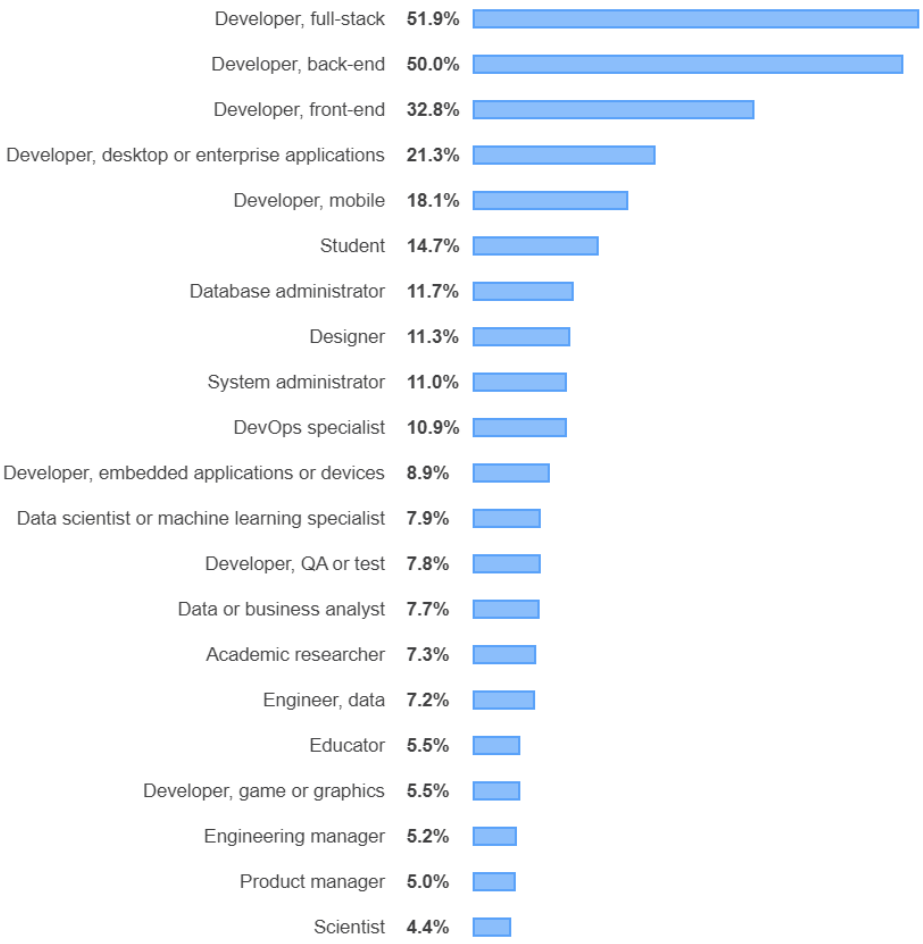


https://www.bigdataschool.ru/blog/big-data-%d1%81-%d1%87%d0%b5%d0%b3%d0%be-%d0%bd%d0%b0%d1%87%d0%b0%d1%82%d1%8c.html

Специалисты в этой области больше всех зарабатывают среди ИТ-специалистов. Например, в 2019 году, согласно ежегодному исследованию Stack OverFlow, годовая зарплата аналитиков, инженеров и исследователей данных в США равнялась 60-70 тысяч долларов, т.е. около 350 тысяч рублей в месяц и спрос на специалистов по данным все время растет по всему миру. Таким образом, большие данные – это очень перспективная и финансово выгодная IT-область.

С зарплатами для Data Engineer все понятно и классно :) Теперь посмотрим на конкуренцию в этом направлении, согласно исследованиям Stack OverFlow.

Распределение количества IT-специалистов по основным направлениям:



https://insights.stackoverflow.com/survey/2019

Как мы видим, только 7,2% IT-специалистов в мире разбираются в области инженерии данных, поэтому спрос на них колоссальный.

В каких отраслях Data Engineer может найти себя?

Data Science стремительно развивается и находит себя во всех сферах деятельности человека, соответственно, поэтому вакансии для Data Engineer можно найти практически в любой отрасли.

А что мы увидим на российском рынке?

Теперь посмотрим на российский рынок,а именно на вакансии на hh.ru, rabota.ru и superjob.ru. В целом, вакансии на всех трех сайтах однотипны, поэтому далее информация только по hh.ru.

Средняя зарплата по вакансии Data Engineer: от 80000-100000 руб.(для Junior Data Engineer) и до 375000 руб(для Senior Data Engineer). Как правило, для позиции Junior Data Engineer(думаю, это важно для нашего потока) требуется меньший список навыков:

- Знание технологий Big Data (Hadoop, Spark, Hive/Impala);
- Знание Python в части написания скриптов для анализа/обработки данных и визуализации результатов;
- Продвинутый уровень SQL (аналитические функции, подзапросы, хранимые процедуры, оптимизация запросов);
- Способность к анализу, систематизации, классификации;
- Способность работать с большими объемами информации; инициативность; способность работать в команде.

Стоит отметить, что данные вакансии популярны только в крупных городах(Москва, Санкт-Петербург, Казань и тд).

Вывод:
Согласно статистике 2019-2020, Data Engineer является профессией, спрос на которую растет быстрее всех прочих, что сказывается на высокой заработной плате инженеров, так как Data engineer играет в организации критически важную роль. Каждый из студентов нашей кафедры может постараться получить необходимые навыки Data Engineering и присоединиться к этому направлению работы :)