Практическая работа № 1

Анализ рынка труда по профессии инженера данных в России.

1. **Цель работы**

Исследовать рынок труда по профессии инженера данных (Data Engineering).

1. **Задание**

* Изучить рынок труда по профессии инженера данных (Data Engineering) и его особенности. Источник данных: hh.ru , www.zarplata.ru и другие сайты по поиску работы.
* Проанализировать требования к языкам, стеку инструментов и опыту.
* Составить отчет. В отчете ответить на вопросы:

1. Рынок труда по профессии инженера данных: количество вакансий, отрасли компаний, размер компаний.
2. Требования: популярные языки, технологии, фреймворки, опыт.
3. Кратко опишите своими словами, для чего используется каждая из популярных технологий, фреймворков (не более 10).
4. Ваши цели на обучение. Перечислите интересующие Вас темы и технологии.
5. **Ход работы**

Источник данных: [Инженера данных (Data Engineering) - hh.ru](https://hh.ru/search/vacancy?text==%28data+OR+дата%29+AND+%28engineer+OR+engineering+OR+инженер%29+OR+%22инженер+данных%22&search_field=name)

* 1. О компаниях

Всего 185 компаний разместили вакансии инженера данных на hh.ru.

В таблице ниже отображено 20 компаний наиболее активных в области Data Engineering на текущий момент

|  |  |  |  |
| --- | --- | --- | --- |
| **Список компаний** | | | |
| № | Работодатель | Всего размещено различных вакансий | Количество вакансий Data Engineering |
| 1 | СБЕР | 5269 | 57 |
| 2 | Тинькофф | 10319 | 36 |
| 3 | Яндекс | 7455 | 19 |
| 4 | МТС | 1973 | 13 |
| 5 | VK | 522 | 11 |
| 6 | Itransition | 91 | 11 |
| 7 | Спортмастер | 1418 | 10 |
| 8 | Деко Системс | 20 | 9 |
| 9 | Ozon | 7888 | 8 |
| 10 | inDrive | 105 | 6 |
| 11 | СИБУР, Группа компаний | 1037 | 5 |
| 12 | билайн | 1338 | 5 |
| 13 | Playrix | 145 | 5 |
| 14 | Газпром-медиа Развлекательное телевидение (ГПМ РТВ) | 119 | 5 |
| 15 | BestDoctor | 60 | 4 |
| 16 | КОРУС Консалтинг | 65 | 3 |
| 17 | Ростелеком | 3166 | 3 |
| 18 | Банк ВТБ (ПАО) | 1737 | 3 |
| 19 | Tele2 | 307 | 3 |
| 20 | Центр финансовых технологий | 139 | 3 |

В следующей таблице представлены 10 лидирующих отраслей, в которых специализируются найденные на предыдущем шаге компании

|  |  |  |  |
| --- | --- | --- | --- |
| **Виды деятельности компаний** | | | |
| № | Вид деятельности | | Количество компаний |
| 1 | Информационные технологии, системная интеграция, интернет | 7.538: Интернет-провайдер  7.539: Системная интеграция, автоматизации технологических и бизнес-процессов предприятия, ИТ-консалтинг  7.540: Разработка программного обеспечения  7.541: Интернет-компания (поисковики, платежные системы, соц.сети, информационно-познавательные и развлекательные ресурсы, продвижение сайтов и прочее) | 81 |
| 2 | Финансовый сектор | 43.641: Аудит, управленческий учет, финансово-юридический консалтинг  43.643: Коллекторская деятельность  43.644: Лизинговые компании  43.645: НПФ  43.646: Страхование, перестрахование  43.647: Банк  43.648: Управляющая, инвестиционная компания (управление активами)  43.650: Финансово-кредитное посредничество (биржа, брокерская деятельность, выпуск и обслуживание карт, оценка рисков, обменные пункты, агентства по кредитованию, инкассация, ломбард, платежные системы) | 29 |
| 3 | Розничная торговля | 41.512: Розничная сеть (drogerie, товары повседневного спроса)  41.515: Розничная сеть (одежда, обувь, аксессуары)  41.517: Розничная сеть (продуктовая)  41.520: Розничная сеть (спортивные товары)  41.521: Розничная сеть (электроника, бытовая техника)  41.525: Интернет-магазин | 11 |
| 4 | Перевозки, логистика, склад, ВЭД | 5.461: Авиаперевозки  5.463: Железнодорожные перевозки  5.465: Транспортно-логистические комплексы, порты (воздушный, водный, железнодорожный)  5.466: Складские услуги  5.467: Курьерская, почтовая доставка  5.468: ВЭД, таможенное оформление | 8 |
| 5 | Телекоммуникации, связь | 9.399: Мобильная связь  9.400: Фиксированная связь  9.401: Оптоволоконная связь | 7 |
| 6 | Нефть и газ | 47.612: Нефтепереработка, нефтехимия (производство)  47.614: Нефтехимия (продвижение, оптовая торговля)  47.615: ГСМ, топливо (продвижение, оптовая торговля)  47.616: ГСМ, топливо (розничная торговля) | 6 |
| 7 | Услуги для бизнеса | 44.390: Безопасность, охранная деятельность  44.391: Кадровые агентства  44.393: Консалтинговые услуги  44.396: Услуги по организации поездок и бронированию  44.397: Контроль качества, экспертиза, испытания и сертификация | 5 |
| 8 | СМИ, маркетинг, реклама, BTL, PR, дизайн, продюсирование | 11.453: Производство мультимедиа, контента, редакторская деятельность  11.456: Теле- и радиовещание  11.459: Маркетинговые, рекламные, BTL, дизайнерские, Event-, PR-агентства, организация выставок | 4 |
| 9 | Услуги для населения | 49.409: Салоны красоты  49.410: Спортивные и фитнес-клубы  49.412: Туристические компании  49.415: Центры развлечения | 4 |
| 10 | Образовательные учреждения | 39.441: Научно-исследовательская, научная, академическая деятельность  39.442: Вуз, ссуз колледж, ПТУ  39.444: Бизнес-образование  39.445: Повышение квалификации, переквалификация  39.446: Тренинговые компании  39.448: Обучение иностранным языкам | 4 |

* 1. О вакансиях

Всего было найдено 440 вакансий в области Data Engineering.

На основе сбора и анализа этих данных были получены следующие результаты:

Из раздела ключевые навыки было отмечено 345 различных навыков. На графике ниже отображено 30 наиболее популярных из них.

|  |  |  |
| --- | --- | --- |
| **Заработная плата**  *(заполнена для 50 вакансий из 440)* | | |
| Валюта | Диапазон | Количество вакансий |
| RUB | от 45 000 до 500 000 | 43 |
| USD | от 5 000 до 11 000 | 3 |
| KZT | от 200 000 | 2 |
| EUR | от 3 000 | 1 |
| UZS | от 7 000 000 | 1 |

* 1. Описание популярных технологий, фреймворков

**Apache Hadoop** - это программная платформа, которая позволяет управлять огромными объемами данных. Hadoop может хранить и обрабатывать огромные объекты структурированных, полуструктурированных и неструктурированных данных, таких как записи посещения в Интернете, журналы веб-серверов, сообщения в соцсетях, электронные письма и данные датчиков IoT. Hadoop поддерживает инициативы расширенной аналитики в хранилищах данных, такие как прогнозный анализ, интеллектуальный анализ данных и машинное обучение. Hadoop состоит следующих основных компонентов: распределенной файловой системы HDFS, в которой хранятся данные; фреймворка для управления ресурсами кластера и менеджмента задач – YARN; MapReduce, разбивающего задачи обработки на несколько задач; и набора утилит Hadoop common, предоставляющего базовые возможности, необходимые другим компонентам Hadoop. К основным преимуществам Hadoop относятся его способность защищать данные в случае сбоев оборудования, его масштабируемость от одного сервера до тысяч машин и его аналитика в реальном времени, которая способствует принятию решений и анализу истории.

**Apache Spark** – это BigData-фреймворк для распределенной обработки данных в оперативной памяти. Spark поддерживает обработку структурированных (в виде таблиц), полуструктурированных (json, yaml, xml и так далее) и неструктурированных данных (тексты и другие медиаформаты). Spark основан на Hadoop MapReduce и расширяет модель MapReduce, чтобы эффективно использовать ее для других типов вычислений, включая интерактивные запросы и потоковую обработку. У Apache Spark есть API для четырех языков программирования: Python, Java, Scala, R. Поскольку Spark в основном написан на Scala, для достижения максимальной эффективности лучше использовать Scala и (в меньшей степени) — Java. Но самый популярный язык для написания Spark Job — Python. Он предоставляет простое и понятное API для взаимодействия со Spark, которое повышает читаемость кода и упрощает его поддержку. Также Python добавляет множество опций для визуализации данных, использование которых невозможно в Scala или Java. PySpark (интерфейс Spark для Python) при использовании основного модуля Spark SQL обычно дает такую же производительность, как и любой другой язык программирования. Apache Spark позволяет эффективно обрабатывать данные. Например, программа WordCount, о которой мы говорили выше, выполняется в нем в 100 раз быстрее, чем в Hadoop MapReduce. Это возможно благодаря распределенной архитектуре и in-memory расчетам. Все вычисления в Spark происходят в оперативной памяти. Apache Spark состоит из четырех компонентов: MLlib - библиотека инструментов для ML-задач; Streaming - инструменты для real-time аналитики; SQL-модуль для работы со структурированными данными; GraphX - библиотека для процессинга графовых структур данных.

**Apache Airflow** – это opensource-инструмент для разработки, менеджмента и мониторинга дата-пайплайнов (ETL-процессов). Вся логика в нем описывается только кодом на Python. Главное преимущество этого - можно использовать Git для версионирования ETL-задач. У Airflow огромное количество коннекторов к внешним источникам данных: файлам, СУБД, API клауд-провайдерам (AWS S3, Azure Storage, GCP Storage), а также к внешним движкам для обработки данных (Apache Spark, Hadoop MapReduce). Если нужного оператора для работы с данными или коннекшена к движку нет, то функциональность Airflow легко расширяется самописными модулями. Airflow изначально проектировался так, чтобы можно было изменять встроенный набор функций под свои задачи. Также у него много дополнительных инструментов, включая встроенное хранилище паролей, переменных, NoSQL-базу данных для хранения метаданных.

* 1. Цели на обучение. Интересующие темы и технологии.

Цель: углубить знания в области больших данных и машинного обучения.

Наиболее интересующие темы:

* Раздел 2. Базы данных: 2.2. Программирование Scala;
* Раздел 3. Обработка данных;
* Раздел 5. Искусственный интеллект.