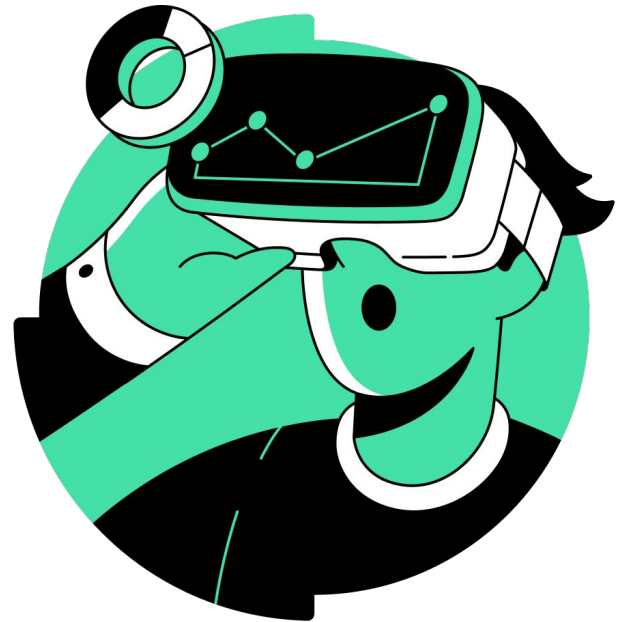


# Дикий Запад

Кто такие дата-инженеры и почему они так сильно отличаются друг от друга.



# Обо мне



**Семён Осипов**

Data Engineer, Semrush



@OsipovSimon



Труба Данных @ohmydataengineer

Обо мне



# О чем мы поговорим?



1. Дата инженер это относительно новая специальность, и требования к ней слишком расплывчаты.
2. Для тех, кто хочет стать дата инженером не понятно, с чего начинать.
3. Уже состоявшиеся специалисты находятся в сложном выборе, куда дальше развиваться.
4. Работодатели закрывают вакансии очень долго, поиск релевантного специалиста осложнен требованиями к позиции.

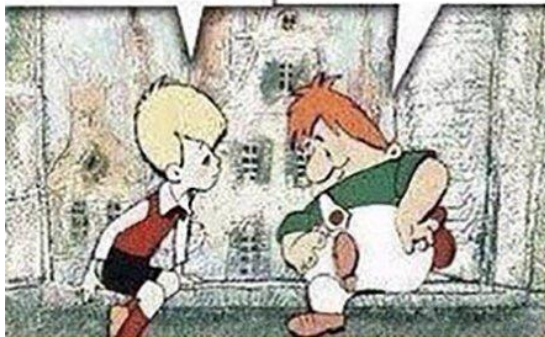


ПИТОНИСТАМ  
МАЛО ПЛАТЯТ...



ЭТО ЗАЧЕМ?

БУДЕШЬ РАЗРАБАТЫВАТЬ  
ИСКУССТВЕННЫЙ  
ИНТЕЛЛЕКТ ЗА 300К/СЕК



СТАНОВИСЬ ДАТА  
САЕНТИСТОМ КАК Я

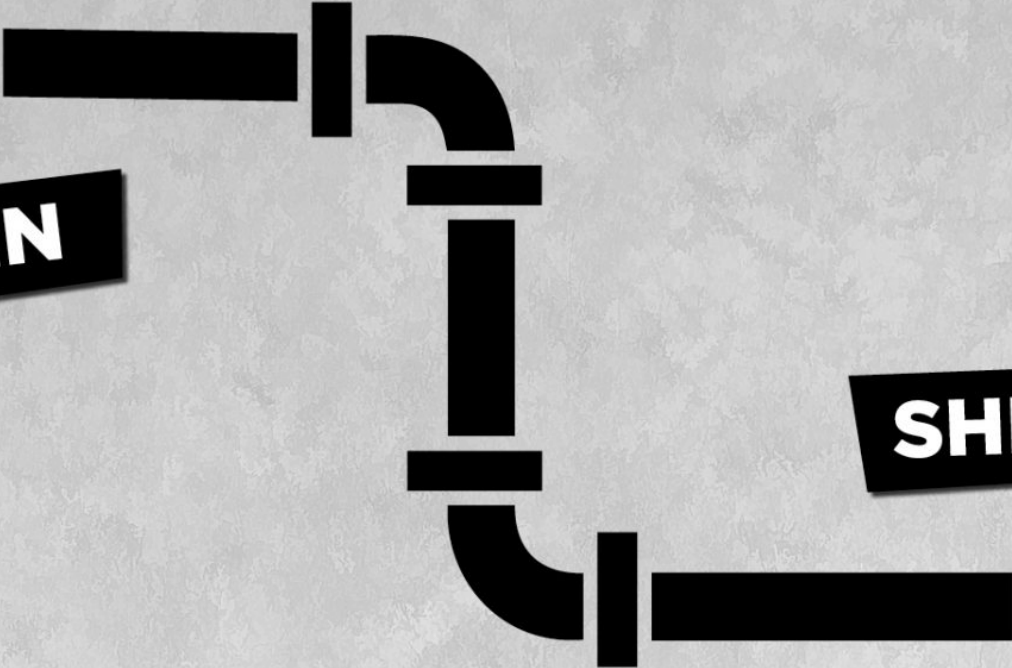


ТЫ ЧЁ ПЁС,  
Я МАТЕМАТИК

ТАК ТЫ ЖЕ ПРОСТО РАНДОМНО  
ПОДБИРАЕШЬ КОЭФФИЦИЕНТЫ  
ПОКА КРОСС-ВАЛИДАЦИЯ  
НЕ ДАСТ НОРМАЛЬНЫЙ РЕЗУЛЬТАТ





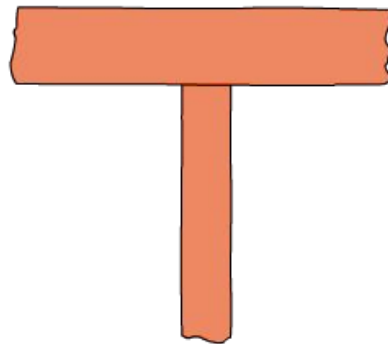


**SHIT IN**

**SHIT OUT**



# Генерализация и специализация



Единорог!





# Немного истории



**Раньше было лучше**



**Раньше было лучше**



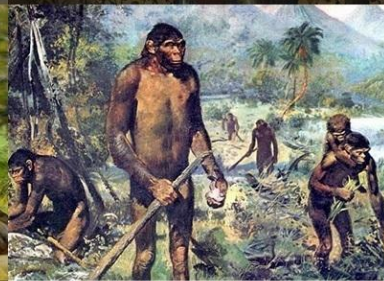
**Раньше было лучше**



**Раньше было лучше**

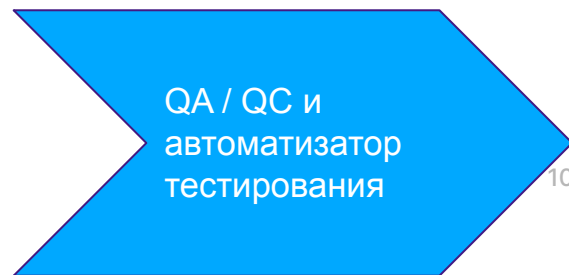
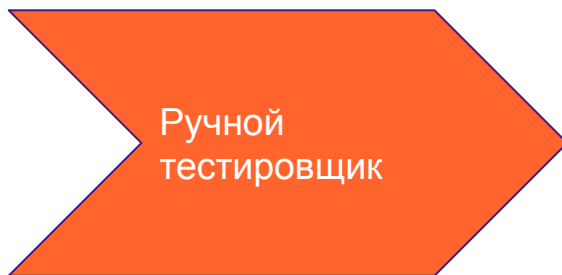
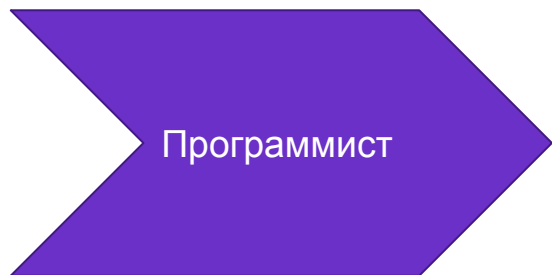


**Раньше было лучше**

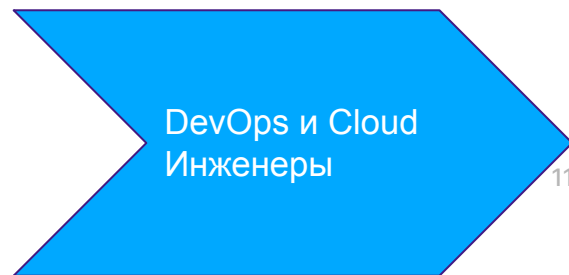
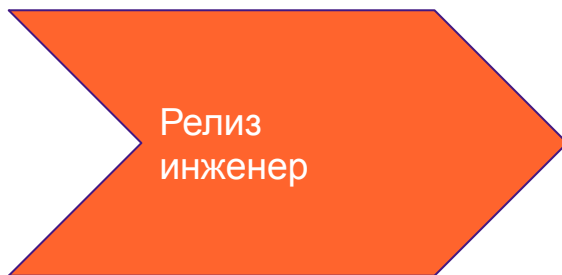
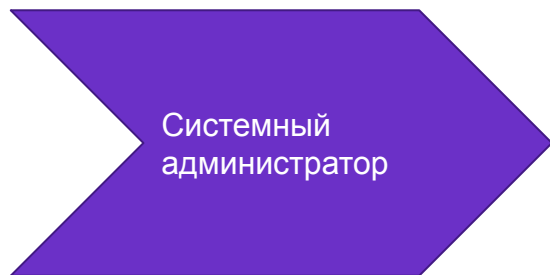


**Во! Теперь ништяк!**

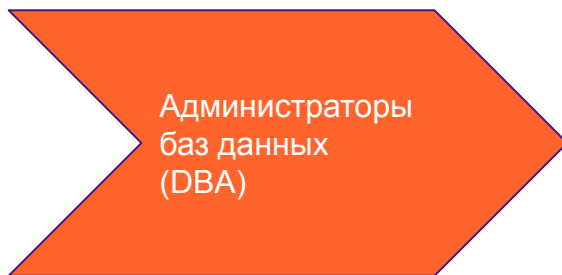
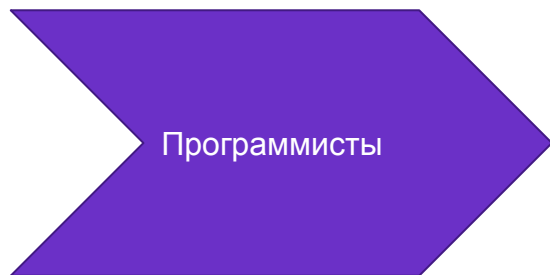
# Помните, были такие...



# А еще были...



# История повторяется..



# Почему так происходит?



- Появляются новые возможности
- С новыми возможностями приходят новые проблемы
- Для решения новых проблем нужны новые подходы и инструменты



# Почему так происходит?



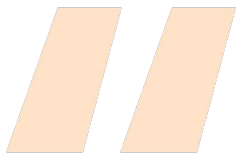
- Инструментов, подходов и задач становится все больше
- Это все тяжело уместить в голову одного специалиста
- Появляется специализация на решение определенного круга новых задач

# Да придут Дата Инженеры





# Так ли это?



Новое модное название для разработчика хранилищ?

Поддерживает наши базы данных и грузит туда из разных источников.

Обслуга для DS

Витриноделы!

Если я приложение написал многопоточное, которое данные туда-сюда гоняет, я DE или Разработчик?

# Кто такой дата инженер?

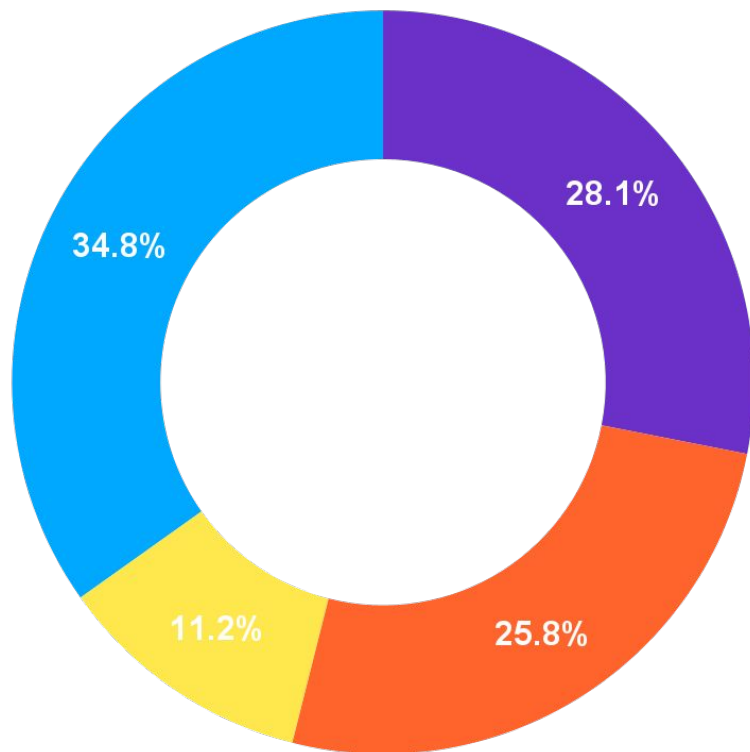


Инженер, который специализируется на работе с системами хранения и загрузки данных.

**Семён Осипов**

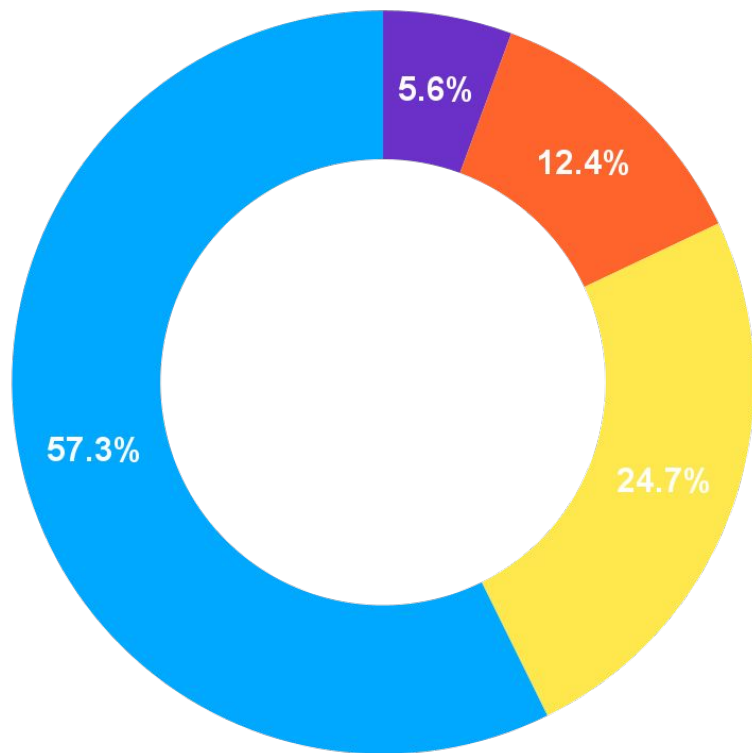
Semrush, Data Engineer

# Опыт в IT до прихода в DE



- От 0 до 1 года
- От 1 до 3 лет
- От 3 до 5 лет
- Больше 5 лет

# Размер компании, в которой работает DE



- До 50 человек
- От 50 до 150
- От 150 до 1000
- Больше 1000

# Топ 5 навыков для DE



Если бы вы сейчас нанимали DE, что бы вы искали в кандидате?





**INFRASTRUCTURE**

**STORAGE**  
 IBM Cloud Storage, Amazon S3, Microsoft Azure Storage, Oracle Cloud Storage, NetScout Systems, Pure Storage, Dell EMC, Veeva, NetScout Systems, Pure Storage, Dell EMC, Veeva

**HADOOP**  
 Cloudera, Hadoop, Amazon EMR, Microsoft Azure HDInsight, Oracle Cloud Infrastructure, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA LAKES**  
 Amazon Redshift, Microsoft Azure Synapse Analytics, Oracle Autonomous Data Warehouse, Snowflake, Databricks, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA WAREHOUSES**  
 Amazon Redshift, Microsoft Azure Synapse Analytics, Oracle Autonomous Data Warehouse, Snowflake, Databricks, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**STREAMING / IN-MEMORY**  
 Amazon Kinesis, Microsoft Azure Stream Analytics, Oracle Streaming Data Integration, Snowflake, Databricks, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**ANALYTICS & MACHINE INTELLIGENCE**

**BI PLATFORMS**  
 Looker, Tableau, Microsoft Power BI, Qlik, Alteryx, SAP Analytics Cloud, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**VISUALIZATION**  
 Tableau, Microsoft Power BI, Qlik, Alteryx, SAP Analytics Cloud, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA ANALYTICS PLATFORMS**  
 Microsoft Azure Synapse Analytics, Oracle Autonomous Data Warehouse, Snowflake, Databricks, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**APPLICATIONS - ENTERPRISE**

**SALES**  
 Salesforce, Microsoft Dynamics 365, Oracle Sales Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**MARKETING - B2B**  
 Adobe Campaign, Oracle Marketing Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**MARKETING - B2C**  
 Adobe Campaign, Oracle Marketing Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**CUSTOMER EXPERIENCE / SERVICE**  
 Oracle Service Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**HUMAN CAPITAL**  
 Oracle HCM Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**NOSQL DATABASES**  
 Amazon DynamoDB, Microsoft Azure Cosmos DB, Oracle NoSQL Database, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**NewSQL DATABASES**  
 Amazon Redshift, Microsoft Azure Synapse Analytics, Oracle Autonomous Data Warehouse, Snowflake, Databricks, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**GRAPH DBs**  
 Amazon Neptune, Microsoft Azure Cosmos DB, Oracle NoSQL Database, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**MPP DBs**  
 Amazon Redshift, Microsoft Azure Synapse Analytics, Oracle Autonomous Data Warehouse, Snowflake, Databricks, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**SERVER-LESS**  
 Amazon Redshift, Microsoft Azure Synapse Analytics, Oracle Autonomous Data Warehouse, Snowflake, Databricks, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**CLUSTER SWKS**  
 Amazon Redshift, Microsoft Azure Synapse Analytics, Oracle Autonomous Data Warehouse, Snowflake, Databricks, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA SCIENCE NOTEBOOKS**  
 Databricks, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA SCIENCE PLATFORMS**  
 Databricks, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**MACHINE LEARNING**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**LEGAL**  
 Lexipol, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**REGTECH & COMPLIANCE**  
 Lexipol, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**FINANCE**  
 Lexipol, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**AUTOMATION & RPA**  
 Lexipol, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**SECURITY**  
 Lexipol, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**ETL / DATA TRANSFORMATION**  
 Informatica, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA INTEGRATION**  
 Informatica, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA GOVERNANCE**  
 Informatica, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA QUALITY**  
 Informatica, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**COMPUTER VISION**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**HORIZONTAL AI**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**SPEECH & NLP**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**ADVERTISING**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**EDUCATION**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**REAL ESTATE**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**GOVT & INTELLIGENCE**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**COMMERCE**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**FINANCE - LENDING**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**INSURANCE**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**MGMT / MONITORING**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA GENERATION & LABELLING**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**AI OPS**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**GPU DBs & CLOUD**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**AI HARDWARE**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**SEARCH**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**LOG ANALYTICS**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**SOCIAL ANALYTICS**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**WEB / MOBILE / COMMERCE ANALYTICS**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**HEALTHCARE**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**LIFE SCIENCES**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**TRANSPORTATION**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**AGRICULTURE**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**INDUSTRIAL**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**OTHER**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**OPEN SOURCE**

**FRAMEWORKS**  
 TensorFlow, PyTorch, Keras, Apache Spark, Hadoop, Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**QUERY / DATA FLOW**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA ACCESS & DATABASES**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**ORCHESTRATION & PIPELINES**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**STREAMING & MESSAGING**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**STAT TOOLS & LANGUAGES**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**AI OPS & INFRA**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**AI / MACHINE LEARNING / DEEP LEARNING**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**SEARCH**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**LOGGING & MONITORING**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**VISUALIZATION**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**COLLABORATION**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**SECURITY**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA MARKETPLACES & DISCOVERY**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**FINANCIAL & ECONOMIC DATA**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**AIR / SPACE / SEA**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA SOURCES & APIs**

**PEOPLE / ENTITIES**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**LOCATION INTELLIGENCE**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**OTHER**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA SERVICES**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**DATA RESOURCES**

**INCUBATORS & SCHOOLS**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

**RESEARCH**  
 Amazon SageMaker, Microsoft Azure Machine Learning, Oracle Analytics Cloud, IBM Watson Analytics, NetScout Systems, Pure Storage, Dell EMC, Veeva

# Разработчик DWH



## Задачи

Разработка хранилища данных внутри компании (таблицы, слои и витрины данных)

## Типовой набор навыков

SQL  
Язык программирования  
Базы данных

## Когда нужен

Когда компания решает строить платформу данных:  
Data Lake, Data Warehouse, Data Lakehouse, Data Swamp и т. д. 😊



# Разработчик ETL



## Задачи

Получение и хранение данных из различных источников.

## Типовой набор навыков

SQL  
Язык программирования  
Базы данных  
Оркестрация запусков

## Когда нужен

Когда нужно автоматизировать загрузку данных

# Инфраструктурный инженер



## Задачи

Поддержание и масштабирование систем получения, обработки и хранения данных.

## Типовой набор навыков

Стек Big Data (Hadoop / Spark)  
Базы данных  
Облака

## Когда нужен

Когда данных становится настолько слишком много.

# ML инженер



## Задачи

Вывод в прод,  
поддержание и  
масштабирование систем  
и моделей машинного  
обучения

## Типовой набор навыков

Стек ML (PyTorch,  
TensorFlow)  
Kubernetes  
Docker  
Язык программирования

## Когда нужен

Когда важна скорость и  
стабильность работы ML  
моделей

# Data Quality инженер



## Задачи

Отслеживание качества данных для работы всех систем

## Типовой набор навыков

SQL  
Язык программирования  
Системы мониторинга

## Когда нужен

Когда обожглись, работая без проверки данных 😊

# Analytics Engineer



## Задачи

Создание и развитие инструментов и платформ для аналитики

## Типовой набор навыков

Комбинация навыков DE + Аналитик + DS

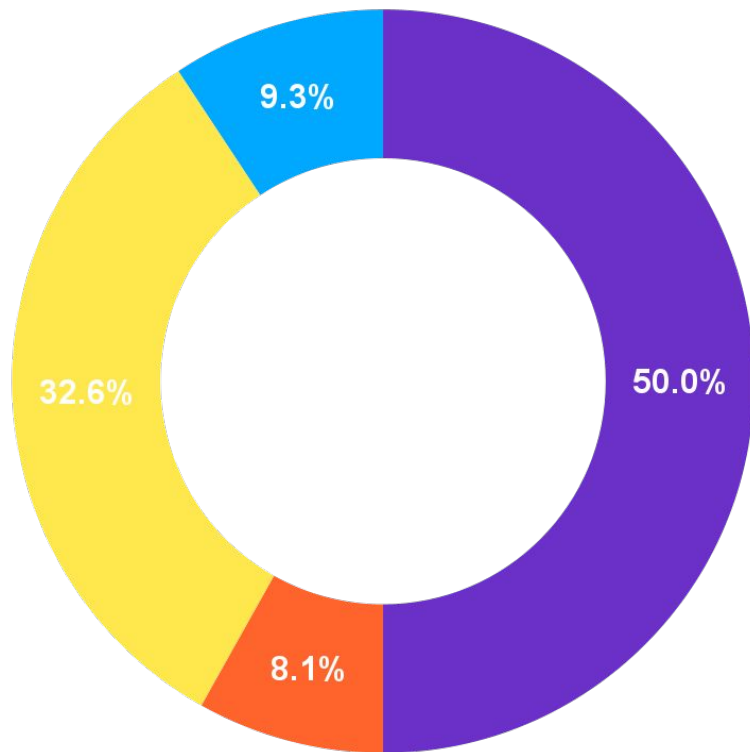
## Когда нужен

Тот самый единорог 😊

Главный холивар - язык программирования! 



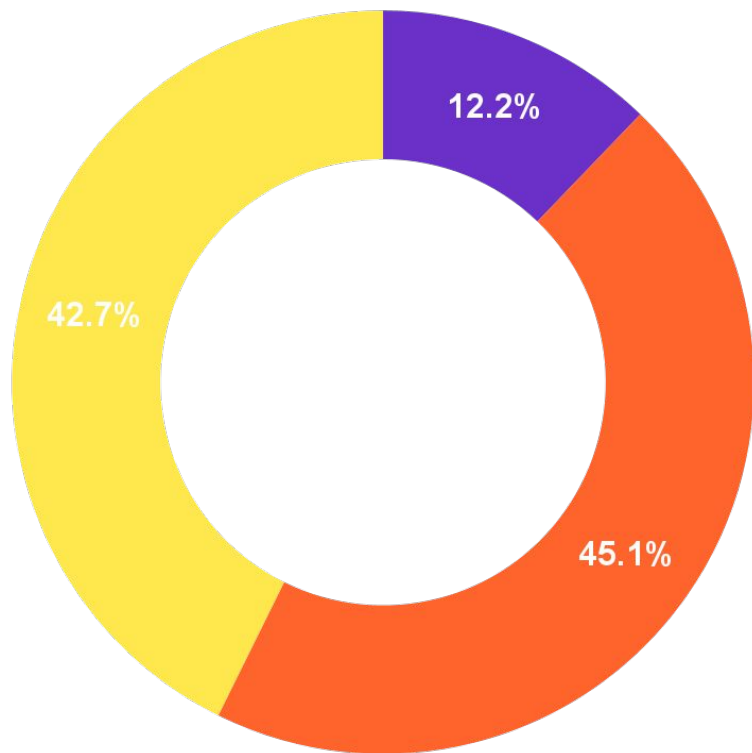
# Главный холивар - язык программирования!



- Python
- Java
- Scala
- SQL



# Но язык программирования не главное



- Язык программирования
- Тулинг и инфраструктура
- SQL

# Что же делать компаниям?



1. Понять, кто им действительно нужен
2. Не пихать все возможные технологии в одно описание вакансии
3. Развивать культуру изнутри, обучать разработчиков или выращивать джунов

# Где искать?



1. Очевидные места: HH и LinkedIn
2. ODS, #jobs / #jobs\_hr канал
3. TG @hadoopusers, а точнее его ответвление @DataJobs

# Задать вопрос?



@OsipovSimon



@SimonOsipov