Лекция_1

Введение в предмет Big data в цифровых медиа

Big Data: с чего начать

Каждый обмен с социальными медиа, каждый цифровой процесс, каждое подключённое устройство генерирует большие данные, которые будут использоваться различными компаниями.

Сегодня компании используют Big Data для углубленного взаимодействия с клиентами, оптимизации операций, предотвращения угроз и мошенничества. За последние два года такие компании, как IBM, Google, Amazon, Uber, создали сотни рабочих мест для программистов и Data science.

Область больших данных слишком размылась на просторах интернета, и это может быть очень сложной задачей для тех, кто начинает изучать большие данные и связанные с ними технологии. Технологии данных многочисленны это может быть огромным препятствием для начинающих. Давайте попробуем разложить все по полочкам.

1. Как начать

В сфере Big Data существует много направлений. Но в широком смысле можно разделить на две категории:

1. Big Data engineering.
2. Big Data Analytics (Scientist).

Эти поля взаимозависимы, но отличаются друг от друга.

Big Data engineering занимается разработкой каркаса, сбора и хранения данных, а также делают соответствующие данные доступными для различных потребительских и внутренних приложений.

У вас хорошие навыки программирования и вы понимаете, как компьютеры взаимодействуют через интернет, но у вас нет интереса к математике и статистике. В этом случае вам больше подойдёт Big data engineering.

В то время как **Big Data Analytics** — среда использования больших объемов данных из готовых систем, разработанных Big data engineering. Анализ больших данных включает в себя анализ тенденций, закономерностей и разработку различных систем классификации и прогнозирования. После магических действий и танцев с бубном Data Analytics (Scientist) интерпретирует результаты.

Если вы хорошо разбираетесь в программировании, за чашкой кофе решаете сложные задачи по высшей математике, понимаете, что такое теория вероятностей, математический анализ, комбинаторики, тогда вам подойдёт Big Data Analytics.

Таким образом, **Big data Analytics** включает в себя расширенные вычисления по данным. В то время как **Big data engineering** включает проектирование и развертывание систем, над которыми должны выполняться вычисления.

Как стать специалистом по большим данным

С направлением определились, теперь давайте разберём, что должен знать Data science, чтобы его рассматривали в качестве будущего кандидата.

Терминология данных

Проект с большими данными имеет два основных понятия — требования к данным и требования их обработке.

Требования к данным

Структурированные данные: хранятся в таблицах или в файлах. Если данные хранятся в предопределённой модели данных (то есть в схемах), это называется структурированными данными.

Неструктурированные: если данные хранятся в файлах и не имеют предопределённой модели, это называется неструктурированными данными.

Источники данных: внутренние (CRM, ERP или любые источники, которые находятся внутри системы) и внешние (соцсети, интернет).

Размер: с размером мы оцениваем количество данных. Типы: S, M, L, XL, XXL, передача потоков.

Пропускная способность: определяет, с какой скоростью данные могут быть приняты в систему. Типы: H, M, L.

Пропускная способность источника: определяет, с какой скоростью данные могут быть обновлены и преобразованы в систему. Типы: H, M, L.

Требования к обработке данных

Время запроса: время, за которое система выполняет запрос. Типы: Long, Medium, Short.

Время обработки: время обработки данных. Типы: длинный, средний, короткий.

Точность: точность обработки данных. Типы: точные или приблизительные, Exact или Approximate.

Учимся проектировать решения

Пример.

Задача — разработать Data lake для эффективного анализа продаж банка.

Данные берём из разных источников.

Внутренние:

- ERP (персональная информация о клиенте, данные о кредитной истории, данные о потенциальных клиентах);
- CRM (данные от колл-центра, данные о продажах,) данные о продуктах, транзакции проведенные через банковскую систему, CRM системы.

Внешние:

- социальные сети (BDSMM);
- интернет;
- веб-аналитика.

Важно понимать, что первым делом нужно рассчитывать, что система должна быть интегрирована со всеми вышеперечисленными источниками и бесперебойно принимать данные.

Определяем конечные цели:

1. Создаем Data lake для объединения данных из нескольких источников.
2. Автоматическое обновление данных через определённые промежутки времени.
3. Доступность данных для анализа (круглосуточно, возможно ежедневно).
4. Архитектура для лёгкого доступа к панели инструментов аналитики.

Теперь, когда мы знаем, каковы наши конечные цели, попробуем сформулировать наши требования более формальными терминами.

Требования к данным

Структура: большая часть данных структурирована и имеет определённую модель. Но источники данных, такие как веб-журналы, взаимодействия с клиентами или данные колл-центра, изображения из каталога продаж, данные рекламы продукта —доступность и требования к изображениям и мультимедийной рекламной информации могут зависеть от компании.

Тип данных: структурированные и неструктурированные данные.

Размер: L или XL.

Пропускная способность: высокая.

Качество: средний.

Полнота данных: неполная.

Требования к обработке

Время запроса: от среднего до длинного.

Время обработки: от среднего до короткого.

Точность: точное.

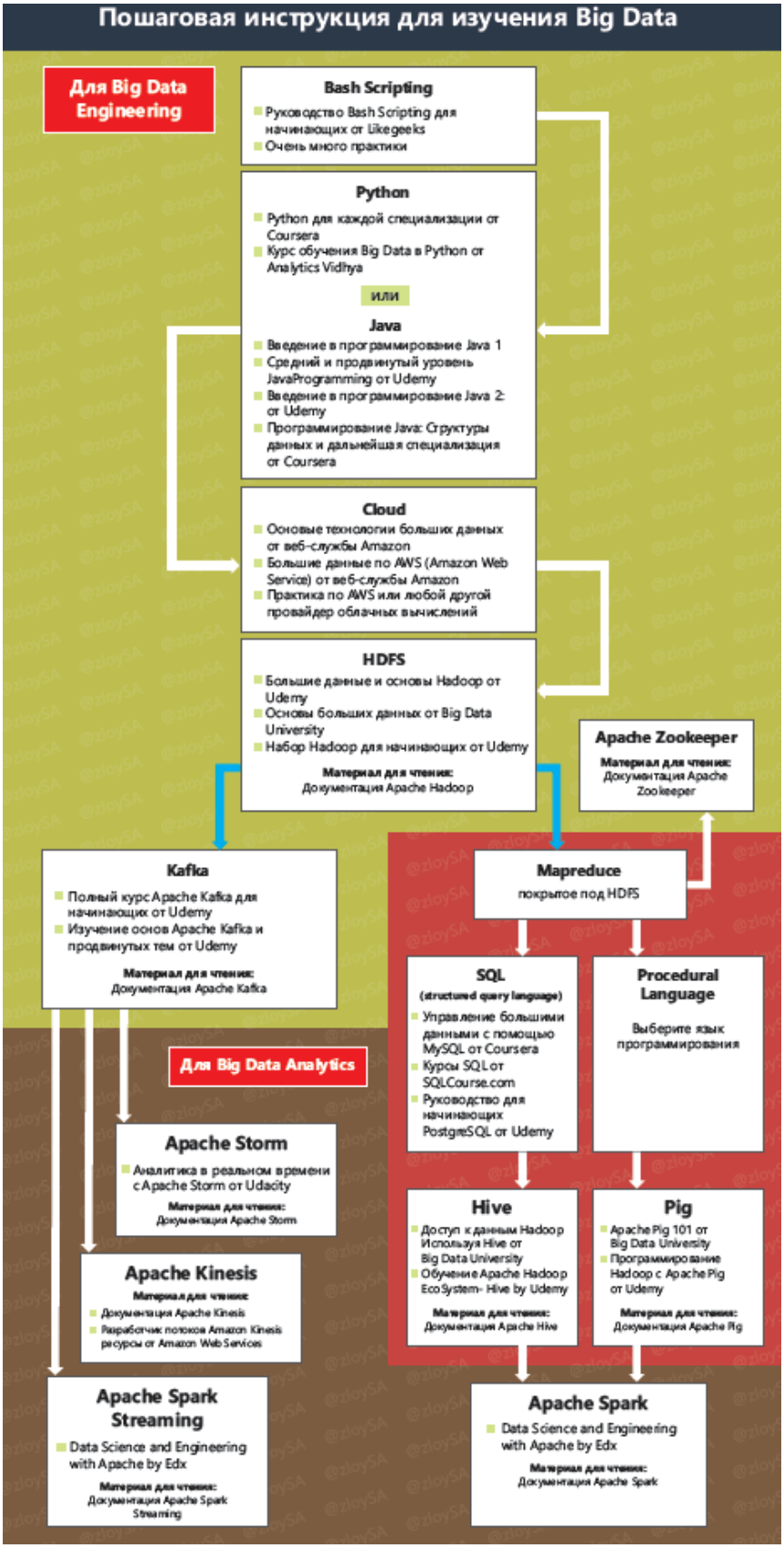
Поскольку несколько источников данных интегрируются, важно отметить, что разные данные поступают в систему с разной скоростью. Например, данные из веб-аналитики будут доступны в непрерывном потоке с высокой степенью детализации.

Основываясь на приведённом выше анализе наших требований к системе, мы можем порекомендовать следующую настройку данных.

Путь обучения работе с большими данными

Теперь рассмотрим, по какой цепочке вам нужно пройти.

Область Big Data разбита на разные технологии. Очень важно, чтобы вы изучали релевантные и совместимые технологии с вашим направлениям работы с данными. Это немного отличается от таких направлений, как машинное обучение, где вы начинаете что-то и пытаетесь завершить всё в этой области.



Одна из основных концепций, которые должен знать любой чувак, который видит себя в этой области, развертывание сервера на Linux, написание скриптов в командной строке Bash Scripting. Это необходимое условие для работы с Big Data.

В основном большая часть технологий данных написана на Java или Scala. Не переживайте, если вы не хотите кодить на этих языках, вы можете выбрать Python или R, потому что большая часть технологий обработки больших данных теперь поддерживает Python и R.

Вы можете начать свой путь с изучения любого из вышеуказанных языков. Я рекомендую выбрать Python или Java.

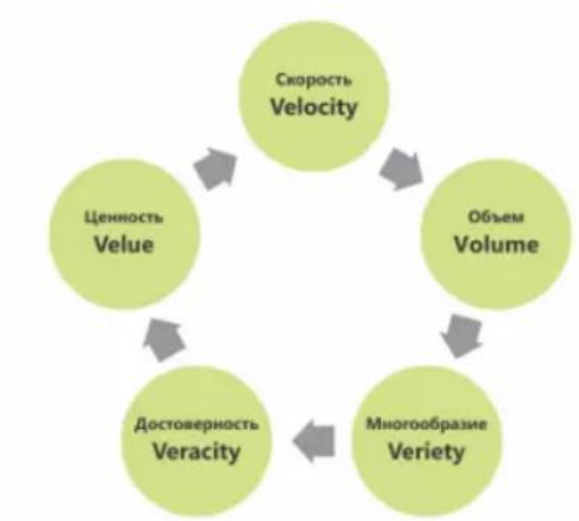
Как мы видели на примере выше, данные могут распределяться на большое количество серверов. Поэтому для быстрого нахождения пакетов и выгрузки массивов потребуется написанию SQL-запросов. Без базового знания SQL вряд ли получится устроиться даже на позицию junior.

Также не помешает поработать с облаком. Попробуйте использовать небольшие наборы данных на AWS, Softlayer или любом другом облачном провайдере. Большинство из них имеют свободный уровень, чтобы учащиеся могли практиковаться. Вы можете пропустить этот шаг сейчас, если хотите, но не забрасывайте в тёмный угол работу с облаком, прежде чем идти на какое-либо собеседование.

Затем вам нужно узнать о распределённой файловой системе. Наиболее популярной DFS является файловая система Hadoop. На этом этапе вы также можете изучить некоторые базы данных NoSQL.

Путь до сих пор является обязательным основанием, которое должен знать каждый специалист.

Теперь решайте, хотите ли вы работать с потоками данных. Это выбор между двумя из четырех V, которые используются для определения больших данных — Volume, Velocity, Variety и Veracity.



Характеристика	Традиционная база данных	База Больших Данных
Объем информации	От гигабайт до терабайт	От петабайт до эксабайт
Способ хранения	От гигабайт до терабайт	От петабайт до эксабайт
Структурированность	Структурирована	Полуструктурирована или неструктурирована
Модель хранения и обработки данных	Вертикальная модель	Горизонтальная модель
Взаимосвязь данных	Сильная	Слабая

Предположим, вы решили работать с потоками данных для разработки систем анализа в реальном времени. Тогда вы должны пойти по ветке Apache Kafka с помощью Mapreduce. Обратите внимание, что в пути Mapreduce вам не нужно изучать Pig и Hive. Достаточно изучить только один из них.

Это не единственный способ получить знания. Вы можете создать свой собственный путь по ходу дела самостоятельно.