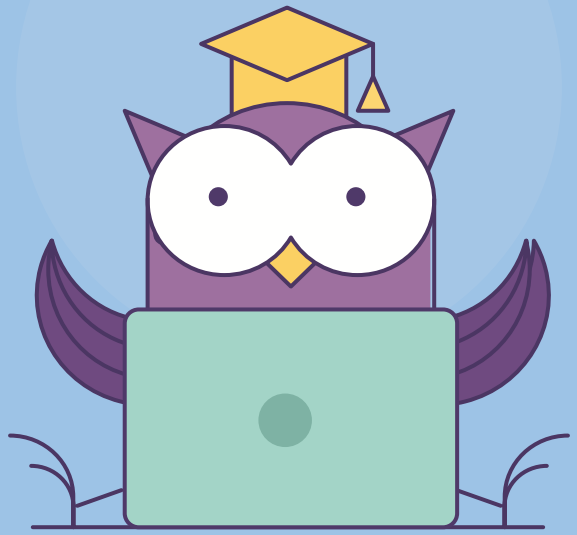
Инженер Данных.
Потребность, навыки, инструменты.

Меня хорошо слышно && видно?



Напишите в чат, если есть проблемы!

Ставьте + если все хорошо
Ставьте - если есть проблемы

Знакомимся: Артемий Козырь

- PwC, Московская Биржа, Сбербанк, СИБУР
- DWH, Reporting, Retail Scoring, Customer segmentation, Next Best Offer, IIoT
- Темы: Intro, Formats, DWH, Notebooks, DQ
- Образование: НИУ ВШЭ

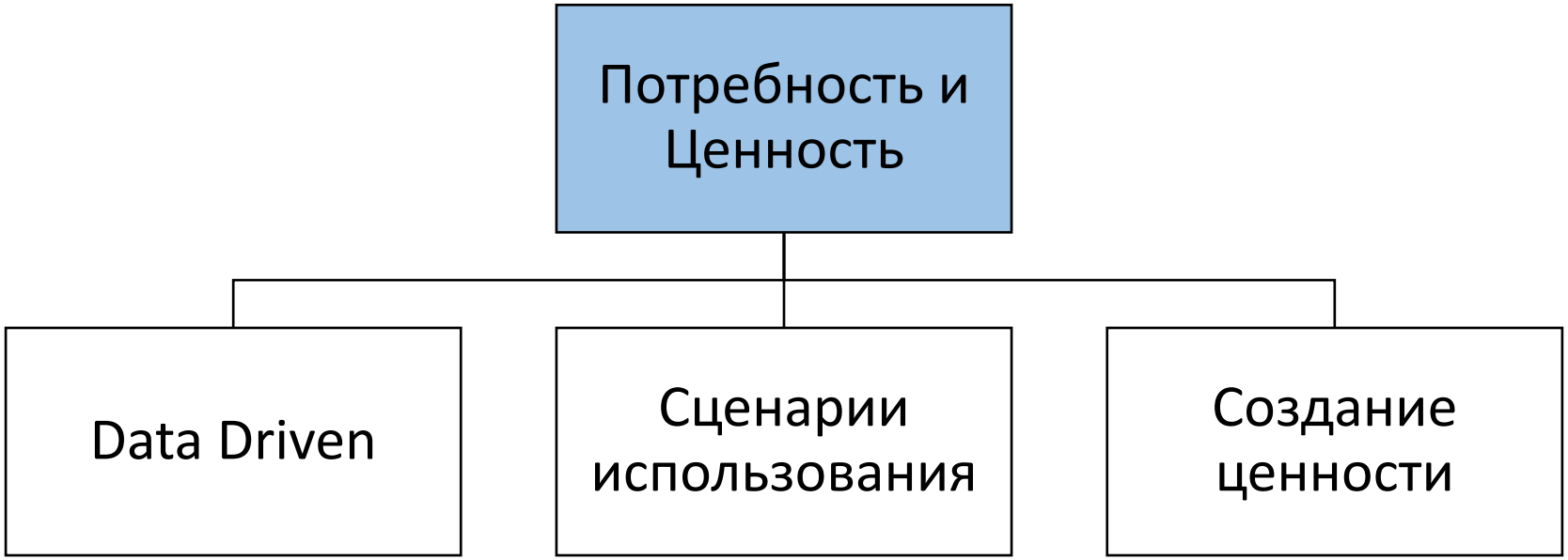
Знакомимся

- Где работаете
- Чем занимаетесь
- Какие из тем знакомы
- Уровень владения
- На чем хотелось бы сделать акцент
- Ожидания от курса в целом

План занятия

Потребность и Ценность	Навыки, Задачи, Инструменты	Приносим пользу	ДЗ + Рефлексия

Потребность и Ценность



Data Driven

- Данные – информация – знание
- Обоснованные и взвешенные решения
- Конкурентные преимущества

Базовые вопросы

- Какую задачу хотим решить?
- Какие данные нужны?
- Что уже есть у компании?
- Как можем получить?
- Как использовать?

Сценарии использования

- **Производство и добыча**: IIoT, Отчетность, Продажи
- **Ритейл**: Продажи, CRM, Аналитика, Логистика
- **Банки**: Скоринг, Антифрод, CRM, Отчетность, РС*
- **Телеком**: CRM, РС*, Аналитика
- **Медиа и Веб**: Реклама, CRM, Аналитика

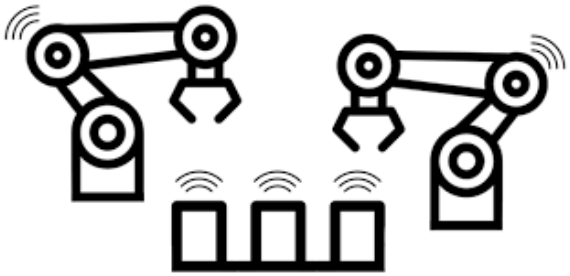
* Рекомендательные системы

Еще сценарии использования

- Выписки по счетам
- Управленческая отчетность, аналитика
- Мониторинг Media, Графы связей
- ClickStream, Веб-аналитика
- Real-time скоринг, антифрод, маркетинг
- Рекомендательные системы
- IIoT: Мониторинг, Прогнозирование

Производство и добыча

Оборудование + качество



План и факт + Рынок и конкуренты

Ритейл

Сегменты + Предпочтения

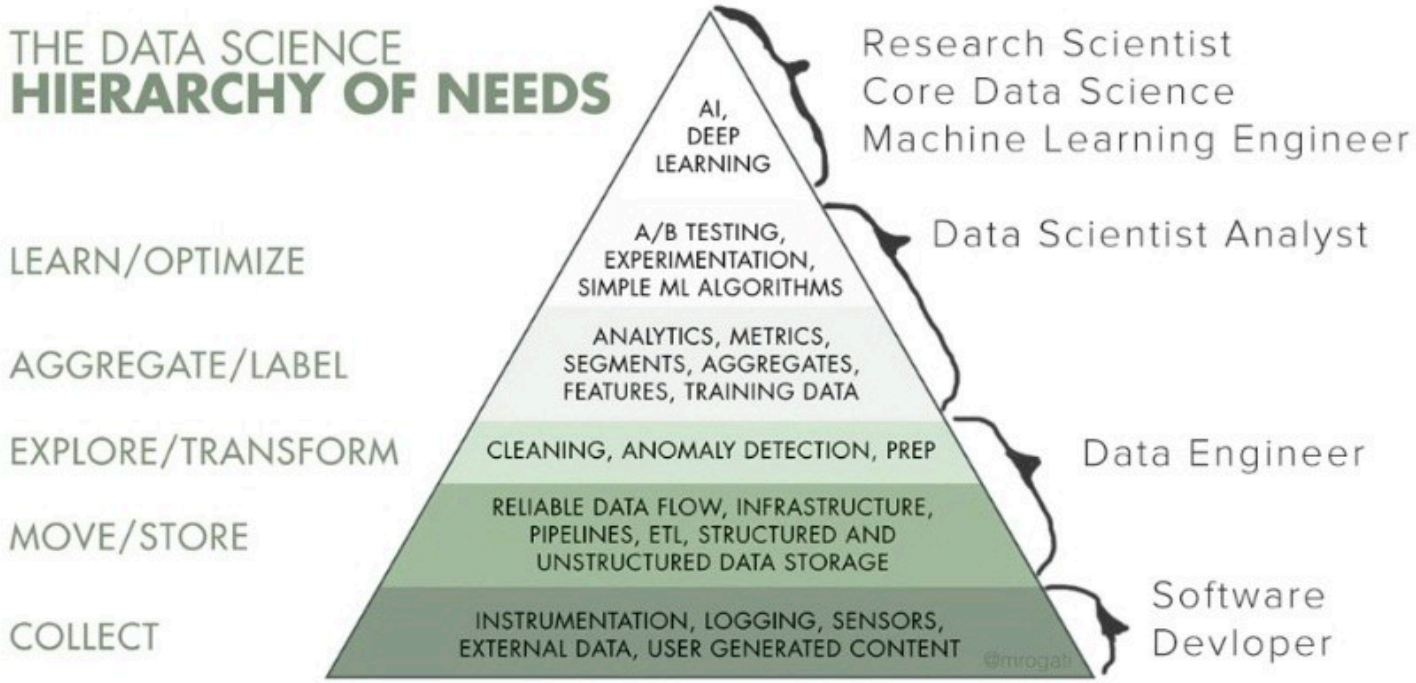
Логистика + запасы

Медиа и Веб

Индивидуальный таргет

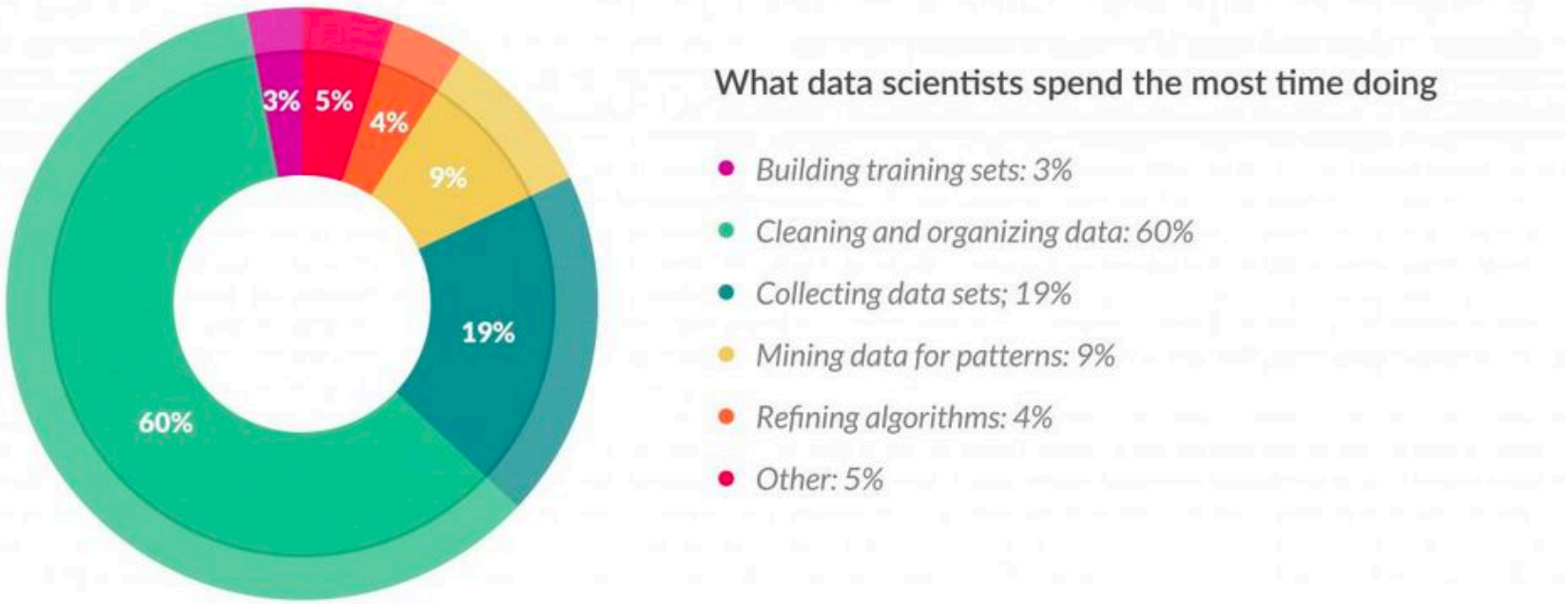
Рекомендации + UX

Иерархия потребностей DS



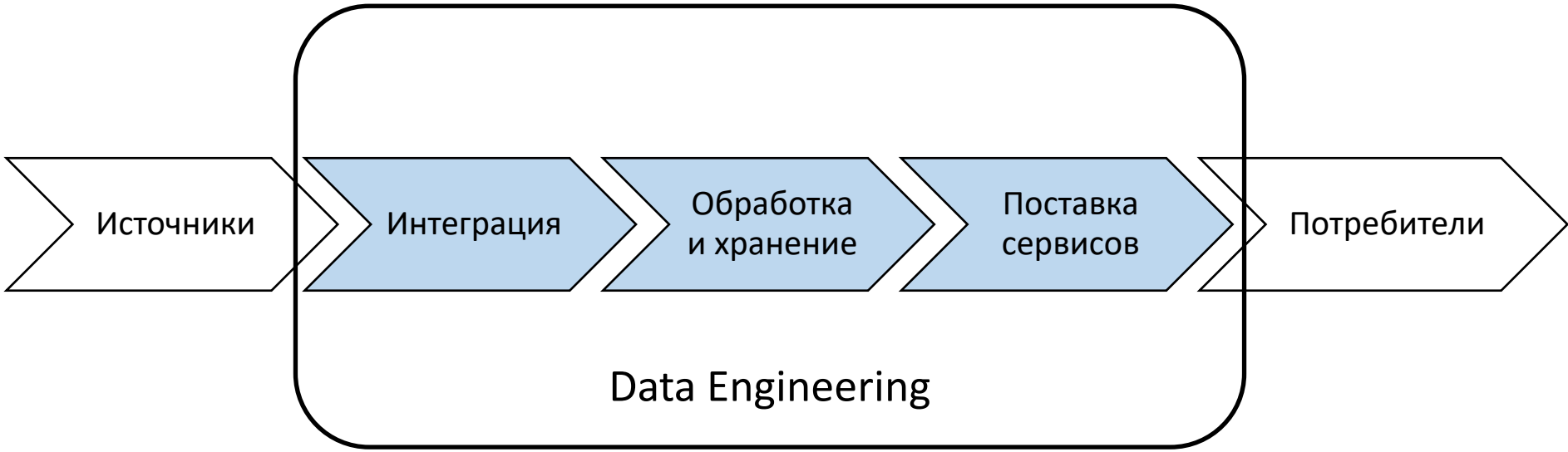
* [A Beginner's Guide to Data Engineering — Part I]

80% усилий тратится на сбор и подготовку*



* Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Создание ценности - Value chain



План занятия

Потребность и Ценность	Навыки, Задачи, Инструменты	Приносим пользу	ДЗ + Рефлексия

Навыки, Задачи, Инструменты

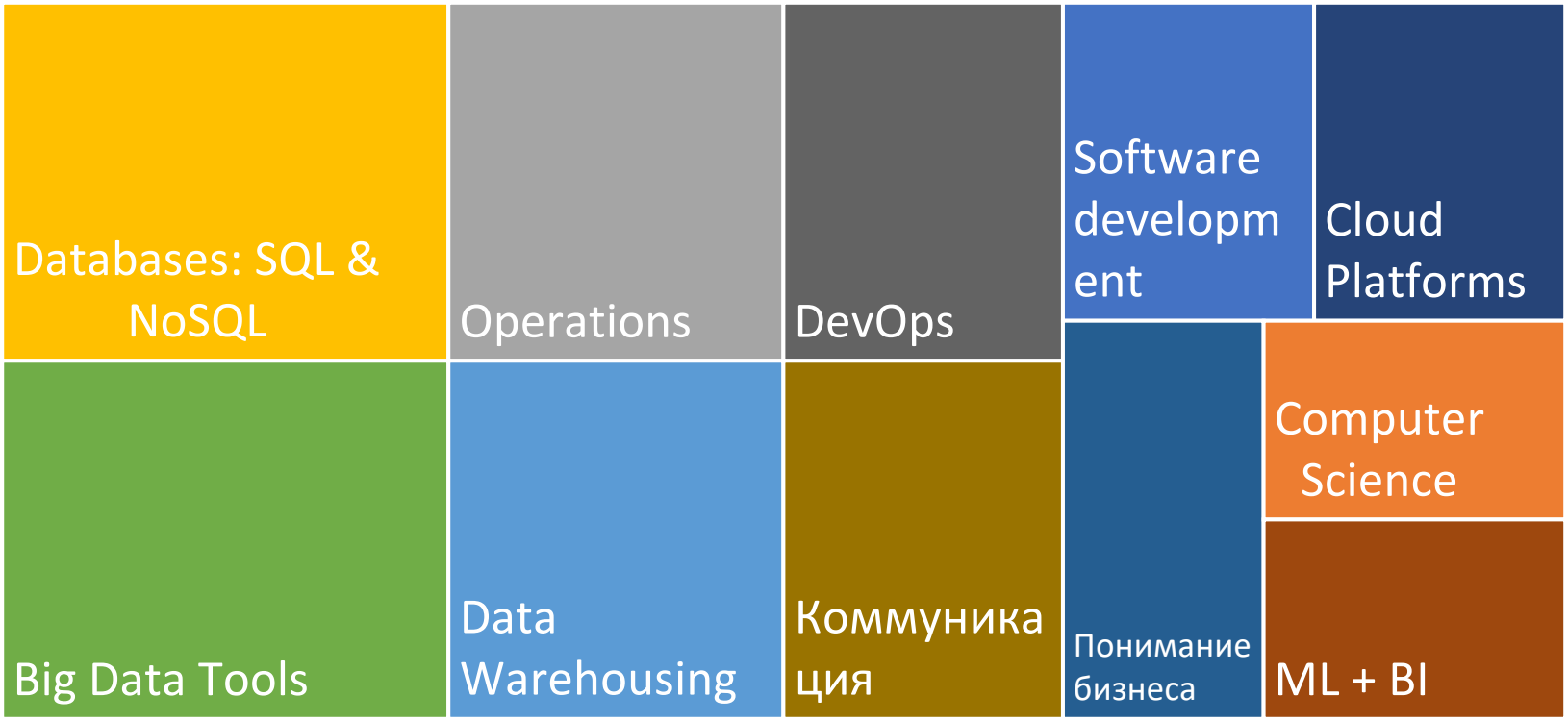
Задачи

- Интеграция
- Хранение
- Обработка (процессинг)
- Поставка **всем** заинтересованным сторонам
- Выбор инструментов, фреймворков и т.д.
- Архитектура и инфраструктура решений

Задачи +

- Выделять дельту (инкремент) из источника
- Обеспечить оптимальное хранение данных
- Подготовить логическую и физическую модель
- Измерить качество данных
- Найти бутылочное горлышко
- Масштабировать решения

Карта навыков



Пример. Базы данных

- Реляционная алгебра + Структуры
- SQL
- Транзакционные алгоритмы: Concurrency control, Recovery, Distributed transactions
- Concurrency control: 2PL, MVCC, Deadlock Detection
- Recovery: Write ahead log, Redo, Undo
- Distributed transactions: 2PC, Distributed Recovery

* Как стать классным спецом по базам данных? / Илья Космодемьянский (Data Egret)

Пример. Мягкие навыки

- Эффективная коммуникация
- Понимать, что говорит и хочет собеседник
- Уметь донести свою мысль
- Делать это на русском и английском языках
- Понимать задачи, которые решает бизнес
- Использовать опыт коллег

Откуда брать данные

- Файлы: CSV, XML, XLS, JSON ...
- Базы данных: SQL, CDC
- Очереди сообщений: ESB, IIoT, ClickStream
- Веб-сервисы: REST, SOAP

Структурированные данные

SELECT * FROM DM_OPERSVOD_STG.OPER_DDS | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	DT	PLANT_PRODUCT_ID	BP	PLAN	PPR	FORECAST	FACT
1	2019-05-01	[MENGE_01_1]	15,240.91600	16,162.41100	16,211.42700	[NULL]	15,382.33000
2	2019-05-01	[MENGE_01_4-_01-M]	147.41600	152.37600	156.47700	[NULL]	143.81100
3	2019-05-01	[MENGE_02_1-_01-P]	17,164.78500	18,033.13500	17,875.37500	[NULL]	16,988.40200
4	2019-05-01	[MENGE_02_1-_01-V]	2,190.42700	2,275.55400	2,257.64500	[NULL]	2,106.97700
5	2019-05-01	[MENGE_02_1]	12,713.05200	13,512.29100	13,399.02300	[NULL]	13,703.00500
6	2019-05-01	[MENGE_02_2-_01-M]	3,058.71500	3,128.26000	3,128.24100	[NULL]	3,277.45200
7	2019-05-01	[MENGE_02_3-_01-P]	0.00000	0.00000	0.00000	[NULL]	28.06800
8	2019-05-01	[MENGE_02_4-_01-P]	7,868.51500	8,363.31600	8,297.60100	[NULL]	8,381.36300

Полу- и неструктурированные данные


```
{
    "message": "Success",
    "status_code": 200,
    "result": [
        {
            "id": "1",
            "gender_category": "Male"
        },
        {
            "id": "2",
            "gender_category": "Female"
        },
        {
            "id": "3",
            "gender_category": "Transgender"
        },
        {
            "id": "4",
            "gender_category": "Others"
        }
    ]
}
```


Institoris, Heinricus [Kramer, Heinrich]. Sprenger, Iacobus [Sprenger, Jakob].

Malleus maleficarum.

Институрис [Крамер], Генрих. Шпренгер, Якоб. Молот ведьм

Nurenberge, 1519

ЯЗЫК: Латинский

ГМИР №: РК-518

АВТОР: Institoris, Heinricus [Kramer, Heinrich]. Sprenger, Iacobus [Sprenger, Jakob].

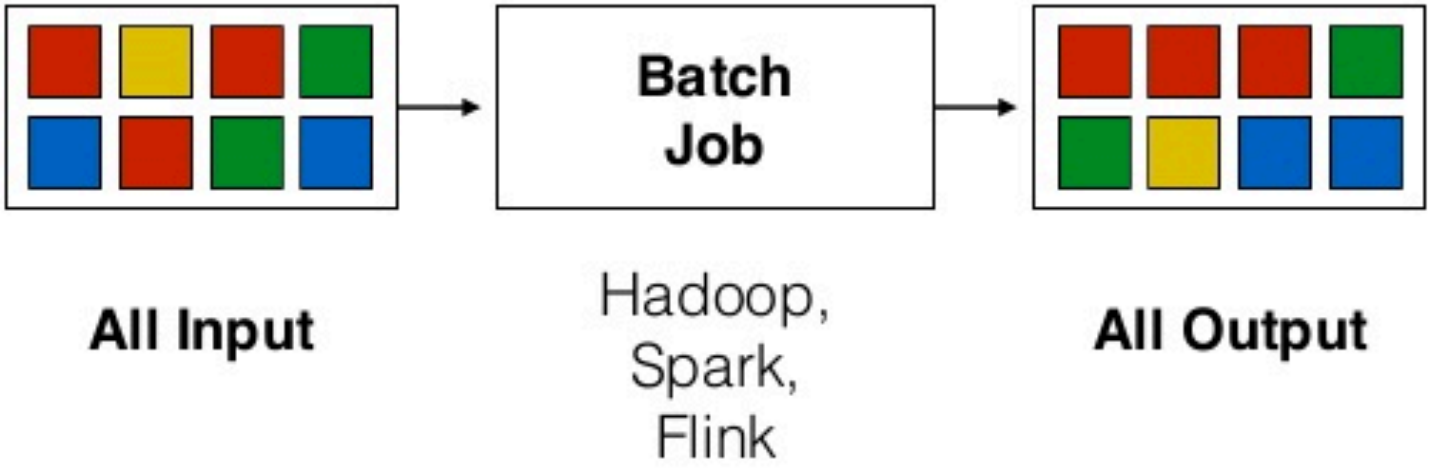
ИЗДАТЕЛЬ, ПЕЧАТНИК: Fredericus Peypus

МЕСТО: Nurenberge

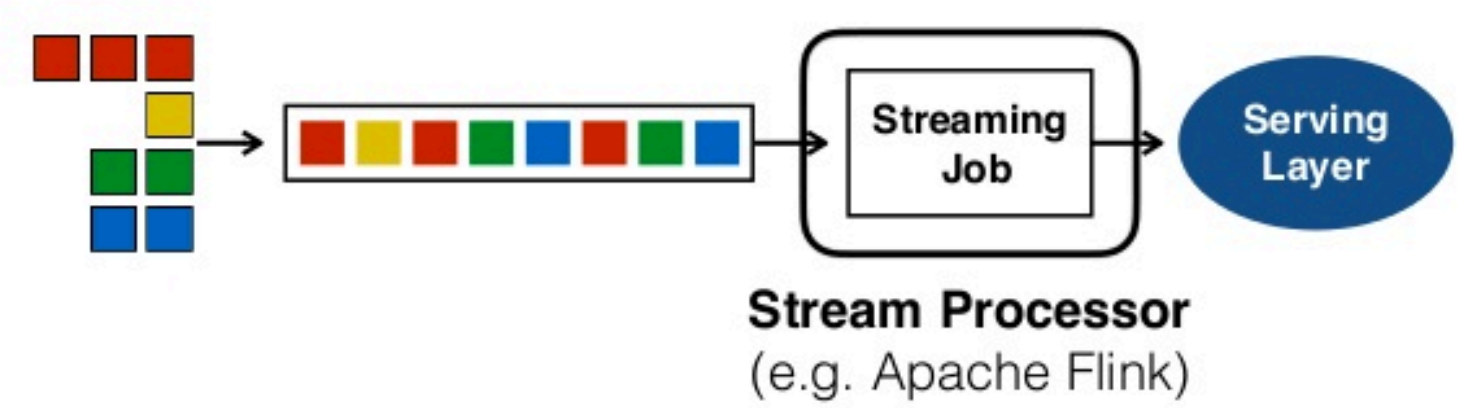
ГОД: 1519

ФОРМУЛА: Л. 1 нн. (тит), 2-152, 1-10 нн.лл.

Пакетная обработка - Batch



Потоковая обработка - Streaming



Batch vs Stream

	Batch	**Stream**
Скоуп	Процессинг целиком	Скользящее окно / самая свежая запись
Масштаб	Пакеты данных	Единичные записи или микро-батчи
Отклик	Минуты-часы	Секунды-мс
Возможности	Доступна сложная логика расчетов	Простые функции, агрегаты, метрики

Инструменты

1. Платформы, дистрибутивы
2. Загрузка и форматы данных. Data Ingestion
3. Очереди сообщений. Хранилища данных. NoSQL
4. Процессинг. Доступ к данным. ML
5. Оркестрация, Мониторинг, Data Quality

Платформы, дистрибутивы

1. Облачные платформы: Google Cloud Platform
2. Дистрибутив: Cloudera CDH

Загрузка и форматы данных

1. Распределенные файловые системы: HDFS
2. Инструменты выгрузки: StreamSets, Fluentd, Sqoop
3. Форматы данных: AVRO, Parquet, ORC

Очереди сообщений. Хранилища данных

1. Очереди сообщений: Kafka, Pub/Sub
2. DWH: BigQuery, Vertica, Clickhouse
3. NoSQL: Cassandra, Elasticsearch
4. SQL engine: Hive, Impala

Процессинг. Доступ к данным. ML

1. Apache Spark
2. Spark Streaming
3. Spark MLlib
4. Ноутбуки: Jupyter, Datalab

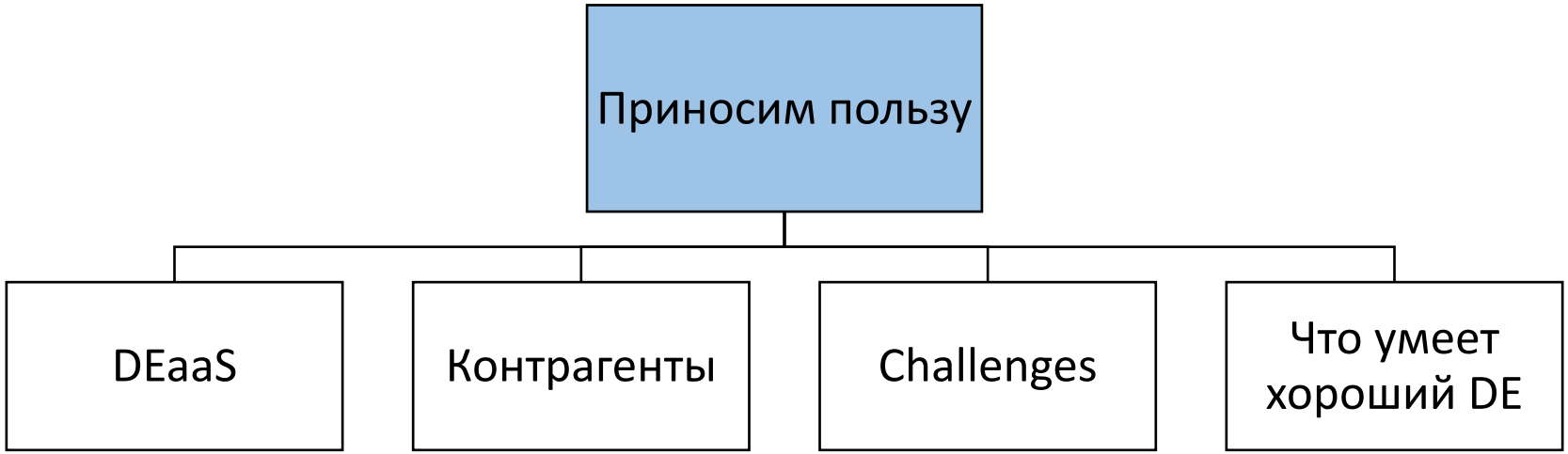
Обеспечивающие системы

1. Оркестрация: Airflow
2. DevOps: Git, Jenkins, Ansible
3. Мониторинг: Prometheus, Graphite, Grafana

План занятия

Потребность и Ценность	Навыки, Задачи, Инструменты	Приносим пользу	ДЗ + Рефлексия

Приносим пользу



DE as a Service

- Предоставлять востребованные и надежные дата-сервисы
- Знать своего клиента и его потребности
- Строить решения согласно потребностям

DE as a Service - Цикл

- Выяснить потребности клиента
- Гипотеза – предположение
- Валидация гипотезы
- Поставка минимальными итерациями

Контрагенты и команда

Challenges - Архитектура

- Интеграция систем
- Выстраивание пайплайнов
- Выбор правильных инструментов
- Оптимизация производительности
- Отказоустойчивость и надежность

Challenges – Коммуникация и процессы

- Понять потребность заказчика
- Если что-то важно для вас, это еще не значит, что это имеет ценность для клиента
- Постоянно меняющиеся требования
- Доступы и внутренняя безопасность

Challenges – Качество данных

- Влияет на качество решений
- Данные бывают corrupted
- Полнота, своевременность, консистентность

Что умеет хороший DE – Hard

- Выбрать правильные инструменты для задачи
- Benchmarking: эксперименты с настройками
- Диагностика: Know your data
- Идентифицировать бутылочное горлышко
- Автоматизировать всё

Что умеет хороший DE – Soft

- Уметь задавать вопросы и получать ответы
- Понять что от него требуется
- Инкрементальная поставка + Обратная связь
- Быстро читать и разбираться в доке

Что умеет хороший DE

Экзамен на DBA в идеальном мире

- На время починить базу
  - В идеале - увиденную в первый раз
  - В чем проблема непонятно, но "все торомозит"
- Экзаменаторов 10 (а лучше 20-30)
- 3 спрашивают "Ну как?" в Slak
- 3 спрашивают "Ну что?" по телефону
- 1 требует залогировать время
- 3 внедряют Scrum здесь и сейчас

* Как стать классным спецом по базам данных? / Илья Космодемьянский (Data Egret)

Домашнее задание

- Анализ рынка: Вакансии, Компании, Требования, Задачи, Инструменты
- География: РФ, USA, EU
- Понять для себя:
  - Где и чем вам хотелось бы заниматься
  - Какие технологии хотелось бы изучить
  - Поставить цели на обучение
- Источники: Glassdoor, HH, Monster, LinkedIn

Рефлексия

- Что вам запомнилось больше всего
- Пройти опрос

Ваши вопросы?