Типичные вопросы на собесах DE

В какой-то момент ты начинаешь уже стрелять, как из пушки на эти вопросы. Это может уже начинать подбешивать)

SQL и базы данных

1. Какие бывают виды джоинов логические (left, right, full...)?
2. Какие бывают физические джоины (hash join, nested loop join, merge sort join)?
3. Как работают физические джоины?
4. Как удалить дубликаты строк в SQL? (Например пронумеровать строки оконкой и удалить все, где больше 2) distinct во вложенном запросе или использовать UNION
5. Какие бывают оконные функции в SQL? **sum Count LAG LEAD dense rank rank cume_dist**
6. Как найти медиану в таблице? (можно с помощью оконных функций)
7. Написание SQL запроса для вывода клиентов, которые в прошлом месяце купили больше продуктов, чем за прошлый год и потратили менее 5000 рублей.

Модели данных и нормализация

1. Что такое нормализация данных? **1нф 2нф 3нф**
2. Какие бывают модели данных (вкратце)? снежинки звезды и Кимбал Имном слои
3. Что такое Data Vault (вкратце)? **HUB SATELITE LINK** (+ есть еще, но я не помню как называются)

Потоковая и пакетная обработка данных

1. Что такое потоковая обработка данных и пакетная обработка данных (вкратце)?
2. real time Kafka таблица наполняется в реальном времени..

3. пакетная - загружаем инкременетально за T-1.

ClickHouse

1. Как работает ClickHouse? колоночный, поэтому очень быстро происходят аггрегация значений. Схлопнуть одну колонку это очень эффективно, вместо фул скана по по каждому столбцу в строковых
2. Что такое колоночная база данных, в чем ее преимущества и недостатки?
3. Отличия OLTP vs OLAP

Python

1. Какие бывают магические функции в Python? **__init__. __main__**
2. Что такое **декораторы** в Python?
3. Что такое **try except else finally**
4. Чем отличается итератор от генератора? генератор он генерирует данные. Итератор он загружает в память всю эту хрень и если дать ему слишком много данных, то памяти может не хватить. Методы iter() и next() для итератора . А для генератора yield
5. `a = [1, 2, 3]`

`b = a`

`b.append(4) Что выведет print(a)`

6. Какие типы данных в Python являются изменяемыми и неизменяемыми?
7. **Как работает память в питоне? Ссылки на объекты в питоне?**
8. **Сортировка пузырьком (уметь писать закрытыми глазами). - Джуна**
9. Быструю сортировку - мидла
10. Сортировка слиянием mergesort () - мидла/джун
11. Можно ли в словарь в key записать изменяемый тип? Почему?
12. Решение алгоритмических задач на Python (например, палиндром, подсчет количества символов, можно ли собрать строку из другой строки)

Apache Spark

1. Почему нельзя использовать Pandas для больших данных, а нужно использовать Spark?

Pandas локально, Spark распредленно. В Spark можно выделять ресурсы , контролировать расчет и много настроек. Spark умеет работать с HDFS, S3, csv **(движок в оперативной памяти).**

1. Минимальное параллелизм в Spark и что это такое

getNumPartitons() = 200 (reparition(column)

1. Что такое RDD в Spark? мелкая часть данных
2. Что такое Dataset и чем отличается от dataframe и RDD?
3. Чем отличается repartition от coalesce в Spark?

в репартишн есть шаффл, а в коалеске его почти нет

1. Какие виды кэширования существуют в Spark и чем они отличаются?

cache и persisit(storage_level)

1. Что такое persist в Spark и какие storage levels существуют?
2. Что делает YARN и зачем он нужен?
3. Какие настройки Spark applications вы используете? Хабр

```
val spark = SparkSession
        .builder()
        .appName("StructStreaming")
        .master("yarn")
        .executors("10")
        .memory("100G")
        .config("hive.merge.mapfiles", "false")
        .config("hive.merge.tezfiles", "false")
        .config("parquet.enable.summary-metadata", "false")
        .config("spark.sql.parquet.mergeSchema","false")
        .config("hive.merge.smallfiles.avgsize", "16000000
0")
        .enableHiveSupport()
        .config("hive.exec.dynamic.partition", "true")
        .config("hive.exec.dynamic.partition.mode", "nonstr
```

```
ict")
        .config("spark.sql.orc.impl", "native")
        .config("spark.sql.parquet.binaryAsString","true")
        .config("spark.sql.parquet.writeLegacyFormat","tru
e")
        //.config(“spark.sql.streaming.checkpointLocation”,
“hdfs://pp/apps/hive/warehouse/dev01_landing_initial_area.d
b”)
        .getOrCreate()
```

1. Сколько гигабайт памяти выделяется на каждую задачу в Spark?
2. **Что такое spill в Spark?**
3. Что такое **broadcast join** в Spark и как его настроить? маленькая таблица котируется на все сервера и там уже просиходит join с данными большой таблицы
4. Что такое ленивые вычисления в Spark?

df = df.where(‘’’ col IS NOT NULL’’’)

.show() .collect() .count()

1. **Что такое Adaptive query execution?**

Apache Airflow

1. Разворачивали ли вы Airflow в Docker? **worker, webserver, sheduler, init**

Celery Executor local executor

1. Сколько типов сервисов в Docker при разворачивании Airflow?
2. Какая база данных используется в Airflow? posgressql
3. **Чем отличается Python оператор от Bash оператора в Airflow?**

HDFS

1. Что такое HDFS блоки и какие у них есть минусы?
2. Как бороться с маленькими файлами в HDFS? Переполнение NameNode