GeekBrains

BigData. Введение в экосистему Hadoop.

Урок 8. DWH

На уроке разбираем как устроены корпоративные хранилища данных, какие сложности приходится решать при их проектировании. Разбираем этапы развития и современные подходы проектирования. А также существующие профессии по работе с данными и их особенности.

Оглавление

- Оглавление

Теоретическая часть

- DataLake
- DWH
- Профессии в BigData

Используемая литература

Теоретическая часть

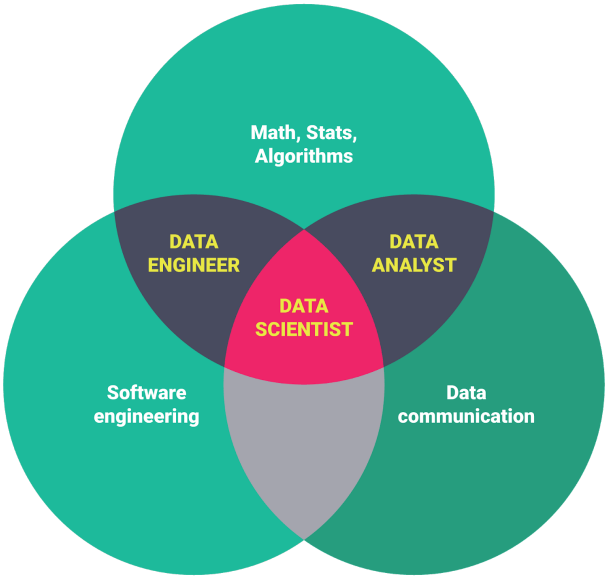
DataLake

Data lake — это огромное хранилище, которое принимает любые файлы всех форматов. Источник данных тоже не имеет никакого значения. Озеро данных может принимать данные из CRM- или ERP-систем, продуктовых каталогов, банковских программ, датчиков или умных устройств — любых систем, которые использует бизнес. Уже потом, когда данные сохранены, с ними можно работать — извлекать по определенному шаблону в классические базы данных или анализировать и обрабатывать прямо внутри data lake.

DWH

DWH -- это система данных, отдельная от оперативной системы обработки данных. В корпоративных хранилищах в удобном для анализа виде хранятся архивные данные из разных, иногда очень разнородных источников. Эти данные предварительно обрабатываются и загружаются в хранилище в ходе процессов извлечения, преобразования и загрузки, называемых ETL. Решения ETL и DWH — это (упрощенно) одна система для работы с корпоративной информацией и ее хранения.

Профессии в BigData



Домашнее задание

1. Финальное задание
2. Нужно либо ответить на вопросы ниже, либо развернуть etl на основе Apache NiFi и Cassandra (но можно и что-нибудь попроще, например PostgreSQL). Если будете создавать etl, жду описание шагов, что как и зачем делали.
3. Что такое Hadoop?
4. Что такое HDFS?
5. *Что такое YARN?
6. Какие минусы или опасные места HDFS?
7. Что такое блок HDFS?
8. Для чего используется NameNode?
9. Для чего используется DataNode?
10. Что будет, если записать много маленьких файлов в HDFS?
11. Что будет, если несколько DataNode внезапно отключатся?
12. Как проадпейдить несколько записи в большом файле на hdfs?
13. *Почему задачи на YARN нестабильны?
14. Что такое Hive?
15. Что хранит HiveMetastore?
16. Чем отличается external table и managed table?
17. *Какие форматы умеет читать Hive?
18. *Чем отличается управление ресурсов в Hive и Impala?
19. Чем отличается колочный формат хранения данных от строчного?
20. Чем отличается parquet/orc от csv?
21. Чем отличается Avro от json?
22. *Чем отличается документориетированный формат данных от реляционного?
23. Чем отличается etl и elt?
24. Далее - не различаем etl и elt:
25. Какие основные челенджы etl?
26. *Какие инструменты etl вы знаете?
27. Для чего нужны key-value СУБД?

28. *Какие сложности стриминга в hdfs?
29. *Какие минусы key-value хранилищ?
30. Из чего состоит хранилище данных?
31. Какие виды хранилищ данных вы знаете?
32. *Основные задачи Data governance?

Задачи со * предназначены для продвинутых учеников, которым мало сделать обычное ДЗ.

Используемая литература

Для подготовки данного методического пособия были использованы следующие ресурсы:

1. https://mcs.mail.ru/blog/chto-takoe-ozera-dannyh-i-zachem-tam-hranyat-big-data
2. https://mcs.mail.ru/blog/chto-takoe-dwh-i-pochemu-bez-nih-dannye-kompanii-bespolezny