1. Что такое Hadoop?

Стек технологий, связанных с обработкой и хранением больших данных.

1. Что такое HDFS?

HDFS – Hadoop Distributed File System. Фйловая система распределённого хранения данных. Суть в том, что данные хранятся на нескольких DataNodes (которые могут являться разными серверами или различными частями одной машины. Система обеспечивает репликацию данных и отслеживание их сохранности.

1. \*Что такое YARN?

YARN – Yet Another Resource Negotiator. Это сервис, который обеспечивает распределённое выполнение задач. Хорошо работает в связке с Hadoop, но может работать и без её. Основными участниками сервиса являются ResouceManager, Application Master и Node Manager.

1. Какие минусы или опасные места HDFS?

Минусы: Скорость чтения, практически полное отсутствие возможности изменения данных (в сущности возможна лишь дозапись). Храние файлов малого размера очень не эффективно.

1. Что такое блок HDFS?

Блок, это ячейка памяти фиксированного размера в HDFS. NameNode хранит информацию именно о ячейках. В Ячейке хранится один файл (или одна часть файла) вне зависимости от его размера.

1. Для чего используется NameNode?

NameNode управляет распределением данных. Она хранит в оперативной памяти метаданные, занимается мониторингом состояния DataNodes, распределяет данные по другим узлам при падении или сбое.

1. Для чего используется DataNode?

DataNode непосредственно хранит данные. Её хранилище поделено на блоки, которые и заполняются файлами (или частями файлов). DataNode с некоторой периодичностью сообщает NameNode о своём состоянии. Также именно DataNode загружает данные от пользователя и «общается» с другими нодами при репликации данных.

1. Что будет, если записать много маленьких файлов в HDFS?

Хранилище быстро забьётся, так как каждый маленький файл будет занимать один блок.

1. Что будет, если несколько DataNode внезапно отключатся?

Об этом узнает NameNode, так как от них прекратится подача отчётов о состоянии. После чего NameNode отправит запрос в оставшиеся узлы о репликации данных.

1. Как проадпейдить несколько записи в большом файле на hdfs?

Выгрузить файл из hdfs, внести изменения, загрузить изменённый файл, удалить старый.

1. \*Почему задачи на YARN нестабильны?
2. Что такое Hive?

Hive это SQL-like служба для взаимодействия с данными в hdfs.

1. Что хранит HiveMetastore?

Hive Metastore хранит информацию о структуре такблиц, партиций, данные о колонках и типах данных в них. Также он содержит информацию о сериалайзере и десериалайзере и данные о HDFS файлах, а также где они хранятся.

1. Чем отличается external table и managed table?

Managed table создаётся по умолчанию при создании таблицы. При этом таблица генерируется в стандартном каталоге hive. External table обрабатывается в том же месте, шде и находится.

1. \*Какие форматы умеет читать Hive?

Csv, json, avro, parquet, orc

1. \*Чем отличается управление ресурсов в Hive и Impala?
2. Чем отличается колочный формат хранения данных от строчного?

В строчном формате данные хранятся построчно с некоторым разделителем. Такие данные хорошо читаются взглядом, однако занимают много места и долго обрабатываются в том случае, если нам надо вытащить только, к примеру одну колонку.

Колоночные форматы хранят информацию в виде колонок, а также хранят некоторую дополнительную информацию (например, количество записей в колонке, или, например, если уникальных данных в колонке немного, то хранится могут только индексы. Такие базы данных могут занимать в разы меньше места, а также быстрее выполнять запросы.

1. Чем отличается parquet/orc от csv?

parquet/orc колоночные бинарные форматы хранения, csv строчный не бинарный.

1. Чем отличается Avro от json?

Avro бинарный аналог json. Занимает намного больше места.

1. \*Чем отличается документориетированный формат данных от реляционного?

Документоориентированные БД позволяют хранить документы в формате, подобном JSON. В отличие от реляционных баз данных, где данные хранятся в фиксированном наборе таблиц с фиксированным набором колонок и фиксированным типом данных, в документных БД структура базы может значительно меняться в процессе разработки.

1. Чем отличается etl и elt?

Различия в порядке действий. В случае с ETL данные сначала выгружаются, затем как-то трансформируются (очищаются, приводятся в соответствие с бизнес моделью и требованиями аналитика. Затем загружаются в хранилище.

При ELT данные выгружаются и сразу в полном объёме сохраняются в хранилище, обработка идёт последним шагом. Плюсы в том, что данные хранятся в неизменном виде, как они были получены, соответственно нет потери данных и можно проводить более полный анализ.

Далее - не различаем etl и elt:

1. Какие основные челенджи etl?
2. \*Какие инструменты etl вы знаете?

Infromatica, Sqoop, Oracle Data Integrator, Apache Airflow

1. Для чего нужны key-value СУБД?

Для быстрого получения значения по ключу.

1. \*Какие сложности стриминга в hdfs?

Необходимо реализовать промежуточное хранение данных, так как в случае передачи большого количества информации маленькими файлами, файловая система быстро забъётся.

1. \*Какие минусы key-value хранилищ?

Не подходят для решения комплексных задач.

1. Из чего состоит хранилище данных?

Из системы извлечения и (в случае ETL) первичной обработки данных, системы записи и хранения данных (например hdfs) и системы выдачи и представления данных.

1. Какие виды хранилищ данных вы знаете?

Хранилище данных со структурированными данными и озеро данных с неструктурированными данными

1. \*Основные задачи Data governance?

Формирование процессов ввода данных согласно установленным стандартам, распределение ролей, и прав доступа между участниками обработки информации. Введение единых стандартов, применяемых во всех бизнес процессах.