**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.**

**Что такое Hadoop?**

Платформа, в которую входят разные библиотеки, и фреймворки которые позволяют работать большими объемами данных и сохранять информацию на кластерах и узлах  
**Что такое HDFS?**

Один из механизмов системы Hadoop, который отвечает за хранение данных на разных кластерах, тем самым создает надежность данных. **\*Что такое YARN?**

Один из механизмов, позволяющий распределять ресурсы между всеми блоками(Hive, spark и др.) системы Hadoop

**Какие минусы или опасные места HDFS?**

+ Надежность;

+ Масштабируемость;

+ Высокая производительность;

- Возможность выполнения операций изменения файлов **Что такое блок HDFS?**

Механизм хранения данных, состоящий из управляющего узла NameNode, и узлов хранения DataNode **Для чего используется NameNode?**

Управляет пространством имен файловой системы. **Для чего используется DataNode?**

Служит рабочем пространством для записи и хранения информации. **Что будет, если записать много маленьких файлов в HDFS?**

Произойдёт загрузка NameNode, и возникнет очередь обработки файлов  **Что будет, если несколько DataNode внезапно отключатся?**

По идеи, все данные имеют репликацию на других DataNode, вероятность потери данных минимальна **Как проадпейдить несколько записи в большом файле на hdfs?**

На сколько я помню надо будет пересохранять этот файл, то есть запись в середину не доступна, только в конце **\*Почему задачи на YARN нестабильны?**

**При большой загрузки образуется очередь запросов**

**Что такое Hive?**

Один из механизмов Hadoop, использующий синтаксис SQL, с помощью технологии MapReduce **Что хранит HiveMetastore?**

Информацию(БД, таблицы) которые были созданы в Hive **Чем отличается external table и managed table?**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Поддержка Truncate | Поддержка ACID | Удаление таблиц и метаданных | Перемещение таблиц в каталог хранилища | Кеширование результатов |
| managed table | да | да | да | Да | да |
| external table | нет | нет | Только метаданные | нет | нет |

**\*Какие форматы умеет читать Hive?**

Avro, ORC, Parquet, SEQUENCEFILE, TextFile **\*Чем отличается управление ресурсов в Hive и Impala?**

-Impala обрабатывает SQL-запросы на лету, реализуя интерактивные вычисления в режиме онлайн, тогда как Hive не походит для OLTP-задач, т.к. работает с пакетами данных не в реальном времени

- Hive написан на Java, тогда как Impala – на C++

- Вычислительная модель Hive-MapReduce, Impala –MMP(быстрее)

- Отказоустойчивость – Hive является отказоустойчивой системой, которая сохраняет все промежуточные результаты. Impala в этом аспекте проигрывает

**Чем отличается колочный формат хранения данных от строчного?**

Способом записи информации, колоночные форматы данных быстрее перезаписываются т.к. поиск информации идет не по всему файлу, а то колонке **Чем отличается parquet/orc от csv?**

См. предыдущий ответ+ csv(строчный) занимает больше места на HD **Чем отличается Avro от json?**

Avro это программная платформа а json формат данных, или я что то путаю?  **\*Чем отличается документориетированный формат данных от реляционного?**

Реляционные базы данных подходят для хранения данных связанных и структурированных + есть ACID, видимо поэтому они и прижились. Документоориентированный формат подходит для хранения объемных данных использует в основном ключ-значение.

**Чем отличается etl и elt?**

* ETL обозначает извлечение, преобразование и загрузку, в то время как ELT обозначает извлечение, загрузку, преобразование.
* ETL загружает данные сначала на промежуточный сервер, а затем в целевую систему, тогда как ELT загружает данные непосредственно в целевую систему.
* Модель ETL используется для локальных, реляционных и структурированных данных, в то время как ELT используется для масштабируемых облачных структурированных и неструктурированных источников данных.
* ETL в основном используется для небольшого количества данных, тогда как ELT используется для больших объемов данных.

**Далее - не различаем etl и elt:  
Какие основные челенджы etl?**

Что в данном вопросе обозначает челендж? Если я правильно понял , то Загрузка данных в модель довольно сложная задача, т.к. исходные данные находятся в неструктурированном виде. **\*Какие инструменты etl вы знаете?**

HDFS, Apache, Qlik, Oracle.

**Для чего нужны key-value СУБД?**

Такие БД очень производительны, просты в обращении и легко масштабируются. **\*Какие сложности стриминга в hdfs?**

Загрузка ETL происходит через Batch, то есть требуется время на синхронизацию. **\*Какие минусы key-value хранилищ?**

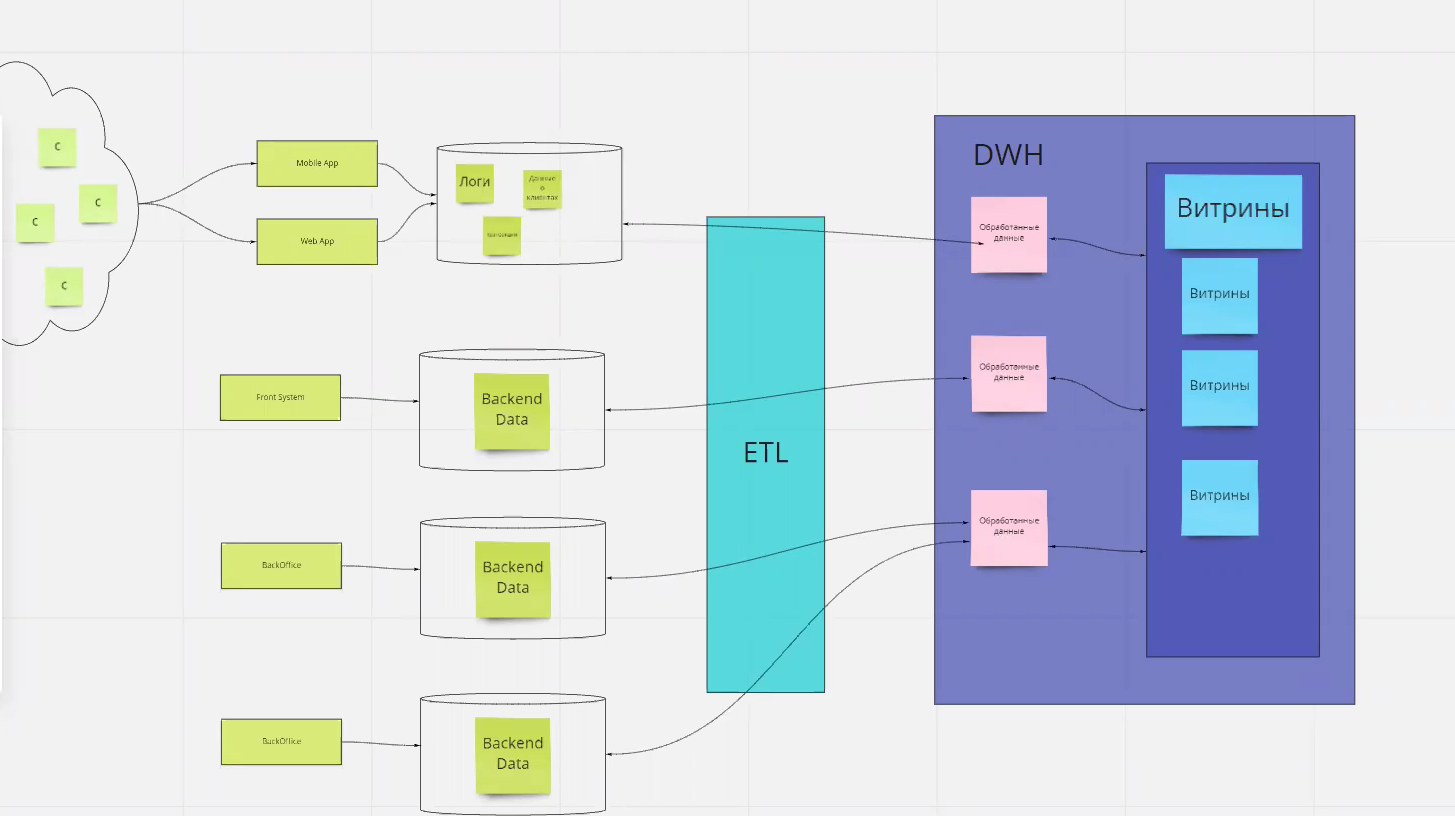
Значения выдаются только одним способом по ключу

**Из чего состоит хранилище данных?**

- Нижний уровень: этот уровень содержит сервер базы данных

- Средний уровень: средний уровень содержит сервер OLAP, который преобразует данные в структуру, лучше подходящую для анализа и сложных запросов

Верхний уровень: верхний уровень — это уровень клиента



**Какие виды хранилищ данных вы знаете?**

* Виртуальное хранилище данных — это набор отдельных баз данных, которые можно использовать совместно, чтобы пользователь мог эффективно получать доступ ко всем данным, как если бы они хранились в одном хранилище данных;
* Модель витрины данных используется для отчетности и анализа конкретных бизнес-линий. В этой модели хранилища – агрегированные данные из ряда исходных систем, относящихся к конкретной бизнес-сфере, такой как продажи или финансы;
* Модель корпоративного хранилища данных предполагает хранение агрегированных данных, охватывающих всю организацию. Эта модель рассматривает хранилище данных как сердце информационной системы предприятия с интегрированными данными всех бизнес-единиц

**\*Основные задачи Data governance?**

* Повышение качества

Низкое качество данных – это проблема не только для сотрудников отделов информационных технологий и аналитики, а бизнеса в целом : дубликаты, привести к одному формату, заполнить пропуски, почистить шумы и т.д.

* Обеспечение целостности и доступности

Возможность использовать облачные хранилища и отказоустойчивые базы данных. Данные могут быть потеряны – случайно или намеренно стёрты с носителя, поэтому необходимо делать резервные копии.

* Контроль

Внедрение процессов Data Governance позволяет получить полный контроль над данными. Можно контролировать, где и в каком формате они хранятся, обеспечивать версионность, поддерживать актуальность данных, определить правила доступа к данным и т.д.

* Обеспечение согласованности

Можно разместить всю информацию в одном общем хранилище, например, в базе данных или облаке. Это позволит сотрудникам работать с актуальной информацией, которая едина для всех.

* Унификация

Накопление информации из множества источников приводит к получению данных в разрозненном виде, их нужно приводить к единому виду и формату. Продуктивнее заранее определить политику работы с данными, стандарт, которому будут следовать все сотрудники. Таким образом, даже новый сотрудник или аналитик на стороне сможет понимать данные, и работа будет эффективнее.