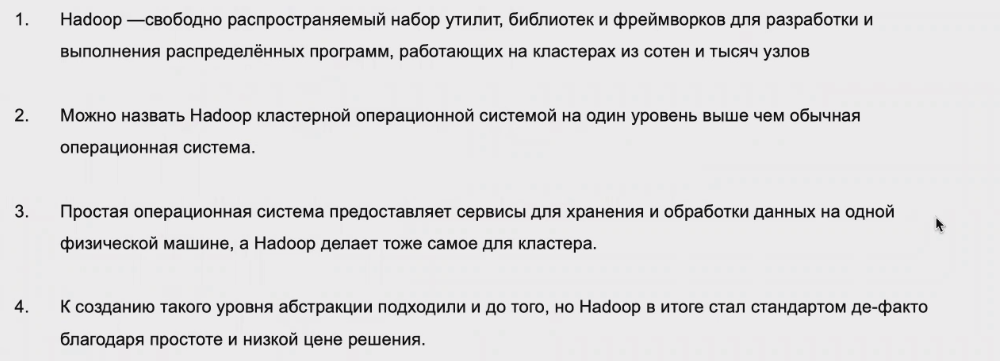
**BigData. Введение в экосистему Hadoop**

(итоговое HW)

**Что такое Hadoop?**



**Что такое HDFS?**

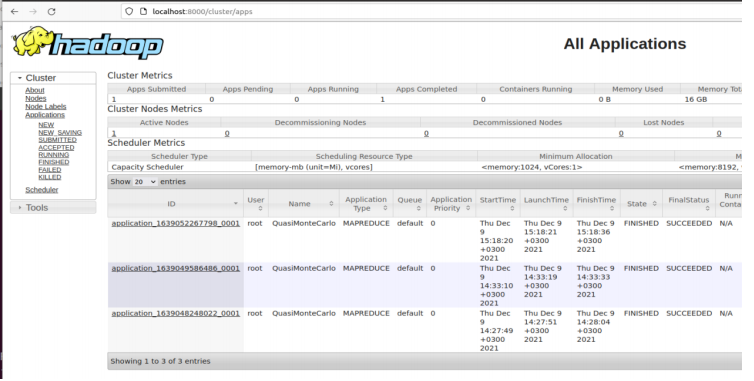
Файловая система HDFS создана для хранения большого объема неструктурированных данных.

HDFS – система для хранения большого объема данных.

**Что такое YARN?**

YARN – это фреймворк управления ресурсами в Hadoop.

YARN отвечает за распределение системных ресурсов между различными приложениями, работающими в кластере Hadoop, и планирование задач, которые должны выполняться на разных узлах кластера. Можно назвать операционной системой на кластерном уровне.  
Настроив веб-морду можно посмотреть информацию по нодам, приложениям



**Какие минусы или опасные места HDFS?**

Нет ACID транзакций, т.е. не обеспечивает надежность при транзакциях (передача маленьких данных с гарантией того что все выполнится или «откатится» к исходному состоянию при отказе – в Hadoop не выполняется).

**Что такое блок HDFS?**

Блок HDFS – это минимальный размер хранимой информации в HDFS. Хранятся куски файлов, кратные 128 или 256 Мб.

**Для чего используется NameNode?**

NameNode хранит метаданные – что-где лежит. Живет в оперативной памяти. Запоминает где какие данные записаны (координаты записанных данных)

Требует резервирования (на Secondary NameNode)

**Для чего используется DataNode?**

Данные хранятся на DataNode.

**Что будет, если записать много маленьких файлов в HDFS?**

Неэффективно будет использоваться железо: запись в DataNode все равно будет записан блоками по 256Мб.

**Что будет, если несколько DataNode внезапно отключатся?**

На время отключения доступ к данным будет прекращен. Новая запись будет осуществляться «мимо» отключенных нод.

**Как проадпейдить несколько записи в большом файле на hdfs?**

Берем блок, в котором содержатся нужные для апдейта записи.

Считываем,

Записываем (добавляем) новые записи

Старые удаляем.

**\*Почему задачи на YARN нестабильны?**

**Что такое Hive?**

Hive - это фреймвок хранилища данных на основе Hadoop, который может работать с реляционными данными в HDFS (один из 3х: Pig Hive Impala)

**Что хранит Hive Metastore?**

Metastore – сущность, хранящая метаданные об объектах (где лежат, какой формат и др.параметры).

**Чем отличается external table и managed table?**

external table –таблица, которая «смотрит» во внешний файл и отображает его данные. Типа представления (view) в SQL. При экстерн создаются метаданные, описывающие файл (при удалении табл – данные остаются)

managed table – таблица чисто внутри Hive (можно вставлять, удалять – данные удалятся)

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

Текстовые (Int, string, FLOAT,)

Колоночные (ORC, parquet)

**\*Чем отличается управление ресурсов в Hive и Impala?**

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

При строчном хранении для получения нужной информации из строки приходится читать всю строку и из нее потом получать нужный кусок.

Колоночные позволяют быстро работать с большим объемом данных (быстрее чем со строчными) + экономит место

В колоночных сжатие за счет формирования хэш-ссылок на данные и запись данных ссылками.

+ уже заранее прописываются сведения, которые наиболее вероятно будут участвовать в select

Колоночные хранят большое количество метаданных, что ускоряет поиск.

**Чем отличается parquet/orc от csv?**

Выше расписано примерно

**Чем отличается Avro от json?**

Avro – оптимизированный json, уменьшенный в несколько раз. Эффективно для передачи по сети, если требуется передача json`ов

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

**Чем отличается etl и elt? (Далее - не различаем etl и elt☺**

При ELT работаем с «оригинальными» - не почищенными данными.

**Какие основные челенджы etl?**

**Какие инструменты etl вы знаете?**

Все, чем можно получить данные, сохранить и далее обработать можем считать ETL

В Hadoop – это Spark и Kafka

**Для чего нужны key-value СУБД?**

Базы данных с использованием пар «ключ‑значение» поддерживают высокую разделяемость и обеспечивают неограниченное горизонтальное масштабирование, недостижимое при использовании других типов баз данных

**Какие сложности стриминга в hdfs?**

**\*Какие минусы key-value хранилищ?**

**Из чего состоит хранилище данных?**

DataNode

SecondryNode

NameNode

**Какие виды хранилищ данных вы знаете?**