Губин Е.В., ИУК4-72Б, вариант 7

Контрольная работа

1. *Опишите назначение HDFS API.*

HDFS API (Hadoop Distributed File System API) — это программный интерфейс, предоставляемый Hadoop для взаимодействия с распределённой файловой системой HDFS.

HDFS API необходим для того, чтобы с распределённой файловой системой можно было взаимодействовать программно: предоставляет методы для создания, чтения, удаления и перемещения файлов.

Этот интерфейс предоставляет возможность работать с данными, словно они лежат в одном месте. Все обращения к различным узлам кластера выполняются под капотом благодаря заранее подготовленной конфигурации кластера.

Так же, при желании, можно получить метаданные с узла NameNode, которые хранят информацию о том, как и на каких узлах хранятся файлы (какие блоки и где хранятся, где была репликация данных).

2. *Опишите стадию Map.*

Стадия Map — это первая фаза выполнения функции MapReduce, используемой в Hadoop для распределённой обработки больших данных.

Стадия Map выполняет преобразование входных данных в набор ключей и значений (key–value pairs), которые затем будут сгруппированы и обработаны на стадии Reduce.

Последовательность действий Map:

1) Входные данные разбиваются на блоки. Каждый блок передаётся отдельному мапперу, что обеспечивает параллельную обработку данных.
2) Для каждого элемента входных данных вызывается функция map() Функция получает:
- key —идентификатор,
- value — смысл для идентификатора (номер строки, количество повторений и тд и тп).
3) Функция map() генерирует промежуточные пары ключ–значение Например, в задаче подсчёта слов:

map("the cat", 0) → ("the", 1), ("cat", 1)

4) Промежуточные пары ключ–значение сортируются по ключам на стороне маппера. Это облегчает последующую передачу данных редьюсерам.
5) Аналогично, как и с Map, происходит разбиение задачи по редьюсерам (для распределённой обработки и данных).

3. *Дайте определение понятию ChainReducer.*

ChainReducer — это специальная утилита в Hadoop, позволяющая объединять несколько Map и Reduce задач в один MapReduce job.

Смысл ChainReducer состоит в том, что данные проходят через несколько Map задач до или после единственной в цепочке Reduce задачи. Каждый Mapper обрабатывает данные по своему:

Mapper -> Mapper -> Mapper -> … Reducer -> Mapper -> Mapper -> …

4. *Раскройте понятие трансформации и назовите основные трансформации, производимые над RDD.*

Трансформация — это операция над RDD (Resalient Distributed Dataset) в Apache Spark, которая принимает один RDD и создаёт на его основе новый RDD. Трансформации происходят параллельно на кластере.

Основные трансформации над RDD:

- map() - Применяет функцию к каждому элементу и возвращает новый RDD.
- flatMap() - То же, что map, но может возвращать несколько элементов для каждого входного.
- filter() - Возвращает элементы, удовлетворяющие условию (predicate).
- mapPartitions() - Применяет функцию к каждому разделу (partition) целиком, а не к отдельным элементам.
- union() - Объединяет два RDD.
- distinct() - Удаляет дубликаты.
- groupBy() - Группирует элементы по произвольному признаку.

5. *Приведите команды для перемещения файлов как внутри HDFS, так и для локальной системы.*

Перемещение файла внутри HDFS

hdfs dfs -mv /old/path/file.txt /new/path/file.txt

Перемещение группы файлов или каталога

hdfs dfs -mv /old/dir/ /new/dir/*

Удаление:

hdfs dfs -rm -r /new/path/file.txt

Перемещение:

hdfs dfs -mv /old/path/file.txt /new/path/

Копирование из HDFS в локальную систему:

hdfs dfs -get /hdfs/path/file.txt /local/path/

Из локальной системы в HDFS:

hdfs dfs -put /local/path/file.txt /hdfs/path/