# 1. Требования

1) RPM-based дистрибутив Linux с ядром версии не ниже 3.10 с доступом в интернет.

2) JDK 1.8

3) Hadoop 2.10.0

4) Пакет snappy-1.1.0

5) maven

Опционально для генератора данных:

6) perl5

7) cpan

8) Библиотека Data::Faker из CPAN

# 2. Сборка

1) Выкачиваем директорию src и файл pom.xml

2) В директории сборки запускаем

mvn install

для тестирования и сборки. Итоговый JAR будет расположен в директории target.

# 3. Запуск

1) Необходимо запустить Namenode командой

#sbin/start-dfs.sh

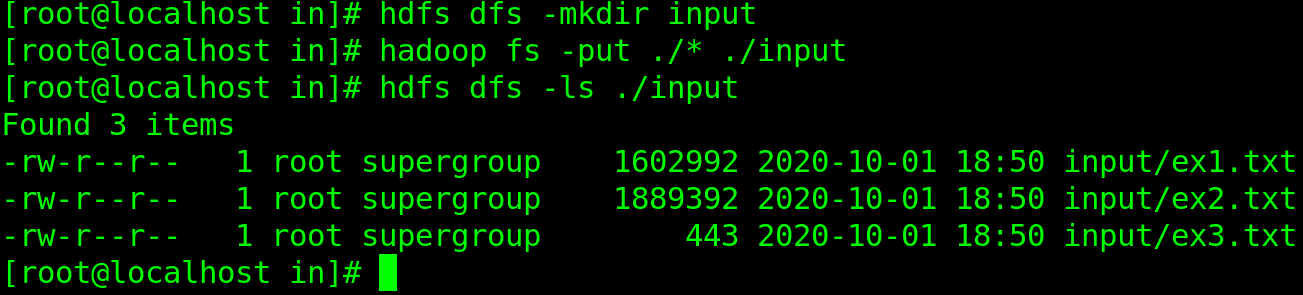
из директории где установлен hadoop.

2) Необходимо загрузить в целевую директорию hdfs входные данные.

Делается это командой:

$hadoop dfs -put <локальный путь> <путь в DFS>

Пример:



3) Для запуска непосредственно приложения выполнить команду:

$hadoop jar ./target/HW1-Haloop-1.0-SNAPSHOT-jar-with-dependencies.jar <входная директория в hdfs> <выходная директория в hdfs>

Например:



# 5. Генератор данных

Для удобства проверки добавлен генератор данных. Для его использование нужно

1) Выкачать директорию fakelog

2) Установить cpan

#yum install cpan

3) Установить Data::Faker

#cpan install Data::Faker

4) Находясь в директории fakelog запустить генератор командой:

$perl ./fakelog.pl <количество строк> <процент неверных строк> <путь выходного файла>

Например:

