- Tạo thư mục lưu dữ liệu đầu vào trên HDFS

```
hdfs dfs -mkdir -p /input
```

Lệnh này tạo thư mục /input trên hệ thống file phân tán HDFS, nơi sẽ chứa các file FASTA cần so khớp.

- Tải dữ liệu FASTA đầu vào lên HDFS

```
hdfs dfs -put query.fasta /input/
```

File *query.fasta* là tệp chứa chuỗi protein cần được so khớp. Sau khi upload, Hadoop sẽ có thể chia tệp này và phân phối cho các mapper xử lý song song.

- Xóa thư mục đầu ra cũ (nếu có)

```
hdfs dfs -rm -r /output
```

Để tránh lỗi khi Hadoop ghi đè lên một thư mục đã tồn tại, thư mục /output được xóa trước nếu tồn tại từ lần chạy trước đó.

- Thực thi job Hadoop BLAST

```
hadoop jar blastjob.jar /input /output
C:\Projects\HadoopBLAST\protein_db\protein_db
```

Trong đó:

  - blastjob.jar: file JAR đã được build, chứa mã MapReduce.
  - /input: đường dẫn thư mục chứa file FASTA trên HDFS.
  - /output: nơi lưu kết quả đầu ra của Hadoop.
  - C:\Projects\HadoopBLAST\protein_db\protein_db: đường dẫn tuyệt đối trên máy cục bộ đến thư mục chứa cơ sở dữ liệu BLAST và binary (đã được nạp thông qua Distributed Cache).
- Tải kết quả từ HDFS về máy cục bộ

```
hdfs dfs -get /output/part-r-00000
```

Sau khi job hoàn tất, kết quả BLAST được lưu trong file *part-r-00000*.