CSC 369: Distributed Computing

Alex Dekhtyar

May The Fourth

Day 13: MapReduce → Hadoop
May The Fourth Be With You
Housekeeping: Labs and Grading
MapReduce...What is it Good for?
MapReduce

Map: $K \times V \rightarrow \{K' \times V'\}$

Reduce $K \times (V)^* \rightarrow K \times (V)^*$
MapReduce

\[\text{Map: } K \times V \rightarrow \{K' \times V'\}\]

\[\text{Reduce } K \times (V)^* \rightarrow K \times (V)^*\]
Map: $K \times V \rightarrow \{K' \times V'\}$

Reduce $K \times (V)^* \rightarrow K \times (V)^*$
Map: $K \times V \rightarrow \{K' \times V'\}$

Record-by-record operations:
- Filter/Selection
- Projection

Reduce $K \times (V)^* \rightarrow K \times (V)^*$
Map: $K \times V \rightarrow \{K' \times V'\}$

Record-by-record operations:

Filter/Selection
Projection

Reduce $K \times (V)^* \rightarrow K \times (V)^*$
Map: $K \times V \rightarrow \{K' \times V'\}$

Record-by-record operations:
- Filter/Selection
- Projection

Reduce $K \times (V)^* \rightarrow K \times (V)^*$

Aggregation
Map: $K \times V \to \{K' \times V'\}$

Record-by-record operations:
- Filter/Selection
- Projection

Reduce $K \times (V)^* \to K \times (V)^*$

Aggregation
Map: $K \times V \rightarrow \{K' \times V'\}$

Record-by-record operations:
- Filter/Selection
- Projection

Reduce $K \times (V)^* \rightarrow K \times (V)^*$

Aggregation
Data Processing Pipeline

- Select/Filter
- Project/Transform
- Group
- Aggregate
- Select/Filter

Map
Shuffle
Reduce
For each day from April 15 to April 30 find the ratio between the number of daily new deaths and the number of daily new cases.

Find all weeks during which the total number of new positive COVID-19 cases in California and Florida exceeded 10,000, and report the largest daily increase in cases for each such week.
Our First Hadoop Steps Will be on such pipelines

- Select/Filter
- Project/Transform
- Group
- Aggregate
- Select/Filter
Resource Manager (YARN)

Distributed Data Store (HDFS)
Distributed Data Store (HDFS)

Resource Manager (YARN)

MapReduce

Spark

Apache Spark™

O.
At first

MapReduce

Resource Manager (YARN)

Distributed Data Store (HDFS)
Distributed Data Store (HDFS)
Distributed Data Store (HDFS)
Distributed Data Store (HDFS)
Distributed Data Store (HDFS)

Node1

Node2

Node3

NodeN

Sharding Replication
Distributed Data Store (HDFS)

Command line interface

```
$ hdfs dfs -ls
Found 3 items
drwxr-xr-x   - dekhtyar hdfs          0 2020-04-02 12:00 .sparkStaging
drwx------   - dekhtyar hdfs          0 2020-05-04 04:05 .staging
drwxr-xr-x   - dekhtyar hdfs          0 2020-05-04 04:05 test
```

```
hdfs dfs -<command> <parameters>
```
Distributed Data Store (HDFS)

Command line interface

```
$ hdfs dfs -ls
Found 3 items
drwxr-xr-x   - dekhtyar hdfs          0 2020-04-02 12:00 .sparkStaging
drwx------   - dekhtyar hdfs          0 2020-05-04 04:05 .staging
drwxr-xr-x   - dekhtyar hdfs          0 2020-05-04 04:05 test
```

hdfs dfs -<command> <parameters>

See handout for hdfs commands
At first

MapReduce

Resource Manager (YARN)

Distributed Data Store (HDFS)
Today: Java Program DEMO

Job runner (main)
Mapper
Reducer

Wednesday: Details

Input