E6893 Big Data Analytics Lecture 2:

Big Data Analytics Platforms

Ching-Yung Lin, Ph.D.
Adjunct Professor, Depts. of Electrical Engineering and Computer Science

September 13th, 2018
TA Office Hours and Submission of Your Interests

Columbia University EECS E6893: Big Data Analytics

Finance
Retail
Media
Energy
Information
Life Sciences
Social Science
Government
Telecom
Transportation
Industry

- Pratyus Pati: Saturday 5-7pm
- Yanbei Pang: Friday 6-8pm
- Vishal Anand: Wednesday 7-9pm
- Zekun Gong: Tuesday 1:30 - 3:30pm
- Chao-Yang Lo: Tuesday 4-6pm
- Zixi Huang: Monday 1-3pm

- Location: CS TA Room
Reading Reference for Lecture 2

Introduction

Part I: Getting Started with Hadoop
Chapter 1: Introducing Hadoop and Seeing What It’s Good For
Chapter 2: Common Use Cases for Big Data in Hadoop
Chapter 3: Setting Up Your Hadoop Environment

Part II: How Hadoop Works
Chapter 4: Storing Data in Hadoop: The Hadoop Distributed File System
Chapter 5: Reading and Writing Data
Chapter 6: MapReduce Programming
Chapter 7: Frameworks for Processing Data in Hadoop: YARN and MapReduce
Chapter 8: Pig: Hadoop Programming Made Easier
Chapter 9: Statistical Analysis in Hadoop
Chapter 10: Developing and Scheduling Application Workflows with Oozie

Part III: Hadoop and Structured Data
Chapter 11: Hadoop and the Data Warehouse: Friends or Foes?
Chapter 12: Extremely Big Tables: Storing Data in HBase
Chapter 13: Applying Structure to Hadoop Data with Hive
Chapter 14: Integrating Hadoop with Relational Databases Using Sqoop
Chapter 15: The Holy Grail: Native SQL Access to Hadoop Data

Part IV: Administering and Configuring Hadoop
Chapter 16: Deploying Hadoop
Chapter 17: Administering Your Hadoop Cluster

Part V: The Part of Tens
Chapter 18: Ten Hadoop Resources Worthy of a Bookmark
Chapter 19: Ten Reasons to Adopt Hadoop
The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

http://hadoop.apache.org
Remind -- Hadoop-related Apache Projects

- **Ambari™**: A web-based tool for provisioning, managing, and monitoring Hadoop clusters. It also provides a dashboard for viewing cluster health and ability to view MapReduce, Pig and Hive applications visually.
- **Avro™**: A data serialization system.
- **Cassandra™**: A scalable multi-master database with no single points of failure.
- **Chukwa™**: A data collection system for managing large distributed systems.
- **HBase™**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive™**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™**: A Scalable machine learning and data mining library.
- **Pig™**: A high-level data-flow language and execution framework for parallel computation.
- **Spark™**: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Tez™**: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases.
- **ZooKeeper™**: A high-performance coordination service for distributed applications.
Four distinctive layers of Hadoop

**Distributed storage:** The Hadoop Distributed File System (HDFS) is the storage layer where the data, interim results, and final result sets are stored.

**Resource management:** In addition to disk space, all slave nodes in the Hadoop cluster have CPU cycles, RAM, and network bandwidth. A system such as Hadoop needs to be able to parcel out these resources so that multiple applications and users can share the cluster in predictable and tunable ways. This job is done by the JobTracker daemon.

**Processing framework:** The MapReduce process flow defines the execution of all applications in Hadoop 1. As we saw in Chapter 6, this begins with the map phase; continues with aggregation with shuffle, sort, or merge; and ends with the reduce phase. In Hadoop 1, this is also managed by the JobTracker daemon, with local execution being managed by TaskTracker daemons running on the slave nodes.

**Application Programming Interface (API):** Applications developed for Hadoop 1 needed to be coded using the MapReduce API. In Hadoop 1, the Hive and Pig projects provide programmers with easier interfaces for writing Hadoop applications, and underneath the hood, their code compiles down to MapReduce.
Common Use Cases for Big Data in Hadoop

- Log Data Analysis  
  - most common, fits perfectly for HDFS scenario: *Write once & Read often.*
- Data Warehouse Modernization
- Fraud Detection
- Risk Modeling
- Social Sentiment Analysis
- Image Classification
- Graph Analysis
- Beyond

D. deRoos *et al*, Hadoop for Dummies, John Wiley & Sons, 2014
Example: Business Value of Log Analysis – “Struggle Detection”

- Carts Abandoned by Month:
  - 60% of carts are regularly abandoned.

- Abandoned Cart Value (SM):
  - 10% Recapture rate = $80M

- Abandoned by Category:
  - Chart showing abandonment rates by category,
    - Categories include: Camping, Climbing, Cycling Gear, Karaoke Gear, Watersport Gear, Skiing, Luggage, Packs, Women's Clothing.

D. deRoos et al, Hadoop for Dummies, John Wiley & Sons, 2014
Remind -- MapReduce example

The overall MapReduce word count process

Input

Deer Bear River
Car Car River
Deer Car Bear

Splitting

Deer Bear River
Car Car River
Deer Car Bear

Mapping

Deer, 1
Bear, 1
River, 1

Car, 1
Car, 1
Car, 1
River, 1

Deer, 1
Bear, 1

Shuffling

Bear, 2
Bear, 1

Car, 1
Car, 1
Car, 1

Deer, 1
Deer, 1

Reducing

Bear, 2
Car, 3

Deer, 2
River, 2

Final result

Bear, 2
Car, 3
Deer, 2
River, 2

http://www.alex-hanna.com
MapReduce Process on User Behavior via Log Analysis

Hadoop makes IP address the key and timestamp and URL the value: The Map Step. Sample Web Log Input

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>201.76.43.67</td>
<td>10:24:48</td>
</tr>
<tr>
<td>123.99.54.76</td>
<td>10:24:21</td>
</tr>
<tr>
<td>201.76.43.67</td>
<td>10:25:15</td>
</tr>
<tr>
<td>231.67.66.84</td>
<td>10:26:03</td>
</tr>
<tr>
<td>201.76.43.67</td>
<td>10:26:13</td>
</tr>
<tr>
<td>201.76.43.67</td>
<td>10:26:23</td>
</tr>
<tr>
<td>101.23.78.92</td>
<td>07:06:56</td>
</tr>
<tr>
<td>114.43.99.75</td>
<td>01:32:43</td>
</tr>
<tr>
<td>109.55.55.81</td>
<td>09:45:31</td>
</tr>
</tbody>
</table>

Shuffle Sort Group

Framework does this automatically

K: 201.76.43.67
V: www.ff.com

10:25:15. www.ff.com/search

The Reduce Step

Ensure something is in cart and determine last page visited.

K: 201.76.43.67
V: www.ff.com/shipping

D. deRoos et al, Hadoop for Dummies, John Wiley & Sons, 2014
Setting Up the Hadoop Environment

- Local (standalone) mode
- Pseudo-distributed mode
- Fully-distributed mode
Data Storage Operations on HDFS

- Hadoop is designed to work best with a modest number of extremely large files.
- Average file sizes → larger than 500MB.

- Write Once, Read Often model.
- Content of individual files cannot be modified, other than appending new data at the end of the file.

- What we can do:
  - Create a new file
  - Append content to the end of a file
  - Delete a file
  - Rename a file
  - Modify file attributes like owner
Remind -- Hadoop Distributed File System (HDFS)

HDFS Architecture

Client -> Metadata ops to Namenode

Namenode -> Metadata (Name, replicas, ...): /home/foo/data, 3, ...

Client -> Read to Datanodes

Datanodes -> Block ops to Namenode

Namenode -> Block ops to Datanodes

Datanodes --> Replication

Datanodes -> Blocks

Rack 1

Write from Client to Datanodes

Rack 2

http://hortonworks.com/hadoop/hdfs/
HDFS blocks

- File is divided into blocks (default: 64MB) and duplicated in multiple places (default: 3)

- Dividing into blocks is normal for a file system. E.g., the default block size in Linux is 4KB. The difference of HDFS is the scale.
- Hadoop was designed to operate at the petabyte scale.
- Every data block stored in HDFS has its own metadata and needs to be tracked by a central server.
HDFS blocks

- Replication patterns of data blocks in HDFS.

- When HDFS stores the replicas of the original blocks across the Hadoop cluster, it tries to ensure that the block replicas are stored in different failure points.
HDFS is a User-Space-Level file system
Interaction between HDFS components

- Command line
- Applications

Client

Data Node 1
Data Node 2
Data Node 3
Data Node 4

Rack 1
Rack 2

NameNode

- fsimage
- edits
disk

Block locations
Block metadata
memory

Master Node 1
HDFS Federation

• Before Hadoop 2.0, NameNode was a single point of failure and operation limitation.
• Before Hadoop 2, Hadoop clusters usually have fewer clusters that were able to scale beyond 3,000 or 4,000 nodes.
• Multiple NameNodes can be used in Hadoop 2.x. (HDFS High Availability feature – one is in an Active state, the other one is in a Standby state).

http://hadoop.apache.org/docs/r2.3.0/hadoop-yarn/hadoop-yarn-site/HDFSHighAvailabilityWithNFS.html
High Availability of the NameNodes

- Active NameNode
- Standby NameNode – keeping the state of the block locations and block metadata in memory -> HDFS checkpointing responsibilities.

- JournalNode – if a failure occurs, the Standby Node reads all completed journal entries to ensure the new Active NameNode is fully consistent with the state of cluster.
- Zookeeper – provides coordination and configuration services for distributed systems.
Several useful commands for HDFS

- All hadoop commands are invoked by the bin/hadoop script.

  ```
  hadoop [--config confdir] [COMMAND]
  [GENERIC_OPTIONS] [COMMAND_OPTIONS]
  ```

- `% hadoop fsck / -files –blocks:
  → list the blocks that make up each file in HDFS.

- For HDFS, the schema name is hdfs, and for the local file system, the schema name is file.

- A file or director in HDFS can be specified in a fully qualified way, such as:
  hdfs://namenodehost/parent/child or hdfs://namenodehost

- The HDFS file system shell command is similar to Linux file commands, with the following general syntax: `hadoop hdfs –file_cmd`

- For instance mkdir runs as:
  `$hadoop hdfs dfs –mkdir /user/directory_name`
Several useful commands for HDFS -- II

For example, to create a directory named “joanna”, run this `mkdir` command:

```
$ hadoop hdfs dfs -mkdir /user/joanna
```

Use the Hadoop `put` command to copy a file from your local file system to HDFS:

```
$ hadoop hdfs dfs -put file_name /user/login_user_name
```

For example, to copy a file named `data.txt` to this new directory, run the following `put` command:

```
$ hadoop hdfs dfs -put data.txt /user/joanna
```

Run the `ls` command to get an HDFS file listing:

```
$ hadoop hdfs dfs -ls .
```
• YARN – Yet Another Resource Negotiator:
  
  – A Tool that enables the other processing frameworks to run on Hadoop.

  – A general-purpose resource management facility that can schedule and assign CPU cycles and memory (and in the future, other resources, such as network bandwidth) from the Hadoop cluster to waiting applications.

➔ YARN has converted Hadoop from simply a batch processing engine into a platform for many different modes of data processing, from traditional batch to interactive queries to streaming analysis.
Four distinctive layers of Hadoop

**Distributed storage:** The Hadoop Distributed File System (HDFS) is the storage layer where the data, interim results, and final result sets are stored.

**Resource management:** In addition to disk space, all slave nodes in the Hadoop cluster have CPU cycles, RAM, and network bandwidth. A system such as Hadoop needs to be able to parcel out these resources so that multiple applications and users can share the cluster in predictable and tunable ways. This job is done by the JobTracker daemon.

**Processing framework:** The MapReduce process flow defines the execution of all applications in Hadoop 1. As we saw in Chapter 6, this begins with the map phase; continues with aggregation with shuffle, sort, or merge; and ends with the reduce phase. In Hadoop 1, this is also managed by the JobTracker daemon, with local execution being managed by TaskTracker daemons running on the slave nodes.

**Application Programming Interface (API):** Applications developed for Hadoop 1 needed to be coded using the MapReduce API. In Hadoop 1, the Hive and Pig projects provide programmers with easier interfaces for writing Hadoop applications, and underneath the hood, their code compiles down to MapReduce.
1. The client application submits an application request to the JobTracker.
2. The JobTracker determines how many processing resources are needed to execute the entire application.
3. The JobTracker looks at the state of the slave nodes and queues all the map tasks and reduce tasks for execution.
4. As processing slots become available on the slave nodes, map tasks are deployed to the slave nodes. Map tasks are assigned to nodes where the same data is stored.
5. The JobTracker monitors task progress. If failure, the task is restarted on the next available slot.
6. After the map tasks are finished, reduce tasks process the interim results sets from the map tasks.
7. The result set is returned to the client application.
Limitation of Hadoop 1

- MapReduce is a successful batch-oriented programming model.

- A glass ceiling in terms of wider use:
  - Exclusive tie to MapReduce, which means it could be used only for batch-style workloads and for general-purpose analysis.

- Triggered demands for additional processing modes:
  - Graph Analysis
  - Stream data processing
  - Message passing

  ➔ Demand is growing for real-time and ad-hoc analysis
  ➔ Analysts ask many smaller questions against subsets of data and need a near-instant response.
  ➔ Some analysts are more used to SQL & Relational databases

YARN was created to move beyond the limitation of a Hadoop 1 / MapReduce world.
Hadoop 2 Data Processing Architecture
YARN’s application execution

- Client submits an application to Resource Manager.
- Resource Manager asks a Node Manager to create an Application Master instance and starts up.
- Application Manager initializes itself and register with the Resource Manager.
- Application manager figures out how many resources are needed to execute the application.
- The Application Master then requests the necessary resources from the Resource Manager. It sends heartbeat message to the Resource Manager throughout its lifetime.
- The Resource Manager accepts the request and queue up.
- As the requested resources become available on the slave nodes, the Resource Manager grants the Application Master leases for containers on specific slave nodes.
- .. only need to decide on how much memory tasks can have.
Remind -- MapReduce Data Flow

MapReduce Use Case Example – flight data

- Data Source: Airline On-time Performance data set (flight data set).
  - All the logs of domestic flights from the period of October 1987 to April 2008.
  - Each record represents an individual flight where various details are captured:
    - Time and date of arrival and departure
    - Originating and destination airports
    - Amount of time taken to taxi from the runway to the gate.
Bi-Annual Data Exposition

Every other year, at the Joint Statistical Meetings, the Graphics Section and the Computing Section join in sponsoring a special Poster Session called The Data Exposition, but more commonly known as The Data Expo. All of the papers presented in this Poster Session are reports of analyses of a common data set provided for the occasion. In addition, all papers presented in the session are encouraged to report the use of graphical methods employed during the development of their analysis and to use graphics to convey their findings.

Data sets

- **2013**: Soul of the Community
- **2011**: Deepwater horizon oil spill
- **2009**: Airline on time data
- **2006**: NASA meteorological data. [Electronic copy of entries](http://stat-computing.org/dataexpo/)
- **1997**: Hospital Report Cards
- **1995**: U.S. Colleges and Universities
- **1993**: Oscillator time series & Breakfast Cereals
- **1991**: Disease Data for Public Health Surveillance
- **1990**: King Crab Data
- **1988**: Baseball
- **1986**: Geometric Features of Pollen Grains
- **1983**: Automobiles

http://stat-computing.org/dataexpo/
### Flight Data Schema

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Year</td>
<td>1987-2008</td>
</tr>
<tr>
<td>2 Month</td>
<td>1-12</td>
</tr>
<tr>
<td>3 DayofMonth</td>
<td>1-31</td>
</tr>
<tr>
<td>4 DayOfWeek</td>
<td>1 (Monday) - 7 (Sunday)</td>
</tr>
<tr>
<td>5 DepTime</td>
<td>actual departure time (local, hhmm)</td>
</tr>
<tr>
<td>6 CRSDepTime</td>
<td>scheduled departure time (local, hhmm)</td>
</tr>
<tr>
<td>7 ArrTime</td>
<td>actual arrival time (local, hhmm)</td>
</tr>
<tr>
<td>8 CRSArrTime</td>
<td>scheduled arrival time (local, hhmm)</td>
</tr>
<tr>
<td>9 UniqueCarrier</td>
<td>unique carrier code</td>
</tr>
<tr>
<td>10 FlightNum</td>
<td>flight number</td>
</tr>
<tr>
<td>11 TailNum</td>
<td>plane tail number</td>
</tr>
<tr>
<td>12 ActualElapsedTime</td>
<td>in minutes</td>
</tr>
<tr>
<td>13 CRSElapsedTime</td>
<td>in minutes</td>
</tr>
<tr>
<td>14 AirTime</td>
<td>in minutes</td>
</tr>
<tr>
<td>15 ArrDelay</td>
<td>arrival delay, in minutes</td>
</tr>
<tr>
<td>16 DepDelay</td>
<td>departure delay, in minutes</td>
</tr>
<tr>
<td>17 Origin</td>
<td>origin IATA airport code</td>
</tr>
<tr>
<td>18 Dest</td>
<td>destination IATA airport code</td>
</tr>
<tr>
<td>19 Distance</td>
<td>in miles</td>
</tr>
<tr>
<td>20 TaxiIn</td>
<td>taxi in time, in minutes</td>
</tr>
<tr>
<td>21 TaxiOut</td>
<td>taxi out time in minutes</td>
</tr>
<tr>
<td>22 Cancelled</td>
<td>was the flight cancelled?</td>
</tr>
<tr>
<td>23 CancellationCode</td>
<td>reason for cancellation</td>
</tr>
<tr>
<td>24 Diverted</td>
<td>1 = yes, 0 = no</td>
</tr>
<tr>
<td>25 CarrierDelay</td>
<td>in minutes</td>
</tr>
<tr>
<td>26 WeatherDelay</td>
<td>in minutes</td>
</tr>
<tr>
<td>27 NASDelay</td>
<td>in minutes</td>
</tr>
<tr>
<td>28 SecurityDelay</td>
<td>in minutes</td>
</tr>
<tr>
<td>29 LateAircraftDelay</td>
<td>in minutes</td>
</tr>
</tbody>
</table>
MapReduce Use Case Example – flight data

- Count the number of flights for each carrier

- Serial way (not MapReduce):

```
Listing 6-1:  Pseudocode for Calculating The Number of Flights By Carrier Serially

create a two-dimensional array
create a row for every airline carrier
  populate the first column with the carrier code
  populate the second column with the integer zero

for each line of flight data
  read the airline carrier code
  find the row in the array that matches the carrier code
    increment the counter in the second column by one

print the totals for each row in the two-dimensional array
```
MapReduce Use Case Example – flight data

- Count the number of flights for each carrier

- Parallel way:

**Listing 6-2: Pseudocode for Calculating The Number of Flights By Carrier in Parallel**

Map Phase:
for each line of flight data
  read the current record and extract the airline carrier code
  output the airline carrier code and the number one as a key/value pair

Shuffle and Sort Phase:
  read the list of key/value pairs from the map phase
  group all the values for each key together
  each key has a corresponding array of values
  sort the data by key
  output each key and its array of values

Reduce Phase:
  read the list of carriers and arrays of values from the shuffle and sort phase
  for each carrier code
  add the total number of ones in the carrier code’s array of values together

print the totals for each row in the two-dimensional array
MapReduce application flow

Determine the exact data sets to process from the data blocks. This involves calculating where the records to be processed are located within the data blocks.

Run the specified algorithm against each record in the data set until all the records are processed. The individual instance of the application running against a block of data in a data set is known as a mapper task. (This is the mapping part of MapReduce.)

Locally perform an interim reduction of the output of each mapper. (The outputs are provisionally combined, in other words.) This phase is optional because, in some common cases, it isn’t desirable.

Based on partitioning requirements, group the applicable partitions of data from each mapper’s result sets.

Boil down the result sets from the mappers into a single result set — the Reduce part of MapReduce. An individual instance of the application running against mapper output data is known as a reducer task.
MapReduce steps for flight data computation

1. **Map Phase**
   - Mapper function $(K1, V1)$
   - $list(K2, V2) = \{PS, 1, PS, 1, ..., C0, 1, C0, 1\}$
   - $k2, list(V2) = \{(AA, [1, 1, ..., 1]), (AS, [1, 1, ..., 1]), ..., (WN, [1, 1, ..., 1])\}$

2. **Shuffle and Sort**
   - Mapper output is sorted and partitioned

3. **Reduce Phase**
   - Reducer function $(K2, list(V2))$
   - $list(K3, V3) = \{(AA, 165121), (AS, 21406), (WN, 61975)\}$

4. **Final output**
   - map: $(K1, V1) \rightarrow list(K2, V2)$
   - reduce: $(K2, list(V2)) \rightarrow list(K3, V3)$
Create FlightsByCarrier.java:

```java
/**
 * The FlightsByCarrier Driver Application
 */
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io. *
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class FlightsByCarrier {
    public static void main(String[] args) throws Exception {
        Job job = new Job();
        job.setJarByClass( FlightsByCarrier.class);
        job.setJobName("FlightsByCarrier");
    }
}
```
```java
    @3
    TextInputFormat.addInputPath(job, new Path(args[0]));
    job.setInputFormatClass(TextInputFormat.class);

    @4
    job.setMapperClass(FlightsByCarrierMapper.class);
    job.setReducerClass(FlightsByCarrierReducer.class);

    @5
    TextOutputFormat.setOutputPath(job, new Path(args[1]));
    job.setOutputFormatClass(TextOutputFormat.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);

    @6
    job.waitForCompletion(true);
```
```java
import java.io.IOException;
import au.com.bytecode.opencsv.CSVParser;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Mapper;

public class FlightsByCarrierMapper extends Mapper<LongWritable, Text, Text, IntWritable> {
    @Override
    protected void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {
        if (key.get() > 0) {
            String[] lines = new CSVParser().parseLine(value.toString());
            context.write(new Text(lines[8]), new IntWritable(1));
        }
    }
}
```

**Listing 6-4: The FlightsByCarrier Mapper Code**

Input file: flight-data.csv

Map Phase

Mapper function (K1, V1)

List(K2, V2)

Shuffle and Sort

K2, List(V2)

Reducer function (K2, list(V2))

List(K3, V3)

Final output: map: (K1, V1) → list(K2, V2) reduce: (K2, list(V2)) → list(K3, V3)
Listing 6-5: The FlightsByCarrier Reducer Code

```java
import java.io.IOException;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Reducer;

public class FlightsByCarrierReducer extends
    Reducer<Text, IntWritable, Text, IntWritable> {
    @Override
    protected void reduce(Text token, Iterable<IntWritable> counts,
                           Context context) throws IOException, InterruptedException {
        int sum = 0;

        for (IntWritable count : counts) {
            sum += count.get();
        }

        context.write(token, new IntWritable(sum));
    }
}
```
Run the code

To run the FlightsByCarrier application, follow these steps:

Go to the directory with your Java code and compile it using the following command:

```
javac -classpath $CLASSPATH MapRed/FlightsByCarrier/*.java
```

Build a JAR file for the application by using this command:

```
jar cvf FlightsByCarrier.jar *.class
```

Run the driver application by using this command:

```
hadoop jar FlightsByCarrier.jar FlightsByCarrier /user/root/airline-data/2008.csv /user/root/output/flightsCount
```
See Result

Show the job’s output file from HDFS by running the command

    hadoop fs -cat /user/root/output/flightsCount/part-r-00000

You see the total counts of all flights completed for each of the carriers in 2008:

    AA  165121
    AS  21406
    CO  123002
    DL  185813
    EA  108776
    HP  45399
    NW  108273
    PA (1) 16785
    PI  116482
    PS  41706
    TW  69650
    UA  152624
    US  94814
    WN  61975
HBase is modeled after Google’s BigTable and written in Java. It is developed on top of HDFS.

It provides a fault-tolerant way of storing large quantities of **sparse data** (small amounts of information caught within a large collection of empty or unimportant data, such as finding the 50 largest items in a group of 2 billion records, or finding the non-zero items representing less than 0.1% of a huge collection).

HBase features compression, in-memory operation, and Bloom filters on a per-column basis.

An HBase system comprises a set of tables. Each table contains rows and columns, much like a traditional database. Each table must have an element defined as a Primary Key, and all access attempts to HBase tables must use this Primary Key. An HBase column represents an attribute of an object.
Characteristics of data in HBase

Sparse data

<table>
<thead>
<tr>
<th>Table 12-1 Traditional Customer Contact Information Table</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Customer ID</strong></td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>00001</td>
</tr>
<tr>
<td>00002</td>
</tr>
</tbody>
</table>

Row Key  Column Family: {Column Qualifier:Version:Value}

00001  CustomerName: {‘FN’: 1383859182496;‘FN’: 1383859182858;‘FN’: 1383859183001;‘FN’: 1383859182915;‘FN’}
       ContactInfo: {‘EA’: 1383859183030;‘EA’: 1383859183073;‘SA’: 1383859183073;‘SA’: 1383859183073}

00002  CustomerName: {‘FN’: 1383859183103;‘FN’: 1383859183163;‘FN’: 1383859183163}
       ContactInfo: {‘SA’: 1383859185577;‘SA’: 1383859185577;‘SA’: 1383859185577}

HDFS lacks **random read and write access**. This is where HBase comes into picture. It's a **distributed, scalable, big data store**, modeled after Google's BigTable. It stores data as key/value pairs.
HBase Architecture

Logical Architecture

Clients
- MasterServer
  - MasterServer (Backup)
- RegionServer
- RegionServer
- RegionServer

Automatic Scalability!

Zookeeper Ensemble for HBase Coordination Services and Fault Recovery

HBase Regions in Detail

Region
- Column Family 1 Store
  - MemStore (Cache for Writes)
  - BlockCache (For Reads)
- Column Family n Store
  - MemStore (Cache for Writes)
  - BlockCache (For Reads)

HDFS
- Write Ahead Log
- HFile
- HFile
- HFile
Creating a table

hbase(main):002:0> create 'CustomerContactInfo', 'CustomerName', 'ContactInfo'
0 row(s) in 1.2080 seconds
HBase Example -- II

**Entering Records**

```
  hbase(main):008:0> put 'CustomerContactInfo', '00001', 'CustomerName:FN', 'John'
   0 row(s) in 0.2870 seconds

  hbase(main):009:0> put 'CustomerContactInfo', '00001', 'CustomerName:LN', 'Smith'
   0 row(s) in 0.0170 seconds

  hbase(main):010:0> put 'CustomerContactInfo', '00001', 'CustomerName:MN', 'T'
   0 row(s) in 0.0070 seconds

  hbase(main):011:0> put 'CustomerContactInfo', '00001', 'CustomerName:MN', 'Timothy'
   0 row(s) in 0.0050 seconds

  hbase(main):012:0> put 'CustomerContactInfo', '00001', 'ContactInfo:EA', 'John.Smith@xyz.com'
   0 row(s) in 0.0170 seconds

  hbase(main):013:0> put 'CustomerContactInfo', '00001', 'ContactInfo:SA', '1 Hadoop Lane, NY 11111'
   0 row(s) in 0.0030 seconds

  hbase(main):014:0> put 'CustomerContactInfo', '00002', 'CustomerName:FN', 'Jane'
   0 row(s) in 0.0290 seconds

  hbase(main):015:0> put 'CustomerContactInfo', '00002', 'CustomerName:LN', 'Doe'
   0 row(s) in 0.0090 seconds

  hbase(main):016:0> put 'CustomerContactInfo', '00002', 'ContactInfo:SA', '7 HBase Ave, CA 22222'
   0 row(s) in 0.0240 seconds
```
Scan Results

hbase(main):020:0> scan 'CustomerContactInfo', {VERSIONS => 2}
ROW       COLUMN+CELL
00001     column=ContactInfo:EA, timestamp=1383859183030, value=JohnSmith@xyz.com
00001     column=ContactInfo:SA, timestamp=1383859183073, value=1 Hadoop Lane, NY 11111
00001     column=CustomerName:FN, timestamp=1383859182496, value=John
00001     column=CustomerName:LN, timestamp=1383859182858, value=Smith
00001     column=CustomerName:MN, timestamp=1383859183001, value=Timothy
00001     column=CustomerName:MN, timestamp=1383859182915, value=T
00002     column=ContactInfo:SA, timestamp=1383859185577, value=7 HBase Ave, CA 22222
00002     column=CustomerName:FN, timestamp=1383859183103, value=Jane
00002     column=CustomerName:LN, timestamp=1383859183163, value=Doe

2 row(s) in 0.0520 seconds
Using the *get* Command to Retrieve Entire Rows and Individual Values

(1) hbase(main):037:0> get 'CustomerContactInfo', '00001'
COLUMN       CELL
ContactInfo:EA  timestamp=1383859183030, value=John.Smith@xyz.com
ContactInfo:SA  timestamp=1383859183073, value=1 Hadoop Lane, NY 11111
CustomerName:FN  timestamp=1383859182496, value=John
CustomerName:LN  timestamp=1383859182858, value=Smith
CustomerName:MN  timestamp=1383859183001, value=Timothy
5 row(s) in 0.0150 seconds

(2) hbase(main):038:0> get 'CustomerContactInfo', '00001',
   {COLUMN => 'CustomerName:MN'}
COLUMN       CELL
CustomerName:MN  timestamp=1383859183001, value=Timothy
1 row(s) in 0.0090 seconds

(3) hbase(main):039:0> get 'CustomerContactInfo', '00001',
   {COLUMN => 'CustomerName:MN',
   TIMESTAMP => 1383859182915}
COLUMN       CELL
CustomerName:MN  timestamp=1383859182915, value=T
1 row(s) in 0.0290 seconds
Creating, Dropping, and Altering DBs in Apache Hive

(1) $ $HIVE_HOME/bin hive --service cli
(2) hive> set hive.cli.print.current.db=true;
(3) hive (default)> USE ourfirstdatabase;
(4) hive (ourfirstdatabase)> ALTER DATABASE ourfirstdatabase SET DBPROPERTIES
('creator'='Bruce Brown', 'created_for'='Learning Hive DDL');
OK
Time taken: 0.138 seconds
(5) hive (ourfirstdatabase)> DESCRIBE DATABASE EXTENDED ourfirstdatabase;
OK
ourfirstdatabase file:/home/biadmin/Hive/warehouse/ourfirstdatabase.db {created_for=Learning Hive DDL, creator=Bruce Brown}
Time taken: 0.084 seconds, Fetched: 1 row(s)
CREATE (DATABASE | SCHEMA) [IF NOT EXISTS]
database_name
(6) hive (ourfirstdatabase)> DROP DATABASE ourfirstdatabase CASCADE;
OK
Time taken: 0.132 seconds
Hive’s operation modes
Reference

Learning Spark
LIGHTNING-FAST DATA ANALYSIS
Holden Karau, Andy Konwinski, Patrick Wendell & Matei Zaharia
Spark Stack

- Spark SQL structured data
- Spark Streaming real-time
- MLlib machine learning
- GraphX graph processing

Spark Core

- Standalone Scheduler
- YARN
- Mesos
Spark Core

Basic functionality of Spark, including components for:

- Task Scheduling
- Memory Management
- Fault Recovery
- Interacting with Storage Systems
- and more

Home to the API that defines resilient distributed datasets (RDDs) - Spark’s main programming abstraction.

RDD represents a collection of items distributed across many compute nodes that can be manipulated in parallel.
Download Spark

http://spark.apache.org/downloads.html

Download Apache Spark™

1. Choose a Spark release: 2.2.0 (Jul 11 2017)
2. Choose a package type: Pre-built for Apache Hadoop 2.7 and later
3. Choose a download type: Direct Download
4. Download Spark: spark-2.2.0-bin-hadoop2.7.tgz
5. Verify this release using the 2.2.0 signatures and checksums and project release KEYS.

Note: Starting version 2.0, Spark is built with Scala 2.11 by default. Scala 2.10 users should download the Spark source package and build with Scala 2.10 support.

Link with Spark

Spark artifacts are hosted in Maven Central. You can add a Maven dependency with the following coordinates:

groupId: org.apache.spark
artifactId: spark-core_2.11
version: 2.2.0
First language to use — Python

```python
# Python 3: Dict comprehensions to n
>>> def f(x): return {i: i**2 for i in range(x)}
>>> f(5)
{0: 0, 1: 1, 2: 4, 3: 9, 4: 16}

>>> fruits = [fruit.upper() for fruit in fruits] while a < n:
>>> print(fruits)
['BANANA', 'APPLE', 'LIME']
>>> print()
```

## Compound Data Types

Lists (known as arrays in other languages) are one of the compound data types that Python understands. Lists can be indexed, sliced, and manipulated with other built-in functions. More about lists in Python 3

Python is a programming language that lets you work quickly and integrate systems more effectively. Learn More
Spark’s Python Shell (PySpark Shell)

```
bin/pyspark
```
Example 2-1. Python line count

```python
>>> lines = sc.textFile("README.md") # Create an RDD called lines

>>> lines.count() # Count the number of items in this RDD
127

>>> lines.first() # First item in this RDD, i.e. first line of README.md
u'# Apache Spark'
```
Core Spark Concepts

• At a high level, every Spark application consists of a **driver program** that launches various parallel operations on a cluster.

• The driver program contains your application’s main function and defines distributed databases on the cluster, then applies operations to them.

• In the preceding example, the driver program was the Spark shell itself.

• Driver programs access Spark through a SparkContext object, which represents a connection to a computing cluster.

• In the shell, a SparkContext is automatically created as the variable called `sc`. 
Driver Programs

Driver programs typically manage a number of nodes called **executors**.

If we run the count() operation on a cluster, different machines might count lines in different ranges of the file.
Example filtering

```python
>>> lines = sc.textFile("README.md")

>>> pythonLines = lines.filter(lambda line: "Python" in line)

>>> pythonLines.first()
'## Interactive Python Shell'
```

`lambda` — define functions inline in Python.

```python
def hasPython(line):
    return "Python" in line

pythonLines = lines.filter(hasPython)
```
Example — word count

```python
import sys
from operator import add
from pyspark import SparkContext

if __name__ == "__main__":
    if len(sys.argv) != 2:
        print >> sys.stderr, "Usage: wordcount <file>"
        exit(-1)
    sc = SparkContext(appName="PythonWordCount")
    lines = sc.textFile(sys.argv[1], 1)
    counts = lines.flatMap(lambda x: x.split(' '))
        .map(lambda x: (x, 1))
        .reduceByKey(add)
    output = counts.collect()
    for (word, count) in output:
        print "%s: %i" % (word, count)

sc.stop()
```
Resilient Distributed Dataset (RDD) Basics

• An RDD in Spark is an immutable distributed collection of objects.

• Each RDD is split into multiple partitions, which may be computed on different nodes of the cluster.

• Users create RDDs in two ways: by loading an external dataset, or by distributing a collection of objects in their driver program.

• Once created, RDDs offer two types of operations: **transformations** and **actions**.

```python
>>> lines = sc.textFile("README.md")  # create RDD

>>> pythonLines = lines.filter(lambda line: "Python" in line)  # transformation

>>> pythonLines.first()  # action
u'## Interactive Python Shell'
```

Transformations and actions are different because of the way Spark computes RDDs. Only computes when something is, the first time, in an action.
Persistance in Spark

- By default, RDDs are computed each time you run an action on them.
- If you like to reuse an RDD in multiple actions, you can ask Spark to persist it using `RDD.persist()`.
- `RDD.persist()` will then store the RDD contents in memory and reuse them in future actions.
- Persisting RDDs on disk instead of memory is also possible.
- The behavior of not persisting by default seems to be unusual, but it makes sense for big data.

**Example 3-4. Persisting an RDD in memory**

```python
>>> pythonLines.persist
>>> pythonLines.count()
2
>>> pythonLines.first()
u'## Interactive Python Shell'
```

To summarize, every Spark program and shell session will work as follows:

1. Create some input RDDs from external data.
2. Transform them to define new RDDs using transformations like `filter()`.
3. Ask Spark to `persist()` any intermediate RDDs that will need to be reused.
4. Launch actions such as `count()` and `first()` to kick off a parallel computation, which is then optimized and executed by Spark.
Spark SQL

- Tableau
- ...
- JDBC/ODBC
- Your Application
- Spark SQL shell
- Hive
- JSON
- Parquet
- ...

Spark SQL
Spark SQL can be built with or without Apache Hive, the Hadoop SQL engine. Spark SQL with Hive support allows us to access Hive tables, UDFs (user-defined functions), SerDes (serialization and deserialization formats), and the Hive query language (HiveQL). Hive query language (HQL) It is important to note that including the Hive libraries does not require an existing Hive installation. In general, it is best to build Spark SQL with Hive support to access these features. If you download Spark in binary form, it should already be built with Hive support. If you are building Spark from source, you should run sbt/sbt -Phive assembly.
Using Spark SQL — Steps and Example

Example 9-5. Python SQL imports

```python
# Import Spark SQL
from pyspark.sql import HiveContext, Row
```

Example 9-8. Constructing a SQL context in Python

```python
hiveCtx = HiveContext(sc)
```

Example 9-11. Loading and querying tweets in Python

```python
input = hiveCtx.jsonFile(inputFile)
# Register the input schema RDD
input.registerTempTable("tweets")
# Select tweets based on the retweetCount
topTweets = hiveCtx.sql("""SELECT text, retweetCount FROM tweets ORDER BY retweetCount LIMIT 10"""")
```
Query testtweet.json

Get it from Learning Spark Github ==>

https://github.com/databricks/learning-spark/tree/master/files

```json

>>> print topTweets.collect()
[Row(text=u'Adventures With Coffee, Code, and Writing.', retweetCount=0)]
```
Machine Learning Library in Spark — MLlib

An example of using MLlib for text classification task, e.g., identifying spammy emails.

1. Start with an RDD of strings representing your messages.
2. Run one of MLlib’s *feature extraction* algorithms to convert text into numerical features (suitable for learning algorithms); this will give back an RDD of vectors.
3. Call a classification algorithm (e.g., logistic regression) on the RDD of vectors; this will give back a model object that can be used to classify new points.
4. Evaluate the model on a test dataset using one of MLlib’s evaluation functions.
Example: Spam Detection

Example 11-1. Spam classifier in Python

```python
from pyspark.mllib.regression import LabeledPoint
from pyspark.mllib.feature import HashingTF
from pyspark.mllib.classification import LogisticRegressionWithSGD

spam = sc.textFile("spam.txt")
normal = sc.textFile("normal.txt")

# Create a HashingTF instance to map email text to vectors of 10,000 features.
tf = HashingTF(numFeatures = 10000)
# Each email is split into words, and each word is mapped to one feature.
spamFeatures = spam.map(lambda email: tf.transform(email.split(" ")))
normalFeatures = normal.map(lambda email: tf.transform(email.split(" ")))

# Create LabeledPoint datasets for positive (spam) and negative (normal) examples.
positiveExamples = spamFeatures.map(lambda features: LabeledPoint(1, features))
negativeExamples = normalFeatures.map(lambda features: LabeledPoint(0, features))
trainingData = positiveExamples.union(negativeExamples)
trainingData.cache() # Cache since Logistic Regression is an iterative algorithm.

# Run Logistic Regression using the SGD algorithm.
model = LogisticRegressionWithSGD.train(trainingData)

# Test on a positive example (spam) and a negative one (normal). We first apply
# the same HashingTF feature transformation to get vectors, then apply the model.
posTest = tf.transform("O M G GET cheap stuff by sending money to ...").split(" "))
negTest = tf.transform("Hi Dad, I started studying Spark the other ...").split(" "))
print "Prediction for positive test example: %g" % model.predict(posTest)
print "Prediction for negative test example: %g" % model.predict(negTest)
```
Feature Extraction Example — TF-IDF

Example 11-7. Using HashingTF in Python

```python
>>> from pyspark.mllib.feature import HashingTF

>>> sentence = "hello hello world"
>>> words = sentence.split()   # Split sentence into a list of terms
>>> tf = HashingTF(10000)   # Create vectors of size S = 10,000
>>> tf.transform(words)
SparseVector(10000, {3065: 1.0, 6861: 2.0})

>>> rdd = sc.wholeTextFiles("data").map(lambda (name, text): text.split())
>>> tfVectors = tf.transform(rdd)   # Transforms an entire RDD
```

Example 11-8. Using TF-IDF in Python

```python
from pyspark.mllib.feature import HashingTF, IDF

# Read a set of text files as TF vectors
rdd = sc.wholeTextFiles("data").map(lambda (name, text): text.split())
tf = HashingTF()
tfVectors = tf.transform(rdd).cache()

# Compute the IDF, then the TF-IDF vectors
idf = IDF()
idfModel = idf.fit(tfVectors)
tfIdfVectors = idfModel.transform(tfVectors)
```
Homework #1: Installation, Data Storage and Processing (due September 20, 5pm)

Questions:

1. Installation and Environment Setup (10 pts)
   Using 3 Screenshots to demonstrate you can successfully run Jupyter Notebook and any two of installed software as described by the HW#1 Installation Instruction Tutorial posted on CourseWorks on Sept. 8.

2. Hadoop (30 pts)
   Submit both screenshots and a script file `hw1_hadoop.sh` containing your commands which:
   a. Demonstrate you can manage your file systems (5)
   b. Upload a file to HDFS (5)
   c. Inspect the last kilobytes of content of the file (10)
   d. Run the mapreduce word count example with the provided 2 text files and find the 3 most frequent words (don’t worry about cleaning punctuations) (10)
   https://drive.google.com/drive/folders/1srgWQiaj_kCZR4BIw8dzPXFLu0ZOy7gs?usp=sharing

3. HBase (20 pts)
   a. Import a table from an external file in HDFS (10)
   b. Display the top 10 rows of content of the table (5)
   c. Display the top 10 rows of content with some specific values (5)
Homework #1: Installation, Data Storage and Processing (due September 20, 5pm)

4. Hive (20 pts)
   a. Import a table from an external file in HDFS (5)
   b. Do five queries and show the results (15)

5. Spark (20 pts)
   Run the Word Count program with your chosen programming language (12)
   https://spark.apache.org/examples.html
   On the provided text, list the top 3 most frequent words (8)
   https://drive.google.com/drive/folders/1sBcQXMQGIHPeuk12J0wwCJXNzNrPTYgz?usp=sharing

The following datasets are what you can use for Question 2-4 (in any order)
IMDB dataset: (title.basics.tsv.gz)
https://datasets.imdbws.com/

Airline on-time performance: (year 2008)
NYC TLC Trip Data (2017 January Green Taxi)
NYPD Motor Vehicle Collisions
https://drive.google.com/file/d/1V1H_-0IwrOwC3CM-n2ZW96ZDudwHblOy/view?usp=sharing
Homework Late Submission Policy

Th 5pm: submission deadline
Fr 5pm: 10% penalty
Sa 5pm: 20% penalty
Su 5pm: 30% penalty
Any submission after Sunday 5pm will not be accepted.
Questions?

HW #1 Questions Tutorial by TA