Analyzing Reddit Data with MapReduce

Introduction

In this project we will be exploring the MapReduce programming model to process data in a distributed manner. With data sets getting larger and larger, it may not be enough to process data with a single system. With the MapReduce programming model we'll be able to speed up processing time by parallelizing work among multiple systems. We'll be using MapReduce to do some simple data processing of a social media website.

Background

Intro to MapReduce

MapReduce is a programming model that applies a split-apply-combine strategy. MapReduce consists of two procedures: a map procedure and a reduce procedure. The map procedure performs filtering, sorting, on data, then the reduce procedure summarizes the output created by the map procedure. During the execution of a MapReduce program, data is split among multiple nodes that apply the map procedure to the initial data input, creating intermediate results. Intermediate results are then shuffled, sorted, and distributed to nodes which apply the reducing procedure to the intermediate results. MapReduce can be used to do distributed string matching, reverse index, count URL access frequency, Log analysis, and overall data analysis.

Learn more about MapReduce here: *https://en.wikipedia.org/wiki/MapReduce*

Intro to Reddit

Reddit is a popular social news website where content is organized into user created boards called "subreddits". Users can add posts to subreddits. Users can subscribe to subreddits to see new posts posted within the subreddit on their homepage. In its basic form, the homepage of Reddit would present a curated feed of popular posts among the subreddits the user is subscribed to. Each post can be "upvoted" or "downvoted". A post can be commented on and each comment can also be "upvoted" or "downvoted". An upvote and downvote can be compared to "like" and "dislike" on Facebook. A user can"upvote" a post if they believe it is contributes to the current "conversation". Likewise, users can "downvote" a post if they believe it does not contribute to the current "conversation". Through the use of upvoting and downvoting mechanisms, users help Reddit algorithms rank and determine quality content that users would be interested in.

Learn more about reddit here: *https://en.wikipedia.org/wiki/Reddit*

Reddit as a Basis for Social Media Research

Within social media, certain content can be posted multiple times. In the case of Reddit, similar content can be posted multiple times in more than one subreddit. In the context of Reddit, when some content

is resubmitted in another post, it is called a "repost". Analysis of reposts help provide a means in understanding social media content placement. Even though two users can submit two different posts using the same content, each post could perform differently. It becomes apparent that there is a multitude of factors that contribute to the performance of a post besides the content such as: the number of user subscribed to the subreddit, how attractive the post title is, the time of day the post was uploaded, and so on. With the large user base of Reddit along with Reddit's board system of subreddits providing simple modeling of social communities, it has been seen to be a good platform to research social media interactions, patterns, and phenomena.

Within the paper "*What's in a name? Understanding the Interplay between Titles, Content, and Communities in Social Media*", researchers Lakkarajue et al. did just this. They collected Reddit data to analyze how certain factors contribute to the popularity of reposts. For their data set, they collected information on reposts of images as it was significantly easy to identify reposted images using reverse image search. Using the data set, they were able to derive models on how time, social context, and presentation affected a post's performance, ultimately helping them understand when, where, and how a post should be presented to maximize its reach.

You can read more about their research here: `http://i.stanford.edu/~julian/pdfs/icwsm13.pdf`

Calculating Image Impact using MapReduce

The goal of this project would be to create a MapReduce program to process the Reddit data Lakkarajue et al. collected in a way that will help us do some data analysis of our own.

Image Impact on Reddit

In this project, we will compare images based on the amount of impact they had on Reddit. We can do this by measuring the amount of engagement each photo induced. This can be done by through simply counting user interactions.

As you've learned, users on Reddit can interact with posts by upvoting, downvoting, or commenting. We can simply define a post's impact to be the number of times it was upvoted, downvoted, or commented on:

$$Impact = upvotes + downvotes + comments$$

This would calculate the impact of a particular post, not a particular image. To calculate the impact of an image, we can simply extend our definition of the impact of a post to define the impact of a particular image to be the sum of all impact values for all the posts associated to a particular image.

$$Impact_{id} = \sum upvotes_p + downvotes_p + comments_p, \forall p \in P_{id}$$

In this equation, p is a post within P, the set all of posts associated with some image with an id id.

Using MapReduce

Since each entry in our data set represents a post with information on the number of upvotes, downvotes, comments, and the image that was posted, this would be a great opportunity to use MapReduce. In this case we can use MapReduce to calculate and map impact values calculated per post to the image id each post is associated with, and then calculate overall impact by summing up (ie. reducing) all the impact values calculated for each image id.

Example:
As a basic example lets say we have three entries representing three posts:

image_id	upvotes	downvotes	comments
123	200	50	6
152	350	100	12
123	450	5	12

We can see the first and third entry are posts associated to an image with the id 123 and the second entry is a post associated with a image 152.

In the end, the result of our program should look like such:

```
123         732
152         462
```

From this we can see more people interacted with posts associated with the image of id 123 compared to the posts associated with image of id 152. We could naively say the image of id 123 had more impact than the image of id 152. If image of id 123 was perhaps an image of a cat and image of id 152 was an image of a shoe, we can predict that if we posted a similar image of a cat to that in image of id 123, it could potentially be more engaging and get more exposure than if we posted a picture of a shoe similar to the image of 152.

The Reddit Data

Now the reddit data we will be using is more involved. It is based on real the Reddit data collected by Lakkarajue et al. for their study.

The following data files are supplied:

- *reddit-data-small.txt* - A small set of reposts associated with 6 images, sorted by image id
- *reddit-data-large.txt* - This contains the whole image reposts, sorted by unix time of submission
- *reddit-cols.txt* - List of values describing the columns of each post entry

The Reddit Data is formatted as follows:

- Each line within the text file represents a Reddit post.
- Each line is formatted as follows:

  ```
  unix_time    image_id    upvotes    downvotes    comments    title
  ```

  where $unix\_time$ is the time the post was submitted, $image\_id$ is the id of the image associated with the post, $upvotes$ is the number of upvotes the post has, $downvotes$ is the number of downvotes the post has, $comments$ is the number of comments the post has, and $title$ is the title of the post.
- An example of a line representing a post of an image of id 12, with 113 upvotes, 10 downvotes, 17 comments, and the title, "Hi, this is my cat.", would be:

  ```
  1542634712    12    113    10    17    Hi,_this_is_my_cat.
  ```

Tab Delimited Files: Bear in mind when working with the Reddit data files, that they are tab-delimited text files, meaning the values in each line are separated by tabs.

Map Reduce Splits: When working with items such as text files (.txt), the Map Reduce frameworks automatically splits the input to each mapper as a line.

Getting Started

To get started, it is important to understand MapReduce semantics and how to run MapReduce. There are plenty of tutorials out there. Look at the example word count program used in the Apache MapReduce tutorial. A link to the tutorial will be included in the end of this document. The word count example goes over MapReduce by making a program that counts the number of occurrences for each word in an input file. Learn how it works, how to compile it, and how to get it running on a Hadoop Cluster. After you get the word count example to work, make a similar program that will use the Reddit data. Start off with the smaller data set so you can see whether or not the values are adding up correctly. When you think your program successfully works with the smaller data set, then you can try running it on the whole data set.

As for Hadoop Clusters, you have multiple options. You can use Amazon Web Service's EMR service which allows user to run their own Hadoop Cluster in the cloud. Rutgers also has a Hadoop Cluster to use.

Use Amazon Web Services Elastic MapReduce Service: It's highly recommended to use AWS EMR to run your MapReduce project. By using AWS EMR you will be able get experience with working with Amazon's popular cloud services being used in industry today along side with experience working with a Hadoop Cluster. The Rutgers Hadoop Cluster can also be used, but it could be unstable due to it being recently started and untested under high loads.

Submission

To submit your project, have ONE of the group members submit two things:

1. **A zip file of your project.**
2. **A pdf document consisting of the following things:**
    - List of group members
    - A short paragraph or two describing what you accomplished.
    - A short paragraph or two describing any issues you may have encountered and how you think you could have solved them. If you didn't have any simply state that you didn't have any issues.

Useful Links and Resources

MapReduce Guides

- *Apache MapReduce Tutorial*
  https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
- *HDFS Overview and Commands*
  https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-common/FileSystemShell.html
- *Running MapReduce on AWS*
  https://github.com/DaveedDomingo/MapReduce-Reddit-Project/blob/master/MapReduceOnAWS.pdf
- *Running MapReduce at Rutgers*
  https://github.com/DaveedDomingo/MapReduce-Reddit-Project/blob/master/MapReduceAtRutgers.pdf

Additional Hints

Tutorials are you friend: Follow online guides and tutorials on youtube to get the main idea of how to start building your web service and do other stuff.

Understand Map Reduce through examples: Look at completed MapReduce projects to get a sense of how MapReduce works and the semantics of a MapReduce program.

Additional Questions

Don't be afraid to ask question! If you have any questions about the project or are having any issues, email me at **David.Domingo@rutgers.edu** or post your question to the class slack channel!