Hadoop and MapReduce

Tiecheng Su
University of Rochester

Abstract

Big data is high volume, high velocity, and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. Hadoop is one of the most important tool the analyze big data.

1. Hadoop Distributed File System(HDFS)

Hadoop store data in HDFS. Hadoop distributed file system is a distributed, scalable, and portable file system for the Hadoop framework. A Hadoop cluster has a single namenode(metadata) and a cluster of datanodes.

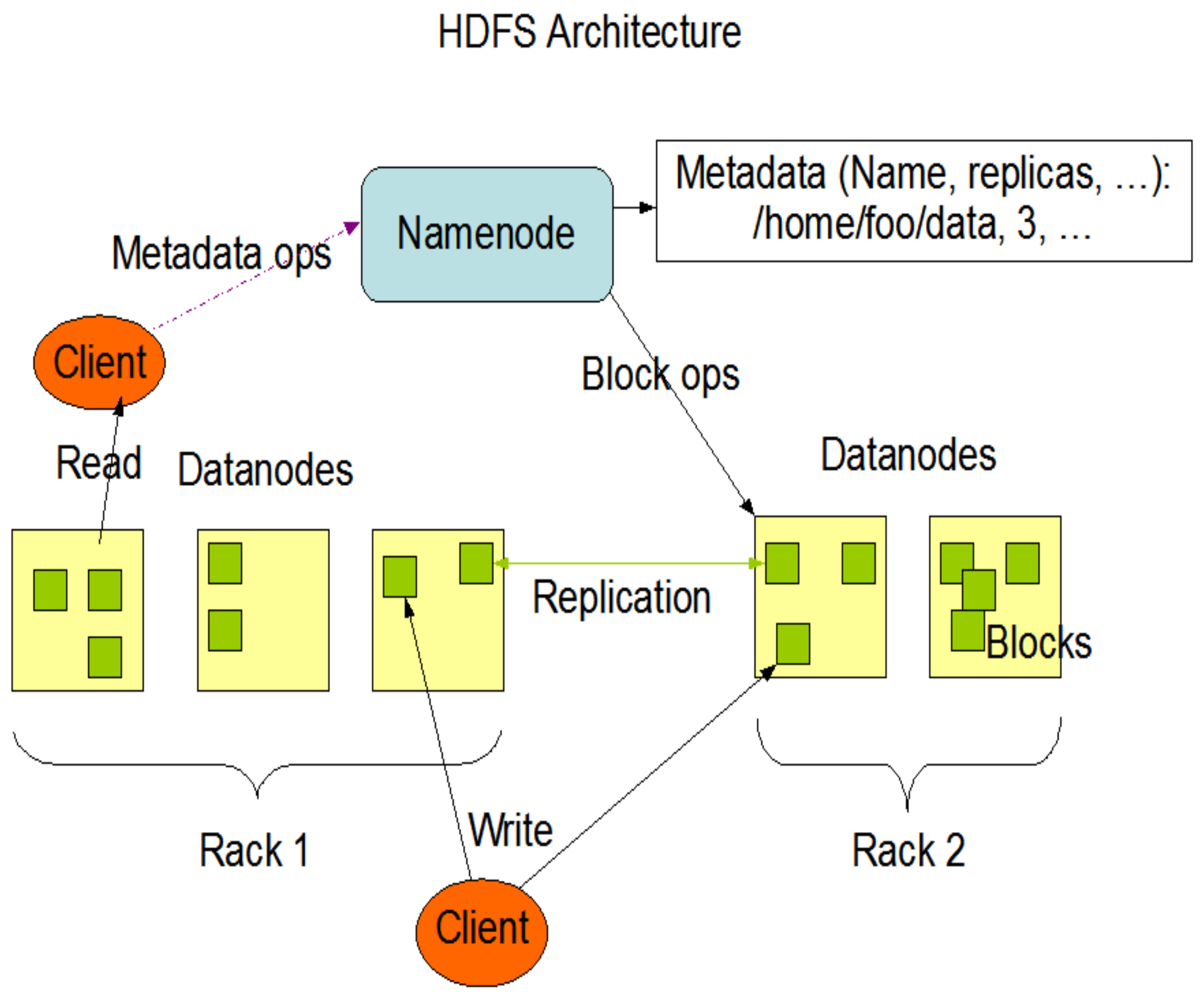
1.1 NameNode

NameNode keeps the directory tree of all files in the file system, and tracks where across the cluster the file data is kept.

NameNode is a Single Point of Failure for the HDFS Cluster. HDFS is not currently a High Availability system. When the NameNode goes down, the file system goes offline. There is an optional SecondaryNameNode that can be hosted on a separate machine.

1.2 DataNode

A DataNode stores data in the HDFS. With the default replication value, 3, data is stored on three nodes; two on the same rack, and one on a different rack.



2. MapReduce

Hadoop process data with MapReduce

2.1 Mapper

In the mapper, user-provided code is executed on each key/value pair from the record reader to produce new key/value pairs, called the intermediate pairs

2.2 Combiner

An optional localized reducer. It takes the intermediate keys from the mapper and applies a user-provided method to aggregate values in the small scope of that one mapper.

2.3 Shuffle and sort

This step takes the output files written by all of the partitioners and downloads them to the local machine in which the reducer is running. These individual data pieces are then sorted by key into one larger data list.

2.4 Reducer

The reducer takes the grouped data as input and runs a reduce function once per key grouping. The function is passed the key and an iterator over all of the values associated with that key.

2.5 JobTracker

The JobTracker is the service within Hadoop that farms out MapReduce tasks to specific nodes in the cluster.

2.6 TaskTracker

A TaskTracker is a node in the cluster that accepts tasks - Map, Reduce and Shuffle operations - from a JobTracker.

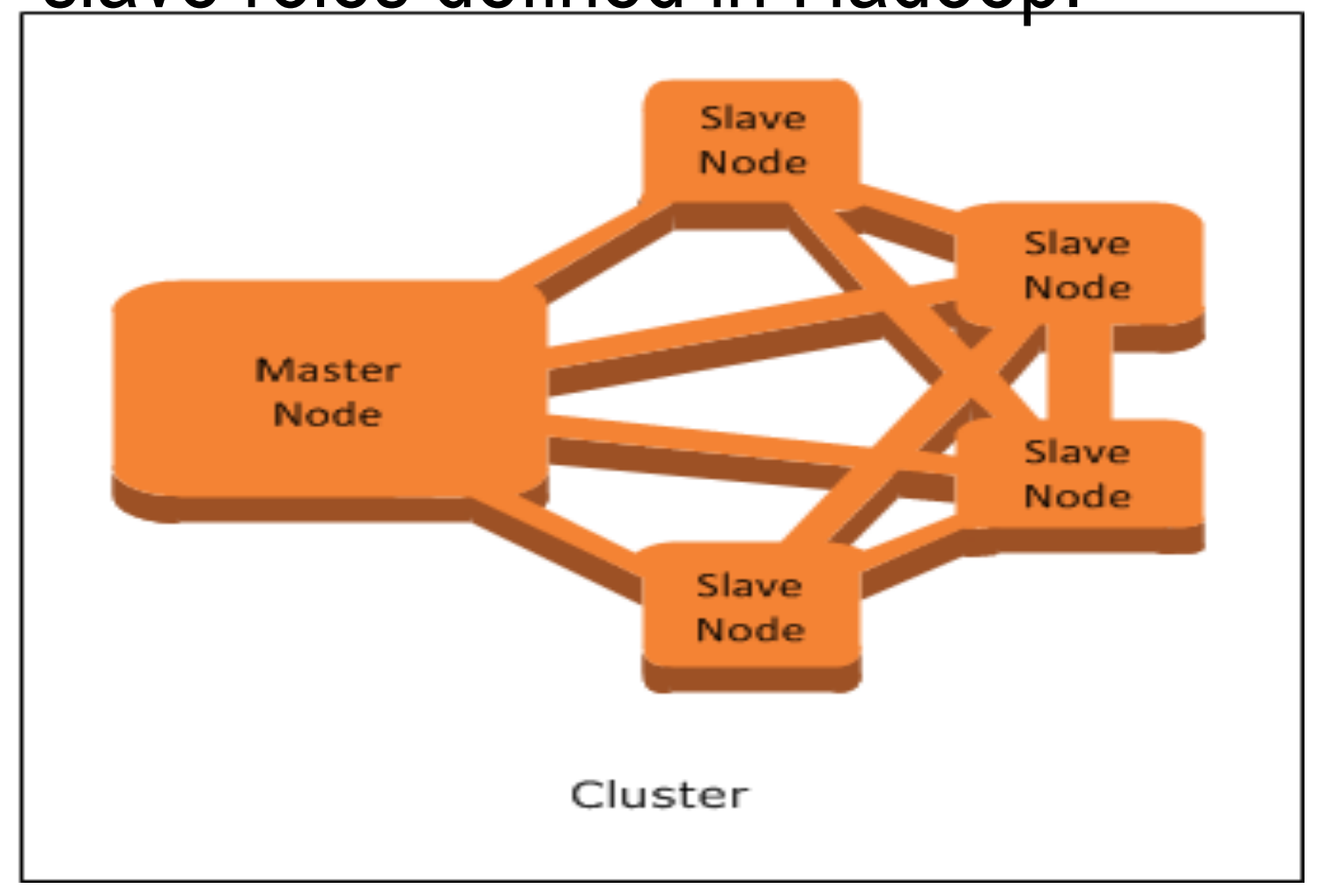
3. Amazon Elastic MapReduce

In this project, I am using Amazon Elastic MapReduce(EMR). Amazon EMR processes big data across a Hadoop cluster of virtual servers on Amazon Elastic Compute(EC2) and Amazon Simple Storage Service(S3). The elastic in EMR's name refers to its dynamic resizing ability, which allows it to ramp up or reduce resource use depending on the demand at any given time.

3.1 Nodes

The central component of Amazon EMR is the cluster. A cluster is a collection of Amazon Elastic Compute Cloud (Amazon EC2) instances. Each instance in the cluster is called a node.

Amazon EMR defines three roles for the servers in a cluster, which are referred to as node types. The Amazon EMR node types map to the master and slave roles defined in Hadoop.



4. Data

In this project, I analyzed a dataset of Reddit comments from May 2015. This dataset comes from Kaggle. The whole dataset is a ~30 GB SQLite database, but I selected the tip 100 subreddits based on comment count, resulting in a ~4.5 GB file.

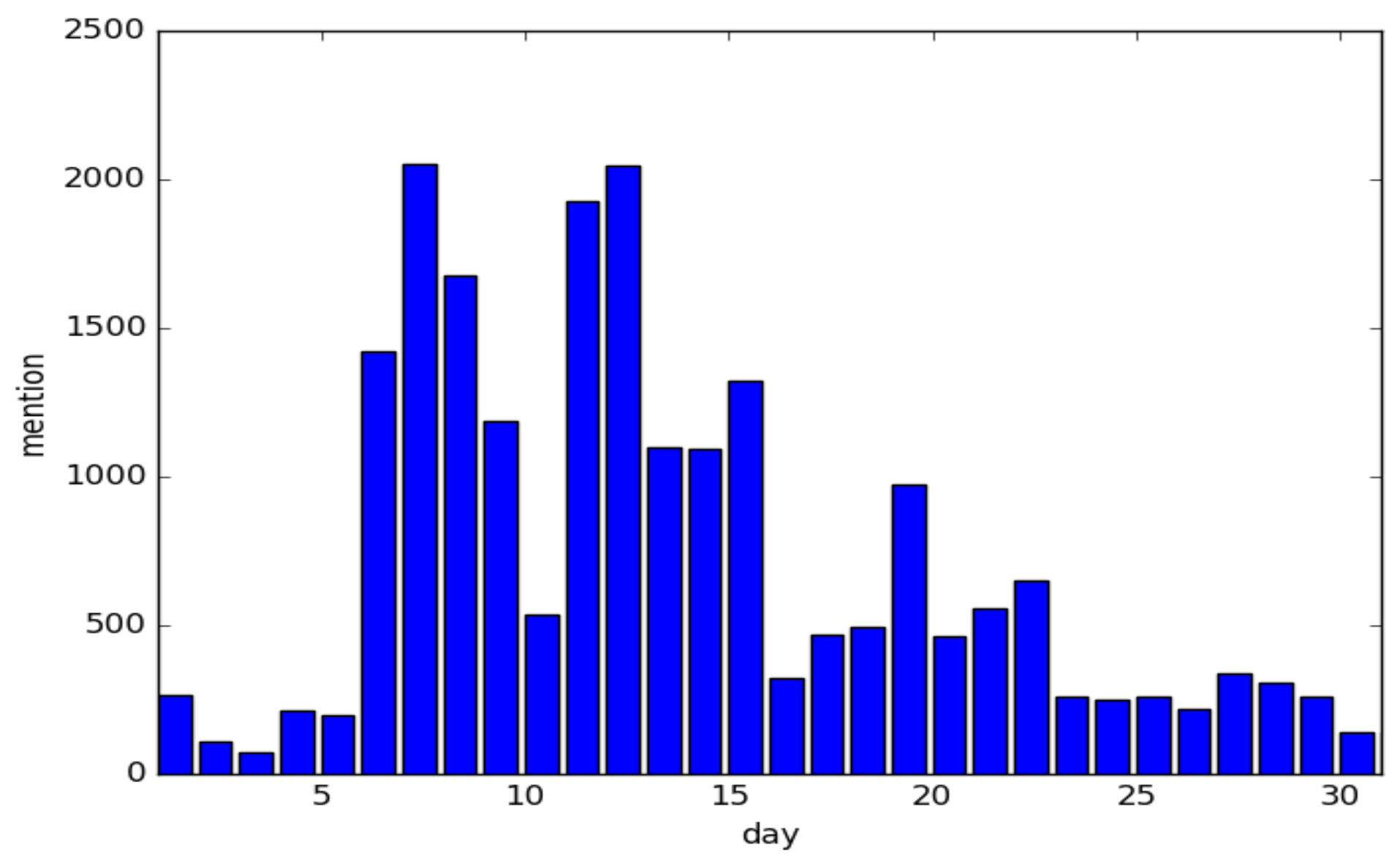
The data scheme is as follow:
subreddit: The subreddit the comment was posted in
author: Username of the comment author
body: Comment text
create_utc: When the comment was posted
ups: Comment upvotes
downs: Comment downvotes
gilded: 1 if the user was given Reddit gold for the comment, 0 otherwise
archived: 1 if the comment was archived, 0 otherwise

5. Reddit comments big data analysis

5.1 Deflategate!

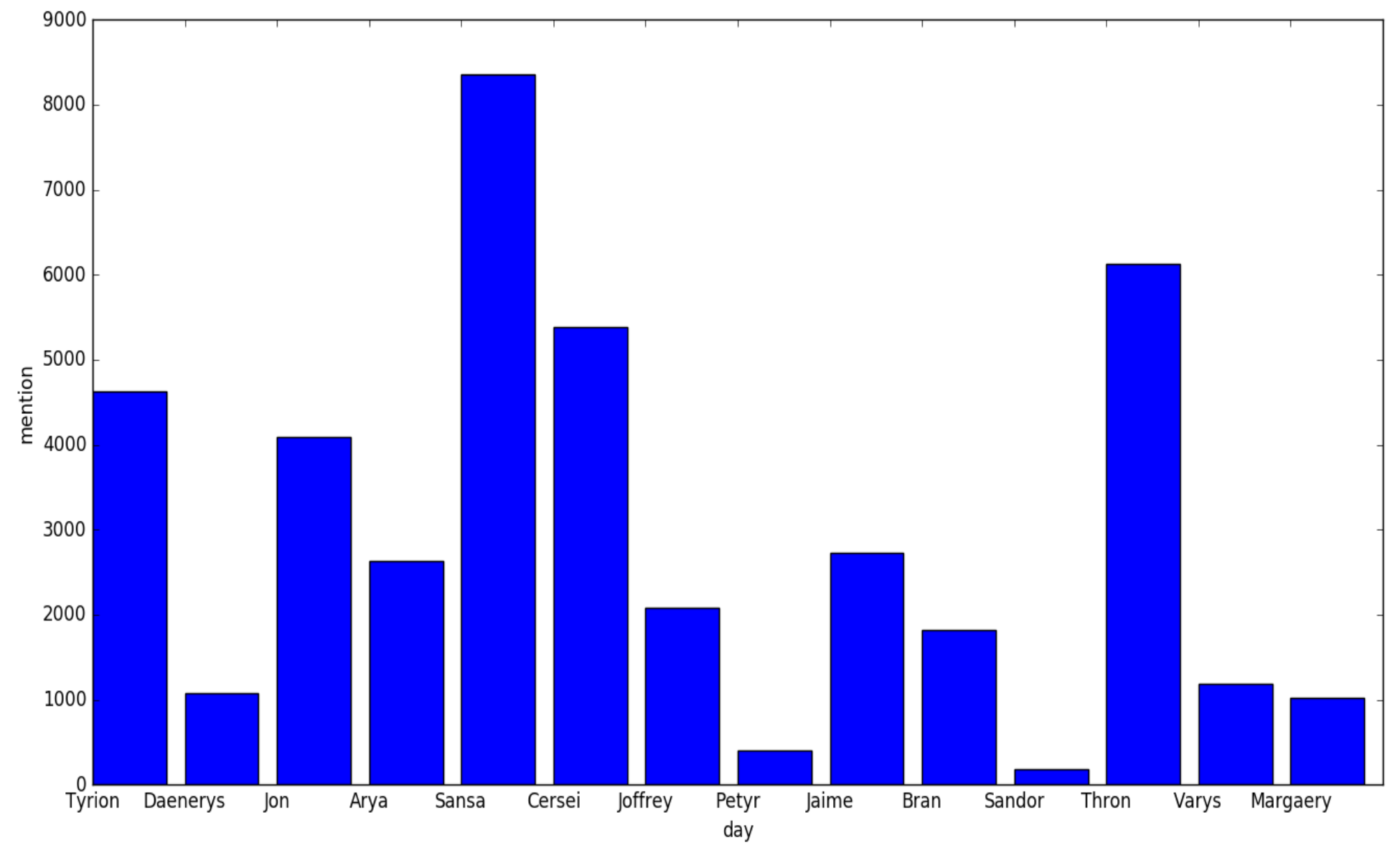
Way back in 2015, the New England Patriots allegedly deflated footballs used in the AFC championship game. Multiple suspensions, a federal judge got involved, dogs and cats living together. I'm curious how much people were talking about it over the month.

I write a mapper and reducer that counts the number of mentions of "Deflategate" or "Tom Brady" for each day of May 2015 in the nfl subreddit



5.2 Game of Thrones

Game of Thrones is definitely one of my favorite TV series, I want to see who's the character that people talk about most on Reddit back in 2015. We can see that Sansa get most votes. Other popular characters are Tyrion, Jon, Cersei and Thron.



5.3 What's popular with Redditors?

Each comment belongs to a particular subreddit, we would like to know for the given time period what were the 10 most popular subreddits? We can see from the result that AskReddit and leagueoflegens are always the top two of May 2015. nba, nfl and funny are also very popular.

week	1	2	3	4
1	AskReddit	AskReddit	AskReddit	AskReddit
2	leagueoflegends	leagueoflegends	leagueoflegends	leagueoflegends
3	nfl	nba	funny	nba
4	pics	funny	nba	funny
5	funny	nfl	pcmasterrace	pics
6	nba	pics	DestinyTheGame	pcmasterrace
7	news	videos	news	DestinyTheGame
8	videos	todayilearned	videos	todayilearned
9	DotA2	worldnews	pics	soccer
10	todayilearned	news	worldnews	videos

Acknowledgements

Thanks for Amazon Web Services for providing Elastic MapReduce