Introduction to Big Data

139

Explore a large-scale dataset

Introduction

In this first project, we'll get a feel for big data by diving head first into analysing a large dataset. We will work with the hourly page view statistics from Wikipedia.

Wikimedia maintains hourly page view statistics for all objects stored in wikimedia servers as publically accessible datasets. The dump goes back all the way to 2007 and at least a couple of Terabytes in size in compressed format.

Exploring the Dataset

Every request made to Wikipedia's servers is serviced by a squid cache proxy which logs the request. One file is written for every hour . Each line of this file corresponds to a single, unique element from the Wikimedia servers. Each line is in the following format:

```
<project name> <page title> <number of accesses> <total data returned in bytes>
```

<project name> has two parts, a language identifier and a subproject suffix. The following abbreviations are used in the subproject suffic:

1. **<no suffix>** : wikipedia
2. **.b** : wikibooks
3. **.d** : wikitionary
4. **.m** : wikimedia
5. **.mw** : wikipedia mobile
6. **.n** : wikinews
7. **.q** : wikiquote
8. **.s** : wikisource
9. **.v** : wikiversity
10. **.w** : mediawiki

For Example, the following line:

```
fr.b Special:Recherche/All_Mixed_Up 1 730
```

denotes that from the French Wikibooks page, the page **Special:Researche/All_Mixed_Up** was accessed once and 730 bytes were transferred in total.

In this project, we will focus on analyzing the page view logs from June 2013. The data has been uploaded to

the s3 location: **`s3://wikipediatraf/201306-gz/`**. You can use **`s3cmd`** or S3 Browser to explore the location. Refer to the page Amazon S3 in the Project Primer for details on accessing S3. The dataset contains 1 file for every hour in June 2013 in gzipped format, 720 files in total.