Practical Data Science with Hadoop® and Spark

Ofer Mendelevitch, Casey Stella, Douglas Eadline

ISBN-13: 978-0-13-402414-1 ISBN-10: 0-13-402414-1

Date: April 19, 2017

Errata:

Chapter 3, page 34 (printed version)

Original:

> #### Distributed File System
>
> The purpose of the secondary NameNode is to perform periodic checkpoints that preserve the status of the NameNode should it fail. For more information about using HDFS see Appendix A.

Corrected:

> #### Distributed File System
>
> The purpose of the secondary NameNode is to perform periodic checkpoints that preserve the status of the NameNode should it fail. For more information about using HDFS see Appendix B.

Chapter 4, page 68 (printed version)

Original:

> #### Using Sqoop V2: A Basic Example
>
> To better understand how to use Sqoop in practice, we're going to demonstrate how to configure and use Sqoop version 2 via a simple example, which can then be …

Corrected:

> #### Using Sqoop V1: A Basic Example
>
> To better understand how to use Sqoop in practice, we're going to demonstrate how to configure and use Sqoop version 1 via a simple example, which can then be …

Chapter 4, page 72 (printed version)

Original:

Step 3: Import Data Using Sqoop:

Since there was only one mapper process, only one copy of the query needed to be run on the database. The results are also reported in single file (`part-m-0000`). Multiple mappers can be used to process the query if the `--split-by` option is used. The split-by option is a way to parallelize the SQL query. Each parallel task runs a subset of the main query with results partitioned by bounding conditions inferred by Sqoop. Your query must include the token `$CONDITIONS`; this is a placeholder for Sqoop to put in unique condition expression based on the `--split-by` option, and Sqoop automatically populates this with the right conditions for each mapper task. Sqoop will try to create balanced sub-queries based on a range of your primary key. However, it may be necessary to split on another column if your primary key is not uniformly distributed.

Corrected:

Step 3: Import Data Using Sqoop:

Since there was only one mapper process, only one copy of the query needed to be run on the database. The results are also reported in single file (`part-m-0000`). Multiple mappers can be used to process the query if the `--split-by` and `-m` option are used. The `-m` specifies how many parallel mapper processes to start. The split-by option is a way to partition the SQL query. If there is no explicit argument for `--split-by`, Sqoop will try to create balanced sub-queries based on a range of your primary key. However, it may be necessary to split on another column if your primary key is not uniformly distributed. Each parallel mapper task runs a subset of the main query with results partitioned by bounding conditions inferred or given to Sqoop. Your query must include the token `$CONDITIONS`; this variable is a placeholder for Sqoop to put in a unique condition expression, based on the `--split-by` option for each independent mapper task.

Chapter 9, page 154 (printed version)

Original:

Distance Functions

In the case where the positive outcome (A=1 or B=1) is more important then the negative outcome, a preferred variant of this metric is

$$d(A,B) = \frac{R+S}{R+S+Q}$$

also known as the Jaccard coefficient.

Corrected:

Distance Functions

In the case where the positive outcome (A=1 or B=1) is more important then the negative outcome, a preferred variant of this metric is

$$d(A,B) = \frac{R+S}{R+S+Q}$$

also known as the Jaccard distance.