APPENDIX A

Additional Information and Resources

In this section, we quickly cover additional information that may be helpful for some readers. We'll look at other types of algorithms, another way to import data into Neo4j, and another procedure library. There are also some resources for finding datasets, platform assistance, and training.

Other Algorithms

Many algorithms can be used with graph data. In this book, we've focused on those that are most representative of classic graph algorithms and those of most use to application developers. Some algorithms, such as coloring and heuristics, have been omitted because they are either of more interest in academic cases or can be easily derived.

Other algorithms, such as edge-based community detection, are interesting but have yet to be implemented in Neo4j or Apache Spark. We expect the list of graph algorithms used in both platforms to increase as the use of graph analytics grows.

There are also categories of algorithms that are used with graphs but aren't strictly graphy in nature. For example, we looked at a few algorithms used in the context of machine learning in Chapter 8. Another area of note is similarity algorithms, which are often applied to recommendations and link prediction. Similarity algorithms work out which nodes most resemble each other by using various methods to compare items like node attributes.

Neo4j Bulk Data Import and Yelp

Importing data into Neo4j with the Cypher query language uses a transactional approach. Figure A-1 illustrates a high-level overview of this process.

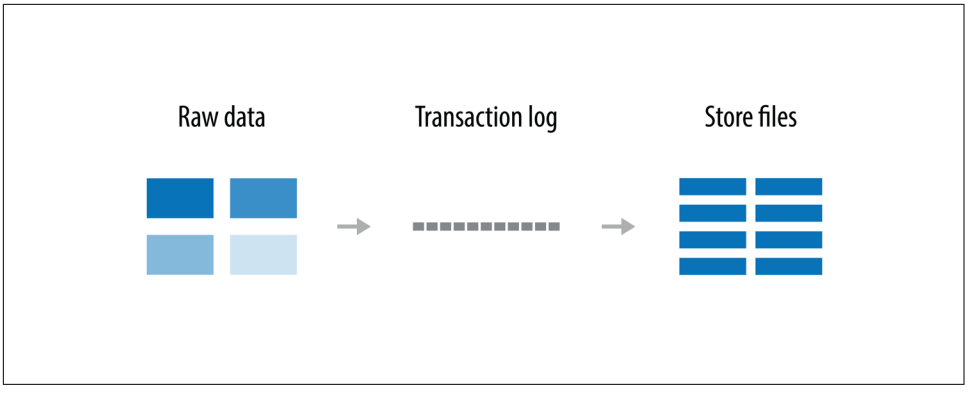

Figure A-1. Cypher-based import

While this method works well for incremental data loading or bulk loading of up to 10 million records, the Neo4j Import tool is a better choice when importing initial bulk datasets. This tool creates the store files directly, skipping the transaction log, as shown in Figure A-2.

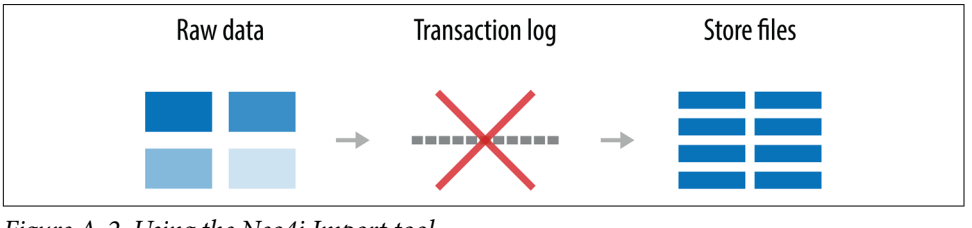

Figure A-2. Using the Neo4j Import tool

The Neo4j Import tool processes CSV files and expects these files to have specific headers. Figure A-3 shows an example of CSV files that can be processed by the tool.

Nodes

id:ID(User)	name
1234	Bob
1235	Alice
1236	Erika

id:ID(Review)	text	stars
678	Awesome	3
679	Mediocre	2
680	Really bad	1

Relationships

:START_ID(User)	:END_ID(Review)
1234	678
1235	679
1236	680

Figure A-3. Format of CSV files that Neo4j Import processes

The size of the Yelp dataset means the Neo4j Import tool is the best choice for getting the data into Neo4j. The data is in JSON format, so first we need to convert it into the format that the Neo4j Import tool expects. Figure A-4 shows an example of the JSON that we need to transform.

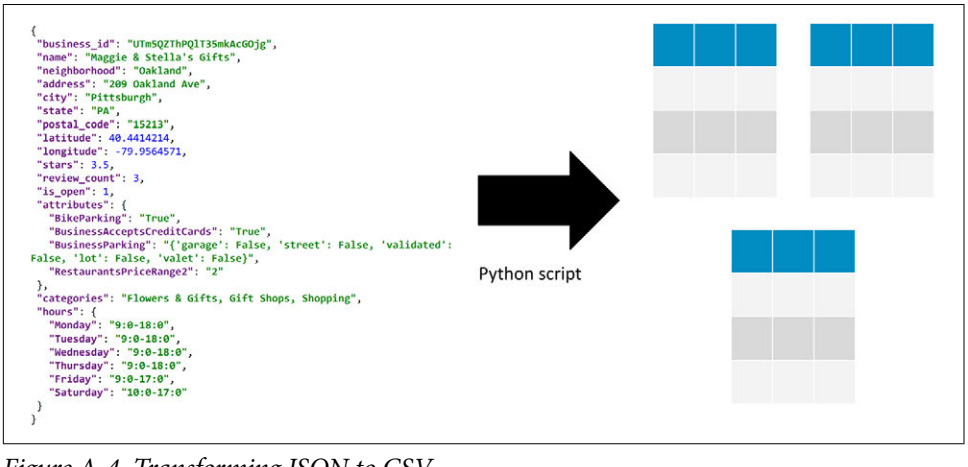

Figure A-4. Transforming JSON to CSV

Using Python, we can create a simple script to convert the data to a CSV file. Once we've transformed the data into that format we can import it into Neo4j. Detailed instructions explaining how to do this are in the book's the book's resources repository.

APOC and Other Neo4j Tools

Awesome Procedures on Cypher (APOC) is a library that contains more than 450 procedures and functions to help with common tasks such as data integration, data cleaning, and data conversion, and general help functions. APOC is the standard library for Neo4j.

Neo4j also has other tools that can be used in conjunction with their graph algorithms library such as an algorithms "playground" app for code-free exploration. These can be found on their developer site for graph algorithms.

Finding Datasets

Finding a graphy dataset that aligns with testing goals or hypotheses can be challenging. In addition to reviewing research papers, consider exploring indexes for network datasets:

- The Stanford Network Analysis Project (SNAP) includes several datasets along with related papers and usage guides.
- The Colorado Index of Complex Networks (ICON) is a searchable index of research-quality network datasets from various domains of network science.
- The Koblenz Network Collection (KONECT) includes large network datasets of various types in order to perform research in network science.

Most datasets will require some massaging to transform them into a more useful format.

Assistance with the Apache Spark and Neo4j Platforms

There are many online resources for the Apache Spark and Neo4j platforms. If you have specific questions, we encourage you to reach out their respective communities:

- For general Spark questions, subscribe to *users@spark.apache.org* at the Spark Community page.
- For GraphFrames questions, use the GitHub issue tracker.
- For all Neo4j questions (including about graph algorithms), visit the Neo4j Community online.

Training

There are a number of excellent resources for getting started with graph analytics. A search for courses or books on graph algorithms, network science, and analysis of networks will uncover many options. A few great examples for online learning include:

- Coursera's Applied Social Network Analysis in Python course
- Leonid Zhukov's Social Network Analysis YouTube series
- Stanford's Analysis of Networks course includes video lectures, reading lists, and other resources
- Complexity Explorer offers online courses in complexity science