Dissertation Project

A Comparison of NoSQL and Indexing Solutions for Big Data

Author: Callum George William Guthrie

Supervisor: Dr. Albert Burger

Heriot Watt University

Edinburgh, Scotland

A dissertation submitted in partial fulfilment of the requirements for the degree of Bachelor of Science.

May 2016

Declaration

I, Callum George William Guthrie confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed:

Date:

Abstract

The era of Big Data is upon us bringing with it a range of new challenges, and encouraging the formulation of new approaches for cleaning, processing and using these enormous amounts of data. These new methods have led to the development of a range of technologies designed to meet the needs of Big Data.

This project focuses on a subset of the new technologies, in particular those products designed to deliver high performance querying of large data sets. It shall compare leading NOSQL solutions (e.g., MongoDB [1], and Neo4j [2]) against modern search and analytics engines (e.g., ElasticSearch [3], and SOLR [4]). The overall goal is to compare and contrast the functionally, performance and analytical capabilities of the different solutions. With the ultimate aim of gaining an understanding of, and insight into, these technologies and their application to Big Data.

The student will produce several versions of a prototype application; with each version employing a different technology (or approach). This work will be undertaken using a real world dataset from the biological environment.

[1] http://www.mongodb.org/about/introduction/

[2] http://www.neo4j.org/

[3] http://www.elasticsearch.org/webinars/introduction-elk-stack/

[4] http://lucene.apache.org/solr/

Contents

Chapter 1

Dataset

The data sources used in the applications are taken from the combined research projects undertaken on the e-Mouse Atlas Project (EMAP) by Dr Duncan Davidson and Prof Richard Baldock.

1.1 EMA

1.2 EMAPA

1.3 EMAGE

Chapter 2

Data Sources

2.1 EMA Database

2.2 OBO

2.3 OWL

Bibliography

[1] Sergey Brin and Lawrence Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833, 2012.

[2] Google Inc. Google home page, 2015.

[3] Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. 1995.