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/

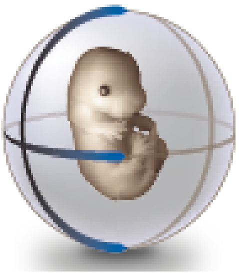
[4] http://lucene.apache.org/solr/

Contents

Chapter 1

Dataset Origins

The dataset used in the prototype applications is a real world dataset taken from the biological environment. The data is constructed by an ontology derived from the combined research projects undertaken on the e-Mouse Atlas Project (EMAP) by Dr Duncan Davidson and Professor Richard Baldock.



The name EMAP carries a certain amount of ambiguity as it is the name of the project that developed the anatomy, and is also the name of the anatomy itself. Therefore with the motivation of clarity, I will refer to the project that developed the anatomy as e-Mouse Atlas (EMA) and the name of the anaotmy as EMAP. Inspired by the findings of Theiler (1989) and Kaufman (1992), EMA uses embryological mouse models to provide a detailed map of mouse development. The EMAP has a developed collection of three dimensional computer models of mouse embryos at the consecutive stages of growth generation with anatomical domians joined by an ontology of anatomical names. The main deliverable of the EMA resource is to provide a comprehensive visualisation of the postimplantation of mouse development and to induct an investiagtion of the gene expression in the postimplantation mouse embryo.

The EMA ontology has several different branch deliverables, each provides an alternative aspect of the evolution of a mouse embryo. The branches which will be utilised for this research project are the timed stage specific structure, EMAP and the aggregated non stage specific e-mouse Atlas Project Abstract (EMAPA) which are resepectively discussed below [REFERENCE]. The EMA dataset's were chosen as the source of data for this research as it is a freely available, rich and substantial data source.

1.1 EMAP

The devised EMAP ontology was originally developed to deliver a structured and controlled vocabulary of stage-specific anatomical structures for the developing laboratory mouse. As the EMA research has progressed, the ontology has followed suit, and is continually under development. The current ontology is in scope for a forthcoming release.

1.1.1 Data

Based on the timed component stage-specific Theiler development of a mouse embryo, the EMAP dataset combined the stage identifier and the anatomical name based on the reaerched information for the respective development stage. The timed components are regarded as the main aspect of the ontology and the abstract, non-stage-specific terms came as a secondary protocol.

A hierarchical structutue for each of the Theiler stages has been developed and are presented in sepereate directed acyclic graphs. This data is availabe in an obo-formatted file; which I will discuss later in the document. [REFERENCE]

1.1.2 Purpose

The intended outcome of this version of the EMAP was to provide information about the shape, gross anatomy and detailed histiological structure of the mouse.

1.2 EMAPA

The EMAPA structure is a refined and developed non-stage specific representation of the mouse anatomy ontology. This enahnced version of the ontology is now considered as the primary EMA anatomy ontology and will form the basis of the dataset for this research. As with the EMAP structure the EMAPA is available in

1.2.1 Data

1.2.2 Purpose

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.