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

# Contents

# Chapter 1
# Introduction

The era of Big Data is upon us bringing with it a range of new challenges ”Without big data, you are blind and deaf and in the middle of a freeway.” (Moore, 2012) The importance which accompany these challenges encouraged the formulation of new approaches for cleaning, processing and using these enormous amounts of data. (Section 2.1)

Data is being collated and stored every second of every day and the value of doing so has never been greater. Billion dollar companies such as Google and Amazon dominate the market in data collection and pride themselves in knowing everything about everything. Former CEO of Google, Eric Schmidt famously said in 2010 ”We know where you are. We know where you’ve been. We can more or less know what you’re thinking...” (Schmidt, 2010) Thus the power of data collection has led to the development of a range of technologies designed to meet the needs of Big Data.

The purpose of the following research, by way of investigation, is to deliver an insightful examination of a subset of new technologies which deliver high performance querying of large datasets. The ultimate aim of the research is to gain an understanding of these technologies and achieve a level of mastery which permits a thorough scrutiny of their application to Big Data.

There are a number of different indexing solutions available however for the means of this project to encapsulate a comprehensive examination a focus will be on leading NoSQL solutions, modern search and analytics engines and for comparative reasons a conventional relational database management system. The following technologies to be used for the project:

1. MongoDB - Document-Oriented Database
2. Neo4j - Graph Database
3. Apache Cassandra - Distributed Database
4. MySQL - Relational Database Management System

## 1.1 Objectives

The key objectives and main intended outcomes for the project are:

- Investigate the strengths and weaknesses of the functionality each technology provides.
- Compare and contrast the analytical capabilities of each technology by way querying prototype models.
- Conduct a comparative analysis to investigate the scalability of each technology.

## 1.2 Project Intentions

Choosing an indexing solution for querying Big Data sets can be difficult as there are so many options with limits on each respective functionalities. Taking the A & O into consideration the project intends to identify the best overall performing indexing solution for this dataset http://www.dataversity.net/nosql-or-rdbms-choosing-the-right-database-for-you/

# Chapter 2

# Background

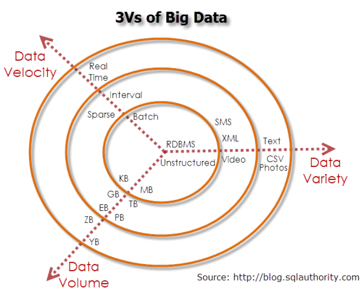
To deliver an all round level of comprehension the following section discusses generic terms which are used throughout the project such as Big Data (Section 2.1) and Extract Transform Load (Section 2.2)

## 2.1 Big Data

Big Data is a broad evolving term bound to a complex and powerful application of analytical insight which over recent years has had a variety of definitions. In simplistic terms Big Data can be described as extremely large datasets that may be studied computationally to reveal patterns, trends, and associations for ongoing discovery and analysis.

### 2.1.1 3vs Model

In 2001, Gartner analyst Doug Laney delivered the original 3vs model which categorises big data in to three dimensions; Volume, Variety and Velocity. The characteristics of each property are defined as **Volume** - The size of the generated data is required to assess whether the dataset is in fact 'big' enough to be categorised as Big Data.




**Variety** - The type of content, and an essential fact that data analysts must know. This helps people who are associated with and analyze the data to effectively use the data to their advantage and thus uphold its importance.**Velocity** - In this context, the speed at which the data is generated and processed to meet the demands and the challenges that lie in the path of growth and development. **Variability** - The inconsistency the data can show at times?-which can hamper the process of handling and managing the data effectively. **Veracity** - The quality of captured data, which can vary greatly. Accurate analysis depends on the veracity of source data. Laney's 2001 publication *3D data management: Controlling data volume, variety and velocity* is still widely recognised today as the expansion of all three properties encapsulate the challenges currently faced of big data management.

## 2.2 Extract Transform Load

A basic definition of the Extract Transform Load (ETL) process is pulling data out of one database, refactoring the composition of the data and putting the data into another database.

### 2.2.1 Process

ETL is a three step procedure which combines database functions into one tool.

#### 2.2.1.1 Extract

is the first step in the ETL procedure in which data is read from a database. The Extract step covers the data extraction from the source system and makes it accessible for further processing. The main objective of the extract step is to retrieve all the required data from the source system with as little resources as possible. The extract step should be designed in a way that it does not negatively affect the source system in terms or performance, response time or any kind of locking.

#### 2.2.1.2 Transform

where the extracted data is manipulated from its previous state and converted into another database format. The transform step applies a set of rules to transform the data from the source to the target. This includes converting any measured data to the same dimension (i.e. conformed dimension) using the same units so that they can later be joined. The transformation step also requires joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules.

#### 2.2.1.3 Load

completes the three step procedure and is where data is written into the target database. During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible. The target of the Load process is often a database. In order to make the load process efficient, it is helpful to disable any constraints and indexes before the load and enable them back only after the load completes. The referential integrity needs to be maintained by ETL tool to ensure consistency.

### 2.2.2 Tool Implementation

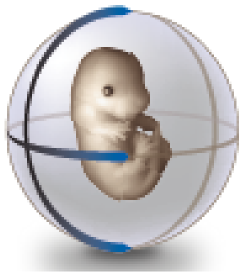
https://jena.apache.org/ There are many ready-to-use ETL tools on the market. The main benefit of using off-the-shelf ETL tools is the fact that they are optimized for the ETL process by providing connectors to common data sources like databases, flat files, mainframe systems, xml, etc. They provide a means to implement data transformations easily and consistently across various data sources. This includes filtering, reformatting, sorting, joining, merging, aggregation and other operations ready to use. The tools also support transformation scheduling, version control, monitoring and unified metadata management.

# Chapter 3

# Dataset Background

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 anatomy 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 domains joined by an ontology of anatomical names. The main deliverable of the EMA resource is to provide a comprehensive visualisation of the post-implantation of mouse development and to induct an investigation of the gene expression in the post-implantation mouse embryo.



## 3.1 Discussion

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 respectively 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.

### 3.1.1 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. 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 researched 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 structure for each of the Theiler stages has been developed and are presented in separate directed acyclic graphs. This data is available in an obo-formatted file; which I will discuss later in the document. [REFERENCE]

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

### 3.1.2 EMAPA

The EMAPA structure is a refined and developed non-stage specific representation of the mouse anatomy ontology. This enhanced 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

## 3.2 Data Sources

### 3.2.1 EMA Database

### 3.2.2 OBO

### 3.2.3 OWL

# Chapter 4

# Technical Literature Discussion