MS4S21 CW-2

Mark Baber - 17076749

Contents

[Section A - 20 Marks 2](#_Toc71291263)

[1.1 – MongoDB (300) 2](#_Toc71291264)

[1.2 – Comparisons (300) 3](#_Toc71291265)

[2 – Data Pipelines (600) 4](#_Toc71291266)

[Section B - 20 Marks 5](#_Toc71291267)

[2.1 – Introduction to Knowledge Graphs 5](#_Toc71291268)

[2.2 – RDF 5](#_Toc71291269)

[2.3 – The use of Knowledge Graphs 5](#_Toc71291270)

[2.4 – Overview of Knowledge Graphs 5](#_Toc71291271)

[References 6](#_Toc71291272)

# Section A - 20 Marks

You are required to write a brief report which should:

1. Provide an introduction to the database called MongoDB in the context of Big Data and Technologies, and
2. Draw comparisons between MongoDB and DynamoDB and their usability in the context of Big Data and technologies.

## 1.1 – MongoDB (300)

MongoDB describes itself as 'A general purpose, document-based, distributed database built for modern application developers and for the cloud era' (MongoDB, 2021) and is a NoSQL database. This has been described as document based and is mostly used to deal with unstructured data, which in comparison to a ‘normal’ database does not require as much time developing.

With NoSQL we store data in documents (like json objects), which allows for the use of Sharding which is where the data is split along multiple computers which is great for replication and allows for scalability. A benefit of NoSQL is the fact the user does not have to structure the data and spend a long time working out the structures and primary/secondary keys, it can all be added as it develops.

Whilst a MongoDB database does not use primary and secondary keys, they do have an ObjectId which is a hexadecimal unique value, which is made up of 4-bytes of timestamp value, 5-byte random value and 3-byte incrementing counter. These would be implemented as certain values within a document-based database would not require a primary key specified, below is an example:

|  |
| --- |
| {  "name": "notebook",  "qty": 50,  "rating": [ { "score": 8 }, { "score": 9 } ],  "size": { "height": 11, "width": 8.5, "unit": "in" },  "status": "A",  "tags": [ "college-ruled", "perforated"]  } |

(Figure 1 – Example of NoSQL JSON document (MongoDB, 2021))

There are 3 different ways to store your data with a NoSQL database, Key value stores which is stores as a key and data as 1 value, Column-oriented stores which is a flat structure with keys within the columns and Document oriented stores which is an organised collection of information on one document (CoreViewSystem, 2021). For an example, if we were to store my student data as a document store, it would all be on one ‘form’ with all the information I have given the uni.

## 1.2 – Comparisons (300)

There are several NoSQL databases to date, ranging from CassandraDB, CoughDB, MongoDB, DynamoDB and many more. The two which will be discussed here are MongoDB and DynamoDB.

MongoDB is a free and open source which allows for anyone to download, see, and edit the code and can deploy their database anywhere. Whereas DynamoDB is a proprietary piece of software you can only use on Amazon Web Services (AWS). Whilst the two companies have different ideas of how to develop a NoSQL database, they also have a few common concepts:

|  |  |
| --- | --- |
| **DynamoDB** | **MongoDB** |
| Table | Collection |
| Item | Document |
| Attribute | Field |
| Secondary Index | Secondary Index |

(Figure 2 – Table showing common concepts)

Both databases also use different data models, with MongoDB storing their data as JSON-like format, (which we saw in Figure 1) which allows for the storage of a random of datatypes, from strings, numbers, dates, timestamps and more, with a maximum size of 16 MB in size. DynamoDB uses a key-value store which adds support for JSON to provide a document-like data structure, which can store one numeric type and does not support dates, with a maximum size of 400KB.

There are also some other differences with MongoDB having a rich ad-hoc query language + IDE built into their cloud platform, whereas DynamoDB is all through CLI or programming languages. With both databases having great performance, with the ability to scale horizontally and vertically with high number of read/write operations (ZenOfAI, 2021).

## 2 – Data Pipelines (600)

1. What are data pipelines and how can they be utilised in cloud services such as AWS. You are required to pick up tasks or experiments from Coursework 1 to theoretically demonstrate the usability AWS data pipeline.

A data pipeline is a process of taking data from task 1, processing the data and outputting this data as the input for the next processing step (say task 2). It follows a flow through each of the steps until the processes are complete, with some steps being able to run at parallel. With common steps within a data pipeline being data transformation, enrichment, filtering, grouping with the ability to run other algorithms on the data (Hazelcast, 2021).

With the cloud becoming more popular, the use of data pipelines within the cloud is growing rapidly. If we were to explore a pipeline for data analytics, it should look something like this:



(Figure 3 – Data Analytics Pipeline (Vergadia, 2021)

With each step being just as important as the last, this could be applied to any service from AWS, Microsoft Azure (Azure) and Google Cloud Platform (GCP). Let us apply this pipeline to AWS and see how each step can be used on that platform.

Starting with capture, this could be an application which is reading user input from a virtual machine (VM) which is up and running on an Amazon EC2 service. This could be an app which is reading tweets for example, before being moved to the next step process. The process part is the reading of the data which was previously captures.

Next would be the store, and for storing data AWS has an Amazon S3 service which is their storage buckets also known as Objective storage which they claim is used to store data for millions of applications. From the smaller storage, this data can be stored/archived in a data warehouse which is usually used for long term storage and analysing, Amazons Redshift is used for this.

Now this data has been collected and stored in a data warehouse, the next step would be to analyse the data which could be used for machine learning algorithms to predict trends. The Amazon service which is mostly used for this is the Amazon ERM which is used to run and scale Apache Spark, Hive and other big data frameworks.

This can be seen as a positive when using a big service like AWS as everything can easily talk to each other, and data pipelines can become a real benefit to businesses and data scientists when working with such customisable tools.

# Section B - 20 Marks

## 2.1 – Introduction to Knowledge Graphs (300)

Introduction to knowledge graphs and their usability in general

## 2.2 – RDF (300)

Overview of RDF and ontologies

## 2.3 – The use of Knowledge Graphs (300)

The usability of knowledge graphs and a use-case or example of their application in either academia or industry.

## 2.4 – Overview of Knowledge Graphs (300)

An overview of a knowledge graph platform used either commercially or in academia with any type of license (open-source or paid).

# References

MongoDB, 2021 *Structure your Data for MongoDB* Available at: <https://docs.mongodb.com/guides/server/introduction/> (Accessed 03/05/2021)

MongoDB, 2021 *The database for modern applications* Available at: <https://www.mongodb.com/> (Accessed 30/04/2021)

CoreViewSystem, 2021 *NoSql Databases – Document Stores* Available at: <https://coreviewsystems.com/nosql-databases-document-stores/> (Accessed 04/05/2021)

ZenOfAI, 2021 *Scaling DynamoDB for Big Data using Parallel Scan* Available at: <https://medium.com/zenofai/scaling-dynamodb-for-big-data-using-parallel-scan-1b3baa3df0d8> (Accessed 07/05/2021)

Hazelcast, 2021 *What is a Data Pipeline* Available at: <https://hazelcast.com/glossary/data-pipeline/> (Accessed 07/05/2021)

Vergadia, P., 2020 *How to Build a Scalable Data Analytics Pipeline* Available at: <https://www.freecodecamp.org/news/scalable-data-analytics-pipeline/> (Accessed 07/05/2021)