Big Data 2019  
Portfolio assignment - part two

Mette Strand Hornnes

Innholdsfortegnelse

[1 Choosing the right storage technology 3](#_Toc23788482)

[1.1 Apache Cassandra 4](#_Toc23788483)

[1.1.1 Om Cassandra: 4](#_Toc23788484)

[1.1.2 Datasett 5](#_Toc23788485)

[1.1.3 Spørringer på dette datasettet 6](#_Toc23788486)

[1.1.4 Beskrivelse av datasettet og hvorfor dette passer 6](#_Toc23788487)

[1.1.5 Ved annen bruk av dataene 6](#_Toc23788488)

[1.2 Neo4j 6](#_Toc23788489)

[1.2.1 Om Neo4j 6](#_Toc23788490)

[1.2.2 Datasett 6](#_Toc23788491)

[1.2.3 Spørringer på dette datasettet 7](#_Toc23788492)

[1.2.4 Beskrivelse av datasettet og hvorfor dette passer 7](#_Toc23788493)

[1.2.5 Ved annen bruk av dataene 7](#_Toc23788494)

[1.3 Riak 7](#_Toc23788495)

[1.3.1 Om Neo4j 7](#_Toc23788496)

[1.3.2 Datasett 7](#_Toc23788497)

[1.3.3 Spørringer på dette datasettet 8](#_Toc23788498)

[1.3.4 Beskrivelse av datasettet og hvorfor dette passer 8](#_Toc23788499)

[1.3.5 Ved annen bruk av dataene 8](#_Toc23788500)

[1.4 Mongo DB 8](#_Toc23788501)

[1.4.1 Om Neo4j 8](#_Toc23788502)

[1.4.2 Datasett 8](#_Toc23788503)

[1.4.3 Spørringer på dette datasettet 8](#_Toc23788504)

[1.4.4 Beskrivelse av datasettet og hvorfor dette passer 9](#_Toc23788505)

[1.4.5 Ved annen bruk av dataene 9](#_Toc23788506)

[2 Practical use of one of the storage technologies 9](#_Toc23788507)

[3 Kildeliste: 11](#_Toc23788508)

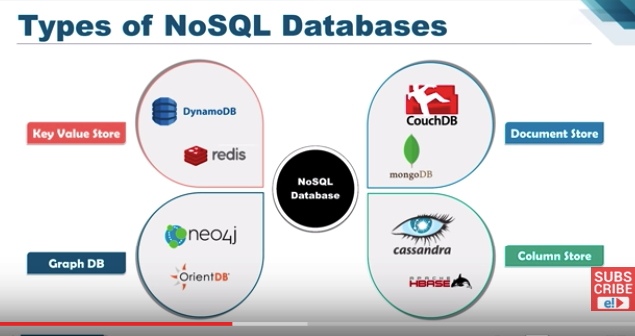
Introduction

In this second, and last, part of your portfolio, you will demonstrate that you

have understood the concept of polygot persistence and are capable of choosing

the right storage technology for datasets you select yourself.

# 1 Choosing the right storage technology



A central tenet in the NoSQL movement is that there is no one-size-fits-all data model and storage technology. In the second part of this course, you have learned about four major data models and database engines for those data models.

In this assignment you will, for each of those four data models, find a dataset in the Kaggle Dataset directory for which that data model is the most suitable.

Keep in mind that choosing the right storage technology is not only about understanding the data itself, but also about how that data is created and used.

**Therefore, for each of the four data models you will:**

## 1.1 Colum and Row Store Database - Apache Cassandra

SKRIVE OM COLUMN AND ROW DATABASE‼‼‼

### 1.1.1 Om Cassandra:

Cassandra er et distribuert databasesystem som er designet for å håndtere store volum med strukturert data. Dataene som brukes kan være fordelt utover ulike maskiner. Dette gir høy skalerbarhet, gjør det mulig å håndtere store mengder data raskt (mye raskere enn for eksempel MySQL), samt gir høy tilgjengelighet. Cassandra er dermed laget med tanke på at feil på en maskin (software eller hardware) kan oppstå og hindrer derfor «Single point of failures».

<https://www.guru99.com/cassandra-tutorial.html>

**Arkiktetur**

Arkitekturen til Cassandra består kun av noder (maskiner/servere) eller clustere (samling av noder) og ingen master. Alle nodene spiller samme rolle og er selvstendige samtidig som de er koblet til de andre nodene med et peer-to-peer-system. Hver node kan skrives til eller leses fra uavhengig av hvor dataene ligger, noe som også hindrer Single point of failues.

Med en såkalt gossip-protokoll deles informasjonen om data gjennom clusteret hvert sekund.   
Det lages også kopier (replicas) av dataene som legges på andre noder, noe som igjen er med på å hindre feil og tap av data.

<https://intellipaat.com/blog/tutorial/cassandra-tutorial/brief-architecture-of-cassandra/>

Skriving av data skjer på følgende måte:

1. Klienten skriver til en av nodene
2. Fra denne skrives replicas til andre noder (antall bestemmes av replication factor som settes ved oppsett)
3. Nodene bekrefter lagring tilbake til den første noden, som igjen bekrefter tilbake til klienten om at lagring er fullført (Acknowledge-melding).
4. Når klienten mottar denne ACK-meldingen skrives info om datalagringen til en commit-logg.

### 1.1.2 Datasett

Find a dataset in Kaggle’s dataset directory (<https://www.kaggle.com/datasets>) and argue for why this data set is best stored and used using this data model.

### 1.1.3 Spørringer på dette datasettet

Describe queries which might be useful to run on that dataset. You do not need to implement those queries to an appropriate query language, but you should be able to describe what the parameters of the queries will be, how the data model can be leveraged to answer those queries, if it had been implemented.

Cassandra benytter Cassandra Query Language (CQL).

Hente data

SELECT \* from tabell WHERE . . .

Where kan kun brukes på kolonner som er primary key ELLER er indexert

### 1.1.4 Beskrivelse av datasettet og hvorfor dette passer

If the process that created the data is described in the dataset description, argue for why this process and throughput makes this data model suitable. If the origin process is not already described, imagine a process that might be a reasonable source of the data, and argue for why this imagined process and expected throughput makes this data model suitable.

### 1.1.5 Ved annen bruk av dataene

There are some situations were long-term storage, day-to-day use and historical analysis of the data presents different requirements for the storage technology, thereby necessitating different choices for each stage in thelife-cycle of the dataset. What are your thoughts concerning this for your datasets? Will a different storage technology be more appropriate for a different use of the same data?

## Neo4j

### 1.2.1 Om Neo4j

. . . .

### 1.2.2 Datasett

Find a dataset in Kaggle’s dataset directory (<https://www.kaggle.com/datasets>) and argue for why this data set is best stored and used using this data model.

### 1.2.3 Spørringer på dette datasettet

Describe queries which might be useful to run on that dataset. You do not need to implement those queries to an appropriate query language, but you should be able to describe what the parameters of the queries will be, how the data model can be leveraged to answer those queries, if it had been implemented.

### 1.2.4 Beskrivelse av datasettet og hvorfor dette passer

If the process that created the data is described in the dataset description, argue for why this process and throughput makes this data model suitable. If the origin process is not already described, imagine a process that might be a reasonable source of the data, and argue for why this imagined process and expected throughput makes this data model suitable.

### 1.2.5 Ved annen bruk av dataene

There are some situations were long-term storage, day-to-day use and historical analysis of the data presents different requirements for the storage technology, thereby necessitating different choices for each stage in thelife-cycle of the dataset. What are your thoughts concerning this for your datasets? Will a different storage technology be more appropriate for a different use of the same data?

## Riak

### 1.3.1 Om Neo4j

. . . .

### 1.3.2 Datasett

Find a dataset in Kaggle’s dataset directory (<https://www.kaggle.com/datasets>) and argue for why this data set is best stored and used using this data model.

### 1.3.3 Spørringer på dette datasettet

Describe queries which might be useful to run on that dataset. You do not need to implement those queries to an appropriate query language, but you should be able to describe what the parameters of the queries will be, how the data model can be leveraged to answer those queries, if it had been implemented.

### 1.3.4 Beskrivelse av datasettet og hvorfor dette passer

If the process that created the data is described in the dataset description, argue for why this process and throughput makes this data model suitable. If the origin process is not already described, imagine a process that might be a reasonable source of the data, and argue for why this imagined process and expected throughput makes this data model suitable.

### 1.3.5 Ved annen bruk av dataene

There are some situations were long-term storage, day-to-day use and historical analysis of the data presents different requirements for the storage technology, thereby necessitating different choices for each stage in thelife-cycle of the dataset. What are your thoughts concerning this for your datasets? Will a different storage technology be more appropriate for a different use of the same data?

## Mongo DB

### 1.4.1 Om Neo4j

. . . .

### 1.4.2 Datasett

Find a dataset in Kaggle’s dataset directory (<https://www.kaggle.com/datasets>) and argue for why this data set is best stored and used using this data model.

### 1.4.3 Spørringer på dette datasettet

Describe queries which might be useful to run on that dataset. You do not need to implement those queries to an appropriate query language, but you should be able to describe what the parameters of the queries will be, how the data model can be leveraged to answer those queries, if it had been implemented.

### 1.4.4 Beskrivelse av datasettet og hvorfor dette passer

If the process that created the data is described in the dataset description, argue for why this process and throughput makes this data model suitable. If the origin process is not already described, imagine a process that might be a reasonable source of the data, and argue for why this imagined process and expected throughput makes this data model suitable.

### 1.4.5 Ved annen bruk av dataene

There are some situations were long-term storage, day-to-day use and historical analysis of the data presents different requirements for the storage technology, thereby necessitating different choices for each stage in thelife-cycle of the dataset. What are your thoughts concerning this for your datasets? Will a different storage technology be more appropriate for a different use of the same data?

# 2 Practical use of one of the storage technologies

Here, you will choose one of the datasets from section 2.1 and use the associated database engine you have learned about in the course, namely, Cassandra, neo4j, Riak or Mongo DB to implement your solution. If you described an extensive solution for that dataset, you don’t need to implement everything but your implementation should support the following:

* Importing the dataset into the database
* Running one of the queries you described
* Adding, removing or modifying the data in the database using procedures with names and behavior which are a natural part of a use-case for the dataset you selected.

Keep in mind that demonstrating an understanding beyond the barest rudiments will be beneficial for your grade. Also, while the examples in the lectures will be using Java, you are free to use either Java, Scala, or Python for your work.

Delivery

The delivery should be a zip file containing the following parts.

3.1 Source code

You should provide the source code you have used in this project as part of the

zip file using a folder and naming scheme which clearly identifies the purpose of

each file.

3.2 Report

You should provide the report for this project as a single PDF file. The contents

of the report are expected to be fairly extensive.

# 3 Kildeliste:

<https://www.guru99.com/cassandra-tutorial.html>

<https://intellipaat.com/blog/tutorial/cassandra-tutorial/brief-architecture-of-cassandra/>