AQA Computer Science A-Level

4.11.1 Big Data

Concise Notes

Specification:

4.11.1 Big Data:

Know that 'Big Data' is a catch-all term for data that won't fit the usual containers. Big Data can be described in terms of:

- Volume – too big to fit into a single server
- Velocity – streaming data, milliseconds to seconds to respond
- Variety – data in many forms such as structured, unstructured, text, multimedia

Know that when data sizes are so big as not to fit on to a single server:

- The processing must be distributed across more than one machine
- Functional programming is a solution, because it makes it easier to write correct and efficient distributed code

Know what features of functional programming make it easier to write:

- Correct code
- Code that can be distributed to run across more than one server

Be familiar with the:

- Fact-based model for representing data
- Graph schema for capturing the structure of the dataset
- Nodes, edges and properties in graph schema

Big Data

- A catch-all term for data that doesn't fit the usual containers.
- The three defining features of big data are "the three Vs":

Volume

- There is too much data for it all to fit on a conventional hard drive or server
- Data has to be stored over multiple servers

Velocity

- Data on the servers is created and modified rapidly
- The servers must respond within a matter of milliseconds

Variety

- The data held on the servers consists of many different types of data.

- The most challenging attribute of big data is its lack of structure
- This makes it difficult to analyse the data
- Conventional databases are not suited to storing big data
- Machine learning techniques must be used to discern patterns in the data
- This allows useful information to be extracted from big data
- The processing associated with big data must be split across multiple machines
- Conventional programming paradigms are not well suited to working across multiple machines

Functional programming

- A solution to the problem of processing data over multiple machines
- Functional programming makes it easier to write correct, efficient, distributed code
- Functional programs are stateless so have no side effects
- Functional programs make use of immutable data structures
- The functional programming paradigm supports higher-order functions

The fact-based model for representing data

- One way of representing big data
- Each individual piece of information is stored as a fact
- Facts are immutable and can't be overwritten
- Stored with each fact is a timestamp
- Using the fact-based model reduces the risk of losing data due to human error
- The model does away with an index for the data
- New data is simply appended to the dataset as it is created

Representing big data using graph schema

- Uses graphs consisting of nodes and edges to graphically represent the structure of a dataset
- Nodes represent entities and can contain properties
- Edges represent relationships and are labelled with a brief description