8.4 Distributed Storage and Processes

The increasing volume and variety of data collected by healthcare enterprises is a challenge to traditional relational database management systems. This increase in data is due both to an increase in computerization of health records, and to an increase in the capture of data from other sources, such as medical instruments (e.g. biometric data from home monitoring equipment), imaging data, gene sequencing, administrative information, environmental data and medical knowledge. The proliferation of large volumes of both structured and unstructured data sets has led to the popularity of the term 'Big data' within the healthcare context. Big data refers to any collection of data sets that is so large and complex that it becomes difficult to process them using traditional data processing applications.

Accommodating and analyzing this expanding volume of diverse data (i.e. 'Big Data') requires distributed database technologies. A distributed database is a federation of loosely coupled data stores with separate processing units, which are controlled by a common distributed database management system. It may be stored in multiple computers located in the same physical location, or dispersed over a network of interconnected computers. Distributed databases may be categorized as either:

- *Homogeneous* – A distributed database with identical software and hardware running on all database instances.
- *Heterogeneous* – A distributed database supported by different hardware, operating system, database management systems and even data models (e.g. using the VHR strategy described in section 8.3 Virtual Health Record).

In both cases, however, the database appears through a single interface as if it were a single database.

Distributed databases are used for Big Data analytics for a number of reasons, including:

- Transparency of querying over heterogeneous data stores (as described in section 8.3 Virtual Health Record)
- Increase in the reliability, availability and protection of data due to data replication
- Local autonomy of data (e.g. each department or institution controls their own data)
- Distributed query processing can improve performance, as the load can be balanced among the servers

A number of tools are available for the distributed storage and processing of big data, including Apache Hadoop. Apache Hadoop is an open-source software framework, which splits files into large blocks and distributes these blocks amongst the nodes in the cluster. To process the data, Hadoop sends code to the nodes that have the required data, and the nodes then process the data in parallel. Hadoop supports horizontal scaling – that is, as data grows additional servers can be added to distribute the load across them.

Many distributed database solutions use NoSQL (Not Only SQL) systems. NoSQL systems are increasingly being used for big data, as they provide a mechanism for storage and retrieval of data in a variety of structures, including relational, key-value, graph or documents. The Oxford University, in collaboration with Kaiser Permanente (case study 13.1.2 Kaiser Permanente) are using a NoSQL database (RDFox) to investigate how to perform complex queries efficiently across extremely large numbers of patient records. RDFox is a highly scalable and performant NoSQL database that is readily distributed across parallel processing units.