Data Science

Kevin O'Brien

May 5, 2013

Contents

1 Coursera's Introduction to Data Science

- **Introduction**

0.1 Examples, data science articulated, history and context, technology landscape

- **Part 1 Data Manipulation, at Scale**

1.1 Databases and the relational algebra

1.2 Parallel databases, parallel query processing, in-database analytics

1.3 MapReduce, Hadoop, relationship to databases, algorithms, extensions, languages

1.4 Key-value stores and NoSQL; tradeoffs of SQL and NoSQL

1.5 Entity resolution, record linkage, data cleaning

1.1 Data Semantics

Data semantics is the study of the meaning and use of specific pieces of data in computer programming and other areas that employ data. When studying a language, semantics refers to what individual words mean and what they mean when put together to form phrases or sentences. In data semantics, the focus is on how a data object represents a concept or object in the real word.

Data semantics is highly subjective. If a person who has never worked with a computer database tries to pull information from it, the words and phrases used to access the database would make no sense. Semantic meaning occurs only when a group agrees on specific definitions for certain data types or words. For others to pick up on these semantic meanings, they cannot change. If the word "dog" referred to a furry, four-legged animal one day and a two-legged bird the next, it would lose its meaning and no one would know what another person meant when she said "dog."

1.2 Data Mapping

Data mapping is the process by which two distinct data models are created and a link between these models is defined. Data models can include either metadata, an atomic unit of data with a precise meaning in regards to semantics, and telecommunications. The system uses the atomic unit system to measure the properties of electricity which contain the information. Data mapping is most readily used in software engineering to describe the best way to access or represent some form of information. It works as an abstract model to determine relationships within a certain domain of interest. This is the fundamental first step in establishing data integration of a particular domain.

The main uses for data mapping include a wide variety of platforms. Data transformation is used to mediate the relationship between an initial data source and the destination in which that data is used. It is useful in identifying parts of data lineage analysis, the way in which data flows from one sector of information to another. Data mapping is also integral in discovering hidden information and sensitive data such as social security numbers when hidden within a different identification format. This is known as data masking.

Certain procedures are put in place when data mapping is conducted. This allows a user to create or transform the information into a form in which the best results can be culled. Commonly, this takes the form of some graphical mapping tool that is able to automatically generate results and execute a transformation of the data. Essentially, a user is able to literally draw a line from one field to another, identifying the correct connection. This is known as manual data mapping.

In regards to the basic mapping technique of a data element, a number of specific formula considerations need to be addressed. The data element itself needs to identified and named, a clear definition of the data needs to be determined and representation of the values are enumerated. In some terms, the identifiers are represented in the form of a database. Standard structures are built with basic units of information, such as names, addresses or ages.

1.2.1 Example

For example, when a company merges with another company, they need to merge data for both sets of customers. Data mapping can be used to track one set of information and cross-reference it with another set of data. This allows both companies to merge the data into one final database.

One of the newest techniques in data mapping involves using statistics simultaneously with two values of divergent data sources. This allows more complex mapping operations between the two data sets. It can be highly valued when it comes to discovering more specialized informational aspects such as substrings.

1.3 Entity Resolution

Entity Resolution is the problem of identifying and linking/grouping different manifestations of of the same real world object. Examples of manifestations and objects:

- Different ways of addressing (names, email addresses, FaceBook accounts) the same person in text.
- Web pages with differing descriptions of the same business.
- Different photos of the same object.

1.4 Sparsity and Density

Sparsity and density are terms used to describe the percentage of cells in a database table that are not populated and populated, respectively. The sum of the sparsity and density should equal 100

A table that is 10% dense has 10% of its cells populated with non-zero values. It is therefore 90% sparse meaning that 90% of its cells are either not filled with data or are zeros.

Because a processor adds up the zeros, sparcity can negatively impact processing time. In a multidimensional database sparsity can be avoided by linking cubes. Instead of creating a sparse cube for data that is not fully available, a separate but linked cube will ensure the data in the cubes remains consistent without slowing down processing.

1.5 Data Classification

Data classification is the categorization of data for its most effective and efficient use. In a basic approach to storing computer data, data can be classified according to its critical value or how often it needs to be accessed, with the most critical or often-used data stored on the fastest media while other data can be stored on slower (and less expensive) media. This kind of classification tends to optimize the use of data storage for multiple purposes - technical, administrative, legal, and economic.

Data can be classified according to any criteria, not only relative importance or frequency of use. For example, data can be broken down according to its topical content, file type, operating platform, average file size in megabytes or gigabytes, when it was created, when it was last accessed or modified, which person or department last accessed or modified it, and which personnel or departments use it the most. A well-planned data classification system makes essential data easy to find. This can be of particular importance in risk management, legal discovery, and compliance with government regulations.

Computer programs exist that can help with data classification, but in the end it is a subjective business and is often best done as a collaborative task that considers business, technical, and other points-of-view.

2 Databases

Database management systems (DBMSs) and, in particular, relational DBMSs (RDBMSs) are designed to do all of these things well. Their strengths are

1. To provide fast access to selected parts of large databases.
2. Powerful ways to summarize and cross-tabulate columns in databases.
3. Store data in more organized ways than the rectangular grid model of spreadsheets and R data frames.
4. Concurrent access from multiple clients running on multiple hosts while enforcing security constraints on access to the data.
5. Ability to act as a server to a wide range of clients.

2.1 Graph Database

A graph database is a database that uses graph structures with nodes, edges, and properties to represent and store data. By definition, a graph database is any storage system that provides index-free adjacency. This means that every element contains a direct pointer to its adjacent element and no index lookups are necessary. General graph databases that can store any graph are distinct from specialized graph databases such as triplestores and network databases.

Compared with relational databases, graph databases are often faster for associative data sets, and map more directly to the structure of object-oriented applications. They can scale more naturally to large data sets as they do not typically require expensive join operations. As they depend less on a rigid schema, they are more suitable to manage ad-hoc and changing data with evolving schemas. Conversely, relational databases are typically faster at performing the same operation on large numbers of data elements.

Graph databases are a powerful tool for graph-like queries, for example computing the shortest path between two nodes in the graph. Other graph-like queries can be performed over a graph database in a natural way (for example graph's diameter computations or community detection).

2.2 PostgreSQL

PostgreSQL (pronounced "post-gress-Q-L") is an open source relational database management system (DBMS) developed by a worldwide team of volunteers. PostgreSQL is not controlled by any corporation or other private entity and the source code is available free of charge.

3 Big Data

3.1 MapReduce

MapReduce is a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers. It was developed at Google for indexing Web pages and replaced their original indexing algorithms and heuristics in 2004.

The framework is divided into two parts:

- **Map** a function that parcels out work to different nodes in the distributed cluster.

- **Reduce** another function that collates the work and resolves the results into a single value.

The MapReduce framework is fault-tolerant because each node in the cluster is expected to report back periodically with completed work and status updates. If a node remains silent for longer than the expected interval, a master node makes note and re-assigns the work to other nodes.

MapReduce is a programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster. The MapReduce concept is fairly simple to understand for those who are familiar with clustered scale-out data processing solutions.
The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job. An example of MapReduce

Lets look at a simple example. Assume you have five files, and each file contains two columns (a key and a value in Hadoop terms) that represent a city and the corresponding temperature recorded in that city for the various measurement days. Of course weve made this example very simple so its easy to follow. You can imagine that a real application wont be quite so simple, as its likely to contain millions or even billions of rows, and they might not be neatly formatted rows at all; in fact, no matter how big or small the amount of data you need to analyze, the key principles were covering here remain the same. Either way, in this example, city is the key and temperature is the value.

```
Toronto, 20
Whitby, 25
New York, 22
Rome, 32
Toronto, 4
Rome, 33
New York, 18
```

Out of all the data we have collected, we want to find the maximum temperature for each city across all of the data files (note that each file might have the same city represented multiple times). Using the MapReduce framework, we can break this down into five map tasks, where each mapper works on one of the five files and the mapper task goes through the data and returns the maximum temperature for each city. For example, the results produced from one mapper task for the data above would look like this:

```
(Toronto, 20) (Whitby, 25) (New York, 22) (Rome, 33)
```

Lets assume the other four mapper tasks (working on the other four files not shown here) produced the following intermediate results:

```
(Toronto, 18) (Whitby, 27) (New York, 32) (Rome, 37)(Toronto, 32) (Whitby, 20)
```

All five of these output streams would be fed into the reduce tasks, which combine the input results and output a single value for each city, producing a final result set as follows:

```
(Toronto, 32) (Whitby, 27) (New York, 33) (Rome, 38)
```

As an analogy, you can think of map and reduce tasks as the way a census was conducted in Roman times, where the census bureau would dispatch its people to each city in the empire. Each census taker in each city would be tasked to count the number of people in that city and then return their results to the capital city. There, the results from each city would be reduced to a single count (sum of all cities) to determine the overall population of the empire. This mapping of people to cities, in parallel, and then combining the results (reducing) is much more efficient than sending a single person to count every person in the empire in a serial fashion.

(source: hadoop.apache.org)
Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

3.2 NoSQL

NoSQL database, also called Not Only SQL, is an approach to data management and database design that's useful for very large sets of distributed data.

NoSQL, which encompasses a wide range of technologies and architectures, seeks to solve the scalability and big data performance issues that relational databases werent designed to address. NoSQL is especially useful when an enterprise needs to access and analyze massive amounts of unstructured data or data that's stored remotely on multiple virtual servers in the cloud.

Contrary to misconceptions caused by its name, NoSQL does not prohibit structured query language (SQL). While it's true that some NoSQL systems are entirely non-relational, others simply avoid selected relational functionality such as fixed table schemas and join operations. For example, instead of using tables, a NoSQL database might organize data into objects, key/value pairs or tuples.

3.3 NoSQL implementations

Arguably, the most popular NoSQL database is Apache Cassandra. Cassandra, which was once Facebooks proprietary database, was released as open source in 2008. Other NoSQL implementations include SimpleDB, Google BigTable, Apache Hadoop, MapReduce, MemcacheDB, and Voldemort. Companies that use NoSQL include NetFlix, LinkedIn and Twitter.

NoSQL is often mentioned in conjunction with other big data tools such as massive parallel processing, columnar-based databases and Database-as-a-Service (DaaS).

3.4 Background of NoSQL

Relational databases were introduced into the 1970s to allow applications to store data through a standard data modeling and query language (Structured Query Language, or SQL). At the time, storage was expensive and data schemas were fairly simple and straightforward. Since the rise of the web, the volume of data stored about users, objects, products and events has exploded. Data is also accessed more frequently, and is processed more intensively for example, social networks create hundreds of millions of customized, real-time activity feeds for users based on their connections' activities.

Even rendering a single web page or answering a single API request may take tens or hundreds of database requests as applications process increasingly complex information. Interactivity, large user networks, and more complex applications are all driving this trend.

In response to this demand, computing infrastructure and deployment strategies have also changed dramatically. Low-cost, commodity cloud hardware has emerged to replace vertical scaling on highly complex and expensive single-server deployments. And engineers now use agile development methods, which aim for continuous deployment and short development cycles, to allow for quick response to user demand for features.

3.4.1 The Need for NoSQL

Relational databases were never designed to cope with the scale and agility challenges that face modern applications and aren't built to take advantage of cheap storage and processing power that's available today through the cloud. Relational database vendors have developed two main technical approaches to address these shortcomings:

3.4.2 Manual Sharding

Tables are broken up into smaller physical tables and spread across multiple servers. Because the database does not provide this ability natively, development teams take on the work of deploying multiple relational databases across a number of machines. Data is stored in each database instance autonomously. Application code is developed to distribute the data, distribute queries, and aggregate the results of data across all of the database instances. Additional code must be developed to handle resource failures, to perform joins across the different databases, for data rebalancing, replication, and other requirements. Furthermore, many benefits of the relational database, such as transactional integrity, are compromised or eliminated when employing manual sharding.

4 Data Scrubbing

Data scrubbing, sometimes called data cleansing, is the process of detecting and removing or correcting any information in a database that has some sort of error. This error can be because the data is wrong, incomplete, formatted incorrectly, or is a duplicate copy of another entry. Many data-intensive fields of business such as banking, insurance, retail, transportation, and telecommunications may use these sophisticated software applications to clean up a database's information.

Errors are in databases can be the result of human error in entering the data, the merging of two databases, a lack of company wide or industry wide data coding standards, or due to old systems that contain inaccurate or outdated data. Before computers had the capabilities to sort through and clean data, most data scrubbing was done by hand. Not only was this time consuming and expensive, but it oftentimes led to even more human error.

The need for data scrubbing is made clear when considering how easily errors can be made. For example, consider a database of names and addresses. One name is Bobby Johnson of Needham, MA. Another name is Bob Johnson of Needham, MA. This variation of names is most likely an error, and is referring to one person. However, a computer would normally deal with the information as though it were two different people. Specialized data scrubbing software is able to distinguish the discrepancy and fix it.

While these small errors may seem like a trivial problem, when merging corrupt or erroneous data into multiple databases, the problem may be multiplied by the millions. This so-called ”dirty data” has been a problem as long as there have been computers, but the problem is becoming more critical as businesses are becoming more complex and data warehouses are merging data from multiple sources. There is no point in having a comprehensive database if that database is filled with errors and disputed information.

Companies using specialized data scrubbing software can either develop it in-house or buy it from a variety of vendors. The software is not cheap and can range anywhere from a price of $20,000to$300,000. It oftentimes also requires some customization so that the software will work to the business' specific needs. The software goes through a process of using algorithms to standardize, correct, match, and consolidate data and is able to work with single or multiple sets of data.

Data scrubbing is sometimes skipped as part of a Data Warehouse implementation but it is one of the most critical steps to having a good, accurate end product. Because mistakes will always be made in data entry, the need for data scrubbing will always be present.

4.0.3 Distributed Caches

A number of products provide a caching tier for database systems. These systems can improve read performance substantially, but they do not improve write performance, and they add complexity to system deployments. If your application is dominated by reads then a distributed cache should probably be considered, but if your application is dominated by writes or if you have a relatively even mix of reads and writes, then a distributed cache may not improve the overall experience of your end users.

NoSQL databases have emerged in response to these challenges and in response to the new opportunities provided by low-cost commodity hardware and cloud-based deployment environments - and natively support the modern application deployment environment, reducing the need for developers to maintain separate caching layers or write and maintain sharding code.

NoSQL encompasses a wide variety of different database technologies but generally all NoSQL databases have a few features in common.

4.0.4 Dynamic Schemas

Relational databases require that schemas be defined before you can add data. For example, you might want to store data about your customers such as phone numbers, first and last name, address, city and state a SQL database needs to know this in advance.

This fits poorly with agile development approaches, because each time you complete new features, the schema of your database often needs to change. So if you decide, a few iterations into development, that you'd like to store customers' favorite items in addition to their addresses and phone numbers, you'll need to add that column to the database, and then migrate the entire database to the new schema.

If the database is large, this is a very slow process that involves significant downtime. If you are frequently changing the data your application stores because you are iterating rapidly this downtime may also be frequent. There's also no way, using a relational database, to effectively address data that's completely unstructured or unknown in advance.

NoSQL databases are built to allow the insertion of data without a predefined schema. That makes it easy to make significant application changes in real-time, without worrying about service interruptions which means development is faster, code integration is more reliable, and less database administrator time is needed.

4.0.5 Auto-Sharding, Replication & Integrated Caching

Because of the way they are structured, relational databases usually scale vertically a single server has to host the entire database to ensure reliability and continuous availability of data. This gets expensive quickly, places limits on scale, and creates a relatively small number of failure points for database infrastructure.

The solution is to scale horizontally, by adding servers instead of concentrating more capacity in a single server. Cloud computing makes this significantly easier, with providers such as Amazon Web Services providing virtually unlimited capacity on demand, and taking care of all the necessary database administration tasks. Developers no longer need to construct complex, expensive platforms to support their applications, and can concentrate on writing application code. In addition, a group of commodity servers can provide the same processing and storage capabilities as a single high-end server for a fraction of the price.

"Sharding" a database across many server instances can be achieved with SQL databases, but usually is accomplished through SANs and other complex arrangements for making hardware act as a single server. NoSQL databases, on the other hand, usually support auto-sharding, meaning that they natively and automatically spread data across an arbitrary number of servers, without requiring the application to even be aware of the composition of the server pool. Data and query load are automatically balanced across servers, and when a server goes down, it can be quickly and transparently replaced with no application disruption.

Most NoSQL databases also support automatic replication, meaning that you get high availability and disaster recovery without involving separate applications to manage these tasks. The storage environment is essentially virtualized from the developer's perspective.

Lastly, many NoSQL database technologies have excellent integrated caching capabilities, keeping frequently-used data in system memory as much as possible. This removes the need for a separate caching layer that must be maintained.

5 Data Analytics

Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organization to make better business decisions and in the sciences to verify or disprove existing models or theories. Data analytics is distinguished from data mining by the scope, purpose and focus of the analysis. Data miners sort through huge data sets using sophisticated software to identify undiscovered patterns and establish hidden relationships. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher.

The science is generally divided into exploratory data analysis (EDA), where new features in the data are discovered, and confirmatory data analysis (CDA), where existing hypotheses are proven true or false. Qualitative data analysis (QDA) is used in the social sciences to draw conclusions from non-numerical data like words, photographs or video. In information technology, the term has a special meaning in the context of IT audits, when the controls for an organization's information systems, operations and processes are examined. Data analysis is used to determine whether the systems in place effectively protect data, operate efficiently and succeed in accomplishing an organization's overall goals.

The term "analytics" has been used by many business intelligence (BI) software vendors as a buzzword to describe quite different functions. Data analytics is used to describe everything from online analytical processing (OLAP) to CRM analytics in call centers. Banks and credit cards companies, for instance, analyze withdrawal and spending patterns to prevent fraud or identity theft. Ecommerce companies examine Web site traffic or navigation patterns to determine which customers are more or less likely to buy a product or service based upon prior purchases or viewing trends. Modern data analytics often use information dashboards supported by real-time data streams. So-called real-time analytics involves dynamic analysis and reporting, based on data entered into a system less than one minute before the actual time of use.

5.1 Supervised and Unsupervised Learning

Supervised learning is tasked with learning a function from labeled training data in order to predict the value of any valid input. Common examples of supervised learning include classifying e-mail messages as spam, labeling Web pages according to their genre, and recognizing handwriting. Many algorithms are used to create supervised learners, the most common being neural networks, Support Vector Machines (SVMs), and Naive Bayes classifiers.

Unsupervised learning is tasked with making sense of data without any examples of what is correct or incorrect. It is most commonly used for clustering similar input into logical groups. Unsupervised learning can be used to reduce the number of dimensions in a data set in order to focus on only the most useful attributes, or to detect trends. Common approaches to unsupervised learning include k-Means, hierarchical clustering, and self-organizing maps.

5.2 Decision Tree Learning

Decision tree learning, used in statistics, data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value.

More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making.

5.3 Data Mining and Machine Learning

Data Mining and Machine Learning are commonly confused, as they often employ the same methods and overlap significantly. They can be roughly defined as follows:

- Machine learning focuses on prediction, based on known properties learned from the training data.

- Data mining (which is the analysis step of ***Knowledge Discovery in Databases***) focuses on the discovery of (previously) unknown properties on the data.

The two areas overlap in many ways: data mining uses many machine learning methods, but often with a slightly different goal in mind. On the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy.

Much of the confusion between these two research communities (which do often have separate conferences and separate journals, ECML PKDD being a major exception) comes from the basic assumptions they work with: in machine learning, performance is usually evaluated with respect to the ability to reproduce known knowledge, while in Knowledge Discovery and Data Mining (KDD) the key task is the discovery of previously unknown knowledge.

Evaluated with respect to known knowledge, an uninformed (unsupervised) method will easily be outperformed by supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data.

5.4 Clustering

Given large data sets, whether they are text or numeric, it is often useful to group together, or cluster, similar items automatically. For instance, given all of the news for the day from all of the newspapers in the United States, you might want to group all of the articles about the same story together automatically; you can then choose to focus on specific clusters and stories without needing to wade through a lot of unrelated ones. Another example: Given the output from sensors on a machine over time, you could cluster the outputs to determine normal versus problematic operation, because normal operations would all cluster together and abnormal operations would be in outlying clusters.

Like CF, clustering calculates the similarity between items in the collection, but its only job is to group together similar items. In many implementations of clustering, items in the collection are represented as vectors in an n-dimensional space. Given the vectors, one can calculate the distance between two items using measures such as the **Manhattan Distance**, **Euclidean distance**, or **cosine similarity**. Then, the actual clusters can be calculated by grouping together the items that are close in distance. There are many approaches to calculating the clusters, each with its own trade-offs. Some approaches work from the bottom up, building up larger clusters from smaller ones, whereas others break a single large cluster into smaller and smaller clusters. Both have criteria for exiting the process at some point before they break down into a trivial cluster representation (all items in one cluster or all items in their own cluster). Popular approaches include k-Means and hierarchical clustering. As I'll show later, Mahout comes with several different clustering approaches.

5.5 Categorization

The goal of categorization (often also called classification) is to label unseen documents, thus grouping them together. Many classification approaches in machine learning calculate a variety of statistics that associate the features of a document with the specified label, thus creating a model that can be used later to classify unseen documents. For example, a simple approach to classification might keep track of the words associated with a label, as well as the number of times those words are seen for a given label. Then, when a new document is classified, the words in the document are looked up in the model, probabilities are calculated, and the best result is output, usually along with a score indicating the confidence the result is correct. Features for classification might include words, weights for those words (based on frequency, for instance), parts of speech, and so on. Of course, features really can be anything that helps associate a document with a label and can be incorporated into the algorithm.

5.6 Collaborative filtering

Collaborative filtering (CF) is a technique, popularized by Amazon and others, that uses user information such as ratings, clicks, and purchases to provide recommendations to other site users. CF is often used to recommend consumer items such as books, music, and movies, but it is also used in other applications where multiple actors need to collaborate to narrow down data.

5.7 Random Forest

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

6 Parallel Computing

Parallel computing is a form of computation in which many calculations are carried out simultaneously,operating on the principle that large problems can often be divided into smaller ones, which are then solved concurrently ("in parallel"). There are several different forms of parallel computing: bit-level, instruction level, data, and task parallelism. Parallelism has been employed for many years, mainly in high-performance computing, but interest in it has grown lately due to the physical constraints preventing frequency scaling.

As power consumption (and consequently heat generation) by computers has become a concern in recent years,parallel computing has become the dominant paradigm in computer architecture, mainly in the form of multicore processors.

Parallel computers can be roughly classified according to the level at which the hardware supports parallelism, with multi-core and multi-processor computers having multiple processing elements within a single machine, while clusters, MPPs, and grids use multiple computers to work on the same task. Specialized parallel computer architectures are sometimes used alongside traditional processors, for accelerating specific tasks.

Parallel computer programs are more difficult to write than sequential ones, because concurrency introduces several new classes of potential software bugs, of which race conditions are the most common. Communication and synchronization between the different subtasks are typically some of the greatest obstacles to getting good parallel program performance.

The maximum possible speed-up of a program as a result of parallelization is known as Amdahl's law.

7 Data Protection - A Summary

What does the Data Protection Act cover? The DPA covers personal data which is defined as information relating to a living individual who can be identified from those data, or from those data and other information which is in the possession of or is likely to come into the possession of, the data controller. Personal data includes expression of opinion and indications of the intentions of the data controller or any other person in respect of the individual.

There is a subsection of personal data known as sensitive personal data, this includes information regarding racial or ethnic origin, political opinions, religious beliefs, membership of trade unions, physical or mental health, sexual life, the commission or alleged commission of any offence, and any related proceedings.

What does the Data Protection Act mean for the University? The Information Commissioners Office (ICO) oversees the Data Protection Act; the University is registered with the ICO and must annually renew this notification. The Data Protection Act regulates how the University can process personal information and sets out 8 principles which must be followed.

8 What are the 8 Data Protection Principles?

The Data Protection Principles outline best practice with regards to processing Personal Data and must be complied with. The principles are:-

1. Personal data shall be processed fairly and lawfully.

2. Personal data shall be obtained only for one or more specified purposes, and shall not be further processed in any manner incompatible with that purpose or those purposes.

3. Personal data shall be adequate, relevant and not excessive.

4. Personal data shall be accurate and where necessary, kept up to date.

5. Personal data processed for any purpose or purposes shall not be kept for longer than is necessary.

6. Personal data shall be processed in accordance with the rights of data subjects under the Act.

7. Appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data.

8. Personal data shall not be transferred to a country outside the Euro-

pean Economic Area unless that country ensures an adequate level of protection.

How does the Data Protection Act affect how the University uses personal data? In addition to the Data Protection principles outlined above the DPA specifies conditions that must be met when processing personal data, the lists below are not exhaustive but contain the conditions that are likely to be relied upon by the University.When processing Personal Data one of the following conditions must be met:

The individual has given consent.
The processing is necessary for the performance of a contract.
The processing is necessary for a legal obligation.
The processing is necessary for the protection of the data subjects vital interests.
The processing in necessary for the exercise of any other functions of a public nature exercised in the public interest.
The processing is necessary for the purposes of legitimate interests pursued by the data controller.
When processing Sensitive Personal Data not only must one of the above apply, but there are additional conditions, at least one of which must be met:

The data subject has given his explicit consent. The processing is necessary for compliance with legal obligations in connection with employment. The processing is necessary to protect the vital interests of the data subject or another person where consent cannot be given by or on behalf of the data subject, and the data controller cannot reasonably be expected to obtain consent The processing in necessary to protect the vital interests of another person, in a case where consent of the data subject has been unreasonably withheld. The personal data has been made public as a result of steps deliberately taken by the data subject. The processing is necessary for the purpose of, or in connection with, any legal proceedings or for the purpose of obtaining legal advice.

The processing is of sensitive personal data consisting of information as to racial or ethnic origin, is for the purpose of identifying or reviewing the existence or absence of equality of opportunity or treatment between persons of different racial or ethnic origins, with a view to enabling such equality to be promoted or maintained, and is carried out with appropriate safeguards for the rights and freedoms of data subjects. What happens if the DPA is

breached?
The Information Commissioner has the authority to carry out Assessments of any Data Controllers against whom he has received complaints, if they are found to be breaching the DPA enforcement notices will be issued to force compliance. Breaches can also be tried in court.

The Act provides for separate personal liability for any of the offences in the Act. If a member of staff consents to an offence committed by the University, or that offence is attributable to any neglect on his/her part, that member of staff can be proceeded against and fined accordingly. Additionally, a data subject has the right to sue for compensation if he/she has suffered damage and/or distress as a result of the Universitys breach of the data protection regulations.

8.1 Offences under the act include

:

Processing without notification Failure to notify the commissioner of changes to notification register entry Failure to comply with an enforcement notice/information notice/special information notice Knowingly or recklessly obtaining or disclosing personal data or the information contained in personal data without the consent of the data subject.

9 some key terms

9.1 Primary Key

9.2 Foreign Key

9.3 Tuple

10 Database Administration

Database administrators design, implement, maintain and repair an organisations database. The role includes developing and designing the database strategy, monitoring and improving database performance and capacity, and planning for future expansion requirements. They may also plan, co-ordinate and implement security measures to safeguard the database.

A database administrator may

- undertake daily administration, including monitoring system performance, ensuring successful backups, and developing/implementing disaster recovery plans
- manage data to give users the ability to access, relate and report information in different ways
- develop standards to guide the use and acquisition of software and to protect valuable information
- modify existing databases or instruct programmers and analysts on the required changes
- test programs or databases, correct errors and make necessary modifications
- train users and answer questions

11 Knowledge Discovery in Databases

Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have created an immense need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions.

11.1 Data Rich, Information Poor

The amount of raw data stored in corporate databases is exploding. From trillions of point-of-sale transactions and credit card purchases to pixel-by-pixel images of galaxies, databases are now measured in gigabytes and terabytes. (One terabyte = one trillion bytes. A terabyte is equivalent to about 2 million books).

For instance, every day, Wal-Mart uploads 20 million point-of-sale transactions to an A&T massively parallel system with 483 processors running a centralized database. Raw data by itself, however, does not provide much information. In today's fiercely competitive business environment, companies need to rapidly turn these terabytes of raw data into significant insights into their customers and markets to guide their marketing, investment, and management strategies.

11.2 Data Warehouses

The drop in price of data storage has given companies willing to make the investment a tremendous resource: Data about their customers and potential customers stored in "Data Warehouses." Data warehouses are becoming part of the technology. Data warehouses are used to consolidate data located in disparate databases. A data warehouse stores large quantities of data by specific categories so it can be more easily retrieved, interpreted, and sorted by users. Warehouses enable executives and managers to work with vast stores of transactional or other data to respond faster to markets and make more informed business decisions. It has been predicted that every business will have a data warehouse within ten years. But merely storing data in a data warehouse does a company little good. Companies will want to learn more about that data to improve knowledge of customers and markets. The company benefits when meaningful trends and patterns are extracted from the data.

11.3 What is Data Mining?

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides.

11.4 What Can Data Mining Do?

Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, heath care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed.

For businesses, data mining is used to discover patterns and relationships in the data in order to help make better business decisions. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty. Specific uses of data mining include:

- Market segmentation - Identify the common characteristics of customers who buy the same products from your company.

- Customer churn - Predict which customers are likely to leave your company and go to a competitor.
- Fraud detection - Identify which transactions are most likely to be fraudulent.
- Direct marketing - Identify which prospects should be included in a mailing list to obtain the highest response rate.
- Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.
- Market basket analysis - Understand what products or services are commonly purchased together; e.g., beer and diapers.
- Trend analysis - Reveal the difference between a typical customer this month and last.

Data mining technology can generate new business opportunities by:

Automated prediction of trends and behaviors: Data mining automates the process of finding predictive information in a large database. Questions that traditionally required extensive hands-on analysis can now be directly answered from the data. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Automated discovery of previously unknown patterns: Data mining tools sweep through databases and identify previously hidden patterns. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Using massively parallel computers, companies dig through volumes of data to discover patterns about their customers and products. For example, grocery chains have found that when men go to a supermarket to buy diapers, they sometimes walk out with a six-pack of beer as well. Using that information, it's possible to lay out a store so that these items are closer.

AT&T, A.C. Nielson, and American Express are among the growing ranks of companies implementing data mining techniques for sales and marketing. These systems are crunching through terabytes of point-of-sale data to aid analysts in understanding consumer behavior and promotional strategies. Why? To gain a competitive advantage and increase profitability!

Similarly, financial analysts are plowing through vast sets of financial records, data feeds, and other information sources in order to make investment decisions. Health-care organizations are examining medical records to understand trends of the past so they can reduce costs in the future.

11.5 The Evolution of Data Mining

Data mining is a natural development of the increased use of computerized databases to store data and provide answers to business analysts.

- Evolutionary Step Business Question Enabling Technology
- Data Collection (1960s) "What was my total revenue in the last five years?" computers, tapes, disks
- Data Access (1980s) "What were unit sales in New England last March?" faster and cheaper computers with more storage, relational databases
- Data Warehousing and Decision Support "What were unit sales in New England last March? Drill down to Boston." faster and cheaper computers with more storage, On-line analytical processing (OLAP), multidimensional databases, data warehouses
- Data Mining "What's likely to happen to Boston unit sales next month? Why?" faster and cheaper computers with more storage, advanced computer algorithms

Traditional query and report tools have been used to describe and extract what is in a database. The user forms a hypothesis about a relationship and verifies it or discounts it with a series of queries against the data. For example, an analyst might hypothesize that people with low income and high debt are bad credit risks and query the database to verify or disprove this assumption. Data mining can be used to generate an hypothesis. For example, an analyst might use a neural net to discover a pattern that analysts did not think to try - for example, that people over 30 years old with low incomes and high debt but who own their own homes and have children are good credit risks.

11.6 How Data Mining Works

How is data mining able to tell you important things that you didn't know or what is going to happen next? That technique that is used to perform these feats is called modeling. Modeling is simply the act of building a model (a set of examples or a mathematical relationship) based on data from situations where the answer is known and then applying the model to other situations where the answers aren't known. Modeling techniques have been around for centuries, of course, but it is only recently that data storage and communication capabilities required to collect and store huge amounts of data, and the computational power to automate modeling techniques to work directly on the data, have been available.

As a simple example of building a model, consider the director of marketing for a telecommunications company. He would like to focus his marketing and sales efforts on segments of the population most likely to become big users of long distance services. He knows a lot about his customers, but it is impossible to discern the common characteristics of his best customers because there are so many variables. From his existing database of customers, which contains information such as age, sex, credit history, income, zip code, occupation, etc., he can use data mining tools, such as neural networks, to identify the characteristics of those customers who make lots of long distance calls. For instance, he might learn that his best customers are unmarried females between the age of 34 and 42 who make in excess of 60,000 per year. This, then, is his model for high value customers, and he would budget his marketing efforts to accordingly.

11.7 Data Mining Technologies

The analytical techniques used in data mining are often well-known mathematical algorithms and techniques. What is new is the application of those techniques to general business problems made possible by the increased availability of data and inexpensive storage and processing power. Also, the use of graphical interfaces has led to tools becoming available that business experts can easily use.

Some of the tools used for data mining are:

- Artificial neural networks - Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Decision trees - Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.
- Rule induction - The extraction of useful if-then rules from data based on statistical significance.
- Genetic algorithms - Optimization techniques based on the concepts of genetic combination, mutation, and natural selection.
- Nearest neighbor - A classification technique that classifies each record based on the records most similar to it in an historical database.

11.8 Real-World Examples

Details about who calls whom, how long they are on the phone, and whether a line is used for fax as well as voice can be invaluable in targeting sales of services and equipment to specific customers. But these tidbits are buried in masses of numbers in the database. By delving into its extensive customer-call database to manage its communications network, a regional telephone company identified new types of unmet customer needs. Using its data mining system, it discovered how to pinpoint prospects for additional services by measuring daily household usage for selected periods. For example, households that make many lengthy calls between 3 p.m. and 6 p.m. are likely to include teenagers who are prime candidates for their own phones and lines. When the company used target marketing that emphasized convenience and value for adults - ”Is the phone always tied up?” - hidden demand surfaced. Extensive telephone use between 9 a.m. and 5 p.m. characterized by patterns related to voice, fax, and modem usage suggests a customer has business activity. Target marketing offering those customers ”business communications capabilities for small budgets” resulted in sales of additional lines, functions, and equipment.

The ability to accurately gauge customer response to changes in business rules is a powerful competitive advantage. A bank searching for new ways to increase revenues from its credit card operations tested a nonintuitive possibility: Would credit card usage and interest earned increase significantly if the bank halved its minimum required payment? With hundreds of gigabytes of data representing two years of average credit card balances, payment amounts, payment timeliness, credit limit usage, and other key parameters, the bank used a powerful data mining system to model the impact of the proposed policy change on specific customer categories, such as customers consistently near or at their credit limits who make timely minimum or small payments. The bank discovered that cutting minimum payment requirements for small, targeted customer categories could increase average balances and extend indebtedness periods, generating more than $25 million in additional interest earned,

Merck-Medco Managed Care is a mail-order business which sells drugs to the country's largest health care providers: Blue Cross and Blue Shield state organizations, large HMOs, U.S. corporations, state governments, etc. Merck-Medco is mining its one terabyte data warehouse to uncover hidden links between illnesses and known drug treatments, and spot trends that help pinpoint which drugs are the most effective for what types of patients. The results are more effective treatments that are also less costly. Merck-Medco's data mining project has helped customers save an average of 10-15% on prescription costs.

11.9 The Future of Data Mining

In the short-term, the results of data mining will be in profitable, if mundane, business related areas. Micro-marketing campaigns will explore new niches. Advertising will target potential customers with new precision.

In the medium term, data mining may be as common and easy to use as e-mail. We may use these tools to find the best airfare to New York, root out a phone number of a long-lost classmate, or find the best prices on lawn mowers.

The long-term prospects are truly exciting. Imagine intelligent agents turned loose on medical research data or on sub-atomic particle data. Computers may reveal new treatments for diseases or new insights into the nature of the universe. There are potential dangers, though, as discussed below.

11.10 Privacy Concerns

What if every telephone call you make, every credit card purchase you make, every flight you take, every visit to the doctor you make, every warranty card you send in, every employment application you fill out, every school record you have, your credit record, every web page you visit ... was all collected together? A lot would be known about you! This is an all-too-real possibility. Much of this kind of information is already stored in a database. Remember that phone interview you gave to a marketing company last week? Your replies went into a database. Remember that loan application you filled out? In a database. Too much information about too many people for anybody to make sense of? Not with data mining tools running on massively parallel processing computers! Would you feel comfortable about someone (or lots of someones) having access to all this data about you? And remember, all this data does not have to reside in one physical location; as the net grows, information of this type becomes more available to more people.

12 Cluster Anlysis

13 Data dredging

Data dredging (data fishing, data snooping) is the inappropriate (sometimes deliberately so) use of data mining to uncover misleading relationships in data. Data-snooping bias is a form of statistical bias that arises from this misuse of statistics. Any relationships found might appear to be valid within the test set but they would have no statistical significance in the wider population.

13.1 Data Dredging

Data dredging and data-snooping bias can occur when researchers either do not form a hypothesis in advance or narrow the data used to reduce the probability of the sample refuting a specific hypothesis. Although data-snooping bias can occur in any field that uses data mining, it is of particular concern in finance and medical research, both of which make heavy use of data mining techniques.

14 Web-mining

More than ever, entities and individuals alike are using the World Wide Web to conduct a host of business and personal transactions. As a result, companies are increasingly employing Web data mining tools and techniques in order to find ways to improve their bottom lines and grow their customer base. Web data mining involves the process of collecting and summarizing data from a Web sites hyperlink structure, page content, or usage log in order to identify patterns. Using Web data mining, a company can identify a potential competitor, improve customer service, or target customer needs and expectations. A government agency may also seek to uncover terrorist threats or other criminal activities through the use of a Web data mining application.

Some common Web data mining techniques include Web content mining, Web usage mining, and Web structure mining. Web content mining examines the subject matter of a Web site. For example, Web content miners may analyze a site's audio, text, images, and video features. Web content miners typically focus on a sites textual information more than other site features. Natural language processing and information retrieval are two data mining techniques often used by Web content miners.

Web usage mining is usually an automated process whereby Web servers collect and report user access patterns in server access logs. A company may, for example, use a Web usage data mining tool to report on server access logs and user registration information in order to create a more effective Web site structure. Web structure mining studies the node and connection structure of Web sites. It can be useful in identifying similarities and relationships that exist among different Web sites. Web structure mining often involves uncovering patterns from hyperlinks or pulling out document structures on a Web page.

Two general data mining techniques that can be employed by Web data miners are data mining association analysis and data mining regression. Data mining association analysis helps uncover noteworthy relationships buried in large data sets. **Data mining regression** is a statistical technique whereby mathematical formulas are used to predict future results, such as profit margins, house values, or sales figures.

Data mining software vendors offer Web data mining tools that can pull out predictive information from large quantities of data. Businesses often use these software mining tools to analyze specific data sets regarding consumer behavior. Using the results of the data analysis, companies are able to forecast future business trends.

Data mining (DMM), also called Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using tools such as classification, association rule mining, clustering, etc. Data mining is a complex topic and has links with multiple core fields such as computer science and adds value to rich seminal computational techniques from statistics, information retrieval, machine learning and pattern recognition.

Data mining has been defined as *"the nontrivial extraction of implicit, previously unknown, and potentially useful information from data"* and *"the science of extracting useful information from large data sets or databases"* . It involves sorting through large amounts of data and picking out relevant information. It is usually used by businesses, intelligence organizations, and financial analysts, but is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. Metadata, or data about a given data set, are often expressed in a condensed data mine-able format, or one that facilitates the practice of data mining. Common examples include executive summaries and scientific abstracts.

Although data mining is a relatively new term, the technology is not. Companies for a long time have used powerful computers to sift through volumes of data such as supermarket scanner data, and produce market research reports. Continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy and usefulness of analysis. Data mining identifies trends within data that go beyond simple analysis. Through the use of sophisticated algorithms, users have the ability to identify key attributes of business processes and target opportunities. The term data mining is often used to apply to the two separate processes of knowledge discovery and prediction. Knowledge discovery provides explicit information that has a readable form and can be understood by a user.

Forecasting, or predictive modeling provides predictions of future events and may be transparent and readable in some approaches (e.g. rule based systems) and opaque in others such as neural networks. Moreover, some data mining systems such as neural networks are inherently geared towards prediction and pattern recognition, rather than knowledge discovery.

The term "data mining" is often used incorrectly to apply to a variety of other processes besides data mining. In many cases, applications may claim to perform "data mining" by automating the creation of charts or graphs with historic trends and analysis. Although this information may be useful and timesaving, it does not fit the traditional definition of data mining, as the application performs no analysis itself and has no understanding of the underlying data. Instead, it relies on templates or pre-defined macros (created either by programmers or users) to identify trends, patterns and differences. A key defining factor for true data mining is that the application itself is performing some real analysis. In almost all cases, this analysis is guided by some degree of user interaction, but it must provide the user some insights that are not readily apparent through simple slicing and dicing. Applications that are not to some degree self-guiding are performing data analysis, not data mining.

15 Predictive analytics

Predictive analytics encompasses a variety of techniques from statistics, data mining and game theory that analyze current and historical facts to make predictions about future events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

Predictive analytics is used in financial services, insurance, telecommunications, retail, travel, healthcare, pharmaceuticals and other fields.

One of the most well-known applications is credit scoring, which is used throughout financial services. Scoring models process a customer's credit history, loan application, customer data, etc., in order to rank-order individuals by their likelihood of making future credit payments on time.

Very simply, the aim of research is to add something of value to an existing body of knowledge.