Data Science - The ABCs of Data Science

The Author

June 18, 2013

Contents

1 Preliminaries

For data scientists, there's SQL, statistics, predictive modeling and programming (probably Python)

2 Apache: List of Products

2.1 Apache Whirr

Apache Whirr is a set of libraries for running cloud services.

3 Apache Hadoop

Apache Hadoop was created out of necessity as data from the web exploded, and grew far beyond the ability of traditional systems to handle it. adoop was initially inspired by papers published by Google outlining its approach to handling an avalanche of data, and has since become the de facto standard for storing, processing and analyzing hundreds of terabytes, and even petabytes of data.

Apache Hadoop is 100% open source, and pioneered a fundamentally new way of storing and processing data. Instead of relying on expensive, proprietary hardware and different systems to store and process data, Hadoop enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits. With Hadoop, no data is too big. And in today's hyper-connected world where more and more data is being created every day, Hadoop's breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless.

Apache Hadoop is a software framework that supports data-intensive distributed applications under a free license.It enables applications to work with thousands of nodes and petabytes of data. Hadoop was inspired by Google's MapReduce and Google File System (GFS) papers.

3.1 Hadoop and MapReduce

Apache Hadoop is a good place to start with the MapReduce segment. Hadoop began conceptually with a paper that emanated from Google in 2004 and described a process for parallelizing the processing of web-based data it called MapReduce. Shortly thereafter, Apache Hadoop was born as an open source implementation of the MapReduce process. The community surrounding it is growing dramatically and producing add-ons that expands Apache Hadoop's usability within corporate data centers.

Apache Hadoop users typically build their own parallelized computing clusters from commodity servers, each with dedicated storage in the form of a small disk array or, more recently, solid-state drive (SSD) for performance. These are commonly referred to as "shared-nothing" architectures. Storage-area network (SAN) and network-attached storage (NAS), while scalable and resilient, are typically seen as lacking the kind of I/O performance these clusters need to rise above the capabilities of the standard data warehouse. Therefore, Hadoop storage is direct-attached storage (DAS). However, the use of SAN and NAS as "secondary" storage is emerging.

A potential Hadoop user is confronted with a growing list of sourcing choices that range from pure open source to highly commercialized versions. Apache Hadoop and related tools are available for free at the Apache Hadoop site. Cloudera Inc. offers a commercial version that includes some Cloudera add-ons and support. Other open source variants, such as the Facebook distribution, are also available from Cloudera. Commercial versions include MapR, which EMC Corp. now incorporates into a Hadoop appliance.

4 Apache Oozie

Oozie is a workflow scheduler system to manage Apache Hadoop jobs.

Oozie Workflow jobs are ***Directed Acyclical Graphs*** (DAGs) of actions.

Oozie Coordinator jobs are recurrent Oozie Workflow jobs triggered by time (frequency) and data availabilty.

Oozie is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs out of the box (such as Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop and Distcp) as well as system specific jobs (such as Java programs and shell scripts).

Oozie is a scalable, reliable and extensible system.

5 Apache Pig

Apache Pig is a high-level procedural language platform developed to simplify querying large data sets in Apache Hadoop and MapReduce. Apache Pig features a ***Pig Latin*** language layer that enables SQL-like queries to be performed on distributed datasets within Hadoop applications. Pig originated as a Yahoo Research initiative for creating and executing map-reduce jobs on very large data sets. In 2007 Pig became an open source project of the Apache Software Foundation.

6 Big Data

Big data (also spelled Big Data) is a general term used to describe the voluminous amount of unstructured and semi-structured data a company creates – data that would take too much time and cost too much money to load into a relational database for analysis. Although Big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data.

A primary goal for looking at big data is to discover repeatable business patterns. It's generally accepted that unstructured data, most of it located in text files, accounts for at least 80% of an organization's data. If left unmanaged, the sheer volume of unstructured data that's generated each year within an enterprise can be costly in terms of storage. Unmanaged data can also pose a liability if information cannot be located in the event of a compliance audit or lawsuit.

Big data are measurements of data that have grown so large that normal databases are unable to contain and work with the massive amount of information. Data come in three sizes: small, medium and big; none of these measurements is strict; instead, each depends more on ease of use and what type of machine can handle the information. Special machines, much larger and complex than those used for ordinary databases, are needed for big data. These types of data are typically found in government and scientific agencies, but some very large websites also contain this large amount of information.

Data come in three standard, but not strict, sizes. Small data are able to fit on a single computer or machine, such as a laptop. Medium data are able to fit on a disk array and are best managed by a database. Databases, no matter how large, are incapable of working with big data, and special systems much be used instead. While there is no strict guideline for what big data are, it typically starts around the terabyte (TB) level and goes up to the petabyte (PB) level.

Attempting to work with big data on a database that is not specialized for this amount of data will cause several substantial problems. The database is not able to handle the amount of information, so some data must be erased. This is like trying to fit 100 gigabytes (GB) on a computer with only 50 GB of hard drive space; it cannot be done. The data left will be unwieldy to both control and manage, because any function would take a long time to complete and the database must be closed off to new submissions.

While it is possible to keep purchasing machines and adding new data to the databases, this creates the unwieldy problem. This is because database software is only made to work with medium data. Larger datasets lead to errors and administrative problems, because the software simply cannot move or work with large data without encountering problems.

Big data are not encountered by most organizations or websites. Defense and military agencies use this amount of information to create models and store test results, and many large scientific agencies need these specialized machines for similar reasons. Some very large websites need large data machines, but websites are not as common as agencies in this market. These organizations need to keep all their data, because it helps to better

analyze future data and make predictions.

6.1 The Big Data appliance

As the interest in Big Data analytics expands into enterprise data centers, the vendor community sees an opportunity to put together Big Data "appliances." These appliances integrate server, networking and storage gear into a single enclosure and run analytics software that accelerates information delivery to users. These appliances are targeted at enterprise buyers who will value the ease of implementation and use characteristics inherent in Big Data appliances. Vendors in this space include EMC with appliances built around the Greenplum database engine, IBM/Netezza, MapR's recently announced commercialized version of Hadoop, Oracle and Teradata with comparable, pre-integrated systems.

7 Big Data analytics

Big Data analytics is an area of rapidly growing diversity. Therefore, trying to define it is probably not helpful. What is helpful, however, is identifying the characteristics that are common to the technologies now identified with Big Data analytics. These include:

- The perception that traditional data warehousing processes are too slow and limited in scalability
- The ability to converge data from multiple data sources, both structured and unstructured
- The realization that time to information is critical to extract value from data sources that include mobile devices, RFID, the web and a growing list of automated sensory technologies

In addition, there are at least four major developmental segments that underline the diversity to be found within Big Data analytics. These segments are MapReduce, scalable database, real-time stream processing and Big Data appliance.

8 Big Data storage for Big Data analytics

The practitioners of Big Data analytics processes are generally hostile to shared storage. They prefer DAS in its various forms, from SSD to high-capacity SATA disk buried inside parallel processing nodes. Shared storage architectures, such as SAN and NAS, are typically perceived as relatively slow, complex and, above all, expensive. These qualities aren't consistent with Big Data analytics systems, which thrive on system performance, commodity infrastructure and low cost.

Real-time or near-real-time information delivery is one of the defining characteristics of Big Data analytics; therefore, latency is avoided whenever and wherever possible. Data in memory is good; data on spinning disk at the other end of a Fibre Channel SAN connection is not. But perhaps worse than anything else, the cost of a SAN at the scale needed for analytics applications is thought to be prohibitive.

There's a case to be made for shared storage in Big Data analytics. Yet storage vendors and the storage community in general, have yet to make that case to practitioners of Big Data analytics. An example can be seen in the integration of the ParAccel's Analytic Database (PADB) with NetApp SAN storage.

Developers of data storage technology are moving away from expressing storage as a physical device and toward the implementation of storage as a more virtual and abstract entity. As a result, the shared storage environment can and should be seen by Big Data practitioners as one in which they can find potentially valuable data services, such as:

1. Data protection and system availability: Storage-based copy functions that don't require database quiescence can create restartable copies of data to recover from system failures and data corruption occurrences.

2. Reduced time to deployment for new applications and automated processes: Business agility is enhanced when new applications can be brought online quickly by building them around reusable data copies.

3. Change management: Shared storage can potentially lessen the impact of required changes and upgrades to the online production environment by helping to preserve an "always-on" capability.

4. Lifecycle management: The evolution of systems becomes more manageable and obsolete applications become easier to discard when shared storage can serve as the database of record.

5. Cost savings: Using shared storage as an adjunct to DAS in a shared-nothing architecture reduces the cost and complexity of processor nodes.

Each of the above mentioned benefits can be mapped to shared-nothing analytics architectures. One can expect to see more storage vendors doing this over time. For example, while it hasn't been announced, EMC could integrate Isilon or Atmos storage with its MapR-based appliance.

Traditional data warehousing is a large but relatively slow producer of information to business analytics users. It draws from limited data resources and depends on reiterative extract, transform and load (ETL) processes. Customers are now looking for quick access to information that is based on culling nuggets from multiple data sources concurrently. Big Data analytics can be defined, to some extent, in relationship to the need to parse large data sets from multiple sources, and to produce information in real-time or near-real-time.

Big Data analytics represents a big opportunity. IT organizations are exploring the analytics technologies outlined above to parse web-based data sources and extract value from the social networking boom. However, an even larger opportunity – the Internet of Things – is emerging as a data source. Cisco Systems Inc. estimates there are approximately 35 billion electronic devices that can connect to the Internet. Any electronic device can be connected (wired or wirelessly) to the Internet, and even automakers are building Internet connectivity into vehicles. "Connected" cars will become commonplace by 2012 and generate millions of transient data streams.

9 Big Memory

A petabyte (derived from the SI prefix peta-) is a unit of information equal to one quadrillion (short scale) bytes, or 1000 terabytes. The unit symbol for the petabyte is PB. The prefix peta (P) indicates the fifth power to 1000:

```
1 PB     = 1000000000000000B

= 10005 B

= 1015 B

= 1 million gigabytes

= 1 thousand terabytes
```

The exabyte (derived from the SI prefix exa-) is a unit of information or computer storage equal to one quintillion bytes (short scale). The unit symbol for the exabyte is EB. The unit prefix exa indicates the sixth power of 1000:

```
1 EB     = 1,000,000,000,000,000,000B

= 1018 bytes

= 1,000,000,000 gigabytes

= 1000000terabytes
```

10 Cassandra

Apache Cassandra is an open source distributed database management system. It is an Apache Software Foundation top-level project[1] designed to handle very large amounts of data spread out across many commodity servers while providing a highly available service with no single point of failure. It is a NoSQL solution that was initially developed by Facebook and powered their Inbox Search feature until late 2010. Jeff Hammerbacher, who led the Facebook Data team at the time, has described Cassandra as a BigTable data model running on an Amazon Dynamo-like infrastructure.

Cassandra provides a structured key-value store with tunable consistency. Keys map to multiple values, which are grouped into column families. The column families are fixed when a Cassandra database is created, but columns can be added to a family at any time. Furthermore, columns are added only to specified keys, so different keys can have different numbers of columns in any given family. The values from a column family for each key are stored together.

11 Classification

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.

12 CouchDB

Apache CouchDB, commonly referred to as CouchDB, is an open source database that focuses on ease of use and on being "a database that completely embraces the web".

It is a NoSQL database that uses JSON to store data, JavaScript as its query language using MapReduce and HTTP for an API. One of its distinguishing features is multi-master replication. CouchDB was first released in 2005 and later became an Apache project in 2008. Unlike in a relational database, CouchDB does not store data and relationships in tables. Instead, each database is a collection of independent documents. Each document maintains its own data and self-contained schema. An application may access multiple databases, such as one stored on a user's mobile phone and another on a server. Document metadata contains revision information, making it possible to merge any differences that may have occurred while the databases were disconnected.

13 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.

13.1 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.

13.2 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.

13.3 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.

13.4 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.

13.5 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.

13.6 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.

13.7 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.

14 Databases : Key Terms

14.1 Primary Key

14.2 Foreign Key

14.3 Tuple

15 Database Administration

Database administrators design, implement, maintain and repair an organisation's 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

16 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.

16.1 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.

17 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.

18 Data Mining

Formally known as Knowledge Discovery in Databases (KDD)

18.1 About Data mining

Data mining uses a relatively large amount of computing power operating on a large set of data to determine regularities and connections between data points. Algorithms that employ techniques from statistics, machine learning and pattern recognition are used to search large databases automatically. Data mining is also known as Knowledge-Discovery in Databases (KDD).

Like the term artificial intelligence, data mining is an umbrella term that can be applied to a number of varying activities. In the corporate world, data mining is used most

frequently to determine the direction of trends and predict the future. It is employed to build models and decision support systems that give people information they can use. Data mining takes a front-line role in the battle against terrorism. It was supposedly used to determine the leader of the 9/11 attacks.

Data miners are statisticians who use techniques with names like near-neighbor models, k-means clustering, holdout method, k-fold cross validation, the leave-one-out method, and so on. Regression techniques are used to subtract irrelevant patterns, leaving only useful information. The term Bayesian is seen frequently in the field, referring to a class of inference techniques that predict the likelihood of future events by combining prior probabilities and probabilities based on conditional events.

Spam filtering is arguably a form of data mining, which automatically brings relevant messages to the surface from a chaotic sea of phishing attempts and Viagra pitches.

Decision trees are used to filter mountains of data. In a decision tree, all data passes through an entrance node, where it faces a filter that separates the data into streams depending on its characteristics. For example, data about consumer behavior is likely to be filtered based on demographic factors. Data mining is not primarily about fancy graphs and visualization techniques, but it does employ them to show what it has found. It is known that we can absorb more statistical information visually than verbally and this format for presentation can be very persuasive and powerful if used in the right context.

As our civilization becomes increasingly data-saturated and sensors are distributed en masse into our local environments, we will inadvertently discover things that might be missed on the first pass over. Data mining will let us correct these mistakes and discover new insights based on past data, giving us more bang for our data storage buck.

18.2 Data mining tools

Data mining tools are software components and theories that allow users to extract information from data. The tools provide individuals and companies with the ability to gather large amounts of data and use it to make determinations about a particular user or groups of users. Some of the most common uses of data mining tools are in the fields of marketing, fraud protections and surveillance.

The manual extraction of data has existed for hundreds of years. However, the automation of data mining has been most prevalent since the dawn of the computer age. During the 20th century, various computer sciences emerged to help support the concept of developing data mining tools.

The overall goal of the utilization of the tools is to uncover hidden patterns. For example, if a marketing company finds that a person takes a monthly trip from New York City to Los Angeles, it becomes beneficial for that company to advertise details of the destination to the individual.

Within the data mining industry, standards have been established to define the parameters of the use of data mining tools. Annually, the Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)

holds a meeting to determine what processes are used. The same group is also responsible for assessing the ethical implications of the analysis of data from individuals and companies. A biannual journal is published by the group entitled SIGKDD Explorations.

The most prevalent tool used in data mining is the process called Knowledge Discovery in Databases (KDD). KDD was developed in 1989 by Gregory Piatetsky-Shapiro. Using this data mining tool, users are able to process raw data, mine the data for information and interpret the various results in the form of information management.

One of the most important forms of data mining tools is used for combating terrorism in the 21st century. In the United States, the National Research Council uses the concepts of pattern mining and subject-based data mining to identify terrorist activity in the large pool of information around the world.

Pattern mining is defined by the process of locating patterns within a large volume of data. Subject-based data mining attempts to identify relationships between individuals. Both techniques can also be utilized in general business practice by defining the mindset of a customer base and the interactive relationship between customers.

18.3 Data Mining Techniques

Data mining generally refers to a method used to analyze data from a target source and compose that feedback into useful information. This information typically is used to help an organization cut costs in a particular area, increase revenue, or both. Often facilitated by a data-mining application, its primary objective is to identify and extract patterns contained in a given data set.

Most importantly, data mining techniques aim to provide insight that allows for a better understanding of data and its essential features. Companies and organizations can employ many different types of data mining techniques. While they may take a similar approach, all usually strive to meet different goals.

The purpose of predictive data mining techniques almost always is to identify statistical models or patterns that can be utilized to predict a response of interest. For example, a financial institution might use it to identify which transactions have the highest probability of fraud. This is the most common data mining technique and one that has become an efficient decision-making tool for mid- to large-sized companies. It also has proven effective at predicting customer behavior, categorizing customer segments, and forecasting various events.

Summary models rely on data mining techniques that respond accordingly to summarized data. For instance, an organization might assign airline passengers or credit card transactions into different groups based on their characteristics extracted from the analytical process. This model also can help businesses gain a deeper understanding of their customer base.

Association models take into account that certain events can occur together on a regular basis. This could be the simultaneous purchasing of items such as a mouse and keyboard or a sequence of events that led to the failure of a particular hardware device. Association models represent data mining techniques used to identify and characterize these associated occurrences.

Network models use data mining techniques to reveal data structures that are in the form of nodes and links. For example, an organized fraud ring might compile a list of stolen credit card numbers, and then turn around and use them to purchase items online. In this illustration, the credit cards and online merchants represent the nodes while the actual transactions act as the links.

Data mining has many purposes and can be used for both positive and malicious gain. More organizations are coming to discover the benefits of merging data mining techniques to form hybrid models. These powerful combinations often result in applications with superior performance. By integrating the key features of different methods into single hybrid solutions, organizations usually can overcome the limitations of individual strategy systems.

18.4 Important Data Mining concepts

The most important data mining concepts are used for the analysis of collected information, most notably in the effort to observe a behavior. Unknown interactions between data are researched in a variety of ways to ascertain critical relationships between subjects and aggregated information. One challenge in data mining is that the actual information collected may not be reminiscent of the whole domain. In an effort to address this fact, correlations between the data can be methodically controlled by the various data mining concepts.

Preprocessing-processing of the information is one of the most important aspects of data mining. The raw data must be mined and interpreted. In order to perform this action, a process must be determined, the target data should be assembled and patterns are found. The process is known as Knowledge Discovery in Databases and was developed by Gregory Piatetsky-Shapiro in 1989.

Four different classes of data mining concepts allow the process to take place. Clustering uses the algorithm created from the data mining process to assemble items into similar groups. Unlike clustering, classification of the information is when the data is assembled into predefined groups and analyzed. Association attempts to find relationships between variables, determining which groups of data are commonly associated. The final type of data mining is regression, based on the method of identifying a function within the data collection.

Validating the information is the final step in discovering what the data mining application represents. When not all algorithms present a valid data set, the patterns that occur can result in a situation called overfitting. To overcome this problem, the data is compared to a test set. This is a concept in which the measurements are aligned with a series of algorithms that would provide a plausible set of data sets. If the acquired information does not line up to the test set, then the assumed patterns in the data must be inaccurate.

Some of the most important data mining concepts occur in a variety of industries. Gaming, business, marketing, science, engineering and surveillance all utilize data mining techniques. By conducting these techniques, each field can determine best practices or better ways to find results.

18.5 Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)

Standards for data mining concepts are enforced by the Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD). This organization publishes the "International Journal of Information Technology and Decision Making" as well as the journal SIGKDD Explorations. Enforcing ethics and basic principles of data mining keeps the industry working efficiently and with limited legal problems.

19 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 Commissioner?s 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.

19.1 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

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

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

7. Appropriate technical and organisational measures shall be taken against unauthorised

8. Personal data shall not be transferred to a country outside the European Economic Area

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 subject?s 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 University?s breach of the data protection regulations.

19.2 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.

20 Data Science

20.1 What is Data Science

Data science incorporates varying elements and builds on techniques and theories from many fields, including math, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products. Data science is a novel term that is often used interchangeably with competitive intelligence or business analytics, although it is becoming more common. Data science seeks to use all available and relevant data to effectively tell a story that can be easily understood by non-practitioners. Some areas of research are:

- Cloud computing
- Databases and information integration
- Learning, natural language processing and information extraction
- Computer vision
- Information retrieval and web information access
- Knowledge discovery in social and information networks

21 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 $20,000 to 300,000$. 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.

22 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."

23 Data Storage

Much of the information available today is contained in some type of database. Blogs use databases to store posts and user information, discussion sites use them to store information about members, and organizations use them to store useful data for their business — from financial records to customer information.

The majority of the databases used today are relational databases that use structured queries to retrieve information and present it to the user. This was not always the case, and flat file databases were created to store information in a non-structured way.

A flat file is a collection of data stored and accessed sequentially. A comma separated values (CSV) sheet in Microsoft Excel is a flat file. There are no application specific formats applied to the data contained within the file and only a comma denotes the end of one field in a record. Each record is written on a line in the file, allowing all data for a single record to be stored separately from other records.

A flat file does not incorporate relationships with other tables that rely on special instructions to be used. The common database used today is a relational database. The data model used for this kind of storage allows information in one table to be related to information in other tables using key fields which exist in each table.

For example, suppose a customer calls an organization to place an order. The customer information is entered and stored. Then the order information is entered and stored. In a flat file, this information would be stored with the information for the order itself to allow the record for the order and/or the customer to be retrieved. Keep in mind that flat file databases do not have to use a single flat file. Information about orders could be stored in one flat file, while information about customers is stored in a different flat file. These files are not related in any way, so the flat file database for customer information has no idea that any information exists about orders.

To make a flat file data model functional, all relevant information about a record needs to be stored in the same file. Flat file databases can quickly become very large and difficult to manage because of the simple way they are organized. Many of today's more advanced data models use tables to organize groups of related data. This makes the data easier to locate and more flexible to work with.

The same customer example given above might look a bit different if a different data model were applied to the scenario. When the customer calls to place an order, his or her information is entered and stored in a customers table within the database. The information for his or her order is stored in two tables, order header and order details. Information like order number, order date, and customer id are stored in the order header table. The items ordered along with quantities and unit costs are stored in the order details table. The order details table also contains the order number, allowing this information to be related back to the order header information. In the order header table for this record, there is a reference to the customer id linking this order to the customer who ordered it.

24 Direct-attached storage (DAS)

Direct-attached storage (DAS) is computer storage that is directly attached to one computer or server and is not, without special support, directly accessible to other ones. The main alternatives to direct-attached storage are network-attached storage (NAS) and the storage area network (SAN).

For an individual computer user, the hard drive is the usual form of direct-attached storage. In an enterprise, providing for storage that can be shared by multiple computers and their users tends to be more efficient and easier to manage.

25 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.

26 Graph Analysis

Graph analysis is among the hottest techniques around for making sense of large datasets, primarily by determining how tightly different data points are related or how similar they are. The term "graph" came into the broader lexicon along with social networks, which built social graphs to assess the relationships among their millions of users, but the technique has much broader uses.

27 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. Graph databases are based on graph theory. Graph databases employ nodes, properties, and edges. Nodes are very similar in nature to the objects that object-oriented programmers will be familiar with.

28 Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS) is a sub-project of the Apache Hadoop project. This Apache Software Foundation project is designed to provide a fault-tolerant file system designed to run on commodity hardware. According to The Apache Software Foundation, the primary objective of HDFS is to store data reliably even in the presence of failures including NameNode failures, DataNode failures and network partitions. The NameNode is a single point of failure for the HDFS cluster and a DataNode stores data in the Hadoop file management system. HDFS uses a master/slave architecture in which one device (the master) controls one or more other devices (the slaves). The HDFS cluster consists of a single NameNode and a master server manages the file system namespace and regulates access to files.

29 Machine Learning

29.1 Machine Learning : Introductory Overview

Machine Learning includes a number of advanced statistical methods for handling regression and classification tasks with multiple dependent and independent variables. These methods include Support Vector Machines (SVM) for regression and classification, Naive Bayes for classification, and k-Nearest Neighbours (KNN) for regression and classification.

Detailed discussions of these techniques can be found in Hastie, Tibshirani, & Freedman (2001); a specialized comprehensive introduction to support vector machines can also be found in Cristianini and Shawe-Taylor (2000).

29.2 Support Vector Machines (SVM)

This method performs regression and classification tasks by constructing nonlinear decision boundaries. Because of the nature of the feature space in which these boundaries are found, Support Vector Machines can exhibit a large degree of flexibility in handling classification and regression tasks of varied complexities. There are several types of Support Vector models including linear, polynomial, RBF, and sigmoid.

29.3 Naive Bayes

This is a well established Bayesian method primarily formulated for performing classification tasks. Given its simplicity, i.e., the assumption that the independent variables are statistically independent, Naive Bayes models are effective classification tools that are easy to use and interpret. Naive Bayes is particularly appropriate when the dimensionality of the independent space (i.e., number of input variables) is high (a problem known as the curse of dimensionality). For the reasons given above, Naive Bayes can often outperform other more sophisticated classification methods. A variety of methods exist for modeling the conditional distributions of the inputs including normal, lognormal, gamma, and Poisson.

29.4 k-Nearest Neighbors

k-Nearest Neighbors is a memory-based method that, in contrast to other statistical methods, requires no training (i.e., no model to fit). It falls into the category of Prototype Methods. It functions on the intuitive idea that close objects are more likely to be in the same category. Thus, in KNN, predictions are based on a set of prototype examples that are used to predict new (i.e., unseen) data based on the majority vote (for classification tasks) and averaging (for regression) over a set of k-nearest prototypes (hence the name k-nearest neighbors).

30 Mahout

Apache Mahout is an Apache project to produce free implementations of distributed or otherwise scalable machine learning algorithms on the Hadoop platform. Mahout is a work in progress; the number of implemented algorithms has grown quickly but there are still various algorithms missing. While Mahout's core algorithms for clustering, classification and batch based collaborative filtering are implemented on top of Apache Hadoop using the map/reduce paradigm, it does not restrict contributions to Hadoop based implementations. Contributions that run on a single node or on a non-Hadoop cluster are welcomed. For example, the 'Taste' collaborative-filtering recommender component of Mahout was originally a separate project and can run stand-alone without Hadoop.

- Collaborative Filtering
- User and Item based recommenders
- K-Means, Fuzzy K-Means clustering
- Mean Shift clustering
- Dirichlet process clustering
- Latent Dirichlet Allocation
- Singular value decomposition
- Parallel Frequent Pattern mining
- Complementary Naive Bayes classifier
- Random forest decision tree based classifier

31 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.

31.1 An example of MapReduce

Let's 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 we've made this example very simple so it's easy to follow.

You can imagine that a real application won't be quite so simple, as it's 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 we're 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)
```

Let's 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) (New York
```

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.

32 MapReduce 2

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.

According to software engineer Mark C. Chu-Carroll:

> "The key to how MapReduce works is to take input as, conceptually, a list of records. The records are split among the different computers in the cluster by Map. The result of the Map computation is a list of key/value pairs. Reduce then takes each set of values that has the same key and combines them into a single value. So Map takes a set of data chunks and produces key/value pairs and Reduce merges things, so that instead of a set of key/value pair sets, you get one result. You can't tell whether the job was split into 100 pieces or 2 pieces...MapReduce isn't intended to replace relational databases: it's intended to provide a lightweight way of programming things so that they can run fast by running in parallel on a lot of machines."

MapReduce is important because it allows ordinary developers to use MapReduce library routines to create parallel programs without having to worry about programming for intra-cluster communication, task monitoring or failure handling. It is useful for tasks such as data mining, log file analysis, financial analysis and scientific simulations. Several implementations of MapReduce are available in a variety of programming languages, including Java, C++, Python, Perl, Ruby, and C.

32.1 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 weren't 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.

32.2 NoSQL implementations

Arguably, the most popular NoSQL database is Apache Cassandra. Cassandra, which was once Facebook's 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).

32.3 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.

32.3.1 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:

33 NewSQL

NewSQL is a class of modern relational database management systems that seek to provide the same scalable performance of NoSQL systems for online transaction processing (read-write) workloads while still maintaining the ACID guarantees of a traditional single-node database system.

34 NoSQL

NoSQL (sometimes expanded to "not only SQL") is a broad class of database management systems that differ from classic relational database management systems (RDBMSes) in some significant ways. These data stores may not require fixed table schemas, usually avoid join operations, and typically scale horizontally.

Academia typically refers to these databases as structured storage, a term that would include classic relational databases as a subset.

MongoDB is an implementation of a NoSQL system.

35 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***.

36 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.

37 Predictive Model Markup Language

The Predictive Model Markup Language (Data Mining Group, 2008) provides such a standard language for representing data mining models. PMML is an XML based standard that is supported, to some extent, by the major commercial data mining vendors and many open source data mining tools.

38 Random Forest

Random forest is a highly versatile machine learning method with numerous applications ranging from marketing to healthcare and insurance. It can be used to model the impact of marketing on customer acquisition, retention, and churn or to predict disease risk and susceptibility in patients.

Random forest is a capable of regression and classification. It can handle a large number of features, and it's helpful for estimating which or your variables are important in the underlying data being modeled.

Random forest is solid choice for nearly any prediction problem (even non-linear ones). It's a relatively new machine learning strategy (it came out of Bell Labs in the 90s) and it can be used for just about anything. It belongs to a larger class of machine learning algorithms called ensemble methods.

Ensemble learning involves the combination of several models to solve a single prediction problem. It works by generating multiple classifiers/models which learn and make predictions independently. Those predictions are then combined into a single (mega) prediction that should be as good or better than the prediction made by any one classifer.

Random forest is a brand of ensemble learning, as it relies on an ensemble of decision trees. More on ensemble learning in Python here: Scikit-Learn docs.

So we know that random forest is an aggregation of other models, but what types of models is it aggregating? As you might have guessed from its name, random forest aggregates Clasification (or Regression) Trees. A decision tree is composed of a series of decisions that can be used to classify an observation in a dataset.

The algorithm to induce a random forest will create a bunch of random decision trees automatically. Since the trees are generated at random, most won't be all that meaningful to learning your classification/regression problem (maybe 99.9% of trees).

39 Real-time Analytics

Real-time analytics is the use of, or the capacity to use, all available enterprise data and resources when they are needed. It consists of dynamic analysis and reporting, based on data entered into a system less than one minute before the actual time of use. Real-time analytics is also known as real-time data analytics, real-time data integration, and real-time intelligence.

Technologies that support real-time analytics include:

- Processing in memory (PIM) – a chip architecture in which the processor is integrated into a memory chip to reduce latency.
- In-database analytics – a technology that allows data processing to be conducted within the database by building analytic logic into the database itself.
- Data warehouse appliances – combination hardware and software products designed specifically for analytical processing. An appliance allows the purchaser to deploy a high-performance data warehouse right out of the box.
- In-memory analytics – an approach to querying data when it resides in random access memory (RAM), as opposed to querying data that is stored on physical disks.
- Massively parallel programming (MPP) – the coordinated processing of a program by multiple processors that work on different parts of the program, with each processor using its own operating system and memory.

39.1 Applications of real-time analytics

In CRM (customer relations management), real-time analytics can provide up-to-the-minute information about an enterprise's customers and present it so that better and quicker business decisions can be made – perhaps even within the time span of a customer interaction. Real-time analytics can support instant refreshes to corporate dashboards to reflect business changes throughout the day. In a data warehouse context, real-time analytics supports unpredictable, ad hoc queries against large data sets. Another application is in scientific analysis such as the tracking of a hurricane's path, intensity, and wind field, with the intent of predicting these parameters hours or days in advance.

The adjective real-time refers to a level of computer responsiveness that a user senses as immediate or nearly immediate, or that enables a computer to keep up with some external process (for example, to present visualizations of Web site activity as it constantly changes).

40 Real-time stream processing

The ability to do real-time analytics on multiple data streams using StreamSQL has been available since 2003. Up until now, StreamSQL has only been able to penetrate some

relatively small niche markets in the financial services, surveillance and telecommunications network monitoring areas. However, with the burgeoning interest in all things Big Data, StreamSQL is bound to get more attention and find more market opportunities.

StreamSQL is an outgrowth of an area of computational research called ***Complex Event Processing (CEP)***, a technology for low-latency processing of real-world event data. Both IBM, with InfoSphere Streams, and StreamBase Systems Inc. have products in this space.

41 Scalable Database

While Hadoop has grabbed most of the headlines because of its ability to process unstructured data in a data warehouse-like environment, there's much more going on in the Big Data analytics space.

Structured data is also getting lots of attention. A vibrant and rapidly growing community surrounds NoSQL, an open source, non-relational, distributed and horizontally scalable collection of database structures that address the need for a web-scale database designed for high-traffic websites and streaming media. Document-oriented implementations available include MongoDB (as in "humongous" DB) and Terrastore.

Another analytics-oriented database emanating from the open source community is SciDB which is being developed for use cases that include environmental observation and monitoring, radio astronomy and seismology, among others.

Traditional data warehouse vendors aren't standing idly by. Oracle Corp. is building its "next-generation" big data platforms that will leverage its analytical platform and in-memory computing for real-time information delivery. Teradata Corp. recently acquired Aster Data Systems Inc. to add Aster Data's SQL-MapReduce implementation to its product portfolio.

42 Sharding

Horizontal partitioning is a database design principle whereby rows of a database table are held separately, rather than splitting by columns (which is what Normalization and Vertical Partitioning do, to differing extents). Each partition forms part of a shard, which may in turn be located on a separate database server or physical location.

There are numerous advantages to this partitioning approach. The total number of rows in each table is reduced. This reduces index size, which generally improves search performance. A database shard can be placed on separate hardware, and multiple shards can be placed on multiple machines.

This enables a distribution of the database over a large number of machines, which means that the database performance can be spread out over multiple machines, greatly improving performance. In addition, if the database shard is based on some real-world segmentation of the data (e.g. European customers vs. American customers) then it may be possible to infer the appropriate shard membership easily and automatically, and query only the relevant shard.

Sharding is in practice far more difficult than this. Although it has been done for a long time by hand-coding (especially where rows have an obvious grouping, as per the example above), this is often inflexible. There is a desire to support sharding automatically, both in terms of adding code support for it, and for identifying candidates to be sharded separately.

Where distributed computing is used to separate load between multiple servers (either for performance or reliability reasons) a shard approach may also be useful.

43 SQL

SQL stands for structured Query Language

43.1 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.

44 Tableau

Data Visualization Tool

45 Visual Analytics

Visual Analytics is the science of analytical reasoning supported by interactive visual interfaces. Today, data is produced at an incredible rate and the ability to collect and store the data is increasing at a faster rate than the ability to analyze it. Over the last decades, a large number of automatic data analysis methods have been developed. However, the complex nature of many problems makes it indispensable to include human intelligence at an early stage in the data analysis process. Visual Analytics methods allow decision makers to combine their human ?exibility, creativity, and background knowledge with the enormous storage and processing capacities of today's computers to gain insight into complex problems. Using advanced visual interfaces, humans may directly interact with the data analysis capabilities of today's computer, allowing them to make well-informed decisions in complex situations.

46 Waikato Environment for Knowledge Analysis (WEKA)

- Developes Machine Learning (ML) techniques for solving real world data mining problems
- Collection of visualization tools and algorithms for data analysis and predictive modelling.
- Weka Provides access to SQL databases using JDBC.

47 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 site?s 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 site?s 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.