BCS THE CHARTERED INSTITUTE FOR IT

BCS HIGHER EDUCATION QUALIFICATIONS

BCS Level 5 Diploma in IT

BIG DATA MANAGEMENT

SAMPLE PAPER AND ANSWER POINTERS

Answer **any** FOUR questions out of SIX. All questions carry equal marks.

Time: TWO hours

Answer any Section A questions you attempt in Answer Book A

Answer any Section B questions you attempt in Answer Book B

The marks given in brackets are **indicative** of the weight given to each part of the question.

Calculators are NOT allowed in this examination.

Section A

Answer Section A questions in Answer Book A

A1.

a) Explain how each of the following **THREE** characteristics of Big Data define the differences between Big Data and other data processing:

i) Volume;

ii) Velocity;

iii) Variety.

(12 marks)

b) Describe **FOUR** major sources of machine generated data that might be found in a Big Data processing task.

(13 marks)

A1. Answer Pointers

[Syllabus coverage 1. Big data fundamentals
1.1 Defining characteristics – Laney's V's and later extensions, veracity, valance and value.
1.2 Categories and examples of Big Data sources – Machines, people and organizations]

a)

i)The volume characteristic for big data differs from other traditional data processing in terms of magnitude. Storage values for big data are measured in massive terms – Terabytes, Petabytes, and greater. An explanation could include reference to big data volumes not being amenable to traditional storage and not being feasibly processed in traditional database management systems. Context could be used in the explanation e.g. Facebook database is estimated to add 600TB daily to its storage, YouTube adds 100 PB of data each year. Volume of big data is increasing in orders of magnitude for a range of sources all of which requires storage in forms that are different from other data processing.

(4 marks)

ii) Velocity characteristic of big data refers to the rate of accumulation of data (and not necessarily storage per se). Big data sources are capable of generating vast amounts of data that require very fast processing, so cannot be done by traditional methods such as batch or scheduled processing on traditional platforms. Context could be used to aid the explanation e.g. Scientific applications such as genome project, the square kilometre array project and similar which generate huge volumes of data in a short time (The square kilometre array produces an estimated 750 TB of data per second). The velocity of big data is not capable of being processed in traditional ways (for instance -a massively parallel data base system could only process about 2 seconds worth of the parallel array project).

(4 marks)

iii) Variety is a characteristic of big data that is distinguished from other data in the multitude of different data types that need to be stored and processed. The variety of big data is structural, semantic and media characterised. It can include structured (traditional) semi structured and unstructured data. Data variety can be video, free text, GPS, or RFID, and industrial sensor data. This data has to be allowed for in a big data processing system and usually needs to be integrated for some form of meaningful analysis. This variety requires forms of processing not feasible in traditional systems.

(4 marks)

b)

Machine generated data can come from a range of sources. The principal sources could be selected from real time industrial sensors, environmental, satellite, weather stations, real time tracking software. In health systems from monitoring devices. The list is wide and these major categories can be used for a description. Game events logs – each user move is tracked and stored. Website click logs. Every time user clicks a link the result is stored and might dictate real time adjustment to site experience or stored. Sensor logs, RFID, GPS, point of sales data from retail each product barcode is recorded frequency of sales recorded etc', provides information on selling patterns of products A description selected from these typical sources, a suitable description will include the source and its context in use.

(13 marks)

A2.

a) Describe **THREE** important skillsets that need to be considered in building a Big Data science team.

(9 marks)

b) Give **TWO** examples of the way in which international privacy laws might affect the use of personally identifiable information.

(10 marks)

c) Compare and contrast data privacy and data protection.

(6 marks)

A2. Answer Pointers

[Syllabus coverage 2.Management issues for big data

2.1 Staffing, skillsets, total cost of ownership

2.3 Security, ethical and legal considerations – personal data, internal data, national law]

a) There are a range of important skillsets required in a big data science team. Dependent on the industry different particular (industry specific skills) skilled experience would be

important. A description of such specific skillsets is acceptable. In general, a big data science team would typically include important skillset roles such as: Business analyst, Platform engineers, Data engineers, Algorithm scientists, Production developers, Web analysts. Descriptions of the roles from three of the above or from industry specific big data roles should include detail clearly defining the skillsets. Examples are: Data engineering specialist involved with the extracting source data, cleaning data, and loading tables with data (ETL skills) data engineering skillsets include familiarity with using big data tools such as Pig. The data engineer skills will include data warehouse table design etc'. Algorithm specialists with skills in mathematics and statistics skilled in scientific programming skilled in using algorithm prototyping tools such Azure ML etc'. Web analyst would have skillsets in web analytic toolkits knowledge of browser technologies and the tagging of user interactions gathering user data a big data web analyst skillset will include expertise in consolidating data from attached data gathering devices and internally stored business data. Three descriptions as above or similar examples.

(9 marks)

b) *International privacy laws will affect the way in which personally identifiable information (PII) is used in a variety of ways. Examples are: For business done in the European Union the General Data Protection Regulations (GDPR) govern the type of data and the use of that data. The GDPR restricts the way in which data can be gathered stored and used. For businesses that are involved with big data projects this means that no matter where they are headquartered breaching the regulations opens liabilities to huge financial penalties (up to 20 million euros)so it becomes a legal and financial risk to deal with data from the EU if you do not understand the conditions imposed on it by the privacy laws. International privacy laws will also vary on the interpretation of what constitutes PII. A passport number is PII but an individual's ethnicity is not. In some legal jurisdictions an ip address is PII but in others it is not. This can have implications for collecting and using cross border data. This affects the way in which internal policies on data access and use have to be set up in any business involved with big data projects. A further possible example is in the way in which international privacy laws are amended and cause change to liabilities. The US 'safe harbour' legislation allowed data to be transferred between business in EU and the US for companies in the US which had data governance standards procedures and processes, this was however rescinded following US data breaches, so restricting access until an alternative set of international regulations was agreed. Any two examples based on the above or similar that show an understanding of appropriate ways that dealing with large amounts of diverse data are affected by law.*

(10 marks)

c) *Data privacy is related to the collection, storage, and subsequent use of data – e.g. Cookies on websites tracking behaviours or hidden speed cameras or discrete video recording in public places. Data privacy is concerned with the appropriate use of such technologies to gather information on people. Data protection is concerned with the proper safeguarding of sensitive information on people that a business may have collected, stored, and analysed. This is data that is legally held (Bank details credit details and medical or insurance information) Data protection issues involve making correct decisions on the processing and sending of such data/information across different jurisdictions and complying with the obligations of the jurisdictions. It is also concerned with appropriate restrictions on the rights to access such data.*

(6 marks)

A3.

a) Describe with the aid of a diagram, the Hadoop distributed file system architecture, and explain how it stores data across a cluster of machines.

(10 marks)

b) Explain how the MapReduce pipeline is used for program execution in the Hadoop system.

(15 marks)

A3. Answer Pointers

[Syllabus coverage 4. Architectures for big data

4.1 The Hadoop ecosystem definition – MapReduce, HDFS and Yarn]

a) Drawn and labelled diagrams explaining the basic architecture are acceptable but must convey the basic architectural description as outlined below and aid the explanation.

The HDFS architecture is made up of a master slave relationship with the top node being the NameNode the slaves being DataNodes. The NameNode contains the metadata which describes files on the data store. The metadata will include filename, size, number of blocks, and location of blocks in the DataNodes. The NameNode server handles the incoming filesystem operations Open, closing, renaming files. The Name Node performs basic load balancing between machines in the DataNodes by checking on the state of each entry in the registry. The HDFS model involves splitting up files into default 64MB blocks (can be modified to larger blocks – thus reducing the size of the file metadata) and placing those blocks on DataNodes (HDFS default value is three DataNodes get a single block). The system architecture provides for multiple data blocks to be replicated across several DataNodes thus supporting fault tolerance of any DataNode machine. The NameNode keeps track of the nodes where data blocks are stored. DataNodes handle read/write requests and DataNodes create, delete and replicate blocks amongst the machine storage (disk drives, SSD etc'.) following instructions from the NameNode. The architecture allows for a secondary NameNode to protect against failure of the name server. The Hadoop cluster continuously monitors the DataNodes to ensure that the node machine is still functioning correctly. The Name Node is sent regular 'heartbeat' signals (default is 3 seconds) to indicate that they are available. The NameNode will wait for up to 10 heartbeats In the absence of a heartbeat in the 30 sec interval the NameNode 'assumes' that the data is lost and the NameNode replicates the lost data blocks to other DataNodes. The secondary NameNode will communicate with the primary NameNode to regularly update copies of the registry.

(10 marks)

b) This question can be answered by reference to a sketch/diagram a suitable form is in fig 1 below. Answers can also be made by reference to Java code snippets or pseudo code for candidates who have a background in software. The explanation will need to acknowledge the process steps as outlined below.

MapReduce facilitates the execution of a program (Java, Python, or similar programs) in the Hadoop system following a master slave model using a Job Tracker and a Task Tracker. The job tracker resides within the NameNode server and the task tracker resides within the

DataNode machines. The job tracker is responsible for program execution across the whole Hadoop cluster and controls the task tracker at each of the node machines. The task tracker maintains contact with job tracker in the same heartbeat cycle as in the NameNode DataNode 3 sec cycle. The job tracker also performs load balancing of a program execution across these data nodes. The data file is split into 64MB blocks and each block would be assigned to three separate data nodes. MapReduce allows the execution program to be sent to each part of the data in the separate blocks allowing parallel execution of the program thus facilitating a high velocity requirement of big data processing. MapReduce makes use of two techniques Map and Reduce. The map function takes input as key-value pairs and maps them according to the mapping function to an output which is also in the form of key-value pairs. The map function works over the split data nodes and thus processes the data in a parallel fashion. All outputs belonging to the same key are processed by a further step in the MapReduce pipeline referred to as shuffle. The shuffle process takes the outputs from any mapper that has finished across any of the nodes thus works in parallel to provide output to the reducer. The reducer performs aggregation on key-value pairs. The reducer provides output on to the HDFS filesystem.

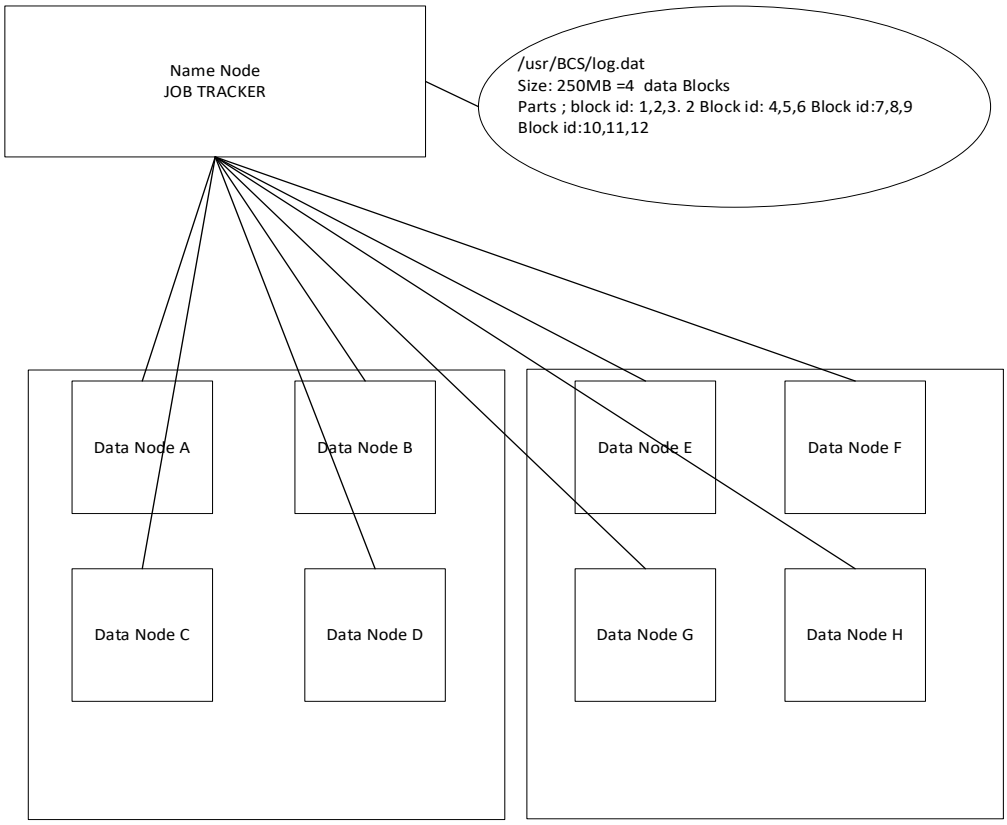


Fig 1.

(15 marks)

Section B

Answer Section B questions in Answer Book B

B4.

a) It is widely acknowledged that Relational Database Management Systems cannot always support the rapid growth in data storage. Give brief comments on why this view is often taken.

(5 marks)

b) Explain why the use of a Hierarchical Data File Store (HDFS) in products like Hadoop can offer significant advantages when processing Big Data.

(10 marks)

c) One method of storing Big Data is within key-value pairs.
Explain key-value pairs and give two scenarios in which the use of key-value pairs would be applicable.

(10 marks)

B4. Answer Pointers

[Syllabus coverage 3. Storage, data and programming interface models used for big data 3.1 Data – key-value, relational, extensible and document]

a) The view is based on acknowledgement that the traditional approach to storage is contingent on a structure that is based on a schema and undertakes processing of data in the form of batches with possibly day or intra-day schedules. These processes have a latency overhead where the limit is quickly reached when dealing with the huge volumes and velocity of big data. The cycle of processing and analysis in a traditional system would make up to date dynamic analysis difficult. The traditional approach also lacks an efficient way to deal with the variety of data types in big data and the physical storage capacities of huge volumes of raw data.

(5 marks)

b) The Hadoop Distributed File System (HDFS) offers the advantage of processing locally stored data in a parallel topography. Hadoop provides a framework for a file system, cluster administration, and data processing that work together. The file system within Hadoop is the basic structure that stores data across multiple computers. A significant advantage is that the file system allows a user to connect to a node which has access to data files distributed across multiple data nodes although the user only 'sees' a seamless access to the data. A further advantage is that the file system is based on multiple machines that will expect (tolerate) failure. The hierarchy of the HDFS file system consists of a top-level name node that holds metadata which identifies a file and its location across data nodes. The data nodes do the creation, deletion, replication, and handle reads/writes of file data according to instructions from the name node. This offers the very significant advantage of being able to have parallel operations on extremely large volumes of data. The name node in the

hierarchy also handles the replication of data blocks across data nodes so that failure of a single data node does not loose data which is a significant advantage when dealing with extremely large volumes data. HDFS also supports large data blocks more suited to big data processing (64 MB blocks) the large data blocks are better optimised for handling the typical large files in big data applications. The HDFS file system is also capable of having the name node fail as it can utilise a secondary name node server, which essentially provides a mirror this offers the advantage of greater fault tolerance.

(10 marks)

c) Key-value pairs consist of a set of two linked items of data. The key is a unique identifier to a location that contains data. The key is the sole search criterion in the database that will locate (retrieve) the data item. Key-value pairs (KVP) are applicable in a situation where the data has a large variety of types. In databases such as NoSQL which are not based on a schema. KVP's such as those used in NoSQL databases are applicable where the data is unstructured such as video, audio files or log files. The key links to the value which is the location where the file resides. KVP's are also applicable in the situation where the growth in data storage is anticipated to require a large degree of horizontal scaling allowing scope for greater distribution. The high velocity of data associated with big data is better handled with KVP's in shopping cart applications a key-value based on user id is paired with shopping cart for various items stored. The data base will be accessed across sessions by using a key-value pair with the user ID quickly able to access the stored cart file. In this scenario KVP's offer the most efficient means of data retrieval.

(10 marks)

B5.

a) Gartner's ascendency model provides **FOUR** levels of analytics for information. Explain the type of analytics required at **EACH** level, and state the most appropriate tools to use at each level.

(15 marks)

b) Give an example of an Artificial Intelligence method that uses an unsupervised learning technique and describe the value that this type of method provides to the analyst and business.

(10 marks)

B5. Answer Pointers

[Syllabus coverage 6. Tools for analysis

6.1 Gartner's ascendency model – descriptive, diagnostic, predictive, and prescriptive

6.3 Use of Artificial intelligence in Big Data applications]

a) The Gartner ascendancy model grades the value of analytic approaches against the difficulty and level of sophistication required to use them.

Level 1 – Descriptive analytics has lowest value to the business the concentration is on past or archived data based on hindsight. It is largely descriptive statistics looking at mean, variance, simple distribution etc. looks at questions that relate to what happened. An example would be looking at retail sales of product A over past week/month. Or looking at downtime of network nodes or machine failures over preceding days and measuring average, variation or similar. Tools of relevance would be Excel, SPSS, and STATA. The database infrastructure typically table based.

Level 2 – Diagnostic analytics has greater value to the organisation and involves greater difficulty in providing skills and tools to perform diagnostics. Diagnostic analytics seek to answer the why of an event. Diagnostic analytics will measure historic/descriptive data against current data. The techniques will be drill-down, looking at statistical correlations and exploring probability models together with pattern matching. The toolkits are similar to level 1 but the skillsets required are more sophisticated. The database is likely to a distributed system using an RDBMS with SQL and proprietary tools with business information visualisation tools. The staffing skillsets will be based on those with ability to write ad hoc queries and produce web based reports and undertake some statistical analysis.

Level 3. Predictive analytics deals with future events. Predicting demand for a product based on statistical analysis or predicting queue lengths for distribution warehouses etc. this has significant value to an organisation and is becoming more difficult to achieve. To provide predictive analytics requires that the first two levels are being well done and data is being archived/stored appropriately. Typically predictive analytics tools will deal with programming with Java, Python, R, and SAS. Modelling tools will be used based on statistics and mathematical probabilities. The databases will be larger and capable of supporting toolsets such as RapidMiner and MATLAB. The processing engines will tend to be powerful math capable machines. At this level big data architectures become appropriate and for larger businesses the use of parallel or Hadoop type architectures

may be needed for predictive work based on large data sets and predicting events far ahead.

Level 4Is the apex of the Gartner tree and provides for Prescriptive analytics. This is of maximum value to an organisation and is the most difficult level to achieve and maintain. Prescriptive analytics seeks to determine what future actions to take and can be based on taking optimal actions to maximise a gain or to minimize loss. This type of analytics deals with risk and its avoidance and produces results which provide corporate management with direction of travel. Prescriptive analytics will make use of systems that provide feedback (neural networks) to modulate models that measure performance of the prescriptive models.

An example might be Netflix, Amazon or Spotify in predicting the next user choice. The tools are similar to level 3 with the extra need for highly skilled staff with domain experience and high-level computer science/mathematical specialisms. Large scale systems will make significant use of cloud services and large warehouses/clusters using Hadoop based tools Spark and in house designed/developed analytic engines based on AI architectures.

(15 marks)

b) *The clustering method is a common method of unsupervised learning. Unsupervised learning using clustering involves the examination of data to detect a regular pattern and to place data into sets that are based on commonalities in the patterns of the attribute. When a new item of data is read the data is checked for a cluster to be assigned to if it does not fit into the regular pattern of clusters or is an outlier in a particular cluster then it is identified as an anomaly. Similarly, if a data item fits well within a cluster it can be readily identified (to some degree as being part of that cluster attribute) The benefit unsupervised learning has for the analyst is that there is no straight mapping of input data attributes to the target attribute. There is therefore no effort expended in time consuming training of input data sets or the need to have specific domain expertise in mapping input attributes to target attributes as would be in using a supervised or other method. The benefit to the business is in having a method to detect customer behaviours and potentially in some business applications detect fraud such as in credit card thefts. When a credit card transaction does not match a cluster for that type of customer the business can be alerted to it. The method can be used by business across a range of application areas and is not domain tied. The method can also have application in matching an input data set to clusters where the assignment indicates that the input is of the type represented by that particular cluster set, an example is detecting disease based on a set of attributes of a patient. If a patient has a set of data attributes that puts the individual as a member of a cluster with a probability of stroke rather than an other cluster then the method is useful to help the patient to be aware of potential risk.*

(10 marks)

B6.

a) In Big Data computing, proprietary distributed file systems which are optimised for maximising efficiency for a particular set of applications can be used. The Google File System (GFS) cluster is an example of such a distributed file system.

With the aid of a diagram, explain the architecture of the GFS cluster and suggest **TWO** ways in which this architecture is optimal for some type of applications.

(15 marks)

b) In a Big Data initiative, you might consider the use of Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) from a public cloud provider. Explain what these services provide and give **TWO** advantages and **TWO** disadvantages you might expect from each service.

(10 marks)

B6. Answer Pointers

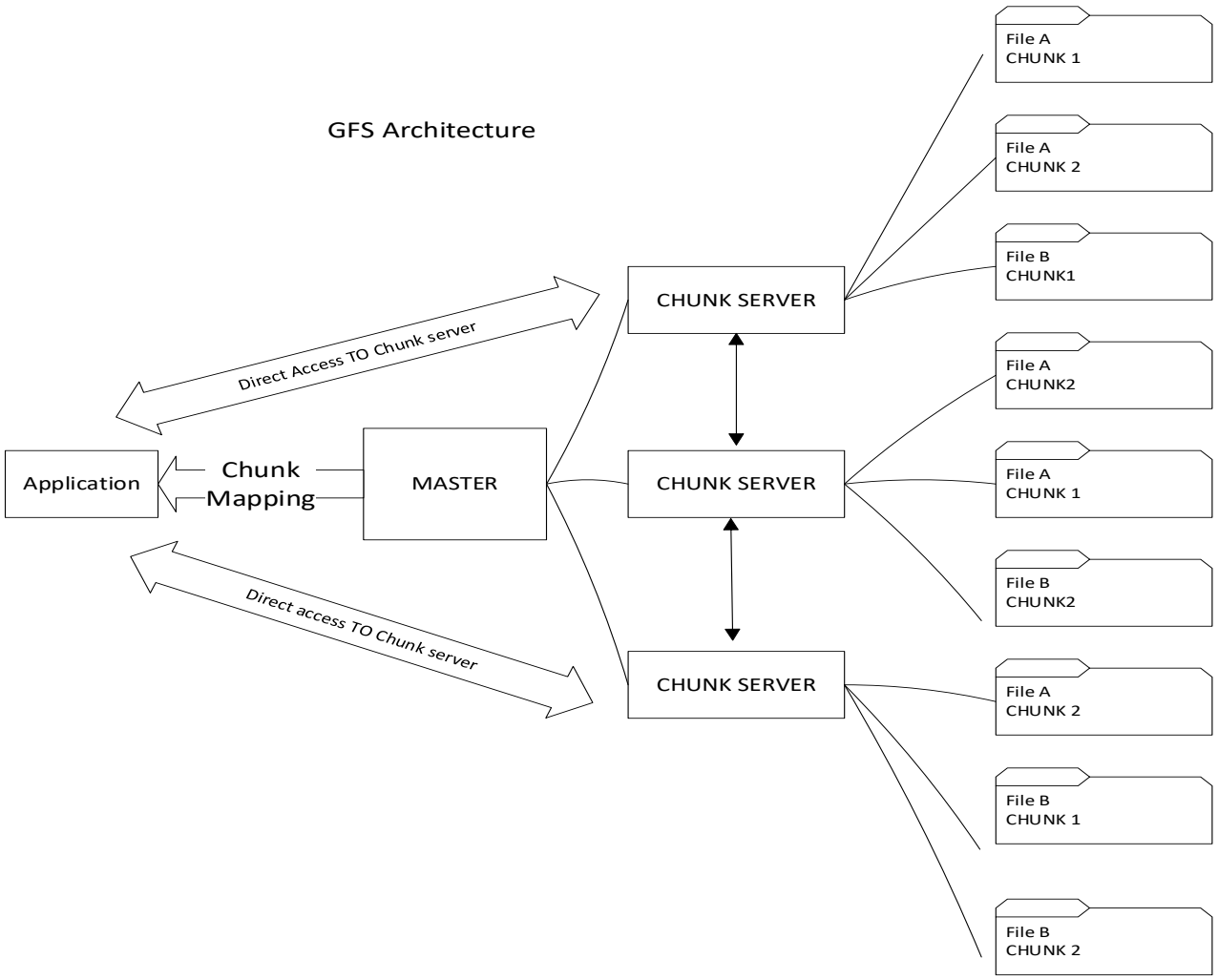
[Syllabus coverage 5. Big data infrastructures

5.1 Networks, cloud based, in-house v outsourced

5.2 Distributed file systems]

a) A suitable diagram will show the basic structure of the GFS architecture similar to below or other variations showing the essential features of master/chunks and associations.

The GFS cluster is made up of a single master (which can be shadowed by a replica for resilience) and multiple chunkservers. Each chunkserver is accessible to a number of client machines. These machines are typically low cost running a user server process in Linux. Files are split into chunks with each chunk of fixed size with a unique ID. The ID is assigned on creation by the master. Chunk servers are stored on local disks as Linux files. Each chunk ca be replicated multiple times – the default is three. The master controls all system metadata including the mapping from files to chunks and chunk location. Google type applications – such as storing indexing and reading massive volumes of fixed data with infrequent appending to files are processed optimally by the GFS cluster in several possible ways. The scalability concern is met with the use of massive scalability on cheap commodity hardware, so space/storage efficiency is not important. Traditional file systems are designed to cope with a range of file sizes to accommodate the typical data file range of size from bytes to Kb. For Google type applications the file size is invariably large so the 64MB chunk size is optimal for storing single large size files. Google applications require almost exclusively to write once read many access (immutable) with only an occasional need to append the file. The GFS architecture is therefore more capable in massive streaming of reads from the files, rather tan a sub optimal traditional system of read write update delete etc.



(15 marks)

b) *IaaS is the provision of hardware resources from the cloud. The typical service contains access via rental of servers and disk storage hardware. Networking, pipeline infrastructure and can include dedicated data centres at a physical location. The advantages of IaaS is the ability to instantly provision for the required hardware as part of an upscaling requirement or to cease rental of hardware such as servers or disk space as part of redundancy in you application requirements. There is no need to provide capital expenditure for a large-scale big data project but the rental model of IaaS provides for operational expenditure on a week to week or shorter basis. It is also an advantage that you can readily make use of up to date hardware components and not incur the cost discarding previous generation hardware. The infrastructure cost of providing disaster recovery is also eliminated in renting cloud infrastructure. Possible disadvantages include security issues with having data hosted by a secondary organisation and complying with international jurisdictions. Providing you services to financial institutions might require stringent audit conditions on traceability of systems and this might be an issue with some cloud providers. Rental of instant infrastructure is not necessarily cheaper in the long run. A mature big data application might well be better hosted in-house when it is stable in scale.*

PaaS refers to the renting of a complete infrastructure and development environment for a business. It includes development tools operating systems programming toolkits and business information reporting and analytic software packages or development environments. In addition, elements from the IaaS will be used to host these platforms. The advantages include having immediate access to big data development tools without requiring expert staff to maintain and install them. PaaS also offer the ability to manage distributed teams making use of geographically dispersed team members using the PaaS infrastructure to network the platforms. The complete development lifecycle can be managed through renting a platform service. The disadvantages are cost these services are not necessarily cheap. The provider might not offer the desired toolkit for a project that matches your team skills. The issue of security is similar to renting infrastructure. An issue can be that PaaS vendors might use open source platforms and development environments which might have restrictions on subsequent deployment or present security anomalies.

(10 marks)