NAME : EDWARD **Course : Big Data Processing**
NIM : 2201741971 **Course Code : COMP6579**
CLASS : LB-08 **Faculty / Department : School of Computer Science**

1. **(10%)** Big Data has been successfully applied in many different domains such as Web, Financial, Healthcare, Environment, Logistic and Transportation, Industry, and Retail. Explore the possibilities Big Data implementation in Higher Education, give a use case of Big Data implementation in a University.

Answer:

Big data itself suggest lots of data. Data means money, that's why every domain is adopting it. **Big data means collecting large amount of data and processing it which gives meaningful insights from it**.

Big Data implementation in Higher Education must be coupled with the business processes to **enhance administrative activities and assist universities in the procurement of creative programs for students**.

The student retention rate may be increased if an early warning system cantered on a Big Data analysis is established and the implementation is properly deployed. Institutions of higher education will **use analytics to optimize multiple operations, including enrollment, education programs, student participation, student support, financial assistance management, and other academic learning and organizational functions**.

For example, in the end of every semester University management takes survey from the all the students about faculty. **These data can be used to determine which faculty is good at which area and which areas need to be improved** i.e. clarifying the doubts, spending enough time to understand the problems of the students.

It is important **to use the more a considerable volume of the student statistics and the data which campuses now have in the different storage facilities to make more knowledgeable decisions**.

Another Example, by collecting the exam marks of each student. By applying big data to the marks we can know which subject he is good at and which subject he/she needs improvement based on the data of interests of the students we can provide customized programs to the students.

Only a limited percentage of universities using Big Data Analytics, there would be a strong demand for guidelines and best practices as more universities have adopted these types of systems.

Thus it is **important for the effective and successful delivery of the initiatives that the universities are all aligned to the initiative and prepared to help its progress**. In addition, guidelines would have to be **established on just how much the data universities should gather and, in turn, what applications will be made of them**.

NAME : EDWARD **Course : Big Data Processing**
NIM : 2201741971 **Course Code : COMP6579**
CLASS : LB-08 **Faculty / Department : School of Computer Science**

2. **(10%)** There are four types of analytics in Big Data i.e. Descriptive, Diagnostic, Predictive, and Prescriptive analysis. What do you know about those type of analytics, give an example for each!

Answer:

Descriptive

➔ This analytics helps to define what happened based on incoming data, using real-time dashboard and email reports.

Example:

Extracting current 12th board rank holder for a particular year.

Diagnostic

➔ This analytics helps to determine why something has happened by using historical data. It gives in depth information for any problem.

Example:

In hospital, old prescribed medications for a patient or College data of old already graduated students.

Predictive

➔ It defines what is likely to happen. It uses findings of descriptive and diagnostic analytics to detect output.

Example:

Weather Forecasting, it basically determines the weather information before the time. But accuracy depends on the stability of the situation.

Prescriptive

➔ It reveals what actions to be taken. Prescriptive analysis uses advanced tools and technologies for example machine learning, business rules and algorithm.

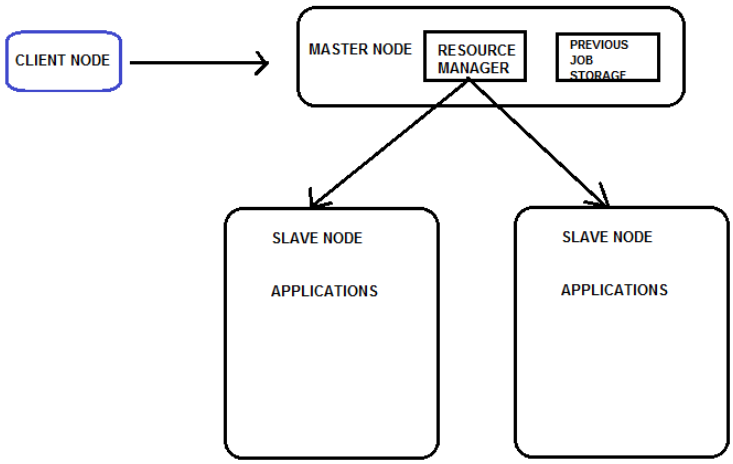
Example:

In Healthcare, its analysis data on patients, treatment, cost and show reimbursement on the cost of checkup whether it is increasing or decreasing or holds same so according to that actual cost will be defined.

NAME : EDWARD **Course : Big Data Processing**
NIM : 2201741971 **Course Code : COMP6579**
CLASS : LB-08 **Faculty / Department : School of Computer Science**

3. **(10%)** Distributed Computing and Distributed File System are the core technologies behind Big Data. Explain how Hadoop Distributed File System (HDFS) works in the Big Data Cluster.

Answer:



Master nodes or Name nodes are the main component in an HDFS that stores the large data sets in the system. These observe the main activities like calculations or processing of data using MapReduce.

The Worker nodes refer to the virtual workstations in the cluster that **perform the processing's or computations in the parallel environment**. These **worker nodes accomplish the task using the Data node and Task Tracker service utilities**. Further, these nodes **obtain data processing or storing instructions from the master nodes**. Data nodes can read-write on files, perform any modification as the instructions given by the client node.

Another element in the cluster, the client nodes include the data sets into the big data cluster. They also get the job finishing reports.

Thus, **HDFS receives data along with client requests and creates segments of these data sets**. The **data are transferred to different operational nodes in blocks**. Therefore, a highly efficient processing interface is built.

NAME : EDWARD **Course : Big Data Processing**
NIM : 2201741971 **Course Code : COMP6579**
CLASS : LB-08 **Faculty / Department : School of Computer Science**

4. **(15%)** There are two programming models in Big Data Ecosystem I.e. Map Reduce and Spark, what are the similarities and differences between them? Why Spark outperform Map Reduce in the execution time?

Answer:

Difference between Spark and Map Reduce:

No.	Spark	Map Reduce
1	Process Data in Random Access Memory.	Persists data back to the disc after a map or reduce.
2	Needs lots of memory.	Does Not Need lots of memory.
3	Faster in the Execution Time.	Not so Fast in the Execution Time.
4	Can be used both for batch processing and real time processing.	Only used for batch processing.
5	Can be written in java, Scala, python, and R.	Supports Only java programming language.
6	Have its own Scheduler.	Dependent on External Scheduler.
7	Support Duplicate Elimination.	Does not Support Duplicate Elimination.
8	Support through Spark SQL.	Support through Hive Query Language.
9	Easy to write and debug codes.	Difficult to write and debug codes.
10	Security Features getting evolving and mature.	More Secured compared to Spark.

Similarities between Spark and Map Reduce:

- Both spark and map reduce can use commodity servers and run on cloud.
- They both have similar hardware requirement.
- They both can integrated with the same data sources and files formats that is spark compatibility with various data types and data sources is the same as map reduce.
- They both have good failure tolerance .
- They both support Batch Processing.
- They both support Java Programming Language.
- They both support Machine Learning.
- They both have scheduler.

Spark outperform map reduce in the execution time **because Spark shows much stronger scalability as the number of cores increases.**

one reason is that spark keeps RDDs in memory which reduce the amount of data to be materialized and **other reason is that per iteration Spark has no framework overhead such as tear down as map reduce**.

NAME : EDWARD **Course : Big Data Processing**
NIM : 2201741971 **Course Code : COMP6579**
CLASS : LB-08 **Faculty / Department : School of Computer Science**

5. **(10%)** Cloud computing already became the most popular infrastructure for data center, including for Big Data Cluster. We could prepare a Big Data Cluster with required number of nodes just in several minutes. **Explain what are the benefits and drawbacks of Big Data in the cloud**?

Answer:

Benefits of Big data in the cloud:

a) Require zero capital expenditure.

➔ It is very costly to keep big data and store big data it need high investments , resources to keep and process data but as-a-service models have allowed companies to practically eliminate its biggest capital expenses by shifting these into operating expenditure column so when need data or when need to set up database servers, would not to make any investment.

b) Enables faster scalability.

➔ Large amount of data requires more clouds provide not only readily available infrastructure but also provide the ability to scale infrastructure very quickly so can manage data.

c) Lowers the cost of analytics.

➔ Mining data in cloud has made the analytics process less costly you can save your cost related to system maintenance and upgrades energy consumption.

Drawbacks of big data in the cloud:

a) Less control over security

➔ Large data set contain lots of personal information such as individual address credit card information and many more personal things while security should not be hindrance to migrating to the cloud you will have less direct control over data to overcome this problem.

b) Less control over compliance

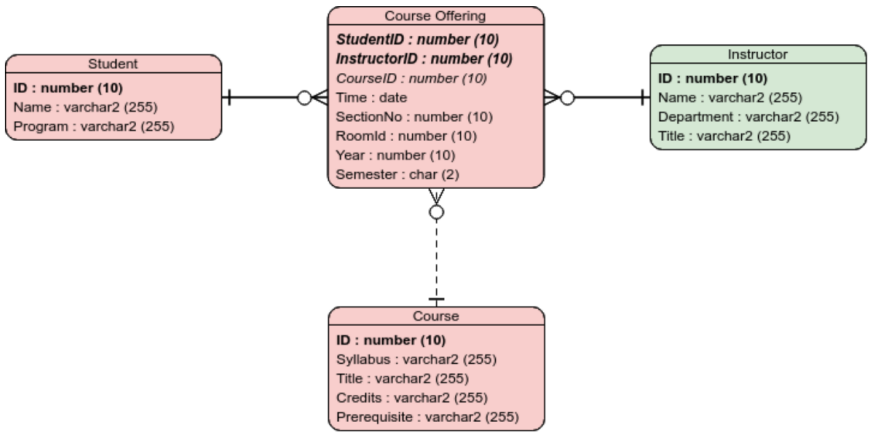
➔ When move data to the cloud make sure that what data is stored and where the data is going to use because we have less control over compliance.

c) Network dependency issue

➔ Easy connectivity of data in the cloud is sometime a big problem because the availability of data is highly reliant on network connection this depend on internet while transferring big data will get interruption in transferring the data.

NAME : EDWARD **Course : Big Data Processing**
NIM : 2201741971 **Course Code : COMP6579**
CLASS : LB-08 **Faculty / Department : School of Computer Science**

6. **(20%)** One of the benefits from Graph Database is significantly faster than traditional Relational Database System (RDBMS), especially for a database with many relations. **Create a graph database of the following RDBMS tables with 3 records for each table**. The Entity Relationship Diagram **represents an Online Course Database**. Use a **graph diagram that consists of vertices and edges to represents the graph database**. **Explain why a database query would be executed significantly faster in the Graph Database than the RDBMS** based on the example database.



Answer:

a) Record 1:

CREATE (s:Student {id:1, name:"Edward", program:"Computer Science"})-
[e:COURSE_ENROLLED]->(c:CourseOffering {courseId:1, time:1030, section:20, room:328, year:2020, semester:"5"})
<-[t:TEACHES]-(i:Instructor {id:10, name:"Fepri Putra", department:"Big Data", title:"Lecturer"})

CREATE (cs:Course {id:22, syllabus:"AB", title:"Big Data Processing", credits:4, prerequisites:"None"})-[:OFFERED]->(c)

b) Record 2:

CREATE (s:Student {id:2, name:"Fiona", program:"Computer Science"})-
[e:COURSE_ENROLLED]->(c:CourseOffering {courseId:2, time:1130, section:21, room:505, year:2020, semester:"3"})
<-[t:TEACHES]-(i:Instructor {id:11, name:"Diaz Santika", department:"Artificial Intellegence", title:"Lecturer"})

CREATE (cs:Course {id:25, syllabus:"AA", title:"Computer Vision", credits:4, prerequisites:22})-[:OFFERED]->(c)

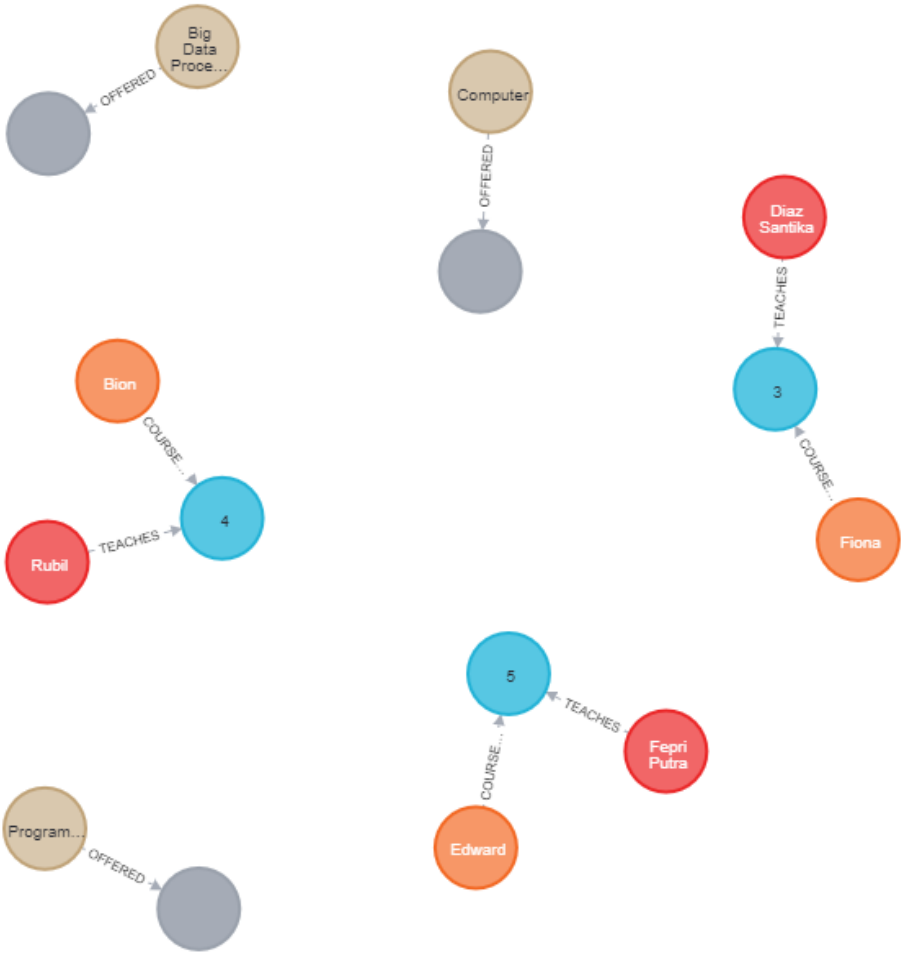
NAME : EDWARD **Course : Big Data Processing**
NIM : 2201741971 **Course Code : COMP6579**
CLASS : LB-08 **Faculty / Department : School of Computer Science**

c) Record 3:

CREATE (s:Student {id:3, name:"Bion", program:"Computer Science"})-
[e:COURSE_ENROLLED]->(c:CourseOffering {courseId:3, time:1230, section:29, room:45, year:2020, semester:"4"})
<-[t:TEACHES]-(i:Instructor {id:26, name:"Rubil", department:"Cyber", title:"Lecturer"})

CREATE (cs:Course {id:28, syllabus:"AC", title:"Programming Language Concepts", credits:2, prerequisites:22})-[:OFFERED]->(c)

Graph Database:



Each node (entity or attribute) in the graph database model directly and physically contains a list of relationship records that represent the relationships to other nodes.

These relationship records are organized by type and direction and may hold additional attributes. Whenever run the equivalent of a JOIN operation, the graph database uses this list, directly accessing the connected nodes and eliminating the need for expensive search-and-match computations.

NAME : EDWARD **Course : Big Data Processing**
NIM : 2201741971 **Course Code : COMP6579**
CLASS : LB-08 **Faculty / Department : School of Computer Science**

Case study:

(25%) Suggest a Big Data Architecture for a company that provides a flight information service to their customers (for example https://flightstats.com/). Read more about flightstats here:

- https://onemileatatime.com/best-website-for-flight-status/
- https://www.dropbox.com/s/yqf2u4nmi5sbya2/Session%2001-03%20Example%20-%20Flight%20data%20management.pdf?dl=0

Answer:

a) Association rules:

This technique is used for identifying interesting correlations between variables.

For Example: A Flight Company chain places product such as Food and Drink next to each other to increase sales.

b) Classification:

It is a **technique in which correctly identified past data is used to identify a new observation**. The past data is called training data and the new observation is called test data.

For Example: Flight classification have numerous Flight labeled as Regular, Business, etc. New classification model made will take care of putting in one of the flight classes.

c) Regression:

This technique **describes how the value of dependent variable changes when the independent variable is changes**. This is mostly used with continues data.

For Example: Predicting the price of the ticket (dependent) given the classification, whether it was recently regular, business. , garage, etc. (independent variables)

d) Clustering:

This is unsupervised technique. It **identifies similar data and groups them into clusters**. Data points belonging to same cluster are like each other whereas data points belonging to different clusters are dissimilar

For Example: customer segmentation based on similar purchase patterns.

e) Anomaly Detection:

It is a **technique in which the data points that stand out in the dataset and do not depict normal behavior are identified**. Such data points are called as outliers. Such model can be either supervised or unsupervised.

For Example: Identifying service quality issues such as pricing glitch on the website.