Project Final Report Requirements

Your team needs to compile a final report that integrates the work documented in the Proposal, and the Phase I, II, and III reports. Please follow these guidelines:

· The final report must be a separate well-written document that uses content from the proposal, and the Phase I, II, and III reports in order to tell a well-structured and clearly-communicated story. (Firmly keep in mind that TIM 158 is meant to satisfy both the comprehensive and the Disciplinary Communication (DC) requirements for the TIM Major.)

· The report should have 5 major sections:

- Introduction (which clearly states the intent and content of the report).
- Traditional Data Center Architecture for Company X (X= your company's name).
- Software-Defined Infrastructure (SDI) for Company X: Virtualization and Cloud Computing.
- Machine Learning Problem solved by the team within the context of the traditional DC and/or SDI.
- Conclusion (which should summarize the work done, the key results and recommendations, and draw meaningful conclusions).

· Each section should be properly sub-divided into sub-sections with the appropriate headings.

· In addition, your report must have a table of contents, and a 1-page meaningful executive summary (ES) at the beginning. This ES should summarize all the key results from the 5 sections of the report.

· The final report is due on one of three dates: Thursday 8/06/17 (in class); Tuesday 8/11/17 (in E2, Room 561, by 5PM); and Thursday 8/13/17 (in E2, Room 561, by 5PM).

TIM 158 Project: Developing IT Infrastructure & Machine Learning Application for Netflix

Group 13
Craig Standley
Cole Teza
Jerzy Barbato
Miguel Calleja
Robert Fazio

Table Of Contents

Executive Summary

Over the past ten weeks, our team has conducted analysis of Netflix's core IT processes in an attempt to define a machine-learning-based software application. This application will be implemented to improve an existing business function within the company. Examining the role of IT in Netflix, we discovered the effectiveness of their recommendation algorithm. With close to 70 per cent of the choices viewers make about what content to watch based on its recommendation, we sought to design an application that would improve or assist this algorithm. This analysis led us to develop a classifier for Netflix to determine whether a movie will get a high recommendation or not, based on its critical rating. This report will outline the ML application we have built, the software and hardware that are used to support it, and the detailed analytics and processes that go hand in hand with our implementation.

In order to efficiently offer their recommendation service to members, Netflix needed a tiered client server architecture for their software and hardware. This is critical as it allows Netflix's application on the client end to communicate with the Netflix web server to recommend content online or allow for other functionality. We identified that a 4 tiered architecture would meet the requirements for the classifier application.

We realized that Netflix had to move away from vertically scaled single points of failure, like relational databases in it's datacenter architecture, and instead went towards highly reliable, horizontally scalable, distributed systems in the cloud. Elasticity of the cloud allows Netflix to add thousands of virtual servers and petabytes of storage within minutes, saving on cost and resources. Leveraging multiple AWS cloud regions, spread all over the world, enables us to dynamically shift around and expand our global infrastructure capacity, creating a better and more enjoyable streaming experience for Netflix members wherever they are.

Our team identified, that a lot of Netflix's competitive edge over other competitors in the industry is due to the effectiveness of its Cine-Match Algorithm. Recommendations is a powerful tool for data mining which helps Netflix determine Genres, Anchoring, Movie Fads, and Rating. By using the information gathered above Netflix can more accurately recommend shows and movies to consumers which helps keep consumer retention and subscription rates up. Developing on this idea we designed a classifier type machine learning application. The objective for this classifier was to train it to categorize movies with a "GOOD" or "Bad" classifier tag based on IMDB ratings, so that members could more easily identify movies regarded highly by critics. This also allows users to sort movies by these classification tags, if for example they wish to ironically watch a "Bad" movie. This classifier was implemented using a Naive Bayes algorithm and was able to be trained to produce a classification accuracy of about 80 percent.

Introduction

Company Selection

Company: Netflix

Company Info

Vision:

"To be the best global entertainment distribution service
Licensing entertainment content around the world
Creating markets that are accessible to filmmakers
Helping content creators around the world to find a global audience"[4]

Mission:

"Our core strategy is to grow our streaming subscription business domestically and globally. We are continuously improving the customer experience, with a focus on expanding our streaming content, enhancing our user interface and extending our streaming service to even more Internet-connected devices, while staying within the parameters of our consolidated net income and operating segment contribution profit targets."[4]

Products:

Streaming Media
Video on Demand

Services:

Film Production
Film Distribution
Television Production

Customers:

Household consumer

Financial Performance:

Netflix Financials	2016				2017
	Q1	Q2	Q3	Q4	Q1
Net Income	$27,658	$40,755	$51,577	$66,748	$178,222
Return on Investment	1.95%	2.09%	2.32%	1.95%	
Compound Annual Growth Rate 2015-2016	27.06%				

Figure 4.
Consolidated Financial numbers from the official Netflix bank sheet and income statement[3]

Competitive Strategy:

The competitive strategy for Netflix is a focus strategy. Netflix is determined to provide a wide array of different genres of movies and television shows. Netflix also funds and creates "Netflix Originals" or shows and movies that can only be found on Netflix. In all Netflix does not worry too much about the cost of creativity as much as it worries about having a large range of products to serve its customers.

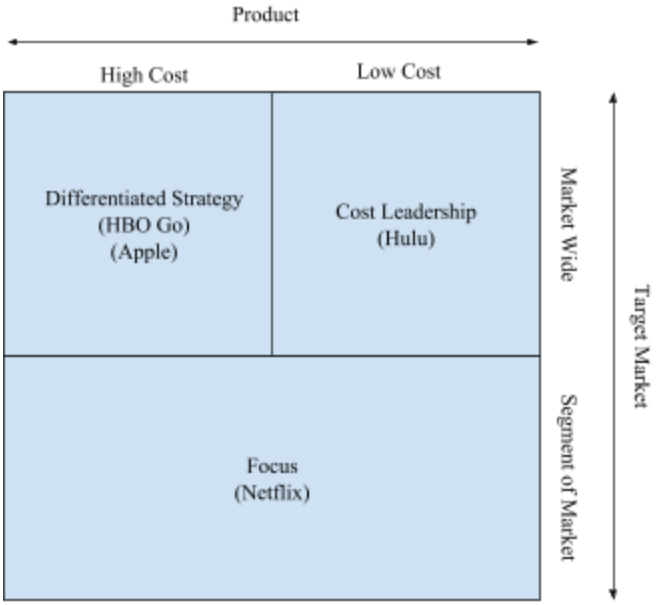


Figure 5. Competitive Strategy for Netflix

2) Information Technology

What is IT's role within the company?	The purpose of IT within Netflix is to make accurate suggestions to subscribers to maximize the number of movies rented by each customer. It also uses AWS for cloud data to store its user info and movie catalog.
How does the company use IT to support or enable its business processes and competitive strategies?	A lot of netflix's competitive edge over other competitors (e.g. Hulu Plus, Amazon Video, HBO) is due to its Cine-Match Algorithm, which gives users movie and television show recommendations. Recommendations is a powerful tool for data mining which helps Netflix determine Film Quality, Genres & Movie Elements, Anchoring, Movie Fads, and Rating. By using the information gathered above Netflix can more accurately recommend shows and movies to consumers which helps keep consumer retention rates up.
What technologies and application software does	The current software that Netflix uses is

it currently use?	Cine-Match.

Figure 6. Netflix IT info

3) Business needs

Problem 1: Friends, A social side to Netflix to sync shows with the people you tend to be watching with. The Netflix algorithm will check what other groups are watching as well to recommend them as well.
Problem 2: Anti Recommended, based off your watch history Netflix would never show you what you don't tend to watch.
Problem 3: Skip Credits, Send a show to a complete state once you reach the credits. This will eliminate the show from one's "Continue Watching" tab.
Problem 4: "I'm Feeling Lucky" Random show button. That still is
Problem 5: Content Creation, based on what is highly rated and viewed by all users, Netflix could tailor new netflix originals to have similar Actors/Directors/etc.

Figure 7. Business needs

4) Project Plan

1. **Proposal:** Identify a business problem within the context of a medium-sized or large company. and defining a machine-learning-based software application to solve this problem
2. **Phase I:** Information technology (IT): Building a traditional IT data-center architecture to support this application
3. **Phase II:** Software Defined Infrastructure (SDI): Building a virtualization and cloud-computing version of the traditional IT architecture defined in phase 2.
4. **Phase III:** Information Science (IS): Developing the actual machine-learning based software application to solve the problem defined in the proposal.
5. **Phase IV:** Integration and project presentation: Integrating the IT and IS phases of the project, and making a compelling presentation to senior management.

TASK	***DUE DATE***	***Roles/ Responsibilities***
Form project teams and choose company	04/13/2017 (In class)	Contributions: Everyone Project Lead: Craig Standley
Formulate Project Proposal	04/18/2017	Contributions: Everyone
Phase 1 (Information Technology)	05/2/2017	Contributions: Everyone
Phase 2 (Software Defined Infrastructure)	05/16/2017	Contributions: Everyone
Phase 3 (Information Science)	05/30/2017	Contributions: Everyone
Phase 4: Closure and Final Report	06/6/2017	Contributions: Everyone

Figure 8 Division of tasks

Phase I

Schedule:

Wednesday 4/26: Create project Schedule and Plan	**Thursday 4/27:** Complete Software Requirements and Vendor selection for Netflix	**Friday 4/28:** Complete Client-Server Software/ Hardware Architecture Design for Netflix	**Saturday 4/29:** Complete Database Design and Network Architecture for Netflix	**Sunday 4/30:** IT Integration for Netflix	**Monday 5/1 :** Define Data Mining problem for Phase III	**Tuesday 5/2:** Check work

Figure 1.1: Schedule for Project Phase I

1. Develop the project plan for Phase 1
    a. List tasks
        A. Software Requirements and Vendor selection for Netflix
        B. Client-Server Software/Hardware Architecture Design for Netflix
        C. Database Design for Netflix
        D. Network Architecture for Netflix
        E. IT Integration for Netflix
        F. Data Mining
    b. Create activity matrix

	A	**B**	**C**	**D**	**E**	**F**
A	A					
B		B				
C		X	C			
D		X	X	D		
E	X	X	X	X	E	
F	X					F

Figure 1.2 Activity Matrix

B depends on A

C depends on B
D depends on B, C
E depends on A, B, C, D
F depends on A

c. Create a Gantt chart

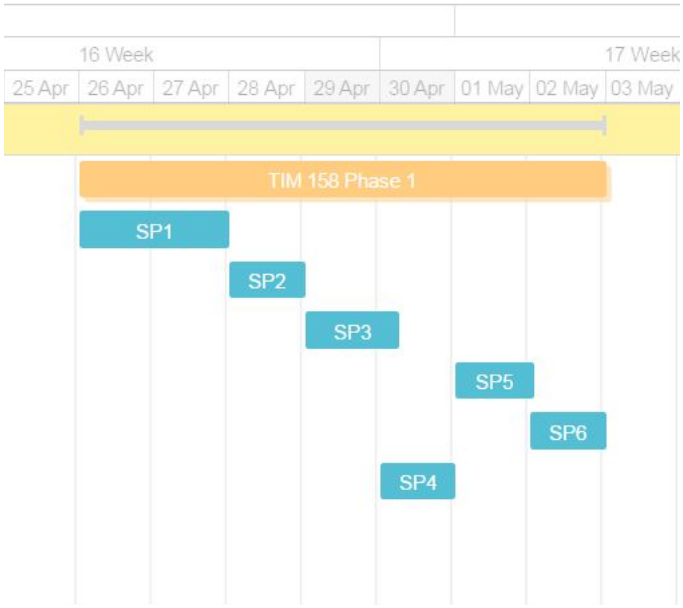


Figure 1.4: Gantt Chart

a. Find critical path with PERT chart

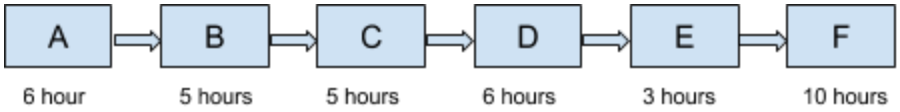


Figure 1.5: PERT Chart

-Critical path: A-B-C-D-E-F

2. Assign Tasks to Group Members

	Task A	Task B	Task C	Task D	Task E	Task F
Cole			lead			lead
Craig	lead					
Jerzy				lead		

Miguel		lead				
Robert					lead	

Figure 1.6: Task Matrix

Software Requirements and Vendors selection for Netflix

(1.) Define The Problem

SP1: Identify at the key business processes.

SP2: Define and list a basic set of requirements for the software to automate these business processes.

SP3: Develop use case for key business process. Use these use-cases to check the list of software requirements.

SP4: Specify software and hardware vendor options for the different tiers in your IT design.

SP5: Create a set of selection criteria to select between the software and hardware options and then use these selection criteria to choose the appropriate software and hardware vendors for each tier of your design.

(2.) Plan

What information is available for solving the problem?

Notes, and class handouts.

What assumptions need to be made to make the solution process manageable?

Assume that Netflix has conventional market and business strategies to maximize profit.

What analysis needs to be performed to resolve the issues defined in Step 1?

The following process will be implemented to resolve the defined sub-problems

a) Identify the main business problem.
b) Identify the main sub-functions which are involved.
c) Create a function tree for the business process and define software to be used for automation.
d) Develop use-cases for sub-function.
e) Specify software vendor options.
f) Create a set of selection criteria
g) Choose the appropriate software vendor

(3.) Execute the Plan

a) Identify the main business problem.

The key business problem that Netflix is proposing to solve is to provide online streaming of different

The key business problem that Netflix is proposed to solve with IT is increasing group watching retention rates.

b) Identify the main sub-functions which are involved.

The two main sub function for the group watching process are shown in the table below.

Sub-Function	**Added benefit to solve problem**

Group User Login	Gives Access to multiple members preferences
"Group" machine learning recommendation system	Retain more users in groups due to showing only movies that fit all groups preference

Figure 2.1: Sub-function Table

c) Create a function tree for the business process and define software to be used for process automation

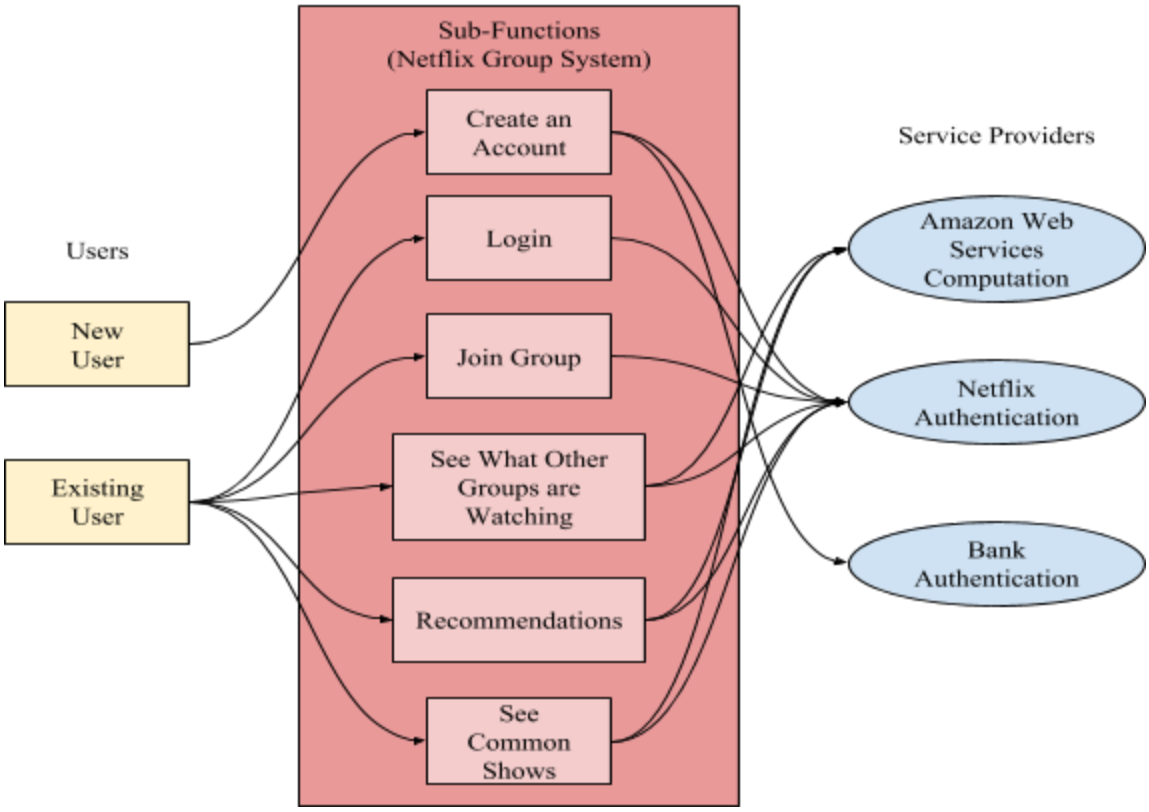


Figure 2.2: Function Map

Automation Software	**Relevant Use-Cases**
Data Analytics	Rate video, Make Recommendation
Transaction Processing System	Billing
Digital Supply Chain Management Software (Content Ops and Stream)	Stream Video

DataBase Management Software	Login, Browse Library

Figure 2.3: Software needed for business process

Use Cases:

Here the main Use-Cases are browsing the netflix library of shows and movies, streaming content through the application, Logging into your Member account, rating content and paying for the service. Netflix is an internet service for streaming movies and TV shows to personal computers and TVs. Anyone can browse the Netflix library, (by title, actor, director, genre, etc.), but the user must have a subscription if they wish to stream any content. A User can activate, suspend, or cancel their membership. A Membership is active as long as it has not been suspended or cancelled. The subscription fee is charged per month, on the monthly anniversary of the day the membership was first activated, this billing period is tracked by a billing timer. Functions needed to complete each sub process are identified using a Use-Case diagram which contains all the Use-Cases, actors, and relations in the process scope. In the diagram below all relations are include relations and are represented with the dashed arrows. The Lines from each actor to the main use-case represent who is involved in that action.

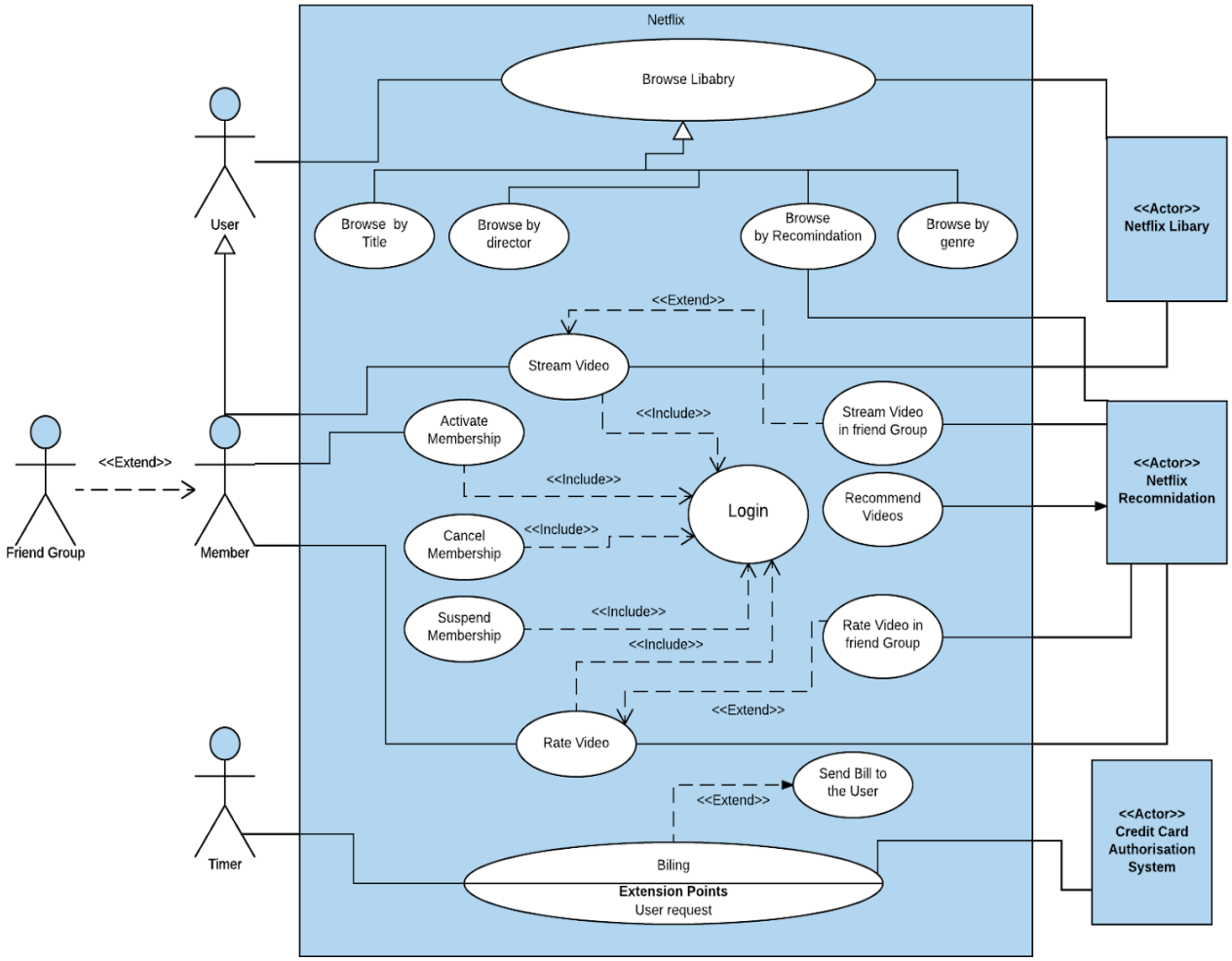


Figure 2.4: General Netflix Use-case

An example Use-Case Description is given for friend group stream video use case,

Use-Case Description	
Name: Netflix Group Recommendation	**ID:** 1
Description: Makes recommendations when watching in groups based on an accumulation of all members preferences.	

Actors: Friend Group Member Netflix Library	**Relationships:** Includes: Extends:
Triggers: Selecting the group option in the menu	
Pre-conditions: All Users in the group are members.	**Post-conditions:**
Main Flow: 1. User logs into registered Membership Account 2. User selects "Watch with Friends" Options from home page 3. User adds friends accounts to group (EX1) 4. User is taking to group member home page 5. Recommendation Algorithm looks a commonly liked shows for all group members 6. Recommendations are listed for user group to chose from 7. User group selects movie from list(EX2) 8. User group watches movie	
Exception Flow: EX1. Friend is not found or doesn't have account 1. Inform User that Friend Couldn't be found 2. Suggest Friends that exist with similar names 3. Suggest Account creation for Non-existent friend 4. Go back to add other friends or complete group EX2. Group doesn't choose any movie made by recommendation 1. Algorithm asks User Group for additional search criteria(Genre, length) 2. Algorithm recalculates recommendation options	

Figure 2.5: Use-Case Description

f) Create a set of selection criteria

The software criteria was selected for each of the four types of software needed to achieve automation for the Group Recommendation Process. The selection criteria for each type was chosen based on what is most need from the software to meet Netflix's needs, for example the Transaction System deals with Users credit/debit card Information so it is imperative that it is secure. The Criteria are rated on a 1-10 scale.

a. Software Selection Criteria:

i. **Data Analytics**

Component	Responsiveness	Accuracy	Reliability	Price
Importance	8/10	9/10	7/10	5/10

Figure 2.6: Data Analytics Selection Criteria

ii. **Transaction Processing System**

Component	Consistance	Ease of Use	Duriblity	Securce	Price
Importance	9/10	7/10	8/10	10/10	7/10

Figure 2.7: TPS Selection Criteria

iii. **Digital Supply Chain Management Software**

Component	Responsiveness	Speed of Computations	Reliability	Efficiency	Price
Importance	8/10	8/10	9/10	8/10	5/10

Figure 2.8: DSCM Selection Criteria

iv. **DataBase Management Software**

Component	Ease of Access	Fast Query	Reliability	Atomicity	Duriblity	Price
Importance	10/10	8/10	9/10	6/10	8/10	5/10

Figure 2.9: DBMS Selection Criteria

g) Choose the appropriate software vendor

Automation Software	Software Vendors	Selected Vendors
Data Analytics	R Programing, Apache, SAS, AWS	AWS

Transaction Processing System	Braintree, paypal, AWS	AWS
Digital Supply Chain Management Software (Content Ops and Stream)	Apache, SAP, KPMG, AWS	AWS
DataBase Management Software	Oracle, SAP, Microsoft, MYSQL, Apache, AWS	AWS

Figure 2.3: Software needed for business process

(4.) Check your Work

The work was completed based on the process outlined in the class notes and seems to be correct. Some of the IT software used by Netflix was found through online research and case studies. The use case diagrams and description were formed by using UML processes taught in TIM 58.

(5.) Learn and Generalize

This problem illustrated how business process can be broken down into sub-functions and use-cases and can be automated with the proper software, the selection of which is made depending on the business needs of a firm.

Client-Server Software/Hardware Architecture Design for Netflix

(1.) Define The Problem

SP1:

A. Research and compile information on the IT hardware and software infrastructure for Netflix
B. Create a FAST diagram to present the gathered information on the system.
C. Design the client server software and hardware architecture to solve your company's key business problem.
D. Specify software and hardware vendor options (alternatives) for the different tiers (layers) in your IT architecture.
E. Create a set of criteria to address the selection of the software and hardware options (alternatives), and then use these selection criteria to choose the appropriate software and hardware vendor for each tier (layer) of your design

(2.) Plan

- Read associated chapters in the textbook
- Search for data on netflix

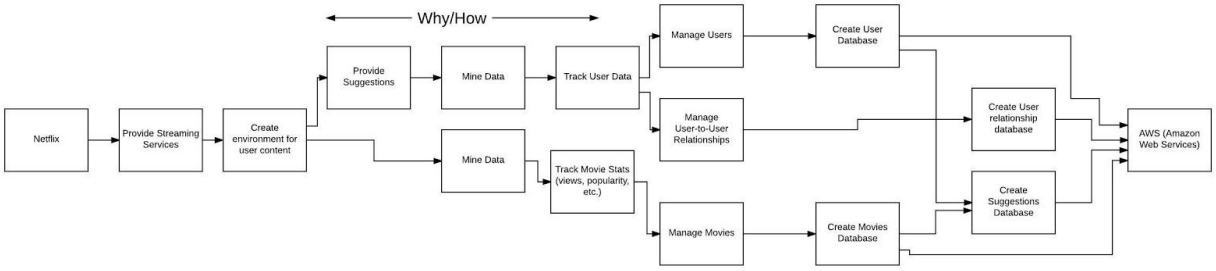
    - Hardware Infrastructure
    - Software Infrastructure
- Analyze Data
- Reference Example FAST Diagrams
- Research software and hardware alternatives
- Analyze
- Check Work
- Complete and Generalize

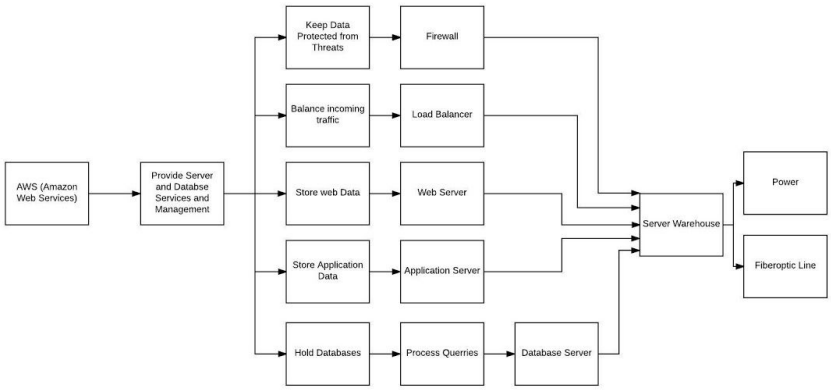
(3.) Execute the Plan

b. Research and compile information on the IT hardware and software infrastructure for Netflix

- IT Hardware Infrastructure: Netflix uses AWS (Amazon Web Service Entirely) This means they have employee machines and a network in their office so they can connect to AWS and modify their databases and manage their information
- IT Software Infrastructure: Data Mining, Build and Delivery tools, Runtime services, DBMS, Content Encoding, Security, and UI.
- Sources:
    - https://aws.amazon.com/solutions/case-studies/netflix/
    - http://www.datacenterknowledge.com/archives/2016/02/11/netflix-shuts-down-final-bits-of-own-data-center-infrastructure/
    - https://netflix.github.io/
    -

c. Create a FAST diagram to present the gathered information on the system.

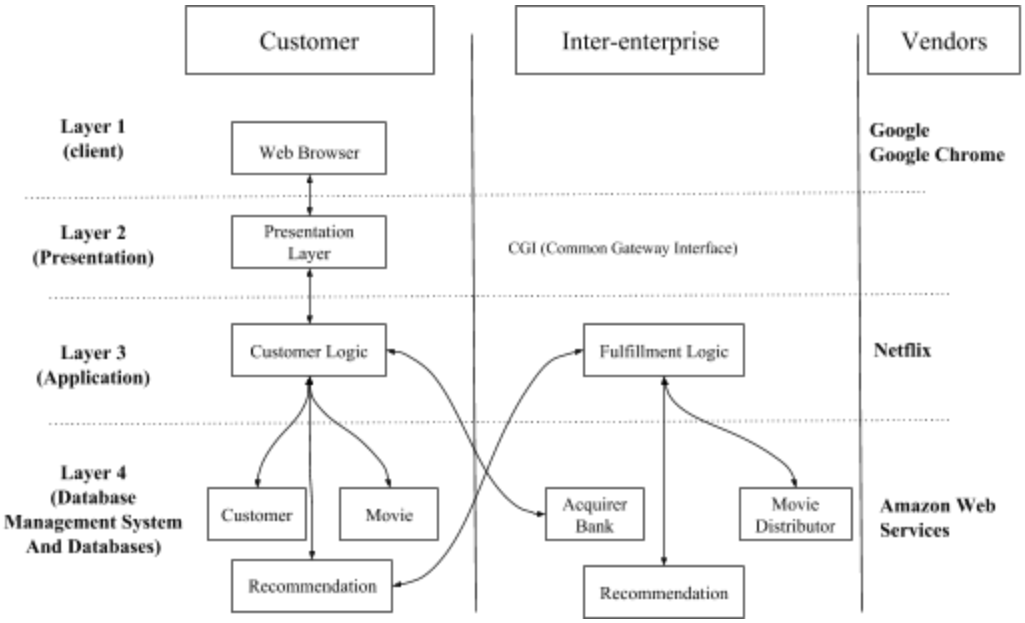




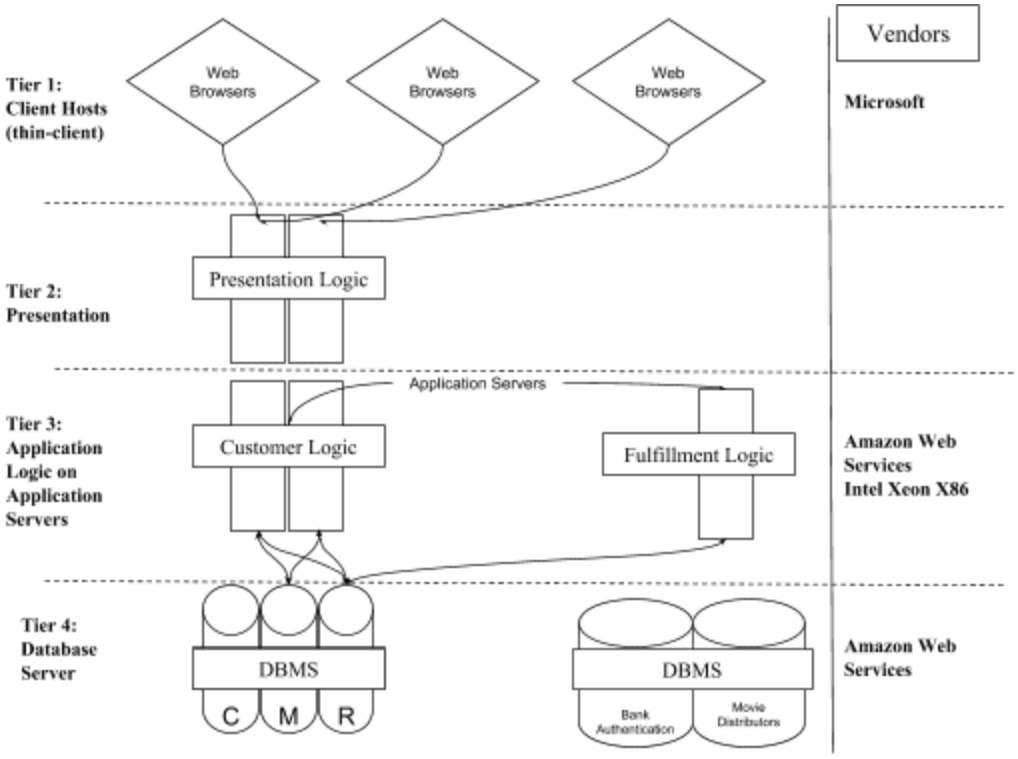
Since Netflix utilizes AWS we made 2 FAST diagrams one showing the software that netflix develops and works with on top of AWS and their software and machines. The second FAST diagram we made shows AWS hardware architecture and they also have various software that helps clients like netflix manage their cloud services.

d. Design the client server software and hardware architecture to solve your company's key business problem.

i. Software Architecture



ii. Hardware Architecture



-

e. Specify software and hardware vendor options (alternatives) for the different tiers (layers) in your IT architecture.

- Software Hardware Package Options
    - AWS Amazon Web Services (Currently Using)
    - Azure Microsoft
    - Google Cloud
    - Rackspace
    - Verizon Cloud
    - In House Computing
        - Servers:
            - Processor: Intel x860 Xeon
            - Storage: Seagate, Sandisk
            - Housing+Motherboard: HP, IBM
        - Software
            - DBMS: Oracle, Microsoft, IBM
            - Security: AVG

f. Create a set of criteria to address the selection of the software and hardware options (alternatives), and then use these selection criteria to choose the appropriate software and hardware vendor for each tier (layer) of your design

Software Selection Criteria:

i. Client Host

Component	Responsiveness	Ease of Use	Reliability
Importance	8/10	7/10	9/10

ii. Presentation

Component	Responsiveness	Ease of Use	Reliability
Importance	8/10	9/10	7/10

iii. Application

Component	Responsiveness	Speed of Computations	Reliability
Importance	8/10	8/10	9/10

iv. DBMS

Component	Ease of Access	Fast Query	Reliability
Importance	10/10	8/10	9/10

Hardware Selection Criteria:

v. Client Host

Component	Low Price	Reliable Specs	Future Proof (5 years)
Importance	5/10	7/10	8/10

vi. Presentation

Component	Responsiveness	Ease of Use	Reliability
Importance	8/10	7/10	9/10

vii. Application

Component	Responsiveness	Ease of Use	Reliability
Importance	8/10	7/10	8/10

viii. Data Base

Component	Storage Capacity	Ease of Use	Reliability
Importance	9/10	7/10	8/10

(4.) Check your Work

a. We checked our work and it makes sense.

(5.) Learn and Generalize

Database Design for Netflix

(1.) Define The Problem

SP1: What is the number of databases required?
SP2: What is the entity type for each data-base?
SP3: What are the attributes for each entity?
SP4: What are the key tables needed for the application, define them.
SP5: For each database states the attributes and its primary key.
SP6: Create an entity relationship diagram
SP7: Show how the information in the databases are related.

(2.) Plan

What information is available for solving the problem?

Notes, and class handouts.

What assumptions need to be made to make the solution process manageable?

Assume that Netflix has conventional market and business strategies to maximize profit.

What analysis needs to be performed to resolve the issues defined in Step 1?

The following process will be implemented to resolve the defined sub-problems

A. Specify the key databases needed for the application.
B. Provide the appropriate entity for each database defined in part A.
C. For each database, state its attributes and its primary key.
D. Create an entity relationship diagram for each database.
E. Explain how the information in the different databases are related (connected) to each other for a typical use-case.
F. Identify a list of potential DBMS vendors and storage vendors for Netflix's IT architecture.

(3.) Execute the Plan

a. Specify the key databases needed for the application.

Relevant Databases
Member Information
Content Library
Content Distributor
Billing and Payment

Figure 3.1 Relevant Data Bases

b. Provide the appropriate entity for each database defined in part A.
c. For each database, state its attributes and its primary key.

The first entity is the Member which uses data from the Member information database and contains the following attributes, Member ID, Membership Type, Member name, Member status, Member Video ratings, Member watched list, and login information. The Primary key for this entity is the Customer ID number since it is unique to every customer and can be used to link the customer with the Library and Payment entities.

Member						
Member ID (Primary key)	Membership Type	Member name	Member Status	Member rating	Member watched list	Login Information
1						
2						
.						
n						

Table 2.2 Member Entity

The second entity is the Library which will use data from the Content Library database and will contain the following attributes, Video ID , Video Name, Video Info, Avg Video rating, and Video size. The Primary key for this entity is the Video ID which is unique to each video in the library database; it can be used to link the Library to the Distributor entity.

Library				
Video ID (Primary key)	Video name	Video Info	Avg Video rating	Video size
1				
2				
.				
n				

Table 2.3 Library Entity

The third entity is the Content Distributor which will use data from the Content Distributor database and will contain the following attributes Distributor ID, Video ID , Distributor Name, Distributor Info, Distributor Video License and Video Price. The Primary key for this entity is the Distributor ID which is unique to each Distributor in the Content Distributor database; it can be used to link the Distributor to the Library entity. Note that Video ID is a secondary key used to link movies to their distributors.

Content Distributor					
Distributor ID (Primary key)	Video ID	Distributor Name	Distributor Info	Distributor Video License	Video License Price
1					
2					
.					
n					

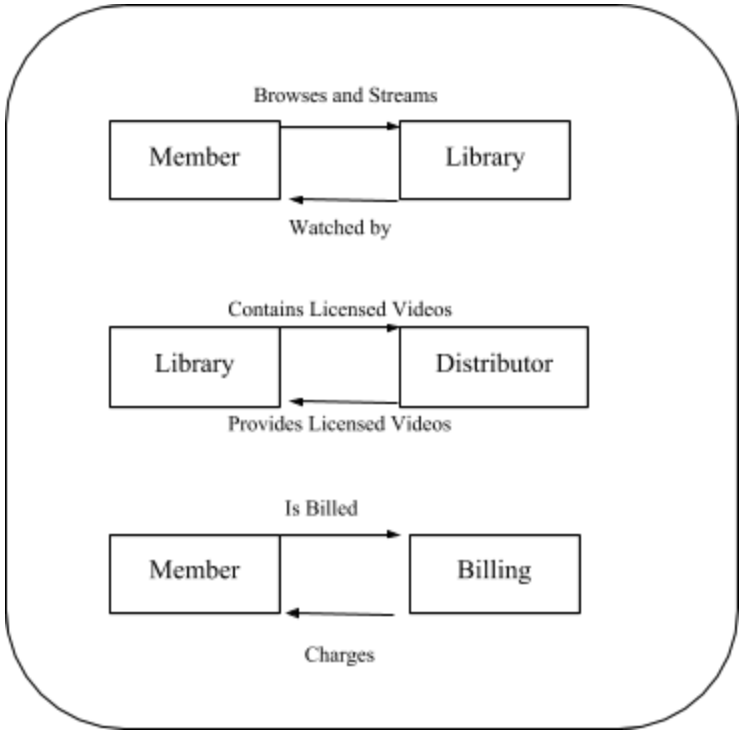
Table 2.4 Content Distributor Entity

The Final entity is Billing which will use data from the Billing and Payment database and will contain the following attributes Billing ID, Member ID, Member Card Info, Billing Status, Billing amount, Billing Date . The Primary key for this entity is the Billing ID which is unique to each Payment entry in the Billing and Payment database; it can be used to link the Billing to the Member entity.

Billing					
Billing ID (Primary key)	Member ID	Member Card Info	Billing Status	Billing amount	Billing Date
1					
2					
.					
n					

Table 2.5 Billing Entity

d. Create an entity relationship diagram for each database.



e. Explain how the information in the different databases are related (connected) to each other for a typical use-case.

- The Membership database contains all the Members Information and settings this is accessed through logging in, this information is used by the application and presentation logic to set up the members streaming portal. It also contains the membership type, which the billing and payment entity uses to calculate how much to charge the member.
- The Library is supplied by movies and TV shows which are added through the content distributors entity

f. Identify a list of potential DBMS vendors and storage vendors for Netflix's IT architecture.

DBMS Vendors	Storage Vendors
Oracle	Intel
SAP	HP
MYSQL	Cisco

	Apache	AWS
Selected Vendors	MYSQL	AWS

Figure 2.3:Vendors for Data Center

(4.) Check your Work

The work was completed based on the process outlined in the lecture notes, all the subproblems have been addressed and the work is correct based on the given outline.

(5.) Learn and Generalize

This problem looked at basic database center design for Netflix and showed how to design database entities that hold useful information for the given business process. It also shows how these entities are linked and related.

Network Architecture for Netflix

(1.) Define The Problem

SP1: Show the network topology to transmit application commands and data back and forth between Netflix and the end-user for a typical use-case.

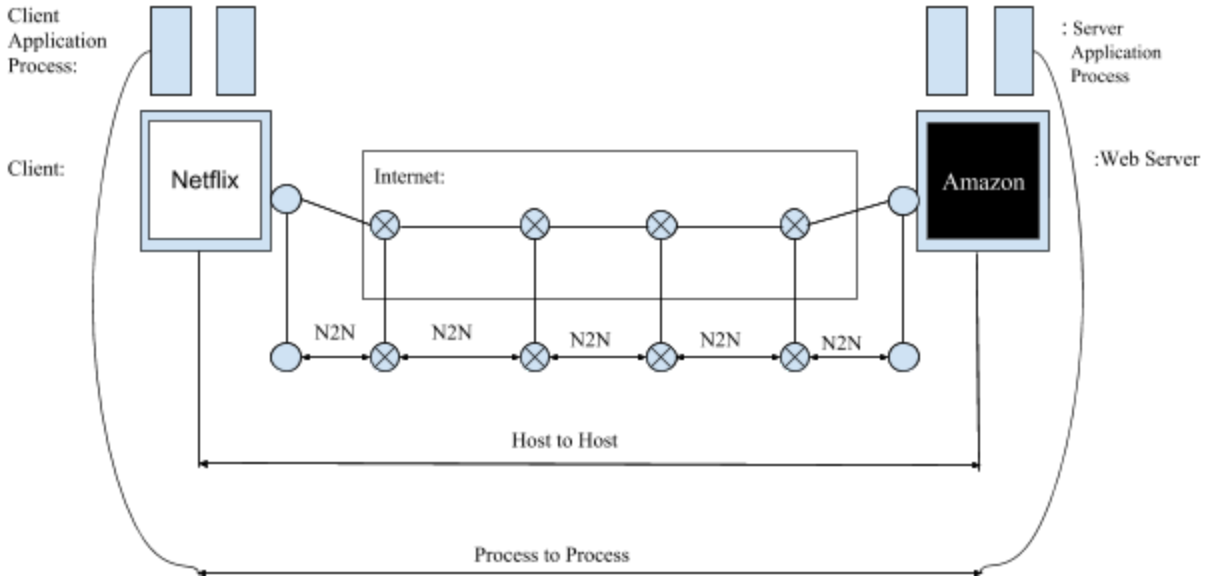
SP2: Clearly explain the information/data flow on the internet across the four primary layers (application, network, etc.) for a typical use-case.

(2.) Plan

A. Using network topology notes from class, construct the Network Architecture that will be used in order to transmit commands between the Netflix web server and the clients . This will contain node to node, host to host, and process to process links as well as switches and routers.
B. Identify the flow of information by breaking the transmission of data into four different layers, consisting of the application layer, transport layer, network layer and the data link layer. Using these four layers, follow the flow of information from the application layer all the way through transmission

(3.) Execute the Plan

a. Show the network topology to transmit application commands and data back and forth between Netflix and the end-user for a typical use-case.



b. Clearly explain the information/data flow on the internet across the four primary layers (application, network, etc.) for a typical use-case.

-The Four Primary Layers:

Layer 1: Data Link Layer

Enables the delivery of frames between two neighboring nodes (node to node = N2N).

Layer 2: Network Layer

Enables the delivery of packets as datagrams between two hosts (host to host = H2H).

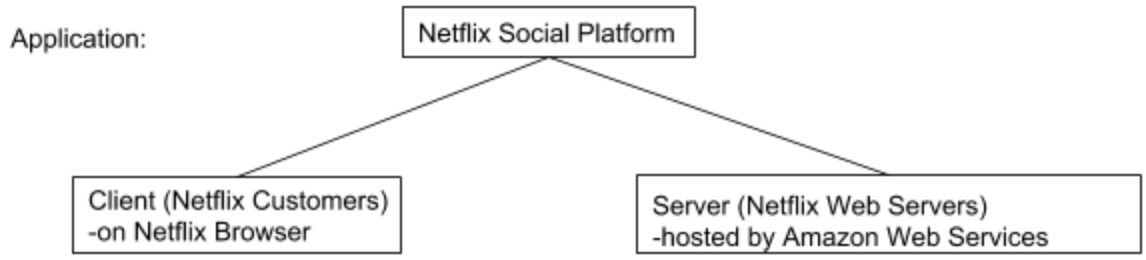
Layer 3: Transport Layer

Is responsible for the process to process delivery of segments (parts of messages) from one process to another.

Layer 4: Application Layer

Is responsible for passing messages between the client process and server process to perform useful tasks.

Information/Data flow across the four primary layers:

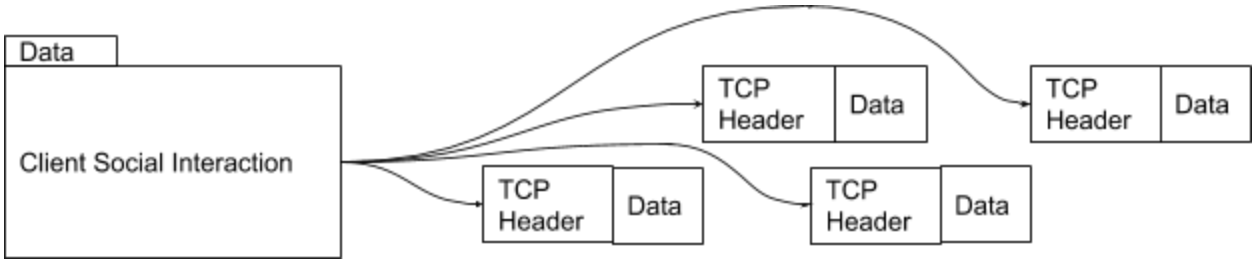


Step 1) Application Process

Converts social information (relations and preferences of the client, such as friend requests/ updates/ deletions, group submissions/updates/deletions, instant messages, likes, dislikes) for transmission across the internet using an application protocol.

Step 2) Transport Layer Process:

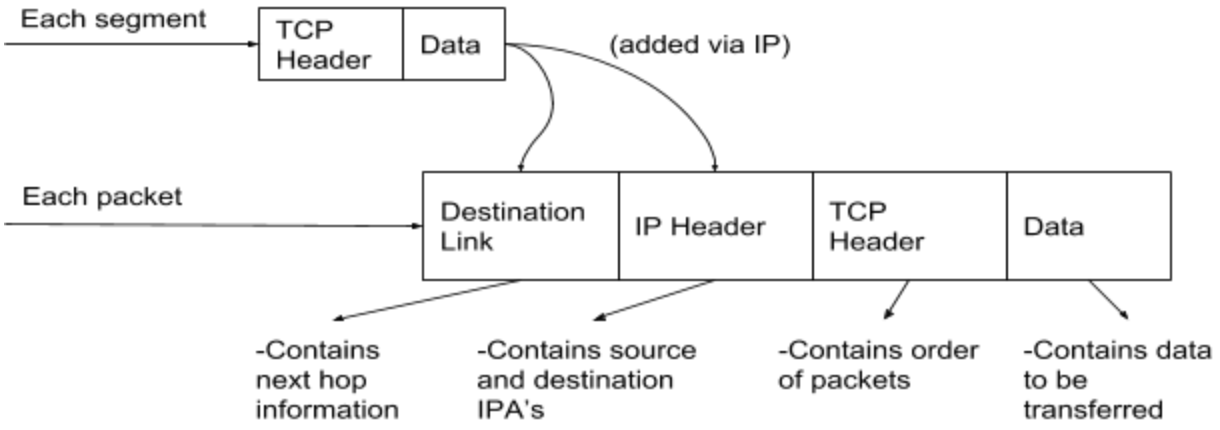
The converted information is passed on to the transport layer, where it is broken up into segments for transmission across the internet using the transport control protocol (TCP). During this step, each segment of data gets a TCP header which gets encapsulated into the data segments, making the order that data segments get transported in irrelevant.



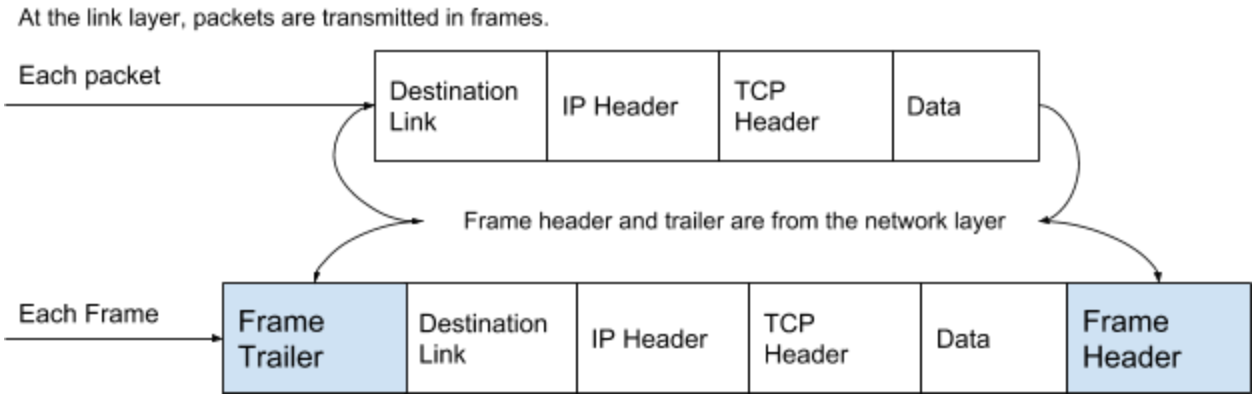
Step 3) Network Layer Process

-The data segments are sent to the network layer where a source address and destination address are added. This utilizes a protocol called IP (internet protocol), which specifies the source and destination

addresses.Once encapsulated with the source and destination addresses, the data segments are considered packets that are ready for transmission.
-Source and destination IPa's are used to create a routing table that will be needed later
-In case of Netflix, the destination is the IP address of Netflix web server

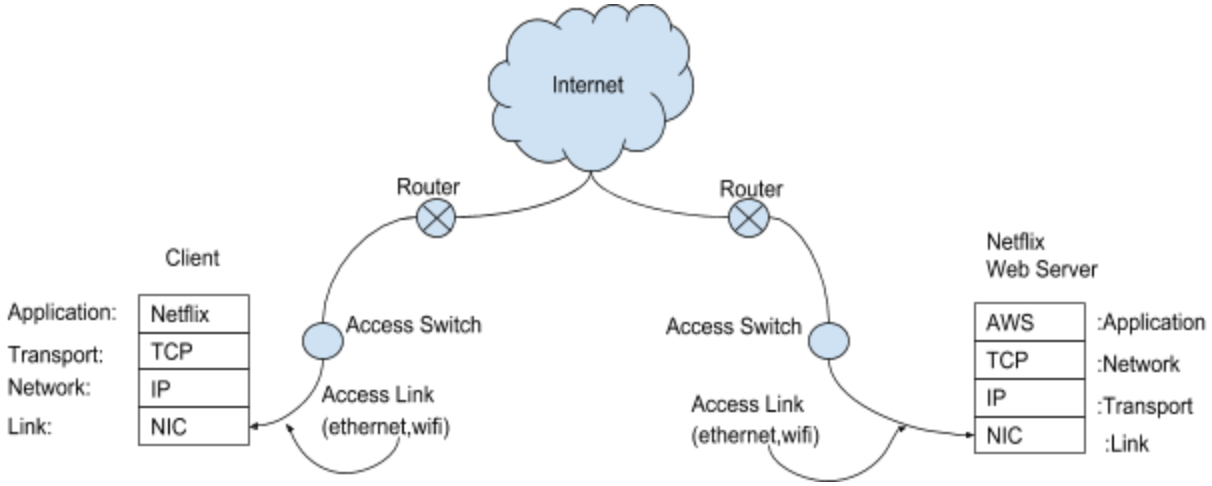


Step 4) Data Link Layer
-The IP and TCP encapsulated data packets are then passed to the link layer for transmission across a single link. In this stage of the process, packets are transmitted into frames for node to node delivery .The first link is sometimes called the "access link"



Step 5) Data Transmission

-The frames of data are now transmitted from link (hop) to link using the routing table that was created by the IP in the network layer (Step 3).
-NIC: Network Interface Card: Enables the packets to be transmitted across a link



(4.) Check your Work
-Using the client-server network architecture discussed in class we constructed the network topology for Netflix. Similarly to the application discussed in class (gmail), The Netflix Social Platform will run a client-server architecture whose main function is to enable the communication and interaction between clients. This ensures that our architecture is accurate for the client-server architecture at hand (Netflix Social Platform) and follows the guidelines discussed in class. The information that gets transmitted will always follow the five step process that we have presented involving all four layers, as this is the process that we reviewed in lecture and learned in CMPE150.

(5.) Learn and Generalize
-In constructing the network architecture and mapping the flow of information across the primary layers, we have learned how to apply a basic knowledge of networks in order to track the delivery of information needed for a use case. This is critical as it allows the Netflix application on the client end to communicate with the Netflix web server in order fulfill social requests or allow certain functionality. Basically, there is a five step process that ensures the transmission of data from the application at hand to the designated web server (host/server application). Starting from the application layer, the data is parsed and passed through to the transport layer, then to the network layer and finally the data link layer where it is transmitted. During each step, the data being transmitted is encapsulated and given a header (IP, TCP, etc.) which is protocoled and used to keep track of the order of packets, destination addresses, source addresses and more.We can use this information and defined process in the future when network problems or questions are presented later in the course.

IT Integration for Netflix

(1.) Define The Problem
SP1: Develop the end-to-end integrated IT architecture for Netflix: client-server architecture, database, storage, and network architecture.

(2.) Plan
Step 1: Define the network devices that are needed
Step 2: Map the requirements to the devices
Step 3: Based on the IT architecture define a set of subnets
Step 4: For each subnet, determine the network topology to connect the servers in that subnet using routers and switches
Step 5: Connect the subnets using r/s to create the tiered architecture and add load balancers and firewalls for each layer of architecture

(3.) Execute the Plan
Develop the end-to-end integrated IT architecture for Netflix: client-server architecture, database, storage, and network architecture.

a) Network Devices Needed:
-Routers/Switches
-Load Balancers
-Fire Walls

b) Requirements of the network devices:

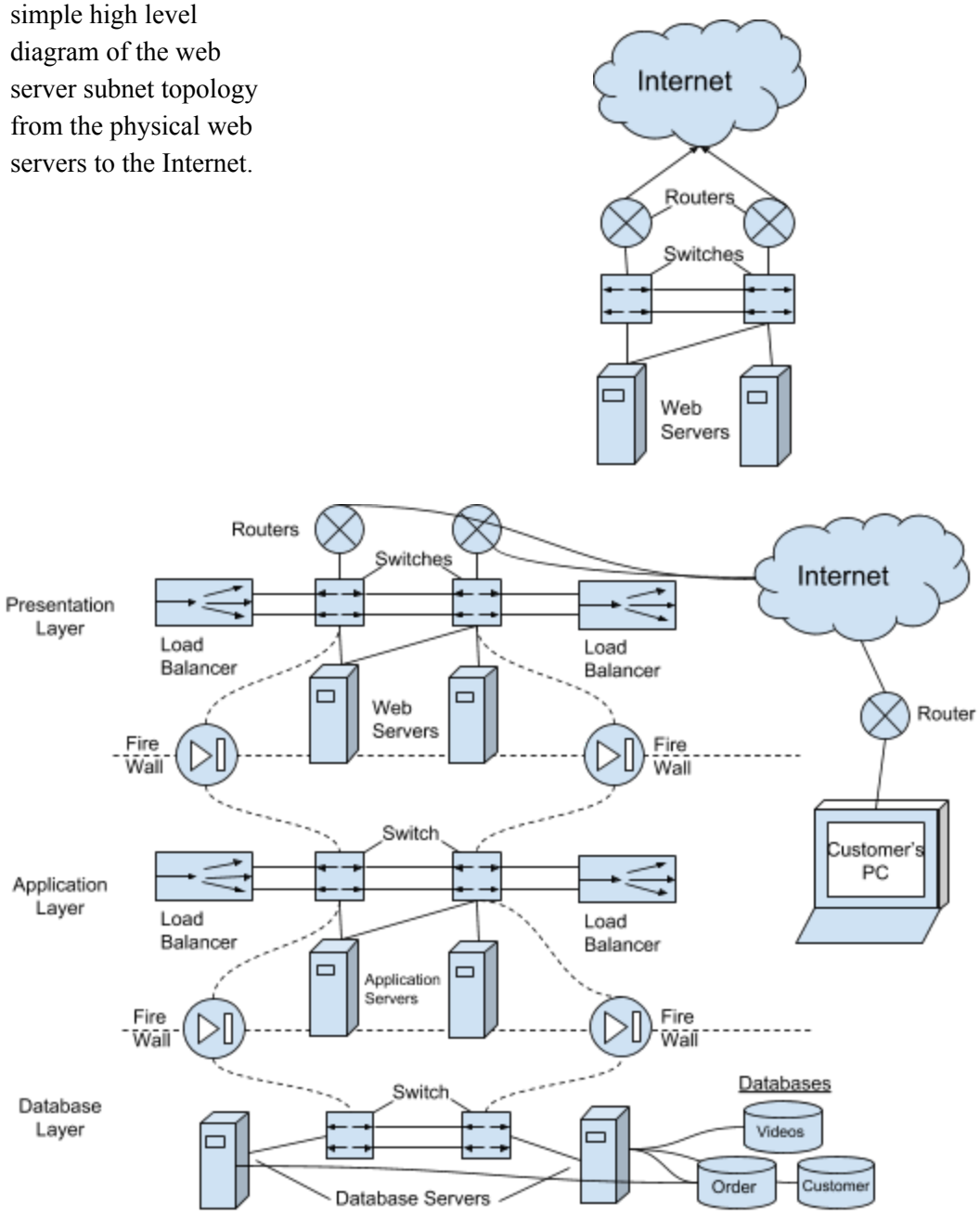
Requirements	Routers/Switches	Load Balancers	Fire Walls
Scalability	X	X	
Reliability		X	
Security			X

Figure 1.1: Shows the requirements of the network architecture in terms of physical devices

c) Necessary Subnets:
-Web Server Layer ⇒ Web Server Subnet
-Application Server Layer ⇒ Application Server Subnet
-Database Server Layer ⇒ Application Server Subnet

d) Server Subnet Topology:

Figure 1.2: This is a simple high level diagram of the web server subnet topology from the physical web servers to the Internet.



The figure above depicts the full IT architecture starting with Amazon's Data Center. The Amazon Data Centers house databases, database servers, application servers, web servers, load balancers, and firewalls for Netflix's application to run.

(4.) Check your Work

Went back to the Cisco network manual handout that was provided to us and checked with the full end-to-end IT infrastructure. Complete.

(5.) Learn and Generalize
Starting with use cases we find what requirements and respective network devices we will need to send data. Then we can map the network topology of the subsequent subnetworks. From there we can use firewalls and load balancers between each layer or subnet to connect up the integrated IT network architecture.

Data Mining

(1.) Define The Problem
SP1: Define the data mining problem that you plan to address in Phase III within the context of Netflix.

(2.) Plan

(3.) Execute the Plan
Define the data mining problem that you plan to address in Phase III within the context of Netflix.

http://bit.ly/2qCA5Ad
Netflix data set: 17,000 movies, 100 Million reviews

https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset
5000 movies with 28 included attributes IMDB

https://archive.ics.uci.edu/ml/datasets/Movie (Backup dataset)

Phase 2

Schedule:

Wednesday 5/10: Take a breather after midterm.	**Thursday 5/11:** Create project Schedule and Plan	**Friday 5/12:** IT Data Center Architecture for Netflix	**Saturday 5/13:** Virtualization for Data Center for Netflix	**Sunday 5/14:** Software Defined Infrastructure (Cloud Computing) for Netflix	**Monday 5/15 :** Machine Learning Problem for Netflix	**Tuesday 5/16:** Check work

Figure 1.1: Schedule for Project Phase I

1. Develop the project plan for Phase 2
    a. List tasks
        A. IT Data Center Architecture
        B. Virtualization for Data Center
        C. Software Defined Infrastructure (Cloud Computing)
        D. Machine Learning Problem
    b. Activity Matrix

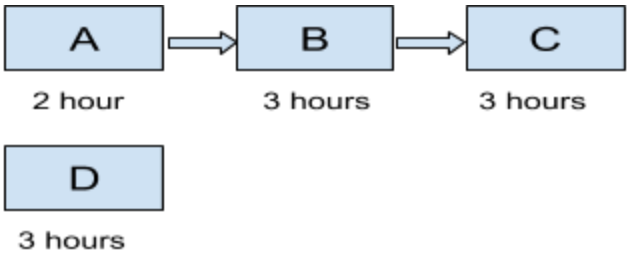
	A	**B**	**C**	**D**
A	A			
B	X	B		
C	X	X	C	
D				D

Figure 1.2 Activity Matrix

B depends on A
C depends on A, B
D independent

c. Gantt Chart

a. Find critical path with PERT chart



Critical path: A-B-C

Figure 1.3: PERT Chart

2. Assign Tasks to Group Members

	Task A	Task B	Task C	Task D
Cole	Check	Assist	lead	Check
Craig	Assist	Assist	Assist	Assist
Jerzy	lead	Check	Check	Assist
Miguel	Assist	lead	Check	Check
Robert	Check	Check	Assist	lead

Figure 1.6: Task Matrix

Complete IT Data Center Architecture for Netflix

DEFINE

Develop the integrated data center for your company: software applications, client-server architecture, database design, and network architecture.

PLAN

1) Revisit tasks 1-5 of Project Phase I
2) Look at the TA's notes on how to improve on data center architecture
3) Design the end-to-end integrated system architecture
    a) Software Application Architecture
    b) Client-server Architecture
    c) Database Architecture
    d) Network architecture
    e) Integrate above

EXECUTE

Develop the integrated data center for your company: software applications, client-server architecture, database design, and network architecture.

1) Revisit tasks 1-5 of Project Phase I
    -Had to make minor adjustments to original diagrams to encompass Netflix's main function (provide streamable, on-demand videos -movies, tv shows, documentaries, etc) rather than a recommendation function
2) Look at the TA's notes on how to improve on data center architecture
    -Did so using the Professor's feedback as well as our returned midterm feedback

3) Design the end-to-end integrated system architecture

Netflix goal is to provide affordable and reliable Streaming to homes around the world. To achieve this we must have a fully integrated IT system which will automate the important business process involved in Content Streaming. Netflix will use the following software vendors to achieve this automation. This software was selected due to its ease of use, its price, the features that are included and their reliability.

a) Software Application Architecture

Automation Software	Software Vendors	Selected Vendors
Data Analytics	R Programing, Apache, SAS, AWS	AWS

Transaction Processing System	Braintree, paypal, AWS	AWS
Digital Supply Chain Management Software (Content Ops and Streaming)	Apache, SAP, KPMG, AWS	AWS
DataBase Management Software	Oracle, SAP, Microsoft, MYSQL, Apache, AWS	AWS

Figure 1.1: Software needed for business process

In order to use this software efficiently, Netflix needed tiered client server architecture. This is critical as it allows Netflix's application on the client end to communicate with the Netflix web server to watch content online or allow for other certain functionality. We identified that a 4 tiered architecture would meet the requirements of the application. The DBMS would be integrated into the Database tier and the Data Analytics, TPS, and DSCM would reside in the application.

b) Client-server Architecture

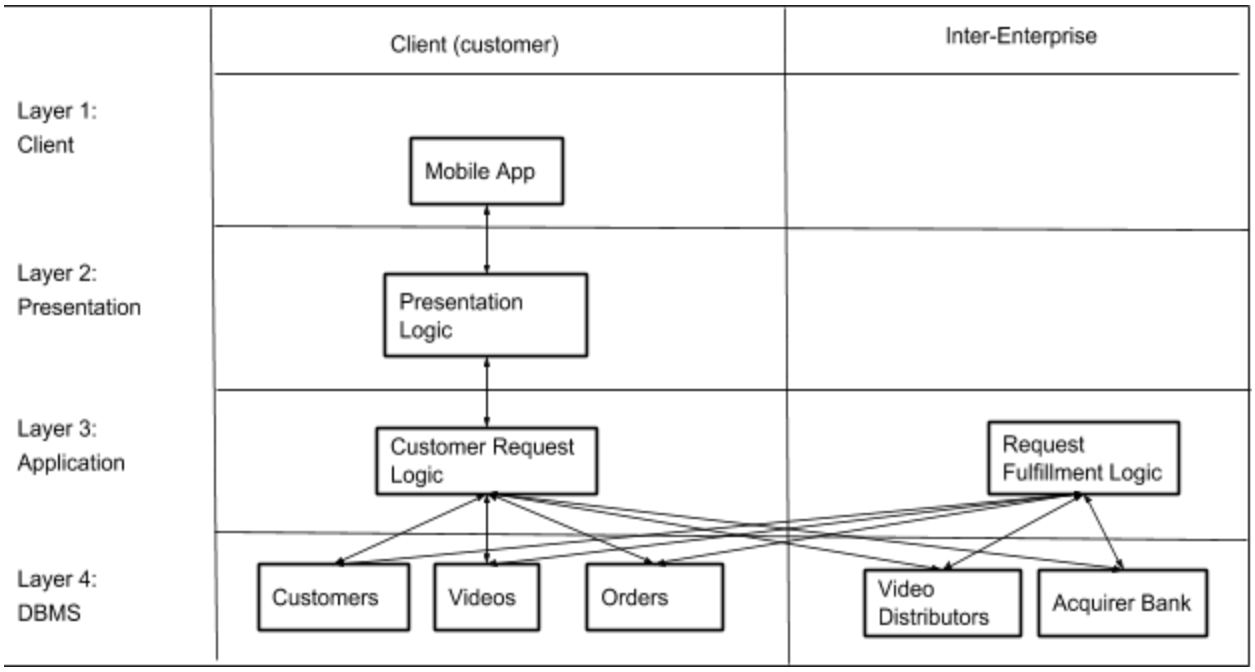


Figure 1.2: Client-Server Software Architecture

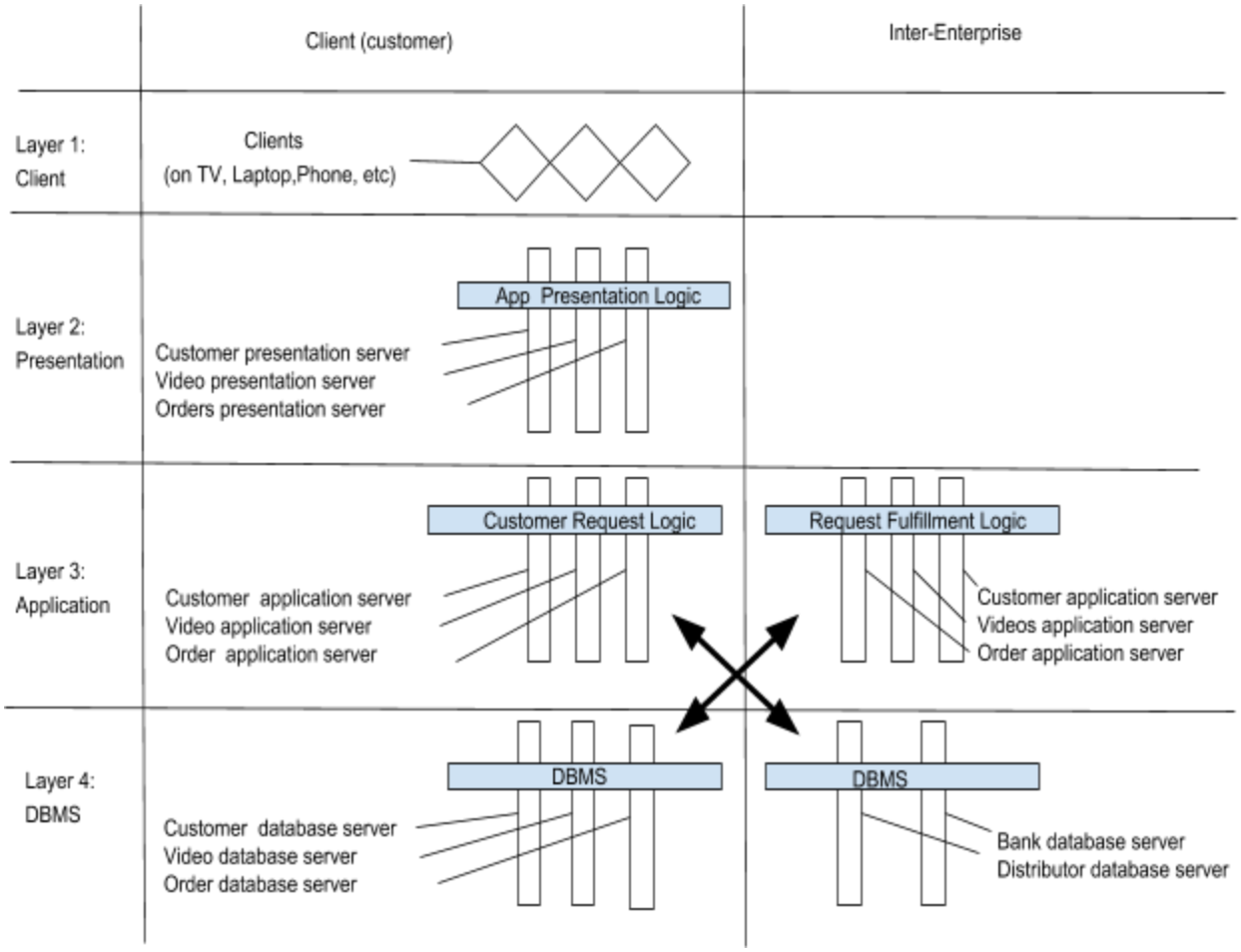


Figure 1.3: Client-Server Hardware Architecture

c) Database Architecture

To provide storage for all the data needed for a Streaming Service application Netflix has 4 main database that are distributed throughout the servers, these are identified in the table below .

Relevant Databases
Customer Information
Content Library
Content Distributor
Billing and Payment

Figure 1.4: Relevant Databases for Netflix Business process

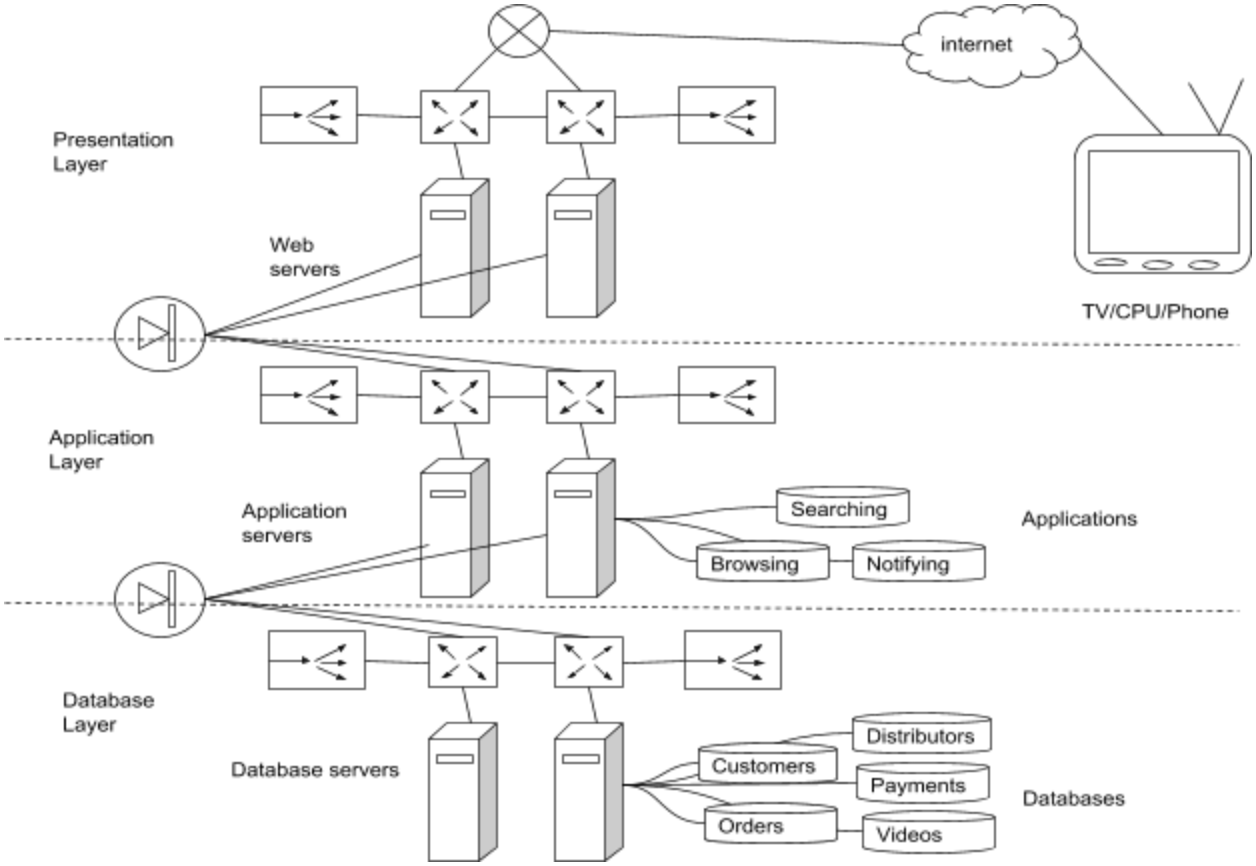


Figure 1.5: Database Network Architecture Diagram

d) Network Architecture

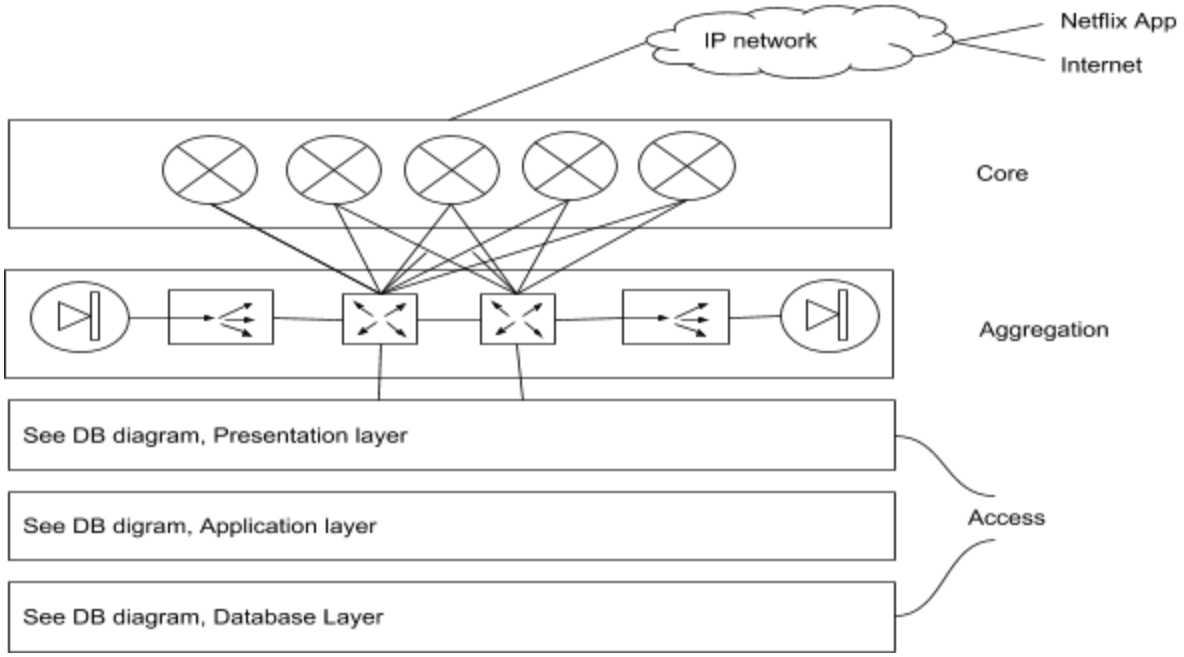


Figure 1.6: Network Topology Diagram

The figure above depicts the full Network architecture of Netflix Information Technology. The full networked IT consist of 4 subnets which are made up of, database servers, application servers, web servers, load balancers, and firewalls for its application to run.

e) Integrated Diagram

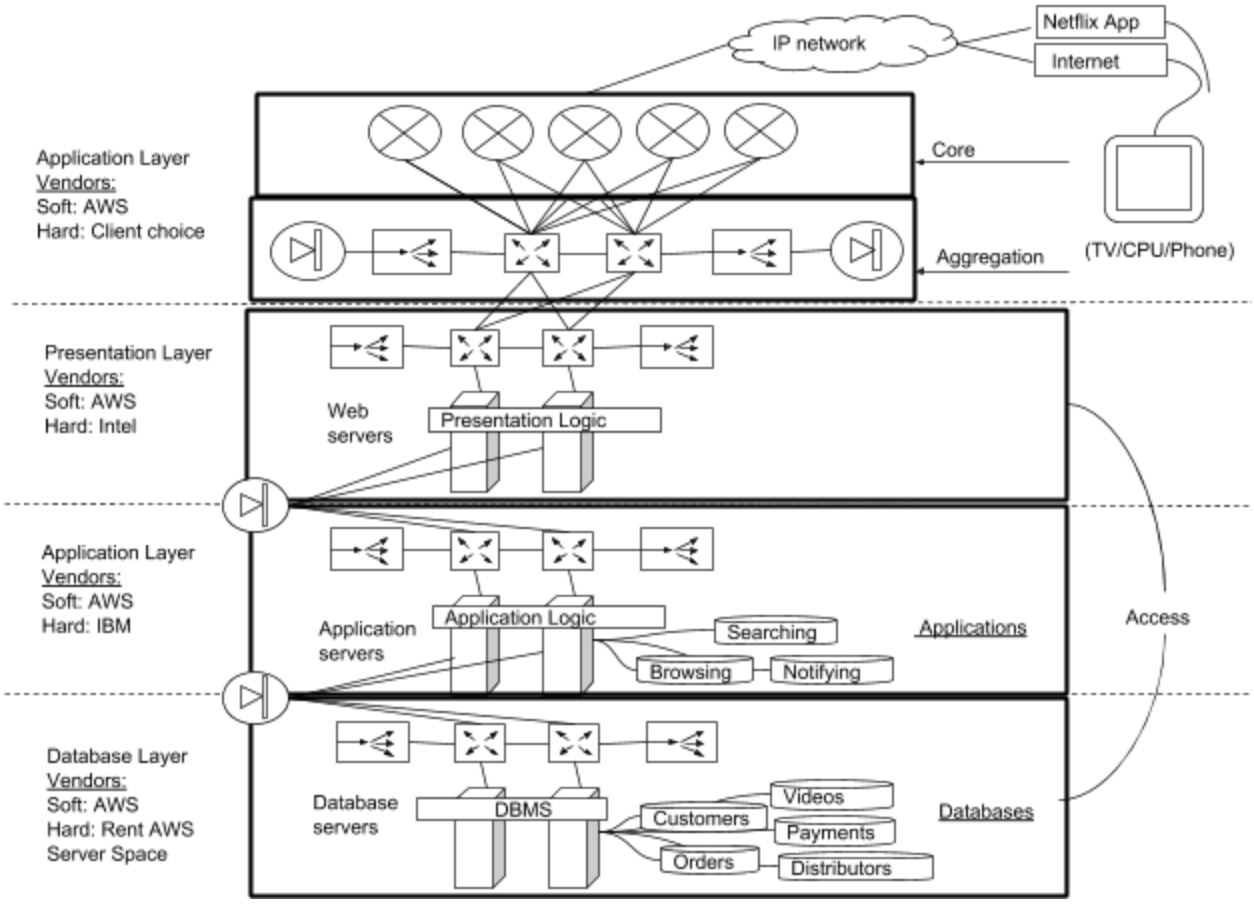


Figure 1.7: Database Network Architecture Diagram

CHECK

The work was taken from the previous phase and simply corrected and combined to get a full and comprehensive solution so it can be assumed it was done correctly.

LEARN

Starting with use cases we find what requirements and respective network devices we will need to send data. Then we can map the network topology of the subsequent subnetworks. From there we can use firewalls and load balancers between each layer or subnet to connect up the integrated IT network architecture.

Virtualization for Netflix's Data Center

DEFINE

Address the issue of virtualizing your company's data center. You will first need to define a process for performing this task, which will include addressing the following issues: addressing which attributes (e.g. processing, memory, storage, network) you plan to virtualize and the appropriate rationale for making your choices; the selection of the appropriate virtualization option based on appropriate (selection) criteria, etc.

PLAN

1) Figure out which type of virtualization is needed for Netflix's data center
    - Type 1 hypervisor runs on bare metal
    - Type 2 hypervisor runs on a host operation system and guest OS runs inside the hypervisor
2) Create the Data Center Use Cases
3) Evaluate your current server workloads
    - (Maximum server workload/ Average server workload) = n
    - Where do I find this information?
4) Select virtualization software and hosting hardware

EXECUTE

1. Virtualization 2 for Netflix's Data Center

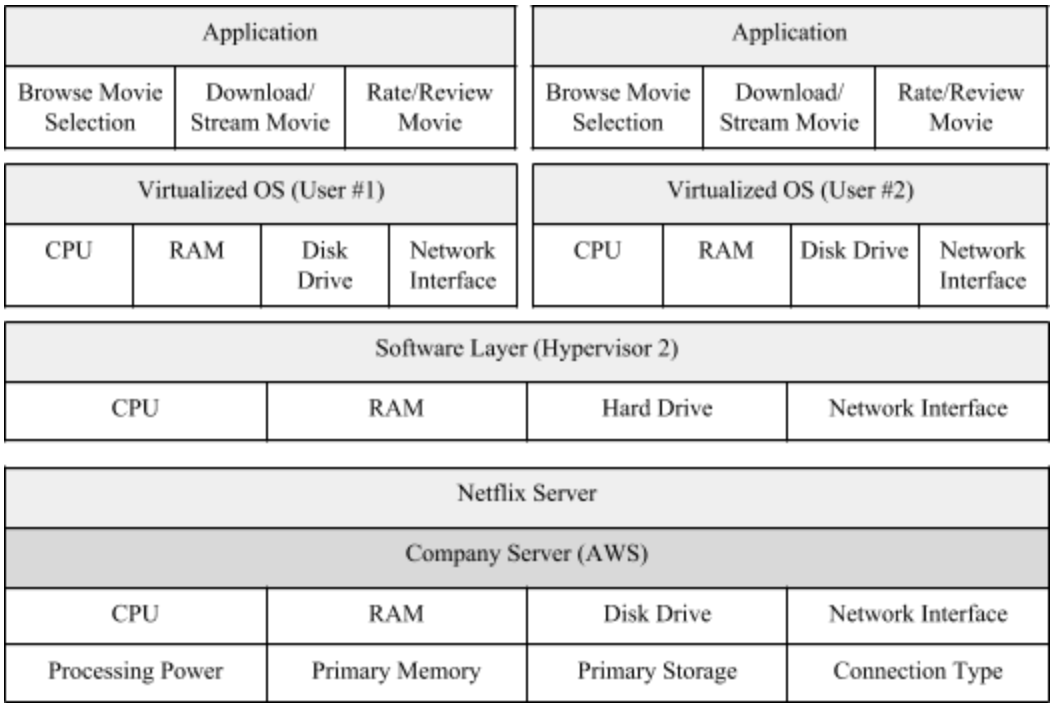


Figure 2.1: Netflix Virtualization

a. The Application layer displays the different actions that a user can do. The two different Virtualized OS shows that there are two users using the same array of applications. Both

users have a CPU, RAM, Disk Drive and Network Interface. All users are connected to the hypervisor. The Netflix Server is being hosted on Amazon Web Services (AWS) which contains all of the CPU, RAM, Disk Drive, Network Interface.

b. The above diagram shows layer by layer the needed functions and components of the virtualization.

2. Evaluate the Current Server Workload
    a. There are millions of users accessing netflix's server everyday
    b. Max Bandwidth= 70% of all internet Traffic
    c. Average Bandwidth = 40% of all internet Traffic
    d. 70%/40% = 1.75
3. Select Virtualization Software and Hosting Hardware
    a. The company that is best suited for hosting Netflix's Virtualization Software is Amazon Web Services (AWS) and it's Amazon Elastic Cloud (EC2) which allows scalable deployment of application by providing a web service through which users can boot a Amazon Machine Image, or instance, from any desired software. There are two different types of Visualizations that are viable. Paravirtualization and Hardware Assisted virtual machine. Netflix can use both of these virtualizations in order to achieve a broader range of clients.

- Options for Virtualization:

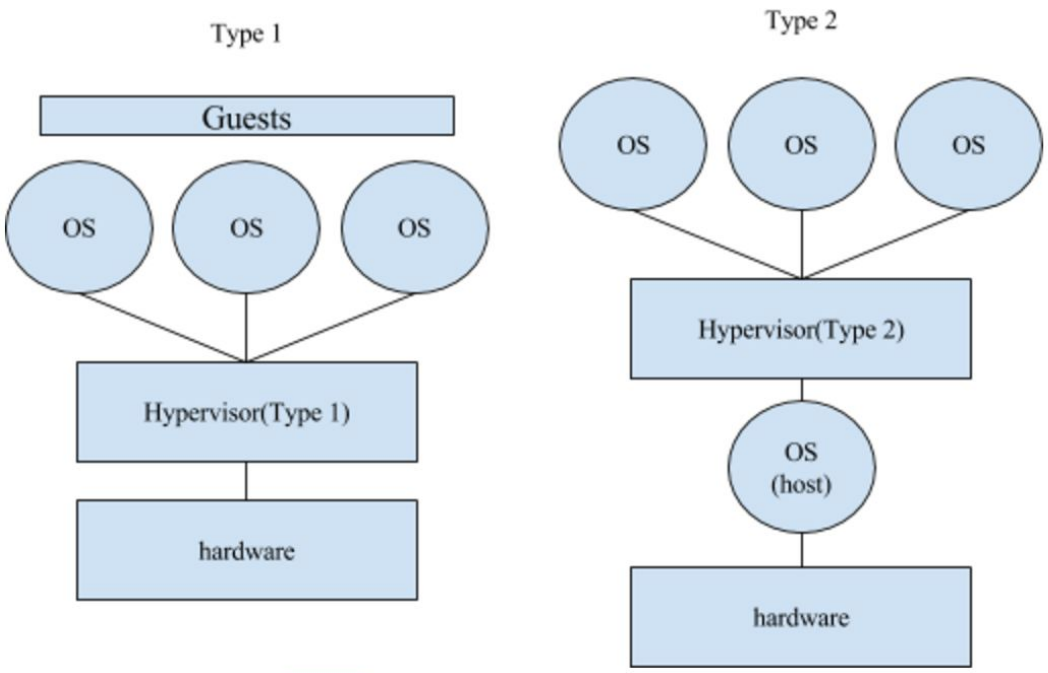


Figure 2.2: Virtualization Options

- Type 1 works strongly with Netflix in order to increase the ability to handle traffic. This means if we virtualize our Network and Processing on a type 1 virtualization method we can maximize the use value of each machine requiring less resources from AWS.

- Network: Network Virtualization creates the software equivalent of a switch allowing for switching to occur within a system and for it be programmed and adjusted to tailor the firm's needs.
- Processing: Process Virtualization utilizes VRAM (virtual ram) and virtual processors to increase the usability of a server. This is an important thing for

CHECK

Based on the information provided in class as well as a variety of internet resources we can confirm that our work is concise and correct and correctly analyzes how to decide on virtualization and implementation for Netflix.

LEARN

We learned how to implement a process for virtualizing Instagram's Data Center. This displays the information in a why which we can see the process in which hardware is con

Software Defined Infrastructure (Cloud Computing) for Netflix

DEFINE

Address the issue of hosting your data center, either all or some parts of it, in a Cloud Computing environment. You will first need to define a process for performing this task, which will include the following issues: addressing which attributes (e.g., application, data, server hardware) you want to host and the appropriate rationale for making your choices; the selection of the appropriate cloud services option based on appropriate (selection) criteria, etc,

PLAN

1. Explain rationale for why implementation of Cloud Computing is useful to Netflix.
    a. Pros and Cons
    b. Netflix needs
2. Provide the Building Blocks for Netflix's Web server.
3. Define the type of Cloud Computing that Netflix will use.
    a. Processing
    b. Networking
4. Define a selection criteria for CC Vendors
5. Select optimal CC Vendors

EXECUTE

1. Explain rationale for why implementation of Cloud Computing is useful to Netflix.

    Cloud Computing can providing IT to users the same way conventional companies provide utilities. The benefits of Cloud Computing are given in the figure below and illustrate how CC can provide extra Elasticity to Netflix's IT system to meet unexpected changes in supply or demand of the network and data center compared to traditional hardware.

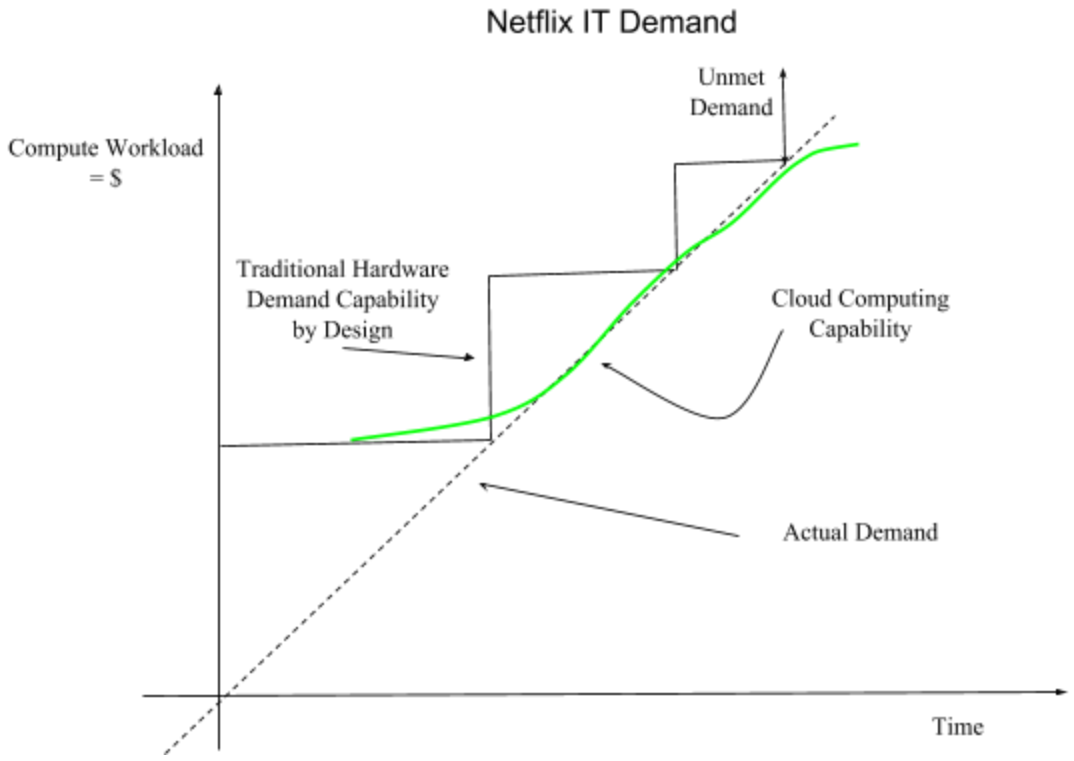


Figure 2.3: Netflix real time demand scaling (CC)

We realized that Netflix had to move away from vertically scaled single points of failure, like relational databases in our datacenter, and instead went towards highly reliable, horizontally scalable, distributed systems in the cloud. We chose Amazon Web Services (AWS) as our cloud provider because it provided us with the greatest scale and the broadest set of services and features.

Elasticity of the cloud allows Netflix to add thousands of virtual servers and petabytes of storage within minutes, making such an expansion possible without wasting time and money on new server racks. On January 6, 2016, Netflix expanded its service to over 130 new countries, becoming a truly global Internet TV network. Leveraging multiple AWS cloud regions, spread all over the world, enables us to dynamically shift around and expand our global infrastructure capacity, creating a better and more enjoyable streaming experience for Netflix members wherever they are.

The cloud allows one to build highly reliable services out of fundamentally unreliable but redundant components. By incorporating the principles of redundancy and graceful degradation in our architecture, and being disciplined about regular production drills using Simian Army software like chaos monkey, it is possible to survive failures in the cloud infrastructure and within our own systems without impacting the member experience.

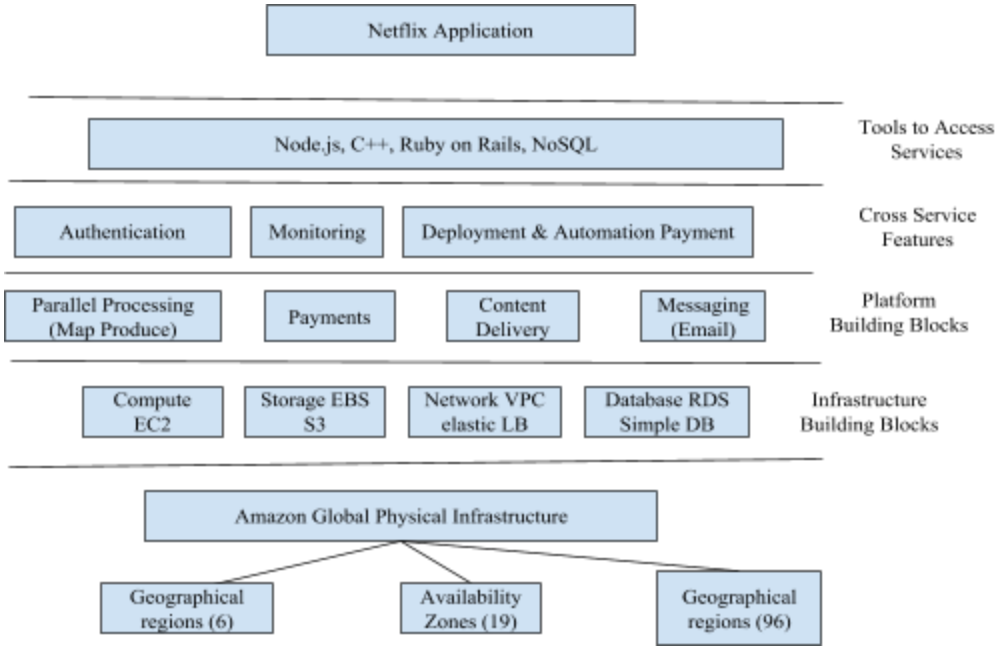


Figure 2.4: Netflix Services using AWS

The diagram above are the necessary building blocks in Netflix's Services via Amazon Web Services. The reason why this is used is because it provides a raw infrastructure so Netflix can do whatever it wants to upgrade its applications. It provides the company multiple options for expansion.

CHECK

Based on the information and insights provided in class as well as our own insights and internet resources we can say that our work is concise and correct and is a strong way to define the software infrastructure

LEARN

Address the issue of hosting your data center, either all or some parts of it, in a Cloud Computing environment. You will first need to define a process for performing this task, which will include the following issues: addressing which attributes (e.g., application, data, server hardware) you want to host and the appropriate rationale for making your choices; the selection of the appropriate cloud services option based on appropriate (selection) criteria, etc,

Phase 3

Machine Learning (ML) problem for Netflix

1. **Define the Problem**

- This problem is an **extension** of Project Phase 2, Task 4; and HW#5, Problem 4.

a. Define three ML problems for your company: an unsupervised learning clustering problem; a supervised learning classification problem; and a supervised learning regression problem. Each problem needs to be first stated in words, and then, more precisely, as a mathematical problem. (These injunctions were explicitly discussed in Lecture #14 on 5/18/17).

b. Clearly describe the stepwise plan or approach for solving each of the three problems in Excel.

2. **Plan the Problem**

- Review Class Lectures
- Review Associate Readings from the text
- Analyze and Organize thoughts
- Develop initial draft
- Finalize
- Check Work

3. **Execute the Problem**

a. Define three ML problems for your company: an unsupervised learning clustering problem; a supervised learning classification problem; and a supervised learning regression problem. Each problem needs to be first stated in words, and then, more precisely, as a mathematical problem. (These injunctions were explicitly discussed in Lecture #14 on 5/18/17).

- **Unsupervised Learning Clustering Problem**:

Description: (what is the Netflix problem we can solve here ??)

- The optimization clustering problem is split into 2 steps.
  (a) the cluster assignment step and (b) the move or find centers step. In step (a) we randomly select the centers for our K clusters. We use the Euclidean distance to measure the distance of a data point from the random centers or centroids and assign them to clusters based on this distance. Then, in part (b) we move the center of the clusters to the new cluster mean based on the data point assignment that we have discovered in part (a). We then repeat this process to find the minimized distance data cluster centers and the clusters to which each data point is assigned.

Math:

- (a) Cluster assignment:
  Randomly assign K clusters and K cluster centers
  K = 5 Clusters = (c_1, c_2, c_3, c_4, c_5) Cluster centers = ($\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$)
  Find which cluster any given data point is closest to by minimal Euclidean distance

$|x_{(i)}^{P} - \mu_{K}^{P}|$: whichever μ yields lowest result is assigned cluster c

- (b) Move and find centers:
  Find the mean of data points assigned to each cluster.
  i.e, if n data points are assigned to cluster c_K then
  $(x^{(i)} + x^{(j)} + x^{(k)} + x^{(l)} + ...) / n$ calculates the new center for each cluster.
  And the overall objective function of the minimization problem is
  $J(c_1, c_2, c_3, c_4, c_5); \mu_1, \mu_2, \mu_3, \mu_4, \mu_5 = \sum_{i=1}^{m} ||x^{(i)} - c^{(i)}||$
  Where J is the total distance from cluster centers to their assigned data points.
  We must find all x in clusters c_1, c_2, c_3, c_4, c_5 and cluster centers $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ that minimize J.

- **Supervised Learning Classification Problem:**
  - **Description:** Consists of inferring a classification based on "training data" given 2 classes ("Recommend" and "Do Not Recommend" .
    - Process
      - Training data is classified into classes
      - x(i) = {Action, Horror, Sci-Fi, Drama, Comedy, etc. (genre); y(i) Recommendation
        - If 1 then recommend film
        - If 0 then do not recommend film
  - **Math:**
    - Probability that user is a fan of another film from that genre
      - P(B/A) = (P(B)*P(A|B))/ (P(A)
    - Probability that user likes the recommended genre.
      - P(B/A) = (P(B)*P(A|B))/ (P(A)

- **Supervised Learning Regression Problem:**
  - **Description:** Consists of inferring what will occur based on "training data". Regression will help to predict the path that something travels based on trend data.

  - **Math:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_n X_n + \varepsilon$$

- Using the regression formula we can predict the retention of viewers for a given movie or show over time based on number of monthly viewers and changes in rating (based on social trends and awareness).

b. Clearly describe the stepwise plan or approach for solving each of the three problems in Excel.

Unsupervised Learning Clustering Problem:

Plan:

Step 1: Read Data Smart and follow steps given in Bag O'Donuts Ch. 2 example
Step 2: Segment the data we've obtained into 5 clusters using a Pivot Table based off of available and relative features (i.e show rating, views, viewer age, time spent on, etc)
Step 3: Run Solver on the data using 5 cluster centers and find out the optimal cluster centers and the optimal cluster center assignments for each data point.

Supervised Learning Classification Problem:

Plan:

Step 1: Read Chapter 3 of Data Smart and steps given in Tweets example
Step 2: Download Required Spreadsheets
Step 3: Walk through Naive-Baye method covered in text book using downloaded inputs
Step 4: Make modification on the spreadsheets, using movie genres instead of tweets and probabilities based on previous selections
Step 5: Determine the recommended shows for a user who has a certain history
-will be based off of relative frequencies, i.e if you watched horror last the odds that you may like an action are more than the odds of a drama and so on.
Step 6: Find a way to integrate show recommendations across multiple users.

Supervised Learning Regression Problem:

Plan:

Step 1: Reread regression chapter in data smart
Step 2: Download Netflix Prize data
Step 3: Assemble the Training Data
Step 4: Create Dummy Variables for Intercepts

4. Check Your Work

All work here matches the work covered in the machine learning lectures. The math equations and excel steps are taken from the "Data Smart" textbook

5. Learn and Generalize

In completing this problem we learned 3 different applications of machine learning within Netflix's core business application. These 3 applications can either be supervised or unsupervised and all have to deal with making predictions. We found it difficult to come up with new ideas because Netflix is

already doing so much with machine learning. In fact they offered a million dollar prize for anyone with a considerably advantageous machine learning algorithm.

Solving the Machine Learning (ML) Problem for Netflix

1. **Define the Problem**

a. Using suitable criteria, select one of the three problems from Task III as the ML problem that you are going to solve in this Phase. Clearly define the problem you plan to solve.
b. Describe the stepwise plan or approach for solving this ML problem,
c. Implement (Execute) your plan using MS Excel.
d. Clearly explain your results and how these results would be used within the organization. Clearly define the users for this ML applications, and provide a suitable use-case.

2. **Plan the Problem**

- By downloading the Netflix Prize Data, which consists of about 100 million moving ratings, we should be able to sort the ratings by movie ID and RatingDate (the date in which the rating was produced by a certain customer for a certain movie). By sorting the elements in this way we can create different clusters of length of popularity. We would be able to see when a movie is first released and it's frequency of ratings until it's last rating.
- Organize the data in a useful way
- Set up a pivot table on movie ID, rating and date.
- Define the Training Examples
-

3. **Execute the Problem**

a. Using suitable criteria, select one of the three problems from Task III as the ML problem that you are going to solve in this Phase. Clearly define the problem you plan to solve.

<u>Selection Criteria:</u>

-Processing Time = duration of processing solutions (high is bad)
-Data Reqs = Data requirements for processing (high is bad)
-Knowledge = Knowledge required to complete problem (high is bad)
-Select the problem with the least total requirements

Clustering Problem

Component	Processing Time	Data Reqs.	Knowledge
Usage	10/10	6/10	6/10

Figure 3.1: Cluster ML Rating

- For the clustering problem, we chose to avoid it because of the time needed for testing. We agreed that this wasn't the best option because there's too much wasted time waiting for results that may be incorrect. We must make use of the time we spend on the project and waiting for cluster processing seemed illogical.

Classification Problem:

Component	Processing Time	Data Reqs.	Knowledge
Usage	5/10	10/10	8/10

Figure 3.2: Classification ML Rating

- For the classification problem, we simply could not get our hands on the data we desired. We still can try to pursue this problem in the future, but we cannot tackle it with the data that we have now, and instead had to use IMDB data. This is the machine learning problem that we wish to apply to Netflix.

Regression Problem:

Component	Processing Time	Data Reqs.	Knowledge
Usage	2/10	7/10	9/10

Figure 3.3: Regression ML Rating

- For the regression analysis problem, we had the data that we needed, and regression is much faster on Excel than clustering is, so processing time is drastically reduced. To add to this, our group is most familiar with regression due to previous classes.

b. Describe the stepwise plan or approach for solving this ML problem,

Plan:

Step 1: Read Chapter 3 of Data Smart and steps given in Tweets example
Step 2: Download Required Spreadsheets
Step 3: Walk through Naive-Baye method covered in text book using downloaded inputs
Step 4: Make modification on the spreadsheets, using movie genres instead of tweets and probabilities based on previous selections
Step 5: Determine popularity of a show or movie based on relative frequencies
Step 6: Find a way to integrate show recommendations across multiple users.

c. Implement (Execute) your plan using MS Excel.

Step 1: Read Chapter 3 of Data Smart and steps given in Tweets example

-Did so using Data Smart Text

Step 2: Download Required Spreadsheets/ Data

We accessed the IMDb (Internet Movie Database) 5000 database from Kaggle, Which contains 5,000 popular films.

Our Database has the following attributes

- Movie Title
- Content Ratings
- Director's Name
- Actor's Name
- Genre
- Plot Words

Step 3: Walk through Naive-Baye method covered in text book using IMDB data inputs

-Followed steps in the text.

Step 4: Make modification on the spreadsheets, using movie genres instead of tweets and probabilities based on previous selections

Given this data we implemented the Naive Bayes classifier to categorize the films by rating, GOOD or BAD (GOOD if IMDB rating ≥ 7.5 and BAD if IMDB rating < 7.5) based on Director, Actors, and movie plot keywords.

Step 5: Determine popularity of a show or movie based on relative frequencies

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	Number	Rating	Prediction	MovieWords													
2	1	BAD	GOOD	Leonardo-DiCaprio	Liam-Neeson	Jim-Broadber	Martin-Scorse	butcher	civil	war	gangster	new	york	city	revenge		
3	2	GOOD	BAD	Christian-Bale	Ni-Ni	Shigeo-Koba	Yimou-Zhang	abusive	stepfather	attempted	rape	food	shortage	sexual	abuse	starving	ch
4	3	BAD	GOOD	Jeff-Bridges	Zooey-Desch	Jon-Heder	Ash-Brannon	chicken	competition	island	penguin	surfing					
5	4	BAD	BAD	Jon-Lovitz	Matthew-Broc	Roger-Bart	Frank-Oz	community	connecticut	fem	bot	tv	producer	writer			
6	5	GOOD	BAD	Ioan-Gruffudd	Sam-Sheparc	Ewen-Bremne	Ridley-Scott	army	helicopter	somali	somalia	warlord					
7	6	BAD	GOOD	Will-Ferrell	Thomas-Midd	Katherine-Laf	Jay-Roach	campaigning	congressman	north	carolina	title	at	the	end	u.s.	cc
8	7	GOOD	BAD	Milla-Jovovich	Bruce-Willis	Gary-Oldman	Luc-Besson	1910s	alien	artificially	created	woman	love	taxi	driver		
9	8	BAD	BAD	Chris-Noth	Liza-Minnelli	Kristin-Davis	Michael-Patrik	abu	dhabi	box	office	hit	muslim	nanny	united	arab	er
10	9	BAD	BAD	Frank-Welker	Rosie-Perez	Elton-John	Bibo-Bergeroi	adventurer	el	dorado	gold	high	priest	implied	sex		
11	10	BAD	BAD	Peter-Dinklage	Josh-Gad	Drake	Steve-Martinc	acorn	herd	iceberg	ocean	pirate					
12	11	BAD	BAD	Hayley-Atwell	Derek-Jacobi	Lily-James	Kenneth-Brar	dress	duke	fairy	godmother	fairy	tale	pumpkin			
13	12	BAD	BAD	Michael-Imperioli	AJ-Michalka	Tom-McCarth	Peter-Jackso	1970s	afterlife	heaven	pedophile	rape					
14	13	GOOD	GOOD	Alexander-Gould	Stephen-Roo	Brad-Garrett	Andrew-Stant	great	barrier	reef	protective	father	separation	from	family	shark	sh
15	14	GOOD	GOOD	Orlando-Bloom	Billy-Boyd	Bernard-Hill	Peter-Jackso	battle	epic	king	orc	ring					
16	15	GOOD	GOOD	Christopher-Lee	Orlando-Bloo	Billy-Boyd	Peter-Jackso	epic	evil	wizard	middle	earth	ring	wizard			
17	16	BAD	BAD	Jeff-Bridges	Djimon-Houn	Olivia-William	Sergey-Bodrc	apprentice	demon	exorcism	master	apprentice	relationship	witch			
18	17	BAD	BAD	Angelina-Jolie-Pitt	Noah-Taylor	Chris-Barrie	Simon-West	illuminati	planetary	alignment	time	tomb	tomb	raider			
19	18	BAD	BAD	Johnny-Depp	Morgan-Freer	Clifton-Collins	Wally-Pfister	artificial	intelligence	consciousnes	power	outage	scientist	technology			
20	19	BAD	BAD	Michael-Jeter	Trevor-Morga	Alessandro-N	Joe-Johnston	dinosaur	island	jurassic	park	paleontologis	search				
21	20	GOOD	BAD	James-Franco	David-Oyelow	Tyler-Labine	Rupert-Wyatt	alzheimer's	disease	ape	chimpanzee	fire	when	animals	attack		
22	21	BAD	BAD	Martin-Short	Tod-Fennell	Joan-Plowrigl	Mark-Waters	actor	playing	multiple	roles	brownie	the	creature	closing	credits	se
23	22	BAD	BAD	Bruce-Willis	Cole-Hauser	Megalyn-Echi	John-Moore	bomb	cia	courthouse	escape	russian					
24	23	BAD	BAD	Dennis-Quaid	Marc-Blucas	Jordi-Mollà	John-Lee-Har	army	dictator	general	texan	texas					
25	24	GOOD	BAD	Holly-Hunter	Craig-T.-Nels	Lou-Romano	Brad-Bird	hero	island	lawsuit	secret	superhero					
26	25	BAD	GOOD	Christopher-Masters	Frank-Langell	Matthew-Mod	Renny-Harlin	latin	pirate	pirate	ship	treasure	treasure	map			
27	26	BAD	BAD	Logan-Lerman	Rosario-Daws	Steve-Coogai	Chris-Columb	greek	lightning	lightning	bolt	poseidon	teenager				
28	27	BAD	BAD	Will-Smith	Rip-Torn	Linda-Fiorenti	Barry-Sonner	alien	box	office	hit	flying	saucer	laser	gun	wisecrack	hu
29	28	GOOD	BAD	Tom-Hanks	John-Ratzenb	Wayne-Knigh	John-Lassete	collector	dog	friend	rescue	toy					
30	29	BAD	BAD	Denzel-Washington	Rosario-Daws	Ethan-Suplee	Tony-Scott	freight	train	race	against	time	runaway	train	train	train	er
31	30	BAD	BAD	Mel-Melançon	John-Lone	Harris-Yulin	Brett-Ratner	boat	gang	hong	kong	triad	vacation				
32	31	BAD	BAD	Harrison-Ford	Amber-Vallett	Miranda-Otto	Robert-Zeme	ghost	haunted	house	research	secret	vermont				
33	32	BAD	BAD	Will-Forte	Bobb'e-J.-Thc	Al-Roker	Phil-Lord	food	giant	food	mayor	sardine	weather				
34	33	BAD	BAD	Denis-Leary	Maile-Flanag	Kelly-Keaton	Carlos-Saldar	egg	lost	world	rescue	squirrel	weasel				
35	34	BAD	BAD	Adam-Scott	Adrian-Martin	Joey-Slotnick	Ben-Stiller	daydream	life	magazine	magazine	photographer	snow	leopard			
36	35	BAD	BAD	Bill-Murray	LL-Cool-J	Kelly-Lynch	McG	booty	shake	box	office	hit	duct	tape	over	mouth	fir
37	36	GOOD	BAD	Leonardo-DiCaprio	Matt-Damon	Ray-Winstone	Martin-Scorse	boston	mole	police	undercover	undercover	cop				
38	37	BAD	BAD	Ming-Na-Wen	Harvey-Fierst	June-Foray	Tony-Bancrof	based	on	poem	based	on	TRUE	story	china	one	w
39	38	BAD	GOOD	Robert-Downey-Jr.	Steve-Coogai	Brandon-T.-Je	Ben-Stiller	film	director	parody	spoof	vietnam	written	and	directed	by	ca
40	39	GOOD	BAD	Robin-Wright	Goran-Visnjic	Joely-Richard	David-Fincher	computer	hacker	hacker	investigation	journalist	punk				
41	40	GOOD	BAD	Bruce-Willis	Aldis-Hodge	Kevin-Chamb	John-McTiern	bomb	detective	new	york	city	police	terrorist			

Figure 3.4: Classifier application output

d. Clearly explain your results and how these results would be used within the organization. Clearly define the users for this ML applications, and provide a suitable use-case.

We can see that after trained, our machine learning program was able to learn what movies would receive an IMDb rating less than 7.5 or more than 7.5. With over 100 samples we got a total of 79% correctness from our Naive Bayes machine. With this data Netflix can see the trend that comes out of combination of actors, directors, and plot type to assist them with picking movies to broadcast on their servers or with creating original content.This will be very useful for Netflix employees and Netflix directors, as they will be able to classify the future success of shows with great accuracy.

4. Check Your Work

We checked our work using methods learned in our text and in class and based on those references we conclude that it is correct and concise. All steps shown here match the steps in the book, and we chose 7.5 as the accuracy of the classifier

5. Learn and Generalize

In completing this problem, we have learned how to apply Naive Bayes Classification to modern, everyday tasks. If Netflix were to use this model, our classifier for successful movies would be able to predict the success of movies to nearly an 80% accuracy. This may be useful for Netflix in the future, as this can help company workers and movie producers determine the possible popularity of their movies or shows before they are even released. Obviously there are many other factors that we could not possibly account for with the data set that we had, but the process and methods can be applied to classify clients or objects in various industries.

Conclusion:

Our Teams Performance

At the start of the group project, our group was not good at communicating and planning. This caused us to fall behind in the early phases and forced us to meet up multiple times a week in order to get back on track. The primary problem was that we attempted to implement an ML learning process for Netflix when we were supposed to define Netflix's basic process. Once we corrected this mistake, the project was much smoother and understandable. By the end of the project, our group was able to meet up in order to contribute deliverables and evaluate each other's performance in a much more efficient way. Although our project has room for improvement, we are all satisfied with the end result and have input countless hours into this project.

Processes Developed

In working together, our group developed several process for database design, creating and editing SDI (software defined infrastructure), placing virtualization and cloud computing within the SDI and implemented all of this knowledge into a machine learning process for Netflix movie and show classification in Excel

Machine Learning Solutions (Successes and Failures)

When attempting to integrate a Machine learning problem into our group's database skeleton of Netflix, we ran into various issues. The first major problem was the fact that data is not easily obtainable or given away for free by firms. This forced us to use proxy data and make assumptions about Netflix's movie database using IMDB's movie database. After finding the suitable proxy data, our group was then faced with the problem of selecting and performing a machine learning algorithm. At first, we decided to do regression but quickly changed it to classification when we learned that 12 out of 14 groups chose regression

Machine Learning Solutions (Analyze Key Results)

If Netflix were to use this model, our classifier for successful movies would be able to predict the success of movies to nearly an 80% accuracy. This may be useful for Netflix in the future, as this can help company workers and movie producers determine the possible popularity of their movies or shows before they are even released. Obviously there are many other factors that we could not possibly account for with the data set that we had, but the process and methods can be applied to classify clients or objects in various industries.

What our team learned

Through our work our team learned not only how to design and implement a full IT diagram or how to implement machine learning to improve business performance but how to work effectively as a unit. Our work phase to phase improved greatly as well as our ability to communicate and adjust to problems as a unit, not as individuals. At the same time we learned the technical skills that a company would utilize to create an efficient IT system given the modern tools of 2017. Hopefully these technical and team skills will be useful in the job field or will be applicable in future data analysis.