Assignment

Cloud Computing And Big Data Analytics for Cost Optimization

NAME : J. Janessa

REGISTER NO : 192424017

COURSE CODE : CSA1057

Assignment - 1

1. Explain how cloud based models reduce capital investment in IT infrastructure justify your answer with real world Examples.

cloud based models reduce capital investment by shifting Expenses from CAPEX to OPEX, instead of purchasing servers, storage, and networking equipment, businesses rent these resources from cloud providers like AWS, Azure or Google cloud this Eliminates large upfront costs, reduces maintenance burden and offers scalable flexible IT solution that grow with demand.

EXAMPLE : startups can launch apps without buying servers - just pay for cloud usage saving money and time

a) compare CAPEX Vs OPEX in the context of cloud computing.

Feature	CAPEX	OPEX
Definition	upfront investment in Physical assets	ongoing Expense to run services
Example	Buying servers, data centers	Paying monthly cloud bills
Flexibility	low - assets are fixed	High - pay as you use
Risk	High due to hardware depreciation	low - scale based on demand.

b) How cloud models Eliminate Hardware Acquisition Costs.

cloud provides own and maintain the physical infrastructure. Businesses access computing resources remotely.

Eg: storage, computer power.

This Eliminates the need for.

=> Purchasing expensive servers or networking gear

=> Building or renting physical data centers

=> Hiring large IT staff for maintenance.

A startup using Google cloud doesn't need to buy a server rack. instead, it provisions a virtual machine in minutes. Paying only for the resources used.

c) Explain subscription and consumption - Based pricing.

=> subscription - Based pricing

users pay a fixed fee for a defined Package of services. it offers predictable costs and is suitable for consistent workloads.

Example :-

microsoft 365 charges a fixed monthly fee per user for access to office apps and cloud storage.

=> Consumption Based Pricing.

users are billed based on actual resource usage like compute time, storage, or bandwidth. ideal for businesses with fluctuating demand.

Example :-

AWS charges per second for EC2 instance usage and per GB for S3 storage.

d) case study: cloud Reducing captial Expenses

=> Netflix migrated from physical data centers to amazon web services

=> This shift helped avoid investing in Expensive server infrastructure.

=> Netflix pays AWS only for the resources it uses during Peak viewing times

=> No hardware maintenance

=> Elastic scaling based on demand

=> Reduced CAPEX, increased Efficiency

e) How Does this Benefit small to mid sized Enterprises.

cloud Computing offers major benefits to SMEs:

=> low initial cost: No need to invest in Expensive hardware

=> scalability: start small and Expand resources as needed.

=> Access to advanced tech: SMEs can use AI analytics and security tools through cloud platforms.

=> Reduced IT overhead: No need to hire large IT teams or manage physical server.

Example:

A small business can run its website on google cloud or use shopify to manage online sales without owning any servers.

Assignment - 2

2. A data center migration for your main Question and all sub Questions regarding risks and strategies in a data center migration to a virtual environment.

During a data center migration to a virtual Environment, Key risks to assess include data loss, downtime, compatibility issues, security vulnerabilities and staff skill gaps. mitigation involves careful planning, robust backups phased migration and updated security and training Protocols.

a) identify Key technical and operational risks in virtualization.

=> Hypervisor Vulnerabilities : Exploits at the virtualization layer

=> Resource contention: multiple VMs competing for limited CPU, memory or storage

=> Configuration errors : misconfigurations in virtual networks or storage settings.

=> Driver or software incompatibility: legacy apps may fail on new virtual platforms.

operational Risks: lack of trained personnel, Backup and recovery Process gaps

b) How can downtime and service disruption be minimized.?

=> Phased migration : move systems gradually to isolate and fix issues quickly.

=> live migration Tools: use tools like VMare Vmotion or Hyper-V live migration

=> load Balancers : Redirect traffic automatically during VM failovers of migration.

=> Redundancy and failover: set up high availability disaster recovery.

=> Pre-migration testing: conduct simulations to detect live migration.

c) suggest strategies for data integrity and compatibility.

=> Regular Backups: Full and incremental backups before and migration

=> checksum validation: verify file-level integrity after transfer occur.

=> Snapshot Rollback: use VM snapshots to restore quickly if Errors

=> Application Assessment: Test apps on virtual platforms before migration

=> use compatibility layers: use virtual hardware abstraction to support legacy systems

d) propose a change management plan for staff adoption.

=> stakeholder communication: inform staff of goals, timelines and Expectations

=> Training sessions: provide technical training on virtualization tools and best practices.

=> Role Redefinition: update job responsibilities to align with virtualized workflow

=> Pilot Teams: Form early adopter teams to guide others and troubleshoot early

=> Feedback mechanism: collect and address staff concerns throughout the transition.

e) How do security policies evolve in virtualized environments:

=> micro segmentation : Isolate workloads at the VM level using virtual firewalls.

=> zero Trust model : verify every user and device using virtual regardless of location

=> hypervisor hardening: Apply strict access controls and patch hypervisors regularly.

=> virtual network monitoring : use tools to inspect traffic between VMs

=> policy Automation: Automate Compliance checks and security Configurations

Assignment - 3

8. Analyze the infrastructure required for secure and reliable access to a cloud hosted library system.

A secure and reliable cloud hosted library system requires robust infrastructure, including identity management role based access, encrypted data transmission, redundant cloud storage and real-time data sync tools. It must also include backup, disaster recovery and continuous monitoring to Ensure availability and security.

a) what are user authentication and role based access requirements?

=> use multi factor authentication for all users

=> integrate with OAuth 2.0, SAML or openID connect.

=> Ensure password policies

=> Define user roles such as admin, librarian, student, Guest

=> Grant permissions based on roles only librarians can add remove content.

=> Enforce least privilege principle - users access only what they need.

b) How can cloud file storage provide high availability?

=> Redundancy: cloud providers replicate data across multiple data centers.

=> Auto healing storage: systems like AWS S3 and Azure Blob storage automatically replace corrupted data.

=> load Balancing: Distributed access requests to Ensure consistent performance.

=> content delivery Networks: improve access speed and availab globally.

=> Uptime SLA Guarantees: use providers offering 99.9% + availability

c) Discuss secure access using HTTPS, VPN and identity federation

HTTPS:

=> Encrypts all communication between users and the library system.

=> protects login credentials and sensitive user data.

VPN:

=> Ensures secure remote access for internal staff

=> shields data from being intercepted on public networks.

identity federation:

=> Allows users to log in using institutional or third party credentials.

=> centralized identity with SSO improves usability and security.

d) suggest tools for real-time sync metadata, and indexing

Real-time sync:

Firebase AWS Appsync. or google drive. API for instant updates across clients.

metadata management:

Apache Tika or Elastic search to Extract and manage metadata from documents.

Indexing:

=> Elastic search or Algolia to provide fast, full text search.

=> Apache solr for advanced indexing and query capabilities.

Explain database backup and recovery strategy in this context.

Regular Automate Backups:

=> Daily backups of library content and user data.

=> Versioning enabled for documents.

multi region Replication

=> copies of data stored in different geographic regions to Ensure disaster resilience.

Recovery plan:

=> use snapshots based recovery

=> Define RTO and RPO to meet service expectations

Tested DR Procedures:-

=> Regularly test backup restoration to Ensure readiness in real failure scenarios.

Assignment - 4

Explain the tools used for big data integration and how they help manage data variety.

Big data integration tools help combine data from diverse sources semi-structured, structured and unstructured formats. Tools like Apache Nifi, Talend and Flume handle data ingestion, transformation, and routing enabling efficient integration across schema mapping, real-time sync and metadata management to reduce complexity and ensure consistency.

a) what is data integration in big data and why is it challenging?

Data integration in Big Data:

it is the Process of collecting, combining and transforming data from multiple sources into a unified view for analysis or storage in big data platforms like Hadoop or cloud data lakes.

challenges:

=> Data variety: Diverse formats (JSON, XML, CSV, SQL etc)

=> High volume & velocity: Real time ingestion and processing is complex.

=> Data Quality & consistency: cleaning and standardizing inconsistent data.

=> schema Evolution: Handling changing structures over time

=> Heterogeneous sources: integrating from databases, APIs, IoT, web, etc.

b) Compare tools like Apache Nifi, Talend and Flume

Feature	Apache Nifi	Talend	Apache Flume
Type	Flow - based Programming, GUI	ELT/ELT tool with GUI	Event based log ingestion.
Strength	Real - time data - flow, drag-and-drop UI	Rich data transformation, big data support	Reliable streaming of log data to HDFS
Use case	IoT, real-time routing & transformation	Complex data Pipelines, batch Processing	Ingesting logs from web servers
Streaming support	yes	limited	yes
Batch support	yes	yes	limited
integration	REST, MQTT, Kafka, DBS, cloud	Hadoop, spark, AWS, DBS	HDFS, Kafka, Hadoop.

c) How do these tools support batch and streaming integration ?

=> Apache Nifi : supports both batch and streaming via flow based architecture, prioritization, Queuing and back pressure mechanisms. ideal for real time data routing and transformation.

=> Talend : primarily strong in batch processing with rich connectors. Talend Real - time Big Data version supports streaming via spark streaming or Kafka.

Apache Flume : specializes in streaming log / event data to Hadoop or HDFS. Not suitable for complex transformation or batch.

describe how schema transformation and mapping work.

=> schema Transformation : converts data from one format or structure to another.

mapping Process :

=> Identify source and target schemas.

=> Define rules or mapping (eg: rename, combine, split fields)

=> Apply data type conversions (eg: string -> integer)

=> use tools like Talend or NiFi processors to apply transformation.

Dynamic schema Handling :

=> Tools like NiFi can adapt to changing schemas using Record based processors and schema registries.

e) Provide an example use case of integrating structured and semi structured data.

use case : online retail Analytics

sources :

=> structured : Relational DB (eg., mysql) containing user orders

=> semi-structured : JSON logs from a website or app tracking user behavior.

Integration steps :

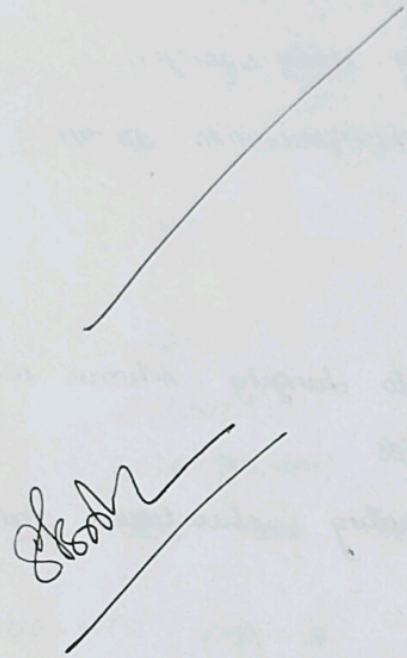
=> NiFi ingests both datasets in real time

=> uses processors to convert JSON logs into a structured format

=> Joins the structured order data with the behavioural logs based on user ID

=> Transforms data into a unified schema

=> Stores the final integrated data in Hadoop HDFS or a cloud data lake.



Assignment - 5

1) What is the role of YARN in managing Hadoop jobs, and how does it improve performance?

YARN is the resource management layer of Hadoop that manages and schedules computing resources across applications. It improves performance by decoupling resource management from job processing allowing multiple data Processing Engines to run simultaneously, leading to better scalability, resource utilization, and fault isolation.

a) Describe the YARN architecture including RM and NM.

=> Resource manager :

=> Central authority that allocates resource to applications

=> scheduler : Allocates resources based on policies

=> Application manager : manages application submissions and tracking

=> Node manager :

=> Runs on each node in the cluster

=> Reports node health and resource usage to the RM

=> Launches and monitors containers where tasks execute

=> ApplicationMaster :

=> launched per application

=> manages tasks Execution and Coordination within the applicat

=> Containers :

=> lightweight environments where actual processing tasks run.

b) How does YARN differ from original mapreduce in scalability?

Features	original mapreduce	YARN
Architecture	monolithic	Decentralized
Resource manager	single job Tracker	Resource manager handles global resources
scalability	limited due to Centralized control	Highly scalable via distributed design
multi-Framework	only mapReduce jobs supported	supports Spark, Tez, Flink etc.
Fault tolerance	low	Improved isolation and recovery mechanisms

c) Explain job scheduling in YARN (Fair, FIFO, Capacity)

1. FIFO scheduler:

=> First in First out

=> Jobs are Queued and Executed in order of submission

2. Fair scheduler:

=> Allocates resources evenly across running jobs

=> Ensures that no single job monopolizes resources.

=> suitable for multi-user environment.

3. capacity scheduler:

=> Divides cluster into queues with minimum resource guarantees

=> Each Queue gets a capacity and allows job priority within it

=> supports multi tenant environments

does YARN support multi framework workloads.

=> YARN is Framework - agnostic it can run map-Reduce, Apache Spark, Tez, Hive, Flink etc.

=> each application launches its own applicationmaster to manage tasks

=> Containers can run different types of workloads

=> YARN's resource sharing allows multiple frameworks to coexist and compete fairly.

e) what are benefits of resource utilization and fault isolation?

Improved Resource utilization:

=> Dynamic allocation of memory and CPU to containers

=> unused resources can be reassigned quickly to other tasks

=> Avoids overprovisioning - leading to cost savings.

Fault Isolation:

=> Failures in one application do not affect others

=> Application Masters are independent so one failing doesn't crash the cluster

=> Better diagnostics and automatic retries per job