cloud computing

1. models – infrastructure as a service, platform as a service, software as a service
2. cloud deployment models – public cloud, private cloud, hybrid cloud, community
3. big data characteristics – volume, velocity, variety, veracity
4. big data cloud technologies – compute, storage, databases, data warehouses, data lake, data lake house
5. cloud data processing – batch processing, stream processingin
6. data processing frameworks – Hadoop, spark, Apache Kafka
7. data pipelines – ELT, ETL
8. cloud- native big data services – amazon EMR, Google Dataflow, Azure HDInsight
9. cloud computing benefits – cost efficiency, scalability, flexibility, disaster recovery, automatic updates, storage, processing, machine learning
10. cloud computing challenges – security concerns, compliance issues, latency, vendor lock – in
11. big data types – structured data, unstructured data, semi – structured data
12. big data challenges – storage, processing power, quality, security and privacy, real time analytics
13. big data technologies- Hadoop (HDFS, MapReduce), Apache Spark, Apache Kafka, NoSQL Databases(Mongo DB, Cassandra)
14. data life cycle – ingestion, storage, processing, analytics, visualization
15. data pipeline challenges – Latency, cost Management, Data integrity, security

cloud native big data services

1. Amazon Web Services

* AWS S3(simple storage service, storage)
* AWS EMR (Elastic MapReduce, processing)
* AWS Redshift(Warehouse)
* AWS Glue(serverless data integration service)
* AWS Lambda (automatic serverless real time data compute service)
* AWS kinesis (collect, process, analyze real time streaming data)
* AWS Data Pipeline (movement and transformation of data between aws compute and storage services

1. Google Cloud Platform

* GCP BigQuery (serverless data warehouse, super fast SQL processing)
* GCP Dataflow (real time analytics on stream data and batch processing data on Apache beam)
* GCP cloud storage (unstructured data, data lake)
* GCP Dataproc (for deploying Apache Hadoop and Apache Spark clusters in minutes)
* GCP Pub/Sub (real time event stream messaging service)

1. Microsoft Azure

* Azure HDInsight (open source service, runs Hadoop spark Kafka
* Azure Synapse Analytics( limitless relational and non-relational data analytics service)
* Azure Data Lake Storage (structure and unstructured data storage)
* Azure event hub (super fast real time and batch processing platform, event ingestion service)
* Azure Data Factory (data integration service, create, schedule and orchestrate ETL and ELT workflows.

1. Others

* Apache NiFi – real time data flow open source tool, data ingestion, transformation, management

Data warehouses

1. Characteristics – structured data, schema on write( only clean, consistent data into system), optimized for complex SQL queries, long term historical data
2. Advantages – performance, data consistency, suitable for business intelligence
3. Disadvantages – cost, rigid structure, predefined schema, unsuitable for semi and unstructured data

Data lake

1. Characteristics- structured, unstructured, semi – structured data, schema on read (raw data is structured only when it is read or queried), scalable storage, real time and batch processing
2. Advantages – Flexibility, cost effective storage, supports diverse use cases
3. Disadvantages – data quality issues, complexity

Data Lakehouse

1. Combined strength of data lake and data warehouse
2. Storage flexibility of data lake with structured querying capabilities of data warehouse.
3. Examples – Databricks, AWS Redshift Spectrum.

Data security and Governance in cloud architecture

1. Data security – confidentiality(authorized users access sensitive data), integrity(accurate consistent data), availability
2. Data governance – management of data security (policies, roles, procedures)

* Data stewardship (responsible management of data)
* Compliance (adherence to laws, regulation, and policies
* Data quality (accurate, consistent, and complete data across organization)

1. Security challenges

* shared responsibility model (cloud, customer),
* data breaches (misconfiguration, week access controls, vulnerabilities),
* data loss
* Insider threat
* Compliance and privacy (general data protection regulation) and HIPAA(health insurance portability and accountability act)

1. Cloud security practices
2. Encryption

* Encryption (encoding data in such a way that only authorized parties can access it
* At rest data encryption, In transit data encryption
* Examples – AWS key Management services (KMS), Google cloud key management

1. Identity and access management (IAM)

* Polices that controls access and actions a user can perform
* Examples – AWS IAM, Azure active directory

1. Network Security

* Secures communication channel between cloud resources and user
* AWS VPC, Azure Network Security Group (NSGs)

1. Monitoring and Logging

* Real time monitoring, automated alerts, maintaining logs of access and activities
* Examples – AWS CloudTrail, Google Cloud Operations Suite

1. Backup and Disaster Recovery:

* Back up critical data and test disaster recovery plans
* AWS Backup, Azure Backup

1. Data Governance Practices
2. Data classification

* (public, internal, confidential, restricted)
* Examples – google cloud data loss prevention (DLP), AWS Macie

1. Data access policies

* Rules and policies to govern access
* Examples (Azure Role-based access control (RBAC), AWS IAM Roles

1. Data retention and deletion policies

* Managing the lifecycle of data
* Examples- Google cloud Retention polices, AWS S3 Object Life cycle Polices

1. Compliance and Auditing

* Ensuring that data management practices comply with regulation like GDRP, HIPAA and PCI-DSS
* Examples AWS Artifact, Google Cloud Compliance Reports

Interconnected cloud Services

1. APIs and Microservices

* APIs (different cloud services or applications communicate by sending and receiving data, enables integrations between services such as storage and databases and machine learning tools)
* Microservices architecture (microservices architecture breaks down applications into smaller independent services that interact through APIs. Each service is focused on specific function such as authentication, data processing.

1. Cloud orchestration and automation

* Cloud orchestration (automated coordination of multiple cloud services and resources, ex- automating data pipeline that extracts data from multiple sources processes it and loads it into a data warehouse.
* Serverless functions (serverless computing allows you to run code in response to events without provisioning or managing servers. Automatically scales based on the workload

1. Real time data streaming and integration

* Stream processing with event driven architectures
* Integration platforms

1. Hybrid cloud and multi cloud architecture

* Hybrid cloud (allows data and application move between public and private cloud environments (AWS Outposts, GCP Anthos ))
* Multi cloud, using multiple public cloud services from different provider to avoid dependency on a single vendor, (Azure Arc, HashiCorp Terraform