**Midterm – Part 1: Theory**

**Q1’s Ans:**

1. **Structured, Semi-structured and unstructured:**
   1. **Structured data:**

The data is said to structured when the values are organized such that they can be accessible or retrievable. In the traditional relational systems, data is structured in the form of tables. For example, Oracle database, MySQL, MS-SQL etc. In Oracle DB, data is represented in rows and columns. There are also other tools where you can see structured data like Microsoft Excel. Structured data is easy to manage, read and migrate during any kind of operations on data. It would be easy to do analysis on structured data when compared to other types of data.

* 1. **Semi-structured Data:**

The data in semi-structured format do not have proper format however it is flexible to format in the way we need. Examples of semi-structured data are HTML, JSON. We can also take Emails which we use on daily basis as an example which contains ‘To’ address, subject line and body however you can add text, attachment or image in the body, which is flexible to us. Since semi-structured data has some format to it, hence it is bit easy to analyze.

* 1. **Unstructured Data:**

The data without any defined structure is said to be unstructured data. 80-90% of the data collected by businesses is unstructured. Doing analysis on real-time data will give you more accurate results. Hence the problem for Bigdata is 80% of data available for analysis is of unstructured type. As this data cannot be stored in relational databases, it has other ways to store and manage like schema-free and schema-on-read kind of databases. NoSQL is best for flexible storage. Apache Hadoop can store both structured and unstructured data.

1. Types of Storages
   1. **Block storage:**

Block storage is a form cloud storage used to store data. Data is stored in the form of blocks and each block is assigned with unique identification number for addressing. Since we have addressing capability, it does not matter file type, which means we can store any kind of data. Block size is determined by the system based on data received. Few benefits of block storage are flexibility, higher efficiency, and compatibility. Few limitations are block storages are expensive as it provides many features and there are few performances limitation when it comes improper partition into blocks. Applications of block storage are Databases, email servers, etc.

* 1. **Object storage:**

Object storage is also a cloud storage which stores data as objects. Data will be stored in the storehouses instead of folders and assigns unique identifier to it. This identifier resembles as URL to find the file in the storage. Data is stored without breaking it which makes easy to organize it among other data. There are benefits like cost-effective, faster data retrieval, highly scalable. Object storage finds its application when it comes to rich quality media storages.

* 1. OLAP: OLAP (Online Analytical Processing) refers to processing huge amounts of data for making key business decisions. It works on data warehouses, which means data will be collected from various sources. The data sizes are approximately in Zettabytes and Petabytes. OLAP is used by senior managers, CEOs, and top-level management. OLAP is crucial for key business decisions and problem solving. OLAP is required for performing AI and ML jobs, business intelligence.
  2. OLTP: OLTP (Online Transactional processing) handles transaction activities on data. Data storing and retrieval will be taken care by OLTP. Query processing will very fast on OLTPs. It does not require huge amount of storage, but Megabytes and Gigabytes are required. OLTP is used by relationship managers, clerks in the business.

1. **GCP Services**
   1. **Cloud Storage:** Cloud storage is the managed service provided by GCP for storing unstructured data. Cloud storage looks very similar to File system in laptops or PCs. We can create cloud storage units named buckets to store data files or folders irrespective of their types. Generally, all the files will be stored as blobs. We can associate cloud storage buckets to the clusters wherein cloud storage acts as file system for clusters. There are wide range of configurations that can be applied on buckets like storage classes, backup mechanism, region, and location etc.

Pros:

1. Enhanced security with configurations
2. Flexibility of usage and fast transfer services
3. Data can be easily applied to AI/ML models

Cons:

1. Not totally control for clients
2. Requires internet to access files
   1. **Dataproc:** It is service provided by GCP for managing Hadoop and Google Kubernetes Engine ecosystem on GCP. Dataproc runs multiple frameworks like Apache Hadoop, Spark etc. Dataproc also manages orchestration of compute engines and maintenance required by Hadoop. Dataproc can be used for easy and quick way to implement Hadoop framework. With the help of Dataproc, you can get access to all Hadoop services. It provides various configuration options like number of nodes, scaling, operating system, etc.

Pros:

* + - 1. Low cost for establishing infrastructure
      2. Improved scalability
      3. Flexibility to choose nodes as you need without overhead

Cons:

1. Unable to change OS of existing cluster
   1. **CloudSQL:** CloudSQL is fully managed PaaS provided by GCP for hosting MySQL, PostgreSQL, and SQL Server. It allows you to choose multiple versions of database, which makes easy to migrate on-premises database to cloud. You can customize the instance as you need from storage required to access control, location, version, etc. It has a capability to perform easy migration into cloud.

Pros:

* + - 1. Database scalability
      2. High security
      3. Low-cost infrastructure

Cons:

1. Limited vendors (only 3 are available now)
2. More latency
   1. **BigQuery:** BigQuery is a fully managed, service-less solution provided by GCP for data warehousing. It processes several Petabytes of data. It is easy to setup and use BigQuery for analytical processing applications. BigQuery supports SQL like queries. It is optimized to process queries quickly even the data size is huge. As it contains datasets, you can easily organize data inside it.

Pros:

* + - 1. Pay-as-Use service
      2. SQL like queries
      3. Easily connects to third-party applications for data operations

Cons:

1. Bit expensive when compared to competitors

**Q2’s Ans: IaaS, PaaS and SaaS**

As the cloud computing technology paced up and various businesses started using it. Cloud service providers started implementing three broadly classified services along with others – IaaS, PaaS, and SaaS

1. **IaaS (Infrastructure as a Service):** IaaS provides servers, operating systems, databases, and networking through virtualization technology. These various services are based at one location and can be used from any parts of the world using cloud technology. These services can be accessed via APIs. The client can have control on infrastructure. Since are services you would need based at same location, the latency will be very low, and infrastructure will secure. The best part is cost – Costs are incurred based on usage of the services.
2. **PaaS (Platform as a Service):** PaaS provides you a platform which can be mainly used to run specific applications. While IaaS, provides you complete infrastructure of all services you need, PaaS mainly limited to an application. This service can be accessed using APIs. This service is simple, scalable, cost-effective, and easy deployment of applications. Clients need not to worry about maintenance of servers.
3. **SaaS (Software as a Service):** SaaS provides complete software which is ready to use for the clients. Customers can buy software for the provider and can start using it. Clients do not have burden of maintaining hardware and software related issues and upgrades.

|  |  |  |
| --- | --- | --- |
| **IaaS** | **PaaS** | **SaaS** |
| Complete infrastructure required for the businesses to run software applications is provided | Server-less platform is provided to run applications by the business. Easy to deploy and run the applications | On-demand Software offered by cloud providers to the businesses which can be used readily |
| Clients have full control on services | Clients do not have full control, but they can manage their applications on the platform | Clients do not have any control on the application or software |
| Services are highly scalable | Platform is highly scalable based on their application | Not so scalable as others |
| Cost will be calculated based on usage of services | Cost will be incurred upon usage of services | Cost-effective and charged per application |
| All services under IaaS are dynamic and flexible based on business functionality | Highly available with integrated services and databases | Accessible over internet, no need to have technical staff |
| Maintenance will be carried out by clients | Maintenance about platform is least concern | Negligible maintenance required |
| E.g.: Complete software solutions with Servers, storage, networking.  GCP, AWS and MS Azure | E.g.: GCP App Engine  Google CloudSQL | E.g.: Business transaction applications, tax software, eCommerce applications |

**Q3’s Ans:** **Difference between Standard, Nearline, Coldline, and Archive storage**

1. **Standard:**

Standard storage class in Google cloud storage is mainly used for frequent access of data files on cloud storage. In standard, the data is stored for the small duration. Data in this class works best if it is placed in same location as clusters. The availability (uptime percentage) of standard storage is more than 99.95%.

1. **Nearline:**

Nearline class is of low cost, high durability. The main purpose of nearline storage class is for infrequent access to the data. The retention period in this class is 30 days. The availability of nearline is more than 99.90% and varies depending on location type.

1. **Coldline:**

This class is very low cost and highly durable which is used for infrequent access to the data. The major difference between Nearline and Coldline is the data in Coldline is less frequent access when compared to Nearline class. The retention period in this class is 90 days. The availability of nearline is more than 99.90% and varies depending on location type.

1. **Archive:**

Archive class used for storing backups, archival data, and files. Archive class plays important role during disaster recovery. Archive class is of lowest cost and highest durability when compared all others. The retention period in this class is 365 days. The availability of nearline is more than 99.90% and varies depending on location type.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Standard** | **Nearline** | **Coldline** | **Archive** |
| **Data access** | very demand access of data | Infrequent access of data | Slightly more infrequent access | Mainly for backups and archival |
| **Cost** | Costlier | Low cost | lower cost than Nearline | Least expensive |
| **Retention period** | No SLA | 30 days | 90 days | 365 days |
| **Availability** | Highly available  (99.95%) | Lower than standard  (99.90%) | Lower than Nearline  (99.90%) | Lower than Coldline  (99.90%) |

**Q4’s Ans: Relational and Non-relational databases**

1. **Relational:** Relational database stores data in the form of tables. It comes under structured data. User can easily read the data from the tables since it has rows and columns. Relational database can be managed using relational database system architecture and principles. We can define various constraints on tables like primary key, unique constraint etc. Data can be managed in the relational database using SQL (Structured query language) for creating, reading, writing, and modifying. There are different kinds of languages for relational database like DDL (data definition language), DML (data manipulation language), DCL (data control language) and TCL (transactional control language). For examples, Oracle SQL, MySQL, etc.
2. **Non-relational database:** The database with no proper structure to it. Which means, there is no tables, rows, and columns. These databases come under semi-structured or unstructured data. These databases are also called as NoSQL databases. Non-relational databases works differently than relational databases. Data in non-relational database will be represented in JSON, XML, etc. Few examples are MongoDB, Apache Cassandra, etc.

|  |  |
| --- | --- |
| **Relational Database** | **Non-Relational Database** |
| Data stored in the form tables | Data stored in different formats based on database for example, JSON, XML, etc. These database also called NoSQL |
| These databases follow Data management system architecture | These databases follow document-like architecture |
| All the relational databases uses Structured query language (SQL) for data operations | There is no single way for data operations. For example, MongoDB uses query documents and Cassandra uses CQL. |
| Advantages:   1. Follows ACID properties 2. High accuracy 3. High security | Advantages:   1. Unstructured data 2. Large data 3. Low cost 4. Performance and faster |
| Disadvantages:   1. Expensive 2. Slow performance | Disadvantages:   1. Not follows ACID properties 2. No standard rules |

**Q5’s Ans:** 5 V’s:

1. **Volume:** In Bigdata world, the data huge and critical for analysis to get best results. According to a report, it is estimated to be 463 Exa Bytes of data generated each day. So, storing, managing and analyze of huge data is crucial. Microsoft and Amazon together can store 1200 Peta Bytes.
2. **Velocity:** The speed at which data receives is called velocity. Velocity can be achieved by establishing high speed media of communication, low latency source systems and higher bandwidths. Now-a-days, optical fiber cables are used to transfer data in seconds. It is estimated that data transferred using optical fiber cables are 200,000 Kilometers per second.
3. **Variety:** Data is available in wide varieties like text, audio and video. When it comes to data analysis, each source of data is important. Hence bigdata need to store, manage all kinds of data.
4. **Veracity:** Veracity is defined as how accurate or truth the data is. These days data is huge, we cannot distinguish true or false data. Hence it is veracity of data is crucial in Bigdata world.
5. **Value:** Even other 4 V’s are in place but nothing can be achieved without value for the business. This is the most important characteristic out of all. Data must be effective to generate accurate results.

**Q6’s Ans:** OLAP and OLTP

**OLAP:** Online Analytical Processing (OLAP) is software for performing multi-dimensional analysis at fast pace on large sizes of data. Businesses use OLAP for performing analysis, problem-solving and decision making on their data coming from multiple source systems. Typically, OLAP is done on data warehouses will sits at backend applications. OLAP is ideal for data mining, business intelligence and complex analytical computations. OLAP is used by data scientists, top-level managers, CEOs and key stakeholders in the business to take critical business decisions. OLAP systems holds historical data and volume of data will be in Terabytes and Petabytes. Query processing will be slower as it involves large amounts of data and requires high computational power. As OLAP have millions of data, it requires huge storage mechanism. OLAP is crucial for healthcare, advertising industries.

**OLTP:** Online Transactional processing (OLTP) is software mainly involves in daily transactions. A transaction can be defined as reading, writing or modifying data in the databases. OLTP is generally available at RDBMS systems. OLTP sits at the front-end of the business. It is mainly used by clerks and managers. OLTP holds current data and processes queries in short period of time. Data processed using OLTP ranges in Megabytes and Gigabytes. OLTP does not need huge storage system as it works for transactional purposes only. OLTP is used in recording customers, dumping data.

|  |  |
| --- | --- |
| **OLAP** | **OLTP** |
| OLAP is used for analysis purposes | OLTP for transactional purposes |
| It works on data warehouses | It works on RDBMS |
| OLAPs are used for machine learning jobs, business intelligence applications | OLTPs are used for data storing and retrieval operations |
| It handles huge of data in terms of Zettabytes and Petabytes | It handles in range of Megabytes and Gigabytes |
| Query processing is slow as it deals with huge data | Query processing is fast |
| It holds both current and historic data | It holds only current data |
| It helps in taking business decisions and forecasting | It helps in running business |
| OLAPs are used by data scientists, business executives, and top-level management | OLTP is used by low-level managers and clerks |
| It is capable to run complex queries | It runs basic queries like SELECT, UPDATE, DELETE |