Cloud Computing

Assignment

Name- Adhyyan Tripathi

Sec/Dept. – C/CSE

Reg. no. – 201700403

Q.1) Security in the cloud, End-user access to the cloud computing.

Ans – As with many aspects of the cloud, security here can actually be better than in an internal datacenter. The ephemeral nature of virtual instances forces you to adopt robust security processes that many traditional hosting environments get away without using, so the move can result in a high-security computing infrastructure. Security in the cloud, can be expressed in three levels –

1. Data Security- This can be achieved by –
    a. Data Control
    b. Encrypting Everything
    c. Regulatory and Standards Compliance
2. Network Security -This can be achieved by –
    a. Managing Firewall Rules
    b. Network Intrusion Detection
3. Host Security- This can be achieved by –
    a. System Hardening
    b. Antivirus Protection
    c. Host Intrusion Detection
    d. Data Segmentation
    e. Credential Management

The functionality “End-User Access to Cloud Computing” enables users to be a “power collaborators”. Some of the most popular software-as-a-Service (SaaS) offerings for consumers are –

YouTube-

On YouTube, people can view first-hand accounts of current events, find videos about their hobbies and interests, and discover the quirky and unusual—all from videos shared by other subscribers. YouTube has become so popular that

it now provides a set of development application programming interfaces (APIs) to enable developers to integrate YouTube functionality into their web sites. The YouTube APIs and tools allow programmers to bring the YouTube experience to their web pages, applications, and devices.

Zimbra-

On September 17, 2007, Yahoo! announced that it had entered into an agreement to acquire Zimbra, Inc., a company specializing in web-based email and collaboration software, for approximately $350 million. The Zimbra email and calendar server is available for Linux, Mac OS X, and virtualization platforms. Zimbra can synchronize with smartphones (such as iPhone and BlackBerry) and desktop clients (such as Outlook and Thunderbird). Yahoo! Zimbra Desktop11 is a free, open source email and calendar client which runs on any Windows, Apple, or Linux desktop computer.

Similar to the above two, a few others who provide end-user access to the cloud computing – Facebook, Zoho, DimDim Collaboration.

Q.2) Cloud security services.

Ans –

1. Identity and Access Management should provide controls for assured identities and access management. Identity and access management includes people, processes and systems that are used to manage access to enterprise resources by assuring the identity of an entity is verified and is granted the correct level of access based on this assured identity. Audit logs of activity such as successful and failed authentication and access attempts should be kept by the application/solution.
2. Data Loss Prevention is the monitoring, protecting and verifying the security of data at rest, in motion and in use in the cloud and on-premises. Data loss prevention services offer protection of data usually by running as some sort of client on desktops/servers and running rules around what can be done. Within the cloud, data loss prevention services could be offered as something that is provided as part of the

build, such that all servers built for that client get the data loss prevention software installed with an agreed set of rules deployed.

3. Web Security is real-time protection offered either on-premise through software/appliance installation or via the cloud by proxying or redirecting web traffic to the cloud provider. This provides an added layer of protection on top of things like AV to prevent malware from entering the enterprise via activities such as web browsing. Policy rules around the types of web access and the times this is acceptable also can be enforced via these web security technologies.
4. E-mail Security should provide control over inbound and outbound e-mail, thereby protecting the organization from phishing and malicious attachments, enforcing corporate policies such as acceptable use and spam and providing business continuity options. The solution should allow for policy-based encryption of e-mails as well as integrating with various e-mail server offerings. Digital signatures enabling identification and non-repudiation are features of many cloud e-mail security solutions.
5. Security Assessments are third-party audits of cloud services or assessments of on-premises systems based on industry standards. Traditional security assessments for infrastructure and applications and compliance audits are well defined and supported by multiple standards such as NIST, ISO and CIS. A relatively mature toolset exists, and a number of tools have been implemented using the SaaS delivery model. In the SaaS delivery model, subscribers get the typical benefits of this cloud computing variant elasticity, negligible setup time, low administration overhead and pay-per-use with low initial investments.
6. Intrusion Management is the process of using pattern recognition to detect and react to statistically unusual events. This may include reconfiguring system components in real time to stop/prevent an intrusion. The methods of intrusion detection, prevention and response in physical environments are mature; however, the growth of virtualization and massive multi-tenancy is creating new targets for intrusion and raises many questions about the implementation of the same protection in cloud environments.
7. Security Information and Event Management systems accept log and event information. This information is then correlated and analyzed to provide real-time reporting and alerting on incidents/events that may

require intervention. The logs are likely to be kept in a manner that prevents tampering to enable their use as evidence in any investigations.

8. Encryption systems typically consist of algorithms that are computationally difficult or infeasible to break, along with the processes and procedures to manage encryption and decryption, hashing, digital signatures, certificate generation and renewal and key exchange.
9. Business Continuity and Disaster Recovery are the measures designed and implemented to ensure operational resiliency in the event of any service interruptions. Business continuity and disaster recovery provides flexible and reliable failover for required services in the event of any service interruptions, including those caused by natural or man-made disasters or disruptions. Cloud-centric business continuity and disaster recovery makes use of the cloud's flexibility to minimize cost and maximize benefits.
10. Network Security consists of security services that allocate access, distribute, monitor and protect the underlying resource services. Architecturally, network security provides services that address security controls at the network in aggregate or specifically addressed at the individual network of each underlying resource. In a cloud/virtual environment, network security is likely to be provided by virtual devices alongside traditional physical devices.

Q.3) Vulnerability assessment tools for cloud computing.

Ans – Vulnerability assessment tools are –

1. Nikto2
2. Netsparker
3. OpenVAS
4. W3AF
5. Arachni
6. Acunetix
7. Nmap
8. OpenSCAP
9. GoLismero
10. Intruder
11. Aircrack
12. Comodo HackerProof
13. Retina CS Community
14. Microsoft Baseline Security Analyzer (MBSA)

15. Nexpose

Q.4) Basics of Apache Pig.

Ans - In 2006, Apache Pig was developed as a research project at Yahoo, especially to create and execute MapReduce jobs on every dataset. In 2007, Apache Pig was open sourced via Apache incubator. In 2008, the first release of Apache Pig came out. In 2010, Apache Pig graduated as an Apache top-level project.

Apache Pig is an abstraction over MapReduce. Apache Pig is a tool or a platform which is used to analyse larger sets of data representing them as data flow. Apache Pig is generally used with Hadoop ands we can perform data manipulation in Hadoop using Apache Pig.

Pig Latin is a high-level language that is provided by Apache Pig which is used to write data analysis programs. Pig Latin provides various operators through which programmers can develop their own functions for reading, writing, and processing data.

To analyse data using Apache Pig, programmers need to write scripts using Pig Latin language. All these scripts are internally converted to Map and Reduce tasks. Apache Pig has a component known as Pig Engine that accepts the Pig Latin scripts as input and converts those scripts into MapReduce jobs.

Advantages of using Apache Pig:

- Using Pig Latin, programmers can perform MapReduce tasks easily without having to type complex codes in Java.
- Apache Pig uses multi-query approach, thereby reducing the length of codes. For example, an operation that would require you to type 200 lines of code (LoC) in Java can be easily done by typing as less as just 10 LoC in Apache Pig. Ultimately Apache Pig reduces the development time by almost 16 times.
- Pig Latin is SQL-like language and it is easy to learn Apache Pig when you are familiar with SQL.
- Apache Pig provides many built-in operators to support data operations like joins, filters, ordering, etc. In addition, it also provides nested data types like tuples, bags, and maps that are missing from MapReduce.

Features of Apache Pig:

- Rich set of operators: It provides many operators to perform operations like join, sort, filer, etc.
- Ease of programming: Pig Latin is similar to SQL and it is easy to write a Pig script if you are good at SQL.
- Optimization opportunities: The tasks in Apache Pig optimize their execution automatically, so the programmers need to focus only on semantics of the language.
- Extensibility: Using the existing operators, users can develop their own functions to read, process, and write data.
- UDF's: Pig provides the facility to create User-defined Functions in other programming languages such as Java and invoke or embed them in Pig Scripts.
- Handles all kinds of data: Apache Pig analyzes all kinds of data, both structured as well as unstructured. It stores the results in HDFS.

Difference between Apache Pig and MapReduce:

- Apache Pig is a data flow language. Whereas, MapReduce is a data processing paradigm.
- MapReduce is low level and rigid but Apache Pig is a high-level language.
- Performing a Join operation in Apache Pig is pretty simple. Whereas, it is quite difficult in MapReduce to perform a Join operation between datasets.
- Any novice programmer with a basic knowledge of SQL can work conveniently with Apache Pig. But, Exposure to Java is must to work with MapReduce.
- MapReduce will require almost 20 times more the number of lines to perform a task. But Apache Pig uses multi-query approach, thereby reducing the length of the codes to a great extent to perform the same task.
- There is no need for compilation in Apache Pig. On execution, every Apache Pig operator is converted internally into a MapReduce job. Whereas, MapReduce jobs have a long compilation process.

Applications of Apache Pig:

A few Applications of Apache Pig are as follows -

- Processes large volume of data
- Supports quick prototyping and ad-hoc queries across large datasets
- Performs data processing in search platforms
- Processes time-sensitive data loads
- Used by telecom companies to de-identify the user call data information

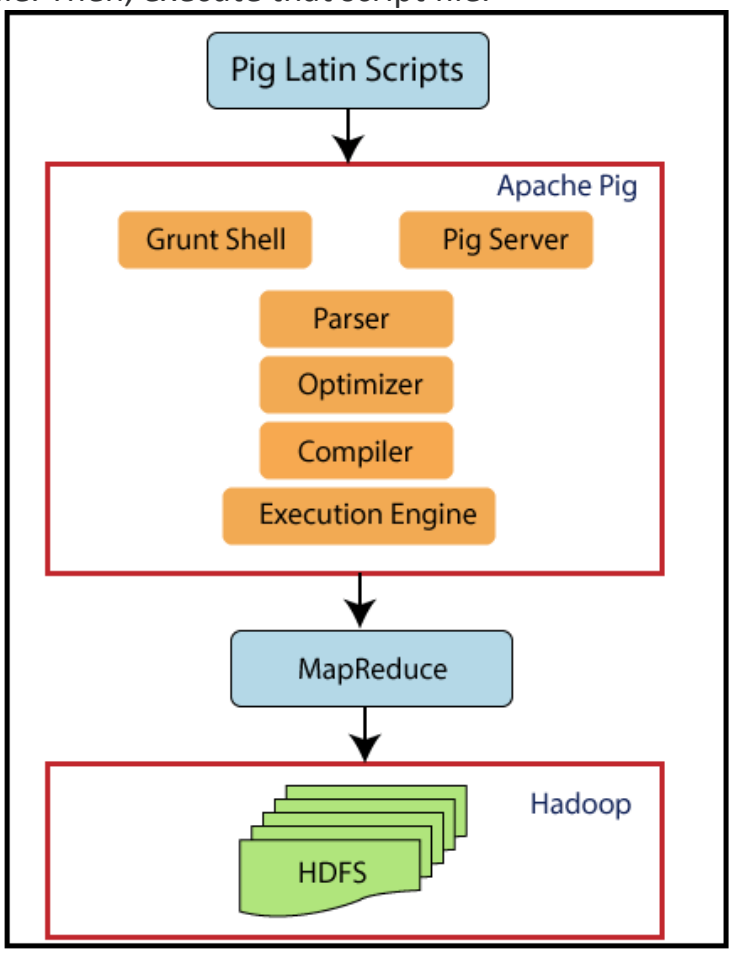
Architecture of Apache Pig:

There are three ways to execute the Pig script -

Grunt Shell: This is Pig's interactive shell provided to execute all Pig Scripts.

Script File: Write all the Pig commands in a script file and execute the Pig script file. This is executed by the Pig Server.

Embedded Script: If some functions are unavailable in built-in operators, we can programmatically create User Defined Functions to bring that functionalities using other languages like Java, Python, Ruby, etc. and embed it in Pig Latin Script file. Then, execute that script file.



Apache Pig Architecture consists of the following major components –

- **Parser:** Parser handles all the Pig Latin statements or commands. Parser performs several checks on the Pig statements like syntax check, type check, and generates a DAG (Directed Acyclic Graph) output. DAG output represents all the logical operators of the scripts as nodes and data flow as edges.

- **Optimizer:** Once parsing operation is completed and a DAG output is generated, the output is passed to the optimizer. The optimizer then performs the optimization activities on the output, such as split, merge, projection, pushdown, transform, and reorder, etc. The optimizer processes the extracted data and omits unnecessary data or columns by performing pushdown and projection activity and improves query performance.
- **Compiler:** The compiler compiles the output that is generated by the optimizer into a series of Map Reduce jobs. The compiler automatically converts Pig jobs into Map Reduce jobs and optimizes performance by rearranging the execution order.
- **Execution Engine:** After performing all the above operations, these Map Reduce jobs are submitted to the execution engine, which is then executed on the Hadoop platform to produce the desired results. You can then use the DUMP statement to display the results on screen or STORE statements to store the results in HDFS (Hadoop Distributed File System).

Pig Latin Data Model –

The data model of Pig Latin is fully nested and it allows complex non-atomic datatypes such as map and tuple.

Atom: Any single value in Pig Latin, irrespective of their data, type is known as an Atom. It is stored as string and can be used as string and number. int, long, float, double, char array, and byte array are the atomic values of Pig. A piece of data or a simple atomic value is known as a field.
Example – 'raja' or '30'

Tuple: A record that is formed by an ordered set of fields is known as a tuple; the fields can be of any type. A tuple is similar to a row in a table of RDBMS.
Example – (Raja, 30)

Bag: A bag is an unordered set of tuples. In other words, a collection of tuples (non-unique) is known as a bag. Each tuple can have any number of fields (flexible schema). A bag is represented by '{}'. It is similar to a table in RDBMS, but unlike a table in RDBMS, it is not necessary that every tuple contain the same number of fields or that the fields in the same position (column) have the same type.
Example − {(Raja, 30), (Mohammad, 45)}

A bag can be a field in a relation; in that context, it is known as inner bag.
Example − {Raja, 30, {9848022338, raja@gmail.com,}}

Map: A map (or data map) is a set of key-value pairs. The key needs to be of type char array and should be unique. The value might be of any type. It is represented by '[]'
Example − [name#Raja, age#30]

Relation: A relation is a bag of tuples. The relations in Pig Latin are unordered (there is no guarantee that tuples are processed in any particular order).