

Programming Languages.

Linux

Linux is a open source, multiuser capability, multi tasking, portable

Users communicate with the kernel through a program known as the shell. The shell is a command line interpreter; it translates commands entered by the user and converts them into a language that is understood by the kernel.

The main concepts of unix/linux system are:

1. Kernel

The kernel is the heart of the operating system. It interacts with the hardware and most of the tasks like memory management, task scheduling and file management.

2. Shell

The shell is the utility that processes your requests. When you type in a command at your terminal, the shell interprets the command and calls the program that you want. The shell uses standard syntax for all commands.

3. Commands and Utilities

There are various commands and utilities which you can make use of in your day to day activities. cp, mv, cat and grep, etc. are few examples of commands and utilities.

4. Files and directories

All the data of Unix is organized into files. All files are then organized into directories. These directories are further organized into a tree-like structure called the filesystem.

Java

- Java is a high-level programming language originally developed by Sun Microsystems and released in 1995.
- Java programming language was originally developed by Sun Microsystems which was initiated by James Gosling and released in 1995 as core component of Sun Microsystems' Java platform (Java 1.0 [J2SE]).
- The latest release of the Java Standard Edition is Java SE 8.

Java is –

- Object Oriented
- Platform Independent
- Simple
- Secure
- Architecture-neutral
- Portable
- Robust
- Multithreaded

- Interpreted
- High Performance
- Distributed
- Dynamic

Classes and Objects

Objects have states and behaviors.

A class can be defined as a template/blueprint that describes the behavior/state that the object of its type support.

Inheritance

Inheritance can be an effective method to share code between classes that have some traits in common, yet allowing the classes to have some parts that are different. Extends is the keyword used to inherit the properties of a class. Following is the syntax of extends keyword.

Polymorphism

Polymorphism is the ability of an object to take on many forms. The most common use of polymorphism in OOP occurs when a parent class reference is used to refer to a child class object.

Abstraction

Abstraction is the quality of dealing with ideas rather than events.

Encapsulation

Encapsulation in Java is a mechanism of wrapping the data (variables) and code acting on the data (methods) together as a single unit. In encapsulation, the variables of a class will be hidden from other classes, and can be accessed only through the methods of their current class. Therefore, it is also known as data hiding.

To achieve encapsulation in Java –

```
Declare the variables of a class as private.  
Provide public setter and getter methods to modify and view the variables values.
```

Interfaces

An interface is a reference type in Java. It is similar to class. It is a collection of abstract methods. A class implements an interface, thereby inheriting the abstract methods of the interface. An interface is similar to a class in the following ways –

```
An interface can contain any number of methods.
```

```
An interface is written in a file with a .java extension, with the name of the interf
```

The byte code of an interface appears in a .class file.

Interfaces appear in packages, and their corresponding bytecode file must be in a dir

However, an interface is different from a class in several ways, including –

You cannot instantiate an interface.

An interface does not contain any constructors.

All of the methods in an interface are abstract.

An interface cannot contain instance fields. The only fields that can appear in an ir

An interface is not extended by a class; it is implemented by a class.

An interface can extend multiple interfaces.

packages

A Package can be defined as a grouping of related types (classes, interfaces, enumerations and annotations) providing access protection and namespace management.

Collection

A collections framework is a unified architecture for representing and manipulating collections. All collections frameworks contain the following –

Interfaces – These are abstract data types that represent collections. Interfaces all

Implementations, i.e., Classes – These are the concrete implementations of the collec

Algorithms – These are the methods that perform useful computations, such as searchir

Exception

An exception (or exceptional event) is a problem that arises during the execution of a program.

Map reduce

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name

MapReduce implies, the reduce task is always performed after the map job.

Map stage

Reduce stage

Terminology

PayLoad - Applications implement the Map and the Reduce functions, and form the core

Mapper - Mapper maps the input key/value pairs to a set of intermediate key/value pairs

NamedNode - Node that manages the Hadoop Distributed File System (HDFS).

DataNode - Node where data is presented in advance before any processing takes place.

MasterNode - Node where JobTracker runs and which accepts job requests from clients.

SlaveNode - Node where Map and Reduce program runs.

JobTracker - Schedules jobs and tracks the assign jobs to Task tracker.

Task Tracker - Tracks the task and reports status to JobTracker.

Job - A program is an execution of a Mapper and Reducer across a dataset.

Task - An execution of a Mapper or a Reducer on a slice of data.

Task Attempt - A particular instance of an attempt to execute a task on a SlaveNode.

Scala

Scala, short for Scalable Language, is a hybrid functional programming language. It was created by Martin Odersky. Scala smoothly integrates the features of object-oriented and functional languages. Scala is compiled to run on the Java Virtual Machine. Scala is object-oriented Scala is functional Scala is statically typed Scala runs on the JVM Scala can Execute Java Code Scala can do Concurrent & Synchronize processing

PHP

PHP started out as a small open source project that evolved as more and more people found out how useful it was. Rasmus Lerdorf unleashed the first version of PHP way back in 1994.

PHP is a recursive acronym for "PHP: Hypertext Preprocessor".

PHP is a server side scripting language that is embedded in HTML. It is used to manage

It is integrated with a number of popular databases, including MySQL, PostgreSQL, Oracle, and Microsoft SQL Server.

PHP is pleasingly zippy in its execution, especially when compiled as an Apache module.

PHP supports a large number of major protocols such as POP3, IMAP, and LDAP. PHP4 also supports SSL.

PHP is forgiving: PHP language tries to be as forgiving as possible.

PHP Syntax is C-Like.

Common uses of PHP

PHP performs system functions, i.e. from files on a system it can create, open, read, write, and delete files.

PHP can handle forms, i.e. gather data from files, save data to a file, through email, etc.

You add, delete, modify elements within your database through PHP.

Access cookies variables and set cookies.

Using PHP, you can restrict users to access some pages of your website.

It can encrypt data.

Characteristics of PHP

Five important characteristics make PHP's practical nature possible –

- Simplicity
- Efficiency
- Security
- Flexibility
- Familiarity

Vb.net

Visual Basic .NET (VB.NET) is an object-oriented computer programming language implemented on the .NET Framework. Although it is an evolution of classic Visual Basic language, it is not backwards-compatible with VB6, and any code written in the old version does not compile under VB.NET.

Like all other .NET languages, VB.NET has complete support for object-oriented concepts. Everything in VB.NET is an object, including all of the primitive types (Short, Integer, Long, String, Boolean, etc.) and user-defined types, events, and even assemblies. All objects inherit from the base class Object.

VB.NET is implemented by Microsoft's .NET framework. Therefore, it has full access to all the

libraries in the .Net Framework. It's also possible to run VB.NET programs on Mono, the open-source alternative to .NET, not only under Windows, but even Linux or Mac OSX.

The following reasons make VB.Net a widely used professional language –

Modern, general purpose.

Object oriented.

Component oriented.

Easy to learn.

Structured language.

It produces efficient programs.

It can be compiled on a variety of computer platforms.

Part of .Net Framework.

Strong Programming Features VB.Net

VB.Net has numerous strong programming features that make it endearing to multitude of programmers worldwide. Let us mention some of these features –

Boolean Conditions

Automatic Garbage Collection

Standard Library

Assembly Versioning

Properties and Events

Delegates and Events Management

Easy-to-use Generics

Indexers

Conditional Compilation

Simple Multithreading

Asp.Net

ASP.NET is a web development platform, which provides a programming model, a comprehensive

software infrastructure and various services required to build up robust web applications for PC, as well as mobile devices.

ASP.NET works on top of the HTTP protocol, and uses the HTTP commands and policies to set a browser-to-server bilateral communication and cooperation.

ASP.NET is a part of Microsoft .Net platform. ASP.NET applications are compiled codes, written using the extensible and reusable components or objects present in .Net framework. These codes can use the entire hierarchy of classes in .Net framework.

The ASP.NET application codes can be written in any of the following languages:

```
C#  
Visual Basic.Net  
Jscript  
J#
```

ASP.NET is used to produce interactive, data-driven web applications over the internet. It consists of a large number of controls such as text boxes, buttons, and labels for assembling, configuring, and manipulating code to create HTML pages.

C#

C# is a modern, general-purpose, object-oriented programming language developed by Microsoft and approved by European Computer Manufacturers Association (ECMA) and International Standards Organization (ISO).

C# was developed by Anders Hejlsberg and his team during the development of .Net Framework.

C# is designed for Common Language Infrastructure (CLI), which consists of the executable code and runtime environment that allows use of various high-level languages on different computer platforms and architectures.

The following reasons make C# a widely used professional language –

```
It is a modern, general-purpose programming language  
It is object oriented.  
It is component oriented.  
It is easy to learn.  
It is a structured language.  
It produces efficient programs.  
It can be compiled on a variety of computer platforms.  
It is a part of .Net Framework.
```

Strong Programming Features of C#

Although C# constructs closely follow traditional high-level languages, C and C++ and being an

object-oriented programming language. It has strong resemblance with Java, it has numerous strong programming features that make it endearing to a number of programmers worldwide.

Following is the list of few important features of C# –

- Boolean Conditions
- Automatic Garbage Collection
- Standard Library
- Assembly Versioning
- Properties and Events
- Delegates and Events Management
- Easy-to-use Generics
- Indexers
- Conditional Compilation
- Simple Multithreading
- LINQ and Lambda Expressions
- Integration with Windows

IDE's

IntelliJ Idea

IntelliJ IDEA is a Java integrated development environment (IDE) for developing computer software. It is developed by JetBrains (formerly known as IntelliJ), and is available as an Apache 2 Licensed community edition,[3] and in a proprietary commercial edition. Both can be used for commercial development.

Eclipse

Eclipse is an integrated development environment (IDE) used in computer programming, and is the most widely used Java IDE.[6] It contains a base workspace and an extensible plug-in system for customizing the environment. Eclipse is written mostly in Java and its primary use is for developing Java applications, but it may also be used to develop applications in other programming languages via plug-ins, including Ada, ABAP, C, C++, C#, COBOL, D, Fortran, Haskell, JavaScript, Julia,[7] Lasso, Lua, NATURAL, Perl, PHP, Prolog, Python, R, Ruby (including Ruby on Rails framework)

Android Studio

Android Studio is the official[7] integrated development environment (IDE) for Google's Android operating system, built on JetBrains' IntelliJ IDEA software and designed specifically for Android development.[8] It is available for download on Windows, macOS and Linux based operating systems.[9][10] It is a replacement for the Eclipse Android Development Tools (ADT) as primary IDE

for native Android application development.

Database and Data-warehouse

MySQL

MySQL is a fast, easy-to-use RDBMS being used for many small and big businesses. MySQL is developed, marketed and supported by MySQL AB, which is a Swedish company. MySQL is becoming so popular because of many good reasons –

MySQL is released under an open-source license. So you have nothing to pay to use it.

MySQL is a very powerful program in its own right. It handles a large subset of the features of the SQL language.

MySQL uses a standard form of the well-known SQL data language.

MySQL works on many operating systems and with many languages including PHP, PERL, C, C++, etc.

MySQL works very quickly and works well even with large data sets.

MySQL is very friendly to PHP, the most appreciated language for web development.

MySQL supports large databases, up to 50 million rows or more in a table. The default configuration can be changed to support even larger databases.

MySQL is customizable. The open-source GPL license allows programmers to modify the MySQL source code.

RDBMS Terminology

Before we proceed to explain the MySQL database system, let us revise a few definitions related to the database.

Database – A database is a collection of tables, with related data.

Table – A table is a matrix with data. A table in a database looks like a simple spreadsheet.

Column – One column (data element) contains data of one and the same kind, for example, names.

Row – A row (= tuple, entry or record) is a group of related data, for example the data for one person.

Redundancy – Storing data twice, redundantly to make the system faster.

Primary Key – A primary key is unique. A key value can not occur twice in one table.

Foreign Key – A foreign key is the linking pin between two tables.

Compound Key – A compound key (composite key) is a key that consists of multiple columns.

Index – An index in a database resembles an index at the back of a book.

Referential Integrity – Referential Integrity makes sure that a foreign key value always

Hive

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce. Hive is not

- A relational database
- A design for OnLine Transaction Processing (OLTP)
- A language for real-time queries and row-level updates

Features of Hive

- It stores schema in a database and processed data into HDFS.
- It is designed for OLAP.
- It provides SQL type language for querying called HiveQL or HQL.
- It is familiar, fast, scalable, and extensible.



Mongo Db

MongoDB is an open-source document database and leading NoSQL database. MongoDB is written in C++. MongoDB is a cross-platform, document oriented database that provides, high performance, high availability, and easy scalability. MongoDB works on concept of collection and document. Document Collection Database

HBase

HBase is a data model that is similar to Google's big table designed to provide quick random access to huge amounts of structured data. HBase is a distributed column-oriented database built on top of the Hadoop file system. It is an open-source project and is horizontally scalable.

HBase is a data model that is similar to Google's big table designed to provide quick random access to huge amounts of structured data. It leverages the fault tolerance provided by the Hadoop File System (HDFS).

It is a part of the Hadoop ecosystem that provides random real-time read/write access to data in the Hadoop File System.

One can store the data in HDFS either directly or through HBase. Data consumer reads/accesses the data in HDFS randomly using HBase. HBase sits on top of the Hadoop File System and provides read and write access.

Data ingestion Tools

Flume

Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events (etc...) from various sources to a centralized data store.

Flume is a highly reliable, distributed, and configurable tool. It is principally designed to copy streaming data (log data) from various web servers to HDFS.

Features of Flume

Some of the notable features of Flume are as follows –

Flume ingests log data from multiple web servers into a centralized store (HDFS, HBase).
Using Flume, we can get the data from multiple servers immediately into Hadoop.
Along with the log files, Flume is also used to import huge volumes of event data produced by applications.
Flume supports a large set of sources and destinations types.
Flume supports multi-hop flows, fan-in fan-out flows, contextual routing, etc.
Flume can be scaled horizontally.

Scoop

Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. It is provided by the Apache Software Foundation.

Sqoop Import

The import tool imports individual tables from RDBMS to HDFS. Each row in a table is treated as a record in HDFS. All records are stored as text data in text files or as binary data in Avro and Sequence files. Sqoop Export

The export tool exports a set of files from HDFS back to an RDBMS. The files given as input to Sqoop contain records, which are called as rows in table. Those are read and parsed into a set of records and delimited with user-specified delimiter.

Frame Works

Apache Hadoop

What is Big Data?

Big data means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks. What Comes Under Big Data?

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

Black Box Data : It is a component of helicopter, airplanes, and jets, etc. It captures data from the black box recorder.

Social Media Data : Social media such as Facebook and Twitter hold information and trends.

Stock Exchange Data : The stock exchange data holds information about the 'buy' and 'sell' orders.

Power Grid Data : The power grid data holds information consumed by a particular node.

Transport Data : Transport data includes model, capacity, distance and availability of vehicles.

Search Engine Data : Search engines retrieve lots of data from different databases.

Big Data

Thus Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

Structured data : Relational data.

Semi Structured data : XML data.

Unstructured data : Word, PDF, Text, Media Logs.

Benefits of Big Data

Big data is really critical to our life and its emerging as one of the most important technologies in modern world. Follow are just few benefits which are very much known to all of us:

Using the information kept in the social network like Facebook, the marketing agencies

Using the information in the social media like preferences and product perception of

Using the data regarding the previous medical history of patients, hospitals are prov

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frameworked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.



Apache Spark

Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application.

Features of Apache Spark

Apache Spark has following features.

Speed – Spark helps to run an application in Hadoop cluster, up to 100 times faster i

Supports multiple languages – Spark provides built-in APIs in Java, Scala, or Python.

Advanced Analytics – Spark not only supports ‘Map’ and ‘reduce’. It also supports SQL




Other Hadoop components

Zoo-keeper

ZooKeeper is a distributed co-ordination service to manage large set of hosts. Co-ordinating and managing a service in a distributed environment is a complicated process. ZooKeeper solves this issue with its simple architecture and API. ZooKeeper allows developers to focus on core application logic without worrying about the distributed nature of the application.

The ZooKeeper framework was originally built at “Yahoo!” for accessing their applications in an easy and robust manner. Later, Apache ZooKeeper became a standard for organized service used by Hadoop, HBase, and other distributed frameworks. For example, Apache HBase uses ZooKeeper to track the status of distributed data.

 Apache ZooKeeper is a service used by a cluster

(group of nodes) to coordinate between themselves and maintain shared data with robust synchronization techniques. ZooKeeper is itself a distributed application providing services for writing a distributed application.

Oozie

Hadoop Distribution

Horton Works

Cloudera

Source Control and build-tools

Git

GitHub

Maven

Jenkins

ANT

Text Operation tools

Regex using Java

Grep

Sed

Awk

Find

Web Service Api's

Rest Api

MicroServices

Command line tools

Curl

Wget

Winscp

Putty

File Formats

Xml

Json

Csv

Storage formats

Sequence file

AVRO

Parquet