Asterix Solution's

Big Data - Hadoop Training Program

A complete Hadoop Development Training Program.

Your Journey to Professional Hadoop Development training starts here!

Hadoop! Hadoop! Hadoop! If you belong to IT field, you must have heard everyone talking about it. So what Hadoop is all about. First thing first Hadoop is not a programming language, it is an API(a set of programs working towards a single goal). So, what is the goal here? To manage the data which is in a big amount. Big here means very BIIIIIIIIIIIG !!

Course Details

You can complete this training in 3 modes

1. Saturday and Sunday| Each day around 4 hours.
2. Saturday Or Sunday | Each day around 4 hours.
3. Weekdays | Each day around 2:30 hours.

Course Content

The proposed Big Data - Hadoop Training Program covers the following modules:

1. Prerequisite Training
    1.1. Linux (Centos) – architecture and commands
2. Introduction to Big Data and Hadoop
3. Hadoop ecosystem concepts
4. Installing hadoop using cloudera quickstart
5. Hadoop Distributed File System(HDFS)
6. MapReduce
7. PIG
8. Hive
9. Sqoop
10. Flume
11. Oozie
12. NOSQL - HBase
13. Apache Spark with Scala (Core Spark and Spark SQL)
14. Real Time tools like Hue, Putty, FileZilla, Cloudera Manager
15. 6 Live Case Studies

Introduction to Big Data and Hadoop

- What is Big Data?
- What are the challenges for processing big data?
- What is Hadoop?
- Why Hadoop?
- History of Hadoop
- Hadoop ecosystem and related projects

Setting up the Development Environment (Cloudera quickstart vm)

- Overview of Big Data Tools, Different vendors providing hadoop and where it fits in the industry
- Setting up Development environment & performing Hadoop Installation on User's laptop
- Hadoop daemons
- Starting stopping daemons using command line and cloudera manager

HDFS

- Understanding the problem statement and challenges persisting to large data and how HDFS solves that problem
- Understanding the HDFS architecture

- Exploring the HDFS using Command-line as well as Web Interface
- Writing files to HDFS
- Reading files from HDFS
- Rack awareness
- HDFS commands

YARN

- Classical version of Apache Hadoop (MRv1)
- Limitations of classical MapReduce
- Addressing the scalability ,resource utilization issue and need to support different programming paradigms
- YARN: The next generation of Hadoop's compute platform (MRv2)
- Architecture of YARN
- Application submission in YARN
- Type of yarn schedulers (FIFO, Capacity and Fair)

MapReduce

- Understanding how the distributed processing solves the big data challenge and how MapReduce helps to solve that problem
- Setting up multiple Eclipse based Java project to have hands-on experience on MapReduce

- Word count problem and solution
- MapReduce flow
- Explain the Driver, Mapper and Reducer code
- Configuring development environment - Eclipse
- Testing, debugging project through eclipse and then finally packaging, deploying the code on Hadoop Cluster
- Input Formats - Input splits & records, text input, binary input
- Output Formats - text output, binary output, lazy output
- MapReduce combiner
- MapReduce partitioner
- Data locality
- Speculative execution
- Job optimization

SQOOP

- Setting up RDBMS Server and creating & loading datasets into RDBMS Mysql.
- Sqoop Architecture
- Writing the Sqoop Import Commands to transfer data from RDBMS to HDFS/Hive/Hbase
- Incremental Imports

- Writing the Sqoop Export commands to transfer data from HDFS/Hive to RDBMS
- Sqoop Jobs to automate those sqoop commands for day to day use

FLUME

- Understanding the Flume architecture and how is it different from sqoop
- Flume Agent Setup
- Setting up data
- Types of sources, channels, sinks Multi Agent Flow
- Different Flume implementations
- Hands-on exercises (configuring and running flume agent to load streaming data from web server)

HIVE

- Hive Architecture and components with typical query flows.
- Creating the table, Loading the datasets & performing analysis on that Datasets
- Types of Hive metastore configurations
- Understanding Hive Data model
- Running the DML commands like Joining tables, writing sub query, saving results to table or HDFS etc.

- Understanding the different File formats and choosing the right one, when to use partitioning, bucketing to optimize query performance

- Writing UDF's to reuse project/Domain specific implementations

PIG

- Pig Architecture and components
- Pig Latin and data model in pig
- Loading structured as well as unstructured data
- Performing Data Transformation by using built-in functions of PIG

  for ex. filter, group, join etc.
- Writing and calling that UDF in PIG Grunt shell
- Creating the PIG script and running the entire script on one go

SPARK (Core Spark and Spark SQL)

- Understanding the Spark architecture and why it is better than Map Reduce
- Working with RDD's.
- Hands on examples with various transformations on RDD
- Perform Spark actions on RDD
- Spark Sql concepts : Dataframes & Datasets
- Hands on examples with Spark SQL to create and work with data frames and datasets
- Create Spark DataFrames from an existing RDD
- Create Spark DataFrames from external files

- Create Spark DataFrames from hive tables
- Perform operations on a DataFrame
- Using Hive tables in Spark

HBASE

- Understanding the Need of NoSQL Database and how is it different from RDBMS
- Understanding the complete architecture of HBASE and how to model the data into column families
- Performing HBASE-HIVE Integration and HBASE-PIG Integration
- Writing the queries to interact with the data stored in HBASE

OOZIE

- Oozie Fundamentals
- Oozie workflow creations
- Oozie Job submission, monitoring, debugging using oozie web console and command line
- Concepts on Coordinators and Bundles
- Hands-on exercises

6 Live Projects

Project #1: Map reduce Use case – Titanic Data Analysis

Data: This Titanic data is publicly available, using this dataset we will perform Analysis and will draw out some insights like finding the average age of male and females died in Titanic, Number of males and females died in each compartment

Problem Statement:

- The average age of the people (both male and female) who died in the tragedy using Hadoop MapReduce.
- How many persons survived – traveling class wise.

Project #2: PIG Use case – Sentiment Analysis on Demonetization

Data: Find out the views of different people on the demonetization by analyzing the tweets from twitter. Here is the dataset where twitter tweets are gathered in CSV format.

We have to analyze the Sentiment for the tweet by using the words in the text. We will rate the word as per its meaning from +5 to -5 using the dictionary AFINN. The AFINN is a dictionary which consists of 2500 words which are rated from +5 to -5 depending on their meaning.

Problem Statement:

- Find all the positive tweets
- Find all the negative tweets

Project #3: Hive Use Case – Pokemon Data Analysis

Data: Pokémon Go is a free-to-play, location-based augmented reality game developed by Niantic for iOS and Android devices. It was released only in July

2016 and only in selected countries. You can download Pokémon for free of cost and start playing. You can also use PokéCoins to purchase Pokéballs, the in-game item you need to be able to catch Pokémon

Problem Statement:

- Find out the average HP (Hit points) of all the Pokémon
- find the count of 'powerful' and 'moderate' Pokemon
- Find out the top 10 Pokémons according to their Hit points
- Find out the top 10 Pokémons based on their Attack stat
- Find out the top 10 Pokémons based on their Defense stat
- Find out the top 10 Pokémons based on their total power.
- Find out the top 10 Pokémons having a drastic change in their attack and sp.attack
- Find out the top 10 Pokémons having a drastic change in their defense and sp.defense
- Find out the top 10 fastest Pokémons

Project #4: Spark Use Case – Breast Cancer Data Analysis

Data: The clinical dataset is released for the awareness of breast cancer. For practice, few problems have been designed with the solution which makes the user understand better.

Problem Statement:

- What is the average age at which initial pathologic diagnosis to be done
- Find the average age of people of each AJCC Stage
- Find out the people with vital status and their count

Project #5: Spark Use Case – Uber Data Analysis

Data: The Uber dataset consists of 4 columns. They are dispatching_base_number, date, active_vehicles and trips

Problem Statement:

· Find the days on which each basement has more trip

Project #6: Spark SQL Use Case – 911 -Emergency Helpline Number Data Analysis

Data: In this post Spark SQL Use Case 911 Emergency Helpine Number Data Analysis, we will be performing analysis on the data provided the callers who had called the emergency helpline number in North America.

Problem Statement:

- What kind of problems are prevalent, and in which state?
- What kind of problems are prevalent, and in which city?

Placements

How our Placement Process Work ?

- We have a dedicated team who looks after placing our students.
- Above 90% of our students trained right now are placed & hence any vacancy in their respective companies we get to know it first.
- Apart from that our company has tie-up with an array of job-portals & IT Job Consultants.
- To sum it up in a line, we have taken utmost care regarding the Job Placements

Top Companies Where Our Students Are Placed

- **Accenture**
- **Wipro**
- **Tata Consultancy Services**
- **Choice India Pvt Ltd**
- **Planfirma Technologies Pvt Ltd**
- **Interact CRM**
- **Infosys**
- **Sudeshi Infotech**
- **HRDXangers Ltd**
- **AD Stringo**
- **Mastek Pvt Ltd**
- **Forbes Technosys Ltd**
- **Wind World**
- **SQWA India**
- **Parametrix Technologies**
- **Reliance JIO**
- **SpiderWeb Technologies**
- **J.NET Pvt Ltd**

Contact

Asterix Solution

Vashi Branch Address: ShopNo 7 & 8, Shivshankar Tower, Opposite Kadambari CHS, Sanpada East,Navi Mumbai-400 705 (Nearest Railway Station: Vashi)

Thane Branch Address: 202 2nd Floor,Rajhans Annexe,Opposite Gaondevi Bus Depot,Above Rajmal Lakhichand Jewellers,Thane West,Mumbai-400602

Contact: 9136507630 / 7715036251

Website: www.asterixsolution.com

Email: welisten@asterixsolution.com