Big Data

Digital Lync

EDUCATION - INNOVATION - INCUBATION

01

Welcome To

Digital Lync

Digital Lync empowers technology seekers by providing world class infrastructure, best quality project based technology education, Research and Development of great products and supports enthusiastic new entrepreneurs.

Our Process

Career Guidance

Expert counsellor advises to match your skills with trending technologies in the industry.

Choose Course

You pick and we guide. We mentor you all along and help you throughout your learning process.

Placement Assistance

We are a link to your awesome future! We recognize, enhance and present your skills to the coolest companies.

www.digital-lync.com

Big Data

Big Data refers to complex and huge sets of data which culminate very fast from multiple sources. Proven technologies like Hadoop, Python convert these voluminous Data Centres into business insights.

DURATION: 80+ HOURS

WHAT YOU NEED TO KNOW

- Analytical Abilities
- Logical Reasoning
- Creative Skills
- Quantitative Skills like Maths and Statistics.
- Basic knowledge of Databases

CAREER OPPORTUNITIES

- Big Data Engineer
- Big Data Analyst

Why Big Data

Big Data is high priority for businesses. Data generation is increasing day by day. Processing of these huge data sets require high end technologies and platform. As a Big Data Developer Specialist, you would work on popular technologies like Hadoop, Python and enable organizations to decipher data for taking strategic decisions.

02

Big Data

Curriculum

INTRODUCTION TO BIG DATA & HADOOP

MODULE 1 :
IMPORTANCE OF DATA & DATA ANALYSIS

- What is Big Data?
- Big Data & its Hype
- Big Data Users & Scenarios
- Structured vs Unstructured Data
- Challenges of Big Data
- How to overcome the challenges?
- Divide & Conquer philosophy
- Overview of Hadoop

MODULE 2 :
HADOOP AND ITS FILE SYSTEM – HDFS – 3 HRS.

- History of Hadoop
- Hadoop Ecosystem
- Hadoop Animal Planet
- What is Hadoop?
- Key Distinctions of Hadoop
- Hadoop Components
- HDFS
- Map Reduce
- Why Distributed File System?

- The Design of HDFS
- Hadoop Distributed File System
- What is a HDFS block?
- Why HDFS block is so large in HDFS?
- Name Node
- Data Node
- Secondary Name Node
- A file in HDFS
- Hadoop Components/Architecture
- Name Node, Job Tracker, Data Node, TaskTracker & Secondary Name node
- Understanding Storage components(Name Node, Data Node & Secondary Name node)
- Understanding Processing components(JobTracker & TaskTracker)
- How Secondary Name node overcomes the failure of the primary Name node
- Anatomy of a File Read
- Anatomy of a File Write

MODULE 3: UNDERSTANDING HADOOP CLUSTER

- Walkthrough of CDH VM setup
- Hadoop Cluster Modes
- Standalone Mode
- Pseudo-Distributed Mode
- Distributed Mode
- Hadoop Configuration files
- core-site.xml
- mapred-site.xml
- hdfs-site.xml
- yarn-site.xml
- Understanding Cluster configuration

MODULE 4:
MAPREDUCE

- Meet MapReduce
- Word Count algorithm – Traditional approach
- Traditional approach on a Distributed system & it's drawbacks
- MapReduce Approach
- Input & Output Forms of a MR Program
- Hadoop Data Types
- Map, Shuffle & Sort, Reduce Phases
- Workflow & Transformation of Data
- Word Count Code Walk through
- Input Split & HDFS Block
- Relation between Split & Block
- MR Flow with Single Reduce Task
- MR flow with multiple Reducers
- Data locality Optimization
- Speculative Execution
- Combiner
- Partitioner

MODULE 5:
PIG

- What is Pig
- Why Pig
- Pig vs Sql
- Execution Types or Modes
- Running Pig
- Pig Data types
- Pig Latin Relational Operators
- Multi Query Execution
- Pig Latin Diagnostic Operators
- Pig Latin Macro & UDF Statements
- Pig Latin Commands
- Pig Latin Expressions
- Schemas

- Pig Functions
- Pig Latin File Loaders
- Pig UDF & executing a Pig UDF
- Pig Use cases

MODULE 6:
HIVE

- Introduction to Hive
- Pig vs. Hive
- Hive Limitations & Possibilities
- Hive Architecture
- Metastore
- Hive Data Organization
- Hive QL
- Sql vs. Hive QL
- Hive Data types
- Data Storage
- Managed & External Tables
- Partitions & Buckets
- Static Partitioning & Dynamic Partitioning
- Storage Formats
- File Formats – Sequence File & RC File
- Using Compression in Hive
- Built-in Serdes
- Importing Data (Using Load Data & Insert Into)
- Alter & Drop Commands
- Data Querying
- Using MR Scripts
- Hive Joins
- Sub Queries
- Views

MODULE 7:
HBASE

- Introduction to NoSql & HBase
- HBase vs. RDBMS
- HBase Use cases
- Row & Column oriented storage
- Characteristics of a huge DB
- What is HBase?
- HBase Data-Model
- HBase Logical Model & Physical Storage
- HBase Architecture
- HBase in operation (put, get, scan & delete)
- Loading Data into HBase
- HBase Shell Commands
- HBase Operations through Java
- HBase Operations through MR

MODULE 8:
ZOOKEEPER & OOZIE

- Introduction to Zookeeper
- Distributed Coordination
- Zookeeper Data Model
- Zookeeper Service

MODULE 9:
SQOOP

- Introduction to Sqoop
- Sqoop design
- Sqoop basic Commands
- Sqoop Table Import flow of execution

- Sqoop Import Commands – to HDFS, Hive & HBase tables
- Sqoop Incremental Import
- Incremental Append
- Incremental Last Modified
- Sqoop export flow of execution
- Sqoop Export Command

MODULE 10:
FLUME

- Flume Architecture
- Flume Components
- Streaming live Twitter data with Flume

SPARK

MODULE 1:

- Introduction & Overview
- Architecture
- Installation of Spark-- Options
- Starting the Spark--- possibilities
- Amazon EMR
- EC2
- Maven
- Standalone mode
- With mesos
- With YARN
- HDinsight
- Spark context & Spark Session

MODULE 2:

- Basics & Spark Shell Applications
- Various possibilities
- Eclipse with Maven
- Eclipse with SBT
- Zeppelin Notebook
- IntelliJ

- Spark Jobs & API's
- Spark Core
- RDD's
- Transformations
- Actions
- Data Frame

MODULE 3:

- Spark with External Data Sources
- From Local file system
- From HDFS
- From Amazon S3
- From Cassandra Spark SQL
- Schema
- Case Classes
- Joins
- Catalyst Optimizer

MODULE 4:

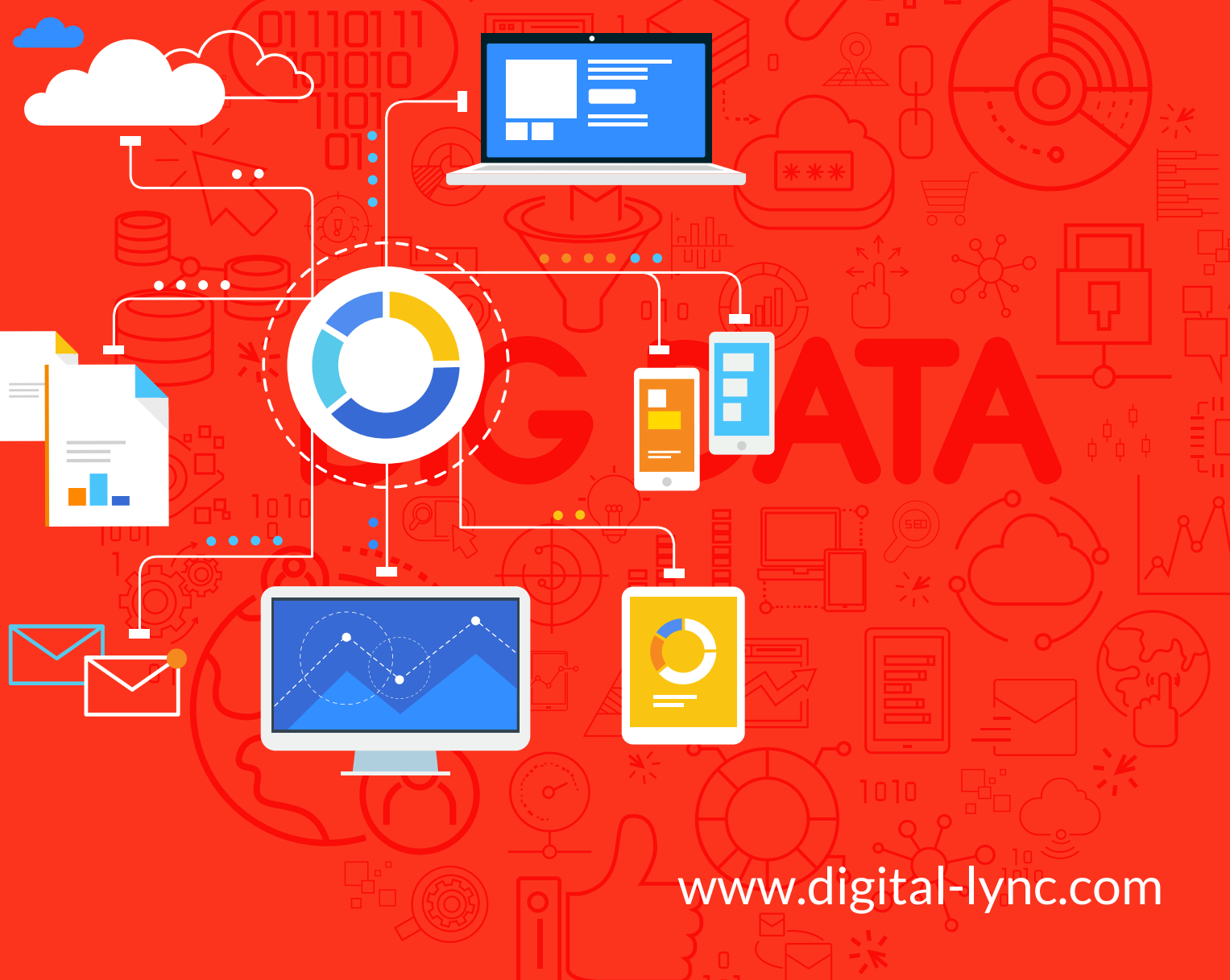
- Spark Streaming
- Spark MLlib
- Spark GraphX
- PySpark

03

Big Data Project : 1

BIG DATA ECOSYSTEM

Flipkart, Amazon, Alibaba all these websites look so cool from the front, but ever wondered about what happens in the background? Well, don't wonder! Using Big Data Structure, we will show you how to make the complete background infrastructure. Gone are the days when we used SQLs. It's time for NOSQLs and that too inside the Big Data Ecosystem. Apply HADOOP HDFS, Spark, HBASE, SQOOP etc and make the complete infrastructure.

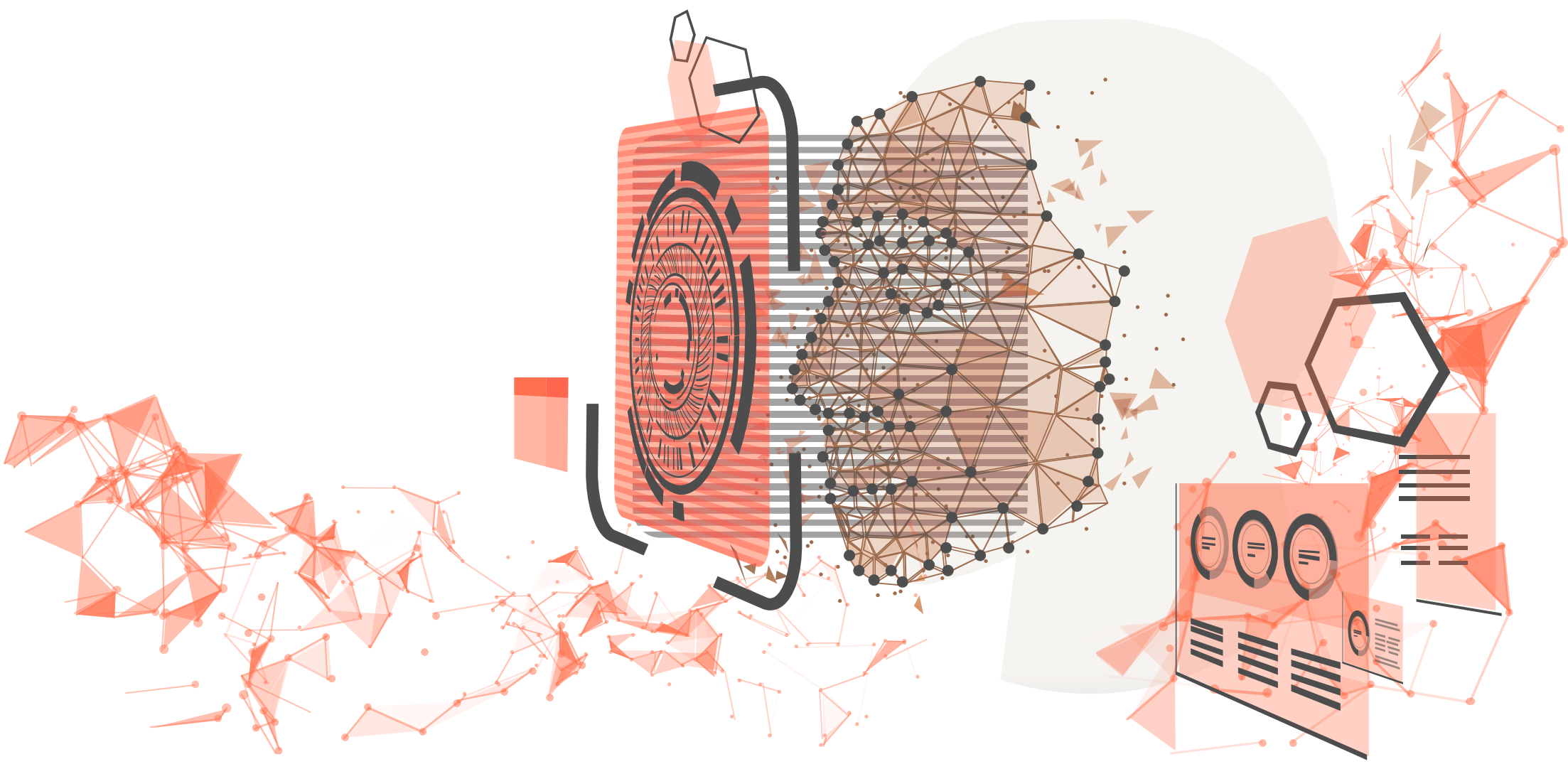


www.digital-lync.com

Project : 2

FACIAL RECOGNITION

Using deep learning to recognise faces is one thing. But, what if your company has more than 50 thousand employees? In that case working using normal server systems will not help. So, why not use Distributed Systems, Incorporate BIG DATA clusters on them and apply the same computer vision and Deep Learning concepts using SPARK, HADOOP, CASSANDRA and other stacks? Won't it be cool to merge two technologies? It is cool, and let's work together to make this unique model!



04

Why Digital Lync

The Best Trainers

Our developers are our trainers! our trainers spend half of their time on products at our innovation center.

Online Classes

Our online experience is as good as in - class experience. Our programs offer end to end personal interaction with students.

We are Committed

We consider your success to be ours. Your achievement is our commitment. We teach until you learn.

Really Affordable

We believe in a collaborative, cool and learn friendly environment. Our promise, is to provide all this for no additional charge.

www.digital-lync.com

Trending Technologies

- Python
- Devops
- AWS
- Azure (Cloud Computing)
- Data Sciences
- Deep Learning
- Artificial Intelligence
- Data Analysis
- Big Data
- FullStack
- Digital Marketing
- Mobile Development
- Blockchain
- Visual Design
- Game Development
- IOT
- Cyber Security