2021-22 Batch

	Big Data Analytics	L	T	P	C
Version 1.0					
Pre-requisites/Exposure					
Co-requisites	–				

Unit 1. Introduction to Big Data Analytics

Big Data overview, Structures of data, Big Data growth story, Big Data sources, Big Data adoption drivers, Need of Big Data, Growth drivers for IT industry, Big Data: Definition, Characteristics of Big Data, Units to measure Big Data, Big Data types, Benefits & barrier of Big Data analytics, Need of Big Data, Big Data process, Big Data framework, Big Data platform and application frameworks, An example of Big Data platform in practice, A Big Data platform manifesto, Big Data technologies, Big Data tools, Big Data & analytics, Merging the traditional and Big Data approaches, More ways: Wide ranging analytics and techniques, The 5 key Big Data use cases, Big Data usage, Use cases, Big Data and complexity in health care, Use cases: Healthcare and life sciences, Use cases: Transportation services, Use cases: Life insurance, IBM's Big Data success story, Data repository analyst view, Business drivers with examples, BI versus data studies, Architecture of modern analytics, Large data drivers, Method to study the evolving Big Data environment, Latest Big Data ecosystem, Large data research explanations, Overview of lifecycle data processing, Major functions for a good research, Information analytics history and summary, Big Data resources, Tackling the question, Primary stakeholders recognition, Analytical supporter interview, Original assumptions creation, Potential databases detection, Preparing evidence for Big Databases, Analytical sandbox planning, ETL execution, The data learning, Data conditioning, Visualize and verification, Popular data preparation instruments, Step 3: Planning of models, Research on model planning in industry verticals, Data exploration, Big Data: Choice of model, Popular model of Big Data, Step 4: Model design, Popular model building tools, Step 5: Results contact, Step 6: Consumption, Key outputs from a successful analytics project, Case study: Digital network for creativity and research (GINA), Step 1: Searching, Step 2: Preparing evidence, Step 3: Planning of models, Step 4: Model design, Step 5: Results contact, Step 6: Consumption.

Unit 2. Hadoop Fundamentals

What is Hadoop? Examples of Hadoop in action: IBM Watson, Examples of Hadoop in action, Introduction to Hadoop, Data distribution, Flat scalability, HDFS (Hadoop Distributed File System), Name nodes, Data nodes, Data nodes with blocks of multiple files with a replica of 2, The data is distributed across nodes at the time of loading, MapReduce, An SQL example of MapReduce, The map function, Sort phase, The reduce function, Combiner and partition functions, Streaming and pipes, MapReduce example: Wordcount, MapReduce co-locating with HDFS, MapReduce processing, Speculative execution, MapReduce: A tale of two APIs, MapReduce anatomy, What is HBase? NoSQL technology, CAP theorem, ACID properties, Why HBase? Important things to keep in mind, HBase vs. RDBMS, For example, Physical view in HBase, Logical to physical view, HBase components, HBase components definitions and roles, Characteristics of HBase tables, HBase is a sorted multidimensional map, Row key design considerations, Column family design considerations, Cluster configuration, HDFS configurations settings, Hadoop site.xml for a single-node configuration, HDFS start, Interact with HDFS, Example of put command, Retrieve data

from HDFS, HDFS command reference, HDFS permissions and security, HDFS additional tasks: Rebalancing blocks, Closing the nodes, Check health file system, Load scenarios, Load solution using Flume, How Flume works, Consolidation, Replicating and multiplexing, Apache Hadoop core components, Why Hadoop Apache? Why is it easier for Hadoop than other distributed computing systems? Where is Hadoop ideal for computing? Daemons of HDFS, Secondary name node, Check-pointing by secondary name node, HDFS architecture, Daemons MapReduce, YARN capital assignments, Node manager with three containers, Resource allocation on another node, Running tasks on m multiple containers, Resource allocation, Fair scheduler per-queue properties, The map reducing job workflow, HDFS daemons with strong disponibility, Height of the block, A file stored in a single block, Abstract block, Name node keeps the block locations, Data flow in replication, Data pipeline in creating block replicas, Under reproduction, Name node metadata, Single namespace HDFS architecture, HDFS federation architecture, Place of data, Replica placement on two racks, HDFS network topology, Table: Block location class methods, How does HDFS store, read, and write files? Data node pipeline in writing a file, Verification of checksum, Data collection and data analysis Hadoop cluster, Hadoop cluster in data storage and processing, Master protocol application, MRv2 cluster operation, Current and existing APIs, Data serialization options, Apache Avro, Sequence and Avro reference files, Apache Thrift, Thrift and protocol buffers comparison, Commands for HDFS shell file system, Select MapReduce work main and importance forms, A mapper's development cycle and areducer's function, The mapper's life cycle in the latest API, A reducer's life-cycle in the old API, The reducer lifecycle in the latest API, Input clues to output clues link, Input/Output Mapper sort, Key/Value types, Input Formats in the old API, Mapper Key/Values Input/Output number, Reducer Input/Output number of K values, Keys and attributes sorting, Combiners, Shuffle, Table: Parameters in compare methods, Table: Configuration properties to tune the sort and shuffle process, MapReduce with shuffle and sort, Settings and submissions for MapReduce job, Table: FileInputFormat<K,V> class Static methods using JobConf, Settings and submissions for MapReduce job, Combiner on reducer, Shuffle transfer number, Speculative performance, Data paths in MapReduce task input and output, Data movement in a MapReduce job in the reduce phase, Data flow provided by the InputFormat, Data flow provided by the OutputFormat, InputFormats for File-Based Input formats, Table: InputFormats for File-Based Input formats, RecordReader, compression andsequence files, LineRecordReader example, RecordReader with FileSplit, Built-In RecordReaders, Sequence files, SequenceFileInputFormat classsubclasses, SequenceFile class nested classes and interfaces, Sequence file header, Compression, Configuration properties for configuring, Codecs supported by Hadoop, Commands for HDFS Shell file system, Administration commands.

Unit 3. Query Languages for Hadoop

What is JAQL? JSON: JavaScript Object Notation, JSON format, where does JAQL fit? MapReduce overview, MapReduce and Hadoop, Starting up the JAQL server & entering JAQL in command line mode, JAQL and MapReduce, Let's do this step by step, JAQL and MapReduce: The rewrite engine and explain, JAQL schema, Data types, JAQL basics, Arrays, Records, Operators, Lazy/late evaluation, Why materialized assignment (:=) ? The -> operator, Expressions, Functions, Why JAQL core operators, Core operators: Expand, Core operators: Group, Core operators: Group format (single), Core operators: Group (single), Core operators: Understanding grouping, Core operators: Co-groups, Core operators: Join outer joins, Core operators: Sort, JAQL SQL, JAQL SQL: Case-sensitivity, JAQL and MapReduce basics, JAQL and MapReduce: Explain, JAQL and MapReduce: Map, MapReduce: Job configuration, JAQL and MapReduce:

Native MR jobs, JAQL I/O, JAQL I/O adapter operations, JAQL I/O: I/O adapters, JAQL I/O: I/O adapters arguments, JAQL I/O: Delimited files, JAQL I/O: Binary sequence files, JAQL I/O: Text sequence files, JAQL I/O: Other adapters.

Unit 4. Hive: Hadoop Reporting and Analysis

History of Hive, Hive components, Hive directory structure, Physical layout: Data in Hive, Database use/drop/alter, Primitive data types, Complex data types, Creating a table, Table partitioning, Managed Vs external tables, Indexes, Drop/alter table, Loading data into Hive: From a file, Loading data from a directory, Select from, Selecting from partitions, Joins, Order by/sort by, Views, Order, CLI (Command Line Interface), Metastore, Real world use cases, Hive use case, Hive command line, Language of Hive Query (HQL), Creating tables, Hive primitive data types, Hive data review, Aggregations and affiliation, HBase, HBase schema, Social media events with sparse columns, HBase timestamp versioning, Importing from MySQL to Hive, Import to HBase from MySQL, Intake of Flume streaming data, Multi-agent Flume data flow, Fan-in Flume data flow, Log ingestion into HDFS.

Unit 5. Pig: Hadoop Reporting and Analysis

What is Pig? Pig versus other tools, Executing Pig, First look at Pig data, Pig Latin statement basics, Input, LOAD operator continued, Accessing data, Case sensitivity, Field reference, Pig data types, Operators, Parameter substitution, Output, MapReduce in Pig, Cascading in Pig, Apache hive and Pig, Pig data form, Complex data types, Map, Schema, Casting, Casting error, Comparison operators, Identifiers, Boolean operators, Invoking the grunt shell, Auto completion, Grunt shell flow, Pig operators and commands, Regex in the file path, Store, Dump, Foreach generate, Flatten, New schema, Nested block, Null, Comparison operators, Assert, SPLIT, Flatten, RANK, Order by, Using the partitioner, Using a shell program, MapReduce program, CUBE, Rollup, Parameter substitution, Advanced JOIN, Equi Joins, Inner Joins, Left outer join, Cogroup, CROSS Join, Functions, Pig storage, HBase Storage, Apache Oozie, Types of Oozie jobs, Set a value to a property, Scheduling a Pig script, Integrating with the workflow, Upload Files to HDFS, Bundle, Oozie user interface.

Unit 6. Data Visualizations

Visualization, Visualization in Big Data? Visualization value of Big Data, Large data visualization issues, Analysis of diagrams, Graphs and network organization, Algorithms for graph analytics and solutions, Dedicated appliances for graph analytics, how to select among different chart types? Pie chart, Doughnut chart, Line chart, Map chart, Tree map chart, Waterfall chart, Scatter plot, Histogram chart.

Unit 7. Sqoop: Hadoop Reporting and Analysis

What is Sqoop? Sqoop connection, Sqoop import, Sqoop import examples, Sqoop exports, Sqoop exports, Additional export information, Distributed systems, Sqoop command.

Unit 8. Flume: Hadoop Reporting and Analysis

What is Flume? Applications of Flume, Flume advantages, Flume features, Log data streaming, HDFS issue, Flume event, Flume agent, Flume channel, Multi-hop flow, Data streaming in Flume, Apache Flume environment, Flume installation, Apache Flume configuration, Apache Flume fetching twitter data,

Starting HDFS for Flume, Verifying HDFS, Apache Flume sequence generator source, Verifying the HDFS, Apache Flume netcat source, Passing data to the source.

Unit 9. Oozie: Hadoop Reporting and Analysis

What is apache Oozie? Use cases of apache oozie, Hue editor for Oozie, Oozie Eclipse Plugin (OEP), Apache Oozie workflow, Workflow code, Apache Oozie property file, Apache Oozie coordinator, Apache Oozie bundle, Apache Oozie CLI and extensions.

Unit 10. NoSQL

What is NoSQL? NoSQL and SQL, Brief history of NoSQL databases, Features of NoSQL, Types of NoSQL databases, Relational Vs Document database, Graph-based, Advantages and disadvantagesof NOSQL, NoSQL's benefits, MongoDB, Features of MongoDB, Why use MongoDB? MongoDB data modeling, MongoDB: Create database, MongoDB drop database, MongoDB create collection, MongoDB drop collection, MongoDB update documents, MongoDB delete documents, MongoDB query documents, MongoDB and SQL similarity, MongoDB text search, Text index, MongoDB shell, How to run the shell, MongoDB shell collection methods, Specifying the collation.

Unit 11. ZooKeeper: Hadoop Reporting and Analysis

What is ZooKeeper? Zookeeper service: Replicated mode, Zookeeper service: Standalone mode, Consistency guarantees, ZooKeeper structure: Data model, Role in hadoop infrastructure, Real world use cases, Problem with unstructured data, Need to harvest unstructured data, Need for structured data, Approach for text analytics, Web tooling overview, Basic components of an extractor.

Unit 12. R: Hadoop Reporting and Analysis

What is open-source R? The R appeal: What attracts users? Companies currently using R, what is the R programming language? Limitations of open-source R, Open source R packages to boost performance, Challenges with running large-scale analytics, 3 key capabilities in big R, Big R architecture, User experience for big R, what's behind running big r's scalable algorithms? Big R machine learning: Scalability and performance, Simple Big R example.

Unit 13. Analytics for Big Data at Rest & in Motion

IBM info-sphere streams, The IBM analytics platform, Traditional computing versus stream computing, Something meaningful is happening, The info sphere streams platform, How streams works: Analysis, How streams works: Scaling, From essential elements to deployed, running jobs, Info sphere streams objects: Runtime view, Info sphere streams objects: Development view, Processing elements, Application resiliency, Monitor applications, track and debug data flow, Administration, Administration: Stream tool, Streams for excel, Samples shipped with the product, A data flow network: Operators and streams, Streams processing language: Highlights, From bottom to top, Provided primitive operators, Utility operators, Operator instances, Streams flow, Content of streams, Attribute types, Further look at composite types, Type definitions, Stream schemas, Form of a simple streams processing language, Form of operator, Merging input streams, Multiple input port, Referencing prior tuples, Side effects, Edge operators, File source operator, File source operator, Different types of operator, Filter operator, Dynamic filter operator, Split operator, Streams windows, Window properties, Policy specifications, Tumbling

window, Partitioning windows, The notion of a database join, Join operator, Parameters for a join, Aggregate operator, Punctor operator, Sort operator, Barrier operator, Pair operator, Delay operator, Throttle operator, Switch, Deduplicate, Collections, Lists, Sets, Maps, Operations on lists, Setting up debugging, Debugging operator port, Streams processing language capabilities, Currently provided by infosphere streams, What is a toolkit and how to use one? Working with toolkit paths, Toolkit structure, Toolkit versioning.

Lab Exercises -

Exercise 1. HDFS commands – Basic
Exercise 2. HDFS commands – Advance
Exercise 3. HDFC commands to handle unstructured datasets
Exercise 4. HDFC commands to handle Semi- unstructured datasets (XML files)
Exercise 5. MapReduce Program – Word count
Exercise 6. MapReduce Program – Find the maximum temperature of city
Exercise 7. MapReduce Program – Weather Data Analysis
Exercise 8. MapReduce Program – Aggregating text fields
Exercise 9. Hive commands
Exercise 10. Hive Joins for Datasets
Exercise 11. Hive Partitioning for Datasets
Exercise 12. Hive Bucketing
Exercise 13. Hive Performance tuning
Exercise 14. Hive MapReduce
Exercise 15. PIG commands
Exercise 16. PIG - Group By, Nested Foreach, Join
Exercise 17. PIG MapReduce - word count
Exercise 18. PIG - Twitter Data analysis
Exercise 19. SQOOP Operations
Exercise 20. Flume - Sentimental analysis
Exercise 21. Apache Storm - Mobile Call Log Analyzer
Exercise 22. MongoDB Operations