DATA MASTERY PROGRAM

COURSE CURRICULUM

BigDataYatra
Data Mastery Awaits You !

1 HADOOP AND HDFS	**2** PYTHON CODING	**3** APACHE SPARK INTERNALS
4 SPARK FUNCTIONS & SQL	**5** SPARK OPTIMIZATION	**6** DATABRICKS & DELTA
7 END TO END PROJECT		

BigDataYatra
Data Mastery Awaits You !

Ultimate Big Data Program

Trainer

Somesh Khatawe

(Senior Data Engineer Fractal Analytics)

Trainer

Shivam Jha

(Senior Data Engineer MasterCard)

What we offer !!

- ✓ Pre requisite for spark
- ✓ Assisted spark setup on IDE and jupyter notebook
- ✓ In depth spark training with practical and assignments
- ✓ Individual progress tracking
- ✓ End to end spark projects

For more Information please refer the course curriculum

+91 7795538424 +91 99106 57939 bigdatayatra@gmail.com

CHAPTER I: INTRODUCTION TO BIGDATA & HADOOP

- WHAT IS BIG DATA
- DIFFERENT V's OF BIG DATA
- STRUCTURED/ SEMI-STRUCTURED/ UN-STRUCTURED FORMATS
- MONOLITHIC VS DISTRIBUTED SYSTEM
- VERTICAL SCALING & HORIZONTAL SCALING
- BIG DATA FRAMEWORK: HADOOP
- HADOOP ECOSYSTEM UNDERSTANDING AND INTERNALS: HDFS, MAPREDUCE, YARN
- HDFS ARCHITECTURE
- HOW A FILE IS STORED IN DISTRIBUTED ENVIRONMENT- HDFS
- DATA NODE AND NAME NODE IN HADOOP SYSTEM
- DEALING WITH DATA NODE AND NAME NODE FAILURE – HEARBEAT, REPLICATION, FSIMAGE, EDIT LOGS, CHECKPOINTING, SECONDARY NAME NODE
- RACK AWARENESS MECHANISM
- EDGE NODE IN HADOOP
- HADOOP INSTALLATION – COMPLETE GUIDE
- HADOOP HANDS-ON & PRACTICAL: DATA NODE, NAME NODE, YARN WALKTHROUGH ALONG WITH IMPORTANT “HDFS COMMANDS”
- MAPREDUCE- WHAT IS MAPREDUCE PROCESSING ENGINE
- WHAT IS MAP PHASE AND REDUCE PHASE
- SHUFFLE AND SORT IN MAPREDUCE
- CONCEPT OF RECORD READER
- HASH PARTITIONING IN MAPREDUCE
- CONCEPT OF COMBINER
- INTERNALS OF MAPREDUCE ENGINE (WITH EXAMPLE)
- YARN – YET ANOTHER RESOURCE NEGOTIATOR

CHAPTER II: PYTHON PROGRAMMING

- PYTHON OVERVIEW
- COMPILED AND INTERPRETED LANGUAGE
- GENERATE OF BYTE CODE – PRACTICAL
- DYNAMIC VS STATIC TYPING
- COMMENTS IN PYTHON – SINGLE AND MULTI LINE
- PYTHON VARIABLES AND ID() FUNCTION
- VARIABLE NAMING CONVENTION
- DATATYPES – NUMBERS, TEXT, BOOLEAN, MAPPING, COLLECTION
- TYPE() FUNCTION
- PRINT FUNCTION
- USE VARIABLE IN PRINT
- ESCAPE SEQUENCE (\, \N, \T)
- IMPORT MODULE IN PYTHON
- KEYWORD AND KWLIST
- OPERATOR (ARITHMETIC, RELATIONAL, LOGICAL, BITWISE, ASSIGNMENT, IDENTITY, MEMBERSHIP)
- INPUT()
- TYPE CONVERSION
- BIN() FUNCTION
- ORD() AND CHR() FUNCTION
- EVAL() IN PYTHON
- DECISION CONTROL IN PYTHON (IF, IF ELSE, IF ELIF LADDER, SINGLE LINE IF ELSE)
- LOOPING IN PYTHON - WHILE AND FOR LOOP
- BREAK AND CONTINUE
- FOR LOOP WITH ELSE AND WHILE LOOP WITH ELSE
- RANGE() CLASS IN PYTHON
- PASS KEYWORDS

- PYTHON DATA STRUCTURES
- LIST AND LIST FUNCTIONS
- PACKING VS UNPACKING IN PYTHON
- LIST COMPREHENSION
- ENUMERATE IN PYTHON
- STRING AND STRING FUNCTIONS
- SLICING OPERATOR
- TUPLE AND TUPLE FUNCTIONS
- SET AND SET FUNCTIONS
- DICTIONARY AND DICTIONARY FUNCTIONS
- FUNCTIONAL PROGRAMMING IN PYTHON (PYTHON FUNCTIONS)
- KEYWORD AGRUMENT IN PYTHON
- VARIABLE ARGUMENTS IN PYTHON (*args and **kwargs)
- RECURSION
- ERROR HANDLING IN PYHTON
- __name == __main__ IN PYTHON
- LAMBDA IN PYTHON
- DECORATOR
- GENERATOR
- ITERATOR
- SEARCHING AND SORTING ALGORITHMS
- 100+ PYTHON PROGRAMS IN LIVE CLASS AND ASSIGNMENTS

CHAPTER III: APACHE SPARK – INTRODUCTION

- WHAT IS APACHE SPARK
- WHY SPARK : SPARK VS MAPREDUCE
- HOW DATA STORED IN SPARK
- RDD IN SPARK
- LAZY EVOLUTION

- IMMUTABILITY
- WHY RDD IS CALLED RESILIENT DISTRIBUTED DATASET
- DAG AND LINEAGE
- PAIR RDD
- WHAT IS SPARK CONTEXT
- CREATING SPARK CONTEXT PROGRAMMATICALLY
- WAYS TO CREATE RDD PROGRAMMATICALLY.
- DEFAULT PARTITIONS IN RDD
- CHECK NUMBER OF PARTITIONS IN A RDD.
- PARTITIONS IN PARALLELIZE RDD VS PARTITIONS IN RDD CREATING FROM A TEXT FILE.
- RDD PROGRAMMING EXAMPLES INVOLVING COMPLEX TRANSFORMATIONS
- SPARK UI BRIEF WALKTHROUGH (EXPLAIN ALL TABS AND INDIVIDUAL COLUMNS – WHAT THEY SIGNIFY WITH EXAMPLE)
- SPARK SHARED VARIABLES: BROADCAST AND ACCUMULATORS
- HOW SPARK EXECUTES PROGRAM ON CLUSTER
- CLIENT MODE, CLUSTER MODE AND LOCAL MODE
- EXECUTOR AND DRIVER IN SPARK
- DATA SHUFFLING IN SPARK
- TRANSFORMATIONS: TYPES OF TRANSFORMATIONS (NARROW VS WIDE TRANSFORMATION)
- ACTIONS IN SPARK
- HOW JOBS, STAGE, TASKS ARE CREATED IN SPARK RDD & DF(SPARK UI DEBUG)
- MAP AND MAP PARTITION
- REDUCEBYKEY VS RECUDE
- REDUCEBYKEY VS GROUPBYKEY (INTERNALS)

CHAPTER IV: APACHE SPARK – STRUCTURED API

- APACHE SPARK ECOSYSTEM
- STRUCTURED APIS VS LOWER LEVEL APIS
- DATAFRAME VS RDD VS DATASET
- SERIALIZATION VS DESERIALIZATION IN SPARK
- WHAT IS SPARK SESSION
- CREATING SPARK SESSION
- VARIOUS DATATYPES IN SPARK
- VARIOUS WAYS OF CREATING DATAFRAME IN SPARK:
- CREATING EMPTY DATAFRAME
- CREATING DATAFRAME FROM RDD
- CREATING DATAFRAME FROM COLLECTION CREATING DATAFRAME BY SPECIFYING SCHEMA USING STRUCTTYPE AND STRUCTFIELD.
- CREATING DATAFRAME BY SPECIFYING SCHEMA USING DDL STRING APPROACH
- CREATING NESTED DATAFRAME
- NULLABLE ARGUMENT – HOW IT WORKS
- EXTRACT TRANSFORM LOAD (ETL) IN SPARK
- DIFFERENT FILE FORMATS IN SPARK (ROW BASED VS COLUMN BASED)
- INTERNALS OF DIFFERENT FILE FORMATS
    - CSV,
    - XML
    - JSON
    - AVRO
    - ORC
    - PARQUET
- LOW LEVEL COMPRESSION TECHNIQUES
    - BIT PACKING
    - RUN LENGTH ENCODING
    - DICTIONARY ENCODING
    - DELTA ENCODING
    - READING MULTIPLE JSON

- READING JSON FILE IN SPARK
- READ JSON FROM RDD OF JSON STRING
    - MULTILINE OPTION
    - READING FROM A DIRECTORY
    - SPECIFY SCHEMA EXPLICITLY WHILE READING JSON FILE.
    - FLATTENING JSON FILES.
- READING CSV FILE IN SPARK
    - VARIOUS OPTIONS WHILE READING CSV
    - DRAWBACK OF INFERSCHEMA
    - SPECIFY SCHEMA EXPLICITLY WHILE READING CSV FILE.
    - READING MULTIPLE CSV FILES
    - READING FROM DIRECTORY
- CORRUPT RECORDS HANDLING IN SPARK (IN JSON AND CSV).
    - PERMISSIVE
    - FAILFAST
    - DROPMALFORMED
- READING TEXT FILE IN SPARK
- READING EXCEL FILES IN SPARK
    - READ FROM SPECIFIC SHEET
    - READ FROM SPECIFIC CELL
- READING PARQUET IN SPARK
    - OPTIONS WHILE READING PARQUET
    - READ MULTIPLE PARQUET WITH SAME DIFFERENT SCHEMA
- READ ORC FILE IN SPARK
- READ AVRO FILE IN SPARK
- TO_AVRO AND FROM_AVRO
- SCENARIO : READING DIRECTORIES OF MULTIPLE FILES.
- DATAFRAME WRITIER API IN SPARK
- WRITE MODES IN SPARK
    - APPEND
    - OVERWRITE (ALL, FEW PARTITIONS)

- ❖ERRORIFEXIST
- ❖IGNORE

❖SCHEMA EVOLUTION IN PARQUET, ORC, AVRO, JSON, CSV FILE FORMATS.
❖WRITE IN EXCEL – APPEND IN EXCEL IN NEW SHEET, OVERWRITE ETC.
❖SCHEMA EVOLUTION IN EXCEL
❖SPARK SQL
❖SPARK TABLES WITH AN OVERVIEW OF HIVE TABLES
❖TEMP VIEW: DIFFERENT WAYS OF CREATING TEMP VIEW

❖CREATING LOCAL AND GLOBALTEMP VIEW IN SPARK
❖SPARK CATALOG
❖USE OF ENABLEHIVESUPPORT() IN SPARK – TO INTERACT WITH HIVE.
❖MANAGED AND EXTERNAL TABLE IN SPARK
❖CREATING MANAGED TABLE – DIFFERENT WAYS (SQL WAY, CTAS, SAVEASTABLE)
❖CREATE YOUR DATABASE AND WRITE YOUR TABLE.
❖CREATING EXTERNAL TABLE
❖DROPPING EXTERNAL TABLE SCENARIO
❖FROM DATAFRAME CREATE TABLE USING SAVEASTABLE() – CREATE EXTERNAL AND MANAGED BOTH
❖COMPRESSION CODECS – READ AND WRITE COMPRESSED FILES

- ❖ LZO
- ❖ SNAPPY
- ❖GZIP
- ❖BZIP2

❖SUMMARY OF ALL FILE FORMATS

CHAPTER V : APACHE SPARK TRANSFORMATIONS AND SQL

- ❖ SELECT – SELECT SINGLE COLUMN, MULTIPLE COLUMNS, COLUMN BY INDEX, ALL COLUMN FROM LIST

- REFERRING COLUMN – COLUMN STRING, COLUMN OBJECT, COLUMN EXPRESSION (EXPR) AND SELECTEXPR
- ALIAS
- DEALING WITH NULL IN SPARK
  - ISNULL
  - ISNOTNULL
  - ISNAN
  - BLANK VS NONE VS “NULL” VS NAN
  - COUNT(*) VS COUNT(1) VS COUNT(‘ABC’) VS COUNT(COL) – WITH NULL.
  - DF.COUNT() AND PYSPARK.SQL.FUNCTIONS.COUNT(COL) IN PYSPARK
  - COUNTDISTINCT(COL OR COLS) WILL NULL
  - COALESCE(*COLS)
  - NA.FILL AND NA.DROP
- COLUMN OPERATORS IN SPARK
  - + / * % > < ==
  - BETWEEN
  - EXPLORATORY DATA ANALYSIS
  - DESCRIBE
  - SUMMARY
- DATETIME FUNCTIONS
  - ADD_MONTHS
  - DATETIME.DATE
  - CURRENT_DATE
  - CURRENT_TIMESTAMP
  - UNIX_TIMESTAMP
  - TO_TIMESTAMP
  - TO_DATE
  - FROM_UNIXTIME
  - DATE_FORMAT
  - DATEDIFF , DATE_SUB , DATE_ADD , DATE_TRUNC
  - YEAR ,MONTH
  - DAYOFMONTH,DAYOFWEEK, WEEKOFYEAR, DAYOFYEAR

- ❖ CASE-WHEN CLAUSE IN PYSPARK AND ADDING NEW COLUMN
- ❖ CONCAT – WITH STRING AND COMPATIBLE ARRAY TYPES (CONCAT WITH NULL)
- ❖ FILTER / WHERE - WITH SINGLE AND MULTIPLE CONDITIONS
- ❖ ISIN, ENDSWITH, STARTSWITH, CONTAINS, ==, !=, <>, FILTER WITH SQL EXPRESSION, LIKE
- ❖ SORT AND ORDERBY
- ❖ ASC() AND DESC()
- ❖ ASC_NULLS_FIRST(), ASC_NULLS_LAST(), DESC_NULLS_FIRST(), DESC_NULLS_LAST()
- ❖ SORTBY AND SORTBYKEY
- ❖ SORTBY VS ORDERBY
- ❖ DROP
- ❖ WITHCOLUMN
- ❖ LIT()
- ❖ CAST() / ASTYPE(DATATYPE) - ASTYPE()
- ❖ STRING FUNCTIONS:
- ❖ SUBSTRING() / SUBSTR()
- ❖ SPLIT()
- ❖ UPPER
- ❖ LOWER
- ❖ INITCAP
- ❖ TRIM
- ❖ RTRIM
- ❖ LTRIM
- ❖ REPLACE (NEW IN 3.5.0)
- ❖ REGEXP_REPLACE
- ❖ CONCAT_WS
- ❖ LOCATE
- ❖ WITHCOLUMNRENAMED (USECASE: RENAME ALL COLUMNS)
- ❖ DISTINCT
- ❖ COUNTDISTINCT
- ❖ DROPDUPLICATES / DROP_DUPLICATES
- ❖ LIMIT

- ❖ IMPORTANT ACTIONS:
  - ❖ HEAD
  - ❖ TAKE
  - ❖ FIRST
  - ❖ SHOW
  - ❖ COUNT
  - ❖ COLLECT

- DATETIME USECASES
  - CONVERT DATETYPE, STRING DATE TO TIMESTAMP TYPE
  - CONVERT TIMESTAMP TYPE TO DATETYPE
  - CONVERT TIMESTAMP AND DATETYPE TO UNIX TIMESTAMP
  - CONVERT UNIX TIMESTAMP TO TIMESTAMP AND CONVERT UNIX TIMESTAMP TO DATE
  - CONVERT STRING DATE OF ANY TYPE TO DESTINATION FORMAT TYPE (DD/MM/YYYY , MM/DD/YYYY , MM-DD-YYYY)
  - CONVERT TIMESTAMP TYPE TO DESIRED FORMAT
- AGGREGATE FUNCTIONS
  - SUM , MIN , MAX , AVG
  - COUNT
  - GROUPBY
  - SINGLE AGGREGATION VS MULTIPLE AGGREGATION
  - AGG (AGGREGATE) IN PYSPARK
  - FIRST , LAST
  - APPROX_COUNT_DISTINCT ,COUNTDISTINCT
  - MEAN
  - ARRAY/COLLECTION FUNCTIONS
  - EXPLODE ,EXPLODE_OUTER
  - POSEXPLODE , POSEXPLODE_OUTER
  - FLATTEN
  - COLLECT_LIST, COLLECT_SET
  - CONTAINS / ARRAY_CONTAINS
  - ARRAY(*COLS)
  - RETRIEVING ELEMENTS FROM ARRAY / MANIPULATION ON ARRAY ELEMENTS IN SPARK
  - ARRAY_DISTINCT(COL)
  - TRANSFORM(COL, F)
  - ARRAY_MAX
  - ARRAY_MIN
  - REVERSE
  - ELEMENT_AT

- JOINS IN PYSPARK
  - INNER JOIN
  - LEFT JOIN / LEFT OUTER JOIN
  - RIGHT JOIN / RIGHT OUTER JOIN
  - FULL JOIN / FULL OUTER JOIN
  - SEMI JOIN / LEFT SEMI JOIN
  - ANTI JOIN / LEFT ANTI JOIN
  - CROSS JOIN / CARTESIAN PRODUCT
  - HANDLING AMBIGUOUS COLUMNS IN JOIN
  - HANDLING AMBIGUOUS COLUMNS IN JOIN
  - ALIAS IN JOIN
  - SELECTING SPECIFIC COLUMNS AFTER JOIN
  - JOIN ON MULTIPLE COLUMNS
  - JOIN MORE THAN TWO DATAFRAMES / TABLES – HOW IT WORKS?
  - JOINS SCENARIOS – WITH NULL / WITHOUT NULL – DIFFERENCE ** IMP INTERVIEW QUESTION
  - INNER, RIGHT, LEFT, FULL JOIN WITH NULL
- MISCELLANEOUS FUNCTIONS
- SHA SHA1
- BASIC MATHS
  - CEIL, COS , LOG , ROUND, SQRT , RAND
- DTYPES , SCHEMA, SCHEMA.FIELDS
- COMPLEX TYPES:
  - UNION /UNIONALL
  - UNIONBYNAME
  - MINUS, INTERSECT
  - TRANSFORM , WITHFIELD
  - WHEN ,OTHERWISE
  - EQNULLSAFE
- CALL FUNCTIONS
  - UDF (USER DEFINED FUNCTION)

- WINDOW FUCNTIONS
  - SUM, MIN , MAX, AVG
  - ROW_NUMBER
  - RANK
  - DENSE_RANK
  - NTILE
  - CUME_DIST
  - LEAD
  - LAG
  - FRAME CLAUSE

CHAPTER VI : SPARK OPTIMIZATION TECHNIQUES / ADVANCED TOPICS:

- TYPES OF OPTIMIZATION – APPLICATION LEVEL AND RESOURCE LEVEL
- OPTIMIZE YOUR SPARK CLUSTER CONFIGURATION
- SPARK CLUSTER AND ITS INTERNAL
- TYPES OF EXECUTOR IN SPARK . FAT EXECUTOR & THIN EXECUTOR
- ON HEAP VS OFF HEAP MEMORY
- HOW TO SELECT NUMBER OF EXECUTOR, NUMBER OF CORES AND MEMORY
- HOW TO SET SPARK SPARK PROPERTIES – METHODS
- RESOURCE ALLOCATION – DYNAMIC VS STATIC
- MEMORY DISTRIBUTION IN APACHE SPARK EXECUTOR
- JAVA HEAP VS EXTERNAL MEMORY
- TOTAL CONTAINER MEMORY
- HOW TO CALCULATE INITIAL NUMBER OF PARTITIONS
- SCENARIO BASED INTERVIEW QUES
- DETERMINE CLUSTER HAS BEEN RESOURCES – OOM OR NOT?
- CALCULATING EXECUTOR CORES AND MEMORY FOR GIVEN REQUIREMENT

- STANDARDIZE FORMULA FOR CORE AND MEMORY CALCULATION
- OPTIMIZE YOUR SPARK CODE
- SHUFFLE PARTITION
- SPARK FILE LAYOUT
- REPARTITION AND COALESCE
- PARTITON SKEW
- WHEN TO INCREASE / DECREASE PARTITIONS
- PARTITIONBY VS BUCKETBY
- CACHE AND PERSIST AND SPARK STORAGE LEVELS
- JOINS IN SPARK
- SPARK JOIN OPTIMIZATION
- FINETUNING VARIOUS SPARK CONFIGURATIONS
- ADAPTIVE QUERY EXECUTION (AQE)
- SPARK EXECUTION PLAN AND EXPLAIN PLAN
- FACT AND DIMENSION
- SLOWLY CHANGING DIMENSION (SCD)
- MONITORING AND DEBUGGING WITH IMPORTANT SPARK CONFIGURATIONS
  - SPARK JOBS NOT STARTING
  - SLOW TASKS - spark.task.cpus
  - SLOW AGGREGATIONS
  - SLOW JOINS
  - SLOW READ AND WRITES
  - DRIVER OOM ERROR
  - EXECUTOR OOM ERROR
  - NO SPACE LEFT ON DISK ERROR
  - SERIALIZATION ERROR
  - DATA SPILL

CHAPTER VII:-AZURE FUNDAMENTALS & STORAGE:

FUNDAMENTALS:

- CLOUD AND ON-PREMISE
- CHARACTERISTICS OF CLOUD
- IAAS PAAS, SAAS
- CLOUD DEPLOYMENT MODELS- PUBLIC, PRIVATE, HYBRID
- CREATE YOUR AZURE ACCOUNT (WORKSPACE)
- AZURE MICROSOFT SERVICES WALKTHROUGH ON PORTAL
- RESOURCE, RESOURCE GROUP, SUBSCRIPTION
- DATA CENTERAZURE REGIONS
- AZURE AVAILABILITY ZONES
- ZONAL SERVICES & ZONE-REDUNDANT SERVICES
- HANDLING DATACENTER FAILURES
- REGION PAIR

STORAGE:

- STORAGE ACCOUNT - BLOB, TABLE, FILE, QUEUE SERVICES
- ACCESS TIERS -DATA ACCESSIBILITY
- LOCALLY REDUNDANT STORAGE(LRS)
- ZONE REDUNDANT STORAGE(ZRS)
- GEO REDUNDANT STORAGE(GRS)
- READ ACCESS GEO REDUNDANT STORAGE
- GEO-ZONE REDUNDANT STORAGE
- READ ACCESS GEO-ZONE REDUNDANT STORAGE
- INTRODUCTION TO DATALAKE
- AZURE DATALAKE STORAGE GEN2 (ADLS GEN2)
- AZURE STORAGE ACCOUNT FEATURES
- ACCESS CONTROL LIST (ACL)
- ACCESS TIERS - HOT, COLD/COOL, ARCHIVE

CHAPTER VIII :- DATABRICKS

- WHAT IS AZURE DATABRICKS
- KEY FEATURES OF DATABRRICKS
- TYPES OF CLUSTERS IN DATABRICKS
  - ALL PURPOSE CLUSTER
  - JOB CLUSTER
  - CLUSTER POOL
- CLUSTER MODES
  - SINGLE NODE
  - STANDARD
  - HIGH CONCURRENCY CLUSTER
- DATABRICKS FILE SYSTEM (DBFS)
- DATABRICKS UTILITIES
  - FILE SYSTEM OPERATIONS (DBUTILS.FS)
  - NOTEBOOK WORKFLOWS (DBUTILS.NOTEBOOK)
  - SECRET MANAGEMENT (DBUTILS.SECRETS)
  - WIDGETS (DBUTILS.WIDGETS)
- DATABRICKS JOBS TASK ORCHESTRATION
  - JOB SCHEDULING
  - TASK DEPENDENCY
  - JOB MONITORING AND LOGGING

CHAPTER IX :- DELTA

- DELTALAKE INTRODUCTION
- LAKEHOUSE ARCHITECTURE
- FEATURES OF DETLA/DELTALAKE
- FILE COMPACTION USING OPTIMIZE
- ZORDER CLUSTERING
- DATA SKIPPING USING STATS
- CACHING WITH DELTA CACHE
- AUTO OPTIMIZATION AND AUTO COMPACTION
- VACUUMING

ADD ON TO BIG DATA JOURNEY

- ONE END TO END SPARK PROJECT
- DOUBT SOLVING SESSIONS
- IMPORTANT INTERVIEW QUESTION
- PRACTICE ASSIGNMENTS
- HOW TO ACE ANY DATA ENGINEERING INTERVIEW – IMPORTANT TIPS AND TRICKS