NEXT IT CAREER

BIG DATA & SPARK

BIG DATA & SPARK COURSE CONTENT

BigData Frameworks and Components	• Hadoop • Spark • Hortonworks • Hive • Impala • Sqoop • PySpark/Scala Spark • SuperSet
Software's and Languages:	• Putty • WinSCP • VIrtualbox/VMware • Hortonworks Sandbox • Python/Scala • SQL • IntelliJ/PyCharm/Anacon da/Jupyter Notebook • Linux • Git
Introduction to BigData and Hadoop	• Introduction to Big Data. • Introduction to Hadoop • Introduction to Spark • Difference between Hadoop and Spark • Common Big Data problems • How Hadoop and Spark • helps to solve the issue
Python	• Python Introduction • Basic Datatypes • Collection Datatypes • Conditional Statements • Functions • Modules • Classes • Exceptions Handling • File Handling • Basic NumPy and Pandas
SQL	• Basic Data types • Collection Data types • Create Table • Alter Table • Drop Table • Insert the data

	• Basic Select • From • Where • Group By • Order By
Hadoop Ecosystem	• What is Hadoop • Hadoop Architecture • What is Namenode • What is Datanode • What is Resource Manager • What is Application Manager
HFDC	• What is HDFS • Advantages and Disadvantages • Scalability • Replication factor • Block details • HDFS commands
Yarn	• Daemons in Yarn • Node Manager • Application Master • Resource Manager • Yarn Commands • How Yarn allocates resources • How Spark/MapReduce running in Yarn
HIVE	• Hive architecture • Difference between SQL and HQL. Will try to cover all the data types and SQL standards. DDL and DML. • Different types of files handling • How to use SerDes • What is Partitioning • What is Bucketing • ORC and Parquet file handling and differences • Limitations in Hive
Sqoop	• Sqoop architecture • Import data from Oracle • Import data from MySQL • Import data from MsSql data • Shell script importance in Sqoop • Import data to Hive • Compression techniques (parquet,

	sequence, Avro) • Best practice
Spark	• Why Spark why not Hadoop? • HDFS/Yarn importance in Spark • Spark architecture • Different types of APIs • RDD (Resilient Distributed Dataset) • Dataframe • Dataset • Why spark faster than MapReduce? • Why /How spark process in Memory? • Why MapReduce Slow?
RDD Internals	• RDD • RDD Properties • Immutability • Laziness • Fault tolerance • SparkContext, SqlContext, SparkSession Internals • Create RDD different ways • Transformations • Action • Commonly used transformations & Actions • Narrow transformations • Wide transformations • Debugging transformations • Spark web UI • Dataframe:
Spark SQL	• Convert RDD to Dataframe • Python Dataframe • Spark dataframe Introduction • Dataframe reader • Dataframe Vs dataset • Process different type data • CSV • Json (complex) • XML • Avro • Orc • Text data • Parquet • Spark vs Hive

	• Spark process Hive data • Process Different Database data • Oracle • MySQL • MySQL data analysis • Sqoop Vs Spark • Data-migration Project • ETL project Vs Spark project • Process different NoSQL Database data • Spark integrate with HBase and Phoenix • Spark Cassandra Integration • Spark MongoDb integration
PySpark Advanced Concepts:	• Dataset Api importance • Spark Memory management • Resource optimization • Spark submit num- executors, --executor- cores, --executor- memory importance • Spark debugging using client mode and web UI. • Get data from S3 and • process using Databricks • How to automate spark using Airflow
Other important topics	• Git commands • Commit your IntelliJ code to GitHub • Resume preparation • Mock Tests • Interview tips
	• ls command • cat command • cp command • mv command • mkdir command • rmdir command • rm command • touch command • locate command • find command • grep command • kill command • ping command • wget command

Linux Basics	• uname command • top command • history command • man command • echo command • zip, unzip commands • hostname command • useradd, userdel commands • apt-get command • nano, vi, jed commands • alias, unalias commands • su command • htop command • ps command • sudo command • pwd command • cd

INTERVIEW QUESTIONS AND ANSWERS

Duration : 90 days
Location : Hyderabad, KPHB

Project Orientation Program

100% JOB ASSISTANCE PROGRAM

Contact: 9542589410, 7995138210