PENTAGON SPACE
Mastering The Future

BIG DATA

BIG DATA TOPICS

What is Big Data?
- Big Data History, Technologies, Use cases
- Benefits of Big Data
- Reasons to Learn Big Data
- Top Big Data Tools
- Why Big Data is Gaining Hype
- Lambda Architecture for Big Data
- Spark Map vs FlatMap
- Lazy Evaluation in Spark
- Fault Tolerance in Spark
- Spark Directed Acyclic Graphs (DAG)
- Spark Cluster Managers
- Spark – Hadoop Compatibility
- Performance Tuning in Spark
- Apache Spark Executor
- Apache Spark Stages
- Limitations & Drawbacks of Spark

SPARK

Introduction to Spark

- What is Spark?
- Spark Environment Setup for Ubuntu
- Spark Installation in Standalone Mode
- Spark Terminologies & Concepts
- Spark Ecosystem Components
- How does Spark Work?- Runtime Architecture
- Why Should I Learn Spark?
- Spark Features
- Spark Shell Commands
- Installing Multinode Clusters for Spark
- SparkContext and its Functions
- Spark RDD, Features, and Operations
- Ways to Create a Spark RDD
- Spark RDD Persistence & Caching
- Spark RDD Features
- Spark RDD Limitations
- Spark Transformations and Actions on RDDs
- Spark RDD Lineage
- Apache Spark Paired RDD
- Create Spark Projects in Eclipse
- Spark Notes for Beginners & Experienced
- Apache Spark Use Cases in Real Time
- Introduction to Spark SQL
- Apache Spark SQL Tutorial
- Spark SQL Features
- DataFrames in Spark SQL
- Spark SQL DataSets
- PySpark SparkFiles & their Class Methods
- PySpark RDD with Operations & Commands
- PySpark SparkConf- Attributes & Applications
- PySpark SparkContext and its Parameters
- PySpark MLlib- Algorithms & Parameters
- PySpark Profiler- Methods & Functions

PySpark Serializers- Marshal & Pickle
PySpark StorageLevel
PySpark StatusTracker(jtracker)
PySpark Broadcast & Accumulator
PySpark Interview Questions
Optimization in Spark SQL
Spark SQL Performance Tuning
RDD vs DataFrame vs DataSet
SparkR DataFrame and Operations
Introduction to Spark Streaming
Spark Streaming- DStreams (Discretized Streams)
Transformations in Spark Streaming
Checkpointing in Spark Streaming
Structured Streaming in SparkR
Creating SparkDataFrames in SparkR
GraphX API in Spark
Spark GraphX Features
Graph Algorithms GraphX
Spark Machine Learning with R
Apache Spark MLlib Algorithm Featurization
Spark MLlib Data Types
Spark Machine Learning Algorithm
Spark Streaming vs Apache Storm
Spark SQL vs Hive

Kafka

Introduction to Apache Kafka
Features of Kafka
Kafka Terminologies
Apache Kafka Pros & Cons
Apache Kafka Applications

Apache Kafka Architecture
Apache Kafka Workflow(Pub-Sub Messagin
Kafka Cluster Setup
Kafka vs RabbitMQ
Kafka vs Storm
Kafka vs RabbitMQ
Kafka Producers
Kafka Consumers
Kafka Brokers
Kafka Queuing- Messaging System
Creating Kafka Clients
Apache Kafka Connect
Kafka-Docker: Kafka using Docker
Kafka Topics
Kafka Tools
Kafka Monitoring Tools
Kafka Operations with Commands
Role of Zookeeper in Kafka
Kafka Streams- Stream & Real-Time Processing
Apache Kafka + Spark Streaming Integration
Kafka + Hadoop Integration
Kafka Optimization- Performance Tuning
Kafka Load Testing with JMeter
Storm-Kafka Integration
Kafka SerDe
Kafka Schema Registry
Security Concepts in Kafka

Hadoop

Hadoop Architecture
Internal Working of Hadoop
Hadoop Commands

Hadoop Clusters
Hadoop High Availability
Hadoop Schedulers
Hadoop Distributed Cache
Hadoop Automatic Failover
Hadoop Limitations & Solutions
Hadoop HBase Compaction & Data Locality in Hadoop
Install Hadoop
Hadoop vs Cassandra
Hadoop vs MongoDB
Hadoop vs Spark vs Flink

HDFS

Introduction to HDFS
HDFS Architecture
Features of HDFS
HDFS Read-Write Operations
HDFS Data Read Operation
HDFS Data Write Operation
HDFS Commands
HDFS Data Blocks
HDFS Rack Awareness
HDFS High Availability
HDFS NameNode High Availability
HDFS Federation- Architecture & Benefits
HDFS Disk Balancer
Erasure Coding in HDFS
Fault Tolerance in HDFS
MapReduce Tutorials

Map Reduce

Introduction to MapReduce
- MapReduce Data Flow
- How Hadoop MapReduce Works
- MapReduce Mapper
- MapReduce Reducer
- MapReduce Key-Value Pairs
- MapReduce InputFormat
- MapReduce InputSplit
- MapReduce RecordReader
- MapReduce Partitioner
- MapReduce Combiner
- Shuffling-Sorting in MapReduce
- MapReduce OutputFormat
- MapReduce InputSplit vs Blocks

Hive

Introduction to Apache Hive
- Apache Hive Architecture
- Apache Hive Data Types
- Apache Hive Built-in Operators
- Built-In Functions in Hive
- User-Defined Functions (UDF) in Hive
- Hive DDL Commands and Types
- Views and Indexes in Hive
- Configuring Hive Metastores
- Developing Data Models in Hive
- Hive Custom and Built-in SerDe
- Hive Data Partitioning
- Bucketing in Hive

Hive Partitioning vs Bucketing
Apache Hive Joins and Types
Map Join in Hive
Bucket Map Join in Hive
Skew Join in Hive
Hive SMB (Sort Merge Bucket) Join
Hive Internal vs External Tables
Configuring Hive Metastore to MySQL
HiveQL (Hive Query Language) Select Statement
HiveQL Group By Clause
HiveQL Order By Clause
7 Best Hive Optimization Techniques
HBase vs Hive
Pig vs Hive
Impala vs Hive
Best Hive Books
Impala Tutorials

Impala

Introduction to Impala
Impala Environment Setup
Features of Impala
Impala Architecture
Impala Use Cases
Impala Built-in Functions
Impala User Defined Functions (UDF)
Impala Data Types
Comments in Impala
Introduction to Impala SQL (Impala Query Language)
Selecting a Database with Hue Browser- Impala SQL
CREATE DATABASE in Impala SQL
DROP DATABASE in Impala SQL

DESCRIBE Statement in Impala SQL
SELECT Statement in Impala SQL
CREATE TABLE Statement in Impala SQL
DROP TABLE Statement in Impala SQL
INSERT Statement in Impala SQL
TRUNCATE TABLE Statement in Impala SQL
SHOW Statement in Impala SQL
CREATE VIEW Statement in Impala SQL
DROP VIEW Statement in Impala SQL
ALTER VIEW Statement in Impala SQL
ALTER TABLE Statement in Impala SQL
ORDER BY Clause in Impala SQL
GROUP BY Clause in Impala SQL
LIMIT Clause in Impala SQL
HAVING Clause in Impala SQL
WITH Clause in Impala SQL
UNION Clause in Impala SQL
OFFSET Clause in Impala SQL
DISTINCT Operator in Impala SQL
Impala Shell Commands

HBase

Introduction to HBase
Features of HBase
HBase Architecture
HBase Pros & Cons
HBase Use Cases
HBase Shell Commands and Usage
HBase Read & Write Operations
HBase Commands to Define and Manipulate Data
HBase Table Management Commands
HBase Data Manipulation Commands- Create, Truncate, Scan

HBase Admin API
HBase Client API
HBase MemStore Configuration and Benefits
HBase Optimization: Performance Tuning
HBase Compaction and Data Locality in Hadoop

Pig

Introduction to Pig
Pig Environment Setup
Apache Pig Features
Apache Pig Architecture
A Comprehensive Guide to Apache Pig
Pros and Cons of Pig
Pig Architecture & Execution Modes
Pig Grunt Shell Commands
Pig Built-in Functions
User-Defined Functions in Pig
Introduction to Pig Latin
Pig Latin Operators and Statements
Executing Apache Pig Scripts
Reading and Storing Pig Data and Operators
Apache Pig Execution Modes and Mechanisms

Sqoop

Introduction to Sqoop
Sqoop Environment Setup
Sqoop Features
Sqoop Architecture
Importing Data from RDBMS to HDFS- Sqoop
Exporting Data from HDFS to RDBMS- Sqoop
Sqoop Eval- Commands and Query Evaluation

Sqoop import-all-tables
Sqoop Validation- Interfaces and Limitations
Sqoop Codegen Arguments and Commands
Combining Datasets with Sqoop Merge
Sqoop Metastore Tool
Sqoop Troubleshooting Tips & Known Issues
Sqoop List Tables and their Arguments
Sqoop List Databases and Syntax
Creating and Executing Jobs in Sqoop
Sqoop Connectors & Drivers (JDBC)
Sqoop Import Mainframe Tool
Databases Supported in Sqoop

ZooKeeper

Introduction to ZooKeeper
ZooKeeper Features
ZooKeeper Architecture
ZooKeeper Workflow
Terminologies of ZooKeeper
ZooKeeper Applications
Pros and Cons of ZooKeeper
ZooKeeper Data Model
ZooKeeper Znode
Leader Election in ZooKeeper
ZooKeeper CLI (Command Line Interface)
ZooKeeper Access Control with ACLs
ZooKeeper API- Java & C Bindings
ZooKeeper Sessions
ZooKeeper Queues- Priority and Producer-Consumer
ZooKeeper Locks- Shared and Recoverable Shared

Ambari

Introduction to Ambari
- Ambari Features
- Ambari Architecture
- Pros of Ambari
- Ambari Views
- Ambari Groups and Users
- Ambari Web UI- Accessing and Troubleshooting
- Ambari Cluster Setup
- Ambari Security Guide- Kerberos
- Ambari Troubleshooting
- Ambari Uses

Apache AVRO

Introduction to AVRO
- Features of AVRO
- AVRO Uses
- AVRO Schema
- AVRO Reference API
- Serialization in AVRO
- AVRO SerDe- Code Generation
- AVRO SerDe- Parsers
- AVRO- SASL Profile

YARN

Introduction to Hadoop YARN
- Hadoop YARN Resource Manager
- Hadoop YARN Node Manager
- Apache Mesos vs Hadoop YARN

CASSANDRA

Introduction to Cassandra
- Cassandra Environment Setup
- Features of Cassandra
- Cassandra Applications
- Cassandra Glossary
- Cassandra Architecture
- Cassandra vs Hadoop
- Cassandra vs MongoDB
- Cassandra vs RDBMS
- Cassandra vs HBase
- Best Cassandra Books
- Cassandra Techniques
- The Data Model in Cassandra
- Cassandra API- References, CQL, Thrift
- 4 Cassandra Monitoring Tools
- Cassandra Clusters and the Cluster Builder
- Cassandra CRUD Operations
- Cassandra Query Language Shell (CQLSH)
- 10 Cassandra Documented Shell Commands
- Data Definition Commands in CQL
- Data Manipulation Commands in CQL
- CQL Clauses- SELECT, WHERE, ORDER BY
- Cassandra Data Types- Built-in, Collection, User-defined
- User-defined Data Types in Cassandra
- Cassandra Collection Data Types- List, Set, Map
- Troubleshooting in Cassandra

MONGODB

Introduction to MongoDB
- MongoDB Environment Setup

Features of MongoDB
Pros & Cons of MongoDB
Comprehansive Guide
Why MongoDB should we learn
The 16 Data Types of MongoDB
Updating Documents in MongoDB
Deleting Documents in MongoDB
Creating a Database in MongoDB
Dropping a Database in MongoDB
Creating & Dropping Collections in MongoDB
MongoDB Data Modeling with Document Structure
Projection in MongoDB and How it Affects Performance
MongoDB Capped Collection
Operations Performed on Document
Backup & Restore Methods
MongoDB Text Search
MongoDB RockMongo
MongoDB GridFS
MongoDB Regular Expression (regex)
How to Execute MongoDB in PHP
MongoDB Vs Hadoop
Limiting Records in MongoDB with skip()
MongoDB Indexes and their Types
MongoDB Relationships & Database Reference
How to Execute MongoDB in Java
MongoDB Covered & Analyzing Query
MongoDB ObjectId & Atomic Operations
Generate Auto Increment Sequence
MongoDB Aggregation
MongoDB Replication & Sharding
Working of MapReduce in MongoDB