BIG DATA TRAINING

Training Session Details
Please review the following detailed training session contents that will be covered during the training.

Day 1: Introduction to Big Data/ Defining Big Data

Session 1: Overview of Big Data
   1. Big Data trends and Industry transformation (including challenges)
   2. Dimensions of Big Data: volume, velocity, variety
   3. Establishing the business importance of Big Data

Session 2: Evolution of Big Data
   1. Big Data history, and how it emerged
   2. Vendors and Tools for Big Data
   3. Big Data Industry scenarios

Session 3: The Big Data Technology Stack
   1. Storage, Hadoop Distributed File System, MapReduce jobs, HIVE and JavaScript Console, Query Stack, Reporting, Dash-boarding etc.
      o Hadoop: HDFS and MapReduce

Session 4: Big Data Considerations
   1. Difference between traditional approaches and Big Data
      o HDFS vs NFS
      o Traditional SQL System vs MPP systems
      o Traditional RDBMS vs MapReduce
      o Big Data Analytics vs Traditional BI
   2. Big Data Adoption Challenges
   3. Benefits of Big Data

Day 2: Deep dive into Big Data Technology Stack

Session 1: Data Management in Big Data
   1. Data types and data models:
      o key value,
      o graph,
omni212

- document,
  - column-family
2. HBase, Hive (overview), MongoDB (Overview)
3. Storage related products/services by Big Data Vendors - Microsoft HDInsight and Azure Storage, Cassandra, Hypertable, Amazon S3, BigTable, DynamoDB and MongoDB

Session 2: Processing Big Data
1. MapReduce Programming – Map and Reduce Concepts, Programming Framework, Executing Hadoop MapReduce jobs
2. RHadoop (Stat Processing)

Session 3: Big Data Development Framework
1. Abstracting Hadoop MapReduce jobs with Pig
2. Querying with Hive (4)
3. Machine Learning Concepts and technologies: Mining data with Mahout
4. Graphs (Pegasus)
5. HCatalog

Session 4: Big Data - BI and Reporting
1. Data Visualization tools and methods
2. Predictive Analytics
3. Reporting tools

Session 5: Big Data Integration and Management
1. Integration Layer: Integrating disparate data stores - Data Place, Connecting and extracting data from storage
   a. SQOOP,
   b. Flume (Log Collector)
2. Management Framework
   a. Ambari (provisioning, managing, and monitoring clusters)
   b. ZooKeeper (Coordination)
   c. Oozie (workflow management)
3. Security and Governance

Day 3: Developing and Deploying a Big Data Solution

Session 1: HDInsight - Microsoft’s Hadoop based service
1. Working with Azure
   o Azure Table Storage and BLOBS
2. Components of Hadoop stack on Microsoft HDInsight - Hadoop Distributed File System, Running MapReduce jobs, HIVE and JavaScript Console, etc.

3. Overview of crime analysis scenario

4. Environment setup related discussions

Practice Session for crime analysis scenario:

a. Creating a new account
b. Creating an Azure Storage container
c. Creating HDInsight Cluster (linked to container)
d. Exploring the cluster: HDFS Console, Hive Console, Cluster management options
e. Running sample map-reduce jobs (from within available samples only)
f. Uploading data into container
g. Creating and running new Map Reduce jobs
h. Viewing status of map-reduce jobs on the cluster
i. Viewing output files in the storage container
j. Creating Hive Tables
k. Loading and querying Hive Tables

Session 2: Explore and visualize data with Microsoft Excel

1. Excel as a Data Visualization Tool- Power Query, Power View, Power Pivot etc.
2. Power BI- An Introduction

Practice Session:

a. HDInsight connectivity configuration
b. Hive ODBC Driver
c. Power Query for fetching data
d. Power Pivot for Data Modelling
e. Power View and Power Map for Visualization

Session 3: Advanced Analytics and Reporting

1. Export and import data using SQOOP
2. Predictive analysis and interactive exploration using Excel Data Mining
3. Reporting against HDInsight data using SQL Server Reporting Services

Practice Session:
a. SQOOP
b. Excel Data Mining Add-in
c. SQL Server Reporting Service
d. Azure Reporting Services