Big Data Summer Training

BigData Analytics-Python Programming

Prepared and Presenting By: Amrit Chhetri (Certified BigData Analyst),
Principal Techno-Functional Consultant and Principal IT Security Consultant

knowledgelab SM

Incubated @ STEP IIT (Kgp)

About Me :

- Me:
  - I’m Amrit Chhetri from Bara Mangwa, West Bengal, India, a beautiful Village/Place in Darjeeling.
  - I am CSCU, CEH, CHFI,CPT, CAD, CPD, IOT & BigData Analyst(University of California), Information Security Specialist(Open University, UK) and Machine Learning Enthusiast (University of California[USA] and Open University[UK]), Certified Cyber Physical System Exert(Open University[UK]) and Certified Smart City Expert.
- Current Position:
  - Principal IT Security Consultant/Instructor, Principal Forensics Investigator and Principal Techno-Functional Consultant/Instructor with RCS
  - BigData Consultant to KnowledgeLab
- Experiences:
  - I was J2EE Developer and BI System Architect/Designer of DSS for APL and Disney World
  - I have played the role of BI Evangelist and Pre-Sales Head for BI System* from OST
  - I have worked as Business Intelligence Consultant for national and multi-national companies including HSBC, APL, Disney, Fidedality , LG(India) , Fidelity, BOR(currently ICICI), Reliance Power. * *Top 5 Indian BI System (by NASSCOM)*

Data Science and BigData Processing with Python Training Session-VI

Agendas/Modules:

- Database Programming-SQLite
- Database Programming-MongoDB
- Database Programming-MySQL
- MapReduce using MRJob-Advanced
- Advanced Web-Data Programming
- BigData Programming with Scala
- Advanced Java Programming-JDBC
- Building Report using BIRT
- Machine Data Analysis using Splunk
- BigData Advanced Analytics-Introduction
- BigData Platforms Preparations
- Introduction to Hive-Syncfusion

Database Programming-SQLite:

- SQLite is an embedded Database system and it is a primary database of Android Applications.
- Python's Sqlite3 module is used to write programs for SQLite Database
- Common Functions are:
  - connect() :
  - cursor() :
  - fetchall() :
  - execute() :
- Alternatives to SQLite is Derby, MySQL embedded version, etc
- SQLite Tables can also be accessed using JDBC programs

Database Programming-MongoDB:

- Mongodb is an Non-SQL Database and it is used to storage and process larger volume of data
- pymongo is a standard module that is used to write python-codes for mongodb
- The common functions of pymongo are:
  - Insert(key: values) Writes values to MongoDB tables ;
  - find_one() :
- The steps to run Python code for mongodb:
  - Install mongodb using default configuration
  - Install mongodb module using pip3.exe install pymongo
  - Write Python code using pymongo
  - Run mongodb as mongod.exe --dbpath D:\MONGODB --storageEngine=mmapv1
- MongoDB can also be accessed using JDBC code

Database Programming-MySQL:

- MySQL is an enterprise-grade Database System from Oracle
- MySQL is one of the standard RDBMS in Hadoop Ecosystem
- MySQL-python important functions :
  - connect() :
  - cursor() :
  - fetchall()
  - execute()
- MySQL Database can also be accessed using JDBC

MapReduce using MRJob-Advanced:

- Marjob performs MapReduce using mapper(), reducer() and combiner()
- Example:

```
from mrjob.job import MRJob ; import re
expression= re.compile(r"[\w']+")
class WordCount(MRJob):
    def mapper(self, _, line):
        for word in expression.findall(line):
            yield (word.lower(), 1)
    def combiner(self, word, counts):
        yield (word, sum(counts))
    def reducer(self, word, counts):
        yield (word, sum(counts))
if __name__ == '__main__':
    WordCount.run()
```

Python Regular Expressions:

- Regular expression is UNIX-style expression using character sequence
- Regular expression is achieved by importing re module
- The common function are:
    - re.match(pattern, string, flag=0) ; re.search(pattern, string)
    - re.findall(pattern, string)
- Example:

```
import re
fo = open("data.txt", "r") ;    line = fo. readline();    words=line.split(" ")
for word in words:
    if re.search("Data", word):
        print("Data is there")
    else:
        print("Data Not Found")
```

Advanced Web Data Programming:

- Streaming of Twitter feeds is performed using Tweepy
- The steps to perform Twitter streaming are:
  - Create a Twitter Account , if does not have
  - Get Consumer Key and API Key
  - Write Python code to accessing Twitter feeds using Tweepy
  - Save the tweets or feeds into Database capable of storage larger volume of data – MongoDB or Cassendra, etc
- The data populated by weepy can be loaded directly into Hive, HBase or MapReuced and saved to HDFS as intermediary or final outcomes
- Twitter Sentimental Analysis is common application possible using Tweepy

Scala Programming Fundamentals:

- Scala is high-level Programming Language and it is used in writing code to access Database Engine
- Steps to write Scala using Eclipse:
  - Install Eclipse and add Scala Plugin
  - Download and install Scala Compiler or Environment
  - Create Scala Project and start developing codes on Scala
- Scala is used for MapReduce Programming and some of some of the components of Hadoop are programmed on Scala
- Scala is also used to call HiveQL Scripts, Pig Scripts and Spark RDD
- Scala follows the programming paradigm of Java with its own programming construct

Advanced Java Programming-JDBC:

- Java JBBC Programs are used to works with Databases inside an application
- JDBC Programs primarily has three components
  - Java Code : Code to logics necessary for selecting, updating, deleting and inserting data
  - JDBC Driver : Program to establish connectivity to Database Server
  - Database Engine : Database Systems like MongoDB, MySQL, Oracle, MSSQL , PostgreSQL, Sybase, Informix, IBM DB2
- Steps to write JDBC in Eclipse:
  - Create Java Project and start MySQL Database or know the Database System to connect
  - Placed Driver's JAR into Build-path
  - Create Java code to perform SQL actions to the Database system
  - To write good program always use PreparedStatement

Building Report using BIRT:

- BIRT is one of the Report Designer for BigData
- BIRT support accessing data from Hive, HDFS and Hbase
- BIRT also supports JDBC Connectivity to various database
- BIRT works fine with Open Source ETL Tools like Pentahoo and Talend Studio
- Designing Report using BIRT:
  - Install BIRT Plugin or get BIRT Report Deisgner
  - Create BIRT Project and create data-sources
  - Create Result-Set using BIRT’s Query Editor
  - Select the type of Report and put the columns of your interest on Report

Machine Data Analysis using Splunk:

- Splunk is Open Source Tool for machine generated data
- It support Search and it has it own search engine and syntaxes and it works on Windows and Linux
- It is also used on Analyzing different types of logs generated by BigData System , including Apache Web Server, Weblogic Application Server

BigData Advanced Analytics Introduction:

- Hadoop Analytics is pointed to extract data from heterogeneous sources in Hadoop System
- The common Data Storage systems in Hadoop are HDFS, Hive, HBase, Logs, No-SQL(Mongo, Cassandra).
- The most effective tools for Statistical Analysis of MapReduce Data are
  - MARLAB
  - Octave
  - R
  - Spark
- The tools which are used to move data in this ecosystem is called ETL , Extraction, Transformation and Load and they
  - Scoop
  - Pentahoo ETL, Talend

BigData Platforms Preparations:

- The distribution of Hadoop are :
  - Deployable Package/Unit: Cloudera ,MapR , Hortonworks , Syncfusion
  - Self-Configured : Apache Hadoop 2.7.2 on Ubuntu 15.04
  - Hadoop-As-A-Service : Qubole, MS Azur, Amazon AWZ EC2
- Hadoop 2.7.2 can be configured on Ubuntu 15.04 for self-made Hadoop Stack
- The major steps of Hadoop configuration(single node):
  - Installation of JDK and installation of SSH
  - Extraction Hadoop archive and moving to a folder
  - Adding new entries inside configuration files
  - Creating HDFS folder inside Hadoop System
  - Starting Hadoop, WOW !

Introduction to Hive-Syncfusion:

- Hive leverages SQL-Engine like Services and it supports HiveQL
- Steps to run Hive on Syncfusion Platforms
  - Install MS .Net Framework and install Syncfusion Studio on Windows Machine
  - Open Syncfusion Studio and run ‘Command Shell’ available at top of Syncfusion Studio
  - Type hive to start the HIVE prompt, hive>
  - Table Create HiveQL : hive> CREATE TABLE PRD(id int, category int);
  - Data Insert HiveQL : INSERT INTO PRD VALUES (1, 100);
  - Select HiveQL : SELECT id, category FROM PRD;
- The common HiveQL statement
  - hive> show databases ;
  - hive> use <database name>;
  - hive> show tables;
  - Hive> describe <table>

THANK YOU ALL

knowledgelab SM

Incubated @ STEP IIT (Kgp)