Training on *Clouder Data Platform and MLOps/AIOps*

Webinar on :

BigData Technology Stacks

Edges To AI Application On Hybrid Data Cloud & Spark Application Performance Tuning

SESSION-01

Prepared and presenting by:
Amrit Chhetri
Cyber Security Architect & CEI(RCS/Rosefinch, Siliguri, West Bengal)
DFIR Expert|AI, Cyber Security & Digital Forensics Researcher
Certified Forensic Psychologist,
Associate Technical Editor(4N6)
Tech Speaker and Forensic Researcher
Member Of: DSCI(Individual) & Nasscom Community

Session from Day 1(8 Hours)

BigData Technologies

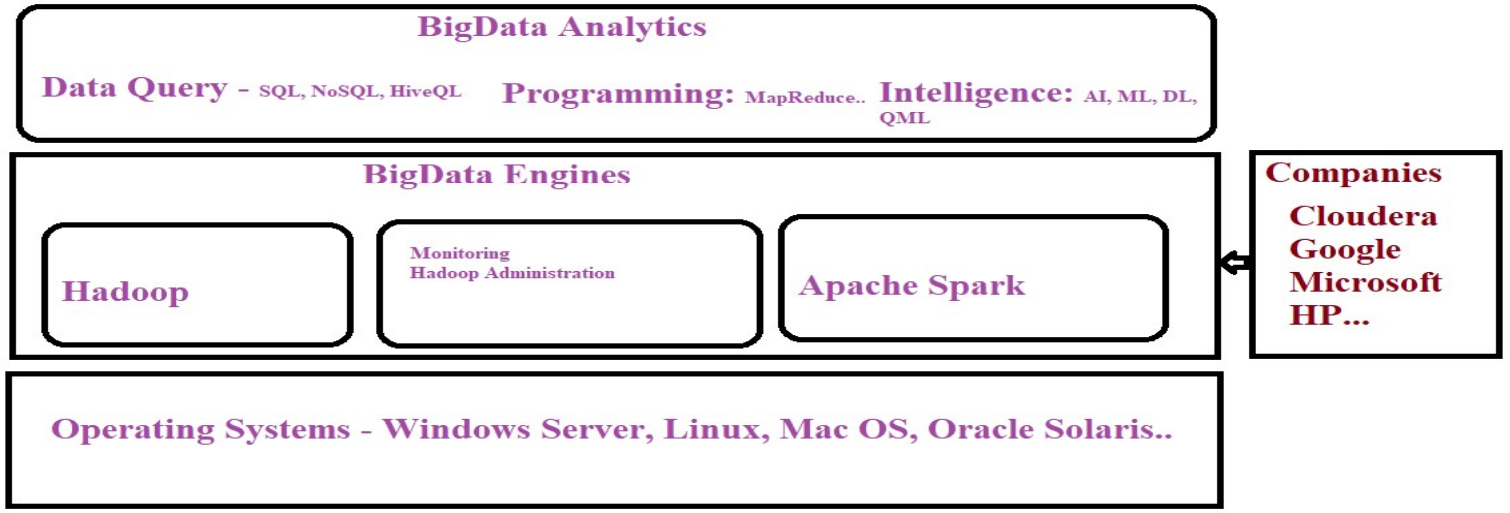
Agendas:

1. BigData Components and Architecture
2. Data Lake and Database Systems
3. BigData and AI on Cloud Platforms
4. CDP Access And Administration(Basic)
5. BigData Developer Platform-Hadoop, Spark & Scala
6. Labs. of Session 1- On BigData Technology
7. BigData Quotient Assessment- Short Hackathons, CTFs & Quiz

BigData Components & Architecture:

(1,Defining BigData and Challenges)

BigData Technology-*2 Minutes Understand*:



Context: BIgData Definition, Comapanies, Programming, DevOps(Development), DevSecOps(Cyber Security), Design Principles(ZTA, Architecture by Domains, Secure By Design, Secure Coding, Cyber Resilience with SOC Technology)

BigData Technology- *6Vs* :

- BigData is a Database Technology for Storing and Processing Huge Volume of Data
- It is large set of data and follows 6Vs- Volume, Variety, Velocity, Value, Veracity, Variability- which are not supported in Traditional Data Processing Systems
- "BigData is a collection of very large set of data which includes structured ,semi-structured and non- structured data and they are processed by non-traditional and parallel data processing engine to produce meaningful insights for Businesses and other Data Analysis Requirements." – Data Scientist

BigData Components & Architecture:

(2, Challenges & Technologies)

Challenges with BigData Solutions- Summary:

- Challenges in BigData: Capture, Citation and Storage, Searching
- Existing Solutions: Query, visualizations and analysis which are handled by Apache Hadoop or Spark in Distributed and Parallel Computation

BigData Technologies:

- Apache Hadoop Stack or Apache Hadoop-based Platforms is the distributed Data Processing Platforms and it solves the issues of BigData
- Apache Hadoop is the main Platform of BigData Hadoop Stack
- BigData and BigData Analytics is the key Software Component of today's Dashboard Analytics

Distributions Of Hadoop:

- MapR Hadoop, Apache Hadoop(Open Source Software)
- Apache Hadoop is distributed or shipped by other BigData companies/vendors too - MapR, Cloudera, Oracle, Google, Microsoft
- MS Azure HDInsight ,Oracle Hadoop, Informatica....

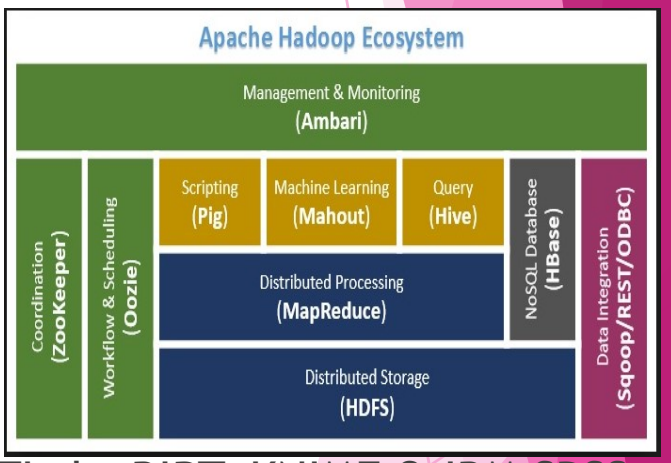
BigData Components & Architecture:

(3, Anallytics)

BigData Analytics Components:

- Two core components of BigData Analytics: Apache Hadoop and HDFS
- BigData Analytics Dashboard Technologies: KNIME, IBM SPSS
- Analytics with AI: IBM SPSS, **RapidMiner,** Apache Pinot(OLAP)

BigData Hadoop System Components :



- Distributed Processing Engine: Hadoop, Spark and Tez
- Distributed File System : HDFS and RDD
- Data Warehouse System : HBase
- Scripting/Query : Pig and Hive
- Database System : NoSQL, Cassandra
- Data Analysis Platforms : Hive, Spark, R/Octave/MATLab, BIRT, KNIME & IBM SPSS
- Monitoring : Apache Amber
- Machine Learning API/Platforms : Mahout, Spark, MATLAB, Google TensorFlow, Quantum Machine Learning, CML(Cloudera Machine Learning)
- Distributed Networks : Blockchain SWARM Intelligence

Data Lake and Database Systems:

(1, Technology)

Data Lake:

- " A Data Lake is centralized repository designed to store, process and secure BigData- large amount of structured, semi-structured and unstructured data"
- Data Warehouse with BigData Technologies
- Engineered for Security Analytics, Business Analytics...

Benefits of Data Lakes:

- Advanced Analytics Support- Machine Learning, QML and Mobile Machine Learning
- Scalability, Heterogeneous Data Formats
- Data Quality Assurance, Democratized Architecture
- Better Performance

Database Systems In DataLakes:

- Relation Database Systems : MySQL, IBM DB2, Oracle Database, MS SQL Server...
- N0-SQL Database Systems : Cassendra, No-SQL
- Embedded Database Systems:
- OLAP Systems : Apache Pinot

Data Lake and Database Systems:

(2, Architecture And Design

Data Lake Platforms:

- BigData Processing Engine- Hadoop, Spark
- Analytics
- ?

Data Lake Architecture(Generic):

Design Frameworks and Standards:

- Secure By Design And DevSecOps:
- Zero Trust Architecture: NIST 300-161, TOGAF :
- BigData System Design Engineering :
- Functional Compliances: PCI-DSS, HIPAA, SOX, GLBA,

Design Tools:

- Secure By Design:

BigData And AI On Cloud Platforms:

BigData On Cloud:

- Cost Effective Architecture
- Less Expensive and Lower TCO
- Customer Support and Community Support
- Availability of SLA

AI Cloud Platforms:

- Easy to Maintain and Deploy
- General Trend....

Native Clouds Platforms:

- Neutanix
- General Trend....

BigData Developer Platform-Hadoop, Spark & Scala:

Labs Environment Options:

- CCA-176 Registration Free – preloaded with Apache Spark
- CDP Trial Access
- CDP In-House
- Open Source BigData Environment- Hadoop, Spark and Scala(Windows and Linux)
  - Hadoop: <URL>
  - Spark : <URL>
  - Scala : <URL>

Programming:

1. Java and MapReduce : Eclipse with Plugins
2. Python for Machine Learning : PyCharm
3. Scala for Data Analysis : IntelliJ
4. Java+Python+Scala : Eclipse

PyTorch On- Windows and Linux:

PyTorch On Windows:

1. Install PyTorch with installation matrices(CPU), details <URL>
2. Install Pilchard Edu(Free) and install PyTorch API
3. Create Example Models on Data Analysis or clone from <>

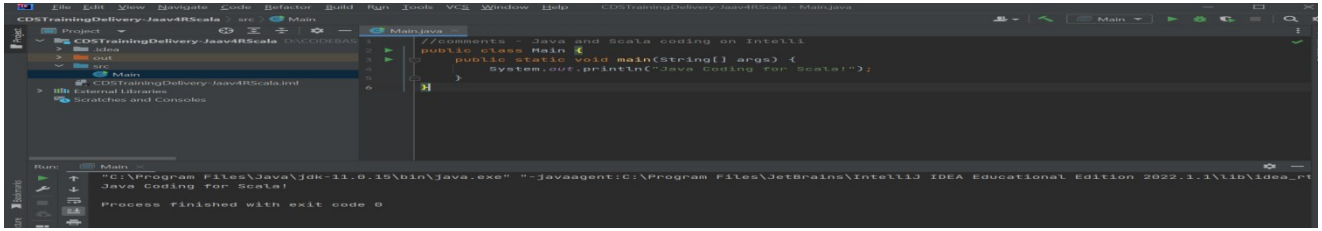
PyTorch On Linux:

1. Install PyTorch with installation matrices(CPU), details <URL>
2. Install Pilchard Edu(Free) and install PyTorch API
3. Create Example Models on Data Analysis or clone from <>

Labs. Of Session 1- On BigData Technology:

Java Basic for Scala Programming:

1. Get IntelliJ from <> and install
2. Create Java Project - CDSTrainingDelivery-Jaav4RScala
3. Create example codes and run

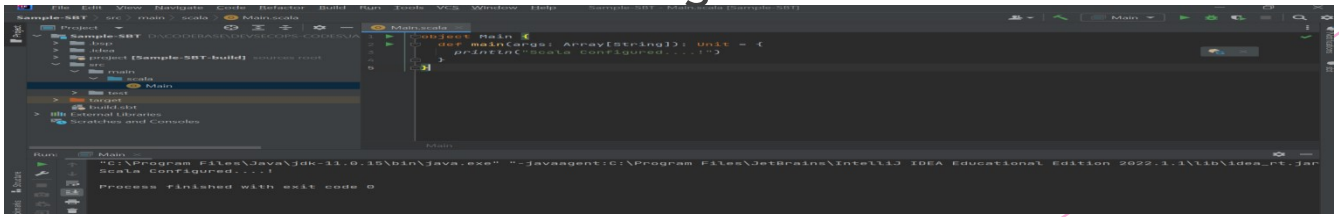


Scala with GoLand IDE:

1. Get GoLand IDE from <> and install
2. Get Scala SDK from
2. Configure Scala IDE

Scala with IntelliJ IDE(Edu Free):

1. Get IntelliJ from <> and install
2. Configure Intellij for Scala – either using sbt or Intellij Build Systems
2. Write Codes and Start coding



Labs. Of Session 2- On BigData Technology:

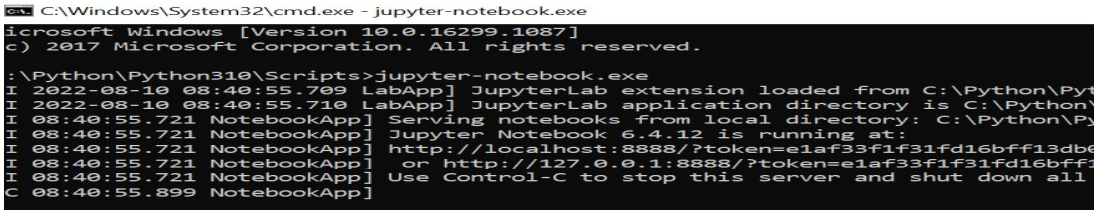
Python Programming:

Python Coding with PyCharm:

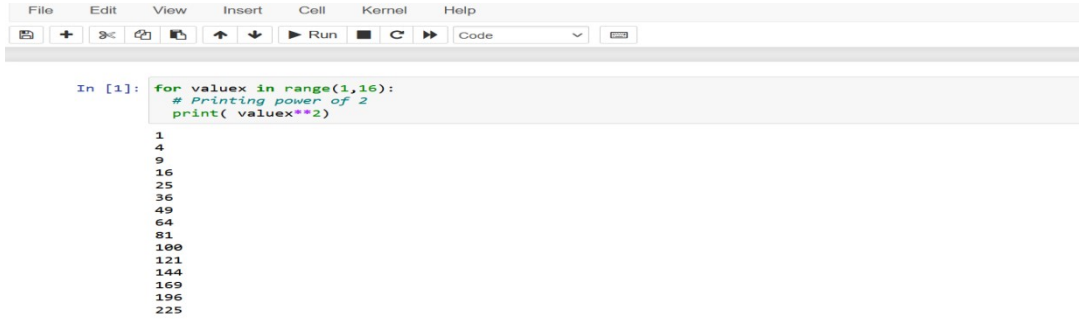
1. Get PyCharm IDE from <> and install
2. Get Python SDK from
2. Create Project and start coding

Python Coding with Jupyter Notebook:

1. Install Jupyter Notebook using details from
2. Start Jupyternotebook jupyter-notebook.exe



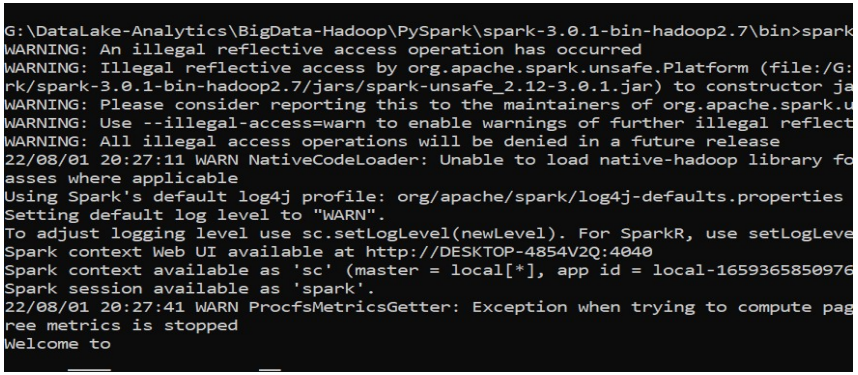
3. Access it on Local Browser and start coding

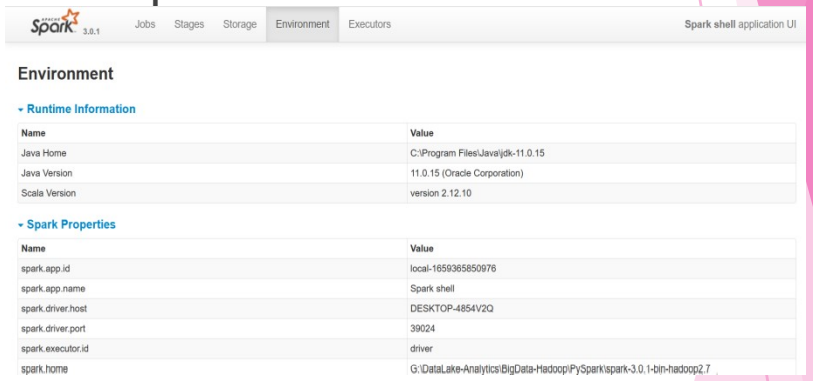


Labs. Of Session 3- On BigData Technology:

PySpark On Windows Installation -Steps:

1. Get PySpark supported version as Support Matrices
2. Create “DataLake-Analytics” folder and unzip tar aka zipped file
3. Create two System Environment Variables – HADOOP_HOME and SPARK_HOME *,X:\DataLake-Analytics\BigData-Hadoop\PySpark\spark-3.0.1-bin-hadoop2.7*
4. Run spark-shell.cmd to start Spark Instance
5. Start Firefox and check the installation with http://<IP Address>:4040





PySpark On Linux Installation -Steps:

1. Get PySpark supported version as Support Matrices
2. Create “DataLake-Analytics” folder and unzip tar aka zipped file
3. Update variable HADOOP_HOME and SPARK_HOME in bashrc or profile file
4. Start Firefox and check the installation with http://<IP Address>:4040

CDP Access And Administration(Basics):

CDP Access by Enterprise:

1. Considerations: Functions and Features, SLA, Data Security Assurance & Frameworks
2. Product Evaluation and purchasing CDP
3. Platform Access

Administration:

1. User and Access Management
2. Data Security Compliance Management
3. System Security, Cyber Insurance and Data Security
4. Patch Management and Maintenances

BigData Quotient Assessment-Short Hackathons,CTFs & Quiz:

1. Hackathon Problem Statement 1:

An Startup specializing in Embedded Machine Learning for IOMT(Internet Of Medical Things) need a Solution to determine Operational Statistics of IOT Devices in Real-Time. They decided to create Digital Model of Physical Sensors to detect Operational Statistics- Temperature, Motor Roations, Chemicals around, Humidity. If you were System Designer, what are top challenges you need to address?

2. Hackathon Problem Statement 2:

An Startup specializing in Embedded Machine Learning for IOMT(Internet Of Medical Things) need a Solution to determine Operational Statistics of IOT Devices in Real-Time. They decided to create Digital Model of Physical Sensors to detect Operational Statistics- Temperature, Motor Rotations, Chemicals around, Humidity. If you were System Designer, what are top challenges you need to address?

3. Rules of Submission:

1.

BigData Quotient Assessment-Short Hackathons,CTFs & Quiz:

1. **The Tool which is used in management of MS SQL Server Objects?**
    1. MS SQL Server Management Studio
    2. Azure Studio
    3. Android Studio
    4. MS Visual Studio

2. **Which Python Framework is used Web Programming?**
    1. Beautiful Soup
    2. PennyLane
    3. TensorFlow
    4. Eclipse

3. Which can be used in ELT while Datasets for Machine Learning?
    1. Pentaho/Talend
    2. PyTorch
    3. OpenVINO Toolkit
    4. Azure Computer Vision

4. Which Android IDE can be used to design Intelligent Mobile Apps?
    1. Android Studio
    2. PyCharm
    3. Netron
    4. Jupytier Notebook

5. What is Data Lake?
    1. Data Warehouse with BigData Technologies
    2. **It reservoir of Data and Services**
    3. **Collection Data from Lakes**
    4. **It is Biotechnology**

BigData Quotient Assessment-Short Hackathons,CTFs & Quiz:

1. Hackathon Problem Statement 1: (Submit by 3 Days):

An Startup specializing in Embedded Machine Learning for IOMT(Internet Of Medical Things) need a Solution to determine Operational Statistics of IOT Devices in Real-Time. They decided to create Digital Model of Physical Sensors to detect Operational Statistics- Temperature, Motor Roations, Chemicals around, Humidity. If you were System Designer, what are top challenges you need to address?

2. Hackathon Problem Statement 2:

Perform Internet-Based to understand “Secure By Design Principles” for BigData Analytics and write few recommendations for “Best Practices”

3. Hackathon Problem Statement 3:

- Download Data Science Hackathon details from <URL>
- Use Template to make/write your Solution to Problem/Ideas Statement
- Create Folder in this Google Drive and submit there/here.

References And Resources:

Certification Resources:

- Cloudera Free Courses: <url>
- Cloudera Developer Certification: <url>

Data Lake with Cloudera Technologies:

- Everyday Ethics for Artificial Intelligence, IBM https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf
- Ethics of Artificial Intelligence and Robotics, Stanford - https://plato.stanford.edu/entries/ethics-ai/#AIRobo

Smart :

- AI Ethics, MIT Press- https://mitpress.mit.edu/books/ai-ethics
- A Practical Guide to Building Ethical AI , Harvard Business Reviews ,https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai

BigData Magazines and Journals:

- AI Ethics, MIT Press- https://mitpress.mit.edu/books/ai-ethics
- A Practical Guide to Building Ethical AI , Harvard Business Reviews ,https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai

Installation and Configurations Details:

- Data Lake Configuration Master: <GitHub/Google Drive URL>
- Machine Learning Configuration Master: GitHub/Google Drive URL>

THANK YOU ALL

Spark and BigData Training Materials

- **Reference Guides:** https://drive.google.com/drive/folders/1bek8gOnDzqvNPdlUY7s4_Qt14epnK4Kz?usp=sharing
- **Presentation : https://drive.google.com/drive/folders/1VvxnbeFMqy_w2zXEbchIx0B_MhffwgPV?usp=sharing**
- **Practivce Codes : <>**