Big Data Summer Training

BigData Analytics- An Introduction

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

knowledgelab SM

Incubated @ STEP IIT (Kgp)

Presentation Topics:

- BigData Introduction
- History of BigData
- Advantages of BigData Solution
- Trends of BigData Analytics
- BigData Adoption Trends
- BigData Software Stacks
- BigData Analytics-Platforms
- BigData Programming Platforms
- Trends of BigData Analytics
- BigData MapReduce Demo- Word Count
- BigData in Telecommunication
- Configuring BigData Platform from Cloudera
- HAAS- Qubole Registration

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.
- 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)*

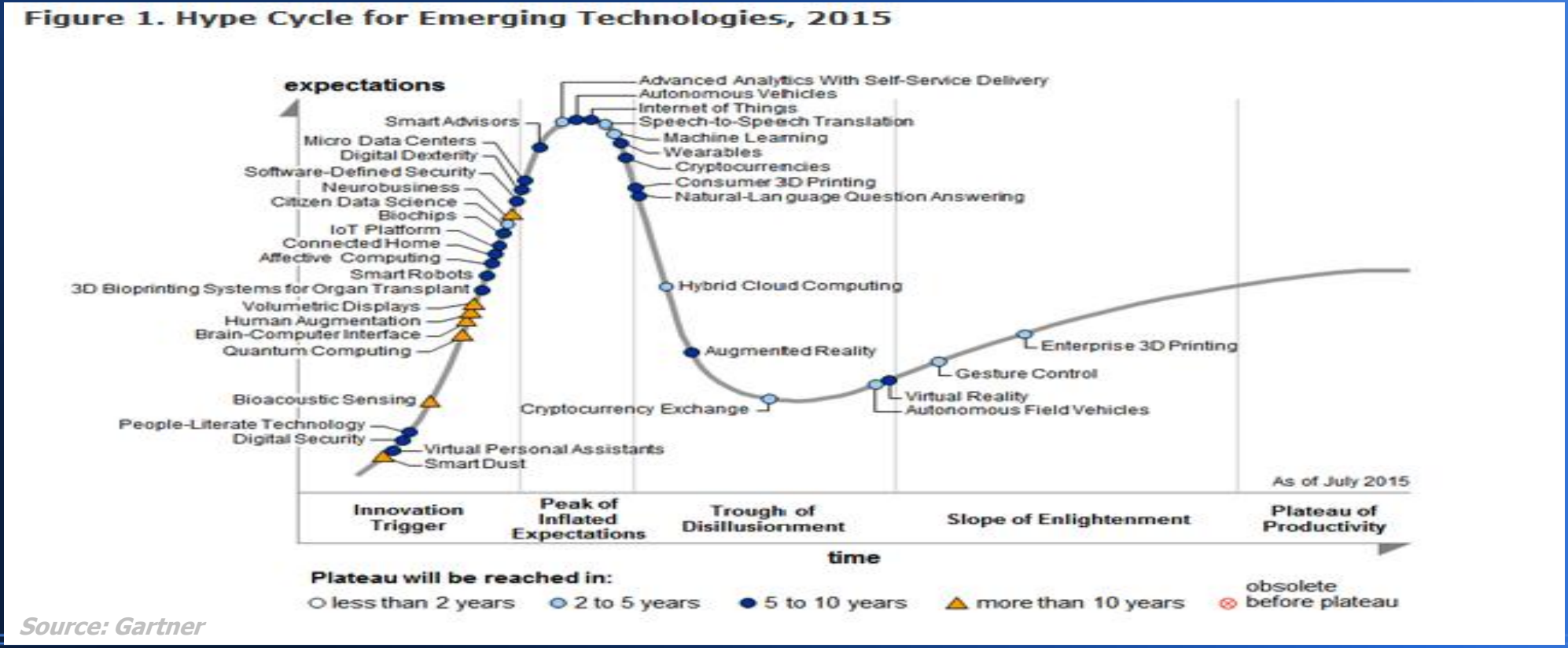
BigData Introduction:

- BigData is a large set of data and it follows 3Vs (Volume, Variety and Velocity) that traditional data processing application does not.
- "BigData is a collection of very large set of data which includes structured semi-structured and non- structured data and it they are processed by non-traditional and parallel-processing data processing system to produce meaning insights." - Amrit Chhetri
- The challenges of BigData are capture, citation and storage, search, query, visualization and analysis are handled by Hadoop.
- Apache Hadoop Stack or Apache Hadoop-based Platforms is the distributed Data Processing Platforms and it solves the issues of BigData.

Contd...

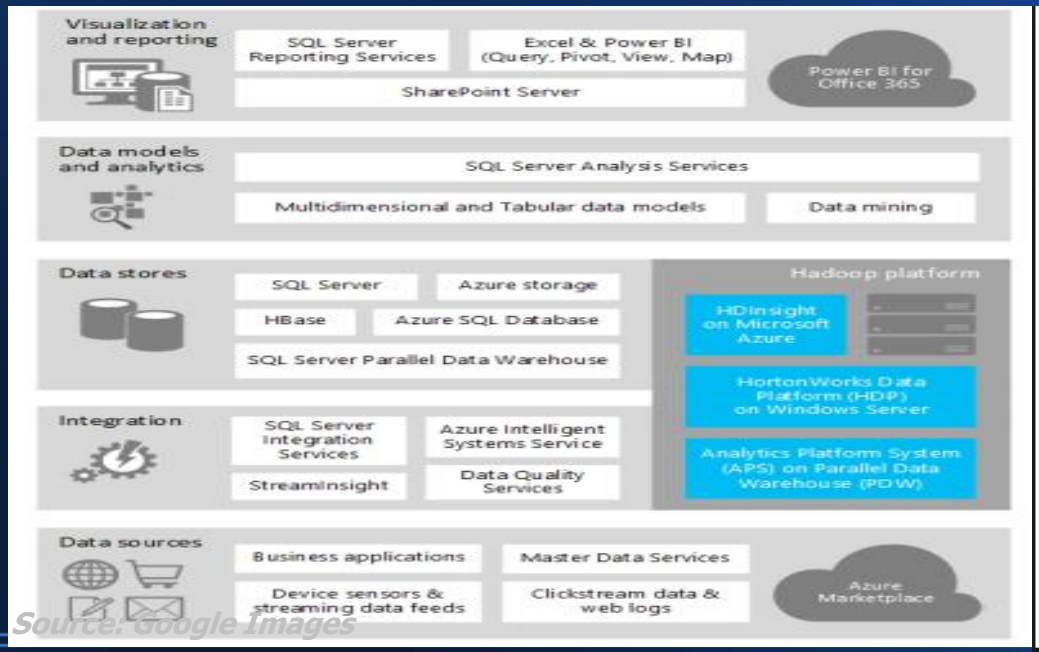
BigData Introduction:

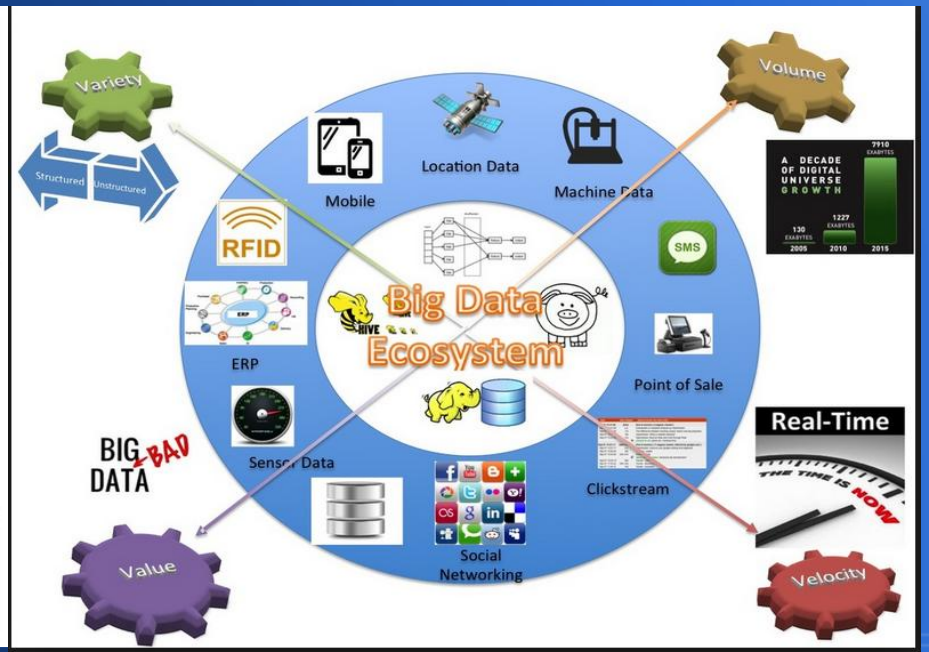
- BigData ranks itself on the top position in Gartner's Hype Cycle 2015 .
- BigData in Gartner's 2015 Hype Cycle 2015:



History of BigData:

- The term 'BigData' was introduced the first time in 2007 .
- Apache Hadoop is the first BigData Solution and the concept was incorporated in 2004
- Hadoop-As-A-Service(HAAS) is the latest trend of BigData Analytics and Qubole is an example.

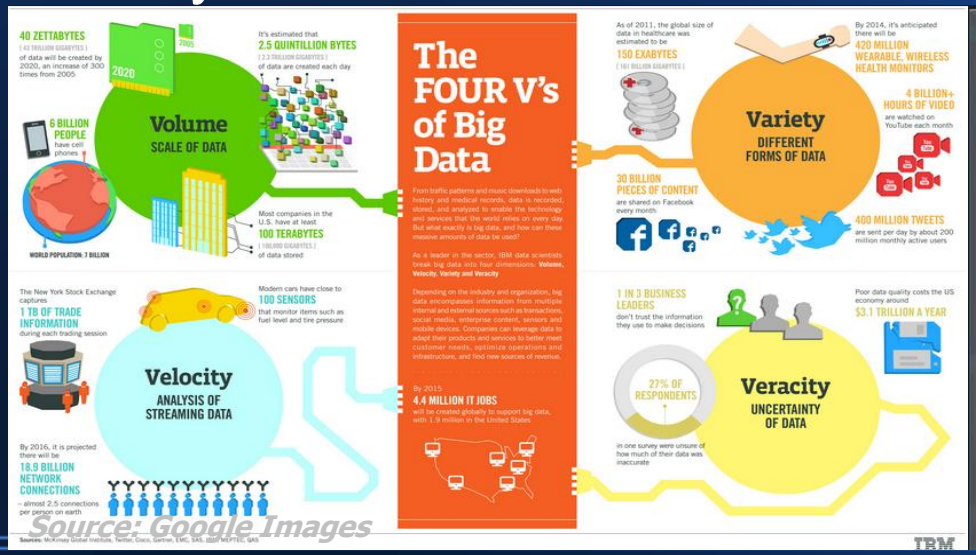




Source: Google Images

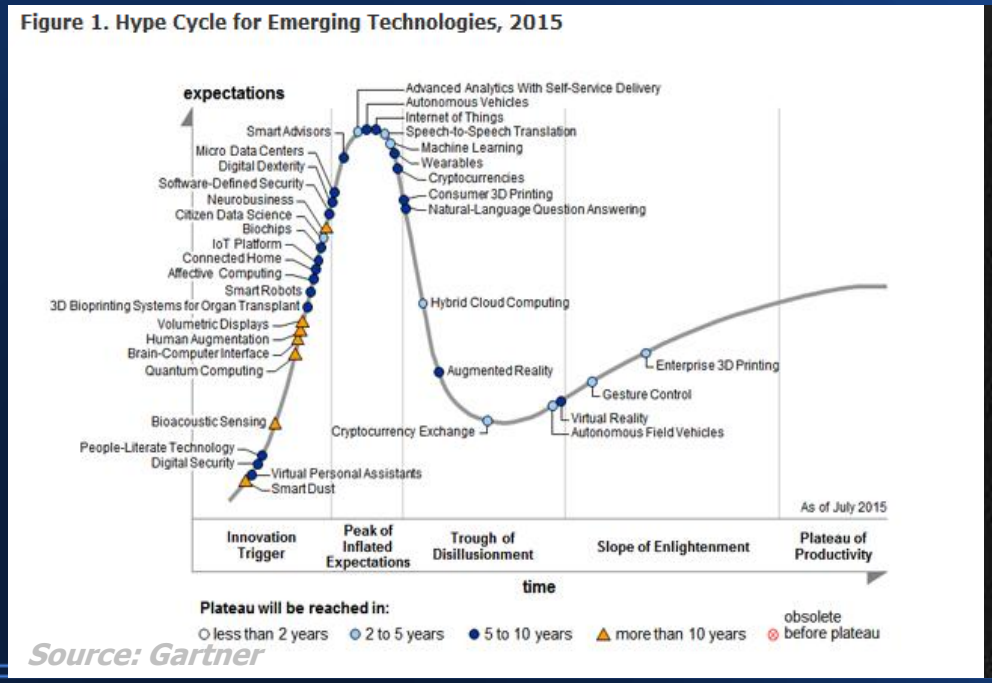
Advantages of BigData Solution:

- BigData is a distributed and Parallel Processing Platform
- BigData Analytics supports heterogeneous Data Sources
- Availability of Open Source Platforms for analyzing large volume of Data is another great advantage
- BigData is meant to address all 3/4 V- Volume, Variety, Velocity and Veracity



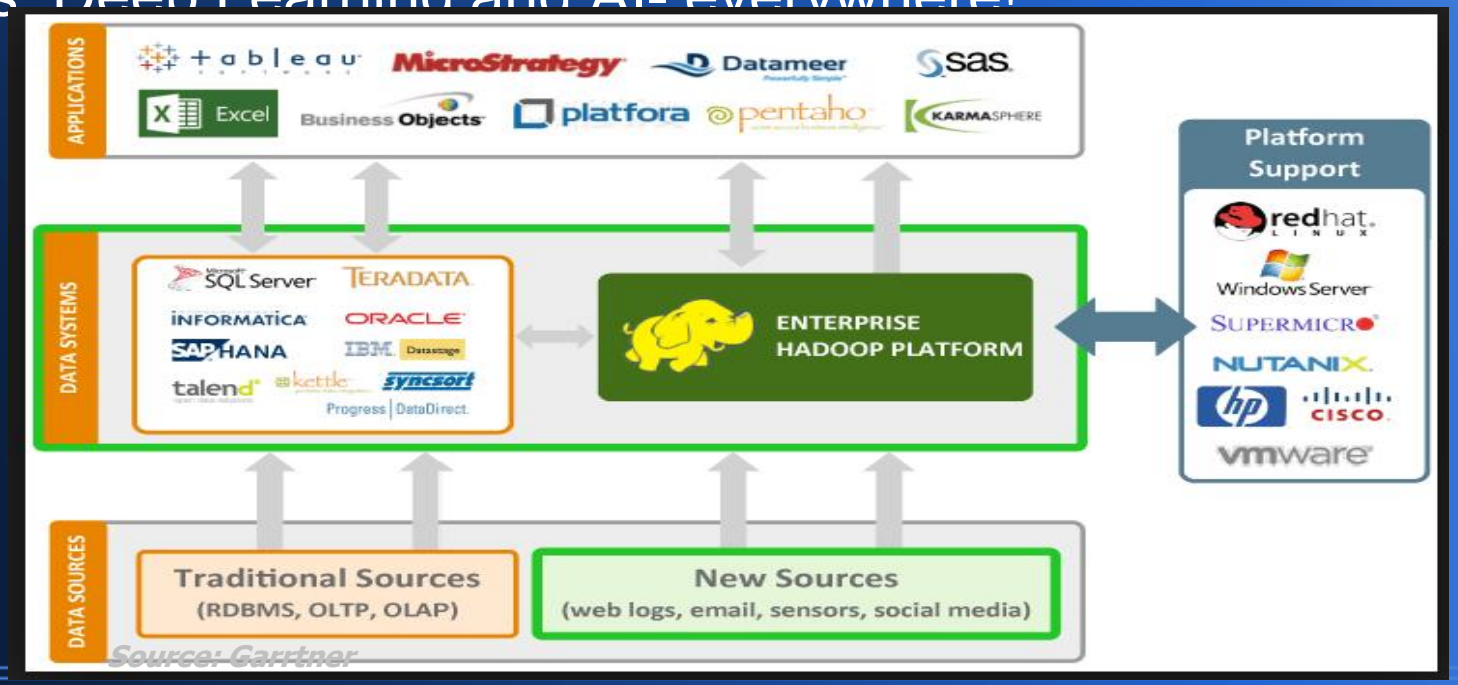
Trends of BigData Analytics:

- Self-Service BigData Analytics using BigData Analytics
- Mobile Analytics for accessing Analytics on Mobile
- Interactive Visuals or Reports to drill-into details
- Machine Learning and AI for Business Forecasting and Monitoring



BigData Adoption Trends:

- Customer Retention-Telecom, Banking, Finance, Healthcare and Infotainment and others .
- Service Quality Improvement-Telecom, Banking, Healthcare and Infotainment
- Predictive Analytics, Deep Learning and AI- everywhere!

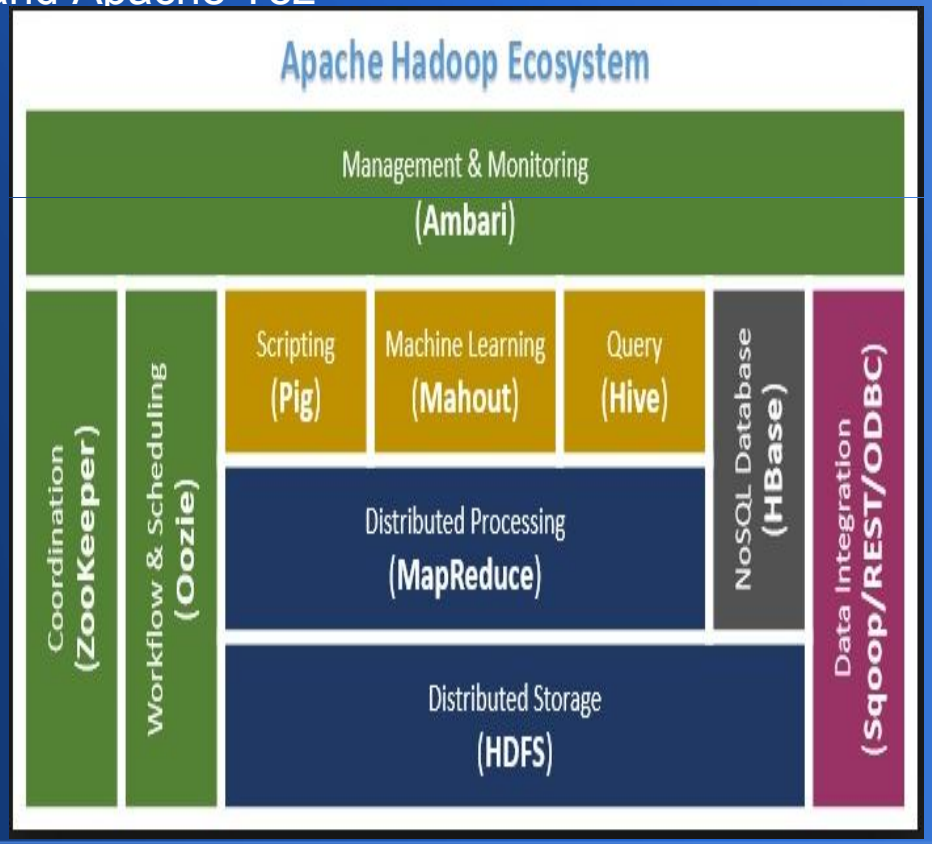


BigData Software Stacks:

- Apache Hadoop is the main Platforms of BigData Hadoop Stack
- Apache Hadoop is distributed or shipped by other BigData brands too including MapR, Cloudera, etc .
- The common distributions of Hadoop are:
  - Cloudera Hadoop
  - MapR Hadoop
  - Hortonworks Hadoop
  - MS Azure HDInsight
  - Oracle Hadoop
  - Syncfusion
  - Qubole and Informatica

BigData Analytics-Platforms:

- Apache Hadoop and HDFS are two core components of BigData Analytics
- BigData Hardtop Analytics comprises of
  - Distributed Processing Engine: Apache Hadoop and Apache Tez
  - Distributed File System : HDFS and RDD
  - Data Warehouse System : HBase
  - Scripting/Quering : Pig and Hive
  - Database System : NoSQL, Cassandra
  - Data Analysis Platforms : Hive, Spark, R/Octave/ MATLab and BI Tools (BIRT)
  - Monitoring : Apache Amber
  - Machine Learning AP : Mahout, Spark, MATLAB, Google TensorFlow



BigData Programming Platforms:

- Programming Language Compatible to BigData:C++, C#, Java, Python, Scala, PHP and Ruby
- BigData Scripting Languages: Pig Latin and HiveQL
- IDEs for Python :Eclipse(PyDev), IDLE, Anaconda and Geany
- API : MapReduce, Pig, Hive, HBase, Spark, MRLib, Mahout
- IDEs for MapReduce : Eclipse, Intellij
- Adoption of Machine API-Mahout, Spark, Octave/R/MATLAB

Trends of BigData Analytics:

- BigData In-Memory Analytics -Tez and Spark
- BigData on Mobile
- Adoption of Machine Learning - Mahout, Spark, Octave/R/MATLAB
- BigData for IOT Ecosystem- Sensors, IOT Protocols, BigData Platforms and Telecommunication Platforms
- Self-Service BI

MapReduce Demo-Word Count:

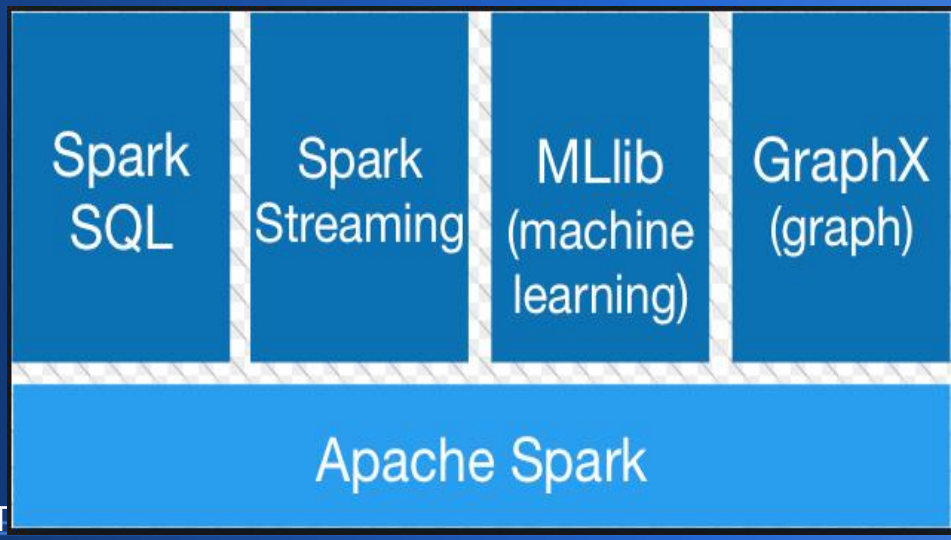
- Get Eclipse, install it on Windows or Linux System
- Install PyDev using ‘Install New Software’ option and install MRJob Python Module , #pip install MRJob
- Write Word Count Code :

```
from mrjob.job import MRJob
import re
re_Compile = re.compile(r"[\w']+")
class WordCount(MRJob):
    def mapper(self, _, line):
        for word in re_Compile.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()
```

```
"is"    "Occured 1 times"
"mid"   "Occured 1 times"
"ms"    "Occured 1 times"
"option"        "Occured 1 times"
"options"       "Occured 1 times"
"personal"      "Occured 1 times"
"service"       "Occured 1 times"
"services"      "Occured 3 times"
"size"  "Occured 1 times"
"small" "Occured 1 times"
"system"        "Occured 1 times"
"the"   "Occured 2 times"
"threat"        "Occured 1 times"
```

BigData in Telecommunication:

- Market Share Analysis and Competitive Analysis
- Customer Retention - Mobile, Phone, Data Card and Other Services
- Customer Behavior Analysis- Demographics, Usage Patterns, Payment/Recharge Patterns
- Location-Based Marketing-Service Analysis by Location, Regions and Seasons
- Real-Time Promotion and Offerings
- MIS Reporting -Call Drops,
- Service Quality Improvement - Network Traffics, Customer, Location, Trend and Demands
- Customer Service Improvement- Appropriate Plans, Billing
- Recommendation System Improvement
- Real-Time Performance Monitoring
- Smart Recommendation System(using Machine Learning and AI)
  - Customer Plan, Services
  - Customized Services,
  - Special Offer Improvement
- IOT (4G/5G) Communication Analysis - Analyzing IoT Networks over T
-



Cloudera BigData Platform:

- Install Windows 2012 Server(64x) or Windows 10 (64x) or Windows 2016(64x)
- Install Vmware Player 12 or higher
- Create folder ‘BigDataTrainining/Vminstances/Cloudera’ on D or E Drive
- Extract zipped file of Cloudera inside BigDataTrainining/Vminstances/Cloudera
- Open VDMX file using Vmware Player and import the necessary configuration
- Start Cloudera VM and open the Cloudera Home page on browser
- Open Hue (username/password: cloudera/cloudera)
- Create table: CREATE TABLE PRD(prd_id int, prd_cat int) ;
- Insert Date: INSERT INTO PRD values(23,23);
- Select Data: SELECT prd_id , prd_cad from PRD;
- WOW! Hive is working on Cloudera!!

HAAS-Qubole Registration:

- Hadoop on Cloud(Public or Hybrid) is called HAAS and it stands for Hadoop-As-A-Service .
- HASS is ready to use Platform model based on SAAS(Software-As-A-Service)
- Qubole is one of the HAAS
- Follow the steps below to run Hive on Qubole
- Register for Qubole.com or log-in into it using Gmail credentials
- Create table: CREATE TABLE PRD(prd_id int, prd_cat int) ;
- Insert Date: INSERT INTO PRD values(23,23);
- Select Data: SELECT prd_id , prd_cad from PRD;
- WOW! Hive is working on Cloudera!!

Day 1- Tasks:

- Registration on Qubole
- Installation of JDK 1.8 on Windows (32 bits or 64 bits)
- Installation of Eclipse
- Running Python on Eclipse, steps :
  - Run Eclipse
  - If Pydev ins not installed, click on Help->Install New Software-> click on ‘Add’ button and enter the URL (http://pydev.org/updates)
  - Restart eclipse
  - Configure Python Interpreter, click on Windows->Preferences-> Pydev-> Interpreter-> select Python Interpreter
  - Create Pydev Project and write Python Code and run it using ‘Run As Python’.

THANK YOU ALL