Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

Associate Professor

Learning Outcomes

After this lecture, you will be able to

- known tools used in big data,
- learn challenges in big data.

Hadoop

- The Apache Hadoop software library is a big data framework.
- It allows distributed processing of large data sets

HPCC

- Developed by LexisNexis Risk Solution.
- It delivers on a single platform, a single architecture and a single programming language for data processing.

Storm

1. Free big data open-source computation system.
2. Best big data tools which offers distributed real-time, fault-tolerant processing system

Qubole

Qubole data is Autonomous Big data management platform. It is a big data open-source tool which is self-managed, self-optimizing and allows the data team to focus on business outcomes.

Cassandra

The Apache Cassandra database is widely used today to provide an effective management of large amounts of data.

Statwing

Statwing is an easy-to-use statistical tool. It was built by and for big data analysts. Its modern interface chooses statistical tests automatically.

CouchDB

CouchDB stores data in JSON documents that can be accessed web or query using JavaScript. It offers distributed scaling with fault-tolerant storage. It allows accessing data by defining the Couch Replication Protocol.

pentaho

Pentaho provides big data tools to extract, prepare and blend data. It offers visualizations and analytics that change the way to run any business. This Big data tool allows turning big data into big insights.

Flink

Apache Flink is one of the best open-source data analytics tools for stream processing big data. It is distributed, high-performing, always-available, and accurate data streaming applications.

Cloudera

Cloudera is the fastest, easiest and highly secure modern big data platform. It allows anyone to get any data across any environment within single, scalable platform.

OpenRefine

OpenRefine is a powerful big data tool. It is a big data analytics software that helps to work with messy data, cleaning it and transforming it from one format into another. It also allows extending it with web services and external data.

RapidMiner

RapidMiner is one of the best open-source data analytics tools. It is used for data prep, machine learning, and model deployment. It offers a suite of products to build new data mining processes and setup predictive analysis.

DataCleaner

DataCleaner is a data quality analysis application and a solution platform. It has strong data profiling engine. It is extensible and thereby adds data cleansing, transformations, matching, and merging.

Kaggle

Kaggle is the world's largest big data community. It helps organizations and researchers to post their data & statistics. It is the best place to analyze data seamlessly.

Hive

Hive is an open-source big data software tool. It allows programmers analyze large data sets on Hadoop. It helps with querying and managing large datasets real fast.

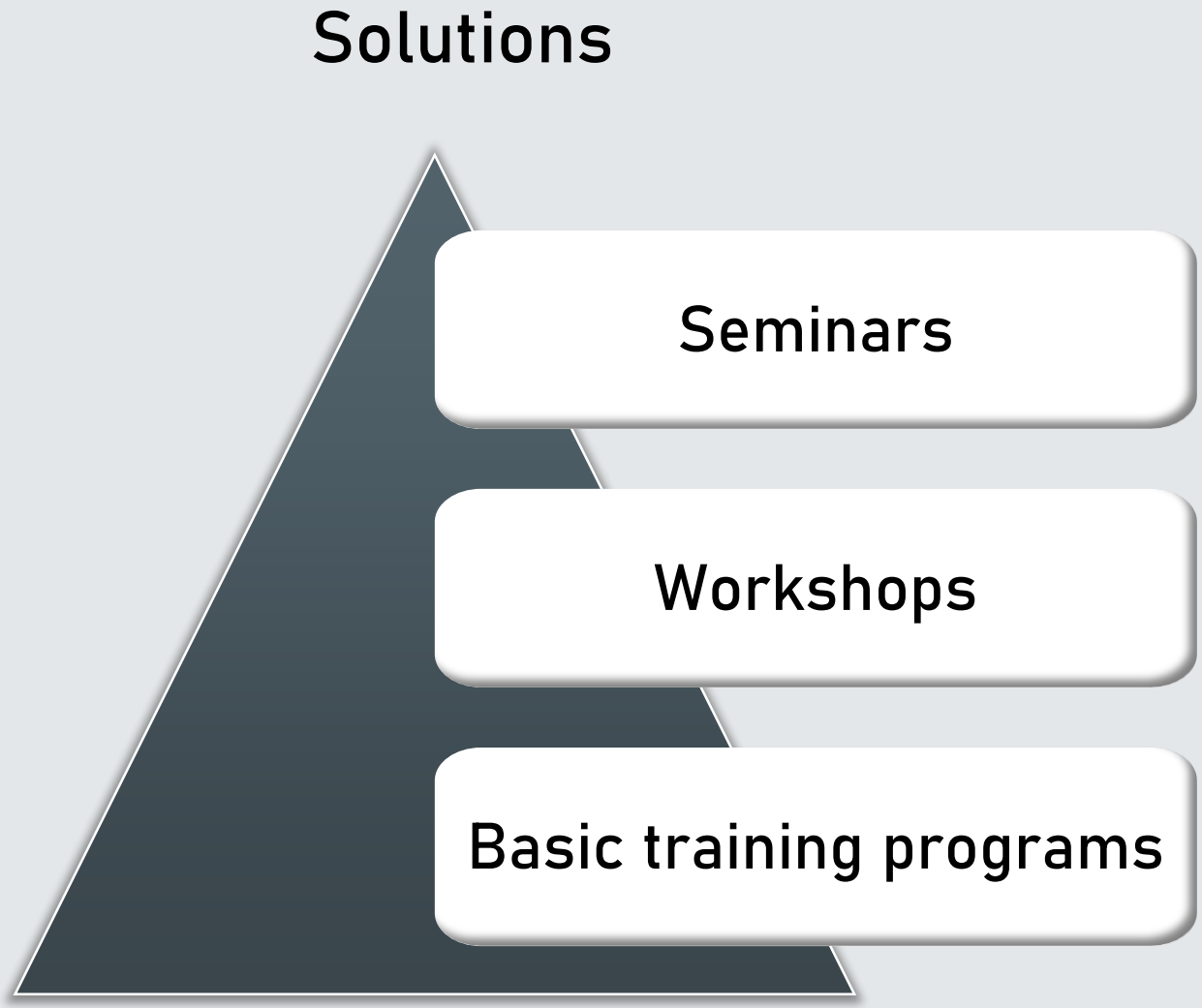
Challenges in Big Data

Challenges in Big Data

- Lack of proper understanding of Big Data
- Data growth issues
- Confusion while Big Data tool selection
- Lack of data professionals
- Securing data
- Integrating data from a variety of sources

Lack of proper understanding of Big Data

Lack of proper understanding of Big Data



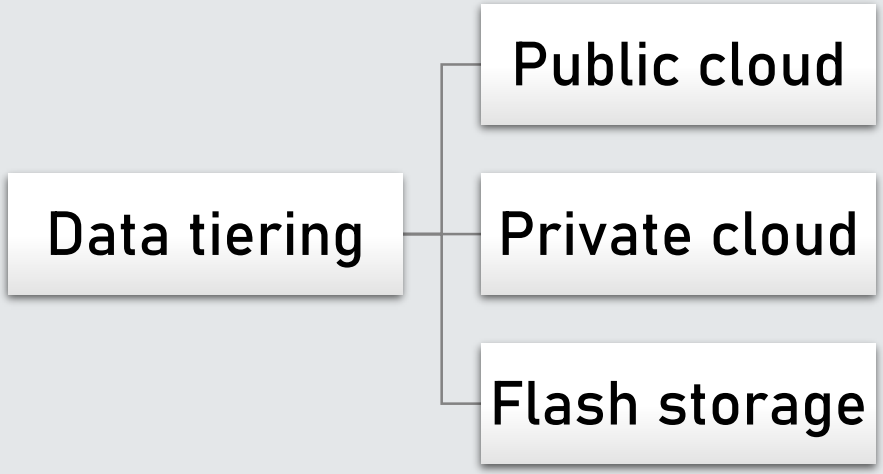
Data growth issues

Solution

compression,

tiering,

and deduplication.



Confusion while Big Data tool selection

Solution

Hire experienced professionals

Big Data consulting

Lack of data professionals

Solution

Skilled professionals

purchase of data analytics solutions

Securing Data

Solution

- Recruiting more cybersecurity professionals
- Data encryption
- Data segregation
- Identity and access control
- Implementation of endpoint security
- Real-time security monitoring
- Use Big Data security tools, such as IBM Guardian

Integrating data from a variety of sources

Solution

- Talend Data Integration
- Centerprise Data Integrator
- ArcESB
- IBM InfoSphere
- Xplenty
- Informatica PowerCenter
- CloverDX
- Microsoft SQL
- QlikView
- Oracle Data Service Integrator


That's all for now...