BDF

Big Data

- Characteristics
  - Volume
    - GB-Exabyte
  - Velocity
    - Resting
    - Streaming
  - Veriety
    - Data Sources
      - Structured
      - Semi-Structure
      - Unstructure

Tools and Technologies

- Hadoop
  - Distributed Storage and Distributed Processing Framework
  - Components
    - 4 Componets
      - Hadoop Common
        - Libs/DLL
      - HDFS
        - Distributed Storage
      - YARN
        - Resource Negotiator
      - MapReduce
        - Distributed Processing
  - Ecosystem Projects
    - Streaming Data Ingestion
      - Kafka
      - Flume
    - Data Ingestion from DMBS
      - Sqoop
    - Processing
      - Hive
      - Impala
      - Pig
      - Spark
    - Storage
      - HBase
      - MongoDB
    - Management
      - Amberi
      - Hue
    - Security
      - Sentry
    - Resource Manage
      - Mesos
  - Installation
    - Standalone
      - Single JVM
    - Pseudo Distribution
      - Multiple JVM
    - Full Distribution
- MongoDB : It's a Document DB
  - Components
    - MongoD (Server)
    - MongoShell (Client)
  - Other Componets
    - MongoConnector : It will connect MongoDB Server from Different Programming Languages
    - MongoDump
    - Mongoresotre
    - MongoImport
    - MongoExport
  - Characteristics
    - Sharding
    - Replication
  - Tools (GUI)
    - Robo Mongo
    - Compass
    - NoSQL Manager
    - Edda
    - PHPMoAdmin

Data Privacy and Ethics

- Depends on Domain
  - PCI DSS
    - Financial
  - HIPAA
    - Heath
  - GDPR
    - Europe

Big Data Project

- C-Suit Members
- Domain Expert
- Scary Data People (Data Analyst)
- IT Professional (Infrastructure)

Execute The Big Data Project

- Ask the Right Question to Data
- Get Support
- Identify the rources
- Develop a Solution
- Generate Insight
- Deploy

Big Data Sources

- Enterprise
  - SAP / MSFT/ ORCLE / ERP
- Data Warehouse
- Unstructured Data
- Meta Data
- Social Media
  - FB
    - Public Feed
    - keyword Insights
    - Graph API
  - Twitter
    - REST API
    - Stream API
- Pubic Data Source (Open Data)

Analyse Data : Tools : Data Mining

- Tools
  - KNIME
  - WEKA
  - R
  - Python
- Concept
  - To identify the Pattern inside Dataset
- Dataset
  - Supervised : When we have Label in our Dataset
    - Classification : When Label is Categorical
      - Algorithm : Logistic Regression/ Decision Tree / Decision Forest/ Neural Net
    - Regression: When Label is Numerical
      - Algorithm : Linear Regression / Polynomial Regression / Decision Tree / KNN
    - Association: Defining the relationship between Label and Features
  - Unsupervised : When we don't have any Label into our Dataset
    - Algorithm: K Means / K Means ++