- It is a general purpose
- It provides high level APIs in java, Scala,
- It also ~~has~~ support high level tools - SparkSQL for SQL & structured data pro
  MLlib for ML, GraphX for graph process
- Faster than MapReduce's batch processing engine.
- Spark is a cluster-computing framework, so its functionality is supported by MapReduce.
- It uses HDFS & Resilient Distributed Datasets (RDDs) & MapReduc
  uses persistent storage.
- Spark provides an interactive mode while Mapreduce does not have this feature.
- Spark offers greater speed, agility & ease of use while MapReduce provides low cost of operation.

Chapter 10: Big~~Data~~ IoT Data Science

Data Science: It is the study of data in a scientific manner, which comprises integrating several disciplines

Data Science Process:
(all other written)

Data exploration, modeling and evaluation.

- EDA is performed to gain a basic intuition & understanding of the data.
- Data description - describe the data and understand its various characteristics
- Sampling the data - useful for quickly seeing the data samples. Possible to extract a predefined percentage of random sample data from dataset.
- Data querying - further explored by making specific queries. It enables to make selections of data based on some conditions.
- Data Reduction - used when the dataset has high dimensionality. Transformations like feature extraction, PCA, LDA multidimensional scaling

Feature Selection - goal of this approach is to select those features that contributes maximum to estimator's accuracy.

- Feature selection is a preprocessing technique, which is useful for building robust predictive models.
- main Contribution of this are:
  - Dimensionality reduction
  - Speeding up the learning process
  - Reduce the model complexity
  - Models with high accuracy.

Approaches for feature selection:

Filter-based methods (based on predefined performance (metric) Ex: information gain, gain ratio, etc.)

Wrapper-based methods (they consider actual induction/modeling algorithm. Ex: Naive Bayes/ SVM)

Embedded methods (where feature selection is embedded in the model construction process itself) Ex: CART, Random forests,

Model Deployment:

- models are deployed into production environment.
- The ML models & other related tools and computational requirements are all integrated in a platform called DataOps.
- It provides automation, data access, integration modules & model deployment & management functionality.

Reporting & Visualization:

- Reporting is of 3 types - static/canned reports, dashboards, alerts.

Canned Reports: can be generated by analysis tool itself, & extracted by users of tools by themselves & send to other end users base on requirements

Dashboards: can have a set of information shown to specialized group of people. different views & each view show diff. persepective.

alerts: Real-time information is usually reported in form of this

Concept of Data lake/swap:

- Data lake: consists of data that is in its raw and unprocessed form and data is gathered irrespective of its quality.

Ex: Hadoop, NoSQL are more relevant

main characteristics:

- Retain all data to ensure that in some future time.
- Support for various data types/formats, data in form of web logs, images, videos, sensor data, social network data, etc.
- support various kinds of users - those need structured data, those seek raw data & combine with other sources, those perform more in depth analysis by integrating various types of data
- Adaptability to changing conditions that require different way.

Feature	Data Warehouse	Data Lake
Data storage	• Pre-processed data is stored for predefined uses • Tables less storage, costly storage infrastructure.	• All the data is stored, no specific use is preconceived. Needs a lot of storage, at low cost
Data model	Structured data with well-defined data models & methodology.	Data is stored as it comes in raw form, unstructured, non-traditional data, etc.
Security	are more mature & sophisticated	are still evolving.

Relation b/w IoT & Big Data:

- This data has characteristics of big data in terms of:
  volume/scale, velocity, variety, heterogeneity.

Big data Analytics in IoT:

- Big data analytics provides a means for analyzing & visualizing data from IoT sensors, actuators, devices & other connected components of IoT system.
- IoT data analytics are useful for:
  - Automating many decision-making processes
  - Increasing the efficiency with which processes can be executed.
  - Condition-based monitoring & predictive maintenance of equipment.
  - Service efficiency that encompasses remote management
  - Reducing overall operational expenditure & increasing revenue.
- The analytics can be in form of:
  Descriptive analysis, Diagnostic analysis, Predictive analysis, Prescriptive analysis.

Real-Time Analytics:

- The approaches for doing analytics on this type of data can be mainly divided into:

1) Event Processing-based Approaches:
   - These are based on methods such as ESP & CEP.
   - Goal is to capture interesting patterns from data coming from a single or multiple IoT devices & able to send alerts, warnings, etc.
   - This requires understanding several filters that operate on streaming data.

2) Data Stream mining Approaches:
   - In this, hidden knowledge is extracted from streaming data.
   - The key challenges are;
     memory bounded: streaming data is continuous & can arrive indefinitely, the system cannot store the entire stream.

single pass : each record is examined only once. Since it cannot be stored, possibility of rewinding & looking back at same data is not possible.

Real-Time response: Time taken for processing each record should be minimum, i.e, rapid processing is a technique

Concept drift : The patterns may not remain consistent over a period of time since new data is arrived. & possibility of underlying data distribution of data is different.

Stream data-mining applications :

- Industrial processes particularly in manufacturing - IIoT
- Real time security monitoring using IoT devices,
- Traffic monitoring, Real time disaster monitoring using IoT sensors.

Algorithms for stream data mining:

Stream frequent pattern analysis, landmark window, Sliding window, Damped window.

Stream clustering: The objective is to find groups of data items that are similar in some way & separate them from other dissimilar data items. These groups are homogeneous & have distinct characteristics. Classified into : Partioning methods, hierarchial methods, density-based methods, grid-based methods,

Stream classification: The objective is to assign data to distinct predefined categories called "classes". Acheived by developing a model and applying on new data to assign class labels. This process is basically divided into 2 steps : Testing, Training.

Offline Analytics:

- These are those that usually performed on highly scalable computing infrastructures such as cloud computing platforms.
- These are required for processing large volume of data, which is mainly stored in a repository on cloud.
- The main classes of algorithms for offline analytics :

Clustering: unsupervised classification technique that separates an unlabeled dataset into a number of distinct groups.

- In IoT, clustering is typically done for data coming from various sensors & there is requirement to capture some form.

Classification: Supervised learning approach in which a set of labeled data also called as training data is used to learn, which has capability to predict the class label of unlabeled data.

Training sample: It is a training dataset that can be used in predictive modeling task.

Regression: Given a set of input variables (X) and output variable (y) learning algorithm is used to learn the mapping function from input to output $Y = f(X)$, where Y is real or continuous value.

Correlation and Pattern analysis: This type of analytics is often exploratory in nature & mainly focused on identification of patterns in data. It gives idea about relationships b/w various variables. The correlation coefficient is a metric that quantifies/measures the strength of relationship b/w pairs of attributes

Big Data Analytics Platforms for IoT:

Microsoft Azure Stream Analytics, Aws IoT Analytics, IBM Watson Analytics, Cisco Data Analytics, Google Cloud IoT

ML & DL Tools:

- **Tensorflow:** open source software library for numerical computation using data flow graphs. Also includes TensorBoard, a data visualization toolkit
- **Theano:** a python library that allows defining, optimizing & evaluating mathematical expressions involving multidimensional arrays efficiently.
- **Keras:** High-level neural networks API, capable of running on top of Tensorflow, CNTK or Theano. Developed with a focus on enabling fast experimentation.
- **Scikit-learn:** is ML library written Python. provides various classification, regression & clustering algorithms. Tightly integrated with python numerical & scientific libraries Numpy & Scipy.