Data Science with big data

1- Introduction into Data science

The size of data started to increase day after day. The data started to come from different devices these days example of cell phone. The traditional data base became very week to handle the large size of structured and un-structured data. For this reason the big data concepts started to appear. Beside the huge size of data, we need different ways to deal with different cases and simplify the data to avoid the performance issues and to get the result with accepted period of time with accuracy. The concepts of data science started to appear. The Data science has to deal with different areas such as un-structure (video, text..), Database (structured data), Statistics, AI (Machine learning & Natural language processing). Data science involves using methods to analyze massive amounts of data and extract the knowledge it contains.

2- Data Science deals with different facts of data

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

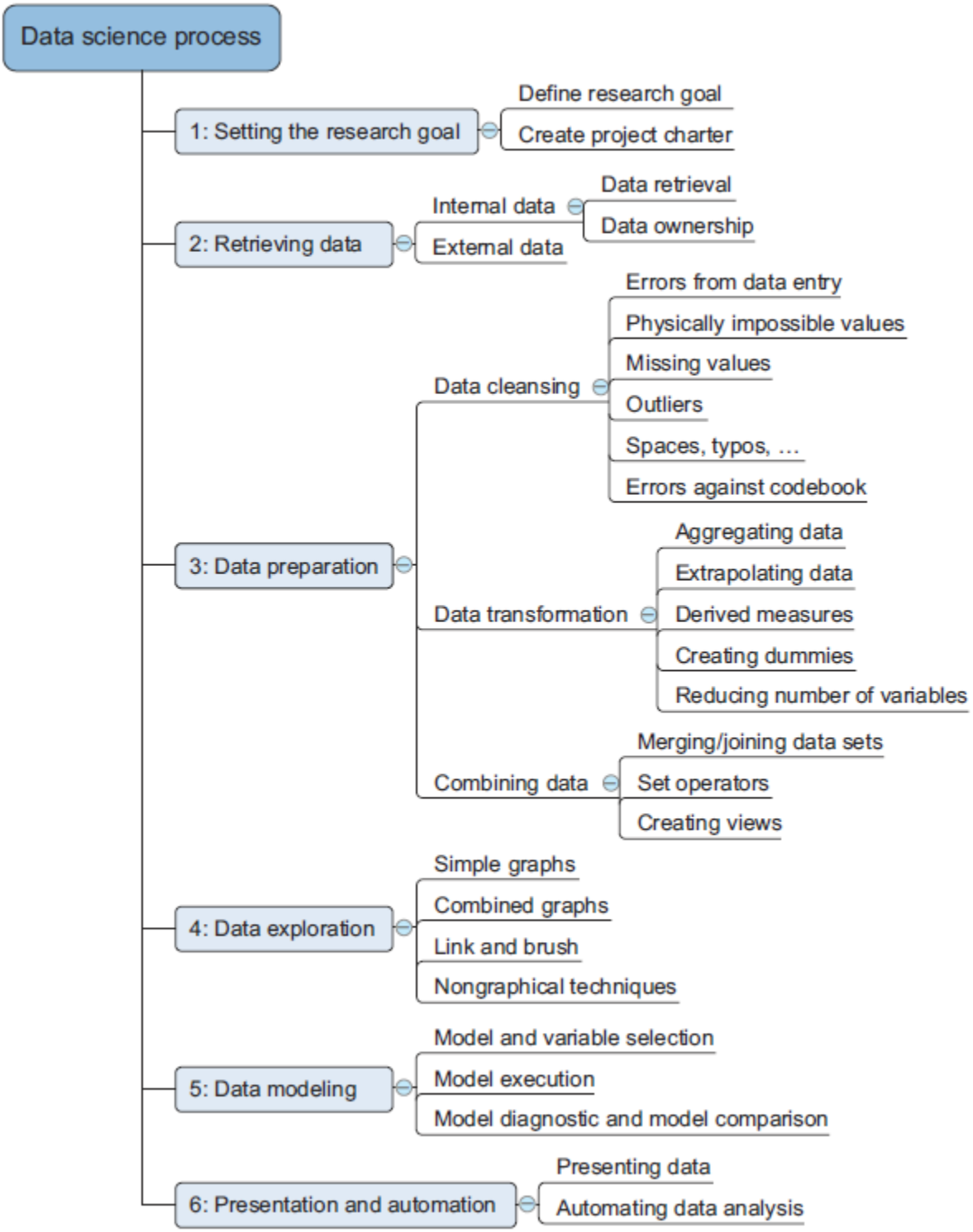
3- The goal of Data scientist

- Data scientist deals with different kinds of data to simplify the data and implement different algorithms to predict the results and produce these results in different format like use data visualization.
- Data scientist deals with
    - traditional databases
    - files
    - distributed file system such as Hadoop File distributed system (HDFS)
    - NO SQL data base like Hive, Cassandra, HBase, Mongo DB
    - MPP (massively parallel processing) database such as Pivotal Greenplum, HP Vertica and Tera data

4- The Data scientist tools

There are many tools for data science and many programming languages. The most usable one is Python. The Python language can be used with single machine data processing and distrusted system such as Hadoop framework system. Also we can use python in machine learning and data visualization.

5- Data Science steps:



A. Research goal: We specify the goal for our project or research. Which area has the project deals with? What kind of data has the project touch? Do we need single machine processing or distributed system of data process?
B. Retrieving data. Retrieving data is the first part of data analytics. We have to define the source of our data. Is this data external or internal? How to deal with the security of the external and internal data. What is the data format?
C. Data preparation. Data preparation is the second part of data analytics. In this part we have to deal with data. Our data will be in dataframe. We will have two kinds of data frame.
   - Pandas dataframe (Single machine data processing)
   - Spark dataframe (Hadoop File Distributed system). Under set of machines that works under one cluster.

   This part deals data filter, Data cleaning, Data aggregation and grouping, Data union, Data join, Data transformation, data calculation, data parsing and data extraction.

D. Data exploration: Data exploration is the third part of data analytics. In this part we have to understand more about the data to build views, use statistics with calculation like (mean, median and Standard deviation). Deals with probabilities. Also this deals with divide the data set into training and test data set.

E. Data modeling: This part is part of Artificial Intelligence. This part deals with machine learning , Natural Language Processing (NLP) and deep learning. In our study, we will concentrate on machine learning. The machine learning contains supervised learning (classification & regression) and unsupervised (clustering)

F. Presentation and automation

   The data visualization is the important part of machine learning. After the data calculation and prediction, we can convert the data output to charts and tables. We can build the data visualization reports manually or we can automate the process. Before visualize the data, we need to prepare the data for 2 dimensions.