Introduction to Big Data and Machine Learning

Dr. Mihail

November 12, 2019

Big Data - What is it?

Data

Data can be defined as information in raw or unorganized form [a]. Broadly defined, the “big” in Big Data” refers to datasets that cannot fit in the resources of a single machine, maybe even a supercomputer.

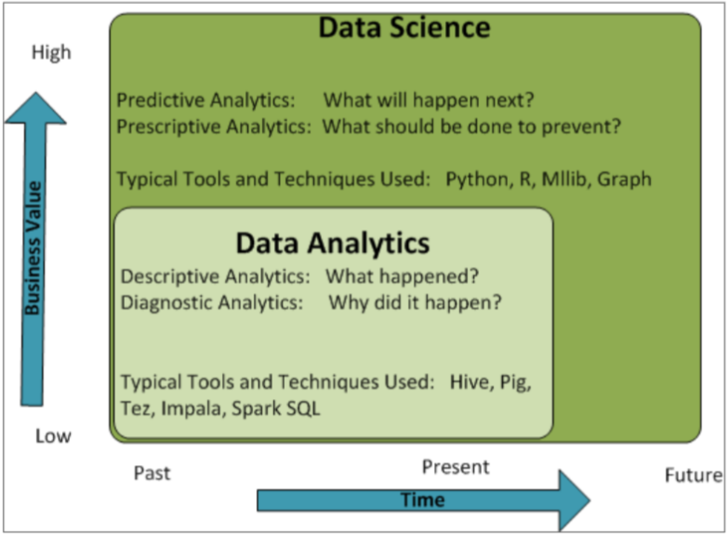
[a] http://www.businessdictionary.com/definition/data.html

Big Data Analytics

The process of analyzing Big Data to provide past, current and future statistics and any other insights useful to decision making. Big Data analytics typically can be categorized in:

- Data Analytics: deals with collection and interpretations, focus on past
- Data Science: deals with predictive and prescriptive analytics, focus on present and future

Analytics vs. Science



Data Science

Predictive Analytics: What will happen next?
Prescriptive Analytics: What should be done to prevent?

Typical Tools and Techniques Used: Python, R, Mllib, Graph

Data Analytics

Descriptive Analytics: What happened?
Diagnostic Analytics: Why did it happen?

Typical Tools and Techniques Used: Hive, Pig, Tez, Impala, Spark SQL

Conventional Data Analytics

Schema

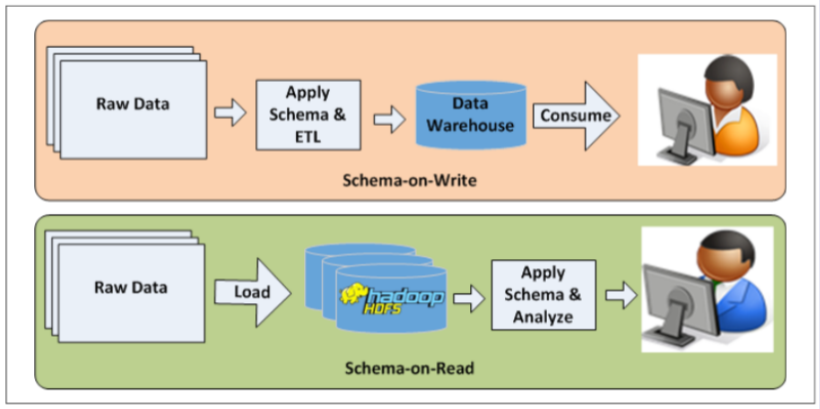
- Relational Database Management Systems (RDBMS)
- Data warehouses and Data marts for analytics used Schema-on-Write approach
- Traditional data warehouses designed for Extract, Transform and Load (ETL) tasks
- Predefined questions are then answered using SQL queries
- ETL pipelines developed to load data into the database in a consumable format

Big Data Analytics

New Era of Data

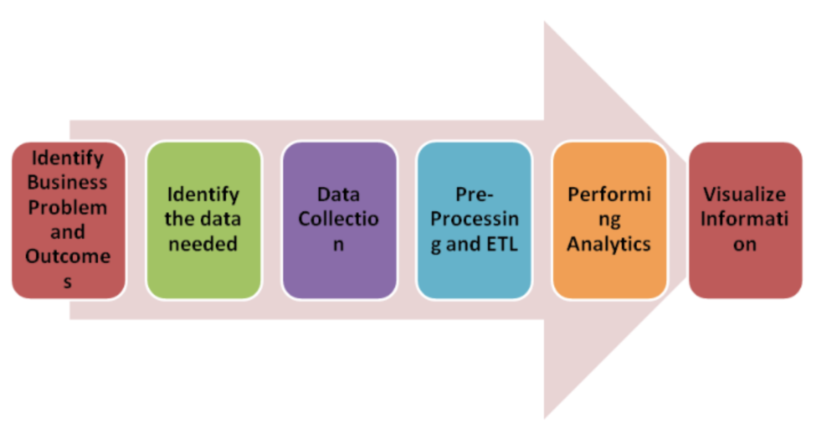
- No longer use schemas. Why?
  - Too much initial work (one to six months)
  - Any change requires developers and forces predefined boundaries
  - Processing structured and unstructured data is challenging in traditional RDBMs (e.g., large binary images or videos)
- Instead use Schema-on-Read (SOR)

Schema-on-Read vs. Schema-on-Write



Big Data Project Lifecycle

Lifecycle



Identify Business Problem and Outcomes

Identify the data needed

Data Collection

Pre-Processing and ETL

Performing Analytics

Visualize Information

Hadoop and Spark

Flexibility

- Large-scale data pre-processing
- Exploration of extremely large sets of data
- Accelerating data-driven innovation by providing schema-on-read approach
- Variety of tools and APIs for data exploration

Data Scientists vs. Software Engineers

Differences

- Software engineers develop general-purpose software for applications based on business requirements
- Data scientists don't develop application software, but they develop software to help them solve problems
- Typically, software engineers use Java, C++, and C-sharp programming languages
- Data scientists tend to focus more on scripting languages such as Python and R