Five P's of Data Science

Data Science is about extracting knowledge from data. At the WorDS Center (words.sdsc.edu), we define data science as a multidisciplinary craft that combines people, process, computational and Big Data platforms, application-specific purpose and programmability. Publications and provenance of the data products leading to these publications are also important for data science, but we start by defining 5 P's that take significant part in the data science activities.

- **Purpose:** The purpose refers to the challenge or set of challenges defined by your big data strategy. The purpose can be related to a scientific analysis with a hypothesis or a business metric that needs to be analyzed based often on Big Data.
- **People**: The data scientists are often seen as people who possess skills on a variety of topics including: science or business domain knowledge; analysis using statistics, machine learning and mathematical knowledge; data management, programming and computing. In practice, this is generally a group of researchers comprised of people with complementary skills.
- **Process**: Since there is a predefined team with a purpose, a great place for this team to start with is a process they could iterate on. We can simply say, People with Purpose will define a Process to collaborate and communicate around! The process of data science includes techniques for statistics, machine learning, programming, computing and data management. A process is conceptual in the beginning and defines the course set of steps and how everyone can contribute to it. Note that similar reusable processes can be applicable to many applications with different purposes when employed within different workflows. Data science workflows combine such steps in executable graphs. We believe that process-oriented thinking is a transformative way of conducting data science to connect people and techniques to applications. Execution of such a data science process requires access to many datasets, Big and small, bringing new opportunities and challenges to Data Science. There are many Data Science steps or tasks, such as Data Collection, Data Cleaning, Data Processing/Analysis, Result Visualization, resulting in a Data Science Workflow. Data Science Processes may need user interaction and other manual operations, or be fully automated.Challenges for the data science process include 1) how to easily integrate all needed tasks to build such a process; 2) how to find the best computing resources and efficiently schedule process executions to the resources based on process definition, parameter settings, and user preferences.
- **Platforms**: Based on the needs of an application-driven purpose and the amount of data and computing required to perform this application, different computing and data platforms can be used as a part of the data science process. This scalability should be made part of any data science solution architecture.
- **Programmability**: Capturing a scalable data science process requires aid from programming languages, e.g., R, and patterns, e.g., MapReduce. Tools that provide access to such programming techniques are key to making the data science process programmable on a variety of platforms.

To summarize, data science can be defined as a craft of using the five pieces identified above. Having a process between the more business driven P's people and purpose and the more technical driven P's platforms and programmability leads to a streamlined approach that starts and ends with a defined business value, team accountability and collaboration in mind.

Source: http://words.sdsc.edu/words-data-science/data-science