Team Data Science Process

Productivity practices for collaborative data science

What?

The Team Data Science Process(TDSP) is an agile, iterative data science methodology to improve collaboration and team learning. It is supported through a lifecycle definition, standard project structure, artifact templates, and tools for productive data science.

Why?

Industry statistics on the success rates of data initiatives are sobering. Companies are often unclear how to setup standard practices for their data teams to ensure collaboration and team productivity. TDSP is a distillation of the best practices and structures from both Microsoft as well as the industry, needed for successful implementation of data science initiatives to help companies fully realize the benefits of their analytics program.

Try TDSP today. Improve your team collaboration and productivity.
For more information:

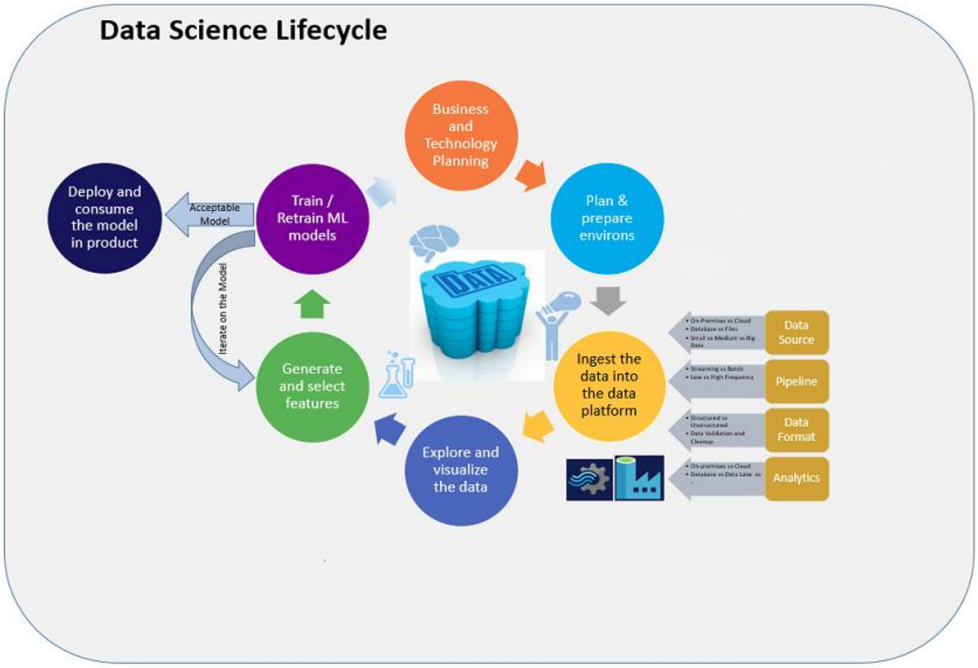
http://aka.ms/tdsp

Data science: A team-oriented process

Are you unsure how to go about building out a productive data science team? Do you feel lack of collaboration or a consistent process is hindering project success? Are your data scientists having to do many routine tasks manually? Do you face challenges in being able to capture and reuse knowledge from your data initiatives across the team? If the answer to any of the above is *yes*, the Team Data Science Process (TDSP) from Microsoft may help you fully realize the promise of data science for your business.

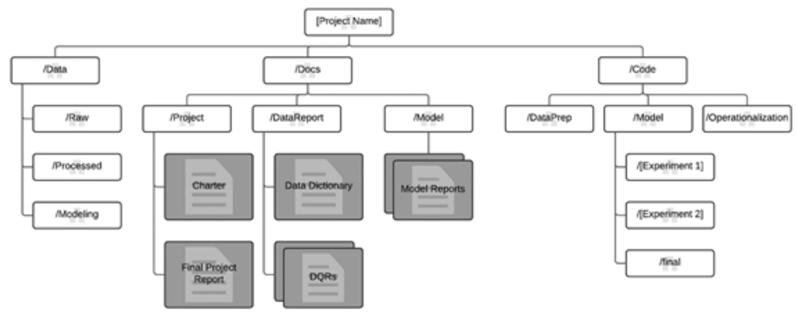
TDSP comprises of the following components (described on next page):

a) A **data science lifecycle** definition
b) A **standard structure for projects** including a well-defined directory hierarchy, list of output artifacts produced along with their document template structure. All code and document artifacts are stored in a versioned repository and project tasks are tracked in a project tracking system.
c) Management of **shared and distributed analytics infrastructure**
d) **Productivity Tools and utilities** for data scientists which also simplifies adherence to the process by automatically producing some of the project artifacts and providing scripts for many of the common tasks like creation and management of repositories, shared analytics resources.

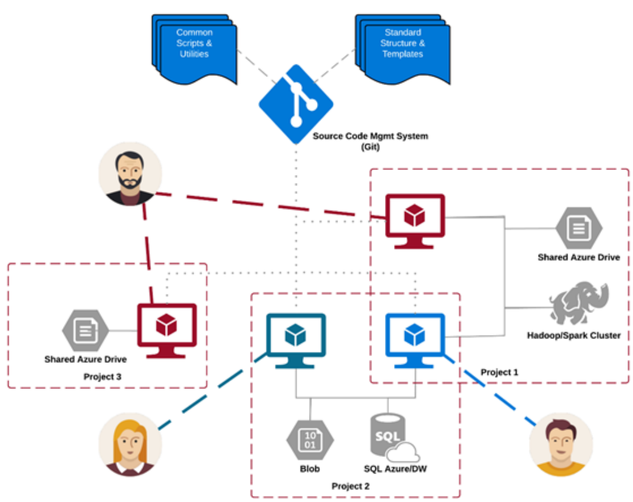


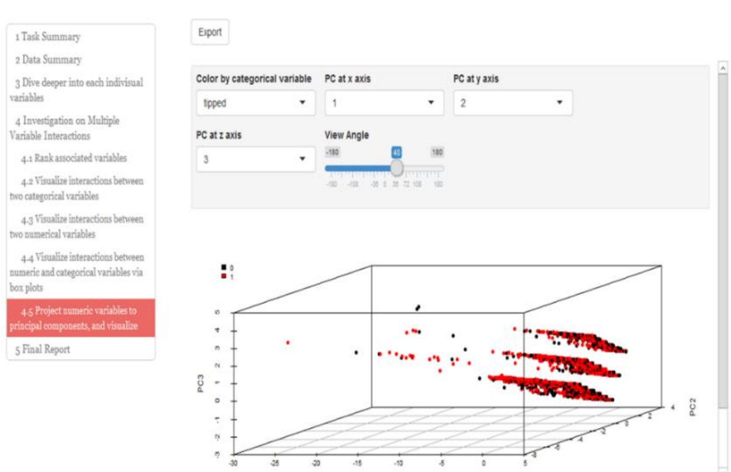
Data Science Lifecycle: The data science lifecycle defines a systematic set of steps starting with the planning step where the business problem / question is framed, to the development of predictive analytics models and their consumption of predictions from intelligent applications. Data Science is a highly iterative discovery process with emphasis on evaluating and validating each step of the way; refining the hypothesis and the predictive models to lead to a sound solution.

Standard Project Structure: Having all projects share the directory structure and project document with similar template makes it easy for the whole team to find information about past projects. It also ensures quality by ensuring all aspects of project as listed in the document template are addressed in a checklist like fashion. All artifacts, code are stored in a version control system (VCS) like Git, TFS, Subversion to allow the team to collaborate. Tracking tasks, features in an Agile project tracking system like Jira, Rally, Visual Studio Team Services and optionally linking them to a VCS allows tracking closer tracking of code to individual features and allows teams to get better at estimate effort. In the data science team at Microsoft, we use Visual Studio Team Services for its Git code repository support, Agile project tasks and sprints tracking.



Shared and distributed analytics infrastructure: TDSP provides recommendations for managing shared analytics and storage infrastructure like cloud file systems for storing datasets, databases, Big Data (Hadoop, Spark) clusters, machine learning services etc. on the cloud or On-premises. This is where raw and processed datasets are stored. This enables reproducible analysis. It also avoids duplication which can lead to inconsistencies and additional infrastructure costs. Scripts are provided to provision the shared resources, track them and allow each team member to connect to those resources securely. Within the Microsoft data science team, we use the Data Science Virtual Machine (http://aka.ms/dsvm) as our development environment on the cloud. This is useful in ensuring consistent configuration across the project team, validating experiments and also saving time for setting up the environment.





Productivity tools and utilities: Introducing processes in organization is challenging. By providing tools for aspects of the process and lifecycle, we not only get the benefit of productivity but also consistency in adherence and adoption of new processes. We will provide an initial set of tools and scripts to jump start adoption of TDSP within the team and automate some of the common tasks in the data science lifecycle like data exploration and baseline modeling. There is a mechanism for individuals to contribute shared tools and utilities into to their team's shared code repository so it can be leveraged by other projects within the team or the organization.