

Velocity

Data is becoming ubiquitous—and the need to leverage data is becoming increasingly acute in a variety of mission spaces. Developed by researchers at Pacific Northwest National Laboratory (PNNL), Velocity is an open source tool designed to enable scientists to rapidly explore large, complex data sets, develop statistical and machine-learning algorithms, and validate their approaches using a familiar desktop programming environment. Velocity automatically manages the complicated tasks of distributed storage and computation, empowering data scientists to do what they do best: tackle critical research and mission objectives by deriving insight from data.

CHALLENGE

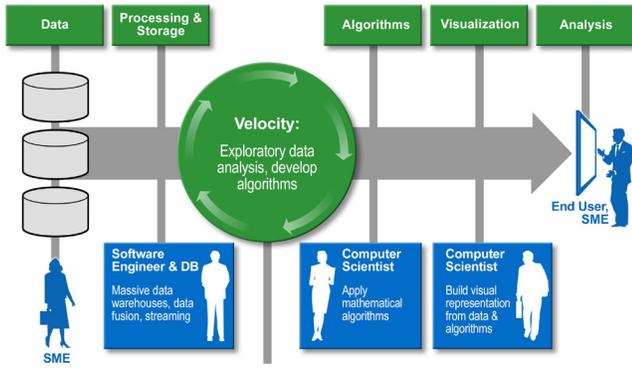
Extracting useful, actionable information from the data explosion is a grand challenge for industry, academia, and government. The impetus to leverage the deluge of data has led to notable advances in hardware, networks,

and software. But big data is much more than tools. It's about empowering people to meet their objectives—be that enhancing threat detection, improving operational efficiency, increasing profitability, or unraveling the mystery of cancer.

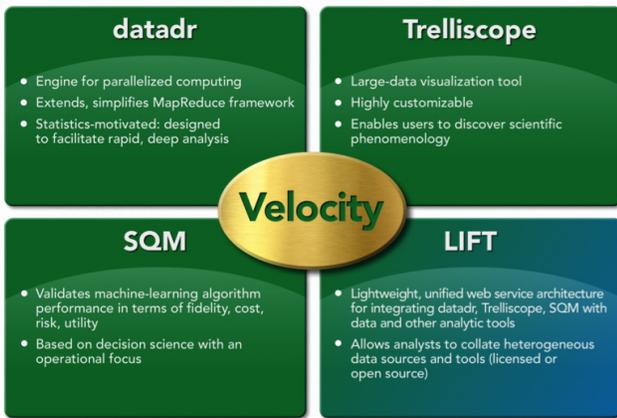
Many data challenges, large or small, are sufficiently complex to require a data scientist—a technical expert trained in statistics, applied mathematics, and computer science. Velocity is a software tool developed by data scientists for data scientists—to enable them to work more efficiently, develop better algorithms, and produce more accurate analyses.

SOLUTION

Velocity consists of four open source software tools developed at PNNL; datadr, Trelliscope, SQM, and LIFT. The combination of these tools makes it easy to rapidly reformat large datasets, explore and visualize data, develop statistical algorithms, and validate



Velocity can play a central role in developing end-to-end operational systems that provide actionable information to end users.



approaches. Velocity makes it easy to execute complex calculations with minimal lines of code and rapidly interact with data. It takes care of distributed storage and computing behind the scenes, allowing data scientists to focus on making sense of complex data to inform mission-critical decisions. Whereas existing commercial solutions often lack the extensibility and flexibility required to address the nuances of demanding and dynamic environments, Velocity addresses data scientists' needs, letting them perform their craft on a much larger scale.

Velocity is not constrained to a particular hardware platform, data architecture, or operating system. It can be used on a single desktop, a small inexpensive cluster, or a high-end data center. Velocity will run

on a distributed memory architecture, where large datasets are stored and manipulated in memory.

IMPACT

Velocity's flexibility makes it well-suited for performing deep analysis in virtually any domain. At PNNL, Velocity has been successfully used to discover patterns and insights in data in a variety of areas including the Electric Power Grid, High Energy Physics (Belle II), Proteomics, and Netflow (cyber data). Because the tools used in Velocity are domain agnostic, the potential application areas are virtually unlimited.

ABOUT PNNL

Interdisciplinary teams at Pacific Northwest National Laboratory address many of America's most pressing issues in energy, the environment and national security through advances in basic and applied science. PNNL employs 4,300 staff, has an annual budget of nearly \$1 billion, and has been managed for the U.S. Department of Energy by Ohio-based Battelle since the laboratory's inception in 1965.

For more information, contact:

Landon Sego, Ph.D.

Statistical Scientist
 landon.sego@pnnl.gov
 (509) 375-2753

Dave Thurman

Manager, Analytics Programs
 National Security Directorate
 dave.thurman@pnnl.gov
 (206) 528-3221