Big Data Studio for Big Data Literacy

Mark Bieraugel
California Polytechnic State University, San Luis Obispo

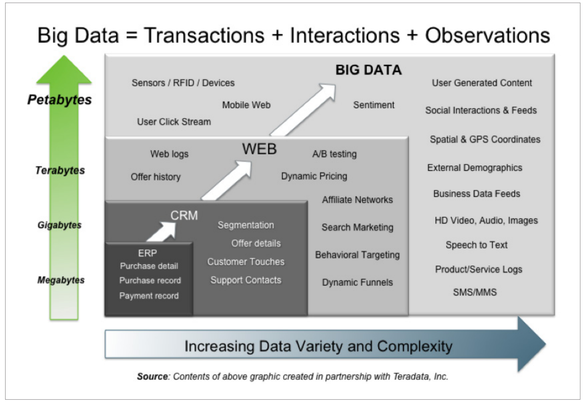
"Data is the new oil"…and big data is the richest of wells

Science graduates from across disciplines will be required to analyze and work with massive data sets. Large data sets, called "Big Data," cannot be handled by conventional database hardware and software, but need special software. Students need a place to work with big data on campus or at the library prior to graduating. The hands-on familiarity of big data will make them better prepared for the world of work.

2. What is big data?

Big data is characterized by three Vs: volume, velocity, and variety.

- **Volume**-huge sets of data
- **Velocity**-real time second by second data from sensors
- **Variety**-videos, emails, documents, photos



Source: *Contents of above graphic created in partnership with Teradata, Inc.*

3."Data is the new oil…" No, not exactly

Data is not rare or hard to find like oil. Data is gushing out of businesses, government agencies, and out of you. We are all data creators when we tweet, share photos on Facebook or Pinterest, or our interests online, especially with our smartphones.

4. Black gold, Texas Tea

"Data has the potential to make hidden relationships crystal clear, to be a common language between people who might never have spoken, to inspire collaboration, to offer metrics for decision making, and to turn seemingly unrelated ideas into powerful insights that can solve the most complex and intractable problems we face." *DataKind.org*

5. What the frack?

There have always been big data sets, but now the data is more accessible. Cheaper hardware and better software can store, retrieve, and analyze big data sets more easily and less expensively then ever before.

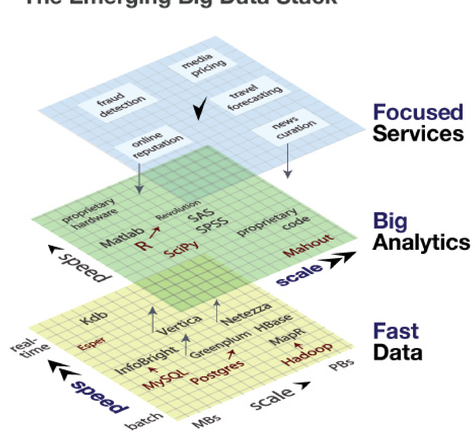
"Big Data is unmistakably revolutionary. For the first time in the technology world, we're thinking about how to collect more data and analyze it, instead of how to reduce data and archive what's left. We're no longer intimidated by data volumes; now we seek out extra data to help us gain even further insight into our businesses, our governments, and our society." *Andrew Burst*

6. Oil sands, tar sands, bituminous sands…

The "Big Data Studio" is a place to educate students in big data literacy by creating a controlled environment, a 'sandbox' where students can work with and analyze these data sets.

7. Elephants to the rescue!

Driving the growth of big data use is Hadoop, an open source software framework for storing and accessing large data sets.



Source: *Michael Driscoll*

8."Data is the new oil…" Well, yes!

"Oil requires extraction and refinement before it becomes the gasoline that fuels our vehicles. Likewise, data requires collection, mining and analysis before we can realize its true value for businesses, governments, and individuals alike." *Alex Yoder*

8. Building the Big Data Studio

Multiple stakeholders from numerous academic departments, cross disciplinary by nature.

Stakeholders:

- Computer Scientists – to build and tune the hardware and software stack
- Statisticians – to program in the R language for analytics and analyze data
- All Disciplines – to work with and analyze data sets specific to their field

Gusher!

Next Steps:

Stakeholder buy-in

Build the stack or work from the cloud?

Open source versus proprietary software and hardware?

Find funding

Acknowledgments
Thanks to Jeanine Scaramozzino, Katherine O'Clair, and Kristen Thorp. "New oil" quote by Andreas Weigand.

For further information
Please contact mbieraug@calpoly.edu