XDATA

Dr. Christopher White
Program Manager
Information Innovation Office

Proposers' Day
XDATA BAA

April 11, 2012

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

Background

"DoD is swimming in sensors and drowning in data"

Data

- Imperfect
- Distributed
- For diverse missions

XDATA - technology to leverage big-data at all stages from analysis to operations

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

XDATA Links and Related

DARPA release

- http://www.darpa.mil/NewsEvents/Releases/2012/03/29.aspx

DARPA BAA with solicitation

- https://www.fbo.gov/index?s=opportunity&mode=form&id=7a77846c73ffc5cb22f9295ffe6cdd55&tab=core&_cview=0

White House

- http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf

White House Fact Sheet

- http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_3_29_2012.pdf

NITRD

- http://www.nitrd.gov/Subcommittee/bigdata.aspx

Other coverage/Blogs

- http://www.wired.com/dangerroom/2012/03/big-data/
- http://www.nytimes.com/2012/03/29/technology/new-us-research-will-aim-at-flood-of-digital-data.html?_r=1
- http://www.cccblog.org/2012/03/29/obama-administration-unveils-200m-big-data-rd-initiative/
- http://www.networkworld.com/news/2012/032912-obama-big-data-257781.html?hpg1=bn
- http://news.idg.no/cw/art.cfm?id=B97173C2-D872-F31B-205A9D3F5A9EF9AA
- http://gigaom.com/cloud/obamas-big-data-plans-lots-of-cash-and-lots-of-open-data/

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

XDATA - Program Goal

Develop computational techniques and software tools for analyzing large volumes of data,

- semi-structured (e.g., tabular, relational, categorical, meta-data), and
- unstructured (e.g., text documents, message traffic).

Challenges

a) developing scalable algorithms for processing imperfect data in distributed data stores, and
b) creating effective human-computer interaction tools for facilitating rapidly customizable visual reasoning for diverse missions.

Open source software toolkits that enable flexible software development supporting users processing large volumes of data in timelines commensurate with mission workflows of targeted defense applications

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

Tech development path

- Principled implementations, scalable to data volume, compatible with distributed computer architectures, and suitable for field trials of timely information processing and analysis
- Minimize design-to-testing time
- Accommodate changing problem spaces and/or collaborative environments

- Open Source software toolkits
  - Flexible software development
  - User customization

- Development-in-process model

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

Technical Areas

- TA1: Scalable analytics and data processing technology
- TA2: Visual user interface technology
- TA3: Research software integration
- TA4: Evaluation

Propose:

- All or parts of any of the four Technical Areas (TA1-4).
- Proposers may submit multiple proposals.
- Multiple TAs (or portions of) not be combined in a single proposal.
- TA4 performer will not be selected to perform in any other TA.

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

TA1: Scalable analytics & data processing technology

- Leveraging the problem structure to create new algorithms to achieve optimal tradeoffs among time/space/stream complexity
- Propagation of uncertainty (i.e., error bars); performance guarantees for loss of precision due to approximations
- Measuring non-linear relationships
- Sampling and estimation techniques for distributed platforms, including compensating for missing/corrupted/incomplete information
- Methods for distributed dimensionality reduction, matrix factorization, matrix completion (not all data in one place)
- Streaming data feeds
- Optimal cloud configurations and resource allocation with asymmetric components (e.g., many standard machines, small number of large memory machines, machines with graphical processing units)

Other compelling ideas welcome!

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

TA2: Visual user interface technology

- Visualization of data for scientific discovery, activity patterns, and summaries
- Expressive visualization and/or query languages and processing that support domain-specific interaction, successive query refinement, multiple looks at data, faceted search, multidimensional queries, and collaborative/interactive search
- Principled design, including menus, query boxes, hover tips, invalid action notifications, layout logic, overview, zoom and filter, and details-on-demand
- Support study and characterization of users, including extraction of relations and history, usage, hover time, click rate, dwell, etc.
- Functions of timeliness, on-line vs. batch processing, meta-information, etc.
- Analytical workflows including data cleaning and intermediate processing
- Tools for rapid domain-specific end-user customization

Other compelling ideas welcome!

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

TA3: Research software integration

- Develop and implement an overarching integration framework for
  - Maintaining, supporting, and connecting the open source toolkits from TA1 and TA2
  - Implementing tests and evaluations developed by TA4

Must have:

- Prior experience in agile development and test environments
- Experience in cloud computing implementations and open platforms (e.g., Hadoop, Cloudbase/Accumulo, Hive, Hbase, Oracle, SQL, etc.), and trade studies
- Cleared personnel in team for discussions with users

Other details in BAA

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

TA4: Evaluation

- Develop and implement an overarching evaluation framework
  - Work with end-users to developing use cases
  - Work with TA1 and TA2 performers in developing performance metrics and testing of components
  - Work with TA3 performer to implement evaluation of integrated systems
  - Must address Human Use issues, as applicable

Must have:

- Knowledge of state-of-the-art benchmarks in cloud computing, scalable computing algorithms, and visual analytics
- Expertise in developing tests and benchmarks for new technology
- Cleared personnel in team for discussions with users

Other details in BAA

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

Government-furnished Resources

- Integration facility in Washington DC area
  - User interaction on site
  - TA3 and TA4 proposers should have staff on site here
  - TA1 and TA2 proposers – quarterly meetings and on site reps during summer (eight to ten weeks between June and September of each year)
- Cloud computing resources, open source data, proxy problems for evaluation

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

XDATA – technology to enable data-oriented decision making

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)

www.darpa.mil

Approved for Public Release, Distribution Unlimited (DISTAR Case 19136)