AGENCY FOR SCIENCE, TECHNOLOGY & RESEARCH
SCIENCE & ENGINEERING RESEARCH COUNCIL
(A*STAR - SERC)

DATA VALUE CHAIN AS A SERVICE

THEMATIC STRATEGIC RESEARCH PROGRAMME
(TSRP)

Science and Engineering Research Council (SERC)
Agency for Science, Technology & Research (A*STAR)
1 Fusionopolis Way, #18-10 Connexis North
Singapore 138632
Tel: (65) 6826-6111
Fax: (65) 6779-8061
1. INTRODUCTION

1.1. The Science and Engineering Research Council (SERC) of the Agency for Science, Technology and Research (A*STAR) announces a call for research proposals in the area of **Data Value Chain as a Service**.

1.2. Manufacturing has been and will continue to be a key growth engine for Singapore. For the nation to sustain its competitiveness, we need to look for and nurture other growth engines. Figure 1 shows Singapore’s GDP breakdown for 2008. Manufacturing’s contribution has shifted from 25% about a decade ago to 20% last year. Service-based activities have stepped up significantly to help shape our GDP. The service sector appears to be a good candidate to be another economic pillar. To make that happen, efforts to improve standards of living, boost productivity, and create jobs must focus increasingly on the service sector. This prompted the question on what new R&D strategies are needed to make this sector more innovative, efficient and competitive. This call for proposal is intended to seed the initial foray into long-term service-related R&D.

![Figure 1: Singapore's 2008 GDP breakdown](image)

1.3. A recent multi-agency brainstorming effort on Science and Technology revealed the potential of technology innovation in advancing the service-based sectors such as government, healthcare, education, retail, transportation, manufacturing and businesses in general. Technology-enabled services can improve the quality of existing services as well as create new ones in our journey toward a service-oriented economy.

1.4. In the service industry, competitive edge can be achieved when corporations possess a very good understanding of their businesses, their customers and the environment around their businesses. The insight enables services to be personalizable, predictive, proactive and adaptive, facilitate participatory engagement, and help users in making better decisions. To do so, companies invest significant resources and time to collect and analyze pertinent information. Advent in IT has made the data collection process easy. The challenge that the enterprises face now is
how to effectively handle and understand information and data that are generated by the services themselves.

1.5. Another key driver for addressing data-related issues is motivated by Singapore’s nationalistic vision for an integrated information infrastructure that enables important data owned by public agencies to be shared for research purposes. The recent announcement in [20] suggests opportunities in the future for multi-domain multi-agency data to be pulled together and analyzed, yielding value-adding insights that can help improve government and industry services. The data could span across different agencies such as transport, healthcare, telecommunications and finance. For instance, Microsoft projected that the shift to data-driven medicine in healthcare will have transformative impact [21]. Massive-scale data analytics will enable real-time tracking of disease and targeted responses to potential pandemics. While the government has vested interest in data sharing, enterprises will also stand to benefit from it. For instance, some businesses are already screening publicly available web data to improve their competitive edge.

1.6. In a world that is highly submerged in data, data analytics and higher form of data exploitation have received a lot of attention from enterprises, government and research communities. Enterprises are compelled to move up the knowledge chain in exploiting data. Standard reporting, simple query and drilldown are no longer sufficient. Analytics has demonstrated potential to help improve the quality of services in sectors such as retail [1], transportation [2], supply chain [3], healthcare [4,5], finance, education and even manufacturing. Data analytics promise abilities to extract the data and analyze it for richer understanding. Further up the value chain is predictive analytics that is useful for providing a sense of what is likely to happen in the near future. The trend is pushing us into a regime where higher level of decision-making is being exercised to explore, investigate, forecast and optimize (IDC #208699). The bottom-line for many businesses is time-to-decision. The quicker they can make sense of the data the better. In some applications, near real-time reactions are necessary.

1.7. The research community and enterprises can easily produce voluminous amount of data. Information on the Internet, coupled with data generated by an ever-Increasing number of sensors, pose challenges to an economy that is heavily knowledge-driven. The service industry needs to collect, store, move, integrate, process and analyze data efficiently and quickly for competitive advantage. During the analysis stage, visualization of the data could prove useful for navigating and understanding massive data. Beyond that, enterprises are increasingly interested in running simulation based on the derived data models. We refer to this entire flow as the DATA VALUE CHAIN (DVC).
1.8. The data value chain is proving to be more challenging for organizations to handle. Numerous factors contribute to the problem:

1.8.1. The data to be collected, managed and analyzed is getting too massive for some enterprises to handle economically and efficiently with its own resources (both hardware and software). Organizations are finding that their IT resources and support structure is not agile and cost-effective enough to tackle the rapidly-changing needs in the DVC;

1.8.2. In addition to the size, data are also increasingly being collected in a distributed fashion, making it difficult for small and medium enterprises to cope. While the distributed data could be moved to a central location, other factors such as confidentiality could prevent this from happening;

1.8.3. In the past, users who handled small-scale data did not encounter major problems in having to manually manage, integrate and process the data. When the data now comes in multiple sources, formats, structures and locations, coupled with an ever-increasing data size, users find that most of their time is spent on non-core activities such as data management and pre-processing. People cost is a line item that companies watch closely.

1.9. Handling the aforementioned new data value chain will require a new perspective as well as new/different enabling technologies and tools. The recent recession has compelled enterprises to think more seriously about the need to lower cost, increase efficiency and conserve cash. IT service outsourcing is one viable alternative. A recent manifestation of this trend is cloud computing – where enterprises can purchase/consume IT functions as services provisioned over the Internet [19]. Cloud also offers a plausible platform for the sharing of data from different domains and organizations. Hewlett Packard has a grander vision of everything-as-a-service. Regardless of the ultimate degree of realization of such a dream, there is no doubt that the world is moving rapidly toward a service-oriented economy.

1.10. This TSRP is established to understand and address the challenges associated with transforming the DVC as services in the cloud. To achieve this, new technologies are required, the traditional approach needs to be refactored to exploit the cloud efficiently, and issues preventing successful adoption of cloud services have to be investigated. There are numerous works on data related services (e.g. Amazon’s S3 on storage, Microsoft’s Azure on database, Hadoop on data processing). In themselves, there are still gaps that need to be addressed. Moreover,
there is little effort that focuses on the larger aspect of dealing with the data value chain. The various stages of the value chain are inter-related. When one stage of the DVC is moved to the cloud, the impact on the rest of the chain cannot be ignored. For instance, storing large amount of data in the cloud would typically render it infeasible to bring all the data back to the enterprise for analysis due to bandwidth latency and cost factors. The rest of the value chain may have to be executed in the cloud too. It is hence important that we understand the implication with respect to service quality, cost, reliability and security. This TSRP also serves as a platform to encourage the community to devote reasonable effort to this area so that we can be ready to tackle the DVC not for just today’s needs but future challenges too. In particular, this programme desires to see outcomes that can proliferate the availability of user-friendly, configurable, scalable and flexible data value chain services.

1.11. When we refer to data value chain as a service, we envisage the paradigm to possess some of the following broad characteristics:

- The services are technology-enabled services, as compared to those such as professional/consultancy services.
- The services can be provisioned (by any compliant providers) and consumed (by any qualified users) dynamically within the service ecosystem (e.g. a service cloud);
- The services are consumed in an on-demand and elastic manner;
- Independent services can be easily put together to deliver the functionalities required by the data value chain;
- The services demonstrate adaptability and scalability in meeting the needs of the enterprises.

2. OBJECTIVES AND SCOPE OF CALL FOR PROPOSAL

2.1. The aim of the call for proposals for full research projects is to provide directed funding to advance collaborative research activities in the domain of Data Value Chain as a Service (DVCaaS) so as to allow the research community to build industry-relevant capabilities in the fields of cloud-based data-oriented services.

2.2. Proposals should generally include activities such as proof of concept, testbeds, applications, as well as the participation of industry collaborators. Proposals should also demonstrate the building of multi-disciplinary/organizational research teams.

2.3. As this is a seeding initiative for research in services, SERC desires to confine the applicable domains to a smaller and manageable set. This TSRP specifically looks for use cases of the DVCaaS paradigm that will benefit services offered in the healthcare, transportation (including
logistics and supply chain) and government sectors. Proposals that fall into these domains would be given priorities. SERC is also open to consider proposals for applications in other sectors that can be illustrated to be impactful and compelling. It is desired that each proposal can validate its relevance through demonstration with at least 2 industry-and/or government-relevant applications.

2.4. As this TSRP is focused on data and relevance to real industry needs, it is imperative that proposals must be able to enumerate and secure representative and realistic data (or data value chain flow) in the proposed service domain(s).

2.5. In this programme, SERC will selectively fund projects that address critical technologies in the following 3 priority thrusts:

   i. Service Framework and Architecture for Data Value Chain
   ii. Cloud-based Data Architecture and Programming Models
   iii. Data Security and Privacy Protection in Shared Service Ecosystem

2.6. Breakthrough concepts and novel ideas are particularly encouraged for this programme. Proposals in these areas may work on novel service architecture, service-oriented middleware, algorithms and tools. It is expected that the collective projects under this TSRP will result in usable and extensible service platform, middleware and tools that can be deployed and used in a cloud environment to offer DVC services. The ensuing technologies should demonstrate the potential to produce revolutionary results that can be further developed into practical solutions.

3. PROGRAM DESCRIPTION

3.1. There are R&D challenges that could be addressed throughout the entire data value chain (e.g. data curation and integration, data warehousing, machine learning, analytics, simulation etc.). These challenges are exciting and meaningful in their own rights. However, the focus of this TSRP is not on solution/technology for each component/stage of the DVC per se. It is, however, expected that the technologies arising from this TSRP will allow for the provision of each DVC stage as services.

3.2. The TSRP programme encourages R&D activities that demonstrate a good understanding of the traditional DVC and investigate novel way of making them more efficient and better through a service delivery paradigm. Proposed projects will investigate techniques to integrate distributed data services throughout the data value chain. It is also expected that the traditional DVC approach might have to be refactored to achieve desirable outcomes [18] in the cloud. Proposals should investigate how each DVC function can be mapped efficiently onto the
cloud and how they should be integrated seamlessly to realize the data value chain specific to a suite of relevant applications.

3.3. For the outcome of this TSRP to be of practical use and to maintain a manageable focus for the programme, SERC has decided to focus on a set of enabling technologies that can be immediately useful in delivering and encouraging the adoption of DVC services. We envisage that this foundational set of technologies will pave the way for extension and enhancement in the long term. We assume that participating projects will actively engaged partners and vendors to incorporate established technologies and solutions to demonstrate the DVC.

3.4. As DVCaaS involves multiple players and technologies, establishing acceptable standards in the interfaces and integration is important. Proposals should be mindful of how the research outcome can contribute to the standards process.

3.5. Service Framework and Architecture for Data Value Chain

Proposals are called for to investigate and develop cloud-based, economy-driven, service-oriented architecture that efficiently handles the characteristics of the data value chain. The technologies pursued should address the following issues:

- The service architecture needs to be autonomic in the way it can self-configure to meet the DVC service level; the manner it can self-optimize to meet business objectives; how it can self-heal during service disruption; and how it can self-adapt to changes in the environment. Automation of the above processes is essential to handle a complex environment and the scale of the cloud. An autonomic service fabric is important as cloud resources and services do come and go and get disrupted. The service platform needs to be able to sense, make real-time decisions and react to unforeseen circumstances. The service architecture has to natively support such autonomic abilities, in particular, at the workflow or component level. As DVC services will be provisioned in a fee-based model, the autonomic framework has to allow economic factors to be considered and optimized for providers as well as users.

- How multi-step workflow and various forms of data management and processing can be supported via dynamic, adaptive and seamless service mash-up and orchestration. The DVC flow involves consumption of potentially distributed and complex service components. The workflow is likely to involve the use of different component services (potentially offered by different providers) from
various locales. The current cloud approach to service composition/orchestration is still somewhat static and hardwired. To achieve seamless mash-up, one other area that needs to be addressed is the lack of standards in the way service is being described and integrated. Within a cloud context, the integration challenge is further complicated by the fact that users expect service reliability and data protection that can be enforced and guaranteed consistently across all providers in the value chain.

In investigating the above challenges, this TSRP desires to achieve an implementation of the proposed service architecture that can be demonstrated to support the specification and execution of a generic distributed DVC workflow (capture, storage, integration, preprocessing, analysis, simulation). It is expected that most existing and third party solutions for each workflow step can be easily incorporated and instantiated using the new service framework. Large datasets present new technological challenges in terms of performance, service quality and cost. Parallelized data analytics software will emerge. The service architecture needs to consider challenges in the seamless integration of such software into the service paradigm.

The cloud community has also realized that a hybrid cloud is likely to be the solution that enterprises will adopt for various reasons such as data confidentiality, bandwidth constraints or legacy application/system etc. In this regards, proposed solutions should investigate foreseeable challenges in putting together an effective DVC within a hybrid cloud model.

The specific proposals are suggested as follows:

- Autonomic cloud-based service architecture and framework that can handle and process the entire data value chain. The proposal should include strategies for dynamic and adaptive DVC service composition, mash-up and orchestration as well as techniques for hybrid cloud integration. The service framework needs to be business-aware when fulfilling a DVC request. Proposals should also consider how existing DVC tools and algorithms should be refactored and optimized for cloud and massive data set.

3.6. Cloud-based Data Architecture and Programming Models

In a cloud environment where the data is likely to be distributed in different locales and systems, a key concern in terms of service quality and usability lies in the data architecture and the way data services are being programmed. For instance, MapReduce-like architecture offers a promising way for data to be distributed in a scalable, fault-tolerant way
amenable to data-intensive processing. This contrasts with the classical SQL-based share-nothing parallel database systems that have established following and numerous existing interfaces with other software and systems. Both approaches have their pros and cons. The ability to combine the advantages of both approaches may boost the adoption of cloud in providing data storage cum processing services. Within the 2 camps, there are opportunities for improving the state-of-the-art (e.g. better scheduling and load distribution for MapReduce, latency-friendly cross-domain implementation etc.). Cloud-based data architecture also needs to consider the re-factoring that is needed for data analytics. Numerous analytics algorithms require linear algebra computations. How would this be done differently for data that is distributed in the cloud? The same challenge applies to data visualization.

To facilitate sustainable adoption, a seamless programming model is desired. Non-IT trained personnel will need an easy-to-use framework where they can effortlessly put together the DVC. There is a strong desire for a single environment that can scale from the desktop to the cluster to cloud without having to use different programming languages, tools or interfaces.

In order for data to be useful, they need to be described and annotated in a way that facilitates easy retrieval and processing. In handling large-scale data, automation is indispensable. Standards-based way to describe and index data (especially unstructured ones) is needed to enable automation. Can data be self-descriptive enough that data processing and analysis tools know automatically how to act on them? Are the ontology and meta-data rich enough to facilitate automatic service mash-up? In handling ontologies, a specific challenge is how to align ontologies from different domains to facilitate service composition and mash-ups. Ontology facilitates reasoning. However, existing formalized ontology-based reasoning (e.g. using Description Logic) is inefficient and has limited expressivity to support real-world reasoning. Moreover, the contextually sensitive nature of data should not be neglected if we desire future services to be context-aware.

The specific proposals are suggested as follows:

- Efficient cloud-based data architecture that exploits the benefits of MapReduce-like framework and parallel database systems
- User-friendly and efficient programming models and user-service interaction environment for distributed DVC services in the cloud
- Data-centric ontologies to facilitate service mash-up and automation of DVC workflow
3.7. Data Security and Privacy Protection in Shared Service Ecosystem

Using services provided by third parties in the cloud will invariably involve the release of enterprise-owned data outside the confines (and control) of the corporation. Enterprises would undoubtedly be concerned about the protection of their data. Many potential cloud users view security as a major stumbling block. Technologies that can substantially mitigate security-related risk and establish service trustworthiness would help tremendously to accelerate the adoption of DVCaaS.

One form of data protection is encryption. Data at rest and those in transit over a communications channel can be protected this way. However, processing of data (while in memory or in virtual machine) in its encrypted form is probably a stronger means of protection. Moreover, if data is to be searched, any proposed encryption-based solution must also allow searching of encrypted data without compromising it. It is known that sandboxing offered by virtualization technologies is susceptible to attacks. These requirements pose tremendous opportunities for research.

In a cloud setting, multiple parties may be involved in providing the DVC services. Services (and associated resources) are also instantiated/consumed on-demand. They disappear when the service objective is met. The dynamics are made more complex when resources can also be reallocated or swapped during the service lifespan. This poses a problem in establishing and validating audit trails. One possible way to establish trustworthiness in a complex multi-tenant environment such as the cloud is data provenance. It has been attempted at the system level but suffers from various limitations. More efficient strategies (e.g. application-level provenance, automated reasoning, non-mutability) to capture the audit trail are needed to further enhance data security and quality.

Another benefit of DVC is the ability to share data among trusted parties [6]. Even in such a scenario, an enterprise may still desire to protect the privacy of the people with whom the data is associated. Privacy-preserving technologies (e.g. anonymization, encryption) would increase the willingness to share data.

It is highly likely that emerging techniques to protect data will invariably introduce overheads to the DVC activities. Hence, any proposed solution must consider the impact on the service quality.

The specific proposals are suggested as follows:

- Data protection technologies to protect data in the cloud as they are being created, stored, accessed, transferred and processed;
- Data provenance in the cloud, with specific emphasis on application-level and automated reasoning techniques;
- Privacy-preserving technologies for data-sharing in the cloud.

References


4. PARTICIPATION AND ELIGIBILITY

4.1. This call for proposals is open to researchers in all local public sector research organizations, universities, polytechnics and A*STAR-funded research institutes and centres.

4.2. Researchers who are in receipt of SERC intramural or extramural grants in physical sciences and engineering may also apply only if the proposal is a new project.

4.3. Researchers can choose freely the subject of their project (cooperation) within the identified areas and choose the industry partners with whom to collaborate. They should detail how the partners can be identified and engaged.

4.4. It is strongly encouraged that proposals should have multi-disciplinary/organizational research teams and industry collaborators.

5. SELECTION PROCESS AND AWARD

5.1. Complete proposals received under the call will be reviewed by a SERC-appointed Expert Review Panel comprising local and international researchers in the field who will recommend complete proposals for funding. The proposals will be reviewed against the objectives of the Data Value Chain as a Service programme, based on the criteria below, in order of importance:

i. Relevance of the proposal
ii. Innovative merit of the scientific approach
iii. Multi-disciplinary/organisational research team
iv. Track record of the Principle Investigator (PI)
v. Clear outcomes and deliverables
vi. Budget requested

5.2. SERC may require proposals to be combined as it sees fit to enhance collaboration, facilitate integration and optimise resources.

5.3. SERC may seek referee reports, written submissions or presentations from some Principle Investigators (PI) to judge, elaborate or clarify aspects of the research approach described in the proposal during the review process. It is expected that the PI’s presentation to the SERC appointed Expert Review Panel would occur sometime in April 2010.
5.4. Notification of awards will be sent to the respective employing organisations and copied to the Principal Investigator within 6 months of the closing date for submissions. SERC’s decision will be **final**.

6. **TERMS AND CONDITIONS**

6.1 The duration of the project should not be more than 3 years. The receipt of the project proposals by A*STAR in no way implies any commitment to support the proposed activities.

6.2 Collaborative research activities should be described in a single proposal in which a single award is requested. Awards will be made to the lead local institution. Distribution/allocation of sub-awards to other institutions must be administrated by the PI and his/her institute, and must be clearly indicated.

6.3 Both direct and indirect cost items will be supported in each proposal. Direct cost is defined as the incremental cost required to execute the project. This excludes contributions in-kind, depreciation cost of existing equipment, cost of existing manpower as well as building cost. Indirect costs will also be supported (except for A*STAR RiS) and can be used for the research project as stipulated under A*STAR’s funding Terms and Conditions. This is pegged at 20% of total approved direct costs.

6.4 No support for Ph.D. students will be allocated under the initiative. Applicants are strongly encouraged to look towards the A*STAR Graduate Scholarships for funding of Ph.D. students.

6.5 SERC will support up to 100% of each line item qualifying as direct cost. Line items are cost items listed under the three votes – expenditure on manpower (EOM), equipment, and other operating expenses (OOE). SERC reserves the right to vary/remove any of the line items during the moderation of the budget proposal.

6.6 All intellectual properties rights resulting from the projects under the programme is subject to A*STAR terms and conditions on intellectual property rights as stated in Terms and Conditions of SERC Grants.

7. **PROGRAM REVIEW AND MONITORING**

7.1 Projects will be reviewed regularly, and research teams are expected to submit quarterly summary reports. Presentation and public seminars may also form part of the project progress review.
7.2 SERC may appoint a Steering Committee comprising local and international experts to oversee the progress of the funded programme. The Steering Committee may adjust and restructure research directions if required.

7.3 It is the Council’s intent to invite representatives from industry to be updated on the activities of the programme and to provide industry perspectives to the programme.

8. SUBMISSION AND FURTHER INFORMATION

8.1 Please use the proposal and budget templates provided on the website (http://www.a-star.edu.sg/?TabId=247) under “Call for Proposals – Data Value Chain as a Service”:
   • TSRP_Proposal_Details_(Services_Jan2010).xls
   • TSRP_Budget_Proposal_(Services_Jan2010).xls
   • TSRP_Capability_Indicators_(Services_Jan2010).xls

8.2 Softcopy submission of the proposal must be made through the SERC’s webbased Proposal Administration, Management and Submission (PAMS) system, (http://pams.a-star.edu.sg/profile/login.asp). Online help and instructions are available on PAMS. Please allow pop-ups from the PAMS website. All proposed team members must enter/update their profile into the PAMS system.

8.3 After successfully logging into the system, on the Home page’s left hand menu, please select “My Proposals – Create Proposal” to upload your documents. Kindly select “TSRP – Data Value Chain as a Service” under Funding Scheme/Programme.

8.4 The deadline for softcopy submission to PAMS is 1 March 2010 (Monday) at 12 noon SST. A “Transmission is Successful” pop-up will appear after the softcopy submission. Please complete the Front Page “The Next Step” and Section 1 “List of appropriate international reviewers recommended by Principal Investigator &/or Team Members” and attach this document to your hardcopy submission for the necessary endorsements.

8.5 All hardcopy submission must be submitted through their respective Office of Research/Principal’s or Director’s Office/Agency Headquarters/RI ED’s Office or equivalent. SERC will not accept proposals from individuals without the appropriate endorsement from his/her employing organisations. This is to ensure that organisational support is clearly associated with the proposed research plan.
8.6 Only FULL proposals will be accepted for review, evaluation and assessment for awards. The deadline for hardcopy submission to SERC is 5 March 2010 (Friday) at 12 noon SST.

8.7 Please submit the signed hardcopy to:
Mr Soh Kok Hoe
Science and Engineering Research Council (SERC)
Agency for Science, Technology & Research (A*STAR)
1 Fusionopolis Way, #18-10 Connexis North,
Singapore 138632

8.8 Proposals and requests for further information should be directed to:
Mr SOH Kok Hoe, Programme Officer, at soh_kok_hoe@a-star.edu.sg
9. **INDICATIVE ACTIVITY SCHEDULE FOR THE CALL FOR PROPOSALS**

<table>
<thead>
<tr>
<th>No</th>
<th>DATE</th>
<th>ACTIVITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 March 2010, 12 noon SST</td>
<td>Deadline for submission of complete proposals to SERC via PAMS (softcopies)</td>
</tr>
<tr>
<td>2</td>
<td>5 March 2010, 12 noon SST</td>
<td>Deadline for submission of complete and endorsed proposals (hardcopies) to SERC</td>
</tr>
<tr>
<td>3</td>
<td>April 2010</td>
<td>Proposal evaluation by appointed expert review panel (presentation to SERC may be required)</td>
</tr>
<tr>
<td>4</td>
<td>July 2010</td>
<td>Announcement of successful proposals</td>
</tr>
</tbody>
</table>