OSDC/PIRE research proposals

The System and Network Engineering (SNE) research group at the University of Amsterdam is one of the international members of the PIRE project. We are very active and well known for our research on smart e-Infrastructures; network and computing performance, sustainability and security all play an important role in these complex environments and are at the center of our scientific focus.

More information on the group's activities can be found at http://sne.science.uva.nl/ and in the homepages of dr. Paola Grosso (http://staff.science.uva.nl/~grosso/) and prof. Cees de Laat (http://staff.science.uva.nl/~delaat/) and dr. Ana Oprescu (http://uva.nl/profile/a.m.oprescu).

Here follow 6 research proposals, grouped into 3 themes, to be performed at the University of Amsterdam during a period of 6 weeks in the early summer of 2014. The work is suitable for several students working on different aspects; the suitable candidates are graduate students in Computer Science or Computer Engineering.

Theme 1: OSDC and data interoperability

1. **OSDC to OSDC**
   Contact person: Ana-Maria Oprescu

Within the OSDC Pire fellowship 2013, the SNE group has initiated the development of the UvA BigDataBus [1], a service dedicated to data access agnostic to database paradigm. We would like to continue this project, by

- Extending the service to encompass file-based data storage.
- Extending the service to support legacy data, i.e. bridging gaps within data formats due to time-based updates of the software/hardware infrastructure

To achieve the above, the researcher will need to become familiar with such technologies as RESTful services, as well as machine learning techniques.

[1] Ana-Maria Oprescu, Paola Grosso, Pedro Bello-Maldonado, Yuri Demchenko, Cees de Laat, "BigDataBus: Towards a Big Data Aggregation and Exchange Platform for eScience", Proceedings of the Cracow '13 Grid Workshop, November 4-6, 2013, Krakow, Poland, page 97-99, Oct 2013

2. **ENVRI to OSDC**
   Contact person: Roberto Cossu, Ana-Maria Oprescu

The work is suitable for one PhD/researcher interested in multidisciplinary applications based on Earth Science data available through the ENVRI platform.

The goal of this research isto define and implement models making use of data available through the ENVRI platform and the functionalities provided by the platform itself for obtaining value-added results of scientific interest.

To achieve the above the researcher will need to familiarize with the ENVRI services and tool, including state-of-the art protocols for data discovery and data access (such as the OGC protocols: OpenSearch, WMS…). The ENVRI consortium partners will provide the necessary support.

More details may be found in Theme 1.2 OSDC and ENVRI.

Theme 2: OSDC and SDN

Contact person: Zhiming Zhao

Cloud and advanced network technologies such as Software Defined Networking (SDN) allow applications to customize the creation of virtual machines and to program their connectivity; these features can optimize the data aggregation and transfer with control of network.

The goal of this project is to investigate the key features of the intelligent data access services that can bridge the gap between application level workflows and the controllability of the advanced network.

The students will use the existing framework developed in the SNE group for planning workflow, semantically linking data and infrastructures, and data delivery and aggregation, and will explore the integration between these systems in the OSDC test bed. The student should be familiar with java programming, semantic web technology and scientific workflow systems.

More details may be found inTheme 2.1 OSDC and SDN

Theme 3: OSDC and Big Data issues

Contact person: Miroslav Zivkovic

1. **Network-level control of BigData application performance**
   In this project we aim to investigate intelligent network that, during each stage of Big Data analysis adaptively scales to adjust for bandwidth/processing requirements of the data transfers. This should result in improved data processing time and improved overall utilization of the system. SDN is an ideal candidate to build such intelligent adaptive network, and it can be used to configure such network on-demand to the optimal size and shape for computing servers. As the SDN Controller has an overview of the underlying network, like network utilization, etc. the developed solutions can accurately translate the Big Data analysis needs by programming the network on demand.

   Project content

The literature review, SDN capabilities inventory, requirements scoping to specifying the initial architecture and design of the enhanced, intelligent SDN Controller.
The ideal candidate should have expertise in two or more of the following: networking, SDN, performance engineering, software engineering.

More details may be found in Theme 3.1 Network-level control of BigData application performance.

2. **Streaming issues of BigData applications**
   In proposed research we want to identify and further specify the main research areas for online data analytics of large streaming data. Some of the envisioned issues to be tackled are continuous queries on data streams may involve different streaming data sources, with different data models (e.g. XML-like), and communication protocols they use (e.g. HTTP). Besides, with the emerging cloud-based heterogeneous systems, the integration of different processing elements (as potential solution for distributed analysis) may be important. Last, but not least, we would like to explore highly scalable algorithms for analysis, e.g. anomaly detection.

   Project content

   The project presents a foundation for the future research in this area. Therefore, the candidate involved should perform a great deal of survey the state-of-the-art, identifying challenges and defining concrete research topics. Based on the outcome of this phase, a solution for selected problem may be derived, and investigated in a laboratory environment.

   More details may be found in Section Detailed Proposals

3. **Data retention policies**
   A large amount of data is processed by any Big Data relating business companies. As the data used for analysis grows, the search time increases, as well as the storage space. In this project we want to investigate the various data retention policies; a simple example could be to remove data that is older than a certain threshold time, or threshold size.
   We want to make an efficient an customizable management software for the problem at hand and implement it using some of the available Big Data platforms (e.g. Splunk).

   More details may be found in Theme 3.2 Streaming issues of BigData applications.

Detailed Proposals

Theme 1.2 OSDC and ENVRI

The work is suitable for one PhD/researcher interested in multidisciplinary applications based on Earth Science data available through the ENVRI platform. The goal of this research is:

- to define and implement models making use of data available through the ENVRI platform and the functionalities provided by the platform itself for obtaining value-added results of scientific interest.

The researcher will have the opportunity to run the proposed model on a huge variety and amount of data available in ENVRI. Data include (but are not limited to) the following collections:

- Argo floats series are available for years 2009, 2010 and 2011;
- Samples of ICOS CO2 and CH4 for Mace Head, Cabauw and Puijo stations;
- MOIST seismic ground motion, gravity, magnetism, seafloor deformation, pore pressure or heat flow for Western Ionian sea, Marmara sea and Gulf of Cadiz;
- EISCAT series from Tromsoe and Svalbard radars;
- Lifewatch alien biocase series ;
- EARLINET Lidar series from Andenes, Athens and Aberystwyth stations;
- ESA Envisat data.

(The candidates are invited to identify additional data of interest for their model. ENVRI partners will evaluate the feasibility of adding such data to the ones already discoverable in the platform.)
Concerning the main ENVRI platform services, i.e., data discovery, access, and processing, the candidates can propose one the following two options (the final decision will be made in agreement with the ENVRI consortium partners):

- he/she will develop the model in a machine external to the ENVRI platform (this can be a machine at the hosting premises or at the home organization of the researcher, or even the laptop of the researcher). In this case the SW developed by the researcher shall use of the ENVRI APIs (based on OpenSearch) to discover and access the data that will be then locally processed (downloading of data may require to accept particular license conditions as defined by the data owners);
- he/she will develop the model in the ENVRI Virtual Research Environment, which is based on D4-Science technology, without the need of downloading data.

  To achieve the above the researcher will need to familiarize with the ENVRI services and tool, including state-of-the art protocols for data discovery and data access (such as the OGC protocols: OpenSearch, WMS…). The ENVRI consortium partners will provide the necessary support.

Theme 2.1 OSDC and SDN

In research infrastructures, data access services are important components for enabling discovery and retrieval of scientific data subject to authorization. The data access services provide applications with abstract interface to select and obtain data content from distributed sources. When data are heterogeneous, the data access services may convert or aggregate different data types (often pulled from a variety of distributed data resources) into uniform representations with uniform semantics. In the meantime, Cloud and advanced network technologies such as Software Defined Networking (SDN) allow applications to customize the creation of virtual machines and to program their connectivity; these features can optimize the data aggregation and transfer with control of network. However, most of current data access services handle the transfer based on specific task in a workflow process and provide applications limited capability to optimize their data transfer from a global view of the workflow. It makes the resource utilization difficult to achieve system level optimal during workflow execution.

The goal of this project is to investigate the key features of the intelligent data access services that can bridge the gap between application level workflows and the controllability of the advanced network. The students will use the existing framework developed in the SNE group for planning workflow, semantically linking data and infrastructures, and data delivery and aggregation, and will explore the integration between these systems in the OSDC test bed. The student should be familiar with java programming, semantic web technology and scientific workflow systems.

Theme 3.1 **Network-level control of BigData application performance**

One of the major requirements for Big Data systems is ability to efficiently process (analyze) large amount of both structured and unstructured data. In order to achieve speed and efficiency the algorithms used for Big Data analytics are usually parallelized and distributed over many clusters of hundreds of servers connected via high-speed (Ethernet) networks. Therefore, data processing speed (one of the biggest bottlenecks for Big Data) can only be as fast as network's capability to transfer data between servers in different phases of the analysis process. A recent study on Facebook traces [1] show that this data transfer between successive stages may account more than 50% of job completion time.

In this project we aim to investigate intelligent network that, during each stage of Big Data analysis adaptively scales to adjust for bandwidth/processing requirements of the data transfers. This should result in improved data processing time and improved overall utilization of the system. SDN is an ideal candidate to build such intelligent adaptive network, and it can be used to configure such network on-demand to the optimal size and shape for computing servers. As the SDN Controller has an overview of the underlying network, like network utilization, etc. the developed solutions can accurately translate the Big Data analysis needs by programming the network on demand.

[1] Mosharaf Chowdhury, Matei Zaharia, Justin Ma, Michael I. Jordan, Ion Stoica, Managing Data Transfers in Computer Clusters with Orchestra.

Project content

The literature review, SDN capabilities inventory, requirements scoping to specifying the initial architecture and design of the enhanced, intelligent SDN Controller.
The ideal candidate should have expertise in two or more of the following: networking, SDN, performance engineering, software engineering.

Theme 3.2 **Streaming issues of BigData applications**

In this project we will look into the challenges of the (near) real-time Big Data analytics of streaming data. It is well known that this presents a very challenging area for many reasons, some of the most important being high data volume (infinite length), lack of capability to store data, (resource) expensive computations, etc. There is a clear need for scalable processing over very large data streams, that would satisfy the requirements of scalable search, statistical analysis and scalable (resource) expensive computations. The processing of data streams should include multiple streaming data sources and data processors.

In proposed research we want to identify and further specify the main research areas for online data analytics of large streaming data. Some of the envisioned issues to be tackled are continuous queries on data streams may involve different streaming data sources, with different data models (e.g. XML-like), and communication protocols they use (e.g. HTTP). Besides, with the emerging cloud-based heterogeneous systems, the integration of different processing elements (as potential solution for distributed analysis) may be important. Last, but not least, we would like to explore highly scalable algorithms for analysis, e.g. anomaly detection.

Project content

The project presents a foundation for the future research in this area. Therefore, the candidate involved should perform a great deal of survey the state-of-the-art, identifying challenges and defining concrete research topics. Based on the outcome of this phase, a solution for selected problem may be derived, and investigated in a laboratory environment.