# Data steaming infrastructure for scalable distributed information sharing platforms

The ViSUS platform involves a series of components that allowed developing a range of tools for a variety of applications including: accessing large data via web services, processing massive images, provide I/O components for massive simulation codes (tested up to 140,000 cores), and providing clients that can run on handheld devices (iPhone/iPad). The image below shows the dependency among the ViSUS core components, the external libraries used (required or optional), and the vertical applications developed in each application area. Scripting interfaces based on swig provide ready access to the ViSUS functionalities using a python interpreter for scripting purposes.
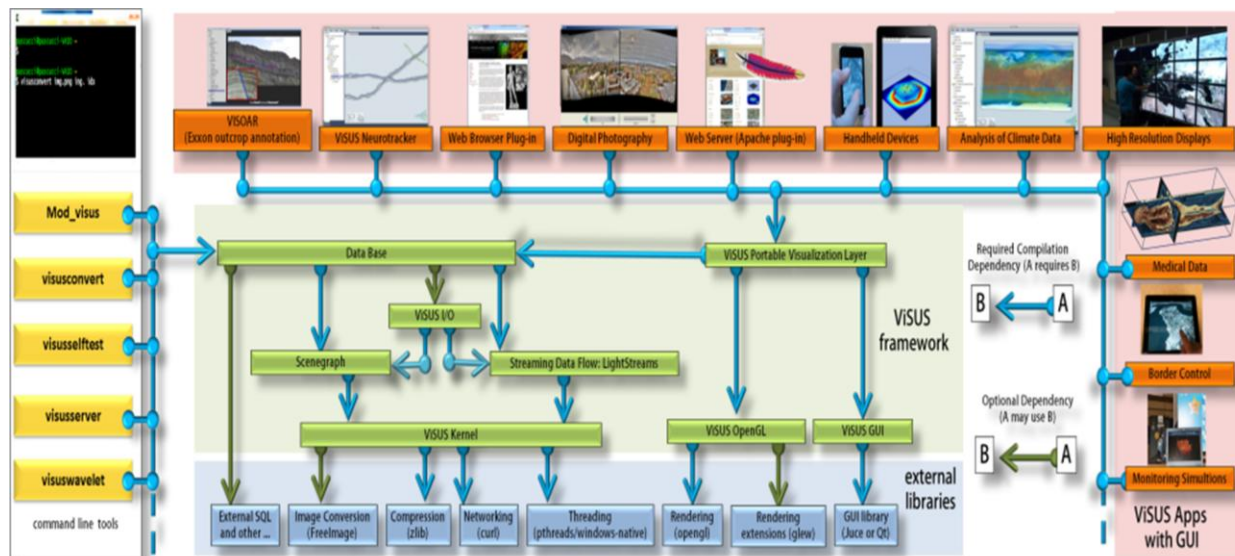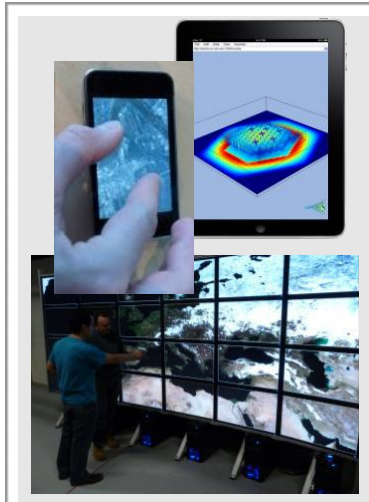

The ViSUS framework introduces an advanced data-flow infrastructure that allows easy data access in a variety of hardware-configuration scenarios. ViSUS is used in a number of practical usage scenarios where the data-streaming infrastructure enables effective connection from real-time scientific simulations to visualization environments. For instance in the adjacent figure ViSUS is shown in use with a turbulent-mixing simulation running on the BlueGene/L supercomputer that streams its data directly to a remote ViSUS client. This type of direct streaming connection between data sources and remote users allows for interactive testing of alternative scenarios and validation and verification of actionable information. Using the ViSUS infrastructure even users using remote, hand-held devices can access and interact with extremely large data.

A key requirement in working with modern distributed, heterogeneous environments designed for massive scale data sharing, is the efficient combination and coordination of data processing and data movement. One of the main tenets of the ViSUS system is in organizing the data in a cache-friendly way to best leverage modern hardware architectures. A particular emphasis of the ViSUS system is the use of *cache oblivious* approaches that optimize cache access patterns without being limited to any particular cache size. Since these techniques are intrinsically scalable, they are optimal for heterogeneous hardware infrastructures without reprocessing data on each platform. This feature enables the highly effective use of a variety of platforms, ranging from portable devices (e.g. iPhone/iPad) to large workstations or powerwall-type displays. The adjacent figure shows the same ViSUS code running efficiently at different scales, without the need to rewrite the visualization or data access components. Recent testing has shown scalability of the ViSUS I/O library (format PIDX) for over 100,000 core platforms (Intrepid supercomputer at Argonne National Laboratory).

A ViSUS client remotely displaying a turbulent-mixing simulation concurrently running on the BlueGene supercomputer at Lawrence Livermore National Laboratory (demonstration on the floor at the IEEE SC Supercomputing

To achieve this scalability, the ViSUS system optimizes the data flow networks connecting the data sources to the user interfaces, as well as the sequences of data transformations. With these optimizations the ViSUS system is naturally highly resilient to latencies resulting from network infrastructure and computing stages. This resilience is ideal for environments that involve demanding user interactions (even small latencies can dramatically impair the effectiveness of the human analytical reasoning process). The fundamental ViSUS approach is to improve the performance of the environment by reducing the granularity of the computations. Smaller computations that are executed quickly produce incomplete yet valid outputs that we can efficiently transferred through the environment at a minimal cost while maximizing opportunities for parallel computation. The generated outputs progressively improve as more of the input information is processed. By varying the granularity of the computation the ViSUS systems optimizes the overall performance of the data flow, avoiding costly bulk data moments and unnecessary data duplications. The image below shows an instantiation of a ViSUS data streaming infrastructure, where each rectangle represents a streaming module that processes data sets in movement along the connecting lines.

This efficient data streaming framework the key to achieving a sophisticated data management, processing, and visualization environment that relies on the concurrent presentation of multiple data abstractions and different levels of detail. Naïve implementations of data movements lead to major bottlenecks. Achieving responsiveness and scalability of the system requires quick access to real-time data segments, while avoiding repeated computations. To accommodate different types of computations and interaction modalities, we define modules that employ two primary data transmission mechanisms: a *pull model*, where data is produced on demand in a "lazy" way, and a *push model*, where input data is immediately processed and transferred to the next module.

Distributed data sharing

Some of the core challenges in developing a distributed data sharing infrastructure involve the access to da large variety of data sources and file formats. The flexible ViSUS infrastructure enables easy deployment of a distributed data sharing infrastructure for general data types, structured and unstructured, such as text, tables, images, etc. When needed, on-the-fly data conversions can be wrapped into data streaming modules that can cache information for repeated access. The infrastructure facilitates large scale data distribution allowing possible compatibility with legacy code/formats. The images below show possible deployment scenarios for making immediately available data acquired at different sources or for cascade distribution with caching a multiple levels.
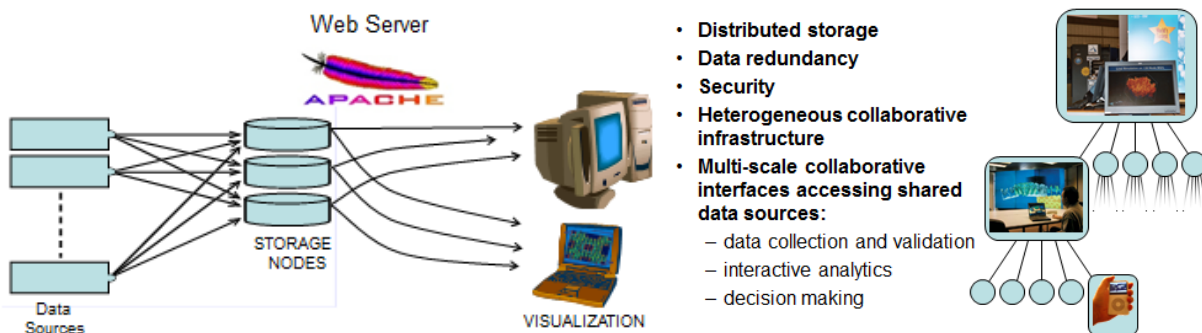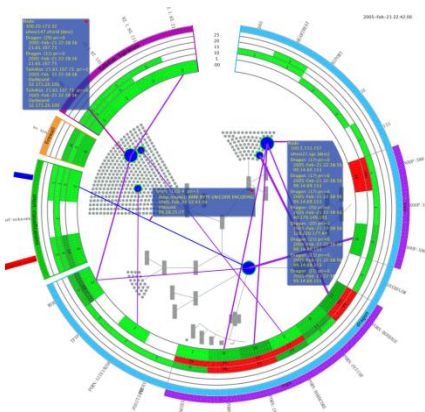
# Scalable User Interfaces for Suggestive Big Data Analytics

Coupled with scalable domain-specific user interfaces the ViSUS data access and processing infrastructure offers major advantages in the development of scalable analytics capabilities that form an agile human-machine collaborative environment. Previous experience shows that these capabilities facilitate and empower a seamless and mutual dialog where both analysts and computers contribute to the solution of a given task. This scalable visualization environment will be uniformly accessible to different classes of users such as analysts and decision makers that require different degrees of data access capabilities and different presentation devices. Using both complex convergent and divergent thinking, forming multiple hypotheses, and rejecting others through a dynamic and seamless human-machine dialog, will immerse the user in a dynamic environment where any visual metaphor on the screen, either textual or graphical, can be grabbed, manipulated and seamlessly moved between disparate interfaces to explore the underlying data and intuitively form complex queries. This agile approach translates the user interaction and available computing resource in a continuously evolving set of requirements, workflows, and processing pipelines.

In the creation of user interfaces we focus on the analytic discourse between users and their data through scalable visual interfaces, immersive interaction paradigms and our powerful data streaming transportation layer. We leverage our innovative work data streaming and visual analytics and create a novel immersive visual experience that cater to the user mental model, reduces cognitive load and free the user to engage in uninterrupted hypotheses generation (see Figures below).

When using multiple coordinated views, the user starts with a global overview of data and drills down to more detailed views, using selection, filtering, and coordinated zoom operations that are continuously mapped to the creation of ViSUS progressive processing pipelines. A user can focus on a data subset by progressively selecting and narrowing in different views, based on known attributes (linking)
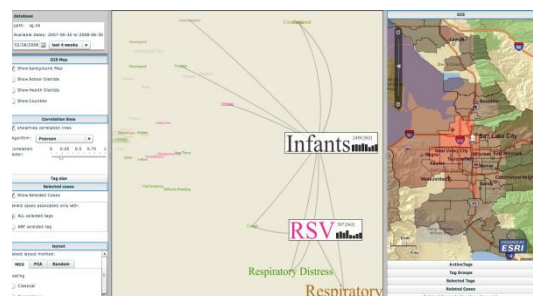
The ViSUS system is designed to take advantage of modern hardware architectures. With technical advances like cache oblivious approaches the same ViSUS system runs on systems as diverse as powerwalls,
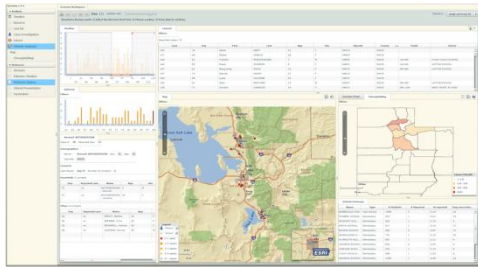
adopting a *convergence thinking* approach, an analytical reasoning process that assembles evidences from the user's ability to gather, see, and understand complex data.

In our environment, we also generalize this approach with the concept of *loosely coordinated multiple views* (see Figure below) by only enforcing a loose coordination between the various views. This enables users to perform brushing, linking, and filtering operations on only a subset of the views or, alternatively, designate d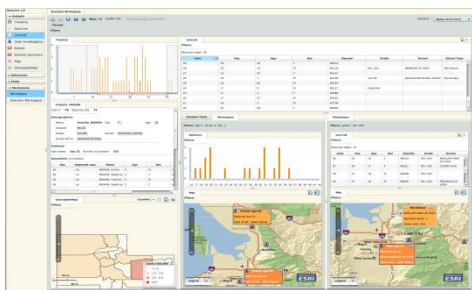ifferent data sources to different views. The framework can then organize views into workspaces that focus on different tasks and aspect of the data.

VisAlert provides a visual correlation of the "when", "where" and "what" to facilitate network intrusion detection and improve situational awareness.

While the traditional coordinated multiple views only supports convergent thinking, the loosely coordinated views approach also supports *divergence thinking*, i.e., the exploration of multiple plausible alternatives and seeking evidences that may support or refute them.

To ease the learning curve and cater to users with a wide range of skills and expertise we employ *progressive disclosure of capabilities*, exposing initial only a subset of the functionalities and modules that are appropriate for novice and non-expert users. Additional and more advance features are introduced as the user gain more experience and require more advanced capabilities.

Overall, we introduce a radical shift from the current interpretation of the "Overview, pan and zoom, then details on demand" strategies of visual analytics systems and adopt and disruptive technology based on (i) scalability of the data management system, (ii) scalability of the data processing environment, (iii) scalability of the user interfaces, and (iv) scalability of the user interactions. This results in agile deployment of data processing and visualization environments that foster fluid and simultaneous visualization and interactions with massive data at multiple levels of abstractions liberating the user of the conceptual and processing inefficiencies of classical batch processing and turn-based data interaction models.

Overall the combination of the Scalable and Extensible User Interfaces with the Large Data capabilities of the ViSUS system provide a highly adaptable system for Large Data Analytics and Visualization.