

**WHITE PAPER** : BRINGING CLOUD COMPUTING ARCHITECTURES TO BIG DATA PROBLEMS

# **BRINGING CLOUD COMPUTING ARCHITECTURES TO BIG DATA PROBLEMS**

September 2011



Appistry enables companies to create automated solutions that transform complex, unstructured data into actionable information

*The Ayrris™ platform by Appistry brings High Performance Computing and Cloud Architectures together for the first time to help companies solve the problems associated with Advanced Analytics applications.*

## Executive Summary

Big Data problems exist in every industry, from financial services and life sciences to health care and defense. And they represent some of the hardest and most important problems affecting anti-terrorism, bioinformatics, financial analysis and scientific research.

Yet the traditional high performance computing (HPC) and analytics systems industry professionals rely on have become increasingly more difficult to implement, deploy, and manage. These systems demand more computation and data storage resources than ever before. To tackle this, high-end software architects typically develop systems for single use implementations or integrate multiple open source projects into a single framework. However, in addition to being costly, these systems tend to be highly problematic. They are difficult to maintain and often overlook situations that can cause errors. What's more, because they rely upon a static master node that assigns work, this produces a single point of failure — and performance nightmares.

The key problems:

- » Traditional architectures that don't scale to today's needs
- » Large amounts of unstructured data from a variety of sources
- » Highly variable and intense computational requirements
- » Complex automation and pipeline management challenges

Application developers and system administrators no longer can depend on the release of faster and larger hardware to handle their processing and storage needs.

**To create next generation analytics systems, system administrators and developers need a better solution.**

They need a platform that:

- » Enables fast data processing rates to improve application performance and cost effectiveness.
- » Is easy to maintain with tools that allow a system administrator to manage a collection of machines as if they were a single system.
- » Handles scaling, distribution, and reliability issues, eliminating points of failure, so developers can spend more time concentrating on the business logic instead of babysitting computers.
- » Simplifies application deployment by offering a scalable and consistent framework capable of providing dynamic membership and software provisioning services.

## THE PROBLEMS

### Moving Data is a Bottleneck

Traditional architectures typically possess a storage array with a computational cluster. For many applications, a large percentage of time is wasted simply moving gigabytes of data from the storage array to the computation cluster for processing.

*Can we rethink how data is stored and maintained? Instead of moving data to the work, what if we moved the work to the data to speed up the process?*

### Scalable Execution is a Difficult, Costly Development Task

Execution environments naturally grow or shrink as machines are added or removed. To begin developing a distributed processing environment for business logic from scratch presents a number of challenges. To accomplish the task quickly, all of the processing nodes are specifically named by the application; scalability becomes cumbersome and relies upon maintaining a development staff capable of expanding the system. To accomplish the task generically, a robust framework must be developed to accommodate the deployment of the business logic independent of the number nodes.

*Is there a way to design a scalable network that distributes labor and resources so that no single machine within the system becomes overloaded with communication or computational requests?*

### Software Distribution and Provisioning is Cumbersome

Traditional large-scale HPC and analytics systems are difficult to deploy and manage. High-end software architects typically handcraft systems for single-use implementations or integrate multiple open source projects into a single framework. As a result, these systems often overlook problems that can cause cascading errors as well as performance bottlenecks. It is cumbersome for developers and system administrators to deploy existing web servers, application servers, software libraries, and service-oriented applications across many machines.

*How can we create a superior platform for HPC and analytics applications that will simplify deployment and system maintenance? How can we free up valuable time of high-end developers and application specialists?*

#### **DISCOVER THE APPISTRY ADVANTAGE: THE AYRRIS™ PLATFORM**

Ayrris applies cloud-computing architectures to provide a scalable, reliable, and adaptable framework for data-intensive, high-performance computing applications.

An Ayrris solution begins with a collection of commodity computers dedicated to a single Ayrris instance. Each of these machines, called a worker, is prepared with the base operating system (Windows or Linux) and the base Ayrris system services. Once the Ayrris system services are installed, the machine becomes a complete, autonomous worker in the Ayrris system. Although able to work alone, the worker immediately starts looking across its network for peers of the same Ayrris instance. When workers find each other, they will begin to share resources and to create the larger system.

This balance between worker self-sufficiency and collaboration provides the Ayrris platform with its greatest strengths. It provides the higher-level system services and infrastructure required to build the large HPC and analytics applications needed to solve today's large data problems.

### The Appistry Solution:

- » Superior Performance
- » Absolute Reliability
- » Simplified Scalability
- » Painless Deployment
- » Unlimited Capacity

## Scalable + Reliable = Cost-Effective Data Delivery

Ayrris was built to directly answer the needs of today's large data HPC and analytics applications, which can easily amass petabytes of data. Traditional SAN, NAS, and RDBMS platforms are expensive to acquire and difficult to manage at these scales. The Ayrris platform offers an affordable storage solution for holding vast amounts of files and binary objects.

To provide dynamic scalability across machines and reliable storage of files, Ayrris unifies the collection of attached storage devices owned by the workers within the Ayrris instance and logically joins them into a single storage repository. This single storage repository can span multiple data centers. Ayrris mirrors and distributes all data files across all cloud storage

servers, providing disaster recovery services through a feature called “territories.” A territory is a logical grouping of workers in the Ayrris platform typically created by administrators based on network or geographic topologies. When territories are enabled, Ayrris will distribute the copies of the file across the territorial boundaries. In the event of a data center outage, access to files is not lost. When the connections are reestablished, Ayrris will examine and correct any version skew of files between data centers.

## Faster Data Delivery

In addition to providing scalable, reliable storage, the Ayrris platform improves data processing rates for HPC and analytic applications through the Computational Storage™ feature. Computational Storage is the unification of a storage system with an execution system that can move work to where the data resides. Unlike traditional clusters, the Computational Storage feature of Ayrris uses information about file locations to guide incoming HPC and analytic work requests to the machine holding the relevant data files. As files are read at disk speed instead of network speed, latency is reduced and data throughput is increased.

As a result, Ayrris can improve data processing rates for HPC-style applications by 10 to 100 times, making analytics results available faster and allowing the delivery of data-intensive applications at unprecedented price points.

## Effortless Execution

A simplified application environment created by combining a robust processing engine with cloud computing architecture makes constructing HPC applications easy for all application developers. The developer simply needs to construct the application code and describe the runtime needs through the processing engine’s orchestration language. The developer does not need to handle the threading, distribution, or reliability of the application as the platform provides these services.

To execute applications within Ayrris, developers present the system with two elements for application construction:

- » The first element is the application binaries. The application binaries can be low-level code libraries (written in C/C++, Java, or .NET), prebuilt applications and utilities, or open source projects. These binaries contain the business logic for the application.
- » The second element for an Ayrris application is the service orchestration. The service orchestration is a metadata representation of the series of actions to execute incoming HPC jobs and/or service-oriented architecture type requests. These applications are deployed to all the machines within the Ayrris instance using a distribution mechanism.

The application orchestrations teach Ayrris about the runtime requirements of applications. Ayrris then guides each individual request to the best machine possible at the moment through its knowledge of the machines available in the system, the application orchestrations, and each machine's current resource utilization. In addition Ayrris instructs multiple machines to be aware of each request. If a worker fails, Ayrris has the ability to restart the request at the failure point instead of starting over from the first step. Workers will automatically self-organize into functional groups. These smaller groups manage communication pathways that would otherwise overwhelm the larger system. To the system administrator and application developer, Ayrris is a single computation and storage resource. They do not need to be aware of the individual machines.

## Simplified Deployment and Provisioning

The Ayrris platform handles software provisioning by providing a mechanism for software developers and administrators to package individual software components with all the metadata necessary to instruct the system on how to manage deployment and provisioning. These software modules, referred to as "FAR" files, contain the binaries, installers, configuration files, and metadata describing the software's name, version, dependencies, and life cycle actions. This metadata provides the Ayrris environment with the information necessary to develop an installation plan for all software components running within an Ayrris instance.

The Ayrris platform's installation planning mechanism simplifies the deployment of large Ayrris instances. In fact, it gracefully updates subsets of machines until the entire system is updated. Users do not have to be aware of the update process and the system is self-healing with respect to application versions. A worker provides its system resources in accordance with its capacities. Further, workers within an Ayrris environment do not have to be homogenous with respect to hardware. That means, workers become aware of each machine's resources and dynamically load balance work and responsibilities accordingly. Workers are constantly watching to detect the addition of new members or the loss of a machine. Changes in membership status are quickly propagated through the system so each machine can recalibrate communications and actions as needed. Workers isolate actions to the smallest subset of workers necessary to complete the action quickly and reliably. If every worker needed to be aware of every action, the system would eventually collapse as the numbers of workers grow. By isolating actions to a small subset of machines, the system can grow to very large sizes.

## Conclusion

The Ayris platform offers system administrators and developers the services necessary to construct next-generation analytics systems. They're able to accelerate development through a robust processing engine and simplify deployment with a scalable, reliable framework.

As a result, companies are able to:

- » **Lower costs through a reliable platform that does not require a dedicated staff to oversee**
- » **Analyze at lightning speed through unprecedented data processing rates**
- » **Achieve quality and performance through a scalable execution that can handle massive amounts of data**

Appistry's products offer a new level of scalability, elasticity and reliability for data-intensive applications, allowing enterprises to exploit the myriad of available data and quickly turn that data into actionable intelligence faster, better and cheaper than ever before.

With a broad customer list that includes FedEx, GeoEye, Lockheed Martin, Northrop Grumman and others, Appistry's platform supports mission-critical, data-intensive applications for some of the world's leading organizations.

