FRIEDA: Flexible Robust Intelligent Elastic Data Management Framework

Devarshi Ghoshal[1], Valerie Hendrix[1], William Fox[1], Sowmya Balasubhramanian[1], and Lavanya Ramakrishnan[1]

[1]Lawrence Berkeley National Lab

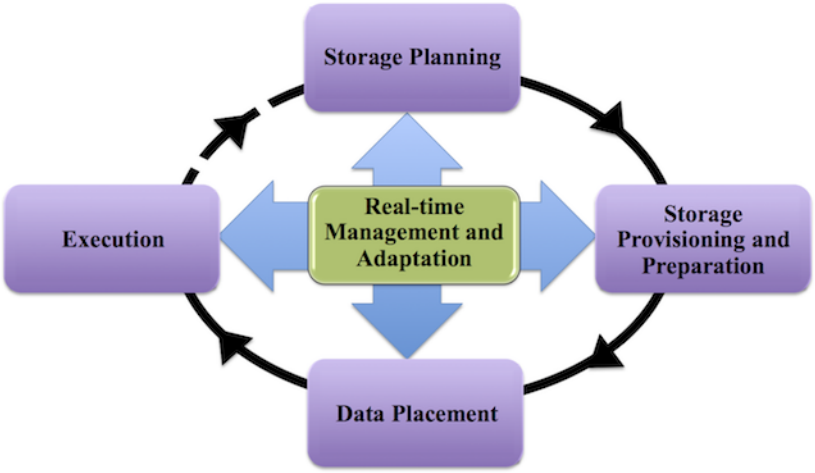
11 Nov 2016

Paper DOI: http://dx.doi.org/10.21105/joss.00164
Software Repository: https://bitbucket.org/dghoshal/frieda
Software Archive: http://dx.doi.org/10.5281/zenodo.290299

Summary

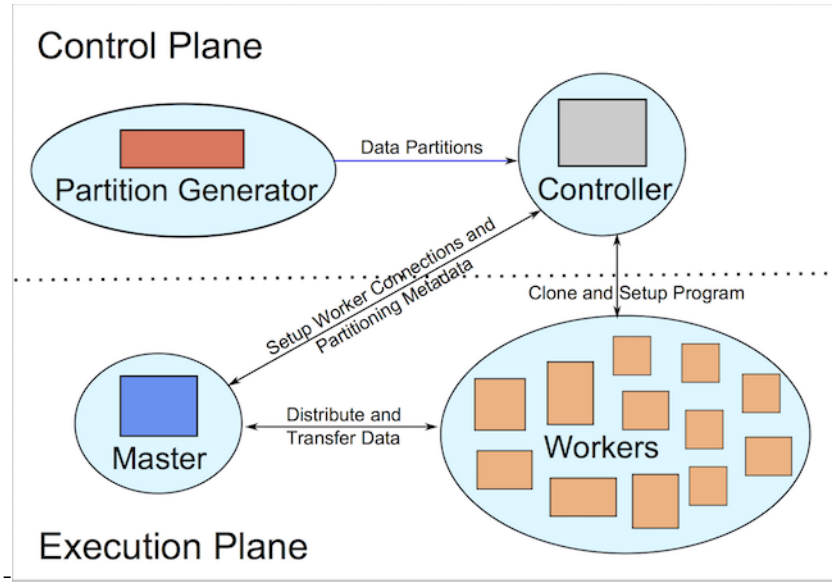
Scientific applications are increasingly using cloud resources for their data analysis workflows. However, managing data effectively and efficiently over these cloud resources is challenging due to the myriad storage choices with different performance, cost trade-offs, complex application choices and complexity associated with elasticity, failure rates in these environments. The different data access patterns for data-intensive scientific applications require a more flexible and robust data management solution than the ones currently in existence. FRIEDA is a Flexible Robust Intelligent Elastic Data Management framework that employs a range of data management strategies in cloud environments (FRIEDA (2016),Ghoshal and Ramakrishnan (2012),Ghoshal and Ramakrishnan (2014)).

FRIEDA can manage storage and data lifecycle of applications in cloud environments (Ramakrishnan et al. (2014)). There are four different stages in the data management lifecycle of FRIEDA – i) storage planning, ii) provisioning and preparation, iii) data placement, and iv) execution.

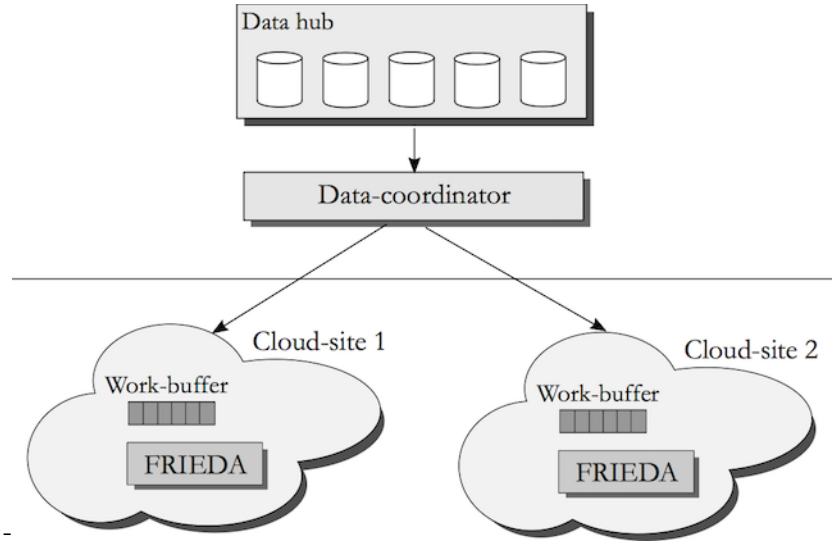


FRIEDA defines a data control plane and an execution plane. The data control plane defines the data partition and distribution strategy, whereas the execution plane manages the execution of the application

using a master-worker paradigm. FRIEDA also provides different data management strategies, either to partition the data in real-time, or pre-determine the data partitions prior to application execution.



FRIEDA also provides a module to manage data across multiple heterogeneous cloud sites, called the FRIEDA data coordinator. The data coordinator is responsible for coordinating data movement to different cloud sites based on the properties of data and cloud instances.



FRIEDA is released on a modified BSD license with an added paragraph at the end. FRIEDA supports cloud platforms like Amazon EC2 and OpenStack. Users need to provide cloud and application configuration information through a YAML file that is used by FRIEDA to setup and manage data in the cloud. FRIEDA provides different data management strategies for different applications. It has been tested to work efficiently with a protein sequence database ($>$ 6 GB), as well as for comparing light source images ($>$ 500 GB). It is only limited by the underlying network limitations of the cloud, but allows users to configure the resources and data distribution as per the requirements. The references include papers that describe the methodologies and the experiments in detail.

References

FRIEDA. 2016. "FRIEDA." http://frieda.lbl.gov.

Ghoshal, Devarshi, and Lavanya Ramakrishnan. 2012. "FRIEDA: Flexible Robust Intelligent Elastic Data Management in Cloud Environments." In *Proceedings of the 2012 Sc Companion: High Performance Computing, Networking Storage and Analysis*, 1096–1105. SCC '12. Washington, DC, USA: IEEE Computer Society. doi:10.1109/SC.Companion.2012.132.

———. 2014. "Provisioning, Placement and Pipelining Strategies for Data-Intensive Applications in Cloud Environments." In *Proceedings of the 2014 Ieee International Conference on Cloud Engineering*, 325–30. IC2E '14. Washington, DC, USA: IEEE Computer Society. doi:10.1109/IC2E.2014.66.

Ramakrishnan, Lavanya, Devarshi Ghoshal, Valerie Hendrix, Eugen Feller, Pradeep Mantha, and Christine Morin. 2014. "Storage and Data Life Cycle Management in Cloud Environments with Frieda." In *Cloud Computing for Data-Intensive Applications*, edited by Xiaolin Li and Judy Qiu, 357–78. New York, NY: Springer New York. doi:10.1007/978-1-4939-1905-5_15.