

WMArchive Performance Service

CERN Summer Student Programme 2016

This report, including detailed weekly reports and documentation on the project, is available in the [WMArchiveAggregation repository](#) on GitHub.

- **Author:** [Nils Leif Fischer](#)
- **Supervisors:**
 - [Valentin Kuznetsov](#), Cornell University
 - [Dr. Dirk Düllmann](#), CERN
- **Date:** June 27, 2016 to September 23, 2016

- **Abstract:**

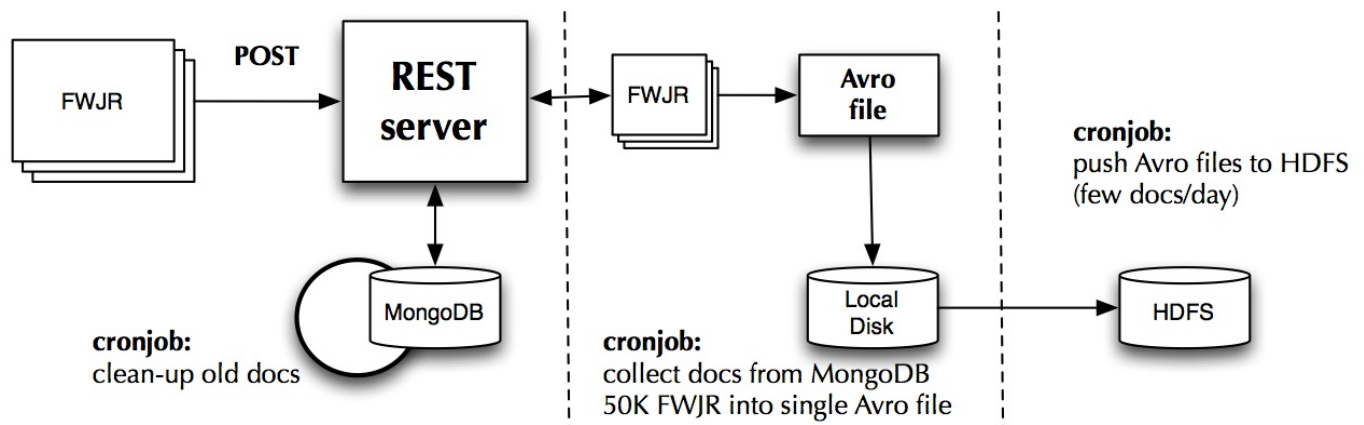
This project is part of the [WMArchive](#) project that provides long-term storage for the CMS workflow and data management *framework job reports (FWJRs)*.

An aggregation pipeline regularly processes the distributed database of FWJRs to collect performance metrics. An [interactive web interface](#) visualizes the aggregated data and provides flexible filters and options to assist the CMS data operators in assessing the performance of the CMS computing jobs.

Overview

The CMS production agents schedule jobs on the computing grid. When a job finishes it generates a FWJR. The agent then posts this FWJR to

the WMArchive REST server that buffers it in a short-term MongoDB database. Then every day the new job reports are converted into a binary Avro file format and migrated to the long-term HDFS storage. This long-term storage is designed to hold millions of these FWJR documents for an indefinite amount of time. This architecture is summarized in the diagram below:

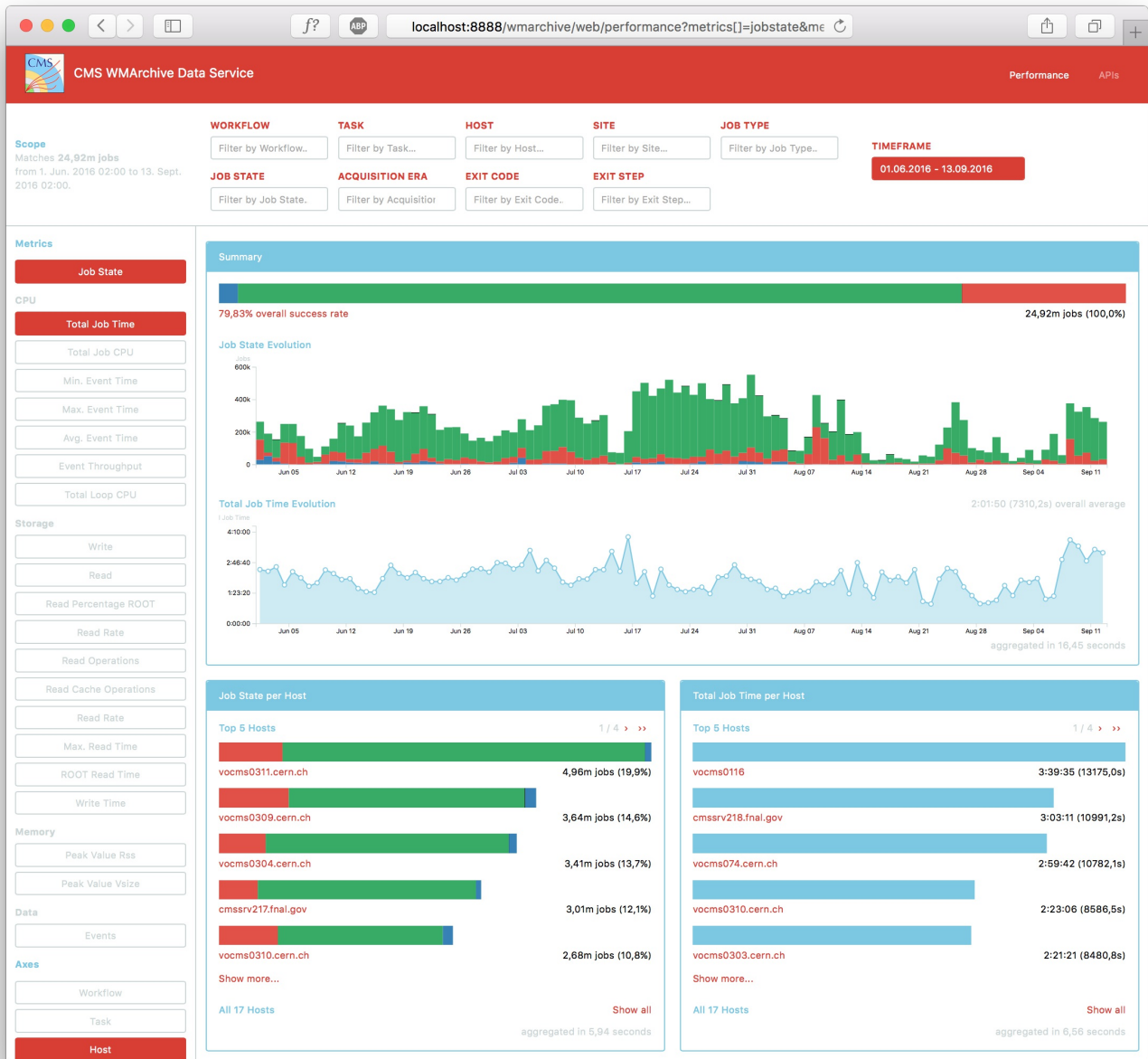


The WMArchive Performance Service project is about retrieving this data stored in the long-term archive and visualizing it in a web interface for CMS data operators to investigate.

Of course to access the data in the long-term HDFS storage it is necessary to schedule jobs that retrieve the data, and those can take a significant amount of time. So to provide a responsive user interface I constructed an aggregation pipeline that regularly processes the distributed database of FWJRs to collect performance metrics and cache the aggregated data back in the MongoDB, where it is quickly accessible by the REST server and the UI. So this cache is not data on each individual job report but instead aggregated data grouped only by a number of attributes data operators may want to filter by, so for example the job success state, its host or its processing site.

The second part of the project is to build a web interface for CMS data operators to visualize and investigate the aggregated data. It provides

flexible filters and options to assist the CMS data operators in assessing the performance of the CMS computing jobs. The screenshot below shows the result of my work on the performance service frontend this summer that is available on the [CMSWeb Testbed](#).



Progress Reports

Every week I documented my progress on the project in reports that are available in the [WMArchiveAggregation repository](#). These reports include detailed discussions on the architecture and design choices,

implementation details and documentation.

Documentation

The [WMArchiveAggregation repository](#) also contains thorough documentation on the project setup, usage, architecture and implementation in addition to code-level documentation in the [WMArchive repository](#).

Outlook and Conclusion

During my stay at CERN for the Summer Student Programme 2016 I produced a working prototype of the WMArchive Performance Service that is available on the [CMSWeb Testbed](#). I outlined the remaining tasks to finalize the aggregation pipeline for production deployment and further pending improvements in the [WMArchiveAggregation repository](#).

This project gave me the unique opportunity to both, learn to apply a variety of modern computing technologies to aid in scientific data processing, and experience the extraordinary intercultural and scientific environment at CERN.

For the former I would like to thank the WMArchive group and particularly Valentin Kuznetsov for their professional support and encouragement during this project.

For the latter I must extend my gratitude to my fellow summer students that filled this summer with cultural and personal differences that were never in need of resolution but instead served only to enrich every participant's experience and appreciation of each other. Remarkably, merely assembling a crowd of aspiring scientists to work in CERN's

stimulating and resourceful environment suffices to produce these most appreciable qualities, thus elevating them to virtues intrinsic to the scientific community. I am therefore particularly grateful for the plethora of contributors to international scientific institutions such as CERN for enabling researchers to work together in our quest to understand the universe we live in.